Abstract
This paper studies the effect of customer abandonment in the economic optimization of a service facility. Specifically, we consider how to jointly set the price and service capacity in order to maximize the steady-state expected profit, when the system is subject to customer abandonment and the price is paid only by customers that eventually receive service. We do this under the assumption that there is a high rate of prospective customer arrivals. Our analysis reveals that the economically optimal operating regime is consistent with the standard heavy traffic regime. Furthermore, we derive the following economical insight: when the capacity cost is sufficiently high, it can be advantageous for the system manager to “underinvest” in capacity and “take advantage of the abandonments” to trim congestion. Lastly, we show the loss in profit when customer abandonments are ignored is on the order of the square-root of the demand rate.

Keywords: Customer Abandonment; Pricing and Capacity Sizing; Profit Maximization, Heavy Traffic

1 Introduction
There is a long history of using price and/or capacity size planning to control congestion in a service facility. Many academic papers (such as Mendelson (1985), Dewan and Mendelson (1990), Stidham (1992), and references in Stidham (2009)) study the service facility under the assumption that all customers will wait for their service as long as necessary, which is
The objective of this paper is to understand the effect of customer abandonment in the economic optimization of service facilities that can be modeled as queues.

Specifically, this paper studies how to jointly set the price and capacity size in order to maximize the steady-state expected profit in a $M/GI/1+GI$ queue (“$+GI$” refers to i.i.d. abandonment (patience) times with a general distribution) with FCFS service when (i) the demand for service is $d(p)$ when the service is priced at $p > 0$, (ii) the cost of capacity that results in average service or processing rate $\mu$ is $c\mu$ for some $c > 0$, and (iii) there is a linear holding cost $h_0 \geq 0$. (The holding cost may be zero because the loss of profit due to customer abandonment still results in a sensible problem formulation.) Then, the objective is to

$$\max_{p,\mu} \mathcal{P}(p, \mu), \quad \text{where} \quad \mathcal{P}(p, \mu) \equiv pd(p)(1 - P_a) - c\mu - h_0 \mathbb{E}[Q(\infty)],$$

(1)

where $P_a$ represents the steady-state fraction of customers that abandon, and $Q(\infty)$ is the random variable that represents the steady-state queue-length.

The first question that arises is: How much does the presence of customer abandonments affect the maximum profit in (1)? If the answer is “not much”, then there is no reason to explicitly model the customer abandonments, and we can continue to use the solution in the aforementioned papers that model the service facility without abandonment. To gain intuition, we first restrict our analysis to the purely exponential setting. Solving (1) for a $M/M/1$ model (that assumes customers will wait as long as necessary for service) is straightforward, because then $P_a = 0$ and there is a convenient closed form expression for $\mathbb{E}[Q(\infty)]$. The analysis for a $M/M/1 + M$ model is more complicated; however, the birth-death structure implies that the steady-state probabilities can be written explicitly, and then can be used to provide expressions for $P_a$ and $\mathbb{E}[Q(\infty)]$ (although those expressions are not simple). Then (1) can be solved numerically.

Figure 1 below displays the solution to (1) in the purely exponential setting, as the abandonment rate varies between 0 and 1, which is equivalent to the steady-state customer abandonment probability varying between 0 and 0.30, for a linear demand function $d(p) = 10 \times \max(1 - p/5, 0)$ for $p > 0$, and $c = h_0 = $1. Figure 1(a) establishes that customer abandonments can result in a significant loss in profit, while Figure 1(b) suggests that the optimal price and capacity size for the $M/M/1 + M$ and $M/M/1$ systems are not too far apart.

Then, the second question arises: Can we use the solution from a system with no abandonment to approximate the solution to a system with abandonment? Despite the seeming promise of Figure 1(b), there are several reasons to worry about such an approach, which motivates developing theory. First, such an approach ignores the relationship between system utilization and the probability a customer abandons, which suggests that the first term in (1) may not be approximated well. Second, the heavy traffic approximation for the steady-state queue-length in a $GI/GI/1$ system has an exponential distribution (cf. Whitt (2002)) whereas that for a $GI/GI/1 + GI$ has a truncated normal distribution (cf. Ward
(a) The maximum profit in a $M/M/1 + M$ queue as compared to that in a $M/M/1$ queue.

(b) The solution to (1) in the purely exponential setting.

In this paper, we directly study (1), which provides the following structural insight: There is one scaling (fluid) under which the solution to (1) is consistent with the solution to the system with no abandonment and one finer scaling (diffusion) under which the two solutions differ.

Our Approach. The problem (1) is analytically intractable when the inter-arrival, service, and abandonment distributions are not all exponential. However, the problem becomes tractable under a large market size assumption. That assumption is most easily interpreted in the context of a linear demand function that has maximum market size $n > 0$ and price responsiveness coefficient $b > 0$; i.e.,

$$d(p) = n \times \max(1 - bp, 0) \text{ for } p \geq 0.$$  \hspace{1cm} (3)

The large market size assumption implies that $n$ is large compared to $b$. In general, we assume demand becomes large in a way that preserves the structure of the demand function. That is, for a given price $p \geq 0$, we assume

$$d(p) = n \times \lambda(p), \text{ with } \lambda(0) = 1,$$  \hspace{1cm} (4)

and we let $n$ grow large while all other parameters remain fixed. The maximum market size $n$ is exactly the demand when the service is offered for free since $d(0) = n$. The linear demand function (3) clearly satisfies (4).
The Proposed Policy. Our proposed policy for setting the price and capacity size is based on the solution to two optimization problems. The first is a static planning problem that ignores stochasticity (that is, a deterministic optimization problem); see Section 2. The second is an optimization problem that uses the steady-state of an underlying diffusion approximation to provide a “stochasticity correction factor”; see Sections 3-5.

- **Example:** Consider a service facility having demand function (3) with \( b = 1 \) and \( n = 5000 \), deterministic service times, and an abandonment distribution that is Uniform(0, \( d \)) for \( d = 1/365, 2/365, 7/365 \). Such a stochastic flow system, modeled as \( M/D/1 + \text{Uniform}(0, d) \) queue, is reasonable when the service facility is a make-to-order production firm having yearly demand \( d(p) \) (thousands of orders per year), measures delivery times in days, and puts together orders from component parts in a standardized fashion so that production times have very little variance. We assume the customer orders are subject to deadlines as \( d = 1, 2, 7 \) days. We focus on maximizing the revenue from selling products, and assume \( h_0 = 0 \) for simplicity. Then, the solution from a system with no abandonment is found by solving (2), which has solution \((p^* , n\mu^*)\) for \( p^* = (1 + c)/2 \) and \( \mu^* = \lambda(p^*) = (1 - c)/2 > 0 \). We assume \( c = 0.1 \). As a concrete example, this could correspond to a customized bicycle shop in which the cost to produce one bike is $0.1K, or $100, the associated selling price \( p^* = $0.55K \) or $550, and the resulting yearly profit is approaching one million (in the $700-$950K range). Figure 2 shows that the proposed policy outperforms the solution that ignores abandonments. The simulation results\(^1\) in Figure 2 show that the proposed policy can increase yearly profit by several percent (on the order of tens of thousands of dollars). We will revisit this example in Section 4 to highlight the key feature of the proposed policy, and again in Section 6 to connect with the asymptotic performance results developed in Sections 3-5.

\[^1\]Note: Without appealing to asymptotic analysis, one might use the purely Markovian model \( M/M/1 + M \) in the face of the general non-Markovian model, e.g., \( M/D/1 + \text{Uniform}(0, d) \) as considered above. Numeric calculation demonstrates that the Markovian counterpart (matching the mean parameters) can yield a large misspecification error of the maximum profit (measured in percentage error) as 23.7%, 27.7%, 60.9%, under \( d = 1/365, 2/365, 7/365 \), respectively. Moreover, a direct numerical approach to the problem (1) cannot provide any useful insights on the optimal policy’s structure.

Our Contributions.

1. We establish that the solution to (1) is consistent with the standard heavy traffic regime. More precisely, for a given demand function having maximum market size \( n \), if \((p^n, \mu^n)\)
solves (1), then for some $\theta \in \mathbb{R}$,

$$\frac{d(p^n) - \mu^n}{\sqrt{n}} = \frac{n\lambda(p^n) - \mu^n}{\sqrt{n}} \to \theta \text{ as } n \to \infty.$$  \hfill (5)

See Theorem 1.

As a consequence, the standard heavy traffic regime is an economically optimal operating regime.

2. We develop two approximating diffusion control problems (DCPs) that follow from (5). One is simpler to solve and implement but is more fragile to abandonment distribution characteristics (for example, is useless if the abandonment distribution density function is 0 or $\infty$ at 0), whereas the other is not. Both give rise to an asymptotically optimal control policy. See Theorems 2 and 3.

This refined analysis reveals that ignoring abandonments (as was done in Figure 2) leads to a loss in profit on the order of the square root of the maximum market size $n$ (a small percentage in the context of Figure 2 since $1/\sqrt{n} = 1/\sqrt{5000} = 1.4\%$).

3. We find that when the cost of capacity is high enough, an asymptotically optimal policy sets $p$ and $\mu$ such that demand exceeds capacity; see Figure 3 in Section 4. This statement can only be made for a model with abandonments; without abandonments, such a policy would give rise to a system without a stationary distribution, rendering the optimization (1) meaningless.

Figure 2: Simulation results of the profit comparison in a $M/D/1+$Uniform$(0,d)$ having $c = 0.1$, yearly demand function given in (3) with $n = 5000$, $b = 1$, and $d = 1/365, 2/365, 7/365$. 

This refined analysis reveals that ignoring abandonments (as was done in Figure 2) leads to a loss in profit on the order of the square root of the maximum market size $n$ (a small percentage in the context of Figure 2 since $1/\sqrt{n} = 1/\sqrt{5000} = 1.4\%$).
When the cost of capacity is high enough, the system manager will want to take advantage of abandonments to trim system congestion.

The remainder of this paper is organized as follows. We first review the most related literature. Then, in Section 2 we set up and solve the deterministic relaxation to (1), also known as the static planning problem (SPP). Section 3 establishes that an optimal policy induces the system to operate in the standard heavy traffic regime. That motivates us in Section 4 to formulate a diffusion control problem (DCP) based on the limiting behavior of the diffusion-scaled profit. Our proposed policy follows from the solution to that DCP. The aforementioned approximating DCP only involves the behavior of the abandonment distribution at 0, and so in Section 5 we provide an alternative approximating DCP that involves the full abandonment distribution and is valid for a broader family of abandonment distributions. Section 6 examines the impact of ignoring customer abandonment on the percentage loss in profit and provides some supporting numerics. The proofs of all results in the main body are found in the Appendix.

Literature Review

The literature related to our work is concerned with determining the economically optimal operating regime for a given service system. The majority of this work establishes that the standard heavy traffic (or, Halfin-Whitt regime for the many server setting) is induced by the optimal solution (rather than the heavy traffic condition being given) from, e.g., selecting optimal price and capacity size. We refer the reader to Armony and Maglaras (2004a,b), Harrison (2003), Maglaras and Zeevi (2003), Plambeck and Ward (2006), and Nair et al. (2016). All of these studies have assumed infinitely patient customers (i.e., no abandonment from the system).

Plambeck (2004) and Plambeck and Ward (2008) show that in manufacturing systems with a high volume of prospective customers, the optimal price and capacity together with optimal leadtime quotation result in heavy traffic. One main conclusion from the paper Armony et al. (2009), in a purely exponential setting, is that it is very important to understand customer impatience when setting capacity. In relation to that work, we establish that such an insight continues to hold in more generality.

The capacity setting papers most closely related to ours are Bassamboo and Randhawa (2010) and Randhawa (2013). Bassamboo and Randhawa (2010) assume the demand rate (no pricing control) and solves for the optimum capacity in both $M/M/N + G$ and $M/M/1 + G$ queueing systems. In particular, in the case of the $M/M/1 + G$ system, the objective of the authors is to solve a cost minimization problem that is equivalent to (1), and has capacity control but not pricing control (so only one decision variable). The authors characterize conditions under which optimal capacity sizing induces the overloaded regime, in which demand exceeds capacity, and the critically loaded regime, in which demand equals capacity. The authors show that fluid approximations are extremely accurate in
the overloaded regime. That insight is also true in Randhawa (2013), which considers a capacity sizing problem when customers balk but do not renege (that is, arriving customers may decide not to join the system, but any customer that joins waits as long as is necessary to obtain service). Very interestingly, and consistent with Remark 3 in Randhawa (2013), we find that when both price and capacity are included as decision variables, the critically loaded regime is the economically optimal operating regime. We then show the stronger result that the economically optimal operating regime is the standard heavy traffic regime. (In the critically loaded regime, the system load approaches one, but the rate is unspecified. The standard heavy traffic regime specifies the rate.)

Our asymptotic analysis relies heavily on past work that has developed heavy-traffic approximations for the queueing systems with abandonment. The heavy traffic limit of the queue-length process and the abandonment probability follow from the heavy traffic limit for the offered waiting time process. The offered waiting time process, introduced in Baccelli et al. (1984), tracks the amount of time an infinitely patient customer must wait for service. Its heavy traffic limit when the abandonment distribution is left unscaled is a reflected Ornstein-Uhlenbeck process (see Ward and Glynn (2005)), and its heavy traffic limit when the abandonment distribution is scaled through its hazard rate is a reflected nonlinear diffusion (see Reed and Ward (2008)).

2 The Static Planning Problem

We introduce the static planning problem (SPP) that ignores stochasticity. We then show that that leads to also ignoring abandonments, as in (2). To do this, we adapt the reasoning in Section 3 in Whitt (2006) from a many server setting to a single server setting. We assume unit maximum market size \((n = 1)\) for explanatory purposes, and then generalize.

Suppose the abandonment distribution is \(F\). Note that when \(\lambda(p)\) is the demand rate and \(\mu\) is the capacity size (service rate), then \(\lambda(p) − \mu\) is the rate at which incoming customers abandon. In this deterministic model, if all customers wait \(w\) time units, then \(\lambda(p)F(w)\) also equals the rate at which incoming customers abandon. Equating these two rates shows that all customers wait for \(w = w(p, \mu)\) time units, defined as the solution to

\[
\lambda(p)F(w) = [\lambda(p) − \mu]^+,
\]

whose solution is unique when \(F\) is continuous and strictly increasing. Next, since \((1 − F(x))\) is the probability a customer that has waited \(x\) time units is still present in the system, it follows that \(\int_0^w \lambda(p)(1 − F(x))dx\) approximates the number of customers waiting. This reasoning results in the following SPP

\[
\pi \equiv \max_{p, \mu} \left\{ p\lambda(p) \left( 1 - \frac{[\lambda(p) - \mu]^+}{\lambda(p)} \right) - c\mu - h_0 \int_0^{w(p, \mu)} \lambda(p)(1 - F(x))dx \right\},
\]

(6)
For $n \geq 1$, the same logic leads to
\[
\max_{p, \mu} \left\{ p n \lambda(p) \left( 1 - \frac{\left[ n \lambda(p) - \mu \right]^+}{n \lambda(p)} \right) - c \mu - h_0 \int_0^{\pi(p, \mu)} n \lambda(p)(1 - F(x))dx \right\},
\]
which has the maximized objective function value $n \overline{\pi}$, and a unique optimal solution $(p^*, n \mu^*)$ if and only if (6) has unique optimal solution $(p^*, \mu^*)$, which is guaranteed by the following sufficient condition.

(A1) The abandonment distribution function $F$ is continuous and strictly increasing. The function $\lambda(p)$ has $\lambda(0) > 0$, and is strictly decreasing, convex, and continuous. Furthermore, $p \lambda(p)$ is strictly concave.

**Lemma 1.** Assume (A1). The unique solution to the SPP (6) is exactly the solution to the following SPP
\[
\max_{p, \mu} \{ p \lambda(p) - c \mu \} \quad \text{subject to } \lambda(p) \leq \mu.
\]
Furthermore, that solution $(p^*, \mu^*)$ has $\lambda(p^*) = \mu^*$, $c < p^*$ and $0 < \mu^*$.

We denote
\[
\pi \equiv p^* \lambda(p^*) - c \mu^*.
\]
The implication of Lemma 1 is that the consequence of ignoring stochasticity is ignoring abandonments. In particular, the solution to the deterministic optimization problem (2) in the Introduction is exactly the solution to (7), scaled by $n$. That is, $(p^*, n \mu^*)$ solves (2), and the resulting maximum profit is $n \overline{\pi}$.

## 3 The Behavior of an Optimal Policy

We now investigate how an optimal policy behaves as the market size $n$ grows. Henceforth, we use $\lambda^n(\cdot)$ to denote the demand function of the system with a maximum market size $n$, namely, let
\[
\lambda^n(p) = n \lambda(p).
\]
When $n$ is regarded as fixed, this is exactly the demand function $d(p)$ specified in (4) in the Introduction. We refer to any process or quantity associated with the $M/GI/1 + GI$ model having demand function $\lambda^n(p)$ by inserting a superscript $n$ into the appropriate symbol.

The model is built from three independent sequences of non-negative i.i.d. random variables $(u_i, i \geq 1), (v_i, i \geq 1), (d_i, i \geq 1)$. We assume $u_1$ is an exponential random
variable with unit mean, $v_1$ has a unit mean and finite variance, and let the inter-arrival and service time sequences be specified as

$$u_i^n \equiv \frac{u_i}{\lambda^n(p)} \quad \text{and} \quad v_i^n \equiv \frac{v_i}{\mu^n}. \quad (9)$$

If the $i$-th customer, who arrives at time $t^n_i \equiv \sum_{j=1}^{i} u^n_j$ and has service time $v^n_i$, must wait longer than $d_i$ time units to reach the server, then s/he departs at time $t^n_i + d_i$ without receiving service. (There can be customers initially present; however, since we are interested in steady-state behavior, we do not specify those.) We assume a mild technical assumption $P(v_1 < u_1) > 0$, which ensures the steady-state abandonment probability and steady-state offered waiting times are well-defined; see Lemma 2 in Baccelli et al. (1984). Also, Theorem 3.2 in Kang and Ramanan (2012) ensures the existence of steady-state queue-length distribution (under a very mild assumption such that the inter-arrival distribution has a density, which is satisfied by an exponential distribution).

A policy $w = \{ (p^n, \mu^n) : n \geq 1 \}$ is a sequence of prices and capacity sizes. We let $\mathcal{P}_w^n$ denote the steady-state expected profit under an admissible policy $w$ with maximum market size $n$ (equivalently, demand function $\lambda^n$); that is, $\mathcal{P}_w^n = \mathcal{P}(p^n, \mu^n)$.

The following result establishes the connection between the problem we would like to solve (1) and the deterministic optimization problem (7) that results from ignoring stochasticity (and, therefore, also ignoring abandonments). Here and throughout $(p^n_{\text{opt}}, \mu^n_{\text{opt}})$ solves (1) when the demand function is $\lambda^n$ and

$$\text{opt} = \{ (p^n_{\text{opt}}, \mu^n_{\text{opt}}) : n \geq 1 \}$$

denotes an optimal policy (which may or may not be unique).

**Proposition 1.** (Optimality of Critical Load) Assume ($\&1$). Under an optimal policy $\text{opt}$,

$$\frac{\mathcal{P}^n_{\text{opt}}}{n} \to \pi, \quad p^n_{\text{opt}} \to p^* \quad \text{and} \quad \frac{\mu^n_{\text{opt}}}{n} \to \mu^* \quad \text{as} \quad n \to \infty.$$ 

Proposition 1 implies that the critically loaded regime is the economically optimal operating regime, because the system load $\lambda^n(p^n_{\text{opt}})/\mu^n_{\text{opt}}$ approaches 1 as $n$ tends to infinity. To see that, note that $\lambda(p^*) = \mu^*$ from Lemma 1.

**Remark 1.** Because $c < p^*$ by Lemma 1, the condition in Bassamboo and Randhawa (2010) that guarantees the critically loaded regime is the economically optimal operating regime is satisfied.

Proposition 1 provides a necessary condition for us to use to develop a proposed policy $v$. Specifically, $v$ should be such that

$$\frac{\mathcal{P}^n_v}{n} \to \pi \quad \text{as} \quad n \to \infty. \quad (10)$$
However, the condition (10) does not differentiate between a system that explicitly models abandonments and a system that ignores them. This motivates us to use a stricter criterion, that takes into account convergence rate. To do that, for any policy $w$, we define the centered and scaled profit
\[ \tilde{P}_n^w \equiv \sqrt{n} \left( \frac{P_n^w}{n} - \pi \right). \] (11)

**Definition 1.** We say that an admissible policy $w$ is asymptotically optimal if $\lim \inf_{n \to \infty} \tilde{P}_n^w \geq \lim \sup_{n \to \infty} \tilde{P}_n^{w'}$, for any other admissible policy $w'$, and also $\lim \inf_{n \to \infty} \tilde{P}_n^w > -\infty$.

Taking the other policy $w'$ to be $\text{opt}$, we see that an asymptotically optimal policy $w$ must have scaled profit $P_n^w/n$ that converges to $\pi$ as fast as an optimal policy, assuming the limits exist.

The question is: How do we translate the asymptotic optimality definition into a stricter necessary condition to use to develop a proposed policy? For this, we first define the capacity imbalance under any policy $w$,
\[ \theta_n^w = \frac{\lambda^n(p^n) - \mu^n}{n}. \]
The capacity imbalance of an optimal policy approaches 0 as $n \to \infty$ by Proposition 1. The following result gives the convergence rate.

**Theorem 1.** (Optimality of Heavy Traffic Regime) Assume (A1). Under any asymptotically optimal policy $w$
\[ \limsup_{n \to \infty} \sqrt{n} |\theta_n^w| < \infty. \]

Since an optimal policy is asymptotically optimal, Theorem 1 implies that the standard heavy traffic regime is an economically optimal operating regime; i.e., (5) holds under $\text{opt}$ (at least on a subsequence). This provides a stricter necessary condition to use to develop a proposed policy, because (5) implies $\lambda^n(p^n)/\mu^n \to 1$ as $n \to \infty$ but not vice versa. Specifically, Theorem 1 motivates proposing a policy $v$ that satisfies the heavy traffic condition (5), repeated here for the reader’s convenience.
\[ \sqrt{n} \theta_n^v = \frac{n \lambda(p^n) - \mu^n}{\sqrt{n}} \to \theta \in \mathbb{R} \text{ as } n \to \infty. \]

4 The Proposed Policy

The proposed policy when the maximum market size is $n$ has the form
\[ p^n = p^* \text{ and } \mu_n = n \mu^* - \sqrt{n} \theta + o(\sqrt{n}) \text{ for some } \theta \in (-\infty, \infty). \] (12)
Under such a policy, the heavy traffic condition (5) holds, because
\[ \sqrt{n} \theta_n = \frac{n \lambda(p^*) - n \mu^* + \sqrt{n} \theta + o(\sqrt{n})}{\sqrt{n}} = \theta + o(1), \]
where the first equality follows by definition and the second follows because \( \lambda(p^*) = \mu^* \).
Recalling that \( p^* \) and \( \mu^* \) are the solution to the static planning problem (7), that was derived by ignoring stochasticity, we interpret the parameter \( \theta \) as a “stochasticity correction” factor.

Our purpose in this section is to understand the performance of a policy \( \nu \) defined by (12) as a function of \( \theta \), in order to find a \( \theta^* \) that results in an asymptotically optimal policy. The first step is to study the behavior of the centered and scaled profit under policy \( v = \nu(\theta) \) as \( n \) becomes large. Substitution, algebra, and using \( \pi = p^* \lambda(p^*) - c \mu^* \), show that

\[
P_{\nu(\theta)}^n = \sqrt{n} \left( \frac{p^* \lambda(p^*)(1 - P^a_\theta) - c \mu^n - h_0 \mathbb{E}[Q^n(\infty)]}{n} - \frac{\pi}{n} \right)
= \sqrt{n} \left( p^* \lambda(p^*)(1 - P^a_\theta) - c \left( \mu^* - \frac{\theta}{\sqrt{n}} + o(\sqrt{n}) \right) - h_0 \mathbb{E}\left[\frac{Q^n(\infty; \theta)}{\sqrt{n}}\right] \right)
= c \theta + o(1) - p^* \lambda(p^*) \sqrt{n} P^a_\theta - h_0 \mathbb{E}\left[\frac{Q^n(\infty; \theta)}{\sqrt{n}}\right]. \tag{13}
\]
We would like to take the limit as \( n \to \infty \) in the above expression (13), in order to arrive at an optimization problem to determine \( \theta \). This requires understanding the asymptotic behavior of \( \sqrt{n} P^a_\theta \) and \( \mathbb{E}[Q^n(\infty)/\sqrt{n}] \), which relies on understanding the offered waiting time process.

The offered waiting time process \( \{V^n(t), t \geq 0\} \) in the system having maximum market size \( n \) tracks the amount of time a hypothetical customer arriving at time \( t \geq 0 \) would have to wait for service, which depends only on the service times of the customers waiting at time \( t \) who eventually receive service (do not abandon). Equivalently, \( V^n(t) \) represents the workload; that is, the time required for the server to process all the jobs in the queue at time \( t \) if no more jobs arrive. We use \( t^n_i \) to represent the workload at the instant the \( i \)-th customer arrives, not including the \( i \)-th customer’s service requirement \( v^n_i \). Under a given policy,
\[ V^n(t) = V^n(0) + \sum_{i=1}^{A^n(t)} v^n_i \mathbf{1}\{V^n(t^n_i) < d_i\} - B^n(t) \geq 0, \]
where \( A^n(t) \equiv \max\{i \geq 0 : \sum_{j=1}^i u^n_j \leq \lambda^n(p^n)t\} \) counts the number of customers that have arrived to the system by time \( t \geq 0 \), \( B^n(t) \equiv \int_0^t \mathbf{1}\{V^n(s) > 0\} ds \) is the cumulative server busy time, and \( V^n(0) \) is the possibly random initial state. The \( i \)-th customer abandons with a probability \( P(V^n(t^n_i) \geq d_i | V^n(0)) \) that depends on the initial conditions. The Markov
chain \( \{V^n(t^n_i), i \geq 1\} \) having state space \([0, \infty)\) has a unique steady-state distribution by Lemma 2 in Baccelli et al. (1984), and we let \( V^n(\infty) \) be a random variable having that steady-state distribution. When the initial state is distributed according to the steady-state distribution, the probability the \( i \)-th customer abandons is a random variable having distribution

\[
P(V^n(t^n_i) \geq d_i) = F(V^n(\infty)) \quad \text{for all } i.
\]

Then, the expected steady-state probability a customer abandons is

\[
P^*_{\text{a}} = \mathbb{E}[F(V^n(\infty))].
\]

We impose the following assumption in this section.

\((A2)\) The abandonment distribution \( F \) is differentiable with bounded derivative, in particular, \( 0 < F'(0) < \infty \).

Theorem 1 in Ward and Glynn (2005), adapted to our setting, establishes that for \( V \) defined in (15) below, when \( V^n(0) = 0 \) and under policy \( v(\theta) \) (which implies Assumption 1 in that same paper holds because the heavy traffic condition (5) holds),

\[
\sqrt{n}V^n(\cdot) \Rightarrow V(\cdot), \quad \text{as} \quad n \to \infty,
\]

in the topology of weak convergence on \( D[0, \infty) \); see, for example, Billingsley (1999) for a discussion of this convergence concept. In the following display, \((V, L)\) is the unique solution to the reflected stochastic differential equation

\[
V(t) = \sigma W(t) + \frac{\theta}{\lambda} t - F'(0) \int_0^t V(s)ds + L(t) \geq 0
\]

subject to: \( L \) is non-decreasing, has \( L(0) = 0 \) and \( \int_0^\infty V(s)dL(s) = 0 \),

where \( \{W(t) : t \geq 0\} \) denotes a one-dimensional standard Brownian motion, \( \lambda \equiv \lambda(p^*) \) and

\[
\sigma^2 \equiv \lambda^{-1}(\text{var}(u_1) + \text{var}(v_1)).
\]

From Sections 18.3 and 18.4.1 in Browne and Whitt (1995), the steady-state distribution \( V(\infty) \) of \( V \) follows a truncated normal distribution on \([0, \infty)\) (see (19) for the mean of \( V(\infty) \)).

In contrast to the offered waiting time process, the queue-length process includes both customers that will eventually receive service and those that will abandon. Still, given a policy under which the probability a customer abandons is small when \( n \) is large, we conjecture that Little’s law almost holds; i.e., that \( \mathbb{E}[Q^n(\infty)] \approx n\lambda \mathbb{E}[V^n(\infty)] \) (equivalently, \( \mathbb{E}[Q^n(\infty)/\sqrt{n}] \approx \mu^* \mathbb{E}[\sqrt{n}V^n(\infty)] \)). That conjecture is consistent with the following asymptotic relationship between the scaled queue-length and offered waiting time process

\[
Q^n(\cdot)/\sqrt{n} - \lambda\sqrt{n}V^n(\cdot) \Rightarrow 0, \quad \text{as} \quad n \to \infty,
\]

12
in \( D(0, \infty) \), which follows from Theorem 3 in Ward and Glynn (2005), adapted to our setting. The process-level convergences in (14) and (17) suggest the weak convergence of the associated steady-state distributions.

**Lemma 2.** Assume (A1) and (A2). Under the policy in (12),

\[
\mathbb{E}[\sqrt{n}V^n(\infty)] \to \mathbb{E}[V(\infty)] \quad \text{and} \quad \mathbb{E}\left[\frac{Q^n(\infty)}{\sqrt{n}}\right] \to \mu^* \mathbb{E}[V(\infty)] \quad \text{as} \quad n \to \infty.
\]

The following result allows us to determine the limit of the centered and scaled profit.

**Proposition 2.** (Limiting Performance Measure Expressions) Assume (A1) and (A2). As \( n \to \infty \), under the policy in (12),

\[
\sqrt{n}P^n_a \to F'(0)\mathbb{E}[V(\infty)] \quad \text{and} \quad \mathbb{E}\left[\frac{Q^n(\infty)}{\sqrt{n}}\right] \to \mu^* \mathbb{E}[V(\infty)].
\]

The limit of \( \tilde{P}^n_{v(\theta)} \) in (13) as \( n \) becomes large is immediate from Proposition 2, and letting

\[
g(\theta) \equiv \mathbb{E}[V(\infty; \theta)] = \frac{\theta}{\lambda F'(0)} + \frac{\sigma}{\sqrt{2F'(0)}} h_z \left(-\frac{\theta}{\lambda \sigma} \sqrt{\frac{2}{F'(0)}}\right),
\]

it is seen that

\[
\lim_{n \to \infty} \tilde{P}^n_{v(\theta)} = c\theta - p^* \lambda (p^*) F'(0) g(\theta) - h_0 \mu^* g(\theta) = c\theta - (p^* F'(0) + h_0) \mu^* g(\theta),
\]

where the second equality follows since \( \lambda(p^*) = \mu^* \). This leads to a sensible trade-off: Increasing \( \theta \) in (20) reduces the capacity in the proposed policy (12), which has marginal savings \( c \), but also incurs the associated congestion cost \( (p^* F'(0) + h_0) \mu^* g(\theta) \) (because \( g(\theta) \) is increasing in \( \theta \); see Lemma 3 below). The convergence in (20) leads to the following optimization problem

\[
\max_{\theta \in \mathbb{R}}\{c\theta - (p^* F'(0) + h_0) \mu^* g(\theta)\}.
\]

**Lemma 3.** We have that \( g(\theta) \equiv \mathbb{E}[V(\infty; \theta)] \) is increasing in \( \theta \) and the optimization problem (21) has a unique maximizer \( \theta^* \in \mathbb{R} \).

Our proposed policy is \( v(\theta^*) \), which substitutes \( \theta^* \) for \( \theta \) in (12).

**Theorem 2.** (Asymptotic Optimality) Assume (A1) and (A2). Any policy

\[
v(\theta^*) \equiv \{(p^*, n\mu^* - \sqrt{n}\theta^* + o(\sqrt{n})), n \geq 1\}
\]

is asymptotically optimal and, furthermore,

\[
\tilde{P}^n_{v(\theta^*)} \to \mathbf{P}^*, \quad \text{as} \quad n \to \infty,
\]

where \( \mathbf{P}^* \equiv c\theta^* - (p^* F'(0) + h_0) \mu^* g(\theta^*) \) is the maximum objective value in (21).
The following corollary to Theorem 2 emphasizes that the asymptotically optimal policy \( v(\theta^*) \) has the same behavior as an optimal policy (because an optimal policy must be asymptotically optimal).

**Corollary 1.** Assume (A1) and (A2). Under an optimal policy \( \text{opt} \), the scaled capacity imbalance \( \sqrt{n} \theta_{\text{opt}}^n \rightarrow \theta^* \) and the diffusion scaled profit \( \tilde{P}_{\text{opt}}^n \rightarrow P^* \) as \( n \rightarrow \infty \).

Our proposed policy results in a demand rate \( \lambda^n(p^*) = n\lambda(p^*) \) that exceeds the capacity size \( n\mu^* - \sqrt{n} \theta^* \) whenever \( \theta^* > 0 \), which occurs when the cost of capacity \( c \) is large. By Corollary 1, this is also true for an optimal policy (i.e., in the solution to (2)). This is in contrast to any policy that could potentially be developed from a system that ignores abandonments, because such a system must have demand strictly less than the capacity size for stability (more specifically, the system must possess a steady-state distribution so that the objective (1) with \( P_a = 0 \) is well-defined). The intuition is that it can be advantageous for the system manager to “underinvest in capacity” and “take advantage of the abandonments” to trim congestion (letting the abandonments stabilize the system).

- **Example:** Figure 3 illustrates the aforementioned phenomenon in a \( M/D/1 + \text{Uniform}(0,d) \) system with \( h_0 = 0 \) and maximum market size \( n = 5000 \) (the same setting as in Figure 2 in Section 1). Figure 3 further shows that as the customers are more patient (as the “deadline” \( d \) increases) the optimal capacity size decreases at a slower rate in \( c \) and the optimal policy gets closer to that of the system without abandonment.

## 5 Incorporating the Entire Abandonment Distribution

The optimization problem (21) that determines the stochasticity correction factor \( \theta^* \) depends on the abandonment distribution only through the value \( F'(0) \). However, \( F'(0) \) is not a very robust statistic (in contrast to mean and variance, for example). Furthermore, (21) is well-defined only when \( 0 < F'(0) < \infty \), which is a restrictive assumption. For example, a gamma distribution with shape parameter \( \alpha > 1 \) has \( F'(0) = 0 \), and one with \( 0 < \alpha < 1 \) has \( F'(0) = \infty \). So, unless the shape parameter \( \alpha = 1 \) (in which case the distribution is exponential) the optimization problem (21) cannot be considered. These observations motivate developing an optimization problem that incorporates more of the structure of the abandonment distribution, and this is our purpose in this Section. The end result is an alternative proposed policy, that can apply even when \( F'(0) = 0 \) or \( F'(0) = \infty \). That alternative proposed policy has the same structure as in (12), but uses an optimization problem that is different than (21) to determine the stochasticity correction factor \( \theta^* \).

More specifically, we apply the hazard rate scaling in Reed and Ward (2008), and to do this, we must allow the abandonment distribution to change with \( n \). Specifically, in the queueing model described in the first paragraph of Section 3, replace the sequence \( \{d_i, i \geq 1\} \) with \( \{d^n_i, i \geq 1\} \), which changes with \( n \). Then, the maximum amount of
Figure 3: The capacity size in the proposed policy is $n\mu^* - \sqrt{n}\theta^*$, as compared to the optimal policy $n\mu^*$ when abandonments are ignored (that is, the solution to (2)), in the setting in Figure 2 in the Introduction. Both polices have demand rate $n\lambda(p^*)$. The proposed policy has capacity size strictly less than the demand rate when the cost of capacity $c$ is large, which cannot happen in a system that ignores abandonments.

The time the $i$-th arriving customer will wait for service is $d_i^n$ (instead of $d_i$). The cumulative distribution function of $d_i^n$ is $F^n$, and its associated hazard rate function is $h^n$.

(A3) Suppose
\[ F^n(x) = 1 - \exp\left(-\int_0^x h^n(u)du\right), \quad x \geq 0. \]
holds and $h^n(x) = h(\sqrt{n}x)$ for some hazard rate function $h(\cdot)$. Furthermore, assume $h(u) \leq K(1 + u^\ell)$ for some $\ell > 0$ and $K > 0$.

Theorem 5.1 in Reed and Ward (2008), adapted to our setting, establishes
\[ \sqrt{n}V^n(\cdot) \Rightarrow V(\cdot) \]as $n \to \infty$ in $D[0, \infty)$, where $(V,L)$ is the unique solution to the reflected stochastic differential equation
\[ V(t) = \sigma W(t) + \frac{\theta}{\lambda} t - \int_0^t \left[ \int_0^V h(u)du \right] ds + L(t) \geq 0 \]
subject to: $L$ is non-decreasing, has $L(0) = 0$ and $\int_0^\infty V(s)dL(s) = 0$.

(24)
In a slight abuse of notation, in this Section, $V(\infty)$ is a random variable that has the steady-state distribution of (24), which is given in Proposition 6.1(i) in Reed and Ward (2008).
Observe that (24) and (15) coincide when the abandonment distribution is exponential with rate $\gamma$, because then $F'(0) = \gamma$ and $h(u) = \gamma$ for all $u \geq 0$. In general, the connection between (24) and (15) can be seen by performing a Taylor series expansion of $h$ about 0 (which follows the intuition provided in Reed and Tezcan (2012) for the GI/M/N + GI model) as follows

$$h(u) = h(0) + \sum_{i=1}^{\infty} \frac{u^i}{i!} h^{(i)}(0).$$  \hspace{1cm} (25)

Then, since $h(0) = F'(0)$, it follows that the linear portion of the infinitesimal drift in (15) arises from the lowest order term of the above Taylor series expansion.

The optimization problem that replaces (21) to determine the stochasticity correction factor $\theta^*$ can be seen as follows. As in the previous subsection, for large $n$, we expect

$$\mathbf{E} \left[ \frac{Q^n(\infty)}{\sqrt{n}} \right] \approx \mu^* \mathbf{E} [V(\infty)].$$

In contrast to the previous subsection, the intuition for the approximation for the abandonment probability follows from (with explanation provided below)

$$\sqrt{n} P^n_a = \mathbf{E} \left[ \sqrt{n} F^n(\infty, V^n) \right]$$

$$= \mathbf{E} \left[ \sqrt{n} \left( 1 - \exp \left( - \int_0^{V^n(\infty)} h^n(u) du \right) \right) \right]$$

$$\approx \mathbf{E} \left[ \sqrt{n} \int_0^{V^n(\infty)} h^n(u) du \right]$$

$$= \mathbf{E} \left[ \int_0^{V^n(\infty)} h(u) du \right].$$

The first equality is by definition and the second by the assumption $(A3)$. The approximation on the third line is from a Taylor series expansion of the exponential. The last equality is by a change of variable. In the remainder of this section, the random variable $V(\infty)$ is now in reference to the diffusion in (24). Then, we have the large $n$ approximation for the abandonment probability and steady-state queue-length as shown below.

**Proposition 3.** (Limiting Performance Measure Expressions) Assume $(A1)$ and $(A3)$. Under the policy in (12),

$$\sqrt{n} P^n_a \to \mathbf{E} \left[ \int_0^{V(\infty)} h(v) dv \right] \text{ and } \mathbf{E} \left[ \frac{Q^n(\infty)}{\sqrt{n}} \right] \to \mu^* \mathbf{E}[V(\infty)] \text{ as } n \to \infty.$$  

**Lemma 4.** We have that $g(\theta) \equiv \mathbf{E}[V(\infty; \theta)]$ is increasing in $\theta$ and the optimization problem (26) below has a unique maximizer $\theta^* \in \mathbb{R}$. 

16
Theorem 3. (An Alternative Asymptotically Optimal Policy) Assume (A1) and (A3). Define $\theta^*_2$ as the unique solution to the optimization problem

$$\max_{\theta \in \mathbb{R}} \left\{ c\theta - p^* \lambda(p^*) \mathbb{E} \left[ \int_0^{V(\infty; \theta)} h(u)du \right] - h_0 \mu^* \mathbb{I}[V(\infty; \theta)] \right\} ,$$

(26)

where $\mathbb{E}[V(\infty; \theta)]$ is the steady-state mean of the diffusion given in (24), and we write $V(\infty) = V(\infty; \theta)$ to emphasize the dependence of the steady-state distribution on $\theta$. Then, Theorem 2 and Corollary 1 hold for the policy

$$v(\theta^*_2) \equiv \{(p^*, n\mu^* - \sqrt{n}\theta^*_2 + o(\sqrt{n})), n \geq 1\},$$

(27)

except with $P^*$ replaced by $P^*_2$, defined as the maximum objective value in (26) as follows

$$P^*_2 \equiv c\theta^*_2 - p^* \lambda(p^*) \mathbb{E} \left[ \int_0^{V(\infty; \theta^*_2)} h(u)du \right] - h_0 \mu^* \mathbb{E}[V(\infty; \theta^*_2)].$$

Remark 2. Combining Proposition 1, Theorems 2 and 3, we obtain the following second order approximation to the optimal profit in the $n$-th system:

$$\mathcal{P}(p^*_n, \mu^*_n) = n\pi + \sqrt{n}P^*_i + o(\sqrt{n}),$$

(28)

for $i = 1, 2$ corresponding to when there is no hazard rate scaling and when there is such a scaling, respectively.

Although using the optimization problem (21) to define the stochasticity correction factor $\theta^*$ is simpler than using the above optimization problem (26), its scope is limited to a narrower family of abandonment distribution. Furthermore, the two limiting regimes produce different approximations of the optimal profit in the $n$-th system. We expect the regime with hazard rate scaling to result in an improved approximation, because that regime captures the entire abandonment distribution. The following Proposition orders the approximations produced by the two limiting regimes.

Proposition 4. Let $P^*_1$ be as in Theorem 2 and $P^*_2$ as in Theorem 3. If $h(\cdot)$ is strictly increasing, then $P^*_1 > P^*_2$, and if $h(\cdot)$ is strictly decreasing, $P^*_1 < P^*_2$.

Under the scaling assumption (A3), an optimal policy satisfies

$$\mathcal{P}^*_\text{opt} \rightarrow P^*_2$$

as $n \rightarrow \infty$.

Supposing $h(\cdot)$ is strictly decreasing, Proposition 4 implies that when the entire abandonment distribution information is used, the approximated optimal profit in the $n$-th system is larger than that based on the information only near the origin.
6 The Consequences of Ignoring Abandonments

In this section, we study the impact of ignoring abandonments on the loss in profit and connect back to the concrete example in the Introduction (recall Figure 2) that shows the proposed policy outperforms the solution that ignores abandonments. Suppose the modeler chooses to ignore abandonments. Then, the policy

\[ p^n = p^* \text{ and } \mu^n = n\mu^* - \sqrt{n}\theta_0 \]  

for

\[ \theta_0 \equiv -\sigma\sqrt{h_0/2c} \]

is known to be asymptotically optimal for the GI/GI/1 system; see (4.5) and Theorem 6.1 in Lee and Ward (2014).

Next, define \( \theta^*_1 \) to be the unique solution to (21) and \( \theta^*_2 \) to be the unique solution to (26). Under either policy, the percentage loss in profit from an optimal policy is \( o(1/\sqrt{n}) \) because for \( i \in \{1, 2\} \)

\[
\sqrt{n} \frac{\mathcal{P}(p^*, n\mu^* - \sqrt{n}\theta^*_i) - \mathcal{P}(p^n_{\text{opt}}, \mu^n_{\text{opt}})}{\mathcal{P}(p^n_{\text{opt}}, \mu^n_{\text{opt}})} = \left( \hat{P}^n_{v(\theta^*_i)} - \hat{P}^n_{\text{opt}} \right) \times \frac{n}{\mathcal{P}^n_{\text{opt}}} \to (P^*_i - P^*_i) \times \frac{1}{\pi} = 0,
\]

as \( n \to \infty \) in the relevant asymptotic regime. The equality in the above display follows from algebra, and the convergence follows from Proposition 1, Theorems 2 and 3 and Corollary 1.

In contrast, under the policy (29), the percentage profit loss no longer is \( o(1/\sqrt{n}) \) because

\[
\sqrt{n} \frac{\mathcal{P}(p^*, n\mu^* - \sqrt{n}\theta_0) - \mathcal{P}(p^n_{\text{opt}}, \mu^n_{\text{opt}})}{\mathcal{P}(p^n_{\text{opt}}, \mu^n_{\text{opt}})} = \left( \hat{P}^n_{v(\theta^*_0)} - \hat{P}^n_{\text{opt}} \right) \times \frac{n}{\mathcal{P}^n_{\text{opt}}} \to (P^0 - P^*_i) \times \frac{1}{\pi} < 0,
\]

for \( i \in \{1, 2\} \) as \( n \to \infty \), where (when there is no hazard rate scaling and recalling \( g(\cdot) \) in (19))

\[ P^0 \equiv c\theta_0 - (p^*F'(0) + h_0) \mu^*g(\theta_0). \]

(The reason why \( P^0 < P^*_i \) is because \( \theta^*_i \) \((i = 1, 2)\) is the unique maximizer of (21) and (26), respectively.) Thus, to understand the impact of ignoring abandonments, we want to analyze the scaling limit of percentage loss in profit. We do this in the simpler limiting regime in which the hazard rate is not scaled, and study

\[ L(c, h_0, F'(0), \sigma) \equiv \frac{P^0 - P^*_1}{\pi}. \]

The more negative \( L(c, h_0, F'(0), \sigma) \), the more harmful is using a model that ignores abandonments.
Connection to the Example in the Introduction. Figure 2 in the Introduction shows the simulation results of the profit comparison in a $M/D/1 + \text{Uniform}(0,d)$ model under the proposed policy based on the optimization problem (21) and the policy ignoring abandonments. In the Example, the cost to produce one product is $c = 0.1K$ and the assumed linear demand (3) has maximum yearly market size $n = 5000$ and price responsiveness coefficient $b = 1$. With this, the SPP (7) yields the solution $p^* = 0.55K$ and $\lambda(p^*) = \mu^* = 0.45$, and this in turn provides $\sigma = 1.49$ according to the relation (16). Assuming the holding cost $h_0 = 0$ and noticing the abandonment rate parameter $F'(0) = 0$ under the Uniform$(0,d)$ distribution, one can easily solve the optimization problem (21) numerically, under different deadline requirements $d = 1/365, 2/365, 7/365$.

Notice that the percentage profit loss in (30) is defined with respect to the optimal profit $P(p_{\text{opt}}, \mu_{\text{opt}})$, however, its exact evaluation is a daunting task. Owing to Theorem 2 and Corollary 1, one can instead use the profit under asymptotically optimal policy (22). Under the Example setting, we find that the percentage profit losses are

$$\frac{P(p^*, n\mu^* - \sqrt{n}\theta_0)}{P(p^*, n\mu^* - \sqrt{n}\theta_1^*)} \approx -4.26\%, -2.75\%, -1.34\%,$$

for $d = 1/365, 2/365, 7/365$, respectively. We also note that the simulation results in Figure 2 (i.e., the graph “% Decrease”) are $-5.04\%, -3.28\%, -1.55\%$ for $d = 1/365, 2/365, 7/365$, respectively. Although there is some degree of discrepancy, the above theoretical predicted values reasonably match with the simulation results.

Lastly, we want to examine the impact of ignoring abandonments by studying the behavior of $L(c, h_0, F'(0), \sigma)$ as any one of the four input parameters changes. An exact analysis of $L(c, h_0, F'(0), \sigma)$, such as taking partial derivatives, was intractable to us because it involves cumbersome derivatives of the normal hazard rate function $h(x) = \phi(x)/(1-\Phi(x))$ in (19) and also its evaluation at the solution $\theta^*$. Alternatively, we could use the known double inequality $(\sqrt{x^2 + 2} + x)/2 < h(x) < (\sqrt{x^2 + 4} + x)/2$ (cf. Yang and Chu (2015)) and analytically check simple qualitative behaviors of $L(c, h_0, F'(0), \sigma)$ in its parameters. In particular, it can be checked that the loss in profit $L(c, h_0, F'(0), \sigma)$ becomes large as each of the input parameters becomes large, and thus the more harmful is using a model that ignores abandonments; see Figures 4 and 5.

Impact of the Variability of Abandonment Distribution. Lastly, we examine the impact of variability of abandonment distribution on the optimal profit. In doing so, we consider the same parameter setting as in Figure 2 in the Introduction, except the abandonment distribution now follows a gamma distribution $G(\alpha)$ having both the scale and shape parameters as given $\alpha > 0$ (hence mean 1). Then, the simpler optimization problem (21) is no longer applicable unless $\alpha = 1$, because $F'(0) = 0$ if $\alpha > 1$ and $F'(0) = \infty$ if $0 < \alpha < 1$. Hence, we need to scale the hazard rate function and consider alternative optimization problem in (26). Figure 6 illustrates that variability lessens profits under the proposed policy in (27).
(a) Impact of $c$ on $L(c,h_0,F'(0),\sigma)$

(b) Impact of $h_0$ on $L(c,h_0,F'(0),\sigma)$

Figure 4

(a) Impact of $F'(0)$ on $L(c,h_0,F'(0),\sigma)$

(b) Impact of $\sigma$ on $L(c,h_0,F'(0),\sigma)$

Figure 5
Figure 6: The abandonment times are distributed according to $G(\alpha)$, a mean 1 gamma distribution having both the scale and shape parameters equalling $\alpha > 0$ (hence variance $1/\alpha$). Unless $G(\alpha) = G(1)$ (i.e., Markovian abandonment), the simpler optimization problem (21) is inapplicable. As the variability of the customer abandonment distribution increases, the maximum profit under the proposed policy (27) decreases. Pretending the abandonment distribution is exponential, and using the solution of (21) (the middle graph under $G(1)$) leads to misspecification of the optimal profit.

Appendix

In the Appendix, we provide the arguments to establish the results in the main body. Section A provides the proofs for propositions and theorems, and Section B provides the proofs for lemmas.

### A Proofs of Propositions and Theorems

**Proof of Proposition 1.** Fix $\triangle > 0$ and consider the policy $w$ defined by

$$p^n = p^\star \text{ and } \mu^n = n\mu^\star + \sqrt{n}\triangle.$$  

(A1)

In comparison to an optimal policy, $\mathcal{P}_{\text{opt}}^n \geq \mathcal{P}_{w}^n$ for each $n$ and so

$$\liminf_{n \to \infty} \frac{\mathcal{P}_{\text{opt}}^n}{n} \geq \limsup_{n \to \infty} \frac{\mathcal{P}_{w}^n}{n}.$$
From substitution of (A1) into (1), the definition of $\lambda^n$ in (8), and algebra,

\[
\frac{P_{n}^{\text{w}}}{n} = \frac{1}{n} \left( p^* n \lambda(p^*) (1 - P_{a}^{n}) - c \left( n \mu^* + \sqrt{n} \Delta \right) \right) - h_0 \frac{1}{n} \mathbb{E} [Q^n(\infty)] = p^* \lambda(p^*) - c \left( \mu^* + \frac{\Delta}{\sqrt{n}} \right) - p^* \lambda(p^*) P_{a}^{n} - h_0 \frac{1}{n} \mathbb{E} [Q^n(\infty)].
\]

Since $\pi = p^* \lambda(p^*) - c \mu^*$, taking the limit in the above display shows that

\[
\liminf_{n \to \infty} \frac{P_{n}^{\text{opt}}}{n} \geq \pi, \quad \text{(A2)}
\]

if we can establish

\[
P_{a}^{n} \to 0 \quad \text{and} \quad \frac{1}{n} \mathbb{E} [Q^n(\infty)] \to 0 \quad \text{as} \quad n \to \infty, \quad \text{(A3)}
\]

which we defer to the last paragraph of this proof. We also from (1) have the upper bound

\[
\frac{P_{n}^{\text{opt}}}{n} \leq p^n \lambda(p^n) - c \frac{\mu^n}{n} - p^n \lambda(p^n) P_{a}^{n}.
\]

Since the rate that non-abandoning customers arrive cannot exceed the rate at which they are served, $\lambda^n(p^n)(1 - P_{a}^{n}) \leq \mu^n$, from which it follows that

\[
P_{a}^{n} \geq \frac{[\lambda(p^n) - \mu^n/n]^+}{\lambda(p^n)}.
\]

We conclude that

\[
\frac{P_{n}^{\text{opt}}}{n} \leq p^n \lambda(p^n) \left( 1 - \frac{[\lambda(p^n) - \mu^n/n]^+}{\lambda(p^n)} \right) - c \frac{\mu^n}{n},
\]

and so optimizing over $(p^n, \mu^n/n)$ in the right-hand side of the above expression shows $P_{n}^{\text{opt}}/n \leq \pi$. Then,

\[
\limsup_{n \to \infty} \frac{P_{n}^{\text{opt}}}{n} - \pi \leq 0, \quad \text{(A4)}
\]

The inequalities (A2) and (A4) imply that

\[
\lim_{n \to \infty} \frac{P_{n}^{\text{opt}}}{n} = \pi,
\]

and this is true if and only if

\[
p_{\text{opt}}^{n} \to p^* \quad \text{and} \quad \frac{\mu_{\text{opt}}^{n}}{n} \to \mu^*,
\]

because the solution to the SPP is unique by Lemma 1.
Finally, to complete the proof, we use the fact that the $GI/GI/1 + GI$ queue is upper bounded by a conventional $GI/GI/1$ queue without abandonment to prove (A3). Suppose $W_C^n(\infty)$ and $Q_C^n(\infty)$ represent the respective steady-state wait time and queue-length in a conventional $GI/GI/1$ queue having inter-arrival and service time sequences specified as in (9), under policy $w$ in (A1) (which are well-defined because the $GI/GI/1$ queue is positive recurrent for each $n$ when $\Delta > 0$). In other words, the demand rate is $n\lambda(p^*)$, the capacity size is $n\mu^* + \sqrt{n}\Delta$, the inter-arrival time variance is $\text{var}(u_1)/(n\lambda(p^*))^2$, and the service time variance is $\text{var}(v_1)/(n\mu^* + \sqrt{n}\Delta)^2$. From (12) in Marshall (1968) (see also Kingman (1962))

$$IE\left[W_C^n(\infty)\right] \leq \frac{n\lambda(p^*)}{2} \left(\frac{\text{var}(u_1)}{n\lambda(p^*)^2} + \frac{\text{var}(v_1)}{(n\mu^* + \sqrt{n}\Delta)^2}\right).$$

Algebra and the fact that $\lambda(p^*) = \mu^*$ imply that the right-hand side of the above expression equals

$$\frac{1}{\sqrt{n}} \times \frac{1}{2} \times \left(\frac{\text{var}(u_1)}{\lambda(p^*)} + \frac{\text{var}(v_1)}{\mu^* + \frac{2\Delta}{\sqrt{n}} + \frac{\Delta}{n\lambda(p^*)}}\right) \times \left(\frac{\mu^* + \Delta/\sqrt{n}}{\Delta}\right),$$

which converges to 0 as $n \to \infty$. Hence $IE\left[W_C^n(\infty)\right] \to 0$ as $n \to \infty$. Since $IE\left[V^n(\infty)\right] \leq IE\left[W_C^n(\infty)\right]$ for each $n$, we have $IE\left[V^n(\infty)\right] \to 0$ as $n \to \infty$, which implies $V^n(\infty) \Rightarrow 0$ as $n \to \infty$, because $V^n(\infty)$ is a non-negative random variable. The continuous mapping theorem then shows $F(V^n(\infty)) \Rightarrow 0$ as $n \to \infty$, and, since $F$ is bounded,

$$P_a^n \equiv IE\left[F(V^n(\infty))\right] \to 0 \text{ as } n \to \infty.$$ 

Little’s law implies

$$\frac{IE\left[Q_C^n(\infty)\right]}{n\lambda(p^*)} = IE\left[W_C^n(\infty)\right] \to 0 \text{ as } n \to \infty.$$ 

Since $IE\left[Q^n(\infty)\right] \leq IE\left[Q_C^n(\infty)\right]$ for each $n$, we conclude

$$\frac{1}{n} IE\left[Q^n(\infty)\right] \to 0 \text{ as } n \to \infty.$$ 

This completes the proof. \qed

**Proof of Theorem 1.** The proof of this result requires the use of the following problem that perturbs the constraint in (7)

$$\pi(\nu) \equiv \max_{p,\mu} \{p\lambda(p) - c\mu\} \quad \text{subject to } \lambda(p) \leq \mu + \nu.$$ (A5)

23
It follows from Lemma 5.2 in Lee and Ward (2014) that the unique solution to (A5) is, for \( \nu \leq \mu^* \),

\[
(p^*(\nu), \mu^*(\nu)) = (p^*, \mu^* - \nu),
\]

so that if \( \pi \) solves (7), then

\[
\pi(\nu) - \pi = c\nu.
\]

The next step is to obtain an upper bound on \( \tilde{P}_u^n \) under any admissible policy \( u = \{(p^n, \mu^n) : n \geq 1\} \). First note that from algebraic manipulation

\[
\tilde{P}_u^n = \frac{p^n n \lambda(p^n)(1 - P^n_a) - c \mu^n - h_0 \mathbb{E}[Q^n(\infty)] - \pi n}{\sqrt{n}}
\]

\[
= \sqrt{n} \left( p^n \lambda(p^n) - c \frac{\mu^n}{n} - \pi \right) - p^n \lambda(p^n) \sqrt{n} P^n_a - h_0 \mathbb{E} \left[ \frac{Q^n(\infty)}{\sqrt{n}} \right].
\]

Under any asymptotically optimal policy

\[
p^n \to p^* \quad \text{and} \quad \frac{\mu^n}{n} \to \mu^* \quad \text{as} \quad n \to \infty,
\]

by Proposition 1, so that \( \theta^n \to 0 \) as \( n \to \infty \), where we have dropped the subscript \( u \) on \( \theta^n \) for convenience in presentation. Then, from the definitions of \( \pi(\theta^n) \) and \( \theta^n \),

\[
p^n \lambda(p^n) - c \frac{\mu^n}{n} - \pi \leq \pi(\theta^n) - \pi = c \theta^n.
\]

Hence for large enough \( n \),

\[
\tilde{P}_u^n \leq c \sqrt{n} \theta^n - p^n \lambda(p^n) \sqrt{n} P^n_a - h_0 \mathbb{E} \left[ \frac{Q^n(\infty)}{\sqrt{n}} \right]. \tag{A6}
\]

Now, suppose

\[
\limsup_{n \to \infty} \sqrt{n} |\theta^n| = \infty,
\]

and consider any subsequence \( \{n_i\} \) on which \( \sqrt{n_i} |\theta^{n_i}| \to \infty \). In the case that \( \sqrt{n_i} \theta^{n_i} \to -\infty \), it follows immediately that \( \tilde{P}_u^{n_i} \to -\infty \). Hence consider the case that \( \sqrt{n_i} \theta^{n_i} \to \infty \).

Since the rate at which customers that eventually enter service must be less than or equal to the capacity size (dropping the subscript \( i \) from \( n_i \)), we get

\[
\lambda^n(p^n)(1 - P^n_a) \leq \mu^n,
\]

or, equivalently

\[
-P^n_a \leq \frac{\mu^n - \lambda^n(p^n)}{\lambda^n(p^n)} = -\frac{\theta^n}{\lambda(p^n)}. \tag{A7}
\]
Hence,
\[ c\sqrt{n}\theta^n - p^n \lambda(p^n)\sqrt{n}P^n_a - h_0 \mathbb{E} \left[ \frac{Q^n(\infty)}{\sqrt{n}} \right] \leq c\sqrt{n}\theta^n - p^n \lambda(p^n)\sqrt{n}P^n_a \]
\[ \leq (c - p^n)\sqrt{n}\theta^n, \quad (A8) \]
where the second inequality follows from (A7). Since from Proposition 1 under any asymptotically optimal policy \( p^n \to p^* \) as \( n \to \infty \), and \( p^* > c \), it follows that for all large enough \( n \), \( c - p^n < 0 \). Hence, \( (c - p^n)\sqrt{n}\theta^n \to -\infty \) and so \( \tilde{P}_u^n \to -\infty \) from the bounds in (A6) and (A8). We conclude that such a policy cannot be asymptotically optimal.

**Proof of Proposition 2.** From Lemma 2 and the fact that \( \lambda \equiv \lambda(p^*) = \mu^* \),
\[ \mathbb{E} \left[ \frac{Q^n(\infty)}{\sqrt{n}} \right] \to \mu^* \mathbb{E}[V(\infty)] \text{ as } n \to \infty. \]
Define \( \tilde{V}^n(\infty) \equiv \sqrt{n}V^n(\infty) \), so that
\[ \tilde{V}^n(\infty) \Rightarrow V(\infty) \text{ as } n \to \infty \quad (A9) \]
from the proof of Lemma 2, and note that
\[ \sqrt{n}P^n_a = \mathbb{E} \left[ \sqrt{n}F \left( \tilde{V}^n(\infty) \sqrt{n} \right) \right]. \]
Also note that if \( \{x_n\} \) is a sequence in \( [0, \infty) \) for which \( x_n \to x \) as \( n \to \infty \), then
\[ \left| \sqrt{n}F \left( \frac{x_n}{\sqrt{n}} \right) - F'(0)x \right| \to 0 \text{ as } n \to \infty \quad (A10) \]
because
\[ \left| \sqrt{n}F \left( \frac{x_n}{\sqrt{n}} \right) - F'(0)x_n \right| = \left| \frac{x_n}{\sqrt{n}} - F'(0)x_n \right| \]
\[ = \left| F'(\xi_n)x_n - F'(0)x_n \right|, \quad \text{for } \xi_n \in (0, x_n/\sqrt{n}), \]
where the first equality follows from \( F(0) = 0 \) and the second is by the mean value theorem (recalling that \( F \) is assumed differentiable). Then,
\[ |F'(\xi_n)x_n - F'(0)x| \leq |F'(\xi_n) - F'(0)||x_n + F'(0)||x_n - x| \to 0 \text{ as } n \to \infty \]
establishes (A10). The generalized continuous mapping theorem (Theorem 3.4.4 in Whitt (2002)), (A9), and (A10) imply
\[ \sqrt{n}F \left( \frac{\tilde{V}^n(\infty)}{\sqrt{n}} \right) \Rightarrow F'(0)V(\infty) \text{ as } n \to \infty. \quad (A11) \]
Next, since for any \( v \geq 0 \),
\[
\sqrt{n}F\left( \frac{v}{\sqrt{n}} \right) = \frac{F(v/\sqrt{n}) - F(0)}{v/\sqrt{n}} v = F'(\xi)v \text{ for } \xi \in (0, v/\sqrt{n})
\]
by the mean value theorem and the fact that \( F \) is differentiable, it follows that
\[
\sqrt{n}F\left( \frac{\tilde{V}^n(\infty)}{\sqrt{n}} \right) \leq \tilde{V}^n(\infty) \sup_{x \in [0, \infty)} F'(x).
\]
Since \( F' \) is bounded by assumption, the family \( \{ \sqrt{n}F(\tilde{V}^n(\infty)/\sqrt{n}) \} \) is dominated by the family \( \{ \tilde{V}^n(\infty) \sup_{x \in [0, \infty)} F'(x) \} \), which is uniformly integrable from the proof of Lemma 2. Hence \( \{ \sqrt{n}F(\tilde{V}^n(\infty)/\sqrt{n}) \} \) is uniformly integrable, which together with (A11), implies
\[
IE\left[ \sqrt{n}F\left( \frac{\tilde{V}^n(\infty)}{\sqrt{n}} \right) \right] \to F'(0)IE[\tilde{V}(\infty)] \text{ as } n \to \infty.
\]
This completes the proof. \( \square \)

**Proof of Theorem 2.** The convergence in (23) follows from Proposition 2, as shown in (20), and the representation for the diffusion-scaled profit in (13). Consider any admissible policy \( w = \{(p^n, \mu^n) : n \geq 1\} \). It follows from Theorem 1 that it is sufficient to evaluate the limiting behavior of this policy on any subsequence \( \{n_i\} \) having \( \sqrt{n_i} \theta^{n_i} \to \theta \in IR \) as \( n_i \to \infty \). Recall from (A6) that
\[
\tilde{P}^n_w \leq c\sqrt{n_i} \theta^{n_i} - p^{n_i} \lambda(p^{n_i}) \sqrt{n_i} P^{n_i} - h_0 \mathbb{E}\left[ \frac{Q^{n_i}(\infty)}{\sqrt{n_i}} \right].
\]
Since it follows from Proposition 2 that on any subsequence \( \{n_i\} \) having \( \sqrt{n_i} \theta^{n_i} \to \theta \in IR \), recalling that \( g(\theta) \equiv IE[\tilde{V}(\infty; \theta)] \)
\[
c\sqrt{n_i} \theta^{n_i} - p^{n_i} \lambda(p^{n_i}) \sqrt{n_i} P^{n_i} - h_0 \mathbb{E}\left[ \frac{Q^{n_i}(\infty)}{\sqrt{n_i}} \right] \to c\theta - p^* \lambda(p^*) F'(0) g(\theta) - h_0 \mu^* g(\theta),
\]
as \( n_i \to \infty \), and also since \( \theta^* \) is a unique maximizer of the right-hand side of the above display, as guaranteed from Lemma 3, we find
\[
\limsup_{n_i \to \infty} \tilde{P}^n_w \leq P^* = c\theta^* - p^* \lambda(p^*) F'(0) IE[\tilde{V}(\infty; \theta^*)] - h_0 \mu^* IE[\tilde{V}(\infty; \theta^*)]. \quad (A12)
\]
The proof is complete because (i) the policy \( v(\theta^*) \) attains the upper bound in (A12), (ii) any policy \( w \) under which \( |\sqrt{n_i} \theta^{n_i}| \to \infty \) as \( n_i \to \infty \) has \( \tilde{P}^n_w \to -\infty \) as \( n \to \infty \) by Theorem 1, and (iii) any policy under which \( \sqrt{n_i} \theta^{n_i} \to \theta \neq \theta^* \) as \( n_i \to \infty \) does not attain the upper bound in (A12). \( \square \)
Proof of Corollary 1. We show the scaled capacity imbalance $\sqrt{n} \theta_{opt}^n \to \theta^*$ as $n \to \infty$. From Theorem 1,

$$\limsup_{n \to \infty} \sqrt{n} |\theta_{opt}^n| < \infty.$$ 

Consider any convergent subsequence $\{n_i\}$ on which

$$\sqrt{n_i} \theta_{opt}^{n_i} \to \theta \in \mathbb{R} \quad \text{as} \quad n_i \to \infty,$$

so that the standard heavy traffic assumption is satisfied. On this subsequence, Proposition 2 shows, recalling that $g(\theta) \equiv \mathbb{E}[V(\infty; \theta)]$,

$$c \sqrt{n_i} \theta_{opt}^{n_i} - p^* \lambda(p^*) \sqrt{n_i} P_{a}^{n_i} + h_0 \mathbb{E} \left[ \frac{Q^{n_i}(\infty)}{\sqrt{n_i}} \right] \to c \theta - p^* \lambda(p^*) F'(0) g(\theta) - h_0 \mu^* g(\theta),$$

as $n_i \to \infty$. The definition of $\theta^*$ as the unique solution to the DCP (21) implies that

$$c \theta - p^* \lambda(p^*) F'(0) g(\theta) - h_0 \mu^* g(\theta) \leq c \theta^* - p^* \lambda(p^*) F'(0) g(\theta^*) - h_0 \mu^* g(\theta^*),$$

with equality if and only if $\theta = \theta^*$. Recall that $opt$ must be asymptotically optimal. Since the proposed policy $v(\theta^*)$ achieves the upper bound in the above expression by Theorem 2, the definition of asymptotic optimality is contradicted unless $\theta = \theta^*$. Hence, $\sqrt{n} \theta_{opt}^n \to \theta^*$ as $n \to \infty$. Next, we show $\lim_{n \to \infty} \tilde{P}^n_{opt} = P^*$. The proof has exactly the same structure as the proof of Corollary 6.2 in Lee and Ward (2014). The same arguments leading to (A12) show that,

$$\limsup_{n \to \infty} \tilde{P}^n_{opt} \leq P^*.$$  (A13)

Also, from (23) in Theorem 2,

$$\lim_{n \to \infty} \tilde{P}^n_{v(\theta^*)} = P^*.$$  (A14)

The policy $opt$ must be asymptotically optimal, and so

$$\liminf_{n \to \infty} \tilde{P}^n_{opt} \geq P^*.$$  (A14)

It follows from (A13) and (A14) that $\lim_{n \to \infty} \tilde{P}^n_{opt} = P^*$. This completes the proof. $\square$

Proof of Proposition 3. The proof structure closely follows those of Lemma 2 and Proposition 2. From Lemma EC.4 of Huang and Gurvich (2018), the $M/GI/1+GI$ systems under hazard-rate scaling (indexed by $n$) belong to a single queue family $Q(H)$ for all sufficiently large $n$. The tightness of the family $\{\sqrt{n} V^n(\infty) : n \geq 1\}$ readily follows from the first part of Lemma 1(i) of Huang and Gurvich (2018). Hence, the process-level convergence of Reed and Ward (2008), together with the aforementioned tightness, implies the steady-state convergence of offered waiting time sequence $\sqrt{n} V^n(\infty) \Rightarrow V(\infty)$ as $n \to \infty$. Then,
the convergence \( \mathbb{E} \left[ \frac{Q^n(\infty)}{\sqrt{n}} \right] \to \mu^* \mathbb{E}[V(\infty)] \) as \( n \to \infty \) follows exactly as in the proofs of Lemma 2 and Proposition 2.

Next, recall

\[
\sqrt{n} P^n_a = \mathbb{E} \left[ \sqrt{n} F^n(V^n(\infty)) \right] = \mathbb{E} \left[ \sqrt{n} \left( 1 - \exp \left( - \int_0^{V^n(\infty)} h^n(u) du \right) \right) \right] = \mathbb{E} \left[ g_n \left( \int_0^{\sqrt{n} V^n(\infty)} h(v) dv \right) \right]
\]

where \( g_n(x) \equiv \sqrt{n} \left( 1 - \exp \left( - \frac{x}{\sqrt{n}} \right) \right), x \geq 0 \). Notice that \( g_n(x) \to x \), as \( n \to \infty \), uniformly on compact sets, which in turn implies that \( g_n(x_n) \to x \) whenever \( x_n \to x \). From a generalized continuous mapping theorem (cf. Theorem 3.4.4 in Whitt (2002)), \( \sqrt{n} V^n(\infty) \Rightarrow V(\infty) \), and the uniform integrability of the family \( \{[\sqrt{n} V^n(\infty)]^k : n \geq 1 \} \) for any \( k \geq 1 \) (following from Lemma 1(i) of Huang and Gurvich (2018)), we have

\[
\sqrt{n} P^n_a = \mathbb{E} \left[ g_n \left( \int_0^{\sqrt{n} V_n(\infty)} h(v) dv \right) \right] \to \mathbb{E} \left[ \int_0^{V(\infty)} h(v) dv \right]
\]
as \( n \to \infty \). This completes the proof. \( \square \)

**Proof of Theorem 3.** The proof follows exactly the same arguments as the proof of Theorem 2. \( \square \)

**Proof of Proposition 4.** The proposition can be easily verified by noting the fact \( h(0) = F'(0) \) and comparing the objective functions in (21) and (26). \( \square \)

### B Lemma Proofs

**Proof of Lemma 1.** Lemma 2.1 in Lee and Ward (2014) shows there exists a unique solution to (7) with the desired properties. Then, from the definition of a unique maximizer, it suffices to check that, for \( (p, \mu) \neq (p^*, \mu^*) \) and \( p \geq 0, \mu \geq 0 \),

\[
p\lambda(p) \left( 1 - \frac{[\lambda(p) - \mu^+]^+}{\lambda(p)} \right) - c\mu - h_0 \lambda(p) \int_0^{\pi(p, \mu)} (1 - F(x)) dx < p^* \lambda(p^*) \left( 1 - \frac{[\lambda(p^*) - \mu^*]^+}{\lambda(p^*)} \right) - c\mu^* - h_0 \lambda(p^*) \int_0^{\pi(p^*, \mu^*)} (1 - F(x)) dx.
\]

From Lemma 2.1 in Lee and Ward (2014), one has \( \lambda(p^*) = \mu^* \) and therefore the right-hand side of the above inequality becomes \( p^* \lambda(p^*) - c\mu^* = \pi \). Since also \( h_0 \geq 0 \), it is enough to
show
\[ p\lambda(p) \left( 1 - \frac{[\lambda(p) - \mu]}{\lambda(p)} \right) - c\mu < \pi. \] (A15)

To show the inequality (A15) holds, first note that in the region \( \{(p, \mu) : \lambda(p) \leq \mu\} \), the left-hand side of (A15) equals to \( p\lambda(p) - c\mu \), which is less than \( \pi \) because the solution to the SPP (7) is unique. Next, observe that in the region \( \{(p, \mu) : \lambda(p) > \mu\} \), if \( p \leq c \), then
\[
p\lambda(p) \left( 1 - \frac{[\lambda(p) - \mu]}{\lambda(p)} \right) - c\mu = (p - c)\mu \leq 0.
\]
Since \( 0 < \pi \) by Lemma 2.1 in Lee and Ward (2014), it is enough to verify (A15) in the region
\[
\{(p, \mu) : \lambda(p) > \mu, p > c\}.
\]
To do this, it is helpful to define \( \nu \equiv \lambda(p) - \mu \),

and use the following equality
\[
p\lambda(p) \left( 1 - \frac{[\lambda(p) - \mu]}{\lambda(p)} \right) - c\mu = p\lambda(p) - c\mu - p\nu.
\]
Then, to complete the proof, we must show
\[
p\lambda(p) - c\mu - p\nu < \pi, \text{ when } \lambda(p) > \mu, p > c.
\] (A16)

We show (A16) by further dividing the parameter regime into the following four subcases:

(i) \( \nu \leq \mu^* \);

(ii) \( \nu > \mu^*, p > p^* \);

(iii) \( \nu > \mu^*, p < p^*, \mu \leq \mu^* \);

(iv) \( \nu > \mu^*, p < p^*, \mu > \mu^* \).

Also, it is helpful to recall that since the revenue function \( R(p) \equiv p\lambda(p) \) is concave by assumption (A1),
\[
R(p) \leq R(p^*) + R'(p^*)(p - p^*).
\] (A17)
Furthermore, recall that the KKT conditions in the proof of Lemma 2.1 in Lee and Ward (2014) guarantee
\[
R'(p^*) = \lambda(p^*) + p^*\lambda'(p^*) = c\lambda'(p^*).
\] (A18)
Case (i): Recall the optimization problem (A5) that perturbs the constraint in the SPP (7), which has solution $\pi(\nu)$ and satisfies $\pi(\nu) = \pi + \nu c$. Then, (A16) follows from
\[ p\lambda(p) - c\mu - p\nu < \pi(\nu) - p\nu = \pi - (p - c)\nu. \]

Case (ii): It follows from (A17) that
\[ p\lambda(p) - c\mu - p\nu \leq R(p^*) + R'(p^*)(p - p^*) - c\mu - p\mu^*. \]
Since $\nu > \mu^*$ and $p > c$, $p\nu > c\mu^*$. Then, also noting that $R(p^*) - c\mu^* = \pi$, we find
\[ R(p^*) + R'(p^*)(p - p^*) - c\mu - p\nu < \pi + R'(p^*)(p - p^*) - c\mu. \]
The term $R'(p^*)(p - p^*)$ is negative because $R$ is concave and $p > p^*$. Hence (A16) follows from the previous two displays.

Case (iii): Simple algebra shows that
\[ p\lambda(p) - c\mu - p\nu = (p - c)\mu < (p^* - c)\mu^* = \pi. \]
(The last equality follows because $\lambda(p^*) = \mu^*$.)

Case (iv): It follows from (A17) and $\nu > \mu^*$ that
\[ p\lambda(p) - c\mu - p\nu \leq R(p^*) + R'(p^*)(p - p^*) - c\mu - p\mu^*. \]
Since $R(p^*) = \pi + c\mu^*$,
\[ R(p^*) + R'(p^*)(p - p^*) - c\mu - p\mu^* = \pi + c(\mu^* - \mu) + R'(p^*)(p - p^*) - p\mu^*. \]
Since $\mu^* < \mu$, to show (A16), it is sufficient to show
\[ R'(p^*)(p - p^*) - p\mu^* < 0. \]
From (A18), the fact that $\lambda(p^*) = \mu^*$, and algebra,
\[ R'(p^*)(p - p^*) - p\mu^* = \left( \lambda(p^*) + p^*\lambda'(p^*) \right)(p - p^*) - p\mu^* = -p^*\left( \lambda(p^*) + p^*\lambda'(p^*) \right) + pp^*\lambda'(p^*) = -p^*c\lambda'(p^*) + pp^*\lambda'(p^*) = p^*\lambda'(p^*) - c < 0, \]
because $p > c$ and $\lambda'(p^*) < 0$ from assumption (A1).

30
Proof of Lemma 2. Under the assumed conditions, we can invoke Lemma EC.3 of Huang and Gurvich (2018), which concludes our $M/GI/1 + GI$ systems (indexed by $n$) belong to a single queue family $Q(H)$ for all sufficiently large $n$. The tightness, as well as the uniform integrability, of $\{\sqrt{n}V^n(\infty) : n \geq 1\}$ readily follows from the first part of Lemma 1(i) of Huang and Gurvich (2018). (Under critically loaded regime, as considered in our setting, it can be seen $\bar{\rho} = 1, \bar{w}_p = 0$ and the scaling parameter $c_p = 1/\sqrt{\lambda n} = 1/\sqrt{n\lambda}$; see Section EC.3.1 of Huang and Gurvich (2018).) Hence, the process-level convergence of Ward and Glynn (2005), together with the aforementioned tightness, implies the steady-state convergence of offered waiting time sequence $\sqrt{n}V^n(\infty) \Rightarrow V(\infty)$ as $n \to \infty$. Then, the uniform integrability implies $\mathbb{E}[\sqrt{n}V^n(\infty)] \to \mathbb{E}[V(\infty)]$ as $n \to \infty$.

Combining the preceding convergence with Corollary 1 of Huang and Gurvich (2018), we conclude the second result $\mathbb{E}\left[\frac{Q^n(\infty)}{\sqrt{n}}\right] \to \lambda\mathbb{E}[W(\infty)]$ as $n \to \infty$. Indeed, this is because the term $\lambda$ of Huang and Gurvich (2018) translates into $\lambda n = n\lambda$ in our setup and their Corollary 1 implies $\mathbb{E}[Q^n(\infty)] = \lambda^n\mathbb{E}[W^n(\infty)]$, where $W^n(\infty) \equiv v \wedge V^n(\infty)$ is the steady-state waiting time and $v$ is the patience threshold drawn from the patience distribution $F$. Therefore,

$$\mathbb{E}\left[\frac{Q^n(\infty)}{\sqrt{n}}\right] = \lambda\sqrt{n}\mathbb{E}[W^n] \to \lambda\mathbb{E}[W(\infty)]$$

as $n \to \infty$, from the fact that $\mathbb{E}[\sqrt{n}V^n(\infty)] \to \mathbb{E}[V(\infty)]$ and $\sqrt{n}v \to \infty$ as $n \to \infty$. Recalling that $\lambda = \lambda(p^*) = \mu^*$ completes the proof.

Proof of Lemma 3. We show $g(\theta) \equiv \mathbb{E}[V(\infty; \theta)]$ is strictly increasing in $\theta$. Recalling

$$\mathbb{E}[V(\infty; \theta)] = \frac{\theta}{\lambda F'(0)} + \frac{\sigma}{\sqrt{2F'(0)}} h_z\left(\frac{-\theta}{\lambda \sigma} \sqrt{\frac{2}{F'(0)}}\right),$$

a straightforward calculation shows that

$$g'(\theta) = \frac{1}{\lambda F'(0)} \left(1 - h_z'(x) \left(-\frac{\theta}{\lambda \sigma} \sqrt{\frac{2}{F'(0)}}\right)\right) > 0. \quad (A19)$$

This follows from a property $h_z'(x) < 1$ for any $x \in \mathbb{R}$; see, e.g., (3) in Sampford (1953).

Next, let

$$f_{\text{obj}}(\theta) \equiv c\theta - (p^* F'(0) + h_0) \mu^* g(\theta)$$

be the objective function in (21), recalling that

$$g(\theta) = \frac{\theta}{\lambda F'(0)} + \frac{\sigma}{\sqrt{2F'(0)}} h_z\left(\frac{-\theta}{\lambda \sigma} \sqrt{\frac{2}{F'(0)}}\right),$$
for $\lambda = \lambda(p^*)$. Then

$$f''_{\text{obj}}(\theta) = -(p^*F'(0) + h_0) \mu^* g''(\theta) < 0,$$

because

$$g''(\theta) = \frac{1}{\lambda^2 \sigma F'(0)} \sqrt{\frac{2}{F'(0)}} h_z^2 \left( -\frac{\theta}{\lambda \sigma \sqrt{F'(0)}} \right) > 0 \quad \text{(A20)}$$

from the fact that the hazard rate function associated with a standard normal distribution is convex (see, for example, (4) in Sampford (1953)). We conclude that $f_{\text{obj}}$ is strictly concave. Therefore, to complete the proof, we must show a critical point exists; that is, there is a unique solution to $f''_{\text{obj}}(\theta) = 0$.

A critical point of $f_{\text{obj}}$ satisfies

$$c = (p^*F'(0) + h_0) \mu^* g'(\theta). \quad \text{(A21)}$$

A unique solution to (A21) exists if and only if

$$(p^*F'(0) + h_0) \mu^* \lim_{\theta \to -\infty} g'(\theta) < c < (p^*F'(0) + h_0) \mu^* \lim_{\theta \to \infty} g'(\theta), \quad \text{(A22)}$$

since $g'(\theta)$ is increasing in $\theta$ from (A20) and continuous. Straightforward calculus shows $h_z'(x) \to 1$ as $x \to \infty$ and $h_z'(x) \to 0$ as $x \to -\infty$. Hence from (A19)

$$\lim_{\theta \to -\infty} g'(\theta) = 0 \quad \text{and} \quad \lim_{\theta \to \infty} g'(\theta) = \frac{1}{\lambda(p^*)F'(0)},$$

and so

$$(p^*F'(0) + h_0) \mu^* \lim_{\theta \to -\infty} g'(\theta) = 0 \quad \text{(A23)}$$

$$(p^*F'(0) + h_0) \mu^* \lim_{\theta \to \infty} g'(\theta) = (p^*F'(0) + h_0) \frac{1}{F'(0)} \geq p^* > c, \quad \text{(A24)}$$

where we use $\lambda(p^*) = \mu^*$ and $p^* > c$ from Lemma 1 to derive (A24). Together (A22), (A23), and (A24) guarantee a unique solution to (A21) exists. \qed

**Proof of Lemma 4.** The proof structure is close to that of Lemma 3. While the first assertion can be checked from a direct calculation of $\frac{d}{d\theta} \mathbb{E}[V(\infty; \theta)]$, our argument here is based on a simple stochastic comparison result. Suppose $\theta_1 \leq \theta_2$. It is straightforward to see that the proof in Proposition 2 in Ward and Glynn (2003) can be extended to show that on each sample path

$$V(t; \theta_1) \leq V(t; \theta_2), \quad \text{for each} \quad t \geq 0, \quad \text{(A25)}$$

where $V(t; \theta_i)$ is as in (24) with $\theta$ replaced by $\theta_i$, $i = 1, 2$. So, $\mathbb{E}[V(t; \theta_1)] \leq \mathbb{E}[V(t; \theta_2)]$ for each $t \geq 0$. It is known from ergodicity of $\{V(t)\}_{t \geq 0}$ that (e.g., Corollary 5.11(2) in
Budhiraja and Lee (2007) \( \mathbb{E}[V(t; \theta)] \to \mathbb{E}[V(\infty; \theta)] \) as \( t \to \infty \) and hence we conclude that \( \mathbb{E}[V(\infty; \theta_1)] \leq \mathbb{E}[V(\infty; \theta_2)] \).

Now, we establish the second assertion. Taking the derivative in the objective function of (26), we determine a condition under which there exists a unique solution \( \theta^* \) to

\[
c = p^* \lambda(p^*) \frac{d}{d\theta} \mathbb{E} \left[ \int_0^V h(u) \, du \right] - h_0 \mu^* \frac{d}{d\theta} \mathbb{E}[V(\infty; \theta)].
\]

(A26)

Recall the process \( V \) in (24) has a unique steady-state \( V(\infty; \theta) \) whose density, as given in Proposition 6.1(i) in Reed and Ward (2008), is

\[
p(x; \theta) = M(\theta) \exp \left( \frac{2}{\sigma^2} \left( \frac{\theta}{\lambda} x - \int_0^x H(s) \, ds \right) \right), \quad x \geq 0,
\]

(A27)

where \( H(s) \equiv \int_0^s h(u) \, du \) is a cumulative hazard function and \( M(\theta) \in (0, \infty) \) is such that \( \int_0^\infty p(x; \theta) \, dx = 1 \). Such an explicit steady-state density formula along with a basic calculus yields that the objective function of (26) is strictly concave in \( \theta \).

Let \( U(\theta) \) denote the right hand side of (A26). A unique solution to (A26) exists if and only if

\[
\lim_{\theta \to -\infty} U(\theta) < c < \lim_{\theta \to \infty} U(\theta).
\]

(A28)

Recall the cost per capacity \( c \) must be such that \( 0 < c < p^* \) from Lemma 1. Now, we verify (A28) always holds. We first show \( \lim_{\theta \to -\infty} U(\theta) = 0 \). Our argument is based on a simple stochastic comparison result. Suppose \( \theta < 0 \). (The case \( \theta \geq 0 \) is immaterial since we are concerning only about \( \theta \to -\infty \).) Let \( R = \{R(t) : t \geq 0\} \) satisfy the stochastic equation \( R(t) = \sigma W(t) + \frac{\theta}{\lambda} t + L(t) \), where \( W \) is the same Brownian motion as the one used to define \( V \) in (24). It is straightforward to see that the proof in Proposition 2 in Ward and Glynn (2003) can be extended to show that on each sample path

\[
V(t) \leq R(t), \quad \text{for each} \quad t \geq 0,
\]

(A29)

and so \( \mathbb{E}[V(t)] \leq \mathbb{E}[R(t)] \) for each \( t \geq 0 \). It is well-known that (cf. Corollary 1.1.1 and Corollary 2.3.1 in Abate and Whitt (1987)) \( \mathbb{E}[R(t)] \to \mathbb{E}[R(\infty; \theta)] = \lambda \sigma^2/(2|\theta|) \) as \( t \to \infty \). Hence \( \{R(t) : t \geq 0\} \) is a uniformly integrable (UI) family, and therefore, from (A29), \( \{V(t) : t \geq 0\} \) is a UI family. Thus, \( \mathbb{E}[V(t; \theta)] \to \mathbb{E}[V(\infty; \theta)] \) as \( t \to \infty \) and we conclude that \( \mathbb{E}[V(\infty; \theta)] \leq \mathbb{E}[R(\infty; \theta)] \). With this bound and a fact \( \mathbb{E}[R(\infty; \theta)] = \lambda \sigma^2/(2|\theta|) \to 0 \) as \( \theta \to -\infty \), one has \( \mathbb{E}[V(\infty; \theta)] \to 0 \) as \( \theta \to -\infty \). Moreover, Corollary 2.3.1 in Abate and Whitt (1987) implies \( \mathbb{E}[(R(\infty; \theta))^2] = \lambda^2 \sigma^4/(2|\theta|^2) \), which tends to zero as \( \theta \to -\infty \). Thus, \( V(\infty; \theta) \to 0 \) a.s. as \( \theta \to -\infty \), which implies \( \mathbb{E} \left[ \int_0^V h(u) \, du \right] \to 0 \) as \( \theta \to -\infty \).

This, together with the fact that \( \mathbb{E}[V(\infty; \theta)] \) is monotonely decreasing as \( \theta \) decreases and \( \lim_{\theta \to -\infty} \mathbb{E}[V(\infty; \theta)] \to 0 \), the mean value theorem concludes \( \lim_{\theta \to -\infty} U(\theta) = 0 \) in (A28).
It remains to show the upper limit of (A28) satisfies \(\lim_{\theta \to \infty} U(\theta) \geq p^*\). (Recall \(c\) must be such that \(0 < c < p^*\) from Lemma 1.) In fact, we will show that if \(h_0 = 0\) then \(\lim_{\theta \to \infty} U(\theta) = p^*\). (Henceforth, assume \(h_0 = 0\).) To do this, we use a variant of Laplace method of asymptotic expansion for integrals, see, e.g., Chapter 3 of Miller (2006). Recall \(H(x) \equiv \int_0^x h(u)du\) and from (A27),

\[
\mathbb{E}\left[\int_{0}^{V(\infty; \theta)} h(u)du\right] = \mathbb{E}[H(V(\infty; \theta))] = \frac{\int_{0}^{\infty} H(x) \exp\left(\frac{2}{\sigma^2} \left(\frac{x}{\lambda} - \int_{0}^{x} H(s)ds\right)\right) dx}{\int_{0}^{\infty} \exp\left(\frac{2}{\sigma^2} \left(\frac{c}{\lambda} - \int_{0}^{c} H(s)ds\right)\right) dx}. \tag{A30}
\]

Now, we examine an asymptotic behavior of \(\mathbb{E}[H(V(\infty; \theta))]\) as \(\theta \to \infty\). For a real valued function \(f(x)\), consider an integral \(I(f(\cdot); \theta)\) defined by

\[
I(f(\cdot); \theta) \equiv \int_{0}^{\infty} f(x) \exp\left(\frac{2}{\sigma^2} \left(\theta x - \int_{0}^{x} H(s)ds\right)\right) dx = \int_{0}^{\infty} f(x) \exp(\theta G(x)) dx,
\]

where

\[G(x) \equiv \frac{2}{\sigma^2 \lambda} x - \frac{2}{\sigma^2 \theta} \int_{0}^{x} H(s)ds.\]

Then, the integral in (A30) is given as \(I(H(\cdot); \theta)/I(1(\cdot); \theta)\), where \(1(\cdot)\) stands for a constant function \(1(x) = 1\). Notice that \(G'(x) = \frac{2}{\sigma^2 \lambda} - \frac{2}{\sigma^2 \theta} H(x)\) and 
\(G''(x) = -\frac{2}{\sigma^2 \theta} h(x) \leq 0\) for \(x \geq 0\) (because the hazard rate \(h(x) \geq 0\)). So, \(G(x)\) has a unique maximizer \(x_0\) such that \(G'(x_0) = 0\), which implies \(H(x_0) = \frac{\theta}{\lambda}\) (recall \(H(x)\) is a cumulative hazard rate and hence increasing in \(x\)). Intuitively, for large values of \(\theta\), the integrand of \(I(f(\cdot); \theta)\) has one narrow sharp peak around \(x_0\) and the integral value \(I(f(\cdot); \theta)\) is completely dominated by that peak. From Section 3.4 of Miller (2006), the ratio of two integrals has an asymptote given as

\[\mathbb{E}[H(V(\infty; \theta))] = \frac{I(\theta(\cdot); \theta)}{I(1(\cdot); \theta)} \sim H(x_0) = \frac{\theta}{\lambda} \quad \text{as} \quad \theta \to \infty,
\]

where \(f(x) \sim g(x)\) as \(x \to \infty\) means \(\lim_{x \to \infty} f(x)/g(x) = 1\). Now, a direct calculation (using the ratio expression in (A30)) shows that \(\frac{d}{d\theta} \mathbb{E}[H(V(\infty; \theta))]\) is continuous and non-decreasing in \(\theta\), which implies, from, e.g., Exercise 4.58 in Francinou et al. (2014), that

\[\frac{d}{d\theta} \mathbb{E}[H(V(\infty; \theta))] \sim \frac{1}{\lambda} \quad \text{as} \quad \theta \to \infty.
\]

Hence, we conclude that if \(h_0 = 0\) then \(\lim_{\theta \to \infty} U(\theta) = p^*\), that is, in (A26), one has \(p^* \lambda(p^*) \frac{d}{d\theta} \mathbb{E}\left[\int_{0}^{V(\infty; \theta)} h(u)du\right]\) converges to \(p^*\) as \(\theta \to \infty\). To sum up, \(\lim_{\theta \to \infty} U(\theta) \geq p^*\) whenever \(h_0 \geq 0\). This completes the proof. \(\square\)

**Acknowledgements:** We would like to thank Ramandeep Randhawa for valuable discussion. We also thank Junfei Huang for a discussion that helped us to confirm the proof of Lemma 2. This work is supported by the William S. Fishman Faculty Research Fund at the University of Chicago Booth School of Business.
References


