Capacity Sizing Under Parameter Uncertainty: Safety Staffing Principles Revisited

Achal Bassamboo*
Northwestern University

Ramandeep S. Randhawa†
University of Southern California

Assaf Zeevi‡
Columbia University

To appear in *Management Science, 2010*

Abstract

We study a capacity sizing problem in a service system that is modeled as a single-class queue with multiple servers and where customers may renege while waiting for service. A salient feature of the model is that the mean arrival rate of work is *random* (in practice this is a typical consequence of forecasting errors). The paper elucidates the impact of uncertainty on the nature of capacity prescriptions, and relates these to well established rules-of-thumb such as the square root safety staffing principle. We establish a simple and intuitive relationship between the incoming load (measured in Erlangs) and the extent of uncertainty in arrival rates (measured via the coefficient of variation) which characterizes the extent to which uncertainty dominates stochastic variability or vice versa. In the former case it is shown that traditional square root safety staffing logic is no longer valid, yet simple capacity prescriptions derived via a suitable newsvendor problem, are surprisingly accurate.

1 Introduction

Background and motivation. Motivated by telephone call center applications, this paper focuses on the general problem of capacity sizing in service systems under parameter uncertainty. In the case of call centers, one natural manifestation of this uncertainty is prediction errors associated with forecasted load. In the language of operations research, this implies that the *arrival rates* are uncertain, which stands in contrast to the bulk of literature that assumes all model primitives are known with certainty, and "noise" is restricted to stochastic variability in *realizations*; see the literature review in section 2.

Determining appropriate processing capacity is a key decision in the management of many modern service systems. The choice of capacity level is often made to strike the "right" balance between basic operating costs (e.g., personnel, infrastructure etc), and customer experience (e.g., waiting

*Kellogg School of Management, e-mail: a-bassamboo@northwestern.edu
†Marshall School of Business, e-mail: ramandeep.randhawa@marshall.usc.edu
‡Graduate School of Business, e-mail: assaf@gsb.columbia.edu

times, abandonments etc). Most academic studies attribute congestion-related effects to stochastic variability in customer arrivals and service requirements, and this view is often echoed by practitioners. As a consequence, a typical rule-of-thumb for capacity planning involves: i) setting a "base capacity" to match mean demand, computed using arrival and service rate parameters; and ii) augmenting that by a "safety capacity" that hedges against variability in *realized* arrivals/services. To make this more concrete, consider a single class multi-server queue with infinite waiting room, and where customers are processed according to the order of their arrival. Arrivals follow a Poisson process with rate λ and services are exponentially distributed with rate μ. Then, the capacity prescription alluded to above is given by $C = \lambda/\mu + \beta\sqrt{\lambda/\mu}$ for some constant β (i.e., this is the prescribed number of servers, ignoring integrality constraints). The second term represents the *variability hedge*, and takes the familiar form of the square-root safety staffing principle which dates back to the seminal work of Erlang (1917). Roughly speaking, since a Poisson random variable has variance equal to its mean, then the second term above is exactly of the same order as stochastic fluctuations in this system. The decision variable β dictates the amount of hedging, and is a result of a "second order" optimization problem that seeks a suitable trade-off between staffing costs and quality-of-service. This form of staffing gives rise to the so-called Quality and Efficiency Driven (QED) regime that has been extensively studied in the literature; see the recent survey paper by Gans, Koole and Mandelbaum (2003) for more on this regime, and the literature review in section 2 for a sample of papers, many of them quite recent, that formulate and study such problems.

The square-root-rule is predicated on precise knowledge of the mean behavior of stochastic primitives, in particular, arrival and service rates. However, in most settings one encounters in practice such key parameters must be inferred or forecasted based on available data, and hence can be quite "noisy." Empirical findings in recent literature have corroborated the presence of such parameter uncertainty; see, e.g., Brown, Gans, Mandelbaum, Sakov, Shen, Zeltyn and Zhao (2005) for an empirical study in a call-center setting. Below is an illustrative example, suggestive of this phenomenon, that uses data from a moderate sized European retail banking call center. Focusing on the time slot of 8 – 10 AM, we use 52 weeks of data to compute estimates of the mean number of arriving calls for each day of the week in this time slot ($\hat{\lambda}$) along with their coefficient of variation (defined as the ratio of standard deviation to the mean, namely, CV [empirical] $= \hat{\sigma}_\lambda/\hat{\lambda}$). Table 1 presents the estimates for the mean and coefficient of variation, along with an estimate of the latter that *assumes* the underlying arrival process is Poisson (CV [Poisson] $= 1/\sqrt{\hat{\lambda}}$). Note that the observed values of coefficient of variation tend to be much greater than under the Poisson assumption. One possible explanation for this discrepancy is the presence of ambiguity in the arrival rate itself. This leads to the following natural question:

> What is the appropriate hedge in setting capacity levels under parameter uncertainty, and how does this relate to the square-root-rule?

Day of Week	Mean no. of arriving calls	CV [empirical] (%)	CV [Poisson] (%)
Mon	943	26.5	3.3
Tue	824	22.3	3.5
Wed	807	26.5	3.5
Thu	778	28.5	3.6
Fri	767	33.5	3.6
Sat	293	61.8	5.8
Sun	139	148.1	8.5

Table 1: **Data from a telephone call center.** Mean number of call arrivals in the time slot 8 – 10AM in 2004. CV [empirical] is a direct estimate of the coefficient of variation (in %) based on the data, and CV [Poisson] is estimated assuming the underlying arrival process were Poisson.

A recent stream of work, initiated by Harrison and Zeevi (2005), addresses the question of capacity planning under parameter uncertainty by proposing the following model: assuming that uncertainty effects dominate stochastic fluctuations, one can ignore the latter and formulate a suitable *newsvendor problem* of capacity planning. The hedge in this case focuses on parameter uncertainty and not variability. In section 3 we consider a widely used stylized model of a telephone call center and a prototypical problem of capacity planning, and develop the aforementioned newsvendor approximation and its solution. Our main goal is to examine the salient properties of this solution and shed light on its relationship to the square-root safety staffing principle. (Further discussion contrasting the present paper with the work of Harrison and Zeevi (2005) is postponed to Section 2.)

Main findings and key qualitative insights. Our analysis gives rise to a simple rule of thumb that identifies two regimes: an *uncertainty-dominated* regime where there is no tangible benefit from a variability hedge as prescribed by the square-root rule; and a *variability-dominated* regime where there may be potential benefits from introducing an additional variability hedge. In particular, despite the crude and simple logic underlying the newsvendor formulation alluded to above, the uncertainty hedge embodied in its solution is extremely accurate. These statements are buttressed with analytical results as well as extensive numerical experiments.

To give a flavor of these findings, let us first explain what is meant by "extremely accurate" in speaking of the newsvendor-based solution: if one plugs this solution into the actual system, then the resulting performance is almost indistinguishable from the best achievable performance. (The latter is associated with the optimal solution to the *original* capacity planning problem, which can rarely be computed analytically.) The mathematical statement that underlies this observation is perhaps even more striking. In systems with a high level of parameter uncertainty, the gap between

the performance of the newsvendor-based solution and that of the optimal solution *does not* scale up as one increases the volume of work flowing through the system. This is surprising in light of the rather "crude" method used to derive the newsvendor-based solution, and considering that ignoring variability effects typically results in performance that deteriorates when the arrival rate scales up: for example, in the absence of a variability hedge the optimality gap is expected to double when the magnitude of arrival rate quadruples (see further discussion and intuition in Section 4).

If arrival rates were known then one would expect the accuracy of the newsvendor-based solution to deteriorate. In that case, it becomes optimal to operate the system in the QED regime where the square-root "safety capacity" corrections become pertinent. Indeed, we analytically prove that if the coefficient of variation of the (random) arrival rate Λ is larger than a threshold proportional to $\sqrt{1/\mathbb{E}\Lambda}$, as is the case for the example in Table 1, then we are in the *uncertainty-dominated* regime. There we show that the performance of the newsvendor-based prescription cannot be improved upon in any significant manner. If the coefficient of variation is smaller than that threshold, then the optimality gap characterizing the newsvendor-based solution is proportional to $\sqrt{\mathbb{E}\Lambda}$, the square root of the mean arrival rate. This is the *variability-dominated* regime mentioned above, and it is here that the square-root "safety capacity" corrections may contribute to further improving performance.

Remainder of the paper. The next section surveys related literature. Section 3 describes the capacity planning problem and the proposed solution. The main findings of the paper are communicated in two sections: numerical results that illustrate the key insights, as well as underlying intuition, are presented in section 4; and the theoretical foundations are derived in section 5. The robustness of the main findings are presented in section 6. Some concluding remarks and future research directions are covered in section 7. Proofs are relegated to Appendix A and B. Further numerical results are presented in Appendix C.

2 Literature Review

Origins and scope of the square-root rule. As alluded to earlier, the bulk of the literature on capacity planning in service operations focuses on the (more traditional) case where all pertinent system parameters are fixed and given. This line of work dates back to Erlang (1917), who was the first to suggest the square-root safety capacity rule in the context of his study of circuit-switched telephony. Much of the literature relevant to the present paper is more recent and by and large motivated by telephone call centers, where capacity planning translates into staffing decisions. Good recent surveys are contained in Gans et al. (2003) and Aksin, Armony and Mehrotra (2007).

The variability hedge interpretation and intuition underlying the square-root rule have already

been discussed briefly in section 1; see Kolesar and Green (1998) for further discussion in the context of a constraint satisfaction capacity planning problem. Halfin and Whitt (1981) provide a more rigorous foundation that builds on a diffusion approximation. Their work was the first to formally identify and characterize the QED operating regime in the context of an $M/M/N$ queue; for this reason it is also referred to as the Halfin-Whitt regime. The square-root rule has since been extended and applied in a variety of settings that go beyond the basic Erlang formulas. Representative examples include Jennings, Mandelbaum, Massey and Whitt (1996) that treat the case of time-varying arrival rates and Garnett, Mandelbaum and Reiman (2002) that deals with the same model as Halfin and Whitt (1981) with the addition of customers reneging according to an exponential patience distribution (i.e., the $M/M/N+M$ model). Extensions to the case of general reneging distributions are treated in Mandelbaum and Zeltyn (2009) and Bassamboo and Randhawa (2009). The paper by Borst, Mandelbaum and Reiman (2004) is among the first to consider an optimization problem that seeks to balance personnel costs and customer delays. Gurvich, Armony and Mandelbaum (2008) covers the case of multiple customer classes and therefore deals with both staffing and dynamic assignment of incoming work to servers; see also Gurvich and Whitt (2009). Gans et al. (2003, Section 4.1.1) provides a detailed discussion of the square-root safety rule in the context of call center staffing, and further pointers to related literature.

Identifying and modeling parameter uncertainty. Whitt (1999) is among the first papers to flag the issue of arrival rate (demand) uncertainty. As indicated earlier, in many service operations it is natural to view forecast errors as a proxy for this uncertainty; see, e.g., Jongbloed and Koole (2001) who propose a method to compute forecast intervals and demonstrate risks of using point estimates. Chen and Henderson (2001) provide empirical evidence of uncertainty in arrival rates and discuss modeling implications. Using empirical data from a banking call center, Brown et al. (2005) execute a detailed empirical study that, among other things, covers also arrival rate properties. Steckley, Henderson and Mehrotra (2009) investigate the presence of forecast errors in service systems and provide empirical evidence that these errors can dominate fluctuations due to stochastic variability. Avramidis, Deslauriers and L'Ecuyer (2004) suggest stochastic models to capture this uncertainty and test these using empirical data. The recent work of Maman (2009) also finds empirical evidence of this uncertainty in both call center applications as well as health care operations, and discusses performance implications (see also further discussion in section 5). The results displayed in Table 1 are consistent with most of the empirical observations made in the aforementioned papers.

Capacity planning under parameter uncertainty and relations to the present paper. Despite the empirical evidence of potentially significant "noise" in arrival rate estimates, the stream of literature that deals with capacity planning in such settings, including some of the work cited in the previous paragraph, is in a relatively nascent stage. Harrison and Zeevi (2005) is among the

first studies to propose a staffing method in queueing models with multiple customer classes and server pools. Their approach is based on reducing the original capacity optimization problem to a multi-dimensional newsvendor problem (see also Robbins and Harrison (2007) for a computational oriented study). The logic underlying this reduction germinates in viewing stochastic variability as a "lower order" effect in comparison with parameter (demand) uncertainty. The method can therefore be viewed as crafting a *stochastic* fluid model of the original stochastic network. Bassamboo, Harrison and Zeevi (2006) and Bassamboo and Zeevi (2009) are examples of subsequent work that treat a joint staffing and dynamic call assignment problem, and use historical call records to create a data-driven implementation of the basic method, respectively.

The present paper is most closely connected to the work of Harrison and Zeevi (2005). In particular, the newsvendor formulation and its fractile solution that we discuss in the next section are special cases of that studied in Harrison and Zeevi (2005) which allows for temporal variation in arrival rates, and treats general parallel server-type network models. The purpose of our present paper is to conduct a more detailed investigation of the efficacy of the newsvendor prescription of Harrison and Zeevi (2005). To that end, we focus on a more restricted queueing model, namely one with only a single customer class and single server pool, trading off generality for tractability. In particular, unlike Harrison and Zeevi (2005) which centers almost exclusively around a numerical study, our work establishes theoretical properties of the newsvendor-based prescription, which in turn allows us to elucidate its connection with the extant square-root rule.

There are two other studies that focus on the single-class/single-pool setting and are related to our work. The first is that of Whitt (2006*b*) that discusses staffing levels in a call center subject to arrival rate uncertainty and absenteeism of servers. The focus of that paper is primarily on numerical investigation of a fluid-based prescription, building on Whitt (2006*a*) that develops fluid approximations for a $G/GI/s+GI$ queue. The other reference is the M.S. thesis of Maman (2009) that is concerned with demand uncertainty and its implications in call centers and healthcare operations. Her work provides further empirical evidence of the "over-dispersion" observed in actual arrival patterns relative to a Poisson distribution, and suggests adjusting the square-root rule to a power that is greater than a $1/2$ based on the excess variability one observes in the data. The main thrust of her work is quite distinct from ours: it does not focus on the newsvendor-based prescription and the study of its theoretical properties, but rather seeks to directly adjust the existing square-root rule to account for added variability. (It will be conducive to postpone further comments on this work to section 5, at which point our main results have been presented.)

3 The Capacity Planning Problem and An Approximate Solution

Problem formulation. We consider customers that arrive to a service system with b statistically identical servers. Customer arrivals are modeled using a doubly stochastic Poisson process: the arrival rate Λ is itself a random variable with distribution F and mean denoted by λ; and conditioned on Λ, the arrival process is a homogenous Poisson process with that rate. Customers have service requirements that are i.i.d. exponential random variables, independent of the arrival process and rate, with mean $1/\mu$. Customers are processed in the order in which they arrive and servers do not idle when there is work waiting to be processed. When all servers are busy, arriving customers wait in an infinite capacity buffer. A customer may abandon the system while waiting for the commencement of service. Specifically, each customer is endowed with an exponentially distributed "impatience" random variable τ that has mean $1/\gamma$; the customer will abandon the system when his waiting time in queue reaches a total of τ time units. This description corresponds to an $M/M/b+M$ queueing model, where the arrival rate is itself random.

There is a cost associated with customers waiting in queue, as well as with each abandoning customer. The holding cost is incurred at a rate h per customer per unit time spent waiting in queue, while the cost of abandonment is p per customer. The system manager decides the capacity level b, i.e., the number of servers, to minimize the system cost. The cost of staffing each server is c per unit time. (For convenience we normalize the length of the planning horizon to unity.) Denoting the number of customers in the system in steady state by the random variable N, the total expected customer cost in steady state is given by $(h+p\gamma)\mathbb{E}[N-b]^+$, where $\mathbb{E}$ denotes the expectation operator. Thus, the optimization problem is: find $b \geq 0$ that minimizes

$$\Pi(b) := (h+p\gamma)\mathbb{E}[N-b]^+ + cb. \tag{1}$$

In the above, and in what follows, we ignore integrality constraints, and assume for simplicity that b is real-valued. Let b^* denote the minimizer and $\Pi^* := \Pi(b^*)$ the corresponding minimum cost.[†]

An approximate solution. Despite apparent simplicity, it is not possible to get an exact solution to (1); this stems from the difficulty in characterizing the distribution of N, which depends intricately on the capacity level b. While one can clearly solve this optimization problem numerically, we will propose a different approach based on an approximation of the objective function in (1). Following Harrison and Zeevi (2005), the main idea is to ignore *stochastic variability* in customer arrivals and service requirements, and instead focus on the *uncertainty* in the arrival rate. In this relaxation, customers arrive in the form of fluid at the rate of Λ per unit time. The processing capacity is fixed at μb. Thus, the difference in these two rates should equal the rate of customer abandonment. Thus the approximate rate of abandonment is $\mathbb{E}[\Lambda-\mu b]^+$. An exact expression

[†]Noting that $\mathbb{E}[N-b]^+$ is convex, we obtain that Π is a convex function, and thus has a well defined minimizer.

for the rate of abandonment is obtained by multiplying the expected steady state queue-length by the abandonment rate γ, and yields $\gamma\mathbb{E}[N-b]^+$. Equating these two expressions, we obtain the approximation for the expected steady-state queue-length

$$\mathbb{E}[N-b]^+ \approx \frac{1}{\gamma}\mathbb{E}[\Lambda-\mu b]^+.$$

Using this approximation, the optimization problem (1) reduces to:

$$\min_{b\geq 0} \left\{\bar{\Pi}(b) := (p+h/\gamma)\mathbb{E}[\Lambda-\mu b]^+ + cb\right\}. \tag{2}$$

This is an instance of the familiar newsvendor problem: optimize over processing capacity $b\mu$ (where b is the decision variable), with a unit capacity cost of c/μ and a unit sales price of $p+h/\gamma$. Thus, we obtain the newsvendor-based prescription as the standard critical fractile solution

$$\bar{b} = \frac{1}{\mu}\bar{F}^{-1}\left(\frac{c/\mu}{p+h/\gamma}\right), \tag{3}$$

where F is the cumulative distribution function of Λ, $\bar{F} := 1-F$ denotes the corresponding tail distribution function, and $\bar{F}^{-1}$ denotes its generalized inverse, i.e., $\bar{F}^{-1}(y) = \inf\{x \in \mathbb{R}_+ : \bar{F}(x) \leq y\}$. Note that if the cost of a unit of capacity is prohibitively high, in particular higher than the unit penalty charge ($c/\mu > p+h/\gamma$), then it follows that the optimal solution to (2) does not install any capacity. This is consistent with (3).

Discussion and main objectives. The prescription defined above was originally proposed by Harrison and Zeevi (2005) (see also Whitt (2006*b*)). Our main objective in what follows is to elucidate salient features of this prescription, draw connections with the extant square-root safety staffing principle, and extract more general qualitative insights that pertain to the management of service operations under parameter uncertainty. Before proceeding to uncover some of the rather surprising characteristics of the newsvendor-based prescription, it is worth pointing out that this approach is predicated on logic that is quite different from most antecedent work in the area of queueing and service systems. In particular, the critical fractile solution embodies within it an *uncertainty hedge* that is proportional to the standard deviation of the arrival rate distribution F (see Proposition 2 in section 5). In contrast, the more traditional *variability hedge*, which underlies the square-root rule, prescribes a safety capacity which is, roughly speaking, proportional to the standard deviation of stochastic fluctuations in the arrival process.

4 Main Results: Numerical Illustration and Intuition

In this section we highlight the essence of our main findings via illustrative numerical examples, and discuss the intuitive foundations that underlie them. Section 5 will then formalize this and provide theoretical justification.

4.1 Illustration of the main findings

In all numerical examples considered in this section, we take the cost parameters to be $c = 1/3$ and $p = h = 1$, and set the mean service time to 1 and the mean impatience time to $1/3$. In this manner the aggregate penalty cost $p + h/\gamma$ and the cost of capacity ensure a non-trivial solution. (Appendix C provides further evidence via more extensive numerical results, which have been cut out of the main text for brevity.)

Performance of the newsvendor-based prescription: Deterministic versus uncertain arrival rates. We consider two numerical examples. In the first, the arrival rates Λ follow a uniform distribution, and for illustrative purposes we focus on three cases: $U[25, 50]$, $U[50, 100]$, and $U[200, 400]$. The realization of this random variable will then govern the mean of the arrival (Poisson) process. The results for this setting are summarized in Table 2.

Arrival rate	Optimal solution		Prescription		Difference		
distribution	b^*	Π^*	$\bar{b}$	$\Pi(\bar{b})$	$\lvert b^* - \bar{b}\rvert$	$\Pi(\bar{b}) - \Pi^*$	$\frac{(\Pi(\bar{b}) - \Pi^*)}{\Pi^*}$
U[25,50]	46	17.34	43	17.49	3	0.15	0.9%
U[50,100]	89	33.13	87	33.20	2	0.07	0.2%
U[200,400]	351	127.08	350	127.08	1	0.007	0.006%

Table 2: **Performance of the newsvendor-based prescription: uncertain arrival rates.** Comparison of the optimal (b^*) and prescribed ($\bar{b}$) capacity levels along with their respective performance. Arrival rates follow a uniform distribution with a coefficient of variation of 19.2%.

It will be useful to contrast these results with a more traditional setting where the arrival rate is deterministic, and this is summarized in Table 3. Specifically, we consider cases where the arrival rate deterministically takes the values $37.5, 75$, and 300, i.e., the arrival rate is set to be the mean of the uniform distributions considered above.

Arrival rate	Optimal solution		Prescription		Difference		
λ	b^*	Π^*	$\bar{b}$	$\Pi(\bar{b})$	$\lvert b^* - \bar{b}\rvert$	$\Pi(\bar{b}) - \Pi^*$	$\frac{(\Pi(\bar{b}) - \Pi^*)}{\Pi^*}$
37.5	42	15.94	37	17.67	5	1.73	10.9%
75	83	29.64	75	31.60	8	1.96	6.6%
300	316	109.01	300	112.4	16	3.39	3.1%

Table 3: **Performance of the newsvendor-based prescription: deterministic arrival rates.** Comparison of the optimal (b^*) and prescribed ($\bar{b}$) capacity levels along with their respective performance. Arrival rates are deterministic.

Before discussing the results in these tables, let us first briefly explain the computational details. For the case with uncertainty, our proposed newsvendor prescription calls for the computation of $\bar{b}$ given in (3), which is straightforward (for the case without uncertainty $\bar{b} = \lambda/\mu$). The performance of this solution, $\Pi(\bar{b})$, is then assessed using the characterization of the steady-state distribution in the underlying Markov chain, conditional on the arrival rate. The expected penalty costs are then computed using numerical integration. The optimal capacity b^* which minimizes $\Pi(\cdot)$ in (1) was computed numerically by evaluating the objective function for a large set of capacity levels, and utilizing the convexity of the objective function. This also gives the *best achievable performance*, $\Pi^* = \Pi(b^*)$. Deterministic arrival rates are clearly a special case of the above.

What stands out in Table 2 is the remarkable accuracy and performance of the newsvendor-based capacity prescription. In particular, the capacity levels prescribed by our newsvendor logic ($\bar{b}$) are extremely close to the optimum (b^*) and the difference in performance is always below 0.15, i.e., a 0.9% optimality gap. The results presented in the table are in fact indicative of a more general phenomenon; we will shortly provide an intuitive explanation, which is then made rigorous in section 5.

In contrast, the results in Table 3, present an optimality gap that is *increasing* as the magnitude of the arrival rate increases. In particular, when the arrival rate quadruples from 75 to 300, the corresponding optimality gap is seen to approximately double. Moreover, the difference between the optimal capacity level b^* and the newsvendor-based prescription $\bar{b}$ seems to diverge as well as the volume of arrivals increases. Specifically, the optimal capacity level is consistently larger than $\bar{b}$; the difference $\beta := b^* - \bar{b}$ can be identified as the safety capacity, stemming from the "variability hedge" alluded to in section 1. A more careful look at these differences reveals that they are roughly equal to square-root the arrival rate (i.e., $\beta \approx 1 \cdot \sqrt{\lambda}$). This is exactly what one would expect based on the square-root safety staffing rule, which is applicable in the setting of deterministic arrival rates: the base capacity $\bar{b}$ is augmented by a *safety capacity* that is proportional to $\sqrt{\lambda}$.

4.2 Intuition

To build intuition, it will be convenient to restrict attention in this section to the case where the abandonment rate γ equals the service rate μ, and the cost of queueing and abandonments $h + p\gamma$ is normalized to unity. (Our main theoretical results will make this intuition rigorous for the general case.)

Connection with the square-root rule. With the above choice of parameters, the number of customers in the system is identical to that in an infinite server queue with service rate μ. Consider first the case where the arrival rate is deterministic, i.e., $\Lambda = \lambda$. The steady-state distribution of the number-in-system is then Poisson with mean λ/μ, and the expected queue-length is

$\mathbb{E}[\text{Poisson}(\lambda/\mu) - b]^+$. Using the normal approximation for a Poisson distribution (see, e.g., Kolesar and Green (1998) for a similar application) we get the following expression for the cost of a capacity level b:

$$\begin{aligned}\Pi(b) &\approx \mathbb{E}[\text{Normal}(\lambda/\mu, \lambda/\mu) - b]^+ \\ &= (\lambda/\mu - b)^+ + K\sqrt{\lambda/\mu}\exp\left(-\frac{(\lambda/\mu - b)^2}{2(\lambda/\mu)}\right) + cb,\end{aligned} \tag{4}$$

which follows by taking the expectation of the truncated normal distribution. Here K is a finite positive constant whose exact value is not pertinent to the arguments below.

Given the above, we have that $\bar{b} = \lambda/\mu$. Let $\beta = b - \bar{b}$ denote the corresponding safety capacity. Then, optimizing (4) over b is equivalent to optimizing

$$-\min\{\beta, 0\} + K\sqrt{\lambda/\mu}\exp\left(-\frac{\beta^2}{2(\lambda/\mu)}\right) + c\beta,$$

over β, which yields $\beta^* \approx C_1\sqrt{\lambda/\mu}$, where the constant C_1 can be solved for explicitly and expressed in terms of K and c; we omit the details. Thus, the optimal capacity satisfies $b^* \approx \bar{b} + C_1\sqrt{\lambda/\mu}$. It follows that

$$\begin{aligned}\Pi(b^*) &\approx c\lambda/\mu + C_2\sqrt{\lambda/\mu} \\ &= \Pi(\bar{b}) - C_3\sqrt{\lambda/\mu},\end{aligned}$$

for positive constants C_2, C_3 that can be explicitly identified. The first term on the right-hand-side is the performance of the newsvendor-based prescription $\bar{b}$. Thus, the above provides a simple derivation of the square-root rule and its performance: both safety capacity and the gap in performance, namely $\Pi(\bar{b}) - \Pi(b^*)$, grow proportionally to $\sqrt{\lambda/\mu}$, i.e., the square-root of the system load measured in Erlangs. This is one of the main observations present in Table 3.

Performance of the newsvendor-based prescription. We now turn to the case where the arrival rate Λ is a random variable. Returning to (4), and using the tower property of conditional expectations, we arrive at

$$\begin{aligned}\Pi(b) &\approx cb + \mathbb{E}[\Lambda/\mu - b]^+ + K\mathbb{E}\left[\sqrt{\Lambda/\mu}\exp\left(-\frac{(\Lambda/\mu - b)^2}{2(\Lambda/\mu)}\right)\right] \\ &= \bar{\Pi}(b) + K\mathbb{E}\left[\sqrt{\Lambda/\mu}\exp\left(-\frac{(\Lambda/\mu - b)^2}{2(\Lambda/\mu)}\right)\right],\end{aligned} \tag{5}$$

where the expectation above is with respect to Λ. The sum of the first two terms on the right-hand-side of (5) is exactly the objective function in our proposed newsvendor problem, and the capacity choice that optimizes this sum is exactly $\bar{b}$, the newsvendor-based prescription. Hence, the performance of $\bar{b}$ is directly linked to the accuracy of the proposed newsvendor objective function

$\bar{\Pi}(\cdot)$ vis-à-vis the actual objective function $\Pi(\cdot)$. The *accuracy gap* can therefore be expressed as follows:

$$\begin{aligned} \Delta(b) &:= \Pi(b) - \bar{\Pi}(b) \\ &\approx K\mathbb{E}\left[\sqrt{\Lambda/\mu}\exp\left(-\frac{(\Lambda/\mu - b)^2}{2(\Lambda/\mu)}\right)\right]. \end{aligned} \tag{6}$$

The results given in Table 2 indicate that the above accuracy gap is very small, and in particular does not grow as the mean arrival volume grows. To see how this follows from (6), note that for any fixed capacity level b, the uncertainty in the arrival rate implies that Λ/μ will never exactly equal b. In particular, when it is significantly above/below b, corresponding to overloaded/underloaded scenarios, the accuracy gap $\Delta(b)$ will be dominated by the exponentially decaying term on the right-hand-side of (6). The critically loaded case, i.e., when realizations of Λ fall in "close proximity" to b, only occurs with "low" probability. Taking expectations, it then follows that $\Delta(b)$ remains bounded as the mean demand level increases. Finally, since the approximation error introduced by using the newsvendor model in lieu of the original performance function is clearly "negligible" (with regard to its dependence on mean arrival rate), it follows that the minimizer of $\Pi(b)$, namely b^*, is essentially obtained by minimizing the newsvendor-part of the right-hand-side of (5). This explains why b^* and $\bar{b}$ are almost indistinguishable in the results displayed in Table 2, and indicates that square-root staffing corrections to $\bar{b}$ will have barely noticeable impact on performance.

4.3 What happens under moderate uncertainty?

The results discussed above explain the performance of the newsvendor-based solution in the case where uncertainty is "dominant" relative to variability. This raises the question of what happens in the case where arrival rate uncertainty is more "modest."

To investigate this further, we perform another numerical experiment, in which we vary the level of uncertainty in the arrival rate and study the performance of the newsvendor-based prescription. The choice of parameters is identical to previous examples, only here we select several uniform distributions that reflect decreasing levels of uncertainty: this is implemented by shrinking their support. The results are displayed in Table 4. Note that as the uncertainty in the arrival rate decreases, the optimality gap associated with the newsvendor-based prescription increases. Nevertheless, the optimality gap is still seen to be quite small for cases where the coefficient of variation exceeds 5%.

Referring to (6), this performance can be seen to depend on two system measures: the amount of uncertainty in the arrival rate relative to the mean, and the offered load. The first measure is the coefficient of variation of the arrival rate CV, where $CV = \sigma/\lambda$, with $\lambda = \mathbb{E}\Lambda$ denoting the mean arrival rate and σ the standard deviation. The second measure is the offered load given by

Arrival rate		Optimal solution		Prescription		Difference		
Distribution	CV (%)	b^*	Π^*	$\bar{b}$	$\Pi(\bar{b})$	$\lvert b^* - \bar{b}\rvert$	$\Pi(\bar{b}) - \Pi^*$	$\frac{(\Pi(\bar{b})-\Pi^*)}{\Pi^*}$
$\Lambda \sim$ U[0,300]	57.73	224	88.34	225	88.34	1	0	0%
$\Lambda \sim$ U[125,175]	9.62	165	59.06	162	59.16	3	0.10	0.2%
$\Lambda \sim$ U[135,165]	5.77	162	57.40	157	57.78	5	0.38	0.7%
$\Lambda \sim$ U[140,160]	3.85	162	56.78	155	57.42	7	0.64	1.1%
$\Lambda \sim$ U[145,155]	1.92	161	56.40	152	57.73	9	1.33	2.4%
$\Lambda = 150$	0	161	56.26	150	58.25	11	1.99	3.5%

Table 4: **Performance of the newsvendor-based prescription: impact of uncertainty magnitude.** Comparison of the optimal (b^*) and prescribed ($\bar{b}$) capacity levels, along with their respective performance.

$\mathcal{E} = \lambda/\mu$ Erlangs. Notice that if the CV is on the order of $1/\sqrt{\mathcal{E}} = \sqrt{\mu/\lambda}$, then the third term on the right-hand-side of (5) behaves essentially as if the arrival rate were deterministic: the accuracy gap is then anticipated to be proportional to the square-root of the load.

Main qualitative insight. The above results suggest the following rule of thumb. If the coefficient of variation of the arrival rate has *larger* magnitude than $1/\sqrt{\mathcal{E}}$, then the newsvendor-based prescription (3) has a very accurate performance. Square-root safety capacity corrections are anticipated to offer limited improvements here. In the other case, where the coefficient of variation of the arrival rate is of *smaller* magnitude than $1/\sqrt{\mathcal{E}}$, one may benefit from refining (3) using a square-root staffing approach. (Note that in Table 4, we have $\mathcal{E} = 150$, and hence $1/\sqrt{\mathcal{E}} \approx 8\%$.) We remark that arrival rate uncertainty, and the coefficient of variation, are in many cases a consequence of forecasting methods: for example, uncertainty corresponds naturally to prediction confidence intervals.

5 Main Results: Theoretical Foundations

Preliminaries. As described earlier, we consider a system with arrivals that follow a doubly stochastic (Poisson) process with constant and random rate Λ such that $\lambda := \mathbb{E}\Lambda > 0$, and standard deviation $\sigma_\lambda < \infty$. Our goal is to study the accuracy of the prescription (3), and we carry this out in a regime where the mean arrival rate λ is large. (This is characteristic of many modern service operations, given considerations of economies-of-scale.) To this end, we index various quantities with a subscript λ, for example, the distribution of the arrival rate will be denoted F_λ, its coefficient of variation $CV_\lambda = \sigma_\lambda/\lambda$ and the offered load $\mathcal{E}_\lambda = \lambda/\mu$. The corresponding optimization problem

is: choose $b \geq 0$ to minimize

$$\Pi_\lambda(b) := (h + p\gamma)\mathbb{E}[N_\lambda - b]^+ + cb, \tag{7}$$

where N_λ denotes the number of customers in system in steady-state. We denote the optimal value of the objective function by Π^*_λ. The newsvendor-based capacity prescription is given by the solution to (7), namely,

$$\bar{b}_\lambda = \frac{1}{\mu}\bar{F}_\lambda^{-1}\left(\frac{c/\mu}{p + h/\gamma}\right). \tag{8}$$

For ease of exposition in the foreground of the paper, we make the following technical assumption on the arrival rates. (Appendix A contains an extension of our main theoretical results that are derived under a more general yet slightly less eye pleasing condition.)

Assumption 1. *The arrival rate for the system indexed by λ can be expressed as $\Lambda = \lambda + \sigma_\lambda X$, where X is a random variable with zero mean, unit variance, and a bounded density function.*

To further ease notation in the front matter, it will be useful to define the following convention which, among other things, allows us to suppress wherever possible the fact that we are dealing with sequences indexed by λ (the statements and proofs given in Appendix A and B will be more explicit and make no attempt to obscure this fact). Given any two non-negative finite-valued real number sequences $\{u_\lambda\}, \{v_\lambda\}$, we say that $u_\lambda \gg v_\lambda$ if u_λ/v_λ increases without bound as $\lambda \to \infty$. We say that a sequence $\{u_\lambda\}$ is *bounded away from zero* if $\liminf_{\lambda\to\infty} u_\lambda > 0$. We write $u_\lambda = \mathcal{O}(v_\lambda)$ to mean u_λ/v_λ is bounded (i.e., $\limsup_{\lambda\to\infty} u_\lambda/v_\lambda < \infty$). Writing $u_\lambda = \mathcal{O}(1)$ simply means that u_λ is bounded as $\lambda \to \infty$. Finally, we write $u_\lambda \asymp v_\lambda$ if u_λ/v_λ is bounded from above and away from zero, that is, $0 < \liminf_{\lambda\to\infty} u_\lambda/v_\lambda \leq \limsup_{\lambda\to\infty} u_\lambda/v_\lambda < \infty$.

Performance of the newsvendor-based prescription: deterministic arrival rates. We investigate the performance of the prescription (8) and how it is impacted by the interplay between arrival rate uncertainty and stochastic variability. As a baseline, we begin with the deterministic case that has been widely studied in the literature. In this case, the prescription equals λ/μ, i.e., the capacity is set to be the offered load.

Proposition 1 (**Deterministic baseline case**). *If the arrival rate is deterministic, i.e., $CV_\lambda = 0$, the capacity level $\bar{b}_\lambda = \lambda/\mu$ is $\mathcal{O}(\sqrt{\mathcal{E}_\lambda})$-optimal. That is,*

$$\Pi_\lambda(\bar{b}_\lambda) = \Pi^*_\lambda + \mathcal{O}(\sqrt{\mathcal{E}_\lambda}).$$

Proposition 1 characterizes the effects of stochastic variability and is consistent with the vast literature on capacity planning in queueing systems with deterministic arrival rates. This result states that for the deterministic case, the newsvendor prescription, which sets the capacity *exactly* equal to the offered load, has an optimality gap proportional to the square-root of the offered load

measured in Erlangs. That is, as the arrival rate increases the optimality gap increases proportional to the square-root of the arrival rate. For example, if the arrival rate roughly quadruples, the optimality gap should roughly double in magnitude. This behavior is consistent with the observations in Table 3.

Performance of the newsvendor-based prescription: uncertainty versus variability. Having established the deterministic baseline case, we now turn to a setting where there is uncertainty in the arrival rates, and study how this affects the performance of the newsvendor prescription. The following theorem, which is the main result of the paper, precisely characterizes the interaction between uncertainty and stochastic variability.

Theorem 1 (**Uncertainty versus stochastic variability**).

(a) (Uncertainty-dominated regime.) If the coefficient of variation of the arrival rate satisfies $CV_\lambda \gg 1/\sqrt{\mathcal{E}_\lambda}$*, then the newsvendor prescription* (8) *is* $\mathcal{O}(1/CV_\lambda)-$*optimal. That is,*

$$\Pi_\lambda(\bar{b}_\lambda) = \Pi^*_\lambda + \mathcal{O}(1/CV_\lambda).$$

(b) (Variability-dominated regime.) If the coefficient of variation of the arrival rate satisfies $CV_\lambda \ll 1/\sqrt{\mathcal{E}_\lambda}$*, then the newsvendor prescription* (8) *is* $\mathcal{O}(\sqrt{\mathcal{E}_\lambda})-$*optimal. That is,*

$$\Pi_\lambda(\bar{b}_\lambda) = \Pi^*_\lambda + \mathcal{O}(\sqrt{\mathcal{E}_\lambda}).$$

(c) If the coefficient of variation of the arrival rate satisfies $CV_\lambda \asymp 1/\sqrt{\mathcal{E}_\lambda}$ *(i.e., variability and uncertainty are balanced), then the newsvendor prescription* (8) *is* $\mathcal{O}(\sqrt{\mathcal{E}_\lambda})-$*optimal. That is,* $\Pi_\lambda(\bar{b}_\lambda) = \Pi^*_\lambda + \mathcal{O}(\sqrt{\mathcal{E}_\lambda})$.

Parts (a) and (b) above characterize two very different operating modes: the former is characteristic of systems in which mean demand forecasting involves errors that dominate stochastic variability; the latter case corresponds to systems where prediction accuracy is quite sharp, and hence the dominant effect is stochastic variability. Regime (c) is one where both effects are comparable in magnitude.

The performance accuracy of the newsvendor-based prescription in the variability dominated regime is similar to what one sees in the absence of uncertainty, i.e., the case studied in Proposition 1. When uncertainty dominates, the optimality gap is proportional to the reciprocal of the coefficient of variation. This makes the prescription more accurate compared to the case studied in Proposition 1. In other words, arrival rate uncertainty *improves* the performance of the newsvendor-based prescription! The intuition has already been spelled out in the previous section, though in a simpler and more restricted setting: the optimality gap is bounded by a term similar to (6),

thus for any realized arrival rate there will be an imbalance between capacity and demand. This implies that stochastic fluctuations are secondary in their effect on costs, and hence the newsvendor prescription has a cost that is within $\mathcal{O}(1/CV_\lambda)$ of optimal. The proof of the theorem makes this intuition rigorous at the level of generality studied in this section.

Best case performance of the newsvendor-based prescription: $\mathcal{O}(1)-$ optimality. The above discussion suggests that the newsvendor prescription becomes progressively more accurate, in terms of performance, as the arrival rate tends to be more uncertain. Pushing this logic to the extreme, when the coefficient of variation of the arrival rate is approximately a constant, i.e., it does not depend on the volume of work, we expect the optimality gap to be bounded regardless of the volume of work flowing through the system. The following result makes this precise.

Theorem 2 (**$\mathcal{O}(1)-$optimality**). *If the coefficient of variation of the arrival rate is bounded away from zero, then the newsvendor prescription* (8) *is $\mathcal{O}(1)-$optimal. That is,*

$$\Pi_\lambda(\bar{b}_\lambda) = \Pi^*_\lambda + \mathcal{O}(1).$$

This type of accuracy and strong notion of optimality is somewhat unexpected given the fairly crude nature of the newsvendor-based logic. To understand what is driving this order-1 accuracy, the reader should refer back to the previous section. Recapping the key ideas, note that for any fixed capacity level the realized arrival rate will result in the system being either: underloaded (offered load lower than capacity); overloaded (offered load exceeding capacity); or critically loaded (offered load approximately equal to capacity). In the underloaded and overloaded cases, stochastic fluctuations become secondary effects: these are precisely the settings where the exponential term in (6) dominates. (See also Bassamboo and Randhawa (2009) for a similar observation in overloaded $M/M/N+G$ systems with deterministic arrival rates.) The critically loaded cases will result in degraded performance, but under Assumption 1 for any reasonable uncertainty and sufficiently large λ, this occurs with low probability. This is the key observation explaining the remarkably accurate performance of the newsvendor prescription. In contrast, in cases where the arrival rate can be predicted with high accuracy, then, roughly speaking, for almost all realizations of the arrival rate the system is critically loaded and Proposition 1 applies; this is the variability-dominated regime.

Interpretation of the newsvendor prescription and relation to the square-root rule. The newsvendor-based prescription can be thought of as consisting of a component catering to the mean demand or offered load $\mathcal{E}_\lambda$, and a safety capacity component to hedge against uncertainty. This decomposition is made precise in the following result.

Proposition 2 (**Newsvendor prescription decomposed into base and safety capacities**). *If the newsvendor-based capacity prescription $\bar{b}_\lambda$ in* (8) *is bounded away from zero, then*

$$\bar{b}_\lambda = \mathcal{E}_\lambda + \mathcal{O}\left(\mathcal{E}_\lambda CV_\lambda\right).$$

This result shows that the capacity prescription consists of a *base capacity* equal to the offered load $\mathcal{E}_\lambda$, and a *safety capacity* that is proportional to $\mathcal{E}_\lambda CV_\lambda$, i.e., the product of the offered load and the coefficient of variation of the arrival rate (which equals σ_λ/μ). As an illustration, consider the case where the standard deviation of the arrival rate is of the form λ^α for $0 < \alpha < 1$, that is, the arrival rate can be expressed as follows: $\Lambda = \lambda + \lambda^\alpha X$. In this case, the capacity prescription (8) equals $\mathcal{E}_\lambda + \mathcal{O}(\lambda^\alpha)$. This should be contrasted with conventional square-root staffing methods which prescribe investing in a base capacity level equal to the offered load and a safety capacity that is proportional to the *square-root* of the offered load. Clearly when $\alpha > 1/2$, the uncertainty hedge dominates the safety capacity considerations. Recent work of Maman (2009, p. 28) studies such models of arrival rate uncertainty and analyzes capacity levels of the form $\lambda + \mathcal{O}(\lambda^\alpha)$, looking at their impact on expected queue-lengths, probability of waiting and other related performance indicators (though it does not study any capacity optimization problem). That work also finds empirical support for the case of $\alpha > 1/2$; what we refer to as the uncertainty-dominated regime.

Main qualitative insight revisited. Theorem 1 provides justification for the rule of thumb described in the previous section: for a given system compute the coefficient of variation of the arrival rate and compare it to the reciprocal of the square-root of the mean load (measured in Erlangs). In the uncertainty-dominated regime, where the coefficient of variation dominates, the newsvendor based prescription is guaranteed to have an excellent performance. In the opposite case, i.e., in the variability-dominated regimes, the newsvendor prescription still performs reasonably well (see numerical results in previous section), but it is here that extra safety capacity may improve performance further by suitably hedging stochastic variability.

6 Robustness of Results

The goal in this section is to establish the robustness of the results derived thus far. In particular, we will develop bounds on how much one can deviate from the newsvendor capacity prescription without any significant loss in performance. Then, we discuss how the newsvendor-based approach can be applied to systems with general service and abandonment distributions, and illustrate its performance using numerical examples.

6.1 Sensitivity of the newsvendor-based prescription

So far we have been assuming that the prescription (3) can be implemented as is, whereas in practice numerous constraints might introduce slight deviations in the way it is put in place. (Examples include shift constraints, training requirements to be completed by the employees, etc.) To this end, the key observation is that arrival rate uncertainty contributes to "flattening" the objective

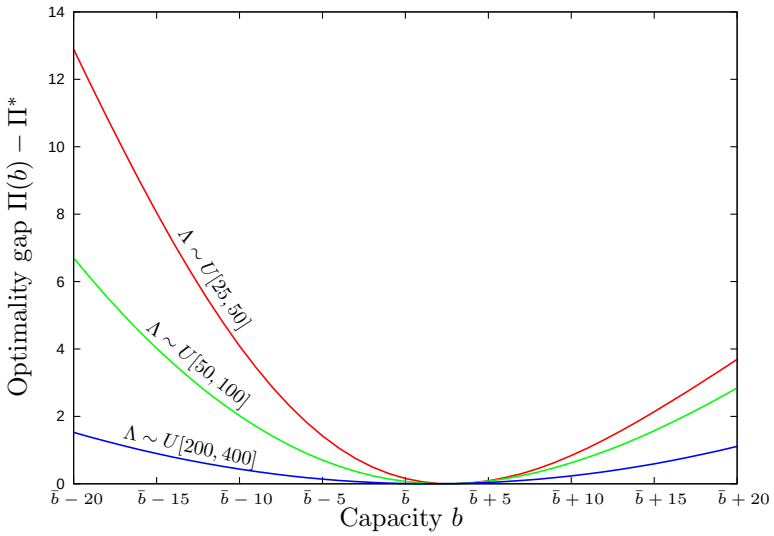

Figure 1: Robustness of the newsvendor prescription $\bar{b}$.

function, and hence small deviations from the newsvendor-based prescription do not lead to significant degradation in performance. This is illustrated in Figure 1 in which we study the performance of different capacity levels for arrival rates $\Lambda \sim U[25, 50], U[50, 100]$, and $U[200, 400]$. Notice that indeed for large mean arrival rates, the objective function is flatter around the prescription. Harrison and Zeevi (2005) have also noted such flatness in their numerical studies, but did not provide supporting theory.

A closer inspection of Figure 1 reveals that for capacity levels that are within approximately $\sqrt{\mathcal{E}_\lambda}$ of the prescription, the performance deterioration is minimal. This latter observation has a theoretical basis that is made precise in the following result:

Proposition 3. *If $\bar{b}_\lambda$ is bounded away from zero, then for any perturbation $|s_\lambda| = \mathcal{O}(\sqrt{\mathcal{E}_\lambda})$, the corresponding cost is $\mathcal{O}\left(\min\left\{\frac{1}{CV_\lambda}, \sqrt{\mathcal{E}_\lambda}\right\}\right)$-optimal. That is, for any staffing level $b_\lambda = \bar{b}_\lambda + s_\lambda$,*

$$\Pi_\lambda(b_\lambda) = \Pi^*_\lambda + \mathcal{O}\left(\min\left\{\frac{1}{CV_\lambda}, \sqrt{\mathcal{E}_\lambda}\right\}\right).$$

Comparing this result with Theorem 1, we find that deviations from the prescription that are of the order of square-root of the system load do not adversely affect the order of the optimality gap. This provides the system manager some leeway in selecting an appropriate capacity levels (so as to ensure near-optimal performance).

6.2 The case of general service and abandonment distributions

Problem formulation. We extend the basic model by considering the case where the service time and abandonment distributions need not be exponential. This setting is less amenable to

analysis and we will only derive the corresponding newsvendor-based prescription and investigate its performance via numerical experiments.

Let F and G denote the service and impatience time (or abandonment) distributions, respectively. We assume that the abandonment distribution has a density g which is strictly positive on $[0, \infty)$. We now characterize the objective function. Fix the capacity at some level b and denote the number of customers in the system in steady state by $N(b)$. Then, as in the Markovian system, the customer holding cost equals $h\mathbb{E}[N(b) - b]^+$. Let $\alpha(b)$ denote the mean number of customer abandonments per unit time at capacity level b. Analogous to (1), the capacity planning problem can be stated as follows:

$$\min_{b \geq 0} h\mathbb{E}[N(b) - b]^+ + p\alpha(b) + cb. \tag{9}$$

An approximate solution. We use a fluid-based approach to compute a capacity prescription for this system. As in the exponential distribution case, the net rate of customer abandonments can be approximated as $\alpha(b) \approx \mathbb{E}[\Lambda - \mu b]^+$. However, approximating the queue-length requires more care, and the rate-based analysis described earlier does not suffice. Whitt (2006*a*) undertakes such an analysis, and finds that the queue-length can be approximated as follows: $\mathbb{E}[N(b) - b|\Lambda]^+ \approx \frac{1}{\gamma}\Lambda\hat{G}([1 - \mu b/\Lambda]^+)$; where $\hat{G}(\cdot) \equiv G_e \circ G^{-1} = G_e(G^{-1}(\cdot))$ with $G_e(x) = \gamma \int_0^x \bar{G}(y)dy$ denoting the stationary excess cdf and G^{-1} denoting the inverse of the distribution function G. Note that this approximation depends on the service time distribution only via its mean, while it depends on the entire impatience distribution.

Using these approximations for the expected queue-length and the rate of customer abandonments in (9), we obtain the following:

$$\min_{b \geq 0} \Pi(b) \approx \min_{b \geq 0} \left[\frac{h}{\gamma} \mathbb{E}\left[\Lambda \hat{G}([1 - \mu b/\Lambda]^+) \right] + p\mathbb{E}[\Lambda - \mu b]^+ + cb \right]. \tag{10}$$

Let $\bar{b}$ denote a minimizer of the above. Notice that this optimization problem reduces to (2) for exponentially distributed service and impatience times. Further, it depends on the service time distribution only via its mean. Thus, if the impatience distribution is exponential, we recover the optimization problem (2) and the corresponding prescription equals (3).

Numerical study of accuracy. We next perform a numerical study to investigate the accuracy of the prescription $\bar{b}$, and the impact of arrival rate uncertainty on its performance. For this study, we consider four distributions, the unit mean lognormal distribution (with a variance of 4) and the unit mean Erlang-2 distribution for the service times, and the lognormal distribution (with mean 1/3 and variance 4/9) and Erlang-2 distribution with mean 1/3 for the abandonment times. (These distributions were also used in the numerical study in Whitt (2006*a*); see Brown et al. (2005) for some empirical evidence.) The cost parameters are set equal to those in the experiments for the Markovian system, that is, $c = 1/3$, $h = 1$ and $p = 1$.

Arrival rate		Optimal solution		Prescription		Difference		
Distribution	CV (%)	b^*	Π^*	$\bar{b}$	$\Pi(\bar{b})$	$\lvert b^* - \bar{b}\rvert$	$\Pi(\bar{b}) - \Pi^*$	$\frac{(\Pi(\bar{b})-\Pi^*)}{\Pi^*}$
Abandonment distribution is Erlang and Service distribution is Lognormal								
$\Lambda \sim$ U[0,300]	57.7%	234	90.49	237	90.54	3	0.05	0.06%
$\Lambda \sim$ U[125,175]	9.6%	168	59.56	168	59.56	0	0	0%
$\Lambda \sim$ U[145,155]	3.8%	163	56.57	154	58.13	9	1.56	2.8%
$\Lambda = 150$	0	163	56.23	150	59.79	13	3.55	6.3%
Abandonment and Service distributions are both Erlang								
$\Lambda \sim$ U[0,300]	57.7%	235	90.44	237	90.45	2	0.007	0.008%
$\Lambda \sim$ U[125,175]	9.6%	168	59.21	168	59.21	0	0	0%
$\Lambda \sim$ U[145,155]	3.8%	161	56.13	154	57.38	7	1.25	2.2%
$\Lambda = 150$	0	161	55.91	150	59.13	11	3.22	5.8%

Table 5: **Performance of the newsvendor-based prescription under general abandonment and service time distributions.** The abandonment distribution is Erlang-2 (mean 1/3) and the service distribution is Erlang-2 (mean 1) and Lognormal (mean 1 and variance 4). The table compares the optimal (b^*) and prescribed ($\bar{b}$) capacity levels along with their respective performance.

For each experiment, we optimize (9) and study the performance of the prescription $\bar{b}$ for four different arrival rate distributions: $\Lambda \sim U[0, 300]$, $U[125, 175]$, $U[145, 155]$, and deterministic $\Lambda = 150$. In each scenario, the optimal capacity is computed by estimating the cost at different capacity levels and performing a search over a large set of capacity levels (as before). Discrete event simulations are used to estimate the mean queue-length and mean number of abandonments. For the case where the arrival rate is distributed according to a uniform distribution, the cost is estimated by a numerical integration in the sense that we divide the interval into 100 equally spaced points, and estimate the cost with arrival rate corresponding to each point on the grid, and then take a weighted sum. For a fixed arrival rate, the performance parameters are estimated using an average over 10 runs, where each run consists of $450,000/\mathbb{E}\Lambda$ time units. For the case with a deterministic arrival rate, the performance is estimated using an average over 10 runs, where each run now consists of $1,500,000/\mathbb{E}\Lambda$ time units. The half-width of the 95% confidence interval was observed to be less than 0.5% in each setting. The results are displayed in Tables 5 and 6 for Erlang-2 and lognormal abandonment distributions respectively. Note that the prescription $\bar{b}$ does not depend on the service distribution.

These results are very similar to those obtained for the Markovian system in Table 4. As in

Arrival rate		Optimal solution		Prescription		Difference		
Distribution	CV (%)	b^*	Π^*	$\bar{b}$	$\Pi(\bar{b})$	$\lvert b^* - \bar{b}\rvert$	$\Pi(\bar{b}) - \Pi^*$	$\frac{(\Pi(\bar{b})-\Pi^*)}{\Pi^*}$
Abandonment and Service distributions are both Lognormal								
$\Lambda \sim$ U[0,300]	57.7%	211	85.70	211	85.70	0	0	0%
$\Lambda \sim$ U[125,175]	9.6%	164	58.93	160	59.03	4	0.10	0.2%
$\Lambda \sim$ U[145,155]	3.8%	161	56.68	152	57.82	9	1.14	2%
$\Lambda = 150$	0	161	56.45	150	57.98	11	1.53	2.7%
Abandonment distribution is Lognormal and Service distributions is Erlang								
$\Lambda \sim$ U[0,300]	57.7%	210	85.64	211	85.65	1	0.01	0.01%
$\Lambda \sim$ U[125,175]	9.6%	163	58.81	160	58.88	3	0.07	0.1%
$\Lambda \sim$ U[145,155]	3.8%	161	56.56	152	57.62	9	1.06	1.9%
$\Lambda = 150$	0	159	56.43	150	57.99	9	1.56	2.8%

Table 6: **Performance of the newsvendor-based prescription under general abandonment and service time distributions.** The abandonment distribution is Lognormal (mean 1/3 and variance 4/9) and the service distribution is Erlang-2 (mean 1) and Lognormal (mean 1 and variance 4). The table compares the optimal (b^*) and prescribed ($\bar{b}$) capacity levels along with their respective performance.

Table 4, we find that when the CV exceeds $1/\sqrt{\mathcal{E}} \approx 8\%$, which is indicative of the uncertainty-dominated regime, the newsvendor-based prescription performs exceptionally well, and the accuracy of the prescription and its performance deteriorates as the uncertainty in the arrival rate decreases.

7 Discussion and Future Work

This paper studies a capacity sizing problem in a queueing system with parameter uncertainty. We focus on systems with a single class of customers and a single pool of servers, where the arrival rate of the customers is the uncertain parameter. We believe that the results derived in this paper are pertinent to many other settings beyond those discussed thus far, and we now discuss such applications and avenues for future work.

(a) **Dynamic control implications in queueing networks:** An important direction for future work is to analyze a more general network consisting of multiple customer classes and multiple server pools. In such systems, apart from deciding capacity levels, one also needs to determine routing rules that dictate how the arriving customers will be handled by the servers. In follow-

up work, Bassamboo, Randhawa and Zeevi (2009), we construct a simple *quasi-static* control policy that is driven by similar logic to the one that underlies the capacity prescriptions studied in the present paper. In particular, when the alluded policy is paired with the newsvendor-based capacity prescription, the resulting performance is nearly optimal in the uncertainty-dominated regime. This suggests that uncertainty in the arrival rate plays a significant role not only in simplifying capacity prescriptions, but also in solving optimal control problems.

(b) **Single server systems:** With a view towards call center applications, this paper focuses on many server systems where the decision variable is the number of operators/servers. One can construct a similar capacity sizing problem in the context of a single server system, with the speed of the server as the capacity decision. In this case, the expected queue-length equals $\mathbb{E}[N-1]^+$ (as there is only one server) and the cost of capacity is $c\mu$, where μ is the server's processing rate. Thus, the optimization problem is $\min\{(p\gamma+h)\mathbb{E}[N-1]^+ + c\mu : \mu \geq 0\}$, which is analogous to (1). Under parameter uncertainty, the newsvendor-based capacity prescription for this system is again analogous to (3) and given by

$$\bar{\mu} = \bar{F}^{-1}\left(\frac{c}{p+h/\gamma}\right).$$

There are two pieces of evidence that suggest that this prescription would perform extremely well. First, are numerical experiments reported on in Appendix C.2. Second, recent work Bassamboo and Randhawa (2009, Section 5) proves that for the single server system operating in the overloaded regime, the fluid model is $O(1)$-accurate. A detailed analysis of this system with the aim of deriving an analog of Theorem 1 is beyond the scope of the present paper.

(c) **Connections to inventory systems:** This paper studies a newsvendor-based prescription for capacity planning in queueing systems, and thus there is an immediate connection to static inventory systems. Furthermore, a similar analysis can be performed for dynamic inventory systems where product demand arrives in form of a Poisson process (with uncertain rate) and there are exogenous delivery lead times. In such systems, the optimization problem is to find the base stock level that minimizes expected holding and backorder costs (see Chapters 6 and 7 of Zipkin 2000).

Analyzing inventory systems with lead times requires characterizing the customer demand that arrives during the length of one lead time; denote this quantity by D. Assuming sufficient capacity, these systems are modeled as $M/G/\infty$ queueing systems. Thus, for a realized customer arrival rate λ, in steady-state D has a Poisson distribution with mean λ times the mean lead time. Denoting the base stock level by b, one obtains the mean backorder level as $\mathbb{E}[D-b]^+$ and the firm's optimization problem can then be written as $\min\{\alpha\mathbb{E}[D-b]^+ + \beta b :$

$b \geq 0\}$, for some coefficients $\alpha, \beta \geq 0$ that depend on the per-unit backordering and holding costs. This optimization problem is identical to the one studied in this paper, (see (2)) and thus the analysis in our paper extends to this setting. In particular, the newsvendor prescription derived in (3) will be subject to an analog of Theorem 1.

A base stock policy entails placing orders/producing units every time the inventory level drops below the base stock level. In the presence of ordering costs or set up times, it may be optimal to produce multiple units in a batch if the inventory level decreases. Such policies are referred to as (r, q) policies in Zipkin (2000). Analyzing these policies is an interesting avenue for future work.

(d) **Lack of knowledge of demand distribution:** In this paper, we assume that the demand arrival rate distribution is known to the system manager. If the demand distribution needs to be estimated from demand data, then there is a further source of error to be dealt with; see for example Bassamboo and Zeevi (2009) for a possible data-driven approach to that problem. Our view of uncertainty in arrival rates is closely tied in to forecasting errors and their implications on system design and management. A careful characterization of the interplay between forecasting errors and the performance of the corresponding newsvendor prescription makes for an important future study and this paper hopefully provides a good starting point for such pursuits.

References

Aksin, Z., Armony, M. and Mehrotra, V. (2007), 'The modern call center: A multi-disciplinary perspective on operations management research', *Production and Operations Management* **16**(6), 665–688.

Avramidis, A., Deslauriers, A. and L'Ecuyer, P. (2004), 'Modeling daily arrivals to a telephone call center', *Management Science* pp. 896–908.

Bassamboo, A., Harrison, J. M. and Zeevi, A. (2006), 'Design and control of a large call center: Asymptotic analysis of an LP-based method', *Operations Research* **54**(3), 419–435.

Bassamboo, A. and Randhawa, R. S. (2009), 'Accuracy of Fluid Models for Capacity Planning in Queueing Systems', *Working paper* .

Bassamboo, A., Randhawa, R. S. and Zeevi, A. (2009), 'Capacity Planning and Scheduling in Queueing Systems under Arrival Rate Uncertainty', *In progress* .

Bassamboo, A. and Zeevi, A. (2009), 'On a data-driven method for staffing large call centers', *Operations Research* **57**(3), 714–726.

Borst, S., Mandelbaum, A. and Reiman, M. I. (2004), 'Dimensioning large call centers', *Operations Research* **52**, 17–34.

Brown, L., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S. and Zhao, L. (2005), 'Statistical Analysis of a Telephone Call Center', *Journal of the American Statistical Association* **100**(469), 36–50.

Chen, B. and Henderson, S. (2001), 'Two issues in setting call centre staffing levels', *Annals of Operations Research* **108**(1), 175–192.

Erlang, A. K. (1917), 'Solution of some problems in the theory of probabilities of significance in automatic telephone exchanges', *Electroteknikeren* **13**, 5–13.

Gans, N., Koole, G. and Mandelbaum, A. (2003), 'Telephone call centers: Tutorial, review, and research prospects', *Manufacturing and Service Operations Management* **5**(2), 79–141.

Garnett, O., Mandelbaum, A. and Reiman, M. (2002), 'Designing a call center with impatient customers', *Manufacturing and Service Operations Management* **4**(3), 208–227.

Gurvich, I., Armony, M. and Mandelbaum, A. (2008), 'Service level differentiation in call centers with fully flexible servers', *Management Science* **54**(2), 279–294.

Gurvich, I. and Whitt, W. (2009), 'Service-level differentiation in many-server service systems: a solution based on fixed-queue-ratio routing', *Operations Research, forthcoming* .

Halfin, S. and Whitt, W. (1981), 'Heavy-traffic limits for queues with many exponential servers', *Operations Research* **29**, 567–588.

Harrison, J. and Zeevi, A. (2005), 'A method for staffing large call centers based on stochastic fluid models', *Manufacturing & Service Operations Management* **7**(1), 20–36.

Jennings, O., Mandelbaum, A., Massey, W. and Whitt, W. (1996), 'Server staffing to meet time-varying demand', *Management Science* **42**(10), 1383–1394.

Jongbloed, G. and Koole, G. (2001), 'Managing uncertainty in call centres using Poisson mixtures', *Applied Stochastic Models in Business and Industry* **17**(4), 307–318.

Kolesar, P. and Green, L. (1998), 'Insights on service system design from a normal approximation to Erlang's delay formula', *Production and Operations Management* **7**(3), 282–293.

Maman, S. (2009), Uncertainty in the Demand for Service: The Case of Call Centers and Emergency Departments, PhD thesis, Field of Statistics, Technion - Israel Institute of Technology, Haifa, Israel.

Mandelbaum, A. and Zeltyn, S. (2009), 'Staffing many-server queues with impatient customers: constraint satisfaction in call centers', *Working paper* .

Motwani, R. and Prabhakar, R. (1995), *Randomized Algorithms*, Cambridge University Press, Cambridge, U.K.

Robbins, T. and Harrison, T. (2007), 'Call center scheduling with uncertain arrivals and global service level agreements'. Working paper.

Steckley, S. G., Henderson, S. G. and Mehrotra, V. (2009), ' Forecast errors in service systems', *Probability in the Engineering and Informational Sciences* **23**(2), 305–332.

Whitt, W. (1981), 'Comparing counting processes and queues', *Advances in Applied Probability* **13**, 207–220.

Whitt, W. (1999), 'Dynamic staffing in a telephone call center aiming to immediately answer all calls', *Operations Research Letters* **24**(5), 205–212.

Whitt, W. (2006*a*), 'Fluid models for multiserver queues with abandonments', *Operations Research* **54**(1), 37–54.

Whitt, W. (2006*b*), 'Staffing a call center with uncertain arrival rate and absenteeism', *Production and Operations Management* **15**(1), 88–102.

Zipkin, P. H. (2000), *Foundations of inventory management*, McGraw-Hill, New York.

A Proofs of Theorems

We state and prove a general version of Theorem 1 below. Theorem 2 follows from Theorem 1 by noting that if the coefficient of variation is bounded away from zero, we have $1/CV_\lambda = \mathcal{O}(1)$.

A generalization of Theorem 1. We consider a sequence of systems indexed by their mean arrival rate λ. Consistent with the notation introduced in the paper, we denote the distribution of the random arrival rate by F_λ, its density by f_λ, and its standard deviation by σ_λ. Focusing on the optimization problem in (7), and the corresponding prescription $\bar{b}_\lambda$ in (8), we next present a general version of Theorem 1 that does not impose any assumptions on the sequence of arrival rates. For this purpose, we introduce the following function:

$$M_\lambda(y) := \sup_{x\in[y-\sqrt{\lambda}\log\lambda, y+\sqrt{\lambda}\log\lambda]} \Big\{\lambda f_\lambda(x) \vee 1\Big\}, \text{ for } y \in \mathbb{R}_+. \tag{11}$$

We then have the following result whose proof will be given shortly thereafter.

Theorem 3 (**Generalization of Theorem 1**).

(a) If $1/M_\lambda(\bar{b}_\lambda\mu) \gg 1/\sqrt{\mathcal{E}_\lambda}$, *then the newsvendor prescription* (8) *is* $\mathcal{O}(M_\lambda(\bar{b}_\lambda\mu))-$*optimal. That is,* $\Pi_\lambda(\bar{b}_\lambda) = \Pi^*_\lambda + \mathcal{O}(M_\lambda(\bar{b}_\lambda\mu))$.

(b) If $1/M_\lambda(\bar{b}_\lambda\mu) \ll 1/\sqrt{\mathcal{E}_\lambda}$, *then the newsvendor prescription* (8) *is* $\mathcal{O}(\sqrt{\mathcal{E}_\lambda})-$*optimal. That is,* $\Pi_\lambda(\bar{b}_\lambda) = \Pi^*_\lambda + \mathcal{O}(\sqrt{\mathcal{E}_\lambda})$.

(c) If $1/M_\lambda(\bar{b}_\lambda\mu) \asymp 1/\sqrt{\mathcal{E}_\lambda}$, *then the newsvendor prescription* (8) *is* $\mathcal{O}(\sqrt{\mathcal{E}_\lambda})-$*optimal. That is,* $\Pi_\lambda(\bar{b}_\lambda) = \Pi^*_\lambda + \mathcal{O}(\sqrt{\mathcal{E}_\lambda})$.

Discussion and Relation to Theorem 1. Theorem 3 states that the accuracy of the prescription $\bar{b}_\lambda$ obtained in (8) depends on the distribution of the arrival rate, as well as the prescription $\bar{b}_\lambda$ itself. In particular, the accuracy of the fluid-based prescription depends on the behavior of the density of the arrival rate in the vicinity of the prescribed capacity, as captured by the function M_λ. Note that if Assumption 1 holds, then X has a bounded density. This implies that $f_\lambda(x) = \mathcal{O}(1/\sigma_\lambda)$, hence $M_\lambda(\bar{b}_\lambda\mu) = \mathcal{O}(1/CV_\lambda)$. So, indeed Theorem 3 reduces to Theorem 1 under Assumption 1.

Proof of Theorem 3. We first prove the result for the case $\mu = \gamma$, and then use that to address the case of $\mu \neq \gamma$. Put $\bar{q}_\lambda(b) := \mathbb{E}[\Lambda - \mu b]^+/\gamma$ and $\bar{\Pi}_\lambda(b) := (p\gamma + h)\bar{q}_\lambda(b) + cb$ for any capacity level b.

Step 1. The case of $\mu = \gamma$.

The following lemma derives a bound on the expected queue-length, and is the main driver behind the proof of the theorem.

Lemma 1. *If* $\mu = \gamma$, *then for any non-negative real-valued sequence of capacity levels* $\{b_\lambda\}$, *the expected queue length is lower bounded as* $\mathbb{E}[N_\lambda - b_\lambda]^+ \geq \bar{q}_\lambda(b_\lambda)$. *Further, if* $b_\lambda = \mathcal{O}(\lambda)$, *then the expected queue length is upper bounded as* $\mathbb{E}[N_\lambda - b_\lambda]^+ \leq \bar{q}_\lambda(b_\lambda) + \mathcal{O}(\min\left\{M_\lambda(b_\lambda\mu), \sqrt{\mathcal{E}_\lambda}\right\})$.

For ease of exposition, the proof of this and other lemmas are postponed to just after this proof. Lemma 1 immediately gives us the following relation for any sequence of capacity levels $\{b_\lambda\}$ such that $b_\lambda = \mathcal{O}(\lambda)$:

$$\bar{\Pi}_\lambda(b_\lambda) \leq \Pi_\lambda(b_\lambda) \leq \bar{\Pi}_\lambda(b_\lambda) + \mathcal{O}(\min\{M_\lambda(\bar{b}_\lambda\mu), \sqrt{\mathcal{E}_\lambda}\}). \tag{12}$$

It then follows that

$$\begin{aligned}
\Pi_\lambda(\bar{b}_\lambda) &\overset{(a)}{\leq} \bar{\Pi}_\lambda(\bar{b}_\lambda) + \mathcal{O}(\min\{M_\lambda(\bar{b}_\lambda\mu), \sqrt{\mathcal{E}_\lambda}\}) \\
&\overset{(b)}{\leq} \bar{\Pi}_\lambda(b^*_\lambda) + \mathcal{O}(\min\{M_\lambda(\bar{b}_\lambda\mu), \sqrt{\mathcal{E}_\lambda}\}) \\
&\overset{(c)}{\leq} \Pi_\lambda(b^*_\lambda) + \mathcal{O}(\min\{M_\lambda(\bar{b}_\lambda\mu), \sqrt{\mathcal{E}_\lambda}\}),
\end{aligned}$$

where: (a) follows by applying the upper bound in (12) for $\bar{b}_\lambda$ noting that $\bar{b}_\lambda = \mathcal{O}(\lambda)$; (b) follows by noting that $\bar{b}_\lambda$ minimizes $\bar{\Pi}_\lambda$; and finally (c) follows by applying the lower bound in (12) for b^*_λ. This completes the proof of Theorem 3 for the case $\mu = \gamma$.

Step 2. The case of $\mu \neq \gamma$.

We begin by rewriting the objective function in terms of the rate at which customers abandon, $A_\lambda(b) = \gamma \mathbb{E}[N_\lambda - b]^+$, rather than the queue-length. Then, the original optimization problem can be written as

$$\min_{b \geq 0} (p + h/\gamma) A_\lambda(b) + cb. \tag{13}$$

We first consider the case $\mu > \gamma$. Consider a sequence of systems with the same parameters as that under consideration except that its abandonment rate and service rate equal μ; that is, systems in this sequence have a higher abandonment rate compared to the original sequence. We refer to this sequence as Sequence II, and our original sequence as Sequence I. We will use the superscript I, II to make explicit the systems to which various quantities are associated with. For systems in Sequence II, we choose the cost parameters to be the following: the server cost remains c; the penalty cost remains p; and the holding cost is set to $h\mu/\gamma$. That is, systems in Sequence II have $\mu^{\text{II}} = \mu$, $\gamma^{\text{II}} = \mu$, $c^{\text{II}} = c$, $p^{\text{II}} = p$, and $h^{\text{II}} = h\mu/\gamma$. This choice leads to the following optimization problem analogous to (13):

$$\min_{b \geq 0} (p^{\text{II}} + h^{\text{II}}/\gamma^{\text{II}}) A^{\text{II}}_\lambda(b) + c^{\text{II}} b = \min_{b \geq 0} (p + h/\gamma) A^{\text{II}}_\lambda(b) + cb.$$

Comparing the underlying Markov chains of systems in sequences I and II for the same mean arrival rate λ, we observe that the number of customers in system in steady state is (stochastically) greater in the system in Sequence I, i.e., $N^{\text{II}}_\lambda(b) \leq_{st} N^{\text{I}}_\lambda(b)$ for $\lambda > 0$.* Thus, we obtain $\mathbb{E}[b - N^{\text{II}}_\lambda(b)]^+ \geq \mathbb{E}[b - N^{\text{I}}_\lambda(b)]^+$ for $\lambda > 0$. That is, the expected number of servers that are idle in steady state is larger in the system in Sequence II. Noting that the rate at which customers abandon can also be written as

$$A^\ell_\lambda(b) = \mathbb{E}[\Lambda - \mu(b - [b - N^\ell_\lambda(b)]^+)], \text{ for } \ell \in \{\text{I}, \text{ II}\},$$

we obtain $A^{\text{II}}_\lambda(b) \geq A^{\text{I}}_\lambda(b)$ for all $\lambda > 0$ and any capacity level $b \geq 0$.

This argument gives us the bound $\Pi^{\text{I}}_\lambda(b) \leq \Pi^{\text{II}}_\lambda(b)$ for all $\lambda > 0$ and $b \geq 0$. As the abandonment rate equals the service rate for systems in Sequence II, we can apply the results of Theorem 3 for the case $\mu = \gamma$ as established in Step 1 to Sequence II to obtain that $\min_{b \geq 0} \Pi^{\text{II}}_\lambda(b) \leq \min_{b \geq 0} \bar{\Pi}^{\text{II}}_\lambda(b) + \mathcal{O}(\min\{M_\lambda(\bar{b}_\lambda \mu), \sqrt{\mathcal{E}_\lambda}\})$, where $\bar{\Pi}^{\text{II}}_\lambda(b) = (p + h/\gamma)\bar{q}_\lambda(b) + cb$. Note that $\bar{\Pi}^{\text{II}}_\lambda(\cdot) = \bar{\Pi}_\lambda(\cdot)$ and we obtain

$$\Pi^{\text{I}}_\lambda(b^*_\lambda) = \min_{b \geq 0} \Pi^{\text{I}}_\lambda(b) \leq \min_{b \geq 0} \Pi^{\text{II}}_\lambda(b) \leq \bar{\Pi}_\lambda(\bar{b}_\lambda) + \mathcal{O}(\min\{M_\lambda(\bar{b}_\lambda \mu), \sqrt{\mathcal{E}_\lambda}\}). \tag{14}$$

*For any two non-negative real valued random variables X and Y, we say $X \geq_{st} Y$ if $\mathbb{P}(X > x) \geq \mathbb{P}(Y > x)$ for all $x \geq 0$.

We now compute the lower bound. To do so, we construct Sequence III that consists of systems with the same arrival rate distribution and abandonment rate as those in Sequence I. However, the service rate now equals the abandonment rate. That is, both the service rate and abandonment rate now equal γ. The cost of a server in these systems is $c\gamma/\mu$, and the holding and penalty costs are the same as in Sequence I. That is, systems in Sequence III have $\mu^{\mathrm{III}} = \gamma$, $\gamma^{\mathrm{III}} = \gamma$, $c^{\mathrm{III}} = c\gamma/\mu$, $p^{\mathrm{III}} = p$, and $h^{\mathrm{III}} = h$. Thus, the optimization problem is given by:

$$\min_{b\geq 0}(p^{\mathrm{III}} + h^{\mathrm{III}}/\gamma^{\mathrm{III}})A_\lambda^{\mathrm{III}}(b) + c^{\mathrm{III}}b = \min_{b\geq 0}(p + h/\gamma)A_\lambda^{\mathrm{III}}(b) + \frac{c\gamma}{\mu}b. \tag{15}$$

We now use the following result which follows from a straightforward Markov chain analysis.

Lemma 2. *The rate at which customers abandon $A_\lambda(k)$ in an $M/M/k+M$ system with service rate α/k for some $\alpha > 0$ and $k \in \mathbb{N}$ is decreasing in k.*

This result gives us the bound $A_\lambda^{\mathrm{III}}(b\mu/\gamma) \leq A_\lambda^{\mathrm{I}}(b)$ for all $\lambda > 0$ and $b \geq 0$. Thus, we have $\Pi_\lambda^{\mathrm{III}}(b\mu/\gamma) \leq \Pi^{\mathrm{I}}(b)$. Applying Theorem 3 for the case $\mu = \gamma$ as established in Step 1 to Sequence III, we obtain

$$\min_{b\geq 0}\Pi_\lambda^{\mathrm{III}}(b) = \bar{\Pi}_\lambda^{\mathrm{III}}(\bar{b}_\lambda^{\mathrm{III}}) + \mathcal{O}(\min\{M_\lambda(\bar{b}_\lambda^{\mathrm{III}}\gamma), \sqrt{\mathcal{E}_\lambda}\}),$$

where $\bar{\Pi}_\lambda^{\mathrm{III}}(b\mu/\gamma) = (p + h/\gamma)\mathbb{E}[\Lambda - \mu b]^+ + cb = \bar{\Pi}(b)$, and $\bar{b}_\lambda^{\mathrm{III}} = \arg\min_{b\geq 0}\{\bar{\Pi}_\lambda^{\mathrm{III}}(b)\} = \bar{b}_\lambda\mu/\gamma$, where $\bar{b}_\lambda = \arg\min_{b\geq 0}\{\bar{\Pi}_\lambda(b)\}$. Thus, we have

$$\Pi_\lambda^{\mathrm{I}}(b_\lambda^*) \geq \Pi_\lambda^{\mathrm{III}}(b_\lambda^*\mu/\gamma) \geq \min_{b\geq 0}\Pi_\lambda^{\mathrm{III}}(b\mu/\gamma) \geq \bar{\Pi}_\lambda^{\mathrm{III}}(\bar{b}_\lambda^{\mathrm{III}}) = \bar{\Pi}_\lambda^{\mathrm{III}}(\bar{b}_\lambda\mu/\gamma) = \bar{\Pi}_\lambda(\bar{b}_\lambda). \tag{16}$$

The result then follows by combining (14) and (16).

The case $\mu < \gamma$ follows in a similar fashion. Systems in Sequence II now have a lower abandonment rate (which equals μ). Arguing analogous to before, we obtain that for the same capacity level b and mean arrival rate λ, the rate at which customers abandon in systems in Sequence II is lower than those in Sequence I, and hence $\Pi_\lambda^{\mathrm{II}}(b) \leq \Pi_\lambda^{\mathrm{I}}(b)$. Thus, we obtain

$$\Pi_\lambda^{\mathrm{I}}(b_\lambda^*) = \min_{b\geq 0}\Pi_\lambda^{\mathrm{I}}(b) \geq \min_{b\geq 0}\Pi_\lambda^{\mathrm{II}}(b) \geq \bar{\Pi}_\lambda^{\mathrm{II}}(\bar{b}_\lambda) = \bar{\Pi}_\lambda(\bar{b}_\lambda), \tag{17}$$

Next, we consider Sequence III to now obtain $A_\lambda^{\mathrm{III}}(b\mu/\gamma) \geq A_\lambda^{\mathrm{I}}(b)$ for all $\lambda > 0$ and $b \geq 0$. We again argue as before (with the inequality reversed) to obtain

$$\begin{aligned}\Pi_\lambda^{\mathrm{I}}(b_\lambda^*) = \min_{b\geq 0}\Pi_\lambda^{\mathrm{I}}(b) &\leq \min_{b\geq 0}\Pi_\lambda^{\mathrm{III}}(b\mu/\gamma) \\ &\overset{(a)}{=} \bar{\Pi}_\lambda^{\mathrm{III}}(\bar{b}_\lambda^{\mathrm{III}}) + \mathcal{O}(\min\{M_\lambda(\bar{b}_\lambda^{\mathrm{III}}\gamma), \sqrt{\mathcal{E}_\lambda}\}) \\ &\overset{(b)}{=} \bar{\Pi}_\lambda(\bar{b}_\lambda) + \mathcal{O}(\min\{M_\lambda(\bar{b}_\lambda\mu), \sqrt{\mathcal{E}_\lambda}\}),\end{aligned} \tag{18}$$

where: (a) follows by applying the results of Theorem 3 for the case $\mu = \gamma$ as established in Step 1 to Sequence III; and (b) follows by noting that $\bar{b}_\lambda^{\mathrm{III}} = \bar{b}_\lambda\mu/\gamma$ and $\bar{\Pi}_\lambda^{\mathrm{III}}(\bar{b}_\lambda^{\mathrm{III}}) = \bar{\Pi}_\lambda(\bar{b}_\lambda)$. The result then follows by combining (17) and (18). ■

Proof of supporting lemmas. We first state and prove the following key result that will be used in proving Lemma 1.

Lemma 3. *If $\mu = \gamma$, then for any capacity level b and realized arrival rate $\theta > 0$, denoting the number of customers in the system in steady-state by N, we have*

$$\left[\frac{\theta}{\mu} - b\right]^+ \leq \mathbb{E}[(N-b)^+ | \Lambda = \theta] \leq \left[\frac{\theta}{\mu} - b\right]^+ + \sqrt{4\pi/\mu}\sqrt{\theta}\exp\left(-\frac{\mu}{4\theta}\left(\frac{\theta}{\mu} - b\right)^2\right) + 1/\log 2.$$

Proof. Note that since $\mu = \gamma$, the steady-state number of customers in the system N conditioned on the realized arrival rate θ is independent of the number of servers and has a Poisson distribution with mean θ/μ. Thus, applying Jensen's inequality, the lower bound follows. For the upper bound, we divide the argument into two cases based on the relative ranking of b and θ/μ. For convenience, we let N^θ denote the number of customers in the system in steady state when the realized arrival rate is θ.

Case I: $b < \theta/\mu$.

We have

$$\begin{aligned}
\mathbb{E}[(N-b)^+ | \Lambda = \theta] &= \int_b^\infty \mathbb{P}(N^\theta > x)dx \\
&= \int_0^\infty \mathbb{P}(N^\theta > x)dx - \int_0^b \mathbb{P}(N^\theta > x)dx \\
&= \mathbb{E}[N^\theta] - \int_0^b [1 - \mathbb{P}(N^\theta \leq x)]dx \\
&= (\mathbb{E}[N^\theta] - b) + \int_0^b \mathbb{P}(N^\theta \leq x)dx.
\end{aligned}$$

We now use the Chernoff bound $\mathbb{P}(N^\theta \leq x) \leq \exp\left(-\frac{\mu(x-\theta/\mu)^2}{2\theta}\right)$ for all $x < \mathbb{E}[N^\theta] = \theta/\mu$ (see Chapter 4 in Motwani and Prabhakar (1995)), which yields

$$\begin{aligned}
\mathbb{E}[N^\theta - b]^+ &\leq \frac{\theta}{\mu} - b + \int_0^b e^{-\frac{\mu(x-\theta/\mu)^2}{2\theta}}dx \\
&= \frac{\theta}{\mu} - b + \sqrt{4\pi\theta/\mu}\int_0^b \sqrt{\frac{\mu}{4\pi\theta}}e^{-\frac{\mu(x-\theta/\mu)^2}{2\theta}}dx \\
&\leq \frac{\theta}{\mu} - b + \sqrt{4\pi\theta/\mu}\,\bar{\Phi}\left(-\sqrt{\frac{\mu}{2\theta}}\left(b - \frac{\theta}{\mu}\right)\right) \\
&\leq \frac{\theta}{\mu} - b + \sqrt{4\pi/\mu}\sqrt{\theta}\exp\left(-\frac{\mu}{4\theta}\left(b - \frac{\theta}{\mu}\right)^2\right).
\end{aligned}$$

The last inequality follows from the straightforward tail bound for the Normal distribution, $\bar{\Phi}(x) \leq e^{-x^2/2}$. This completes the proof.

Case II: $b > \theta/\mu$.

We again use the following Chernoff bound for a Poisson distribution,

$$\mathbb{P}(N^\theta > x) \le \exp\left(-\frac{\mu(x-\theta/\mu)^2}{4\theta}\right) + \exp\left(-(x-\theta/\mu)\log 2\right),$$

for all $x > \mathbb{E}[N^\theta] = \theta/\mu$ (see Chapter 4 in Motwani and Prabhakar (1995)). This gives us

$$\begin{aligned}\int_b^\infty \mathbb{P}(N^\theta > x)dx &\le \int_b^\infty \left[\exp\left(-\frac{\mu(x-\theta/\mu)^2}{4\theta}\right) + \exp\left(-(x-\theta/\mu)\log 2\right)\right] dx \\ &= \int_b^\infty \exp\left(-\frac{\mu(x-\theta/\mu)^2}{4\theta}\right) dx + \int_b^\infty \exp\left(-(x-\theta/\mu)\log 2\right) dx \\ &\le K + 1/\log 2,\end{aligned}$$

where we use $\int_b^\infty \exp\left(-(x-\theta/\mu)\log 2\right) dx = \exp(-(b-\theta/\mu)\log 2)/\log 2 \le 1/\log 2$ and

$$\begin{aligned}K &:= \int_b^\infty e^{-\frac{\mu(x-\theta/\mu)^2}{4\theta}} dx \\ &= \sqrt{4\pi\theta/\mu}\ \bar{\Phi}\left(\sqrt{\frac{\mu}{2\theta}}(b-\theta/\mu)\right) \\ &\le \sqrt{4\pi/\mu}\sqrt{\theta}\exp\left(-\frac{\mu}{4\theta}\left(b-\frac{\theta}{\mu}\right)^2\right),\end{aligned}$$

where the last inequality uses bound $\bar{\Phi}(x) \le e^{-x^2/2}$. This completes the proof. ■

Proof of Lemma 1. For a given realization of arrival rate $\Lambda = \theta$, let N_λ^θ denote the number of customers in the system in steady state. Note that since $\mu = \gamma$, the steady-state number of customers in the system is independent of the number of servers and has a Poisson distribution with mean θ/μ. Thus, if the capacity level is b_λ, then the queue length is given by $[N_\lambda^\theta - b_\lambda]^+$.

Noting the fact that

$$\mathbb{E}[N_\lambda^\theta - b_\lambda]^+ = \int_0^\infty \mathbb{E}[N_\lambda^\theta - b_\lambda]^+ f_\lambda(\theta)d\theta,$$

we apply Lemma 3 to obtain

$$\mathbb{E}[N_\lambda^\theta - b_\lambda]^+ \le \int_0^\infty \left[\frac{\theta}{\mu} - b_\lambda\right]^+ f_\lambda(\theta)d\theta + \int_0^\infty K_1\sqrt{\theta}\exp\left(-\frac{K_2}{\theta}\left(\frac{\theta}{\mu} - b_\lambda\right)^2\right) f_\lambda(\theta)d\theta + K_3,$$

where $K_1 = \sqrt{4\pi/\mu}$, $K_2 = \mu/4$ and $K_3 = 1/\log 2$. We now show that the second term is bounded as follows:

$$\int_0^\infty K_1\sqrt{\theta}\exp\left(-\frac{K_2}{\theta}\left(\frac{\theta}{\mu} - b_\lambda\right)^2\right) f_\lambda(\theta)d\theta = \mathcal{O}(\min\{(M_\lambda(b_\lambda\mu), \sqrt{\mathcal{E}_\lambda}\}). \tag{19}$$

To see this, start with

$$\begin{aligned}
&\int_0^\infty K_1\sqrt{\theta}\exp\left(-\frac{K_2}{\theta}\left(\frac{\theta}{\mu}-b_\lambda\right)^2\right)f_\lambda(\theta)d\theta \\
&=\int_0^{b_\lambda\mu-\sqrt{\lambda}\log\lambda} K_1\sqrt{\theta}\exp\left(-\frac{K_2}{\theta}\left(\frac{\theta}{\mu}-b_\lambda\right)^2\right)f_\lambda(\theta)d\theta \\
&+\int_{b_\lambda\mu-\sqrt{\lambda}\log\lambda}^{b_\lambda\mu+\sqrt{\lambda}\log\lambda} K_1\sqrt{\theta}\exp\left(-\frac{K_2}{\theta}\left(\frac{\theta}{\mu}-b_\lambda\right)^2\right)f_\lambda(\theta)d\theta \\
&+\int_{b_\lambda\mu+\sqrt{\lambda}\log\lambda}^{\infty} K_1\sqrt{\theta}\exp\left(-\frac{K_2}{\theta}\left(\frac{\theta}{\mu}-b_\lambda\right)^2\right)f_\lambda(\theta)d\theta. \qquad (20)
\end{aligned}$$

Considering the first term, we have

$$\begin{aligned}
&\int_0^{b_\lambda\mu-\sqrt{\lambda}\log\lambda} K_1\sqrt{\theta}\exp\left(-\frac{K_2}{\theta}\left(\frac{\theta}{\mu}-b_\lambda\right)^2\right)f_\lambda(\theta)d\theta \\
&\overset{(a)}{\leq}\left(\int_0^{b_\lambda\mu-\sqrt{\lambda}\log\lambda} K_1^2\theta f_\lambda(\theta)d\theta\right)^{\frac{1}{2}} \\
&\times\left(\int_0^{b_\lambda\mu-\sqrt{\lambda}\log\lambda}\exp\left(-\frac{2K_2}{\theta}\left(\frac{\theta}{\mu}-b_\lambda\right)^2\right)f_\lambda(\theta)d\theta\right)^{\frac{1}{2}} \\
&\overset{(b)}{\leq}\left(K_1\sqrt{\lambda}\right)\left(\exp\left(-\frac{2K_2}{(b_\lambda\mu-\sqrt{\lambda}\log\lambda)\mu^2}\lambda(\log\lambda)^2\right)\int_0^\infty f_\lambda(\theta)d\theta\right)^{\frac{1}{2}} \\
&\overset{(c)}{\leq}\left(K_1\sqrt{\lambda}\right)\left(\exp\left(-K_4(\log\lambda)^2\right)\right) \\
&=o(1), \qquad (21)
\end{aligned}$$

where: $K_4>0$ is a finite constant; (a) follows using the Cauchy-Schwarz inequality; (b) follows by noting that the exponential term is increasing in θ on $[0, b_\lambda\mu-\sqrt{\lambda}\log\lambda]$, and hence we can obtain an upper bound by setting $\theta=b_\lambda\mu-\sqrt{\lambda}\log\lambda$; and (c) follows by noting that $b_\lambda=\mathcal{O}(\lambda)$.

Turning to the second term in (20), we have

$$\begin{aligned}
&\int_{b_\lambda\mu-\sqrt{\lambda}\log\lambda}^{b_\lambda\mu+\sqrt{\lambda}\log\lambda} K_1\sqrt{\theta}\exp\left(-\frac{K_2}{\theta}\left(\frac{\theta}{\mu}-b_\lambda\right)^2\right)f_\lambda(\theta)d\theta \\
&\overset{(a)}{\leq}\frac{M_\lambda(b_\lambda\mu)}{\lambda}\int_{b_\lambda\mu-\sqrt{\lambda}\log\lambda}^{b_\lambda\mu+\sqrt{\lambda}\log\lambda} K_1\sqrt{b_\lambda\mu+\sqrt{\lambda}\log\lambda}\exp\left(-\frac{K_2}{b_\lambda\mu+\sqrt{\lambda}\log\lambda}\left(\frac{\theta}{\mu}-b_\lambda\right)^2\right)d\theta \\
&=M_\lambda(b_\lambda\mu)\int_{b_\lambda\mu-\sqrt{\lambda}\log\lambda}^{b_\lambda\mu+\sqrt{\lambda}\log\lambda} K_1\frac{\sqrt{b_\lambda\mu+\sqrt{\lambda}\log\lambda}}{\lambda}\exp\left(-\frac{K_2}{\mu^2(b_\lambda\mu+\sqrt{\lambda}\log\lambda)}(\theta-b_\lambda\mu)^2\right)d\theta \qquad (22) \\
&\overset{(b)}{\leq}M_\lambda(b_\lambda\mu)\int_{b_\lambda\mu-\sqrt{\lambda}\log\lambda}^{b_\lambda\mu+\sqrt{\lambda}\log\lambda}\frac{K_5}{\sqrt{\lambda}}\exp\left(-\frac{K_6}{\lambda}(\theta-b_\lambda\mu)^2\right)d\theta \\
&\overset{(c)}{=}\mathcal{O}(M_\lambda(b_\lambda\mu)),
\end{aligned}$$

where: $K_5, K_6 > 0$ are finite constants; (a) follows by using $f_\lambda(\theta)\lambda \leq M_\lambda(b_\lambda\mu)$, $\sqrt{\theta} \leq \sqrt{b_\lambda\mu + \sqrt{\lambda}\log\lambda}$ and $-1/\theta \leq -1/(b_\lambda\mu + \sqrt{\lambda}\log\lambda)$; (b) follows by using $b_\lambda = \mathcal{O}(\lambda)$; and (c) follows by noting that the integral can be represented in terms of the probability that a normal random variable with mean $b_\lambda\mu$ and variance $\lambda/(2K_6)$ lies in the interval $[b_\lambda\mu - \sqrt{\lambda}\log\lambda, b_\lambda\mu + \sqrt{\lambda}\log\lambda]$.

Further, we can also write the second term in (20) as

$$\begin{aligned}
&\int_{b_\lambda\mu-\sqrt{\lambda}\log\lambda}^{b_\lambda\mu+\sqrt{\lambda}\log\lambda} K_1\sqrt{\theta}\exp\left(-\frac{K_2}{\theta}\left(\frac{\theta}{\mu}-b_\lambda\right)^2\right) f_\lambda(\theta)d\theta \\
&\overset{(a)}{\leq} \int_{b_\lambda\mu-\sqrt{\lambda}\log\lambda}^{b_\lambda\mu+\sqrt{\lambda}\log\lambda} K_1\sqrt{b_\lambda\mu+\sqrt{\lambda}\log\lambda} f_\lambda(\theta)d\theta \\
&\overset{(b)}{=} \mathcal{O}(\sqrt{\lambda}),
\end{aligned} \tag{23}$$

where (a) follows by bounding the exponential term by unity and using $\sqrt{\theta} \leq \sqrt{b_\lambda\mu + \sqrt{\lambda}\log\lambda}$, and (b) follows by noting that $b_\lambda = \mathcal{O}(\lambda)$.

We finally consider the third term in (20). We have

$$\begin{aligned}
&\int_{b_\lambda\mu+\sqrt{\lambda}\log\lambda}^{\infty} K_1\sqrt{\theta}\exp\left(-\frac{K_2}{\theta}\left(\frac{\theta}{\mu}-b_\lambda\right)^2\right) f_\lambda(\theta)d\theta \\
&\overset{(a)}{\leq} \left(\int_{b_\lambda\mu+\sqrt{\lambda}\log\lambda}^{\infty} K_1^2\theta f_\lambda(\theta)d\theta\right)^{\frac{1}{2}} \\
&\times \left(\int_{b_\lambda\mu+\sqrt{\lambda}\log\lambda}^{\infty} \exp\left(-\frac{2K_2}{\theta}\left(\frac{\theta}{\mu}-b_\lambda\right)^2\right) f_\lambda(\theta)d\theta\right)^{\frac{1}{2}} \\
&\overset{(b)}{\leq} \left(K_1\sqrt{\lambda}\right)\left(\exp\left(-\frac{2K_2}{(b_\lambda\mu+\sqrt{\lambda}\log\lambda)\mu^2}\lambda(\log\lambda)^2\right)\int_0^{\infty} f_\lambda(\theta)d\theta\right)^{\frac{1}{2}} \\
&\overset{(c)}{\leq} \left(K_1\sqrt{\lambda}\right)\left(\exp\left(-K_7(\log\lambda)^2\right)\right) \\
&= o(1),
\end{aligned} \tag{24}$$

where: $K_7 > 0$ is a finite constant; (a) follows by using Cauchy-Schwartz inequality; (b) follows by noting that the exponential term is decreasing in θ for $\theta > b_\lambda\mu$, hence we can upper bound the expression by setting $\theta = b_\lambda\mu + \sqrt{\lambda}\log\lambda$; and (c) follows by using $b_\lambda = \mathcal{O}(\lambda)$.

Combining (21)-(24) in (20) completes the proof of (19) and using this result and the definition of $\bar{q}_\lambda(b_\lambda) = \int_0^\infty \left[\frac{\theta}{\mu} - b_\lambda\right]^+ f_\lambda(\theta)d\theta$, the result follows. This completes the proof. ■

Proof of Lemma 2. Fix the arrival rate and abandonment rates at λ and γ respectively. Now, consider System A with k_a servers, each providing service at rate α/k_a, and System B with k_b servers, each providing service at rate α/k_b, where $k_b > k_a$. The result follows if we show that the

rate at which customers abandon in System A is greater than that in System B. This follows by a straightforward pathwise argument comparing the two birth-death chains, and we only provide a sketch of the argument.

Let $\beta^i(m)$ and $\delta^i(m)$ denote the birth (arrival) and death (departure) rates respectively, at state m in System $i = A, B$. (The state here refers to the number of customers in the system.) Then, the we have $\beta^A(\cdot) = \beta^B(\cdot)$, i.e., the two systems have the same birth rates. We now compare the death rates. For any $m \geq 0$, if there are $m + k_a$ customers in System A, then the death rate $\delta^A(m + k_a) = \alpha + m\gamma = \delta^B(m + k_b)$, which is the death rate in System B when it has $m + k_b$ customers. Also, note that as $k_b > k_a$, System B has a lower service rate, and thus we have $\delta^A(k_a - m) < \delta^B(k_b - m)$ for $1 \leq m \leq k_a$. If we construct these chains on the same probability space, then it follows that the number of customers waiting in the queue in System A is always greater than that in System B. This result is similar in flavor to results in Whitt (1981). Noting that the abandonment rate is the same in the two systems, the result follows. ■

B Proofs of Propositions

Proof of Proposition 1. We establish the result for the case $\mu = \gamma$, and the general case follows using the bounding argument in Step 2 of the proof of Theorem 1.

For the case $\mu = \gamma$, using Lemma 3 with $\theta = \lambda$, we obtain the following bound on the queue-length for any capacity level $b \geq 0$:

$$\left[\frac{\lambda}{\mu} - b\right]^+ \leq \mathbb{E}[N_\lambda - b]^+ \leq \left[\frac{\lambda}{\mu} - b\right]^+ + K_1\sqrt{\lambda}\exp\left(-\frac{K_2}{\lambda}\left(\frac{\lambda}{\mu} - b\right)^2\right) + K_3,$$

where $K_1 = \sqrt{4\pi/\mu}$, $K_2 = \mu/4$ and $K_3 = 1/\log 2$. This bound translates to the following bound on the objective function:

$$\bar{\Pi}_\lambda(b) \leq \Pi_\lambda(b) \leq \bar{\Pi}_\lambda(b) + K_1'\sqrt{\lambda}\exp\left(-\frac{K_2}{\lambda}\left(\frac{\lambda}{\mu} - b\right)^2\right) + K_3', \tag{25}$$

where $K_1' = (p\gamma + h)K_1$ and $K_3' = (p\gamma + h)K_3$. Thus, we have

$$\begin{aligned}
\bar{\Pi}_\lambda(\bar{b}_\lambda) = \min_{b\geq 0}\bar{\Pi}_\lambda(b) \leq \Pi_\lambda^* &= \min_{b\geq 0}\{\Pi_\lambda(b)\} \\
&\leq \min_{b\geq 0}\bar{\Pi}_\lambda(b) + \max_{b\geq 0}K_1'\sqrt{\lambda}\exp\left(-\frac{K_2}{\lambda}\left(\frac{\lambda}{\mu} - b\right)^2\right) + K_3' \\
&= \min_{b\geq 0}\bar{\Pi}_\lambda(b) + \mathcal{O}(\sqrt{\mathcal{E}_\lambda}) \\
&= \bar{\Pi}_\lambda(\bar{b}_\lambda) + \mathcal{O}(\sqrt{\mathcal{E}_\lambda}).
\end{aligned}$$

Applying (25) to $b = \bar{b}_\lambda$, we obtain $\Pi_\lambda(\bar{b}_\lambda) = \bar{\Pi}_\lambda(\bar{b}_\lambda) + \mathcal{O}(\sqrt{\mathcal{E}_\lambda})$. This completes the proof. ■

Proof of Proposition 2. Fix the value of the critical fractile at $\alpha = \frac{c/\mu}{p+h/\gamma}$. We will demonstrate that $\sup_{F\in\mathcal{F}_{\lambda,\sigma_\lambda}} |\bar{b} - \mathcal{E}_\lambda|/(\mathcal{E}_\lambda CV_\lambda) \leq K < \infty$ for some finite constant $K > 0$, where $\mathcal{F}_{\lambda,\sigma_\lambda}$ is the set of distribution functions with mean λ and variance σ_λ^2, and $\bar{b} = \bar{F}^{-1}(\alpha)/\mu$ is the newsvendor-based capacity prescription. Noting that $\mathcal{E}_\lambda CV_\lambda = \sigma_\lambda/\mu$ for all distributions in the set $\mathcal{F}_{\lambda,\sigma_\lambda}$, we focus on computing $\sup_{F\in\mathcal{F}_{\lambda,\sigma_\lambda}} |\bar{b} - \mathcal{E}_\lambda|$.

For any distribution $F \in \mathcal{F}_{\lambda,\sigma_\lambda}$, we decompose it into two additional distributions F^1 and F^2 as follows:

$$F^1(x) = \begin{cases} \frac{F(x)}{(1-\alpha)} & \text{for all } x \leq \bar{b}, \\ 1 & \text{otherwise,} \end{cases}$$

$$F^2(x) = \begin{cases} \frac{F(x)-(1-\alpha)}{\alpha} & \text{for all } x > \bar{b}, \\ 0 & \text{otherwise.} \end{cases}$$

Thus F is a mixture distribution of F^1 and F^2, specifically $F = (1-\alpha)F^1 + \alpha F^2$. Let (m_1, σ_1^2) and (m_2, σ_2^2) denote the mean and variance of F^1 and F^2, respectively. Since $F \in \mathcal{F}_{\lambda,\sigma_\lambda}$, we have the following:

$$(1-\alpha)m_1 + \alpha m_2 = \lambda, \quad (\text{using } \mathbb{E}\Lambda = \lambda) \tag{26}$$

$$(1-\alpha)(m_1^2 + \sigma_1^2) + \alpha(m_2^2 + \sigma_2^2) = \lambda^2 + \sigma_\lambda^2, \quad (\text{using } \mathbb{E}\Lambda^2 = \lambda^2 + \sigma_\lambda^2). \tag{27}$$

These equations can be re-expressed as follows:

$$m_1(\sigma_1, \sigma_2) = \lambda - \frac{\sqrt{\alpha(1-\alpha)(\sigma_\lambda^2 - (1-\alpha)\sigma_1^2 - \alpha\sigma_2^2)}}{1-\alpha} \tag{28}$$

$$m_2(\sigma_1, \sigma_2) = \lambda + \frac{\sqrt{\alpha(1-\alpha)(\sigma_\lambda^2 - (1-\alpha)\sigma_1^2 - \alpha\sigma_2^2)}}{\alpha}. \tag{29}$$

Thus, the optimization problem is equivalent to

$$\begin{aligned} \sup_{\sigma_1,\sigma_2\geq 0} \; & |\bar{b} - \mathcal{E}_\lambda| \\ \text{s.t. } & \bar{b} \leq m_2(\sigma_1, \sigma_2)/\mu \\ & \bar{b} \geq m_1(\sigma_1, \sigma_2)/\mu \\ & m_1(\sigma_1, \sigma_2), m_2(\sigma_1, \sigma_2) \geq 0. \end{aligned} \tag{30}$$

It follows that the optimal solution must have $\bar{b} = m_1(\sigma_1, \sigma_2)/\mu$ or $\bar{b} = m_2(\sigma_1, \sigma_2)/\mu$. Noting that m_1 is increasing in σ_1, σ_2, while m_2 is decreasing in σ_1, σ_2, it follows that the optimal solution to (30) is $\sigma_1 = \sigma_2 = 0$. Noting that $m_1(0,0) = \lambda - \sigma_\lambda\sqrt{\alpha/(1-\alpha)}$ and $m_2(0,0) = \lambda + \sigma_\lambda\sqrt{(1-\alpha)/\alpha}$, we obtain that

$$\sup_{F\in\mathcal{F}_{\lambda,\sigma_\lambda}} |\bar{b} - \mathcal{E}_\lambda| = \max\left\{\sqrt{\frac{1-\alpha}{\alpha}}, \sqrt{\frac{\alpha}{1-\alpha}}\right\}\frac{\sigma_\lambda}{\mu} = K\mathcal{E}_\lambda CV_\lambda,$$

where $K \geq 0$ is a finite constant. This completes the proof. ■

Proof of Proposition 3. We prove the result for the case $\mu = \gamma$. The general case follows by a bounding argument analogous to that developed in the proof of Theorem 3, and we omit these details.

Consider any real-valued sequence $\{s_\lambda\}$ such that $|s_\lambda| = \mathcal{O}(\sqrt{\lambda})$. Then, applying Lemma 1 and Assumption 1, we obtain

$$\Pi_\lambda(\bar{b}_\lambda + s_\lambda) \leq \bar{\Pi}_\lambda(\bar{b}_\lambda + s_\lambda) + \mathcal{O}(\min\{1/CV_\lambda, \sqrt{\mathcal{E}_\lambda}\}). \tag{31}$$

Noting that $\bar{\Pi}_\lambda(\cdot)$ is a convex function whose slope is bounded above by c and bounded below by $c - (p + h/\gamma)\mu$, we have that for any $\lambda > 0$

$$\bar{\Pi}_\lambda(\bar{b}_\lambda + s_\lambda) \leq \bar{\Pi}_\lambda(\bar{b}_\lambda) + c|s_\lambda|. \tag{32}$$

Now applying a Taylor series expansion around $\bar{b}_\lambda$, and noting that $\bar{\Pi}'_\lambda(\bar{b}_\lambda) = 0$, we obtain

$$\bar{\Pi}_\lambda(\bar{b}_\lambda + s_\lambda) \leq \bar{\Pi}_\lambda(\bar{b}_\lambda) + s_\lambda^2 \sup_{\xi \in A} \bar{\Pi}''(\xi)/2, \tag{33}$$

where $A = [\bar{b}_\lambda, \bar{b}_\lambda + s_\lambda]$ if $s_\lambda \geq 0$, and $A = [\bar{b}_\lambda + s_\lambda, \bar{b}_\lambda]$, if $s_\lambda < 0$. Noting that $\bar{\Pi}'_\lambda(x) = c - (p + h/\gamma)\mu\bar{F}_\lambda(\mu x)$, we obtain that

$$\sup_{\xi \in A} \bar{\Pi}''(\xi) \leq \sup_{\xi \in A}(p + h/\gamma)\mu^2 f_\lambda(\mu\xi) \leq \frac{K_1}{\lambda CV_\lambda}, \text{ for } \lambda \text{ sufficiently large,}$$

where $K_1 > 0$ is a finite constant and we use the fact that Assumption 1 implies $f_\lambda(x) = \mathcal{O}(1/CV_\lambda)$. Using this relation in (33), we obtain

$$\bar{\Pi}_\lambda(\bar{b}_\lambda + s_\lambda) = \bar{\Pi}_\lambda(\bar{b}_\lambda) + \mathcal{O}(s_\lambda^2/(\lambda CV_\lambda)).$$

Combining this with (32), we obtain

$$\bar{\Pi}_\lambda(\bar{b}_\lambda + s_\lambda) = \bar{\Pi}_\lambda(\bar{b}_\lambda) + \mathcal{O}\left(\min\{s_\lambda^2/(\lambda CV_\lambda), |s_\lambda|\}\right).$$

Comparing this with (31), we obtain that for $|s_\lambda| = \mathcal{O}(\sqrt{\mathcal{E}_\lambda})$, $\Pi_\lambda(\bar{b}_\lambda + s_\lambda) \leq \bar{\Pi}(\bar{b}_\lambda) + \mathcal{O}(\min\{1/CV_\lambda, \sqrt{\mathcal{E}_\lambda}\})$, and the result follows by noting that $\Pi_\lambda(\bar{b}_\lambda) = \Pi^*_\lambda + \mathcal{O}\left(\min\left\{1/CV_\lambda, \sqrt{\mathcal{E}_\lambda}\right\}\right)$ (cf. Theorem 1). ■

C Additional Numerical Results

In this section, using numerical experiments, we study the accuracy of the newsvendor-based prescription in two settings: a) when the system size is small; and b) when there is a single server and the decision variable is the rate at which this server works.

C.1 Effect of system scale

We set the system parameters identical to those in the numerical experiments of Section 4.1, i.e., $c = 1/3, h = 1, p = 1$, unit mean service times, and mean impatience time of 1/3. Table 7 displays the results for uniform arrival rate distributions with means ranging from 1.5 up to 300. As expected, we observe a degradation in performance for small mean arrival rates, however, we note that even for the case of U[10,20] the performance is quite good.

Arrival rate	Optimal solution		Prescription		Difference		
distribution	b^*	Π^*	$\bar{b}$	$\Pi(\bar{b})$	$\lvert b^* - \bar{b}\rvert$	$\Pi(\bar{b}) - \Pi^*$	$\frac{(\Pi(\bar{b})-\Pi^*)}{\Pi^*}$
U[1,2]	3	1.43	1	5.26	2	3.83	267.8%
U[2,4]	5	2.19	3	3.33	2	1.14	52.1%
U[5,10]	11	4.32	8	5.01	3	0.71	16.4%
U[10,20]	20	7.68	17	8.01	3	0.33	4.3%
U[15,30]	29	10.95	26	11.12	3	0.17	1.6%
U[20,40]	37	14.15	35	14.26	2	0.10	0.7%
U[25,50]	46	17.34	43	17.49	3	0.15	0.9%
U[50,100]	89	33.13	87	33.20	2	0.07	0.2%
U[200,400]	351	127.08	350	127.08	1	7×10^{-3}	0.006%

Table 7: **Performance of the newsvendor-based prescription over different system sizes.** Comparison of the optimal (μ^*) and prescribed ($\bar{\mu}$) capacity levels along with their respective performance for different system sizes. Arrival rates follow a uniform distribution with a coefficient of variation of 19.2%.

C.2 Single server systems

In this section, we consider the setting of a single server system where the decision variable is the service rate. In this case, the expected queue-length equals $\mathbb{E}[N-1]^+$ (as there is only one server) and the cost of capacity is $c\mu$, where μ is the server's processing rate. Thus, the optimization

problem is

$$\min_{\mu \geq 0} \ (p\gamma + h)\mathbb{E}[N-1]^+ + c\mu,$$

which is analogous to (1). Under parameter uncertainty, the newsvendor-based capacity prescription for this system is again analogous to (3) and given by

$$\bar{\mu} = \bar{F}^{-1}\left(\frac{c}{p + h/\gamma}\right).$$

To test the accuracy of the above prescription, we perform a numerical study. In particular, we set the parameters $c = 2/3$, $p = h = 1$ and the abandonment rate $\gamma = 1$. The results of this study are depicted in Table 8. As in the many server case, we find that the newsvendor prescription performs extremely well, and its error does not increase with system scale.

Arrival rate	Optimal solution		Prescription		Difference		
distribution	μ^*	Π^*	$\bar{\mu}$	$\Pi(\bar{b})$	$\lvert\mu^* - \bar{\mu}\rvert$	$\Pi(\bar{\mu}) - \Pi^*$	$\frac{(\Pi(\bar{b}) - \Pi^*)}{\Pi^*}$
U[25,50]	29.58	29.44	33.33	29.67	3.75	0.24	0.8%
U[50,100]	62.42	57.50	66.67	57.67	4.25	0.17	0.3%
U[200,400]	260.92	224.46	266.67	224.54	5.75	0.08	0.04%

Table 8: **Performance of the newsvendor-based prescription for a single server system.** Comparison of the optimal (μ^*) and prescribed ($\bar{\mu}$) capacity levels along with their respective performance. Arrival rates follow a uniform distribution with a coefficient of variation of 19.2%.