Three Sojourns in Queueing Theory

Jacob Mason Bergquist

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2023

Abstract

Three Sojourns in Queueing Theory

Jacob Mason Bergquist

In this thesis, we present three works on queues. In chapter 1, we analyze two non-work-conserving variations of the M/G/1 preemptive LIFO queue, focusing on deriving expressions for the limiting distribution of workload and related quantities. In the first model, preempted customers return to the front of the queue with a new service time, while in the second, they return with their original service time. We use queueing theory methods such as the Rate Conservation Law, PASTA, regenerative process theory and Little's Law. Our results include stability and heavy-traffic limits, as well as tail asymptotics for stationary workload. In chapter 2, we analyze a queueing model with price-sensitive customers, where the service provider aims to maximize revenue and minimize the average queue length. Customers arrive according to a Poisson process, join the queue if their willingness-to-pay exceeds the offered price, and are served in a first-in first-out manner with exponential service times. Our model is applicable to cloud computing, make-to-order manufacturing, and food delivery. We provide performance guarantees for a class of static pricing policies that can achieve a constant fraction of the optimal revenue with a small increase in expected queue length. We present results for the single-server, multi-server, and multi-class cases and provide numerical findings to demonstrate the empirical performance of our policies. In chapter 3, we analyze the Adaptive Non-deterministic Transmission Policy (ANTP), a technique addressing the Massive Access Problem (MAP) in telecommunications, which involves delaying packets at the points of origin to reduce congestion.

We frame these delays as time spent at a "cafe" before proceeding to the service facility. We present sample-path results, giving conditions under which ANTP does not change the total sojourn time of packets, and results under a general stochastic framework, focusing on stability and constructing proper stationary versions of the model. We prove Harris recurrence of an underlying Markov process and find positive recurrent regeneration points under i.i.d. assumptions.

Table of Contents

List of Figures

List of Tables

Acknowledgements

There are so many people I would like to thank. I will start from the beginning. My mom and dad for raising me, my brother and sister for helping me grow. Karl, of course, for advising me since the very beginning. Karl's wife and kids for sharing him with me. Adam for treating me no differently than any other senior member of his lab and giving me five years worth of advising in three semesters. The many friends who helped me along the way: my cohortmates, particularly Luc, Achraf, Harsh, and Ruizhe, who I am honored to be associated with. Older and younger PhDs who have enriched my time: Goutam, Allen, Mali, Oussama, Shatian, Agathe, Tugce, Yuki, Steven, Yuan, Xuan, Enrique, Julian, Camilo, Sudeep, Rachitesh, Aapeli, Ayoub, Abdellah, Matias, Noémie, Madhu, Darshan, Jiaqi. Liz, who helped me so many times and by her guidance helped make this whole enterprise worthwhile. Kristen for letting me bother her. Winsor for helping me across the finish line. Aubrianna for soothing my core course anxieties. My friends turned roommates Daniel, Nico, and Tom for providing respite (sometimes in canine form via Rocco and Leo). Pranay, Saeed, Chelsey, Ilina, and Z for buttressing my confidence. Max, Jayden, Sam, Charlie, Luc, Jonathan, Tyler, Geoff, and Eli for weekly adventures in Ambrosia and Theros. Sebastian for keeping me grounded. Austin for providing solace. I am truly blessed to be surrounded by such great people. Carrying out the math contained in this thesis was amazing, but the best gift of all was the social web I was fortunate enough to be a part of these past five years.

Dedication

To Mom and Dad.

Introduction

People have been waiting in lines for a long time. The mathematical treatment of queues began more recently, in 1909, when a Dane named Agner Krarup Erlang published an innocuous, seven page paper titled "The Theory of Probabilities and Telephone Conversations" ([1]). At the time, Erlang was working at the Copenhagen Telephone Company, and he had been presented with the classic problem of determining how many circuits and telephone operators were needed to guarantee that calls would be serviced in a timely manner. The models he developed and analyzed to tackle these questions in this and later works ([2], [3]) became foundational elements of modern telecommunication network studies. But, as he presciently hinted at in [1][1], the potential applications of his work extended far beyond this original telecommunication application. In understanding that "a special knowledge of telephonic problems is not at all necessary for the understanding thereof," Erlang realized that his analysis had successfully abstracted away from the particulars of the original application. What was left was an instance of a new mathematical model - a queue - and with that, queueing theory was born. Now, more than a century later, we present these three fresh sojourns into the theory of queues.

In the first chapter, we analyze two non-work-conserving variations of the M/G/1 preemptive last-in first-out (LIFO) queue with emphasis on deriving explicit expressions for the limiting (stationary) distribution of workload and other related quantities of interest. In the first model, known as preemptive-repeat different (PRD), preempted customers are returned to the front of the queue with a new independent and identically distributed service time. In the second, known as

[1] "As it is my belief that some point or other from this work may be of interest, and as a special knowledge of telephonic problems is not at all necessary for the understanding thereof, I shall give an account of it below."

preemptive-repeat identical (PRI), they are returned to the front of the queue with their original service time. Our analysis is based on queueing theory methods such as the Rate Conservation Law, PASTA, regenerative process theory and Little's Law ($l = \lambda w$). Along the way, we obtain stability results as well as explicit expressions for the limiting distribution of the service time found in service by an arrival. For the second model we even derive the joint distribution of age and excess of such a service time, and find they are quite different from what is found in standard work-conserving models. For example, in the M/M/1 case, they are independent exponentials but with different rates. We also give heavy-traffic limits and tail asymptotics for stationary workload for both models, as well as deriving an implicit representation for the distribution of sojourn time by introducing an alternative *effective* service time distribution.

In the second chapter, we consider a general queueing model with price-sensitive customers in which the service provider seeks to balance two objectives, maximizing revenue and minimizing the average queue length. Customers arrive according to a Poisson process, observe an offered price, and decide to join the queue if their willingness-to-pay exceeds the offer. The queue is operated first-in first-out, and the service times are exponential. Our model represents applications in areas like cloud computing, make-to-order manufacturing, and food delivery.

The optimal solution for our model is *dynamic*; the price changes as the state of the system changes. However, such dynamic pricing policies may be undesirable for a variety of reasons. In this work, we provide performance guarantees for a simple and natural class of *static* pricing policies which charge a fixed price up to a certain occupancy threshold and then allow no more customers into the system. We provide a series of results showing that such static policies can simultaneously achieve a constant fraction of the optimal revenue with at most a constant factor increase in expected queue length. For example, in the single-server case, we show that a static pricing policy can always achieve at least half of the optimal revenue while at most matching the expected queue length of the optimal solution. We also furnish results for the multi-server and multi-class cases and provide numerical findings demonstrating the empirical performance of our static policies.

In chapter three, we analyze a recently proposed technique for addressing the Massive Access Problem (MAP), an issue in telecommunications which arises when too many devices transmit packets to a gateway in quick succession. This technique, the Adaptive Non-deterministic Transmission Policy (ANTP), involves delaying some packets at the points of origin to alleviate congestion at the routers. In this work, to clarify the approach and to move beyond the original telecommunications application, we frame these potential delays as time spent at a "cafe" before proceeding to the service facility.

We present both sample-path results and results under a stationary ergodic stochastic framework. In the sample-path realm, we give conditions that ensure ANTP will not change the total sojourn time of any packet as compared to what that packet would have experienced in the original FIFO model without delays. The difference is that, under ANTP, some of that sojourn is spent at the cafe instead of in the buffer at the service facility. In a stochastic framework, our focus is on stability and constructing proper stationary versions of the model including the ANTP point process. Under i.i.d. assumptions we dig deeper by proving Harris recurrence of an underlying Markov process, and explicitly find positive recurrent regeneration points.

Chapter 1: Stationary workload for some non-work-conserving M/G/1 preemptive LIFO queues

In this chapter we analyze a queueing model which involves preemption. This chapter is based on the article [4] written in collaboration with Karl Sigman and published in *Stochastic Systems* in May 2022. We thank Peter Glynn for helpful comments and suggestions.

1.1 Introduction

In this paper we consider two *non-work-conserving* variations of the M/G/1 preemptive LIFO (PL) queue. As with the classic/standard (work-conserving) M/G/1 PL model, when a new customer arrives they immediately bump out any customer in service and start service themselves with their own independent and identically distributed (i.i.d) service time, while the preempted customer returns to the front of the queue. But in the classic model, the preempted customer retains its *remaining service time* and thus the model is work-conserving; in particular, the workload process is identical sample-path by sample-path to the standard first-in-first-out (FIFO) M/G/1 model. In the two models we analyze here, the preempted customer either receives a new i.i.d. service time (known as preemptive-repeat different (PRD)), or retains its original service time (known as preemptive-repeat identical (PRI)), hence losing any progress that had been made.

In preemption applications, the PRI model would be suitable for when customers are viewed as bringing service times with them, so there only is one service time in play, whereas, the PRD model would be more suitable for when the server (or system) is viewed as handing out the service (processing) times whenever a customer enters service while the customers are all bringing identical tasks. Jobs arriving to a CPU or a printer for example would naturally be PRI, whereas arrivals wanting access to a particular website would more naturally be PRD.

These two models were presented and analyzed in the recent paper of Asmussen and Glynn [5] in which the focus was on determining the moments of sojourn time and establishing the stability conditions. Their methods of analysis involve branching processes, Galton-Watson family trees and stochastic fixed point equations.

Our focus in the present paper, however, is on deriving the entire limiting (stationary) *distribution of workload*, as well as using workload to derive other quantities of interest. We do so by giving an explicit random variable representation for the workload very much in the spirit of the classic Pollaczek-Khinichine formula for the standard M/G/1 queue in which the limiting distribution of the workload is expressed as a geometric sum of i.i.d. random variables endowed with the *equilibrium* (stationary excess) distribution of service. (See for example, Pages 386-387 in [6].) As part of this we take advantage of the fact that for both models, the limiting distribution of the number of customers in the system (as found by an arrival) is geometric. This fact was first established for the standard work-conserving PL model by Fakinos[7], [8] and Yamazaki[9],[10] and extended to general preemption models (which includes PRD and PRI) in Shanthikumar and Sumita[11],Theorem 2.2.

But as we discover by explicitly computing them, the parameters of these geometric distributions involve the entire distribution of service, not just its mean, and the distribution of the i.i.d. random variables is not the equilibrium distribution, nor the stationary *spread* distribution as found in the inspection paradox.

Along the way we also obtain (in two different ways) the stability results found in [5]–but with more of a 'queueing' interpretation–as well as explicit expressions for the limiting distribution of the service time found in service by an arrival. For PRI we even derive the joint distribution of age and excess (remaining service time) of such a service time, and discover that it is quite different from what is found in standard work-conserving models. Our analysis is based on queueing theory methods such as Rate Conservation Law, PASTA, regenerative process theory and Little's Law ($l = \lambda w$). We also present, for both models, heavy-traffic limits and tail asymptotics for the stationary workload. Section 1.2 deals with PRD, and Section 1.3 with PRI; an implicit sojourn time representation and Laplace transform are given in Section 1.3.11. An Appendix is included at the end containing some of the proofs of our results.

1.1.1 Basic M/G/1 model notation and set up

The M/G/1 queue has a Poisson point process of customer arrival times $\{t_n : n \geq 1\}$ at rate λ, with i.i.d. exponentially distributed interarrival times $T_n = t_{n+1} - t_n$ at rate λ, and (independently) i.i.d. service times $\{S_n : n \geq 1\}$ brought by each customer distributed as a general distribution $G(x) = P(S \leq x),\ x \geq 0$, where S denotes a generic such service time. We assume that $0 < E(S) = 1/\mu < \infty$. T denotes a generic interarrival time distributed as exponential at rate λ, and then we define $\rho \overset{\text{def}}{=} \lambda/\mu$. (A priori all we can say about ρ is that $0 < \rho < \infty$).

The workload $V(t)$ at time t is the sum of all remaining or whole service times in the system at time t: the sum of all service times of customers in the queue plus the remaining service time of the customer in service (if any). $\{V(t) : t \geq 0\}$ then denotes the workload stochastic process; its sample paths are continuous from the right with left hand limits. $V(t_n-)$ denotes the amount of work *found in the system* by the n^{th} arrival, and $V(t_n+)$ is the amount of work right after they arrive. For example, for work-conserving disciplines $V(t_n+) = V(t_n-) + S_n$.

The times at which an arrival finds the system empty, $V(t_n-) = 0$, serve as regeneration points with i.i.d. cycles. In the case when the cycle length distribution is proper and has finite first moment–the positive recurrent case–we are ensured the existence of a (proper) limiting (stationary) distribution of workload, and that is what we mean by *stability* in the present paper. We let V denote a random variable with this distribution and can define the distribution via w.p. 1 limits:

$$P(V \leq x) = \lim_{t\to\infty} \frac{1}{t} \int_0^t I\{V(s) \leq x\} ds,\ x \geq 0.$$

Because we are assuming Poisson arrivals we can use *Poisson Arrivals See Time Averages (PASTA)* (see for example Theorem 6, Page 294 in [6]) to also express this distribution as a w.p. 1

customer average:

$$P(V \leq x) = \lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} I\{V(t_j-) \leq x\}, x \geq 0. \tag{1.1}$$

Of crucial importance in the present paper is computing

$$p_0 \overset{\text{def}}{=} P(V = 0), \tag{1.2}$$

which, because of PASTA, can be viewed as both the long-run proportion of time that the system is empty and as the long-run proportion of arriving customers who find the system empty. We can use the above limits to *determine* stability in advance: it is precisely the case when p_0 exists and satisfies $0 < p_0 < 1$ (we avoid the trivial case of $p_0 = 1$).

For work-conserving M/G/1 models it is well known that $0 < p_0 < 1$ if and only if $0 < \rho < 1$ in which case $p_0 = 1 - \rho$, but in the present paper with non-work-conservation in play, we will see that this does not hold in general. Moreover, in the work-conserving case workload forms a Markov process, but not here. We will refer to the classic work-conserving model as the *classic* (or standard or regular) M/G/1 model, see Chapters 8 and 10 of [6], for example, for an overview.

Related to (1.1), we will also be considering the stationary distributions of the age $B(t)$, excess $S_r(t)$ (remaining service time) and whole length $S^*(t) = B(t) + S_r(t)$ of a service time in service at time $t \geq 0$ (defined to be 0 if the system is empty); for example,

$$P(S_r \leq x) = \lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} I\{S_r(t_j-) \leq x\}, x \geq 0, \tag{1.3}$$

with the other two B and S^* defined similarly. Of particular interest in the present paper is determining their conditional distributions given they are positive; we let $\hat{B} = (B \mid B > 0)$, $\hat{S}_r = (S_r \mid S_r > 0)$, $\hat{S} = (S^* \mid S^* > 0)$ denote random variables with such distributions, and note in passing that $P(B = 0) = P(S_r = 0) = P(S^* = 0) = p_0$.

1.2 The PRD model

We begin with the M/G/1 LIFO preemptive *repeat-different* queue (PRD). We recall this is the preemptive LIFO model in which each time a customer in service is preempted, they go to the front of the queue with a new i.i.d. service time S distributed as G.

Recalling $\hat{S}_r = (S_r \mid S_r > 0)$ from Equation (1.3), PASTA and the memoryless property of the exponential distribution immediately yield the following lemma, where $\overset{\mathrm{d}}{=}$ denotes "equal in distribution".

Lemma 1.2.1. *For the M/G/1 PRD model*

$$\hat{S}_r \overset{\mathrm{d}}{=} (S - T \mid S > T). \tag{1.4}$$

We next compute the mean:

Lemma 1.2.2. *For the M/G/1 PRD model*

$$E(\hat{S}_r) = \frac{\frac{1}{\mu}}{1 - E(e^{-\lambda S})} - \frac{1}{\lambda}. \tag{1.5}$$

Proof.

$$E(S - T \mid S > T) = \frac{E(S - T; T < S)}{P(T < S)} = \frac{E(S - T) + E(T - S; T > S)}{P(T < S)},$$

where the denominator is computed as $1 - E(e^{-\lambda S})$, and via the memoryless property of T the numerator as $\frac{1}{\mu} - \frac{1}{\lambda}(1 - E(e^{-\lambda S}))$; the result follows. □

We now derive a formula for p_0. Using the Rate Conservation Law (RCL) (see for example Theorems 5.5 and 5.6, Page 116 in [12], or [13]) with $\{X(t)\}$ as the stochastic process $\{V(t)\}$, we obtain the following proposition.

Proposition 1.2.1. *When the M/G/1 PRD model is stable,*

$$p_0 = \frac{1 - 2\rho + \lambda E(\hat{S}_r)}{1 - \rho + \lambda E(\hat{S}_r)}. \tag{1.6}$$

Proof. For $X(t) = V(t)$, we have $X'(t) = -I\{V(t) > 0\}$ (right derivative) w.p. 1. Meanwhile, jumps occur at customer arrival times $\{t_j\}$ at rate λ, but we break them up into two disjoint streams (type 0 and 1, respectively): customers who find the system empty which arrive at rate $\lambda_0 \overset{\text{def}}{=} \lambda p_0$, and those that find the system busy who arrive at rate $\lambda_1 \overset{\text{def}}{=} \lambda(1 - p_0)$. This leads to

$$P(V > 0) = \lambda_0 E^0(-J(0)) + \lambda_1 E^1(-J(1)), \tag{1.7}$$

where $-J(0)$ denotes a jump amount $(V(t_j+) - V(t_j-))$ of type 0, and $-J(1)$ denotes a jump amount of type 1. E^i, $1 \leq i \leq 2$ denotes expected value under the limiting distribution with respect to arrivals of type i. We obtain $E^0(-J(0)) = E(S)$ and $E^1(-J(1)) = E(S_1 + S_2 - \hat{S}_r)$ (where S, S_1 and S_2 are i.i.d. distributed as G and independent of $\hat{S}_r$).

Then Equation (1.7) becomes

$1 - p_0 = \rho p_0 + 2\rho(1 - p_0) - \lambda(1 - p_0)E(\hat{S}_r)$, which solving for p_0 yields Equation (1.6). □

As a sanity check, using Equation (1.6) on the special case when G is exponential at rate μ (the M/M/1 case), we know that $\hat{S}_r \sim exp(\mu)$ and hence $\lambda E(\hat{S}_r) = \rho$ and Equation (1.6) yields $p_0 = 1 - \rho$ as it should. (And of course we must have that $\rho < 1$ in this special case.)

Remark 1.2.1. For the M/M/1 PRD model, when a job is interrupted, both $\hat{S}_r$ (removed) and the new i.i.d. S (placed in queue) are distributed as exponential at rate μ, but have different sample paths. Thus the stochastic process $\{V(t) : t \geq 0\}$ has a different distribution than for the standard M/M/1. But for each *fixed* t, $V(t)$ has the same distribution as the standard M/M/1, hence they share the same limiting distribution as $t \to \infty$.

Remark 1.2.2. Equation (1.6) still remains valid for general renewal arrivals at rate λ except then we no longer have a closed form expression for $\hat{S}_r$: Equation (1.4) no longer holds in general. In our use of RCL, $\hat{S}_r$ must have the stationary remaining service time distribution with respect to customer arrivals finding the system busy.

1.2.1 Relating p_0 to the stability condition for PRD

Clearly, Equation (1.6) makes sense (proportion of time) and yields stability (ruling out null recurrence, $p_0 = 0$) only if $0 < p_0 < 1$. Since $\rho > 0$ in Equation (1.6), $0 < p_0 < 1$ if and only if

$$1 - 2\rho + \lambda E(\hat{S}_r) > 0, \tag{1.8}$$

which from Equations (1.4) and (1.2.2) yields the stability condition

$$E(e^{-\lambda S}) > \frac{1}{2}; \tag{1.9}$$

we obtain the same condition as in Theorem 5 on Page 15 in [5] (equality is null recurrence).

Furthermore, plugging Equation (1.2.2) back into Equation (1.6) furnishes a closed form solution for p_0:

Proposition 1.2.2. *For a stable M/G/1 PRD model*

$$p_0 = \frac{2E(e^{-\lambda S}) - 1}{E(e^{-\lambda S})}. \tag{1.10}$$

Remark 1.2.3. As an interesting PRD example, consider the case when service times are a mixture of a point mass at 0 and an exponential at rate 0.01: $S \overset{\mathrm{d}}{=} (0.99)\delta_0 + (0.01)exp(0.01)$; $E(S) = 1 = \mu$, and $\rho = \lambda$. Clearly, $\hat{S}_r \overset{\mathrm{d}}{=} exp(0.01)$ since a 0 is never found in service; $E(\hat{S}_r) = 100$. Thus Equation (1.6) becomes $p_0 = \frac{1-2\lambda+\lambda(100)}{1-\lambda+\lambda(100)} = \frac{1+98\lambda}{1+99\lambda}$, and we see that the system is stable for all values of λ, and as $\lambda \to \infty$, p_0 decreases monotonically to $98/99$. This is quite intuitive: any non-zero remaining service time that gets preempted is replaced by a 0 service time 99% of the time.

1.2.2 Deriving the distribution of V for PRD

With p_0 explicitly in hand via Proposition 1.2.2, and the distribution of $\hat{S}_r$ (distributed as in Equation (1.4)) also in hand, we now obtain an explicit random variable expression for V. To prepare, let N denote, under stationarity, the number of customers in the system; it has a geometric distribution with success probability p_0:

$$P(N = n) = (1 - p_0)^n p_0,\ n \geq 0. \tag{1.11}$$

(That N is geometric for various PL models goes back to [7], [8], [9],[10] and Theorem 2.2 in [11].)

We let $Q = (N - 1)^+$ conditional on $\{N \geq 1\}$, it is identically distributed with N, geometric, because N is geometric. It represents the number in queue (line) given that $N \geq 1$.

The random variable Q and the i.i.d. service time sequence $\{S_j\}$ distributed as G, and the random variable $\hat{S}_r$ (distributed as in Equation (1.4)), are taken as independent in what follows.

Proposition 1.2.3. *For the stable M/G/1 PRD model,*

$$(V \mid V > 0) \overset{\mathrm{d}}{=} \hat{S}_r + \sum_{j=1}^{Q} S_j. \tag{1.12}$$

Thus the distribution of V, F_V, is a mixture

$$F_V = p_0\delta_0 + (1 - p_0)F_{\hat{V}}$$

where $F_{\hat{V}}$ denotes the distribution of $(V \mid V > 0)$ *given in Equation (1.12), and δ_0 denotes the point mass at* 0.

Proof. Given that $V > 0$, there is a customer in service and Q customers in queue. Those in queue have i.i.d. service times distributed as G (and are independent of Q,$\hat{S}_r$) by definition of PRD. The one in service has remaining service time $\hat{S}_r$. That $\hat{S}_r$ is independent of what is in queue follows from exactly the same arguments found in [7], and [10], for example, for the standard M/G/1 PL model, where it is shown that the stationary remaining service time is independent of the system state. □

Remark 1.2.4. One can now compute moments such as $E(V) = (1 - p_0)E(\hat{S}_r) + \frac{1}{\mu}E(Q)$

1.2.3 Heavy-traffic limits for stationary workload V for PRD

We now give a simple argument to characterize the asymptotic behavior of stationary workload in the heavy-traffic regime. $\Longrightarrow$ denotes convergence in distribution, while $\overset{p}{\Longrightarrow}$ denotes convergence in probability in what follows. Here we consider a heavy-traffic regime analogous to how it is considered for a regular (work-conserving) M/G/1 queue with G fixed (with tail denoted by $\overline{G}(x),\ x \geq 0$) in which by letting $\lambda \uparrow \mu$ (equivalently $\rho \uparrow 1$) it holds that $(1-\rho)V \Longrightarrow exp(\alpha)$ (the exponential distribution at rate α). In that classic case $\alpha^{-1} = E(S_e) = E(S^2)/2E(S)$, where S_e has density $g_e(x) = \mu\overline{G}(x)$, the *equilibrium* (or stationary excess) distribution of G. For the PRD model we replace $1-\rho$ by $p_0 = p_0(\lambda)$ from Equation (1.10) and consider what happens to $p_0 V$ as $p_0 \to 0$. Stability for PRD is that $E(e^{-\lambda S}) > 1/2$, and thus by increasing λ to the value λ_2 such that $E(e^{-\lambda_2 S}) = 1/2$ results in $p_0 \downarrow 0$; *heavy-traffic*. The proof of the following is in the Appendix.

Theorem 1.2.1. *Let $\lambda_2 > \lambda$ be the solution to $E[e^{-\lambda_2 S}] = 1/2$ (which exists by the monotone convergence theorem). Then as $\lambda \uparrow \lambda_2$,*

$$p_0 V \Longrightarrow exp(\mu),$$

where $\mu^{-1} = E(S)$.

1.2.4 Average sojourn time for PRD

Since N has a geometric distribution as explained in the proof of Proposition 1.2.3, we obtain the time average number in system as $l = E(N) = \frac{1-p_0}{p_0}$. Using our solution to p_0 from Equation (1.10) then yields

$$l = \frac{\rho}{1 - 2\rho + \lambda E(\hat{S}_r)} = \frac{1 - E(e^{-\lambda S})}{2E(e^{-\lambda S}) - 1} = -1 + \frac{E(e^{-\lambda S})}{2E(e^{-\lambda S}) - 1}.$$

From Little's law ($l = \lambda w$) we thus can also solve for average sojourn time $w = l/\lambda$:

$$w = \frac{1}{\lambda}\Big[-1 + \frac{E(e^{-\lambda S})}{2E(e^{-\lambda S}) - 1}\Big]. \tag{1.13}$$

Remark 1.2.5. In [5], the formula given for w in Proposition 6, Page 17 is incorrect as has been confirmed by the authors (private communication). The error is only an algebraic one; carrying out the computation as they suggest, one indeed obtains our Equation (1.13).

Remark 1.2.6. In all the PL models, the distribution of sojourn time W is identical to that of a busy period. So Equation (1.13) is also the expected value of a busy period for PRD. Moreover, from PASTA, p_0 is equal to the long-run proportion of arrivals who begin a busy period, hence starting a regenerative cycle. Thus by regenerative process theory in discrete-time, $p_0 = \frac{1}{E(N_B)}$, where N_B denotes the number of customers served during a busy period. Using Proposition 1.2.2, we thus can solve for $E(N_B)$ yielding

$$E(N_B) = \frac{1}{p_0} = \frac{E(e^{-\lambda S})}{2E(e^{-\lambda S}) - 1}. \tag{1.14}$$

1.2.5 Tail asymptotics for stationary workload V for PRD

Here we give some results for the asymptotics of the tail, $P(V > x)$, as $x \longrightarrow \infty$. We use the notation $a(x) \sim b(x)$ to denote asymptotic equivalence of two non-negative functions as $x \longrightarrow \infty$:

$$\lim_{x\to\infty} \frac{a(x)}{b(x)} = 1.$$

We consider the case of heavy-tailed service times; they have an infinite moment generating function.

The easiest and cleanest case is when the service-time distribution G is from a subclass of heavy-tailed distributions called *subexponential* distributions:

Definition 1.2.1. *A probability distribution $G(x) = P(S \leq x)$, $x \geq 0$, that satisfies $P(S > x) > 0$, $\forall x > 0$, is called subexponential if for i.i.d. copies $S, S_1, S_2, \ldots$ it holds that for each $n \geq 1$, $P(S_1 + \ldots + S_n > x) \sim nP(S > x)$, as $x \longrightarrow \infty$.*

The reader is referred to [14] for basic results on heavy-tailed distributions including subexponential distributions (in the context of queueing theory).

Proposition 1.2.4 (heavy-tailed case)**.** *For the stable PRD model, suppose that service-time distribution $G(x) = P(S \leq x)$ is subexponential. Then*

$$P(V > x) \sim \frac{1}{p_0} P(S > x),$$

Proof. From Proposition 1.2.3, $(V \mid V > 0) \overset{\mathrm{d}}{=} \hat{S}_r + \sum_{i=1}^{Q} S_i$, where $P(Q = n) = (1 - p_0)^n p_0$, $n \geq 0$. We need to verify that both pieces, $\sum_{i=1}^{Q} S_i$, and $\hat{S}_r$ are subexponential when the S_i are, and hence so is the (independent) sum of them. To this end,

$$P\Big(\sum_{i=1}^{Q} S_i > x\Big) \sim E(Q) P(S > x),$$

is subexponential from Proposition 2.9, Page 266 in [14]. Moreover, since T is light-tailed, $P(S - T > x) \sim P(S > x)$ and hence $\hat{S}_r = (S - T \mid S > T)$ is subexponential with

$$P(\hat{S}_r > x) \sim (1 - E(e^{-\lambda S}))^{-1} P(S > x).$$

Thus

$$P\Big(\hat{S}_r + \sum_{i=1}^{Q} S_i > x\Big) \sim ((1 - E(e^{-\lambda S}))^{-1} + E(Q)) P(S > x).$$

Since $P(V > 0) = 1 - p_0$, and $E(Q) = \frac{1-p_0}{p_0}$, we finally obtain

$$P(V > x) \sim c P(S > x),$$

where

$$c = (1 - p_0)\Big[(1 - E(e^{-\lambda S}))^{-1} + \frac{1 - p_0}{p_0}\Big] = \frac{1}{p_0},$$

with the last equality following from algebra by use of Proposition 1.2.2.

□

Remark 1.2.7. From Equation (1.14), it is interesting to note that the tail asymptotic in Proposition 1.2.4 can be re-written as $P(V > x) \sim E(N_B)P(S > x)$.

1.3 PRI Model

We now consider the LIFO *repeat-identical* M/G/1 queue (PRI) as introduced in [5]. In this model, whenever a customer is preempted, it retains its identical whole original service time S that it arrived with, as opposed to a new i.i.d. one when it gets sent to the front of the queue. This model is more complicated to analyze than PRD as we shall see.

We let V denote stationary workload, $p_0 = P(V = 0)$, N stationary number in system, and $\hat{B}, \hat{S}_r, \hat{S} = \hat{B} + \hat{S}_r$, the various stationary service times defined around Equation (1.3). (We will solve for p_0 later.)

Unlike $\hat{S}_r$ for PRD in Equation (1.6), deriving the distribution of such things for PRI appears to be much more challenging since the service time found in service now depends on its preemptions (if any) from the past. In stationarity, it is now $\hat{S}$ that gets placed back in queue when preempted; it is biased in a complicated way and not (in general) distributed as G or as one might guess via the inspection paradox.

1.3.1 Expressing p_0 in terms of $E(\hat{B})$

Using RCL on workload $V(t)$ similar to the derivation of Equation (1.6) yields the following equation

$$1 - p_0 = \rho + \lambda(1 - p_0)E(\hat{B}). \tag{1.15}$$

When a customer is preempted in stationarity, it is $\hat{S}$ that gets placed back in queue, causing workload to jump up by the net amount $\hat{B}$ (rate $\lambda(1 - p_0)$). Meanwhile every new customer (rate λ) brings a new i.i.d. S causing another jump up; $\lambda E(S) = \rho$.

Solving for p_0 yields

$$p_0 = \frac{1 - \rho - \lambda E(\hat{B})}{1 - \lambda E(\hat{B})}. \tag{1.16}$$

In order to compute $E(\hat{B})$ (in the next section) we will need to first determine the distribution of the number of times that an arrival enters service before completing service and departing.

Proposition 1.3.1. *For a fixed service time S distributed as G, of an arriving customer, let $\tau = \tau(S) \geq 1$ denote the total number of times that it enters service before completion, hence $K \overset{\text{def}}{=} \tau - 1 \geq 0$ denotes the total number of times it was preempted. Then, independent of S, letting $\{T_n : n \geq 1\}$ denote i.i.d. exponential random variables at rate λ, τ can be written in distribution as*

$$\tau \overset{\text{d}}{=} \min\{n \geq 1 : T_n \geq S\}. \tag{1.17}$$

Thus conditional on S, the distribution of τ is geometric with success probability $e^{-\lambda S}$ and hence

$$E(\tau) = E(E(\tau \mid S)) = E(e^{\lambda S}). \tag{1.18}$$

Proof. By the memoryless property of the exponential distribution in the Poisson arrival process, each time the service time enters service it will be preempted if, after an independent exponential (at rate λ) amount of time, it is still in service, and thus

$$\begin{aligned} P(\tau = 1) &= P(T_1 \geq S) \\ P(\tau = n) &= P(S > T_1, \ldots, S > T_{n-1},\ T_n \geq S),\ n \geq 2. \end{aligned}$$

Thus, conditional on S,

$$P(\tau = n \mid S) = (1 - e^{-\lambda S})^{n-1} e^{-\lambda S},\ n \geq 1,$$

and hence

$$E(\tau \mid S) = \frac{1}{e^{-\lambda S}} = e^{\lambda S};$$

Equation 1.18 follows. □

Remark 1.3.1. The proof of Proposition 1.3.1 contains the important fact/observation that for carrying out certain derivations (such as the derivation of the distribution of τ), we can, for each customer with their service time S, use a *new i.i.d.* $\{T_n\}$ *sequence, independent of* S, as a Poisson process just for that S, and immediately place the customer back in service with S each successive time $T_n < S$, instead of treating the interruptions as the arrival of new customers.

1.3.2 Computing $E(\hat{B})$ and determining stability for PRI

Here we introduce a discrete-time regenerative process method for deriving the various distributions of $\hat{B}$, $\hat{S}_r$, and $\hat{S}$. We focus here on its use for $\hat{B}$, but the method will be used later on for other derivations.

From Proposition 1.3.1, the expected number of times a customer is interrupted is given by $E(K) = E(e^{\lambda S}) - 1$, and we note that each time a service time S is interrupted, its *age* at that point is an i.i.d. length T_j distributed as exponential at rate λ conditional on $T_j < S$. Moreover, conditional on S and K, the T_j up to $j = K$ are i.i.d. distributed as that conditional distribution–given S. We will use this fact in what follows. From PASTA w.p. 1 it holds that (for all non-negative measurable functions f)

$$E(f(B)) = \lim_{t\to\infty} \frac{1}{t}\int_0^t f(B(s))ds = \lim_{n\to\infty} \frac{1}{n}\sum_{j=1}^{n} f(B(t_j-)).$$

To compute $E(\hat{B})$, we use $f(b) = b$ and focus on the customer average and let $N^1(n)$ denote the number of arrivals out of the first n for which $B(t_j-) > 0$, and express $E(\hat{B})$ as the w.p. 1 average

$$\begin{aligned}\lim_{n\to\infty}\frac{1}{N^1(n)}\sum_{j=1}^{n}B(t_j-)I\{B(t_j-)>0\} &= \lim_{n\to\infty}\frac{n}{N^1(n)}\times\frac{1}{n}\sum_{j=1}^{n}B(t_j-)I\{B(t_j-)>0\} \\ &= \frac{E(B\ ;\ B>0)}{P(B>0)} \\ &= E(B\mid B>0) \\ &= E(\hat{B}).\end{aligned}\tag{1.19}$$

Assuming that the long-run departure rate equals the arrival rate λ thus ensuring that each customer completes service, as is ensured by stability, each arriving customer will be included in the limit in Equation (1.19) until they finally depart.

As explained in Remark 1.3.1, we can thus take (reorder) each customer independently as being alone sequentially, one at a time, with their independent service time S and their own i.i.d. sequence of interarrival times $\{T_n\}$ and sum up their ages when preempted as if they form one regenerative cycle which ends at interarrival time K. Doing this sequentially for each of the i.i.d. service times $\{S_n : n \geq 1\}$, the cycles are i.i.d. and thus can be used to form a regenerative process in discrete time (equivalently a renewal reward process).

The sum over a cycle is the sum of the K ages $T_1 + \cdots + T_K$ (defined to be 0 if $K = 0$; the service was not preempted.) The cycle length is K.

Using classical regenerative process theory (see for example [15]), we thus can express, for any non-negative measurable function f,

$$E(f(\hat{B})) = \frac{E\Big[\sum_{j=1}^{K} f(T_j)\Big]}{E(K)}. \tag{1.20}$$

Recalling Equation 1.17, $K = \tau - 1$ is not a stopping but τ is. Thus we can compute the numerator of $E(f(\hat{B}))$ (expected sum over a cycle) using Wald's equation and subtracting the last piece:

$$E\Big[\sum_{j=1}^{K} f(T_j)\Big] = E(\tau)E(f(T)) - E(f(T_\tau)). \tag{1.21}$$

For $f(b) = b$ this then yields

$$E(\hat{B}) = \frac{\frac{1}{\lambda}E(e^{\lambda S}) - E(T_\tau)}{E(e^{\lambda S}) - 1}. \tag{1.22}$$

We next compute $E(T_\tau)$:

Lemma 1.3.1.

$$E(T_\tau) = \frac{1}{\mu} + \frac{1}{\lambda} = E(S + T). \tag{1.23}$$

Proof. By the definition of τ using Equation 1.17: Given S, T_τ has the conditional distribution of an i.i.d. (exponential λ) r.v. T conditional on $T > S$. Thus by the memoryless property T_τ given S is equal to S plus an independent exponential rate λ overshoot; $E(T_\tau \mid S) = S + E(T - S \mid T > S, S) = S + E(T)$, where we are using the memoryless property on T. Thus $E(T_\tau) = E(E(T_\tau \mid S)) = E(S + E(T)) = E(S) + E(T) = \frac{1}{\mu} + \frac{1}{\lambda}$. □

Inserting Equation (1.23) into Equation (1.22) then yields

Proposition 1.3.2. *For the M/G/1 PRI model*

$$E(\hat{B}) = \frac{1}{\lambda} - \frac{\frac{1}{\mu}}{E(e^{\lambda S}) - 1} \tag{1.24}$$

From Equation (1.24) for $E(\hat{B})$ we plug it in and obtain

Proposition 1.3.3. *For the M/G/1 PRI model*

$$p_0 = 2 - E(e^{\lambda S}).$$

Stability $0 < p_0 < 1$ *is thus*

$$E(e^{\lambda S}) < 2.$$

(In particular, S must have a finite moment generating function in a neighborhood of 0*; it must be light-tailed.)*

Remark 1.3.2. Note that in the M/M/1 case $E(e^{\lambda S}) = \frac{\mu}{\mu-\lambda}$; the stability condition is thus $\rho < 1/2$. Moreover, for the example given in Remark 1.2.3, stability becomes $\lambda < 1.98/199 \approx 0.0099$. These two simple examples illustrate just how different PRI is from PRD or the standard M/G/1 model.

1.3.3 Computing the distribution of $\hat{B}$ for PRI

Using the same Wald's equation method from Equation (1.21) we can determine the probability distribution of $\hat{B}$. For a fixed $x \geq 0$ we use $f(b) = I\{b > x\}$ to first compute the tail $P(\hat{B} > x)$, then obtain the cdf via $1 - P(\hat{B} > x)$. We also obtain the density, denoted by $f_{\hat{B}}(x)$.

Proposition 1.3.4. *For the M/G/1 PRI model, the cumulative distribution function (cdf), $F_{\hat{B}}(x)$ of $\hat{B}$ is given by*

$$F_{\hat{B}}(x) = \frac{E(e^{\lambda S}) - e^{-\lambda x}E(e^{\lambda S}; S > x) - G(x)}{E(e^{\lambda S}) - 1}, \ x \geq 0.$$

$\hat{B}$ always has a density (it is always a continuous r.v.) given by

$$f_{\hat{B}}(x) = \frac{\lambda e^{-\lambda x}E(e^{\lambda S}; S > x)}{E(e^{\lambda S}) - 1}, \ x \geq 0.$$

Proof. From Equation (1.21), we have

$$\begin{aligned} P(\hat{B} > x) &= \frac{E\left[\sum_{j=1}^{K+1} I\{T_j > x\}\right] - P(T_\tau > x)}{E(K)} && (1.25) \\ &= \frac{E(e^{\lambda S})e^{-\lambda x} - P(T_\tau > x)}{E(e^{\lambda S}) - 1}. && (1.26) \end{aligned}$$

Similar to the proof of Lemma 1.3.1 we obtain

$$\begin{aligned}
P(T_\tau > x) &= E(P(T_\tau > x \mid S)) \\
&= E(P(S + T > x \mid S)) \\
&= E(I\{S > x\}) + E(P(T > x - S; S \le x \mid S)) \\
&= E(I\{S > x\}) + E(e^{-\lambda(x-S)}; S \le x) \\
&= P(S > x) + e^{-\lambda x} E(e^{\lambda S}; S \le x).
\end{aligned}$$

Since $E(e^{\lambda S}) - E(e^{\lambda S}; S \le x) = E(e^{\lambda S}; S > x)$, when we subtract $P(T_\tau > x)$ in Equation (1.26) we obtain

$$P(\hat{B} > x) = \frac{e^{-\lambda x} E(e^{\lambda S}; S > x) - P(S > x)}{E(e^{\lambda S}) - 1}. \tag{1.27}$$

Using $F_{\hat{B}}(x) = 1 - P(\hat{B} > x)$ then yields the cdf. Because of the presence of $G(x) = P(S \le x)$, it appears that we need G to have a density $g(x)$ to ensure that $\hat{B}$ has one. We instead shall initially assume G has a density $g(x)$, so as to get the formula for $f_{\hat{B}}(x)$, then we will show that the existence of $g(x)$ is not required. Writing out

$$E(e^{\lambda S}; S \le x) = \int_0^x e^{\lambda s} g(s) ds,$$

we observe its derivative with respect to x is $e^{\lambda x} g(x)$ so $f_{\hat{B}}(x) = F'_{\hat{B}}(x)$ yields the density in our Proposition. But this formula for the density does not contain $g(x)$ in it. So we will integrate our density formula for $f_{\hat{B}}(x)$ and see that we always get back out our cdf formula thus proving that the existence $g(x)$ is not required. For $c = E(e^{\lambda S}) - 1$, due to non-negativity, we can use Fubini's (Tonelli's) Theorem via

$$\begin{aligned}\int_0^x cf_{\hat{B}}(y)dy &= \int_0^x \lambda e^{-\lambda y}E(e^{\lambda S})I\{S > y\}dy \\ &= E(e^{\lambda S}\int_0^{\min\{x,S\}} \lambda e^{-\lambda y}dy) \\ &= E(e^{\lambda S}(1 - e^{-\lambda \min\{x,S\}})) \\ &= E(e^{\lambda S}) - e^{-\lambda x}E(e^{\lambda S}; S > x) - P(S \leq x) \\ &= cF_{\hat{B}}(x);\end{aligned}$$

$f_{\hat{B}}(x)$ indeed is the density; $g(x)$ is not required. □

Interestingly, when G is exponential at rate μ (M/M/1 case), we get

$$f_{\hat{B}}(x) = \mu e^{-\mu x}; \tag{1.28}$$

$\hat{B}$ is exponential at rate μ, the same as G. Hence $E(\hat{B}) = \frac{1}{\mu}$.

1.3.4 Computing $E(\hat{S})$ and $E(\hat{S}_r)$ for PRI

We now compute $E(\hat{S})$ which represents the mean (in stationarity) of the total service time found in service by a customer who preempts them.

Proposition 1.3.5. *For the M/G/1 PRI model*

$$E(\hat{S}) = \frac{E(Se^{\lambda S}) - \frac{1}{\mu}}{E(e^{\lambda S}) - 1} \tag{1.29}$$

Proof. We compute using the method of Equation 1.20,

$$E(\hat{S}) = \frac{E\left[\sum_{j=1}^{K} S\right]}{E(K)} = \frac{E(KS)}{E(K)}.$$

Conditioning first on S yields $E(KS \mid S) = E(K \mid S)S = (e^{\lambda S} - 1)S = Se^{\lambda S} - S$. Taking expected values then dividing by $E(K) = E(e^{\lambda S}) - 1$ yields the result. □

Finally since $E(\hat{S}) = E(\hat{B}) + E(\hat{S}_r)$ we can compute the mean remaining service time $E(\hat{S}_r) = E(\hat{S}) - E(\hat{B})$ using Equations (1.29) and (1.24):

Proposition 1.3.6. *For the M/G/1 PRI model*

$$E(\hat{S}_r) = \frac{E(Se^{\lambda S})}{E(e^{\lambda S}) - 1} - \frac{1}{\lambda}. \tag{1.30}$$

When we apply Proposition 1.3.5 to the M/M/1 case, recalling Equation (1.28), we get

$$E(\hat{S}) = \frac{1}{\mu} + \frac{1}{\mu - \lambda},$$

which implies that

$$E(\hat{S}_r) = \frac{1}{\mu - \lambda};$$

in particular, this shows that $\hat{B}$ and $\hat{S}_r$ do not have the same distribution. We will determine the distribution of $\hat{S}_r$ and even the joint distribution of $(\hat{B}, \hat{S}_r)$ next.

1.3.5 Deriving the distributions of $\hat{S}$, $\hat{S}_r$ and the joint distribution of $(\hat{B}, \hat{S}_r)$ for PRI

Proposition 1.3.7. *The cumulative distribution function (cdf) of $\hat{S}$, $F_{\hat{S}}(x) = P(\hat{S} \leq x)$ is given by*

$$F_{\hat{S}}(x) = \frac{E(e^{\lambda S}; S \leq x) - G(x)}{E(e^{\lambda S}) - 1}, x \geq 0.$$

In particular, if G has a density $g(x)$ then so does $\hat{S}$ and it is given by (via differentiation)

$$f_{\hat{S}}(x) = \frac{g(x)(e^{\lambda x} - 1)}{E(e^{\lambda S}) - 1}, x \geq 0.$$

Proof. Following the proof of Proposition 1.3.5, we can express

$$F_{\hat{S}}(x) = \frac{E(E(K \mid S)I\{S \leq x\})}{E(e^{\lambda S}) - 1}, x \geq 0.$$

The numerator becomes

$$E(e^{\lambda S}; S \le x) - G(x);$$

the cdf follows. With density $g(x)$,

$$E(e^{\lambda S}; S \le x) = \int_0^x g(s) e^{\lambda s} ds,$$

thus differentiating the numerator yields

$$e^{\lambda x} g(x) - g(x) = g(x)(e^{\lambda x} - 1).$$

□

We now proceed to obtain the distribution of $\hat{S}_r$ and the joint distribution of $(\hat{B}, \hat{S}_r)$. $\overline{G}(x) = P(S > x)$, denotes the tail of G.

Proposition 1.3.8. *The cumulative distribution function (cdf) of $\hat{S}_r$, $F_{\hat{S}_r}(x) = P(\hat{S}_r \le x)$ is given by*

$$F_{\hat{S}_r}(x) = \frac{e^{\lambda x}\overline{G}(x) + E(e^{\lambda S}; S \le x) - 1}{E(e^{\lambda S}) - 1}, \ x \ge 0.$$

$\hat{S}_r$ always has a density (it is always a continuous r.v.) and it is given by

$$f_{\hat{S}_r}(x) = \frac{\lambda e^{\lambda x}\overline{G}(x)}{E(e^{\lambda S}) - 1}, x \ge 0.$$

Proof. Here we follow the Wald's equation method used in Proposition 1.3.2. We have

$$P(\hat{S}_r \le x) = \frac{E\left[\sum_{j=1}^{K} I\{S - T_j \le x\}\right]}{E(K)}.$$

We note that *conditional on S*, the sum up to $\tau = K + 1$, is a stopping time sum of i.i.d. random variables and hence has conditional (given S) expected value

$$
\begin{aligned}
E(\tau \mid S)P(T_j \geq S - x \mid S) &= e^{\lambda S} P(T_j \geq S - x \mid S) \\
&= e^{\lambda x} I\{S > x\} + e^{\lambda S} I\{S \leq x\},
\end{aligned}
$$

hence expected value $e^{\lambda x}\overline{G}(x) + E(e^{\lambda S}; S \leq x)$. We now need to subtract $P(S - T_\tau \leq x)$, which equals 1 because by definition of τ it must hold that $S \leq T_\tau$. This then yields then numerator which we then divide by $E(K) = E(e^{\lambda S}) - 1$. To obtain the density, we use the same trick that we used in the proof of Proposition 1.3.4: We first assume that G has a density $g(x)$ so as to obtain our formula for the density $f_{\hat{S}_r}(x)$ by differentiation, and then by integration show that $g(x)$ is not required.

□

In the M/M/1 case, the above yields that $\hat{S}_r$ is exponential at rate $\mu - \lambda$. Recalling that in this case we also have that $\hat{B}$ is exponential at rate μ, it begs the question of whether in this M/M/1 case, $\hat{S}_r$ and $\hat{B}$ are independent. We will answer that question in the affirmative next by computing, in general, the joint distribution of $(\hat{B}, \hat{S}_r)$.

Proposition 1.3.9. *For the M/G/1 PRI model,*

$$
P(\hat{B} > x,\ \hat{S}_r > y) = \frac{e^{-\lambda x} E(e^{\lambda S}; S > x + y) - e^{\lambda y} P(S > x + y)}{E(e^{\lambda S}) - 1},\ x \geq 0, y \geq 0.
$$

Thus, if G has a density g, then the joint density of $(\hat{B}, \hat{S}_r)$ exists and is given by

$$
f_{(\hat{B}, \hat{S}_r)}(x, y) = \frac{\lambda e^{\lambda y} g(x + y)}{E(e^{\lambda S}) - 1},\ x \geq 0, y \geq 0.
$$

Proof.

$$
P(\hat{B} > x,\ \hat{S}_r > y) = \frac{E\Big[\sum_{n=1}^{K+1} I\{T_n > x, S - T_n > y\}\Big] - P(T_\tau > x, S - T_\tau > y)}{E(K)}.
$$

We note that $P(T_\tau > x, S - T_\tau > y) = 0$ since $T_\tau > S$ by definition of τ. Thus the numerator conditional on S becomes

$$e^{\lambda S} P(x < T < S - y \mid S).$$

It must hold that $S > x + y$ or else $P(x < T < S - y \mid S) = 0$; thus we end up with

$$e^{\lambda S}(e^{-\lambda x} - e^{-\lambda(S-y)}) I\{S > x + y\} = e^{-\lambda x} e^{\lambda S} I\{S > x + y\} - e^{\lambda y} I\{S > x + y\}.$$

Taking expected values and dividing by $E(K)$ then yields the joint tail.

To obtain the density we compute it as $\frac{\partial}{\partial y}\frac{\partial}{\partial x} P(\hat{B} > x,\ \hat{S}_r > y)$: Letting $c = E(e^{\lambda S}) - 1$, we have (after a cancellation via $-e^{\lambda y} g(x + y) + e^{\lambda y} g(x + y) = 0$)

$$c\frac{\partial}{\partial x} P(\hat{B} > x,\ \hat{S}_r > y) = -\lambda e^{-\lambda x} E(e^{\lambda S}; S > x + y).$$

Then

$$-\frac{\partial}{\partial y} \lambda e^{-\lambda x} E(e^{\lambda S}; S > x + y) = \lambda e^{\lambda y} g(x + y).$$

□

As promised, in the M/M/1 case Proposition 1.3.9 yields

$$P(\hat{B} > x,\ \hat{S}_r > y) = e^{-\mu x} \times e^{-(\mu - \lambda) y};$$

$\hat{B}$ and $\hat{S}_r$ are independent exponentials. (The stability condition for the M/M/1 PRI is easily seen to be $\rho < 1/2$, i.e. $\lambda < \mu/2$.)

We now present some relationships between the tails of $\hat{S}_r$ and $\hat{B}$.

Corollary 1.3.1. *For a stable M/G/1 PRI model it holds that*

$$P(\hat{S}_r > x) = e^{\lambda x} P(\hat{B} > x),\ x \geq 0, \tag{1.31}$$

and hence

1. $\hat{S}_r$ is stochastically larger than $\hat{B}$; that is,

$$P(\hat{S}_r > x) \geq P(\hat{B} > x), \; x \geq 0.$$

2. If $P(S > x) > 0$ for all x, then

$$\frac{P(\hat{S}_r > x)}{P(\hat{B} > x)} \to \infty \text{ as } x \to \infty;$$

the tail of $\hat{S}_r$ is heavier the the tail of $\hat{B}$.

Proof. Using Proposition 1.3.9 to compute the individual tails, we have

$$P(\hat{S}_r > x) = \frac{E[e^{\lambda S}; S > x] - e^{\lambda x} P(S > x)}{E(e^{\lambda S}) - 1},$$

while

$$P(\hat{B} > x) = \frac{e^{-\lambda x} E[e^{\lambda S}; S > x] - P(S > x)}{E(e^{\lambda S}) - 1},$$

immediately yielding Equation (1.31), and from which the next two statements directly follow. □

1.3.6 $\hat{S}_r$ for PRI can be heavy-tailed even though S must be light-tailed

Recall from Proposition 1.3.3 that for PRI stability, S must be light-tailed; $E(e^{\lambda S}) < 2$. What about $\hat{S}_r$? We give here an example where $\hat{S}_r$ is heavy-tailed; $E(e^{s\hat{S}_r}) = \infty, \; s > 0$.

Let $M_X(s) = E(e^{sX}), \; s \geq 0$, for a non-negative r.v. X, and note that it alternatively can be computed via using the relation

$$\int_0^\infty e^{sx} P(X > x) dx = \frac{1}{s}(M_X(s) - 1). \tag{1.32}$$

Using the density for $\hat{S}_r$ from Proposition 1.3.8, together with Equation (1.32) yields

$$M_{\hat{S}_r}(s) = b(M_S(s + \lambda) - 1), \tag{1.33}$$

where

$$b = \frac{\lambda}{(s+\lambda)(E(e^{\lambda S})-1)}.$$

For our example, we thus need an S such that $M_S(\lambda) < 2$ but $M_S(\lambda + s) = \infty,\ s > 0$. We consider density functions for S of the form $g(x) = c(p)e^{-\lambda x}(1+x)^{-p},\ p > 1$, where $c(p)$ is the normalizing constant, given explicitly by $c(p) = (p-1)e^{\lambda^2/2}$. Clearly it always holds that $M_S(\lambda) < \infty$ while $M_S(\lambda + s) = \infty,\ s > 0$. But we need to ensure stability, that is, find a value of λ such that

$$M_S(\lambda) = c(p)\int_0^\infty (1+x)^{-p}dx = e^{\lambda^2/2} < 2.$$

We note that $e^{\lambda^2/2} < 2$ if and only if $\lambda < \sqrt{\ln(4)} \approx 1.177$.

Remark 1.3.3. By using the relation in Equation (1.33) and the fact that $P(\hat{B} > x) = e^{-\lambda x}P(\hat{S}_r > x)$ (recall Corollary 1.3.1), it is straightforward to prove that $M_{\hat{B}}(s) < \infty,\ s < \lambda$; stationary age is always light-tailed. But interestingly, $M_{\hat{B}}(\lambda) = 1 + \lambda E(\hat{S}_r)$, and hence involves $E(Se^{\lambda S})$, from Proposition 1.3.6, and thus could be infinite: $M_{\hat{B}}(\lambda) < \infty$ if and only if $E(\hat{S}_r) < \infty$; λ is a critical value.

1.3.7 Deriving w and $E(N_B)$ for PRI

As was derived for the PRD model in Section 1.2.4, and Remark 1.2.6, we can use Little's law ($l = \lambda w$) to obtain expected sojourn time, and expected number served in a busy period using Proposition 1.3.3:

$$w = \frac{1}{\lambda}\Big[\frac{1-p_0}{p_0}\Big] = \frac{1}{\lambda}\Big[\frac{1}{2-E(e^{\lambda S})} - 1\Big]. \tag{1.34}$$

$$E(N_B) = \frac{1}{p_0} = \frac{1}{2-E(e^{\lambda S})}. \tag{1.35}$$

We note that the above formula for w (using different methods) was given as Theorem 2 on Page 6 in [5].

1.3.8 Deriving the distribution of V for PRI

Here we provide a representation of V for PRI analogous to Proposition 1.2.3 for PRD. The proof (left out) is the same. In this case, $p_0 = 2 - E(e^{\lambda S})$, and the the i.i.d. service times $\{\hat{S}_j\}$, and the random variable $\hat{S}_r$, are taken as independent with their distributions provided in Propositions 1.3.7 and 1.3.8.

Proposition 1.3.10. *For the stable M/G/1 PRI model,*

$$(V \mid V > 0) \overset{\mathrm{d}}{=} \hat{S}_r + \sum_{j=1}^{Q} \hat{S}_j. \tag{1.36}$$

Thus the distribution of V*,* F_V*, is a mixture*

$$F_V = p_0\delta_0 + (1 - p_0)F_{\hat{V}}$$

where $F_{\hat{V}}$ *denotes the conditional distribution of* $(V \mid V > 0)$ *given in Equation 1.36, and* δ_0 *denotes the point mass at* 0*.*

From the above, we can now compute $E(V)$, for example, by utilizing the solved value of p_0 and the various expected value results and the fact that Q is geometric:

$$E(V) = (1 - p_0)\Big[E(\hat{S}_r) + E(Q)E(\hat{S})\Big]$$

$$= (E(e^{\lambda S}) - 1)\Big[\frac{E(Se^{\lambda S})}{E(e^{\lambda S}) - 1} - \frac{1}{\lambda} + \big[\frac{1}{2 - E(e^{\lambda S})} - 1\big]\big[\frac{E(Se^{\lambda S}) - \frac{1}{\mu}}{E(e^{\lambda S}) - 1}\big]\Big].$$

For the M/M/1 case this becomes

$$E(V) = \frac{\rho}{1 - \rho}\Big[\frac{1}{\mu - \lambda} + \frac{\rho}{1 - 2\rho}(\frac{1}{\mu} + \frac{1}{\mu - \lambda})\Big].$$

1.3.9 Heavy-traffic limits for stationary workload V for PRI

Here we include a result for PRI in the same spirit as in Theorem 1.2.1 but requiring more care in its proof. Of importance too is the fact that unlike PRD, there may not be a solution $\lambda_2 > \lambda$ to $p_0(\lambda) = 0$ (equivalently a solution to $E(e^{\lambda S}) = 2$); recall the counterexample given in Section 1.3.6. A proof of the following is given in the Appendix.

Theorem 1.3.1. *Suppose that there exists a $\lambda_2 > \lambda$ such that as $\lambda \uparrow \lambda_2$, $E(e^{\lambda S}) \rightarrow E(e^{\lambda_2 S}) = 2$, with $E(Se^{\lambda_2 S}) < \infty$. Then as $\lambda \uparrow \lambda_2$*

$$p_0 V \Longrightarrow exp(\alpha),$$

where $\alpha^{-1} = E(\hat{S}(\lambda_2)) = E(Se^{\lambda_2 S}) - 1/\mu$.

1.3.10 Tail asymptotics for stationary workload V for PRI

PRI requires for stability that S be light-tailed; $E(e^{\lambda S}) < 2$. So one might think at first sight that the tail asymptotics for PRI will be light-tailed. But lurking is the possibility that $\hat{S}_r$ and hence $\hat{S}$ (since $\hat{S} \geq \hat{S}_r$) can be heavy-tailed even though S is not as we illustrated earlier in Section 1.3.6. Given that in Proposition 1.3.10, both $\hat{S}_r$ and $\hat{S}$ are present, the tail asymptotics for PRI are not immediate. We will be satisfied to give a light-tailed asymptotic.

We start by giving a light-tailed result for work in queue (line) V_Q, then follow up by adding in $\hat{S}_r$ to get V, but showing that it does not contribute under suitable conditions and hence that $(V > x) \sim P(V_Q > x)$. (We assume for simplicity of proof a density for S to ensure a density for $\hat{S}$.) The proof is in the Appendix.

Proposition 1.3.11. *For a stable PRI model, suppose that S has a density $g(x)$, and there exists a (Cramér-Lundberg) constant $\gamma > 0$ such that*

$$E(e^{\gamma \hat{S}}) = (1 - p_0)^{-1}, \tag{1.37}$$

and additionally

$$R \overset{\text{def}}{=} E(\hat{S}e^{\gamma\hat{S}}) < \infty.$$

Then letting stationary work in queue (line) be denoted by (from Proposition 1.3.10) $V_Q = \sum_{i=1}^{Q} \hat{S}_i$,

$$P(V_Q > x) \sim Ce^{-\gamma x},$$

where

$$C = \frac{p_0}{R\gamma(1-p_0)}.$$

To use our V_Q tail asymptotics result to obtain our tail asymptotic for V we will need the following Lemma so as to handle adding in $\hat{S}_r$ (Proof in the Appendix).

Lemma 1.3.2. *For a stable M/G/1 PRI model, with G having a density g: If there exists constants* $c > 0$ *and* α *such that*

$$P(S > x) \sim ce^{-\alpha x}, \tag{1.38}$$

then $P(\hat{S}_r > x) \sim (1-p_0)^{-1}\frac{\lambda c}{\alpha-\lambda}e^{-(\alpha-\lambda)x}$, *and a Cramér-Lundberg constant* $\gamma > 0$ *exists for* $\hat{S}$,

$$E(e^{\gamma\hat{S}}) = (1-p_0)^{-1}, \tag{1.39}$$

and it additionally satisfies $\gamma < \alpha - \lambda$. *Hence as* $x \to \infty$,

$$\frac{P(\hat{S}_r > x)}{e^{-\gamma x}} \to 0.$$

Note that any such α in Equation 1.38 must satisfy $\alpha > \lambda$ because $E(e^{\lambda S}) < \infty$ by stability. Lemma 1.3.2 implies that if in addition to the assumptions of Proposition 1.3.11, S has an asymptotic exponential tail, then $\hat{S}_r$ also has an asymptotic exponential tail but with a rate that is smaller than γ and hence is asymptotically negligible in the tail of the sum $\hat{S}_r + V_Q$; the tail asymptotic of V_Q dominates. Together with Proposition 1.3.11 this immediately leads to

Corollary 1.3.2. *For a stable PRI model, suppose that S has a density $g(x)$, satisfies Equation 1.38, and also*

$$R \overset{\text{def}}{=} E(\hat{S}e^{\gamma\hat{S}}) < \infty.$$

Then

$$P(V > x) \sim P(V_Q > x) \sim Ce^{-\gamma x},$$

where

$$C = \frac{p_0}{R\gamma(1-p_0)},$$

with γ as the Cramér-Lundberg constant from Lemma 1.3.2.

1.3.11 An alternative proof of stability, and a representation for the distribution of sojourn time

We focus on PRI; PRD is similar (see Remark 1.3.4). Let τ be defined as in Proposition 1.3.1, with $K = \tau - 1$, and $\{T_n\}$ independent of service time S distributed as G. Define the *effective* service time as

$$\mathbb{S} = S + \sum_{j=1}^{K} T_j, \tag{1.40}$$

that is, S itself plus all the ages added on due to preemption; it is distributed as the total amount of work (servicing) required by a customer in the system. *Note that K and the T_j in the sum depend on S, and given S and $K = k$, the k random variables $T_1, \ldots, T_k$ are i.i.d. distributed as $(T \mid T < S)$.*

Utilizing Equations 1.21 and 1.23, yields

$$E(\mathbb{S}) = \frac{1}{\lambda}E(e^{\lambda S}) - \frac{1}{\lambda} = \frac{1}{\lambda}(E(e^{\lambda S}) - 1). \tag{1.41}$$

Because $E(\mathbb{S})$ is the average customer sojourn time *in service*, we apply Little's Law to the server to obtain:

Proposition 1.3.12. *The long-run proportion of time that the PRI server is busy is given by*

$$\min\{\lambda E(\mathbb{S}), 1\}.$$

Thus the system is stable when $\lambda E(\mathbb{S}) < 1$.

Plugging in Equation 1.41 and recalling Proposition 1.3.3, indeed yields, as it should, $p_0 = 1-\lambda E(\mathbb{S})$, and thus Proposition 1.3.12 provides an alternative derivation of p_0 and proof of stability.

We now give an implicit random variable representation for PRI sojourn time W, in the same spirit as for the classic $M/G/1$ busy periods (see for example Section 8.4, Page 388 in [6]), but more intricate. When a customer C_0 arrives to an empty PRI system (hence begins a busy period) they depart exactly at the end of the busy period hence ending it; W is distributed as a busy period. Now we re-write it: W = the effective service time of C_0 + an i.i.d. copy of W for each of the K preemptions that C_0 faced. Whenever C_0 is preempted, C_0 will reenter service after an i.i.d. copy W (the sojourn time of the customer who preempted them). Summarizing:

Proposition 1.3.13. *Let W denote a sojourn time, and let $\{W_i\}$ denote i.i.d. copies of W independent of K. Then*

$$W \overset{\mathrm{d}}{=} \mathbb{S} + \sum_{i=1}^{K} W_i. \tag{1.42}$$

Taking expected values yields $w = E(\mathbb{S}) + E(K)w$, and hence $w = \frac{E(\mathbb{S})}{1-E(K)}$. Plugging in the values for $E(\mathbb{S})$ (Equation 1.41) and $E(K) = E(e^{\lambda S}) - 1$ yields w as in Equation 1.34.

We next derive an implicit Laplace transform of W by using Proposition 1.3.13. Conditional on S, let $\hat{T}(s)$ denote the Laplace transform of $(T \mid T < S)$;

$$\hat{T}(s) = \frac{E(e^{-sT}; T < S, S)}{1 - e^{-\lambda S}} = \frac{\lambda}{\lambda + s}\Big[\frac{1 - e^{-(s+\lambda)S}}{1 - e^{-\lambda S}}\Big],\ s \geq 0.$$

Proposition 1.3.14. *For PRI, let $c(s) = E(e^{-sW})$ denote the Laplace transform of W. Then*

$$c(s) = E\Big[\frac{e^{-(s+\lambda)S}}{1 - c(s)\hat{T}(s)(1 - e^{-\lambda S})}\Big],\ s \geq 0. \tag{1.43}$$

Proof. From Equation 1.42, conditioning on K and S, and since $\mathbb{S}$ and $\sum_{i=1}^{K} W_i$ are conditionally independent, it follows that

$$E(e^{-sW} \mid S, K) = e^{-sS}(c(s)\hat{T}(s))^K.$$

For a geometric distribution, $P(N = n) = (1-p)^n p,\ n \geq 0$, its discrete moment generating function is computed (for $|z| \leq 1$) as $E(z^N) = p\sum_{n=0}^{\infty}(z(1-p))^n = \frac{p}{1-z(1-p)}$. Thus conditioning only on S yields, since then K is geometric with $p = e^{-\lambda S}$, and using $z = c(s)\hat{T}(s)$,

$$\begin{aligned} E(e^{-sW} \mid S) &= E(E(e^{-sW} \mid S, K) \mid S)) \\ &= e^{-sS}\frac{e^{-\lambda S}}{1 - c(s)\hat{T}(s)(1 - e^{-\lambda S})} \\ &= \frac{e^{-(s+\lambda)S}}{1 - c(s)\hat{T}(s)(1 - e^{-\lambda S})}. \end{aligned}$$

Taking expected values yields $c(s)$. □

It is easily shown that $w = -c'(0) = \frac{E(\mathbb{S})}{1-E(K)}$, and (a bit of work) one also obtains

$$E(W^2) = c''(0) = 2(1 - E[K])^{-1}(\frac{1}{\lambda} + w)\Big(\frac{1}{\lambda}E[e^{\lambda S} - 1] + (\frac{1}{\lambda} + w)E[(e^{\lambda S} - 1)^2] - E[Se^{\lambda S}]\Big),$$

where w is as in Equation 1.34. This agrees with what one gets if utilizing the other methods explained on the top of Page 9 in [5].

Remark 1.3.4. For PRD, we can define the effective service time distribution as

$$\mathbb{S} \overset{\mathrm{d}}{=} (S \mid S < T) + \sum_{j=1}^{K}(T_j \mid T_j < S_j), \tag{1.44}$$

where the $S, T, \{T_j\}, \{S_j\}$ and K are independent, and K (the number of preemptions) is geometric, $P(K = n) = (1-p)^n p,\ n \geq 0$, where $p = P(S \leq T) = E(e^{-\lambda S})$. One can then proceed to obtain for PRD, similar results as we did here for PRI.

Remark 1.3.5. We note that for both the PRD and PRI models, the stability condition can be expressed as

$E(K) = E(\tau - 1) < 1$, revealing a nice intuitive interpretation of stability:

> *The PRD and PRI models are stable if and only if the expected number of times a customer is preempted is strictly less than* 1.

Chapter 2: Static Pricing for Queueing Systems

In this chapter we analyze a queueing model involving prices. It is based on an upcoming article written in collaboration with Adam Elmachtoub.

2.1 Introduction

There are many business applications which can be naturally modeled as queueing systems with prices. One prototypical example is a firm which provides make-to-order goods such as custom electronics or vehicles. In this application, customers approach the firm with a request, the firm quotes a price, and, if the price is acceptable, the customer submits their order and joins a virtual queue. Another important modern application is in cloud computing. Here, the arrival process of customers seeking cloud computing resources is modulated by the price posted by the service provider, and customers accepting the price will put their jobs in queue to await service. The food delivery industry is also based on customers deciding whether or not to order based on delivery fees, and joining a virtual queue to wait for their food if they make a transaction. In such applications, a tradeoff often exists between the congestion the service provider allows and the revenue they earn. In the short term, one may be able to earn extra revenue by allowing a lot of congestion, but large queues may have a long term negative effect on demand. To manage this tradeoff, one approach is to price *dynamically*, changing the price as the state of the system changes. Dynamic pricing allows the service provider to control the congestion in the system without sacrificing much revenue: as the number of customers in the queue increases, the service provider can charge higher prices to reduce the rate of purchasing customers that need to be serviced. Indeed, in almost any theoretical pricing model, the optimal policy will be dynamic.

In practice, however, there are many downsides to dynamic pricing. The strategy may be unappealing to customers, and in response they might begin to exhibit strategic behavior. Implementing a dynamic pricing strategy also requires increased operational complexity and a very good demand model to be effective. In addition, the number of states can grow exponentially in the number of classes, making the optimal policy difficulty to compute and store. To address these issues, *static* pricing may be an attractive alternative. Static pricing is advantageous not only because it provides transparency for the customer, it is also simpler and more tractable for service providers to implement. However, when we restrict ourselves to static pricing policies, some loss of revenue or increased congestion may have to be endured. In this work, we quantify the magnitude of these losses in the worst-case, furnishing universal, non-asymptotic guarantees on the performance of static pricing policies relative to the optimal dynamic policy.

We consider a classic and general model which has applications in a wide range of areas. We assume that customers arrive according to a Poisson process with a rate depending on the price offered by the firm, service times are exponentially distributed, customers' valuation distributions are regular. We consider single and multi-server settings, as well as single and multiple classes of customers (each with different arrival, service, and valuation parameters). The service provider seeks to optimize two objectives, maximizing revenue while minimizing congestion. To reduce this multi-objective problem to a single-objective one, we assume the service provider seeks to maximize their expected revenue rate minus some constant times the average number of customers in the system. This constant captures the desired penalty the service provider wishes to associate with congestion in the system. For example, in the context of cloud computing, the congestion penalty may come from the cost of memory on the servers which hold the jobs in queue and from the reputational loss incurred when delays are too long. We also consider a related model in which the congestion penalty associated with a policy is proportional to the long-run average sojourn times under that policy.

The static pricing policies we consider use a single price for each customer class, and stop selling when the number of customers in the queue reaches a specified threshold. These threshold-static pricing strategies are simple to implement and allow us to directly trade off revenue and congestion via the threshold parameter. If we consider static policies which do not have a threshold, universal guarantees cannot be furnished as the performance can be very poor. We now summarize the main contributions of the paper.

- In Theorem 1, we establish that for a system with C servers and multiple customer classes, a static pricing policy can always attain at least $1 - \frac{\frac{1}{C!}(C^C)}{\sum_{n=0}^{C} \frac{1}{n!} C^n}$ of the multi-objective optimal value. This bound is smallest when there is a single server, in which case a static pricing policy can always attain at least half of the optimal value. We also observe that the advantage of dynamic pricing over static pricing vanishes as the number of servers increase: with 10 servers, our guarantee is at least 78.5%. These simple static pricing policies are of particular interest in the multi-class setting, where the optimal dynamic policy suffers from the curse of dimensionality.

- In the single-class case with an arbitrary number of servers, Theorem 2 furnishes bi-criteria approximations on the revenue and congestion objectives as a function of the cutoff threshold (the maximum number of customers allowed in the system). Service providers can use this result to understand the benefit of additional servers and changing the cutoff threshold in the context of managing revenue and congestion with static pricing.

- In the single-server and single-class case, which corresponds to the classic M/M/1 queue, Theorem 3 gives *closed-form* guarantees on the approximation ratios of our static policies compared to the dynamic optimal. For example, we provide bi-criteria approximations of $(0.5, 1), (0.66, 1.16), (0.75, 1.54)$ and $(0.8, 2)$ for the revenue and congestion objectives, respectively. Moreover, we provide a class of instances proving the tightness of our analysis.

- We report the results of numerical experiments over a wide test bed demonstrating the empirical performance of our policies. These experiments illustrate the high performance of static pricing policies: in the worst-case instance we found, the optimal static pricing policy is able to recover 89.70% of the optimal multi-objective value.

- Theorem 4 extends our results for a related model in which we penalize long-run average sojourn times instead of average occupancy to measure congestion. In this case, we give bi-criteria guarantees for the revenue and congestion objectives.

2.1.1 Literature Review

To understand where the present work lies in this landscape, it is helpful to break the related literature down along three different dimensions: (i) whether pricing is static or dynamic, (ii) whether the analysis is asymptotic or exact, and (iii) the manner in which the state of the queue impacts the arrival process. Along dimension (iii), the literature has primarily considered three different possibilities (in order of their first appearance):

- The state of the queue does not directly impact the arrival process, which we call the Low queue since it was first introduced in [16].
- Observable queues, in which customers observe the state of the queue and use this information to inform their decisions.
- Unobservable queues, in which customers do not observe the state of the queue but have an expectation based on past experience of what their waiting time will be.

In the present work, we provide exact (non-asymptotic) analysis on the trade-off between static and dynamic pricing for Low queues.

To our knowledge, the model we consider was first introduced in [16] and followed up with [17], works in which the author studies the structure of the optimal dynamic pricing policy and gives an algorithm to compute it assuming either a finite or compact set of possible prices, respectively. Low's work originally allowed for only one class of customers; we generalize his model to allow for customers which are differentiated by their demand functions and service rates. The Low queue, albeit restricted to a single-server, was again considered in [18]. They consider two-price policies and, for a given pair of prices, compute the optimal policy under two different criteria: either maximizing revenue rate for a given upper bound on congestion, or minimizing congestion for a given lower bound on revenue rate. This contrasts with our approach in which we optimize a linear combination of the two objectives and allow a different price for each state in the dynamic policy. More recently, [19] considers the problem of optimizing the static price in our model in the case of a single server. They also consider the case when the cutoff threshold (the maximum number of customers allowed in the system) is a decision variable as well, a perspective which is important to the present work, and even provide some results for general service times. However, their work is focused solely on static pricing policies in the case of a single server and single customer class; in contrast, our aim in the present work is to compare the performance of static and dynamic policies with an arbitrary number of servers and customer classes. One more work which extends Low's model is [20], which considers the case when the service rate is variable and can be picked by the service provider. They show numerically that in that context dynamic pricing has significant gains over static pricing.

We now discuss works which consider observable queues. The foundational and seminal work in this line is [21], which introduced observable queues and analyzed them under the assumptions of deterministic customer valuations and static pricing. The most salient point emerging from their exact analysis is that the static price which maximizes social welfare is strictly smaller than the revenue-maximizing price. The previous three works do not allow for dynamic pricing, a case first studied in [22], in which the authors generalize the observable model in [21] (and their work in [23]) to allow for state-dependent pricing. Their work shows that a threshold policy is optimal but does not give explicit expressions for the threshold. The work of [24] provides explicit expressions for the optimal threshold in terms of the Lambert-W function. [25] analyze the value of dynamic pricing in the asymptotic regime, where the arrival and service rates are taken to infinity. They show that dynamic pricing has a significant effect on reducing the impact of stochasticity on accrued revenue, though most of the benefit can be attained by simple two-price policies. Examples of related work in this line are [26], [27], [28], [29], and [30].

We now discuss a few which assume the queue is unobservable. The earliest study of this case is [31], where they show that the revenue maximizing and social welfare maximizing prices are the same. They also generalize the model to allow for multiple classes of customers. Another work in the unobservable case is [32] which considers a multi-class model and provides static pricing mechanisms which maximize total social welfare. [33] carries out an exact analysis assuming an unobservable queue with static prices to identify when a subscription-based pricing scheme (in which users pay a monthly fee for an arbitrary number of uses) may outperform a per-use approach. Again assuming static pricing, [34] analyzes the loss of optimality that must be suffered when a price must be set without knowledge of the arrival rate.

The preceding works and the literature more generally have mostly dealt with the case where the queue is either completely observed or totally opaque to the customer; another more recent avenue is the question of interpolating between these two extremes by investigating optimal signaling, first addressed in [35]. In this work, the authors show that with respect to a certain objective, the optimal signaling policy falls in the middle of the two extremes.

The foregoing works are all unified in that the underlying queueing model is essentially an M/M/C queue. We continue by reviewing works which consider models slightly further afield. Most similar to the present work is [36], which provides universal performance guarantees for static pricing policies when the model is an Erlang loss system. In the same vein and motivated by a similar question is the work [37]. They introduce a highly stylized model involving pricing and queues that is tailored for telecommunication applications. As in the present work, they furnish a lower bound on the ratio between the revenue earned from their simple pricing rule and the maximum possible, which they denote the “price of simplicity". Their model assumes no disutility is incurred for long delays in service. [38] shows that when such disutility is assumed to be present, the “price of simplicity" can be very high. Since in typical applications high delays will be undesirable, this insight demonstrates the importance of considering congestion in these models and motivates our centering of it in the present work. Examples of other important works which analyze pricing in queueing models that are tailored for more specific applications are [39], [40], [41], [42], [43], [44], and [45].

2.1.2 Organization

Our work is organized as follows. In Section 2.2, we provide the general model and describe our static pricing policies. In Section 2.3, we provide approximation guarantees for our static policy with respect to the optimal multi-objective value. We give refined bi-criteria approximation guarantees on revenue and congestion in the single-class case in Section 2.4. In the single-server case, we are able to provide closed-form expressions for these bounds in Section 2.5. We carry out numerical experiments in Section 2.6. We extend our results to consider the expected sojourn times as the congestion objective in Section 2.7. Finally, we conclude our paper and offer some future directions in Section 2.8.

2.2 Model

We consider a model where a service provider seeks to optimize a linear combination of the expected revenue and the expected congestion of the system. There are C servers that serve M types of price-sensitive customers in a first-come, first-serve manner. Customers of type $m \in \{1,2,\ldots,M\}$ arrive according to a Poisson process with rate $\Lambda_m > 0$. Each customer of type m has a valuation drawn i.i.d. from a distribution F_m. We make the standard assumption in the revenue management literature that F_m is a regular distribution for each m (also known as Myerson's regularity, see [46]). When a customer of type m arrives, the provider observes their type and offers service at some price p^m which may depend on the state of the system; the customer decides to join the queue if their valuation is at least p^m. This gives rise to the *effective* arrival rate of customers of type m, which we denote by $\lambda^m(p^m) := \Lambda_m \overline{F}_m(p^m)$. We assume that there is a one-to-one correspondence between prices p^m and effective arrival rates λ^m so that $\lambda^m(p^m)$ has a unique inverse, denoted by $p^m(\lambda^m)$. Thus we can equivalently view the effective arrival rates λ^m as the decision variables, an insight which is critical to our analysis. By our assumption that F_m is regular, we have that $\lambda^m p^m(\lambda^m)$ is concave for each m.

A customer of type m has a service time which follows an exponential distribution with mean

$1/\mu^m$. We assume that the service times are i.i.d. across customers and independent of customer valuations. We also assume the firm incurs a congestion penalty of c_m per time unit per customer of type m in the system. The value c_m can be viewed as a parameter which captures the magnitude of the penalty the service provider wishes to associate with congestion of type m customers. This penalty structure amounts to penalizing a policy proportionally to the expected number of customers of each type in the system. A related quantity, the expected sojourn time of a customer of type m in steady-state, is also a natural choice for a penalty in many applications. We will consider this model as well in Section 2.7.

We now describe this model as a Markov Decision Process by describing its state and action spaces. By the memoryless property of the arrival and service processes, the state of the system is the order and types of all customers in the system (the arrival times and time in service can be ignored.) Thus, the state space S can be formulated as the set of all tuples of arbitrary length that can be formed by picking elements with replacement from the set of customer types $\{1, 2, \ldots, m\}$. For convenience, for $s \in S$, we let s_m denote the number of type m customers in state s and we let $s(i)$ denote the type of the ith customer in line. Turning now to the action space A, we assume that we observe the type of the arriving customer before offering a price, so an arbitrary pricing policy π will be represented by a mapping from the state space S to the set of all non-negative M-tuples of real numbers. In other words, for each state $s \in S$, $\pi(s)$ is an M-tuple of non-negative numbers $(p^1, p^2, \ldots, p^M)$ which gives the prices p^m offered when a customer of type m arrives in state s.

The set of admissible policies $\mathbf{\Pi}$ is the set of non-anticipating policies and those which induce stable queues. We can restrict ourselves, without loss of optimality, to considering policies which induce stable queues, as those which do not induce stability incur infinitely large costs. We denote the optimal policy by π^*. The stationary probabilities of a policy $\pi \in \mathbf{\Pi}$ are denoted by $\mathbb{P}_s(\pi)$ for each state $s \in S$. We will only have closed-form expressions for $\mathbb{P}_s(\pi)$ in the single-class case.

The static policies we consider are those non-anticipating policies which fix a price for each customer class up to a certain occupancy threshold (potentially infinite) and then disallow arrivals after that point. In other words, we consider the class of policies $\mathbf{\Pi}^s \subset \mathbf{\Pi}$ that fix an arrival rate λ^m

for class m customers in every state at or below a certain cutoff point $\gamma \in \{0, 1, 2, \dots\}$ and blocks arrivals after that point by setting the arrival rates to 0 (infinite price). An arbitrary static policy is notated $\pi^{(\lambda^1,\dots,\lambda^M),\gamma}$, which refers to the policy which picks the arrival rate λ^m for customers of type m whenever there are at most than γ customers in the system, and otherwise sets all arrival rates to 0. In the single-class case, we simply write $\pi^{\lambda,\gamma}$.

2.2.1 Objectives

For a given policy $\pi \in \mathbf{\Pi}$, we define $\mathcal{R}(\pi)$ as the average expected revenue rate attained attained by π, i.e.,

$$\mathcal{R}(\pi) := \sum_{s\in S} \sum_{m\in M} p_s^m \lambda^m(p_s^m) \mathbb{P}_s(\pi).$$

Similarly, we define $C(\pi)$ as the "congestion penalty" incurred by policy π. When the congestion penalty is proportional to the number of customers in the system, we have

$$C(\pi) := \sum_{s\in S} \sum_{m\in M} c_m s_m \mathbb{P}_s(\pi) = \sum_{m\in M} c_m \sum_{s\in S} s_m \mathbb{P}_s(\pi)$$

where we recall that s_m denotes the number of type m customers in the system in state s. We consider the congestion penalty corresponding to expected sojourn time in Section 2.7.

We also introduce the stationary distribution of type m customers induced by policy π, which we denote $L^m(\pi)$. We can compute its distribution using the stationary probabilities as

$$P(L^m(\pi) = k) = \sum_{s\in S : s_m = k} \mathbb{P}_s(\pi).$$

We can then compute the expectation as

$$E[L^m(\pi)] = \sum_{k=0}^{\infty} k P(L^m(\pi) = k) = \sum_{k=0}^{\infty} k \sum_{s\in S : s_m = k} \mathbb{P}_s(\pi) = \sum_{k=0}^{\infty} \sum_{s\in S : s_m = k} s_m \mathbb{P}_s(\pi) = \sum_{s\in S} s_m \mathbb{P}_s(\pi).$$

Thus we can express also express our congestion penalty in terms of this expectation, i.e.,

$$C(\pi) = \sum_{m \in M} c_m E[L^m(\pi)]$$

Now we can write our overall objective function, $\mathcal{Z}(\pi)$, as

$$\mathcal{Z}(\pi) := \mathcal{R}(\pi) - C(\pi) = \sum_{s \in S} \left(\sum_{m \in M} p_s^m \lambda^m(p_s^m) - c_m s_m \right) \mathbb{P}_s(\pi).$$

The optimal policy π^* can now be formalized as

$$\pi^* := \arg\max_{\pi \in \Pi} \mathcal{Z}(\pi).$$

We denote by λ_s^{m*} the optimal arrival rate picked in state s for class m customers by policy π^*.

2.2.2 Little's Law and Static Policies

The well-known Little's Law is integral to our work and helps motivate the static policy construction we furnish to prove our guarantees. Suppose we fix an instance of our model and consider an optimal policy π^*. For each customer class m, the policy π^* induces a distribution $L^m(\pi^*)$ of the number of type m customers in the system in stationarity. Moreover, when a customer of type m joins the system in stationarity, they will experience a sojourn time distributed according to a random variable we denote by $W^m(\pi^*)$. Little's Law allows us to relate the first moments of these two distributions, i.e.,

$$E[L^m(\pi^*)] = \tilde{\lambda}^m E[W^m(\pi^*)]$$

where $\tilde{\lambda}^m$ denotes the long-run average arrival rate of customers of type m under π^*. For convenience, we use $\tilde{\lambda}^m$ to introduce the typical traffic intensity notation: for each customer type m, we can compute its traffic intensity as $\tilde{\rho}^m := \frac{\tilde{\lambda}^m}{\mu^m}$.

We can compute this average arrival rate of type m customers under the optimal policy π^* using the stationary probabilities, i.e.,

$$\tilde{\lambda}^m = \sum_{s \in S} \lambda_s^{m*} \mathbb{P}_s(\pi^*)$$

Our key idea is to use these average arrival rates to construct good static policies. In any state in which our static policy sells, we pick the arrival rate $\tilde{\lambda}^m$ for type m customers. We note that this differs than the static price used in [36] to attain performance guarantees for the Erlang loss model, and the analysis is quite different due to our mixed objective function $\mathcal{Z}(\pi)$.

2.3 Multi-class Static Pricing Guarantees

We now present our first main result in Theorem 1 below, which gives approximation guarantees for static pricing for arbitrary numbers of servers C and classes M. We choose the static arrival rates to be $\tilde{\lambda}^m$ and set the cutoff threshold to be $C - 1$, so there are at most C customers in the system.

Theorem 1. *The static pricing policy* $\pi^{(\tilde{\lambda}^1,\ldots,\tilde{\lambda}^M),C-1}$ *guarantees at least* $1 - \frac{\frac{1}{C!}(C^C)}{\sum_{n=0}^{C} \frac{1}{n!} C^n}$ *of the value of the optimal dynamic pricing policy. Equivalently,*

$$\mathcal{Z}(\pi^{(\tilde{\lambda}^1,\ldots,\tilde{\lambda}^M),C-1}) \geq \left(1 - \frac{\frac{1}{C!} C^C}{\sum_{n=0}^{C} \frac{1}{n!} C^n}\right) \mathcal{Z}(\pi^*)$$

Moreover, our analysis is tight when $C = 1$.

The tightness of Theorem 1 will be shown using a class of instances described in Section 2.5.2.

One insight we can derive immediately from Theorem 1 is how the performance of static pricing changes as we increase the number of servers. We observe the guarantees are strictly improving: for example, if $C = 10$, the theorem ensures the existence of a static policy which attains at least 78.5% of the optimal value (instead of the 50% guarantee for $C = 1$). We note that one can optimize the choice of the static price and cutoff threshold, and the result still holds since our proof is by construction. This implies that finding the optimal dynamic policy is not a precursor to finding a good static policy.

To prove Theorem 1, we use two lemmas. Lemma 1 bounds the sum of the traffic intensities of stable policies, and Lemma 2 helps us relate the revenue rate of the optimal policy to the revenue rate of the static policy we construct.

Lemma 1. *For a system with C servers and any stable policy π, we have $\sum_{m \in M} \tilde{\rho}^m(\pi) \leq C$.*

Proof. Proof. Our proof is by contradiction. Suppose there exists a policy π which induces a proper stationary distribution for which $\sum_{m \in M} \tilde{\rho}^m > C$. We first uniformize the continuous-time Markov Chain in the standard way with the dominating transition rate $\sigma = \sum_{i=1}^{M} \Lambda_i + C \max(\mu^1, \ldots, \mu^M)$ and consider the associated discrete-time Markov Chain, which is characterized by a transition matrix we denote by P. By construction, this uniformized discrete-time Markov Chain has the same stationary distribution given by $\mathbb{P}_s(\pi)$. Suppose we sample $a \sim \mathbb{P}_s(\pi)$ and simulate one transition to arrive at a state b. By stationarity, the distribution of b must be the same as that of a, which in particular implies that the expected workloads present in the two states are the same. We let $E[V(s)]$ denote the expected workload in a state s, which is easily computed as the sum of all expected service times of customers in the system in state s.

To arrive at the desired contradiction, we now compute the expected difference in workloads between states a and b by conditioning on a and b, i.e.,

$$
\begin{aligned}
E[V(b) - V(a)] &= \sum_{s \in S} E[V(b) - V(a) | a = s] \mathbb{P}_s(\pi) \\
&= \sum_{s \in S} \sum_{s' \in S} E[V(b) - V(s) | b = s'] P_{s,s'} \mathbb{P}_s(\pi) \\
&= \sum_{s \in S} \left(\sum_{m \in M} \frac{1}{\mu^m} \frac{\lambda_s^m}{\sigma} - \sum_{i=1}^{\min(|s|,C)} \frac{1}{\mu_{s(i)}} \frac{\mu_{s(i)}}{\sigma} \right) \mathbb{P}_s(\pi) \\
&= \frac{1}{\sigma} \sum_{s \in S} \left(\left(\sum_{m \in M} \frac{\lambda_s^m}{\mu^m} \right) - \min(|s|, C) \right) \mathbb{P}_s(\pi) \\
&\geq \frac{1}{\sigma} \sum_{s \in S} (\sum_{m \in M} \tilde{\rho}^m - C) > 0.
\end{aligned}
$$

where we recall that $s(i)$ denotes the ith element in the state tuple s, i.e. the type of the ith customer in service. The first and second equalities follow from conditioning on the initial state s_0 and the next state s_1, respectively. The third equality comes from analyzing the difference in expected workload that arises depending on the transition. The difference will either be $\frac{1}{\mu^m}$ if a type m customer joins the queue, which occurs with probability $\frac{\lambda_s^m}{\sigma}$, or will be $-\frac{1}{\mu_{s(i)}}$ if the ith customer completes service, which occurs with probability $\frac{\mu_{s(i)}}{\sigma}$. The first inequality follows from the definition of $\tilde{\rho}^m$ and the fact that there at most C customers in service. The second inequality follows from assuming the contradiction of the lemma. After one transition, we see a strictly positive increase in expected workload, which contradicts stationarity and thus completes the proof. □

We now give one more key lemma which is used to relate the revenue rate of the optimal policy to the revenue rate attained by the static policies we construct.

Lemma 2. *For any number of servers C and customer classes M, we have*

$$\mathcal{R}(\pi^*) \leq \sum_{m \in M} \tilde{\lambda}^m p^m(\tilde{\lambda}^m).$$

Proof. Proof. First, we recall that since the valuation distribution is regular, the revenue rate function $\lambda^m p^m(\lambda^m)$ is concave in λ^m. By applying Jensen's inequality to the revenue rate function and recalling that $\tilde{\lambda}^m = \sum_{s \in S} \lambda_s^m \mathbb{P}_s(\pi^*)$, we can bound expected revenue rate of the optimal policy by

$$\mathcal{R}(\pi^*) = \sum_{s \in S} \sum_{m \in M} \lambda_s^m p^m(\lambda_s^m) \mathbb{P}_s(\pi^*) = \sum_{m \in M} \sum_{s \in S} \lambda_s^m p^m(\lambda_s^m) \mathbb{P}_s(\pi^*) \leq \sum_{m \in M} \tilde{\lambda}^m p^m(\tilde{\lambda}^m).$$

□

2.3.1 Proof of Theorem 1

The revenue rate of the static policy is constant whenever we are selling, which is when the number of customers is strictly less than γ, and 0 when the number of customers is γ. Thus the expected revenue of the static policy is

$$\begin{aligned}\mathcal{R}(\pi^{(\tilde{\lambda}^1,\ldots,\tilde{\lambda}^M),C-1}) &= \sum_{m\in M} \tilde{\lambda}^m p^m(\tilde{\lambda}^m)\left(1 - \sum_{s\in S:|s|=C} \mathbb{P}_s(\pi^{(\tilde{\lambda}^1,\ldots,\tilde{\lambda}^M),C-1})\right)\\ &\geq \mathcal{R}(\pi^*)\left(1 - \sum_{s\in S:|s|=C} \mathbb{P}_s(\pi^{(\tilde{\lambda}^1,\ldots,\tilde{\lambda}^M),C-1})\right), \qquad (2.1)\end{aligned}$$

where the inequality follows from Lemma 2.

We now turn to the congestion penalty. Suppose that L^{m*} denotes the number of class m customers in the system under π^*, and W^{m*} is the respective sojourn time. We can compute the congestion cost of π^* as

$$C(\pi^*) = \sum_{m\in M} c_m E[L^{m*}] = \sum_{m\in M} c_m \tilde{\lambda}^m E[W^{m*}] \geq \sum_{m\in M} c_m \tilde{\rho}^m, \qquad (2.2)$$

where the second equality follows from Little's Law and the inequality follows by observing that the average sojourn time is at least the average service time.

Suppose that $\tilde{L}^m$ denotes the number of class m customers in the system under π^*, and $\tilde{W}^m$ is the respective sojourn time. The congestion cost of the static policy can be bounded by

$$\begin{aligned}C(\pi^{(\tilde{\lambda}^1,\ldots,\tilde{\lambda}^M),C-1}) &= \sum_{m\in M} c_m E[\tilde{L}^m]\\ &= \sum_{m\in M} c_m \left(1 - \sum_{s\in S:|s|=C} \mathbb{P}_s(\pi^{(\tilde{\lambda}^1,\ldots,\tilde{\lambda}^M),C-1})\right)\tilde{\lambda}^m E[\tilde{W}^m]\\ &= \sum_{m\in M} c_m \tilde{\rho}^m \left(1 - \sum_{s\in S:|s|=C} \mathbb{P}_s(\pi^{(\tilde{\lambda}^1,\ldots,\tilde{\lambda}^M),C-1})\right)\\ &\leq C(\pi^*)\left(1 - \sum_{s\in S:|s|=C} \mathbb{P}_s(\pi^{(\tilde{\lambda}^1,\ldots,\tilde{\lambda}^M),C-1})\right), \qquad (2.3)\end{aligned}$$

where the second equality follows from Little's law and the fact that the average arrival rate of type m customers for the static policy is $\tilde{\lambda}^m$ times the probability of selling, which is the probability of having less than $C = \gamma + 1$ customers. The third equality follows since there is never any wait time when $\gamma = C - 1$, and thus $E[\tilde{W}^m] = \frac{1}{\mu^m}$. The inequality follows from (2.2).

As proved in [47], after we pick fixed arrival rates for each type which do not vary by state, the stationary probabilities can be written

$$\mathbb{P}_s(\pi) = \frac{\prod_{m \in M} \frac{(\rho^m)^{s_m}}{s_m!}}{\sum_{s \in S} \prod_{m \in M} \frac{(\rho^m)^{s_m}}{s_m!}}$$

where $\rho^m = \frac{\lambda^m}{\mu^m}$ and s_m denotes the number of type m customers in the system in state s. Using this expression, and after some algebraic manipulation, we can express

$$\sum_{s \in S : |s| = C} \mathbb{P}_s(\pi^{(\tilde{\lambda}^1, \ldots, \tilde{\lambda}^M), C-1}) = \frac{\frac{1}{C!}((\sum_{m \in M} \tilde{\rho}^m)^C)}{\sum_{n=0}^{C} \frac{1}{n!}(\sum_{m \in M} \tilde{\rho}^m)^n} \tag{2.4}$$

which allows us to observe that this blocking probability is determined by the value of $\sum_{m \in M} \tilde{\rho}^m$.

Finally, we can bound the cost of the static policy by

$$\begin{aligned}
\mathcal{Z}(\pi^{(\tilde{\lambda}^1, \ldots, \tilde{\lambda}^M), C-1}) &= \mathcal{R}(\pi^{(\tilde{\lambda}^1, \ldots, \tilde{\lambda}^M), C-1}) - C(\pi^{(\tilde{\lambda}^1, \ldots, \tilde{\lambda}^M), C-1}) \\
&\geq \mathcal{Z}(\pi^*)(1 - \sum_{s \in S : |s| = C} \mathbb{P}_s(\pi^{(\tilde{\lambda}^1, \ldots, \tilde{\lambda}^M), C-1})) \\
&= \mathcal{Z}(\pi^*)\left(1 - \frac{\frac{1}{C!}((\sum_{m \in M} \tilde{\rho}^m)^C)}{\sum_{n=0}^{C} \frac{1}{n!}(\sum_{m \in M} \tilde{\rho}^m)^n}\right) \\
&\geq \mathcal{Z}(\pi^*)\left(1 - \frac{\frac{1}{C!}(C^C)}{\sum_{n=0}^{C} \frac{1}{n!} C^n}\right)
\end{aligned}$$

where the first inequality follows from (2.1) and (2.3), the second equality follows from (2.4), and the last inequality follows from Lemma 1.

2.4 Single-class Static Pricing Guarantees

In this section, we give bi-criteria approximation guarantees for static pricing with an arbitrary number of servers C and a single customer class in Theorem 2 below. Whereas Theorem 1 considers policies with a particular cutoff (namely $\gamma = C - 1$), Theorem 2 leverages this cutoff point γ as an additional parameter to index a family of guarantees.

Before we proceed, we note that in the single-class case we can make a simplifying assumption. The data for our model depends upon the units we choose for currency and time. By picking the units in a certain way, we can transform an arbitrary instance into one with $\mu = c = 1$ without impacting the ratios of interest. Hence we assume without loss of generality that $\mu = c = 1$. In this case, we can now express the stationary probabilities in a standard closed-form (see e.g. [48])

$$\mathbb{P}_i(\pi) = \frac{\prod_{k=0}^{i-1} \frac{\lambda_k}{\mu_{k+1}}}{1 + \sum_{n=0}^{\infty} \prod_{k=0}^{n} \frac{\lambda_k}{\mu_{k+1}}} \tag{2.5}$$

where $\mu_k = \min(C, k)$.

For a fixed number of servers C, we analyze static policies of the form $\pi^{\lambda,\gamma}$ for $\gamma \geq C - 1$. For such policies, it is a straightforward exercise to arrive at the following expression for the blocking probability

$$\mathbb{P}_{\gamma+1}(\pi^{\lambda,\gamma}) = \frac{\frac{\lambda^{\gamma+1}}{C!C^{\gamma+1-C}}}{\sum_{l=0}^{C} \frac{1}{l!}\lambda^l + \frac{1}{C!}\lambda^C \sum_{l=1}^{\gamma+1-C} \frac{1}{C^l}\lambda^l}. \tag{2.6}$$

From this expression we see the intuitive fact that the blocking probability is increasing in the arrival rate λ.

We now proceed with our second main result which provides approximation guarantees for static pricing for an arbitrary number of servers C. We again choose the static arrival rate to be $\tilde{\lambda}$ but allow the cutoff point γ to vary.

Theorem 2. *The static pricing policy $\pi^{\tilde{\lambda},\gamma}$ guarantees at least $1 - \frac{\frac{C^{\gamma+1}}{C!C^{\gamma+1-C}}}{\sum_{l=0}^{C} \frac{1}{l!}C^l + \frac{1}{C!}C^C(\gamma+1-C)}$ of the revenue rate and incurs a congestion penalty less than $g(\gamma, C)$ times that of the optimal dynamic pricing policy, where*

$$g(\gamma, C) = \max_{\tilde{\lambda} \in [0,C]} \frac{\sum_{i=1}^{\gamma+1} i\mathbb{P}_i(\pi^{\tilde{\lambda},\gamma})}{\tilde{\lambda}}.$$

Equivalently,

$$\mathcal{R}(\pi^{\tilde{\lambda},\gamma}) \geq \left(1 - \frac{\frac{C^{\gamma+1}}{C!C^{\gamma+1-C}}}{\sum_{l=0}^{C} \frac{1}{l!}C^l + \frac{1}{C!}C^C(\gamma + 1 - C)}\right)\mathcal{R}(\pi^*)$$

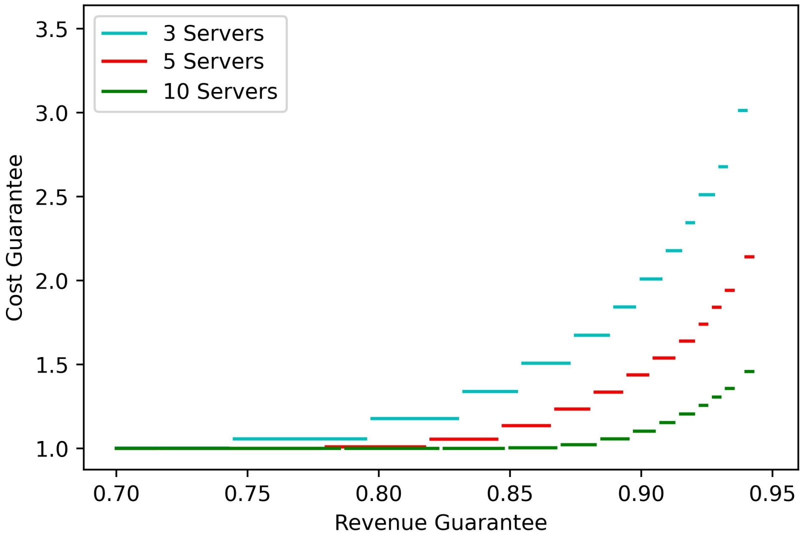
and

$$C(\pi^{\lambda,\gamma}) \leq g(\gamma, C)C(\pi^*).$$

Theorem 2 provides a family of guarantees, parameterized by the choice of γ. This allows the practitioner to select a value of γ suitable to their application: for scenarios in which revenues are expected to be very large compared to the congestion penalties, a higher value of γ is warranted, and vis versa when congestion penalties are expected to be large relative to revenues.

When the number of servers is taken to be very large, it is not surprising that the advantage of dynamic pricing over static pricing diminishes. Theorem 2 gives insight into the rate at which the advantage declines. Figure 1 graphically shows the guarantees of Theorem 2 when we have $3, 5,$ or 10 servers.

Figure 2.1: A graphical representation of the guarantees of Theorem 2



2.4.1 Proof of Theorem 2

In the single-class case, Lemma 2 gives

$$\mathcal{R}(\pi^*) \leq \tilde{\lambda} p(\tilde{\lambda}). \tag{2.7}$$

Now we can compute the revenue rate of the policy $\pi^{\tilde{\lambda},\gamma}$ and proceed to arrive at a lower bound

$$\begin{aligned}
\mathcal{R}(\pi^{\tilde{\lambda},\gamma}) &= \tilde{\lambda} p(\tilde{\lambda}) \sum_{i=0}^{\gamma} \mathbb{P}_i(\pi^{\tilde{\lambda},\gamma}) \\
&= \tilde{\lambda} p(\tilde{\lambda})(1 - \mathbb{P}_{\gamma+1}(\pi^{\tilde{\lambda},\gamma})) \\
&\geq \tilde{\lambda} p(\tilde{\lambda})(1 - \mathbb{P}_{\gamma+1}(\pi^{C,\gamma})) \\
&\geq (1 - \mathbb{P}_{\gamma+1}(\pi^{C,\gamma}))\mathcal{R}(\pi^*) \\
&= \left(1 - \frac{\frac{C^{\gamma+1}}{C! C^{\gamma+1-C}}}{\sum_{l=0}^{C} \frac{1}{l!} C^l + \frac{1}{C!} C^C (\gamma + 1 - C)}\right) \mathcal{R}(\pi^*)
\end{aligned}$$

where the first inequality uses the fact that $\tilde{\lambda} < C$ from Lemma 1 and that the blocking probability is increasing in $\tilde{\lambda}$, the second inequality uses Eq. (2.7), and the final equality from the blocking probability found in Eq. (2.6).

We use Little's Law to bound the congestion penalty of the optimal dynamic policy

$$C(\pi^*) = \tilde{\lambda} E[W^*] \geq \tilde{\lambda} \tag{2.8}$$

where the inequality follows from the fact that $E[W^*] \geq 1$ because a sojourn always includes a service time. We can also show that

$$C(\pi^{\tilde{\lambda},\gamma}) = \sum_{i=1}^{\gamma+1} i \mathbb{P}_i(\pi^{\tilde{\lambda},\gamma}) \leq C(\pi^*) \frac{\sum_{i=1}^{\gamma+1} i \mathbb{P}_i(\pi^{\tilde{\lambda},\gamma})}{\tilde{\lambda}}$$

where the inequality uses Eq. (2.8). Now by finding the maximum value of the coefficient of $C(\pi^*)$ over $\tilde{\lambda} \in [0, C]$, we can get a bound which does not depend upon $\tilde{\lambda}$. Indeed, letting $f(\gamma, C, \tilde{\lambda}) =$

$\frac{\sum_{i=1}^{\gamma+1} i\mathbb{P}_i(\pi^{\tilde{\lambda},\gamma})}{\tilde{\lambda}}$, we can write $g(\gamma, C) = \max_{\tilde{\lambda}\in[0,C]} f(\gamma, C, \tilde{\lambda})$ so that we have

$$C(\pi^{\tilde{\lambda},\gamma}) \leq C(\pi^*)f(\gamma, C, \tilde{\lambda}) \leq C(\pi^*)g(\gamma, C),$$

which completes the proof.

2.5 Single-server, Single-class Static Pricing Guarantees

In this section, we focus on the single-class, single-server setting (M/M/1), where we are able to give closed-form expressions for the congestion ratio bounds. Theorem 2 reduces to the following result when we take $C = 1$.

Theorem 3. *The static pricing policy $\pi^{\tilde{\lambda},\gamma}$ guarantees at least $\frac{\gamma+1}{\gamma+2}$ of the revenue rate and incurs a congestion penalty less than $g(\gamma)$ times that of the optimal dynamic pricing policy, where*

$$g(\gamma) = \begin{cases} 1, & \textit{if } \gamma = 0 \\ \frac{2\sqrt{3}}{3} \approx 1.155, & \textit{if } \gamma = 1 \\ 1.532^{1}, & \textit{if } \gamma = 2 \\ \frac{\gamma+1}{2}, & \textit{if } \gamma \geq 3 \end{cases}$$

Equivalently,

$$\mathcal{R}(\pi^{\tilde{\lambda},\gamma}) \geq \frac{\gamma+1}{\gamma+2}\mathcal{R}(\pi^*)$$

and

$$C(\pi^{\tilde{\lambda},\gamma}) \leq g(\gamma)C(\pi^*).$$

Moreover, this analysis is tight.

[1] The exact, closed-form expression for $g(2)$ involves the root of a quartic and would take a few pages to relay.

The tightness of Theorem 3 is shown using a class of instances described in Section 2.5.2.

The fact that we can provide exact values for the function $g(\gamma)$ is a stroke of some good luck. As we shall see in the proof, finding these worst-case congestion penalty bounds involves solving an optimization problem over $\tilde{\lambda} \in [0,1]$. For $\gamma = 0$, the maximizer is at $\tilde{\lambda} = 0$ and for $gamme \geq 3$, the maximizer is at $\tilde{\lambda} = 1$. For $\gamma = 1$ and $\gamma = 2$, the maximizer lies strictly within the interval $[0,1]$; for $\gamma \geq 3$, the function is monotonically increasing on the interval and hence the maximum occurs when $\tilde{\lambda} = 1$. This is fortunate because finding the *unconstrained* maxima of the relevant ratio involves finding the roots higher and higher degree polynomials: for $\gamma = 1$, a degree 2 polynomial arises, for $\gamma = 2$, a degree 4 polynomial arises, and for general γ a degree 2γ polynomial arises. If the unconstrained maximum when $\gamma = 3$ were to lie in the interval $[0,1]$, it is unlikely we could find a closed-form expression.

2.5.1 Proof of Theorem 3

The revenue guarantee follows directly from Theorem 2. We now prove the congestion penalty guarantees of Theorem 3. Just as in the proof of Theorem 2, we can write

$$C(\pi^{\tilde{\lambda},\gamma}) \leq C(\pi^*)g(\gamma)$$

where

$$g(\gamma) = \max_{\tilde{\lambda}\in[0,1]} \frac{\sum_{i=1}^{\gamma+1} i\mathbb{P}_i(\pi^{\tilde{\lambda},\gamma})}{\tilde{\lambda}}$$

and the steady state probabilities are given by (2.5). For $\gamma < 3$ it is straightforward to arrive at the claimed values of $g(\gamma)$. When $\gamma = 0$, $g(\gamma) = \frac{1}{\tilde{\lambda}+1}$ and thus the worst case is when $\tilde{\lambda} = 0$ and $g(0) = 1$. When $\gamma = 1$, we take the derivative and compute its numerator as

$$1 - 2\tilde{\lambda} - 2\tilde{\lambda}^2,$$

from which we arrive at the positive root $\tilde{\lambda} = \frac{\sqrt{3}-1}{2}$. Plugging this value back into the objective function gives $g(2) = \frac{2}{\sqrt{3}}$ as claimed. When $\gamma = 2$, finding the maximizer reduces to finding the real positive root of the equation

$$-3\tilde{\lambda}^4 - 4\tilde{\lambda}^3 - 2\tilde{\lambda}^2 + 4\tilde{\lambda} + 1$$

which can be accomplished via the well-known quartic formula [49].

For $\gamma \geq 3$, we proceed by showing that the objective is increasing in $\tilde{\lambda}$ over the interval $[0, 1]$. Let $f_\gamma(\tilde{\lambda})$ denote the objective function for a fixed γ and express it equivalently as

$$f_\gamma(\tilde{\lambda}) = \frac{\sum_{j=1}^{\gamma+1} j\tilde{\lambda}^j}{\sum_{i=1}^{\gamma+2} \tilde{\lambda}^i}$$

We show the derivative $f'_\gamma(\tilde{\lambda})$ is nonnegative for $0 \leq \tilde{\lambda} \leq 1$ and $\gamma \geq 3$. We take the derivative and ignore the denominator by looking only at the signs

$$\begin{aligned}
\operatorname{sgn}(f'_\gamma(\tilde{\lambda})) &= \operatorname{sgn}\left(\sum_{i=1}^{\gamma+2}\sum_{j=0}^{\gamma+2} j^2\tilde{\lambda}^{i+j-1} - \sum_{i=1}^{\gamma+2}\sum_{j=0}^{\gamma+2} ji\tilde{\lambda}^{i+j-1}\right) \\
&= \operatorname{sgn}\left(\sum_{i=0}^{\gamma+1}\sum_{j=0}^{\gamma+2} j^2\tilde{\lambda}^{i+j} - \sum_{i=0}^{\gamma+1}\sum_{j=0}^{\gamma+2} j(i+1)\tilde{\lambda}^{i+j}\right) \\
&= \operatorname{sgn}\left(\sum_{i=0}^{\gamma+1}\sum_{j=0}^{\gamma+1}(j^2 - j(i+1))\tilde{\lambda}^{i+j}\right)
\end{aligned}$$

We show this is nonnegative for $0 \leq \tilde{\lambda} \leq 1$ when $\gamma \geq 3$ by showing that the coefficients of the lower-order terms (those with degrees lesser or equal to $\gamma + 1$) are all positive, and that the sum of all coefficients is positive as well. This suffices because for $\tilde{\lambda} \leq 1$, the lower-order terms dominate.

We can use the above expression to write the coefficient of $\tilde{\lambda}^k$ for $2 \leq k \leq \gamma + 1$ as

$$\sum_{j=0}^{k}(j^2 - j(k - j + 1)) = \frac{(k-1)k(k+1)}{6}$$

which we observe is indeed positive for $k \geq 2$. Proceeding similarly for the higher-order terms, for $\gamma + 1 < k \leq 2\gamma + 2$ the coefficient of $\tilde{\lambda}^k$ is

$$\sum_{j=k-\gamma+1}^{\gamma+1} (j^2 - j(k - j + 1)) = \frac{(2\gamma - k + 3)(4\gamma^2 - 4\gamma k + 12\gamma + k^2 - 9k + 8)}{6}.$$

To conclude, we show that the sum of all the coefficients is nonnegative for $\gamma \geq 3$. Indeed,

$$\sum_{k=2}^{\gamma+1} \frac{(k-1)k(k+1)}{6} + \sum_{k=\gamma+2}^{2\gamma+2} \frac{(2\gamma - k + 3)(4\gamma^2 - 4\gamma k + 12\gamma + k^2 - 9k + 8)}{6} = \frac{1}{12}(\gamma + 1)(\gamma + 2)^2(\gamma - 3)$$

from which we can immediately see the desired nonnegativity. With this monotonicity in hand, we can upper bound

$$g(\gamma) \leq \frac{\sum_{i=1}^{\gamma+1} i\mathbb{P}_i(\pi^{1,\gamma})}{1} = \sum_{i=1}^{\gamma+1} \frac{i}{\gamma + 2} = \frac{\gamma + 1}{2},$$

thus completing the proof.

2.5.2 A class of instances proving tightness

We now describe a class of instances which we use to show that Theorem 3 is tight and that Theorem 1 is tight when $C = 1$. Moreover, the following examples also show that static policies with no threshhold can perform arbitrarily poorly.

Consider an instance with linear demand (uniform valuation distribution), i.e., where $\lambda = b - ap$ for some positive a and b. Take $a < b < 2a$, so that the expected value from accepting a customer when one is in service is always negative: the largest price a customer will pay is $b/a < 2$, while the expected congestion penalty from accepting a customer while one is in service is 2 (2 customers). Thus, the optimal policy is in fact a static policy with a cutoff point of 0. In this case, it is a straightforward single-variable optimization problem and we can arrive at the following expression for the optimal policy

$$\pi^* = \{\sqrt{b-a+1}-1, 0, 0, \dots\}$$

To now show the tightness of the objective value guarantee of Theorem 1 when $C = 1$, we write $b = \kappa a$ for $1 < \kappa < 2$ and take the limit as $a \to \infty$. In the limit, $\sqrt{\kappa a - a + 1} - 1 \to \infty$, and thus $\tilde{\lambda} \to 1$. Then we have

$$\mathcal{Z}(\pi^{\tilde{\lambda},0}) = \tilde{\lambda} p(\tilde{\lambda}) \mathbb{P}_0(\pi^{\tilde{\lambda},0}) - \mathbb{P}_1(\pi^{\tilde{\lambda},0}) = \tilde{\lambda} p(\tilde{\lambda}) \frac{1}{1+\tilde{\lambda}} - \frac{\tilde{\lambda}}{1+\tilde{\lambda}} = \tilde{\lambda}\frac{(\kappa a - \tilde{\lambda})}{a(1+\tilde{\lambda})} - \frac{\tilde{\lambda}}{1+\tilde{\lambda}} \to \frac{\kappa - 1}{2}$$

On the other hand, we can see that the optimal price converges to κ

$$p(\lambda_0^*) = \frac{\kappa a - \sqrt{\kappa a - a + 1} + 1}{a} \to \kappa$$

and thus

$$\mathcal{Z}(\pi^*) = \lambda_0^* p(\lambda_0^*) \mathbb{P}_0(\pi^*) - \mathbb{P}_1(\pi^*) = \frac{\lambda_0^*}{1+\lambda_0^*} p(\lambda_0^*) - \frac{\lambda_0^*}{1+\lambda_0^*} \to \kappa - 1$$

which proves the desired tightness for $\gamma = 0$. Moreover, since the expected value of allowing a customer to queue is always negative, policies $\pi^{\tilde{\lambda},\gamma}$ with $\gamma > 0$ will perform worse, and thus we observe that in this instance the best static policy which fixes an arrival rate $\tilde{\lambda}$ attains only half the optimal value. However, it is worth noting that in this instance the optimal policy is in fact a static policy, just one which does not pick the arrival rate $\tilde{\lambda}$.

The same limiting instance shows the revenue guarantees of Theorem 3 are also tight. For any γ, we have

$$\mathcal{R}(\pi^{\tilde{\lambda},\gamma}) = \tilde{\lambda} p(\tilde{\lambda})(1 - \mathbb{P}_{\gamma+1}(\pi^{\tilde{\lambda},\gamma})) \to \kappa \frac{\gamma+1}{\gamma+2}$$

On the other hand, as we computed above,

$$\mathcal{R}(\pi^*) = \lambda_0^* p(\lambda_0^*) \mathbb{P}_0(\pi^*) \to \kappa$$

which proves the claimed tightness of the revenue guarantees.

The same class of instances will serve to prove our congestion penalty guarantees are tight, but we will not use the limiting instance. Instead, note that the only inequality used to furnish our congestion penalty guarantees is Eq. (2.8). Since for instances in this class the optimal policy π^* does not allow a queue to form, this inequality is tight, and hence so are our ensuing guarantees.

Finally, these same instances can be used to show that unthresholded static policies can perform arbitrarily poorly. Note that computing the optimal unthresholded policy for a given instance is a simple single-variable optimization problem. By taking a large and κ close to 1, we can exhibit instances where the unthresholded static policy recovers only an ϵ proportion of the optimal value, for ϵ arbitrarily small. For example, with $a = 1000$ and $\kappa = 1.05$, the optimal policy earns 0.37 while the optimal unthresholded policy earns 0.0006, which gives a performance ratio of $\frac{0.0006}{0.37} = 1.6\%$.

2.6 Numerical Experiments

In this section, we present the results of numerical experiments on our static policies. Specifically, we consider linear ($\lambda = b - ap$), exponential ($\lambda = be^{-ap}$), and logistic ($\lambda = \frac{b(1+e^{-ap^0})}{1+e^{a(p-p^0)}}$) demand functions. For each number of servers $C \in \{1, 3, 5, 10\}$, we sample a uniformly from $[0.1, 5]$, b uniformly from $[.5, 10]$, and p_0 uniformly from $[0, 20]$. Under linear demand, when $a > b$, the optimal policy does not sell and thus achieves an objective value of 0. Though this is also attainable by a static policy, for simplicity we reject these instances and resample. For each C and each demand function, we perform 1000 replications and report the worst-case and average approximation ratios for the overall objective, revenue, and congestion penalties in Tables 2.1 and 2.2.

We report results for two different static policies. First, we consider the optimal static pricing policy. In other words, for each instance, we are simultaneously optimizing both the static price and the cutoff we choose. The notation λ^* in the column headers refers to these policies. Second, we consider static policies which are constrained to use the price which induces the same average arrival rate as that of the optimal policy (these are the policies considered by our theorems). For each instance, we find the optimal dynamic policy, compute the average arrival rate, fix the price which induces this arrival rate, and then find the optimal cutoff for that fixed price. The notation $\tilde{\lambda}$ in the column headers refers to these policies.

Table 2.1: Worst-case approximation ratios of static pricing policies

		Objective		Revenue		Congestion	
	C	λ^*	$\tilde{\lambda}$	λ^*	$\tilde{\lambda}$	λ^*	$\tilde{\lambda}$
Linear	1	93.0%	75.7%	80.8%	69.1%	111.6%	114.0%
	3	95.4%	95.1%	92.1%	87.8%	109.5%	119.0%
	5	98.0%	97.9%	97.2%	97.2%	108.3%	121.0%
	10	99.9%	99.9%	99.9%	99.9%	100.0%	100.0%
Exponential	1	89.7%	89.0%	82.6%	75.7%	105.4%	119.0%
	3	97.4%	97.1%	96.6%	98.5%	103.2%	117.0%
	5	99.6%	99.6%	99.5%	99.9%	101.2%	103.0%
	10	99.9%	99.9%	99.9%	99.9%	99.9%	100.0%
Logistic	1	89.1%	73.0%	81.4%	64.2%	118.9%	159.0%
	3	93.3%	87.9%	89.1%	84.4%	106.4%	121.0%
	5	96.2%	94.1%	94.9%	93.1%	100.0%	104.0%
	10	99.2%	98.9%	99.4%	98.9%	104.5%	104.0%

Note. The table compares optimal static pricing policies (attained by jointly optimizing both the rate λ^* and the cutoff) and static policies which pick the rate induced by the optimal policy $\tilde{\lambda}$ (and choose an optimal cutoff for this rate).

On the whole, we see that static policies perform fairly well. Even in the worst-case instance found in our experiments, the static policy is able to attain 89.1% of the optimal dynamic policy's objective value. Looking at the average performance, the difference becomes smaller still: even with only 1 server, static policies can typically attain between 96% and 98% of the optimal dynamic policy's value.

Moreover, the empirics echo our theoretical results: we see the advantage of dynamic over

Table 2.2: Average approximation ratios of static pricing policies

		Objective		Revenue		Congestion	
	C	λ^*	$\tilde{\lambda}$	λ^*	$\tilde{\lambda}$	λ^*	$\tilde{\lambda}$
Linear	1	98.0%	87.0%	96.8%	81.4%	95.8%	78.8%
	3	97.8%	97.7%	97.5%	96.6%	98.8%	96.8%
	5	99.5%	99.5%	99.5%	99.7%	99.7%	100.5%
	10	99.9%	99.9%	99.9%	99.9%	100.0%	100.0%
Exponential	1	97.4%	97.1%	96.7%	93.7%	96.5%	92.7%
	3	99.8%	99.7%	99.7%	99.9%	99.6%	100.4%
	5	99.9%	99.9%	99.9%	99.9%	99.9%	100.0%
	10	99.9%	99.9%	99.9%	99.9%	99.9%	100.0%
Logistic	1	96.0%	85.2%	95.6%	86.7%	95.0%	99.0%
	3	97.8%	94.6%	98.1%	95.3%	99.6%	101.0%
	5	96.2%	94.1%	94.9%	93.1%	100.0%	104.0%
	10	99.9%	99.9%	99.9%	99.9%	100.5%	100.3%

Note. The table compares optimal static pricing policies (attained by jointly optimizing both the rate λ^* and the cutoff) and static policies which pick the rate induced by the optimal policy $\tilde{\lambda}$ (and choose an optimal cutoff for this rate).

static pricing decreases rapidly as we increase the number of servers. For instances with 10 or more servers, the performance of the optimal static policy is barely distinguishable from the dynamic optimum. This further lends credence to our practical insight that dynamic pricing may not be necessary for queueing systems with many servers.

The empirics also give some insight into how the static policies we construct in our proofs compare to the optimal static policies. The difference is sensitive to the form of the demand function. For exponential demand, the difference is fairly marginal: the largest discrepancy we observe in the worst-case objective ratios is 0.7% when $C = 1$. For linear and logistic demand functions, on the other hand, there are much larger discrepancies. In any case, the discrepancies diminish rapidly as we increase the number of servers.

2.7 Sojourn Time Penalty

Up to this point, the congestion penalty of a policy in our model has been given by the long-run average number of customers in the system under that policy. In this section, we analyze a new model where the congestion penalty of a policy is instead given by the long-run average sojourn times. As usual, the sojourn time is defined as the total time a customer spends in the system, including both the waiting time and the service time. Hence, this model is of practical importance as it is relates directly to the quality of service experienced by the customers.

The operation of this model is identical to the first. We shall focus on the single class setting. Potential customers arrive according to a Poisson process at rate Λ and join the system if the offered price exceeds their valuation drawn i.i.d. from some regular distribution F. Letting $W(\pi)$ denote the steady-state sojourn time distribution of a customer who joins the system under policy π, the value rate $\mathcal{Z}(\pi)$ of a policy π is now given by

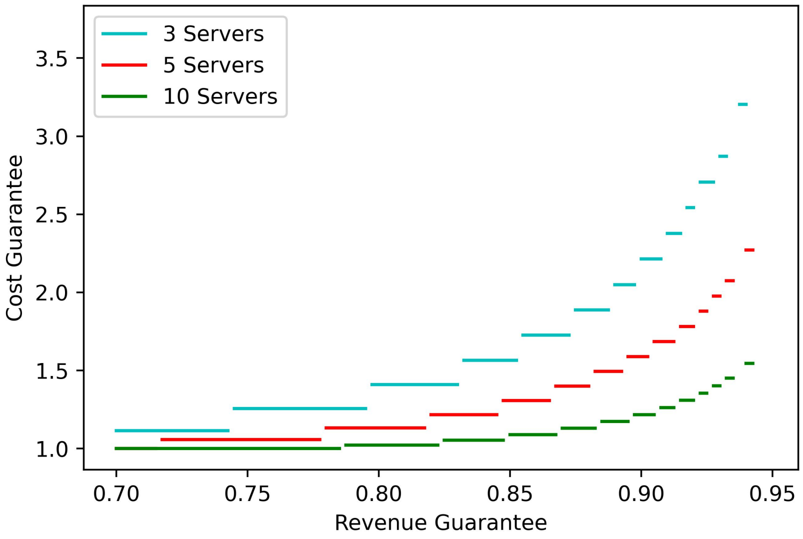
$$\begin{aligned}\mathcal{Z}(\pi) &= \sum_{i=0}^{\infty} p_i \lambda(p_i) \mathbb{P}_i(\pi) - E[W(\pi)] \\ &= \sum_{i=0}^{\infty} p_i \lambda(p_i) \mathbb{P}_i(\pi) - \frac{E[L(\pi)]}{\tilde{\lambda}(\pi)} \\ &= \sum_{i=0}^{\infty} \left(p_i \lambda(p_i) - \frac{i}{\tilde{\lambda}(\pi)} \right) \mathbb{P}_i(\pi)\end{aligned}$$

where the second equality uses Little's Law. We use the notation $\tilde{\lambda}(\pi)$ to emphasize that $\tilde{\lambda}$ is induced by π, and we remark that now this average arrival rate $\tilde{\lambda}(\pi)$ arises naturally when formulating the problem.

In this case, objective value guarantees on $\mathcal{Z}$ for the class of static policies which pick the arrival rate $\tilde{\lambda}$ are no longer possible. This fact is verified with an instance from the class described in Section 2.5.2. If we take an instance with linear demand ($\lambda = b - ap$) with $b = 6000$ and $a = 5000$, we can observe that $\mathcal{Z}(\pi^*) \approx 0.17 > 0$, whereas $\mathcal{Z}(\pi^{\tilde{\lambda},0}) \approx -0.4 < 0$. For these instances, taking a cutoff higher than 0 degrades performance further. Thus no result of the flavor of Theorem 1 is possible, and hence we must resort to providing bi-criteria bounds on revenue and congestion in the spirit of Theorem 2.

We present Theorem 4 below for the sojourn time penalty model for an arbitrary number of servers C. We observe that the congestion guarantees of the sojourn time model are not as strong as the guarantees we can furnish for the congestion model. Thus we are led to the insight that the control of sojourn times with static pricing is a marginally more difficult objective than just controlling congestion. The relative difference is most pronounced when looking at policies with small cutoff points, particularly $\gamma = 1$ or $\gamma = 2$. In the single-server case, the congestion guarantee reduces to simply $\frac{\gamma+2}{2}$ and the revenue guarantee reduces to $\frac{\gamma+1}{\gamma+2}$ for all $\gamma \geq 0$. We provide Figure 3 to graphically illustrate the guarantees of Theorem 4 when we have $3, 5$, or 10 servers.

Figure 2.2: A graphical representation of the guarantees of Theorem 4



Theorem 4. *The static pricing policy* $\pi^{\tilde{\lambda},\gamma}$ *guarantees at least* $1 - \frac{\frac{C^{\gamma+1}}{C!C^{\gamma+1-C}}}{\sum_{l=0}^{C} \frac{1}{l!}C^l + \frac{1}{C!}C^C(\gamma+1-C)}$ *of the revenue rate and incurs a congestion penalty less than* $\frac{\sum_{i=1}^{C} \frac{1}{(i-1)!}C^i + \frac{1}{C!}C^C \sum_{i=C+1}^{\gamma+1} i}{\sum_{l=0}^{C} \frac{1}{l!}C^{l+1} + \frac{\gamma-C}{C!}C^C}$ *times that of the optimal dynamic pricing policy. Equivalently,*

$$\left(1 - \frac{\frac{C^{\gamma+1}}{C!C^{\gamma+1-C}}}{\sum_{l=0}^{C} \frac{1}{l!}C^l + \frac{1}{C!}C^C(\gamma+1-C)}\right)\mathcal{R}(\pi^*) \leq \mathcal{R}(\pi^{\tilde{\lambda},\gamma})$$

and

$$\frac{\sum_{i=1}^{C}\frac{1}{(i-1)!}C^{i}+\frac{1}{C!}C^{C}\sum_{i=C+1}^{\gamma+1}i}{\sum_{l=0}^{C}\frac{1}{l!}C^{l+1}+\frac{\gamma-C}{C!}C^{C}}C(\pi^{*})\geq C(\pi^{\tilde{\lambda},\gamma}).$$

Proof. Proof. As the revenue portion of our model remains unchanged, the revenue guarantee follows immediately from Theorem 2. Thus, it suffices to prove the congestion guarantees. We begin by analyzing the congestion penalty of $\pi^{\tilde{\lambda},\gamma}$,

$$C(\pi^{\tilde{\lambda},\gamma})=E[W(\pi^{\tilde{\lambda},\gamma})]\leq C(\pi^{*})E[W(\pi^{\tilde{\lambda},\gamma})] \tag{2.9}$$

where the inequality follows since $C(\pi^{*})=E[W(\pi^{*})]\geq 1$ because a sojourn always includes at least a service time. To remove the dependence on the value of $\tilde{\lambda}$, we seek the worst-case (largest) value of $E[W(\pi^{\tilde{\lambda},\gamma})]$ over $\tilde{\lambda}\in[0,C]$. In contrast to the previous model, this coefficient of $C(\pi^{*})$ is non-decreasing in $\tilde{\lambda}$: clearly the expected sojourn time of a customer increases if we increase the arrival rate to the system. Thus we can plug in $\tilde{\lambda}=C$ to get a universal bound.

$$C(\pi^{\tilde{\lambda},\gamma})\leq C(\pi^{*})E[W(\pi^{\tilde{\lambda},\gamma})]\leq C(\pi^{*})E[W(\pi^{C,\gamma})]=C(\pi^{*})\frac{\sum_{i=1}^{\gamma+1}i\mathbb{P}_{i}(\pi^{C,\gamma})}{C(1-\mathbb{P}_{\gamma+1}(\pi^{C,\gamma}))}$$

where the first inequality is Eq. (2.9), the second is by the fact that sojourn times are non-decreasing in $\tilde{\lambda}$, and the equality uses Little's Law to compute $E[W(\pi^{C,\gamma})]$. Finally, we can use the expression for the multi-server steady-state probabilities in (2.5) to arrive at the claimed congestion guarantees. □

2.8 Conclusion

In this work, we have furnished universal performance guarantees for static pricing policies in the context of a general queueing model with price-sensitive customers. We observe that the advantage of dynamic over static pricing decays rapidly as the number of servers increases. We also provide bi-criteria guarantees on the revenue and cost objectives, with closed-form guarantees in the M/M/1 setting. In applications where controlling sojourn times rather than congestion is the more salient objective, we provided similar guarantees on the performance of static policies. For the practitioner wondering if implementing dynamic pricing is worth the additional complexity and potential adversarial customer behavior it might induce, our results provide useful insights.

There are many possible avenues to extend our results. One promising direction of interest would be to relax the assumption that the service times are exponential. With general i.i.d. service times, the optimal pricing policy now depends not only on the number in system but also on the remaining service times of the customers in service. There are many subtleties that must be accounted for to prove similar result, but at the heart of our argument are two very robust tools: Little's Law and Jensen's inequality.

Lastly, our theoretical results have focused on a particular class of static policies for which we can provide good theoretical guarantees. The question, for example, of which instance of our model exhibits the largest gap between optimal dynamic and static performance remains open. Our theoretical results have shown that the worst-case objective value ratio is at least $\frac{1}{2}$, and our empirical results have shown that it is at most 0.891. Trying to narrow this interval down and find the value of this constant remains a question for future work.

Chapter 3: On an Adaptive Non-deterministic Transmission Process Queueing Model

This chapter studies an application of queueing in telecommunications. It is based on an upcoming article written in collaboration with Karl Sigman and Erol Gelenbe.

3.1 Introduction

The widespread proliferation of the Internet of Things (IoT) has brought about new challenges in the field of telecommunications, particularly in the area of network access. One of the major challenges is the Massive Access Problem (MAP), which occurs when too many IoT devices simultaneously transmit data to a base station or IoT gateway, causing congestion and potentially untenable delays. When a packet is delayed for too long, its data may no longer be of use; thus it is of interest to develop techniques to ameliorate the MAP and ensure packets are transmitted in a timely manner.

A variety of techniques to address the MAP have been proposed. Early work focused on implementing adaptive routing to reduce congestion in networks with multiple available paths ([50], [51]). More recently, many reactive techniques which try to adapt to realized traffic patterns have been proposed and analyzed under the assumption that the traffic arrives in a random manner. Solutions in this class include Access Class Barring ([52], [53], [54]), Cognitive Machine-to-Machine communication ([55], [56]), and device clustering ([57], [58]).

On the other hand, one can do more than simply react to the realized traffic stream: proactive prediction of traffic patterns is also possible. These techniques attempt to use observed traffic to predict future traffic, and then use that prediction to perform scheduling of packet generation times. Examples of proactive techniques include Joint Forecasting-Scheduling (JFS) and Priority based on Average Load (PAL), in which channel resources are allocated to IoT devices based on traffic characteristic ([59], [60]). However, these techniques involve time-consuming machine learning methods and require a significant amount of communication over the network to implement.

A simpler technique to address the MAP is that of traffic shaping, where packets are delayed at the origin. One particular way of performing traffic shaping, the Quasi-Deterministic Transmission Process (QDTP), was introduced and empirically investigated in the context of IoT devices in [61]. Following that, QDTP and the more general ANTP (which is the focus of the present work) were more formally analyzed using queueing theory in [62]. The authors show that the delay sequences of customers at this cafe facility under QDTP and ANTP satisfy an expression similar to Lindley's recursion. Moreover, it is shown that, as long as the delays D_n satisfy a certain condition, the response time of a packet will not increase under ANTP. Though ANTP was originally motivated by problems in telecommunications which demand reducing congestion, here we also consider other applications in which notions such as utility make sense in the measuring of the benefits of the reduction in congestion. That is why we call the initial delay facility a *cafe*; one would prefer to spend most of the delay in a cafe rather than in a line somewhere.

In the present work, we furnish deeper theoretical results characterizing the performance of ANTP in a stochastic setting. Our focus is on obtaining stability conditions and stationary versions under general stochastic assumptions. We then delve into the special case of the GI/GI/1 queue and show Harris recurrence of the underlying Markovian structure. Our work builds on the previous study by providing a deeper understanding of ANTP and offers insights that can inform future work in this area.

3.2 Model

In this paper, we analyze an ANTP single-server queueing model. It consists of two connected components, an initial delay facility called the *cafe*, to which the arrivals first attend at arrival times $\{a_n\}$, followed by a FIFO single-server queue called the *service facility* where they attend at times $t_n \geq a_n$ and have corresponding service requirements/times $\{S_n\}$. With $t_0 = a_0 = 0$, the point process $\{t_n\}$ is defined recursively as a function of the original arrival point process $\{a_n\}$:

$$t_{n+1} = \max\{t_n + D_n, a_{n+1}\},\ n \geq 0, \tag{3.1}$$

where D_n can be, for example, a deterministic constant D or a random function of S_n. The interarrival times are then $T_n = t_{n+1} - t_n,\ n \geq 0$, as opposed to $A_n = a_{n+1} - a_n,\ n \geq 0$. The purpose of ANTP is to reduce congestion faced by the customers as compared to simply joining the service facility (without attending the cafe) at the original arrival times $\{a_n\}$ and with the service times $\{S_n\}$.

The times spent at the cafe, $W_n = t_n - a_n$, are the *initial delays* experienced by the customers. (When $t_n = a_n$ there is no such delay and the customer goes directly to the service facility). V_n denotes the delay in queue by the n^{th} arrival to the ANTP service facility before entering service; it satisfies the classic FIFO single-server recursion for delay in queue (line), $V_{n+1} = (V_n + S_n - T_n)^+,\ n \geq 0$. And thus the total delay, which we denote by Z_n is computed as $Z_n = W_n + V_n$.

We first present some sample-path results including that when $D_n < S_n,\ n \geq 0$, it follows that the Z_n are the same as if there was no cafe and the customers attended only the service facility at times a_n; in other words, if $D_n < S_n,\ n \geq 0$, then ANTP does not increase customer sojourn times. The difference is that congestion at the service facility is reduced (shared) by spending time at the cafe.

In a stochastic framework, we assume that $\{(a_n, S_n, D_n) : n \geq 0\}$ is a stationary ergodic marked point process (mpp) with arrival rate $0 < \lambda < \mu$ where $0 < E(S) = \frac{1}{\mu} < \infty$. We let $\{(a_n^0, S_n^0, D_n^0) : n \in \mathbb{Z}\}$ denote (by extension) a two-sided point-stationary ergodic version of the input. We show that when $E(D) < \frac{1}{\lambda}$, then the arrival rate of $\{t_n\}$ is also λ and Z_n has a proper limiting distribution π to which Z_n converges to in total variation as $n \to \infty$. Furthermore, in the special case when $\{a_n\}$ is a renewal process, and independently $\{(S_n, D_n)\}$ is an iid sequence of vectors, then we show that $\{(W_n, V_n) : n \geq 0\}$ forms a positive Harris recurrent Markov chain (hence must be a regenerative process). We demonstrate that the chain might never hit $(0, 0)$, but nonetheless we find the regeneration points even in such a case.

3.3 Notation and definitions

1. C_n denotes the n^{th} customer.

2. a_n is the time at which C_n arrives at the cafe; $a_0 = 0$. The interarrival times are denoted by $A_n = a_{n+1} - a_n,\ n \geq 0$.

3. $L_{n+1} = (L_n + S_n - A_n)^+,\ n \geq 0$ defines the delay (in line/queue) sequence for the *nominal* FIFO $G/G/1$ model. This is the model that C_n would attend *if there was no cafe* and instead customers went directly to the service facility at times $\{a_n\}$.

4. With $t_0 = a_0 = 0$, the arrival process to the ANTP service facility is defined recursively:

$$t_{n+1} = \max\{t_n + D_n, a_{n+1}\},\ n \geq 0. \tag{3.2}$$

$T_n = t_{n+1} - t_n,\ n \geq 0$ are the interarrival times for the ANTP service queue. Note that they are bounded from below by D_n, since $t_{n+1} \geq t_n + D_n$;

$$T_n \geq D_n.$$

5. $W_n = t_n - a_n,\ n \geq 0$ is the *initial delay* process; $W_0 = 0$. C_n arrives at time a_n to the cafe and remains there for W_n units of time before joining the ANTP service facility queue at time $t_n = a_n + W_n$. (Entering means that they join the line if there is one or enter service immediately if there is not.)

6. $V_{n+1} = (V_n + S_n - T_n)^+,\ n \geq 0$ with $V_0 = 0$ defines the delay (in line/queue before starting service) sequence at the ANTP service facility queue.

7. C_n has a total delay (before starting service) of $Z_n = W_n + V_n$, and a response (sojourn) time of $R_n = Z_n + S_n = W_n + V_n + S_n$. C_n arrives at time a_n and departs at time $a_n^d = a_n + R_n$.

As comparison, if all customers only attended the nominal queue (no cafe), at the times $\{a_n\}$, then (recalling Point 3 above) the delay and sojourn time of C_n would be L_n and $L_n + S_n$ respectively. We wish to compare the ANTP Z_n and $Z_n + S_n$ with the nominal L_n and $L_n + S_n$, so as to determine if and how ANTP reduces congestion for customers.

8. When we consider a stochastic framework, we then assume that the input $\{(a_n, S_n, D_n) : n \geq 0\}$ is a stationary ergodic marked point process (mpp) with $0 < \lambda < \mu$ where where λ is the arrival rate and $0 < E(S) = \frac{1}{\mu} < \infty$. (We let $\{(a_n^0, S_n^0, D_n^0) : n \in \mathbb{Z}\}$ denote (by extension) a two-sided point-stationary ergodic version of the input.)

3.4 Some basic properties of ANTP

Here we present several sample-path results. When we give a result in a stochastic setting, we are assuming the stationary ergodic assumptions for $\{(a_n, S_n, D_n) : n \geq 0\}$ mentioned as Point 8 in Section 3.3.

Proposition 3.4.1. *The initial delay sequence* $W_n = t_n - a_n$, $n \geq 0$, *satisfies the recursion*

$$W_{n+1} = (W_n + D_n - A_n)^+, \ n \geq 0; \tag{3.3}$$

the delay in line (queue) recursion of a FIFO $G/G/1$ *queue*[1]. *If a rate* $0 < \lambda < \infty$ *exists for* $\{a_n\}$, *that is, if* $a_n/n \longrightarrow \frac{1}{\lambda}$ *as* $n \longrightarrow \infty$, *and an average* d *exists for the* D_n, *that is*

$$\frac{1}{N}\sum_{i=1}^{N} D_n \longrightarrow d, \ as\ N \longrightarrow \infty,$$

where $0 < d < \frac{1}{\lambda}$, *then the rate of* $\{t_n\}$ *exists and is also equal to* λ; $t_n/n \longrightarrow \frac{1}{\lambda}$ *as* $n \longrightarrow \infty$.

Proof. We prove the recursion by induction on n. $W_0 = 0$ since $t_0 = a_0 = 0$. $A_0 = a_1 - a_0 = a_1$.

[1] But its meaning is different: If we view it as a FIFO single-server queueing model with "service times" D_n and interarrival times A_n, then W_n represents the delay in queue (line) of the n^{th} customer arrival (not the sojourn time which would be $W_n + D_n$). But in our model, W_n is the amount of time that the n^{th} customer spends in the Cafe, that is, the sojourn time in the Cafe.

Thus $(W_0 + D_n - A_0)^+ = (D_n - a_1)^+$ and from Equation 3.2 we have $t_1 = D$ if $D_n > a_1$ in which case $(D_n - a_1)^+ = D_n - a_1 = t_1 - a_1 = W_1$, and $t_1 = a_1$ if $D_n \leq a_1$ in which case $t_1 - a_1 = 0 = (D_n - a_1)^+ = W_1$. Thus the result holds for $n = 1$.

Now suppose that for some $n \geq 1$, the recursion 3.3 holds up to $W_n = (W_{n-1} + D_{n-1} - A_{n-1})^+$. We will show it holds for W_{n+1} as well: By the induction hypothesis $W_n = t_n - a_n$ and so

$$\begin{aligned} (W_n + D_n - A_n)^+ &= (t_n - a_n + D_n - (a_{n+1} - a_n))^+ \\ &= (t_n + D_n - a_{n+1})^+. \end{aligned}$$

From Equation 3.2, if $t_n + D_n > a_{n+1}$, then $t_{n+1} = t_n + D_n$ and so $(t_n + D_n - a_{n+1})^+ = t_{n+1} - a_{n+1} = W_{n+1}$; whereas if $t_n + D_n \leq a_{n+1}$, then $t_{n+1} = a_{n+1}$ and so $W_{n+1} = t_{n+1} - a_{n+1} = 0 = (t_n + D_n - a_{n+1})^+$; thus $t_{n+1} - a_{n+1} = W_{n+1}$ as was to be shown.

For the rate result: $t_n = a_n + W_n$ and it is well know that under the conditions assumed, $W_n/n \longrightarrow 0$ as $n \longrightarrow \infty$ (see for example, Lemma 6.1 on Page 134 in [12]), and hence t_n/n and a_n/n have the same limit. □

Remark 3.4.1. Recalling Point 3 in Section 3.3, a possible practical approach to defining D_n is to fix an appropriate parameter b, and define $D_n = bS_n$, for then if the average of the S_n exists, denote it by $1/\mu$, then so does d, and $d = b/\mu$.

Proposition 3.4.2. *If $Z_0 = L_0$, then*

1. *If $D_n \leq S_n$, $n \geq 0$, then $Z_n = L_n$, $n \geq 0$: total delay (before entering service) in ANTP is identical to that in the nominal FIFO $G/G/1$ model ($L_{n+1} = (L_n + S_n - A_n)^+$, $n \geq 0$); thus sojourn times are identical also.*

2. *If $D_n = S_n$, $n \geq 0$, and if $V_0 = 0$, then $W_n = L_n$, $n \geq 0$, and thus from 1, $V_n = 0$, $n \geq 0$: Every customer enters service immediately when arriving at the ANTP service facility; they spend no time delayed in the queue.*

3. *If $D_n < S_n,\ n \geq 0$, then for any $n \geq 1$, if $W_n > 0$ then $V_n > 0$ (equivalently if $V_n = 0$ then $W_n = 0$, i.e., $t_n = a_n$). Moreover, if $V_0 = 0$, then*

$$V_n \leq \sum_{i=0}^{n-1}(S_i - D_i),\ n \geq 1. \tag{3.4}$$

The point of the first assertion is that delay at the service facility is reduced by the amount $L_n - V_n = W_n$; the difference in delay time is spent at the cafe instead of in queue. The point of the second assertion is that when $D_n = S_n,\ n \geq 0$, then *all* of the total delay of C_n is spent in the cafe; none is spent in the line of the ANTP service facility. The point of the third assertion is that when $D_n < S_n,\ n \geq 0$, the server is never wasting time idle while a customer is at the cafe waiting; the ANTP system is efficient.

Proof. For the 1st assertion it suffices (since by assumption $Z_0 = L_0$) to prove that if $Z_n = L_n$ for a given $n \geq 0$, then $Z_{n+1} = L_{n+1}$. Noting that $T_n = t_{n+1} - t_n = A_n + W_{n+1} - W_n,\ n \geq 0$ we have, assuming that $Z_n = W_n + V_n = L_n$ for some n:

$$\begin{aligned}
Z_{n+1} &= W_{n+1} + V_{n+1} \qquad (3.5)\\
&= [W_n + D_n - A_n]^+ + [V_n + S_n - T_n]^+\\
&= [W_n + D_n - A_n]^+ + [W_n + V_n + S_n - A_n - W_{n+1}]^+\\
&= [W_n + D_n - A_n]^+ + [Z_n + S_n - A_n - W_{n+1}]^+\\
&= [W_n + D_n - A_n]^+ + [L_n + S_n - A_n - W_{n+1}]^+.
\end{aligned}$$

Case A) $W_{n+1} = W_n + D_n - A_n > 0$. Then starting with the last line of (3.5)

$$\begin{aligned}
Z_{n+1} &= W_n + D_n - A_n + [L_n + S_n - A_n - (W_n + D_n - A_n)]^+ \\
&= W_n + D_n - A_n + [W_n + V_n + S_n - W_n - D_n]^+ \\
&= W_n + D_n - A_n + [V_n + S_n - D_n]^+ \\
&= W_n + D_n - A_n + [V_n + S_n - D_n] \ \ (D_n < S_n \text{ by assumption}) \\
&= W_n + V_n + S_n - A_n \\
&= L_n + S_n - A_n \ \ (W_n + V_n = Z_n = L_n \text{ by assumption}) \\
&= L_{n+1}.
\end{aligned}$$

Case B) $W_{n+1} = [W_n + D_n - A_n]^+ = 0$. Then starting with the last line of (3.5) immediately yields $Z_{n+1} = [L_n + S_n - A_n]^+ = L_{n+1}$.

Thus in both cases $Z_{n+1} = L_{n+1}$, and the proof of the first assertion is complete.

For the 2nd assertion: We already are assuming that $Z_0 = L_0$. Thus if also $V_0 = 0$, then $0 = V_0 = Z_0 = W_0$ from 1 and so the recursions for $\{L_n\}$ and $\{W_n\}$ both start at 0 and hence yield identical processes $L_n = W_n$, $n \geq 0$. Thus from 1 it follows that $V_n = 0$, $n \geq 0$.

For the 3rd assertion, suppose that $0 < W_n = W_{n-1} + D_{n-1} - A_{n-1}$. Then

$$\begin{aligned}
V_n &= (V_{n-1} + S_{n-1} - A_{n-1} - W_n + W_{n-1})^+ \\
&= (V_{n-1} + S_{n-1} - A_{n-1} - (W_{n-1} + D_{n-1} - A_{n-1}) + W_{n-1})^+ \\
&= (V_{n-1} + S_{n-1} - D_{n-1})^+ \\
&= V_{n-1} + S_{n-1} - D_{n-1} > 0,
\end{aligned}$$

where the positivity at the end is due to the assumption that $D_{n-1} < S_{n-1}$, $n \geq 1$. The inequality in Equation (3.4) follows immediately since $T_n \geq D_n$, $n \geq 0$ implies that $S_n - T_n \leq S_n - D_n$, $n \geq 0$, and thus starting with $V_1 \leq S_0 - D_0$, the result follows recursively by induction: $V_{n+1} = (V_n + S_n - T_n)^+ \leq$

$(V_n + S_n - D_n)^+$.

□

Remark 3.4.2. Note that Equation (3.4) still holds (same proof) when $D_n = S_n$ and hence offers an alternative proof to the 2nd assertion in Proposition 3.4.2.

3.5 A stochastic framework

Here we assume that $\{(A_n, S_n, D_n) : n \geq 0\}$ forms stationary ergodic sequence, equivalently that $\{(a_n, (S_n, D_n))\} : n \geq 0\}$, forms a point-stationary ergodic marked point process. A proof of the following is based on the classic Loynes' Lemma, and can be found, for example on Pages 131-137, Lemma 6.1, and Theorem 6.1 in [12].

Proposition 3.5.1. *If $E(D) < E(A) = \frac{1}{\lambda}$, then there exists a (2-sided) jointly stationary ergodic (proper) version of $\{(W_n, A_n, D_n)\}$ denoted by $\{(W_n^0, A_n^0, D_n^0)\} = \{(W_n^0, A_n^0) \ : n \in \mathbb{Z}\}$, such that*

$$W_{n+1}^0 = (W_n^0 + D_n^0 - A_n^0)^+, \ n \in \mathrm{Z}. \tag{3.6}$$

(This is the stationary delay sequence for a $G/G/1$ queue.) W_n converges in total variation, as $n \longrightarrow \infty$ to the distribution of W_0^0, regardless of initial conditions, $W_0 = x \geq 0$. If $E(D) > E(A) = \frac{1}{\lambda}$, then $\{W_n\}$ is unstable; W_n diverges as $n \longrightarrow \infty$.

The above Proposition 3.4.1 allows us to construct a stationary ergodic version of the point process $\{t_n\}$:

Corollary 3.5.1. *If $E(D) < E(A) = \frac{1}{\lambda}$, then $t_n^0 = a_n^0 + W_n^0$ defines a point-stationary ergodic version of $\{t_n\}$, that is, $T_n^0 = t_{n+1}^0 - t_n^0$ defines a stationary ergodic sequence of interarrival times. Moroever,*

$$E(T^0) = \frac{1}{\lambda}.$$

Proof. Defining $t_n^0 = W_n^0 + a_n^0$, so that $T_n^0 = t_{n+1}^0 - t_n^0 = A_n^0 + W_{n+1}^0 - W_n^0$ indeed yields a stationary ergodic sequence of interarrival times; $\{t_n^0\}$ is a point-stationary ergodic version of $\{t_n\}$.

$E(T_n^0) = E(A_n^0) + E(W_{n+1}^0 - W_n^0) = \frac{1}{\lambda} + 0 = \frac{1}{\lambda}$. □

Remark 3.5.1. $|W_{n+1}^0 - W_n^0| \leq |D_n^0 - A_n^0| < \infty$; thus even if $E(W^0) = \infty$, then $E(W_{n+1}^0 - W_n^0) = 0$ holds.

3.6 Stability of ANTP and the existence stationary versions

We assume that $E(D) < \frac{1}{\lambda}$ and that $\lambda < \mu$ in what follows.

From Proposition 3.5.1 and Corollary 3.5.1 we can replace $\{(W_n, A_n, T_n, S_n, D_n)\}$ by a two-sided stationary ergodic joint version, $\{(W_n^0, A_n^0, T_n^0, S_n^0, D_n^0)\}$, in the following total delay recursion so that it jointly uses uses stationary ergodic versions of input:

$$Z_{n+1} = (W_n^0 + D_n^0 - A_n^0)^+ + (V_n + S_n^0 - T_n^0)^+, \ n \geq 0. \tag{3.7}$$

The first piece on the right already forms a stationary ergodic sequence; it is from Equation 3.6. We now deal with the second piece. Recalling from Corollary 3.5.1 that $E(T_n^0) = \frac{1}{\lambda}$, and our assumptions that $\lambda < \mu$, we can analogously obtain, using Proposition 3.5.1 methods, on the second piece, a jointly stationary ergodic pair $\{(W_n^0, V_n^0) : n \in \mathrm{Z}\}$, yielding a stationary ergodic version $\{Z_n^0\}$ of $\{Z_n\}$ satisfying

$$Z_{n+1}^0 = (W_n^0 + D_n^0 - A_n^0)^+ + (V_n^0 + S_n^0 - T_n^0)^+, \ n \in \mathrm{Z}\}. \tag{3.8}$$

We can also throw in $\{S_n^0\}$ to obtain a stationary ergodic sojourn time sequence via $R_n^0 = Z_n^0 + S_n^0$. Summarizing:

Theorem 3.6.1. *The ANTP model with stationary ergodic input is always stable when* $E(D) < \frac{1}{\lambda}$ *and* $\lambda < \mu$*; there always exists a unique stationary ergodic version of total delay and sojourn time. Moreover, coupling occurs: All sample paths of* $\{(W_n, V_n) : n \geq 0\}$ *with initial condition* $(W_0, V_0) = (x, y) \geq 0$ *(hence all sample paths of* $\{Z_n\}$ *with initial condition* $Z_0 = z \geq 0$*) become* *identical to those for which* $(W_0, V_0) = (0, 0)$*; thus* (W_n, V_n) *converges in total variation to the joint distribution of* (W^0, V^0) *regardless of initial conditions, and* Z_n *converges in total variation to the distribution of* $W_0 + V_0$*, regardless of initial conditions.*

3.7 I.I.D. input case; Harris recurrence of ANTP

Here we focus on the special case when each of the following two input sequences, $\{A_n\}$ and $\{(S_n, D_n)\}$, are iid and independent. Note that we are allowing S_n and D_n to be dependent for each n, because in applications (as we suggested earlier) the D_n might naturally be taken as a function of S_n.

These assumptions, which we will refer to as *the iid input case*, in particular imply that the nominal FIFO queueing model ($L_{n+1} = (L_n + S_n - A_n)^+$) forms a FIFO $GI/GI/1$ queue. Moreover, the cafe recursion, $W_{n+1} = (W_n + D_n - A_n)^+$, too can be "viewed" as that of a FIFO $GI/GI/1$ queue (but recall Footnote 1). $\{W_n : n \geq 0\}$ forms a Markov chain.

Since $T_n = t_{n+1} - t_n = A_n + W_{n+1} - W_n$ we can re-write the recursion for $\{V_n\}$ by using the Markov chain $\{W_n\}$ to drive it:

$$\begin{aligned} V_{n+1} &= (V_n + S_n - T_n)^+ && (3.9) \\ &= \Big((V_n + S_n - A_n - (W_{n+1} - W_n)\Big)^+, && (3.10) \\ &= \Big(V_n + S_n - A_n - ((W_n + D_n - A_n)^+ - W_n)\Big)^+, \ n \geq 0. && (3.11) \end{aligned}$$

Focusing on Equation (3.11), and recalling the iid assumptions, it is immediate that for $M_n \overset{\text{def}}{=} (W_n, V_n)$,

$$\{M_n \ : n \geq 0\}, \text{ forms a Markov chain on } \mathbb{R}^2_+ \qquad (3.12)$$

We next dive deeper, in the following two Propositions we exhaustively find regenerations points of two kinds. We will be invoking results for Harris recurrent Markov chains; for a reference, we refer the reader to Chapter VII, Section 3, Page 198 in [63].

Proposition 3.7.1. *[Regeneration Type I] For the stable iid input case ($\lambda < \mu$, $E(D) < \frac{1}{\lambda}$), the Markov chain $M_n = (W_n, V_n)$ forms a Harris ergodic Markov chain. If*

$$P(A_n > \max\{S_n, D_n\}) > 0, \tag{3.13}$$

then the successive times when $M_n = (0,0)$ can be chosen as positive recurrent regeneration points. In particular, both total delay, $W_n + V_n$, and total response time $W_n + V_n + S_n$ form positive recurrent regenerative processes.

Proof. From Theorem 3.6.1, $\{M_n\}$ is ergodic and converges in total variation to a limiting stationary probability distribution π, regardless of initial conditions on M_0. Thus for $A \subset \mathbb{R}^2_+$, if $\pi(A) > 0$, then regardless of initial conditions, by ergodicity,

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} I\{M_n \in A\} = \pi(A) > 0, \text{ wp1.},$$

A is visited infinitely often. Thus π serves as a recurrence measure; $\{M_n\}$ is positive Harris recurrent by definition.

Since the recursion for $\{W_n\}$ describes a stable $GI/GI/1$ queue, we know that $P_\pi(W_0 = 0) > 0$. Thus there exists a $B > 0$ such that $P_\pi(W_0 = 0,\ V_0 \leq B) > 0$. By Harris recurrence, the event $\{W_n = 0, V_n \leq B\}$ thus occurs infinitely often and does so a positive proportion of time. From the condition $p = P(A_n > \max\{S_n, D_n\} + \delta) > 0$, if we define $k = [B/\delta]$ (the smallest integer $\geq B/\delta$), and define the event

$$F_n^k = \{A_{n+i} > \max\{S_{n+i}, D_{n+i}\} + \delta,\ 0 \leq i \leq k-1\},$$

then whenever the event $\{W_n = 0, V_n \leq B\}$ occurs, the event F_n^k is independent of it and will occur with probability $p^k = P(F_n^k) > 0$.

Using Equation (3.11), suppose that for some n, both events $\{W_n = 0, V_n \leq B\}$, and F_n^k occur.

Then since $W_{n+1} = (W_n + D_n - A_n)^+$, we conclude that $W_{n+i} = 0,\ 0 \le i \le k$. This then implies that

$$\begin{aligned} V_{n+1} &= \Big(V_n + S_n - A_n - (W_{n+1} - W_n) \Big)^+ \\ &= (V_n + S_n - A_n)^+ \\ &\le \Big(B - \delta\Big)^+, \end{aligned}$$

and we can continue step-by-step to obtain $V_{n+2} \le (B - 2\delta)^+, \cdots, V_{n+k} \le (B - k\delta)^+ = 0$. Thus we have $W_{n+k} = V_{n+k} = 0$. By the Borel-Cantelli Lemma, the event $\{W_n = 0,\ F_n^k\}$ will occur infinitely often with a positive proportion of times $\ge p^k P_\pi(W_n = 0, V_n \le B) > 0$. □

Remark 3.7.1. Although the two (stability) conditions $E(A) > E(S)$, $E(A) > E(D)$ imply $P(A > S) > 0$ and $P(A > D) > 0$, they are not strong enough to imply $P(A > \max\{S, D\}) > 0$, when S and D are dependent. Counterexample: Choose $P(A = 2.6) = 1$ and choose

$$(S, D) = \begin{cases} (2, 3) & \text{w.p. } 0.5\ , \\ (3, 2) & \text{w.p. } 0.5. \end{cases}$$

Then $P(A > S) = P(S = 2) = 0.5$, $P(A > D) = P(D = 2) = 0.5$. But $P(A > \max\{S, D\}) = P(A > 3) = 0$. (We return to this example in Section 3.8.2.)

A natural sufficient condition for obtaining $P(A > \max\{S, D\}) > 0$, would be that the interarrival time distribution has unbounded support, $P(A > x) > 0,\ x \ge 0$.

Remark 3.7.2. The regenerative cycle length distribution is aperiodic because given that $M_n = 0$, there is a positive probability $P(A_n > \max\{S_n, D_n\})$, that $M_{n+1} = 0$ as well.

3.8 What if $P(A_n > \max\{S_n, D_n\}) = 0$?

Individually, each of $\{V_n\}$ and $\{W_n\}$ will empty infinitely often, a positive proportion of times. The problem is that they might not do so at the same time n. So we need to derive more involved regeneration points in such a case. There are other examples of this sort of phenomena: the classic

FIFO $GI/GI/c$ queue with $c \geq 2$ can be stable but such that the system will never be found empty by an arrival. For $c = 2$, for example, just take $A_n = 1.5,\ n \geq 0,\ S_n = 2,\ n \geq 0$. Then $\rho = \lambda/\mu = 4/3 < 2$, so stability holds. But all arriving customers (after $n = 0$) will find one server free, but the other busy. Nonetheless, for any stable ($\rho < c$) FIFO $GI/GI/c$ queue, regeneration points can be found (see, for example, Chapter VII, Section 2, Page 344 in [63]).

Here we introduce another condition (other than the stability conditions):

Proposition 3.8.1. *[Regeneration Type II]. Assume that in addition to the stability conditions, the following condition holds:*

$$P(D > S) > 0, \tag{3.14}$$

Then positive recurrent regeneration points can be found for $\{M_n\}$ of the form $(X, 0)$ where the construction of X is given explicitly in Section 3.8.1 below.

Proof. First, note that since in general, $T_n \geq D_n,\ n \geq 0$, we have

$$V_{n+1} = (V_n + S_n - T_n)^+ \leq (V_n + S_n - D_n)^+,\ n \geq 0.$$

Then we define a new random variable $\hat{V}_n$ driven by the recursion

$$\hat{V}_{n+1} = (\hat{V}_n + S_n - D_n)^+,\ n \geq 0, \tag{3.15}$$

for which we have the following bound

$$V_n \leq \hat{V}_n,\ n \geq 0, \text{ if } V_0 = \hat{V}_0. \tag{3.16}$$

Now choose $B > 0$ sufficiently large so that $P_\pi(W_0 = 0,\ V_0 \leq B) > 0$ which implies the event $\{W_n = 0,\ V_n \leq B\}$ will happen infinitely often. Choose a $\delta > 0$ such that $P(D > S + \delta) > 0$ ensured by Condition (3.14). Define $k = \lceil B/\delta \rceil$, and define the event

$$F_n^k = \{\{D_{n+i} > S_{n+i} + \delta\},\ 0 \le i \le k-1\}.$$

Now suppose that for some n, both the events $\{W_n = 0,\ V_n \le B\}$ and F_n^k occur. Then similar to the proof of Proposition 3.7.1 (we use Equations (3.15) and (3.16) and set $\hat{V}_n = V_n$), we have $\hat{V}_{n+k} = 0$ and hence $V_{n+k} = 0$.

Meanwhile, W_{n+k} was constructed from only iid $\{(D_{n+i}, A_{n+i}) : 0 \le i \le k-1\}$, *conditional on* F_n^k, and is independent of all else; that is how M_n regenerates.

We can construct (via an algorithm) such a regeneration as follows:

3.8.1 Regeneration construction

1. Let $\{(S_i, D_i) : 0 \le i \le k-1\}$ denote k iid pairs conditional on each pair satisfying $F_0^k = \{D_i > S_i + \delta\},\ 0 \le i \le k-1$.

2. Independently, let $\{A_i : 0 \le i \le k-1\}$ be iid. The A_i are not biased since Equation (3.16) does not in any way depend on their values; for any such iid sequence of A_i, $V_{n+k} = 0$ because $T_i \ge D_i$ regardless of their values; it is only the (S_i, D_i) that are biased.

3. Use as input $\{(A_i, D_i : 0 \le i \le k-1\}$ (starting with $W_0 = 0$) in the recursion
$W_{n+1} = (W_n + D_n - A_n)^+,\ 0 \le n \le k-1$.

4. Let $X = W_k$. Then when a regeneration occurs for M_n, it is distributed as $(X, 0)$.

□

3.8.2 The two conditions $P(A_n > \max\{S_n, D_n\}) > 0$ or $P(D > S) > 0$ cover all ground

If $P(D > S) > 0$ does not hold, then $P(D \le S) = 1$ and hence we have $\max\{S, D\} = S$. Since $P(A > S) > 0$ always holds by stability ($E[A] > E[S]$), we then have $P(A > \max\{S, D\}) > 0$ and we can use Proposition 3.7.1 to furnish regeneration points. Thus we have now characterized the regeneration points for *all* stable systems: If $P(A_n > \max\{S_n, D_n\}) > 0$, then regeneration points are furnished by Proposition 3.7.1, otherwise $P(D > S) > 0$ must hold and the regeneration points are given by Proposition 3.8.1

An explicit example of when $M_n \neq (0,0)$ for $n > 0$

We show here that the example introduced in Remark 3.7.1 is a case where the model never empties, and thus the regeneration points are of the more complex character in Proposition 3.8.1. To see that $M_n \neq (0,0)$ for $n > 0$, we will show that W_n and V_n move in opposite directions. Suppose $W_{n+1} - W_n \leq 0$ for some n. Observe that $W_{n+1} = (W_n + .4)^+$ when $D_n = 3$ and $W_{n+1} = (W_n - .6)^+$ when $D_n = 2$, and so $W_{n+1} - W_n \leq 0$ happens only when $(S_n, D_n) = (3, 2)$. Now by the definition of T_n we have

$$T_n = A_n + (W_{n+1} - W_n) = 2.6 + W_{n+1} - W_n \leq 2.6$$

and thus

$$V_{n+1} = (V_n + S_n - T_n)^+ = (V_n + 3 - T_n)^+ \geq (V_n + .4)^+ = V_n + .4$$

which implies that $V_{n+1} - V_n \geq 0.4$. Thus we have shown that if $W_{n+1} - W_n \leq 0$, then $V_{n+1} - V_n > 0$, and thus we also have the contrapositive: if $V_{n+1} - V_n \leq 0$, then $W_{n+1} - W_n > 0$. Thus when one coordinate of M_n decreases to 0, the other must increase to some positive value; $M_n \neq (0,0)$ for $n > 0$.

To explicitly characterize the regeneration points described in Proposition 3.8.1, we choose a $b > 0$ such that $P_\pi(W_0 = 0, V_0 \leq b) > 0$. Our objective is to find a minimal such b. Suppose the event $\{W_n = 0, V_n \leq b\}$ occurs. We can then condition on alternating $\{(S_{n+i}, D_{n+i}) : 0 \leq i \leq m-1\} = \{(2,3), (3,2), (2,3), \ldots, (3,2)\}$, for any length m, which occurs with positive probability $(1/2)^m$. Using the recursion $W_{n+1} = (W_n + D_n - A_n)^+$, we can compute that $W_{n+1} = 0.4, W_{n+2} = 0, W_{n+3} = 0.4, \ldots$ and so on, with W_{n+i} alternating between 0.4 and 0.

With the values of W_{n+i} in hand, we can use the equation $T_{n+i} = A_{n+i} + (W_{n+i+1} - W_{n+i})$ to see that $T_{n+i} = 3$ for even i and $T_{n+i} = 2.2$ for odd i. Now considering the V_n, we can compute

$$\begin{aligned}
V_{n+1} &= (V_n - 1)^+ \\
V_{n+2} &= V_{n+1} + 0.8 \\
V_{n+3} &= (V_{n+2} - 1)^+
\end{aligned}$$

and so on, with V_n going down by 1 and up by 0.8 until we have $V_{n+i} = 0$ for some i. If $V_{n+i} = 0$, we must have $W_{n+i} = 0.4$ (since $M_n \neq (0,0)$ for $n > 0$), and hence 0.4 is the minimal value such that $P_\pi(W_0 = 0, V_0 \leq 0.4) > 0$ holds. Thus, though the system does not empty, we can still define regeneration points: we take those consecutive times n such that $M_n = (0.4, 0)$.

References

[1] A. K. Erlang, “Sandsynlighedsregning og telefonsamtaler,” *Nyt tidsskrift for matematik*, vol. 20, pp. 33–39, 1909.

[2] ——, “Solution of some problems in the theory of probabilities of significance in automatic telephone exchanges,” *Post Office Electrical Engineer’s Journal*, vol. 10, pp. 189–197, 1917.

[3] ——, “Telefon-ventetider. et sykke sandsynlighedsregning,” *Matematisk Tidsskrift B*, vol. 31, pp. 25–42, 1920.

[4] J. Bergquist and K. Sigman, “Stationary workload and service times for some nonwork-conserving M/G/1 preemptive lifo queues,” *Stochastic Models*, vol. 38, pp. 515 –544, 2022.

[5] S. Asmussen and P. W. Glynn, “On preemptive-repeat LIFO queues,” *Queueing Systems*, vol. 87, pp. 1–22, 2017.

[6] R. W. Wolff, *Stochastic Modeling And The Theory Of Queues*. Prentice Hall: New Jersey, 1989.

[7] D. Fakinos, “The G/G/1 queueing system with a particular queue discipline,” *J.R. Statist. Soc.*, vol. B43, pp. 190–196, 1981.

[8] ——, “On the single-server queue with the preemptive-resume last-come- first-served queue discipline,” *Journal of Applied Probability*, vol. 23, pp. 243–248, 1986.

[9] G. Yamazaki, “The GI/G/1 queue system with last-come-first-served,” *Ann. Inst. Staist. Math.*, vol. 34, pp. 599–604, 1982.

[10] ——, “Invariance relations for GI/GI/1 queueing systems with preemptive-resume last-come-first-served queue discipline,” *J. Opns. Res. Soc. of Japan*, vol. 27, pp. 338–346, 1984.

[11] J. G. Shanthikumar and U. Sumita, “On G/G/1 queues with LIFO-P service discipline,” *J. Operations Research Society of Japan*, vol. 29, pp. 220–230, 1986.

[12] K. Sigman, *Stationary Marked Point Processes: An Intuitive Approach*. Chapman and Hall: New York, 1995.

[13] K. A Sigman, “Note on a sample-path rate conservation law and its relationship with $H = \lambda G$,” *Advances in Applied Probability*, vol. 23, no. 3, pp. 662–665, 1991.

[14] K. A. A. Sigman, “Primer on heavy-tailed distributions,” *Queueing Systems*, vol. 33, pp. 261–275, 1999.

[15] K. Sigman and R. W. A Wolff, “Review of regenerative processes,” *SIAM Review*, vol. 35, no. 2, pp. 269–288, 1993.

[16] D. W. Low, “Optimal dynamic pricing policies for an m/m/s queue,” *Operations Research*, vol. 22, no. 3, pp. 545–561, 1974.

[17] D. W. Low, “Optimal pricing for an unbounded queue,” *IBM Journal of Research and Development*, vol. 18, no. 4, pp. 290–302, 1974.

[18] G. Latouche, “On the trade-off between queue congestion and server’s reward in an m/m/1 queue,” *European Journal of Operational Research*, vol. 4, no. 3, pp. 203–214, 1980.

[19] I. Maoui, H. Ayhan, and R. D. Foley, “Optimal static pricing for a service facility with holding costs,” *European Journal of Operational Research*, vol. 197, no. 3, pp. 912–923, 2009.

[20] B. Ata and S. Shneorson, “Dynamic control of an M/M/1 service system with adjustable arrival and service rates,” *Management Science*, vol. 52, no. 11, pp. 1778–1791, 2006. eprint: `https://doi.org/10.1287/mnsc.1060.0587`.

[21] P. Naor, “The regulation of queue size by levying tolls,” *Econometrica*, vol. 37, no. 1, pp. 15–24, 1969.

[22] H. Chen and M. Z. Frank, “State dependent pricing with a queue,” *IIE Transactions*, vol. 33, no. 10, pp. 847–860, 2001. eprint: `https://doi.org/10.1080/07408170108936878`.

[23] H. Chen and M. Frank, “Monopoly pricing when customers queue,” *EconWPA, Industrial Organization*, vol. 36, Jan. 1995.

[24] C. Borgs, J. T. Chayes, S. Doroudi, M. Harchol-Balter, and K. Xu, “The optimal admission threshold in observable queues with state dependent pricing,” *Probability in the Engineering and Informational Sciences*, vol. 28, no. 1, 101–119, 2014.

[25] J. Kim and R. S. Randhawa, “The value of dynamic pricing in large queueing systems,” *Operations Research*, vol. 66, no. 2, pp. 409–425, 2018. eprint: `https://doi.org/10.1287/opre.2017.1668`.

[26] I. Paschalidis and J. Tsitsiklis, “Congestion-dependent pricing of network services,” *IEEE/ACM Transactions on Networking*, vol. 8, no. 2, pp. 171–184, 2000.

[27] S. Yoon and M. Lewis, "Optimal pricing and admission control in a queueing system with periodically varying parameters," *Queueing Systems: Theory and Applications*, vol. 47, Feb. 2004.

[28] C. Maglaras and A. Zeevi, "Pricing and capacity sizing for systems with shared resources: Approximate solutions and scaling relations," *Management Science*, vol. 49, no. 8, pp. 1018–1038, 2003. eprint: `https://doi.org/10.1287/mnsc.49.8.1018.16402`.

[29] C. Maglaras, "Revenue management for a multiclass single-server queue via a fluid model analysis," *Operations Research*, vol. 54, no. 5, pp. 914–932, 2006. eprint: `https://doi.org/10.1287/opre.1060.0305`.

[30] E. Cil, F. Karaesmen, and E. Örmeci, "Dynamic pricing and scheduling in a multi-class single-server queueing system," *Queueing Systems*, vol. 67, pp. 305–331, Apr. 2011.

[31] N. M. Edelson and D. K. Hilderbrand, "Congestion tolls for poisson queuing processes," *Econometrica*, vol. 43, no. 1, pp. 81–92, 1975.

[32] H. Mendelson and S. Whang, "Optimal incentive-compatible priority pricing for the m/m/1 queue," *Operations Research*, vol. 38, no. 5, pp. 870–883, 1990.

[33] G. P. Cachon and P. Feldman, "Pricing Services Subject to Congestion: Charge Per-Use Fees or Sell Subscriptions?" *Manufacturing & Service Operations Management*, vol. 13, no. 2, pp. 244–260, 2011.

[34] M. Haviv and R. S. Randhawa, "Pricing in queues without demand information," *Manufacturing & Service Operations Management*, vol. 16, no. 3, pp. 401–411, 2014. eprint: `https://doi.org/10.1287/msom.2014.0479`.

[35] D. Lingenbrink and K. Iyer, "Optimal signaling mechanisms in unobservable queues," *Operations Research*, vol. 67, no. 5, pp. 1397–1416, 2019. eprint: `https://doi.org/10.1287/opre.2018.1819`.

[36] O. Besbes, A. N. Elmachtoub, and Y. Sun, "Technical note—static pricing: Universal guarantees for reusable resources," *Operations Research*, vol. 70, no. 2, pp. 1143–1152, 2022. eprint: `https://doi.org/10.1287/opre.2020.2054`.

[37] S. Shakkottai, R. Srikant, A. Ozdaglar, and D. Acemoglu, "The price of simplicity," *IEEE Journal on Selected Areas in Communications*, vol. 26, no. 7, pp. 1269–1276, 2008.

[38] D. Lee, J. Mo, G. Jin, and J. Park, "Price of simplicity under congestion," *IEEE Journal on Selected Areas in Communications*, vol. 30, no. 11, pp. 2158–2168, 2012.

[39] N. Gans and S. Savin, "Pricing and capacity rationing for rentals with uncertain durations," *Management Science*, vol. 53, no. 3, pp. 390–407, 2007. eprint: https://doi.org/10.1287/mnsc.1060.0651.

[40] H. Xu and B. Li, "Dynamic cloud pricing for revenue maximization," *IEEE Transactions on Cloud Computing*, vol. 1, no. 2, pp. 158–171, 2013.

[41] A. Waserhole and V. Jost, "Pricing in vehicle sharing systems: Optimization in queuing networks with product forms," *EURO Journal on Transportation and Logistics*, vol. 5, Jan. 2013.

[42] S. Banerjee, R. Johari, and C. Riquelme, "Pricing in ride-sharing platforms: A queueing-theoretic approach," in *Proceedings of the Sixteenth ACM Conference on Economics and Computation*, ser. EC '15, Portland, Oregon, USA: Association for Computing Machinery, 2015, p. 639, ISBN: 9781450334105.

[43] S. Banerjee, D. Freund, and T. Lykouris, "Pricing and optimization in shared vehicle systems: An approximation framework," in *Proceedings of the 2017 ACM Conference on Economics and Computation*, ser. EC '17, Cambridge, Massachusetts, USA: Association for Computing Machinery, 2017, p. 517, ISBN: 9781450345279.

[44] C.-A. Lin, K. Shang, and P. Sun, "Wait time–based pricing for queues with customer-chosen service times," *Management Science*, vol. 0, no. 0, null, 2022. eprint: https://doi.org/10.1287/mnsc.2022.4474.

[45] S. Benjaafar and X. Shen, "Pricing in on-demand and one-way vehicle-sharing networks," *Operations Research*, 2023.

[46] R. B. Myerson, "Optimal auction design," *Mathematics of Operations Research*, vol. 6, no. 1, pp. 58–73, 1981. eprint: https://doi.org/10.1287/moor.6.1.58.

[47] J. S. Kaufman, "Blocking in a shared resource environment," *IEEE Trans. Commun.*, vol. 29, pp. 1474–1481, 1981.

[48] A. O. Allen, "Appendix c - queueing theory formulas," in *Probability, Statistics, and Queuing Theory with Computer Science Applications (Second Edition)*, ser. Computer Science and Scientific Computing, Second Edition, San Diego: Academic Press, 1990, pp. 665–714, ISBN: 978-0-08-057105-8.

[49] G. Cardano, "Ars magna," 1545.

[50] E. Gelenbe and E. Ngai, "Adaptive qos routing for significant events in wireless sensor networks," *2008 5th IEEE International Conference on Mobile Ad Hoc and Sensor Systems*, pp. 410–415, 2008.

[51] ——, “Adaptive random re-routing for differentiated qos in sensor networks,” *The Computer Journal*, vol. 53, no. 7, pp. 1052–1061, 2010.

[52] F. Ghavimi and H.-H. Chen, “M2m communications in 3gpp lte/lte-a networks: Architectures, service requirements, challenges, and applications,” *IEEE Communications Surveys & Tutorials*, vol. 17, no. 2, pp. 525–549, 2015.

[53] S.-Y. Lien, T.-H. Liau, C.-Y. Kao, and K.-C. Chen, “Cooperative access class barring for machine-to-machine communications,” *IEEE Transactions on Wireless Communications*, vol. 11, no. 1, pp. 27–32, 2012.

[54] T.-M. Lin, C.-H. Lee, J.-P. Cheng, and W.-T. Chen, “Prada: Prioritized random access with dynamic access barring for mtc in 3gpp lte-a networks,” *IEEE Transactions on Vehicular Technology*, vol. 63, no. 5, pp. 2467–2472, 2014.

[55] A. Aijaz and A. H. Aghvami, “Cognitive machine-to-machine communications for internet-of-things: A protocol stack perspective,” *IEEE Internet of Things Journal*, vol. 2, no. 2, pp. 103–112, 2015.

[56] A. Aijaz, S. Ping, M. R. Akhavan, and A.-H. Aghvami, “Crb-mac: A receiver-based mac protocol for cognitive radio equipped smart grid sensor networks,” *IEEE Sensors Journal*, vol. 14, no. 12, pp. 4325–4333, 2014.

[57] I. Park, D. Kim, and D. Har, “Mac achieving low latency and energy efficiency in hierarchical m2m networks with clustered nodes,” *IEEE Sensors Journal*, vol. 15, no. 3, pp. 1657–1661, 2015.

[58] L. Liang, L. Xu, B. Cao, and Y. Jia, “A cluster-based congestion-mitigating access scheme for massive m2m communications in internet of things,” *IEEE Internet of Things Journal*, vol. 5, no. 3, pp. 2200–2211, 2018.

[59] M. Nakip, V. Rodoplu, C. Güzeliş, and D. T. Eliiyi, “Joint forecasting-scheduling for the internet of things,” *2019 IEEE Global Conference on Internet of Things (GCIoT)*, pp. 1–7, 2019.

[60] V. Rodoplu, M. Nakip, R. Qorbanian, and D. T. Eliiyi, “Multi-channel joint forecasting-scheduling for the internet of things,” *IEEE Access*, vol. 8, pp. 217 324–217 354, 2020.

[61] E. Gelenbe, M. Nakıp, D. Marek, and T. Czachorski, “Diffusion analysis improves scalability of iot networks to mitigate the massive access problem,” *2021 29th International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS)*, pp. 1–8, 2021.

[62] E. Gelenbe and K. Sigman, “Iot traffic shaping and the massive access problem,” *IEEE International Conference on Communications*, pp. 2732–2737, 2022.

[63] S. Asmussen, *Applied Probability and Queues (2nd Edition)*. Springer, 2003.

[64] S. Asmussen and P. W. Glynn, *Stochastic Simulation*. Springer-Verlag, 2007, New York.

[65] H. Schmidli, *Risk Theory*. Springer Actuarial Series, 2017.

[66] S. Karlin and H. M. Taylor, *A First Course in Stochastic Processes*, 2nd Edition. Academic Press, 1975.

Appendix A: Appendix A: Stationary workload for some non-work-conserving M/G/1 preemptive LIFO queues

Proof of Theorem 1.2.1

Proof. Let $N(\lambda) = N$ denote the stationary number-in-system. By Proposition 1.2.3, and Equation (1.10) $N(\lambda)$ is a geometric random variable with success probability $p_0(\lambda) = \frac{2E(e^{-\lambda S})-1}{E(e^{-\lambda S})}$. Since $p_0(\lambda) \downarrow 0$ as $\lambda \uparrow \lambda_2$, it is easily seen that

$$p_0(\lambda)N(\lambda) \Longrightarrow exp(1), \tag{A.1}$$

as $\lambda \uparrow \lambda_2$. Because $p_0 \to 0$, it follows from Proposition 1.2.3 that we only need to consider $\hat{V} \overset{\mathrm{d}}{=} (V \mid V > 0)$ and we have

$$p_0(\lambda)\hat{V}(\lambda) \overset{\mathrm{d}}{=} p_0(\lambda)\hat{S}_r(\lambda) + p_0(\lambda)\sum_{i=1}^{N(\lambda)} S_i. \tag{A.2}$$

By Markov's Inequality and Lemma 1.2.2, $p_0(\lambda)\hat{S}_r(\lambda) \overset{p}{\Longrightarrow} 0$ as $\lambda \uparrow \lambda_2$ and hence can be ignored. Now to handle the second term, we write

$$p_0(\lambda)\sum_{i=1}^{N(\lambda)} S_i = \frac{\sum_{i=1}^{N(\lambda)} S_i}{N(\lambda)} p_0(\lambda)N(\lambda), \tag{A.3}$$

and note that

$$\frac{\sum_{i=1}^{N(\lambda)} S_i}{N(\lambda)} \overset{p}{\Longrightarrow} E[S] = \frac{1}{\mu}, \tag{A.4}$$

as $\lambda \uparrow \lambda_2$ because $\frac{1}{n}\sum_{i=1}^{n} S_i \overset{a.s.}{\to} E[S]$ as $n \to \infty$ and $N(\lambda) \overset{p}{\Longrightarrow} \infty$ as $\lambda \uparrow \lambda_2$.

Thus it follows from Equations (A.1) and (A.3) that

$$p_0(\lambda) \sum_{i=1}^{N(\lambda)} S_i \Longrightarrow exp(\mu)$$

□

Proof of Theorem 1.3.1

Proof. Using the representation (1.36), the proof follows the same steps as in Theorem 1.2.1, but there is additional complexity in establishing the convergence in probability as in Equation (A.4) due to the fact that $\hat{S} = \hat{S}(\lambda)$ depends on λ, and i.i.d. copies of it are included in the sum; we need to establish

$$\frac{\sum_{i=1}^{N(\lambda)} \hat{S}_i(\lambda)}{N(\lambda)} \overset{p}{\Longrightarrow} E(\hat{S}_i(\lambda_2)), \tag{A.5}$$

as $\lambda \uparrow \lambda_2$.

To this end, we first construct the random variables by using the inverse transform method from simulation to couple them (see for example, Pages 37-38 in [64]): Let $U, U_1, U_2 \ldots$ denote i.i.d. copies of $Unif(0,1)$ random variables (one can take, for example, as the probability space for a uniform U, the interval $(0,1)$ under Lesbesgue measure and take $U(\omega) = \omega$, the identity map). We use the same U for constructing $N(\lambda)$ for all $\lambda < \lambda_2$, and we use U_i for constructing $\hat{S}_i(\lambda)$ for all $\lambda \le \lambda_2$. Then we re-write

$$\frac{\sum_{i=1}^{N(\lambda)} \hat{S}_i(\lambda)}{N(\lambda)} = \frac{\sum_{i=1}^{N(\lambda)} \hat{S}_i(\lambda_2)}{N(\lambda)} + A, \tag{A.6}$$

where

$$A = \frac{1}{N(\lambda)} \sum_{i=1}^{N(\lambda)} (\hat{S}_i(\lambda_2) - \hat{S}_i(\lambda)),$$

the error term. If we can show that A converges to 0 in probability as $\lambda \uparrow \lambda_2$, hence can be ignored, then in the right-hand side of Equation (A.6) the i.i.d. summands only depend on λ_2, and we get the result exactly as was done for the PRD case in the proof of Theorem 1.2.1. To this end, observe that since $N(\lambda)$ is independent of the i.i.d. $\hat{S}_i(\lambda_2) - \hat{S}_i(\lambda)$, we have (via first conditioning on $N(\lambda)$)

that

$$E(|A|) \leq E(|\hat{S}(\lambda_2) - \hat{S}(\lambda)|),$$

where $\hat{S}(\lambda_2) - \hat{S}(\lambda)$ denotes $\hat{S}_1(\lambda_2) - \hat{S}_1(\lambda)$.

It thus suffices to show that $E(|\hat{S}(\lambda_2) - \hat{S}(\lambda)|) \to 0$.

We now note that due to our coupling from the inverse transform method (and the fact that from Proposition 1.3.7, it follows that $\hat{S}(\lambda)$ converges in distribution to $\hat{S}(\lambda_2)$ as $\lambda \uparrow \lambda_2$), that $\hat{S}(\lambda) \to \hat{S}(\lambda_2)$ w.p. 1 as $\lambda \uparrow \lambda_2$, and thus $|\hat{S}(\lambda_2) - \hat{S}(\lambda)| \to 0$, w.p. 1.

It thus suffices to show that the collection $\{|\hat{S}(\lambda_2) - \hat{S}(\lambda)|\}$ is uniformly integrable (in $\lambda \leq \lambda_2$). We do so by noting further that $|\hat{S}(\lambda_2) - \hat{S}(\lambda)| \leq X(\lambda) = \hat{S}(\lambda_2) + \hat{S}(\lambda)$, and $\{X(\lambda)\}$ is uniformly integrable: It is a non-negative collection such that $X(\lambda) \to X(\lambda_2) = 2\hat{S}(\lambda_2)$ w.p. 1, and $E(X(\lambda)) \to E(X(\lambda_2)) = 2E(\hat{S}(\lambda_2)) < \infty$ (finite by assumption). □

Proof of Proposition 1.3.11

Proof. We follow the key renewal theorem type proof that can be used to prove the *Cramér-Lundberg* theorem/approximation for the standard M/G/1 queue (or in the context of ruin theory), see for example Theorem 5.7 Page 95 in [65]. Because Q is geometric, letting $\hat{S}$ be distributed as in Proposition 1.3.7 with density $f_{\hat{S}}(x)$ and independent of V_Q, we have

$$(V_Q \mid V_Q > 0) \overset{\mathrm{d}}{=} \hat{S} + V_Q, \tag{A.7}$$

and thus $P(V_Q > x) = (1 - p_0)P(\hat{S} + V_Q > x)$, which in turn yields

$$P(V_Q > x) = (1 - p_0)P(\hat{S} > x) + \int_0^x P(V_Q > x - y)(1 - p_0) f_{\hat{S}}(y)dy. \tag{A.8}$$

Although $(1-p_0)f_{\hat{S}}(y)$ is not a density function, $g_\gamma(y) = (1-p_0)e^{\gamma y} f_{\hat{S}}(y)$ is by the definition of γ. Multiplying each side of Equation (A.8) by $e^{\gamma x}$ thus yields a renewal equation:

$$e^{\gamma x}P(V_Q > x) = (1-p_0)e^{\gamma x}P(\hat{S} > x) + \int_0^x e^{\gamma(x-y)}P(V_Q > x-y)g_\gamma(y)dy. \tag{A.9}$$

We now can apply the key renewal theorem, justified by: S is assumed to have a density thus so does $\hat{S}$ (in particular it is non-lattice), $(1-p_0)R$ (the mean of g_γ) is finite, and the function $h(x) = (1-p_0)e^{\gamma x}P(\hat{S} > x) = \int_x^\infty g_\gamma(y)dy - \gamma\int_x^\infty h(y)dy$, the difference of two decreasing Riemann integrable (hence directly Riemann integrable) functions, hence is directly Riemann integrable (see Equation 5.2 Page 190 in [66]). Thus

$$\begin{aligned}
\lim_{x\to\infty} e^{\gamma x}P(V_Q > x) &= \frac{1}{(1-p_0)R}\int_0^\infty (1-p_0)e^{\gamma x}P(\hat{S} \geq x)dx \\
&= \frac{1}{R}\int_0^\infty e^{\gamma x}P(\hat{S} \geq x)dx \\
&= \frac{1}{R}E\Big(\int_0^S e^{\gamma x}dx\Big) \\
&= \frac{p_0}{R\gamma(1-p_0)} = C.
\end{aligned}$$

□

Proof of Lemma 1.3.2

Proof. Using L'Hôpital's rule on

$$\frac{P(\hat{S}_r > x)}{e^{-(\alpha-\lambda)x}},$$

we take the limit of the ratios of the densities:

$$(1-p_0)^{-1}\frac{\lambda e^{\lambda x}P(S > x)}{(\alpha-\lambda)e^{-(\alpha-\lambda)x}} \sim (1-p_0)^{-1}\frac{\lambda e^{\lambda x}ce^{-\alpha x}}{(\alpha-\lambda)e^{-(\alpha-\lambda)x}} \sim (1-p_0)^{-1}\frac{\lambda c}{\alpha-\lambda}.$$

Thus $P(\hat{S}_r > x) \sim (1-p_0)^{-1}\frac{\lambda c}{\alpha-\lambda}e^{-(\alpha-\lambda)x}$.

Now we prove the existence of the Lundberg constant γ, $E(e^{\gamma\hat{S}}) = (1-p_0)^{-1}$, by first finding an equivalent condition for its existence. By Proposition 1.3.3, we have $E(e^{\lambda S}) - 1 = 1 - p_0$, and then using Proposition 1.3.7, we have

$$E(e^{\gamma \hat{S}}) = (1-p_0)^{-1} \int_0^{\infty} e^{\gamma x} g(x)(e^{\lambda x} - 1)dx = (1-p_0)^{-1}\left[E(e^{(\gamma+\lambda)S}) - E(e^{\gamma S})\right].$$

Thus $E(e^{\gamma \hat{S}}) = (1-p_0)^{-1}$ if and only if $E(e^{(\gamma+\lambda)S}) - E(e^{\gamma S}) = 1$.

Since we assume that $\alpha > \lambda$, the exponential tail asymptotic Equation 1.38 implies that $E(e^{sS}) < \infty$ for all $0 \leq s < \alpha$, and $E(e^{\alpha S}) = \infty$. Thus the function $H(s) = E(e^{(s+\lambda)S}) - E(e^{sS})$ is strictly increasing and continuous in $s \geq 0$ such that $H(0) = E(e^{\lambda S}) - 1 = 1 - p_0$ implying that $0 < H(0) < 1$. Therefore as long as $E(e^{(s+\lambda)S}) < \infty$ for sufficiently large s, a solution γ to $H(s) = 1$ exists. But $H(s) < \infty$ for $s < \lambda - \alpha$ and tends to ∞ as $s \uparrow \lambda - \alpha$; it must hit the value 1 for some s. □