ANALYSIS AND COMPARISON OF QUEUES WITH DIFFERENT LEVELS OF DELAY INFORMATION

by

Pengfei Guo

Department of Business Administration
Duke University

Date: ______________________________

Approved:

____
Paul H. Zipkin, Supervisor

____
Vidyadhar G. Kulkarni

____
Pranab Majumder

____
Jing-Sheng (Jeannette) Song

Dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy
in the Department of Business Administration
in the Graduate School of
Duke University

2007

ABSTRACT

ANALYSIS AND COMPARISON OF QUEUES WITH DIFFERENT LEVELS OF DELAY INFORMATION

by

Pengfei Guo

Department of Business Administration
Duke University

Date: ______________________
Approved:

____
Paul H. Zipkin, Supervisor

____
Vidyadhar G. Kulkarni

____
Pranab Majumder

____
Jing-Sheng (Jeannette) Song

An abstract of a dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy
in the Department of Business Administration
in the Graduate School of
Duke University

2007

Abstract

Information about delays can enhance service quality in many industries. Delay information can take many forms, with different degrees of precision. Different levels of information have different effects on customers and so on the overall system. The goal of this research is to explore these effects. We first consider a queue with balking under three levels of delay information: No information, partial information (the system occupancy) and full information (the exact waiting time). We assume Poisson arrivals, independent, exponential service times, and a single server. Customers decide whether to stay or balk based on their expected waiting costs, conditional on the information provided. By comparing the three systems, we identify some important cases where more accurate delay information improves performance. In other cases, however, information can actually hurt the provider or the customers. We then investigate the impacts on the system of different cost functions and weight distributions. Specifically, we compare systems where these parameters are related by various stochastic orders, under different information scenarios. We also explore the relationship between customer characteristics and the value of information. The results here are mostly negative. We find that the value of information need not be greater for less patient or more risk-averse customers. After that, we extend our analysis to systems with phase-type service times. Our analytical and numerical results indicate that the previous conclusions about systems with exponential service times still hold for phase-type service times. We also show that service-time variability degrades the system's performance. At last, we consider two richer models of information: In the first model, an arriving customer learns an interval in which the system occupancy falls. In the second model, each customer's service time is the sum of a geometric number of i.i.d. exponential phases, and an arriving customer learns

the total number of phases remaining in the system. For each information model, we compare two systems, identical except that one has more precise information. We study the effects of information on performance as seen by the service provider and the customers.

Acknowledgements

My journey exploring the effects of delay information on service systems has been long and exciting. There are many people and things I shall forever remember and appreciate. Six years ago, I was attracted to the operations field mainly by deterministic optimization models. Soon, I was led to a more exciting field – stochastic processes – by my advisor, Prof. Paul Zipkin, who always emphasizes that the world is stochastic. After taking courses from Prof. Shaler Stidham and Prof. Vidyadhar Kulkarni (both experts on stochastic processes), I stepped in this unknown field by starting a small research project with Prof. Zipkin as early as the winter break of 2002. In that project, we considered an $M/M/1$ queue and three levels of delay information. From 2002 to 2007, the project kept growing, and finally its results formed a dissertation. Now, coming to the end of this journey, I am deeply indebted to my advisor, Prof. Paul Zipkin, who gave me tremendous support and direction, leading me through this tough research project. From him, I learned knowledge, skills, and the spirit of persistence to do scientific research. Special thanks are due to my committee members, Pranab Majumder, Vidyadhar Kulkarni and Jing-Sheng (Jeannette) Song, for their advice, help and support throughout the process. I thank other Duke Operations faculty members, Fernando Bernstein, Francis de Vericourt, Otis Jennings, Gürhan Kök and Kevin Shang for their good advice. I would also like to thank my fellow Ph.D. students, Suri Gurumurthi, Isilay Talay, Lei Xie and Zhengliang Xue. I especially enjoyed taking courses and discussing challenging questions with them. Finally, I would like to thank my wife Yulan and my son Allan for their support and encouragement.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Nowadays there are increasing possibilities, enabled by new technologies, to provide useful delay information to customers, to enhance the values of the services they receive. For example, in a call center, the provider can announce the expected waiting time to each caller; standard call-center software can do this automatically. In transportation and e-shopping, a customer can easily learn the order status and the estimated shipping time. A quote for production services normally includes a lead-time estimate. In a busy hospital emergency room, information about the anticipated wait is important to an anxious patient. Delay notification is also widely used in traffic-flow control and has even been suggested for control of computer networks (Kelly, 2000).

With developments in technology and managerial practice, it is becoming easy for the provider to acquire and convey to customers fairly accurate information about anticipated delays due to congestion. Such information can directly affect customer satisfaction and also influence customers' behavior. Information can take many forms, with different degrees of precision. Different levels of information have different effects on customers' decisions and thus the overall arrival process. A basic research question arising is: Is more information good?

Since there are two parties in the system, the service provider and the customers, the above question actually has two sides: Is more information good for the service provider? And, is more information good for customers? These two parties' interests may conflict: A server could try to increase his profit by overloading his capacity and thus hurt customers' welfare. A further research question is: Are the provider's and the customers' incentives aligned with more information? These are basic research questions which the dissertation aims to address.

An overview of customer psychology in waiting situations, including the impact of uncertainty, can be found in Maister (1984). There is some empirical evidence about customers' reactions to delays. Taylor (1994) shows that delays affect customers' overall service evaluations. Hui and Tse (1996) and Kumar et al. (1997) study the relationship between information and customer satisfaction. Munichor and Rafaeli (2005) examine the impact on customers' waiting-time perceptions of the use of various waiting-time fillers. Zhou and Soman (2003) examine the determinants of reneging behavior.

There is a substantial literature on queues with impatient customers. Models with balking and reneging (leaving after waiting for some time) can be found in many books, e.g. Kulkarni (1995). Recent works on this topic include Bae, Kim and Lee (2001), Zohar et al. (2002), and Ward and Glynn (2003).

The literature on customers influenced by delay information begins with Naor (1969), who studies a system like ours with partial information, but with identical customers and linear waiting cost. Also, the cost depends on the whole sojourn time, not just the delay. He points out that this system with its self-selecting customers suffers from *externalities*; an arriving customer who stays imposes delays on later customers, but ignores them in making his decision. Consequently, too many customers stay. If everyone were altruistic and acted to maximize the average utility, some of those customers would leave. He shows that a price can steer the system to this "social" optimum. He points out, however, that if the price is determined by the provider to maximize revenue, the provider becomes a monopolist and behaves like one. He sets the price higher than the socially optimal one and thus serves too few customers.

These features lead to additional peculiarities. Edelson and Hildebrandt (1975) mention that a revenue-maximizing service provider may make socially wrong decisions about service capacity, either in the service rate or the number of servers. They also show that these effects disappear when balking is not allowed. Schroeter (1982) considers non-identical customers with uniformly distributed costs. Gavish and Schweitzer (1973) study a full-information system under assumptions similar to Naor's. See Stidham (1985) and Hassin and Haviv (2003) for a survey on such research.

Hassin (1986) shows that the server may sometimes prefer less information to more. More recently, Mandelbaum and Shimkin (2000) and Shimkin and Mandelbaum (2004) discuss equilibria with respect to reneging in a setting similar to our no-information model, with linear and nonlinear waiting cost functions, respectively. Afèche and Mendelson (2004) study revenue-maximizing and socially optimal equilibria under uniform pricing with no information. The paper also discusses priority auctions. Whitt (1999) studies two systems corresponding roughly to our models with no and partial information. His customer-choice mechanism, however, is quite different from ours, and so are his findings about the impact of information on performance. There, information always reduces both waiting and throughput. Thus, despite their similar motivation, his models represent very different behavior from ours.

Armony and Maglaras (2004a, 2004b) analyze systems where an arriving customer learns some delay information and then can choose to balk, wait, or leave a message, in which case the provider calls back within a guaranteed time. That guaranteed time is an estimate based on a heavy-traffic approximation. In that heavy-traffic regime, indeed, such estimates become nearly precise. In Armony and Maglaras (2004a) customers' choices are based on the equilibrium waiting time, as in our no-information system. Customers employ a utility function to assess delays, but the function's argument is the expected delay, a constant. This approach is justified in the heavy-traffic limit, but it thereby suppresses the risk-reduction role of information. In Armony and Maglaras (2004b) each arriving customer receives more information, a point estimate of delay based on the system occupancy. The customer treats the estimate as exact, again based on the heavy-traffic limit. Comparing the results with the other system, they show that more information improves performance on several dimensions. Thus, their modeling approach is quite different from ours, and their findings support a more optimistic view of the role of information.

Another stream of research explores lead-time quotation in production, e.g., Duenyas and Hopp (1995) and Spearman and Zhang (1999). The system studied by Dobson and Pinker (2000) is related to ours. The provider quotes a number, the nominal lead-time, to every arriving customer. This is understood by all parties to mean a certain fixed fractile of the lead-time distribution. The provider himself may use different levels of information to assess this distribution and hence the fractile, but customers see only the nominal leadtime. Their responses are determined by a demand function – more customers stay when the quoted lead-time is shorter. The paper shows that the impact of more information depends on the shape of this demand function, a notion related to our findings about the cost-scale distribution. It only examines the impact on the provider, however, not the customers. Whang (1993) obtains similar results for a somewhat different formulation. He shows that sharing information is beneficial for the provider, if the inverse demand function is convex.

Recently, Armony, et al. (2005) study the impacts of state-dependent delay information on many-server queue in the fluid approximation framework. Altman and Hassin (2002) study an admission control problem in a one-server queueing system with general service times where customers are identical and are informed of the queue length upon arrival. They provide a counter-example for the optimality of the threshold-type policy and obtain an ϵ-optimal policy. Collins and Brooms (2005) study the equilibrium for a Bernoulli feedback queueing system, where customers' service times are affected by future arrivals. Hassin (2006) studies pricing and information issues in a balking queue.

Throughout the dissertation, to assess the effects of information consistently, we posit a customer-decision mechanism common to all levels of information. A given function, $c(w)$, measures the *basic* cost of delay. Different customers, however, value time differently. Each customer arrives with a specific parameter θ, which scales the basic cost function. Upon arrival the customer receives information, which affects his estimate of the distribution of delay. Based on his scale parameter and the information, the customer computes his *expected* delay cost. If that is more than the reward, r, which he anticipates from receiving service, he balks, and if not, he stays. In summary, customers' expected utility function is $U = r - \theta E[c(W)]$. In this scheme, then, different levels of information lead to different delay distributions in the expected-cost calculations, and those in turn affect everything else.

The cost-scale parameter θ has a cumulative distribution function (cdf) H. As explained later, the elasticity of H is crucial in determining whether information is good or not for customers or the service provider. Also, we characterize customers' risk attitudes towards delay uncertainty through the basic waiting-cost function. For example, a linear cost function ($c(w) = 1 + w$) represents risk-neutral customers, and a quadratic function ($c(w) = 1 + w^2$) risk-averse customers. A square-root cost function ($c(w) = 1 + \sqrt{w}$) represents risk-seeking customers. It turns out, interestingly, that customers' risk-attitudes are not important in deciding information's qualitative effects.

Social optimization aims to achieve the best outcome for everyone. It usually requires some central coordination scheme, such as admission control. There, more information is always better than less (absent information-processing costs), regardless of the objective. The controller can choose to ignore additional information, so the less-information solution is feasible for the more-information case. In our system, with its individual decisions, the matter is less obvious. We show for some important cases that more information does improve performance, for the provider or the customers or both, but for other cases it does not.

The contents of the dissertation are divided into four parts. In Chapter 2, we consider a single server queueing system with exponential service times. We assume potential customers arrive in a Poisson process. We consider three typical types of delay information: none, the system occupancy, and the exact waiting time. With *no information*, customers still estimate their waiting times, but these estimates are based only on long-term (equilibrium) experience, not real-time information. The occupancy provides *partial information*; the remaining uncertainty comprises the actual service times of the waiting customers. The exact waiting time gives the customer *full information.* We show how to obtain the equilibrium system behavior for each level of information. In some cases, the solution can be found in closed form. Then, we compare the three systems and obtain several important results about the fundamental questions above.

In Chapter 3, we conduct sensitivity analysis of the cost function and H. Under each level of delay information, we consider two systems with different cost and H functions. We consider certain first-order and second-order stochastic relationships between these two systems, and compare their performance measures. In this part, we also explore the relationship between the value of information and the level of customers' risk aversion.

In Chapter 4, we extend the analysis to a system with phase-type service times.

We derive the equilibrium of the systems under no, partial and full delay information. We take analytical comparison of no and full information systems. Then we carry out a numerical study of these systems by considering exponential, hyper-exponential, and generalized-Erlang distributions of service times.

In Chapter 5, we consider two richer models of delay information. In the first model, an arriving customer learns an interval in which the system occupancy falls. In the second model, each customer's service time is the sum of a geometric number of i.i.d. exponential phases, and an arriving customer learns the total number of phases remaining in the system. For each information model, we compare two systems, identical except that one has more precise information. We study the effects of information on performances as seen by the service provider and customers.

Chapter 2

$\cdot/M/1$ Queue and Delay Information

In this chapter, we consider a queueing system with exponential service times. This chapter is organized as follows: Section 1 introduces the basic formulation. Sections 2-4 develop the models for no information, partial information, and full information, respectively. Section 5 compares the three systems analytically. Section 6 treats an important extension. Section 7 presents some concluding remarks. A supplement containing proofs and other technical material is included in the Appendix.

2.1 Formulation and Preliminaries

Potential customers arrive according to a Poisson process with rate λ. There is a single server, and the service times are independent and exponentially distributed with mean $1/\mu$. The system uses the FCFS discipline.

2.1.1 Customer Behavior

We suppose that a customer's utility equals a reward for receiving service minus a waiting cost. This waiting cost depends on a customer-specific parameter and the expectation of a common function of the waiting time. (In some applications, such as production, the total sojourn time is more important than the delay. The approach can be extended to that case.) Specifically, define

- W = waiting time in queue
- θ = customer-type parameter, indicating the importance of time, $\theta \in [0, 1]$
- H = cumulative distribution function of θ, continuous on $[0, 1]$, with density h

- $c(w) =$ basic cost to wait time w, a positive, increasing, unbounded, continuous function

- $r =$ reward to the customer for receiving service, $r > 0$.

The service reward r is the same for all customers (this assumption is convenient but not essential). Customers differ in the importance of time. This difference is expressed by the customer type θ. Each customer's type is independent of all other events. The customer assesses the distribution of waiting time W, based on the available information. The expected waiting cost $\theta E[c(W)]$ is a function of that information. The utility U for the customer to stay is then

$$U = r - \theta E[c(W)].$$

The customer remains in the system if U is non-negative and otherwise balks.

We can rescale r and c so that $r = 1$. Suppose $c(0) > 1$. Then, some potential customers, precisely those with $\theta > 1/c(0)$, always balk. We can simply ignore them and scale down λ and c to represent the other customers. Thus, we can assume $c(0) \leq 1$. For convenience, we assume for now that $c(0) = 1$. Thus, the reward is just large enough to attract the most sensitive customers when there is no delay. We defer the case $c(0) < 1$ to Section 6.

Different cost functions express different sensitivities to risk, just like utility functions for wealth in finance and economics. A strictly convex cost means a strong aversion to risk, while a linear cost expresses indifference to risk. On the other hand, it is easy to think of situations where the marginal cost of waiting decreases, so the cost is concave. ("We've already waited a whole hour, a few more minutes won't make any difference.")

2.1.2 Average Utility and Throughput

Let I stand for information, a random variable with possible values i. Given information $I = i$, an arriving customer computes the expected waiting cost, which we can write $E_W[c(W)|I = i]$. The customer stays, then, precisely when his θ is less than or equal to the critical level $\theta_i = 1/E_W[c(W)|I = i]$. In these terms, the overall fraction of customers who stay is $E_I[H(\theta_I)]$, the throughput is $\lambda E_I[H(\theta_I)]$, and the probability the server is busy is $(\lambda/\mu)E_I[H(\theta_I)]$.

Define

$$J(\theta) = (1/\theta) \int_0^\theta H(\phi) d\phi.$$

The average expected utility for customers is

$$\begin{aligned}
u &= E[U^+] = E_{\theta,I}[\ \ [\ 1 - \theta E_W[c(W)|I]\ ]^+\ \ ] \\
&= E_I\left[\int_0^{\theta_I} (1 - \phi E_W[c(W)|I])\, h(\phi) d\phi\right] \\
&= E_I\left[H(\theta_I) - (1/\theta_I)\int_0^{\theta_I} \phi h(\phi) d\phi\right] \\
&= E_I[J(\theta_I)].
\end{aligned}$$

2.2 No Information

First consider the situation where the queue is invisible, and the provider tells the customer nothing. Suppose W is the equilibrium waiting time, and all customers know its distribution. An arriving customer has two choices, stay or balk. The customers who stay are precisely those with $\theta \leq \theta_-$, where $\theta_- = 1/E[c(W)]$. Consequently, the fraction of customers who stay is $H(\theta_-)$, and the effective arrival process is Poisson with rate $\lambda_- = \lambda H(\theta_-)$. This effective arrival rate affects W, hence $E[c(W)]$, and hence θ_-. We assume that these parameters arrive at consistent, i.e., equilibrium values. (For discussions of such equilibria in related models, see Stidham 1985 and Hassin and Haviv 2003.)

In sum, the equilibrium arrival rate λ_- solves

$$\lambda_- = \lambda H\left(\frac{1}{E[c(W|-)]}\right). \tag{2.1}$$

Here, $E[c(W|-)]$ indicates the expected cost given λ_-. Assume that it is finite for any $\lambda_- < \mu$.

Proposition 1. *For no information, there exists a unique equilibrium arrival rate* λ_-*.*

(The proof of this and the other results are in the supplement.) The assumption of finite $E[c(W|-)]$ is necessary. Consider $c(w) = e^{\beta w}$ with $\beta \geq \mu$ and exponential service times. For any $\lambda_- > 0$, $E[c(W|-)] = \infty$, while for $\lambda_- = 0$, $E[c(W|-)] = 1$. So, there is no equilibrium. In words, if nobody comes, everybody wants to come, but if anybody comes, nobody wants to come.

Since all customers receive the same information (none), the system behaves as an $M/M/1$ queue. Thus, W has the truncated exponential (or impulse-exponential) distribution with rate $\mu(1-\rho_-)$ and mass $1-\rho_-$ at 0, where $\rho_- = \lambda_-/\mu$. Consequently,

$$E[c(W|-)] = (1-\rho_-) + \rho_- \left\{\mu(1-\rho_-)\tilde{c}[\mu(1-\rho_-)]\right\},$$

where $\tilde{c}$ denotes the Laplace transform of c. Thus, the equilibrium ρ solves

$$\rho = (\lambda/\mu)H\left(\frac{1}{(1-\rho) + \rho\left\{\mu(1-\rho)\tilde{c}[\mu(1-\rho)]\right\}}\right). \tag{2.2}$$

The average utility is then $J(\theta_-)$, where $\theta_- = 1/E[c(W|-)]$.

Example 1: Uniform customers with linear cost

Suppose H is the uniform distribution, and $c(w) = 1+w$. Then, $\tilde{c}(s) = 1/s+1/s^2$, and (4.1) becomes

$$\rho = \frac{\lambda/\mu}{1+\rho/[\mu(1-\rho)]},$$

or

$$(1-\mu)\rho^2 + (\mu+\lambda)\rho - \lambda = 0, \tag{2.3}$$

a quadratic equation. For $\mu = 1$, the root of this equation is

$$\rho = \frac{\lambda}{1+\lambda}.$$

For $\mu \neq 1$, the positive root less than 1 is

$$\rho = \frac{-(\mu+\lambda) + \sqrt{(\mu+\lambda)^2 + 4\lambda(1-\mu)}}{2(1-\mu)}.$$

Similarly, for quadratic cost $c(w) = 1 + w^2$, (4.1) becomes a cubic equation with a single root between 0 and 1.

2.3 Partial Information

Suppose the provider observes and tells the customer $N(t)$, the system occupancy at the moment of arrival. The customer computes $c_n = E[c(W)|N(t) = n]$ and elects to stay if $\theta \leq \theta_n$, where $\theta_n = 1/c_n$. So, given $N(t) = n$, the effective arrival process is Poisson with rate $\lambda_n = \lambda H(\theta_n)$. Thus, $N(t)$ is a birth-death process; the birth rate in state n is $\lambda_n = \lambda H(\theta_n)$, and the death rate is μ. Let N denote the equilibrium occupancy and $p_n = \Pr\{N = n\}$. Then, by the standard analysis of birth-death processes,

$$p_n = \left(\prod_{m=0}^{n-1} \lambda_m/\mu\right) p_0 = \Theta_n(\lambda/\mu)^n p_0,$$

where

$$\Theta_n = \prod_{m=0}^{n-1} H(\theta_m),\ n > 0.$$

Let

$$\Theta = \sum_{n>0} \Theta_n (\lambda/\mu)^n.$$

Then,

$$p_0 = \frac{1}{1+\Theta}.$$

Note that Θ is finite, because $\theta_n \to 0$. The customers' decisions ensure a stable system, even for cases like $c(w) = e^{\beta w}$, where the no-information system fails to reach equilibrium.

Example 2: Uniform customers with linear cost

$$\begin{aligned}
\Theta_n(\lambda/\mu)^n &= \left(\prod_{m=0}^{n-1} \frac{1}{1+m/\mu}\right)(\lambda/\mu)^n \\
&= \frac{\lambda^n}{\mu(\mu+1)\cdots(\mu+n-1)} \\
&= \frac{\Gamma(\mu)}{\Gamma(\mu+n)}\lambda^n
\end{aligned}$$

$$\Theta = \sum_{n=1}^{\infty} \frac{\Gamma(\mu)}{\Gamma(\mu+n)}\lambda^n = \lambda^{1-\mu} e^{\lambda} \gamma(\mu, \lambda)$$

$$\begin{aligned}
p_0 &= \frac{1}{1+\gamma(\mu,\lambda)\lambda^{1-\mu}e^{\lambda}} \\
p_n &= p_0 \frac{\Gamma(\mu)}{\Gamma(\mu+n)}\lambda^n, n > 0,
\end{aligned}$$

where $\Gamma(\mu) = \int_0^\infty t^{\mu-1}e^{-t}dt$ is the gamma function, and $\gamma(\mu,\lambda) = \int_0^\lambda t^{\mu-1}e^{-t}dt$ is the lower incomplete gamma function; see Abramowitz and Stegun (1965). We have obtained the distribution of N in closed form. This is a generalization of the Poisson distribution. The Poisson is the special case with $\mu = 1$. (Ward and Glynn 2003 derive a similar distribution for a different system, one with reneging.) The expected occupancy can be expressed in a simple form (see the supplement):

$$E\left[N\right] = \lambda - \left(\mu - 1\right)\left(1 - p_0\right). \tag{2.4}$$

2.4 Full Information

Now, suppose the provider observes and tells each arriving customer the exact waiting time. (For instance, each arriving customer brings the realization of his service time. This is nearly true in some production systems.) So W is a constant for the customer. The workload (or virtual waiting time) $V(t)$ is the total time needed to complete the service of all customers currently in the system. Then, the waiting time for an incoming customer at t equals $V(t)$. The effective arrival rate given workload v is $\lambda(v) = \lambda H(\theta_v)$, where $\theta_v = 1/c(v)$. The sample path of $V(t)$ works in the usual way: When $V(t) > 0$, it decreases at constant rate -1; when a customer joins the system, $V(t)$ increases by the service time of that customer.

Denote the density function of the equilibrium workload V by $f(v), v > 0$, and let p_0 be its mass at 0. By a level-crossing argument (Brill and Posner 1977), these quantities uniquely satisfy the integral equation

$$f(v) = \lambda p_0 e^{-\mu v} + \int_0^v \lambda H\left[1/c(w)\right] e^{-\mu(v-w)} f(w)dw \tag{2.5}$$

and the normalization condition

$$p_0 + \int_0^\infty f(v)dv = 1. \tag{2.6}$$

One can easily check that the solution is as follows: Define

$$C(v) = \int_0^v H[1/c(t)]dt.$$

Then,

$$f(v) = \lambda p_0 e^{\lambda C(v) - \mu v},$$

where

$$p_0 = \frac{1}{1 + \lambda \int_0^\infty e^{\lambda C(v) - \mu v} dv}.$$

We now have the solution in closed form, up to the evaluation of these integrals. (It is not hard to show that the integrals are finite. Again, the customers' decisions ensure stability.)

Example 3: Uniform customers with linear cost

$$C(v) = \int_0^v \frac{1}{1+t} dt = \ln(1+v).$$

Thus,

$$\begin{aligned} \int_0^\infty e^{\lambda C(v) - \mu v} dv &= \int_0^\infty (1+v)^\lambda e^{-\mu v} dv \\ &= e^\mu \mu^{-(\lambda+1)} \int_\mu^\infty y^\lambda e^{-y} dy \\ &= e^\mu \mu^{-(\lambda+1)} \Gamma(\lambda+1, \mu), \end{aligned}$$

where $\Gamma(\lambda + 1, \mu) = \int_\mu^\infty y^\lambda e^{-y} dy$ is the upper incomplete gamma function; see Abramowitz and Stegun (1965). So,

$$\begin{aligned} p_0 &= \frac{1}{1 + \lambda e^\mu \mu^{-(\lambda+1)} \Gamma(\lambda+1, \mu)}, \\ f(v) &= \lambda p_0 (1+v)^\lambda e^{-\mu v}, \ v > 0. \end{aligned}$$

This is a truncated gamma distribution.

2.5 Comparisons

This section compares the three information models. We find that, perhaps surprisingly, the primary driver of whether information is good or bad for the provider and the customers is the shape of H, the distribution of the customer cost-scale parameter, *not* the cost function c.

2.5.1 Cost-Scale Distributions

We first identify some important classes of distributions H.

Consider a *power distribution* $H(\theta) = \theta^\alpha$ for $\alpha > 0$. (The uniform distribution is the case with $\alpha = 1$.) Here,

$$J(\theta) = \frac{1}{\alpha+1}\theta^\alpha.$$

Thus, J is proportional to H. Consequently, *the average utility* $u = E[J(\theta_I)]$ *is proportional to the throughput* $\lambda E[H(\theta_I)]$. (Clearly, these are the only distributions with this property.)

This is a striking result. One simple performance measure, the busy probability or throughput, serves to characterize *both* the server's profit and the customers' average utility. The provider's and the customers' objectives are perfectly aligned. Put another way, one need not separately measure customer satisfaction. Just count the money. As we'll see shortly, more information is better for all parties in this case.

This is not always so, however. To see that *some* condition is needed, consider the case of identical customers with linear cost. For sufficiently small λ, under no information, all customers choose to stay, so the throughput is λ. Specifically, this happens when

$$\theta\left(1 + \frac{1}{\mu-\lambda} - \frac{1}{\mu}\right) \leq 1,$$

or

$$\lambda \le \mu \left(1 - \frac{1}{1 + \mu[(1/\theta) - 1]}\right).$$

Under partial information, however, there is a cutoff point $\bar{n}$, such that all customers stay when $N \le \bar{n}$, but they all balk when $N > \bar{n}$. Thus, the throughput is $< \lambda$. Here, information *reduces* throughput. If the queue is readily visible, the provider may try to hide it. (This point is due to Hassin 1986.) The same thing happens under full information.

Information can also hurt customers, due to externalities. Suppose there are two types of customers, A and B. Customers of each type are identical, but the types have different θ's, say $\theta_B \ll \theta_A$. An arriving customer is of type T with probability h_T. Suppose that, in the no-information system, all B customers stay, while all A customers leave. In the partial-information system, some of those A's stay, when they encounter an empty or near-empty system. They get only slightly positive utility, but they take it. So, the B's suffer lower utilities, and the loss may overwhelm the benefit to the A's. Here, the customers may prefer to hide the queue. To illustrate, consider the following case: $\theta_B = 1/8$, $\theta_A = 63/64$, $h_A = h_B = 1/2$, $\lambda = \mu = 1$. No information yields average utility 0.375, while partial information yields 0.358.

To understand when such behavior occurs, we distinguish two broad classes of distributions H. As we'll see, the following condition ensures that more information increases throughput.

Condition 2. *The function $H(1/x)$ is convex in $x \ge 1$.*

This means that the cost-scale distribution is spread out, so customers are heterogeneous, in a certain sense. It is equivalent to

$$-\frac{\theta h'(\theta)}{h(\theta)} \le 2$$

(assuming the derivative h' exists). The left-hand side is the elasticity of the density h. The condition posits that h not be too elastic, that is, that customers not be too concentrated. More precisely, it rules out a sharp *decrease* in h. It is a one-sided spread condition.

To see why this condition matters, notice that information has two main effects on customers' decisions. Compared to the no-information scenario, (1) a high-congestion signal (small θ_I) drives away some customers who would otherwise stay, while (2) a low-congestion signal (large θ_I) attracts some customers who would otherwise leave. The key question for the service provider is, which of these effects is greater? When h decreases sharply, it's possible that the first effect is stronger, so the net impact of more information is fewer customers.

For example, consider the beta density

$$h(\theta) = \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

with parameters α, $\beta > 0$, where $B(\alpha, \beta)$ is the beta function. This satisfies the condition, if and only if $\beta \leq 1$. The customers are too concentrated if $\beta > 1$.

Next, assume that H is strictly increasing, so it has a well-define inverse H^{-1}. The following condition guarantees that more information benefits customers.

Condition 3. *The function $J \circ H^{-1}$ is convex on* $[0, 1]$.

One can show that this condition is equivalent to

$$-\frac{\phi h'(\phi)}{h(\phi)} \geq 2 - \frac{\phi h(\phi)}{[H(\phi) - J(\phi)]}.$$

Again, we have a restriction on the elasticity of h, but here it's a *lower* bound, a variable one. It requires that h not *rise* too sharply. For a beta distribution, the condition holds precisely for $\beta \geq 1$.

Note that a power distribution is a beta distribution with $\beta = 1$. Among the beta distributions, these are the only ones that satisfy both conditions.

The following tables show that, even for smooth beta distributions, when these conditions are violated, more information can degrade performance. Table 2.1, 2.2 and 2.3 show the busy probabilities for systems with linear cost ($c(w) = 1 + w$), quadratic cost ($c(w) = 1 + w^2$) and square-root cost ($c(w) = 1 + \sqrt{w}$), respectively. Here, $\beta > 1$, so Condition 1 is violated. The **bold** numbers indicate where more information hurts the provider. Table 2.4, 2.5, 2.6 display average utilities in the same format. These cases have $\beta < 1$, so they violate Condition 2.

Table 2.1: Busy Probability with Linear Cost Function
λ=0.5, $\mu = 1.5$

	β=4			β=6			β=8		
α	no	partial	full	no	partial	full	no	partial	full
0.5	0.3329	**0.3306**	**0.3300**	0.3333	**0.3325**	**0.3319**	0.3333	**0.3330**	**0.3327**
1	0.3321	**0.3266**	**0.3257**	0.3333	**0.3309**	**0.3298**	0.3333	**0.3323**	**0.3315**
2	0.3284	**0.3168**	**0.3169**	0.3329	**0.3261**	**0.3243**	0.3333	**0.3299**	**0.3280**
4	0.3144	**0.2967**	**0.3021**	0.3302	**0.3129**	**0.3127**	0.3329	**0.3217**	**0.3192**
6	0.2958	**0.2803**	**0.2919**	0.3236	**0.2988**	**0.3029**	0.3315	**0.3113**	**0.3107**
8	0.2770	**0.2683**	0.2847	0.3134	**0.2858**	**0.2954**	0.3282	**0.3004**	**0.3034**

Table 2.2: Busy Probability with Quadratic Cost Function
λ=0.5, $\mu = 1.5$

	β=4			β=6			β=8		
α	no	partial	full	no	partial	full	no	partial	full
0.5	0.3306	**0.3264**	**0.3269**	0.3330	**0.3296**	**0.3292**	0.3333	**0.3311**	**0.3304**
1	0.3257	**0.3193**	**0.3217**	0.3320	**0.3254**	**0.3254**	0.3331	**0.3282**	**0.3275**
2	0.3116	**0.3062**	0.3139	0.3277	**0.3166**	**0.3191**	0.3321	**0.3219**	**0.3223**
4	0.2800	0.2847	0.3040	0.3108	**0.3000**	**0.3102**	0.3253	**0.3093**	**0.3143**
6	0.2529	0.2693	0.2980	0.2900	**0.2851**	0.3043	0.3122	**0.2973**	**0.3087**
8	0.2308	0.2596	0.2937	0.2704	0.2724	0.3000	0.2967	**0.2856**	0.3045

Observe that information hurts the provider only for quite large β, so the customers are quite concentrated, and for light traffic (small λ). Likewise, information

Table 2.3: Busy Probability with Square-root Cost Function
λ=0.5, $\mu = 1.5$

	β=4			β=6			β=8		
α	no	partial	full	no	partial	full	no	partial	full
0.5	0.3331	**0.3314**	**0.3310**	0.3333	**0.3329**	**0.3327**	0.3333	**0.3332**	**0.3331**
1	0.3324	**0.3280**	**0.3274**	0.3333	**0.3319**	**0.3314**	0.3333	**0.3329**	**0.3327**
2	0.3298	**0.3183**	**0.3178**	0.3331	**0.3281**	**0.3269**	0.3333	**0.3315**	**0.3307**
4	0.3184	**0.2961**	**0.2983**	0.3313	**0.3145**	**0.3135**	0.3331	**0.3245**	**0.3225**
6	0.3016	**0.2780**	**0.2836**	0.3266	**0.2982**	**0.2997**	0.3323	**0.3130**	**0.3115**
8	0.2833	**0.2658**	**0.2737**	0.3185	**0.2834**	**0.2882**	0.3302	**0.2999**	**0.3006**

Table 2.4: Average Utility with Linear Cost Function
α=2, μ=2

	β=0.15			β=0.1			β=0.05		
λ	no	partial	full	no	partial	full	no	partial	full
0.5	0.0570	0.0594	0.0612	0.0400	0.0405	0.0418	0.0215	**0.0207**	**0.0214**
1	0.0511	0.0523	0.0552	0.0365	0.0356	0.0377	0.0201	**0.0182**	**0.0193**
2	0.0436	**0.0431**	0.0472	0.0319	**0.0294**	0.0323	0.0182	**0.0151**	**0.0167**
4	0.0351	**0.0335**	0.0384	0.0265	**0.0230**	0.0266	0.0159	**0.0119**	**0.0139**
8	0.0262	**0.0251**	0.0301	0.0206	**0.0176**	0.0215	0.0131	**0.0092**	**0.0115**

Table 2.5: Average Utility with Quadratic Cost Function
α=2, μ=2

	β=0.15			β=0.1			β=0.05		
λ	no	partial	full	no	partial	full	no	partial	full
0.5	0.0565	0.0594	0.0635	0.0398	0.0405	0.0434	0.0214	**0.0207**	0.0222
1	0.0500	0.0523	0.0588	0.0360	**0.0356**	0.0402	0.0199	**0.0182**	0.0207
2	0.0417	0.0430	0.0520	0.0310	**0.0294**	0.0359	0.0179	**0.0151**	0.0186
4	0.0323	0.0332	0.0436	0.0250	**0.0229**	0.0308	0.0153	**0.0118**	0.0164
8	0.0231	0.0246	0.0344	0.0187	**0.0173**	0.0254	0.0123	**0.0092**	0.0142

Table 2.6: Average Utility with Square-root Cost Function
α=2, μ=2

	β=0.15			β=0.1			β=0.05		
λ	no	partial	full	no	partial	full	no	partial	full
0.5	0.0556	0.0588	0.0594	0.0392	0.0400	0.0405	0.0211	**0.0205**	**0.0208**
1	0.0496	0.0513	0.0523	0.0355	**0.0349**	0.0357	0.0196	**0.0178**	**0.0183**
2	0.0424	**0.0417**	0.0432	0.0310	**0.0284**	**0.0295**	0.0177	**0.0145**	**0.0151**
4	0.0345	**0.0318**	**0.0337**	0.0259	**0.0218**	**0.0231**	0.0154	**0.0112**	**0.0119**
8	0.0264	**0.0235**	**0.0255**	0.0205	**0.0163**	**0.0178**	0.0128	**0.0085**	**0.0093**

hurts the customers only for very small β and large λ. Even in such cases, the effects are small.

2.5.2 Comparison Results

Let us use superscripts *no*, *part* and $full$ to indicate the different information levels. We first compare the performance of the no-information and partial-information systems, then the no-information and full-information systems, and finally the partial-information and full-information systems.

No information and partial information:

Proposition 4. *If* $p_0^{part} \geq p_0^{no}$, *then* $u^{part} > u^{no}$.

Thus, more information helps *someone* – if not the provider, then the customers.

Proposition 5. *Under Condition 1 [*$H(1/x)$ *is convex],* $p_0^{part} \leq p_0^{no}$.

Thus, for certain shapes of the cost-scale distribution H, more information increases throughput and so helps the provider. For power H, therefore, information also increases customers' average utility.

To compare utilities more generally, we need a different condition on the shape of H.

Proposition 6. *Under Condition 2 [*$J \circ H^{-1}$ *is convex],* $u^{part} > u^{no}$. *Moreover, if* $p_0^{part} < p_0^{no}$, *then*

$$\frac{u^{part}}{u^{no}} \geq \frac{1 - p_0^{part}}{1 - p_0^{no}}.$$

For such customer distributions, then, average utility improves with information. The throughput may or may not increase. If it does increase, then utility increases even more, proportionally.

Finally, we mention an interesting fact about the special case considered in the examples above.

Proposition 7. *For uniform H and linear cost, the relation between $E\left[N^{no}\right]$ and $E\left[N^{part}\right]$ is the same as that between μ and 1. That is, they are equal for $\mu = 1$, $E\left[N^{no}\right] > E\left[N^{part}\right]$ for $\mu > 1$, and $E\left[N^{no}\right] < E\left[N^{part}\right]$ for $\mu < 1$.*

We have already seen that throughput and utility both increase with more information in this case. However, the standard performance measures such as $E[N]$ need not improve. But such measures are not the most relevant ones in this context. The customers here place different weights on delays. Information allows them to filter themselves, so that those who care more about delays wait less.

No information and full information: The systems with no and full information are related in exactly the same ways. For example, if $p_0^{full} \geq p_0^{no}$, then $u^{full} > u^{no}$.

Partial information and full information: It is harder to compare the partial- and full-information systems. We have only the following result:

Proposition 8. *Under Condition 1 [$H(1/x)$ is convex], $p_0^{full} \leq p_0^{part}$.*

Thus, again, with this shape of H, more information increases throughput. And, in case H is a power distribution, the full-information system has higher utility.

We *conjecture* that the other results above, comparing no and partial information, also describe the relation between partial and full information.

2.6 Extension

Now, suppose $c(0) < 1$. Some of the results above remain valid in this case, but others don't.

Some θ_i are now greater than 1. Let us extend the domain of H and J to all $\theta \geq 0$, setting $H(\theta) = 1$ for $\theta \geq 1$. The definition of J in terms of H remains the same. With this understanding, the throughput is still $\lambda E_I[H(\theta_I)]$, and the average utility is still $E_I[J(\theta_I)]$. The solutions for the three information models remain the same.

Turning to comparisons, let us focus on the relation between the no- and partial-information systems. The other comparisons are similar.

It is no longer true that, for a power distribution H, average utility is proportional to throughput. The incentives of the provider and the customers are *not* perfectly aligned.

Propositions 2 and 4 hold as stated. Thus, more information always helps someone, and under Condition 2, it helps customers.

Proposition 3 is no longer valid. The situation here is identical to the case of identical customers, discussed in §6.1 above. For *any* H, and sufficiently small λ, all customers stay in the no-information system. But, with partial information, some customers balk, so the throughput is lower.

We can obtain a qualified version of the result, however: Under Condition 1, for sufficiently large λ, $p_0^{part} \leq p_0^{no}$. In fact, this holds even under a weaker condition on H, namely, $H(1/x)$ is convex for sufficiently large x. (This covers all beta distributions, even those with $\beta > 1$.) Thus, under this condition, more information may hurt the provider with light traffic, but not with heavy traffic.

2.7 Summary

In this chapter, we considered service systems with three levels of customer-delay information. Customers use that information to determine their expected waiting costs, and so to decide whether to stay and receive service or leave (balk). We obtained closed-form solutions for some cases and nearly closed-form solutions for others. In comparing these systems, we found that the form of the cost-scale distribution plays a crucial role. For one important class, average utility is proportional to throughput; so the provider's and customers' objectives coincide; those measures improve as information increases. More broadly, we found sufficient conditions to ensure that more information helps the provider or the customers. In other cases, however, more information can actually hurt one or the other.

Chapter 3

Sensitivity Analysis of Cost Function and Weight Distribution

Delay, in most cases, is an unhappy experience for customers. Worse yet, to join a queue is usually risky, since customers don't know exactly how long they will have to wait. Nowadays there are increasing possibilities, enabled by new technologies, to provide useful delay information to customers, to enhance the values of the services they receive. Customers can thus make better-informed decisions upon arrival.

Chapter 2 describes numerous examples. It studies a single-server queue with three levels of delay information, none, partial (the system occupancy) and full (the exact waiting time). Each customer decides whether to stay or leave, based on this information and his own sensitivity to delays. The results there show that information's impact on the whole system is not always positive: It can reduce the server's throughput and even hurt customers' average utility. The most important factor determining these qualitative effects is the shape of the distribution of customers' delay-sensitivity weights.

In this chapter, we carry out sensitivity analysis of system performance with respect to customer characteristics, especially delay sensitivity and risk attitude. We also explore the relationship between the value of information and customer characteristics. In the latter part, we aim to answer questions such as: Is the value of information greater for less patient customers? Is it larger for more risk-averse customers?

There has been much research on comparison of systems with different input streams. Ross (1978) conjectures that a more regular arrival process leads to better performance. Some counter examples are provided by Heyman (1982). See Rolski (1990), Shaked and Shanthikumar (1994) and Müller and Stoyan (2002) for surveys. Chao and Dai (1995) and Dai and Chao (1996) show that the conjecture holds for a single-server loss system in random environment. Recently, Whitt (2006) analyzes the sensitivity of performance to changes in model parameters in an $M/M/s$ queue with customer abandonment. He shows that performance can be quite sensitive to changes in the arrival and service rates, but relatively insensitive to the abandonment rate.

In our context, the arrival process is formed by a stream of customers who make their own individual decisions, based on their utilities and the available information about the system's state. The overall impact of customers' characteristics is thus far from clear. Understanding this matter is important for the design of systems in different markets with different types of customers.

Intuitively, information should be more valuable to a more risk-averse decision maker. This is true in a static sense, but it may not reflect dynamic behavior. See Hilton (1981), Freixas and Kihlstrom (1984), Willinger (1989) and Nadiminti, et al.(1996). For example, Freixas and Kihlström show that the demand for information may decrease with the level of risk aversion. *Ex post*, information reduces risk, however, *ex ante*, information gathering itself is a risky activity that risk-averse decision makers are less willing to bear. These works consider only a single decision maker facing an exogenous risk. In our system, there is a group of decision makers (customers). Each one's joining decision affects the delays for later-arriving customers, so the system exhibits *negative externality*. The delay risk here is endogenous. The relationship between information and risk aversion is thus even more complex.

The remainder of the chapter is organized as follows: Section 1 reviews the basic formulation and reviews stochastic orders. Sections 2 briefly summarizes the three models in Chapter 2 and provides general stochastic comparison results. Section 3-4 develop the sensitivity analysis of system performance with respect to the change of weight distribution and the change of cost function. Section 5 studies the relationship between the value of information and customers' characteristics. Section 6 concludes.

3.1 Formulation and Preliminaries

3.1.1 Notation and Utility

As in Chapter 2, we assume a single-server queue with exponential service times. Potential customers arrive in a Poisson process. We suppose that a customer's utility equals a reward for receiving service minus a waiting cost. This waiting cost depends on a customer-specific weight and the expectation of a (common) function of the waiting time. Denote

- λ = arrival rate of potential customers
- W = waiting time in queue
- θ = customer weight for delay, $\theta \in [0, 1]$
- H = cumulative distribution function of θ, assumed continuous on $[0, 1]$, with density h
- $c(w)$ = basic cost to wait time w, a positive, increasing, unbounded, continuous function
- r = reward to the customer for receiving service, $r > 0$
- $u(w, \theta)$ = utility for a customer with weight θ to wait time w
- $u(\theta|I)$ = expected utility for a customer with θ, given delay information I

- u = average utility for a whole group of customers

The service reward is the same for all customers (this assumption is convenient but not essential). Customers differ in the importance of time. This difference is expressed by the customer weight θ. Each customer's weight is independent of all other events; it follows the common distribution function $H(\theta)$. The shape of $H(\cdot)$ characterizes the heterogeneity of customers in delay sensitivity. A very concentrated distribution indicates nearly identical customers and a dispersive distribution means that customers are different: Some are patient while others are impatient.

A customer with delay-sensitivity θ has a utility function for receiving service but waiting time w, which is expressed as $u(w, \theta) = r - \theta c(w)$. The waiting time w can be a random variable. The customer assesses the distribution of this random variable, based on the available information. Let I denote the information variable. The expected waiting cost $c_I = E[c(W)|I]$ is a function of that information. The expected utility for the customer to remain in the system is thus $u(\theta|I) = r - \theta c_I$. The customer remains in the system if $u(\theta|I)$ is non-negative and otherwise balks. We assume that there is no reneging.

We normalize $r = 1$ and assume $c(0) = 1$. Under this assumption, a customer seeing an empty system will always join. Define $\theta_I = 1/c_I$. The effective arrival rate given delay I is hence $\lambda H(\theta_I)$. Define the function

$$J(\theta) = \frac{\int_0^\theta H(\phi) d\phi}{\theta}.$$

Chapter 2 shows that the average utility for all customers, u, equals $E_I[J(\theta_I)]$. Also, $J(1/x)$ is decreasing and convex in x.

3.1.2 Stochastic Orders

Consider two systems, identical except for H and c. Use the superscript $k = 1, 2$ to index the systems. We consider different stochastic orders between $\theta^k, k = 1, 2$. Most of these concepts can be found in Shaked and Shanthikumar (1994) and Müller and Stoyan (2002). For discussion of single crossing, see Athey (2002). We shall also consider analogous relations between the cost functions.

If $H^1(x) \geq H^2(x)$, $x \in [0, 1]$, θ^1 is *stochastically smaller* than θ^2 (denoted $\theta^1 \preceq_{st} \theta^2$), or H^2 is said to dominate H^1 according to *first-order stochastic dominance.* A stronger condition is that the ratio h^1/h^2 is monotonically decreasing, a condition called *monotone likelihood ratio.* Then θ^1 is said to to be smaller than θ^2 in the *likelihood-ratio order* (denoted $\theta^1 \preceq_{lr} \theta^2$).

If $\int_v^1 \bar{H}^1(x)dx \leq \int_v^1 \bar{H}^2(x)dx$, for all $v \in [0, 1]$, θ^1 is smaller than θ^2 in the *increasing convex order* (denoted $\theta^1 \preceq_{icx} \theta^2$). If $E[\theta^1] = E[\theta^2]$ and $\int_v^1 \bar{H}^1(x)dx \leq \int_v^1 \bar{H}^2(x)dx$, for all $v \in [0, 1]$, θ^1 is smaller than θ^2 in the *convex order* (denoted $\theta^1 \preceq_{cx} \theta^2$), or H^1 is said to dominate H^2 according to *second-order stochastic dominance.* This condition implies that $var[\theta^1] \leq var[\theta^2]$. Intuitively, it means that system 1's customers are less heterogeneous than system 2's. If the ratio H^1/H^2 is decreasing, a condition called *monotone probability ratio*, θ^1 is said to be smaller than θ^2 in the *reverse hazard-rate order* (denoted $\theta^1 \preceq_{rh} \theta^2$).

The theory of total positivity (see, Karlin 1968) shows that monotone ratios are preserved under integration. Thus, decreasing h^1/h^2 implies decreasing H^1/H^2. Decreasing H^1/H^2, in turn, implies decreasing J^1/J^2.

Let X^1 and X^2 be two random variables with pdfs f^1 and f^2, respectively. If the ratio f^1/f^2 is unimodal, X^1 is said to be *uniformly less variable* than X^2 (denoted $X^1 \preceq_{uv} X^2$). If the ratio f^1/f^2 is log-concave, X^1 is said to be *log-concave relative* to X^2 (denoted $X^1 \preceq_{lc} X^2$). This is a stronger condition than $X^1 \preceq_{uv} X^2$.

3.2 General Comparison Results

Chapter 2 considers three levels of delay information. With *no information*, customers still estimate their waiting times, but these estimates are based only on long-term (equilibrium) experience, not real-time information. The occupancy provides *partial information*; the remaining uncertainty comprises the actual service times of the waiting customers. The exact waiting time gives the customer *full information.* We briefly summarize the models and solutions and then give our general stochastic comparison results.

3.2.1 No Information

Under no information, the resulting system is an $M/M/1$ queue with the equilibrium arrival rate λ_-, which solves

$$\lambda_- = \lambda H\left(\frac{1}{E[c(W|-)]}\right). \tag{3.1}$$

Here, $E[c(W|-)]$ indicates the expected cost given λ_-. Hence the equilibrium system is still an $M/M/1$ queue with effective arrival rate λ_-.

Proposition 9. *Under no information, if $\lambda_-^1 \geq \lambda_-^2$, then $N^1 \succeq_{lr} N^2$.*

Proof. Whitt (1999) (Theorem 4.1) shows that, for any pair of birth-death processes, $N^1 \succeq_{lr} N^2$, provided

$$\frac{\lambda_n^1}{\mu_{n+1}^1} \geq \frac{\lambda_n^2}{\mu_{n+1}^2}, n \geq 0.$$

Here, each system k is an $M/M/1$ queue with arrival rate λ_-^k and service rate μ. Hence the above condition is satisfied. □

3.2.2 Partial Information

Under partial information, N, the system occupancy at the moment of arrival, can be modeled as a birth-death process. The birth rate in state n is $\lambda_n = \lambda H(\theta_n)$, where $\theta_n = 1/c_n = 1/E[c(W)|N = n]$, and the death rate is μ. The equilibrium distribution of N can be expressed as

$$p_n = \left(\prod_{m=0}^{n-1} \lambda_m/\mu\right) p_0 = \Theta_n(\lambda/\mu)^n p_0,$$

where

$$\Theta_n = \prod_{m=0}^{n-1} H(\theta_m), \ n > 0.$$

Let

$$\Theta = \sum_{n>0} \Theta_n(\lambda/\mu)^n.$$

Then,

$$p_0 = \frac{1}{1+\Theta}.$$

Define the cumulative effective arrival rate $\Lambda_n = \lambda \prod_{m=0}^{n} H(1/c_m)$.

Proposition 10. *Under partial information,*

1. *if $\Lambda_n^1 \geq \Lambda_n^2$, then $p_0^1 \leq p_0^2$;*

2. *if $\lambda_n^1 \geq \lambda_n^2$ for all $n = 0, 1, 2, ...$, then $N^1 \succeq_{lr} N^2$;*

3. *if λ_n^1 crosses λ_n^2 once from above (i.e., there exists $\hat{n} > 0$ such that $\lambda_n^1 \geq \lambda_n^2, \ n \leq \hat{n}$, and the inequality is reversed for $n > \hat{n}$), then $N^1 \preceq_{uv} N^2$;*

4. *if the ratio λ_n^1/λ_n^2 is decreasing in n, then $N^1 \preceq_{lc} N^2$.*

Proof. For part 1, if $\Lambda_n^1 \geq \Lambda_n^2$ for all n, then $\Theta_n^1 \geq \Theta_n^2$ for all n, and hence $p_0^1 \leq p_0^2$.

Part 2 follows by the same argument as in Proposition 9.

For part 3, if λ_n^1 crosses λ_n^2 once from above, consider the ratio

$$\frac{p_{n+1}^1/p_{n+1}^2}{p_n^1/p_n^2} = \frac{\lambda_n^1}{\lambda_n^2}.$$

For $n \leq \hat{n}$, this fraction is greater than 1, thus p_n^1/p_n^2 is increasing; for $n > \hat{n}$, similarly, p_n^1/p_n^2 is decreasing. Thus p_n^1/p_n^2 is unimodal. Hence, $N^1 \preceq_{uv} N^2$.

Finally, for part 4, if λ_n^1/λ_n^2 is decreasing in n,

$$\frac{p_{n+1}^1/p_{n+1}^2}{p_n^1/p_n^2} = \frac{\lambda_n^1}{\lambda_n^2}.$$

is decreasing in n. So p_n^1/p_n^2 is log-concave, and $N^1 \preceq_{lc} N^2$. □

3.2.3 Full Information

Define the cumulative effective arrival rate $\Lambda(v) = \int_0^v \lambda(v)dv$. Under full information, the pdf for the equilibrium workload V, $f(v), v > 0$, solves the integral equation

$$f(v) = \lambda p_0 e^{-\mu v} + \int_0^v \lambda H\left[1/c(w)\right] e^{-\mu(v-w)} f(w)dw \tag{3.2}$$

with the normalization condition

$$p_0 + \int_0^\infty f(v)dv = 1. \tag{3.3}$$

The solution is

$$f(v) = \lambda p_0 e^{\Lambda(v) - \mu v}, \tag{3.4}$$

where

$$p_0 = \frac{1}{1 + \lambda \int_0^\infty e^{\Lambda(v)-\mu v} dv}. \tag{3.5}$$

Proposition 11. *Under full information,*

1. *if* $\Lambda^1(v) \geq \Lambda^2(v)$*, then* $p_0^1 \leq p_0^2$*;*

2. *if* $\lambda^1(v) \geq \lambda^2(v)$ *for all* $v \geq 0$*, then* $V^1 \succeq_{lr} V^2$*;*

3. *if* $\lambda^1(v)$ *crosses* $\lambda^2(v)$ *once from above, then* $V^1 \preceq_{uv} V^2$*;*

4. *if the ratio* $\lambda^1(v)/\lambda^2(v)$ *is decreasing in* v*, then*

$$\frac{\ln(f^1(v))'}{\ln(f^2(v))'}$$

is decreasing in v*.*

Proof. Part 1 follows immediately from (3.5).

For part 2, if $\lambda^1(v) \geq \lambda^2(v)$ for all $v \geq 0$, then $\Lambda^1(v) - \Lambda^2(v)$ is positive and increasing. So, $p_0^1 \leq p_0^2$, and

$$\frac{f^1(v)}{f^2(v)} = \frac{p_0^1}{p_0^2} \exp\left\{\left[\Lambda^1(v) - \Lambda^2(v)\right]\right\}$$

is increasing. Hence, $V^1 \succeq_{lr} V^2$.

For part 3, if $\lambda^1(v)$ crosses $\lambda^2(v)$ once from above, $\Lambda^1(v) - \Lambda^2(v)$ is positive and increasing for $v \leq \hat{v}$, but decreasing for $v \geq \hat{v}$. That is, $\Lambda^1(v) - \Lambda^2(v)$ is unimodal, and therefore so is $f^1(v)/f^2(v)$.

Finally, for part 4, if the ratio $\lambda^1(v)/\lambda^2(v)$ is decreasing in v, then

$$\frac{\ln(f^1(v))'}{\ln(f^2(v))'} = \frac{\Lambda^{1'}(v)}{\Lambda^{2'}(v)} = \frac{\lambda^1(v)}{\lambda^2(v)}$$

is decreasing in v. □

3.3 Specific Conditions on Weight Distributions

In this section, we assume $c^1 = c^2 = c$ and consider specific conditions on the H^k.

3.3.1 First-Order Stochastic Dominance

Assume $\theta^1 \preceq_{st} \theta^2$. This means that system 1's customers are stochastically more patient than system 2's. We have the following conclusion about the system occupancy and the workload.

Proposition 12. *If $\theta^1 \preceq_{st} \theta^2$, then $N^1 \succeq_{lr} N^2$ under no or partial information, and $V^1 \succeq_{lr} V^2$ under full information.*

Proof. The condition $\theta^1 \preceq_{st} \theta^2$ means $H^1(\theta) \geq H^2(\theta)$, for all θ in $[0, 1]$. Hence, $H^1(1/c(v)) \geq H^2(1/c(v))$ for all $v \geq 0$ and $H^1(1/c_n) \geq H^2(1/c_n)$ for all $n \geq 0$. So, $\lambda^1(v) \geq \lambda^2(v)$, and $\lambda^1_n \geq \lambda^2_n$. Also, $\lambda^1_- \geq \lambda^2_-$ from (3.1). From Propositions 9, 10 and 11, the conclusion follows. □

Next, we compare the average utilities for the two systems. By the definition of J^k, we have $J^1 \geq J^2$.

Proposition 13. *If $\theta^1 \preceq_{st} \theta^2$, then $u^1 \geq u^2$ under no information.*

Proof. By Proposition 12, $\theta^1_- \geq \theta^2_-$. Thus, $u^1 = J^1(\theta^1_-) \geq J^1(\theta^2_-) \geq J^2(\theta^2_-) = u^2$. □

For partial and full information, first consider the special case of the power distribution, $H(\theta) = \theta^\alpha$ for constant $\alpha > 0$. Note that, for two distributions of this form, $\alpha^1 \leq \alpha^2$ implies $H^1 \geq H^2$.

Proposition 14. *If each H^k is a power distributions with $\alpha^1 \leq \alpha^2$, then $u^1 \geq u^2$ under partial or full information.*

Proof. In this case, suppressing k for the moment,

$$J(\theta) = \frac{1}{\alpha+1}\theta^{\alpha} = \frac{1}{\alpha+1}H(\theta).$$

so the average utility $u = E[J(\theta_I)] = \frac{1}{\alpha+1}E[H(\theta_I)] = \frac{1}{\alpha+1}(1-p_0)\mu/\lambda$. Under either partial or full information, $1 - p_0^1 \geq 1 - p_0^2$, by Proposition 12. Thus, $u^1 \geq u^2$. □

Beyond this special case, the relation between u^1 and u^2 is not clear. A stronger condition is needed to conclude that customers on average are better off in one system than the other.

3.3.2 Monotone Likelihood Ratio

Condition $\theta^1 \preceq_{lr} \theta^2$ means that $h^1(x)/h^2(x)$ is decreasing in x. This condition is stronger than $\theta^1 \preceq_{st} \theta^2$. We now consider the relation between u^1 and u^2 under partial and full information.

Proposition 15. *If $\theta^1 \preceq_{lr} \theta^2$, then*

$$\frac{u^1}{u^2} \geq \frac{J^1(1)p_0^1}{J^2(1)p_0^2},$$

under partial or full information.

Proof. We demonstrate the result for partial information. The proof for full information is similar. Recall that decreasing $h^1(x)/h^2(x)$ implies decreasing $H^1(x)/H^2(x)$, which, in turn, implies decreasing $J^1(x)/J^2(x)$. Hence $J^1(\theta_n)/J^2(\theta_n)$ is increasing in n. From proposition 12, we know that $N^1 \succeq_{lr} N^2$ under partial information, that is, p_n^1/p_n^2 is increasing in n. Hence,

$$\begin{aligned} \frac{u^1}{u^2} &= \frac{\sum_{n\geq 0} J^1(\theta_n)p_n^1}{\sum_{n\geq 0} J^2(\theta_n)p_n^2} \\ &\geq \frac{J^1(\theta_0)p_0^1}{J^2(\theta_0)p_0^2} = \frac{J^1(1)p_0^1}{J^2(1)p_0^2}. \end{aligned}$$

□

Since $p_0^1 \leq p_0^2$ and $J^1(1) \geq J^2(1)$, it is not necessarily true that $[J^1(1)p_0^1]/[J^2(1)p_0^2] \geq 1$. But at least we obtain a lower bound on u^1/u^2.

3.3.3 Stochastic Monotonicity and Convexity

In this subsection, we discuss the relationship between the conditions $\theta^1 \preceq_{icx} \theta^2$ and $\Lambda^1(v) \geq \Lambda^2(v)$ and $\Lambda_n^1 \geq \Lambda_n^2$. First, the condition $\theta^1 \preceq_{icx} \theta^2$ is not a sufficient condition for $\Lambda^1(v) \geq \Lambda^2(v)$. The condition $\theta^1 \preceq_{icx} \theta^2$ means that $\int_x^1 H^1(y)dy \geq \int_x^1 H^2(y)dy$, $x \in [0,1]$. Let $y = 1/c(t)$, $dy/dt = -c'(t)/c^2(t)$. Then this condition becomes

$$\int_v^0 H^1(1/c(t))(-c'(t)/c^2(t))dt \geq \int_v^0 H^2(1/c(t))(-c'(t)/c^2(t))dt.$$

This condition is very different from

$$\int_0^v H^1(1/c(t))dt \geq \int_0^v H^2(1/c(t))dt.$$

Hence, the condition $\theta^1 \preceq_{icx} \theta^2$ doesn't imply $\Lambda^1(v) \geq \Lambda^2(v)$. Similarly, $\theta^1 \preceq_{icx} \theta^2$ is not a sufficient condition for $\Lambda_n^1 \geq \Lambda_n^2$. The former depends on the integrals of the H^k on the whole interval $[0, v]$, while the latter depends on the products of the discrete values of the H^k on $\{\theta_0, \theta_1, \theta_2, ...\}$.

Hence, more concentrated customers need not imply a larger cumulative effective arrival rate.

3.3.4 Single Crossing

Consider the condition that H^1 crosses H^2 once from above in $(0, 1)$. Denote the crossing point by $\hat{\theta}$. Then $H^1(1/c(v))$ crosses $H^2(1/c(v))$ once from below. Hence,

$\lambda^1(v)$ crosses $\lambda^2(v)$ once from below. The crossing point is $\hat{v} = c^{-1}(1/\hat{\theta})$. From Propositions 10 and 11, we derive the following conclusion.

Proposition 16. *If H^1 crosses H^2 once from above, then $N^1 \preceq_{uv} N^2$ under partial information, and $V^1 \preceq_{uv} V^2$ under full information.*

3.3.5 Reverse Hazard-Rate Ordering

The condition $\theta^1 \succeq_{rh} \theta^2$ means that $H^1(\theta)/H^2(\theta)$ is increasing. Hence

$$H^1(1/c(x))/H^2(1/c(x))$$

is decreasing in x. Thus $\lambda^1(v)/\lambda^2(v)$ is decreasing in v. By Propositions 10 and 11, we have the following proposition.

Proposition 17. *If $\theta^1 \succeq_{rh} \theta^2$, then $N^1 \preceq_{lc} N^2$ under partial information and*

$$\frac{\ln(f^1(v))'}{\ln(f^2(v))'}$$

is decreasing in v under full information.

3.4 Specific Conditions on Cost Functions

In this section, we fix $H^1 = H^2 = H$ but consider different conditions on the c^k.

3.4.1 Inequality

Condition $c^1 \leq c^2$ means that customers in system 1 care less about waiting than those in system 2. We have the following conclusion about the system occupancy and the workload.

Proposition 18. *If $c^1 \leq c^2$, then $N^1 \succeq_{lr} N^2$ under no or partial information, and $V^1 \succeq_{lr} V^2$ under full information.*

Proof. From the condition $c^1 \le c^2$, we get $H(1/c^1(v)) \ge H(1/c^2(v))$ for all $v \ge 0$ and $H(1/c^1_n) \ge H(1/c^2_n)$ for all $n = 0, 1, 2....$ Hence, $\lambda^1(v) \ge \lambda^2(v)$ and $\lambda^1_n \ge \lambda^2_n$. Also we can derive that $\lambda^1_- \ge \lambda^2_-$ from (3.1). From Proposition 9, 10 and 11, we obtain the conclusion. □

About average utilities, we have the following conclusions.

Proposition 19. *If* $c^1 \le c^2$*, then* $u^1 \ge u^2$ *under no information.*

Proof. By Proposition 18, $\lambda^1_- \ge \lambda^2_-$, so $\theta^1_- \ge \theta^2_-$. Also, $J^k = J$. Thus, since J is increasing, $u^1 = J(\theta^1_-) \ge J(\theta^2_-) = u^2$. □

This last conclusion need not hold for partial and full information, however. On one hand, since $c^1 \le c^2$, $\theta^1_i \ge \theta^2_i$ and thus $J(\theta^1_i) \ge J(\theta^2_i)$ since $J(\theta)$ is increasing in θ. On the other hand, since $J(\theta_i)$ is decreasing in i, $I^1 \succeq_{st} I^2$ implies $E[J(\theta^k_{I^1})] \le E[J(\theta^k_{I^2})]$ for $k = 1, 2$. Thus, it is unclear whether $E[J(\theta^1_{I^1})]$ or $E[J(\theta^2_{I^2})]$ is larger.

Intuitively, *given* information I, system 2 has larger expected utility, but its larger arrival rate pushes the system to a more congested state, which decreases the utilities. The overall effect is unclear. Tables 3.1 and 3.2 show that either effect can dominate the other. All the cases considered have linear costs: $c^k(w) = 1 + \gamma^k w$. In Table 3.1 the average utility decreases with c^k, while in Table 3.2 it increases for partial and full information.

Table 3.1: Compare Linear Cost Functions with Beta H
$\alpha = \beta = 2$, $\lambda = 2$, $\mu = 2$

	busy probability			average utility		
γ	no	partial	full	no	partial	full
0.1	0.8601	0.8922	0.8959	0.3613	0.3956	0.4001
0.5	0.6972	0.7668	0.7848	0.2750	0.3324	0.3460
1.0	0.6026	0.6954	0.7259	0.2315	0.3038	0.3222

Table 3.2: Compare Linear Cost Functions with Beta H
$\alpha = \beta = 2$, $\lambda = 8$, $\mu = 2$

	busy probability			average utility		
γ	no	partial	full	no	partial	full
0.1	0.9768	1.0000	1.0000	0.0870	0.0893	0.0894
0.5	0.9002	0.9998	1.0000	0.0798	0.0899	0.0898
1.0	0.8284	0.9940	0.9996	0.0732	0.0908	0.0907

In a system with no balking, where all customers stay, the distribution of W is just that of the standard $M/M/1$ system with arrival rate λ. The average waiting cost for system k is $E[\theta^k]E[c^k(W)]$, which is larger for system 1. Thus, the customers in system 1 get lower average utility. In the system allowing balking but with no information, this conclusion is still true, by Propositions 13 and 19. However, in the system with balking and either partial or full information, the average utility in system 1 can be larger. Here, customers make their own decisions to maximize their expected utilities, and this leads to less congestion in system 1. Less congestion in turn increases the average utility for those served customers.

3.4.2 Single Crossing

Consider the condition that $c^1(v)$ crosses $c^2(v)$ once from below. Then $H(1/c^1(v))$ crosses $H(1/c^2(v))$ once from above. Hence, $\lambda^1(v)$ crosses $\lambda^2(v)$ once from above. We get the following conclusion.

Proposition 20. *If $c^1(v)$ crosses $c^2(v)$ once from below, then $N^1 \preceq_{uv} N^2$ under partial information, and $V^1 \preceq_{uv} V^2$ under full information.*

3.4.3 Monotone Ratio

Consider the condition that the ratio $c^1(x)/c^2(x)$ is increasing in x. Hence,

$$d \ln \left[c^1(x)/c^2(x)\right]/dx \geq 0$$

or

$$\frac{c^{1'}(x)}{c^1(x)} \geq \frac{c^{2'}(x)}{c^2(x)}.$$

Since $c^1(0) = c^2(0) = 1$, $c^1(x) \geq c^2(x)$ for all $x \geq 0$.

Consider the elasticity of H at x, $xh(x)/H(x)$. Assume H has a decreasing elasticity. Then,

$$\frac{h(1/c^1(x))[1/c^1(x)]}{H(1/c^1(x))} \geq \frac{h(1/c^2(x))[1/c^2(x)]}{H(1/c^2(x))}.$$

Thus

$$\frac{d}{dx}\ln\left(\frac{H(1/c^1(x))}{H(1/c^2(x))}\right) = -\frac{h(1/c^1(x))[1/c^1(x)]}{H(1/c^1(x))}\frac{c^{1'}(x)}{c^1(x)} + \frac{h(1/c^2(x))[1/c^2(x)]}{H(1/c^2(x))}\frac{c^{2'}(x)}{c^2(x)} \leq 0.$$

Hence, $H(1/c^1(x))/H(1/c^2(x))$ is decreasing in x and so $\lambda^1(v)/\lambda^2(v)$ is decreasing in v. We have the following conclusion.

Proposition 21. *If $c^1(x)/c^2(x)$ is increasing in x and H has a decreasing elasticity, $N^1 \preceq_{lc} N^2$ under partial information and*

$$\frac{\ln(f^1(v))'}{\ln(f^2(v))'}$$

is decreasing in v under full information.

3.4.4 Risk Aversion

In this subsection, we consider a different condition on c^k which indicates the degree of customers' risk aversion. For a utility function of one variable $u(x)$ which is increasing in x, the Arrow-Pratt measure of absolute risk aversion is $A(x) = -u''(x)/u'(x)$. Here, customers have an increasing *disutility* function $c(\cdot)$, hence we use $A(w) = c^{''}(w)/c^{'}(w)$ to measure risk aversion. Let $A^k(w)$ denote the risk aversion of the cost function c^k. We suppose customers in system 1 are more risk averse than those of system 2. As we shall see, this condition is related to others we have seen above.

Proposition 22. *If $A^1(w) \geq A^2(w)$ for all w and $c^{1'}(0) \geq c^{2'}(0)$, then $c^1(w) \geq c^2(w)$; if $A^1(w) \geq A^2(w)$ for all w and $c^{1'}(0) < c^{2'}(0)$, then $c^2(w)$ crosses $c^1(w)$ once from above.*

Proof. According to Pratt (1964), the condition $A^1(w) \geq A^2(w)$ is equivalent to the condition that function $\tau = c^2 \circ (c^1)^{-1}$ be increasing and concave; that is, c^2 is an increasing and concave transform of c^1. In out context, $c^2(0) = c^1(0) = 1$. Hence, $\tau(1) = 1$. Consider the derivative of τ,

$$\frac{d\tau}{dx} = \frac{c^{2'} \circ (c^1)^{-1}(x)}{c^{1'} \circ (c^1)^{-1}(x)}.$$

Condition $c^{2'}(0) \leq c^{1'}(0)$ is equivalent to $c^{2'} \circ (c^1)^{-1}(1) \leq c^{1'} \circ (c^1)^{-1}(1)$ or $\frac{d\tau}{dx}(1) \leq 1$. Since $\tau(1) = 1$ and τ is increasing and concave, it follows that $\tau(x) \leq x$. Insert $x = c^1(w)$, and we get $c^2(w) \leq c^1(w)$, $w \geq 0$.

Similarly, the condition $c^{2'}(0) > c^{1'}(0)$ is equivalent to $c^{2'} \circ (c^1)^{-1}(1) > c^{1'} \circ (c^1)^{-1}(1)$ or $\frac{d\tau}{dx}(1) > 1$. Since $\tau(1) = 1$ and τ is increasing and concave, it follows that the graph of $\tau(x)$ starts from $(1, 1)$ and crosses the diagonal line $y = x$ once from above. Insert $x = c^1(w)$, and we get $c^2(w)$ crosses $c^1(w)$ once from above. □

Thus, greater risk aversion implies either no crossing or single crossing of the two cost functions, depending on their derivatives at 0. By Proposition 20, single crossing of the two cost functions implies nothing about the relation between the throughputs. Hence, a system with more risk-averse customers need not have a smaller throughput.

3.5 The Value of Information and Risk Aversion

In this section, we discuss the relationship between the value of information and customer characteristics, namely, the weight on delay and the degree of risk aversion. We first give the expression of the value of information, then give two general conclusions and at last we give numerical computation results.

A simple and direct measure of the value of information is the difference of the average utilities under more and less information. For a customer with weight θ, his expected utility under no, partial and full information is denoted as $u^{no}(\theta)$, $u^{part}(\theta)$ and $u^{full}(\theta)$, respectively. Define $VI^{fn}(\theta) = u^{full}(\theta) - u^{no}(\theta)$. Then $VI^{fn}(\theta)$ measures the value of full over no information for a customer with weight θ. Similarly, define $VI^{pn}(\theta) = u^{part}(\theta) - u^{no}(\theta)$ and $VI^{fp}(\theta) = u^{full}(\theta) - u^{part}(\theta)$.

Under no information, if $\theta > \theta_-$, the customer will balk and gets utility 0. Hence

$$u^{no}(\theta) = 0.$$

If $\theta \leq \theta_-$, the customer will join and obtain a nonnegative expected utility $1 - \theta E[c(W^{no})]$ and hence

$$u^{no}(\theta) = 1 - \theta E[c(W^{no})].$$

Under partial information, the customer will join if $1 - \theta c_n \geq 0$ or $n \leq c^{-1}(1/\theta)$. Denote n^* to be n satisfying $c_{n^*} = 1/\theta$ (n^* may not be an integer). The expected utility for the customer u^{part} is

$$u^{part}(\theta) = \sum_{n \leq n^*} (1 - \theta c_n) p_n^{part}.$$

Now define a new variable $\tilde{N}^{part}$ such that $\tilde{p}_n^{part} = p_n^{part}$ for all $n < n^*$ and $\tilde{p}_{n^*}^{part} = 1 - \sum_{n<n^*} p_n$. Then we can write

$$u^{part}(\theta) = 1 - \theta E[c_{\tilde{N}^{part}}].$$

Under full information, if $v \leq c^{-1}(1/\theta)$, customer will join and obtain the utility $1 - \theta c(v)$; otherwise , he will leave with utility 0. Hence, the customer's expected utility is

$$u^{full}(\theta) = \int_0^{c^{-1}(1/\theta)} (1 - \theta c(v)) dF^{full}(v).$$

Define a new distribution $\tilde{F}^{full}$ such that $\tilde{F}^{full}(v) = F^{full}(v)$, for $v < c^{-1}(1/\theta)$ and $\tilde{F}^{full}(v) = 1$, for $v \geq c^{-1}(1/\theta)$. Denote a random variable with this distribution by $\tilde{V}^{full}$. One can easily verify that

$$\tilde{V}^{full} \preceq_{st} V^{full}. \tag{3.6}$$

Then we can write

$$u^{full}(\theta) = 1 - \theta E[c(\tilde{V}^{full})]$$

We have two propositions about the value of information.

Proposition 23. *If $p_0^{full} \geq p_0^{no}$, then for all θ, $VI^{fn}(\theta) \geq 0$. Similarly, if $p_0^{part} \geq p_0^{no}$, then for all θ, $VI^{pn}(\theta) \geq 0$.*

Proof. We prove the result for full versus no information. The proof for partial versus no information is similar.

Chapter 2 shows that $V^{full} \preceq_{st} W^{no}$ if $p_0^{full} \geq p_0^{no}$. From (3.6), we derive that $\tilde{V}^{full} \preceq_{st} W^{no}$. Hence $E[c(W^{no})] \geq E[c(\tilde{V}^{full})]$, and thus $u^{no}(\theta) \leq u^{full}(\theta)$ for $\theta \leq \theta_-$. For $\theta > \theta_-$, the customer has utility 0 under no information and nonnegative expected utility under full information, hence, $u^{no}(\theta) \leq u^{full}(\theta)$. □

That is, information benefits every individual customer, if it hurts the server.

Proposition 24. *When $\theta > \theta_-$, $VI^{fn}(\theta)$ is decreasing in θ. That is, the value of information is smaller for less patient customers. When $\theta \leq \theta_-$, if $VI^{fn}(\theta) > 0$, then $VI^{fn}(\theta)$ is increasing with θ. Otherwise, if $VI^{fn}(\theta) < 0$, $VI^{fn}(\theta)$ is decreasing with*

θ. That is, the value of information is larger for less patient customers, if it benefits those customers. Similar conclusions hold for $VI^{pn}(\theta)$.

Proof. We present the proof for full versus no information. The case of partial versus no information is similar.

When $\theta > \theta_-$, let's first write

$$VI^{fn} = u^{full}(\theta) = \int_0^{c^{-1}(1/\theta)} (1 - \theta c(v)) f^{full}(v)dv + (1-\theta)p_0^{full}.$$

We have

$$\frac{dVI^{fn}}{d\theta} = -\int_0^{c^{-1}(1/\theta)} c(v) f^{full}(v)dv - p_0^{full} < 0.$$

When $\theta \le \theta_-$,

$$\begin{aligned} VI^{fn} &= u^{full}(\theta) - u^{no}(\theta) \\ &= \int_0^{c^{-1}(1/\theta)} (1 - \theta c(v)) f^{full}(v)dv + (1-\theta)p_0^{full} - (1 - \theta E[c(W^{no})]). \end{aligned}$$

We have

$$\begin{aligned} \frac{dVI^{fn}}{d\theta} &= -\int_0^{c^{-1}(1/\theta)} c(v) f^{full}(v)dv - p_0^{full} + E[c(W^{no})] \\ &= -E[c(\tilde{V}^{full})] + E[c(W^{no})]. \end{aligned}$$

When $u^{full}(\theta) > u^{no}(\theta)$, $E[c(\tilde{V}^{full})] < E[c(W^{no})]$ and hence $dVI^{fn}/d\theta > 0$; when $u^{full}(\theta) \le u^{no}(\theta)$ then $dVI^{fn}/d\theta \le 0$. □

Thus, the value of information is not necessarily larger for less patient customers.

3.5.1 CARA Utility Function

In this subsection, we study the relationship between the value of information and the degree of risk aversion. We assume H is a uniform distribution and consider the cost function

$$c(w) = e^{\gamma w}$$

for $\gamma > 0$. This function is increasing and convex. To guarantee the stability of the no-information model, we restrict $\gamma < \mu$.

Here, $A(w) = c''(w)/c'(w) = \gamma$. Thus, a larger γ means a higher level of risk-aversion. This risk aversion measure is independent of w, a property called *constant absolute risk aversion* (CARA). Also, the cost function is increasing with γ.

For the numerical computation, we consider $\theta = 0.1, 0.5, 0.9$, which represent very patient, moderately patient and impatient customers, respectively. We fix $\mu = 2$ and change λ over $\{0.5, 1, 2, 3\}$. Figures 3.1, 3.2 and 3.3 show the value of full versus no information with $\theta = 0.1, 0.5, 0.9$, respectively. Figures 3.4, 3.5 and 3.6 show the value of partial versus no information, and Figures 3.7, 3.8 and 3.9 show the value of full versus partial information.

First, we observe that the relationship between the value of information and the level of customers' risk aversion is not monotone. When $\theta = 0.1$, the value of information roughly increases with the level of risk aversion. However, for $\theta = 0.5$ and 0.9, it doesn't. For very patient customers, they join the system under no and full information in most of time. What they care about is congestion. The cost function $c(w) = e^{\gamma w}$ is increasing in γ. So, when γ increases, the system becomes less congested with and without information. However, it is likely that the congestion in the system with information is lessened more than the one without information. Hence, for patient customers, the value of information tends to increase with γ, since they usually join the system, and less congestion with information benefits them most.

However, customers with moderate to severe impatience often leave the system with and without information, and thus the value of information is less promising to them.

In summary, information here not only helps one customer to make his decision, but also affects the congestion of the system itself. Hence, the risk here is endogenous. In this situation, the value of customers' cost function, instead of the shape of it, is an important factor deciding customers' delay risk. And there is no monotone relationship between the value of information and the degree of risk aversion.

We also observe that the value of full versus no information, and that of partial versus no information, decrease with the system's utilization. However, the value of full versus partial information need not behave in this way.

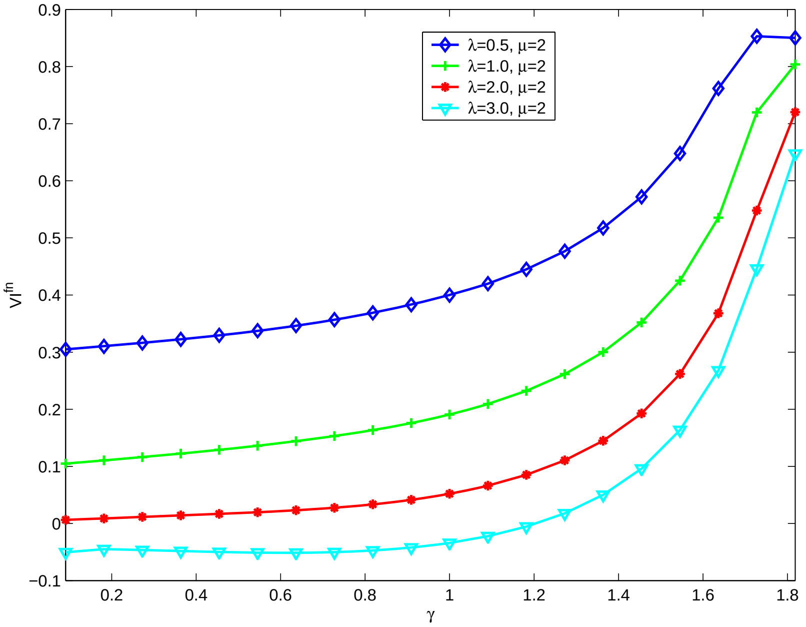

Figure 3.1: The Value of Full/No Information for Customers with $\theta = 0.1$

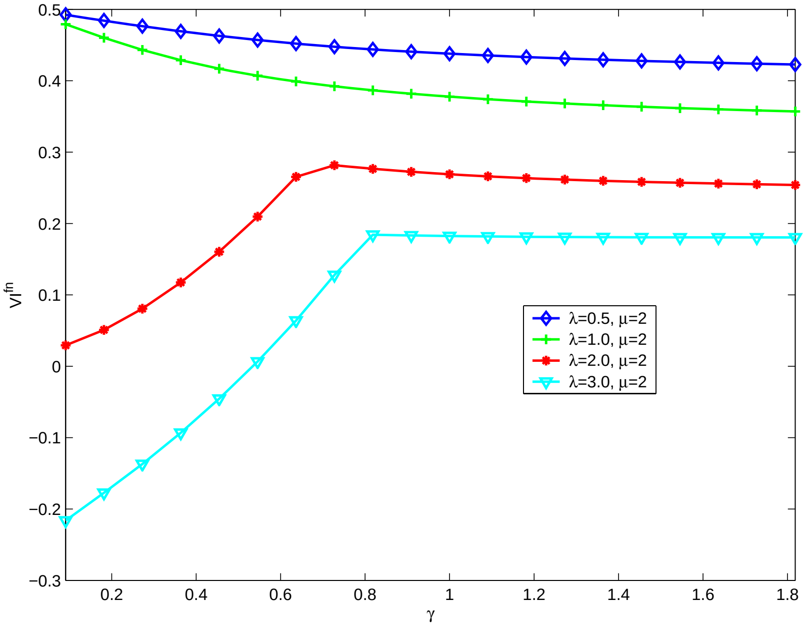

Figure 3.2: The Value of Full/No Information for Customers with $\theta = 0.5$

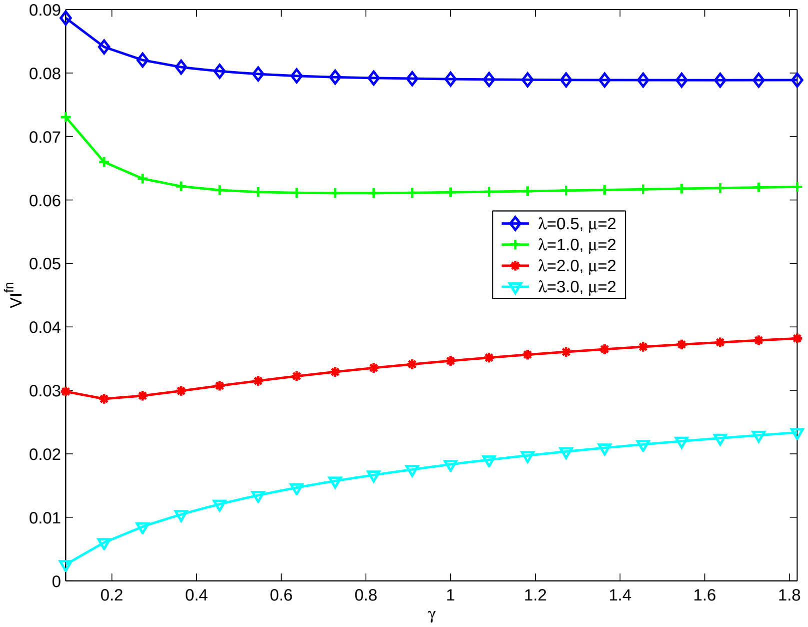

Figure 3.3: The Value of Full/No Information for Customers with $\theta = 0.9$

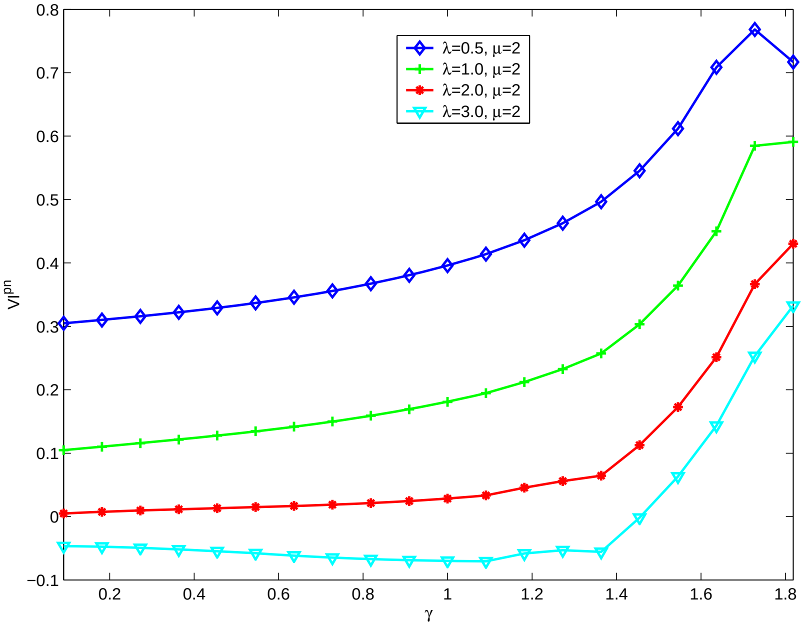

Figure 3.4: The Value of Partial/No Information for Customers with $\theta = 0.1$

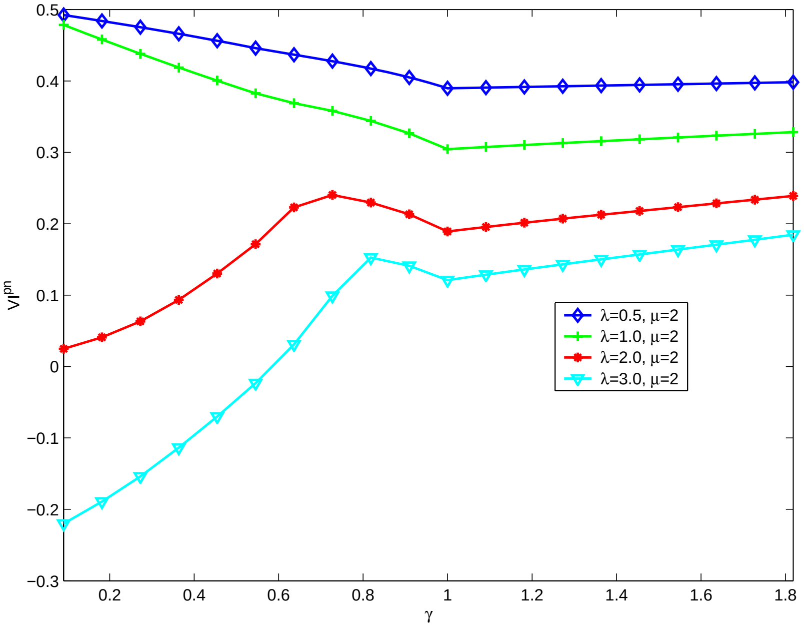

Figure 3.5: The Value of Partial/No Information for Customers with $\theta = 0.5$

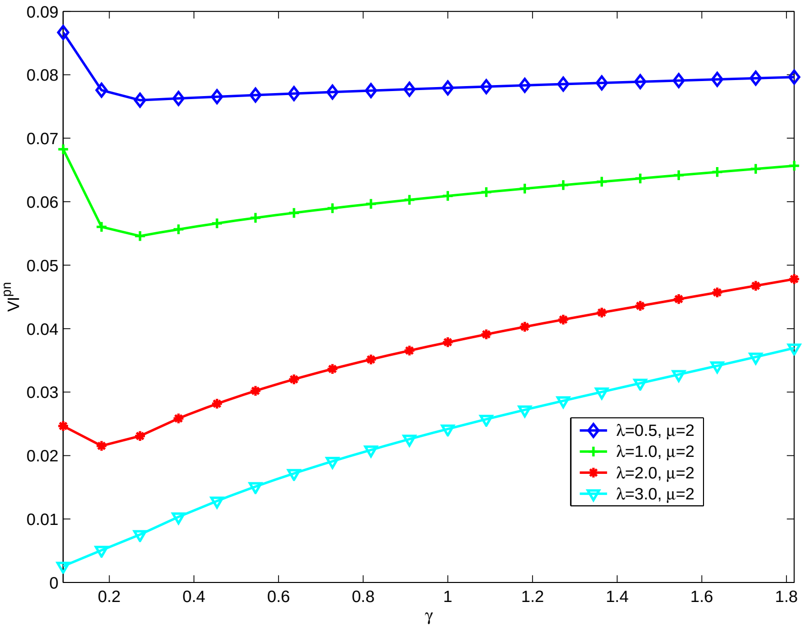

Figure 3.6: The Value of Partial/No Information for Customers with $\theta = 0.9$

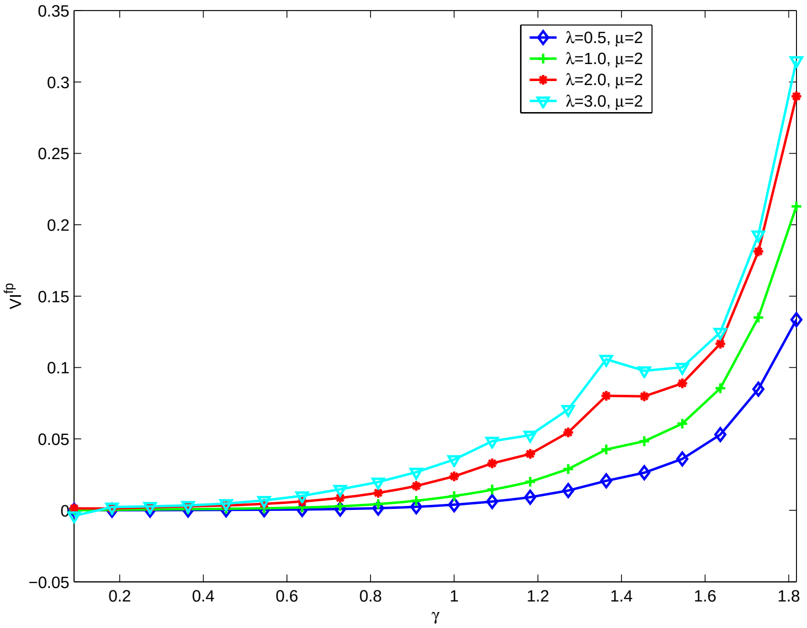

Figure 3.7: The Value of Full/Partial Information for Customers with $\theta = 0.1$

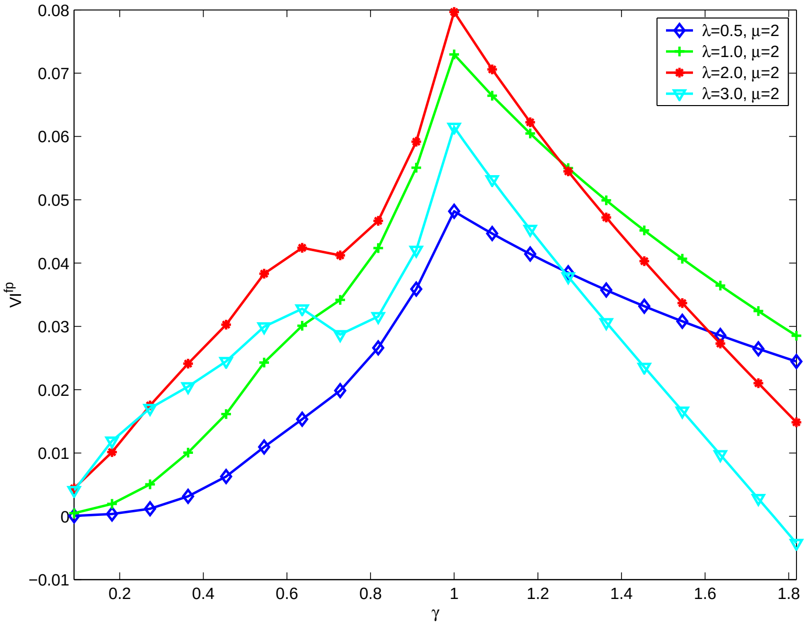

Figure 3.8: The Value of Full/Partial Information for Customers with $\theta = 0.5$

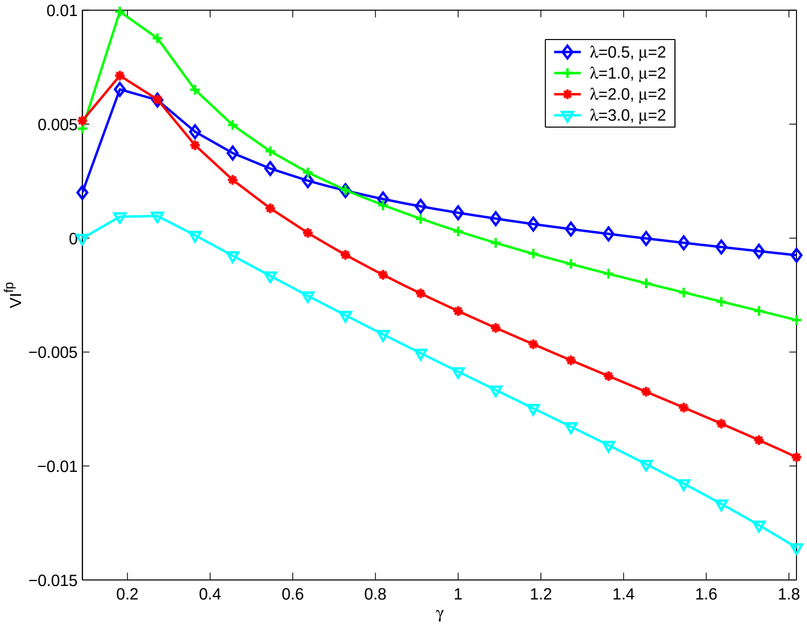

Figure 3.9: The Value of Full/Partial Information for Customers with $\theta = 0.9$

3.6 Conclusions

In this chapter, we explore the impacts of customers' delay sensitivities and risk attitudes on balking queues. We first give stochastic comparison results when the effective arrival rates exhibit certain relationships under different information scenarios. Under partial information, a sufficient condition for a system's throughput to be larger than the other's is that the system have larger cumulative effective arrival rate. (1) If a system always has a larger effective arrival rate than the other system does, the system's occupancy is stochastically larger than the other's, in the sense of likelihood-ratio order; (2) if a system's effective arrival rate crosses the other's once from above, then the system's occupancy is uniformly less variable than the other's; (3) if the ratio of a system's effective arrival rate to the other's is increasing, then the system's occupancy is logconcave relative to the other's. We obtain similar conclusions for the full-information system.

We then consider different orders between H^1 and H^2 and discuss their relationships with orders between the effective arrival rates. For example, we show that when system 1's customers are more sensitive to delay than system 2's in the usual stochastic order, system 1 has smaller effective arrival rates. We also discuss the relationship between c^1 and c^2 and discuss the relationship between them and the effective arrival rates. For example, when H has a decreasing elasticity, increasing c^1/c^2 is a sufficient condition for increasing $\lambda^1(v)/\lambda^2(v)$. We also explore risk-aversion relations between cost functions. We show that a system with more risk-averse customers need not have a smaller throughput.

With regard to customers' average utility, if system 1's customers have larger costs and larger weights on the cost, they have a smaller utility for a fixed waiting time. Hence, in a system without balking, the customers in system 1 get lower average utility. In our system with balking, but without information about delay, this conclusion is still true. However, in a system with balking and information, the average utility in system 1 can be larger.

Finally, we examine the relationship between the value of information and customers' characteristics. We show that, when information hurts the server, it always benefits each customer. However, the value of information need not increase with the weight on delay cost. We also explore numerically the value of information for different delay-sensitive customers with CARA utility functions. We show that there is no simple relationship between the value of information and the degree of customers' risk aversion.

Chapter 4

$\cdot/PH/1$ Queue and Delay Information

In this chapter, we explore information's effects on the server's profit and customers' average utility in a system with phase-type service times. With general service times, information can have some odd effects. A longer queue may indicate a shorter waiting time, as discussed by Altman and Hassin (2002). A similar phenomenon can appear in a two-server system; Whitt (1985) shows that, therefore, it is not always optimal to join the shortest queue.

Section 1 introduces notation and assumptions and briefly summarizes the important conclusions in Chapter 2. Sections 2-4 model systems with no, partial and full information, respectively. Section 5 provides some analytical results comparing the performance of various systems. Section 6 gives numerical results. Section 7 provides some concluding remarks.

4.1 Formulation and Preliminaries

As in Chapter 2, we assume potential customers arrive in a Poisson process. We suppose that a customer's utility equals a reward for receiving service minus a waiting cost. This waiting cost depends on a customer-specific parameter and the expectation of a (common) function of the waiting time. Let S_r denote the residual service time for the customer in service.

4.1.1 Reneging

Suppose for the moment that customers can renege, based on the same calculation as above. When would they do so? With full information, the answer is clearly never. With less information, however, they might. First, consider the case of no information and exponential service times. Suppose an arriving customer decides to stay but immediately finds himself waiting in the queue. He now has more information than he started with; he knows that $W > 0$. At that moment, he may choose to renege. If not, however, he will never leave, because $(W|W > 0)$ is exponential.

With general service times, even this may not hold. For example, in a $M/G/1$ queue with *decreasing-failure-rate* (DFR) service times, the stationary waiting time too is DFR (see, e.g., Shanthikumar, 1988), as is $(W|W > 0)$. In this case, a customer in the queue estimates a larger remaining waiting time as time passes. So, he may choose to renege.

4.1.2 Phase-type Service Times

The phase-type (PH) distribution is discussed in detail by Neuts (1981). The cdf of the PH distribution of service times is expressed as

$$G(x) = 1 - \boldsymbol{\beta}^T e^{\mathbf{B}x}\mathbf{1}, \quad x \geq 0,$$

where $\mathbf{1}$ is a column m-vector of ones, $\boldsymbol{\beta}^T = (\beta_1, \beta_2, ..., \beta_m)$ is a non-negative m-dimensional row vector, and $\mathbf{B}$ is an $m \times m$ matrix with $B_{ij} \geq 0, i \neq j$; $B_{ii} < 0, i = 1, ..., m$; and $\mathbf{B}$ is nonsingular. The general definition allows $\boldsymbol{\beta}^T\mathbf{1} \leq 1$. Here we restrict attention to $\boldsymbol{\beta}^T\mathbf{1} = 1$. We say the distribution has parameters $(\boldsymbol{\beta}, \mathbf{B})$.

4.1.3 Summary of Conclusions in Chapter 2

We summarize the important conclusions obtained in Chapter 2. We use superscripts no, $part$ and $full$ to indicate the different information levels. Also, define the function

$$J(\theta) = \frac{\int_0^\theta H(\phi)d\phi}{\theta}.$$

The expected utility, u, equals $E_I[J(\theta_I)]$, where I denotes the information variable. Denote by p_0 the idle probability. A power distribution has cdf $H(\theta) = \theta^\alpha$ for $\alpha > 0$. (The uniform distribution is the case with $\alpha = 1$.)

1. If $p_0^{part} \geq p_0^{no}$, then $u^{part} > u^{no}$. Thus, more information helps *someone* – if not the provider, then the customers.

2. If $H(1/x)$ is convex in x, then $p_0^{part} \leq p_0^{no}$.

3. If $J \circ H^{-1}$ is convex, then $u^{part} > u^{no}$.

Chapter 2 also shows that, for the special case of a power distribution, the customers' average utility is proportional to the throughput. In this case, the provider's and the customers' objectives are perfectly aligned. That argument holds for general service times, not just exponential ones.

The same relations also hold between no and full information. And the 2nd relation holds between partial and full information. It is conjectured that the 1st and the 3rd hold too.

The condition that $H(1/x)$ is convex in x in property (2) means that the cost-scale distribution is spread out, so customers are heterogeneous, in a certain sense. It is equivalent to

$$-\frac{\theta h'(\theta)}{h(\theta)} \leq 2$$

(assuming the derivative h' exists). The left-hand side is the elasticity of the density h. The condition posits that h not be too elastic, that is, that customers not be too concentrated. More precisely, it rules out a sharp *decrease* in h. It is a one-sided spread condition. For a beta distribution with parameters (a, b), the condition holds, if and only if $b \leq 1$.

The condition that $J \circ H^{-1}$ is convex in property (3) is equivalent to

$$-\frac{\phi h'(\phi)}{h(\phi)} \geq 2 - \frac{\phi h(\phi)}{[H(\phi) - J(\phi)]}.$$

This is another restriction on the elasticity of h, but here it's a *lower* bound, a variable one. It requires that h not *rise* too sharply. For a beta distribution with parameters (a, b), this condition holds, if and only if $b \geq 1$.

The power distribution is the special case of a beta distribution with parameter $b = 1$. So, in this case more information helps everyone.

Here, we investigate whether these properties continue to hold with phase-type service times.

4.2 No Information

Under no information, customers use the equilibrium distribution to estimate W. Thus, W affects the arrivals, which, in turn, affect W. When the system is in equilibrium, there exists a cut-off level, θ_-, for customers' sensitivity parameter θ. At θ_-, the customer has utility 0 and thus is indifferent between joining and balking. Obviously, $\theta_- = 1/E[c(W|-)]$. Below that, customers have positive utility and join; above that, customers get negative utility and leave. Thus the fraction of customers joining the system is $H(\theta_-)$. Those customers form the effective arrival process. The effective arrival rate is $\lambda_- = \lambda H(\theta_-)$. The resulting system is an $M/PH/1$ queue with an effective arrival rate λ_-.

Denote $\rho_- = \lambda_-/\mu$, where $\mu = 1/E[S]$ is the average service rate. The distribution of waiting time W for such a queue is also of phase type (see Neuts Neuts) with representation $(\boldsymbol{\gamma}, \boldsymbol{\Gamma})$, where

$$\boldsymbol{\gamma} = \rho_- \boldsymbol{\delta}, \quad \boldsymbol{\Gamma} = \mathbf{B} + \rho_- \mathbf{B^m} \boldsymbol{\Delta}^{\mathbf{o}},$$

and $\mathbf{B^m}$ is an $m \times m$ matrix with identical columns $\mathbf{B^o}$, $\mathbf{B^o} = -\mathbf{B1}$,

$$\mathbf{\Delta^o} = diag(\delta_1, ..., \delta_m),$$

and $\boldsymbol{\delta}$ is the stationary probability vector of $\mathbf{B} + \mathbf{B^o}\boldsymbol{\beta^T}$.

The equilibrium λ_- solves

$$\lambda_- = \lambda H\left(\frac{1}{E[c(W|-)]}\right). \tag{4.1}$$

The expected utility is shown in Chapter 2 to be

$$u = E[U^+] = J(\theta_-).$$

where $\theta_- = 1/E[c(W|-)]$.

For some cost functions, for example, the linear and quadratic, the expected waiting cost can be expressed as the function of moments of waiting time. The moments of $PH(\boldsymbol{\gamma}, \mathbf{\Gamma})$ can be expressed as

$$E[W^n] = (-1)^n n! \boldsymbol{\gamma}^T \mathbf{\Gamma}^{-n} \mathbf{1}.$$

Generally, the moments of W for the $M/G/1$ queue can be computed using the Pollaczek-Khintchine formula (Heyman and Sobel, 1986). For example, the expected waiting time is

$$E[W] = \frac{\lambda_- E[S^2]}{2(1-\rho_-)},$$

where $E[S^2]$ is the second moment of S.

4.3 Partial Information

4.3.1 Discussion

In Chapter 2, partial information means each customer learns N, the system occupancy. He then estimates his waiting time as $W = S^{(N)}$. Here, this approach is problematic. For $N > 0$, the waiting time is $W = S_r + S^{(N-1)}$. How does the customer estimate S_r? This problem is quite intricate.

Consider a standard $M/G/1$ system. Denote the failure rate of N by r_n, and let p_n be the stationary probability $Pr\{N = n\}$, i.e.,

$$r_n = \frac{p_n}{\sum_{i=n}^{\infty} p_i}.$$

Mandelbaum and Yechiali (1979) and Fakinos (1982) show that the mean conditional residual service time in an $M/G/1$ system is

$$E[S_r|N = n] = \frac{1-\rho}{\lambda p_n}\left(1 - \sum_{i=0}^{n} p_i\right) = \frac{1-\rho}{\lambda r_n} - \frac{1-\rho}{\lambda}, \quad n = 1, 2, 3, ... \tag{4.2}$$

where ρ is the system utilization. This depends on λ as well as n.

This formula (4.2) assumes a constant arrival rate. Here, the arrival rate is state-dependent. But it is likely that $(S_r|N)$ is at least as complex. Altman and Hassin (2002) present an example where $(W|N)$ actually can decrease in N.

Consequently, we examine two alternative models of partial information. First, we assume that the server tells customers the current phase of the service process as well as the system occupancy. Second, we assume that customers learn only N, but estimate S_r in a simple way.

4.3.2 Models

Assume the incoming customer is told the current occupancy, N, and the phase of the customer in service, L. Given $L = l$, the residual service time, denoted by S_{rl} , is the service time starting from state l. This has the PH distribution with parameter $(\mathbf{1}_l, \mathbf{B})$, where $\mathbf{1}_l$ is the unit column vector with l'th coordinate 1. Given the information $(N = n, L = l)$, the waiting time for the incoming customer is

$(W|N = n, L = l) = S^{(n-1)} + S_{rl}$. The convolution of PH distributions is itself a PH distribution. Hence, $(W|N = n, L = l)$ has a PH distribution.

Define the effective arrival rate $\lambda(n, l) = \lambda H(1/E[c(W|N = n, L = l)])$. The system can be modelled as a homogeneous Markov process with state space

$$\{(0); (n, l), \quad n = 1, 2, ..., \quad l = 1, 2, ..., m\},$$

where the state (0) describes an empty system, and (n, l) means there are n customers present and the current service is in phase l. Denote the stationary probabilities by p_0 and p_{nl}. Define the row vector $\mathbf{p}_n^T = (p_{n1}, ..., p_{nm})$.

This system is similar to what Neuts (1981) calls a quasi-birth-death process. Define $\mathbf{\Lambda}_n$ as a $m \times m$ diagonal matrix with the element at (l, l) equal $\lambda(n, l)$. We can write the balance equations in matrix form as

$$\begin{aligned} 0 &= -\lambda(0)p_0 + \mathbf{p}_1^T\mathbf{B^o} && (4.3)\\ \mathbf{0}^T &= \mathbf{p}_1^T(-\mathbf{\Lambda}_1 + \mathbf{B}) + \lambda(0)p_0\boldsymbol{\beta}^T + \mathbf{p}_2^T\mathbf{B^o}\boldsymbol{\beta}^T && (4.4)\\ \mathbf{0}^T &= \mathbf{p}_n^T(-\mathbf{\Lambda}_n + \mathbf{B}) + \mathbf{p}_{n-1}^T\mathbf{\Lambda}_{n-1} + \mathbf{p}_{n+1}^T\mathbf{B^o}\boldsymbol{\beta}^T, \quad n = 2, 3, ... && (4.5) \end{aligned}$$

Multiplying (4.4) and (4.5) by the vector $\mathbf{1}$ over $n = 1, 2, ...$ and summing successively yields

$$\mathbf{p}_n^T\mathbf{\Lambda}_n\mathbf{1} = \mathbf{p}_{n+1}^T\mathbf{B^o}, \quad n = 0, 1, 2, ... \tag{4.6}$$

Substituting (4.6) back to (4.4) and (4.5), we obtain

$$\begin{aligned} 0 &= -\lambda(0)p_0 + \mathbf{p}_1^T\mathbf{B^o}\\ \mathbf{0}^T &= \mathbf{p}_1^T(-\mathbf{\Lambda}_1 + \mathbf{B}) + \lambda(0)p_0\boldsymbol{\beta}^T + \mathbf{p}_1^T\mathbf{\Lambda}_1\mathbf{1}\boldsymbol{\beta}^T\\ \mathbf{0}^T &= \mathbf{p}_n^T(-\mathbf{\Lambda}_n + \mathbf{B}) + \mathbf{p}_{n-1}^T\mathbf{\Lambda}_{n-1} + \mathbf{p}_n^T\mathbf{\Lambda}_n\mathbf{1}\boldsymbol{\beta}^T, \quad n = 2, 3, ... \end{aligned}$$

Introducing the matrix $\tilde{\mathbf{B}}_n = \mathbf{\Lambda}_n(\mathbf{I} - \mathbf{1}\boldsymbol{\beta}^T) - \mathbf{B}, n = 1, 2, ...,$ we can rewrite the above as:

$$\begin{aligned}\mathbf{p}_1^T\tilde{\mathbf{B}}_1 &= \lambda(0)p_0\boldsymbol{\beta}^T \\ \mathbf{p}_n^T\tilde{\mathbf{B}}_n &= \mathbf{p}_{n-1}^T\boldsymbol{\Lambda}_{n-1}, \quad n = 2, 3, ...\end{aligned}$$

For a quasi-birth-death process, $\tilde{\mathbf{B}}_n$ reduces to $\lambda\mathbf{I} - \lambda\mathbf{1}\boldsymbol{\beta}^T - \mathbf{B}$, which is shown in Neuts (1981) to be nonsingular. Here we show that $\tilde{\mathbf{B}}_n$ is also nonsingular.

Lemma 25. *$\tilde{\mathbf{B}}_n$ is nonsingular.*

Proof. Define $\mathbf{D}_n = -\boldsymbol{\Lambda}_n + \mathbf{B}$ and $\boldsymbol{\lambda}_n = \boldsymbol{\Lambda}_n\mathbf{1}$. Then $-\tilde{\mathbf{B}}_n = \mathbf{D}_n + \boldsymbol{\lambda}_n\boldsymbol{\beta}^T$. Since $\mathbf{B}$ is stable (all its eigenvalues have negative real parts), $\mathbf{D}_n$ too is stable and hence nonsingular. By Lemma 2.8.2 of Bocharov, et al. (2004), the following inverse exists,

$$\tilde{\mathbf{B}}_n^{-1} = -(\mathbf{D}_n + \boldsymbol{\lambda}_n\boldsymbol{\beta}^T)^{-1} = -\mathbf{D}_n^{-1} + \frac{\mathbf{D}_n^{-1}\boldsymbol{\lambda}_n\boldsymbol{\beta}^T\mathbf{D}_n^{-1}}{1 + \boldsymbol{\beta}^T\mathbf{D}_n^{-1}\boldsymbol{\lambda}_n},$$

provided $1 + \boldsymbol{\beta}^T\mathbf{D}_n^{-1}\boldsymbol{\lambda}_n \neq 0$. But,

$$\begin{aligned}1 + \boldsymbol{\beta}^T\mathbf{D}_n^{-1}\boldsymbol{\lambda}_n &= 1 - \boldsymbol{\beta}^T(\boldsymbol{\Lambda}_n - \mathbf{B})^{-1}\boldsymbol{\lambda}_n \\ &= 1 - \boldsymbol{\beta}^T[\boldsymbol{\Lambda}_n(\mathbf{I} - \boldsymbol{\Lambda}_n^{-1}\mathbf{B})]^{-1}\boldsymbol{\lambda}_n \\ &= 1 - \boldsymbol{\beta}^T(\mathbf{I} - \boldsymbol{\Lambda}_n^{-1}\mathbf{B})^{-1}\boldsymbol{\Lambda}_n^{-1}\boldsymbol{\Lambda}_n\mathbf{1} \\ &= 1 - \boldsymbol{\beta}^T(\mathbf{I} - \boldsymbol{\Lambda}_n^{-1}\mathbf{B})^{-1}\mathbf{1}\end{aligned}$$

This is the Laplace-Stieltjes transform of a PH distribution with representation $(\boldsymbol{\beta}, \boldsymbol{\Lambda}_n^{-1}\mathbf{B})$ at the value 1. Thus, it must be positive. Therefore, the matrix $\tilde{\mathbf{B}}_n^{-1}$ exists. □

Define $\mathbf{W}_1 = \lambda(0)\tilde{\mathbf{B}}_1^{-1}$ and $\mathbf{W}_n = \boldsymbol{\Lambda}_{n-1}\tilde{\mathbf{B}}_n^{-1}, n = 2, 3, ...$ Then the stationary distribution admits the representation

$$\mathbf{p}_n^T = p_0\boldsymbol{\beta}^T\prod_{i=1}^{n}\mathbf{W}_n, \quad n = 1, 2, ...$$

By the normalization condition

$$p_0 + \sum_{n=1}^{\infty} \mathbf{p}_n^T \mathbf{1} = 1,$$

we can now solve for p_0.

Define $\mathbf{J}_n = [J(\theta_{n1}), ..., J(\theta_{nm})]'$, where $\theta_{nl} = 1/E[c(W|N = n, L = l)]$. The expected utility is:

$$u = E[U^+] = \sum_{n=1}^{\infty} \mathbf{p}_n^T \mathbf{J}_n + p_0 J(1).$$

Here is a simpler and more realistic partial-information model. Arriving customers learn only N, and they approximate the residual service time by S_e, the equilibrium distribution of S. According to Theorem 2.2.3. of Neuts (1981), S_e has a phase-type distribution with representation $(\boldsymbol{\delta}, \mathbf{B})$.

Then, given information $N = n$, a customer estimates the waiting time as

$$(W|N = n) = S^{(n-1)} + S_e, \quad n \geq 1.$$

The conditional arrival rate is now $\lambda_n = \lambda H(1/E[c(W|N = n)])$. The balance equations and solutions are the same as the above, except replacing $\mathbf{\Lambda}_n$ with $\lambda_n \mathbf{I}$.

4.4 Full Information

For work on queues with workload-dependent arrival rate, see Hu and Zazanis (1993), Bekker, et al (2004) and Bekker (2005). Recently, Liu and Kulkarni (2006) analyze the $M/G/1$ queue with workload-based balking. Their model assumes identical customers.

The effective arrival rate given workload v is $\lambda(v) = \lambda H(\theta_v)$, where $\theta_v = 1/c(v)$. Using a level-crossing argument, e.g. Brill and Posner (1977), we derive an integral equation for the workload with workload-dependent balking and non-identical customers in Chapter 2

$$f(v) = \lambda p_0 \bar{G}(v) + \int_0^v \lambda H(\theta_w) \bar{G}(v-w) f(w) dw \tag{4.7}$$

and the normalization condition

$$p_0 + \int_0^\infty f(v) dv = 1, \tag{4.8}$$

where $f(v), v > 0$ is the density function of the equilibrium workload V, and p_0 is its mass at 0. (A sufficient condition for positive recurrence is that $\lambda(v) \to 0$ as $v \to \infty$ and that G has finite mean. See, Perry and Asmussen (1995).)

Here, service times have the phase-type distribution with parameters $(\boldsymbol{\beta}, \mathbf{B})$. Then,

$$\begin{aligned} f(v) &= \lambda p_0 \left(\boldsymbol{\beta}^T e^{\mathbf{B}v} \mathbf{1} \right) + \int_0^v \lambda H(\theta_w) \left(\boldsymbol{\beta}^T e^{\mathbf{B}(v-w)} \mathbf{1} \right) f(w) dw \\ &= \boldsymbol{\beta}^T \left[\lambda p_0 e^{\mathbf{B}v} + \int_0^v \lambda H(\theta_w) e^{\mathbf{B}(v-w)} f(w) dw \right] \mathbf{1} \\ &= \boldsymbol{\beta}^T \mathbf{f}(v), \end{aligned} \tag{4.9}$$

where

$$\mathbf{f}(v) = \left[\lambda p_0 e^{\mathbf{B}v} + \int_0^v \lambda H(\theta_w) e^{\mathbf{B}(v-w)} f(w) dw \right] \mathbf{1}. \tag{4.10}$$

Equation (4.9) can be solved explicitly. The result is given in the following proposition.

Proposition 26.

$$f(v) = \lambda p_0 \boldsymbol{\beta}^T e^{\lambda C(v) \mathbf{1} \boldsymbol{\beta}^T + \mathbf{B}v} \mathbf{1}, \tag{4.11}$$

where $C(v) = \int_0^v H(\theta_w) dw$.

Proof. From (4.10), we have the initial condition $\mathbf{f}(0) = \lambda p_0 \mathbf{1}$. Also,

$$\begin{aligned}
\mathbf{f}'(v) &= \left[\lambda p_0 \left(\mathbf{B}e^{\mathbf{B}v}\right) + \lambda C'(v) f(v)\mathbf{I} + \int_0^v \lambda C'(w)\left(\mathbf{B}e^{\mathbf{B}(v-w)}\right) f(w)dw\right]\mathbf{1} \\
&= \lambda C'(v) f(v)\mathbf{1} + \mathbf{B}\left[\lambda p_0 e^{\mathbf{B}v} + \int_0^v \lambda C'(w) e^{\mathbf{B}(v-w)} f(w)dw\right]\mathbf{1} \\
&= \lambda C'(v)[\boldsymbol{\beta}^T \mathbf{f}(v)]\mathbf{1} + \mathbf{B}\mathbf{f}(v) \\
&= \lambda C'(v)\left(\mathbf{1}\boldsymbol{\beta}^T\right)\mathbf{f}(v) + \mathbf{B}\mathbf{f}(v) \\
&= \left[\lambda C'(v)\mathbf{1}\boldsymbol{\beta}^T + \mathbf{B}\right]\mathbf{f}(v).
\end{aligned}$$

It is easy to verify that the proposed solution indeed satisfies these relations. □

Generally, to solve (4.7), we can use numerical methods; see Perry and Asmussen (1995) and Bekker et al. (2004).

The expected utility is

$$u = E[U^+] = \int_0^\infty J(\theta_v) f(v)dv + p_0 J(1).$$

4.5 Some Comparisons

In this section, for no and full information, we compare two systems with different service times. We use superscript $1, 2$ to indicate the two systems. We consider the situation $S^1 \preceq_{cx} S^2$, where $\preceq_{cx}$ means the convex order. X is less than Y in convex order, if $E[f(X)] \le E[f(Y)]$ for all convex functions f such that the expectations exist. $S^1 \preceq_{cx} S^2$ implies $E[S^1] = E[S^2]$ and $var[S^1] \le var[S^2]$. For example, O'Cinneide (1991) shows that a PH-distribution with an order m representation is always larger than the order m Erlang distribution of the same mean, in the sense of the convex order.

We also compare the systems with no and full information and identical S.

4.5.1 Equilibrium Service Time

The distribution of the equilibrium service time S_e is given by

$$G_e(x) = \int_0^x \frac{\bar{G}(t)}{E[S]} dt, \quad x \geq 0.$$

Two random variables are related by $S^1 \preceq_{cx} S^2$ means that $E[S^1] = E[S^2]$ and $\int_t^\infty \bar{G}^1(x)dx \leq \int_t^\infty \bar{G}^2(x)dx$ for all t. This implies $G_e^1(x) \geq G_e^2(x)$ for all x, or $S_e^1 \preceq_{st} S_e^2$. Also, $E[S_e] = E[S^2]/(2E[S])$.

4.5.2 No Information

Consider two systems with no information and service times S^1 and S^2.

Proposition 27. *For no information, if $S^1 \preceq_{cx} S^2$, then $p_0^1 \leq p_0^2$.*

Proof. $S^1 \preceq_{cx} S^2$ implies $S_e^1 \preceq_{st} S_e^2$. For the $M/G/1$ queue, the waiting time is the geometric sum of S_e (See Ross 1983), i.e.,

$$W =_{st} \sum_{q=1}^{Q} X_q,$$

where $X_1, X_2, ...$ are i.i.d. with distribution function G_e and Q is geometrically distributed with parameter ρ_-.

Now assume $\rho_-^1 < \rho_-^2$, then $Q^1 \preceq_{st} Q^2$. By Theorem 4.3.5. of Müller and Stoyan (2002), $Q^1 \preceq_{st} Q^2$ and $X_q^1 \preceq_{st} X_q^2$ lead to $W^1 \preceq_{st} W^2$. Since $c(\cdot)$ is an increasing function, $E[c(W^1|-)] \leq E[c(W^2|-)]$. So, from equation (4.1), $\rho_-^1 \geq \rho_-^2$, a contradiction. Hence, $\rho_-^1 \geq \rho_-^2$ or $p_0^1 \leq p_0^2$. □

Proposition 28. *For no information, if $S^1 \preceq_{cx} S^2$, then $u^1 \geq u^2$.*

Proof. By Proposition 27, $S^1 \preceq_{cx} S^2$ implies $p_0^1 \leq p_0^2$. Hence, $E[c(W^1|-)] \leq E[c(W^2|-)]$. Since $J(1/x)$ is decreasing in x, we obtain that

$$\begin{aligned} u^1 &= J(1/E[c(W^1|-)]) \\ &\geq J(1/E[c(W^2|-)]) = u^2. \end{aligned}$$

□

Thus, a more variable service time reduces the system's throughput and decreases customers' average utility.

4.5.3 Full Information

Lemma 29. *For full information,*

$$p_0 = 1 \Bigg/ \left(1 + \sum_{n=1}^{\infty} \zeta_n\right), \tag{4.12}$$

where

$$\zeta_n = (\lambda E[S])^n E\left[H\left(1/c(S_e^{(1)})\right) H\left(1/c(S_e^{(2)})\right) \cdots H\left(1/c(S_e^{(n-1)})\right)\right].$$

Proof. The solution of the integral equation (4.7) has the following general form (e.g. Perry and Asmussen, 1995):

Denote $\lambda(v) = \lambda H(\theta_v)$, and $K(x,y) = \lambda(y)\bar{G}(x-y)$. Define

$$K^{(1)} = K, \quad K^{(n+1)}(x,y) = \int_y^x K(x,z)K^{(n)}(z,y)dz, \quad K^* = \sum_{n=1}^{\infty} K^{(n)}.$$

Then,

$$p_0 = 1 \Big/ \left(1 + \int_0^{\infty} K^*(x,0)dx\right), \quad f(x) = p_0 K^*(x,0).$$

Rewrite $\int_0^\infty K^*(x,0)dx$ as $\sum_{n=1}^\infty \int_0^\infty K^{(n)}(x,0)dx$. Denote $\zeta_n = \int_0^\infty K^{(n)}(x,0)dx$. In the following, we consider ζ_n for each n.

$$\zeta_1 = \int_0^\infty K^{(1)}(x,0)dx = \lambda(0)\int_0^\infty \bar{G}(x)dx = \lambda E[S].$$

$$\begin{aligned}
\zeta_2 &= \int_0^\infty K^{(2)}(x,0)dx = \int_0^\infty \left(\int_0^x \lambda(z)\bar{G}(x-z)\lambda(0)\bar{G}(z)dz\right)dx \\
&= \lambda\int_0^\infty \lambda(z)\left(\int_z^\infty \bar{G}(x-z)\bar{G}(z)dx\right)dz \\
&= \lambda E[S]\int_0^\infty \lambda(z)\bar{G}(z)dz \\
&= (\lambda E[S]^2)\int_0^\infty \lambda(z)(\bar{G}(z)/E[s])dz \\
&= (\lambda E[S])^2 E\left[H\left(1/c(S_e^{(1)})\right)\right].
\end{aligned}$$

The last step comes from the fact that $\bar{G}(x)/E[s]$ is the pdf of S_e.

$$K^{(3)}(x,y) = \int_y^x \lambda(z)\bar{G}(x-z)K^{(2)}(z,y)dz.$$

Thus,

$$\begin{aligned}
\zeta_3 &= \int_0^\infty K^{(3)}(x,0)dx \\
&= \int_0^\infty \left[\int_0^x \lambda(z)\bar{G}(x-z)\left(\int_0^z \lambda(t)\bar{G}(z-t)\lambda(0)\bar{G}(t)dt\right)dz\right]dx \\
&= \lambda\int_0^\infty\int_0^x\int_0^z \lambda(z)\bar{G}(x-z)\lambda(t)\bar{G}(z-t)\bar{G}(t)dtdzdx \\
&= \lambda\int_0^\infty\int_0^z\int_z^\infty \lambda(z)\bar{G}(x-z)\lambda(t)\bar{G}(z-t)\bar{G}(t)dxdtdz \\
&= \lambda E[S]\int_0^\infty\int_0^z \lambda(z)\lambda(t)\bar{G}(z-t)\bar{G}(t)dtdz \\
&= \lambda E[S]\int_0^\infty\int_t^\infty \lambda(z)\lambda(t)\bar{G}(z-t)\bar{G}(t)dzdt \\
&= \lambda E[S]\int_0^\infty \lambda(t)\left(\int_0^\infty \lambda(z+t)\bar{G}(z)dz\right)\bar{G}(t)dt \\
&= \lambda E[S]^3\int_0^\infty \lambda(t)\left(\int_0^\infty \lambda(z+t)(\bar{G}(z)/E[S])dz\right)(\bar{G}(t)/E[S])dt \\
&= (\lambda E[S])^3 E\left[H\left(1/c(S_e^{(1)})\right)H\left(1/c(S_e^{(2)})\right)\right].
\end{aligned}$$

Continuing in this way yields the conclusion. □

We have the following proposition.

Proposition 30. *For full information, if $S^1 \preceq_{cx} S^2$, then $p_0^1 \leq p_0^2$.*

Proof. $S^1 \preceq_{cx} S^2$ implies $S_e^1 \preceq_{st} S_e^2$. Thus $(S_e^1)^{(n)} \preceq_{st} (S_e^2)^{(n)}$. Since $H(1/c(\cdot))$ is a decreasing function,

$$\begin{aligned}
&E\left[H\left(1/c((S_e^1)^{(1)})\right)\cdots H\left(1/c((S_e^1)^{(n-1)})\right)\right] \\
&\geq E\left[H\left(1/c((S_e^2)^{(1)})\right)\cdots H\left(1/c((S_e^2)^{(n-1)})\right)\right].
\end{aligned}$$

This leads to $p_0^1 \leq p_0^2$ from Lemma 29. □

The conclusion here is the same as Proposition 27's. Again, variability of the service times hurts the server.

4.5.4 Comparison Between No and Full Information

In order to compare the no and full information systems, we introduce another birth-death process, solely to help in the analysis.

Consider a system with exponential service times with mean $E[S]$. Customers learn the system occupancy N at arrival epochs. However, in customers' beliefs, S is replaced by S_e. That is, given $N = n$, the arriving customer estimates his waiting time to be $S_e^{(n)}$. Hence, the effective arrival rate is $\lambda_n = \lambda H\left(1/E[c(S_e^{(n)})]\right)$. The system occupancy N can be modeled as a birth-death process as in Chapter 2. Denote the stationary probabilities for this system by p_n^{BD}. The standard analysis yields

$$p_0^{BD} = 1 \Bigg/ \left(1 + \sum_{n\geq 1} \theta_n\right), \tag{4.13}$$

where

$$\theta_n = (\lambda E[S])^n H\left(1/E[c(S_e^{(1)})]\right) H\left(1/E[c(S_e^{(2)})]\right) \cdots H\left(1/E[c(S_e^{(n-1)})]\right).$$

Next, consider a similar system with exponential service times with mean $E[S]$ and no information. Still, in customers' beliefs, S is replaced by S_e. The distribution of N here is geometric. Suppose the utilization is ρ'_{-}. Then the waiting time, in customers belief, is a geometric sum of S_e, i.e.,

$$W =_{st} \sum_{q=1}^{Q} X_q,$$

where $X_1, X_2, ...$ are i.i.d. with distribution function G_e and Q is geometrically distributed with parameter ρ'_{-}. The effective arrival rate solves (4.1). This is exactly the effective arrival rate in the real system under no information. Hence, $\rho'_- = \rho_-$, and the idle probability for this system is p_0^{no}.

According to Chapter 2, in the system with exponential service times, if $H(1/x)$ is convex in x, $p_0^{part} \leq p_0^{no}$. One can easily verify that the proof for this conclusion still holds, if customers in both systems replace S with S_e in their beliefs. But the analogue of the partial-information system here is precisely the birth-death process above. Hence,

Lemma 31. *If $H(1/x)$ is convex in x, $p_0^{BD} \leq p_0^{no}$.*

Furthermore, if $H(1/x)$ is convex in x, we have

$$\begin{aligned}
\zeta_n &= (\lambda E[S])^n E\left[H\left(1/c(S_e^{(1)})\right) H\left(1/c(S_e^{(2)})\right) \cdots H\left(1/c(S_e^{(n-1)})\right)\right] \\
&\geq (\lambda E[S])^n E\left[H\left(1/c(S_e^{(1)})\right)\right] E\left[H\left(1/c(S_e^{(2)})\right)\right] \cdots E\left[H\left(1/c(S_e^{(n-1)})\right)\right] \\
&\geq (\lambda E[S])^n H\left(1/E[c(S_e^{(1)})]\right) H\left(1/E[c(S_e^{(2)})]\right) \cdots H\left(1/E[c(S_e^{(n-1)})]\right) \\
&= \theta_n. \qquad (4.14)
\end{aligned}$$

The first inequality follows from the positive correlation between $S_e^{(m)}$ and $S_e^{(n)}$; the second inequality is by Jensen's inequality. By (4.12), (4.13) and (4.14), we obtain the following lemma

Lemma 32. *If $H(1/x)$ is convex in x, $p_0^{full} \leq p_0^{BD}$.*

Combining Lemmas 31 and 32 yields the following result:

Proposition 33. *If $H(1/x)$ is convex in x, $p_0^{full} \leq p_0^{no}$.*

4.6 Numerical Results

4.6.1 Impact of Information

We consider three service time distributions:

- Exponential: $\mu = 4/3$;
- Generalized Erlang: $\mu_1 = 2, \mu_2 = 4$;
- Hyper-exponential: $\mu_1 = 1, \mu_2 = 2, \beta_1 = 0.5, \beta_2 = 0.5$.

All three distributions have the same mean 3/4. The variance is 0.5625 for the exponential, 0.3125 for the generalized Erlang, and 0.6857 for the hyper-exponential.

Let H be a beta distribution with parameters (a, b). We fix $a = 2$ and change b over $\{0.3, 0.5, 1, 2, 4, 8, 16\}$. We change the arrival rate λ over $\{0.5, 1, 1, 2, 4, 8\}$.

Table (4.1) shows the busy probability under different scenarios, and Table (4.2) shows the average utility. We can see from Table (4.1) that, for $b < 1$, the busy probability under more information is always larger than that under less information, while for $b > 1$, this is not true: The bold numbers in Table (4.1) show the cases where more information reduces utilization. These results verify that property (2) holds here. By comparing Tables (4.1) and (4.2), we see that, when the busy probability under more information is smaller, the utility is larger. So, property (1) holds too.

According to property (3), if $b > 1$, the average utility under more information should be larger than that under no information. Table (4.2) shows that this property holds. The bold numbers in the table indicate that, when the condition in property (3) is violated, the average utility under more information can be lower than that under less information.

Also we can see that when $b = 1$ (H is a power distribution), both the busy probability and average customers' utility both increase with information accuracy. Thus, information helps both the server and customers. This too is consistent with the finding in Chapter 2.

Table 4.1: Busy Probability with Linear Cost Function and Beta H

		Exponential			Generalized Erlang			Hyper Exponential		
b	λ	no	part	full	no	part	full	no	part	full
0.3	0.5	0.1338	0.2796	0.2864	0.1438	0.2832	0.2882	0.1297	0.2794	0.2858
	1	0.2058	0.4454	0.4627	0.2233	0.4543	0.4671	0.1986	0.4450	0.4614
	2	0.2979	0.6321	0.6654	0.3253	0.6488	0.6735	0.2867	0.6314	0.6630
	4	0.4042	0.7970	0.8430	0.4415	0.8196	0.8529	0.3890	0.7960	0.8400
0.5	0.5	0.1757	0.2846	0.2941	0.1884	0.2902	0.2968	0.1704	0.2843	0.2933
	1	0.2665	0.4577	0.4819	0.2890	0.4713	0.4885	0.2573	0.4569	0.4798
	2	0.3755	0.6547	0.7002	0.4090	0.6795	0.7118	0.3618	0.6533	0.6967
	4	0.4915	0.8265	0.8839	0.5334	0.8568	0.8961	0.4741	0.8247	0.8802
1	0.5	0.2434	0.2976	0.3095	0.2588	0.3063	0.3137	0.2368	0.2967	0.3081
	1	0.3656	0.4891	0.5202	0.3941	0.5105	0.5308	0.3537	0.4869	0.5169
	2	0.4962	0.7092	0.7661	0.5358	0.7453	0.7830	0.4797	0.7056	0.7609
	4	0.6165	0.8868	0.9440	0.6599	0.9211	0.9555	0.5980	0.8833	0.9404
2	0.5	0.3140	0.3209	0.3297	0.3279	0.3306	0.3352	0.3076	0.3187	0.3278
	1	0.4781	0.5451	0.5722	0.5099	0.5706	0.5866	0.4646	0.5399	0.5674
	2	0.6252	0.7958	0.8465	0.6658	0.8349	0.8663	0.6078	0.7886	0.8400
	4	0.7365	0.9541	0.9861	0.7748	0.9760	0.9914	0.7196	0.9500	0.9842
4	0.5	0.3635	**0.3497**	**0.3505**	0.3688	**0.3560**	**0.3557**	0.3604	**0.3464**	**0.3484**
	1	0.5904	0.6185	0.6304	0.6203	0.6403	0.6463	0.5772	0.6101	0.6245
	2	0.7458	0.8918	0.9226	0.7814	0.9216	0.9397	0.7301	0.8828	0.9165
	4	0.8357	0.9922	0.9990	0.8642	0.9976	0.9996	0.8226	0.9904	0.9987
8	0.5	0.3748	**0.3684**	**0.3657**	0.3749	**0.3708**	**0.3689**	0.3746	**0.3665**	**0.3641**
	1	0.6844	**0.6825**	**0.6831**	0.7062	**0.6978**	**0.6970**	0.6741	0.6750	0.6774
	2	0.8427	0.9609	0.9747	0.8691	0.9769	0.9839	0.8305	0.9548	0.9709
	4	0.9057	0.9997	1.0000	0.9240	1.0000	1.0000	0.8970	0.9996	1.0000
16	0.5	0.3750	**0.3742**	**0.3727**	0.3750	**0.3746**	**0.3738**	0.3750	**0.3738**	**0.3720**
	1	0.7402	**0.7238**	**0.7213**	0.7466	**0.7322**	**0.7301**	0.7359	**0.7193**	**0.7170**
	2	0.9098	0.9928	0.9963	0.9269	0.9972	0.9984	0.9017	0.9906	0.9952
	4	0.9490	1.0000	1.0000	0.9595	1.0000	1.0000	0.9438	1.0000	1.0000

Table 4.2: Average utility with Linear Cost Function and Beta H

		Exponential			Generalized Erlang			Hyper Exponential		
b	λ	no	part	full	no	part	full	no	part	full
0.05	0.5	0.0198	**0.0188**	0.0198	0.0204	**0.0194**	**0.0200**	0.0196	**0.0188**	0.0197
	1	0.0179	**0.0156**	**0.0170**	0.0186	**0.0164**	**0.0173**	0.0176	**0.0156**	**0.0170**
	2	0.0156	**0.0121**	**0.0140**	0.0164	**0.0132**	**0.0144**	0.0153	**0.0120**	**0.0139**
	4	0.0130	**0.0089**	**0.0112**	0.0138	**0.0103**	**0.0117**	0.0126	**0.0088**	**0.0111**
0.1	0.5	0.0360	0.0368	0.0386	0.0373	0.0379	0.0390	0.0354	0.0368	0.0385
	1	0.0315	**0.0305**	0.0332	0.0330	**0.0321**	0.0338	0.0308	**0.0305**	0.0331
	2	0.0262	**0.0235**	0.0271	0.0278	**0.0256**	0.0278	0.0255	**0.0234**	0.0269
	4	0.0206	**0.0172**	0.0213	0.0221	**0.0197**	0.0221	0.0200	**0.0171**	0.0212
0.5	0.5	0.1298	0.1566	0.1635	0.1373	0.1610	0.1651	0.1266	0.1562	0.1631
	1	0.1027	0.1295	0.1389	0.1101	0.1355	0.1409	0.0996	0.1290	0.1384
	2	0.0750	0.0973	0.1075	0.0811	0.1038	0.1093	0.0725	0.0968	0.1072
	4	0.0505	0.0664	0.0748	0.0546	0.0717	0.0753	0.0488	0.0659	0.0750
1	0.5	0.2163	0.2646	0.2756	0.2301	0.2723	0.2786	0.2105	0.2638	0.2749
	1	0.1625	0.2174	0.2317	0.1752	0.2269	0.2351	0.1572	0.2164	0.2310
	2	0.1103	0.1576	0.1707	0.1191	0.1656	0.1724	0.1066	0.1568	0.1707
	4	0.0685	0.0985	0.1052	0.0733	0.1023	0.1035	0.0664	0.0981	0.1064
2	0.5	0.3479	0.4070	0.4225	0.3701	0.4192	0.4279	0.3382	0.4052	0.4211
	1	0.2472	0.3306	0.3517	0.2668	0.3451	0.3574	0.2391	0.3289	0.3505
	2	0.1535	0.2242	0.2405	0.1644	0.2330	0.2409	0.1489	0.2237	0.2414
	4	0.0874	0.1225	0.1257	0.0922	0.1234	0.1207	0.0853	0.1229	0.1285
4	0.5	0.5267	0.5671	0.5828	0.5543	0.5832	0.5915	0.5140	0.5633	0.5802
	1	0.3578	0.4574	0.4835	0.3854	0.4781	0.4933	0.3463	0.4540	0.4812
	2	0.1995	0.2815	0.2981	0.2107	0.2890	0.2959	0.1946	0.2816	0.3003
	4	0.1043	0.1323	0.1327	0.1082	0.1312	0.1259	0.1026	0.1331	0.1364
8	0.5	0.7101	0.7208	0.7293	0.7300	0.7361	0.7400	0.7002	0.7149	0.7255
	1	0.4925	0.5898	0.6120	0.5296	0.6144	0.6272	0.4772	0.5827	0.6075
	2	0.2405	0.3189	0.3285	0.2500	0.3205	0.3225	0.2362	0.3194	0.3321
	4	0.1169	0.1339	0.1342	0.1195	0.1329	0.1272	0.1156	0.1343	0.1383

4.6.2 Impact of Variability of Service Times

In this subsection, we fix the mean of S but change the coefficient of variation.

First, consider a two-phase hyper-exponential service time. We fix $\mu_1 = 1$, and set $E[S] = 0.5$ and $\rho = 0.9$, then $\lambda = \rho/E[S] = 1.8$. We change the squared coefficient defined by $cv^2 = var[S]/E[S]^2$ over the set $\{1, 1.5, 2, 2.5\}$. Given $\mu_1, E[S]$ and cv^2, the other two parameters μ_2, β_1, $(\beta_2 = 1 - \beta_1)$ can be computed easily. The results are shown in Table 4.3. We can see that both the busy probability and the average utility decrease with the coefficient of variation.

Next, consider a m-phase Erlang distribution, where each phase has an exponential distribution with rate μ. Then $E[S] = m/\mu$ and $var[S] = m/\mu^2$. Thus $cv^2 = var[S]/E[S]^2 = 1/m$. Still, we set $E[S] = 0.5$, $\rho = 0.9$, and $\lambda = \rho/E[S] = 1.8$. We change cv^2 over the set $\{1/4, 1/3, 1/2, 1\}$. Given $E[S]$ and cv^2, the parameter μ can be computed as $\mu = 1/(cv^2 E[S])$. The results are shown in Table 4.4. Again, the busy probability and the average utility decrease with the coefficient of variation.

Thus, service-time variation hurts the system's performance for the customers and the server. This finding confirms the results of the previous section. The effect here is the same as for conventional queue-performance measures like expected delay.

4.7 Summary

This paper extends the discussion of Chapter 2 to systems with phase-type service times. Under no information, we show that the equilibrium is an $M/PH/1$ queue with a smaller effective arrival rate. Under partial information, where the service provider informs customers about the system occupancy and the phase of the customer in service, we show that the system is similar to a quasi-birth-death process with state-dependent arrival rates, and we obtain an explicit expression for the stationary vector.

Under full information, we model the workload distribution with a Volterra integral equation of type two, and we obtain a nearly closed-form solution.

We then carry out stochastic comparisons of the two systems with different service times, under no and full information. We show that the variability of service times hurts the system's throughput and customers' average utility. We also compare no- and full-information systems and show that information's effect on the system's throughput depends on the shape of the distribution function of customers' delay-sensitivity parameter.

The numerical results are consistent with the analytical results. They also verify that all the important results in Chapter 2 still hold here. In particular, information's effect is mainly determined by the shape of H, the distribution function of customers' delay-sensitivity parameter. More information does not always improve performance. We also test the influence of the variance of service times. We find that performance declines with the variance.

Table 4.3: Influence of Coefficient of Variation of Hyper-exponential Service Times

	Busy Probability			Average Utility		
cv^2	no	part	full	no	part	full
1.0	0.5766	0.6527	0.6781	0.2486	0.3206	0.3397
1.5	0.5456	0.6400	0.6666	0.2331	0.3171	0.3382
2.0	0.5198	0.6240	0.6537	0.2205	0.3122	0.3346
2.5	0.4978	0.6041	0.6395	0.2100	0.3048	0.3281

Table 4.4: Influence of Coefficient of Variation of Erlang Service Times

	Busy Probability			Average Utility		
cv^2	no	part	full	no	part	full
0.250	0.6389	0.7026	0.7106	0.2811	0.3450	0.3505
0.333	0.6307	0.6960	0.7062	0.2767	0.3415	0.3490
0.500	0.6153	0.6838	0.6981	0.2686	0.3353	0.3463
1.000	0.5766	0.6527	0.6781	0.2486	0.3206	0.3397

Chapter 5

Other Models of Delay Information

Chapter 2 studies a single-server queue with three levels of delay information, none, partial (the system occupancy) and full (the exact waiting time). Each customer decides whether to stay or leave, based on this information and his own sensitivity to delays. Information's effects on the server's profit and customers' average utility are mainly determined by the spread of the distribution of customers' sensitivities to delays.

This chapter considers two other stylized models of information. The first model is specified by a *consecutive partition* of the nonnegative integers, a partition in which each subset consists of consecutive integers. When a customer arrives, he learns which subset the system occupancy is in. A finer partition means more precise information. (Chapter 2's no- and partial-information scenarios are extreme cases.) In the second information model, each customer's service time consists of a random number of small phases. The server observes the number of phases brought by each customer and keeps track of the total remaining phases in the system. That information is communicated to each arriving potential customer. The mean size of a phase, a continuous parameter, determines the precision of the information. (This model lies between Chapter 2's partial- and full-information scenarios.) The goal of this chapter is to test the conclusions of Chapter 2 in these two new information scenarios.

Some systems actually work in this manner, or nearly so: In a call center, when a customer makes a phone call, he immediately knows whether the system occupancy is 0 or positive (a busy line). This is an example of the partition model. In modern digital communication networks, a message is broken down into small packets which are processed separately. In some job-shop production systems, each customer brings a certain number of identical units to be worked. These are examples of the phase model.

For each information model, we compare two systems, identical except that one has more precise information. In many cases, better information increases throughput and thus benefits the service provider. But this is not always so. The effect depends on the shape of the distribution describing customers' sensitivities to delays. We also study the effects of information on performance as seen by customers. In the partition information model, when more accurate information hurts the server, it must benefit customers. In both models, when information benefits the server, although we cannot ensure that customers' utilities increase, the *ratio* of the average utilities in the two systems is bounded below by the ratio of their idle probabilities.

The reminder of this chapter is organized as follows: Section 1 presents the solution of the partition model and provides comparison results of two systems with different levels of information. Section 2 presents the solution of the phase model and comparison results. Section 3 briefly summarizes the results.

5.1 Partition Information

5.1.1 Model and Solution

In this section, we consider the first model of information described above. The model is specified by a consecutive partition of the nonnegative integers, a partition in which each subset consists of consecutive integers. When a customer arrives, he learns which subset the system is in. Note that Chapter 2's no- and partial-information models are extreme cases.

A consecutive partition can be defined by a subsequence of nonnegative integers starting with 0. Each integer in the subsequence is the first element of the corresponding subset. (If the subsequence is finite, then the last subset in the partition is infinite.) Let $\mathcal{N} = \{\mathcal{N}^k : k = 0, 1, ...K\}$ denote the partition. (K can be finite or ∞.)

Let λ^k denote the arrival rate given $N \in \mathcal{N}^k$, and $\rho^k = \lambda^k/\mu$. (For the moment, these are unknowns.) Given $\rho^k = \rho$,

$$p_{n|k}(\rho) = \Pr\{N = n | N \in \mathcal{N}^k\} = \frac{\rho^n}{\sum_{m \in \mathcal{N}^k} \rho^m} \quad , \quad n \in \mathcal{N}^k.$$

Let $c^k = c^k(\rho) = E[c(W)|N \in \mathcal{N}^k]$. Then,

$$c^k(\rho) = \sum_{n \in \mathcal{N}^k} p_{n|k}(\rho) c_n.$$

An arriving customer seeing $N \in \mathcal{N}^k$ stays precisely when his $\theta \leq \theta^k = 1/c^k$. Thus, ρ^k solves the equation

$$\rho = (\lambda/\mu) H\left(1/c^k(\rho)\right). \tag{5.1}$$

Lemma 34. *ρ^k is unique and decreases in k.*

Proof. $c^k(\rho)$ increases in ρ, so $H\left(1/c^k(\rho)\right)$ decreases, and so (5.1) has a unique solution. Also, for any ρ, $c^k(\rho) \leq c^{k+1}(\rho)$, and thus $\rho^k \geq \rho^{k+1}$. □

This is consistent with both the no- and partial-information models of Chapter 2.

This solution depends only on $\mathcal{N}^k$; the calculation for each subset is independent of the others. Having determined λ^k for each k, the arrival rate in state n becomes $\lambda_n = \lambda^k$, $n \in \mathcal{N}^k$. Now we have a birth-death process, which can be analyzed in the standard way.

5.1.2 Comparison

Consider two consecutive partitions, one a refinement of the other. The refined partition provides customers with more information. Any such refinement can be constructed by a sequence of simple refinements, each of which splits a single subset into two. Let's focus on such a simple refinement. Starting with a partition $\mathcal{N}$, we construct a new one $\overline{\mathcal{N}}$ by splitting one subset, say $\mathcal{N}^{ko}$, into the two subsets $\mathcal{N}^{k-}$ and $\mathcal{N}^{k+}$, where the numbers in $\mathcal{N}^{k-}$ are smaller than those in $\mathcal{N}^{k+}$. Let p_n denote the steady-state probabilities for the original partition and $\bar{p}_n$ those for the refined one. Define

$$p^k = \sum_{n \in \mathcal{N}^k} p_n = \Pr\{N \in \mathcal{N}^k\},$$

and define $\bar{p}^k$ similarly. Also, set $\bar{p}^{ko} = \bar{p}^{k-} + \bar{p}^{k+}$.

Lemma 35. $\rho^{k-} \geq \rho^{ko} \geq \rho^{k+}$.

Proof. $c^{ko}(\rho^{ko})$ is a weighted average of $c^{k-}(\rho^{ko})$ and $c^{k+}(\rho^{ko})$, and

$$c^{k-}(\rho^{ko}) \leq c^{ko}(\rho^{ko}) \leq c^{k+}(\rho^{ko}).$$

Therefore,

$$\begin{aligned} \rho^{ko} &\leq H\left[1/c^{k-}(\rho^{ko})\right] \\ \rho^{ko} &\geq H\left[1/c^{k+}(\rho^{ko})\right]. \end{aligned}$$

The assertion thus follows by (5.1). □

Let n^k denote the first element of subset $\mathcal{N}^k$. Consider what happens to the ratio $\eta_n = \bar{p}_n/p_n$. For $n \leq n^{k-}$, $\bar{p}_n$ and p_n change by the same rates, so η_n remains constant at η_0. For $n^{k-} < n \leq n^{k+}$, since $\rho^{k-} \geq \rho^{ko}$, η_n increases. Likewise, for $n^{k+} < n \leq n^{ko+1}$, η_n decreases. Finally, for $n \geq n^{ko+1}$, η_n remains constant at $\eta_{n^{ko+1}}$. Consequently, the sequence η_n is unimodal. Its largest value is at $n = n^{k+}$.

Lemma 36. *If $\eta_0 \geq 1$, then $\bar{N} \preceq_{st} N$.*

Proof. In this case, $\eta_0 \geq 1$ for all $n \leq n^{k+}$. Normalization then requires that $\eta_0 \leq 1$ for all $n \geq n^{ko+1}$. Thus, $\bar{p}_n$ crosses p_n just once, from above. □

We can write the average arrival rates for the original system and the refined one as follows:

$$\begin{aligned} E[\lambda^K] &= \sum_{k \neq ko} p^k \lambda H(\theta^k) + p^{ko} \lambda H(\theta^{ko}) \\ E[\bar{\lambda}^K] &= \sum_{k \neq ko} \bar{p}^k \lambda H(\theta^k) + \bar{p}^{k-} \lambda H(\theta^{k-}) + \bar{p}^{k+} \lambda H(\theta^{k+}). \end{aligned}$$

The following is a sufficient condition to compare the idle probabilities:

Lemma 37. *If $\bar{p}^{k-} H(\theta^{k-}) + \bar{p}^{k+} H(\theta^{k+}) > \bar{p}^{ko} H(\theta^{ko})$, then $p_0 > \bar{p}_0$.*

Proof. Suppose $p_0 \leq \bar{p}_0$, then $\bar{N} \preceq_{st} N$ by Lemma 36. Hence, The condition $\bar{p}^{k-} H(\theta^{k-}) + \bar{p}^{k+} H(\theta^{k+}) > \bar{p}^{ko} H(\theta^{ko})$ implies that

$$\begin{aligned} E[\bar{\lambda}^K] &= \sum_{k \neq ko} \bar{p}^k \lambda H(\theta^k) + \bar{p}^{k-} \lambda H(\theta^{k-}) + \bar{p}^{k+} \lambda H(\theta^{k+}) \\ &> \sum_{k \neq ko} \bar{p}^k \lambda H(\theta^k) + \bar{p}^{ko} \lambda H(\theta^{ko}) \\ &\geq \sum_{k \neq ko} p^k \lambda H(\theta^k) + p^{ko} \lambda H(\theta^{ko}) = E[\lambda^K]. \end{aligned}$$

The second inequality follows from the fact that $H(\theta^k)$ is decreasing in k and $\bar{N} \preceq_{st} N$. Therefore,

$$\begin{aligned} p_0 &= 1 - E[\lambda^K]/\mu \\ &> 1 - E[\bar{\lambda}^K]/\mu = \bar{p}_0, \end{aligned}$$

a contradiction. □

Thus, as long as the refined system has a larger average arrival rate on the split subset, it has larger throughput overall.

Next, let's compare the average utilities. For the original partition and the refined one, these are, respectively,

$$\begin{aligned} u &= \sum_{k \neq ko} p^k J(\theta^k) + p^{ko} J(\theta^{ko}) \\ \bar{u} &= \sum_{k \neq ko} \bar{p}^k J(\theta^k) + \bar{p}^{k-} J(\theta^{k-}) + \bar{p}^{k+} J(\theta^{k+}). \end{aligned}$$

Here is a sufficient condition:

Lemma 38. *If*

$$\bar{p}^{k-} J(\theta^{k-}) + \bar{p}^{k+} J(\theta^{k+}) \geq \bar{p}^{ko} J(\theta^{ko}),$$

then $\eta_0 \geq 1$ *implies* $\bar{u} \geq u$.

Proof. Note that $J(\theta^k)$ is decreasing in k. The proof is similar to that of Lemma 37. □

5.1.3 A Special Case

The comparison results above are expressed as sufficient conditions that are not easily checked in terms of the original data. To do that, we need to focus on a special case. There is one particular nonnegative integer k. When a customer arrives and $N < k$, he learns the exact value of N. When $N \geq k$, however, the customer learns only that fact, i.e., $N \geq k$. The information sets here are $\mathcal{N}^l = \{l\}$ for $l = 0, 1, 2, ..., k-1$ and $\mathcal{N}^{ko} = \{k, k+1, k+2, ...\}$.

This scenario is realistic in many settings. In some cases, the service provider may perceive an incentive to avoid bad news; in other cases, there may be technical restrictions. For example, in the call center mentioned earlier, an arriving customer immediately knows whether N is 0 or positive.

Now, consider a refined system, where $\mathcal{N}^{ko}$ is decomposed into $\mathcal{N}^{k-} = \{k\}$ and $\mathcal{N}^{k+} = \{k+1, k+2, ...\}$.

The average cost on the set $\mathcal{N}^{ko}$, c^{ko}, can be expressed as

$$\begin{aligned} c^{ko} &= \frac{\rho^{ko} c_k + \left(\rho^{ko}\right)^2 c_{k+1} + \left(\rho^{ko}\right)^3 c_{k+2} + ...}{\rho^{ko} + \left(\rho^{ko}\right)^2 + \left(\rho^{ko}\right)^3 + ...} \\ &= \frac{c_k + \rho^{ko} c_{k+1} + \left(\rho^{ko}\right)^2 c_{k+2} + ...}{1 + \rho^{ko} + \left(\rho^{ko}\right)^2 + ...} \qquad (5.2) \end{aligned}$$

c^{k-} and c^{k+} are defined similarly. Define $\alpha = \bar{p}^{k-}/\bar{p}^{ko}$.

Lemma 39. $\alpha c^{k-} + (1-\alpha) c^{k+} \leq c^{ko}$.

Proof. For the refined system, we have the balance equations:

$$\bar{p}_{k+1} = \bar{p}_k \rho^{k-}$$

and

$$\bar{p}_{k+i} = \bar{p}_{k+i-1} \rho^{k+} = \bar{p}_k \rho^{k-} (\rho^{k+})^{i-1}, \ i = 2, 3, ...$$

Hence,

$$\alpha = \frac{\bar{p}^{k-}}{\bar{p}^{ko}} = \frac{\rho^{k-}}{\rho^{k-} + \rho^{k-}[\rho^{k+} + (\rho^{k+})^2 + ...]}.$$

So

$$\begin{aligned} \alpha c^{k-} + (1-\alpha) c^{k+} &= \alpha c^{k-} + (1-\alpha) \frac{\rho^{k+} c_{k+1} + \left(\rho^{k+}\right)^2 c_{k+2} + \left(\rho^{k+}\right)^3 c_{k+3} + ...}{\rho^{k+} + \left(\rho^{k+}\right)^2 + \left(\rho^{k+}\right)^3 + ...} \\ &= \frac{\rho^{k-} c_k + \rho^{k-} \rho^{k+} c_{k+1} + \rho^{k-} (\rho^{k+})^2 c_{k+2} + ...}{\rho^{k-} + \rho^{k-} \rho^{k+} + \rho^{k-} (\rho^{k+})^2 + ...} \\ &= \frac{c_k + \rho^{k+} c_{k+1} + (\rho^{k+})^2 c_{k+2} + ...}{1 + \rho^{k+} + (\rho^{k+})^2 + ...} \qquad (5.3) \end{aligned}$$

We have $\rho^{ko} \geq \rho^{k+}$ by Lemma 35. Comparing (5.2) and (5.3), the conclusion follows.

□

We obtain several main conclusions, which are consistent with those of Chapter 2:

Proposition 40. *If $H(1/x)$ is strictly convex in x, then $p_0 > \bar{p}_0$.*

Proof.

$$\begin{aligned} \alpha H(1/c^{k-}) + (1-\alpha)H(1/c^{k+}) &> H\left(\frac{1}{\alpha c^{k-} + (1-\alpha)c^{k+}}\right) \\ &\geq H(1/c^{ko}) \end{aligned}$$

Hence, $\bar{p}_0 < p_0$ by Lemma 37. □

This convexity condition, which characterizes the shape of H, is fundamental also in Chapter 2. It means that customers are heterogeneous in a certain sense.

Proposition 41. *If $\bar{p}_0 \geq p_0$, then $\bar{u} \geq u$.*

Proof. Since $J(1/x)$ is convex in x,

$$\begin{aligned} \alpha J(1/c^{k-}) + (1-\alpha)J(1/c^{k+}) &\geq J\left(\frac{1}{\alpha c^{k-} + (1-\alpha)c^{k+}}\right) \\ &\geq J(1/c^{ko}) = J(\theta^{ko}). \end{aligned}$$

By Lemma 38, it follows that $\bar{u} \geq u$. □

Hence, when information hurts the server, it benefits customers.

Proposition 42. *If $H(1/x)$ is strictly convex in x,*

$$\frac{\bar{u}}{u} > \frac{\bar{p}_0}{p_0}.$$

Proof. If $H(1/x)$ is strictly convex in x, then $p_0 > \bar{p}_0$ according to Proposition 40. The condition $\bar{p}_0 < p_0$ implies that

$$\rho^{k-} + \rho^{k-}[\rho^{k+} + (\rho^{k+})^2 + (\rho^{k+})^3 + ...] > \rho^{ko} + (\rho^{ko})^2 + (\rho^{ko})^3 + ...$$

Denote $\rho_i = \lambda_i/\mu$. We have

$$\begin{aligned}
\frac{\bar{u}}{\bar{p}_0} &= J(1/c_0) + \rho_0 J(1/c_1) + \left(\prod_{i=0}^{1}\rho_i\right) J(1/c_2) + ... + \left(\prod_{i=0}^{k-2}\rho_i\right) J(1/c_{k-1}) \\
&\quad + \left(\prod_{i=0}^{k-2}\rho_i\right) \left[\rho^{k-} J(1/c^{k-}) + \rho^{k-}[\rho^{k+} + (\rho^{k+})^2 + (\rho^{k+})^3 + ...] J(1/c^{k+})\right] \\
&= J(1/c_0) + \rho_0 J(1/c_1) + \left(\prod_{i=0}^{1}\rho_i\right) J(1/c_2) + ... + \left(\prod_{i=0}^{k-2}\rho_i\right) J(1/c_{k-1}) \\
&\quad + \left(\prod_{i=0}^{k-2}\rho_i\right) \left[\rho^{k-} + \rho^{k-}[\rho^{k+} + (\rho^{k+})^2 + ...]\right] \left[\alpha J(1/c^{k-}) + (1-\alpha) J(1/c^{k+})\right] \\
&> J(1/c_0) + \rho_0 J(1/c_1) + \left(\prod_{i=0}^{1}\rho_i\right) J(1/c_2) + ... + \left(\prod_{i=0}^{k-2}\rho_i\right) J(1/c_{k-1}) \\
&\quad + \left(\prod_{i=0}^{k-2}\rho_i\right) \left[\rho^{ko} + (\rho^{ko})^2 + ...\right] J\left(\frac{1}{\alpha c^{k-} + (1-\alpha) c^{k+}}\right) \\
&= J(1/c_0) + \rho_0 J(1/c_1) + \left(\prod_{i=0}^{1}\rho_i\right) J(1/c_2) + ... + \left(\prod_{i=0}^{k-2}\rho_i\right) J(1/c_{k-1}) \\
&\quad + \left(\prod_{i=0}^{k-2}\rho_i\right) \left[\rho^{ko} + (\rho^{ko})^2 + (\rho^{ko})^3 + ...\right] J(1/c^{ko}) \\
&= \frac{u}{p_0}.
\end{aligned}$$

□

In this case, it still may not be true that the refined system has higher utility than the original system. But at least its utility cannot be too much lower.

Chapter 2 also shows that, if $J \circ H^{-1}$ is convex, then the average utility under partial information is indeed larger than that under no information. However, the corresponding result here need not hold. Here is a counterexample: Assume H is a beta distribution with parameter (α, β). Chapter 2 shows that $J \circ H^{-1}$ is convex, if and only if $\beta \geq 1$. Set $\alpha = 2, \beta = 8, \mu = 2, \lambda = 4$ and $c(w) = 1 + w$. Table 5.1 shows that the average utility first increases with k then decreases. Hence, more information can decrease utility, even when $J \circ H^{-1}$ is convex.

Table 5.1: Compare Average Utility with Beta H

k	1	3	5	7	9	11
u	0.2760	0.2836	0.2840	0.2830	0.2821	0.2814

5.2 Phase Information

5.2.1 Assumptions and Notation

In this section, we consider another model of information. This one lies between Chapter 2's partial and full information. Assume each customer's service time consists of a random number K of phases. The times of these phases are i.i.d. random variables with the exponential distribution of rate ν. Thus, conditional on K, the service time has the Erlang distribution with parameters (K, ν). Suppose that K has a geometric distribution with parameter q, i.e., $\Pr\{K = k\} = (1-q)q^{k-1},\ k > 0$. Then, the unconditional service time has the exponential distribution with rate $\mu = (1-q)\nu$.

When a customer arrives, the provider observes his K but not the times of the individual phases. The provider keeps track of the *total* remaining number of phases of all remaining customers. Call this M. When a new customer arrives, the provider tells him M. That is more information than the system occupancy N but less than the workload V. Given $M = m$, the customer assesses the distribution of waiting time W. The expected basic waiting cost is $c_m = E[c(W)|M = m]$, a function of that information. Hence, given $M = m$, the probability that an arriving customer stays is $H(1/c_m)$, and so the effective arrival rate is

$$\lambda_m = \lambda H(1/c_m). \tag{5.4}$$

The process M can be modeled as a continuous-time Markov chain.

Note that different q and ν can give the same μ. For fixed μ, a larger ν and q means the server can differentiate customers' service times into smaller phases. That in turn means more information.

5.2.2 Model and Solution

Balance equation

This system is like a queue with state-dependent batch or bulk arrivals. The process M is a continuous-time Markov chain with state space $\{0, 1, 2, ...\}$. Let $p_m = \Pr\{M = m\}$. If we cut between state m and $m+1$, we obtain the global balance equation:

$$\nu p_{m+1} = \sum_{k=0}^{m} \lambda_k p_k \Pr\{K \geq m+1-k\}, \; m \geq 0, \tag{5.5}$$

where λ_k is defined in (5.4).

For $m = 0$,

$$\nu p_1 = \lambda_0 p_0.$$

For $m \geq 1$,

$$\begin{aligned} \nu p_{m+1} &= \sum_{k=0}^{m} \lambda_k p_k \Pr\{K \geq m+1-k\} \\ &= \sum_{k=0}^{m} \lambda_k p_k q^{m-k} \\ &= q \sum_{k=0}^{m-1} \lambda_k p_k q^{m-1-k} + \lambda_m p_m \\ &= (q\nu + \lambda_m) p_m. \end{aligned} \tag{5.6}$$

These equations can be solved just like those of a birth-death process. Denote $\xi_0 = \lambda_0/\nu$ and $\xi_m = q + \lambda_m/\nu$, $m > 0$. Then,

$$p_n = \Theta_n p_0, \ n > 0,$$

where

$$\Theta_n = \prod_{m=0}^{n-1} \xi_m.$$

Thus,

$$p_0 = \frac{1}{1+\Theta},$$

where

$$\Theta = \sum_{n>0} \Theta_n. \tag{5.7}$$

Closed-form solutions for a special case

For linear cost $c(w) = 1 + w$ and uniform H, $\lambda_k = \lambda/(1 + k/\nu)$, and we can obtain closed-form expressions for the system performance measures. See the appendix for the derivations.

The idle probability for the system, p_0, can be expressed as

$$p_0 = \left\{ 1 + \frac{\lambda B\left(q; \nu, \lambda/q + 1\right)}{q^{\nu}\left(1-q\right)^{\lambda/q+1}} \right\}^{-1},$$

where $B(x; a, b) = \int_0^x t^{a-1}(1-t)^{b-1}dt, a > 0, b > 0$ is the incomplete beta function.

The Laplace-Stieltjes transform of the stationary distribution of the waiting time W can be expressed as

$$\tilde{W}(s) = A(s)/A(0),$$

where

$$A(s) = 1 + \frac{\frac{\lambda}{1+s/\nu} B\left(\frac{q}{1+s/\nu}; \nu, \lambda/q + 1\right)}{\left(\frac{q}{1+s/\nu}\right)^{\nu}\left(1 - \frac{q}{1+s/\nu}\right)^{\lambda/q+1}}.$$

The mean waiting time can be expressed as

$$E[W] = \frac{\lambda + (1-\mu)(1-p_0)}{\mu}. \tag{5.8}$$

(5.8) indicates that $E[W]$ is not always decreasing in p_0. This depends on the relationship between μ and 1. A similar conclusion is obtained in Chapter 2.

5.2.3 Comparison

In this section, we compare the performance of two systems with different levels of information, an original system as above, and a refined system with larger q and ν but the same μ. The refined system can differentiate customers' service times into smaller pieces and hence provide customers with more precise information. The refined system can be constructed by decomposing each phase in the original system into the sum of a geometric number of smaller phases. Denote this geometric random variable by L and its parameter by ζ. Let $\tilde{\nu}$ and $\tilde{q}$ be the refined system's parameters. Then,

$$\tilde{\nu} = \nu/(1-\zeta), \quad \tilde{q} = \zeta + q(1-\zeta).$$

Note that $(1-\tilde{q})\tilde{\nu} = (1-q)\nu = \mu$.

Let $\tilde{M}$ be the total number of small phases in the refined system. Define the other corresponding parameters as follows: $\tilde{c}_m = E[c(\tilde{W})|\tilde{M} = m]$, $\tilde{\lambda}_m = \lambda H(1/\tilde{c}_m)$, $\tilde{\xi}_0 = \tilde{\lambda}_0/\tilde{\nu}$ and $\tilde{\xi}_m = \tilde{q} + \tilde{\lambda}_m/\tilde{\nu}$, $m > 0$. Note that $\tilde{c}_0 = c_0$ and $\tilde{\lambda}_0 = \lambda_0$. Also, define $\hat{\xi}_0 = \tilde{\lambda}_0/\nu$ and $\hat{\xi}_m = q + \tilde{\lambda}_m/\nu$, $m > 0$.

The following subsections compare the throughput and average utility for the two systems.

Throughput

We start with two lemmas. Denote the m-fold convolution of L by $L^{(m)}$.

Lemma 43. *The idle probability in the refined system can be expressed as*

$$\tilde{p}_0 = \frac{1}{1+\tilde{\Theta}},$$

where

$$\tilde{\Theta} = \sum_{n>0} \tilde{\Theta}_n$$

$$\tilde{\Theta}_n = E[\hat{\xi}_0 \hat{\xi}_{L^{(1)}} \cdots \hat{\xi}_{L^{(n-1)}}],$$

Proof. By (5.7),

$$\begin{aligned}
\tilde{\Theta} &= \sum_{n>0} \prod_{m=0}^{n-1} \tilde{\xi}_m \\
&= \left(\tilde{\lambda}_0/\tilde{\nu}\right) \left\{ 1 + \sum_{n\geq 1} \prod_{m=1}^{n} (\tilde{q} + \tilde{\lambda}_m/\tilde{\nu}) \right\} \\
&= [(\lambda_0/\nu)(1-\zeta)] \left\{ 1 + \sum_{n\geq 1} \prod_{m=1}^{n} [\zeta + (1-\zeta)q + (1-\zeta)\tilde{\lambda}_m/\nu] \right\} \\
&= \hat{\xi}_0 \left\{ (1-\zeta) + (1-\zeta) \sum_{n\geq 1} \prod_{m=1}^{n} [\zeta + (1-\zeta)\hat{\xi}_m] \right\}.
\end{aligned}$$

Expand the expression in brackets. First, consider terms which include none of the $\hat{\xi}_i$. The sum of such terms is 1.

Second, consider terms which include exactly one $\hat{\xi}_i$. The coefficient of $\hat{\xi}_i$ is $(1-\zeta)\zeta^{i-1}$. This is $\Pr\{L = i\}$. Hence, the sum of all such terms is

$$\sum_{i\geq 1} (1-\zeta)\zeta^{i-1}\hat{\xi}_i = E[\hat{\xi}_{L^{(1)}}].$$

Third, consider terms which include the product of two different $\hat{\xi}_i$. The coefficient of $\hat{\xi}_i\hat{\xi}_j$ with $i < j$ can be shown to equal $(1-\zeta)^3(\zeta^{j-2}+\zeta^{j-1}+\zeta^j+\cdots) = (1-\zeta)^2\zeta^{j-2}$.

Moreover, for two independent random variables L_1 and L_2 with the same distribution as L,

$$\begin{aligned}
\Pr\{L_1 = i, L_2 = j-i\} &= \Pr\{L_1 = i\}\Pr\{L_2 = j-i\} \\
&= (1-\zeta)\zeta^{i-1}(1-\zeta)\zeta^{j-i-1} \\
&= (1-\zeta)^2\zeta^{j-2}.
\end{aligned}$$

Hence, the sum of all such terms is $E[\hat{\xi}_{L^{(1)}}\hat{\xi}_{L^{(2)}}]$.

Similarly, one can show that the coefficient for the product of n different $\hat{\xi}_i$ can be expressed as the joint probability of $\{L_1, L_2, L_3, ..., L_n\}$ where $L_j, j = 1, ..., n$ are i.i.d. random variables with the same distribution as L. Hence, the sum of these terms is $E[\hat{\xi}_{L^{(1)}} \cdots \hat{\xi}_{L^{(n)}}]$. □

In Chapter 2, for full information, an expression analogous to $\tilde{\Theta}_n$ equals

$$(\lambda/\mu)^n E[H(1/c(S^0)H(1/c(S^1)\cdots H(1/c(S^{n-1})].$$

This is a limiting case of this lemma.

Note that the random variable $\left[\tilde{W}|\tilde{M} = L^{(m)}\right]$ equals $[W|M = m]$. Therefore, $E[\tilde{c}_{L^{(m)}}] = c_m$.

Lemma 44. *If $H(1/x)$ is strictly convex, $E[\hat{\xi}_{L^{(n)}}] > \xi_n$, $n > 0$.*

Proof. Since $H(1/x)$ is convex in x, for $n > 0$,

$$\begin{aligned}
E[\hat{\xi}_{L^{(n)}}] &= q + \lambda E[H(1/\tilde{c}_{L^{(n)}})]/\nu \\
&> q + \lambda H(1/E[\tilde{c}_{L^{(n)}}])/\nu \\
&= q + \lambda H(1/c_n)/\nu \\
&= q + \lambda_n/\nu = \xi_n.
\end{aligned}$$

□

Using this lemma, we have

Proposition 45. *If $H(1/x)$ is strictly convex, $\tilde{p}_0 < p_0$.*

Proof.

$$\begin{aligned}\tilde{\Theta}_n &= E[\hat{\xi}_0\hat{\xi}_{L^{(1)}}\cdots\hat{\xi}_{L^{(n-1)}}] \\ &\geq E[\hat{\xi}_0]E[\hat{\xi}_{L^{(1)}}]\cdots E[\hat{\xi}_{L^{(n-1)}}] \\ &> \xi_0\xi_1\cdots\xi_{n-1} = \Theta_n.\end{aligned}$$

The first inequality follows from the positive correlation between $L^{(i)}$ and $L^{(j)}$; the second inequality follows from Lemma 44. □

Utility

Denote $J_n = J(1/c_n)$. In the original system,

$$u = \sum_{n\geq 0} p_n J_n = p_0 \left\{ J_0 + \sum_{n\geq 0}\prod_{m=0}^{n} \xi_m J_{n+1} \right\}. \tag{5.9}$$

Similarly, define $\tilde{J}_n = J(1/\tilde{c}_n)$ for the refined system. The average utility in the refined system $\tilde{u}$ can be expressed as follows.

Lemma 46.

$$\tilde{u} = \tilde{p}_0 \left\{ \tilde{J}_0 + \sum_{n\geq 0} E[\hat{\xi}_0\hat{\xi}_{L^{(1)}}\cdots\hat{\xi}_{L^{(n)}}\tilde{J}_{L^{(n+1)}}] \right\} \tag{5.10}$$

Proof.

$$\begin{aligned}\tilde{u} &= \tilde{p}_0 \left\{ \tilde{J}_0 + \sum_{n\geq 0}\prod_{m=0}^{n} \tilde{\xi}_m \tilde{J}_{n+1} \right\} \\ &= \tilde{p}_0 \left\{ \tilde{J}_0 + \hat{\xi}_0(1-\zeta)\left[\tilde{J}_1 + \sum_{n\geq 1}\prod_{m=1}^{n} [\zeta + (1-\zeta)\hat{\xi}_m]\tilde{J}_{n+1} \right] \right\}\end{aligned}$$

Expand the expression in brackets as in the proof of Lemma 43. The result then follows. □

Then we have the following proposition.

Proposition 47. *If $H(1/x)$ is strictly convex,*

$$\frac{\tilde{u}}{u} > \frac{\tilde{p}_0}{p_0}.$$

Proof. Due to the convexity of $H(1/x)$ and $J(1/x)$, by Jensen's inequality we have $E[\hat{\xi}_{L^{(n)}}] \geq \xi_n$ and $E[\tilde{J}_{L^{(n+1)}}] \geq J_{n+1}$. Hence,

$$\begin{aligned} E[\hat{\xi}_0 \hat{\xi}_{L^{(1)}} \cdots \hat{\xi}_{L^{(n)}} \tilde{J}_{L^{(n+1)}}] &\geq E[\hat{\xi}_0] E[\hat{\xi}_{L^{(1)}}] \cdots E[\hat{\xi}_{L^{(n)}}] E[\tilde{J}_{L^{(n+1)}}] \\ &> \xi_0 \xi_1 \cdots \xi_n J_{n+1}. \end{aligned}$$

The first inequality follows from the positive correlation between $L^{(i)}$ and $L^{(j)}$ and the fact that both $\hat{\xi}$ and $\tilde{J}$ are monotone functions in the same direction. By (5.9) and (5.10), we get $u/p_0 \leq \tilde{u}/\tilde{p}_0$. □

Next, we give two counterexamples: The first one shows that, even when $J \circ H^{-1}$ is convex, the average utility need not increase with information; the second one shows that $\tilde{p}_0 \geq p_0$ need not imply $\tilde{u} \geq u$. In both examples, we assume H is a beta distribution with parameter (α, β) and we set $\alpha = 2, \beta = 8$. We assume $\mu = 2$ and a linear cost function $c(w) = 1 + w$. $\lambda = 4$ in the first example, and $\lambda = 0.5$ in the second. Table 5.2 shows that the average utility decreases with ν when $\beta > 1$. Hence, more information can decrease the average utility, even though $J \circ H^{-1}$ is convex. Table 5.3 shows that the average utility can decrease, even when the idle probability increases. Hence, more information can decrease the average utility and the throughput simultaneously. Thus, the effects of information here are less straightforward than in the systems of Chapter 2.

Table 5.2: Compare Average Utility with Beta H and $\lambda = 4$

ν	2	4	6	8	10	12	14
u	0.2477	0.2314	0.2262	0.2237	0.2222	0.2212	0.2205

Table 5.3: Compare Average Utility and Idle Probability with Beta H and $\lambda = 0.5$

ν	2	4	6	8	10	12	14
p_0	0.7502	0.7505	0.7506	0.7506	0.7506	0.7507	0.7507
u	0.8265	0.8165	0.8132	0.8116	0.8106	0.8100	0.8095

5.3 Conclusion

In this chapter, we consider two stylized models of information about service delays, partition information and phase information. In the first model, we consider that the arriving customer can learn the information on the range of system occupancy. We derived the effective arrival rate for each information range. In the second model, we assume that each customer's service time consists of a geometric random sum of phases which have independent and identical exponential distributions. The server informs the incoming customer the total number of phases in the system. We model the process of this number as a continuous-time Markov chain and we derive the balance equations for it. For a special case with uniform customers and a linear cost function, we derive closed-form expressions for key system performance measures.

Under each information model, we compare two systems, one with sharper information than the other. We show that for both information models, providing customers more accurate delay information may benefit the server, but it may not. This effect depends on the shape of the distribution function of customers' delay-sensitivities. We also show that, in the partition information model, when more accurate information hurts the server, it must benefit customers. In both information models, when information benefits the server, the ratio of the average utility of more-information system to that of less-information system is greater than the ratio of their idle probabilities.

Chapter 6

Conclusion

In this dissertation, we start our analysis from an $\cdot/M/1$ queue with three levels of delay information. Customers use that information to determine their expected waiting costs, and so to decide whether to stay and receive service or leave (balk). We obtained closed-form solutions for some cases and nearly closed-form solutions for others. In comparing these systems, we found that the form of the cost-scale distribution plays a crucial role. For one important class, average utility is proportional to throughput; so the provider's and customers' objectives coincide; those measures improve as information increases. More broadly, we found sufficient conditions to ensure that more information helps the provider or the customers. In other cases, however, more information can actually hurt one or the other. We also carry out the sensitivity analysis of cost function and the cost-scale distribution. We consider first-order and second-order stochastic conditions and we obtain some comparative-statics results. In the latter chapters, we extend our analysis in two directions: In the one direction, we extend the three information models to the systems with general service times. In the other, we consider richer models of information. Our generalization shows that most of previous conclusions still hold.

These perverse phenomena occur mainly in extreme conditions, however. Information can make a bad system worse. It would be worth investigating a larger model, where capacity is a decision variable, and/or there are other levers, like prices, to manage demand. We suspect that the strange behavior seen here would be less marked in a broader setting.

The utility-based approach forces us to revise our notions of good performance. Most peculiar is the concept that customers actually may prefer one system to another when its probability of delay and average delay are larger. Of course, this does not mean that the customers want to wait. Rather, it shows that these standard measures do not capture everything that matters to customers.

Numerous extensions are worth pursuing, for example, alternative queue disciplines, inventory, etc. It would be interesting also to explore various pricing schemes, following the lead of Naor.

Information can modulate waiting costs in subtler ways than our model envisions. For example, given a delay estimate, a call-center customer may turn attention to other tasks while waiting. In general, information affects people's expectations, and those expectations affect the overall experience of waiting. (Carmon et al. 1995 pose a framework for such effects.) It would be interesting to study how delay information is acquired and used in various situations and the resulting effects on overall system behavior.

.1 Appendix for Chapter 2

.1.1 No Information

Assume that $E[c(W|-)]$ is finite for any $\lambda_- < \mu$.

Proposition 48. *For no information, there exists a unique equilibrium arrival rate* λ_-*.*

Proof. By assumption, c is increasing and continuous. Thus, $\lambda/E[c(W|-)]$ is a decreasing, continuous function of λ_- mapping the interval $[0, \min\{\lambda, \mu\}]$ into itself. Thus, (3.1) has a unique solution. $\square$

.1.2 Partial Information

Verification of formula (2.4) for uniform H and linear c: Observe that

$$\begin{aligned}
\frac{d}{d\lambda}\gamma(\mu,\lambda) &= \lambda^{\mu-1}e^{-\lambda} = \frac{d}{d\lambda}\left(\lambda^{\mu-1}e^{-\lambda}\sum_{n=1}^{\infty}\frac{\Gamma(\mu)}{\Gamma(\mu+n)}\lambda^n\right) \\
&= (\mu-1-\lambda)(\lambda^{\mu-2}e^{-\lambda})\sum_{n=1}^{\infty}\frac{\Gamma(\mu)}{\Gamma(\mu+n)}\lambda^n + (\lambda^{\mu-1}e^{-\lambda})\sum_{n=1}^{\infty}\frac{\Gamma(\mu)}{\Gamma(\mu+n)}n\lambda^{n-1} \\
&= (\mu-1-\lambda)\lambda^{\mu-2}e^{-\lambda}\Theta + (\lambda^{\mu-2}e^{-\lambda}/p_0)E\left[N\right].
\end{aligned}$$

Thus,

$$\begin{aligned}
\lambda &= (\mu-1-\lambda)\Theta + E\left[N\right]/p_0 \\
&= (\mu-1-\lambda)(1-p_0)/p_0 + E\left[N\right]/p_0,
\end{aligned}$$

or

$$E\left[N\right] = \lambda - (\mu-1)\left(1-p_0\right).$$

.1.3 Cost-Scale Distributions

We verify the assertions about beta distributions. One can directly compute

$$-\frac{\theta h'(\theta)}{h(\theta)} = -(\alpha - 1) + (\beta - 1)\frac{\theta}{1-\theta}.$$

Condition 1 stipulates that this quantity be no more than 2 for all θ. This is clearly true, if and only if $\beta \leq 1$.

The other condition is more intricate. In general,

$$(J \circ H^{-1})''(\phi) = \frac{\phi h^2(\phi) - [2h(\phi) + \phi h'(\phi)][H(\phi) - J(\phi)]}{\phi^2 h^3(\phi)}.$$

Condition 2 means that the numerator be nonnegative. For a beta distribution, the numerator can be written as

$$\frac{\phi^{\alpha-2}(1-\phi)^{\beta-2}}{B(\alpha,\beta)}\left\{\frac{\phi^{\alpha+1}(1-\phi)^{\beta}}{B(\alpha,\beta)} - [(\alpha+1) - (\alpha+\beta)\phi]\frac{\alpha}{\alpha+\beta}H(\phi;\alpha+1,\beta)\right\},$$

where $H(\phi;\alpha+1,\beta)$ denotes the cdf but with parameter $\alpha+1$ instead of α. Using a standard series representation of H, this becomes

$$\frac{\phi^{2\alpha-1}(1-\phi)^{2\beta-2}}{(\alpha+1)B^2(\alpha,\beta)}\left\{(\alpha+1) - [(\alpha+1) - (\alpha+\beta)\phi]\left[1 + \frac{B(\alpha+2,1)}{B(\alpha+\beta+1,1)}\phi + ...\right]\right\}.$$

For $\beta = 1$

$$1 + \frac{B(\alpha+2,1)}{B(\alpha+\beta+1,1)}\phi + ... = \frac{1}{1-\phi}$$

and

$$[(\alpha+1) - (\alpha+\beta)\phi] = (\alpha+1)(1-\phi),$$

so the numerator reduces to 0. For $\beta < 1$ each coefficient in the power series above is > 1, so

$$1 + \frac{B(\alpha+2,1)}{B(\alpha+\beta+1,1)}\phi + ... > \frac{1}{1-\phi}.$$

Also,

$$[(\alpha+1)-(\alpha+\beta)\phi] > (\alpha+1)(1-\phi).$$

Thus, $(J \circ H^{-1})''(\phi) < 0$. Similarly, for $\beta > 1$, $(J \circ H^{-1})''(\phi) > 0$. Thus, the condition holds for $\beta \geq 1$ but not for $\beta < 1$.

The function J plays an important role in the following. Observe that, since H is increasing, so is J, and hence $J(1/x)$ is decreasing in x. Also,

$$\begin{aligned} \frac{d^2}{dx^2}J(1/x) &= -(1/x^2)H(1/x) + (1/x^2)H(1/x) + (1/x^3)h(1/x) \\ &= (1/x^3)h(1/x) > 0. \end{aligned}$$

Thus, $J(1/x)$ is strictly convex in $x \geq 1$.

.1.4 No Information and Partial Information

We start with two preliminary results.

Lemma 49. *p_n^{part} is log-concave in n.*

Proof.

$$\frac{p_{n+1}^{part}}{p_n^{part}} = (\lambda/\mu)H(\theta_n),$$

which is decreasing in n. □

This implies that $\bar{P}_n^{part}$ (the complementary cdf of N^{part}) is also log-concave. (See, e.g., Karlin 1968.) This property means precisely that N^{part} has increasing failure rate. Also, since p_n^{no} and $\bar{P}_n^{no}$ are geometric (log-linear) sequences, the ratios p_n^{part}/p_n^{no} and $\bar{P}_n^{part}/\bar{P}_n^{no}$ are also log-concave. In particular, these ratios are unimodal.

Lemma 50. *If $p_0^{part} \geq p_0^{no}$, then $N^{part} \preceq_{st} N^{no}$.*

Proof. The sequence p_n^{part}/p_n^{no} is log-concave and hence unimodal. If it starts above 1, then it must cross 1 exactly once, by normalization. It follows that $\bar{P}_n^{part} \leq \bar{P}_n^{no}$ for all n. That is, $N^{part} \preceq_{st} N^{no}$. □

Now we prove the first main result.

Proposition 51. *If $p_0^{part} \geq p_0^{no}$, then $u^{part} > u^{no}$.*

Proof. In this case, $N^{part} \preceq_{st} N^{no}$. Since c_n is increasing, $E[c_{N^{part}}] \leq E[c_{N^{no}}]$. Since $J(1/x)$ is decreasing and strictly convex, by Jensen's inequality,

$$\begin{aligned} u^{no} &= J(\theta_-) = J(1/E[c_{N^{no}}]) \\ &\leq J(1/E[c_{N^{part}}]) \\ &< E[J(1/c_{N^{part}})] \\ &= E[J(\theta_{N^{part}})] = u^{part}. \end{aligned}$$

□

Here is the proof of the next result.

Proposition 52. *Under Condition 1 [$H(1/x)$ is convex], $p_0^{part} \leq p_0^{no}$.*

Proof. Suppose to the contrary that $p_0^{part} > p_0^{no}$. Then, $N^{part} \preceq_{st} N^{no}$. Thus,

$$\begin{aligned} H(\theta_-) &= H(1/E[c_{N^{no}}]) \\ &\leq H(1/E[c_{N^{part}}]) \\ &\leq E[H(1/c_{N^{part}})] \\ &= E[H(\theta_{N^{part}})]. \end{aligned}$$

(The second inequality uses Jensen's inequality and the convexity of $H(1/x)$.) Therefore,

$$\begin{aligned} p_0^{no} &= 1 - (\lambda/\mu)H(\theta_-) \\ &\geq 1 - (\lambda/\mu)E[H(\theta_{N^{part}})] = p_0^{part}, \end{aligned}$$

a contradiction. □

Next, we prove the result comparing utilities.

Proposition 53. *Under Condition 2 [$J \circ H^{-1}$ is convex], $u^{part} > u^{no}$. Moreover, if $p_0^{part} < p_0^{no}$, then*

$$\frac{u^{part}}{u^{no}} \geq \frac{1 - p_0^{part}}{1 - p_0^{no}}.$$

Proof. We already know that $u^{part} > u^{no}$ for the case $p_0^{part} \geq p_0^{no}$. Otherwise, if $p_0^{part} < p_0^{no}$, let

$$\tau = \frac{1 - p_0^{part}}{1 - p_0^{no}} = \frac{E[H(\theta_{N^{part}})]}{H(\theta_-)}.$$

We have $\tau > 1$, and so

$$\begin{aligned} \tau u^{no} &= \tau J(\theta_-) = \tau J \circ H^{-1}[H(\theta_-)] \\ &\leq J \circ H^{-1}(\tau H(\theta_-)) \\ &= J \circ H^{-1}(E[H(\theta_{N^{part}})]) \\ &\leq E[J \circ H^{-1}(H(\theta_{N^{part}}))] \\ &= E[J(\theta_{N^{part}})] = u^{part}. \end{aligned}$$

(The first inequality follows from the fact that $J \circ H^{-1}$ is increasing and the second from Jensen's inequality.) □

Finally, we prove the result about $E[N]$ for a special case.

Proposition 54. *For uniform H and linear cost, the relation between $E\left[N^{no}\right]$ and $E\left[N^{part}\right]$ is the same as that between μ and 1. That is, they are equal for $\mu = 1$, $E\left[N^{no}\right] > E\left[N^{part}\right]$ for $\mu > 1$, and $E\left[N^{no}\right] < E\left[N^{part}\right]$ for $\mu < 1$.*

Proof. From (2.4),

$$E\left[N^{part}\right] = \lambda - (\mu - 1)\left(1 - p_0^{part}\right).$$

For $\mu = 1$, therefore, $E\left[N^{part}\right] = \lambda = E\left[N^{no}\right]$.

So, assume $\mu \neq 1$. We have

$$E\left[N^{no}\right] = \frac{\rho^{no}}{1-\rho^{no}},$$

or

$$\rho^{no} = \frac{E\left[N^{no}\right]}{E\left[N^{no}\right]+1}.$$

We know that ρ^{no} satisfies (2.3). Thus,

$$(1-\mu)\left[E\left[N^{no}\right]\right]^2 + (\mu+\lambda)E\left[N^{no}\right]\left[E\left[N^{no}\right]+1\right] - \lambda\left[E\left[N^{no}\right]+1\right]^2 = 0,$$

or

$$\left[E\left[N^{no}\right]\right]^2 + E\left[N^{no}\right](\mu-\lambda) - \lambda = 0. \tag{7}$$

Since $1-p_0^{part} = \rho^{part} > \rho^{no}$, we have $(1-\mu)(\rho^{part})^2 + (\mu+\lambda)\rho^{part} - \lambda > 0$. Thus, again using (2.4),

$$\left[E\left[N^{part}\right]\right]^2(1-\mu) + E\left[N^{part}\right](1-\mu)(\mu-\lambda) - \lambda(1-\mu) > 0.$$

For $\mu < 1$ this reduces to

$$\left[E\left[N^{part}\right]\right]^2 + E\left[N^{part}\right](\mu-\lambda) - \lambda > 0.$$

Comparing this to (7), we see that $E\left[N^{no}\right] < E\left[N^{part}\right]$. Similarly, the opposite conclusion holds for $\mu > 1$. □

.1.5 No Information and Full Information

Observe that

$$\frac{d\ln f^{full}(v)}{dv} = \lambda H\left[\frac{1}{c(v)}\right] - \mu,$$

which is decreasing in v. Thus, f^{full} is log-concave, and so the ratio f^{full}/f^{no} is too. Also, $f^{full}(0^+) = \lambda p_0^{full}$ and

$$\begin{aligned} f^{no}(0^+) &= \rho_-(1-\rho_-)\mu = \lambda\theta_-(1-\rho_-) \\ &\leq \lambda(1-\rho_-) = \lambda p_0^{no}. \end{aligned}$$

So, if $p_0^{full} \geq p_0^{no}$, then $f^{full}(0^+) \geq f^{no}(0^+)$, and so $V^{full} \preceq_{st} V^{no}$.

With this key fact established, the results are entirely analogous to those above. We provide a complete proof only for the first one, which asserts that, if $p_0^{full} \geq p_0^{no}$, then $u^{full} > u^{no}$.

Proof. As above, $p_0^{full} \geq p_0^{no}$ implies $V^{full} \preceq_{st} V^{no}$. Therefore, $E[c(V^{full})] \leq E[c(V^{no})]$. Also, the equilibrium condition can be expressed as $\theta_- = 1/E[c(V^{no})]$. Thus,

$$\begin{aligned} u^{no} &= J(\theta_-) = J(1/E[c(V^{no})]) \\ &\leq J(1/E[c(V^{full})]) \\ &< E[J(1/c(V^{full}))] \\ &= E[J(\theta_{V^{full}})] = u^{full}. \end{aligned}$$

□

.1.6 Partial Information and Full Information

It appears harder to compare partial and full information. In general f^{full}/f^{part} is not log-concave nor even unimodal. (The case of two types of customers provides a counterexample.) We have the analogue to only one of the results above.

First, we establish some preliminary results. Denote $a = \lambda/\mu$, and define

$$\Upsilon(a) = \frac{1}{p_0^{full}} - 1 = a\mu \int_0^\infty e^{-\mu v} \exp\left[a\mu C(v)\right] dv.$$

We analyze the power-series reprentation of $\Upsilon(a)$. We have

$$\begin{aligned} \Upsilon(a) &= a\mu \int_0^\infty e^{-\mu v} \sum_{n=0}^\infty \frac{[a\mu C(v)]^n}{n!} dv \\ &= \sum_{n=0}^\infty \frac{a^{n+1}}{n!} \int_0^\infty \mu e^{-\mu v} [\mu C(v)]^n dv \\ &= \sum_{n=1}^\infty \frac{a^n}{n!} n \int_0^\infty \mu e^{-\mu v} [\mu C(v)]^{n-1} dv. \end{aligned}$$

Thus, $\Upsilon(0) = 0$, and for $n > 0$,

$$\Upsilon^{(n)}(0) = n \int_0^\infty \mu e^{-\mu v} \left[\mu C(v)\right]^{n-1} dv = nE[\{\mu C(S)\}^{n-1}]. \tag{8}$$

In particular, $\Upsilon'(0) = 1$.

The following is an alternative representation of the higher coefficients.

Lemma 55. *For $n > 1$*

$$\frac{\Upsilon^{(n)}(0)}{n!} = E\left[C'(S^{(1)}) C'(S^{(2)}) \cdots C'(S^{(n-1)})\right].$$

Proof. Integrate (8) by parts to obtain

$$\begin{aligned} \frac{\Upsilon^{(n)}(0)}{n} &= \left\{(-e^{-\mu v}) \left[\mu C(v)\right]^{n-1}\right\}_0^\infty + \int_0^\infty e^{-\mu v} (n-1) \mu C'(v) \left[\mu C(v)\right]^{n-2} dv \\ &= (n-1) \int_0^\infty \mu e^{-\mu v} C'(v) \left[\mu C(v)\right]^{n-2} dv, \end{aligned}$$

or

$$\frac{\Upsilon^{(n)}(0)}{n(n-1)} = E\left[C'(S) \left[\mu C(S)\right]^{n-2}\right].$$

In particular,

$$\frac{\Upsilon^{(2)}(0)}{2!} = E\left[C'(S)\right].$$

Thus, the result holds for $n = 2$.

Next, for $n > 2$,

$$\begin{aligned}
\frac{\Upsilon^{(n)}(0)}{n(n-1)} &= \int_0^\infty \mu e^{-\mu v} C'(v) \left[\mu C(v)\right]^{n-2} dv \\
&= \int_0^\infty \mu e^{-\mu v} C'(v) \left[\mu C(v)\right]^{n-3} \mu \int_0^v C'(t_1) dt_1 dv \\
&= \int_0^\infty \mu C'(t_1) \int_{t_1}^\infty \mu e^{-\mu v} C'(v) \left[\mu C(v)\right]^{n-3} dv dt_1 \\
&= \int_0^\infty \mu e^{-\mu t_1} C'(t_1) \int_0^\infty \mu e^{-\mu t_2} C'(t_1 + t_2) \left[\mu C(t_1 + t_2)\right]^{n-3} dt_2 dt_1,
\end{aligned}$$

or

$$\frac{\Upsilon^{(n)}(0)}{n(n-1)} = E\left[C'(S^{(1)}) C'(S^{(2)}) \left[\mu C(S^{(2)})\right]^{n-3}\right].$$

In particular,

$$\frac{\Upsilon^{(3)}(0)}{3!} = E\left[C'(S^{(1)}) C'(S^{(2)})\right],$$

which is the result for $n = 3$.

Now, for $n > 3$,

$$\begin{aligned}
&\frac{\Upsilon^{(n)}(0)}{n(n-1)} \\
= \;& \int_0^\infty \mu e^{-\mu t_1} C'(t_1) \int_0^\infty \mu e^{-\mu t_2} C'(t_1+t_2) \left[\mu C(t_1+t_2)\right]^{n-3} dt_2 dt_1 \\
= \;& \int_0^\infty \mu e^{-\mu t_1} C'(t_1) \int_0^\infty \mu e^{-\mu t_2} C'(t_1+t_2) \left[\mu C(t_1+t_2)\right]^{n-4} \int_0^{t_1+t_2} \mu C'(v) dv dt_2 dt_1 \\
= \;& \int_0^\infty \mu e^{-\mu t_1} C'(t_1) \int_0^\infty \mu e^{-\mu t_2} C'(t_1+t_2) \left[\mu C(t_1+t_2)\right]^{n-4} \\
& \bullet \left(\int_0^{t_1} \mu C'(v) dv + \int_{t_1}^{t_1+t_2} \mu C'(v) dv \right) dt_2 dt_1 \\
= \;& \int_0^\infty \mu C'(v) \int_v^\infty \mu e^{-\mu t_1} C'(t_1) \int_0^\infty \mu e^{-\mu t_2} C'(t_1+t_2) \left[\mu C(t_1+t_2)\right]^{n-4} dt_2 dt_1 dv \\
& + \int_0^\infty \mu e^{-\mu t_1} C'(t_1) \int_{t_1}^\infty \mu C'(v) \int_{v-t_1}^\infty \mu e^{-\mu t_2} C'(t_1+t_2) \left[\mu C(t_1+t_2)\right]^{n-4} dt_2 dv dt_1 \\
= \;& 2 \int_0^\infty \mu e^{-\mu s_1} C'(s_1) \int_0^\infty \mu e^{-\mu s_2} C'(s_1+s_2) \\
& \bullet \int_0^\infty \mu e^{-\mu s_3} C'(s_1+s_2+s_3) \left[\mu C(s_1+s_2+s_3)\right]^{n-4} ds_3 ds_2 ds_1 \\
= \;& 2E\left[C'(S^{(1)}) C'(S^{(2)}) C'(S^{(3)}) \left[\mu C(S^{(3)})\right]^{n-4} \right].
\end{aligned}$$

Continuing in this manner, we obtain for $n > k \geq 3$

$$\frac{\Upsilon^{(n)}(0)}{n(n-1)} = (k-1)! E\left[C'(S^{(1)}) \cdots C'(S^{(k)}) \left[\mu C(S^{(k)})\right]^{n-k-1} \right]$$

In particular, for $k = n-1$,

$$\frac{\Upsilon^{(n)}(0)}{n!} = E\left[C'(S^{(1)}) \cdots C'(S^{(n-1)}) \right],$$

as asserted. □

Lemma 56. *For $n > 1$,*

$$\begin{aligned} & E\left[C'\left(S^{(1)}\right) C'\left(S^{(2)}\right) \cdots C'\left(S^{(n-1)}\right)\right] \\ \geq \;\; & E\left[C'\left(S^{(1)}\right)\right] E\left[C'\left(S^{(2)}\right)\right] \cdots E\left[C'\left(S^{(n-1)}\right)\right]. \end{aligned}$$

Proof. The two expressions are identical for $n = 2$. Consider the case $n = 3$.

$$E\left[C'\left(S^{(1)}\right) C'\left(S^{(2)}\right)\right] = E_{S_1}\left[C'\left(S_1\right) E_{S_2}\left[C'\left(S_1 + S_2\right)\right]\right].$$

Both $C'\left(S_1\right)$ and $E_{S_2}\left[C'\left(S_1 + S_2\right)\right]$ are decreasing as functions of S_1. These two random variables are therefore positively correlated (e.g., Casella and Berger, 2001). Thus,

$$\begin{aligned} E\left[C'\left(S^{(1)}\right) C'\left(S^{(2)}\right)\right] & \geq E_{S_1}\left[C'\left(S_1\right)\right] E_{S_1}\left[E_{S_2}\left[C'\left(S_1 + S_2\right)\right]\right] \\ & = E\left[C'\left(S^{(1)}\right)\right] E\left[C'\left(S^{(2)}\right)\right]. \end{aligned}$$

The general case follows similarly. □

Now we are ready to prove

Proposition 57. *Under Condition 1 [$H(1/x)$ is convex], $p_0^{full} \leq p_0^{part}$.*

Proof. We have

$$\frac{\Upsilon^{(n)}(0)}{n!} = E\left[C'\left(S^{(1)}\right) C'\left(S^{(2)}\right) \cdots C'\left(S^{(n-1)}\right)\right], n > 1.$$

By Lemma (56),

$$\frac{\Upsilon^{(n)}(0)}{n!} \geq E\left[C'\left(S^{(1)}\right)\right] E\left[C'\left(S^{(2)}\right)\right] \cdots E\left[C'\left(S^{(n-1)}\right)\right].$$

By Jensen's inequality and the convexity of $H(1/x)$, each

$$E\left[C'\left(S^{(m)}\right)\right] = E\left[H\left\{\frac{1}{c\left(S^{(m)}\right)}\right\}\right] \geq H\left\{\frac{1}{E[c\left(S^{(m)}\right)]}\right\} = H\left(\frac{1}{c_m}\right).$$

Thus,

$$\frac{\Upsilon^{(n)}(0)}{n!} \geq H\left(\frac{1}{c_1}\right) H\left(\frac{1}{c_2}\right) \cdots H\left(\frac{1}{c_{n-1}}\right) = \Theta_n, n > 1.$$

This is the corresponding factor in Θ for partial information. □

.1.7 Extension

For the case $c(0) < 1$, average utility and throughput are no longer proportional, because J is no longer proportional to H over their extended domain. Unlike H, J is not constant for $\theta \geq 1$.

The proofs of Proposition 2 and the preceding lemmas go through as is.

Proposition 4 requires more care. The case $p_0^{part} \geq p_0^{no}$ is fine, as before. For the other case, $p_0^{part} < p_0^{no}$, H^{-1} exists for the original H, but not the extended one, so the proof needs to be modified.

First, assume $\eta = h(1) > 0$. Define

$$\hat{H}(\theta) = \begin{cases} H(\theta) & , \quad \theta \leq 1 \\ 1 + \eta[1 - (1/\theta)] & , \quad \theta > 1. \end{cases}$$

This function is strictly increasing everywhere, and so has an inverse, $\hat{H}^{-1}$. Also, $\hat{H}$ is continuously differentiable at $\theta = 1$, and so is $\hat{H}^{-1}$. By Condition 2, $J \circ \hat{H}^{-1}(\phi)$ is convex for $\phi < 1$. For $\phi > 1$,

$$J \circ \hat{H}^{-1}(\phi) = 1 - [1 - J(1)][1 - (1/\eta)(\phi - 1)].$$

This is linear, and so convex. Finally, $J \circ \hat{H}^{-1}$ too is continuously differentiable at $\phi = 1$. So, $J \circ \hat{H}^{-1}$ is convex overall. Now, define

$$\hat{\tau} = \frac{E[\hat{H}(\theta_{N^{part}})]}{\hat{H}(\theta_-)}.$$

A proof just like that of Proposition 4, with $\hat{H}$ replacing H, shows that $u^{part}/u^{no} \geq \hat{\tau}$. Moreover, $\hat{H}(\theta) \geq H(\theta)$, and $p_0^{part} < p_0^{no}$ implies $\theta_- \leq 1$, so $\hat{H}(\theta_-) = H(\theta_-)$.

Therefore,

$$\hat{\tau} \geq \frac{E[H(\theta_{N^{part}})]}{H(\theta_{-})} = \frac{1 - p_0^{part}}{1 - p_0^{no}}.$$

This completes the proof, assuming $h(1) > 0$.

For the case $h(1) = 0$, we can use a limit argument. Construct a sequence of H's, each with $h(1) > 0$, which converges to the original H. Apply the argument above to each item in the sequence. Then use continuity.

As for Proposition 3, the crucial step is the use of Condition 1 and Jensen's inequality to get

$$H(1/E[c_{N^{part}}]) \leq E[H(1/c_{N^{part}})].$$

For large λ, $c_{N^{part}}$ is nearly always > 1. It spends most of its time in the region where $H(1/x)$ is convex. Moreover, since $H(1/x)$ decreases from 1 to 0 and never becomes negative, there must be places where it's strictly convex or kinked. By combining these facts carefully, one can verify the inequality above.

.2 Appendix for Chapter 5

Consider the special case above, with linear c and uniform H. Here we derive the closed-form expressions for p_0 and $E[W]$ mentioned earlier.

Idle probability

$$\Theta_1 = \frac{\lambda}{\nu}.$$

For $n > 1$,

$$\begin{aligned}
\Theta_n &= \frac{\lambda}{\nu}\prod_{m=1}^{n-1}\left(q\nu + \frac{\lambda}{1+m/\nu}\right)/\nu \\
&= \frac{\lambda}{\nu}\prod_{m=1}^{n-1}\left(q + \frac{\lambda}{\nu+m}\right) \\
&= \frac{\lambda}{\nu}\prod_{m=1}^{n-1} q\left(\frac{\nu+\lambda/q+m}{\nu+m}\right) \\
&= \frac{\lambda}{\nu}q^{n-1}\frac{\Gamma(\nu+\lambda/q+n)/\Gamma(\nu+\lambda/q+1)}{\Gamma(\nu+n)/\Gamma(\nu+1)} \\
&= \frac{\lambda}{\nu}q^{n-1}\frac{\Gamma(\nu+\lambda/q+n)/(\Gamma(\nu+\lambda/q+1)\Gamma(n-1))}{\Gamma(\nu+n)/(\Gamma(\nu+1)\Gamma(n-1))} \\
&= \frac{\lambda}{\nu}q^{n-1}\frac{B(\nu+1,n-1)}{B(\nu+\lambda/q+1,n-1)},
\end{aligned}$$

where $B(x, y)$ is the beta function. Thus

$$\begin{aligned}
\Theta &= \sum_{n>0}\Theta_n \\
&= \frac{\lambda}{\nu}\left[1 + \sum_{n>1} q^{n-1}\frac{B(\nu+1,n-1)}{B(\nu+\lambda/q+1,n-1)}\right]
\end{aligned}$$

The incomplete beta function $B(x; a, b) = \int_0^x t^{a-1}(1-t)^{b-1}dt, a > 0, b > 0$ can be expressed as follows (see, Press, et al. 1992)

$$B(x; a, b) = \frac{x^a(1-x)^b}{a}\left[1 + \sum_{n=0}^{\infty}\frac{B(a+1,n+1)}{B(a+b,n+1)}x^{n+1}\right].$$

Using this expression, we can rewrite Θ as

$$\Theta = \frac{\lambda B\left(q;\nu,\lambda/q+1\right)}{q^{\nu}(1-q)^{\lambda/q+1}}.$$

Thus

$$p_0 = \left\{ 1 + \frac{\lambda B\left(q; \nu, \lambda/q+1\right)}{q^{\nu}\left(1-q\right)^{\lambda/q+1}} \right\}^{-1}.$$

Mean waiting time

Based on the distribution of M, we can obtain the Laplace-Stieltjes transform of the stationary distribution of the waiting time W.

$$\begin{aligned}
\tilde{W}(s) &= \int_0^\infty e^{-st} dF_W(t) \\
&= p_0 + \sum_{n=1}^{\infty} \left(\frac{\nu}{\nu+s}\right)^n p_n \\
&= p_0 + \sum_{n=1}^{\infty} \left(\frac{\nu}{\nu+s}\right)^n \Theta_n p_0 \\
&= p_0 \left[1 + \frac{\lambda}{\nu+s} + \sum_{n=2}^{\infty} \left(\frac{\nu}{\nu+s}\right)^n \frac{\lambda}{\nu} q^{n-1} \frac{B(\nu+1, n-1)}{B(\nu+\lambda/q+1, n-1)} \right] \\
&= p_0 \left\{ 1 + \frac{\lambda}{\nu+s} \left[1 + \sum_{n>1} \left(\frac{q}{1+s/\nu}\right)^{n-1} \frac{B(\nu+1, n-1)}{B(\nu+\lambda/q+1, n-1)} \right] \right\} \\
&= p_0 \left[1 + \frac{\frac{\lambda\nu}{\nu+s} B\left(\frac{q}{1+s/\nu}; \nu, \lambda/q+1\right)}{\left(\frac{q}{1+s/\nu}\right)^{\nu} \left(1 - \frac{q}{1+s/\nu}\right)^{\lambda/q+1}} \right] \\
&= A(s)/A(0),
\end{aligned}$$

where

$$A(s) = 1 + \frac{\frac{\lambda}{1+s/\nu} B\left(\frac{q}{1+s/\nu}; \nu, \lambda/q+1\right)}{\left(\frac{q}{1+s/\nu}\right)^{\nu} \left(1 - \frac{q}{1+s/\nu}\right)^{\lambda/q+1}}.$$

Denote $x(s) = q/(1+s/\nu)$. The mean waiting time is thus

$$
\begin{aligned}
E[W] &= -\tilde{W}'(0) \\
&= -p_0 \left[1 + \left(\frac{\lambda\nu}{\nu+s}\right)\left(\frac{B(x(s);\nu,\lambda/q+1)}{x(s)^\nu(1-x(s))^{\lambda/q+1}}\right)\right]' \Bigg|_{s=0} \\
&= -p_0 \left[-\frac{\lambda}{\nu}\frac{B(q;\nu,\lambda/q+1)}{q^\nu(1-q)^{\lambda/q+1}} + \lambda\left(\frac{B(x(s);\nu,\lambda/q+1)}{x(s)^\nu(1-x(s))^{\lambda/q+1}}\right)'\right]\Bigg|_{s=0} \\
&= p_0\frac{1}{\nu}\left(\frac{1}{p_0}-1\right) \\
&\quad -p_0\lambda\left[\frac{x(s)'}{q(1-q)} - \frac{B(q;\nu,\lambda/q+1)x(s)'}{q^\nu(1-q)^{\lambda/q+1}}\left(\frac{\nu}{q}-\frac{\lambda/q+1}{1-q}\right)\right]\Bigg|_{s=0} \\
&= p_0\frac{1}{\nu}\left(\frac{1}{p_0}-1\right) \\
&\quad -p_0\left[\lambda\frac{\mu-\nu}{\nu^2}\frac{\nu^2}{(\nu-\mu)\mu} - \left(\frac{1}{p_0}-1\right)\frac{\mu-\nu}{\nu^2}\left(\frac{\nu^2}{\nu-\mu} - \frac{\lambda\nu^2}{\mu(\nu-\mu)} - \frac{\nu}{\mu}\right)\right] \\
&= \frac{1}{\nu}(1-p_0) + \frac{\lambda p_0}{\mu} + (1-p_0)\left(-1+\frac{\lambda}{\mu}-\frac{1}{\nu}+\frac{1}{\mu}\right) \\
&= \frac{\lambda+(1-\mu)(1-p_0)}{\mu}.
\end{aligned}
$$

References

[1] Abramowitz, M. and I. Stegun. 1965. *Handbook of Mathematical Functions.* Dover. New York.

[2] Afèche, P. and H. Mendelson. 2004. Pricing and priority auction in queueing systems with a generalized delay cost structure. *Management Sci.* 50 869-882.

[3] Altman, E. and R. Hassin. 2002. Non-threshold equilibrium for customers joining an M/G/1 queue. In: *Proceedings of the 10th International Symposium of Dynamic Games*, Saint-Petersburg Russia.

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

[5] Armony, M. and C. Maglaras. 2004b. Contact center with a call-back option and real-time delay information. *Oper. Res.* 52 527-545.

[6] Armony, M., N. Shimkin and W. Whitt. 2005. The impacts of delay announcements in many-server queues with abandonment. Working Paper.

[7] Athey, Susan. 2002. Monotone comparative statics under uncertainty. *The Quarterly J. Economics.* 117 187-223.

[8] Bae, J., S. Kim and E. Lee. 2001. The virtual waiting time of the M/G/1 queue with impatient customers. *Queuing Systems.* 38 485-489.

[9] Bekker, R., S. Borst, O. Boxma and O. Kella. 2004. Queues with workload-dependent arrival and service rates. *Queuing Systems.* 46 537-556.

[10] Bekker, R. 2005. Finite-buffer queues with workload-dependent service and arrival rates. *Queueing Systems.* 50 231-253.

[11] Bocharov, P., C. D'Apice, A. Penchinkin and S. Salerno. 2004. *Queueing Theory.* VSP. Utrecht, Boston.

[12] Brill, P. and M. Posner. 1977. Level crossings in point process applied to queues: single-server case. *Oper. Res.* 25 662-674.

[13] Carmon, Z., J. Shanthikumar and T. Carmon. 1995. A psychological perspective on service segmentation models: The significance of accounting for consumers' perceptions of waiting and service. *Management Sci.* 41 1806-1815

[14] Casella, G. and R. Berger. 2001. *Statistical Inference. 2nd ed.* Duxbury.

[15] Chao, X. and L. Dai. 1995. A monotonicity result for a single-server loss system. *J. Applied Prob.* 32 1112-1117.

[16] Collins, E., and A. Brooms. 2005. The Bernoulli feedback queue with balking: Stochastic order results and equilibrium joining rules. University of London. Working paper.

[17] Dai, L. and X. Chao. 1996. Comparing single-server loss systems in random environment. *IEEE Transaction on Automatic Control.* 41 1079-1083.

[18] Dobson, G., J. Pinker. 2006. The value of sharing lead time information. *IIE Transactions.* 38 171-183.

[19] Duenyas, I. and W. Hopp. 1995. Quoting lead times. *Management. Sci.* 41 43-57.

[20] Edelson, M. and K. Hildebrand. 1975. Congestion tolls for Poisson queueing processes. *Econometrica.* 43 81-92.

[21] Fakinos, D. 1982. The expected remaining service time in a single server queue. *Oper. Res.* 30 1014-1018.

[22] Freixas, X., and R. Kihlstrom. 1984. Risk aversion and information demand. In M. Boyer and R. Kihlstrom, eds., *Bayesian Models in Economic Theory.* Amsterdam: North Holland.

[23] Gavish, B. and P. Schweitzer. 1973. Queue regulation policies using full information. Working paper, Israel Scientific Center.

[24] Hassin, R. 1986. Consumer information in markets with random products quality: The case of queues and balking. *Econometrica.* 54 1185-1195.

[25] Hassin, R. and M. Haviv. 2003. *To Queue or Not to Queue: Equilibrium Behavior in Queuing Systems.* Kluwer Academic Publishers. Boston/Dordrecht/London.

[26] Heyman, D. 1982. On Ross's conjectures about queues with non-stationary Poisson arrivals. *J. Appl. Prob.* 19 245-249.

[27] Heyman, D. and M. Sobel. 1986. *Stochastic Models in Operations Research.* Vol.1. McGraw-Hill. New York.

[28] Hilton, R. 1981. The determinants of information value: Synthesizing some general results. *Management Science.* 1981 57-64.

[29] Hui, M. and D. Tse. 1996. What to tell customer in waits of different lengths: an integrative model of service evaluation. *J. Marketing.* 60 81-90.

[30] Hu, J. and M. Zazanis. 1993. A sample path analysis of M/GI/1 queues with workload restrictions. *Queueing Systems.* 14 203-213.

[31] Karlin, S. 1968. *Total Positivity. Vol. 1.* Stanford University Press.

[32] Kelly, F. 2000. Models for self-managed Internet. *Phi. Trans. R. Soc. Lond. A* 358 2335-2348.

[33] Kijima, M. 1997. *Markov Processes for Stochastic Modeling.* Chapman & Hall.

[34] Kulkarni, V. 1995. *Modeling and Analysis of Stochastic Systems.* Chapman & Hall.

[35] Kumar, P., M. Kalwani and M. Dada. 1997. The impact of waiting time guarantees on customers' waiting experiences. *Marketing Science.* 16 295-314.

[36] Liu, L. and V. Kulkarni. 2006. M/G/1 queues with workload-based balking. *Queueing Systems.* 52 251-260.

[37] Maister, D. 1985. The psychology of waiting lines, in *The Service Encounter: Managing Empolyee/Customer Interaction in Service Businesses.* Czepiel, J., M. Solomon, and C. Suprenant (eds.) Lexington Books, Lexington, MA. 113-123.

[38] Mandelbaum, A. and N. Shimkin. 2000. A model for rational abandonments from invisible queues. *Queueing Systems.* 36 141-173.

[39] Mandelbaum, A. and U. Yechiali. 1979. The conditional residual service time in the M/G/1 queue. Working Paper.

[40] Müller, A. and D. Stoyan. 2002. *Comparison methods for stochastic models and risks*, John Wiley & Sons.

[41] Munichor, N., A. Rafaeli. 2007. Numbers or apologies? Customer reactions to telephone waiting time fillers. *Journal of Applied Psychology* 92 511-518.

[42] Nadiminti, R., T. Mukhopadhyay and C. Kriebel. 1996. Risk version and the value of information. *Decision Support Systems.* 16 241-254.

[43] Naor, P. 1969. The regulation of queue size by levying tolls. *Econometrica* . 37 15-24.

[44] Neuts, M. 1981. *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach.* The Johns Hopkins University Press.

[45] O'Cinneide, C. 1991. Phase-Type distributions and majorization. *The Annals of Applied Probability.* 1 219-227.

[46] Press, W., S. Teukolsky, W. Vetterling and B. Flannery. 1992. *Numerical Recipes in Fortran*, 2nd ed. Cambridge University Press.

[47] Perry, D. and S. Asmussen. 1995. Rejection rules in the M/G/1 queue. *Queueing Systems.* 19 105-130.

[48] Rolski, T. 1990. Queues with nonstationary inputs. *Queueing Systems.* 5 113-130.

[49] Ross, S. 1978. Average delay in queues with nonstationary arrivals. *J. Appl. Prob.* 15 602-609.

[50] Ross, S. 1983. *Stochastic Process.* Wiley: New York.

[51] Schroeter, R. 1982. The costs of concealing the customer queue. Working paper EC-118, Bureau of Business and Economic Research, Arizona State University.

[52] Shaked, M. and J. Shanthikumar. 1994. *Stochastic Orders and Their Applications.* Academic Press, Boston, San Diego, New York, London, Sydney, Tokyo, Toronto.

[53] Shanthikumar, J. 1988. DFR property of first-passage times and its preservation under geometric compounding. *The Annals of Probability* 16 397-406.

[54] Shimkin, N. and A. Mandelbaum. 2004. Rational abandonment from tele-queues: nonlinear waiting cost with heterogeneous preferences. *Queueing Systems.* 47 117-146.

[55] Smith, R. and W. Whitt. 1981. Resource sharing for efficiency in traffic systems. *Bell System Tech. J.* 60 39-55.

[56] Spearman, M. and R. Zhang. 1999. Optimal lead time policies. *Management Sci.* 45 290-295.

[57] Stidham, S. 1985. Optimal control of admission to a queuing system. *IEEE Trans. Auto. Control.* AC 30 705-13.

[58] Stoyan, D. and D. Daley. 1983. *Comparison Methods for Queues and Other Stochastic Models.* John Wiley & Sons.

[59] Taylor, S. 1994. Waiting for service: The relationship between delays and evaluations of service. *J. Marketing.* 58 56-69.

[60] Ward, A. and P. Glynn. 2003. A diffusion approximation for a Markovian queue with reneging. *Queueing Systems*, 43 103-128.

[61] Whang, S. 1993. Analysis of interorganizational information sharing. *Journal of Organizational Computing.* 3 257-277.

[62] Whitt, W. 1985. Deciding which queue to join: Some counterexamples. *Management Sci.* 34 55-62.

[63] Whitt, W. 1999. Improving service by informing customer about anticipated delays. *Management Sci.* 45 192-207.

[64] Whitt, W. 2006. Sensitivity of performance in the Erlang-A queueing model to changes in the model parameters. *Operations Research.* 54 247-260.

[65] Willinger, M. 1989. Risk aversion and the value of information. *Journal of Risk and Insurance.* 56 320-328.

[66] Wolff, R. 1989. *Stochastic Modeling and the Theory of Queues.* Prentice-Hall, Englewood Cliffs, NJ.

[67] Zhou, R. and D. Soman. 2003. Looking back: Exploiting the psychology of queueing and the effect of the number of people behind. *J. Consumer Research.* 29 517-529.

[68] Zohar, E., A. Mandelbaum and N. Shimkin. 2002. Adaptive behavior of impatient customers in tele-queues: theory and empirical support. *Management Sci.* 48 566-583.

Biography

Birth

Hubei Province, P.R.China on December 15, 1974

Education

Duke University, Fuqua School of Business
Durham, NC, USA
Doctoral of Philosophy in Business Administration, 2007

Shanghai Jiao Tong University, School of Management
Shanghai, China
Master of Science in Management Science & Engineering, 2000

Xi'an Jiao Tong University, School of Management
Xi'an, China
Bachelor of Engineering in Information Systems, 1997

Publications

Guo, P. and P. Zipkin. 2003. Analysis and comparison of queues with different levels of delay information. *Management Science.* Forthcoming.

Guo, P. and P. Zipkin. 2006. The effects of information on a queue with balking and phase-type service times. *Queueing Systems.* Under review.

Guo, P. and P. Zipkin. 2007. The impacts of customers' delay-risk sensitivities on a queue with balking. *Probability in the Engineering and Informational Sciences.* Under review.

Guo, P. and P. Zipkin. 2007. Information and congestion in a service system with balking. *European Journal of Operational Research.* Under review.