INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

Bell & Howell Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

UMI®

Timing Information in Data Networks

Anand Sharad Bedekar

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

University of Washington

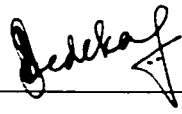
2000

Program Authorized to Offer Degree: Electrical Engineering

UMI Number: 9983457

Bell & Howell Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

In presenting this dissertation in partial fulfillment of the requirements for the Doctoral degree at the University of Washington, I agree that the Library shall make its copies freely available for inspection. I further agree that extensive copying of this dissertation is allowable only for scholarly purposes, consistent with "fair use" as prescribed in the U.S. Copyright Law. Requests for copying or reproduction of this dissertation may be referred to Bell and Howell Information and Learning, 300 North Zeeb Road, Ann Arbor, MI 48106-1346, to whom the author has granted "the right to reproduce and sell (a) copies of the manuscript in microform and/or (b) printed copies of the manuscript made from microform."

Signature 

Date AUGUST 17, 2000

University of Washington

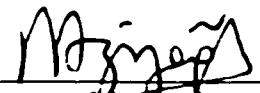
Graduate School

This is to certify that I have examined this copy of a doctoral dissertation by

Anand Sharad Bedekar

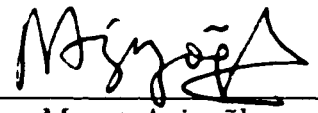
and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Chair of Supervisory Committee:

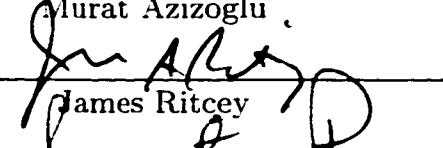


Murat Azizoğlu

Reading Committee:



Murat Azizoğlu



James Ritcey

Sumit Roy

Date: AUGUST 17, 2000

University of Washington

Abstract

Timing Information in Data Networks

by Anand Sharad Bedekar

Chair of Supervisory Committee:

Associate Professor Murat Azizoğlu
Electrical Engineering

The role of timing information is considered as the key to an information theoretic understanding of communication networks. In this dissertation, we focus on the analysis of problems related to timing information in data networks from an information theoretic perspective.

One problem is that of identifying the timing capacity of network components, i.e. the maximum information rate achievable by encoding information in the timing of packets. We first consider the timing capacity of discrete-time queues in which at most one packet may arrive or finish service in a slot, and demonstrate the extremal nature of the geometric service time distribution among such queues. We then analyze a discrete-time queueing model with batch arrivals and a batch service mechanism that is related to the leaky bucket flow control system. Within this model, we obtain a closed form expression for the timing capacity of the queue that can serve a geometrically distributed number of packets in a slot. We also establish a connection between the extremal nature of the geometric server and a queueing theoretic property of such queues. Finally, we obtain an upper bound to the timing capacity of a queueing system with multiple servers having i.i.d. geometrically distributed packet

service times.

Another problem is that of quantifying the amount of information about the times at which messages arrive at the source that must be transmitted to the destination to enable it to decode the messages within a finite amount of time. Suppose that messages are transmitted one at a time, and are decoded in the same order they arrived at the source. By viewing the message arrival and decoding processes as the arrival and departure processes from a hypothetical single-server queue, we can associate a service time with each message. The rate-distortion function of the message arrival process with message service time as the distortion measure is a lower bound for the amount of information about message arrival times that the receiver must receive. We explicitly obtain this rate-distortion function for the Poisson message arrival process.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

I would like to express my sincere thanks to my advisor, Prof. Murat Azizoğlu, for his thoughtful and patient guidance throughout the course of this dissertation. He has been very supportive in allowing me a lot of freedom to pursue my own direction of work, and his encouragement has been a great motivator for me.

I have been fortunate to have had a very enthusiastic group of friends who have made my stay in Seattle thoroughly enjoyable. They are too numerous to name, but I must mention Aby, Amol, Mohit, Niranjan, Rani, Sachin, Siva, Sugar, Suresh, and Ufilya. My friends Ahmed, Gokhan, Hamed, and Todd in the Fundamentals of Networking Laboratory have also been very helpful and encouraging, and I would like to thank all of them.

Most of all, I would like to express my heartfelt gratitude to my parents Padma and Sharad and my brother Aditya for their boundless love and unwavering patience, without which I would never have made it this far. Their wonderful outlook towards life has taught me much and helped me maintain an even keel. I dedicate this dissertation to them.

DEDICATION

To my parents, and to Aditya.

Chapter 1

INTRODUCTION

Since the development of the first data networks in the 1960s, the field of communication networks has evolved steadily into one of the main disciplines of communications. The last decade has witnessed a tremendous spurt in the use of networks, with the growth of the Internet as a global backbone for information transfer far surpassing all predictions. On the one hand, the development of wireless networks has allowed for greater mobility and access to users of the Internet. On the other hand, the development of all-optical networks has spurred a tremendous growth in the backbone capacity of wide-area networks.

In the general field of communications, information theory has provided a thorough characterization of the fundamental performance limits of communication channels, as well as of efficient techniques for achieving these limits. The results of information theory are now extensively being applied to the development of wireless systems, error correction coding, encryption, and data compression for digital voice and video transmission. In the field of communication networks, however, the understanding of fundamental performance limits has lagged behind the rapid implementation of ever more complex network systems. In particular, the impact of information theory in providing a theoretical foundation for the understanding of networks has been minimal. Information theoretic analysis of multiuser channels, beginning with Shannon's analysis of the canonical two-way channel [39], has largely consisted of generalizing the methods used in the analysis of single-user channels to channels with multiple

transmitters and multiple receivers. Much of the work in this direction is surveyed in [44, 22, 46]. However, although this work has succeeded in characterizing the achievable information rates on various multi-user channels, it has not managed to capture the essence of the behavior of sources on a network or of dynamic resource allocation schemes in networks.

A recent review [14] of the relationship between the fields of information theory and communication networks by leading experts in these fields bemoaned the "unconsummated union" between the two fields. Part of the reason for this is that the need for rapid and simple implementation of network systems and the need to make them interoperable for easy integration have led to the haphazard development of network protocols that are not based on any fundamental theoretical foundation or demonstrated optimality. However, as explained in [14], the principal reasons for information theory's lack of impact on the understanding of networks have to do with the approach taken by information theory itself. First, the classical information theory model of a communication system largely ignores the bursty nature of message sources, i.e. the fact that a source only has messages to send intermittently at random times, rather than having an infinite reservoir of data to send. This burstiness is a key phenomenon underlying the dynamic allocation of resources in a network, whereas the information theoretic analyses of single-user and multiuser channels depend on the infinite reservoir assumption. Secondly, information theory's main theorems characterize information rates that are achievable asymptotically, in the sense that the allowable delay for getting a message to its destination is assumed to be infinite. In networks, however, since resources are allocated dynamically and since the sources are bursty, the delay incurred in transmitting a message from its source to its destination is required to be finite, and indeed is a fundamental measure of the performance of resource allocation schemes. A third reason is that in the classical communication channel models studied by information theory, the random times taken by information packets to traverse a network are not well modeled. These random times, together

with the random message generation times at the sources and the dynamic allocation of network resources, give rise to queueing delays within the network. These queueing delays have an important influence on network performance, but are largely ignored in the information theoretic analyses of multiuser channels.

Performance analysis of networks has traditionally been based on queueing theory. The queueing theoretic approach is to assume an appropriate stochastic structure for the random message generation times at the sources and for the random delays experienced by the messages under certain loading conditions on the network, and to obtain steady state results about the average values of common measures of network performance such as message delay, fraction of packets that get dropped due to buffer overflows, etc. Thus the analysis is network-centric, i.e. conducted from the network's point of view, in the sense that the focus is on evaluating the performance of network resource allocation schemes assuming that the statistical behavior of the sources is known, rather than on efficient strategies that transmitters should follow to make the most out of a particular network resource allocation strategy. In contrast, an information theoretic approach to analyzing networks, like its analysis of communication channels, would be source-centric, in the sense that the network's resource allocation scheme would be assumed known, while the focus would be on finding efficient strategies that the transmitters should follow to make the most of the resources allocated to them by the network.

The first major information theoretic insight into the role of timing in networks was provided by Gallager's analysis [17] of the bursty arrivals of messages at random times at a source. Gallager reasoned that if a receiver is able to decode messages within a certain average delay after their arrival at the source, the receiver must have received some amount of information about the message arrival times at the source. For instance, before a receiver can decode a particular message, it must first be informed about the presence of the message at the source. A common way to do this is to use an idle symbol to indicate the periods when the source has no message to

send. The receiver is usually only interested in the contents of the messages and not in their arrival times *per se*. However this information about message arrival times must unavoidably be transmitted along with the information about the contents of the messages, and forms part of what is commonly called *protocol information* in networks terminology. The smaller the allowed delay, the quicker the receiver must be informed of the message arrivals at the source, and the greater the amount of such protocol information that must be transmitted. In [17], Gallager obtained a lower bound to the amount of such information in the case when the message arrival process at the source is Poisson, subject to a constraint on the average allowed delay per message. This bound is independent of the underlying network, and also does not assume any constraints on the order in which the messages are delivered to their destination. Although [17] was published nearly twenty-five years ago and has been acknowledged as a path-breaking work, there has not been much further analysis in this direction. It remains a tantalizing glimpse into a little understood but fundamental characteristic of all networks.

More recently, another aspect of timing in data networks was illuminated by Anantharam and Verdu [4]. They observed that in a packet network, information can be encoded not only in the contents of the packets but also in the timing. By transmitting the packets at appropriately chosen times, a source could convey information to the receiver across a network. The random delays experienced by the packets in crossing the network distorts this timing information. By modeling the random delays experienced by packets in traversing the network as the random service times in a single server queue, they quantified the maximum information rate, which we call the timing capacity, that is reliably achievable by encoding information purely in the timing of packets. They established the remarkable result that among all single server queues with a given mean service time, the $\cdot/M/1$ queue (i.e. the queue with i.i.d. exponentially distributed service times) has the least timing capacity. Perhaps the most important contribution of this paper was the use of queueing theoretic results

to establish information theoretic coding theorems.

The focus of this dissertation is to build on and establish connections between these seminal ideas, and to give perspective to the role of timing information in data networks. In this dissertation, we will identify some problems that capture the essential features of some important high-speed network components, and yet are tractable enough to permit a reasonable analysis of the role of timing information in the performance of these components. In the process, we will need to develop some new queueing theoretic results for these components to facilitate information theoretic analysis. We will also uncover a connection between the analyses of Gallager and of Anantharam and Verdu discussed above.

In the remainder of this chapter, we review these and other relevant analyses in greater detail to lay the groundwork for the rest of the dissertation. In Section 1.1, we review the work of Anantharam and Verdu [4] on the information rate that can be reliably sustained by encoding information in the timing of packets in a single-server queue. In Section 1.2, we examine the work of Gallager [17] on the amount of information about message arrival times that a receiver must receive to be able to decode the messages within a given finite average delay after their arrival at the source. Queueing theory forms the basis of much of the traditional performance analyses of networks. Queueing results play an important role in [4], and will also play an important role in the work presented in this dissertation. In Section 1.3, we review the queueing models that we will use later in this dissertation, especially discrete-time queueing models, and discuss relevant queueing theoretic results for these models. In Section 1.4, we survey some other intriguing aspects of timing information that have been examined in the literature, but will not be of primary interest to us in this dissertation. We will conclude this chapter with an overview of the organization of this dissertation and an outline of its main contributions in Section 1.5.

1.1 Information-Theoretic Capacity of Queues

In a recent paper, Anantharam and Verdu [4] analyzed the information theoretic capacity of continuous time queues. In this section, we take a closer look at the work of Anantharam and Verdu [4] on the amount of information that can be encoded in the timing of packets in a continuous time single-server queue.

Consider a source node and a destination node communicating over a network. Suppose that all the packets are routed over the same path, and reach the destination in order with no packets dropped at intermediate nodes. The flow of packets between the source and the destination can then be modeled as a single-server queue with an infinite buffer, with the random delays experienced by the packets over the network represented by the random service times of the packets.

In [4], Anantharam and Verdu observed that information can be transmitted over a queue not only through the contents of the packets but also through the times at which the packets arrive at the queue. The sequence of packet arrival times can be used as a code to convey information to the receiver, which observes the times at which the packets depart from the queue and attempts to infer the transmitted codeword. The random service times experienced by these packets act as noise in distorting the transmitted codeword.

In [4], Anantharam and Verdu analyzed the maximum rate at which information can be transmitted with arbitrarily small probability of error over a continuous-time single-server queue, using only the timing of the packets. We refer to this maximum rate as the *timing capacity* of the queue, and to the maximum information rate achievable using codes with an average packet rate of λ packets/s as the *λ-timing capacity*. For queues with arbitrary service time distributions, upper and lower bounds to the λ-timing capacity were established in [4]. It was also shown that among all queues with an average service rate of μ packets/s, the queue with exponentially distributed

service time has the least λ-timing capacity, which is given by $C(\lambda) = \lambda \log(\mu/\lambda)$.[1] This capacity can be achieved by a random code generated by independent, exponentially distributed interarrival times. The timing capacity was shown in [4] to be the supremum of λ-timing capacities over $0 \leq \lambda \leq \mu$. Consequently, the timing capacity of the queue with exponential service time distribution is $C = (\mu \log_2 e)/e$ bits/s. It was also shown that for the exponential server queue, the timing capacity does not increase even in the presence of instantaneous feedback to the transmitter about the departure times of the packets.

A distinctive feature of the proofs in [4] was the use of queueing theoretic results to establish information theoretic results for the timing capacity. For instance, Burke's theorem [6, Section 3.7] asserts that the departure process of the $M/M/1$ queue in steady state is a Poisson process, i.e. the sequence of interdeparture times are i.i.d. exponentially distributed random variables. This fact plays an important role in analyzing the input-output mutual information of the queue with exponentially distributed service times.

In conventional networks, information is encoded in the contents of the packets. The contents of the packet may also be corrupted by noise, which can be assumed independent of the timing of the packets. If the service time of a packet is independent of the contents of the packet, then the total information transmitted is the sum of the information in the contents of the packets and the information in their timing. If the contents of each packet can carry C_0 bits of information, the total information capacity C_I of the queue with information-bearing packets can be written as the sum of the timing capacity and the capacity of the packet contents, i.e.

$$C_I = \sup_{0 \leq \lambda \leq \mu} C(\lambda) + \lambda C_0.$$

If each packet contains a single bit of information ($C_0 = 1$ bit/packet), and the service time is exponentially distributed with rate μ packets/s, the total information capacity

[1] The capacity will be in units of bits/s if the base of the logarithm is 2.

evaluates to $2\mu e^{-1} \log_2 e = 1.0615\mu$ bits/s. This surprising result of [4] says that the information capacity of a queue serving single-bit packets with exponential packet service time distribution with rate μ packets/s is actually more than μ bits/s.

The work of [4] can be interpreted in terms of the separation of network functions into *layers* [6, Ch. 2] in the following way. The job of routing packets over the network is performed by what is known as the network layer. The transport layer has the responsibilities of splitting messages into packets, end-to-end error correction, rearrangement of out-of-order packets using sequence numbers, retransmissions of missing packets, and estimation of the congestion in the network to control the rate of packet transmission. There is usually some sort of feedback, in the form of acknowledgments about the receipt of packets, that allows the transmitter to estimate the time taken by the packets to reach the destination, and thus form an estimate of the congestion in the network. The times at which packets are transmitted are chosen by the transport layer at the transmitting node based on this estimate of the network congestion. The work of [4] can be interpreted as an information theoretic analysis of the packet channel provided by the transport layer, with the times at which packets are transmitted chosen based on the contents of the message itself, rather than just in response to the network state.

This work makes the conventional assumption of information theory that the source has an infinite reservoir of data to send, and ignores the issue of the random generation times of messages at the source. It also does not address the issue of finite allowable message delay, and its methods make use of arbitrarily long codes in establishing results on achievable information rates.

Within the framework of [4], there are several factors that have not been considered. Many computer networks, such as ATM networks, are inherently discrete time systems. That is, there is a natural time unit in the system, called a slot, and events are synchronized to occur at integer multiples of this time unit. This can introduce complications such as the possibility that multiple packets can be served in the same

time slot. In a network, packets can be routed over different paths which may be beyond the transmitter's control, and can arrive out of order at the destination. Nevertheless, the work of [4] provides an important starting point for the analysis of the maximum information rate that can be sustained between a source-destination pair on a network with random packet delays.

1.2 Protocol Information in Networks

In a network, messages are typically generated intermittently at the sources at random instants of time, and have to be transmitted across the network within a finite average delay. Consider a source A communicating with a receiver B. Suppose that messages arrive at A at random times $X_1, X_2, \cdots, X_n$ to be transmitted to B, and are decoded at B at times $Y_1, Y_2, \cdots, Y_n$, where $Y_i > X_i$, $i = 1, \cdots, n$. The fact that B is able to decode message i at time Y_i means that by time Y_i, B must have been informed that message i arrived at A. Thus B must receive some information about the times at which messages arrived at A, apart from the information about the contents of the messages. This additional information may be referred to as *protocol information*, although that term usually also includes other control information transmitted by each source, such as addressing information, routing information, etc. The point is that this information must be transmitted over the network along with the message content information, and thus must consume network resources. This amount of information about message arrival times that the receiver must receive was first analyzed in a seminal paper by Gallager [17].

Gallager observed that the particular transmission and decoding strategy, coupled with the underlying random network delays, induces a conditional probability distribution $P(Y_1, \cdots, Y_n | X_1, \cdots, X_n)$ which is 0 if $Y_i < X_i$ for any $i = 1, \cdots, n$. He reasoned that the amount of information about message arrival times that the receiver must have received is at least $I(X_1, \cdots, X_n; Y_1, \cdots, Y_n)$, where the mutual

information $I(\cdot\,;\cdot)$ is computed using the probability distribution of the message arrival process $P(X_1, \cdots, X_n)$ and the conditional probability distribution induced by the particular strategy. Suppose that it is required that this conditional distribution, the average delay per message $(1/n)\sum_{i=1}^{n} E(Y_i - X_i)$ be no more than d. It would then be of interest to minimize the mutual information per message $(1/n)I(X_1, \cdots, X_n; Y_1, \cdots, Y_n)$ over all conditional distributions subject to the constraint $(1/n)\sum_{i=1}^{n} E(Y_i - X_i) \leq d$. Gallager made the observation that $Y_1, \cdots, Y_n$ could be thought of as a reproduction of the message arrival sequence $X_1, \cdots, X_n$, and the total delay $\sum_{i=1}^{n}(Y_i - X_i)$ can be considered as a measure of the distortion between the sequences $X_1, \cdots, X_n$ and $Y_1, \cdots, Y_n$. With this point of view, the above minimization of $(1/n)I(X_1, \cdots, X_n; Y_1, \cdots, Y_n)$ amounts to finding the rate distortion function for source coding of the message arrival times with message delay as the distortion measure.

This rate distortion problem can be formally posed as follows. Consider a source code for the message arrival time process that encodes a sequence of n message arrival times $X^n = (X_1, \cdots, X_n)$ into one of M_n codewords. The decoding of this codeword yields a reproduction sequence $Y^n = (Y_1, \cdots, Y_n)$. The allowed source codes are those whose reproduction sequence satisfies $Y_i \geq X_i, 1 \leq i \leq n$. The distortion δ between the sequence of message arrival times and the sequence of reproduced times is defined as

$$\delta(X_1, \ldots, X_n; Y_1, \ldots, Y_n) = \begin{cases} \sum_{i=1}^{n}(Y_i - X_i), & Y_i \geq X_i, \\ \infty & \text{otherwise.} \end{cases}$$

This distortion measure ensures that for any finite distortion encoding, the reproduced time of each message is always later than the original message generation time. The rate of such an encoding is $(\log M_n)/n$ bits/message. A source coding rate R is *achievable at average distortion d* if for any $\gamma > 0$, there exist source codes for which $(1/n)\log M_n < R + \gamma$ and $(1/n)E[\delta(X^n; Y^n)] \leq d$ for all sufficiently large n.

The rate distortion function $R(d)$ is the smallest rate achievable at average distortion d. It is known in rate distortion theory that $R(d)$ is no more than $\liminf_{n\to\infty} (1/n)I(X^n;Y^n)$ for any sequence of conditional distributions $P(Y^n|X^n)$ satisfying $(1/n)E[\delta(X^n;Y^n)] \leq d$. Thus the rate distortion function with message delay as the distortion measure is a lower bound on the amount of information about message arrival times that must be transmitted to the receiver.

For the important special case of the Poisson message arrival process with rate λ, Gallager obtained the following lower bound on this rate distortion function:

$$R(d) \geq -\log(1 - e^{-\lambda d}) \quad \text{bits/message.}$$

One point to be noted in the above formulation is that it is not required that the messages be decoded at the receiver in the same order they arrived at the source, i.e. it is not required that $Y_i \geq Y_{i-1}$. Thus the possible means of information transmission from the source to the destination include networks in which packets are routed along possibly different paths and can reach the destination out of order. In such networks, there must be some additional information, such as sequence numbers, that help the receiver identify which particular message each packet belongs to. This information is not accounted for in the above formulation. Another unresolved issue is whether the lower bound to the rate distortion function for the Poisson message arrival process is actually achieved. The optimal source encoding mechanism that achieves the rate distortion function is unknown, and would also be of interest. Of course, the question of how to efficiently encode this information for transmission over specific networks remains open.

1.3 Queueing Theory

Queueing theory [25][26] has provided models for statistical prediction of the performance of data networks. By assuming certain stochastic properties for the message

generation process at each node in the network and on the time that the communication resources are occupied by each message, the theory aims to predict such measures of performance as the average delay experienced by a message for a given throughput level, the sizes of the buffers that must be used to keep the buffer overflow probability below a given level, the number of users that can be simultaneously supported while maintaining a given response time, etc. A network link can be modeled as a queue with a buffer, and the network can be viewed as a network of queues. With certain simplifying assumptions about the nature of the packet and message arrival processes at each node and about the distribution of the time taken to transmit a packet on a given link, queueing theory provides results on the steady state behavior of networks, such as the moments of the time spent by a message in the system, the buffer occupancy, the nature of the departure processes from the various nodes, etc. In some simple cases, exact distributions of these stochastic quantities can also be obtained.

For a communication session between two nodes in a network, packet transmission from the source to the destination can be modeled by a queue. Due to the random demands imposed on the network's resources by the other nodes in the network, each packet takes a random time to reach the destination. Packets may have to be buffered at the source node until previous packets have been transmitted, and some packets may have to be discarded if the buffer overflows. Packets may be transmitted one at a time or in batches, depending on the network access protocol. Depending on the routing mechanisms, packets may reach the destination in the same order they were transmitted at the source, or they may be out of order. All of these factors can be modeled by queues of different types.

Various types of continuous-time queueing systems [25, 26, 13, 48] and discrete-time systems [27, 42, 7, 50] have been analyzed in the literature. The following examples give an indication of the type of results provided by queueing theory that will be useful to us later in our analysis.

Example 1.1: Consider a continuous-time single-server first-in-first-out queue with

an infinite buffer and i.i.d. packet service times. If the service times are exponentially distributed, the queue is denoted as a $\cdot/M/1$ queue. Let the arrival process to an $M/M/1$ queue be a Poisson process with mean interarrival time $1/\lambda$, and let the mean service time be $1/\mu$ seconds. In steady state, the number N of packets in the system has the geometric distribution

$$P_N(k) = \rho^k(1-\rho), k = 0, 1, 2, \ldots,$$

where $\rho = \lambda/\mu$ [6, Section 3.3.1]. The waiting time W of a packet in the buffer before it gets into service has the mean value

$$E[W] = \frac{\rho}{\mu - \lambda}.$$

The average delay per packet is

$$E[T] = \frac{1}{\mu - \lambda}.$$

Burke's theorem [6, Section 3.7] asserts that in steady state, the departure process of this queue is also a Poisson process with mean interdeparture time $1/\lambda$, and moreover, the number of packets left behind in the queue by a departing packet is independent of all previous inter-departure times. ◁

Queues may have multiple servers. If a server is idle when a packet arrives, it takes up service of that packet immediately. A queue with m servers, each of which has an exponential service time distribution, is denoted as $\cdot/M/m$. Queues with m servers can be used to model networks with routing mechanisms in which each packet can follow one of m independent routes through the network. If the buffer at a node has a finite size K, any packet that finds the buffer full upon arrival is discarded. Such queues are denoted as $\cdot/M/m/K$. There are also queues in which a server takes a vacation for a random time after each packet service, so that the next packet has to wait until the vacation ends before it comes into service. All these queueing systems are analyzed in [6, 25].

Discrete-time queueing systems are becoming increasingly important because they provide an accurate model for high-speed computer communication networks like ATM [7]. In ATM networks, all the packets are of fixed size, and thus the transmission time of a single packet provides a natural time unit, called *slot*, in terms of which the operation of the system can be described. All arrivals and departures are assumed to occur at integer-valued epochs, or slots.

Example 1.2: In the discrete-time analogue of the continuous-time queues mentioned in Example 1.1, the service time of the packets are i.i.d. positive integer valued random variables, and at most one arrival and one departure can occur in a slot. We will denote a discrete-time queue for which the service time has a geometric distribution as a $\cdot/Geo^{(s)}/1$ queue, with the superscript indicating that at most a single arrival and a single departure can occur in a slot. An arrival process which has independent, geometrically distributed interarrival times is called a *Bernoulli* process. When a $\cdot/Geo^{(s)}/1$ queue is driven by a Bernoulli arrival process, in steady state, the departure process is also a Bernoulli process [42, p. 13]. ◁

The discrete-time queueing system with K independent servers, each of which has i.i.d. geometrically distributed service times, is denoted as $\cdot/Geo^{(s)}/K$. Discrete-time queues with multiple servers, finite buffers, and servers with vacations are analyzed in [42]. In discrete-time queues, depending on the way the time axis is slotted, there may be the possibility of a batch of packets arriving or departing in a slot. Note that in the $\cdot/Geo^{(s)}/K$ system, multiple departures can occur in a slot. The following example describes a type of batch-arrival batch-service discrete-time queue that will be used in our analysis in Chapter 3.

Example 1.3: Consider the following model of a batch-arrival batch-service discrete-time queue. Let Y_i be the number of packets in the queue at the beginning of slot i. Let A_i be the number of arriving packets, D_i the number of departing packets, and $X_i = Y_i + A_i$ the number of packets in the queue just after the arrivals, in slot i. In slot i, the queue can serve a random number S_i of packets. If there are S_i or more

packets in the queue, S_i of them depart. If there are less than S_i packets waiting in the queue in slot i, then all the packets in the queue depart. That is, $D_i = \min(X_i, S_i)$. S_i are assumed to be i.i.d. random variables. The equations governing this queueing system are

$$\begin{aligned} X_i &= Y_i + A_i \\ D_i &= \min(X_i, S_i) \\ Y_{i+1} &= (X_i - S_i)^+. \end{aligned} \tag{1.1}$$

This model is motivated by a packet switch [7]. In each slot, the switch serves some packets from the buffer at each input port. From the point of view of a particular virtual circuit connection over the switch, the number of packets served in a slot can be modeled as a random variable S_i, with the randomness is caused by the random number of packets that are ready for service at the other input ports of the switch. The above model can also serve as an approximation for a single virtual circuit connection that is regulated by a leaky bucket scheme, with S_i representing the number of tokens available in slot i. The assumption that S_i are i.i.d. may be a good approximation in the situation in which the leaky bucket actually regulates the aggregate traffic of several virtual circuits out including the one under consideration. From the point of view of the connection under consideration, the number of tokens available can then be considered approximately independent from slot to slot.

Consider a queue within this model for which S_i are are geometrically distributed. We will refer to this queue as a $\cdot/Geo^{(b)}/1$ queue, with the superscript indicating that a batch of packets is served in each time slot. Suppose the number of arrivals A_i are independent and geometrically distributed. Some queueing theoretic results for this queue are obtained in Appendix B. These include steady state distributions for D_i and X_i, and in particular, the result that in steady state, the departures D_i are independent and geometrically distributed. It is also shown that for queues in which S_i have an arbitrary distribution, if A_i are i.i.d. and geometrically distributed, then

the steady state distribution of the queue size is geometric. ◁

Queueing theoretic results are primarily steady state results on the behavior of the network for certain kinds of packet arrival processes at the nodes in the network. From the point of view of a particular source-destination pair, however, the interesting question is that of choosing the packet sequence and the times at which to send the packets to the network for each message, so as to achieve the desired information rate while satisfying the reliability and delay requirements. In [4] as well as in this dissertation, queueing theoretic steady state results are used to establish limits on the information rates that can be achieved over single-server queues.

1.4 Other Related Work

There are some other information theoretic aspects of timing that have been pointed out in the literature, but these will not be of primary interest in this dissertation. We present a brief review of these for completeness.

Gallager's work [17] prompted a considerable amount of related work in the late 1970s, which did not result in any major fundamental new ideas, however. Among these are a short paper by Camrass and Gallager [8] which discusses two different packetization schemes for data link layers and their relation to an optimal encoding scheme for messages with geometrically distributed lengths [21], and another by Humblet [24], which points out some information theoretic aspects of addressing for multiple access channels. A paper by Gallager [18] discusses some information theoretic aspects of the use of flags, efficiency of ARQ protocols, and other techniques used in data link layers. Much of the material of this paper is also substantially covered in [6, Ch. 2].

Another problem in which timing plays a crucial role is the classical multiple-access collision channel. The collision channel was first introduced as a model for a wireless packet radio network by Abramson [2], and has become a canonical model

for random accessing of a single shared channel by multiple transmitters who have no information about the presence of messages at any other transmitter. In this model of a multiple-access channel, each station transmits an information packet as soon as it has a message available to transmit. If multiple transmitters access the channel simultaneously, a *collision* occurs, and all the packets are lost. If no other transmitter transmits while a particular station is transmitting a packet, that packet is received error-free. The conventional assumptions in the analysis of this channel consist of a tractable stochastic model for the message arrival times at each transmitter, with each message considered as being equivalent to single packet transmitted on the channel, and the availability of instant feedback to all transmitters regarding the status of each packet transmission. A large amount of analysis [9, 43, 3, 35] has been conducted on efficient algorithms to resolve collisions and schedule retransmissions for packets that experience collisions, and on the stability of these algorithms. Much of this analysis is reviewed in [6, Ch. 4]. This classical analysis has more of a queueing theoretic and dynamic systems flavor, rather than an information theoretic style. A juxtaposition of these studies with multiple-user information theory was presented in [19].

The assumption that each transmitter transmits a message as soon as one is available treats the timing of the transmissions as a factor that is beyond the transmitters' control. In [30], Massey and Mathys showed that if the transmitters were allowed to choose the packet transmission times, the sequence of packet transmission times could be chosen so that the identities, or addresses, of the transmitters of all successfully received packets could be determined by the receiver purely from the times at which the channel was accessed. They made the information theoretic assumption that each of the users of an N-user collision channel had an infinite reservoir of data to send. They further assumed that there was complete lack of synchronism between the users, and that there was no feedback available. They demonstrated a construction of an explicit set of sequences of channel accessing times, one for each user, with the property that the transmitter of each successfully transmitted packet can be correctly

identified, irrespective of the actual time differences between the users. Their work raises intriguing questions about the relation between the amount of addressing information that must be supplied by users of a multiple-access channel, and the amount of information that can be encoded in the times at which users access the channel.

The connection between timing and addressing is also evident in time-division multiplexing (TDM) systems, in which a particular sequence of slots is allotted to pre-assigned to all sources that send their packets to a multiplexer. The particular slot in which a packet is transmitted immediately identifies the address of the source and the destination, and no further addressing overhead is required within the packet itself. The work of [30] is thus an important extension of this idea to multiple-access collision channels.

Timing information has also been analyzed in the context of covert communication between users of a resource-sharing computer system. Consider a user A of such a system communicating with another user B of the system in the following manner. B submits a job to the (shared) CPU of the system, and measures the amount of time taken by the CPU to complete the job. A can influence the response time of the CPU to B's job by submitting a job of its own. The distribution of the completion time of B's job depends on whether or not A submitted a job to the system at that time. Thus the submission of a job by A, or lack thereof, allows A to communicate a '0' or a '1' to B. Such a channel can be called a *timing channel*, because the output symbols of the channel observed by the receiver B are the times taken for a job to be completed, and the particular job itself is irrelevant. It is assumed that the receiver B interrupts its job as soon as it makes a decision on the symbol that A is attempting to send, and B immediately submits another job. It is also assumed that there is instantaneous feedback to A about this action being taken by B, so that A is immediately free to transmit the next symbol to influence the response time to B's new job. In [32], the inputs to the timing channel are 0 and 1, while the output is a non-negative real-valued time, and the distribution of the output time is a shifted

exponential distribution with different shifts depending on whether the input is 0 or 1. The timing channel considered in [31] also has inputs 0 and 1, but the output time values are restricted to the set $\{1, 2\}$. If the input is 0, the output time is always 2, while if the input is 1, the output may be either 1 or 2.

The capacity of these covert timing channels is analyzed in [32, 31] in terms of bits per transmitted symbol, as well as bits per unit time. The key assumption that the receiver interrupts its job as soon as it has decided the transmitter's symbol, and that the transmitter knows the times at which the receiver has performed this action. This assumption helps remove the time correlation between completion times of successive jobs, and converts the channel into a discrete-time channel with variable output times. It also eliminates the possibility that more than one of A's symbols can simultaneously influence the response time of a particular job submitted by B. In contrast, in the problem of the timing capacity of queueing systems, neither the transmitter nor the receiver is aware of the number of packets in the queue at any time. Thus the analysis of the capacity of covert timing channels in [32, 31] is weaker than the analysis of the timing capacity of the single-server queue in [4] and that presented in this dissertation. A more difficult problem would be the situation where the transmitter A does not have any feedback about the times at which the receiver B makes the decision on a particular transmitted symbol. This would give rise to interference between the various symbols transmitted by A. It is not hard to see that the problem of communicating reliably in this situation is the processor-sharing analogue of the problem of the timing capacity of the single-server queue analyzed in [4].

1.5 Contributions and Outline of the Dissertation

We have seen in the preceding sections that although information theory has largely been unable to capture the essence of the interplay between timing and network

performance, some seminal ideas have been proposed that hold the promise of illuminating the role of timing in an information theoretic foundation for network analysis. One of these is the idea of encoding information in the timing of packets on a network. Another is the fact that a certain amount of information about message arrival times perforce has to be transmitted in order to enable a receiver to decode messages within a finite delay. In this dissertation, we build on and establish connections between these seminal ideas about the role of timing information in data networks. Our focus will be on analyzing more complex models and examining the implications of these ideas in scenarios that arise in current high-speed networks.

High-speed network systems like ATM that use fixed-size packets or cells are well-modeled as discrete-time systems because the packet transmission time provides a convenient unit of time, called a *slot*, in terms of which the operation of the system can be described [7]. In Chapters 2 and 3, we focus on an information theoretic analysis of discrete-time queueing systems. Consider a particular virtual circuit connection over an ATM switch that switches packets arriving at its input ports to the appropriate output ports and buffers cells at each input port. From the point of view of the source and destination associated with the connection under consideration, the switch behaves like a discrete-time queue to which packets or cells are sent by the source. The cell at the head of the queue has to wait for a random amount of time before it gets switched to its appropriate destination. As described in Examples 1.2 and 1.3, discrete-time queues can have batches of arrivals and departures in a slot, and in some cases the arriving and departing batch sizes may be restricted to 1.

Following the classical information theory model and the work of [4], we assume that the source has an infinite reservoir of data to send, and concentrate on the effect of the random times spent by packets in the buffer. In Chapter 2, we analyze the timing capacity of discrete-time queues in which at most one packet can arrive and one packet can depart in a slot, and each packet experiences i.i.d. service times. We show that for a given packet departure rate, among all such queues having a given

mean service time, the queue with geometric service time distribution has the least λ-timing capacity, and we explicitly obtain this capacity. This result is the discrete-time analogue of the extremal nature of the single-server exponential queue established in [4]. It is also of interest in its own right as it provides an explicit formula of the capacity of a certain binary channel with infinite memory.

In Chapter 3, we analyze the timing capacity of the discrete-time queues with batch-arrivals and batch service within the queueing model described in Example 1.3. The behavior of such queues is characterized by the number of packets that the server can serve in each slot. A natural way to transmit information through such queues is by encoding it in the number of packets that arrive at the queue in each slot, rather than in the interarrival times of the packets. The receiver observes the number of packets served by the queue in each slot and attempts to infer the transmitted codeword. The randomness in the number of packets served in each slot distorts the transmitted information. We analyze the amount of information that can be reliably encoded in the number of packets sent to the queue. We establish an upper bound on the λ-timing capacity of queues within this model, and show that this bound is tight in the case of the queue that serves a geometrically distributed number of packets in a slot. For queues with an arbitrary service distribution within this model, we obtain a lower bound to the λ-timing capacity in terms of a certain parameter that determines the steady state distribution of the queue length in such queues for a certain class of arrival processes. We show that the extremal nature of the geometric server within this model is related to this parameter, rather than to the mean number of packets that the queue can serve per slot. Thus, while the analysis of the discrete-time queues with single-arrivals is similar to the analysis of continuous-time queues in [4], the analysis of the discrete-time queues with batch arrivals presents new challenges, and gives quite different results.

In Section 3.8, we consider the generalization of the discrete-time single-server queues analyzed in Chapter 2 to multiple servers, each of which can serve at most one

packet in each slot. It was noted in [14] that the timing capacity of queues with multiple servers is significantly more difficult to analyze than that of single server queues. For the case of K servers with i.i.d. geometrically distributed service times, we obtain an upper bound to the λ-timing capacity in terms of the capacity of the discrete memoryless *K-output binomial channel* with an output constraint. This upper bound can be evaluated numerically for finite K. For the case $K = \infty$, we establish the existence of a distribution that achieves the capacity of the infinite output binomial channel subject to the output constraint. We also obtain a necessary and sufficient condition that characterizes this distribution using the generalized Kuhn-Tucker theorem. The actual solution of these conditions to find the capacity achieving distribution for the infinite output binomial channel remains elusive.

In Chapter 4, we turn to the problem of quantifying the information about random message arrival times that must be transmitted to enable the receiver to decode the messages within a finite amount of time, and its relation to the transmission of information through the timing of packets. Gallager's analysis [17] had allowed for the possibility that the messages may not be decoded at the receiver in the same order they arrived at the source. We focus on the situation in which messages are transmitted one at a time in order and are decoded at the receiver in the same order they arrived at the source, for instance a virtual circuit network or a point-to-point channel. In this situation, the sequence of message decoding times can be considered as the departure process from a first-in-first-out single server queue whose arrival process is the message arrival process. With this point of view we can associate a service time with each message, and the average service time per message can be thought of as a measure of the distortion between the message arrival and decoding processes. If the messages are transmitted one at a time over the point-to-point channel, then the service time of a message is precisely the time available for transmitting useful information about that message to the receiver, and thus directly influences the achievable probability of message error. In this sense the service time distortion measure does

have a physical significance. Consider any such strategy for which the probability that the service time per message being more than d is vanishingly small. We show that the amount of information about message arrival times that the receiver must receive is lower bounded by the rate-distortion function for encoding the sequence of message arrival times so as to generate an order-preserving reproduction such that the service time distortion per message between the two sequences is limited to d. We explicitly obtain this rate distortion function for the case where the message arrival process is Poisson, and show that the optimal source encoding is the single server queue with i.i.d. exponentially distributed service times. This rate distortion function is actually identical to the rate distortion function of the Poisson process under a different distortion measure identified by Verdu [45], who had also pointed out its similarities to the canonical rate-distortion function of a sequence of i.i.d. Gaussian random variables under a squared error distortion measure.

We conclude the dissertation with a summary, directions for future work, and an interesting conjecture in Chapter 5.

Chapter 2

THE TIMING CAPACITY OF DISCRETE-TIME QUEUES WITH SINGLE ARRIVALS

In this chapter, we analyze the timing capacity of the discrete-time analogue of the continuous-time single-server queue analyzed in [4]. In this queueing model, at most one packet can arrive and at most one packet can depart in each time slot. Service times of packets are independent, identically distributed (i.i.d.) integer-valued random variables. Such a discrete-time single server queue can be used to model a single virtual circuit connection across an ATM switch [7], in which packets are delivered to the destination port in the same order they arrive at the source port.

We obtain upper and lower bounds to the λ-timing capacity of such queues. In particular, we show that among all queues having a given mean service time within this model, the queue with geometric service time distribution has the least λ-timing capacity[1], which is given by

$$C(\lambda) = h(\lambda) - \frac{\lambda}{\mu} h(\mu)$$

where $h(\cdot)$ is the binary entropy function. For the geometric server, it is shown, using the results of [47], that the λ-timing capacity can be achieved by a random code generated by independent, geometrically distributed packet interarrival times. The timing capacity is shown to be the supremum of λ-timing capacities over $0 \leq \lambda \leq \mu$, and is given by

$$C = \log \left[1 + \mu(1-\mu)^{\frac{1-\mu}{\mu}} \right].$$

[1] For discrete-time queues, the units of capacity will be bits/slot if the base of the logarithm is 2.

2.1 The Model

In this model of a single-server queue, arrivals and departures can occur only at integer-valued epochs, called slots. At most one packet can arrive and at most one packet can depart in each slot. Let A_i denote the interarrival time (in slots) between the arrival of the $(i-1)^{\text{th}}$ and the i^{th} packet, D_i the interdeparture time between the departure of the $(i-1)^{\text{th}}$ and the i^{th} packet, and S_i the service time of the i^{th} packet. Since at most one arrival and one departure can occur in each slot, $A_i \geq 1, D_i \geq 1$. Service times S_i are assumed to be i.i.d. A packet that arrives in slot i can depart no earlier than slot $i+1$, so $S_i \geq 1$, and $\mu = 1/E[S_i] \leq 1$. Let W_i be the time in slots for which the server is idle after the departure of the $(i-1)^{\text{th}}$ packet until the arrival of the i^{th} packet. $W_i = 0$ if the i^{th} packet arrives before the departure of the $(i-1)^{\text{th}}$ packet. Then $D_i = W_i + S_i$.

We will denote the geometric distributions on the non-negative and positive integers by $\text{Geo}(x)$ and $\text{Geo}^+(x)$ respectively, defined in Notations A.2 and A.3 in Appendix A.

Example 2.1: Consider a queue for which S_i have the $\text{Geo}^+(\mu)$ distribution. We will refer to this queue as a $\cdot/Geo^{(s)}/1$ queue, with the superscript indicating that at most a single packet can be served in a time slot. The steady state queueing behavior of this queue is analyzed in [42]. If the arrival process to this queue has independent $\text{Geo}^+(\lambda)$ interarrival times, then it can be shown [42, p. 13] that in steady state, the departure process also has independent, $\text{Geo}^+(\lambda)$ interdeparture times. ◁

We will assume at present that the queue is initially empty. In Section 2.4 it will be shown that, as in the case of the continous time queue [4], the timing capacity with the queue initially in equilibrium is no more than the timing capacity with the queue initially empty.

2.2 Timing Codes

An (n, M, T, ϵ) timing code consists of M codewords with the following properties.

(i) Each codeword is a sequence of n packets, arriving at the queue with interarrival times $A_1, \cdots, A_n$.

(ii) The decoder observes the sequence of interdeparture times $D_1, \cdots, D_n$, and attempts to infer the transmitted codeword.

(iii) The average departure time of the n^{th} packet is T, i.e. $E\left[\sum_{i=1}^{n} D_i\right] = T$, where the average is taken over all codewords as well as over the service times of the packets.

(iv) The average probability of error for the code is ϵ.

An (n, M, T, ϵ) timing code has information rate $R = (\log M)/T$. Rate R is said to be *achievable* if there exists a sequence $(n, M_n, T_n, \epsilon_n)$ of timing codes that has a subsequence $(n_k, M_{n_k}, T_{n_k}, \epsilon_{n_k})$ for which

$$\lim_{k\to\infty} \frac{\log M_{n_k}}{T_{n_k}} = R \quad \text{and} \quad \lim_{k\to\infty} \epsilon_{n_k} = 0.$$

The timing capacity C is the supremum of the set of achievable rates.

Rate R is said to be *achievable at departure rate* λ if there exists a sequence $(n, M_n, T_n, \epsilon_n)$ of timing codes that has a subsequence $(n_k, M_{n_k}, T_{n_k}, \epsilon_{n_k})$ for which

$$\lim_{k\to\infty} \frac{\log M_{n_k}}{T_{n_k}} = R, \ \lim_{k\to\infty} \frac{n_k}{T_{n_k}} = \lambda, \ \text{and} \ \lim_{k\to\infty} \epsilon_{n_k} = 0. \tag{2.1}$$

The λ-timing capacity $C(\lambda)$ is the supremum of all rates that are achievable at departure rate λ.

These definitions are slightly different from the definitions used in [4]. We have defined an (n, M, T, ϵ) timing code as one for which the average departure time of

the n^{th} packet is equal to T, whereas in [4], T is an upper bound on the average departure time of the n^{th} packet. The latter seems to be somewhat unnatural and does not appear to have any bearing on the results. Under that definition, for instance, an (n, M, T, ϵ) timing code is also an $(n, M, 2T, \epsilon)$ timing code. Also in [4], rate R is defined to be achievable at departure rate λ if it is achievable by a sequence of $(n, M, n/\lambda, \epsilon)$ timing codes, whereas we require that the departure rate be λ only in the limit. This seems to be a reasonable requirement, and also affords some simplification in the proof of Proposition 2.1 below.

Following Anantharam and Verdu [4], we first show that to analyze the timing capacity of a queue, it is sufficient to analyze the λ-timing capacity, by the following result.

Proposition 2.1. *For a queue with an arbitrary service time distribution with mean $1/\mu$ slots.*

$$C = \sup_{0 \le \lambda \le \mu} C(\lambda). \tag{2.2}$$

Proof. From the definitions, it is clear that $C \ge C(\lambda)$ for all λ. Now there exists a sequence of $(n, M_n, T_n, \epsilon_n)$ timing codes that achieves C, i.e. that has a subsequence $(n_k, M_{n_k}, T_{n_k}, \epsilon_{n_k})$ such that

$$\lim_{k \to \infty} \frac{\log M_{n_k}}{T_{n_k}} = C \text{ and } \lim_{k \to \infty} \epsilon_{n_k} = 0.$$

The departure time of the last packet is at least equal to the sum of the service times of all the packets, hence $\sum_{i=1}^{n_k} D_i \ge \sum_{i=1}^{n_k} S_i$. This implies that $T_{n_k} \ge \sum_{i=1}^{n_k} E[S_i] = n_k/\mu$, and therefore $n_k/T_{n_k} \le \mu$. Hence $l = \limsup_{k \to \infty}(n_k/T_{n_k})$ exists, and $0 \le l \le \mu$. But since the limsup is itself a limit point, there is a further subsequence $\{n_{k_j}\}$ of $\{n_k\}$, for which

$$\lim_{j \to \infty} \frac{n_{k_j}}{T_{n_{k_j}}} = l, \ \lim_{j \to \infty} \frac{\log M_{n_{k_j}}}{T_{n_{k_j}}} = C, \text{ and } \lim_{j \to \infty} \epsilon_{n_{k_j}} = 0.$$

Therefore by definition, C is achievable at departure rate l for some $0 \leq l \leq \mu$, i.e. $C \leq C(l)$. Hence $C = \sup_{0 \leq \lambda \leq \mu} C(\lambda)$. □

2.3 Converse Theorem

Proposition 2.2. *Consider a queue whose service time distribution is F, and let $S \sim F$. Let $\mathcal{A}(x)$ be the set of probability distributions on the non-negative integers having mean x. For any probability distribution Q, let W_Q be a random variable that is independent of S, and such that $W_Q \sim Q$. Then for $\lambda \leq \mu$, the λ-timing capacity $C(\lambda)$ of the queue satisfies*

$$C(\lambda) \leq \lambda \sup_{Q \in \mathcal{A}(\frac{1}{\lambda} - \frac{1}{\mu})} I(W_Q; W_Q + S). \tag{2.3}$$

Proof. Consider a sequence of timing codes $(n, M_n, T_n, \epsilon_n)$ for which $\lim_{n\to\infty}(n/T_n) = \lambda$ and $\lim_{n\to\infty} \epsilon_n = 0$. For the timing code using n packets, let $A_1, \cdots, A_n$ be the interarrival times of packets $1, \cdots, n$, and let $D_1, \cdots, D_n$ be the respective interdeparture times. By Fano's lemma [10, Sec. 2.11] and the data processing theorem [10, Sec. 2.8], we have

$$\begin{aligned}
\log M_n &\leq \frac{1}{1-\epsilon_n}\left[\log 2 + I(A_1, \cdots, A_n; D_1, \cdots, D_n)\right] \\
&= \frac{1}{1-\epsilon_n}\left[\log 2 + H(D_1, \cdots, D_n)\right. \\
&\quad \left. - H(D_1, \cdots, D_n | A_1, \cdots, A_n)\right] \\
&\leq \frac{1}{1-\epsilon_n}\left[\log 2 + \sum_{i=1}^{n} H(D_i) - \sum_{i=1}^{n} H(D_i | A_1, \cdots, A_n, D_1, \cdots, D_{i-1})\right].
\end{aligned}$$

The interdeparture time D_i does not depend on the interarrival times $A_{i+1}, \cdots, A_n$. Given $A_1, \cdots, A_i$ and $D_1, \cdots, D_{i-1}$, the idle time W_i of the server between the departure of the $(i-1)^{\text{th}}$ packet and the arrival of the i^{th} packet is determined. Also W_i is a sufficient statistic of $A_1, \cdots, A_i, D_1, \cdots, D_{i-1}$ for D_i. Hence

$$H(D_i | A_1, \cdots, A_n, D_1, \cdots, D_{i-1}) = H(D_i | W_i).$$

Therefore

$$\begin{aligned}\frac{\log M_n}{T_n} &\leq \frac{1}{1-\epsilon_n}\frac{n}{T_n}\left[\frac{1}{n}\sum_{i=1}^{n} I(W_i; D_i) + \frac{\log 2}{n}\right] \\ &= \frac{1}{1-\epsilon_n}\frac{n}{T_n}\left[\frac{1}{n}\sum_{i=1}^{n} I(W_i; W_i + S_i) + \frac{\log 2}{n}\right].\end{aligned}$$

Define

$$\beta_S(x) = \sup_{Q \in \mathcal{A}(x)} I(W_Q; W_Q + S),$$

where $\mathcal{A}(x)$ is as defined earlier. First we show that $\beta_S(\cdot)$ is a concave function. Let $X_1, X_2 \geq 0$ be integer-valued random variables, independent of S, with means x_1 and x_2, and probability distributions P_1 and P_2 respectively. For $0 \leq \alpha \leq 1$, consider a random variable X that is independent of S and has distribution $P = \alpha P_1 + (1-\alpha)P_2$. Then $E[X] = \alpha x_1 + (1-\alpha)x_2$. For a given random variable S, the mutual information function $I(X; X+S)$ is a concave function of the distribution of X [16, Section 4.4], hence $I(X; X+S) \geq \alpha I(X_1; X_1+S) + (1-\alpha) I(X_2; X_2+S)$. The concavity of $\beta_S(\cdot)$ follows.

From the definition of $\beta_S(\cdot)$, $I(W_i; W_i + S_i) \leq \beta_S(E[W_i])$. Using this and the concavity of $\beta_S(\cdot)$, we have

$$\frac{1}{n}\sum_{i=1}^{n} I(W_i; W_i + S_i) \leq \beta_S\left(\frac{1}{n}\sum_{i=1}^{n} E[W_i]\right)$$

Also, the fact that the average departure time of the last packet is T_n puts a constraint on $E[W_i]$:

$$\frac{1}{n}\sum_{i=1}^{n} E[W_i] = \frac{T_n}{n} - \frac{1}{\mu}.$$

Therefore

$$\frac{\log M_n}{T_n} \leq \frac{1}{1-\epsilon_n}\frac{n}{T_n}\left[\beta_S\left(\frac{T_n}{n} - \frac{1}{\mu}\right) + \frac{\log 2}{n}\right].$$

Taking the limit as $n \to \infty$ we obtain (2.3). □

Next we address the question of evaluating the function $\beta_S(\cdot)$.

Let $\overline{S}$ be a random variable having the Geo$^+$ (μ) distribution. Consider an integer-valued non-negative random variable $\overline{W}$ that is independent of $\overline{S}$, and has the distribution

$$P_{\overline{W}}(k) = \begin{cases} \rho, & k = 0 \\ (1-\rho)(1-\lambda)^{k-1}\lambda, & k = 1, 2, \cdots \end{cases}$$

where $\rho = \lambda/\mu$. Thus $E[\overline{W}] = \frac{1}{\lambda} - \frac{1}{\mu}$. It can be easily verified, e.g. using probability generating functions, that $\overline{W} + \overline{S}$ has the Geo$^+$ (λ) distribution. The following proposition shows that this distribution on $\overline{W}$ maximizes $I(W; W+\overline{S})$ over all distributions on W with mean $\frac{1}{\lambda} - \frac{1}{\mu}$.

Proposition 2.3. *Let $\overline{S}$ and $\overline{W}$ be as defined above. Let W be any integer-valued non-negative random variable independent of $\overline{S}$ with $E[W] = \frac{1}{\lambda} - \frac{1}{\mu}$. Then*

$$I(W; W+\overline{S}) \leq I(\overline{W}; \overline{W}+\overline{S}) = \frac{h(\lambda)}{\lambda} - \frac{h(\mu)}{\mu}.$$

Proof. Among all positive integer-valued random variables with a given mean, the geometric distribution has the maximum entropy [10, Sec. 11.1]. Since $\overline{W} + \overline{S}$ is geometrically distributed, $H(\overline{W} + \overline{S}) \geq H(W + \overline{S})$. Hence

$$\begin{aligned} I(W; W+\overline{S}) &= H(W+\overline{S}) - H(W+\overline{S}|W) \\ &= H(W+\overline{S}) - H(\overline{S}) \\ &\leq H(\overline{W}+\overline{S}) - H(\overline{S}) \\ &= I(\overline{W}; \overline{W}+\overline{S}) \\ &= \frac{h(\lambda)}{\lambda} - \frac{h(\mu)}{\mu}. \end{aligned}$$

The last equality can be easily verified by direct computation. □

In conjunction with Proposition 2.2, this gives the following upper bound on the λ-timing capacity of a $\cdot/Geo^{(s)}/1$ queue with mean service time $1/\mu$ slots:

$$C(\lambda) \leq h(\lambda) - \frac{\lambda}{\mu} h(\mu). \tag{2.4}$$

2.4 *Achievability Theorem*

We will now show that for a queue with geometric service time distribution, the upper bound on the λ-timing capacity given by (2.4) can be achieved. We will also show that among all queues with a given mean service time, the queue with geometric service time distribution has the least λ-timing capacity.

To show the achievability, we will use the device developed in [4] for showing that the λ-timing capacity of the queue with an initially empty buffer is achievable even if the queue is initially in equilibrium with respect to some arrival process. The definitions in Section 2.2 can be generalized to a queue initially in equilibrium. For an (n, M, T, ϵ) code with the queue initially in equilibrium, n is the number of packets sent by the transmitter. Since there are a random number of packets initially in the queue, the receiver may at first receive some spurious packets. Let us assume at present that the packets contain some marker that allows the receiver to distinguish between these spurious packets and those actually sent by the transmitter. The receiver then keeps track of the departure times of the packets initially in the queue, as well as those of the n packets sent by the transmitter. With this convention, for an (n, M, T, ϵ) timing code, T refers to the average departure time of the n^{th} packet sent by the transmitter. As before, M is the number of messages that can be transmitted, and ϵ the average probability of error for the code. For T and ϵ, the averages are taken over all the codewords as well as over the initial state of the queue.

Suppose that rate R is achievable at departure rate λ when the queue is initially in equilibrium with respect to a rate λ' arrival process, for some $\lambda' < \mu$. That is, with the queue initially in equilibrium, there exists a sequence of timing codes $(n, M_n, T_n, \epsilon_n)$

that has a subsequence $(n_k, M_{n_k}, T_{n_k}, \epsilon_{n_k})$ for which (2.1) is satisfied. In equilibrium, the queue is empty with probability $1 - \rho'$, where $\rho' = \lambda'/\mu$. Let $(\epsilon_{\text{empty}})_{n_k}$ be the probability of error of the n_k^{th} timing code from the above subsequence given that the queue is initially empty, and let $(\epsilon_{\text{non-empty}})_{n_k}$ be the probability of error of the code given that the queue is initially non-empty. Then $\epsilon_{n_k} = (1 - \rho')(\epsilon_{\text{empty}})_{n_k} + \rho'(\epsilon_{\text{non-empty}})_{n_k}$, hence $(\epsilon_{\text{empty}})_{n_k} \leq \epsilon_{n_k}/(1 - \rho')$. Hence if the same timing codes are used with the queue initially empty, on the subsequence n_k for which $\epsilon_{n_k} \to 0$, we have $(\epsilon_{\text{empty}})_{n_k} \to 0$.

Let $(T_{\text{empty}})_{n_k}$ be the average departure time of the n_k^{th} packet sent by the transmitter when the code is used with the queue initially empty. Clearly this is no more than the average departure time of the last packet when the queue is in equilibrium, i.e. $(T_{\text{empty}})_{n_k} \leq T_{n_k}$. Let T_0 be the average departure time of the last packet initially in the queue when the queue is in equilibrium. Then $T_{n_k} - T_0 \leq (T_{\text{empty}})_{n_k} \leq T_{n_k}$, hence

$$\lim_{k\to\infty} \frac{T_{n_k} - T_0}{n_k} \leq \lim_{k\to\infty} \frac{(T_{\text{empty}})_{n_k}}{n_k} \leq \lim_{k\to\infty} \frac{T_{n_k}}{n_k} = \frac{1}{\lambda}.$$

Since T_0 is finite, the limiting packet departure rate if the queue is initially empty is also λ. It follows that rate R is also achievable at departure rate λ when the queue is initially empty.

Let us now suppose that the packets do not have any marker, so that the receiver cannot distinguish the transmitter's packets from those initially in the queue. The receiver then keeps track of the departure times of the first n packets that depart from the queue. Thus the receiver may initially get some spurious packets and may ignore some of the transmitter's packets. For an (n, M, T, ϵ) timing code for this case, T refers to the average departure time of the n^{th} packet departing from the queue. It is easy to show that if rate R is achievable at packet rate λ in this situation, it is also achievable at packet rate λ when the receiver is able to distinguish the transmitter's packets from the packets initially in the queue, and hence also achievable at packet

rate λ when the queue is initially empty. We will use this to lower-bound the λ-timing capacity of the queue with geometric service time distribution in the following proposition.

Proposition 2.4. *For the $\cdot/Geo^{(s)}/1$ queue with mean service time $1/\mu$ slots and $0 < \lambda < \mu$,*

$$C(\lambda) \geq h(\lambda) - \frac{\lambda}{\mu} h(\mu). \tag{2.5}$$

Proof. Let the queue initially be in equilibrium with respect to a rate λ process with independent geometrically distributed interarrival times. Let the receiver begin observing the departures from time $t = 1$, and keep track of the first n departures from the queue. Due to the randomness in the number of packets initially in the queue, the receiver may initially receive some spurious packets not sent by the transmitter.

We want to show that the λ-timing capacity is at least $h(\lambda) - \frac{\lambda}{\mu} h(\mu)$. From the results of [47], it suffices to show that for some arrival process $\{A_i\}$ and the corresponding departure process $\{D_i\}$, the liminf in probability of the normalized input-output information density is at least $h(\lambda)/\lambda - h(\mu)/\mu$, i.e. to show that for some arrival process $\{A_i\}$,

$$P\left[\frac{1}{n} i_{A^n;D^n}(A^n; D^n) < \frac{h(\lambda)}{\lambda} - \frac{h(\mu)}{\mu}\right] \longrightarrow 0 \tag{2.6}$$

as $n \to \infty$, where $A^n = (A_1, \cdots, A_n)$, $D^n = (D_1, \cdots, D_n)$, and

$$i_{A^n;D^n}(A^n; D^n) = \log \frac{P_{D^n|A^n}(D_1, \cdots, D_n | A_1, \cdots, A_n)}{P_{D^n}(D_1, \cdots, D_n)}.$$

Our choice of arrival process $\{A_i\}$ has independent, $\text{Geo}^+(\lambda)$ interarrival times. It is known in queueing theory that with this arrival process, in steady state, the interdeparture times D_i are also independent with the $\text{Geo}^+(\lambda)$ distribution [42, p. 13].

Let the number of packets initially in the queue be X_0. We have

$$\begin{aligned} P_{D^n|A^n}(D^n|A^n) &= \sum_{k=0}^{\infty} P_{X_0}(k) P_{D^n|A^n,X_0}(D^n|A^n,k) \\ &\geq P_{X_0}(0) P_{D^n|A^n,X_0}(D^n|A^n,0) \\ &= P_{X_0}(0) \prod_{i=1}^{n} P_{D_i|A^i,D^{i-1},X_0}(D_i|A^i,D^{i-1},0) \\ &= P_{X_0}(0) \prod_{i=1}^{n} P_{D_i|W_i}(D_i|W_i), \end{aligned}$$

where W_i is the idle time of the server after the departure of the $(i-1)^{\text{th}}$ packet until the arrival of the i^{th} slot, computed assuming that the queue is initially empty ($X_0 = 0$). Now $P_{D_i|W_i}(j|k) = (1-\mu)^{j-k-1}\mu$. Hence

$$\begin{aligned} \frac{1}{n} i_{A^n;D^n}(A^n;D^n) &\geq \frac{1}{n} \log \frac{P_{X_0}(0) \prod_{i=1}^{n} (1-\mu)^{S_i-1}\mu}{\prod_{i=1}^{n} (1-\lambda)^{D_i-1}\lambda} \\ &= \log \frac{\mu}{\lambda} + \log \frac{1-\lambda}{1-\mu} + \log(1-\mu) \left(\frac{1}{n} \sum_{i=1}^{n} S_i \right) \\ &\quad - \log(1-\lambda) \left(\frac{1}{n} \sum_{i=1}^{n} D_i \right) + \frac{1}{n} \log P_{X_0}(0). \end{aligned} \tag{2.7}$$

As $n \to \infty$, $\frac{1}{n}\sum_{i=1}^{n} D_i \to 1/\lambda$, and $\frac{1}{n}\sum_{i=1}^{n} S_i \to 1/\mu$ in probability. Also, since $P_{X_0}(0) > 0$, $\frac{1}{n}\log P_{X_0}(0) \to 0$. Hence as $n \to \infty$, the right side of (2.7) converges in probability to

$$\begin{aligned} &\log \frac{\mu}{\lambda} + \log \frac{1-\lambda}{1-\mu} + \frac{1}{\mu}\log(1-\mu) - \frac{1}{\lambda}\log(1-\lambda) \\ &= \frac{h(\lambda)}{\lambda} - \frac{h(\mu)}{\mu}. \end{aligned}$$

This establishes (2.6) and the proposition. □

Combining (2.4) and (2.5), we obtain the λ-timing capacity of the $\cdot/Geo^{(s)}/1$ queue with mean service time $1/\mu$ slots:

$$C(\lambda) = h(\lambda) - \frac{\lambda}{\mu} h(\mu). \tag{2.8}$$

From Proposition 2.1, the timing capacity is the supremum of the λ-timing capacities over $0 \leq \lambda \leq \mu$. This supremum occurs at

$$\lambda^* = \frac{1}{1 + \left[\mu(1-\mu)^{\frac{1-\mu}{\mu}}\right]^{-1}},$$

and the timing capacity is given by

$$C = \log\left[1 + \mu(1-\mu)^{\frac{1-\mu}{\mu}}\right]. \tag{2.9}$$

The following proposition shows that the λ-timing capacity of a queue with arbitrary service time distribution is at least as much as that of the $\cdot/Geo^{(s)}/1$ queue with the same mean service time.

Proposition 2.5. *For a queue having an arbitrary service time distribution with mean $1/\mu$ slots and $0 < \lambda < \mu$,*

$$C(\lambda) \geq h(\lambda) - \frac{\lambda}{\mu}h(\mu). \tag{2.10}$$

Proof. The proof follows the one given in [4] for establishing the analogous result with continuous-time queues. Consider the queue to be initially in equilibrium with a rate λ process having independent geometrically distributed packet arrivals. We want to show that for some arrival process $\{A_i\}$ and the corresponding departure process $\{D_i\}$,

$$\lim_{n\to\infty} P\left[\frac{1}{n} i_{A^n;D^n}(A^n; D^n) < \frac{h(\lambda)}{\lambda} - \frac{h(\mu)}{\mu}\right] \longrightarrow 0. \tag{2.11}$$

Let $Q_{D^n}(D^n)$ and $Q_{D^n|A^n}(D^n|A^n)$ be, respectively, the joint distribution of the departures and the joint distribution of the departures conditioned on the arrivals. By definition,

$$i_{A^n;D^n}(A^n; D^n) = \log \frac{Q_{D^n|A^n}(D_1, \cdots, D_n|A_1, \cdots, A_n)}{Q_{D^n}(D_1, \cdots, D_n)}.$$

Let us continue using the notation $P_{D^n}(D^n)$ and $P_{D^n|A^n}(D^n|A^n)$ for the corresponding distributions for the $\cdot/Geo^{(s)}/1$ queue with mean $1/\mu$ slots.

Our choice of arrival process will again be the process that has independent, $\text{Geo}^{-}(\lambda)$ interarrival times. Now

$$\begin{aligned}
\frac{1}{n} i_{A^n;D^n}(A^n;D^n) &= \frac{1}{n}\log\frac{Q_{D^n|A^n}(D^n|A^n)}{Q_{D^n}(D^n)} \\
&= \frac{1}{n}\log\frac{Q_{D^n|A^n}(D^n|A^n)P_{D^n}(D^n)}{P_{D^n|A^n}(D^n|A^n)Q_{D^n}(D^n)} + \frac{1}{n}\log\frac{P_{D^n|A^n}(D^n|A^n)}{P_{D^n}(D^n)} \\
&= \frac{1}{n}\log\frac{Q_{A^n|D^n}(A^n|D^n)}{P_{A^n|D^n}(A^n|D^n)} + \frac{1}{n}\log\frac{P_{D^n|A^n}(D^n|A^n)}{P_{D^n}(D^n)}.
\end{aligned}$$

The liminf in probability of the first term is non-negative, because

$$\begin{aligned}
P\left[\frac{1}{n}\log\frac{Q_{A^n|D^n}(A^n|D^n)}{P_{A^n|D^n}(A^n|D^n)} \le -\delta\right] &= \sum_{a^n}\sum_{d^n} Q_{A^n|D^n}(a^n|d^n)Q_{D^n}(d^n)\cdot \\
&\qquad \mathbf{1}\left[Q_{A^n|D^n}(a^n|d^n) \le \exp(-n\delta)P_{A^n|D^n}(a^n|d^n)\right] \\
&\le \sum_{a^n}\sum_{d^n} \exp(-n\delta)P_{A^n|D^n}(a^n|d^n)Q_{D^n}(d^n) \\
&= \exp(-n\delta).
\end{aligned}$$

For the second term, it is true that

$$P\left[\frac{1}{n}\log\frac{P_{D^n|A^n}(D^n|A^n)}{P_{D^n}(D^n)} < \frac{h(\lambda)}{\lambda} - \frac{h(\mu)}{\mu}\right] \longrightarrow 0.$$

This can be shown in the same way as in the proof of Proposition 2.4, because in that proof we only made use of the facts that $\frac{1}{n}\sum_{i=1}^{n} D_i \to 1/\lambda$ and $\frac{1}{n}\sum_{i=1}^{n} S_i \to 1/\mu$, and these are true for any stable queue. This establishes (2.11) and (2.10). □

2.5 *Discussion*

Note that encoding information in the arrival times of the packets is equivalent to encoding information in the presence or absence of a packet in a slot. Consider a discrete-time channel in which the input in a slot is 1 if a packet arrives at the $\cdot/Geo^{(s)}/1$ queue in that slot, and 0 if no packet arrives. The output of the channel is 1 in those slots in which a packet departs, and 0 in the other slots. This is a binary-input binary-output channel with infinite memory, and its capacity is the same as the timing capacity (2.9) of the $\cdot/Geo^{(s)}/1$ queue.

Figure 2.1 shows a plot of the timing capacity of the $\cdot/Geo^{(s)}/1$ queue, given by (2.9), as a function of the mean service rate. The timing capacity is an increasing function of the service rate μ, and reaches a maximum of 1 bit/slot at $\mu = 1$ packet/slot.

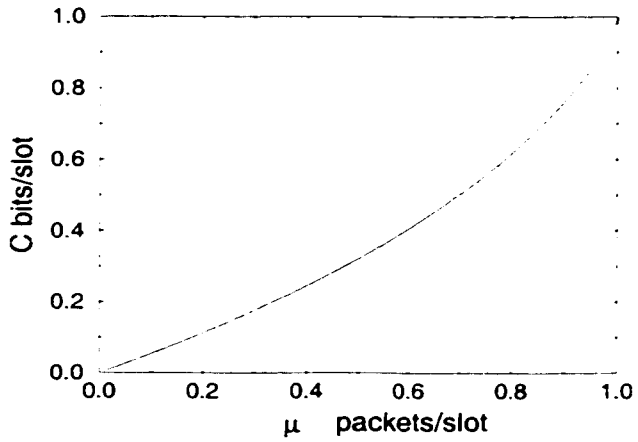


Figure 2.1: Timing capacity as a function of service rate for the $\cdot/Geo^{(s)}/1$ queue.

In (2.8), if we let the slot size go to zero, and examine the λ-timing capacity per unit time, we recover the expression for the λ-timing capacity of the continuous-time queue with exponential service time distribution obtained in [4]:

$$C(\lambda) = \lambda \log \frac{\mu}{\lambda}.$$

Thus the result for the discrete-time $\cdot/Geo^{(s)}/1$ queue agrees with the result for the continuous time queue with exponential service time distribution in the limit of small slot size.

For the $\cdot/Geo^{(s)}/1$ queue, the timing capacity is a convex function of μ. This gives rise to the following interesting possibility. Suppose that a server of total service rate μ is to be shared between two independent users, in such a way that the first user sees a service rate of μ_1, and the second user sees a service rate of μ_2, where $\mu_1 + \mu_2 = \mu$. Suppose further that each user sees a geometric service time distribution independent

of the other user. Then the maximum information rate that the two users can achieve are $C(\mu_1)$ and $C(\mu_2)$ respectively. Suppose the system is to be shared in such a way that the total information rate achievable is maximum, i.e. we want to assign μ_1 and μ_2 so as to maximize $C(\mu_1) + C(\mu_2)$, subject to $\mu_1 + \mu_2 = \mu$. Because C is a convex non-negative function with $C(0) = 0$, we get $C(\mu_1) + C(\mu_2) \leq C(\mu)$, so that the optimal assignment is $\mu_1 = 0, \mu_2 = \mu$ or $\mu_1 = \mu, \mu_2 = 0$. That is, only one of the users uses the system at a given time, and they time-share by using the system at the full rate μ but only for an appropriate fraction of the time. When user 1 has the use of the system, the queue is driven at the optimal rate λ^* corresponding to μ, and at the end of the transmission, leaves the system in equilibrium for user 2. User 2 then uses a timing code assuming that the queue is in equilibrium, and there is no loss of capacity when the queue is initially in equilibrium.

Sharing the system by rates μ_1 and μ_2 is in a sense equivalent to frequency division multiplexing (FDM), whereas letting only one user use the system at a given time corresponds to time division multiplexing (TDM). Thus for this system, TDM is better than FDM as far as timing capacity is concerned.

2.6 Conclusion

In this chapter, we analyzed the timing capacity of discrete-time queues in which at most one packet may arrive and at most one packet may finish service in a slot. We established the canonical role of the queue with geoemtric service time distibution for such queues. In the next chapter, we will examine the timing capacity of queues that have batch arrivals and a batch service mechanism. We will also look at the generalization of the single-service queues considered in this chapter to queues that have multiple independent servers.

Chapter 3

THE TIMING CAPACITY OF DISCRETE-TIME QUEUES WITH BATCH ARRIVALS AND DEPARTURES

In this chapter, we first analyze the timing capacity of a discrete-time queueing model that incorporates batch packet arrivals and a batch service mechanism related to the leaky bucket flow control system. We establish an upper bound on the λ-timing capacity of queues within this model, and show that this bound is tight in the case of the queue that can serve a geometrically distributed number of packets in a slot. For queues with an arbitrary service distribution within this model, we obtain a lower bound to the λ-timing capacity in terms of a certain parameter that relates to the steady-state queue length distribution in such queues. We examine the extremal nature of the geometric server based on this parameter. We then consider the case of queues with multiple servers, each of which can serve at most one packet per slot. We obtain an upper bound on the timing capacity of the queueing system with K servers having i.i.d. geometrically distributed service times in terms of the capacity of the K-output binomial channel subject to a constraint. This upper bound can be evaluated numerically for finite K. For the case $K = \infty$, although the bound cannot be evaluated numerically, we establish the existence of an input distribution that achieves the capacity of the infinite output binomial channel, and obtain a necessary and sufficient condition that characterizes this distribution.

In Section 3.1, we describe the queueing model with batch arrivals and departures that will be the focus of the major part of this chapter. In Section 3.2, we demonstrate a duality between this model and the queueing model analyzed in Chapter 2. Timing

codes and achievable rates for this model are defined in Section 3.3. In Section 3.4, we obtain an upper bound on the λ-timing capacity of queues within this model. In Section 3.5, we establish the achievability of this upper bound for the geometric server, and analyze the extremal nature of the geometric server. In Section 3.8, we analyze the timing capacity of queues with multiple servers.

3.1 The Model

In this section, we describe a first-in first-out queueing model that incorporates batch arrivals and a particular batch service mechanism that approximates the leaky bucket flow control system. The behavior of the queue is governed by the number of packets it serves in each slot. Let Y_i be the number of packets in the queue at the beginning of slot i. Let A_i be the number of arriving packets, D_i the number of departing packets, and $X_i = Y_i + A_i$ the number of packets in the queue just after the arrivals, in slot i. Figure 3.1 explains this notation.

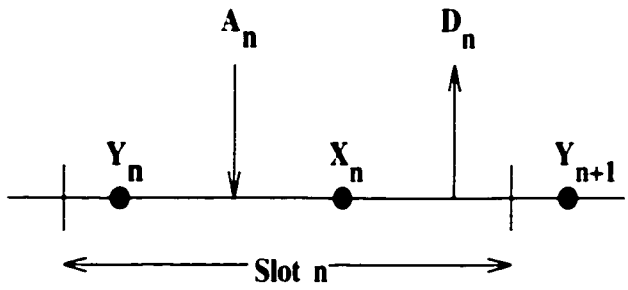


Figure 3.1: Arrivals and departures in a discrete-time batch-arrival, batch-service queue.

In general, the distribution of D_i depends on the value of X_i. For example, there are discrete-time queues called *S-queues* [50, Sec. 4.7], in which the conditional distribution of D_i given X_i has a specific form. S-queues are quasi-reversible, and networks of S-queues have product form distributions with Poisson-distributed arrivals [50, Sec.

5.2]. We consider the special case in which D_i depends on X_i as follows. In slot i, the queue can serve a random number S_i of packets. If there are S_i or more packets in the queue, S_i of them depart. If there are less than S_i packets waiting in the queue in slot i, then all the packets in the queue depart. That is, $D_i = \min(X_i, S_i)$. The equations governing this queueing system are

$$\begin{aligned} X_i &= Y_i + A_i \\ D_i &= \min(X_i, S_i) \\ Y_{i+1} &= (X_i - S_i)^+. \end{aligned} \tag{3.1}$$

S_i are assumed to be i.i.d. random variables. Let us denote the distribution function of the S_i as F_S and the corresponding probability mass function as P_S.

This model is motivated by a packet switch [7]. In each slot, the switch serves some packets from the buffers at the input ports. Consider a single virtual circuit connection between a certain input port and a certain output port on the switch. From the point of view of that connection, the switch can be thought of as a server that serves a random number of packets in each slot, depending on the random number of packets of all the other connections on the switch that are ready for service. This queueing model is related to the leaky bucket flow control system in the following way. Consider a particular virtual circuit that is one of several that share the same input port on an ATM switch. Suppose the aggregate flow of packets over all virtual circuits using that port is regulated by a leaky bucket mechanism. If the total number of virtual circuits is large, and the particular virtual circuit under consideration is a low-priority connection, the number of tokens available for that connection can be approximated to be i.i.d. from slot to slot. If S_i is the number of tokens available for that virtual circuit in slot i, and X_i is the number of packets waiting, then the number of packets switched in slot i is $D_i = \min(X_i, S_i)$, as in (3.1).

This queueing model is different from the model considered in Chapter 2 in that successive packets do not have independent service times. In fact, because there is

no single packet in service at any given time, the service time of a packet is not well-defined in this model. The natural means for conveying information through this queue is the number of packets that arrive at and depart from the queue in each slot, rather than the interarrival times of the packets. Thus a codeword for this "channel" is a sequence of integers $A_1, \ldots, A_n$, where A_i is the number of packets arriving in slot i. The receiver observes $D_1, \ldots, D_n$, where D_i is the number of packets that depart in slot i, and attempts to infer the transmitted codeword. The "noise" in this channel is the randomness in the number of packets that are served in each slot.

We will continute to denote the geometric distributions on the non-negative and positive integers by $\mathrm{Geo}(x)$ and $\mathrm{Geo}^+(x)$ respectively, as defined in Notations A.2 and A.3 in Appendix A.

Example 3.1: Consider a queue for which $S_i \sim \mathrm{Geo}(\mu/(1+\mu))$, so that $E[S_i] = \mu$. We will refer to this queue as a $\cdot/Geo^{(b)}/1$ queue, with the superscript indicating that a batch of packets is served in each time slot. Suppose the number of arrivals $A_i \sim \mathrm{Geo}(\lambda/(1+\lambda))$ and are i.i.d., so that $E[A_i] = \lambda$. We will refer to this arrival process as a *rate* λ *geometric process*. Some queueing theoretic results for this queue are derived in Appendix B. These include steady-state distributions for D_i and X_i, and in particular, the result that in steady state, the departures $D_i \sim \mathrm{Geo}(\lambda/(1+\lambda))$ and are i.i.d. ◁

For queues in which S_i have an arbitrary distribution, it is shown in Appendix B that when the arrival process is a geometric process, the steady-state distribution of the queue size is geometric.

3.2 Duality between Single-arrival and Bulk-arrival Queues

In this section we will demonstrate a duality between the single-arrival single-service queues described in Section 2.1 and the batch-arrival batch-service queueing systems dscribed in Section 3.1. For any single-arrival single-service queueing system, there

is a dual batch-arrival batch-service queueing system in which the service times of the single-arrival system play the role of the arriving batch sizes of the batch-arrival system. This duality mapping is invertible, and hence queueing properties of a single-arrival single-service system can be inferred from the properties of its dual.

Consider a single-arrival single-service queue as described in Section 2.1. Let the interarrival times of the single-arrival queue be $A_1, A_2, \cdots$, and the sequence of service times be $S_1, S_2, \cdots$. Let the interdeparture times be $D_1, D_2, \cdots$. Let the delay of packet i be $\delta_i, i = 1, 2, \cdots$, with δ_0 taken to be 0. Let the idle time of the queue immediately prior to the arrival of packet i be W_i. Then the following recursive relations hold for any $i = 1, 2, \cdots$.

$$\delta_i = (\delta_{i-1} - A_i)^+ + S_i \tag{3.2}$$

$$W_i = (A_i - \delta_{i-1})^+ \tag{3.3}$$

Note that (3.2) defines a discrete-time batch-arrival batch-service queueing system as described in Section 3.1, with the service time S_i of the i^{th} packet of the single-service queue playing the role of the arriving batch size in slot i. The number of packets that can be potentially served in slot i by the batch-service queue is the interarrival time A_{i+1} between packets i and $i+1$ of the single-arrival queue. The delay δ_i of packet i in the single-arrival system plays the role of the effective queue size X_i in slot i in the batch-arrival system, and the size of the departing batch in slot i in the latter system is $\min(\delta_i, A_{i+1})$. The number of leftover packets Y_i at the end of slot i in the batch-arrival system is $(\delta_{i-1} - A_i)^+$. The batch-arrival batch-service system thus constructed can be thought of as a *dual* of the single-arrival single-service system, with the service times of the latter playing the role of the arriving batch sizes of the former, and the interarrival times of the single-arrival queue corresponding to the maximum allowed service per slot of the batch-service queue.

Along the same lines, given a batch-arrival queue of the type described in Section 3.1, we can construct a dual single-arrival single-service queue in which arrivals

to the batch queue correspond to the service times in the single-arrival queue, and the service random variables in the batch queue play the role of interarrival times in the dual single-arrival system.

Thus it is clear that given the sequence of arriving batch sizes and the maximum allowed service sequence of a batch-arrival batch-service system following the model of Section 3.1, there is a unique single-arrival single-service system whose dual is the given batch-arrival batch-service system. That is, the duality map is invertible.

Properties of the original queue can thus be inferred from the properties of the dual queue. For example, Proposition B.2 in Appendix B asserts that for a batch-arrival batch-service queue within the model of Section 3.1, if the arriving batch sizes are i.i.d. and geometrically distributed, then the distribution of the effective queue size is geometric with a parameter γ which is the solution of a certain equation involving the characteristic function of the service random variable of the batch queue. The dual of this batch-arrival system is the single-arrival single-service queue with i.i.d. geometrically distributed service times, and the effective queue size of the batch queue corresponds to packet delays in the dual single-arrival system. This implies that in the single-arrival single-service queues with geometric service times, packet delays are geometrically distributed with parameter γ as described in Proposition B.2.

3.3 Timing Codes

In the following, we assume that the queue is initially empty. As in the first model, we will show later that the timing capacity with the queue initially in equilibrium is no more than the timing capacity with the queue initially empty.

An (n, M, N, ϵ) code consists of M codewords with the following properties.

(i) Each codeword is a sequence of non-negative integers $A_1, \ldots, A_n$, where A_i is the number of packets arriving at the queue in slot i.

(ii) The decoder observes the sequence $D_1, \ldots, D_n$, where D_i is the number of packets that depart from the queue in slot i, and attempts to infer the transmitted codeword.

(iii) The average number of packets that depart from the queue in n slots, averaged over all codewords, is N, i.e. $E\left[\sum_{i=1}^{n} D_i\right] = N$.

(iv) The average probability of error for the code is ϵ.

An (n, M, N, ϵ) timing code has information rate $R = (\log M)/n$.

Rate R is said to be *achievable* if there exists a sequence $(n, M_n, N_n, \epsilon_n)$ of timing codes that has a subsequence $(n_k, M_{n_k}, N_{n_k}, \epsilon_{n_k})$ for which

$$\lim_{k\to\infty} \frac{\log M_{n_k}}{n_k} = R \text{ and } \lim_{k\to\infty} \epsilon_{n_k} = 0.$$

The timing capacity C is the supremum of the set of achievable rates.

Rate R is said to be *achievable at departure rate* λ if there exists a sequence of timing codes $(n, M_n, N_n, \epsilon_n)$ that has a subsequence $(n_k, M_{n_k}, N_{n_k}, \epsilon_{n_k})$ for which

$$\lim_{k\to\infty} \frac{\log M_{n_k}}{n_k} = R, \lim_{k\to\infty} \frac{N_{n_k}}{n_k} = \lambda, \text{ and } \lim_{k\to\infty} \epsilon_{n_k} = 0. \tag{3.4}$$

The λ-timing capacity $C(\lambda)$ is the supremum of all rates that are achievable at departure rate λ.

As in Section 2.2, in the definition of achievability at departure rate λ, we require that the packet departure rate be λ only in the limit, rather than for each code.

As before, we first show that to analyze the timing capacity of a queue in this model, it suffices to analyze the λ-timing capacity, by the following result.

Proposition 3.1. *For a queue with an arbitrary service distribution with mean μ packets per slot,*

$$C = \sup_{0\le\lambda\le\mu} C(\lambda).$$

Proof. From the definitions, it is clear that $C \geq C(\lambda)$ for all λ. Now there exists a sequence of $(n, M_n, N_n, \epsilon_n)$ codes that achieves C, i.e. that has a subsequence $(n_k, M_{n_k}, N_{n_k}, \epsilon_{n_k})$ such that

$$\lim_{k\to\infty} \frac{\log M_{n_k}}{n_k} = C \text{ and } \lim_{k\to\infty} \epsilon_{n_k} = 0.$$

The total number of packets departing in n slots is no more than the maximum number of packets that the queue could have served, i.e. $\sum_{i=1}^{n_k} D_i \leq \sum_{i=1}^{n_k} S_i$. This implies that $N_{n_k} \leq \sum_{i=1}^{n_k} E[S_i] = n_k\mu$, and therefore $N_{n_k}/n_k \leq \mu$. Hence $l = \limsup_{k\to\infty}(N_{n_k}/n_k)$ exists, and $0 \leq l \leq \mu$. But since the limsup is itself a limit point, there is a further subsequence $\{n_{k_j}\}$ of $\{n_k\}$, for which

$$\lim_{j\to\infty} \frac{N_{n_{k_j}}}{n_{k_j}} = l, \lim_{j\to\infty} \frac{\log M_{n_{k_j}}}{n_{k_j}} = C, \text{ and } \lim_{j\to\infty} \epsilon_{n_{k_j}} = 0.$$

So by definition, C is achievable at departure rate l for some $0 \leq l \leq \mu$, i.e. $C \leq C(l)$. Therefore

$$C = \sup_{0\leq\lambda\leq\mu} C(\lambda).$$

□

3.4 Converse Theorem

Proposition 3.2. *Consider a queue with service distribution F_S, and let $S \sim F_S$. For any probability distribution Q, let $X_Q \sim Q$ be a random variable independent of S. Let $\mathcal{B}_S(x)$ be the set of probability distributions Q on the non-negative integers for which $E[\min(X_Q, S)] = x$. Then for $0 \leq \lambda \leq \mu$,*

$$C(\lambda) \leq \sup_{Q\in\mathcal{B}_S(\lambda)} I(X_Q; \min(X_Q, S)). \tag{3.5}$$

Proof. Consider a sequence of timing codes $(n, M_n, N_n, \epsilon_n)$ for which $\lim_{n\to\infty} \epsilon_n = 0$ and $\lim_{n\to\infty}(N_n/n) = \lambda$. For the n^{th} code in the sequence, let $A_1, \ldots, A_n$ be the

number of packets sent to the queue in slots $1, \ldots, n$ respectively, and let $D_1, \ldots, D_n$ be the number of departing packets in the respective slots. By Fano's lemma [10, Sec. 2.11] and the data processing lemma [10, Sec. 2.8],

$$\begin{aligned} \log M_n &\leq \frac{1}{1-\epsilon_n}\left[\log 2 + I(A_1, \ldots, A_n; D_1, \ldots, D_n)\right] \\ &\leq \frac{1}{1-\epsilon_n}\left[\log 2 + \sum_{i=1}^{n} H(D_i) - \sum_{i=1}^{n} H(D_i|A_1, \ldots, A_n, D_1, \ldots, D_{i-1})\right]. \end{aligned}$$

The number of packets D_i departing in slot i does not depend on the arrivals in slots $i+1, \ldots, n$. Given $A_1, \ldots, A_i$ and $D_1, \ldots, D_{i-1}$, the number of packets X_i in the queue in slot i is determined. Also X_i is a sufficient statistic of $A_1, \ldots, A_i, D_1, \ldots, D_{i-1}$ for D_i. Hence $H(D_i|A_1, \ldots, A_n, D_1, \ldots, D_{i-1}) = H(D_i|X_i)$. Therefore

$$\begin{aligned} \frac{\log M_n}{n} &\leq \frac{1}{1-\epsilon_n}\left[\frac{1}{n}\sum_{i=1}^{n} I(X_i; D_i) + \frac{\log 2}{n}\right] \\ &= \frac{1}{1-\epsilon_n}\left[\frac{1}{n}\sum_{i=1}^{n} I(X_i; \min(X_i, S_i)) + \frac{\log 2}{n}\right]. \end{aligned}$$

Define

$$\eta_S(x) = \sup_{Q \in \mathcal{B}_S(x)} I(X_Q; \min(X_Q, S)),$$

where $\mathcal{B}_S(x)$ and X_Q are as defined in the statement of the proposition. Using arguments similar to those in the proof of Proposition 2.2, it can be shown that $\eta_S(\cdot)$ is a concave function.

From the definition of $\eta_S(\cdot)$, $I(X_i; \min(X_i, S_i)) \leq \eta_S(E[\min(X_i, S_i)])$. In conjunction with the concavity of $\eta_S(\cdot)$, this implies

$$\frac{1}{n}\sum_{i=1}^{n} I(X_i; \min(X_i, S_i)) \leq \eta_S\left(\frac{1}{n}\sum_{i=1}^{n} E[D_i]\right).$$

Thus

$$\frac{\log M_n}{n} \leq \frac{1}{1-\epsilon_n}\left[\eta_S\left(\frac{1}{n}\sum_{i=1}^{n} E[D_i]\right) + \frac{\log 2}{n}\right].$$

Taking the limit as $n \to \infty$ we obtain (3.5). □

Next we address the evaluation of the function $\eta_S(\cdot)$ for a given service distribution S. The example below considers this for the case of a Bernoulli server, and also establishes a connection between the queueing models of Sections 3.1 and 2.1.

Example 3.2: Let S have a Bernoulli distribution with mean μ, i.e. $P_S(1) = \mu, P_S(0) = 1 - \mu$. Clearly in this case, for any non-negative integer-valued random variable X, $\min(X, S)$ takes only the values 0 and 1. Let X be independent of S and such that $E[\min(X, S)] = \lambda$. Now $P_{\min(X,S)}(1) = \mu P(X \geq 1)$. Hence $P(X \geq 1) = \lambda/\mu$ and $P_X(0) = 1 - \lambda/\mu$. Therefore

$$\begin{aligned} I(X; \min(X, S)) &= H(\min(X, S)) - H(\min(X, S)|X) \\ &= h(\lambda) - P_X(0)H(\min(X, S)|X = 0) \\ &\quad - \sum_{k=1}^{\infty} P_X(k)H(\min(X, S)|X = k) \\ &= h(\lambda) - (1 - \frac{\lambda}{\mu}) \cdot 0 - \sum_{k=1}^{\infty} P_X(k)h(\mu) \\ &= h(\lambda) - \frac{\lambda}{\mu}h(\mu). \end{aligned}$$

This is true for any X such that $E[\min(X, S)] = \lambda$, and hence

$$\eta_S(\lambda) = h(\lambda) - \frac{\lambda}{\mu}h(\mu).$$

For a queue with Bernoulli(μ) service, each packet is in service for a geometrically distributed number of slots. Thus this queue is in fact the $\cdot/Geo^{(s)}/1$ queue analyzed in Chapter 2. The λ-timing capacity of such a queue is precisely $h(\lambda) - \frac{\lambda}{\mu}h(\mu)$, as given by (2.8). ◁

The following proposition establishes a min-max saddle point property of the geometric distribution, and obtains $\eta_S(\lambda)$ for the case where S has a geometric distribution.

Proposition 3.3. *Let $\overline{S} \sim Geo(b)$. Let $\overline{X} \sim Geo(a/b)$ be independent of $\overline{S}$, where $0 < a < b$. Let X be any non-negative integer-valued random variable independent*

of $\overline{S}$ such that $E[\min(X,\overline{S})] = a/(1-a)$. Let S be any non-negative integer-valued random variable independent of $\overline{X}$, such that $E[\min(\overline{X},S)] = a/(1-a)$. Then

$$I(X;\min(X,\overline{S})) \le I(\overline{X};\min(\overline{X},\overline{S})) \le I(\overline{X};\min(\overline{X},S)), \tag{3.6}$$

so that

$$\eta_{\overline{S}}(\lambda) = I(\overline{X};\min(\overline{X};\overline{S})) = \frac{a}{1-a}\left[\frac{h(a)}{a} - \frac{h(b)}{b}\right].$$

Proof. First note that $\overline{D} = \min(\overline{X},\overline{S})$ is Geo (a), since

$$P(\min(\overline{X},\overline{S}) \ge k) = P(\overline{X} \ge k)P(\overline{S} \ge k) = a^k.$$

Consider any non-negative integer-valued random variable X that is independent of $\overline{S}$ and such that $E[\min(X,\overline{S})] = a/(1-a)$. Let $D = \min(X,\overline{S})$. Since the geometric distribution has the maximum entropy among all non-negative integer valued random variables with a given mean [10, Sec. 11.1], $H(D) \le H(\overline{D}) = h(a)/(1-a)$. Therefore

$$I(X;D) = H(D) - H(D|X) \le \frac{h(a)}{1-a} - H(D|X). \tag{3.7}$$

Now $H(D|X) = E_{X,D}[-\log P_{D|X}(D|X)]$, where

$$\begin{aligned} P_{D|X}(j|k) &= \begin{cases} b^j(1-b), & j < k \\ b^j, & j = k \end{cases} \\ &= b^j(1-b)^{U(k-j)}, \end{aligned}$$

where $U(n)$ is 0 if $n = 0$ and 1 if $n > 0$. Then

$$\begin{aligned} H(D|X) &= E[-D\log b] + E[-\log(1-b)U(X-D)] \\ &= -\frac{a}{1-a}\log b - \log(1-b)P(\overline{S} < X) \\ &= -\frac{a}{1-a}\log b - \log(1-b)\left[1 - \sum_{j=1}^{\infty} P_X(j)b^j\right]. \end{aligned} \tag{3.8}$$

Now $E[\min(X, \overline{S})] = a/(1-a)$. Also

$$\begin{aligned} E[\min(X, \overline{S})] &= \sum_{k=1}^{\infty} P(\min(X, \overline{S}) \geq k) \\ &= \sum_{k=1}^{\infty} b^k P(X \geq k) \\ &= \sum_{j=1}^{\infty} P_X(j) \sum_{k=1}^{j} b^k \\ &= \frac{b}{1-b}\left[1 - \sum_{j=0}^{\infty} P_X(j) b^j\right]. \end{aligned} \tag{3.9}$$

Hence

$$1 - \sum_{j=0}^{\infty} P_X(j) b^j = \frac{a(1-b)}{b(1-a)}. \tag{3.10}$$

Combining (3.7), (3.8), and (3.10), we obtain

$$I(X; \min(X, \overline{S})) \leq \frac{a}{1-a}\left[\frac{h(a)}{a} - \frac{h(b)}{b}\right],$$

with equality iff X is Geo (a/b).

To prove the second inequality in (3.6), let $\overline{D} = \min(\overline{X}, \overline{S})$ as above, and let $Y = \min(\overline{X}, S)$. Then we have

$$\begin{aligned} I(\overline{X}; \min(\overline{X}, S)) &= E_{\overline{X},Y}\left[\log \frac{P_{\overline{X}|Y}(\overline{X}|Y)}{P_{\overline{X}}(\overline{X})}\right] \\ &= E_Y E_{\overline{X}|Y}\left[\log\left(\frac{P_{\overline{X}|Y}(\overline{X}|Y)}{P_{\overline{X}|\overline{D}}(\overline{X}|Y)} \frac{P_{\overline{X}|\overline{D}}(\overline{X}|Y)}{P_{\overline{X}}(\overline{X})}\right)\right] \\ &= E_Y\left[D(P_{\overline{X}|Y} || P_{\overline{X}|\overline{D}}|Y)\right] + E_Y E_{\overline{X}|Y}\left[\log \frac{P_{\overline{X}|\overline{D}}(\overline{X}|Y)}{P_{\overline{X}}(\overline{X})}\right]. \end{aligned}$$

The first term on the right side is non-negative, and is strictly positive unless S is

geometrically distributed. A slight rewriting of the second term yields

$$
\begin{aligned}
& I(\overline{X}; \min(\overline{X}, S)) \\
& \quad \geq \quad E_{\overline{X},Y}\left[\log \frac{P_{\overline{D}|\overline{X}}(Y|\overline{X})}{P_{\overline{D}}(Y)}\right] \\
& \quad = \quad \sum_{k=0}^{\infty} P_{\overline{X}}(k)\left[\sum_{j=0}^{k-1} P_{S|\overline{X}}(j|k) \log \frac{P_{\overline{S}}(j)}{P_{\overline{D}}(j)} + P(S \geq k|\overline{X}=k) \log \frac{P(S \geq k|\overline{X}=k)}{P_{\overline{D}}(k)}\right] \\
& \quad = \quad \sum_{k=0}^{\infty} P_{\overline{X}}(k)\left[\sum_{j=0}^{k-1} P_{S|\overline{X}}(j|k) \log \frac{b^j(1-b)}{a^j(1-a)} + P(S \geq k|\overline{X}=k) \log \frac{b^k}{a^k(1-a)}\right] \\
& \quad = \quad \log \frac{1}{1-a} + \left(\log \frac{b}{a}\right)\left[\sum_{k=0}^{\infty} P_{\overline{X}}(k)\left\{\sum_{j=0}^{k-1} j P_{S|\overline{X}}(j|k) + k P(S \geq k|\overline{X}=k)\right\}\right] \\
& \qquad + \log(1-b)\left[\sum_{k=0}^{\infty} P_{\overline{X}}(k) \sum_{j=0}^{k-1} P_{S|\overline{X}}(j|k)\right] \\
& \quad = \quad \log \frac{1}{1-a} + \left(\log \frac{b}{a}\right) E[\min(\overline{X}, S)] \\
& \qquad + \log(1-b)\left[\sum_{k=0}^{\infty} P_{\overline{X}}(k)\left\{1 - P(S \geq k|\overline{X}=k)\right\}\right]
\end{aligned}
$$

To simplify the last term on the right side, note that since S is independent of $\overline{X}$,

$$
\begin{aligned}
\sum_{k=0}^{\infty} P_{\overline{X}}(k) P(S \geq k|\overline{X}=k) &= \sum_{k=0}^{\infty} \left(\frac{a}{b}\right)^k \left(1 - \frac{a}{b}\right) P(S \geq k) \\
&= \left(1 - \frac{a}{b}\right) \sum_{k=0}^{\infty} P(\overline{X} \geq k) P(S \geq k) \\
&= \left(1 - \frac{a}{b}\right) \sum_{k=0}^{\infty} P(\min(\overline{X}, S) \geq k) \\
&= \left(1 - \frac{a}{b}\right)(1 + E[\min(\overline{X}, S)]),
\end{aligned}
$$

and $E[\min(\overline{X}, S)] = a/(1-a)$ by hypothesis. Hence we have

$$
\begin{aligned}
I(\overline{X}; \min(\overline{X}, S)) &\geq \log \frac{1-b}{1-a} + \frac{a}{1-a} \log \frac{b}{a} - \log(1-b)\left(1 - \frac{a}{b}\right)\left(\frac{1}{1-a}\right) \\
&= \frac{a}{1-a}\left[\frac{h(a)}{a} - \frac{h(b)}{b}\right].
\end{aligned}
$$

□

Let $\overline{C}_\mu(\lambda)$ denote the λ-timing capacity of the $\cdot/Geo^{(b)}/1$ queue whose service distribution is Geo $(\mu/(1+\mu))$. Proposition 3.3 in conjunction with the converse result of Proposition 3.2 gives the following upper bound on $\overline{C}_\mu(\lambda)$:

$$\overline{C}_\mu(\lambda) \leq (1+\lambda)h\left(\frac{\lambda}{1+\lambda}\right) - \lambda\left(\frac{1+\mu}{\mu}\right)h\left(\frac{\mu}{1+\mu}\right). \tag{3.11}$$

3.5 *Achievability Theorem*

We will now show that for the $\cdot/Geo^{(b)}/1$ queue, the upper bound (3.11) on the λ-timing capacity is achievable. We will also derive a lower bound to the λ-timing capacity of a queue with an arbitrary service distribution, based on a queueing theoretic property of the steady-state queue length in such queues.

In the proof of Proposition 2.4, we needed to use results from queueing theory, such as the fact that for the $\cdot/Geo^{(s)}/1$ queue, if the interarrival times are independent and geometrically distributed, then so are the interdeparture times. Appendix B establishes some queueing theoretic results for the batch-arrival batch-service queues described in Section 3.1. Proposition B.1 shows that in steady state, for the $\cdot/Geo^{(b)}/1$ queue, if the arrival process has independent, geometrically distributed number of arrivals in each slot, then the departure process also has independent, geometrically distributed number of departures in each slot. For a queue with an arbitrary service distribution, Proposition B.2 shows that if the arrival process is geometric, then the steady-state distribution of the queue size X_i just after the arrivals is geometric with a certain parameter γ. We will use these queueing theoretic properties in establishing achievability results, and the parameter γ will turn out to play a significant role.

In the proof of Proposition 2.4, we followed the technique used in [4] of showing that achievability with the queue initially in equilibrium with respect to some arrival process implies achievability with the queue initially empty. That assertion used the fact that for a queue with single arrivals and single service, the steady-state probability that the queue is empty is non-zero. Proposition B.2 in Appendix B shows that for

the batch-arrival batch-service queues described in Section 3.1, if the arrival process is a rate λ geometric process, the steady-state probabilty that the queue is empty at the beginning of a slot is $(1-\gamma)(1+\lambda) > 0$. Using this, we can show for these queues that if a certain information rate can be achieved with the queue initially in equilibrium with respect to an appropriate geometric arrival process, then that rate can also be achieved with the queue initially empty, as follows.

The definitions in Section 3.3 can be generalized to a queue initially in equilibrium with respect to a geometric arrival process. Thus we have (n, M, N, ϵ) timing codes with the queue initially in equilibrium with respect to a geometric arrival process, and the corresponding definitions of rate, achievability, etc. Note that since the queue is initially in equilibrium, there is a random number of packets initially in the queue. Hence the receiver may initially receive some spurious packets not sent by the transmitter. For (n, M, N, ϵ) timing codes with the queue initially in equilibrium, we assume that N, the average number of departures in n slots, and ϵ, the average probability of error, are computed by averaging over all codewords as well as over the initial state of the queue.

Suppose that rate R is achievable at departure rate λ with the queue initially in equilibrium with respect to a rate λ geometric arrival process. That is, with the queue initially in equilibrium, there exists a sequence of timing codes $(n, M_n, N_n, \epsilon_n)$ that has a subsequence $(n_k, M_{n_k}, N_{n_k}, \epsilon_{n_k})$ for which (3.4) is satisfied. Suppose that the same codes are used with the queue initially empty. Noting that the steady-state probability that the queue is empty at the beginning of a slot is $(1-\gamma)(1+\lambda)$, and using an argument similar to the one used in Section 2.4, we can show that the probability of error $(\epsilon_{\text{empty}})_{n_k} \to 0$.

Also the number of packets that would depart in n_k slots if the queue were initially empty is no more than the number of departing packets when the queue is initially in equilibrium, i.e. $(N_{\text{empty}})_{n_k} \leq N_{n_k}$. With the queue initially in equilibrium, let Y_1 be the number of packets initially in the queue, i.e. at the beginning of slot 1. Then

$N_{n_k} - E[Y_1] \leq (N_{\text{empty}})_{n_k} \leq N_{n_k}$, so that

$$\lim_{k\to\infty} \frac{N_{n_k} - E[Y_1]}{n_k} \leq \lim_{k\to\infty} \frac{(N_{\text{empty}})_{n_k}}{n_k} \leq \lim_{k\to\infty} \frac{N_{n_k}}{n_k} = \lambda.$$

Since $E[Y_1]$ is finite, it follows that rate R is also achievable at departure rate λ packets/slot when the queue is initially empty.

We now obtain a lower bound to the λ-timing capacity $\overline{C}_\mu(\lambda)$ of the $\cdot/Geo^{(b)}/1$ queue with mean service rate μ.

Proposition 3.4. *For a $\cdot/Geo^{(b)}/1$ queue with the $Geo(\mu/(1+\mu))$ service distribution, for $0 < \lambda < \mu$,*

$$\overline{C}_\mu(\lambda) \geq (1+\lambda)h\left(\frac{\lambda}{1+\lambda}\right) - \lambda\left(\frac{1+\mu}{\mu}\right)h\left(\frac{\mu}{1+\mu}\right). \tag{3.12}$$

Proof. As noted above, it suffices to show the achievability when the queue is initially in equilibrium with respect to a rate λ geometric arrival process.

With $a = \lambda/(1+\lambda)$ and $b = \mu/(1+\mu)$, the right side of (3.12) can be written as $\frac{a}{1-a}\left[\frac{h(a)}{a} - \frac{h(b)}{b}\right]$. From the results of [47], it suffices to show that for some arrival process $\{A_i\}$ and the corresponding departure process $\{D_i\}$, the liminf in probability of the normalized input-output information density is at least $\frac{a}{1-a}\left[\frac{h(a)}{a} - \frac{h(b)}{b}\right]$, i.e. to show that for some arrival process $\{A_i\}$,

$$P\left[\frac{1}{n} i_{A^n;D^n}(A^n;D^n) < \frac{a}{1-a}\left(\frac{h(a)}{a} - \frac{h(b)}{b}\right)\right] \longrightarrow 0 \tag{3.13}$$

as $n \to \infty$, where $D^n = (D_1, \ldots, D_n)$, $A^n = (A_1, \ldots, A_n)$, and

$$i_{A^n;D^n}(A^n;D^n) = \log \frac{P_{D^n|A^n}(D_1, \ldots, D_n | A_1, \ldots, A_n)}{P_{D^n}(D_1, \ldots, D_n)}.$$

Our choice of the arrival process $\{A_i\}$ here is a rate λ geometric arrival process. It is shown in Appendix B that for this arrival process, in steady state, the departures D_i are also independent and have the Geo $(\lambda/(1+\lambda))$ distribution.

Let Y_1 be the number of packets initially in the queue, i.e. at the beginning of slot 1. We have

$$\begin{aligned} P_{D^n|A^n}(D^n|A^n) &= \sum_{k=0}^{\infty} P_{Y_1}(k) P_{D^n|A^n,Y_1}(D^n|A^n,k) \\ &\geq P_{Y_1}(0) P_{D^n|A^n,Y_1}(D^n|A^n,0) \\ &= P_{Y_1}(0) \prod_{i=1}^{n} P_{D_i|A^i,D^{i-1},Y_1}(D_i|A^i,D^{i-1},0) \\ &= P_{Y_1}(0) \prod_{i=1}^{n} P_{D_i|X_i}(D_i|X_i), \end{aligned}$$

where X_i is the number of packets in the queue just after the arrivals in the i^{th} slot, given that the queue is initially empty ($Y_1 = 0$). Now

$$\begin{aligned} P_{D_i|X_i}(j|k) &= P(\min(X_i, S_i) = j | X_i = k) \\ &= \begin{cases} b^j & \text{if } j = k \\ b^j(1-b) & \text{if } 0 \leq j < k \end{cases}. \end{aligned}$$

This can be rewritten as

$$P_{D_i|X_i}(d_i|x_i) = b^{d_i}(1-b)^{U(x_i - d_i)},$$

where $U(k)$ is 0 if $k = 0$ and 1 if $k > 0$. Hence

$$\begin{aligned} \frac{1}{n} i_{A^n;D^n}(A^n;D^n) &\geq \frac{1}{n} \log \frac{P_{Y_1}(0) \prod_{i=1}^{n} b^{D_i}(1-b)^{U(X_i - D_i)}}{\prod_{i=1}^{n} a^{D_i}(1-a)} \\ &= \frac{1}{n} \log P_{Y_1}(0) + \left(\frac{1}{n} \sum_{i=1}^{n} D_i \right) \log \frac{b}{a} - \log(1-a) \\ &\quad + \left(\frac{1}{n} \sum_{i=1}^{n} U(X_i - D_i) \right) \log(1-b). \end{aligned} \tag{3.14}$$

As $n \to \infty$, $(1/n)\sum_{i=1}^{n} D_i \to \lambda = \frac{a}{1-a}$ in probability. Also $\frac{1}{n}\sum_{i=1}^{n} U(X_i - D_i) \longrightarrow P(X > D) = P(S < X)$ in probability, where the last two are steady-state probabilities for the queue. $P(S < X)$ is the steady-state probability that the queue is non-empty at the end of a slot, which is shown in Proposition B.1 in Appendix

B to be $\frac{a(1-b)}{b(1-a)}$. Thus $(1/n)\sum_{i=1}^{n} U(X_i - D_i) \rightarrow \frac{a(1-b)}{b(1-a)}$ in probability. Moreover, $P_{Y_1}(0) = \frac{b-a}{b(1-a)} > 0$, so $(1/n)\log P_{Y_1}(0) \rightarrow 0$. Hence the right side of (3.14) converges in probability to

$$\begin{aligned} &\frac{a}{1-a}\log\frac{b}{a} - \log(1-a) + \frac{a(1-b)}{b(1-a)}\log(1-b) \\ &\quad = \frac{a}{1-a}\left[\frac{h(a)}{a} - \frac{h(b)}{b}\right]. \end{aligned}$$

This establishes (3.13) and the proposition. □

Combining (3.11) and (3.12), we obtain the λ-timing capacity of the $\cdot/Geo^{(b)}/1$ queue:

$$\overline{C}_{\mu}(\lambda) = (1+\lambda)h\left(\frac{\lambda}{1+\lambda}\right) - \lambda\left(\frac{1+\mu}{\mu}\right)h\left(\frac{\mu}{1+\mu}\right). \tag{3.15}$$

From Proposition 3.1, the timing capacity is the supremum of the λ-timing capacity over $0 \leq \lambda \leq \mu$. The supremum occurs at

$$\lambda^* = \frac{1}{\alpha - 1}$$

where $\alpha = \mu^{-1}(1+\mu)^{\frac{1+\mu}{\mu}}$, and the timing capacity is given by

$$\overline{C} = \log\left[\frac{\alpha}{\alpha - 1}\right]. \tag{3.16}$$

3.6 The Extremal Nature of the $\cdot/Geo^{(b)}/1$ Queue

We now turn to the question of the extremality of the $\cdot/Geo^{(b)}/1$ queue. Unlike the $\cdot/Geo^{(s)}/1$ queue in Chapter 2, the $\cdot/Geo^{(b)}/1$ queue does not have the least λ-timing capacity among all queues with a given mean service rate within the model of Section 3.1, as shown by the following counterexample.

Example 3.3: Consider a queue for which S takes value $K > \mu$ with probability $p = \mu/K$ and value 0 with probability $1-p$, so that $E[S] = \mu$. From Proposition 3.2, $C_S(\lambda) \leq \sup I(X; \min(X, S))$, where the supremum is over all distributions on X for

which X and S are independent and $E[\min(X,S)] = \lambda$. For any such random variable X, let $p_k = P_{\min(X,S)}(k)$, and let $q = 1 - p_0$. Thus $p_0 \geq 1 - p$ and $q \leq p$. Now

$$\begin{aligned} I(X; \min(X,S)) &\leq H(\min(X,S)) \\ &= -p_0 \log p_0 - \sum_{k \geq 1} p_k \log p_k \\ &= -p_0 \log p_0 - q \sum_{k \geq 1} \frac{p_k}{q} \log \frac{p_k}{q} - q \log q. \end{aligned}$$

Since $E[\min(X,S)] = \sum_{k \geq 1} k p_k = \lambda$, we have $\sum_{k \geq 1} k p_k / q = \lambda / q$. The geometric random variable maximizes entropy among positive integer-valued random variables with a given mean x, and this maximum entropy is $x h(1/x)$. Hence

$$-\sum_{k \geq 1} \frac{p_k}{q} \log \frac{p_k}{q} \leq \frac{\lambda}{q} h(\frac{q}{\lambda}).$$

If $p < \lambda/2$, then $h(q/\lambda) \leq h(p/\lambda)$. Now $p_0 \geq 1 - p$, so if $p < 1 - 1/e$, then $-p_0 \log p_0 \leq -(1-p)\log(1-p)$. Also $q \leq p$, so if $p < 1/e$, then $-q \log q \leq -p \log p$. So when $p < \min(\lambda/2, 1/e)$, we have

$$\begin{aligned} I(X; \min(X,S)) &\leq -(1-p)\log(1-p) + \lambda h(\frac{p}{\lambda}) - p \log p \\ &= h(\frac{\mu}{K}) + \lambda h(\frac{\mu}{K\lambda}). \end{aligned} \tag{3.17}$$

For a given λ and μ, by taking K sufficiently large, the right side of (3.17) can be made less than $\overline{C}_\mu(\lambda)$. Thus the $\cdot/Geo^{(b)}/1$ queue does not have the least λ-timing capacity among all queues with a given mean service rate within the model of Section 3.1. ◁

However, we can find a lower bound to the λ-timing capacity of a queue with an arbitrary service distribution in terms of a certain parameter γ, which is defined by the following lemma.

Lemma 3.1. *Let S be a non-negative integer-valued random variable with mean μ and probability generating function $\phi_S(\cdot)$, and let $0 < \lambda < \mu$. Then there is a unique solution $\gamma \in \left(\frac{\lambda}{1+\lambda}, 1\right)$ of the equation*

$$\phi_S(x) = 1 + \lambda - \frac{\lambda}{x}. \tag{3.18}$$

Proof. See Appendix B. □

The value of γ depends on λ and on the the distribution of S, not just on the mean μ of S. For some common distributions, a closed form expression for γ can be found. For example, when $S \sim \text{Geo}(\mu/(1+\mu))$, $\gamma = \frac{\lambda(1+\mu)}{\mu(1+\lambda)}$, and when S is Bernoulli, $\gamma = \lambda/\mu$.

Proposition 3.5. *Consider a queue whose service distribution has mean μ and generating function $\phi_S(\cdot)$. For $0 < \lambda < \mu$, let γ be the solution in Lemma 3.1. Then*

$$C(\lambda) \geq (1+\lambda)\left[h\left(\frac{\lambda}{1+\lambda}\right) - \gamma h\left(\frac{\lambda}{\gamma(1+\lambda)}\right)\right]. \tag{3.19}$$

Proof. Consider the queue to be initially in equilibrium with respect to a rate λ geometric arrival process. Proposition B.2 in Appendix B shows that the steady-state distribution of the queue size $\{X_i\}$ is $\text{Geo}(\gamma)$. We will show that for this arrival process $\{A_i\}$ and the corresponding departure process $\{D_i\}$, the liminf in probability of the normalized input-output information density $(1/n)i_{A^n;D^n}(A^n;D^n)$ is at least $\frac{a}{1-a}\left(\frac{h(a)}{a} - \frac{h(a/\gamma)}{a/\gamma}\right)$, where $a = \frac{\lambda}{1+\lambda}$. Let $Q_{D^n}(\cdot)$ and $Q_{D^n|A^n}(\cdot|\cdot)$ denote, respectively, the joint distribution of the departures and the joint distribution of the departures conditioned on the arrivals. By definition,

$$i_{A^n;D^n}(A^n;D^n) = \log \frac{Q_{D^n|A^n}(D_1,\ldots,D_n|A_1,\ldots,A_n)}{Q_{D^n}(D_1,\ldots,D_n)}.$$

Let $P_{D^n}(\cdot)$ and $P_{D^n|A^n}(\cdot|\cdot)$ be the corresponding probabilities when the service distribution is $\text{Geo}(a/\gamma)$. As in the proof of Proposition 2.5,

$$\begin{aligned}\frac{1}{n}i_{A^n;D^n}(A^n;D^n) &= \frac{1}{n}\log\frac{Q_{A^n|D^n}(A^n|D^n)}{P_{A^n|D^n}(A^n|D^n)} \\ &+ \frac{1}{n}\log\frac{P_{D^n|A^n}(D^n|A^n)}{P_{D^n}(D^n)}.\end{aligned}$$

The liminf in probability of the first term on the right is non-negative, by the same

argument used in that proof. As in the proof of Proposition 3.4, we can show that

$$\frac{1}{n}\log\frac{P_{D^n|A^n}(D^n|A^n)}{P_{D^n}(D^n)} \geq \frac{1}{n}\log P_{Y_1}(0) + \left(\log\frac{1}{\gamma}\right)\left(\frac{1}{n}\sum_{i=1}^{n} D_i\right) - \log(1-a) + \log(1-\frac{a}{\gamma})\left(\frac{1}{n}\sum_{i=1}^{n} U(X_i - D_i)\right) \tag{3.20}$$

where $U(k) = 0$ if $k = 0$ and 1 if $k > 0$. As $n \to \infty$, $(1/n)\sum_{i=1}^{n} D_i \to \lambda = \frac{a}{1-a}$ in probability. Also $\frac{1}{n}\sum_{i=1}^{n} U(X_i - D_i) \to Q(X > D) = Q(S < X)$ in probability, where the last two probabilities are the steady-state probabilities for the queue. In steady state, $Q(S \geq X)$ is the probability that the queue is empty at the end of a slot. It is shown in Proposition B.2 in Appendix B that this is given by $(1-\gamma)/(1-a)$, so that $Q(S < X) = (\gamma - a)/(1-a)$. Thus $(1/n)\sum_{i=1}^{n} U(X_i - D_i) \to (\gamma - a)/(1-a)$ in probability. Hence as $n \to \infty$, the right side of (3.20) converges in probability to

$$\frac{a}{1-a}\log\frac{1}{\gamma} - \log(1-a) + \frac{\gamma - a}{1-a}\log(1-\frac{a}{\gamma}) = \frac{a}{1-a}\left[\frac{h(a)}{a} - \frac{h(a/\gamma)}{a/\gamma}\right].$$

Therefore as $n \to \infty$,

$$P\left[\frac{1}{n} i_{A^n;D^n}(A^n; D^n) < \frac{a}{1-a}\left(\frac{h(a)}{a} - \frac{h(a/\gamma)}{a/\gamma}\right)\right] \longrightarrow 0$$

Replacing a by $\lambda/(1+\lambda)$, we obtain (3.19). □

For the case when $S \sim \text{Geo}\,(\mu/(1+\mu))$, $\gamma = \frac{\lambda(1+\mu)}{\mu(1+\lambda)}$, and the lower bound (3.19) is the λ-timing capacity $\overline{C}_\mu(\lambda)$ as given by (3.15).

Consider a $\cdot/Geo^{(b)}/1$ queue whose service distribution is Geo $\left(\frac{\lambda}{\gamma(1+\lambda)}\right)$. Note that if $S_\gamma \sim$ Geo $\left(\frac{\lambda}{\gamma(1+\lambda)}\right)$, then the solution of (3.18) for S_γ is γ. From (3.15), the right side of (3.19) is in fact the λ-timing capacity $C_{S_\gamma}(\lambda)$ of this queue. In conjunction with Proposition 3.2, this gives the following bounds on the λ-timing capacity of a queue with an arbitrary service distribution:

$$C_{S_\gamma}(\lambda) \leq C_S(\lambda) \leq \sup_{Q \in \mathcal{B}_S(\lambda)} I(X_Q; \min(X_Q, S)), \tag{3.21}$$

where $\mathcal{B}_S(\lambda)$ and X_Q are as defined in Proposition 3.2.

3.7 Discussion

An interesting interpretation of the first inequality in (3.21) is the following. Consider the class of service distributions for which the solution of (3.18) for a given λ is a particular value γ. Then in this class, the $\cdot/Geo^{(b)}/1$ queue with service distribution Geo $\left(\frac{\lambda}{\gamma(1+\lambda)}\right)$ has the least λ-timing capacity. Further, note that $C_{S_\gamma}(\lambda)$ is a decreasing function of γ for $\gamma \in \left(\frac{\lambda}{1+\lambda}, 1\right)$. Hence, for a queue with service random variable S for which the solution γ' is less than γ, we have $C_{S_\gamma}(\lambda) \leq C_{S_{\gamma'}}(\lambda) \leq C_S(\lambda)$. That is, among all the service distributions for which the solution of (3.18) is no more than a given γ for a given λ, the $\cdot/Geo^{(b)}/1$ queue with service distribution Geo $\left(\frac{\lambda}{\gamma(1+\lambda)}\right)$ has the least λ-timing capacity. Thus, while the $\cdot/Geo^{(b)}/1$ queue does not have the least λ-timing capacity among all queues with a given mean service rate, it does have the least λ-timing capacity among all queues with a given upper bound on the solution of (3.18).

However, there is a restricted class of queues with the same mean service rate in which the $\cdot/Geo^{(b)}/1$ queue does have the least λ-timing capacity. For the $\cdot/Geo^{(b)}/1$ queue of rate μ, the solution of (3.18) is $\gamma_0 = \frac{\lambda(1+\mu)}{\mu(1+\lambda)}$. For a queue with an arbitrary service distribution S and mean service rate μ packets/slot, if the solution of (3.18) satisfies $\gamma \leq \gamma_0$, then $C_{S_{\gamma_0}}(\lambda) \leq C_{S_\gamma}(\lambda) \leq C_S(\lambda)$. Thus, in the class of queues having a given mean service rate μ for which the solution γ of (3.18) for a given λ satisfies $\gamma \leq \frac{\lambda(1+\mu)}{\mu(1+\lambda)}$, the $\cdot/Geo^{(b)}/1$ queue has the least λ-timing capacity.

In fact, since the timing capacity is the supremum of the λ-timing capacity over $0 \leq \lambda \leq \mu$, we can also assert the following. Suppose that for a particular service distribution, it is true that $\gamma \leq \frac{\lambda(1+\mu)}{\mu(1+\lambda)}$ for all $0 < \lambda \leq \mu$. Then the timing capacity of the queue with that service distribution is at least as much as the timing capacity of the $\cdot/Geo^{(b)}/1$ queue with mean μ.

As an example, for the queue with Bernoulli service distribution with mean μ, we have $\gamma = \frac{\lambda}{\mu} < \frac{\lambda(1+\mu)}{\mu(1+\lambda)}$. It was noted in Example 3.2 that this queue is in fact the

$\cdot/Geo^{(s)}/1$ queue. Thus the $\cdot/Geo^{(s)}/1$ queue has a higher λ-timing capacity than the $\cdot/Geo^{(b)}/1$ queue with the same mean service rate. Since this is true for all $0 \leq \lambda \leq \mu$, the same holds for the timing capacities as well. As another example, for the queue with Poisson service distribution with mean μ, it can be shown that $\gamma \leq \frac{\lambda(1+\mu)}{\mu(1-\lambda)}$ for all $0 \leq \lambda \leq \mu$. Hence this queue also has a higher timing capacity than the $\cdot/Geo^{(b)}/1$ queue with the same service rate.

Note, however, that it is possible to construct a service distribution with mean μ that satisfies $\gamma \leq \frac{\lambda(1+\mu)}{\mu(1+\lambda)}$ for some values of λ, but not for other values of λ. As a simple example, take $\mu = 1$. Let $S_1 = 1$ with probability 1, so that $\phi_{S_1}(x) = x$, and let $S_2 = K$ with probability $1/K$ and 0 with probability $1 - 1/K$. Let $\phi_0(\cdot)$ be the generating function of the geometric distribution with mean 1. Then $\phi_0(x) \geq \phi_{S_1}(x)$ for $0 \leq x \leq 1$, and for $K \geq 3$, $\phi_0(x) \leq \phi_{S_2}(x)$. Now construct a random variable S whose distribution is a convex combination of the distributions of S_1 and S_2. We can choose an appropriate convex combination so that the graph of the generating function of the distribution of S intersects the graph of $\phi_0(\cdot)$ at a single point in $(0, 1)$. Then there will be some values of λ for which $\gamma \geq \gamma_0$, and some values for which $\gamma \leq \gamma_0$. In this sense the extremal nature of the geometric service distribution is restricted.

Figure 3.2 shows a plot of the timing capacity of the $\cdot/Geo^{(b)}/1$ queue, given by equation (3.16), as a function of the mean service rate. The timing capacity is an increasing function of the service rate μ, and has the value 0.415 bits/slot at $\mu = 1$ packet/slot, which is less than the timing capacity (1 bit/slot) of the $\cdot/Geo^{(s)}/1$ queue with $\mu = 1$ obtained in Chapter 2. The timing capacity of the $\cdot/Geo^{(b)}/1$ queue is a concave function of μ, whereas the timing capacity of the $\cdot/Geo^{(s)}/1$ queue is a convex function.

Figure 3.3 shows a comparison of the timing capacities of the $\cdot/Geo^{(s)}/1$ queue, the $\cdot/Geo^{(b)}/1$ queue, and the continuous time queue with the exponential service time distribution, as a function of the mean service rate μ. Note that for the $\cdot/Geo^{(s)}/1$

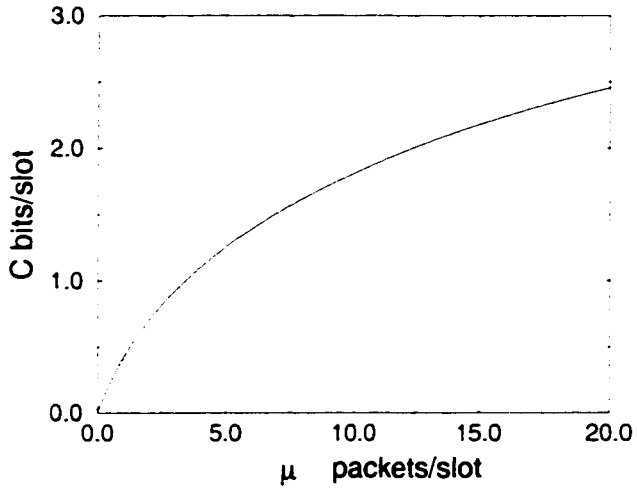


Figure 3.2: Timing capacity as a function of service rate for the $\cdot/Geo^{(b)}/1$ queue.

queue, $0 \leq \mu \leq 1$, so we restrict attention to this range for the comparison. Also the slot size is assumed to be one second, so that the timing capacity of the queue with exponential service time distribution, in bits/sec, can be compared with the timing capacities of the two other queues, which are in bits/slot. The timing capacity of the $\cdot/Geo^{(b)}/1$ queue is less than the timing capacity of the other two queues. For small μ, the slopes of all three curves are the same, and equal to $(\log_2 e)/e$. We observed in Example 3.2 that the $\cdot/Geo^{(s)}/1$ queue analyzed in Chapter 2 is a special case of the model of Section 3.1, and as noted earlier, has a higher capacity than the $\cdot/Geo^{(b)}/1$ queue.

3.8 Discrete-Time Queues with Multiple Servers

A natural question to consider following the analysis of single-arrival and batch-arrival queues is that of the timing capacity of queues with multiple servers. Multiple servers can be used to model the situation where packets are routed along the first available path out of K independent paths, with the delay experienced by a packet over each

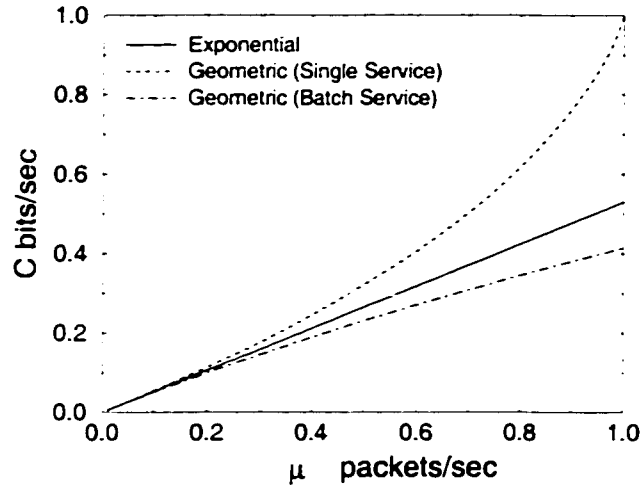


Figure 3.3: Comparison of timing capacities for various queues (assuming 1 slot = 1 second).

individual path modeled as the service time of the packet in a server. It was noted in [14] that the timing capacity of queues with multiple servers appears to be significantly harder to analyze than that of single-server queues. In this section, we consider the timing capacity of discrete-time queues with multiple servers and a common infinite buffer. We obtain a single-letter upper bound on the timing capacity of the $\cdot/Geo^{(s)}/K$ queueing system, i.e. the queueing system with K servers, each of which has i.i.d. geometrically distributed service times. This upper bound is actually equal to the capacity subject to a constraint of the K-output binomial channel, a memoryless channel for which the conditional output probabilities are binomial distributions. The bound can be evaluated numerically for any finite K. For $K = \infty$, we obtain a necessary and sufficient condition that an input probability distribution must satisfy to attain this upper bound, and establish that such a distribution does exist.

The queueing model for the $\cdot/Geo^{(s)}/K$ is the generalization of the single-arrival single-service model of Chapter 2 to multiple servers, each of which can serve at most one packet per slot. Arriving packets are stored in a common buffer and are served in

order by the next available server. Note that there can be multiple departures from the system in a slot if more than one server finishes service in that slot. We allow for batch arrivals in a slot as well. Service times of all packets are assumed to be i.i.d. Information can be encoded in the number of packets arriving at the buffer in each slot. The receiver observes the number of packets departing in each slot and attempts to infer the transmitted codeword. We assume that the receiver only observes the total number of departures in each slot, and cannot identify the server that served each packet.

Example 3.4: Consider a $\cdot/Geo^{(s)}/K$ system whose service times have the $\text{Geo}^{+}(\mu)$ distribution. Suppose that arrivals occur at the end of a slot, and departures occur at the beginning of a slot. Let X_i be the number of packets in the system at the beginning of slot i, D_i the number of departures in slot i, and A_i the number of new arrivals. The dynamic evolution of the queue is defined by

$$X_{i+1} = X_i - D_i + A_i.$$

The number of packets actually in service is $\min(X_i, K)$, and each of these packets can independently depart in slot i with probability μ. Thus the conditional probability mass function of D_i given X_i is

$$\begin{aligned} P_{D_i|X_i}(i|j) &= B(\min(j,K), i, \mu) \\ &= \binom{\min(j,K)}{i} \mu^i (1-\mu)^{\min(j,K)-i}, \quad 0 \le i \le \min(j,K). \end{aligned} \tag{3.22}$$

where $B(n, m, p)$ is the probability mass at m of the binomial distribution with parameters n and p. For the $\cdot/Geo^{(s)}/\infty$ queueing system, the corresponding conditional probability is

$$P_{D_i|X_i}(k|n) = B(n,k,b) = \binom{n}{k} \mu^k (1-\mu)^{n-k}, \quad 0 \le k \le n. \tag{3.23}$$

◁

The conditional probability mass function given by (3.22) can also be considered as the channel transition probability of a discrete memoryless channel. We will refer

to this channel as the *K-output binomial channel.* The conditional probability (3.23) also defines a memoryless channel, but the input and output alphabets are not finite. We will refer to this channel as the *infinite-output binomial channel.*

3.8.1 Timing Codes for $\cdot/Geo^{(s)}/K$

Timing codes and achievable rates for the $\cdot/Geo^{(s)}/K$ (for finite or infinite K) are defined in the same way as for the batch-arrival batch-service queues described in Section 3.3. An $(n, M_n, N_n, \epsilon_n)$ timing code consists of M_n codewords, each of which is a sequence $A_0, \cdots, A_{n-1}$ of non-negative integers, where A_i represents the number of arrivals to the $\cdot/Geo^{(s)}/K$ system in slot i. The receiver observes the number of departures $D_1, \cdots, D_n$ in slots $1, \cdots, n$ and attempts to infer the transmitted codeword. The expected total number of packets departing in slots $1, \cdots, n$ is N_n, where the expectation is over all the codewords and over the random service times of the packets in the $\cdot/Geo^{(s)}/K$ system.

Rate R is said to be achievable if there is a sequence $(n, M_n, N_n, \epsilon_n)$ of timing codes that has a subsequence $(n_k, M_{n_k}, N_{n_k}, \epsilon_{n_k})$ such that

$$\lim_{k\to\infty} \frac{\log M_{n_k}}{n_k} = R \text{ and } \lim_{k\to\infty} \epsilon_{n_k} = 0.$$

The timing capacity C_K of the $\cdot/Geo^{(s)}/K$ is the supremum of the set of achievable rates. The subscript K indicates the number of servers in the system.

Rate R is said to be achievable at departure rate λ if there exists a sequence of timing codes $(n, M_n, N_n, \epsilon_n)$ that has a subsequence $(n_k, M_{n_k}, N_{n_k}, \epsilon_{n_k})$ for which

$$\lim_{k\to\infty} \frac{\log M_{n_k}}{n_k} = R, \lim_{k\to\infty} \frac{N_{n_k}}{n_k} = \lambda, \text{ and } \lim_{k\to\infty} \epsilon_{n_k} = 0. \tag{3.24}$$

The λ-timing capacity $C_K(\lambda)$ of the $\cdot/Geo^{(s)}/K$ is the supremum of all rates that are achievable at departure rate λ. Again, the subscript K indicates the number of servers.

In the case of the single-arrival queues of Section 2.1 and the batch-arrival queues of Section 3.1, the packets depart in the order of their arrival. This allowed us to express the overall capacity of queues with information-carrying packets as the sum of the timing capacity and the capacity of the packet contents. In the case of multiple servers, however, the packets can depart out of order. This necessitates that some part of the packet contents be devoted to carry additional information such as sequence numbers. Moreover, the amount of such additional information depends on the timing code used, since the order of departure is dependent on the arrival times of the packets. This makes the problem of the overall capacity of multiple-server queues with information-bearing packets difficult to analyze. In this section, we will focus only on the timing capacity, will not consider the case of information-bearing packets.

Before we proceed to the main results of this section, we mention a simple upper bound to the λ-timing capacity $C_K(\lambda)$ of the $\cdot/Geo^{(s)}/K$ system in terms of the λ-timing capacity of K independent $\cdot/Geo^{(s)}/1$ queues.

Example 3.5: Consider a system of K identical independent parallel $\cdot/Geo^{(s)}/1$ queues having separate buffers for each server. We remarked earlier that the $\cdot/Geo^{(s)}/K$ can be considered as a model for a network with K paths between the source and the receiver, with packets waiting in a common buffer at the transmitter and being routed over the first available path. The K parallel $\cdot/Geo^{(s)}/1$ queues model the situation where the transmitter chooses the path that each packet gets routed over, and maintains separate queues for each path. The encoder can encode information in the timing of packets by controlling the arrival process to each of the K queues. In the case of the $\cdot/Geo^{(s)}/K$, however, the encoder can only control the combined arrival process to the common buffer and has no control over which server each packet gets served by. Moreover, for the $\cdot/Geo^{(s)}/K$, the receiver only observes the total number of packets departing in each slot and cannot identify the server for each packet, whereas in the case of the system with K parallel queues, the receiver can separately observe the departures from the K servers. It is thus clear that for a given

overall departure rate λ and a given service rate μ per server, the λ-timing capacity of the system of K parallel $\cdot/Geo^{(s)}/1$ is at least as much as that of the $\cdot/Geo^{(s)}/K$ system. It is easy to verify that for a given overall departure rate λ, the λ-timing capacity of the system with K parallel $\cdot/Geo^{(s)}/1$ queues is given by $KC_1(\lambda/K)$, and the timing capacity is simply $KC_1(\mu)$. Thus the timing capacity of the $\cdot/Geo^{(s)}/K$ system is no more than than K times the timing capacity of the $\cdot/Geo^{(s)}/1$ system. The upper bound that we obtain in the following subsection is tighter than this simple bound. ◁

3.8.2 An Upper Bound to the Timing Capacity

Along the same lines as Proposition 3.1, we can show the following relation between the timing capacity and the λ-timing capacity for queues with multiple servers.

Proposition 3.6. *For a $\cdot/Geo^{(s)}/K$ queue in which the service rate of each server is μ packets/slot, for finite K,*

$$C_K = \sup_{0 \le \lambda \le K\mu} C_K(\lambda).$$

The proof is almost identical to the proof of Proposition 3.1, and is omitted.

Using Fano's lemma, we can obtain the following upper bound to the λ-timing capacity of the $\cdot/Geo^{(s)}/K$ queueing system. The derivation is entirely analogous to the converse result of Section 3.4, and is omitted.

Proposition 3.7. *For any probability mass function Q on the non-negative integers, let $X_Q \sim Q$. Let D_Q be a random variable whose conditional probability mass function given X_Q is (3.22) for finite K, and (3.23) for $K = \infty$. Let $\mathcal{W}(\lambda; \mu, K)$ be the set of probability mass functions Q on the non-negative integers for which $E[D_Q] = \lambda$. Then the λ-timing capacity of the $\cdot/Geo^{(s)}/K$ queue satisfies*

$$C_K(\lambda) \le C'_K(\lambda) = \sup_{Q \in \mathcal{W}(\lambda;\mu,K)} I(X_Q; D_Q). \tag{3.25}$$

Example 3.6: Consider the case $K = 1$, i.e. the $\cdot/Geo^{(s)}/1$ queue. For a given input probability mass function Q, the output D_Q defined by the conditional probability (3.22) can be written as $D_Q = \min(X_Q, S)$, where $S \sim \text{Ber}(\mu)$ is a Bernoulli random variable that is independent of the input X_Q. With this point of view, the situation is identical to that described in Example 3.2, which establishes the relation of the $\cdot/Geo^{(s)}/1$ queue to the batch-service queues described in Section 3.1. In the light of Example 3.2, the upper bound of (3.25) is tight for the case $K = 1$. ◁

In the remainder of this section, we will focus on evaluating the upper bound $C'_K(\lambda)$ to the λ-timing capacity $C_K(\lambda)$ of the $\cdot/Geo^{(s)}/K$ queueing system.

We first make some observations on this upper bound. Note that the constraint $E[D_Q] = \lambda$ on the output D_Q in the definition of $\mathcal{W}(\lambda; \mu, K)$ in Proposition 3.7 is actually equivalent to a constraint on the input, since $E[D_Q] = \mu E[\min(X_Q, K)]$. For $K = \infty$, this constraint simply becomes $E[D_Q] = \mu E[X_Q]$. Moreover, it is easy to see that for finite K, $C'_K(\lambda)$ is the capacity, subject to this constraint, of the K-output binomial channel, i.e. the discrete memoryless channel whose transition probabilities are given by the conditional distribution (3.22). For $K = \infty$, $C'_\infty(\lambda)$ is clearly an upper bound to the capacity of the infinite output binomial channel with transition probability (3.23), subject to the constraint $E[D_Q] = \mu E[X_Q]$. For the case of channels with infinite input and output alphabets, however, it is not always true that the supremum in (3.25) is actually achieved. We will return to this question later.

Consider the channel transition probability (3.22) of the K-output binomial channel for finite K. For any input probability mass function Q to the binomial channel, let Q' be the probability mass function such that $Q'(i) = Q(i)$, $i = 0, \cdots, K-1$, and $Q'(K) = \sum_{i=K}^{\infty} Q(i)$. Q' is a truncation of Q to $\{0, 1, \cdots, K\}$ that accumulates all the mass of Q beyond K at K. It is easy to see that the input-output mutual information of the binomial channel using Q' as the input probability mass function is the same as that obtained using Q, i.e. $I(X_{Q'}; D_{Q'}) = I(X_Q; D_Q)$. Thus, to evaluate the upper

bound of (3.25) for finite K, we need only restrict attention to those input probability mass functions that put all the mass on $\{0, 1, \cdots, K\}$.

Moreover, for finite K, it is easy to see from the definition of $C'_K(\lambda)$ in (3.25) that $C'_K(0) = C'_K(K\mu) = 0$, and that $C'_K(\cdot)$ is a concave function over $[0, K\mu]$. Hence there exists $\lambda^* \in [0, K\mu]$ such that

$$\begin{aligned} C'_K(\lambda) &= \sup_{0 \le x \le \lambda} C'_K(x), \quad 0 < \lambda \le \lambda^*, \\ C'_K(\lambda) &= \sup_{\lambda \le x \le K\mu} C'_K(x), \quad \lambda^* < \lambda \le K\mu. \end{aligned}$$

Furthermore, $C'_K(\lambda^*)$ is the maximum value of $C'_K(\cdot)$ over $[0, K\mu]$. Thus we have

$$C'_K(\lambda) = \sup_{Q: \mu E[\min(X_Q, K)] \le \lambda} I(X_Q; D_Q), \quad 0 \le \lambda \le \lambda^*, \tag{3.26a}$$

$$C'_K(\lambda) = \sup_{Q: \lambda \le \mu E[\min(X_Q, K)] \le K\mu} I(X_Q; D_Q), \quad \lambda^* < \lambda < K\mu. \tag{3.26b}$$

Moreover, for any probability mass function Q, $\mu E[\min(X_Q, K)] \le K\mu$. Hence the set of probability mass functions $\{Q : \lambda \le \mu E[\min(X_Q, K)] \le K\mu\}$ that satisfy the constraint in (3.26b) can be represented simply as $\{Q : \mu E[\min(X_Q, K)] \ge \lambda\}$, or equivalently, as $\{Q : K\mu - \mu E[\min(X_Q, K)] \le K\mu - \lambda\}$. Thus

$$C'_K(\lambda) = \sup_{Q: \mu E[\min(X_Q, K)] \le \lambda} I(X_Q; D_Q), \quad 0 \le \lambda \le \lambda^*, \tag{3.27a}$$

$$C'_K(\lambda) = \sup_{Q: K\mu - \mu E[\min(X_Q, K)] \le K\mu - \lambda} I(X_Q; D_Q), \quad \lambda^* < \lambda < K\mu. \tag{3.27b}$$

For $K = \infty$, it is clear that $C'_\infty(0) = 0$, and $C'_\infty(\lambda)$ is concave and increasing in λ. In this case, for any $\lambda > 0$,

$$C'_\infty(\lambda) = \sup_{0 \le x \le \lambda} C'_\infty(x).$$

Thus we have,

$$C'_\infty(\lambda) = \sup_{Q: \mu E[X_Q] \le \lambda} I(X_Q; D_Q). \tag{3.28}$$

With these observations, the upper bound of (3.25) takes the form of a maximization of the mutual information over a set of input distributions defined by inequality

constraints, as given by (3.27) for finite K, and by (3.28) for $K = \infty$. The maximization of (3.27) for finite K is of the well known form (see, for instance, [12, Section 2.3.1])

$$\overline{C}(\lambda) = \sup_{Q:E[f(X)]\leq\lambda} I(Q), \tag{3.29}$$

where $I(Q)$ is the input-output mutual information of a given discrete memoryless channel (in this case, the K-output binomial channel defined by (3.22)) when the input distribution is Q. In the case of (3.27a), the constraint function is $f(x) = \min(x, K)$, while in the case of (3.27b), $f(x) = K\mu - \mu\min(x, K)$. From [12, Lemma 2.3.1] that

$$\overline{C}(\lambda) = \min_{\gamma\geq 0} [F(\gamma) + \gamma\lambda],$$

where

$$F(\gamma) = \max_{Q} [I(Q) - \gamma E_Q[f(X)].$$

For finite K, for any fixed γ, $F(\gamma)$ can be evaluated using a version of the Arimoto-Blahut algorithm with an input constraint due to Csiszar [12, Theorem 2.3.3, p. 140], which works as follows. Let Q_1 be an arbitrary input distribution such that $Q_1(x) > 0$ for all input letters x. For each letter x of the input alphabet, define

$$Q_{i+1} = A_{i+1}^{-1}\ Q_i(x)\ \exp\left\{D(\ P_{D|X}(\cdot|x)\ ||\ P_{D_{Q_i}}) - \gamma f(x)\right\}, \tag{3.30}$$

where A_{i+1} is defined by the condition that $\sum_x Q_{i+1}(x) = 1$, $P_{D_{Q_i}}$ is the probability mass function of the output when the input distribution is Q_i, and $D(P_1||P_2)$ is the Kullback-Leibler distance between two probability distributions. Then the sequence of probability mass functions $Q_i(\cdot)$ converges to a mass function Q^* such that

$$F(\gamma) = I(Q^*) - \gamma E_{Q^*}[f(X)].$$

Q^* attains the supremum of the mutual information in (3.29) for $\lambda = E_{Q^*}[f(X)]$. The special case $\gamma = 0$ corresponds to the absolute maximum of $\overline{C}(\lambda)$ over all possible

values of λ. Thus by tracing out the curve $F(\gamma) + \gamma E_{Q^*}[f(X)]$ for various γ, we can obtain a plot of $\overline{C}(\lambda)$.

Figure 3.4 shows a plot of the upper bound $C'_K(\lambda)$ as a function of λ for various

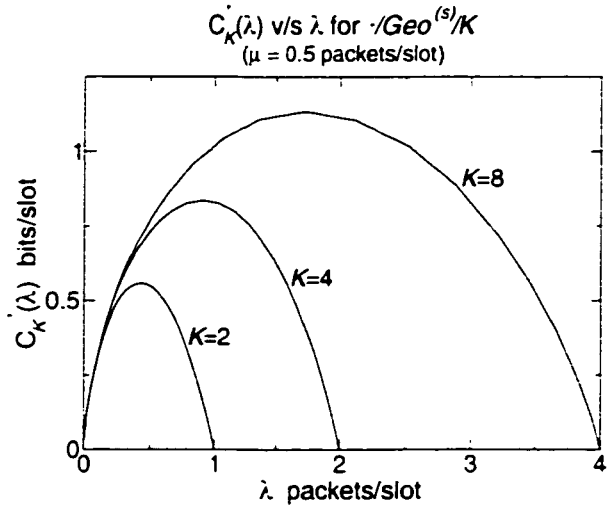


Figure 3.4: Upper bound to the λ-timing capacity of $\cdot/Geo^{(s)}/K$ as a function of λ, for various K with $\mu = 0.5$.

values of K, evaluated using the iterative algorithm (3.30). The service rate per server is fixed at $\mu = 0.5$ in each case. As expected, $C'_K(\cdot)$ is a concave function of λ over $[0, K\mu]$.

Figure 3.5 shows a plot of the upper bound C'_K to the timing capacity of the $\cdot/Geo^{(s)}/K$ system as a function of the service rate μ for various values of K, evaluated using this algorithm. It is seen that the curves for $K > 2$ are neither concave nor convex. There does not appear to be any interesting interpretation of this lack of convexity and concavity.

It is clear that the iterative algorithm defined by (3.30) works only for finite K, because the computation of A_{i+1} can only be accomplished for finite K. For $K = \infty$, there does not appear to be any way to numerically evaluate the upper bound of

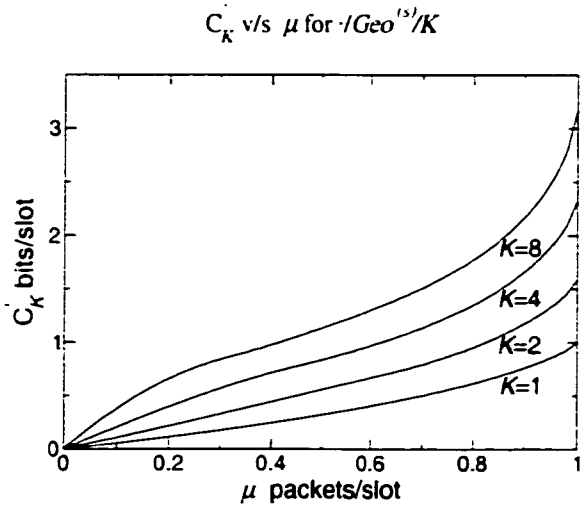


Figure 3.5: Upper bound to the timing capacity of $\cdot/Geo^{(s)}/K$ as a function of the service rate μ for various K.

(3.28). In fact, it is not clear a priori whether the supremum in (3.28) is actually achieved by some input distribution. In Appendix C, we prove that such an input distribution does exist. The proof draws on some standard results from functional analysis and measure theory that define a suitable notion of convergence on the space of input distributions of the channel, i.e. the space of probability measures on the non-negative integers. This notion of convergence is equivalent to the weak convergence of probability measures. We show that the input-output mutual information of the infinite output binomial channel, considered as a functional on the space of input probability measures, is *weak* continuous*, and the set of input distributions satisfying the constraint in (3.28) is *weak* compact*. A standard result of functional analysis asserts that a weak* continuous functional attains its maximum on a weak* compact set. The proof of the weak* continuity of the mutual information functional is somewhat tedious, and we relegate it to Appendix C.

For channels with finite input and output alphabets, it is also known [12, Problem

2.3.2, p. 147] that for the optimization of (3.29), a necessary and sufficient condition for an input distribution Q^* to achieve the supremum in (3.29) is that there should exist constants $\gamma \geq 0$ and $M \geq 0$ such that

1. $D(P_{D|X}(\cdot|x)||P_{D_{Q^*}}) - \gamma f(x) \leq M$, with equality for all x such that $Q^*(x) > 0$, and

2. $\gamma(E_{Q^*}[f(X)] - \lambda) = 0$.

In Appendix C, we also establish a similar set of necessary and sufficient conditions that characterize a distribution that achieves the supremum in (3.28) for the infinite output binomial channel. The necessary and sufficient conditions for an input probability mass function Q^* to attain the supremum in (3.28) are that there should exist $\gamma \geq 0$ and $\overline{C}$ such that $\gamma(E_{Q^*}[X] - \lambda/\mu) = 0$ and

$$\left[B(n,i,\mu)\log\frac{B(n,i,\mu)}{P_{D_{Q^*}}(i)}\right] - \gamma\mu n \leq \overline{C} - \gamma\lambda \quad \forall n,$$

with equality for all n such that $Q^*(n) > 0$. The results used to establish these conditions are the Lagrange multiplier theorem and the Kuhn-Tucker theorem, both common tools in the theory of convex optimization. The approach to establishing the existence of the optimal distribution and the conditions that characterize it closely follows that used in [1] and [40] to characterize the capacity achieving distributions of, respectively, the discrete-time Rayleigh Fading Channel with an input power constraint and the additive Gaussian noise channel with an input amplitude constraint.

Unfortunately, however, there does not appear to be any systematic way to solve these equations. The Poisson and geometric distributions, which have interesting queueing theoretic properties for the $\cdot/Geo^{(s)}/\infty$ system, do not satisfy these conditions.

3.9 Conclusion

In this chapter, we first analyzed the timing capacity of a model of batch-arrival batch-service queues motivated by ATM switches and the leaky bucket flow control schemes. We obtained a closed-form expression for the capacity of the queue with geometric service distribution within this model. We established a lower bound on the timing capacity of queues with arbitrary service distribution within this model based on a general queueing theoretic property of such queues, and examined the extremal nature of the geometric server in terms of this property. We then analyzed the timing capacity of queues with multiple servers, each of which can serve at most one packet per second. We obtained an upper bound on the timing capacity of such queues in terms of the capacity of the binomial channel subject to output constraints. For a finite number of servers, this upper bound can be evaluated numerically. For an infinite number of servers, we established the existence of an input distribution that achieves the capacity of the infinite output binomial channel subject to a certain constraint. We also obtained a set of necessary and sufficient conditions that this input distribution must satisfy.

Chapter 4

ON THE INFORMATION ABOUT MESSAGE ARRIVAL TIMES REQUIRED FOR IN-ORDER DECODING

4.1 Introduction

Consider a source A communicating with a receiver B. Suppose that messages arrive at A according to a point process at random times $X_1, X_2, \cdots, X_n$ to be transmitted to B, and are decoded at B at times $Y_1, Y_2, \cdots, Y_n$, where $Y_i > X_i$, $i = 1, \cdots, n$. The fact that B is able to decode message i at time Y_i means that by time Y_i, B must have been informed that message i arrived at A. Thus B must receive some information about the times at which messages arrived at A, apart from the information about the contents of the messages. This amount of information about message arrival times that the receiver must receive was first analyzed by Gallager [17]. He reasoned that this information must be at least equal to $I(X_1, \cdots, X_n; Y_1, \cdots, Y_n)$, although the actual amount of such information provided by each individual strategy may be higher. Gallager took the approach that $Y_1, \cdots, Y_n$ can be treated as a reproduction of the message arrival times $X_1, \cdots, X_n$, and the average delay per message $(1/n)\sum_{i=1}^{n}(Y_i - X_i)$ can be considered as a measure of the distortion between $X_1, \cdots, X_n$ and $Y_1, \cdots, Y_n$. The minimum information required for any source encoding of $X_1, \cdots, X_n$ that can create a reproduction $Y_1, \cdots, Y_n$ with average delay per message no more than d is the rate-distortion function $R(d)$ for the message arrival process $\{X_i\}$ with average message delay as the distortion measure. In [17], Gallager showed that if $\{X_i\}$ is a Poisson process of rate λ, the rate-distortion function satisfies

$$R(d) = \lim_{n\to\infty} \inf_{\mathcal{P}_n} \frac{1}{n} I(X_1, \cdots, X_n; Y_1, \cdots, Y_n) \geq -\log\left(1 - e^{-\lambda d}\right) \text{ bits/message}, \quad (4.1)$$

where the infimum is over the set $\mathcal{P}_n$ of conditional distributions of $(Y_1, \cdots, Y_n)$ given $(X_1, \cdots, X_n)$ such that $Y_i > X_i$, $i = 1, \cdots, n$, and $(1/n)\sum_{i=1}^{n} E[Y_i - X_i] \leq d$. This lower bound on the amount of information about the message arrival times that the receiver must have received is independent of the channel or network between the source and the receiver, and of the particular strategy used for communication. Gallager considered the case of a network rather than a point-to-point channel, and thus allowed for the possibility that the messages may not be decoded at the receiver in the same order they arrived at the source. Thus in his formulation, although $Y_i > X_i$, it is not required that $Y_{i+1} > Y_i$. In the review of [17] in [14], it was observed that a possible source encoding for the message arrival process is to consider it as the arrival process to a $\cdot/M/1$ queue (i.e. a single-server queue with exponentially distributed service times), and treat the departure process of the queue as an order-preserving reproduction of the message arrival process. For the Poisson message arrival process with rate λ, if the service rate of the $\cdot/M/1$ queue is $\lambda + 1/d$, then the average message delay is d, and the input-output mutual information per message, using the results of [4], is $\log(1 + 1/\lambda d)$. It was noted in [14] that this is greater than the lower bound of (4.1), although the two become essentially the same as $\lambda d \to 0$.

In this chapter we focus on the situation in which messages are transmitted one at a time in order and are decoded at the receiver in the same order they arrived at the source, for instance a virtual circuit or a point-to-point channel. If the transmission time for each message is non-zero, the message decoding times will satisfy $Y_{i+1} > Y_i$ and $Y_i > X_i$. So we can view $Y_1, Y_2, \cdots$ as the departure process from a hypothetical first-in-first-out single server queue whose arrival process is $X_1, X_2, \cdots$. With this point of view we can associate with each message i a *service time* $S_i = Y_i - \max(Y_{i-1}, X_i)$, and the average service time per message $(1/n)\sum_{i=1}^{n} S_i$ can be thought of as a measure of the distortion between $X_1, \cdots, X_n$ and $Y_1, \cdots, Y_n$. If the messages are transmitted one at a time over the point-to-point channel, then S_i is precisely the time available for transmitting useful information about message i to the

receiver, and thus directly influences the achievable probability of message error. In this sense the service time distortion measure does have a physical significance. Consider any such strategy that lets the decoder make its decoding decision with an average service time per message no more than d. The amount of information about message arrival times that the receiver must receive is at least $I(X_1, \cdots, X_n; Y_1, \cdots, Y_n)$, where the conditional probability $P(Y_1, \cdots, Y_n | X_1, \cdots, X_n)$ is zero unless $Y_{i+1} > Y_i$ and $Y_i > X_i$ for all $i = 1, \cdots, n$, and satisfies $(1/n) \sum_{i=1}^{n} E[S_i] \leq d$. As in [17], the information about message arrival times that the receiver must receive will be lower bounded by the rate-distortion function for encoding the sequence of message arrival times $X_1, \cdots, X_n$ to generate an order-preserving reproduction $Y_1, \cdots, Y_n$, such that the average service time distortion between the two sequences is limited to d. If the message arrival process is Poisson with rate λ, we show that this rate-distortion function is given by

$$R(d) = \begin{cases} \log(1/\lambda d) \quad \text{bits/message}, & 0 < d < (1/\lambda) \\ 0, & d \geq (1/\lambda). \end{cases} \tag{4.2}$$

The optimal source encoding for the Poisson process $\{X_i\}$ with the service time distortion results when the reproduction $\{Y_i\}$ is the output process of a $\cdot/M/1$ queue whose arrival process is $\{X_i\}$ and average service time is d. Thus although, as noted by [14], the $\cdot/M/1$ queue does not meet the lower bound of (4.1) for source encoding of Poisson processes subject to a mean delay constraint, it does possess an extremal character as the optimal encoding mechanism for order-preserving encoding of the Poisson process with the service time distortion measure.

It may also be noted that the expression for $R(d)$ given by (4.2) is also identical to the timing capacity (measured per packet rather than per unit time) of a $\cdot/M/1$ queue with mean service time d, obtained in [4], when restricted to timing codes whose average packet departure rate is λ.

An alternative single-letter distortion criterion for point processes was introduced

by Verdu [45]. The rate distortion function for Poisson processes under this criterion, obtained in [45], is in fact identical to the rate distortion function for Poisson processes under the service time distortion measure given by (4.2). The similarities between this rate distortion function and that of a sequence of i.i.d. Gaussian random variables with the squared-error distortion measure have been pointed out in [45].

We also show the following analogous result for discrete-time message arrival processes. If the message arrival process is Bernoulli with rate λ ($0 < \lambda < 1$), then for $1 \leq d < 1/\lambda$,

$$R(d) = \frac{h(\lambda)}{\lambda} - dh\left(\frac{1}{d}\right) \tag{4.3}$$

where $h(\cdot)$ is the binary entropy function. In this case, the optimal source encoding mechanism resembles the $\cdot/Geo^{(s)}/1$ queue (i.e. the discrete-time single server FIFO queue with i.i.d. geometrically distributed service times) having mean service time d.

Some other rate distortion functions of the Poisson process under alternative distortion measures have been studied. In [36], the rate distortion function of a Poisson sequence under a single letter magnitude error criterion is analyzed. In [37], this work is extended to cover various other types of reproductions of the Poisson process, including reproduction of the number of events in a given interval, and the times between successive events. A similar study is presented in [38]. The reproductions associated with all these distortion measures are non-causal, in contrast to the distortion measures considered in [17] and in this chapter.

4.2 *Rate Distortion Function of the Poisson Process for Service-Time Distortion*

Consider a source at which messages arrive at random times $X_1, \cdots, X_n$. Let the contents of the messages be $U_1, \cdots, U_n$. Denote the decoded messages as $V_1, \cdots, V_n$, and the times at which the the decoding decisions are made as $Y_1, \cdots, Y_n$. Let the message interarrival times be $A_i = X_i - X_{i-1}$, and the times between successive

message decodings be $D_i = Y_i - Y_{i-1}$. Then the amount of information the decoder must receive is

$$I(U_1, \cdots, U_n, X_1, \cdots, X_n; V_1, \cdots, V_n, Y_1, \cdots, Y_n).$$

Using standard identities about mutual information, it is easy to show that

$$\begin{aligned} & I(U_1, \cdots, U_n, X_1, \cdots, X_n; V_1, \cdots, V_n, Y_1, \cdots, Y_n) \\ \geq\ & I(U_1, \cdots, U_n; V_1, \cdots, V_n) + I(X_1, \cdots, X_n; Y_1, \cdots, Y_n) \\ =\ & I(U_1, \cdots, U_n; V_1, \cdots, V_n) + I(A_1, \cdots, A_n; D_1, \cdots, D_n) \end{aligned} \tag{4.4}$$

The first term on the right side of (4.4) represents the amount of information about the message contents that the receiver must have received. This depends on the probability of error of the decoding scheme, and in the case where the decoded messages are perfectly error-free, is at least equal to the entropy of the message contents, i.e.

$$I(U_1, \cdots, U_n; V_1, \cdots, V_n) \geq nH(M).$$

The first term of (4.4) is not of primary interest to us, however. Our interest is in the second term on the right side of (4.4).

The second term in (4.4) can be interpreted as the amount of information that must be transmitted to inform the receiver of the message arrival times. If the decoder decodes the message i at time Y_i, then it must have received enough information, explicitly or implicitly, to decide that message i arrived before time Y_i. Thus if the decoder makes its decoding decisions on messages $1, \cdots, n$ at times $Y_1, \cdots, Y_n$, it must have obtained some information about the message arrival process $X_1, \cdots, X_n$. The amount of such information is quantified by the second term of (4.4).

Suppose that the communication between the source and the receiver is over a network in which packets can be routed along different paths and can arrive out of order at the receiver. In this case, it may be that the messages are not decoded

in the order of their arrival, depending on the order in which the packets reach the destination. Thus although $Y_i \geq X_i$, it is not necessary that $Y_i > Y_{i-1}$. This was the situation considered in [17], wherein a lower bound to the second term in (4.4) was obtained. On the other hand, if the communication is over a a point-to-point channel or a virtual circuit network in which the messages are decoded at the destination at the same order they arrived at the source, then

$$Y_i \geq X_i, \text{ and } Y_i \geq Y_{i-1}, \quad i = 1, \cdots, n, \tag{4.5}$$

setting $Y_0 = 0$ by convention. This is the situation that we will be primarily interested in. In this case, the sequence of decoding times $Y_1, \cdots, Y_n$ can be considered as the departure process of a single-server queue whose arrival process is $X_1, \cdots, X_n$. Viewed this way, there is a "service time" associated with each message, which is the time that message would spend in a single server queue whose arrival process is $X_1, \cdots, X_n$ and from which messages depart at times $Y_1, \cdots, Y_n$. More precisely, the service time is given by

$$S_i = Y_i - \max(X_i, Y_{i-1}), \quad i = 1, \cdots, n, \tag{4.6}$$

with $Y_0 = 0$. Note that until time $\max(X_i, Y_{i-1})$, either message i has not arrived at the source, or the previous message $i-1$ is still in transmission. At time Y_i, the decoding decision about message i is made, and no further information about message i can affect this decision. Thus the service time S_i defined by (4.6) is precisely the amount of time available to the transmitter to transmit useful information about message i to the receiver. The probability of error of message i depends directly on the amount of information transmitted within this time, and thus in turn on the amount of such time available.

Another interpretation of the second term on the right side of (4.4) arises from the idea that the sequence of message decoding times $Y_1, \cdots, Y_n$ can be considered as a reproduction of the sequence of message arrival times $X_1, \cdots, X_n$. Put another way,

the sequence of times $D_1, \cdots, D_n$ between successive message decodings can be considered as a reproduction of the sequence of message interarrival times $A_1, \cdots, A_n$. This reproduction is obviously not exact, and we can define a suitable measure of distortion between the two sequences in terms of the service times $S_1, \cdots, S_n$ defined by (4.6). The total service time $\sum_{i=1}^{n} S_i$ can be considered as a measure of distortion between the sequence of message arrival times $X_1, \cdots, X_n$ and its reproduction $Y_1, \cdots, Y_n$, or equivalently, between $A_1, \cdots, A_n$ and $D_1, \cdots, D_n$. Since the available service time S_i of message i influences the probability of error of message i, the average service time $(1/n) \sum_{i=1}^{n} S_i$ can be considered a measure of the average probability of error per message. In this sense, this service time distortion measure does have a physical significance.

The rate distortion function $R(d)$ of the message arrival process under this distortion measure is the minimum amount of information required to generate a reproduction of the sequence of arrival times such that the distortion is no more than d. Informally, suppose that we want to encode $A^n = (A_1, \cdots, A_n)$ using nR bits, so that the rate of the source encoding is R bits per message. We want this encoding to be capable of creating a reproduction $D^n = (D_1, \cdots, D_n)$ of A^n that satisfies (4.5), and such that the probability that the distortion between A^n and D^n being more than n times a given number d vanishes with n. The rate distortion function $R(d)$ of the source with this distortion measure is the smallest source encoding rate at which such a reproduction is possible.

Alternatively, we could require that the *expected* distortion per message be no more than a given number d, i.e. $(1/n) \sum_{i=1}^{n} E[S_i] \leq d$. However, note that there is no limit on the values of the service times, so that the distortion measure is not a bounded distortion measure. For non-bounded distortion measures, there are technical difficulties with using the limiting expected distortion instead of the probability of the distortion being more than a given value. We will use the latter interpretation.

It is well known [5] that the rate distortion function $R(d)$ satisfies

$$R(d) \leq (1/n)I(A_1, \cdots, A_n; D_1, \cdots, D_n)$$

for any joint distribution on (A^n, D^n) that satisfies the constraints of the encoding, in this case (4.5). Thus the rate distortion function $R(d)$ for the service time distortion measure is a lower bound to the second term in the right side of (4.4).

In the following, we will formally define the rate distortion function and some related concepts, and evaluate the rate distortion function for the special case in which the message arrival process is a Poisson process.

4.2.1 Definitions

In this section we formally define the service time distortion measure, source codes whose fidelity is measured according to this distortion measure, achievable source coding rates, and the rate distortion function under the service time distortion measure. For simplicity of notation, we will denote n-sequences by superscripts and infinite sequences by boldface letters, e.g. $a^n = (a_1, \cdots, a_n)$ and $\mathbf{a} = (a_1, a_2, \cdots)$. As usual, random variables will be denoted by capital letters, so that $\mathbf{A}$ stands for the infinite sequence of random variables $(A_1, A_2, \cdots)$ or a discrete random process, while A^n represents the n-tuple of random variables $(A_1, \cdots, A_n)$.

Definition 4.1: For given n-sequences $a^n = (a_1, a_2, \cdots, a_n)$ and $d^n = (d_1, d_2, \cdots, d^n)$ such that $a_i, d_i \geq 0 \ \forall i = 1, 2, \cdots, n$, let $x_i = \sum_{j=1}^{i} a_j$, $y_i = \sum_{j=1}^{i} d_j$, and $y_0 = 0$. The service time distortion ρ_n between $a^n = (a_1, \cdots, a_n)$ and $d^n = (d_1, \cdots, d_n)$ is defined as

$$\rho_n(a^n; d^n) = \begin{cases} \sum_{i=1}^{n} [y_i - \max(y_{i-1}, x_i)] \text{ if } y_i \geq \max(y_{i-1}, x_i) \ \forall i, \\ \infty \quad \text{otherwise.} \end{cases} \tag{4.7}$$

The service time distortion measure $\rho_n(\cdot\,;\cdot)$ is not a metric, as it is obviously not symmetric. However, it can be verified that $\rho_n(\cdot\,;\cdot)$ does satisfy the triangle

inequality. The interpretation of the triangle inequality for $\rho_n(\cdot\,;\cdot)$ is interesting. It asserts that the average service time for a tandem of two single-server queues is no more than the sum of the individual average service times. Conventional measures of distortion analyzed in rate-distortion theory are usually *single letter* distortion measures, in which the distortion between a^n and d^n is of the form $\sum_{i=1}^n D(a_i, d_i)$, where $D(a_i, d_i)$ depends only on a_i and d_i. The service time distortion is of the form $\rho_n(a^n, d^n) = \sum_{i=1}^n s_i$, where s_i depends not only on a_i and d_i but also on $a_1, \cdots, a_{i-1}$ and $d_1, \cdots, d_{i-1}$ as well. Thus the service time distortion measure is not a single letter distortion measure.

Definition 4.2: An (n, M, d, ϵ) distortion code for a discrete random process $\mathbf{A} = (A_1, A_2, \cdots)$ consists of a source encoder map

$$f : (\mathbb{R}^+)^n \to \{1, 2, \cdots, M\}$$

and a decoder map

$$g : \{1, 2, \cdots, M\} \to (\mathbb{R}^+)^n$$

such that the probability of the distortion being greater than nd is less than ϵ, i.e.

$$P_{A^n}\left(a^n : \frac{1}{n}\rho_n(a^n; g[f(a^n)]) > d\right) < \epsilon.$$

Definition 4.3: R is an *ϵ-achievable source coding rate at distortion d for a process* $\mathbf{A}$ iff, for any $\gamma > 0$, there exist $(n, \exp(n(R+\gamma)), d, \epsilon)$ distortion codes for $\mathbf{A}$ for all sufficiently large n. R is an *achievable source coding rate at distortion d for* $\mathbf{A}$ if R is ϵ-achievable for all $\epsilon > 0$. The infimum of all achievable rates for $\mathbf{A}$ is denoted by $R_{\mathbf{A}}(d)$.

Definition 4.4: For processes $\mathbf{A}$ and $\mathbf{D}$ with distributions $P_{\mathbf{A}}$ and $P_{\mathbf{D}}$ respectively, let $\mathcal{P}(P_{\mathbf{A}}, P_{\mathbf{D}})$ be the class of all joint distributions $Q_{\mathbf{A},\mathbf{D}}$ on $(\mathbf{A}, \mathbf{D})$. For any such joint distribution $Q_{\mathbf{A},\mathbf{D}}$, the *limsup in probability* of $(1/n)\rho_n(A^n; D^n)$ is defined as

$$\rho(Q_{\mathbf{A},\mathbf{D}}) = \inf\left\{h : \lim_{n\to\infty} Q_{A^n,D^n}\left((a^n, d^n) : \frac{1}{n}\rho_n(a^n; d^n) > h\right) = 0\right\}.$$

The ρ_s *distortion* between $P_{\mathbf{A}}$ and $P_{\mathbf{D}}$ is defined as

$$\rho_s(P_{\mathbf{A}}, P_{\mathbf{D}}) = \inf_{Q_{\mathbf{A},\mathbf{D}} \in \mathcal{P}(P_{\mathbf{A}}, P_{\mathbf{D}})} \rho(Q_{\mathbf{A},\mathbf{D}})$$

In Sections 2.4 and 3.5, we used the notion of the *information density* for a joint density Q_{A^n,D^n} with marginals P_{A^n} and P_{D^n}, defined as the function

$$i_{A^n,D^n}(a^n; d^n) = \log \frac{Q_{A^n,D^n}(a^n, d^n)}{P_{A^n}(a^n) P_{D^n}(d^n)}.$$

Definition 4.5: The *sup-information rate* $\overline{I}(\mathbf{A}; \mathbf{D})$ of a joint process $(\mathbf{A}, \mathbf{D})$ with joint distribution $Q_{\mathbf{A},\mathbf{D}}$ is defined as the limsup in probability of the sequence of random variables $(1/n) i_{A^n,D^n}(A^n; D^n)$, i.e.

$$\overline{I}(\mathbf{A}; \mathbf{D}) = \inf \left\{ h : \lim_{n \to \infty} Q_{A^n,D^n} \left((a^n, d^n) : \frac{1}{n} i_{A^n,D^n}(a^n; d^n) > h \right) = 0 \right\}.$$

The *inf-information rate* $\underline{I}(\mathbf{A}; \mathbf{D})$ of a joint process $(\mathbf{A}, \mathbf{D})$ with joint distribution $Q_{\mathbf{A},\mathbf{D}}$ is analogously defined as the liminf in probability of the sequence of random variables $(1/n) i_{A^n,D^n}(A^n; D^n)$, i.e.

$$\underline{I}(\mathbf{A}; \mathbf{D}) = \sup \left\{ h : \lim_{n \to \infty} Q_{A^n,D^n} \left((a^n, d^n) : \frac{1}{n} i_{A^n,D^n}(a^n; d^n) < h \right) = 0 \right\}.$$

The utility of the inf-information rate in establishing the achievability results of Sections 2.4 and 3.5 obtained from the result of [47] that a rate R is achievable over a certain channel only if there exists an input process for which the inf-information rate over the channel is at least R. A similar result was established in [41] about the relation of the sup-information rate to the achievable source coding rates for a given process. It was shown in [41] that the smallest achievable source coding rate $R_{\mathbf{A}}(d)$ for a process $\mathbf{A}$ is equal to the infimum of the sup-information rates $\overline{I}(\mathbf{A}; \mathbf{D})$ of all joint processes $(\mathbf{A}, \mathbf{D})$ for which the probability of a distortion greater than d asymptotically vanishes. The statement of the following theorem is a combination of [41, Thm. 1] and [41, Thm. 10(b)].

Theorem 4.1. *For any process* $\mathbf{A}$ *having distribution* $P_{\mathbf{A}}$ *and any sequence of distortion measures* $\{\rho_n(\cdot,\cdot)\}_{n\geq 1}$,

$$R_{\mathbf{A}}(d) = \inf_{Q_{\mathbf{A},\mathbf{D}} \in \mathcal{T}(d)} \bar{I}(\mathbf{A};\mathbf{D}),$$

where $\mathcal{T}(d)$ *is the class of all joint distributions* $Q_{\mathbf{A},\mathbf{D}}$ *on processes* $(\mathbf{A},\mathbf{D})$ *having* $\mathbf{A}$ *marginal* $P_{\mathbf{A}}$ *such that*

$$\rho(Q_{\mathbf{A},\mathbf{D}}) \leq d.$$

Thus to show for a given process $\mathbf{A}$ that $R_{\mathbf{A}}(d) \leq r$, it suffices to find a process $\mathbf{D}$ and joint distribution $Q_{\mathbf{A},\mathbf{D}}$ on $(\mathbf{A},\mathbf{D})$ satisfying $\rho(Q_{\mathbf{A},\mathbf{D}}) \leq d$ for which $\bar{I}(\mathbf{A};\mathbf{D}) \leq r$. i.e. to show that for that joint distribution $Q_{\mathbf{A},\mathbf{D}}$,

$$\lim_{n\to\infty} Q_{A^n,D^n}\left((a^n,d^n) : \frac{1}{n} i_{A^n,D^n}(a^n;d^n) > r\right) = 0.$$

4.2.2 Achievability of R(d)

We first show the achievability of (4.2) for the rate distortion function of the rate λ Poisson process with the service time distortion measure, by examining the sup-information rate for a particular joint process that is closely related to the canonical $M/M/1$ queue. Recall that the λ-timing capacity of the $\cdot/M/1$ queue was established in [4] to be $\log(\mu/\lambda)$ bits/packet, by showing that the inf-information rate of the $M/M/1$ queue in equilibrium was at least equal to $\log(\mu/\lambda)$. We will show that for $0 < d < 1/\lambda$, the rate distortion function $R(d)$ of the rate λ Poisson process with the service time distortion measure satisfies

$$R(d) \leq \log\frac{1}{\lambda d} \quad \text{bits/message}, \quad 0 < d < \frac{1}{\lambda}. \tag{4.8}$$

We will accomplish this by showing that the sup-information rate of a certain joint process that mimics an $M/M/1$ queue in equilibrium is no more than $\log(1/\lambda d)$.

Let $\mathbf{A}$ be the rate λ Poisson process. By the result of Theorem 4.1, to show that $R_{\mathbf{A}}(d) \leq \log(1/\lambda d)$, it suffices to find a process $\mathbf{D}$ and joint distribution $Q_{\mathbf{A},\mathbf{D}}$ on

$(\mathbf{A}, \mathbf{D})$ satisfying $\rho(Q_{\mathbf{A},\mathbf{D}}) \leq d$, such that

$$\lim_{n\to\infty} Q_{A^n,D^n}\left((a^n, d^n) : \frac{1}{n} i_{A^n,D^n}(a^n; d^n) > \log\frac{1}{\lambda d}\right) = 0. \tag{4.9}$$

We will choose the joint distribution $Q_{\mathbf{A},\mathbf{D}}$ on $(\mathbf{A}, \mathbf{D})$ to mimic the distribution that is induced by treating the Poisson process $\mathbf{A}$ as the arrival process to an $M/M/1$ queue with average service time d in *steady state*, as follows. The reproduction process $\mathbf{D}$ will be the departure process from this $M/M/1$ queue. Let us denote $\mu = 1/d$ for consistency with standard queueing theoretic notation for $M/M/1$ queues. A well-known result for $M/M/1$ queues [6, Section 3.3.1] is that in steady state, the distribution of the number of messages N in the queue is Geo (λ/μ), i.e.

$$P_N(k) = \left(\frac{\lambda}{\mu}\right)^k \left(1 - \frac{\lambda}{\mu}\right), \; k \geq 0.$$

Moreover, by the *PASTA* property [6, Section 3.3.2], since the arrival process to this $M/M/1$ queue is Poisson, the number of packets waiting in a queue at the instant of a new packet arrival is the same as the steady state distribution.

The joint distribution $Q_{\mathbf{A},\mathbf{D}}$ is constructed as follows. Suppose we treat the message arrival process $\mathbf{A}$ as the arrival process to an $M/M/1$ queue with service rate $\mu = 1/d > \lambda$, and suppose that the first message, arriving at the queue at time A_1, finds the queue in steady state. Let $\mathbf{D}$ be the departure process from this queue beginning with the departure of the first message. Then the time from A_1 up to the instant D_1 of departure of the first packet would be the sum of the service times of the N messages in the queue at the time of the arrival of the first message, plus the service time of the first message. This is the sum of $(N + 1)$ i.i.d. Exp (μ) random variables, where $N \sim$ Geo (λ/μ), and thus $D_1 - A_1$ would have the Exp $(\mu - \lambda)$ distribution. It is also easy to show that at the instant of departure of the first message, the number left behind in the queue again has the same Geo (λ/μ) distribution as the steady state number. To mimic this behavior with the joint distribution $Q_{\mathbf{A},\mathbf{D}}$, we choose the distribution of the service time $D_1 - A_1$ of the first message to be Exp $(\mu - \lambda)$

independent of A_1, and choose the distributions of the service times of all successive messages arriving after the first one to be i.i.d. Exp (μ) independent of all other arrival and departure times. Thus by choosing the distribution of the departure time of the first message appropriately, we can ensure that the queue behaves as if it remains in steady state after the departure of the first message. Note that since D_1 is the sum of two independent random variables A_1 and $D_1 - A_1$ having the Exp (λ) and Exp $(\mu - \lambda)$ distributions respectively, its density is given by [33. Ex. 6-7. p. 137]

$$P_{D_1}(d) = \begin{cases} \frac{\lambda(\mu-\lambda)}{\mu-2\lambda}\left(e^{-\lambda d} - e^{-(\mu-\lambda)d}\right), d \geq 0, & \mu \neq 2\lambda, \\ \lambda^2 d e^{-\lambda d}, \; d \geq 0 & \mu = 2\lambda. \end{cases} \tag{4.10}$$

The conditional probability density of the departure process **D** given the arrival process **A** can thus be described as follows.

$$P_{D_1|A_1}(d_1|a_1) = (\mu - \lambda)\exp\{-(\mu - \lambda)(d_1 - a_1)\}, \; d_1 \geq a_1, \tag{4.11a}$$

$$P_{D_i|A^i,D^{i-1}}(d_i|a_1, \cdots, a_i, d_1, \cdots, d_{i-1}) = \mu \exp(-\mu s_i), \quad d_i \geq \max(a_i, d_{i-1}), \quad i \geq 2. \tag{4.11b}$$

where $s_i = \sum_{j=1}^{i} d_j - \max(\sum_{j=1}^{i} a_i, \sum_{j=1}^{i-1} d_j)$ is the service time of message i. Since this choice of the conditional distribution of **D** given **A** in (4.11) mimics the behavior of a $M/M/1$ queue in steady state, Burke's theorem [6, Section 3.7] is valid for the departure process of this queue *after* the departure of the first packet. Thus the interdeparture times after the departure of the first packet are i.i.d. Exp (λ) random variables, i.e.

$$P_{D^n}(d_1, \cdots, d_n) = P_{D_1}(d_1) \prod_{i=2}^{n} \left[\lambda e^{-\lambda d_i}\right]. \tag{4.12}$$

Let us examine the sup-information rate of the joint process $(\mathbf{A}, \mathbf{D})$ under this joint distribution.

The information density of $A^n = (A_1, \cdots, A_n)$ and $D^n = (D_1, \cdots, D_n)$ is the random variable

$$\begin{aligned} i_{A^n,D^n}(A^n; D^n) &= \log \frac{Q_{A^n,D^n}(A^n, D^n)}{P_{A^n}(A^n)P_{D^n}(D^n)} \\ &= \log \frac{P_{D^n|A^n}(D_1, \cdots, D_n|A_1, \cdots, A_n)}{P_{D^n}(D_1, \cdots, D_n)}. \end{aligned}$$

For any valid sequences $(A_1, \cdots, A_n)$ of arrival times and $(D_1, \cdots, D_n)$ of departure times, the service times of the n messages are given by (4.6). Using (4.11) and (4.12), we obtain

$$\begin{aligned} \frac{1}{n} i_{A^n,D^n}(A^n; D^n) &= \frac{1}{n} \log \frac{e_{\mu-\lambda}(D_1 - A_1) \prod_{i=2}^n e_\mu(S_i)}{P_{D_1}(D_1) \prod_{i=2}^n e_\lambda(D_i)} \\ &= \frac{n-1}{n} \log \frac{\mu}{\lambda} - \frac{\mu}{n} \sum_{i=2}^n S_i + \frac{\lambda}{n} \sum_{i=2}^n D_i + \frac{1}{n} \log \frac{e_{\mu-\lambda}(D_1 - A_1)}{P_{D_1}(D_1)}. \end{aligned}$$

where $e_x(\cdot)$ is the exponential density with mean $1/x$ as defined in Notation A.6 in Appendix A. From the distribution (4.10) of D_1, the last term goes to 0 as $n \to \infty$, unless $D_1 = 0$ which happens with probability 0. Moreover, by the law of large numbers,

$$\frac{1}{n} \sum_{i=2}^n D_i \to \frac{1}{\lambda} \text{ in probability, and } \frac{1}{n} \sum_{i=2}^n S_i \to \frac{1}{\mu} \text{ in probability.}$$

Thus we see that with this choice of joint distribution $Q_{\mathbf{A},\mathbf{D}}$, (4.9) is satisfied. Thus the sup-information rate of $(\mathbf{A}, \mathbf{D})$ is no more than $\log(\mu/\lambda) = \log(1/\lambda d)$, establishing (4.8).

To show that $R(d) = 0$ for $d > 1/\lambda$, we again resort to the $\cdot/M/1$ queue with service rate $\mu = 1/d$ as the distorting mechanism, except this time we assume that the queue is initially empty. For $\mu < \lambda$, the birth-death Markov chain associated with the Markov process of the queue size of the $M/M/1$ queue is transient [20, Section 5.3]. Let F_{00} be the probability that the system ever becomes empty again after its initial empty state. The transience of the chain means that $F_{00} < 1$, and in fact, from the result of [20, Problem 5.2, p. 181], $F_{00} = 2\mu/(\lambda + \mu)$. Let us denote by M the

number of returns to state 0, so that $P_M(0) = F_{00}$. As before, denote the arrival and departure processes of the queue as **A** and **D** respectively, and their joint distribution as $Q_{\mathbf{A},\mathbf{D}}$. We will show that the sup-information rate of the joint process $(\mathbf{A}, \mathbf{D})$ is no more than 0.

The sup-information rate of $(\mathbf{A}, \mathbf{D})$ is

$$\begin{aligned} i_{A^n;D^n}(A^n; D^n) &= \log \frac{Q_{D^n|A^n}(D^n|A^n)}{Q_{D^n}(D^n)} \\ &= \log \frac{Q_{D^n|A^n}(D^n|A^n)}{Q_{D^n|M=0}(D^n)F_{00} + Q_{D^n|M>0}(D^n)(1-F_{00})} \\ &\leq \log \frac{Q_{D^n|A^n}(D^n|A^n)}{Q_{D^n|M=0}(D^n)F_{00}}. \end{aligned} \tag{4.13}$$

Now note that if $M = 0$, then no message after the first one ever finds the queue empty on arrival. Thus conditional on $M = 0$, the successive departure times $D_2, D_3, \cdots$ are simply the service times of those messages, and are thus i.i.d. with the Exp (μ) distribution. The departure time of the first message D_1, of course, is the sum of its arrival time A_1 and its service time S_1, whose respective distributions are Exp (λ) and Exp (μ). Thus the probability density of the departure time of the first packet is again given by [33, Ex. 6-7, p. 137]

$$Q_{D_1}(d) = \frac{\lambda\mu}{\lambda - \mu}\left(e^{-\mu d} - e^{-\lambda d}\right), \quad d \geq 0. \tag{4.14}$$

The joint density of the departure times given that $M = 0$ is thus

$$Q_{D^n|M=0}(D_1, \cdots, D_n) = Q_{D_1}(D_1)\prod_{i=2}^{n} \mu \exp(-\mu D_i). \tag{4.15}$$

The conditional density $Q_{D^n|A^n}$ of the interdeparture times given the interarrival times is given by

$$Q_{D^n|A^n}(d_1, \cdots, d_n|a_1, \cdots, a_n) = \prod_{i=1}^{n} Q_{D_i|A_1,\cdots,A_n,D_1,\cdots,D_{i-1}}(d_i|a_1, \cdots, a_n, d_1, \cdots, d_{i-1}),$$

where

$$Q_{D_i|A^n,D^{i-1}}(d_i|a_1, \cdots, a_n, d_1, \cdots, d_{i-1} = \mu \exp(-\mu s_i), \quad i = 1, \cdots, n, \tag{4.16}$$

with $s_1, \cdots, s_n$ defined by (4.6). Hence (4.13) becomes

$$\begin{aligned} \frac{1}{n} i_{A^n;D^n}(A^n; D^n) &\leq \frac{1}{n} \log \frac{\prod_{i=1}^n \mu \exp(-\mu S_i)}{Q_{D_1}(D_1) \prod_{i=2}^n \mu \exp(-\mu D_i)} \\ &= \frac{1}{n} \log \mu - \frac{1}{n} \log Q_{D_1}(D_1) + \frac{\mu}{n} \sum_{i=2}^n D_i - \frac{\mu}{n} \sum_{i=1}^n S_i \end{aligned} \quad (4.17)$$

The first two terms clearly converge to 0 in probability. The difference of the last two terms can be written as

$$\frac{\mu}{n} \sum_{i=2}^n D_i - \frac{\mu}{n} \sum_{i=1}^n S_i = \frac{\mu}{n} \sum_{i=2}^n W_i - \frac{\mu}{n} S_1,$$

where $W_i = D_i - S_i$ is the idle time of the server immediately preceding the i^{th} arrival to the queue. Recall our assumption that the queue is initially empty. Let P_0^i be the probability that the queue is again empty immediately preceding the i^{th} arrival. Since the Markov chain associated with the $M/M/1$ queue is transient for $\lambda > \mu$, the empty state of the queue is transient, so by [20, Lemma 5.2, p. 152],

$$\sum_{i=2}^{\infty} P_0^i < \infty.$$

Hence by the first Borel-Cantelli lemma [49, p. 27], the probability that the queue becomes empty infinitely often is zero. That is, with probability 1, all but a finite number of the idle times $W_2, W_3, \cdots$ are zero, so that

$$\lim_{n \to \infty} \frac{1}{n} \sum_{i=2}^n W_i = 0 \quad \text{in probability.}$$

Since $S_1 \sim \text{Exp}(\mu)$, S_1 is also finite with probability 1, so that $S_1/n \to 0$. Thus the right side of (4.17) converges to 0 in probability. Hence

$$\lim_{n \to 0} Q_{A^n, D^n}\left(i_{A^n;D^n}(A^n; D^n) > 0\right) = 0,$$

thus establishing that $R(d) = 0$ for $d > 1/\lambda$.

4.2.3 Converse Result

We now show that the upper bound (4.8) on the rate distortion function $R(d)$ of a rate λ Poisson process under the service time distortion measure with allowed distortion $d < 1/\lambda$ is tight. Consider any joint process $(\mathbf{A}, \mathbf{D})$ with joint distribution $Q_{\mathbf{A},\mathbf{D}}$ whose $\mathbf{A}$ marginal distribution is that of the rate λ Poisson process and which satisfies

$$\rho(Q_{\mathbf{A},\mathbf{D}}) \leq d. \tag{4.18}$$

We will show that for $0 < d < 1/\lambda$, the sup-information rate of any such joint process $(\mathbf{A}, \mathbf{D})$ is at least $\log(1/\lambda d)$, i.e.

$$\bar{I}(\mathbf{A}; \mathbf{D}) \geq \log \frac{1}{\lambda d}. \tag{4.19}$$

First, for any joint process $(\mathbf{A}, \mathbf{D})$ with joint distribution $Q_{\mathbf{A},\mathbf{D}}$ that satisfies (4.18), let the sequence of service times be $S_1, S_2, \cdots$. From Definition 4.4, the sequence of service times satisfies

$$\lim_{n\to\infty} Q\left(\frac{1}{n}\sum_{i=1}^{n} S_i > d\right) = 0. \tag{4.20}$$

From the definitions of the inf-information rate and the sup-information rate, it is clear that the inf-information rate of a given joint process $(\mathbf{A}, \mathbf{D})$ is no more than its sup-information rate. Hence, to show that the sup-information rate of a certain joint process is at least $\log(1/\lambda d)$, it suffices to show that the inf-information rate is at least $\log(1/\lambda d)$. Recall that in the achievability result of Proposition 2.5 on the λ-timing capacity of a discrete-time $\cdot/G/1$ queue, we established a lower bound on the sup-information rate of the joint arrival and departure processes of a discrete-time $\cdot/G/1$ queue driven by a Bernoulli arrival process. This proof was along the same lines as the achievability result for the λ-timing capacity of the continuous time $\cdot/G/1$ queue in [4, Theorem 7]. In the latter proof, it was shown that for a $\cdot/G/1$ queue with service rate μ driven by a rate λ Poisson arrival process with $\lambda < \mu$, the inf-information rate of the arrival and departure processes is at least $\log(\mu/\lambda)$. Thas

is, for $d = 1/\mu < 1\lambda$, the $\cdot/G/1$ queue with i.i.d. service times satisfies (4.19). Thus the result of [4, Theorem 7] establishes (4.19) for all joint processes for which the service times $S_1, \cdots, S_n$ are i.i.d. Of course, the class of processes that satisfies (4.20) is significantly larger, and includes processes in which the service times are dependent on each other and also on the arrival times.

Consider any joint process $(\mathbf{A}, \mathbf{D})$ with joint distribution $Q_{\mathbf{A},\mathbf{D}}$ whose $\mathbf{A}$ marginal distribution is a rate λ Poisson process and which satisfies (4.20). We will continue to denote the conditional distributions defined by (4.11) as $P_{D^n|A^n}$, with $\mu = 1/d$. These conditional distributions $P_{D^n|A^n}$, together with the marginal distribution for $\mathbf{A}$, define a joint distribution $P_{\mathbf{A},\mathbf{D}}$. As in the proof of Proposition 2.5 and [4, Theorem 7],

$$\begin{aligned}\frac{1}{n} i_{A^n;D^n}(A^n; D^n) &= \frac{1}{n} \log \frac{Q_{D^n|A^n}(D^n|A^n)}{Q_{D^n}(D^n)} \\ &= \frac{1}{n} \log \frac{Q_{A^n|D^n}(A^n|D^n)}{P_{A^n|D^n}(A^n|D^n)} + \frac{1}{n} \log \frac{P_{D^n|A^n}(D^n|A^n)}{P_{D^n}(D^n)}.\end{aligned}$$

As in those proofs, the liminf in probability of the first term is non-negative. For the second term, using (4.11) and (4.12),

$$\begin{aligned}&\frac{1}{n} \log \frac{P_{D^n|A^n}(D^n|A^n)}{P_{D^n}(D^n)} \\ &= \frac{n-1}{n} \log \frac{\mu}{\lambda} - \frac{\mu}{n} \sum_{i=2}^{n} S_i + \frac{\lambda}{n} \sum_{i=2}^{n} D_i + \frac{1}{n} \log \frac{e_{\mu-\lambda}(D_1 - A_1)}{P_{D_1}(D_1)}.\end{aligned}$$

The first term converges to $\log(\mu/\lambda)$, and the last term converges to 0 as $n \to \infty$. Consider the middle two terms,

$$\frac{\lambda}{n} \sum_{i=2}^{n} D_i - \frac{\mu}{n} \sum_{i=2}^{n} S_i. \tag{4.21}$$

In the proof of Proposition 2.5, we argued that $(1/n) \sum_{i=1}^{n} D_i$ converges in probability to $1/\lambda$, because the service times of the $\cdot/G/1$ queue are i.i.d. However, in this case, the assumption that the sequence of service times satisfies (4.20) may not be enough to assert the convergence of $(1/n) \sum_{i=2}^{n} D_i$. Note, however, that since $D_1 = A_1 + S_1$,

$$\sum_{i=1}^{n} D_i \geq \sum_{i=1}^{n} A_i \Longrightarrow \sum_{i=2}^{n} D_i \geq \sum_{i=2}^{n} A_i - S_1,$$

so that

$$\begin{aligned}\frac{\lambda}{n}\sum_{i=2}^{n} D_i - \frac{\mu}{n}\sum_{i=2}^{n} S_i &\geq \frac{\lambda}{n}\left[\sum_{i=2}^{n} A_i - S_1\right] - \frac{\mu}{n}\sum_{i=2}^{n} S_i \\ &= \frac{\lambda}{n}\sum_{i=2}^{n} A_i - \frac{\lambda}{n} S_1 - \frac{\mu}{n}\sum_{i=2}^{n} S_i \\ &\geq \frac{\lambda}{n}\sum_{i=2}^{n} A_i - \frac{\mu}{n}\sum_{i=1}^{n} S_i,\end{aligned}$$

where in the last step we used the fact that $\lambda < \mu$. Now $(\lambda/n)\sum_{i=2}^{n} A_i$ converges in probability to 1, while (4.20) asserts that the liminf in probability of $-\frac{\mu}{n}\sum_{i=1}^{n} S_i$ is at least -1. Thus the liminf in probability of (4.21) is non-negative. Hence

$$\lim_{n\to\infty} Q_{A^n,D^n}\left(i_{A^n;D^n}(A^n;D^n) < \log\frac{\mu}{\lambda}\right) = 0.$$

Thus the inf-information rate of any joint process $Q_{\mathbf{A},\mathbf{D}}$ satisfying (4.18) is at least $\log(\mu/\lambda)$, establishing (4.19).

4.3 Discussion

In the derivations of the achievability result of Section 4.2.2 and the converse result of Section 4.2.3, the joint distribution $P_{\mathbf{A},\mathbf{D}}$ that played a special role was chosen to mimic the $M/M/1$ queue *in equilibrium*, by choosing the service time distribution of the first packet appropriately in (4.11a). It is thus clear that the encoding that achieves the rate distortion function of the Poisson process under the service time distortion measure is the $M/M/1$ queue in equilibrium.

In [45], the rate distortion function of the Poisson process was analyzed for the following distortion measure . Let $\mathbf{A} = (A_1, A_2, \cdots)$ be the interarrival times between successive events of a Poisson process. $\{A_i\}$ are i.i.d. with the Exp (λ) distribution. Let $\hat{A}_1, \hat{A}_2, \cdots$ be reproductions of $A_1, A_2, \cdots$, and consider the following distortion

measure.

$$\rho(A_i, \hat{A}_i) = \begin{cases} \hat{A}_i - A_i, & \text{if} \quad \hat{A}_i \leq A_i + d \\ \infty & \text{otherwise.} \end{cases} \tag{4.22}$$

This distortion measure essentially enforces that the reproduction $\hat{A}_i$ be no greater than $A_i + d$. The sequence $\hat{A}_1, \hat{A}_2, \cdot$ is not necessarily a causal reproduction of the sequence $A_1, A_2, \cdots$. That is, if $X_i = \sum_{j=1}^{i} A_j$ and $\hat{X}_i = \sum_{j=1}^{i} \hat{A}_j$, it is not required that $\hat{X}_i \geq X_i$. The rate distortion function of the rate λ Poisson process under this distortion measure, obtained in [45], is in fact identical to the rate distortion function under the service time distortion that we obtained in this chapter. This appears to be coincidental, as there is no similarity between the distortion measures. However, there is a curious connection between the optimal encodings for the two distortion measures. For the distortion measure of (4.22), it was shown in [45] that the transformation that attains $R(d)$ for $0 < d < 1/\lambda$ is equivalent to the representation

$$A_i = \hat{A}_i - d + N, \quad \text{where} \quad N \sim \text{Exp}\,(1/d)\,.$$

In this representation, setting $W = \hat{A}_i - d$, the rate distortion function is equal to $I(W; W + N)$. This is the same type of transformation that attains the λ-timing capacity of the $\cdot/M/1$ queue, and in the discrete-time case, of the $\cdot/Geo^{(s)}/1$ queue. Our result thus exposes an unexplored connection between two seemingly different encodings of the Poisson process.

We can also obtain discrete-time analogues of the results of this chapter. Consider a slotted time system in which messages arrive at the source and are decoded by the receiver at integer-valued epochs called slots. A service time distortion can be defined for discrete-time message arrival and decoding processes as well. The discrete-time analogue of the Poisson process is the Bernoulli process, and results analogous to those of Section 4.2 can be obtained for the rate distortion function of the Bernoulli process under the service time distortion measure. In particular, we can show that

for a rate λ Bernoulli process with the service time distortion,

$$R(d) = \begin{cases} \frac{1}{\lambda}h(\lambda) - dh\left(\frac{1}{d}\right), & d < \frac{1}{\lambda} \\ 0, & d > \frac{1}{\lambda}. \end{cases} \tag{4.23}$$

The optimal source encoding that achieves this rate distortion function is the discrete-time $\cdot/Geo^{(s)}/1$ queue in equilibrium.

We can use these results to analyze the efficiency of using an idle symbol on point-to-point channels to indicate periods in which the transmitter has no message to transmit. Suppose the source and receiver are communicating over a discrete memoryless K-ary erasure channel with an idle symbol and immediate feedback output about each transmission. An example of this is the stop-and-wait ARQ protocol used for error correction on the data link layer of point-to-point links. In stop-and-wait ARQ, for each message, the transmitter transmits one of the K non-idle symbols repeatedly until it is received error-free. When all waiting messages have finished transmission, the transmitter transmits the idle symbol until a new message arrives at the source. The use of the idle symbol in this manner exactly identifies the source idle periods for the receiver. So the amount of information about the message arrival times that the receiver gets per message can be shown to be $H(W)$, where W represents the idle time of the transmitter immediately following the completion of each message. This is greater than (4.3), and the reason is that the receiver gets information about the presence of a new message at a transmitter as soon as it possibly can. A more efficient strategy would be to mix up the use of the idle symbol for representing the contents of the message and mask the idle periods.

Carrying this one step further, consider the multiple access collision channel. On this channel, the strategy employed by each transmitter is essentially the same as the stop-and-wait ARQ protocol, except that errors are not caused due to random noise, but rather due to collisions resulting from simultaneous transmissions by multiple users. On multiple access channels, apart from informing the receiver of the presence

of a message, the transmitter must also inform the receiver of its identity, or *address*, to enable the receiver to distinguish the transmissions of the various users. This is typically done by devoting a few bits out of each K-ary packet to identify the address of the transmitter. We noted the work [30] in Section 1.4, in which Massey and Mathys demonstrated a construction of a set of accessing sequences for various users that allowed the receiver to correctly identify the transmitter of each successfully received packet. An interesting direction of work would be to examine how much information about message arrival times must be transmitted using a service time distortion criterion for each sender, and how much of this information can actually be transmitted through an intelligent choice of channel accessing schemes.

Chapter 5

CONCLUDING REMARKS

We conclude with a summary of the dissertation, an outline of possible extensions and directions for future work, and an interesting conjecture arising out of the problems considered in this dissertation.

5.1 Summary

Two major aspects of timing have been identified in the literature as central to laying an information-theoretic foundation for data networks. One is the fact that information can be encoded in the times at which packets are transmitted over the network [4]. This timing information is distorted due to the random delays experienced by the packets in traversing the network, thus limiting the amount of information that can be reliably encoded in the timing of the packets. The second aspect is that the fact that messages arrive intermittently at the source at random times requires the transmission of a certain amount of information about the message arrival times in order to inform the receiver of the presence of a message at the source [17]. A portion of the network's capacity must be devoted to transmitting this information, thus reducing the available capacity for transmitting the information about the contents of the messages.

High-speed network systems like ATM which use fixed-size packets or cells are well-modeled as discrete-time systems because the packet transmission time provides a convenient unit of time in terms of which the operation of the network can be described. From the point of view of a single source and receiver communicating over

a virtual-circuit ATM network, the transport of packets from the source to the receiver can be modeled as a discrete-time first-in first-out (FIFO) queue. In Chapter 2, we analyzed the timing capacity of discrete-time queues in which at most one packet can arrive and at most one packet can complete service in a slot. Analogous to the results of [4] about continuous-time queues with exponentially distributed service times, we showed that among all discrete-time single-arrival single-service queues with i.i.d. service times with a given mean service time, the $\cdot/Geo^{(s)}/1$ queue, i.e. the queue with geometrically distributed service times, has the least timing capacity, and we obtained an explicit formula this capacity. This result provides a closed form expression for the capacity of a certain binary channel with infinite memory, and thus is of interest in its own right in information theory.

In Chapter 3, we analyzed the timing capacity of a particular model of discrete-time queues in which multiple packets can arrive or depart in a slot. This model can be considered as an approximation of a leaky bucket scheme in which the number of tokens available in each slot is an i.i.d. sequence. We obtained an explicit expression for the timing capacity of the queue in which a geometrically distributed number of packets can be served in a slot. For queues with an arbitrary service distribution, we established a lower bound to the λ-timing capacity based on a new queueing theoretic property of such queues. We also considered the timing capacity of discrete-time queues with multiple servers, each of which can serve at most one packet per slot with i.i.d. packet service times. For the $\cdot/Geo^{(s)}/K$ queueing system, we obtained an upper bound to the λ-timing capacity in terms of the capacity of the K-output binomial channel with an input constraint. This upper bound can be evaluated numerically for finite K. For the case $K = \infty$, we established the existence of a distribution that achieves the capacity of the infinite-output binomial channel subject to an input constraint, and obtained a necessary and sufficient condition to characterize this distribution using a generalized form of the Kuhn-Tucker theorem.

In Chapter 4, we addressed the problem of quantifying the amount of information

about message arrival times that must perforce be transmitted to a receiver if messages arrive intermittently at random times at the source. We considered the situation in which messages are transmitted one at a time in order and are decoded at the receiver in the same order they arrived at the source, for instance a virtual circuit or a point-to-point channel. We argued that in this situation, the message decoding times can be considered as the departure process of a first-in first-out queue whose arrival process is the message arrival process. With this point of view, one can associate a service time with each message, which is in fact the amount of time that is available for transmitting useful information about that message and thus directly influences the achievable probability of error. The total service time can be considered as a distortion measure between the sequence of message arrival times and message decoding times. The rate-distortion function of the message arrival process with this service time distortion measure is a lower bound to the amount of information about message arrival times that the receiver must receive in order to decode the messages in order. We obtained an explicit formula for the rate-distortion function for a Poisson message arrival process under the service time distortion, and established the role of the single-server exponential queue as the optimal source encoding mechanism for this service time distortion.

5.2 Future Goals

The results of the analysis of timing capacity presented in this dissertation have shown that the use of timing provides a small but important amount of information over a packet channel with timing distortion. The use of timing for conveying information is significant in the context of covert channels, as analyzed in [32, 31], and the work of the timing capacity of discrete-time queueing systems in this dissertation can be considered an important step in that direction. In the broader context of data networks, the use of timing may not provide large gains in capacity. However, we

believe that timing information has a significant role to play in conveying various types of overhead information, such as addressing and sequence numbering. In fact, timing information is closely related to a fundamental understanding of the required amount of such information. In the following, we elaborate on some of these ideas and examine the work involved in exploring these connections.

There is an explicit connection between timing and addressing in time division multiplexed (TDM) systems. In such systems, a periodic sequence of time slots is allotted to each of the source-destination pairs accessing a common channel, and each source can transmit only within its allotted slots. With such a pre-assigned allocation of time slots, the time of transmission immediately identifies the source and destination, and no additional addressing information is required. In the case of a statistical multiplexer, however, each source can send packets to a common buffer at any time, and each packet carries addressing bits to identify its source. In between these two extremes would be intermediate schemes in which each source sends packets according to some convention (i.e. a timing code). That is, unlike TDM systems, there is no pre-determined order of transmission, nor is the transmission of packets totally asynchronous as in a pure statistical multiplexer. Rather, there is a certain number of access sequences, or timing codewords, allotted to each source. The amount of addressing information that is required could then be reduced by letting the timing carry part or all of the addressing overhead. In TDM systems and in systems in which each packet carries the address of the source, the destination can determine the identity of the source of each packet as soon the packet is received. With the use of a timing code, however, the destination would wait until a certain sufficiently large number of packets were received, before making a joint decision on the messages transmitted by all the sources. An interesting direction of work would be to characterize the reduction in addressing overhead by the use of suitable timing information. A good starting point would be the analysis of a system of two sources that transmit packets to a common buffer from which the packets are served in order

with i.i.d. service times. A related idea is the work of Massey and Mathys [30] on the use of timing information to convey addressing in the context of the multiple access collision channel, which we discussed in Section 1.4. The work of [30] demonstrates a method of constructing a particular set of timing codes for the the collision channel that obviates the need for any addressing information. However, the broader question of the timing capacity region of the collision channel remains unexplored.

Another form of overhead information that can potentially be reduced by the use of timing information is sequence numbering. In networks with datagram routing, packets may follow different paths and may arrive at the destination out of order. Each packet usually carries a sequence number to help the destination to re-order the packets, and the number of bits required to indicate the sequence number can be quite large. It would be interesting to characterize the achievable reduction in the amount of sequence numbering information by appropriate use of timing information. Again, if a timing code is used, the receiver would wait until a sufficiently large number of packets have arrived before making a decision on re-ordering the packets based on the times at which they reach the destination.

In the analysis of the timing capacity of various queueing systems, we discovered certain underlying structures that correspond to optimal ways of encoding information in the timing of packets. It would be interesting to be construct concrete coding schemes that make use of these optimal structures. For constructing timing codes on multiple-user systems such as statistical multiplexers, an important feature to have in a coding scheme would be scalability, i.e. the total achievable rate should scale linearly with the number of users.

Another interesting question concerns the situation in which a source receives feedback about the times at which packets reach the destination. Protocols like TCP use such feedback to help sources estimate the congestion in the network. The feedback is in the form of acknowledgment packets, which also experience random timing delays in traversing the network. It would be useful to explore the possible

benefits of using appropriate timing codes to convey feedback information in the presence of timing noise on the feedback channel. The construction of such timing codes for providing feedback has direct implications for congestion control protocols. The scalability of such codes with a large number of users is also an important issue.

As far as the larger question of an information theoretic foundation for communication networks is concerned, the holy grail of a union between the fields of information theory and communication networks is still far from our grasp. We hope that the analysis of timing information in data networks presented in this dissertation has built a foundation and exposed new relations between these fields that will provide a stimulus for further exploration.

BIBLIOGRAPHY

[1] I. C. Abou-Faycal, M. D. Trott, and S. Shamai (Shitz), "The capacity of discrete-time memoryless Rayleigh fading channels," preprint, Jan. 1998, see also "The capacity of discrete-time Rayleigh fading channels," in *Proc. IEEE Int. Symp. Inform. Theory*, Ulm, Germany, July 1997, p. 473.

[2] N. Abramson, "The ALOHA system – another alternative for computer communications," in *AFIPS Conf. Proc.*, 1970, vol. 37, pp. 281–285.

[3] V. Anantharam, "The stability region of the finite-user slotted ALOHA protocol," *IEEE Trans. Inform. Theory*, vol. 37, no. 3, pp. 535–540, May 1991.

[4] V. Anantharam and S. Verdu, "Bits through queues," *IEEE Trans. Inform. Theory*, vol. 42, no. 1, pp. 4–18, Jan. 1996.

[5] T. Berger, *Rate Distortion Theory: A Mathematical Basis for Data Compression*, Englewood Cliffs, NJ: Prentice-Hall, 1971.

[6] D. Bertsekas and R. Gallager, *Data Networks*, Englewood Cliffs, NJ: Prentice-Hall, 2nd edition, 1992.

[7] H. Bruneel and B. G. Kim, *Discrete-Time Models for Communication Systems Including ATM*, Boston, MA: Kluwer, 1993.

[8] R. J. Camrass and R. G. Gallager, "Encoding message lengths for data transmission," *IEEE Trans. Inform. Theory*, vol. IT-24, no. 4, pp. 495–496, July 1978.

[9] J. I. Capetanakis, "Tree algorithms for packet broadcast channels," *IEEE Trans. Inform. Theory*, vol. IT-25, no. 5, pp. 505–515, Sept. 1979.

[10] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, New York: Wiley, 1991.

[11] H. Cramer, *Random Variables and Probability Distributions*, Cambridge Tracts in Mathematics, No. 36. London, UK: Cambridge University Press, 3rd edition, 1970.

[12] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, New York: Academic Press, 1981.

[13] J. N. Daigle, *Queueing Theory for Telecommunications*, Reading, MA: Addison-Wesley, 1992.

[14] A. Ephrimedes and B. Hajek, "Information theory and communication networks: An unconsummated union," *IEEE Trans. Inform. Theory*, vol. 44, no. 6, pp. 2416–2434, Oct. 1998.

[15] W. Feller, *An Introduction to Probability Theory and its Applications*, vol. II, New York: Wiley, 1966.

[16] R. G. Gallager, *Information Theory and Reliable Communication*, New York: Wiley, 1968.

[17] R. G. Gallager, "Basic limits on protocol information in data communication networks," *IEEE Trans. Inform. Theory*, vol. IT-22, no. 4, pp. 385–398, July 1976.

[18] R. G. Gallager, "Applications of information theory to data communication networks," in *New Concepts in Multi-user Communication*, J. Skwirzynski, Ed., number 43 in E, pp. 63–82. Alphen aan den Rijn, The Netherlands: NATO Advanced Study Institutes (Sijthoff and Noordhoff), 1981.

[19] R. G. Gallager, "A perspective on multiaccess channels," *IEEE Trans. Inform. Theory*, vol. 31, no. 2, pp. 124–142, Mar. 1985.

[20] R. G. Gallager, *Discrete Stochastic Processes*, Boston, MA: Kluwer, 1996.

[21] R. G. Gallager and D. C. VanVoorhis, "Optimal source codes for geometrically distributed integer alphabets," *IEEE Trans. Inform. Theory*, vol. IT-21, no. 2, pp. 228–230, Mar. 1975.

[22] A. A. El Gamal and T. M. Cover, "Multiple user information theory," in *Proc. IEEE*, Dec. 1980, vol. 68, pp. 1466–1483.

[23] R. L. Graham, D. E. Knuth, and O. Patashnik, *Concrete Mathematics*, Reading, MA: Addison-Wesley, 2nd edition, 1994.

[24] P. A. Humblet, "Data communication networks and information theory," in *Proc. IEEE Nat. Telecommun. Conf.*, Dec. 1980, pp. 20.3/1–20.3/5.

[25] L. Kleinrock, *Queueing Systems, Volume I: Theory*, New York: Wiley, 1975.

[26] L. Kleinrock, *Queueing Systems, Volume II: Computer Applications*, New York: Wiley, 1976.

[27] H. Kobayashi and A. G. Konheim, "Queueing models for computer communications system analysis," *IEEE Trans. Commun.*, vol. COM-25, no. 1, pp. 2–28, Jan. 1977.

[28] M. Loève, *Probability Theory*, Princeton, NJ: Van Nostrand, 2nd edition, 1963.

[29] D. G. Luenberger, *Optimization by Vector Space Methods*, New York: Wiley, 1969.

[30] J. L. Massey and P. Mathys, "The collision channel without feedback," *IEEE Trans. Inform. Theory*, vol. IT-31, no. 2, pp. 192–204, Mar. 1985.

[31] I. S. Moskowitz, S. J. Greenwald, and M. H. Kang, "An analysis of the timed z-channel," *IEEE Trans. Inform. Theory*, vol. 44, no. 7, Nov. 1998.

[32] I. S. Moskowitz and A. R. Miller, "The channel capacity of a certain noisy timing channel," *IEEE Trans. Inform. Theory*, vol. 38, no. 4, July 1992.

[33] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, New York: McGraw-Hill, 1991.

[34] K. R. Parthasarathy, *Introduction to Probability and Measure*, New York: Springer-Verlag, 1977.

[35] R. Rao and A. Ephremides, "On the stability of interacting queues in a multiple-access system," *IEEE Trans. Inform. Theory*, vol. 34, no. 5, pp. 918–930, Sept. 1988.

[36] I. Rubin, "Information rates for Poisson sequences," *IEEE Trans. Inform. Theory*, vol. IT-19, no. 3, pp. 283–294, May 1973.

[37] I. Rubin, "Information rates and data-compression schemes for Poisson processes," *IEEE Trans. Inform. Theory*, vol. IT-20, no. 2, pp. 200–210, Mar. 1974.

[38] H. Sato and T. Kawabata, "Information rates for Poisson point processes," *Trans. IEICE*, vol. E70, no. 9, pp. 817–822, Sept. 1987.

[39] C. E. Shannon, "Two-way communication channels," in *Proc. 4th Berkeley Symp. Math. Stat. and Prob.*, 1961, vol. 1, pp. 611–644, Reprinted in *Key Papers in the Development of Information Theory*, D. Slepian, Ed., New York: IEEE Press, 1974, pp. 339-372.

[40] J. G. Smith, "The information capacity of amplitude and variance-constrained scalar Gaussian channels," *Inform. Contr.*, vol. 18, pp. 203–219, 1971.

[41] Y. Steinberg and S. Verdu, "Simulation of random processes and rate-distortion theory," *IEEE Trans. Inform. Theory*, vol. 42, no. 1, pp. 63–86, Jan. 1996.

[42] H. Takagi, *Queueing Analysis: A Foundation of Performance Evaluation, Vol. 3: Discrete-Time Systems*, Amsterdam: North Holland, 1993.

[43] B. S. Tsybakov and W. Mikhailov, "Ergodicity of slotted ALOHA systems," *Probl. Inform. Transmission*, vol. 15, no. 4, pp. 301–311, Oct.–Dec. 1979.

[44] E. C. van der Meulen, "A survey of multi-way channels in information theory," *IEEE Trans. Inform. Theory*, vol. IT-23, no. 1, pp. 1–37, 1977.

[45] S. Verdu, "The exponential distribution in information theory," *Probl. Inform. Transmission*, vol. 32, no. 1, pp. 86–95, Jan.–Mar. 1996.

[46] S. Verdu, "Fifty years of Shannon theory," *IEEE Trans. Inform. Theory*, vol. 44, no. 6, pp. 2057–2078, Oct. 1998.

[47] S. Verdu and T. S. Han, "A general formula for channel capacity," *IEEE Trans. Inform. Theory*, vol. 40, no. 4, pp. 1147–1157, July 1994.

[48] J. Walrand, *An Introduction to Queueing Networks*, Englewood Cliffs, NJ: Prentice-Hall, 1988.

[49] D. Williams, *Probability with Martingales*, Cambridge, UK: Cambridge University Press, 1991.

[50] M. E. Woodward, *Communication and Computer Networks: Modelling with Discrete-Time Queues*, Los Alamitos, CA: IEEE Computer Society Press, 1994.

Appendix A

NOTATION

In this appendix we state the notation used in this dissertation for certain common probability distributions. We use the notation $X \sim F$ to mean that random variable X has the distribution F. We denote the probability mass function of an integer-valued random variable X by $P_X(\cdot)$. The conditional probability mass function of an integer-valued random variable Y given random variable X is denoted by $P_{Y|X}(\cdot|\cdot)$.

Notation A.1: A random variable X that takes only values 0 and 1 will be called a Bernoulli random variable. We will say that X has the Bernoulli distribution with parameter p, or $X \sim \text{Ber}(p)$, iff

$$P_X(0) = 1 - p, \quad P_X(1) = p.$$

Notation A.2: The geometric distribution with parameter a on the non-negative integers is denoted as $\text{Geo}(a)$. Thus $X \sim \text{Geo}(a)$ iff

$$P_X(i) = a^i(1-a), \quad i = 0, 1, 2, \cdots$$

If $X \sim \text{Geo}(a)$, then $E[X] = a/(1-a)$.

Notation A.3: The geometric distribution with parameter a on the positive integers is denoted as $\text{Geo}^+(a)$. Thus $X \sim \text{Geo}^+(a)$ iff

$$P_X(i) = (1-a)^{i-1}a, \quad i = 1, 2, \cdots$$

If $X \sim \text{Geo}^+(a)$, then $E[X] = 1/a$.

Notation A.4: The Poisson distribution with parameter λ is denoted as $\text{Poi}(\lambda)$. $X \sim \text{Poi}(\lambda)$ iff

$$P_X(i) = \frac{e^{-\lambda}\lambda^i}{i!}, \quad i = 0, 1, 2, \cdots$$

Notation A.5: The binomial distribution with parameters n, p is denoted as Bin (n, p). The probability mass at i of Bin (n, p) is denoted as $B(n, i, p)$, $i = 0, 1, \cdots, n$. Thus if $X \sim$ Bin (n, p), then

$$P_X(i) = B(n, i, p) = \binom{n}{i} p^i (1-p)^{n-i}, \quad i = 0, 1, \cdots, k$$

Notation A.6: The exponential distribution with parameter λ is denoted as Exp (λ). If $X \sim$ Exp (λ), then the probability distribution function of X is given by

$$F_X(x) = 1 - e^{-\lambda x}, \quad x \geq 0,$$

and $E[X] = 1/\lambda$. We denote the probability density function of an Exp (λ) random variable by $e_\lambda(\cdot)$, i.e.

$$e_\lambda(x) = \lambda e^{-\lambda x}, \quad x \geq 0.$$

Appendix B

SOME QUEUEING RESULTS FOR A DISCRETE-TIME QUEUE WITH BATCH ARRIVALS AND BATCH SERVICE

In this appendix, we obtain some basic results about the steady-state behaviour of queues within the discrete-time queueing model described in Section 3. The model allows batch arrivals and batch service, and the queue is modelled as being capable of serving S_i packets in slot i. If there are at least S_i packets in the queue, S_i packets depart, whereas if there are less than S_i packets in the queue, all of them are served. S_i are assumed to be i.i.d. The dynamic evolution of this queueing system is governed by (3.1). We will follow the notation of Section 3 in this appendix. Notations A.2 and A.3 in Appendix A.

Proposition B.1. *Consider a queue with a Geo(b) service distribution driven by an arrival process with independent, Geo(a) arrivals in each slot.*

(i) The steady-state distribution of the system occupancy $\{X_i\}$ just after arrivals is Geo(a/b).

(ii) The steady-state distribution of $\{Y_i\}$ is

$$P_Y(0) = \frac{b-a}{b(1-a)}, \tag{B.1a}$$

$$P_Y(k) = \frac{a(1-b)}{b(1-a)} \left(\frac{a}{b}\right)^{k-1} \left(1-\frac{a}{b}\right), \quad k \geq 1. \tag{B.1b}$$

(iii) The steady-state distribution of the departure process $\{D_i\}$ is Geo(a).

(iv) In steady state, the departures occurring in successive slots are independent, i.e. for any n,

$$P_{D_n|D_1,\ldots,D_{n-1}}(k_n|k_1,\ldots,k_{n-1}) = P_{D_n}(k_n). \tag{B.2}$$

Proof. $\{Y_i\}$ clearly forms a Markov chain which is irreducible and aperiodic, so the stationary distribution of Y_i is also the steady-state distribution [9, Sec. 5.1]. It is easy to verify that (B.1) is the stationary distribution of Y_i. Since $X_i = Y_i + A_i$, with Y_i independent of A_i, it can be verified using generating functions that the steady-state distribution of $\{X_i\}$ is Geo(a/b). Since $D_i = \min(X_i, S_i)$, it is easily verified that the steady-state distribution of $\{D_i\}$ is Geo(a).

Let $D^n = (D_1,\ldots,D_n)$ and $Y^n = (Y_1,\ldots,Y_n)$. Given Y_n, D_n is independent of D^{n-1}. We will show that Y_n is independent of D^{n-1} by induction on n. Clearly, if Y_n is independent of D^{n-1}, then so are X_n and D_n.

First consider $n = 2$. We have,

$$P_{Y_2|D_1}(j|k) = \sum_{i=0}^{\infty} P_{Y_2|D_1,X_1}(j|k,i)P_{X_1|D_1}(i|k)$$

Since $X_1 = Y_2 + D_1$, we have $P_{Y_2|D_1,X_1}(j|k,i) = 1$ if $i = j + k$ and 0 otherwise. Hence

$$\begin{aligned} P_{Y_2|D_1}(j|k) &= P_{X_1|D_1}(j+k|k) \\ &= \frac{P_{D_1|X_1}(k|k+j)P_{X_1}(k+j)}{P_{D_1}(k)} \end{aligned}$$

Noting that

$$P_{D_1|X_1}(k|k+j) = \begin{cases} P_{S_1}(k), & j \neq 0 \\ P(S_1 \geq k), & j = 0 \end{cases}$$

and using the steady-state distributions of S_1, D_1 and X_1, and (B.1), we obtain

$$P_{Y_2|D_1}(j|k) = P_{Y_2}(j),$$

so that the result is true for $n = 2$.

Now assume that the result is true for $n-1$, i.e. Y_{n-1} is independent of D^{n-2}. Since $X_{n-1} = Y_n + D_{n-1}$, we have

$$\begin{aligned} & P_{Y_n|D^{n-1}}(j|k_1,\ldots,k_{n-1}) \\ &= P_{X_{n-1}|D^{n-1}}(j+k_{n-1}|k_1,\ldots,k_{n-1}) \\ &= \frac{P_{D_{n-1}|X_{n-1},D^{n-2}}(k_{n-1}|j+k_{n-1},k_1,\ldots,k_{n-2})}{P_{D_{n-1}|D^{n-2}}(k_{n-1}|k_1,\ldots,k_{n-2})} \cdot \\ & \quad P_{X_{n-1}|D^{n-2}}(j+k_{n-1}|k_1,\ldots,k_{n-2}) \end{aligned}$$

As shown before, if Y_{n-1} is independent of D^{n-2}, then so are X_{n-1} and D_{n-1}. Also, D_{n-1} is independent of D^{n-2} given X_{n-1}. Hence we get

$$\begin{aligned} & P_{Y_n|D^{n-1}}(j|k_1,\ldots,k_{n-1}) \\ &= \frac{P_{D_{n-1}|X_{n-1}}(k_{n-1}|j+k_{n-1})P_{X_{n-1}}(j+k_{n-1})}{P_{D_{n-1}}(k_{n-1})} \\ &= \frac{P_{D_1|X_1}(k_{n-1}|k_{n-1}+j)P_{X_1}(k_{n-1}+j)}{P_{D_1}(k_{n-1})} \\ &= P_{Y_1}(j) \end{aligned}$$

as shown earlier. Thus Y_n is independent of D^{n-1}, and so are X_n and D_n. This completes the proof. □

For a queue with arbitrary service distribution, it can be shown that when the arrival process $\{A_i\}$ has independent, geometrically distributed number of arrivals in each slot, the steady-state distribution of $\{X_i\}$ is also geometric. The proof is based on the following lemma, which was also stated in Section 3.

Lemma B.1. *Let S be a non-negative integer-valued random variable with mean μ and probability generating function $\phi_S(\cdot)$, and let $0 < \lambda < \mu$. Then there is a unique solution $\gamma \in \left(\frac{\lambda}{1+\lambda}, 1\right)$ of the equation*

$$\phi_S(x) = 1 + \lambda - \frac{\lambda}{x}. \tag{B.3}$$

Proof. Let $g(x) = 1 + \lambda - \frac{\lambda}{x}$. On (0,1], $\phi_S(\cdot)$ is a convex, increasing function with $0 \leq \phi_S(0) \leq 1$, $\phi_S(1) = 1$, and $\phi'_S(1) = \mu$, while $g(\cdot)$ is a concave, increasing function with $g(\frac{\lambda}{1+\lambda}) = 0$, $g(1) = 1$, and $g'(1) = \lambda < \mu$. Hence the graphs of $\phi_S(\cdot)$ and $g(\cdot)$ must have a unique intersection point $\gamma \in (\frac{\lambda}{1+\lambda}, 1)$. □

In the case when S is geometrically distributed with mean μ, we have $\gamma = \frac{\lambda(1+\mu)}{\mu(1-\lambda)}$.

Proposition B.2. *Consider a queue whose service distribution has mean μ and generating function $\phi_S(\cdot)$, driven by a rate λ geometric arrival process ($\lambda < \mu$). Let γ be the solution in Lemma B.1. Then*

(i) The steady-state distribution of $\{X_i\}$ is Geo(γ).

(ii) The steady-state distribution of $\{Y_i\}$ is

$$P_Y(k) = \begin{cases} (1-\gamma)(1+\lambda), & k = 0 \\ (\gamma + \gamma\lambda - \lambda)\gamma^{k-1}(1-\gamma), & k \geq 1 \end{cases} \tag{B.4}$$

Proof. It can be shown that when the arrival process $\{A_i\}$ has i.i.d. geometrically distributed number of arrivals in each slot, then the Markov chain $\{X_i\}$ is irreducible and aperiodic. Hence if a stationary distribution exists, it must be the steady-state distribution. It is easy to verify that if X_n has the Geo(γ) distribution, then $Y_{n+1} = (X_n - S_n)^+$ has the distribution (B.4), and since $X_{n+1} = Y_{n+1} + A_{n+1}$, X_{n+1} has the same distribution Geo(γ) as X_n. Hence the steady-state distribution of $\{X_i\}$ must be Geo(γ). It follows that (B.4) is the steady state distribution of $\{Y_i\}$. □

Appendix C

ON THE UPPER BOUND TO THE TIMING CAPACITY OF THE $\cdot/Geo^{(s)}/\infty$ SYSTEM

As noted earlier, the upper bound (3.28) on the λ-timing capacity of the $\cdot/Geo^{(s)}/\infty$ system is also an upper bound on the capacity, subject to a certain constraint, of the infinite output binomial channel. However, for channels with infinite input and output alphabets, it is not always true that the supremum in (3.28) is actually achieved by some input probability distribution. In this appendix, we make use of some results from functional analysis to establish the existence of such a distribution for the infinite output binomial channel. We also make use of some results from the theory of convex optimization to obtain a necessary and sufficient of condition that an input distribution to the binomial channel must satisfy to attain this supremum. Our approach for this closely follows the approach used in [1] and [40] to characterize the capacity achieving distributions of, respectively, the discrete-time Rayleigh Fading Channel with an input power constraint and the additive Gaussian noise channel with an input amplitude constraint.

The transition probability $P_{D|X}(\cdot|\cdot)$ of the infinite output binomial channel is given by (3.23). Consider an input distribution F whose probability mass function is Q. Denote by $P_{D_F}(\cdot)$ or $P_{D_Q}(\cdot)$ the marginal probability mass function of the output induced by the input distribution F. For any input distribution F and any input n, define $i(n;F)$ as

$$i(n;F) = \sum_{j=0}^{\infty} P_{D|X}(j|n) \log \frac{P_{D|X}(j|n)}{P_{D_F}(j)}. \tag{C.1}$$

The input-output mutual information I can be considered as a non-negative, real-

valued functional on the space $\mathcal{F}$ of input distributions of the channel, defined by

$$\begin{aligned} I : \mathcal{F} &\rightarrow \mathbb{R} \\ F &\rightarrow E_F[i(X;F)] \end{aligned} \tag{C.2}$$

It is well known (see, e.g., [12, Lemma 3.5, p. 50]) that I is a concave function of the input distribution.

The upper bound of (3.28) is of the general form

$$\overline{C} = \sup_{F \in \mathcal{F}: E_F[f(X)] \leq a} I(F), \tag{C.3}$$

with $f(x) = \mu x$. We will first examine conditions under which the supremum in (C.3) is actually achieved by some input distribution, and then establish that these conditions hold for the particular case (3.28). We will then establish a necessary and sufficient condition that characterizes the distribution that achieves the supremum in (C.3), and examine the form of these conditions in the case of (3.28).

One way to show that the supremum in (C.3) is actually achieved by some input distribution is to define a suitable metric on the space of input distributions under which the mutual information is a continuous function, and the set of input distributions defined by the constraint is compact. To do this, we draw on some standard results from functional analysis. Let $\overline{\mathbb{R}} = [-\infty, \infty]$ be the extended real line, and let $C_b(\overline{\mathbb{R}})$ be the set of bounded continuous functions on $\overline{\mathbb{R}}$. By the Riesz representation theorem [34, Sect. 40, p. 195], the dual space $C_b(\overline{\mathbb{R}})^*$ of $C_b(\overline{\mathbb{R}})$ is the space of totally finite signed measures on the Borel σ-algebra $\mathcal{B}(\overline{\mathbb{R}})$ of $\overline{\mathbb{R}}$, and thus contains the probability measures as a subset. A notion of convergence called weak* convergence can be defined on the dual space $C_b(\overline{\mathbb{R}})^*$ [29, Sect 5.10, p. 127], which when restricted to sequences of probability measures on $(\overline{\mathbb{R}}, \mathcal{B}(\overline{\mathbb{R}}))$, is identical to weak convergence of probability measures [49, Sect. 17.4, p. 183]. We can define weak* continuity of real-valued functionals on the space of probability distributions, and weak* compactness of sets of probability distributions using this notion of weak* convergence, or

equivalently, using the weak convergence of probability measures [29, Sect. 5.10, p. 127]. A set is weak* compact if every sequence in the set contains a weak* convergent subsequence. A functional J is weak* continuous if for any weak* convergent sequence $x_n \to x$, $J(x_n) \to J(x)$. The following theorem from [29, Section 5.10] states that on a set that is compact in the weak* topology, a weak* continuous functional achieves its maximum.

Theorem C.1. *If J is a real-valued weak* continuous functional on a weak* compact set $\Omega \subseteq \mathcal{F}$, then J achieves its maximum on Ω.*

The following lemma asserts that the set of distributions defined by the constraint in (C.3) is weak* compact under some simple conditions on the constraint function $f(\cdot)$. This result was proved in [1] for the special case $f(x) = x^2$ in the situation where the probability distributions under consideration are on the entire real line, rather than on just the non-negative integers. However, the proof given there can be easily modified to the following, and is omitted.

Lemma C.1. *Let $f : \mathbb{R}^+ \to \mathbb{R}$ be any non-negative continuous function on the non-negative real numbers. Consider the set of probability distributions on the non-negative integers $\Omega_f = \{F : E_F[f(X)] \leq a\}$. Suppose that f is non-decreasing, $f(0) = 0$, and f is unbounded. Then Ω_f is weak* compact.*

Lemma C.1 ensures that the set of input distributions satisfying the constraint in (3.28) is weak* compact. Thus if we can establish that the mutual information functional I for the infinite output binomial channel is weak* continuous, then Theorem C.1 guarantees the existence of an input distribution that achieves the supremum in (3.28).

For a given input distribution F to the infinite output binomial channel, let the entropy of the output distribution be denoted by $H_D(F)$, and the conditional entropy of the output given the input be denoted by $H_{D|X}(F)$. Then the mutual information

functional is $I(F) = H_D(F) - H_{D|X}(F)$. Thus to show that the mutual information $I(\cdot)$ is weak* continuous over the set of input distributions satisfying the constraint, it suffices to show that the functionals $H_D(\cdot)$ and $H_{D|X}(\cdot)$ are weak* continuous.

To show that $H_D(\cdot)$ is weak* continuous, we have to show that if a sequence of distributions $\{F_n\}$ converges weakly to a distribution F, then $H_D(F_n) \to H_D(F)$. Let the probability mass function of the distribution F_n be P_{X_n}, and that of F be P_X. Since all the distributions are on the non-negative integers, if $\{F_n\}$ converges weakly to F, then $P_{X_n}(i) \to P_X(i)$ for all integers $i \geq 0$. Let the output probability mass function for the input distribution F_n be P_{D_n}, and that of F be P_D. Using (3.23), we have

$$\begin{aligned} P_{D_n}(i) &= \sum_{j=i}^{\infty} P_{X_n}(j) B(j, i, \mu), \\ P_D(i) &= \sum_{j=i}^{\infty} P_X(j) B(j, i, \mu). \end{aligned}$$

We will first show that if $\{F_n\}$ converges weakly to F, then $P_{D_n}(i) \to P_D(i)$ for all i, so that the sequence of output distributions $\{F_{D_n}\}$ converges weakly to F_D. Note that

$$\begin{aligned} P_{D_n}(i) &= \sum_{j=i}^{\infty} P_{X_n}(j) \binom{j}{i} \mu^i (1-\mu)^{j-i} \\ &= \mu^i \sum_{k=0}^{\infty} P_{X_n}(i+k) \binom{i+k}{i} (1-\mu)^k. \end{aligned}$$

Since $P_{X_n}(i+k) \to P_X(i+k)$, we have

$$P_{X_n}(i+k) \binom{i+k}{i} (1-\mu)^k \to P_X(i+k) \binom{i+k}{i} (1-\mu)^k.$$

Further, for all n,

$$\left| P_{X_n}(i+k) \binom{i+k}{i} (1-\mu)^k \right| \leq \binom{i+k}{i} (1-\mu)^k = g_i(k),$$

where, for a given i, $g_i(k)$ satisfies (see [23, p. 199])

$$\sum_{k=0}^{\infty} g_i(k) = \sum_{k=0}^{\infty} \binom{i+k}{i} (1-\mu)^k = \frac{1}{\mu^{i+1}} < \infty.$$

Therefore, by the dominated convergence theorem using $g_i(k)$ as the dominating function, as $P_{X_n}(i+k) \to P_X(i)$,

$$\sum_{k=0}^{\infty} P_{X_n}(i+k)\binom{i+k}{i}(1-\mu)^k \to \sum_{k=0}^{\infty} P_X(i+k)\binom{i+k}{i}(1-\mu)^k,$$

so that $P_{D_n}(i) \to P_D(i)$ for all i. Since the function $h(x) = -x\log x$ is continuous over $[0,1]$, we have

$$\lim_{n\to\infty}(-P_{D_n}(i)\log P_{D_n}(i)) = (-P_D(i)\log P_D(i)) \quad \forall i.$$

So to show that $H_D(F_n) \to H_D(F)$, i.e. to show that

$$\sum_{i=0}^{\infty}(-P_{D_n}(i)\log P_{D_n}(i)) \to \sum_{i=0}^{\infty}(-P_D(i)\log P_D(i)), \tag{C.4}$$

it suffices to find a dominating function $\beta(i)$ such that $|P_{D_n}(i)\log P_{D_n}(i)| \le \beta(i)\forall i$, and $\sum_{i=0}^{\infty}\beta(i) < \infty$. Using the inequality ([16, p. 530])

$$\binom{j}{i} < e^{jh(i/j)},$$

where $h(\cdot)$ is the binary entropy function, we have

$$\begin{aligned} P_{D_n}(i) &= \sum_{j=i}^{\infty} P_{X_n}(j)\binom{j}{i}\mu^i(1-\mu)^{j-i} \\ &\le \sum_{j=i}^{\infty} P_{X_n}(j)e^{jh(i/j)}\mu^i(1-\mu)^{j-i} \\ &= \sum_{j=i}^{\infty} P_{X_n}(j)e^{-jD(\mathrm{Ber}(i/j)||\mathrm{Ber}(\mu))}, \end{aligned}$$

where Ber (p) is the Bernoulli distribution with parameter x (see Notation A.1 in Appendix A), and $D(P||Q)$ is the Kullback-Leibler distance [10] between two probability distributions. Let $p_{ij} = i/j$, so that

$$\begin{aligned} jD(\mathrm{Ber}\,(i/j)\,||\mathrm{Ber}\,(\mu)) &= \frac{i}{p_{ij}}D(\mathrm{Ber}\,(p_{ij})\,||\mathrm{Ber}\,(\mu)) \\ &= i\log\frac{1}{\mu} + i\left\{\frac{h(p_{ij})}{p_{ij}} - \frac{1-p_{ij}}{p_{ij}}\log(1-\mu)\right\} \\ &\ge i\log\frac{1}{\mu}, \end{aligned}$$

Hence

$$\begin{aligned} P_{D_n}(i) &\leq \sum_{j=i}^{\infty} P_{X_n}(j) e^{-jD(\mathrm{Ber}(i/j)||\mathrm{Ber}(\mu))} \\ &\leq \sum_{j=i}^{\infty} e^{i\log\mu} P_{X_n}(j) \\ &\leq \mu^i. \end{aligned} \tag{C.5}$$

Further, note that $|x \log x| \leq \sqrt{x}$, so that $|P_{D_n}(i) \log P_{D_n}(i)| \leq \sqrt{P_{D_n}(i)}$. From (C.5), we can choose $\beta(i) = (\sqrt{\mu})^i$ as a dominating function, to establish (C.4). This in turn establishes the weak* continuity of $H_D(\cdot)$.

To show that the conditional entropy $H_{D|X}(\cdot)$ of the output given the input is weak* continuous, note that from (3.23), for an input distribution F whose probability mass function is P_X,

$$H_{D|X}(F) = \sum_{n=0}^{\infty} P_X(n) H(Bin(n,\mu)), \tag{C.6}$$

where $H(Bin(n,\mu))$ is the entropy of the binomial distribution $Bin(n,\mu)$ with parameters n and μ. Denote this entropy as $g(n) = H(Bin(n,\mu))$. From [28, Theorem 11.4 A, p. 183], if a sequence of distributions F_n converges weakly to F, and if $g(\cdot)$ is *uniformly integrable* in $\{F_n\}$, then

$$\int g\, dF_n \to \int g\, dF.$$

If $P_{X_n}(\cdot)$ resp. $P_X(\cdot)$ is the probability mass function of F_n resp. F, then if $g(\cdot)$ is uniformly integrable in $\{F_n\}$,

$$\lim_{n\to\infty} \sum_{k=0}^{\infty} g(k) P_{X_n}(k) = \sum_{k=0}^{\infty} g(k) P_X(k), \tag{C.7}$$

which is equivalent to showing that $H_{D|X}(\cdot)$ is weak* continuous. Now $g(\cdot)$ is uniformly integrable in $\{F_n\}$ iff

$$\int g\, dF_n < \infty \quad \text{and} \quad \lim_{b\to\infty} \int_b^{\infty} g(x)\, dF_n = 0 \text{ uniformly in } n,$$

i.e.

$$\sum_{k=0}^{\infty} g(k) P_{X_n}(k) < \infty \quad \text{and} \quad \lim_{b \to \infty} \sum_{k=b}^{\infty} g(k) P_{X_n}(k) = 0 \text{ uniformly in } n. \tag{C.8}$$

There is no closed form expression for $g(k)$. However, since the geometric distribution has the greatest entropy among all the distributions on the non-negative integers having the same mean, the entropy of $Bin(k, \mu)$ is no more than the entropy of the geometric distribution on the non-negative integers with mean $k\mu$. Thus

$$g(k) \leq H\left[\text{Geo}\left(\frac{k\mu}{1+k\mu}\right)\right] = (1+k\mu)h\left(\frac{k\mu}{1+k\mu}\right) = \alpha(k), \tag{C.9}$$

where $\alpha(k)$ satisfies the following properties.

$$\alpha(k) \leq \frac{1+k\mu}{\log(2)}, \tag{C.10a}$$

$$\frac{\alpha(k)}{k} \text{ is monotonically decreasing in } k, \text{ and} \tag{C.10b}$$

$$\lim_{k \to \infty} \frac{\alpha(k)}{k} = 0. \tag{C.10c}$$

From (C.9) and (C.10a), we have

$$\sum_{k=0}^{\infty} g(k) P_{X_n}(k) \leq \sum_{k=0}^{\infty} (1+k\mu) P_{X_n}(k) \leq 1 + \lambda < \infty,$$

using the input constraint $\mu E_{P_{X_n}}[X] \leq \lambda$ in (3.28). Further,

$$\begin{aligned} \sum_{k=b}^{\infty} g(k) P_{X_n}(k) &\leq \sum_{k=b}^{\infty} \alpha(k) P_{X_n}(k) \\ &\leq \frac{\alpha(b)}{b\mu} \sum_{k=b}^{\infty} k\mu P_{X_n}(k) \\ &\leq \frac{\alpha(b)}{b} \lambda, \end{aligned} \tag{C.11}$$

using (C.9), (C.10b), and (3.28). The bound on the right side of (C.11) goes to zero by (C.10c) independent of P_{X_n}, and thus by (C.8), $g(\cdot)$ is uniformly integrable in $\{F_n\}$. This establishes the weak* continuity of $H_{D|X}(\cdot)$.

Since $I(F) = H_D(F) - H_{D|X}(F)$ and both functionals $H_D(X)$ and $H_{D|X}$ are weak* continuous, so is the mutual information functional $I(\cdot)$. Theorem C.1 and Lemma C.1 then assert the existence of an input distribution that achieves the supremum in (3.28).

We will now proceed to develop some results from the theory of convex optimization to establish a necessary and sufficient condition to characterize the distribution that achieves the supremum in (C.3). We will first make use of the Lagrange multiplier theorem [29, Sect. 8.3, Theorem 1], to characterize this information-maximizing distribution. We state this theorem below with slight modifications from its statement in [29, Sect. 8.3]. In essence the theorem converts an optimization over a subset defined by a constraint to global optimization over the entire space.

Theorem C.2. *Let X be a linear vector space, Z a normed space, $\mathcal{F}$ a convex subset of X, and P the positive cone in Z. Assume that P contains an interior point. Let f be a real-valued concave functional on $\mathcal{F}$ and g a convex mapping from $\mathcal{F}$ to Z. Assume the existence of a point $F_1 \in \mathcal{F}$ such that $g(F_1) < 0$. Let*

$$\overline{C} = \sup_{F \in \mathcal{F}: g(F) \leq 0} f(F) \tag{C.12}$$

and assume $\overline{C}$ is finite. Then there is an element $z_0^* \geq 0$ in Z^* such that

$$\overline{C} = \sup_{F \in \mathcal{F}} \{f(F) - \langle g(F), z_0^* \rangle\}. \tag{C.13}$$

Furthermore, if the supremum is achieved in (C.12) by F_0, then it is achieved by F_0 in (C.13) and

$$\langle g(F_0), z_0^* \rangle = 0.$$

The interpretation of this theorem in the case of (C.3) is as follows. X is the linear vector space of totally finite signed measures on $\overline{\mathbb{R}}$. Its convex subset $\mathcal{F}$ is the set of distributions on the non-negative integers. Z is the set of real numbers $\mathbb{R}$ (so

that its positive cone P obviously contains at least one point). The functional g is defined as $g(F) = E_F[f(X)] - a$. The element F_1 such that $g(F_1) < 0$ is the unit step distribution, that puts all the mass at 0. The dual space Z^* of $Z = \mathbb{R}$ is $\mathbb{R}$ itself. By the theorem, there exists $\gamma \geq 0$ (corresponding to z_0^*) such that

$$\overline{C} = \sup_{F \in \mathcal{F}} \{I(F) - \gamma g(F)\}. \tag{C.14}$$

Since the capacity is achieved for some F_0, the supremum in (C.14) is also achieved by F_0. Moreover, $\gamma g(F_0) = 0$ by the theorem, so that either $E_{F_0}[f(X)] = a$ or $\gamma = 0$. In either case, $\overline{C} = I(F_0)$.

The concept of *Gateaux derivative* [29, Sect. 7.2] is useful for characterizing the optima of functionals defined on a vector space. However, the mutual information is defined only over the set of probability distributions, which is not a vector space itself, but rather is a convex subset of the vector space of totally finite signed measures. The concept of *weak derivative*, defined by Smith [40], is actually the restriction of the Gateaux derivative to a convex set of a vector space.

Definition C.1: Let J be a functional on a convex set $\mathcal{F}$. Let F_0 be a fixed element of $\mathcal{F}$, and $\theta \in [0, 1]$. Suppose that

$$J'_{F_0}(F) = \lim_{\theta \downarrow 0} \frac{J(\theta F + (1-\theta)F_0) - J(F_0)}{\theta} \quad \forall F \in \mathcal{F}. \tag{C.15}$$

is well defined for each $F \in \mathcal{F}$. Then J is said to be *weakly differentiable* in $\mathcal{F}$ at F_0, and the map $J'_{F_0} : \mathcal{F} \to \mathbb{R}$ defined by (C.15) is called the *weak derivative of J in $\mathcal{F}$ at F_0*. If J is weakly differentiable in $\mathcal{F}$ at F_0 for all $F_0 \in \mathcal{F}$, then J is said to be *weakly differentiable in $\mathcal{F}$*, or simply weakly differentiable.

The weak derivative at F_0 is a directional derivative, with the "direction" defined by the element F and F_0 in the convex space. Note that if the space $\mathcal{F}$ was the real numbers $\mathbb{R}$, then for a fixed $F_0 \in \mathbb{R}$, if J is weakly differentiable at F_0 in $\mathbb{R}$, $J'_{F_0}(F)$ is in fact the same for all $F > F_0$ and is called the right derivative in the usual sense. Similarly it is the same for all $F < F_0$ and is the usual left derivative. If both these

are equal then J is differentiable and the weak derivative map $J'_{F_0}(F)$ evaluates to the same value for all F, and this value is the derivative of J at F_0 in the usual sense.

Like the Gateaux derivative, the utility of the weak derivative for characterizing the optimum of concave functions obtains from the following result proved by Smith [40].

Theorem C.3. *Consider a weakly differentiable functional J on a convex set $\mathcal{F}$.*

1. *If J achieves a maximum at F_0 then $J'_{F_0}(F) \leq 0$ for all $F \in \mathcal{F}$.*

2. *If J is concave, then $J'_{F_0}(F) \leq 0$ for all $F \in \mathcal{F}$ implies that J achieves its maximum at F_0.*

The functional J we are interested in is $I - \gamma g$. It was shown in [1] that the weak derivative of I is given by

$$I'_{F_0}(F) = E_F[i(X; F_0)] - I(F_0),$$

whenever each term on the right side of the above is finite, where $i(\cdot; F_0)$ is defined by (C.1), and the weak derivative of $g(F) = E_F[f(X)] - a$ is given by

$$g'_{F_0}(F) = g(F) - g(F_0).$$

The weak derivative of $I - \gamma g$ in $\mathcal{F}$ at F_0 is $I'_{F_0} - \gamma g'_{F_0}$. Further $I - \gamma g$ is concave because both I and g are concave. From Theorem C.3, $I - \gamma g$ achieves its supremum over $\mathcal{F}$ at $F_0 \in \mathcal{F}$ if and only if

$$\begin{aligned} I'_{F_0}(F) - \gamma g'_{F_0}(F) &\leq 0 \; \forall F \in \mathcal{F}, \\ \iff \{E_F[i(X; F_0)] - I(F_0)\} - \gamma \{g(F) - g(F_0)\} &\leq 0 \; \forall F \in \mathcal{F}, \\ \iff E_F[i(X; F_0) - \gamma f(X)] &\leq \overline{C} - \gamma a \; \forall F \in \mathcal{F}, \quad \text{(C.16)} \end{aligned}$$

where in the last step, we have used the facts that $g(F) = E_F[f(X)] - a$, $\overline{C} = I(F_0)$, and $\gamma g(F_0) = 0$.

Theorem C.3 can be further strengthened into the following set of necessary and sufficient conditions for a distribution F_0 to maximize $I - \gamma g$. These conditions were established in [1, Theorem 4] for the particular case of $f(x) = x^2$, but the proof given there generalizes almost verbatim to the following for an arbitrary function $f(\cdot)$, and is omitted here.

Theorem C.4. *Let E_0 be the points of increase of a distribution F_0. Then (C.16) holds for all $F \in \mathcal{F}$ if and only if*

$$i(x; F_0) - \gamma f(x) \leq \overline{C} - \gamma a \quad \forall x \tag{C.17a}$$

$$i(x; F_0) - \gamma f(x) = \overline{C} - \gamma a \quad \forall x \in E_0 \tag{C.17b}$$

Let us summarize the above results. Theorem C.1 and Lemma C.1 ensure the existence of an input distribution that maximizes the mutual information over a set of input distributions defined by an inequality constraint, under certain conditions on the input space or the constraint function. Theorem C.2, the Lagrange multiplier theorem, converts the problem from an optimization over a subset of distributions satisfying the given constraint to a global optimization of a different functional, called the Lagrangian, over the set of all input distributions. It also asserts that if the constrained supremum is achieved by an input distribution F_0, then F_0 also achieves the global supremum of the Lagrangian. Theorems C.3 and C.4 obtain a necessary and sufficient condition for a distribution F_0 to attain this global supremum.

From Theorems C.2, C.3, and C.4, an input distribution Q^* achieves the supremum in (3.28) if and only if there exist constants $\gamma \geq 0$ and $\overline{C}$ such that

1. $\gamma(\mu E_{Q^*}[X_{Q^*}] - \lambda) = 0$, and

2.

$$\sum_{i=0}^{n} B(n, i, \mu) \log \frac{B(n, i, \mu)}{P_{D_{Q^*}}(i)} - \gamma \mu n \leq \overline{C} - \gamma \lambda \quad \forall n, \tag{C.18}$$

with equality for all n such that $Q^*(n) > 0$.

The existence of such a distribution Q^* has already been established. The constant $\overline{C}$ is identical to $C'_\infty(\lambda)$ in (3.28).

Appendix D

A CONJECTURE ON THE DECOMPOSITION OF EXPONENTIALLY DISTRIBUTED RANDOM VARIABLES

In this appendix, we make a conjecture about decomposing an exponentially distributed random variable into two independent non-negative component random variables, that is closely related to several of the problems considered in this dissertation. Analogous to a result about Gaussian random variables, we conjecture that there is an essentially unique way to represent an exponentially distributed random variable as the sum of two independent non-negative random variables. A similar conjecture can be made for geometrically distributed random variables as well.

Let $X \sim \text{Exp}\,(1/(a+b))$, so that $E[X] = a+b$. Suppose that $X = W + N$, where W and N are non-negative independent random variables with $E[W] = a$, $E[N] = b$. One possible choice of the distributions for N and W is the following.

$$\begin{aligned} P(N \le x) &= 1 - \exp(-bx), \quad x \ge 0, \\ P(W = 0) &= \frac{b}{a+b}, \\ P(W \le x | W > 0) &= 1 - \exp\left(-\frac{x}{a+b}\right), \quad x > 0. \end{aligned} \tag{D.1}$$

That is, $N \sim \text{Exp}\,(1/b)$, while W is a mixture of a constant at 0 with probability $b/(a+b)$ and $\text{Exp}\,(1/(a+b))$ with probability $a/(a+b)$. This partitioning has played an important role in several problems discussed in this thesis, such as

- the λ-timing capacity of the $\cdot/M/1$ queue obtained in [4], and the closely related capacity of the additive exponential noise channel [45],

- the rate-distortion functions of the Poisson process under the following distortion measures:
  - the average delay distortion measure discussed in [17],
  - the distortion measure identified in [45],
  - the service time distortion measure discussed in Chapter 4.

The similarities of the results obtained for exponential random variables and Poisson processes using this partition of an exponential random variable to analogous results for Gaussian random variables have been pointed out in [45]. The corresponding results about Gaussian random variables are based on the fact that a Gaussian random variable can be expressed as the sum of two independent Gaussian random variables. A result by Cramer [11] (see also [15, p. 498]) states that the converse of this is also true. Suppose that a Gaussian random variable X can be represented as $X = X_1 + X_2$, where X, X_1, X_2 are all zero mean, $E[X_1^2] + E[X_2^2] = E[X^2]$, and X_1 and X_2 are independent. Cramer's result says that the only non-trivial solution for X_1 and X_2 is that both are Gaussian random variables.

We conjecture that if X is an exponential random variable, any non-trivial partition of X as a sum of two non-negative independent random variables W and N with given expected values is equivalent to (D.1), with the roles of W and N in (D.1) possibly reversed.

An affirmative resolution of this conjecture has interesting consequences. For instance, it would prove that the among all the $M/G/1$ queues, the $M/M/1$ queue is the only one whose departure process in steady state is Poisson. Recall that the interdeparture times D_i can be written as $D_i = W_i + S_i$, where S_i is the service time of packet i, and W_i is the idle time of the server immediately prior to the arrival of packet i. Thus W_i and S_i are independent. For D_i is to be exponentially distributed, if the above conjecture is true, then it is easy to see that S_i must be exponentially

distributed.

For integer valued random variables, a similar conjecture can be made about partitioning $\text{Geo}^{+}(a)$ and $\text{Geo}(a)$ random variables into two independent components. We state the respective partitions below for completeness.

Let $X \sim \text{Geo}^{+}(a)$. We want to express $X = W + N$, where W and N independent non-negative integer-valued random variables such that $E[N] = 1/b$ and $E[W] = (1/a) - (1/b)$, where $0 < b < a < 1$. We conjecture that any such partition is equivalent to the following, possibly with the roles of W and N reversed.

$$\begin{aligned} P(N = k) &= (1-b)^{k-1}b, \quad k \geq 1 \\ P(W = k) &= \begin{cases} \dfrac{b}{a}, & k = 0 \\ (1 - \dfrac{b}{a})(1-a)^{k-1}a, & k \geq 1. \end{cases} \end{aligned} \tag{D.2}$$

Let $X \sim \text{Geo}(u)$, and let $X = W + N$, where W and N are independent non-negative random variables. Again, we conjecture that any such partition is equivalent to

$$\begin{aligned} P(N = k) &= v^k(1-v), \quad k \geq 0 \\ P(W = k) &= \begin{cases} \dfrac{1-u}{1-v}, & k = 0 \\ (\dfrac{u-v}{1-v})u^{k-1}(1-u), & k \geq 1. \end{cases} \end{aligned} \tag{D.3}$$

where $0 < v < u < 1$.

It may be noted that there are interesting ways of expressing a geometric random variable as the sum of two *dependent* random variables that are related to the form expressed by (D.3). Let $X \sim \text{Geo}(u)$, and let W be defined by the conditional probability distribution

$$P_{W|X}(k|n) = \binom{n}{k} b^k (1-b)^{n-k}, \quad 0 \leq k \leq n.$$

It can be verified that W and $N = X - W$ are both geometrically distributed, with

$$W \sim \text{Geo}\left(\frac{ub}{1-u+ub}\right), \quad N \sim \text{Geo}\left(\frac{u(1-b)}{1-ub}\right).$$

Thus a geometrically distributed random variable can be represented as a sum of two geometrically distributed, but dependent, random variables.

Another interesting way to represent a geometrically distributed random variable as the sum of two dependent random variables is based on Proposition B.2, which we used to prove a lower bound for the λ-timing capacity of discrete-time batch-arrival batch-service queues in Proposition 3.5. For $0 < u < 1$, let $X \sim \mathrm{Geo}(u)$, and let S be any non-negative integer-valued random variable independent of X. Let

$$v = \frac{u\{1 - \phi_S(u)\}}{1 - u\phi_S(u)},$$

where $\phi_S(z) = E[z^S]$ is the probability generating function of S. Note that $0 < v < u < 1$. From Proposition B.2, it can be verified that $W = (X - S)^+$ has the same distribution as given in (D.3). W and N are dependent in general, but if S is geometrically distributed, then $N = X - W = \min(X, S)$ is also geometrically distributed and independent of W, in which case this partition becomes equivalent to (D.3).

The above partition also leads to a analogous partition of an exponential random variable into two dependent parts. Let $X \sim \mathrm{Exp}(1/u)$, and let S be any non-negative random variable independent of X. Let $\phi_S(x) = E[e^{-xS}]$. Let $a = u\phi_S(1/u)$ and $b = u(1 - \phi_S(1/u))$, so that $u = a + b$. Then $W = (X - S)^+$ has the same distribution as given in (D.1), although $N = X - W$ is not independent of W. Note that if S is exponentially distributed, then $N = X - W = \min(X, S)$ is also exponentially distributed and independent of W.

VITA

Anand Sharad Bedekar is a native of Bombay, India. He received the B. Tech. degree in Electrical Engineering from the Indian Institute of Technology, Bombay, in 1989, and the M. S. degree in Electrical Engineering from the University of Washington, Seattle, Wash., in 1995. He was an intern at the Mathematics of Networks and Systems group of Lucent Technologies' Bell Laboratories, Murray Hill, New Jersey, from June to September 1998, and at Teledesic Corp., Bellevue, Wash., from October to December 1999.