INFORMATION-THEORETIC ANALYSIS OF HUMAN-MACHINE MIXED SYSTEMS

BY

DAEWON SEO

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Electrical and Computer Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2019

Urbana, Illinois

Doctoral Committee:

Assistant Professor Lav R. Varshney, Chair
Professor Pierre Moulin
Professor Rayadurgam Srikant
Professor Venugopal V. Veeravalli

Abstract

Many recent information technologies such as crowdsourcing and social decision-making systems are designed based on (near-)optimal information processing techniques for machines. However, in such applications, some parts of systems that process information are humans and so systems are affected by bounded rationality of human behavior and overall performance is suboptimal. In this dissertation, we consider systems that include humans and study their information-theoretic limits. We investigate four problems in this direction and show fundamental limits in terms of capacity, Bayes risk, and rate-distortion.

A system with queue-length-dependent service quality, motivated by crowdsourcing platforms, is investigated. Since human service quality changes depending on workload, a job designer must take the level of work into account. We model the workload using queueing theory and characterize Shannon's information capacity for single-user and multiuser systems.

We also investigate social learning as sequential binary hypothesis testing. We find somewhat counterintuitively that unlike basic binary hypothesis testing, the decision threshold determined by the true prior probability is no longer optimal and biased perception of the true prior could outperform the unbiased perception system. The fact that the optimal belief curve resembles the Prelec weighting function from cumulative prospect theory gives insight, in the era of artificial intelligence (AI), into how to design machine AI that supports a human decision.

The traditional CEO problem well models a collaborative decision-making problem. We extend the CEO problem to two continuous alphabet settings with general rth power of difference and logarithmic distortions, and study matching asymptotics of distortion as the number of agents and sum rate grow without bound.

To my wife and children

Acknowledgments

I would like to thank my advisor, Professor Lav R. Varshney, for his continued encouragement and guidance. He was always creative and patient, and the freedom he gave me allowed me to explore diverse fields. This dissertation would not be possible without his unlimited support. In particular, I was fortunate to be the initial member of his group. The time that I spent with him will be a priceless asset to me. I would also like to express my gratitude to Professors Pierre Moulin, Rayadurgam Srikant, and Venugopal V. Veeravalli for serving on my doctoral committee and their feedback. I am also very thankful to Professor Negar Kiyavash, now at Georgia Institute of Technology, for serving on my prelim committee.

I would also like to thank my collaborators for introducing me to some exciting topics. Working with (now Professor) Avhishek Chatterjee greatly broadened my research spectrum, and central ideas on queue-length-dependent channels (Chapters 2 and 3) were jointly done with him. The work on social learning (Chapter 4) was inspired by Professor Vivek K Goyal and Joong Bum Rhim and carried out with Ravi Kiran Raman of our group.

Also, I cannot help mentioning my lab mates and UIUC friends who made or are making an effort in research. Talking with them about miscellaneous school work, coursework and research was always beneficial to me. Interaction with these excellent friends has shaped me. An incomplete list follows (and I apologize for anyone I have omitted): Haizi Yu, Linjia Chang, Ting-Yi Wu, Xiou Ge, Sourya Basu, Yongjune Kim, Jonathan Ligo, Yuheng Bu, Meghana Bande, Srilakshmi Pattabiraman, Aditya Deshmukh, Juho Kim, Minji Kim, Minwoo Kim, Jaeho Lee, and Sijung Yang.

My deepest thanks should go to my wife, Hye Young Lee. She has been my closest friend and I would have never complete the dissertation without her. Also my children, Yoonjae and Yoonji, have given much meaning to my life since their birth. Finally, I would like to thank my parents and parents-in-law for their unlimited support.

I gratefully acknowledge financial support provided by Systems on Nanoscale Information fabriCs (SONIC), one of the six SRC STARnet Centers, sponsored by SRC and DARPA and the National Science Foundation grant CCF-1623821 and CCF-1717530.

Table of Contents

Chapter 1

Introduction

Since its initiation by Shannon [1], the main goal of information theory has traditionally been to understand fundamental limits of machines such as communication devices or storage systems. There have been many efforts to determine the fundamental limits of such devices, and now our understanding of the obstacles we face is clear, although state-of-the-art information theory still does not give complete answers.

However, human-inspired models are not well-studied yet. As the historian of information theory Ronald Kline has forcefully argued [2], there was an initial euphoria in the late 1940s and early 1950s surrounding information-theoretic approaches to human-oriented problems, but this quickly dissipated within the mainstream of research, cf. [3, 4]. Humans have bounded rationality that is hard to model mathematically, perhaps due to cognitive limitations of minds or the time available to make the decision. However, it is true that there is consistency in human behavior. One might wonder whether human behavior is consistent enough to warrant analysis through (perhaps stochastic) mathematical models, the way physical communication channels and information sources seem to. Many long-standing descriptions of people from psychology are consistent and dependable, displaying test-retest reliability, inter-rater reliability, parallel-forms reliability, and internal consistency reliability [5].

Prior works in statistical signal processing and in psychology have separately and independently considered technological limitations and human limitations, but jointly considering the informational and attentional limitations of both humans and machines will be critical in engineering future sociotechnical systems. Hence, this dissertation investigates human-machine mixed systems through an information-theoretic lens.

1.1 Motivation and Prior Work

We study three problems motivated by human behavior, each of which models a distinct aspect of people: 1) workload and work performance, 2) decisions based on previous decisions, and 3) estimation from human's biased belief.

1.1.1 Queue-Length-Dependent Channels

The first topic is workload impact on service quality. Unlike machines or computers, the quality of service by a human worker depends on his/her workload. It is known that overloading a person with work often negatively impacts their quality of work as noted in psychology [6, 7]. Similarly, it is known that when doctors are facing a long queue of patients, they feel rushed and make more mistakes [8, 9].

A typical application scenario of such workload-dependent service quality is crowdsourcing. The organization (e.g., Visipedia [10]) submits jobs to a crowdsourcing platform (e.g., Samasource [11]), and the platform dispatches them to the crowdworker to whom jobs are assigned in the server. Error-correcting codes can be developed for difficult human computations, as described in [12]. Crowdsourcing platforms, like other large-scale systems including distributed storage and cloud computing systems, currently use simple and queue-length-agnostic job assignment policies. Thus crowdsourcing platforms acting as dispatchers can be assumed to be independent of and agnostic about the worker load.

Multimedia communication is another motivating scenario. When a user is in a live video or VoIP call over a multiple-access network, the access point—e.g., WiFi router or base station—has to contend for wireless resources to send the information packets. This results in an accumulation of packets at the MAC buffer of the access point. When the buffer is close to overflow, the access point either drops them [13], sends their corresponding low-quality versions [14] (assuming multiresolution coding [15]), or packs multiple MAC packets in the available time slot using higher coding/modulation. All of these scenarios can be modeled by queue-dependent service quality. We are interested in the maximum rate for reliable data transmission in this system. As multimedia communication uses open-loop transport layer protocols like UDP, the packet dispatcher (application) is agnostic and independent of the load at the server (WiFi access point or BS).

As far as we know, most of the information-theoretic literature focuses on timing capacity results [16–19]. Although we use some proof techniques related to those used in these works, we are not concerned with information encoded in the timing between packets, only in the information in the symbols. There are some interesting works addressing what is called *age of information*, relating loss of information to queuing delays [20], but those settings are different as they are concerned with perishable information.

The study of information-theoretic limits of queuing multiple-access channels was pioneered by Telatar [21], and further explored in [22, 23]. This line of work is essentially concerned with the reliable transmission of bursty sources [24], as we are here. A recent study of microbial communication also had a kind of self-interference called *channel clogging* [25].

1.1.2 Social Learning

Team decision-making typically involves individual decisions influenced by private observations and the opinions of the rest of the team. The *social learning* setting is one such context where decisions of individual agents are influenced by preceding agents in the team [26, 27]. We consider the setting in which individual agents are selfish and aim to minimize their perceived Bayes risk, according to their beliefs as reinforced by the decisions of preceding agents. Social learning, also referred to as observational learning, has been widely studied and we provide a non-exhaustive listing of some of the relevant works.

Aspects of conformism and "herding" were studied in [28–30], where an incorrect decision may cascade for the rest of the agents once agents at the beginning make incorrect decisions. The concept of herding is a consequence of boundedly informative private signals [31]. For example, assume the private signals are binary and give true or false information, each with positive probability. It can happen that a couple of the first agents receive false private signals and thus choose wrong actions. Then, the effect of these actions on the beliefs of subsequent agents can be so great as to cause them to ignore their private signals and follow their precedent agents. The private signals are bounded so that they are not strong enough to overcome the effect of the wrong actions. Further convergence properties of actions taken under social learning have been explored under imperfect information [32]. The notion of sequential social learning has been generalized to learning from neighbors in networks [33], and explored in generality [34]. Social learning has also been explored under quantization of priors [35], and distributed detection with symmetric fusion [36].

Such a learning problem has also been studied under the name of distributed inference or learning. The traditional setup assumes a central fusion node that aggregates all the information from distributed nodes and makes the final decision [37, 38], where the links between distributed nodes and fusion center could be rate-limited [39] or imperfect [40–42]. It is also common to consider such a learning problem in a distributed manner over networks. By repeatedly updating local information without complete knowledge of network connectivity, it is shown that all nodes can identify the true hypothesis [43–45]. Recently, independent works [46] and [47] proposed a similar update rule and convergence result for fixed networks and time-varying networks, respectively. In [48], binary hypothesis testing with time-varying means according to Gaussian process is studied and minimal expected stopping times are derived. In [49], the setup where the entire hypotheses are locally indistinguishable, but globally identifiable is considered and large deviation convergence rate is provided.

Rhim and Goyal's work [50] differs from the aforementioned literature in the sense that they consider unbounded private signals so that there is no herding behavior. In addition, they focus largely on the effects of prior probability and private signal strength. Information is only propagated along the chain once so there is no iterative belief update. Unlike sequential decision-making [51] where all agents know the true prior, agents may have *beliefs* that do not match the true prior.

Human actions are typically affected by individual perceptions of the underlying context. Cumulative prospect theory [52–54] seeks to provide a psychological understanding of human behaviors under risk. It introduces the notion of probability reweighting functions to explain boundedly rational human behaviors. Among reweighting functions, the Prelec reweighting function [55] has significant empirical support and satisfies a majority of the axioms of prospect theory.

In the era of AI (Artificial Intelligence or Augmented Intelligence), a sequential decision-making model has a particular motivation since it captures the nature of collaboration in human-AI teams with either the AI system advising the human who makes the final decision or, less typically, a human advising an AI system that makes the final decision [56, p. 56]. Examples of the former include AI-assisted physicians or chess players (called centaur chess), and of the latter, human-in-the-loop AI systems such as crowdsourcing systems and collaborative filtering mechanisms. Our work proves the interesting conclusion that a team of suboptimal human-AI could beat the team of individually optimal human-AI, if it is well-composed.

1.1.3 Generalized CEO Problems

The last topic is the CEO problem. Consider a particular motivating scenario that there is a sequence of probabilities of successes $\{X(t)\}_{i=1}^{\infty}$. The CEO (chief executive officer) of an organization is interested in $\{X(t)\}_{i=1}^{\infty}$, but does not observe it directly. Instead, there are L agents of the organization who make noisy perceptions (or observations); the ith agent has noisy version $\{Y_i(t)\}_{i=1}^{\infty}$ by its own model such as copula or independent additive noise. The agents must convey their observations to the CEO without convening, but the CEO has cognitive constraints that limit the information rate she can receive from agents, requiring each agent to discretize his observation under rate constraints $\{R_i\}_{i=1}^{L}$. The CEO declares $\{\widehat{X}(t)\}_{t=1}^{\infty}$ that minimizes a distortion (or risk) function $d(X(t), \widehat{X}(t)) = |X(t) - \widehat{X}(t)|^r$ in a long-term average sense.

As we will see, this scenario generalizes existing CEO problem literature in two aspects: source-observation model and distortion function. The first CEO problem by Berger et al. [39] was with discrete alphabets, so the Hamming distortion was considered. Later a jointly Gaussian setting was studied with quadratic distortion [57], where the asymptotic tradeoff between sum rate and distortion was investigated. The quadratic Gaussian CEO problem was further studied in [58–60], finding the exact rate region for finite agents. Under the logarithmic distortion, the exact rate region for general setting was found [61] and we gave the rate region for the jointly Gaussian case explictly using quadratic-logarithmic distortion duality [62]. In contrast to the jointly Gaussian CEO problem, non-Gaussian and non-quadratic CEO problems have received less attention due to limited analytic tractability compared with the Gaussian case. A non-regular source-observation pair such as copula model or truncated Gaussian noise was considered under quadratic distortion [63], and a general continuous source with additive Gaussian noise was considered under quadratic distortion and general distortion [64]. Toward generalization of source-observation pair, it was shown that Gaussianity is in fact the worst [65].

Although [63] considers copula models, the distortion is still quadratic so our scenario belongs to none of aforementioned literature. Regarding the distortion measure, the absolute distortion ($r = 1$) is in particular important when our estimation is *consistent* or *asymptotically consistent*, i.e., the estimate converges to the true value as the number of observations increases, so $|x - \widehat{x}|$ is small with high probability. To illustrate the importance, recall the Maclaurin approximation: a non-decreasing difference distortion function $d_{\mathsf{gen}}(x, \widehat{x}) = d_{\mathsf{gen}}(|x - \widehat{x}|) : \mathbb{R}_+ \mapsto \mathbb{R}_+$ can be expanded around small $|x - \widehat{x}|$ as (assuming

$d_{\mathsf{gen}}(0) = 0)$[1]

$$d_{\mathsf{gen}}(|x - \widehat{x}|) = d'_{\mathsf{gen}}(0)|x - \widehat{x}| + \frac{d''_{\mathsf{gen}}(0)}{2!}|x - \widehat{x}|^2 + \frac{d'''_{\mathsf{gen}}(0)}{3!}|x - \widehat{x}|^3 + \cdots ,$$

where $d'_{\mathsf{gen}}(0), d''_{\mathsf{gen}}(0), d'''_{\mathsf{gen}}(0)$ are right derivatives of d_{gen}. Suppose that the estimator $\widehat{X}$ is consistent. Under appropriate assumptions,[2] the linear term dominates the distortion function as

$$\begin{aligned} d_{\mathsf{gen}}(|x - \widehat{x}|) &= d'_{\mathsf{gen}}(0)|x - \widehat{x}| + o(|x - \widehat{x}|) \\ \implies \mathbb{E}\left[d_{\mathsf{gen}}(|X - \widehat{X}|)\right] &= d'_{\mathsf{gen}}(0)\mathbb{E}\left[|X - \widehat{X}|\right] + o\left(\mathbb{E}\left[|x - \widehat{x}|\right]\right) \\ \implies D_{\mathsf{gen}} &= d'_{\mathsf{gen}}(0)D_{\mathsf{abs}} + o(D_{\mathsf{abs}}), \end{aligned}$$

which shows that the absolute difference distortion D_{abs} is a dominant portion of the general difference distortion function.

We will also discuss an extension of quadratic Gaussian CEO problem [57] to general regular source-observation model with rth power of difference and logarithmic distortions [61].

There are two classical asymptotic approaches that have been developed for CEO problems. The first takes asymptotics in the number of agents [39, 57], where the number of agents grows without bound keeping individual coding rate fixed. In this regime, the nature of detection (for discrete alphabet) or estimation (for continuous alphabet) plays a key role. The second takes asymptotics in individual coding rate with fixed number of agents [66], which highlights the nature of compression. Note that distortion asymptotics of the two regimes in terms of sum rate are different even for a common model. In this work, we will take the first approach.

[1] This approximation for one-sided function is not well defined, but we may think of an extension of d_{gen} on small neighborhood around origin such that all left derivatives agree with their right counterparts at the origin. Then, the Maclaurin series is well defined for the extended function.

[2] Note that $\mathbb{E}\left[d(|X - \widehat{X}|)\right] - d'(0)\mathbb{E}\left[|X - \widehat{X}|\right] = \sum_{k=2}^{\infty} \frac{d^{(k)}(0)}{k!}\mathbb{E}\left[|X - \widehat{X}|^k\right]$. Hence, the condition for the approximation to be valid is equivalent to the fact that the infinite series on the right side vanishes with the number of observations. For example, if all $d^{(k)}(0)$ are absolutely bounded by a constant, and the estimator is consistent and has a sub-Gaussian tail, then the series vanishes with the number of observations.

1.2 Dissertation Outline and Contributions

This dissertation is organized as follows.

Chapter 2 introduces the queue-length-dependent channel for discrete-time queues and discusses its capacity. First, the capacity expression for general queues is developed using information spectrum method and ergodicity of queues. Then, restricting two special types of arrivals, say Type I and Type II, we study two special types of queues, $\mathsf{G}/\mathsf{geo}/1$ and $\mathsf{geo}/\mathsf{G}/1$, for which stationary distributions are available in closed form, so we are able to optimize arrival and service processes.

Chapter 3 studies the single-user capacity of continuous-time queue-length-dependent channels, and extremal properties are derived for two special types of queues, $\mathsf{GI}/\mathsf{M}/1$ and $\mathsf{M}/\mathsf{GI}/1$. Then, the multiaccess capacity is studied using point processes. In particular, when the number of transmitters is large and each is sparse, the superposition of arrivals approaches a Poisson point process. In characterizing the Poisson approximation, we show that the capacity of the multiuser system converges to the capacity of a single-user $\mathsf{M}/\mathsf{GI}/1$ queue-length-dependent system.

Chapter 4 generalizes the social learning problem of Rhim and Goyal [50] with agents having diverse expertise. In addition, we introduce the Prelec weighting function from cumulative prospect theory and study its (near-)optimality and suboptimality depending on expertise levels. A self-organizing team construction is also discussed. This work emphasizes that suboptimal advising could be more helpful for human decision-making than the optimal advising when human belief is biased.

Chapter 5 extends existing CEO problems to two continuous alphabet settings with general rth power of difference and logarithmic distortions, and studies asymptotics of distortion as the number of agents and sum rate grow without bound. The first setting is called a regular source-observation model, such as jointly Gaussian, with difference distortion, and we show that the distortion decays at $R_{\mathsf{sum}}^{-r/2}$ up to a multiplicative constant. The other setting is called a non-regular source-observation model, such as copula or uniform additive noise models, with difference distortion for which estimation-theoretic regularity conditions do not hold. The optimal decay $R_{\mathsf{sum}}^{-r}$is obtained. Lastly, we provide a condition for the regular model, under which quadratic and logarithmic distortions are asymptotically equivalent by entropy power relation as the number of agents grows.

Chapter 6 concludes this dissertation and notes future research directions.

For clarity and readibility, we have deferred proofs to the appendices if they are noncrucial.

1.3 Bibliographical Note

Parts of Chapter 2 appear in the paper:

- A. Chatterjee, D. Seo, and L. R. Varshney, "Capacity of Systems with Queue-Length Dependent Service Quality," in *Proceedings of the International Symposium on Information Theory and Its Applications*, Oct.-Nov. 2016.

and in the journal paper:

- A. Chatterjee, D. Seo, and L. R. Varshney, "Capacity of Systems with Queue-Length Dependent Service Quality," *IEEE Transactions on Information Theory*, Jun. 2017.

Parts of Chapter 3 appear in the paper:

- D. Seo, A. Chatterjee, and L. R. Varshney, "On Multiuser Systems with Queue-Length Dependent Service Quality," in *Proceedings of the IEEE International Symposium on Information Theory*, Jun. 2018.

Parts of Chapter 4 appear in the paper:

- D. Seo, R. K. Raman, and L. R. Varshney, "Probability Reweighting in Social Learning: Optimality and Suboptimality," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Apr. 2018.

Parts of Chapter 5 appear in the paper:

- D. Seo and L. R. Varshney, "The CEO Problem with rth Power of Difference and Logarithmic Distortions," to appear in *Proceedings of the IEEE International Symposium on Information Theory*, July 2019.

and in the journal manuscript:

- D. Seo and L. R. Varshney, "The CEO Problem with rth Power of Difference and Logarithmic Distortions," arXiv:1812.00903 [cs.IT].

Chapter 2

Capacity of Systems with Queue-Length-Dependent Service Quality

In this chapter, we define the capacity of systems with queue-length-dependent service quality as the number of bits reliably processed per unit time, and we characterize this measure it in terms of queuing system parameters. In particular, discrete-time queues are considered with two different types of arrivals, say Type I (at most one arrival per time slot) and Type II (multiple arrivals per time slot), and the capacity theorem is studied separately. In addition, for a $\mathsf{geo}/\mathsf{G}/1$ queue, it turns out that deterministic service is the best and cramming and idle service is the worst. Similarly, for a $\mathsf{G}/\mathsf{geo}/1$ queue, deterministic arrivals and bursty arrivals are the best and the worst, respectively.

2.1 Problem Formulation

We keep the standard transmitter-receiver structure equipped with encoder and decoder, but the channel is modeled as having quality of service that is queue-length-dependent in nature. As usual, a transmitter and a receiver *a priori* agree on a set of possible sequences of symbols (or codebook). The transmitter sends a sequence of symbols corresponding to a message to the dispatcher. The dispatcher sends these symbols to a server according to some stochastic process. The server services these symbols, which are then received by the receiver. The receiver then tries to decode the message based on the received symbols.

The server works like a single first-in first-out (FIFO) queue with i.i.d. service requirements for each job. Jobs correspond to symbols from a finite field $\mathbb{F}$. In the model, servicing a job involves reading the symbol and outputting it. The server may make random errors during these steps and send out erroneous symbols. We are interested in the information capacity of such a system, which we refer to as a *queue-channel*.

2.1.1 Queuing Discipline

We consider a discrete-time system, $t \in \mathbb{Z}_+ := \{0, 1, 2, \ldots\}$. Define $A_i \in \mathbb{Z}_+, D_i \in \mathbb{N}$ to be the inter-arrival time and inter-departure time of the ith job. The service time of the ith job is denoted S_i, and is strictly positive and i.i.d. drawn from a distribution P_S on $\mathbb{N}$.

We use the following convention. Arrivals at time t, if any, happen at the beginning of time slot t. Departures from the queue at time slot t, if any, happen at the end of the time slot. This implies that a job arriving at time slot t may receive and possibly finish its service at time t.

Let $Q(t)$ be the number of jobs in the queue at the end of time slot t and Q_i be the number of jobs in the system when the ith job departs. As $S_i \geq 1$ for all i, at a time slot t, at most one job can depart.

We consider two basic types of arrival processes (also called dispatch processes) into the queue: Type I and Type II.[1] In a Type I process, there is at most one arrival in any time slot and the times between two consecutive arrivals are i.i.d. with distribution P_A on $\mathbb{N}$. Hence the support of A_i is $\mathbb{N}$ for Type I, i.e., $P_A(0) = 0$. In a Type II process, the numbers of arrivals $A(t)$ in time slot $t \geq 1$ are i.i.d. with distribution m_A on $\mathbb{Z}_+$. The service rate and arrival rate are μ and λ, respectively, satisfying $\mathbb{E}_{P_S}[S] = 1/\mu$ and $\mathbb{E}_{P_A}[A] = 1/\lambda$ or $\mathbb{E}_{m_A}[A] = \lambda$, respectively. For stability of the queue, we assume $\lambda < \mu$. We assume P_S, P_A, and m_A have finite second moments. For Type I systems, we assume either P_A or P_S has a support that spans $\mathbb{N}$. For Type II systems, we assume $m_A(1) > 0$.

2.1.2 Service Noise

Transmission of symbols from a finite field $\mathbb{F}$ over the queue-channel happens in two stages. Mapping the message, the transmitter sends symbols $\{X_i \in \mathbb{F} : 1 \leq i \leq n\}$ to a dispatcher, which in turn sends the symbols to the server according to a stochastic process of arrival rate λ.

The symbol corresponding to the ith symbol is $X_i \in \mathbb{F}$, and the output symbol corresponding to the ith symbol is $Y_i \in \mathbb{F}$. They are related through the additive noise variable $Z_i \in \mathbb{F}$ representing work error, such that $Y_i = X_i \oplus Z_i$. The distribution of the errors Z_i depends on Q_i. For any i, given Q_i, Z_i is independent of any other processes or variables, and has a distribution ψ_q (on $\mathbb{F}$) for $Q_i = q$.

[1] Type I and II are analytically tractable sub-classes of the arrival processes with i.i.d. inter-arrival times and i.i.d. numbers of arrivals at each arrival epoch.

An n-length transmission over the queue-channel is denoted as follows. Inputs are $\{X_i : 1 \leq i \leq n\}$, channel realizations are $\{Z_i : 1 \leq i \leq n\}$, and outputs are $\{Y_i : 1 \leq i \leq n\}$. Throughout, a k-dimensional random vector is denoted by $U^k = (U_1, U_2, \ldots, U_k)$.

All logarithms in the chapter have base 2 so that information is measured in bits.

2.2 Capacity of Queue-Channel

We are interested in the information capacity of unreliable server systems, i.e., the queue-channel described above. In this section, we present results that are generic, i.e., are true for both Type I and II arrivals.

2.2.1 Definition

Let $M, \widehat{M} \in \mathcal{M}$ be the message to be transmitted and decoded, respectively.

Definition 1. *An $(n, \widetilde{R}, T)$ code consists of the encoding function $X^n = f(M)$ and the decoding function $\widehat{M} = g(X^n, A^n, D^n)$, where the cardinality of the message set $|\mathcal{M}| = 2^{n\widetilde{R}}$, and for each codeword, the expected total time for all symbols to reach the receiver is less than T.*

Definition 2. *If the decoder chooses $\widehat{M}$ with average probability of error less than ϵ, that code is said to be ϵ-achievable. For any $0 < \epsilon < 1$, if there exists an ϵ-achievable code $(n, \widetilde{R}, T)$, the rate $R = \widetilde{R}/T$ is said to be achievable.*

Definition 3. *For an arrival process with distribution P_A (Type I) or m_A (Type II), the information capacity of the queue-channel is defined as the supremum over all achievable rates, which is denoted by $C(P_A)$ or $C(m_A)$ in bits per unit time.*

Since the transmitter sends symbols to the dispatcher first, we assume the transmitter knows the arrival process statistics, but not the realizations. Contrarily, the receiver knows the realized arrival and departure times of each job.

2.2.2 Coding Theorem

Here, the transmitter does not observe $\{A_i, D_i\}$, whereas the receiver observes these. Thus the queue-channel has inputs $\{X_i\}$ and outputs $\{Y_i, A_i, D_i\}$. As dispatch is independent of job-design, the channel transition probability factors as

$$\mathbb{P}(Y^n, A^n, D^n | X^n) = \mathbb{P}(A^n, D^n)\mathbb{P}(Y^n | X^n, A^n, D^n).$$

The transmitter chooses $\{X_i\}$ and hence can choose any joint distribution for the codebook described by $\{X_i\}$. Note that $\{Y_i, A_i, D_i\}$ depends on $\{X_i\}$, as well as on the arrival and service processes. In general, $\{Y_i, A_i, D_i\}$ may not be a stationary process. This means that the queue-channel is not necessarily an information-stable channel [67], but the capacity formula can nevertheless be found using the information spectrum approach [68,69]. Let the information density be $i(\cdot)$, the normalized information density be

$$\frac{1}{n} i(X^n; Y^n, A^n, D^n) = \frac{1}{n} \log \frac{\mathbb{P}(Y^n, A^n, D^n | X^n)}{\mathbb{P}(Y^n, A^n, D^n)},$$

and the inf-information rate $\underline{\mathbf{I}}(\mathbf{X}; \mathbf{Y}, \mathbf{A}, \mathbf{D})$ be the *lim-inf in probability* of the normalized information density, i.e., the largest $\alpha \in \mathbb{R} \cup \{\pm\infty\}$ such that for all $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left[\frac{1}{n} i(X^n; Y^n, A^n, D^n) \le \alpha - \epsilon\right] = 0.$$

Then, capacity in bits per unit time of the queue-channel is given by

$$C(P_A) \text{ (and } C(m_A)) = \lambda \sup_{\mathbb{P}(\mathbf{X})} \underline{\mathbf{I}}(\mathbf{X}; \mathbf{Y}, \mathbf{A}, \mathbf{D}), \tag{2.1}$$

where λ is the arrival rate of P_A (or m_A), and the supremum is over all input processes $\mathbf{X} = (X_1, X_2, \ldots)$.

This capacity expression is not easy to handle due to the various possibilities of (A^n, D^n) that can arise; however, the next proposition allows us to characterize the distribution of $i(\cdot)$ (and hence, $\underline{\mathbf{I}}$) in a simpler form, in terms of the distributions of X^n, Y^n, and Q^n.

Proposition 1. *The capacity expression (2.1) can be represented by using Q^n,*

$$C(P_A) \textit{ (and } C(m_A)) = \lambda \sup_{\mathbb{P}(\mathbf{X})} \underline{\mathbf{I}}(\mathbf{X}; \mathbf{Y} | \mathbf{Q}).$$

Proof. It suffices to show that

$$i(X^n; Y^n, A^n, D^n) = i(X^n; Y^n | Q^n). \tag{2.2}$$

Note that the additive noise Z^n depends only on $Q^n = \phi_n(A^n, D^n)$, where $\phi_n(\cdot)$ is a function that computes the number of symbols in the queue. Hence, $\mathbb{P}(Y^n|A^n, D^n, X^n) = \mathbb{P}(Y^n|Q^n, X^n)$. Also, X^n is independent of (A^n, D^n).

$$\begin{aligned}
\frac{\mathbb{P}(Y^n, A^n, D^n|X^n)}{\mathbb{P}(Y^n, A^n, D^n)} &= \frac{\mathbb{P}(A^n, D^n|X^n)\mathbb{P}(Y^n|A^n, D^n, X^n)}{\mathbb{P}(A^n, D^n)\mathbb{P}(Y^n|A^n, D^n)} \\
&= \frac{\mathbb{P}(Y^n|A^n, D^n, X^n)}{\mathbb{P}(Y^n|A^n, D^n)} = \frac{\mathbb{P}(Y^n|Q^n, X^n)}{\mathbb{P}(Y^n|A^n, D^n)} \\
&= \frac{\mathbb{P}(Y^n|Q^n, X^n)}{\sum_{X^n} \mathbb{P}(Y^n, X^n|A^n, D^n)} \\
&= \frac{\mathbb{P}(Y^n|Q^n, X^n)}{\sum_{X^n} \mathbb{P}(X^n|A^n, D^n)\mathbb{P}(Y^n|A^n, D^n, X^n)} \\
&= \frac{\mathbb{P}(Y^n|Q^n, X^n)}{\sum_{X^n} \mathbb{P}(X^n|Q^n)\mathbb{P}(Y^n|Q^n, X^n)} \\
&= \frac{\mathbb{P}(Y^n|Q^n, X^n)}{\mathbb{P}(Y^n|Q^n)}.
\end{aligned}$$

Taking logarithm and normalizing yield $i(X^n; Y^n, A^n, D^n) = i(X^n; Y^n|Q^n)$. □

Thus, it follows that the distribution of $i(\cdot)$ depends only on the joint distribution of (X^n, Y^n, Q^n).

Based on this, we can give a single-letter characterization of the capacity of the queue-channel. In the proof of the forthcoming coding theorem, the converse part is essentially due to basic properties of information quantities [69, 70]. The direct part follows by choosing an appropriate input process $\mathbf{X}$ to lower bound $\sup_{\mathbb{P}(\mathbf{X})} \underline{\mathbf{I}}(\mathbf{X}; \mathbf{Y}|\mathbf{Q})$. In this regard, this proof is structurally similar to earlier work that applied information spectrum techniques, e.g. [17, 71].

The proof of the coding theorem also implicitly depends on the following lemma which characterizes the process $\{Q_i\}$.

Lemma 1. *Under the assumptions in Sec. 2.1 and $\lambda < \mu < 1$, there exists a unique distribution π such that if $Q_1 \sim \pi$, then $Q_i \sim \pi$ for all $i \geq 1$, and the process $\{Q_i\}$ is ergodic, i.e., for any $f : \mathbb{Z}_+ \to \mathbb{R}$ with finite $\mathbb{E}_\pi f$, almost surely $\frac{1}{n}\sum_{i=1}^{n} f(Q_i) \to \mathbb{E}_\pi f$ as $n \to \infty$. Moreover, for any initial distribution of Q_1, Q_i converges to π in distribution and $\pi(q) > 0$ for all $q \in \mathbb{Z}_+$.*

Proof. See Appendix A.1. □

Now the capacity theorem.

Theorem 1. *For a given arrival process distribution P_A (or m_A) with $\lambda < \mu < 1$ which follows the assumption in Sec. 2.1, there exists a distribution π such that $\pi(q) > 0$ for all $q \in \mathbb{Z}_+$ and $\mathbb{P}(Q_n) \to \pi$ as $n \to \infty$. The capacity of this queue-channel is $\lambda(\log|\mathbb{F}| - \sum_q \pi(q)H(\psi_q))$, where $H(\psi_q)$ is the entropy of a distribution $\psi_q(Z)$ on any finite set of size $|\mathbb{F}|$.*

Proof. From properties of limit superior and inferior [69],

$$\underline{\mathbf{I}}(\mathbf{X};\mathbf{Y}|\mathbf{Q}) \leq \overline{\mathbf{H}}(\mathbf{Y}|\mathbf{Q}) - \overline{\mathbf{H}}(\mathbf{Y}|\mathbf{X},\mathbf{Q}).$$

Since $\overline{\mathbf{H}}(\mathbf{Y}|\mathbf{Q}) \leq \log|\mathbb{F}|$ by Thm. 1.7.2 in [69] for any $\mathbb{P}(\mathbf{Y})$,

$$\underline{\mathbf{I}}(\mathbf{X};\mathbf{Y}|\mathbf{Q}) \leq \log|\mathbb{F}| - \overline{\mathbf{H}}(\mathbf{Y}|\mathbf{X},\mathbf{Q}).$$

Note that $\overline{\mathbf{H}}(\mathbf{Y}|\mathbf{X},\mathbf{Q})$ is the lim-sup in probability of $\frac{1}{n}\log\frac{1}{\mathbb{P}(Y^n|X^n,Q^n)}$, i.e., the smallest $\beta \in \mathbb{R} \cup \{\pm\infty\}$ such that

$$\lim_{n\to\infty} \Pr\left[\frac{1}{n}\log\frac{1}{\mathbb{P}(Y^n|X^n,Q^n)} \geq \beta + \epsilon\right] = 0$$

for any $\epsilon > 0$. Since noise is additive, by Lem. 1,

$$\begin{aligned}&\frac{1}{n}\log\frac{1}{\mathbb{P}(Y^n|X^n,Q^n)} = \frac{1}{n}\log\frac{1}{\psi_{Q_i}(Z_i)}\\&\to \mathbb{E}_{\pi_Q,Z}[-\log\psi_Q(Z)] \text{ almost surely as } n\to\infty\\&= \sum_q \pi_q H(\psi_q).\end{aligned}$$

Therefore we obtain the converse bound that

$$\underline{\mathbf{I}}(\mathbf{X};\mathbf{Y}|\mathbf{Q}) \leq \log|\mathbb{F}| - \sum_q \pi_q H(\psi_q).$$

On the other hand, we also have

$$\underline{\mathbf{I}}(\mathbf{X};\mathbf{Y}|\mathbf{Q}) \geq \underline{\mathbf{H}}(\mathbf{Y}|\mathbf{Q}) - \overline{\mathbf{H}}(\mathbf{Y}|\mathbf{X},\mathbf{Q}).$$

The second term converges to $\sum_q \pi_q H(\psi_q)$ by Lem. 1. We pick X^n i.i.d. uniformly at random from $\mathbb{F}$. Note that as $\mathbb{F}$ is a field, for any element $y_i \in \mathbb{F}$, $y_i - X_i$ spans all elements in $\mathbb{F}$.

Hence, $\sum_{X\in\mathbb{F}} \psi_{Q_i}(Y_i - X_i) = 1$. Thus,

$$\begin{aligned}\mathbb{P}(Y_i|Q_i) &= \sum_{X_i\in\mathbb{F}} \mathbb{P}(Y_i, X_i|Q_i) = \sum_{X_i\in\mathbb{F}} \mathbb{P}(X_i|Q_i)\mathbb{P}(Y_i|X_i, Q_i) \\ &= \sum_{X_i\in\mathbb{F}} \frac{1}{|\mathbb{F}|}\psi_{Q_i}(Y_i - X_i) = \frac{1}{|\mathbb{F}|},\end{aligned}$$

and then when $\mathbb{P}(\mathbf{X})$ is uniform:

$$\underline{\mathbf{H}}(\mathbf{Y}|\mathbf{Q}) = \log|\mathbb{F}|.$$

Hence

$$\underline{\mathbf{I}}(\mathbf{X};\mathbf{Y}|\mathbf{Q}) \geq \log|\mathbb{F}| - \sum_q \pi_q H(\psi_q),$$

and multiplying by the arrival rate λ completes the proof. □

Note that the expressions in Thm. 1 are given for additive noise channels, but since the coding theorem (Prop. 1) and ergodic lemma (Lem. 1) hold for general queue-length-dependent channels, Thm. 1 can easily be generalized.

With the coding theorems developed in this section in hand, Secs. 2.3 and 2.4 study the capacity of a few interesting classes of discrete-time queues. This results in insights regarding the dispatch and service processes that have the best and worst information processing rates.

2.2.3 Comments

Before studying specific classes of queuing systems, we comment on the relation between the maximum packet throughput and the maximum information throughput (the notion of capacity defined here) of a queuing system. Packet throughput of a queuing system is the maximum rate of packet arrivals that can be served without instability; hence the packet throughput increases with λ on $[0, \mu)$. Though the expression for capacity (information throughput) has λ as a multiplicative factor, this does not mean that information throughput increases with λ. In typical queuing systems, the survival function corresponding to the stationary probability is increasing in λ. Thus, an increase in λ also has a negative impact on the terms involving π. Hence, in typical queuing systems, there is an optimal $\lambda \in (0, \mu)$ that maximizes information throughput. Fig. 2.1 shows an example.

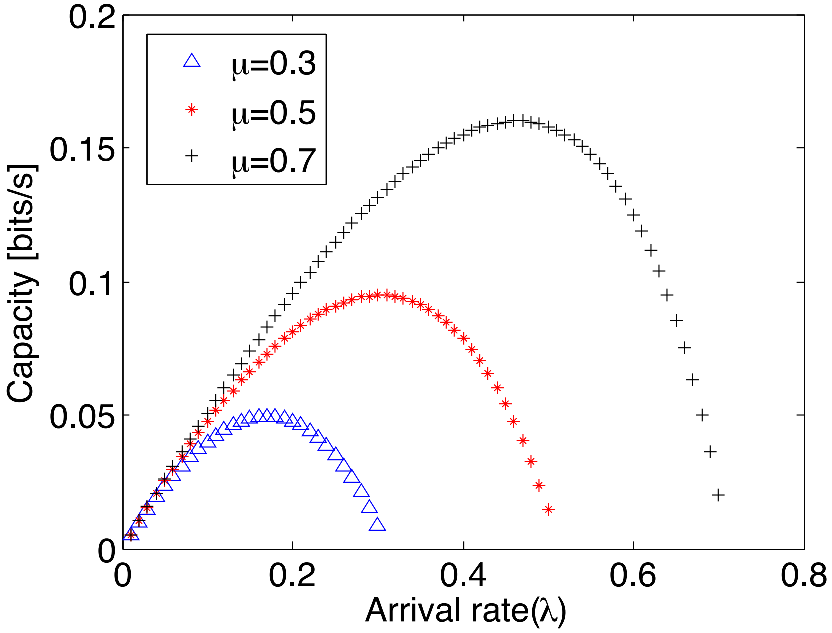


Figure 2.1: Capacity of geo/geo/1 queue is plotted against arrival rate (for different service rates) for $\mathbb{F} = \{0, 1\}$ and noise distribution $\mathbb{P}(Z = 1) = 0.1$ for $q = 0$, otherwise $\mathbb{P}(Z = 1) = 0.4$.

2.3 Queues with Type I Arrival

This section is devoted to understanding the capacity of a queue with a Type I arrival process and its dependence on the distribution of service times and inter-arrival times. First, we find the capacity of a queue with geometric service time and arbitrary arrival process, and characterize the capacity-optimizing arrival distributions. Then, we study the capacity of a queue with geometric inter-arrival time and find the capacity-optimizing service time distribution. Capacity has a saddle-point behavior around the geometric distribution.

In the application scenarios such as crowdsourcing and multimedia communication, server performance deteriorates with increasing queue-length. Deterioration of server performance with increasing queue-length is captured by a $\{\psi_q\}$ whose entropy is non-decreasing with q. A $\{\psi_q\}$ of practical interest is a threshold behavior of the error-entropy with increasing queue-length: $H(\psi_q) = h_0$ for $q \leq b$ and $H(\psi_q) = h_{b+1}$ for $q \geq b + 1$, for some $b \in \mathbb{Z}_+$.

Threshold behavior captures a state of server panic based on workload, suitable for human servers and wireless access points with small MAC buffer. The special case of $b = 0$ describes a human server that is distracted by any waiting job or a bufferless MAC. The special case of $b = 1$ corresponds to a human server being distracted if more than one job is waiting.

2.3.1 Discrete-time $\mathsf{G/geo/1}$ Queue

For a $\mathsf{G/geo/1}$ queue, the service time distribution is geometric with an expected service time $\frac{1}{\mu}$, $\mu < 1$. The arrival process is Type I, with the inter-arrival times distributed as P_A and the expected time between arrivals $\frac{1}{\lambda}$, $\lambda < \mu$. Since this queueing system satisfies the assumptions in Sec. 2.1, its capacity can be obtained from Thm. 1. For any arrival distribution P_A, the capacity of $\mathsf{G/geo/1}$ queue is given by the following theorem.

Theorem 2. *The capacity of the* $\mathsf{G/geo/1}$ *queue-channel is* $\lambda(\log|\mathbb{F}| - (1-\sigma)\sum_q \sigma^q H(\psi_q))$, *where* σ *is the unique solution of the equation* $x = \sum_{n=0}^{\infty} P_A(n)(1-\mu+x\mu)^n$ *in* $(0,1)$.

Proof. See Appendix A.2. □

Proof of this theorem involves obtaining the steady-state distribution π of the queue-lengths seen by the departures. Towards this, techniques similar to that in the analysis of continuous-time $\mathsf{GI/M/1}$ queues [72] are extended to the discrete-time setting. The closed-form expressions here differ to some extent from that in $\mathsf{GI/M/1}$. Also, note that some of the intermediate steps in the proof of Thm. 2 are used to prove some later results.

Based on the capacity characterization of the $\mathsf{G/geo/1}$ queue, we explore the space of arrival distributions. This leads to the following result about the best and worst (in terms of capacity) arrival distribution for a $\mathsf{G/geo/1}$ queue.

Proposition 2. *For* $\mathsf{G/geo/1}$ *queue with thresholded noise such that* $H(\psi_0) = \cdots = H(\psi_b) < H(\psi_{b+1}) = \cdots$ *for some* $b \in \{0,1,\ldots\}$, *deterministic inter-arrival time maximizes capacity among all arrival distributions with the same* λ, *for* $\frac{1}{\lambda} \in \mathbb{Z}_+$.

Proof. Proof of this result builds on the property of the fixed point equation $x = \sum_{t=0}^{\infty}(1-\mu+\mu x)^t P_A(t)$, and uses an intermediate result in the proof of Thm. 2.

First see that for any arrival distribution P_A, $\pi(q) = (1-\sigma)\sigma^q$, and capacity is $\log|\mathbb{F}| - \sum_q \pi(q)H(\psi_q)$, which is maximized when $(1-\sigma)\sum_q \sigma^q H(\psi_q)$ is minimized. Since noise is thresholded at b, i.e., $h_0 = H(\psi_0) = \cdots = H(\psi_b)$ and $h_{b+1} = H(\psi_{b+1}) = \cdots$, then the latter term may be written as

$$\begin{aligned}(1-\sigma)\sum_q \sigma^q H(\psi_q) &= h_0(1-\sigma^{b+1}) + h_{b+1}(1-(1-\sigma^{b+1})) \\ &= h_0 + (h_{b+1}-h_0)\sigma^{b+1}.\end{aligned}$$

Hence, for a given $\{\psi_q\}$, capacity is maximized when σ is minimized.

Next, note that the curves $\widetilde{A}(\sigma) = \sum_{t=1}^{\infty} P_A(t)(1-\mu+\mu\sigma)^t$ are convex and increasing with σ, and $\widetilde{A}(0) > 0$ (see Lem. 17 in Appendix). Also, there is a unique fixed point in $(0, 1)$. Thus, for these classes of curves, the curve that lower bounds a set of curves crosses the line $y = \sigma$ at the smallest value of σ among that set of curves. Similarly, the curve that upper bounds a set of curves crosses the line $y = \sigma$ at the largest value of σ.

For any $0 < \alpha < 1$ and any distribution P_A with mean $\frac{1}{\lambda}$,

$$\sum_{t=0}^{\infty} \alpha^t P_A(t) \geq \alpha^{\frac{1}{\lambda}},$$

by Jensen's inequality, as α^t is convex. Thus for any $\sigma \in (0, 1)$ and P_A with mean $\frac{1}{\lambda}$,

$$\begin{aligned}\widetilde{A}(P_A, \sigma) &= \sum_{t=0}^{\infty} (1-\mu+\mu\sigma)^t P_A(t) \\ &\geq (1-\mu+\mu\sigma)^{\frac{1}{\lambda}} \\ &= \widetilde{A}(\mathsf{det}, \sigma),\end{aligned}$$

where the equality can be attained by a deterministic inter-arrival time. This implies that the curve $\widetilde{A}(\mathsf{det}, \sigma)$ is a lower-bounding curve for all other curves corresponding to different P_A. □

Proposition 3. *For the* $\mathsf{G/geo/1}$ *queue with* $\{\psi_q\}$ *such that* $H(\psi_0) = \cdots = H(\psi_b) < H(\psi_{b+1}) = \cdots$ *for some* $b \in \mathbb{Z}_+$*,* $\widetilde{p}_A(t, \epsilon)$ *asymptotically minimizes the capacity among all arrival processes as* $\epsilon \to 0$*, where*

$$\widetilde{p}_A(t, \epsilon) = \begin{cases} 1-\epsilon, & t = 1 \\ \epsilon, & t = N(\epsilon), \end{cases}$$

for $\epsilon > 0$ *and* $N(\epsilon)$ *is chosen to satisfy the mean constraint* $1/\lambda$*.*

Proof. It is sufficient to show that $\widetilde{A}(P_A, \sigma)$ is asymptotically maximized by $\widetilde{p}_A(t, \epsilon)$ as $\epsilon \to 0$.

Consider developing an upper bound of $\widetilde{A}(P_A, \sigma)$ first. Using the fact that for $\alpha \in (0, 1)$, α^t

is decreasing,

$$\begin{aligned}\widetilde{A}(P_A, \sigma) &= \sum_{t=0}^{\infty}(1-\mu+\mu\sigma)^t P_A(t) \\ &\leq \sum_{t=0}^{\infty}(1-\mu+\mu\sigma) P_A(t) = (1-\mu+\mu\sigma).\end{aligned}$$

On the other hand, $\widetilde{A}(P_A, \sigma)$ evaluated at $\widetilde{p}_A(t, \epsilon)$ is:

$$\widetilde{A}(\widetilde{p}_A(t, \epsilon), \sigma) = (1-\mu+\mu\sigma)(1-\epsilon) + (1-\mu+\mu\sigma)^N \epsilon,$$

which approaches the upper bound as $\epsilon \to 0$, but has a fixed-point solution in $(0, 1)$. The pmf $\widetilde{p}_A(t, \epsilon)$ asymptotically maximizes the fixed-point solution as $\epsilon \to 0$, thus minimizing the capacity. □

The results of Prop. 2 and 3 agree with our intuition. Deterministic arrivals in Prop. 2 give enough time to the server with a given service rate, so that each job sees the lowest queue length behind it on average. On the other hand, a typical realization of $\widetilde{p}_A(t, \epsilon)$ is that jobs arrive every time slot (corresponding to $t = 1$) for some time interval but then the next job arrives a very long time later corresponding to $t = N(\epsilon)$. The server will be busiest during the first interval, but will be almost idle until the next job. It yields the worst performance.

In crowdsourcing, it is common for the arrival process to come from some kind of job pre-processing. Since this pre-processing system itself could be serial or parallel chains of servers with exponentially distributed random delays, we are interested in classes of arrival processes that are certain geometric families of distributions.

Let $\{A_i, 1 \leq i \leq I\}$ be independent geometric random variables with means $\frac{1}{\lambda_i}$. Then define A^s to be a sum-of-geometric random variable and to be $\mathcal{A}^s$ the set of such probability distributions with mean $\frac{1}{\lambda}$, i.e.,

$$\begin{aligned}A^s &= \sum_i A_i, \\ \mathcal{A}^s &= \left\{ P_{A^s} : \mathbb{E}[A^s] = \frac{1}{\lambda} \right\}.\end{aligned}$$

Also define A^m to be a mixture of geometric random variables such that $A^m = A_i$ with probability mass $\{c_i\}$ whose support is $\{1 \leq i \leq I\}$, with $\mathcal{A}^m$ as the set of such probability distributions with mean $\frac{1}{\lambda}$, i.e.,

$$A^m = A_i \text{ with probability } c_i,$$
$$\mathcal{A}^m = \left\{ P_{A^m} : \mathbb{E}[A^m] = \frac{1}{\lambda} \right\}.$$

Then the next lemma follows.

Lemma 2. *For any $P_{A^s} \in \mathcal{A}^s$,*

$$\widetilde{A}(P_{A^s}, \sigma) \leq \widetilde{A}(\mathsf{geo}, \sigma).$$

On the other hand, for any $P_{A^m} \in \mathcal{A}^m$.

$$\widetilde{A}(P_{A^m}, \sigma) \geq \widetilde{A}(\mathsf{geo}, \sigma).$$

Proof. See Appendix A.3. □

Proposition 4. *For any $\mathsf{G}/\mathsf{geo}/1$ queue-channel with thresholded noise at b, geometric inter-arrival times minimize and maximize capacity among all arrival distributions of $\mathcal{A}^s$ and $\mathcal{A}^m$, respectively.*

Proof. Following the arguments in the proof of Prop. 2, we only need to show that geometric inter-arrival times maximizes (resp. minimize) σ for a given λ among $\mathcal{A}^s$ (resp. $\mathcal{A}^m$). Then the proposition follows from Lem. 2. □

There is an important takeaway from this result in the context of job pre-processing for crowdsourcing. In crowdsourcing systems all jobs are pre-processed to make them suitable for crowd workers, and the inter-arrival (inter-dispatch) time in our model corresponds to this pre-processing time. The above theorem implies it is best to have a deterministic pre-processing time. However, if pre-processing times are highly variable due to some system issues (geometric is the most entropic), then instead of having a single pre-processing step it is better to have a series of sub-steps (corresponding to sum-of-geometric) for pre-processing.

A corollary is the capacity extrema representation of the $\mathsf{G}/\mathsf{geo}/1$ queue-channel among the class of sum- (resp. mixture-) of-geometric distributions.

Corollary 1. *For a given arrival rate λ and given service rate μ, the minimum (resp. maximum) capacity of $\mathsf{G}/\mathsf{geo}/1$ queue-channel among the class of sum- (resp. mixture-) of-geometric inter-arrival distributions is $\lambda(\log|\mathbb{F}| - (1-\sigma^*)\sum_q \sigma^{*q} H(\psi_q))$, where $\sigma^* = \frac{\lambda(1-\mu)}{\mu(1-\lambda)}$.*

Proof. From the proof of Prop. 4, we know that geometric arrival achieves capacity extrema for arrival rate λ,

$$\sum_{t=0}^{\infty} \alpha^t P_A(t) = \frac{\frac{\alpha}{1-\alpha}}{\frac{1}{\lambda} + \frac{\alpha}{1-\alpha}}.$$

Letting $\alpha = 1 - \mu + \mu\sigma$ and solving the fixed point equation

$$\widetilde{A}(\mathsf{geo}, \sigma) = \frac{\frac{1-\mu+\mu\sigma}{\mu-\mu\sigma}}{\frac{1}{\lambda} + \frac{1-\mu+\mu\sigma}{\mu-\mu\sigma}} = \sigma,$$

we have the unique solution $\sigma^* = \frac{\lambda(1-\mu)}{\mu(1-\lambda)}$. □

2.3.2 Discrete-time $\mathsf{geo/G/1}$ Queue

In this section we consider another important class of queues, for which the arrival process is Bernoulli, i.e., inter-arrival times are geometric, but the service times have a general distribution.

Define $\binom{n}{k} = 0$ if $n < k$. By characterizing the stationary distribution seen by departures, we prove the following capacity result.

Theorem 3. *For $j \in \{0, 1, \ldots\}$, let $k_j = \sum_{t=0}^{\infty} \binom{t}{j} (1-\lambda)^{t-j} \lambda^j P_S(t)$, and for all complex z with $|z| < 1$, $K(z) = \sum_{j=0}^{\infty} z^j k_j$, then capacity of this system is $\lambda(\log|\mathbb{F}| - \sum_q \pi_q H(\psi_q))$ for $\pi_0 = 1 - \frac{\lambda}{\mu}$, and $\pi_k = \lim_{z\to 0} \frac{\Pi(z) - \sum_{j=0}^{k-1} \pi_j z^j}{z^k}$, where $\Pi(z) = (1 - \frac{\lambda}{\mu}) \frac{(z-1)K(z)}{z - K(z)}$.*

Proof. See Appendix A.4. □

Derivation of the stationary distribution here follows similar steps as the derivation of the stationary distribution of $\mathsf{M/G/1}$ queue [72].

Next, we investigate the service time distributions that respectively maximize or minimize the capacity of a $\mathsf{geo/G/1}$ queue-channel. First, we consider the case of threshold error-entropy behavior for threshold $b = 0$. The following corollary is a direct consequence of Thm. 3.

Corollary 2. *If $\{\psi_q\}$ are such that $H(\psi_0) < H(\psi_1) = H(\psi_2) = \cdots$, then the capacity of the $\mathsf{geo/G/1}$ queue is the same for all P_S.*

Proof. The capacity in this case depends on π only through π_0, but π_0 is the same for all service distributions with mean $\frac{1}{\mu}$, as $\pi_0 = 1 - \frac{\lambda}{\mu}$. □

For the case with threshold $b = 1$, it can be shown that different service time distributions result in different capacities.

Proposition 5. *If $H(\psi_0) = H(\psi_1) < H(\psi_2) = H(\psi_3) = \cdots$, then the capacity of the* $\mathsf{geo/G/1}$ *queue is maximized by a deterministic service time (for $\frac{1}{\mu} \in \mathbb{Z}_+$) and is asymptotically minimized by $\widetilde{p}_S(t, \epsilon)$ as $\epsilon \to 0$, where*

$$\widetilde{p}_S(t, \epsilon) = \begin{cases} 1 - \epsilon, & t = 1 \\ \epsilon, & t = N(\epsilon), \end{cases}$$

for $\epsilon > 0$ and $N(\epsilon)$ is chosen to satisfy the mean constraint $1/\mu$. Among the class of sum-of-geometric random variables, capacity is minimized by the geometric service time distribution.

Proof. Let $h_0 = H(\psi_0)$ and $h_2 = H(\psi_2)$, $h_0 < h_2$. Then, by Thm. 1, the capacity of the system is $\lambda(\log |\mathbb{F}| - h_0(\pi_0 + \pi_1) - h_2(1 - \pi_0 - \pi_1))$.

It is clear from the capacity expression that it is maximized (resp. minimized) when $\pi_0 + \pi_1$ is maximized (resp. minimized). Hence, it is enough to prove that deterministic service time maximizes $\pi_0 + \pi_1$, and geometric service time minimizes $\pi_0 + \pi_1$ among the class of sum-of-geometric random variables.

Note that

$$\pi_1 = \lim_{z \to 0} \frac{\Pi(z) - \pi_0}{z},$$

which, after a few steps of algebra using the expression for $\Pi(z)$ and the fact that $\pi_0 = 1 - \frac{\lambda}{\mu}$, gives

$$\pi_1 = (1 - \tfrac{\lambda}{\mu}) \frac{1 - K(0)}{K(0)}.$$

Thus, $\pi_0 + \pi_1 = (1 - \frac{\lambda}{\mu}) \frac{1}{K(0)}$. Using the definition that $K(0) = k_0 = \mathbb{P}(\text{no arrivals in } S)$, the capacity is minimized when k_0 is maximized and vice-versa. After decomposing k_0 for all t,

$$k_0 = \sum_{t=1}^{\infty} (1 - \lambda)^t P_S(t).$$

The conclusion follows from the proofs of Lem. 2 and Prop. 3: the deterministic arrival with mass at $\frac{1}{\mu} \in \mathbb{Z}_+$ minimizes k_0 by Jensen's inequality, and $\widetilde{p}_S(t, \epsilon)$ asymptotically maximizes k_0 as $\epsilon \to 0$. In addition, k_0 is maximized by geometric distribution among the class of sum-of-geometric random variables by Lem. 2. □

Prop. 5 says that handling works with regularity in time yields the least queue length on average; cramming and staying idle is the worst. For thresholded noise behavior we observe the following. For a geometric service time, the worst dispatch process among the sum-of-geometric distributions is geometric. On the other hand, for a geometric arrival process, the geometric service time is the worst among the sum-of-geometric distributions. Thus, if we visualize the capacity function of a single server queue for a given arrival and service rate plotted against arrival and service distributions (restricted to sum-of-geometric), there is a minimum where both distributions are geometric.

In the context of crowdsourcing, this means it is always better to split highly variable pre-processing (corresponding to the dispatch process) or human work (corresponding to the service process) steps into a series of sub-steps. That is, it is always better to take a job by parts (if coordination costs are not too high [73]).

2.4 Queues with Type II Arrivals

In this section, we study the queue-capacity of systems with Type II arrivals. An equivalent capacity expression holds for Type II arrivals, i.e., possibly multiple arrivals in a time slot. Let N_i be a random variable counting the number of arrivals at time i. Thus, the $\{N_i\}$ are i.i.d. with distribution m_A.

Theorem 4. *The queue-channel capacity of a queue with Type II arrivals distributed as m_A, and service time distributed as P_S is given by $\lambda(\log|\mathbb{F}| - \sum_q \pi_q H(\psi_q))$, for $\pi_0 = 1 - \frac{\lambda}{\mu}$, and $\pi_k = \lim_{z\to 0} \frac{\Pi(z) - \sum_{j=0}^{k-1} \pi_j z^j}{z^k}$, where $\Pi(z) = (1-\frac{\lambda}{\mu})\frac{(z-1)K(z)}{z-K(z)}$, $k_j = \sum_{t=1}^{\infty} \mathbb{P}(\sum_{i=1}^{t} N_i = j) P_S(t)$.*

Proof. The probability of j arrivals within a service time is $\sum_{t=1}^{\infty} \mathbb{P}(\sum_{i=1}^{t} N_i = j) P_S(t)$. The remainder of the proof follows the same approach as the proof of Thm. 3. □

2.4.1 Effects of Service Processes

First, we characterize the effect of different service processes on the capacity, for a given arrival process. As the following results show, deterministic service is best and bursty service is worst, as in Prop. 5.

Proposition 6. *Suppose that $H(\psi_0) = H(\psi_1) < H(\psi_2) = \cdots$. For a given Type II arrival process m_A, the maximum capacity is achieved by deterministic service time over all service* *time distributions. The minimum capacity is asymptotically achieved by $\widetilde{p}_S(t,\epsilon)$ as $\epsilon \to \infty$, where*

$$\widetilde{p}_S(t,\epsilon) = \begin{cases} 1-\epsilon, & t = 1 \\ \epsilon, & t = N(\epsilon), \end{cases}$$

for $\epsilon > 0$ and $N(\epsilon)$ is chosen to satisfy the mean constraint $1/\mu$. In addition, the minimum capacity among the class of sum-of-geometric random variables is achieved by geometric service time distribution.

Proof. Following the proof of Prop. 5, we only need to prove that k_0 is minimized and asymptotically maximized by deterministic service time and $\widetilde{p}_S(t,\epsilon)$, respectively. Further, k_0 needs to be maximized by geometric service time among the class of sum-of-geometric random variables.

Note that $k_0 = \sum_{t=1}^{\infty} \mathbb{P}(\sum_{i=1}^{t} N_i = 0) P_S(t) = \sum_{t=1}^{\infty} (m_A(0))^t P_S(t)$, where $0 < m_A(0) < 1$. Hence, the results follow from the proof of Prop. 5. □

2.4.2 Effects of Arrival Processes

Next, we are interested in understanding the effect of arrival processes on the capacity for the worst service time distribution. Specifically, we are interested in finding the arrival processes that maximize and minimize the capacity.

Analogous to Cor. 2 and Prop. 5 for Type I systems, we have the following results.

Corollary 3. *Consider the queue with given arrival rate λ and service distribution P_S. If $H(\psi_0) < H(\psi_1) = H(\psi_2) = \cdots$, the capacity of the queue with Type II arrival is the same for all arrival distributions.*

Proposition 7. *For $H(\psi_0) = H(\psi_1) < H(\psi_2) = H(\psi_3) = \cdots$, for a given arrival rate λ and a service distribution P_S, the capacity of the queue-channel over all Type II arrival processes with finite support $\{0, 1, \ldots, B\}$ is lower-bounded by*

$$C_L = \lambda \left(\log |\mathbb{F}| + (H(\psi_2) - H(\psi_0))(1 - \frac{\lambda}{\mu}) \frac{1}{k_0} - H(\psi_2) \right),$$

where $k_0 = \sum_t (1 - \frac{1}{B\lambda})^t P_S(t)$.

Proof. By an argument similar to the proof of Prop. 5, the minimum is obtained when k_0 is maximized. Hence, it is sufficient to show the maximum value of k_0, thus the maximum of $m_A(0)$.

Towards this, we first show that the distribution

$$m_A^*(t) = \begin{cases} 1 - \frac{1}{B\lambda}, & t = 0 \\ \frac{1}{B\lambda}, & t = B \end{cases}$$

maximizes $m_A(0)$ among all discrete distributions with bounded support $\{0, \dots, B\}$ and mean $1/\lambda$.

This can be proved by contradiction. Suppose there is another distribution m'_A with mean $1/\lambda$ and $m'_A(0) > m_A^*(0)$. Now

$$\begin{aligned} \mathbb{E}_{m'_A}[X] &= \sum_{t=0}^{B} t m'_A(t) \\ &\leq B \sum_{t=1}^{B} m'_A(t) = B(1 - m'_A(0)) \\ &< B(1 - m_A^*(0)) = \frac{1}{\lambda}, \end{aligned}$$

which contradicts the assumption that m'_A has expectation $1/\lambda$. Hence, there exists no m'_A on $\{0, \dots, B\}$ with $m'_A(0) > m_A^*(0)$.

Although the maximal k_0, k_0^*, is attained by $m_A^*(t)$, the induced Markov chain Q is not irreducible because $m_A^*(1) = 0$. Instead, we use the approximate probability mass function $\widetilde{m}_A(t)$, which has a nonzero mass at $t = 1$. Define

$$\widetilde{m}_A(t) = \begin{cases} 1 - \frac{1}{B\lambda} - \epsilon\left(1 - \frac{1}{B\lambda}\right), & t = 0 \\ \epsilon, & t = 1 \\ \frac{1}{B\lambda} - \frac{\epsilon}{B\lambda}, & t = B. \end{cases}$$

Then, we need to show $\widetilde{m}_A(t)$ approximates k_0 arbitrarily close to k_0^*. Note that $k_0 = \sum_t (m_A(0))^t P_S(t)$, which is a continuous function of $m_A(0)$. Thus, the conclusion follows.

Finally, the lower bound of capacity is computed as in the proof of Prop. 5,

$$C_L = \lambda\left(\log|\mathbb{F}| + (H(\psi_2) - H(\psi_0))(1 - \frac{\lambda}{\mu})\frac{1}{k_0} - H(\psi_2)\right),$$

where $k_0 = \sum_t (1 - \frac{1}{B\lambda})^t P_S(t)$. $\square$

Proposition 8. *For $H(\psi_0) = H(\psi_1) < H(\psi_2) = H(\psi_3) = \cdots$, for a given arrival rate λ and a service distribution P_S, the maximum capacity of the queue-channel over all Type II arrival processes with finite support $\{0, 1, \ldots, B\}$ is*

$$C_U = \lambda \left(\log |\mathbb{F}| + (H(\psi_2) - H(\psi_0))(1 - \frac{\lambda}{\mu})\frac{1}{k_0} - H(\psi_2) \right),$$

where $k_0 = \sum_t (1 - \frac{1}{\lambda})^t P_S(t)$, attained by the Bernoulli arrival process, i.e., $m_A(0) = 1 - 1/\lambda$ and $m_A(1) = 1/\lambda$.

Proof. By a similar argument as above, the maximum is obtained when k_0 is minimized. This is reached when $m_A(0)$ is minimum. Hence, it is sufficient to prove that among all discrete distributions, Bernoulli achieves it.

Again, the proof is by contradiction. Let us assume there is another distribution m'_A with the same mean, i.e., $\sum_t t m'_A(t) = 1/\lambda$ for which $m'_A(0) < m_A(0)$.

$$\begin{aligned} \sum_t t m'_A(t) &\geq \sum_{t \geq 1} m'_A(t) = (1 - m'_A(0)) \\ &> (1 - m_A(0)) = \frac{1}{\lambda}, \end{aligned}$$

which is a contradiction.

The capacity expression follows by substituting in the expression for k_0 for Bernoulli arrival. $\square$

This proposition implies that having at most one arrival per time slot is better. In other words, burstiness in the arrival process hurts performance.

2.5 Without Timing Information

So far, we have assumed that the received or processed jobs have timestamps on dispatch time and completion time. Though this assumption is valid in many wireless settings (MAC timestamps are part of the protocols) and crowdsourcing scenarios (e.g., Samasource maintains timestamps), this information may not always be available. In this section we study the setting where the decoder does not have knowledge of A^n, D^n.

Here, the decoder no longer observes (Y^n, A^n, D^n), but only observes Y^n. Using the information spectrum technique it immediately follows that the capacity is

$$C(P_A) = \lambda \sup_{\mathbb{P}(\mathbf{X})} \underline{\mathbf{I}}(\mathbf{X}; \mathbf{Y}).$$

The following theorem characterizes the capacity of the system based on the queue parameters and noise distributions. Proof follows similar steps as the proof of Thm. 1.

Theorem 5. *For a given arrival (dispatch) process distribution P_A (or m_A) with $\lambda < \mu < 1$ which follows the assumption in Sec. 2.1, there exists a distribution π such that $\pi(q) > 0$ for all $q \in \{0, 1, \ldots\}$ and $\mathbb{P}(Q_n) \to \pi$ as $n \to \infty$. The capacity of this queue-channel is $\lambda(\log |\mathbb{F}| - H(\sum_q \pi_q \psi_q))$, where $\pi_q \psi_q$ is a mixture of distributions $\{\psi_q\}$.*

Proof. First we show the converse. Using the standard information spectrum method,

$$\begin{aligned} \underline{\mathbf{I}}(\mathbf{X}; \mathbf{Y}) &\leq \overline{\mathbf{H}}(\mathbf{Y}) - \overline{\mathbf{H}}(\mathbf{Y}|\mathbf{X}) \\ &\leq \log |\mathbb{F}| - \overline{\mathbf{H}}(\mathbf{Y}|\mathbf{X}) \\ &= \log |\mathbb{F}| - \overline{\mathbf{H}}(\mathbf{Z}). \end{aligned}$$

Without timing information, note that $Z_i \sim \sum_q \mathbb{P}(Q_i = q)\psi_q$. Since $\Pr(Q_i) \to \pi$ and $\{Q_i\}$ is ergodic, by Lem. 1,

$$\begin{aligned} \underline{\mathbf{I}}(\mathbf{X}; \mathbf{Y}) &= \log |\mathbb{F}| - \overline{\mathbf{H}}(\mathbf{Z}) \\ &\to \log |\mathbb{F}| - H\left(\sum_q \pi_q \psi_q\right). \end{aligned}$$

For achievability, like the proof of Thm. 1 we pick a uniform and i.i.d. $\mathbb{P}(X^n) = \prod_{i=1}^n \mathbb{P}(X_i)$ and show that

$$\begin{aligned} \underline{\mathbf{I}}(\mathbf{X}; \mathbf{Y}) &\geq \underline{\mathbf{H}}(\mathbf{Y}) - \overline{\mathbf{H}}(\mathbf{Y}|\mathbf{X}) \\ &= \log |\mathbb{F}| - \overline{\mathbf{H}}(\mathbf{Z}) \\ &\to \log |\mathbb{F}| - H\left(\sum_q \pi_q \psi_q\right). \end{aligned}$$

Therefore multiplying by λ, we obtain the capacity expression in the theorem. □

Next, we consider some queuing systems to find the best dispatch and service processes in this setting. In the case of thresholded noise behavior the following result holds for a $\mathsf{G/geo/1}$ system.

Proposition 9. *Suppose* $\mathbb{F} = \{0,1\}$, $\mathbb{P}(Z_i = 1|q) \leq 0.5$ *for all* q *and* $H(\psi_0) = \cdots = H(\psi_b) < H(\psi_{b+1}) = \cdots$ *for some* $b \in \{0, 1, \ldots\}$. *Then, for a* $\mathsf{G/geo/1}$ *system with no timestamps, the queue-channel capacity is maximized by deterministic inter-arrival (for* $\frac{1}{\lambda} \in \mathbb{Z}_+$*) and is minimized by geometric inter-arrival among the class of sum-of-geometric random variables.*

Proof. In this system

$$H\left(\sum_q \pi_q \psi_q\right) = H\left(\psi_0 \sum_{q \leq b} \pi_q + \psi_{b+1} \sum_{q \geq b+1} \pi_q\right).$$

Note that $\pi_q = (1-\sigma)\sigma^q$ where σ is the fixed-point solution in Thm. 2. Hence,

$$H\left(\sum_q \pi_q \psi_q\right) = H\left(\psi_0(1-\sigma^{b+1}) + \psi_{b+1}\sigma^{b+1}\right).$$

Now, as $\mathbb{P}(Z_i = 1|q) \leq 0.5$, by monotonicity of binary entropy over $[0, 0.5]$, it follows that the above expression is maximized (minimized) when $\sum_{q \geq b+1} \pi_q$ is maximized (minimized), which in turn happens when σ is maximized (minimized).

The remainder of the argument follows as in the proof of Prop. 2 and 4, because for $\mathsf{G/geo/1}$, deterministic arrival minimizes σ, while geometric arrival maximizes among the class of sum-of-geometric random variables. □

Proposition 10. *Suppose* $\mathbb{F} = \{0,1\}$, $\mathbb{P}(Z_i = 1|q) \leq 0.5$ *for all* q *and* $H(\psi_0) = \cdots = H(\psi_b) < H(\psi_{b+1}) = \cdots$ *for some* $b \in \{0, 1, \ldots\}$. *Then, for a* $\mathsf{geo/G/1}$ *system with no timestamps, the queue-channel capacity is maximized by a deterministic service time and (for* $\frac{1}{\lambda} \in \mathbb{Z}_+$*) and is minimized by geometric service time among the class of sum-of-geometric random variables.*

Proof. By the same argument as in proof of Prop. 9, the maximum is achieved when $\pi_0 + \pi_1$ is maximized. The remaining argument follows the proof of Prop. 5. □

2.6 Chapter Summary

In this chapter, we consider a queue-length-dependent channel, where service quality depends on the queue-length of jobs. We define the capacity of such queuing systems to be the maximum rate at which jobs can be processed with arbitrarily small error probability, and characterize it in terms of queuing parameters. It has several engineering applications including crowdsourcing, multimedia communication.

We study Type I and Type II arrivals separately for analytic tractability. In Type I arrivals, for a $\mathsf{G}/\mathsf{geo}/1$ queue with a step-increasing noise, jobs arriving deterministically maximize capacity while bursty arrivals minimize capacity. Similarly, for a $\mathsf{geo}/\mathsf{G}/1$ queue with a step-increasing noise, deterministic service maximizes capacity, but bursty service minimizes capacity. Type II arrivals give similar results except that Bernoulli arrivals maximize capacity for the $\mathsf{G}/\mathsf{geo}/1$ queue.

Chapter 3

On Multiuser Systems with Queue-Length-Dependent Service Quality

In the previous chapter, we studied the capacity of single-user queue-length-dependent quality in discrete time and further optimized the server of a $\mathsf{geo}/\mathsf{GI}/1$ queue or the dispatcher of a $\mathsf{GI}/\mathsf{geo}/1$ queue, under given reliability assumptions.

Beyond the formulation in the previous chapter, there are often multiple input streams in the motivating applications rather than just one, e.g., due to *multihoming*, so here we consider a scenario where multiple transmitter-destination pairs want to send information reliably and therefore dispatch coded symbols on arrival processes. A particular motivational setting is driver-assisted autonomous trucks [74], where a human driver remotely monitors multiple semi-autonomous trucks and steps in (i.e., processes information) only when the autonomous algorithm cannot handle the task.

Fig. 3.1 presents such a multiple-access setting, where before entering a single central processor, the multiple arrival processes are superposed. Once coded symbols arrive at the central queue processor, they are served in a first-in first-out (FIFO) manner, and returned to the intended receiver. Note that if there is a single central receiver, the topology reduces to a multiple-access channel. As before, a distinguishing aspect of this chapter is that reliability of the central server depends on queue-length arising from the superposed arrival process.

Here, we consider the superposition of multiple arrival processes in a continuous-time

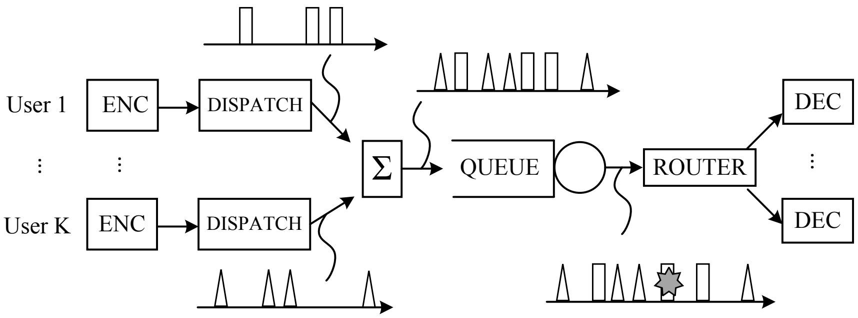


Figure 3.1: Block diagram of the system: only two out of K point processes are illustrated for brevity. We use Σ to denote superposition operation.

setting. Before proceeding, we first study the capacity of the continuous-time single-user case, and also specify the best and worst dispatch processes for a $\mathsf{GI/M}/1$ queue, and service processes for a $\mathsf{M/GI}/1$ queue with additional conditions. Then, the capacity expression of the multiple-access setting is given in terms of the stationary distribution of queue-length seen by each user's departures. Surprisingly, our results show there is no loss in capacity due to multiple-access interference.

As superposition of non-Poisson arrivals is in general intractable, we also consider the large-user asymptotic by introducing a random marked point process (RMPP, or simply PP) approach [75, 76] and apply the superposition convergence to a Poisson point process [77]. The latter states the superposition of a large number of *sparse* arrivals is approximately Poisson. Building on this result, we prove that the capacity for $\sum_k \mathsf{GI}_k/\mathsf{GI}/1$ queues, where Σ_k stands for the superposition, converges to that for single-user $\mathsf{M/GI}/1$ queues. In other words, even though individuals are non-Poisson arrivals, sending information as if a single-user $\mathsf{M/GI}/1$ queue is asymptotically optimal. It also implies the best and worst services obtained for a single-user $\mathsf{M/GI}/1$ queue are preserved.

3.1 Preliminaries and System Model

3.1.1 Point Process

We use a PP approach to queueing systems, enabling us to derive analytical properties. Let us define an RMPP $\Phi = \Phi(t)$ as follows.

Definition 4. *Let $\mathfrak{B}$ be the Borel σ-algebra of $\mathbb{R}$. Given a mark space $\mathcal{M}$ and its sigma-algebra $\sigma(\mathcal{M})$, consider a marked counting measure $N(B \times M)$ where $B \in \mathfrak{B}$ and $M \in \sigma(\mathcal{M})$ such that $N(B \times \mathcal{M}) < \infty$ for any bounded B. Let $\mathcal{N}, \sigma(\mathcal{N})$ be the set of all such counting measures and its smallest σ-algebra, respectively. Then, a random marked point process (RMPP, or simply a point process (PP)), $\Phi(t)$ is a random element from $(\Omega, \mathcal{F}, P)$ to $(\mathcal{N}, \sigma(\mathcal{N}))$.*

For queueing applications, the mark usually denotes a random service time at the server or the time required to finish each job. Hence, $\mathcal{M} = \mathbb{R}_+$ and since only the $\cdot/\mathsf{GI}/1$ queue is considered in this work, each mark is i.i.d. from some distribution P^S. Since all randomness from arrival and service times is captured in the RMPP, any queue response such as queue-length or waiting time is a deterministic function of the RMPP.

Two equivalent representations of a PP are especially useful in this chapter. Suppose the mark space is empty, i.e., $\mathcal{M} = \emptyset$ for illustration. However, the following representations can be easily extended to RMPPs with a non-empty mark space. The first representation is to use an inter-arrival time representation, induced by Dirac delta functions.

Letting $\{T_i \in \mathbb{R}_+\}_{i\in\mathbb{Z}}$ be a non-decreasing random sequence,

$$\Phi(t) \Leftrightarrow \sum_{i=-\infty}^{\infty} \delta_{T_i} \Leftrightarrow (\ldots, A_{-1}, A_0, A_1, \ldots,),$$

where $A_i := T_i - T_{i-1} \geq 0$. So T_i indicates the time epoch when the ith arrival comes. The case for i.i.d. A_i is called a *renewal* process, which arises in Sec. 3.2.

The other representation is by a random counting measure, which is useful especially in Sec. 3.3. Note that

$$N(B) = \int \sum_{i=-\infty}^{\infty} \mathbf{1}_B \delta_{T_i} dt,$$

that is, the number of arrivals in B, for any bounded $B \in \mathfrak{B}$, uniquely determines $\Phi(t)$. Here $\mathbf{1}_B = \mathbf{1}_B(t)$ is the indicator function with criterion $\{t \in B\}$ and we write $\mathbf{1}_B\Phi$ to stand for the restricted RMPP on B.

A time shift operation is denoted by $T_\tau\Phi(t) = \Phi(t+\tau)$, enabling definitions of stationarity and ergodicity.

Definition 5 (Stationarity, Def. 1.2.1 [75])**.** *An RMPP Φ is* stationary *if the probability measure P is invariant with respect to the time shift T_τ, i.e., for any set $Z \in \sigma(\mathcal{N})$,*

$$P(T_\tau Z) = P(Z) \text{ for all } \tau \in \mathbb{R}.$$

Definition 6 (Ergodicity, Def. 1.2.5 [75])**.** *A stationary RMPP Φ (or its probability measure P) is* ergodic *if any set $Z \in \sigma(\mathcal{N})$ satisfying $T_\tau Z = Z$ for all $\tau \in \mathbb{R}$ implies either $P(Z) = 0$ or 1.*

3.1.2 System Model

Multiple users intend to send messages to respective targeted receivers. To do that, the kth user picks an encoded sequence of symbols[1] $X_{(k)}^n$—each symbol is drawn from finite space $\mathcal{X}$—and dispatches it over an independent stationary renewal arrival process with inter-arrival time distribution $P_k^A(t)$. Those arrivals are superposed just before entering a $\cdot/\mathsf{GI}/1$ queue. The server follows FCFS service discipline with i.i.d. service time according to P^S. Assume that the waiting room is unlimited.

Since the server is unreliable, the symbol is corrupted to $Y_{(k)}^n \in \mathcal{Y}^n$ randomly, where $\mathcal{Y}$ is also finite. The transition probability, denoted by $W = W_Q$, is dependent on Q, the queue-length at the moment just before the symbol's departure, excluding the job being serviced. That is, the channel at time t is $W_Q := P_{Y|X,Q}$, where Q is the queue-length seen by departure. In this sense we say the system is *queue-length-dependent*. Departing symbols are labeled and delivered to the intended receiver. Since symbols are encoded against channel noise, receivers can decode the sequence to recover the original information. We assume there is a central coordination mechanism that reveals each transmitter's dispatching process to all other transmitters, but not realizations.

We use $\sum(\cdot)$ to denote superposition, so the queue of interest is written as $\sum_k \mathsf{GI}_k/\mathsf{GI}/1$. The queues are assumed always stable, i.e., superposed arrival rate λ and service rate μ satisfy traffic intensity $\rho := \frac{\lambda}{\mu} < 1$. Also we suppose some technical assumptions on arrivals and service: 1) arrivals and service processes are *simple*, i.e., $P_k^A(0) = 0$ for all k, $P^S(0) = 0$; 2) at least one of $\{P_k^A(t)\}_{k=1}^K$ and $P^S(t)$ is continuous and strictly positive on $\mathbb{R}$.

We assume causal knowledge of arrival and departure realizations, i.e., the encoders do not know them, but the decoders do. Also all P_k^A are available to transmitters, but not their realizations.

3.2 Continuous-time Single-user Queue-channel

This section investigates the capacity of single-user queue-length-dependent channels like [78], but in continuous-time.

[1] Throughout this chapter, *symbol* (common in information theory) and *job* (or customer, common in queueing theory) are interchangeable.

3.2.1 Coding Theorem for $\mathsf{GI}/\mathsf{GI}/1$ Queues

Consider a simple renewal arrival process $\Phi(t)$ with arrival rate λ, i.e., the ith inter-arrival time $A_i \sim P^A$ i.i.d. with $\lambda = 1/\mathbb{E}[A_1]$. Recall that the service quality (channel) of the ith job depends only on the queue-length seen by the ith departure (i.e., just before ith departure), denoted Q_i. We first express capacity using the information spectrum [68, 69]; see [68, 69] for notation of various information functionals.

Proposition 11. *For a simple renewal PP $\Phi(t)$ with rate $\lambda = 1/\mathbb{E}[A_1]$,*

$$\begin{aligned} C(\Phi) &= \sup_{P(\mathbf{X})} \underline{\mathbf{I}}(\mathbf{X};\mathbf{Y}|\mathbf{Q}) \quad \textit{[bits/sym]} \qquad (3.1)\\ &= \sup_{P(\mathbf{X})} \lambda\underline{\mathbf{I}}(\mathbf{X};\mathbf{Y}|\mathbf{Q}) \quad \textit{[bits/time].} \end{aligned}$$

Proof. See [78, Prop. 1]. □

Lemma 3. *For each simple renewal PP Φ, there exists a unique stationary distribution π such that if $Q_1 \sim \pi$, then any $Q_i \sim \pi$. Furthermore, for any measurable $f : \mathbb{Z}_+ \mapsto \mathbb{R}_+$, $\frac{1}{n}\sum_{i=1}^{n} f(Q_i) \to \mathbb{E}_\pi[f(Q)]$ as $n \to \infty$ almost surely.*

Proof. Consider an arrival time instance when the system is empty, i.e., no job in the queue, no job in the server at the instance of an arrival. At this instance, a new cycle of queueing begins from the empty state. So let us consider the queue-length process seen by arrivals, $\{\widehat{Q}_i\}_{i\in\mathbb{Z}}$. In $\mathsf{GI}/\mathsf{GI}/1$ queues, the cycles are i.i.d. and so are called *regenerative* cycles [79, Chap. VI], denoted by $\{R_i \in \mathbb{Z}_+\}_{i\in\mathbb{Z}}$. Also, $\rho < 1$ implies $\mathbb{E}[R] < \infty$ and these cycles are repeated infinitely many times. We know the limiting distribution of $\widehat{Q}$, say $\widehat{\pi}$, exists and is ergodic so for any measurable nonnegative function f,

$$\mathbb{E}_{\widehat{\pi}}[f(\widehat{Q})] = \frac{1}{\mathbb{E}[R]}\mathbb{E}\left[\sum_{i:\text{inside of } R} f(\widehat{Q}_i)\right] = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} f(\widehat{Q}_i).$$

Next, suppose the queue is in steady-state. Since the beginning and end of cycles are empty-state, whenever there is an arrival, there is a corresponding departure in the cycle.

Thus, $\widehat{Q} \overset{d}{=} Q$, i.e., $\widehat{\pi} = \pi$. Therefore, for any measurable nonnegative function f,

$$\begin{aligned}\mathbb{E}_{\widehat{\pi}}[f(\widehat{Q})] &= \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} f(\widehat{Q}_i) \\ &= \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} f(Q_i) = \mathbb{E}_{\pi}[f(Q)].\end{aligned}$$

This completes the proof. □

Combining Prop. 11 and Lem. 3, we have a simpler capacity expression in terms of expectation over Q, or equivalently in terms of stationary distribution $\pi(Q)$.

Theorem 6. *For* $\mathsf{GI}/\mathsf{GI}/1$ *queues, the capacity formula* (3.1) *can be further simplified to*

$$C(\Phi) = \sup_{P_X} \mathbb{E}\left[I(P_X, W_Q)\right] = \sup_{P_X} \sum_{q=0}^{\infty} \pi(q) I(P_X, W_q) \tag{3.2}$$

in bits per job, and

$$C(\Phi) = \lambda \sup_{P_X} \mathbb{E}\left[I(P_X, W_Q)\right] = \sup_{P_X} \lambda \sum_{q=0}^{\infty} \pi(q) I(P_X, W_q) \tag{3.3}$$

in bits per time. Therefore, it is easy to see that the capacity over all renewal PPs with stable arrival rate $\lambda < \mu$ *is*

$$C = \sup_{\lambda\in(0,\mu)} \sup_{P^A} \sup_{P_X} \lambda \mathbb{E}\left[I(P_X, W_Q)\right] \quad \textit{[bits/time]}.$$

Proof. See [78, Thm. 1] with generalization to general discrete channels. □

Remark 1. *In this work, we assume a simple transmitter that does not know arrival and departure realizations, which implies channel state information is unavailable. If the channel state information is available without delay, the capacity formula follows immediately as*

$$C(\Phi) = \lambda \mathbb{E}\left[\sup_{P_X} I(P_X, W_Q)\right] \quad \textit{[bits/time]}. \tag{3.4}$$

Thus, we can see that when the capacity-achieving distributions are all identical with some P_X^**, such as binary symmetric channels or binary erasure channels, the transmitter simply picks* P_X^* *even without the channel state information and achieves* (3.4) *by the codebook* *identical with no channel state information. Channel state feedback even without delay does not improve capacity in this case.*

A closed-form expression of $\pi(Q)$ is unknown in general, but is known for some special types of queues. Let us rewrite (3.2) for two special types of queues $\mathsf{GI/M/1}$ and $\mathsf{M/GI/1}$, and consider per symbol capacity since per time capacity follows by multiplying by λ.

Theorem 7 ($\mathsf{GI/M/1}$ queues)**.** *Let $A^*(\cdot)$ be the Laplace-Stieltjes transform of $P^A(t)$ and define σ^* as the unique solution of $\sigma = A^*(\mu(1-\sigma))$ in $(0,1)$. Then, the capacity of $\mathsf{GI/M/1}$ queues is given by*

$$C(\Phi) = \sup_{P_X} \mathbb{E}[I(P_X, W_Q)] \quad \textit{[bits/sym]},$$

where $\pi(q) = (1-\sigma^*)(\sigma^*)^q$.

Proof. See Appendix B.1. □

Theorem 8 ($\mathsf{M/GI/1}$ queues)**.** *The capacity of $\mathsf{M/GI/1}$ queues is given by*

$$C(\Phi) = \sup_{P_X} \mathbb{E}[I(P_X, W_Q)] \quad \textit{[bits/sym]},$$

where $\pi(q)$ is obtained from the inverse of probability generating function

$$\Pi(z) = \frac{(1-\rho)(1-z)K(z)}{K(z)-z},$$

and $K(z)$ is the probability generating function of k_q with

$$k_q = \int_0^\infty P^S(t) \frac{e^{-\lambda t}(\lambda t)^q}{q!} dt.$$

Proof. See Appendix B.2. □

Example: Consider an $\mathsf{M/M/1}$ queue and a binary symmetric channel (corresponding to binary classification) with queue-length-dependent transition probability ϵ_q. Then, we know that $\pi(q) = (1-\rho)\rho^q$ and Thm. 7 shows that

$$C = \lambda \sum_{q=0}^{\infty} \pi(q)(1 - H_2(\epsilon_q)) \quad \text{[bits/time]},$$

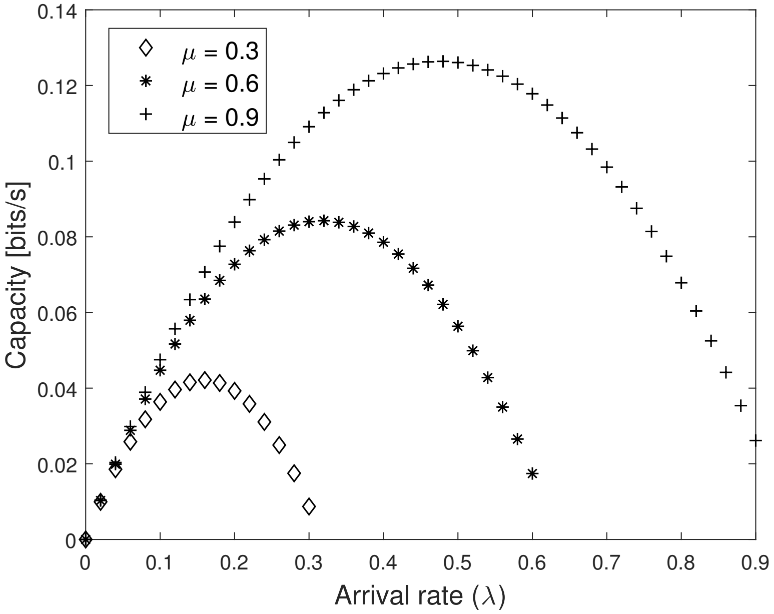


Figure 3.2: Capacity of $\mathsf{M}/\mathsf{M}/1$ queue (for different service rates) with binary symmetric channel is plotted. $\mathbb{P}[X \neq Y] = 0.1$ for $q = 0$, $\mathbb{P}[X \neq Y] = 0.4$ otherwise. It shows that setting a proper workload maximizes per time capacity.

where $H_2(\cdot)$ is the binary entropy function. Fig. 3.2 shows the capacity curves for different rates.

3.2.2 Optimization of Capacity

This subsection considers optimization of the capacities for $\mathsf{GI}/\mathsf{M}/1$ and $\mathsf{M}/\mathsf{GI}/1$ queues given in Thms. 7 and 8. To do so, we impose two conditions such that

1. P_X^* achieves the capacity for all W_q.
2. At such P_X^*, the system becomes more unreliable as q increases in a step-down manner, i.e., for some $b \in \mathbb{Z}_+$,

$$I(P_X^*, W_0) = \cdots = I(P_X^*, W_b) > I(P_X^*, W_{b+1}) = \cdots .$$

Note that condition 1 covers $|\mathbb{F}|$-ary symmetric or $|\mathbb{F}|$-ary erasure channels since P_X^* is uniform. Such channels model multi-label classification via crowdsourcing platform [12] in that events $\{X \neq Y\}$ in a symmetric channel and $\{Y = \text{erasure}\}$ in an erasure channel model 'misclassification' and 'I don't know' answers of a crowdworker, respectively. In particular, introducing the step change in noise allows us to find the best and worst server behaviors explicitly. It is natural in applications (including non-human applications) for the server to be more unreliable as the queue gets longer; see [78] for modeling details.

Corollary 4 (GI/M/1 queue)**.** *Fix arrival rate λ. For* GI/M/1 *queues, the best inter-arrival distribution is deterministic, i.e., $P^A(t)$ only has a unit point mass at $t = \lambda^{-1}$.*

Proof. For the sake of brevity, let $c_b := I(P_X^*, W_b)$ and $c_{b+1} := I(P_X^*, W_{b+1})$. Then, the capacity is written as

$$\begin{aligned} C(\Phi) &= \sum_{q=0}^{\infty} \pi(q) I(P_X^*, W_q) \\ &= \sum_{q=0}^{\infty} (1-\sigma^*)(\sigma^*)^q I(P_X^*, W_q) \\ &= \sum_{q=0}^{b} (1-\sigma^*)(\sigma^*)^q c_b + \sum_{q=b+1}^{\infty} (1-\sigma^*)(\sigma^*)^q c_{b+1} \\ &= c_b(1-(\sigma^*)^{b+1}) + c_{b+1}(\sigma^*)^{b+1} \\ &= c_b - (\sigma^*)^{b+1}(c_b - c_{b+1}). \end{aligned}$$

As $c_b > c_{b+1}$, maximizing $C(\Phi)$ with given λ is equivalent to minimizing σ^*. Note that σ^* is the unique fixed point of $\sigma = A^*(\mu(1-\sigma))$ and at $\sigma = 0$ and 1,

$$\begin{aligned} \int_0^{\infty} P^A(t) e^{-\mu t(1-\sigma)} dt \Big|_{\sigma=0} &= \int_0^{\infty} P^A(t) e^{-\mu t} dt > 0 \\ \int_0^{\infty} P^A(t) e^{-\mu t(1-\sigma)} dt \Big|_{\sigma=1} &= \int_0^{\infty} P^A(t) dt = 1. \end{aligned}$$

Furthermore, $A^*(\mu(1-\sigma))$ is strictly convex in σ since

$$\frac{\partial}{\partial \sigma} A^*(\mu(1-\sigma)) > 0, \quad \frac{\partial^2}{\partial^2 \sigma} A^*(\mu(1-\sigma)) > 0.$$

Due to Jensen's inequality, we obtain

$$\begin{aligned} A^*(\mu(1-\sigma)) &= \int_0^{\infty} P^A(t) e^{-\mu t(1-\sigma)} dt \\ &\geq e^{-\mu \mathbb{E}[A](1-\sigma)} = e^{-\frac{\mu}{\lambda}(1-\sigma)}, \end{aligned} \tag{3.5}$$

where the equality is attained only when $A = \lambda^{-1}$ almost surely. It means that when P^A is deterministic, the curve $A^*(\mu(1-\sigma)) = e^{-\frac{\mu}{\lambda}(1-\sigma)}$ lower bounds all other curves so that achieves the smallest fixed point. Therefore, the deterministic inter-arrival distribution achieves the greatest capacity. □

Corollary 5 (GI/M/1 queue)**.** *Fix arrival rate λ. For* GI/M/1 *queues, cramming inter-arrivals asymptotically minimize the capacity, i.e., $P^A(t;\epsilon,\delta)$ asymptotically achieves the smallest capacity as $\epsilon,\delta \to 0$, where*

$$P^A(t;\epsilon,\delta) = \begin{cases} 1-\epsilon & \textit{if } t = \delta \\ \epsilon & \textit{if } t = \frac{\frac{1}{\lambda}-\delta(1-\epsilon)}{\epsilon} \\ 0 & \textit{otherwise.} \end{cases}$$

Proof. Similar to the proof of Cor. 4, it is sufficient to show that σ^* is maximized, i.e., when P^A is cramming $A^*(\mu(1-\sigma))$ upper bounds all other curves. We know that for any P^A,

$$\begin{aligned} A^*(\mu(1-\sigma)) &= \int_0^\infty P^A(t)e^{-\mu t(1-\sigma)}dt \\ &\leq \int_0^\infty P^A(t)dt = 1. \end{aligned} \tag{3.6}$$

On the other hand, note that the cramming inter-arrival distribution asymptotically achieves the upper bound as $\epsilon,\delta \to 0$ so that it maximizes the fixed point solution σ^*. Also notice that the location of ϵ point mass is determined to satisfy mean constraint $\mathbb{E}[A] = \lambda^{-1}$. □

Corollary 6 (M/GI/1 queue)**.** *Fix service rate μ. For* M/GI/1 *queues with service quality stepping down at $b = 0$, i.e.,*

$$I(P_X^*, W_0) > I(P_X^*, W_1) = I(P_X^*, W_2) = \cdots,$$

the capacity is constant among all service distributions.

Proof. When the threshold $b = 0$, let $c_0 := I(P_X^*, W_0)$ and $c_1 := I(P_X^*, W_1)$. Since the capacity is given by

$$C(\Phi) = \pi(0)c_0 + (1-\pi(0))c_1 = c_1 + \pi(0)(c_0 - c_1),$$

so $\pi(0)$ completely determines the capacity. On the other hand, by the inverse Z-transform relation,

$$\pi(0) = \Pi(0) = 1 - \rho.$$

Thus, the capacity is constant over all P^S of service rate μ. $\square$

Corollary 7 (M/GI/1 queue)**.** *Fix service rate μ. For* M/GI/1 *queues with service quality stepping down at $b = 1$, i.e.,*

$$I(P_X^*, W_0) = I(P_X^*, W_1) > I(P_X^*, W_2) = \cdots ,$$

the capacity is maximized when the service is deterministic. On the other hand, the capacity is asymptotically minimized by cramming service, i.e., $P^S(t; \epsilon, \delta)$ asymptotically minimizes the capacity as $\epsilon, \delta \to 0$, where

$$P^S(t; \epsilon, \delta) = \begin{cases} 1 - \epsilon & \textit{if } t = \delta \\ \epsilon & \textit{if } t = \frac{\frac{1}{\mu} - (1-\epsilon)\delta}{\epsilon} \\ 0 & \textit{otherwise.} \end{cases}$$

Proof. Let $c_0 := I(P_X^*, W_0), c_2 := I(P_X^*, W_2)$ for simplicity. Then the capacity is given by

$$C = (\pi(0) + \pi(1))c_0 + (1 - \pi(0) - \pi(1))c_2.$$

Since $c_0 > c_2$, it is apparent that the capacity is maximized (resp. minimized) when $\pi(0) + \pi(1)$ is maximized (resp. minimized). Also note that

$$\begin{aligned} \pi(0) &= 1 - \rho = 1 - \frac{\lambda}{\mu}, \\ \pi(1) &= \frac{\Pi(z) - \pi(0)}{z}\bigg|_{z=0} = \frac{\frac{(1-\rho)(1-z)K(z)}{K(z)-z} - \pi(0)}{z}\Bigg|_{z=0} \\ &= \frac{\frac{(1-\rho)(1-z)K(z)}{K(z)-z} - (1-\rho)}{z}\Bigg|_{z=0} \\ &= \frac{(1-\rho)(1-K(z))}{K(z) - z}\bigg|_{z=0} = \frac{(1-\rho)(1-K(0))}{K(0)}. \end{aligned}$$

Since $\pi(0) + \pi(1) = \frac{1-\rho}{K(0)}$, the best (resp. the worst) service distribution should minimize (resp. maximize) $K(0) = k_0$. Recall the expression of k_0,

$$k_0 = \int_0^\infty P^S(t) e^{-\lambda t} dt.$$

The same arguments of (3.5) and (3.6) imply that the deterministic service distribution $P^S(t) = \delta_{\mu^{-1}}$ maximizes the capacity, and

$$P^S(t; \epsilon, \delta) = \begin{cases} 1-\epsilon & \text{if } t = \delta \\ \epsilon & \text{if } t = \frac{\frac{1}{\mu}-(1-\epsilon)\delta}{\epsilon} \\ 0 & \text{otherwise} \end{cases}$$

asymptotically minimizes the capacity as $\epsilon, \delta \to 0$. □

Cor. 7 is also of interest when the number of users is large and each arrival process is sparse, see Sec. 3.3.

3.3 Multiuser Input: $\sum_k \mathsf{GI}_k/\mathsf{GI}/1$ Queues

Recall the system model in Sec. 3.1.2. Since K users simultaneously dispatch encoded symbols, each user sees a different queue-length distribution from that for single-user systems; thus, capacity changes. We characterize the individual and sum capacities for the K-user scenario in terms of $\pi_{Kk}(Q)$, the stationary queue-length distribution seen by user k's departures. Note that K, k denote total number of users and a specific kth user, respectively. Since the superposition process is in general intractable, we obtain asymptotics of capacity using Poisson approximation when component PPs are independent and sparse.

For a common setup, consider a triangular array of independent, stationary, and renewal (thus, ergodic) PPs $\Phi_{Kk}, K \in \mathbb{Z}_+, k \in [1:K]$. Also suppose each PP has an inter-arrival distribution P^A_{Kk} with arrival rate λ_{Kk}, not necessarily identical. Let us also assume second-moment finiteness of inter-arrival times, which is necessary to prove Lem. 6:

$$\mathbb{E}_{P^A_{Kk}}[A^2] < \infty \text{ for all } k \in [1:K]. \tag{3.7}$$

3.3.1 Coding Theorem for K-user Channels

Let Φ_K be the superposition arrival process of Kth-row components, i.e, $\Phi_K := \sum_{k=1}^{K} \Phi_{Kk}$. Note that the component PPs are stationary and ergodic.

The next lemma proves the superposition process is stationary and ergodic as well.

Lemma 4. *Suppose each $\Phi_{Kk}, k \in [1 : K]$ is independent, stationary, and ergodic. Then, Φ_K is also stationary and ergodic.*

Proof. First prove the stationarity. Take an arbitrary bounded Borel set B and let $B' = T_\tau B$ be the time-shifted set by $\tau \in \mathbb{R}$. Consider the counting measure representation; then due to independence, $N_K(B) = \sum_k N_{Kk}(B)$ and

$$N_K(B) = \sum_k N_{Kk}(B) \overset{(a)}{=} \sum_k N_{Kk}(B') \overset{(b)}{=} N_K(B'),$$

where (a) is due to the stationarity of individual PPs and (b) is due to independence of individual PPs. As $\tau \in \mathbb{R}$ is arbitrary, stationarity is shown.

Next show the ergodicity. Suppose Φ_K is not ergodic: then, by Def. 6, there exists a $Z \in \sigma(\mathcal{N})$ such that for any $\phi_K \in Z$ and $\tau \in \mathbb{R}$, it holds that $T_\tau \phi_K \in Z$, however, $0 < P_K[Z] < 1$. As Z is closed under any time-shift operation, we can write for $\phi_K \in Z$,

$$\begin{aligned} \phi_K &= \left(\sum_k \phi_{Kk} \right) \in Z \Leftrightarrow \\ \phi'_K &:= T_\tau \phi_K = T_\tau \sum_k \phi_{Kk} = \left(\sum_k T_\tau \phi_{Kk} \right) \in Z \ \forall \tau \in \mathbb{R}. \end{aligned} \tag{3.8}$$

Now consider $P_K[Z]$. Let Z_k be the collection of ϕ_{Kk} consisting some $\phi \in Z$. As ϕ_{Kk} is a component of ϕ_K, $T_\tau \phi_{Kk}$ is also a component of ϕ'_K by (3.8) so that Z_k is also closed. Since each Φ_{Kk} is stationary and ergodic, $P_{Kk}[Z_k]$ is either 0 or 1. However, because $0 < P_K[Z] = \prod_k P_{Kk}[Z_k] < 1$ by independence, there is a contradiction. Therefore, Φ_K is ergodic. □

Let $Q_i^{(K)}$ be the queue-length process seen by the superposed departures. The next lemma further guarantees that the stationary distribution π_K exists and $Q_i^{(K)}$ is ergodic since Φ_K is stationary and ergodic from Lem. 4.

Lemma 5 (Sec. 2.2 [75])**.** *If the input PP Φ of the queue $\cdot/\mathsf{GI}/1$ with traffic intensity $\rho < 1$ is stationary and ergodic, then the queue-length distribution seen by departures is also stationary and ergodic. Furthermore, the stationary distribution is independent of the initial state.*

Now let us consider individual 'seen by departures' processes. Let $Q_i^{(Kk)}, \pi_{Kk}$ be the queue-length process seen by user k's departures and its stationary distribution. The following lemma proves the existence of π_{Kk} and its ergodicity.

Lemma 6. *Suppose* (3.7) *holds. Then, for each $k \in [1 : K]$, the stationary distribution π_{Kk} exists. Furthermore, for any measurable $f : \mathbb{Z}_+ \mapsto \mathbb{R}_+$, $\frac{1}{n}\sum_{i=1}^n f(Q_i^{(Kk)}) \to \mathbb{E}_{\pi_{Kk}}[f(Q)]$ as $n \to \infty$ almost surely.*

Proof. See Appendix B.3. □

As before, Lem. 6 allows a simpler capacity expression. Let $C_{\mathsf{ind}}(\Phi_{Kk}), C_{\mathsf{sum}}(\Phi_K)$ be the kth user's individual capacity and their sum capacity. The following theorem only describes per job capacity, but per time capacity is immediate by multiplying by individual and sum arrival rates, respectively.

Theorem 9.

$$C_{\mathit{ind}}(\Phi_{Kk}) = \mathbb{E}_{\pi_{Kk}}[I(P_X, W_Q)] \quad \textit{[bits/sym]},$$
$$C_{\mathit{sum}}(\Phi_K) = \mathbb{E}_{\pi_K}[I(P_X, W_Q)] = \sum_{k=1}^{K} w_k C_{\mathit{ind}}(\Phi_{Kk}) \quad \textit{[bits/sym]},$$

where $w_k := \lambda_{Kk}/\sum_j \lambda_{Kj}$.

Proof. Since individual $\{\pi_{Kk}\}$ are stationary and ergodic, the first statement follows.

To show the second statement, notice that

$$C_{\mathsf{sum}}(\Phi_K) \le \mathbb{E}_{\pi_K}[I(P_X, W_Q)]$$

holds. In addition, since π_K is the weighted average of π_{Kk}, i.e., $\pi_K(q) = \sum_k w_k \pi_{Kk}(q)$, the equality holds. □

Unlike typical multiple-access settings, it is interesting to note that the per time sum capacity is simply a sum of per time individual capacities, which means that greedy individuals do not degrade optimality in sum information rate. This follows since once arrival processes are fixed, symbol noise levels are also fixed by queue-length. The server processes one symbol at a time; therefore, adding more (or reducing) information in a user's codeword does not increase (or decrease) interference levels.

3.4 Poisson Approximation

In the previous subsection, we obtained the multiple-access capacity formula for general $\sum_k \mathsf{GI/GI/1}$ queues. However, a more explicit expression is unavailable even for an $|\mathbb{F}|$-ary symmetric channel or an erasure channel, unless the queue is $\sum_k \mathsf{M/GI/1}$. This is because the superposition of K independent renewal PPs is not necessarily renewal and is renewal if and only if individual PPs are Poisson [80] (thus, the superposition process is also Poisson). So the tractability of the superposition process is limited. Although it is intractable, when K is large and individual PPs are *sparse* (formally defined in Def. 7 below) we can approximate the superposition process by a Poisson PP.

Consider a triangular array of i.i.d., stationary, ergodic, and renewal PPs, $\{\Phi_{Kk}\}$, where $K \in \mathbb{Z}_+$ and $k \in [1:K]$. Individual processes are assumed to be sparse as given below. The superposition process of row PPs is denoted by $\Phi_K := \sum_k \Phi_{Kk}$ with corresponding probability measure P_K. Let $N_{Kk}(B)$ be the counting measure corresponding to Φ_{Kk}, i.e., the number of events of $\Phi_{Kk}(t)$ in $B \in \mathfrak{B}$. Also let $N_K(B)$ be the number of events of Φ_K in B, so $N_K(B) = \sum_k N_{Kk}(B)$. Then we can derive that $N_K(B)$ converges to the Poisson distribution of intensity measure $\lambda|B|$ where $|\cdot|$ is the Lebesgue measure, or equivalently, $\Phi_K(t)$ converges to the Poisson process, say $\Phi^*(t)$, with probability measure P^*, under the sparsity condition. Let N^* be the counting measure for the Poisson PP, i.e., for any bounded $B \in \mathfrak{B}$,

$$\mathbb{P}[N^*(B) = j] = \frac{1}{j!}(\lambda|B|)^j e^{-\lambda|B|}.$$

Definition 7. *For a given bounded $B \in \mathfrak{B}$, the triangular processes are said to be sparse with sum rate $\lambda_K := \sum_k \lambda_{Kk} + \frac{g_1(K,B)}{|B|}$ if*

$$\lambda_{Kk} := \frac{\mathbb{P}[N_{Kk}(B) = 1]}{|B|}, \tag{3.9}$$

$$g_1(K,B) := \sum_{k=1}^{K}\sum_{j=2}^{\infty} j\mathbb{P}[N_{Kk}(B) = j] \to 0 \text{ as } K \to \infty,$$

$$g_2(K) := \max_{k\in[1:K]} \lambda_{Kk} \to 0 \text{ as } K \to \infty.$$

The next lemma shows that Φ_K *locally* converges to Φ^* on B in total variation sense. The lemma holds for any bounded $B \in \mathfrak{B}$, but we focus on a bounded interval $B = [a,b]$. Proof is based on so called Poisson approximation and available in various forms, e.g., [77], but we give a more detailed proof with explicit convergence speed. Let $\lambda_K^* := \sum_k \lambda_{Kk}$.

Lemma 7. *Fix a bounded $B \in \mathfrak{B}$ of interest and let Φ_K^* be Poisson PP with intensity $\lambda_K^*|B|$. Suppose individual PPs of the triangular array are sparse with sum rate λ_K. Then, $N_K(B) \to N_K^*(B)$ in total variation. Furthermore, the speed of convergence is $O(g(K,B))$, where $g(K,B) := \max\{g_1(K,B), |B|^2 g_2(K)\}$.*

Proof. See Appendix B.4. □

The next corollary is especially useful in the next subsection, where each user sends symbols on i.i.d. renewal arrivals.

Corollary 8. *Further, suppose component PPs in a row of the triangular array are identically distributed, and $\lambda_K^* = \lambda$ for all K, i.e., Poisson PPs corresponding to each row are identical. Then, $d_{\mathsf{TV}}(N_K(B), N^*(B)) \to 0$ as $K \to \infty$ with speed $O(g_1(K,B), |B|^2 K^{-1}\})$, where N^* is the counting measure for the Poisson PP with intensity λ.*

3.4.1 Capacity Approximation

We reformulate input processes of the queue as two-sided RMPPs to streamline proofs and arguments. Recall that the mark space $\mathcal{M} = \mathbb{R}_+$ and service times are drawn i.i.d. from P^S. Suppose that the RMPPs begin at $t = -T$ for large $T > 0$ and the queue is initially empty. Since all randomness of queueing is captured by the RMPP, any queue-state process is a deterministic function of $\Phi(t)$ and initial queue state θ_{-T}. For example, discrete-time queue state processes, such as queue-length seen by arrivals or departures, can be expressed as $z(i, \Phi, \theta_{-T})$ for some deterministic function z.

As we have seen previously, the process of queue-length seen by departures $\{Q_i\}_{i\in\mathbb{Z}}$ is of interest. Note that

$$Q_i(\Phi) = h(i, \Phi, \theta_{-T}) \text{ for some deterministic function } h.$$

Consider the case of Cor. 8, where users' individual arrivals are i.i.d. and corresponding Poisson sum rate is identically $\lambda_K^* = \lambda$ for all K. As corresponding Poisson PPs are identically distributed regardless of row K, row index K for Poisson related quantities is dropped. Let $Q_i^{(K)}$ be the queue-length process seen by ith departure of the K-user superposition process. Similarly let Q_i^* be the corresponding process for the Poisson PP $\Phi^*(= \Phi_K^*$ for all $K)$. Then, the continuity theorem holds due to the local convergence property above. Here, $\overset{\mathsf{TV}}{\to}$

denotes local convergence of PP on $B \in \mathfrak{B}$ in total variation. For random variables, $\overset{\mathsf{TV}}{\to}$ is the usual total variational convergence.

Lemma 8. *For any $\epsilon > 0$, we can take a large interval $B = B(\epsilon) \in \mathfrak{B}$ that yields*

$$d_{\mathsf{TV}}(Q_k^{(K)}, Q_i^*) \leq 2\epsilon + O(g(K,B)),$$

where $g(K,B) = \max\{g_1(K,B), |B|^2 g_2(K)\}$. In other words, $Q_k^{(K)} \overset{\mathsf{TV}}{\to} Q_i^*$.

Proof. See Appendix B.5. □

Recall notations that π_{Kk}, π_K denote the stationary queue-length distributions seen by individual user's and superposed departures, respectively. As individual users are symmetric, π_{Kk} are identical and in addition $\pi_{Kk} = \pi_K$ for all k.

Since each arrival has only a few arrivals on B (with high probability), we implicitly suppose the transmission is repeated many times to achieve block code performance.

Let $c_{\mathsf{max}} := \sup_q \max_{P_X} I(P_X, W_q)$, which is $c_{\mathsf{max}} \leq \log|\mathcal{X}|$ clearly. The final approximation follows.

Theorem 10. *Let $C(\Phi^*)$ be the single-user capacity of $\mathsf{M/GI/1}$ queue with arrival rate λ, derived in Thm. 8. Consider K users with sparse individual PPs Φ_{Kk}. Then, under superposition, the sum capacity $C_{\mathsf{sum}}(\Phi_K)$ at arrival rate λ is approximated by the single-user capacity $C(\Phi^*)$ as*

$$|C_{\mathsf{sum}}(\Phi_K) - C(\Phi^*)| \leq c_{\mathsf{max}}\,(4\epsilon + O(g(K,B))) \;\; \textit{[bits/sym]},$$
$$|C_{\mathsf{sum}}(\Phi_K) - C(\Phi^*)| \leq \frac{g_1(K,B)}{|B|} c_{\mathsf{max}} + \lambda c_{\mathsf{max}}\,(4\epsilon + O(g(K,B))) \;\; \textit{[bits/time]}.$$

Proof. As $\pi_K = \pi_{Kk}$ for all k, individuals can send information at rate

$$C(\Phi_{Kk}) = \sum_q \pi_{Kk}(q) I(P_X, W_q) \quad \text{[bits/sym]},$$

the sum rate is also $C(\Phi_{Kk})$ in bits per symbol sense. On the other hand, the stationary distribution π_K differs from the stationary distribution for Poisson, say π^*, at most $2\epsilon +$

$O(g(K,B))$ in total variation. This implies

$$|C_{\mathsf{sum}}(\Phi_K) - C(\Phi^*)| = \left|\sum_{q=0}^{\infty}(\pi^*(q) - \pi_K(q))I(P_X, W_q)\right|$$

$$\leq c_{\mathsf{max}}\left|\sum_{q=0}^{\infty}(\pi^*(q) - \pi_K(q))\right|$$

$$\leq c_{\mathsf{max}}\sum_{q=0}^{\infty}|\pi^*(q) - \pi_K(q)| = c_{\mathsf{max}} \cdot 2d_{\mathsf{TV}}(Q_k^{(K)}, Q_i^*)$$

$$\leq c_{\mathsf{max}}(4\epsilon + O(g(K,B))).$$

To obtain the second bound, recall that actual sum arrival rate of the superposition process deviates from λ by $\frac{g_1(k,B)}{|B|}$. Therefore,

$$|C_{\mathsf{sum}}(\Phi_K) - C(\Phi^*)|$$

$$= \left|\left(\lambda + \frac{g_1(K,B)}{|B|}\right)\sum_q \pi_K(q)I(P_X, W_q) - \lambda\sum_q \pi^*(q)I(P_X, W_q)\right|$$

$$\leq \frac{g_1(K,B)}{|B|}c_{\mathsf{max}} + \lambda c_{\mathsf{max}} \cdot 2d_{\mathsf{TV}}(Q_k^{(K)}, Q_i^*)$$

$$\leq \frac{g_1(K,B)}{|B|}c_{\mathsf{max}} + \lambda c_{\mathsf{max}}\left(4\epsilon + O(g(K,B))\right) \quad \text{[bits/time]}$$

□

Thm. 10 only considers the sum capacity; however, it is clear from the proof that individual per symbol capacity remains unchanged, and per time capacity is properly scaled, i.e.,

$$\left|C_{\mathsf{ind}}(\Phi_{Kk}) - \frac{C(\Phi^*)}{K}\right| \leq \frac{g_1(K,B)}{K|B|}c_{\mathsf{max}} + \frac{\lambda}{K}c_{\mathsf{max}}\left(4\epsilon + O(g(K,B))\right) \text{ [bits/time]}.$$

Therefore, the best and worst server results in Cor. 7 also apply to the superposition arrivals asymptotically as $K \to \infty$.

Corollary 9. *Suppose the conditions in Sec. 3.2.2 hold. Then, for the K-user setting with sparse individuals, the results in Cor. 7 still hold asymptotically; that is, when the service quality steps down at $b = 1$, the sum and individual capacities are maximized when the service is deterministic. On the other hand, the sum and individual capacities are asymptotically minimized by cramming service.*

3.5 Chapter Summary

In this chapter, we extend the single-user results in Chap. 2 to a multiaccess setting. We first obtain the single-user capacity for continuous-time queues in a single letter form. Also similarly to the single-user case, stationarity and ergodicity of queueing process provide the multiuser capacity expression. Unlike usual multiuser or multiaccess problems, information rate in codewords does not change others' performance. This is because others' jobs affect channels only through arrival processes, but not through information symbols on them. Furthermore, when the number of users is large and each is sparse, the individual and sum capacities are asymptotically close to the single-user capacity of $\mathsf{M}/\mathsf{GI}/1$ queues, and thus the best (resp. the worst) service in the single-user setup is also the best (resp. the worst) in multiuser setup.

Chapter 4

Beliefs in Decision-Making Cascades

Team decision-making typically involves individual decisions that are influenced by private observations and the opinions of the rest of the team. The *social learning* setting is one such context where decisions of individual agents are influenced by preceding agents in the team [26, 27]. We consider the setting in which individual agents are selfish and aim to minimize their perceived Bayes risk, according to their beliefs as reinforced by the decisions of preceding agents.

4.1 Problem Description

Consider an N-agent cascading decision-making problem, as illustrated in Fig. 4.1. The underlying hypothesis, $H \in \{0,1\}$, is a binary signal with prior $\mathbb{P}\left[H=0\right] = p_0$ and $\mathbb{P}\left[H=1\right] = 1-p_0$. There are N agents that sequentially detect the state in a predetermined order. The nth agent has a *private* signal Y_n generated according to the likelihood $f_{Y_n|H}$, which is not necessarily identical for all n. Let the decision made by the nth agent be $\widehat{H}_n$. In addition to the private signal, the nth agent also observes the decisions made by preceding agents, $\{\widehat{H}_1, \ldots, \widehat{H}_{n-1}\}$, to make a decision $\widehat{H}_n$.

However, the nth agent believes the prior probability of the null hypothesis is $q_n \in (0,1)$ as against the true prior probability p_0. We call this the *belief* of the agent in order to distinguish it from the prior. Agent n is also aware of her own likelihood $f_{Y_n|H}$ that defines her private signal. However, she also perceives the likelihoods and beliefs of the other agents to be the same as hers, i.e., she thinks $f_{Y_j|H} = f_{Y_n|H}, q_j = q_n$ for all $j \neq n$, even though they could be different and unknown to her. We assume that the likelihood ratio of each agent is

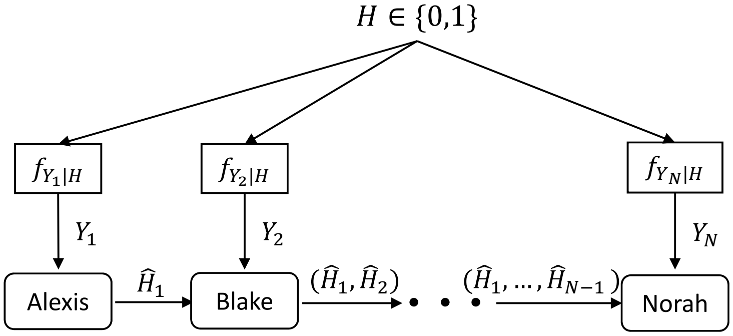


Figure 4.1: A cascading decision-making model with N agents.

an increasing function in y,[1] i.e., for all agents

$$\mathcal{L}_n(y) := \frac{f_{Y_n|H}(y|1)}{f_{Y_n|H}(y|0)}$$

is an increasing function of y.

Several numerical examples are given for private signals defined with independent additive Gaussian noise. The desired monotonicity also holds for many non-additive models, such as exponential distribution with mean H^{-1}, $H \in \mathbb{R}_+$, binomial distribution with success probability $H \in [0,1]$, and Poisson distribution with rate $H \in \mathbb{R}_+$ are members of such family, where H could take two values.

Our performance analysis focuses on the last agent (Nth agent, Norah) and her decision $\widehat{H}_N$. Upon observing her private signal Y_N and the $(N-1)$ preceding decisions, she determines her decision rule. The relative importance of correct decisions and errors can be abstracted as a cost function. For simplicity, we assume correct decisions have zero cost and use the shorthand notations $c_{10} = c(1,0)$ as the cost for false alarm or Type I error (choosing $\widehat{H} = 1$ when $H = 0$), and $c_{01} = c(0,1)$ as the cost for missed detection or Type II error (choosing $\widehat{H} = 0$ when $H = 1$). In addition, we assume that agents have the same costs; they are a team in the sense of Radner [81]. Then the Bayes risk is

$$R_N = c_{10} p_0 p_{\widehat{H}_N|H}(1|0) + c_{01}(1-p_0) p_{\widehat{H}_N|H}(0|1). \tag{4.1}$$

As $\widehat{H}_n$ depends on the previous decisions, the computation of (4.1) also depends on

[1] This property is particularly useful in uniformly most powerful (UMP) tests.

$(\widehat{H}_1, \ldots, \widehat{H}_{N-1})$, and the Bayes risk can be expanded as

$$\begin{aligned} R_N = \sum_{\widehat{h}_1, \ldots, \widehat{h}_{N-1}} & c_{10} p_0 p_{\widehat{H}_N, \widehat{H}_{N-1}, \ldots, \widehat{H}_1 | H}(1, \widehat{h}_{N-1}, \ldots, \widehat{h}_1 | 0) \\ & + c_{01}(1 - p_0) p_{\widehat{H}_N, \widehat{H}_{N-1}, \ldots, \widehat{H}_1 | H}(0, \widehat{h}_{N-1}, \ldots, \widehat{h}_1 | 1). \end{aligned} \tag{4.2}$$

We determine the optimal set of beliefs of the agents $\{q_n^*\}_{n=1}^N$ that minimize (4.2).

In our model, the nth agent minimizes her *perceived* Bayes risk, which is the Bayes risk with prior probability p_0 replaced by her belief q_n. In other words, for all $n = 1, \ldots, N$, the nth agent adopts the decision rule that minimizes her perceived Bayes risk R_n, and her decision is revealed to other agents as a public signal. The decisions $\{\widehat{H}_1, \ldots, \widehat{H}_{n-1}\}$ of the earlier-acting agents reveal information about H and thus should be incorporated into the decision-making process by agent n. As mentioned earlier, since she believes q_n is the true prior, she aggregates information under the assumption that $q_1 = q_2 = \cdots = q_n$.

It is important to note that every agent is selfish and rational; the agents do not adjust their decision rules for Norah's sake. The novelty in the model (and hence in the conclusions) comes from agent n having the limitation of using a private initial belief q_n in place of the true prior probability p_0.

4.1.1 Prospect Theory

Let us also formally introduce the Prelec reweighting function from cumulative prospect-theoretic models of human behavior. It spans a family of open- and closed-minded beliefs (will be clarified later), and thus the optimal beliefs that emerge in the following sections could be approximated by a function in the Prelec family.

Definition 8 ([55]). *For $\alpha, \beta > 0$, the Prelec reweighting function $w : [0,1] \mapsto [0,1]$ is*

$$w(p; \alpha, \beta) = \exp(-\beta(-\log p)^\alpha).$$

The function $w(p; \alpha, \beta)$ is:

1. strictly increasing,

2. has a unique fixed point $w(p; \alpha, \beta) = p$ at $p^* = \exp(-\exp(\log \beta/(1-\alpha)))$, and

3. spans a class of open-minded beliefs when $\alpha < 1$, i.e., overweights (underweights) small (high) probability, and vice versa when $\alpha > 1$.

A more generic form, termed composite Prelec weighting function, has been defined in [82].

4.1.2 Notations

Throughout the chapter, we use f for continuous probability density functions and p for discrete probability mass functions. All logarithms are natural logarithms. We use $\mathcal{N}(\mu, \sigma^2)$ to denote a Gaussian distribution with mean μ and variance σ^2, and $\phi(x; \mu, \sigma^2)$ to denote its density function, i.e.,

$$\phi(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Also in the case of the standard Gaussian, $\phi(x) := \phi(x; 0, 1)$ for simplicity. $Q(x)$ is defined as the complementary cumulative distribution function of the standard Gaussian,

$$Q(x) = \int_x^\infty \phi(t) dt.$$

4.2 Belief Update and Sequential Decision-Making

Our model assumes unbounded private signals. Thus, unlike in [28, 29], it is always possible that a subsequent agent may not follow previous decisions; that is, herding happens with arbitrarily low probability. We now discuss using both a decision history and private signals for Bayesian binary hypothesis testing. The decision rule can be interpreted as each agent updating her posterior belief based on the decision history and then applying a likelihood ratio test to her private signal.

4.2.1 Alexis, the First Agent

Since Alexis has no prior decision history, she follows usual binary hypothesis testing. She uses the following likelihood ratio test with her initial belief q_1, with ties broken arbitrarily:

$$\mathcal{L}_1(y_1) = \frac{f_{Y_1|H}(y_1|1)}{f_{Y_1|H}(y_1|0)} \underset{\widehat{H}_1=0}{\overset{\widehat{H}_1=1}{\gtrless}} \frac{c_{10} q_1}{c_{01}(1 - q_1)}. \tag{4.3}$$

Since we assume the likelihood ratio is increasing in y_1, the rule simplifies to comparing the private signal with an appropriate decision threshold:

$$y_1 \underset{\widehat{H}_1=0}{\overset{\widehat{H}_1=1}{\gtrless}} \lambda_1(q_1), \tag{4.4}$$

where $\lambda_i(q)$ denotes the decision threshold λ that satisfies

$$\mathcal{L}_i(\lambda) = \frac{f_{Y_i|H}(\lambda|1)}{f_{Y_i|H}(\lambda|0)} = \frac{c_{10}q}{c_{01}(1-q)}. \tag{4.5}$$

4.2.2 Blake, the Second Agent

Blake observes Alexis's decision $\widehat{H}_1 = \widehat{h}_1$ and evaluates the likelihood ratio for $(\widehat{H}_1, Y_2)$, using his initial belief q_2 as

$$\frac{f_{Y_2,\widehat{H}_1|H}(y_2,\widehat{h}_1|1)}{f_{Y_2,\widehat{H}_1|H}(y_2,\widehat{h}_1|0)} \underset{\widehat{H}_2=0}{\overset{\widehat{H}_2=1}{\gtrless}} \frac{c_{10}q_2}{c_{01}(1-q_2)}. \tag{4.6}$$

The private signals Y_1 and Y_2 are independently conditioned on H, so $\widehat{H}_1$ and Y_2 are also independently conditioned on H. Hence, the left side of (4.6) is

$$f_{Y_2,\widehat{H}_1|H}(y_2,\widehat{h}_1|h) = f_{Y_2|H}(y_2|h)p_{\widehat{H}_1|H}(\widehat{h}_1|h).$$

So we can rewrite (4.6) as[2]

$$\frac{f_{Y_2|H}(y_2|1)}{f_{Y_2|H}(y_2|0)} \underset{\widehat{H}_2=0}{\overset{\widehat{H}_2=1}{\gtrless}} \frac{c_{10}q_2}{c_{01}(1-q_2)}\frac{p_{\widehat{H}_1|H}(\widehat{h}_1|0)_{[2]}}{p_{\widehat{H}_1|H}(\widehat{h}_1|1)_{[2]}}. \tag{4.7}$$

The likelihood ratio test (4.7) can be interpreted as Blake updating his initial belief upon observing Alexis's decision $\widehat{H}_1$. Combined with q_2, his initial belief is updated according to

[2] The subscript [2] in the term $p_{\widehat{H}_1|H}(\widehat{h}_1|h)_{[2]}$ indicates the value of $p_{\widehat{H}_1|H}(\widehat{h}_1|h)$ that Blake (the second agent) thinks. We specify this because Blake does not know Alexis's belief q_1. Thus, he interprets her decision based on his belief q_2. The value is different from the true value of $p_{\widehat{H}_1|H}(\widehat{h}_1|h) = p_{\widehat{H}_1|H}(\widehat{h}_1|h)_{[1]}$. Of course, it will also be different from what Chuck, the third agent, perceives, which is denoted by $p_{\widehat{H}_1|H}(\widehat{h}_1|h)_{[3]}$. This will be explained in the next subsection.

$p_{\widehat{H}_1|H}(\widehat{h}_1|h)_{[2]}$, from q_2 to $q_2^{\widehat{h}_1}$:

$$\frac{q_2^{\widehat{h}_1}}{1-q_2^{\widehat{h}_1}} = \frac{q_2}{1-q_2}\frac{p_{\widehat{H}_1|H}(\widehat{h}_1|0)_{[2]}}{p_{\widehat{H}_1|H}(\widehat{h}_1|1)_{[2]}}. \tag{4.8}$$

The posterior belief is

$$\begin{aligned} q_2^{\widehat{h}_1} &= \frac{q_2 p_{\widehat{H}_1|H}(\widehat{h}_1|0)_{[2]}}{q_2 p_{\widehat{H}_1|H}(\widehat{h}_1|0)_{[2]} + (1-q_2) p_{\widehat{H}_1|H}(\widehat{h}_1|1)_{[2]}} \\ &= \frac{p_{\widehat{H}_1,H}(\widehat{h}_1,0)_{[2]}}{p_{\widehat{H}_1,H}(\widehat{h}_1,0)_{[2]} + p_{\widehat{H}_1,H}(\widehat{h}_1,1)_{[2]}} \\ &= p_{H|\widehat{H}_1}(0|\widehat{h}_1)_{[2]}. \end{aligned} \tag{4.9}$$

It should be noted that the true $p_{\widehat{H}_1|H}(\widehat{h}_1|h)$ is given by

$$\begin{aligned} p_{\widehat{H}_1|H}(0|h) = p_{\widehat{H}_1|H}(0|h)_{[1]} &= \mathbb{P}\left[Y_1 \le \lambda_1(q_1) | H = h\right] \\ &= \int_{-\infty}^{\lambda_1(q_1)} f_{Y_1|H}(y|h)dy, \\ p_{\widehat{H}_1|H}(1|h) &= \int_{\lambda_1(q_1)}^{\infty} f_{Y_1|H}(y|h)dy. \end{aligned}$$

But Blake evaluates Alexis's decision $\widehat{H}_1$ as if it were made based on q_2 and the likelihood $f_{Y_2|H}(\cdot)$, as against $q_1, f_{Y_1|H}(\cdot)$ respectively. Thus the probability $p_{\widehat{H}_1|H}(\widehat{h}_1|h)$ is computed based on $\lambda_2(q_2)$, instead of $\lambda_1(q_2)$:

$$p_{\widehat{H}_1|H}(0|h)_{[2]} = \int_{-\infty}^{\lambda_2(q_2)} f_{Y_2|H}(y|h)dy, \tag{4.10a}$$

$$p_{\widehat{H}_1|H}(1|h)_{[2]} = \int_{\lambda_2(q_2)}^{\infty} f_{Y_2|H}(y|h)dy. \tag{4.10b}$$

An interesting observation is that Alexis's belief q_1 does not affect Blake's belief update as observed in (4.9) and (4.10). That is, for any belief q_1 that Alexis might hold, Blake, who does not know this belief, presumes that the conditional probabilities are computed according to (4.10) and updates his belief as in (4.9) which depends only on Blake's initial belief and Alexis's decision.

However, Alexis's initial belief implicitly affects Blake's performance since her biased belief changes the resulting decisions whose probabilities are embedded in the probability of Blake's decision:

$$\begin{aligned}p_{\widehat{H}_2|H}(\widehat{h}_2|h) &= \sum_{\widehat{h}_1\in\{0,1\}} p_{\widehat{H}_2,\widehat{H}_1|H}(\widehat{h}_2,\widehat{h}_1|h)\\ &= p_{\widehat{H}_2|\widehat{H}_1,H}(\widehat{h}_2|0,h)_{[2]} \times p_{\widehat{H}_1|H}(0|h)_{[1]}\\ &\quad + p_{\widehat{H}_2|\widehat{H}_1,H}(\widehat{h}_2|1,h)_{[2]} \times p_{\widehat{H}_1|H}(1|h)_{[1]}.\end{aligned}$$

Thus, Alexis's biased belief changes the probability of not only her decision but also of Blake's decision.

4.2.3 Chuck, the Third Agent

Chuck's detection process is similar to Blake's. He observes both Alexis's and Blake's decisions and also updates his initial belief q_3 like in (4.8):

$$\begin{aligned}\frac{q_3^{\widehat{h}_1,\widehat{h}_2}}{1-q_3^{\widehat{h}_1,\widehat{h}_2}} &= \frac{q_3}{1-q_3}\frac{p_{\widehat{H}_2,\widehat{H}_1|H}(\widehat{h}_2,\widehat{h}_1|0)_{[3]}}{p_{\widehat{H}_2,\widehat{H}_1|H}(\widehat{h}_2,\widehat{h}_1|1)_{[3]}}\\ &= \left(\frac{q_3}{1-q_3}\frac{p_{\widehat{H}_1|H}(\widehat{h}_1|0)_{[3]}}{p_{\widehat{H}_1|H}(\widehat{h}_1|1)_{[3]}}\right)\frac{p_{\widehat{H}_2|\widehat{H}_1,H}(\widehat{h}_2|\widehat{h}_1,0)_{[3]}}{p_{\widehat{H}_2|\widehat{H}_1,H}(\widehat{h}_2|\widehat{h}_1,1)_{[3]}}.\end{aligned} \tag{4.11}$$

Note that $\widehat{H}_1$ and $\widehat{H}_2$ are not conditionally independent given H as Blake's decision $\widehat{H}_2$ depends on Alexis's decision $\widehat{H}_1$.

Chuck's belief update can be understood as a two-step process. The first step is to update his belief according to Alexis's decision:

$$\frac{q_3^{\widehat{h}_1}}{1-q_3^{\widehat{h}_1}} = \frac{q_3}{1-q_3}\frac{p_{\widehat{H}_1|H}(\widehat{h}_1|0)_{[3]}}{p_{\widehat{H}_1|H}(\widehat{h}_1|1)_{[3]}}. \tag{4.12}$$

The second step is to update it from $q_3^{\widehat{h}_1}$ based on Blake's decision:

$$\frac{q_3^{\widehat{h}_1,\widehat{h}_2}}{1-q_3^{\widehat{h}_1,\widehat{h}_2}} = \frac{q_3^{\widehat{h}_1}}{1-q_3^{\widehat{h}_1}}\frac{p_{\widehat{H}_2|\widehat{H}_1,H}(\widehat{h}_2|\widehat{h}_1,0)_{[3]}}{p_{\widehat{H}_2|\widehat{H}_1,H}(\widehat{h}_2|\widehat{h}_1,1)_{[3]}}. \tag{4.13}$$

Again, Chuck is aware of neither Alexis's nor Blake's initial beliefs or likelihoods. Thus, Chuck computes all probabilities based on his own belief q_3 and likelihood $f_{Y_3|H}$, which is indicated by the subscript [3] in (4.12) and (4.13).

Details of computations of (4.12) and (4.13) are as follows:

$$
\begin{aligned}
p_{\widehat{H}_1|H}(0|h)_{[3]} &= \int_{-\infty}^{\lambda_3(q_3)} f_{Y_3|H}(y|h)dy, \\
p_{\widehat{H}_1|H}(1|h)_{[3]} &= \int_{\lambda_3(q_3)}^{\infty} f_{Y_3|H}(y|h)dy.
\end{aligned}
$$

Similar to Blake (4.8), Chuck computes $q_3^{\widehat{h}_1}$ for $\widehat{H}_1 = 0$ and $\widehat{H}_1 = 1$ respectively as:

$$
q_3^0 = \frac{q_3}{q_3 + (1-q_3)\frac{\int_{-\infty}^{\lambda_3(q_3)} f_{Y_3|H}(y|1)dy}{\int_{-\infty}^{\lambda_3(q_3)} f_{Y_3|H}(y|0)dy}}, \tag{4.14a}
$$

$$
q_3^1 = \frac{q_3}{q_3 + (1-q_3)\frac{\int_{\lambda_3(q_3)}^{\infty} f_{Y_3|H}(y|1)dy}{\int_{\lambda_3(q_3)}^{\infty} f_{Y_3|H}(y|0)dy}}. \tag{4.14b}
$$

Then,

$$
p_{\widehat{H}_2|\widehat{H}_1,H}(0|\widehat{h}_1, h)_{[3]} = \int_{-\infty}^{\lambda_3(q_3^{\widehat{h}_1})} f_{Y_3|H}(y|h)dy, \tag{4.15a}
$$

$$
p_{\widehat{H}_2|\widehat{H}_1,H}(1|\widehat{h}_1, h)_{[3]} = \int_{\lambda_3(q_3^{\widehat{h}_1})}^{\infty} f_{Y_3|H}(y|h)dy. \tag{4.15b}
$$

Even though the value of $\widehat{h}_1$ does not appear in (4.15), it is implicit in $q_3^{\widehat{h}_1}$ and affects the computation results. Chuck's posterior belief $q_3^{\widehat{h}_1,\widehat{h}_2}$ is obtained by substituting (4.14) and (4.15) in (4.13).

4.2.4 Norah, the Nth Agent

Norah, the Nth agent, observes Y_N and $\{\widehat{H}_1, \ldots, \widehat{H}_{N-1}\}$. Paralleling the arguments in the preceding subsections, her initial belief update is a function of q_N as well as $\{\widehat{H}_1, \ldots, \widehat{H}_{N-1}\}$, but not of $\{q_1, \ldots, q_{N-1}\}$. Generalizing (4.11), we have

$$\frac{q_N^{\widehat{h}_1,\ldots\widehat{h}_{N-1}}}{1-q_N^{\widehat{h}_1,\ldots\widehat{h}_{N-1}}} = \frac{q_N}{1-q_N}\frac{p_{\widehat{H}_1|H}(\widehat{h}_1|0)_{[N]}}{p_{\widehat{H}_1|H}(\widehat{h}_1|1)_{[N]}}\times \\ \prod_{n=2}^{N-1}\frac{p_{\widehat{H}_n|\widehat{H}_{n-1},\ldots,\widehat{H}_1,H}(\widehat{h}_n|\widehat{h}_{n-1},\ldots,\widehat{h}_1,0)_{[N]}}{p_{\widehat{H}_n|\widehat{H}_{n-1},\ldots,\widehat{H}_1,H}(\widehat{h}_n|\widehat{h}_{n-1},\ldots,\widehat{h}_1,1)_{[N]}}. \tag{4.16}$$

Combining all observations, we obtain the following theorem. Define the initial belief update function for Nth agent, U_N as

$$q_N^{\widehat{h}_1\ldots\widehat{h}_{N-1}} = U_N(q_N, \widehat{h}_1, \widehat{h}_2, \ldots, \widehat{h}_{N-1}; N).$$

Theorem 11. *The function $U_n, n \leq N$ yielding the posterior belief of Nth agent has the following recurrence relation:*

- *For $n = 1$, $U_1(q; N) = q$.*
- *For $n > 1$,*

$$U_n(q, \widehat{h}_1, \ldots, \widehat{h}_{n-2}, 0; N) \\ = \frac{\tilde{q}}{\tilde{q} + (1-\tilde{q})\frac{\int_{-\infty}^{\lambda_N(\tilde{q})} f_{Y_N|H}(y|1)dy}{\int_{-\infty}^{\lambda_N(\tilde{q})} f_{Y_N|H}(y|0)dy}}, \tag{4.17a}$$

$$U_n(q, N, \widehat{h}_1, \ldots, \widehat{h}_{n-2}, 1; N) \\ = \frac{\tilde{q}}{\tilde{q} + (1-\tilde{q})\frac{\int_{\lambda_N(\tilde{q})}^{\infty} f_{Y_N|H}(y|1)dy}{\int_{\lambda_N(\tilde{q})}^{\infty} f_{Y_N|H}(y|0)dy}}, \tag{4.17b}$$

where $\tilde{q} = U_{n-1}(q, \widehat{h}_1, \ldots, \widehat{h}_{n-2}; N)$.

Note that capital N in (17a) and (17b) indicate the recursive updates are computed from the value that the Nth agent thinks.

Fig. 4.2 depicts the function $U_4(q_4, \widehat{h}_1, \widehat{h}_2, \widehat{h}_3; 4)$ for $N = 4$ for eight possible combinations of Alexis's, Blake's, and Chuck's decisions $(\widehat{h}_1, \widehat{h}_2, \widehat{h}_3)$. An interesting property of U_N is that the posterior belief is much more dependent on the most recent decision $\widehat{h}_{N-1}$ than on the earlier decisions $(\widehat{h}_1, \ldots, \widehat{h}_{N-2})$. In this sense, we can interpret that recent decisions give more information than earlier decisions. This is especially the case when the $(N-1)$th agent has not followed precedent. This is because the Nth agent rationally concludes that the

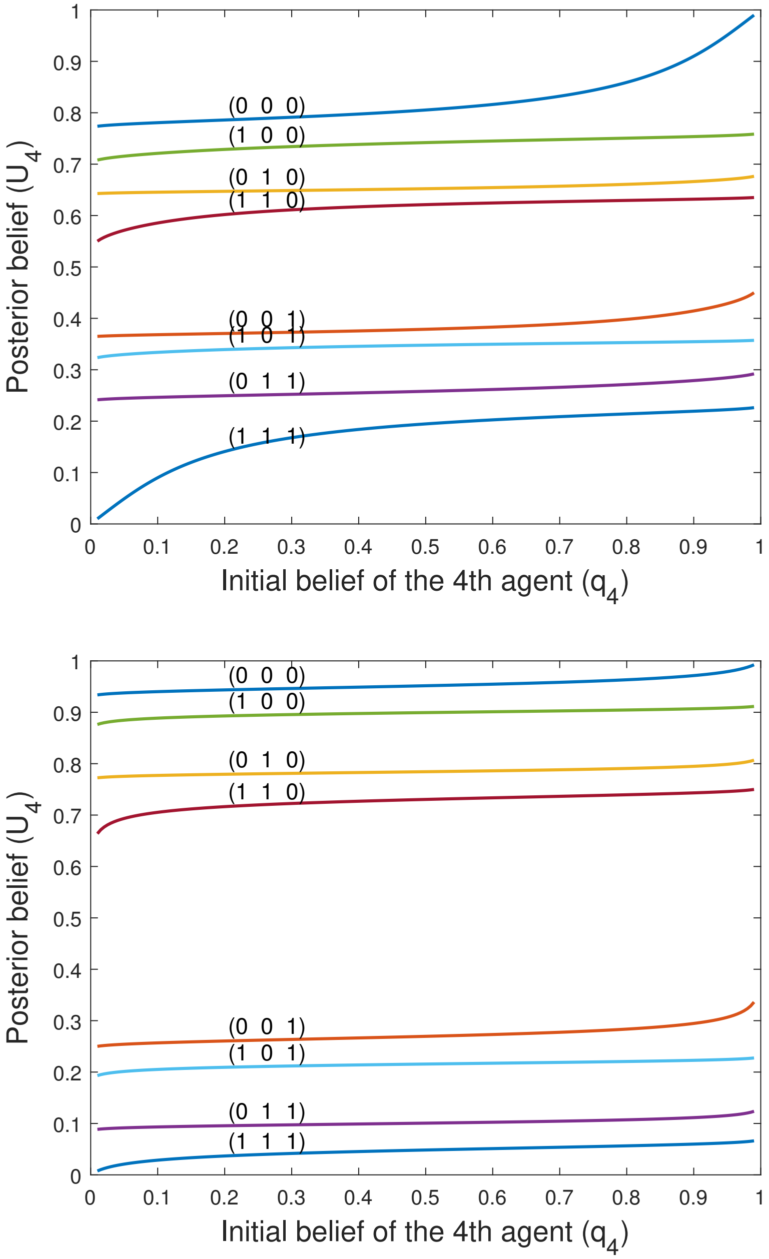


Figure 4.2: The function $U_4(q_4, \widehat{h}_1, \widehat{h}_2, \widehat{h}_3; 4)$—posterior belief of the fourth agent $(q_4^{\widehat{h}_1,\widehat{h}_2,\widehat{h}_3})$—for each possible combination of Alexis's, Blake's, and Chuck's decisions $[\widehat{h}_1, \widehat{h}_2, \widehat{h}_3]$ when $c_{10} = c_{01} = 1$ and private signals are distorted by additive Gaussian noise with two noise levels. The posterior belief is mostly dependent on Chuck's decision; the top four curves are for $\widehat{h}_3 = 0$ and the bottom four curves are for $\widehat{h}_3 = 1$.

$(N-1)$th agent observed strong evidence to justify a deviation from precedent. For example, if the decision history of the first five agents is $(0,0,0,0,1)$ then the sixth agent takes the last decision 1 seriously even though the first four agents chose 0. A reversal of an arbitrarily long precedent sequence may occur because we assume unbounded private signals; if private signals are bounded [28,29], then the influence of the precedent can reach a point where agents cannot receive a signal strong enough to justify a decision running counter to precedent. Another interesting point is that smaller noise variance changes beliefs more. It is clear from (4.17), but also from common sense, that when the variance is smaller, the Nth agent trusts and is more inclined towards previous decisions. Note that even though the prior updates of Norah in Fig. 4.2 do not depend on $\{q_1, \ldots, q_{N-1}\}$ and their corresponding likelihoods, the probability of prior decisions depends on them and, implicitly, so does Norah's decision.

As we can see in Fig. 4.2, the dominant previous decision for agent N is the decision of agent $(N-1)$. We can prove that observing the $(N-1)$th agent's decision 0 (or decision 1), the Nth agent's posterior belief becomes larger (or smaller), which in turn implies that the decision threshold of Nth agent becomes larger (or smaller) so that she is more likely to declare decision 0 (or 1) as well.

Theorem 12. *Suppose that noises are independent and additive, and have continuous densities. Fix some prior decisions $\{\widehat{h}_1, \ldots, \widehat{h}_{N-2}\}$ and let $\tilde{q}_N, \tilde{q}_N^0, \tilde{q}_N^1$ denote the posterior beliefs of the Nth agent given the $(N-2)$ decisions only, the $(N-2)$ decisions with $\widehat{h}_{N-1} = 0$, and the $(N-2)$ decisions with $\widehat{h}_{N-1} = 1$. Then,*

$$\tilde{q}_N^1 < \tilde{q}_N < \tilde{q}_N^0.$$

Proof. We know that $\tilde{q}_N, \tilde{q}_N^0, \tilde{q}_N^1$ differ only by the last multiplicative term of (4.16). Since $\frac{q}{1-q}$ is monotone increasing, the statement is equivalent to showing:

$$\frac{\int_{\lambda_N(\tilde{q}_N)}^{\infty} f_{Y_N|H}(y|0)dy}{\int_{\lambda_N(\tilde{q}_N)}^{\infty} f_{Y_N|H}(y|1)dy} < 1 < \frac{\int_{-\infty}^{\lambda_N(\tilde{q}_N)} f_{Y_N|H}(y|0)dy}{\int_{-\infty}^{\lambda_N(\tilde{q}_N)} f_{Y_N|H}(y|1)dy}.$$

Since the noise is independent and additive, $f_{Y_N|H}(y|1) = f_{Y_N|H}(y-1|0)$ so the term on the left side

$$\frac{\int_{\lambda_N(\tilde{q}_N)}^{\infty} f_{Y_N|H}(y|0)dy}{\int_{\lambda_N(\tilde{q}_N)}^{\infty} f_{Y_N|H}(y|1)dy} = \frac{\int_{\lambda_n(\tilde{q}_N)}^{\infty} f_{Y_N|H}(y|0)dy}{\int_{\lambda_N(\tilde{q}_N)-1}^{\infty} f_{Y_N|H}(y|0)dy}$$
$$= \frac{\int_{\lambda_N(\tilde{q}_N)}^{\infty} f_{Y_N|H}(y|0)dy}{\int_{\lambda_N(\tilde{q}_N)-1}^{\lambda_N(\tilde{q}_N)} f_{Y_N|H}(y|0)dy + \int_{\lambda_N(\tilde{q}_N)}^{\infty} f_{Y_N|H}(y|0)dy} < 1.$$

The right inequality can be shown similarly. □

Considering the complicated relationships that individual decisions have on the evolution of initial beliefs, it is also important to verify if the belief evolution preserves the ordering, given the same set of subsequent decisions. That is, given two beliefs $q_L < q_R$ at some point of the recursive update and the same sequence of following d decisions, then it is important to characterize the likelihoods for which the the ordering is preserved in the resulting posterior beliefs, given the sequence of decisions, which is described in the following theorem.

Theorem 13. *Suppose that noise is independent and additive, and has a continuous density. Consider two beliefs $q_L < q_R$. Then, for any given later-acting decisions d, the posterior belief satisfies $q_L^d < q_R^d$ if and only if*

$$g_1(q) := \frac{q}{1-q} \frac{\int_{-\infty}^{\lambda_N(q)} f_{Y_N|H}(y|0)dy}{\int_{-\infty}^{\lambda_N(q)} f_{Y_N|H}(y|1)dy}, \tag{4.18}$$

$$g_2(q) := \frac{q}{1-q} \frac{\int_{\lambda_N(q)}^{\infty} f_{Y_N|H}(y|0)dy}{\int_{\lambda_N(q)}^{\infty} f_{Y_N|H}(y|1)dy} \tag{4.19}$$

are both increasing in q.

Proof. Note that once observing decision 0, beliefs are updated as

$$\frac{q_L^0}{1-q_L^0} = \frac{q_L}{1-q_L} \frac{\int_{-\infty}^{\lambda_N(q_L)} f_{Y_N|H}(y|0)dy}{\int_{-\infty}^{\lambda_N(q_L)} f_{Y_N|H}(y|1)dy},$$
$$\frac{q_R^0}{1-q_R^0} = \frac{q_R}{1-q_R} \frac{\int_{-\infty}^{\lambda_N(q_R)} f_{Y_N|H}(y|0)dy}{\int_{-\infty}^{\lambda_N(q_R)} f_{Y_N|H}(y|1)dy},$$

and so if (4.18) holds, $q_L^0 < q_R^0$. Similarly, (4.19) can be shown by updating after decision 1. □

Let us state some properties of Mills ratio [83, 84], which is about Gaussian distribution, and we will see that $g_1(q), g_2(q)$ are both increasing if likelihood is Gaussian.

Lemma 9 ([84]). *Define $\eta(x) := \phi(x)/Q(x)$, the inverse of Mills ratio. Then, for any $x \in \mathbb{R}$, it is true that $0 < \eta'(x) < 1$ and $\eta''(x) > 0$.*

Corollary 10. *Consider a Gaussian likelihood, i.e., $Y_N = H + Z_N$, where Z_N are independent and identically drawn from $\mathcal{N}(0, \sigma^2)$, for some $\sigma^2 > 0$. Then $g_1(q), g_2(q)$ are both increasing in q.*

Proof. Let us consider $g_2(q)$ first. For the binary hypothesis test with Gaussian noise, we know that the decision threshold for the likelihood ratio test is given by

$$\lambda_N(q) = \frac{1}{2} + \sigma^2 \log\left(\frac{c_{10}q}{c_{01}(1-q)}\right).$$

Then, we have

$$g_2(q) = \frac{q}{1-q} \frac{Q\left(\frac{\lambda_N(q)}{\sigma}\right)}{Q\left(\frac{\lambda_N(q)-1}{\sigma}\right)}.$$

Letting $x := \log \frac{c_{10}q}{c_{01}(1-q)}$, it is sufficient to show that

$$\begin{aligned}\tilde{g}(x) &:= \log\left(\frac{c_{10}}{c_{01}} g_2(q)\right) \\ &= x + \log\left(Q\left(\sigma x + \tfrac{1}{2\sigma}\right)\right) - \log\left(Q\left(\sigma x - \tfrac{1}{2\sigma}\right)\right),\end{aligned}$$

is increasing in x since c_{10}, c_{01} are positive constants, $\log(\cdot)$ is a monotonically increasing function, and x is a strictly increasing function of q.

The first derivative of $\tilde{g}$ is given by

$$\tilde{g}'(x) = 1 - \sigma\eta\left(\sigma x + \tfrac{1}{2\sigma}\right) + \sigma\eta\left(\sigma x - \tfrac{1}{2\sigma}\right). \tag{4.20}$$

Since $\eta(\cdot)$ is a continuous function, using the mean value theorem, there exists y in $\left(\sigma x - \frac{1}{2\sigma}, \sigma x + \frac{1}{2\sigma}\right)$, such that

$$\sigma\eta\left(\sigma x + \tfrac{1}{2\sigma}\right) - \sigma\eta\left(\sigma x - \tfrac{1}{2\sigma}\right) = \sigma\eta'(y)\frac{1}{\sigma} = \eta'(y). \tag{4.21}$$

From the first property of Lem. 9, $0 < \eta'(y) < 1$, we have

$$\eta\left(\sigma x + \tfrac{1}{2\sigma}\right) - \eta\left(\sigma x - \tfrac{1}{2\sigma}\right) < 1.$$

Thus, from (4.20), it follows that $\tilde{g}'(x) > 0$ for all x, indicating that $\tilde{g}(\cdot)$ is an increasing function of x. This in turn implies that $g_2(\cdot)$ is also an increasing function.

To prove the result for g_1, it is sufficient to observe that by the symmetry of error probabilities:

$$g_1(q) = \frac{1}{g_2(1-q)}.$$

□

4.3 Optimal Belief

We described the initial belief evolution and decision-making model in Sec. 4.2. In this section, we investigate the set of initial beliefs that minimize the Bayes risk. We consider the case of two agents for analytical tractability although the broad nature of the arguments extends to multi-agent systems. Note that the Bayes risk of the system with $N = 2$ is the same as Blake's Bayes risk because his decision is adopted as the final decision.

Let us recapitulate the computation of Blake's Bayes risk. Alexis chooses her decision threshold as $\lambda_1 := \lambda_1(q_1)$. Her probabilities of error are given by

$$P_{e,1}^{\mathrm{I}} = p_{\widehat{H}_1|H}(1|0) = \int_{\lambda_1}^{\infty} f_{Y_1|H}(y|0)dy,$$
$$P_{e,1}^{\mathrm{II}} = p_{\widehat{H}_1|H}(0|1) = \int_{-\infty}^{\lambda_1} f_{Y_1|H}(y|1)dy.$$

Blake however presumes Alexis uses the decision threshold $\lambda_{1,[2]} := \lambda_2(q_2)$ and computes her probabilities of error accordingly[3]:

$$P_{e,1,[2]}^{\mathrm{I}} = p_{\widehat{H}_1|H}(1|0)_{[2]} = \int_{\lambda_{1,[2]}}^{\infty} f_{Y_2|H}(y|0)dy,$$
$$P_{e,1,[2]}^{\mathrm{II}} = p_{\widehat{H}_1|H}(0|1)_{[2]} = \int_{-\infty}^{\lambda_{1,[2]}} f_{Y_2|H}(y|1)dy.$$

[3] Recall that the subscript [2] denotes the quantity 'seen by' Blake.

When Alexis decides $\widehat{H}_1 = 0$, Blake updates his belief q_2 to the posterior q_2^0:

$$\begin{aligned} \frac{q_2^0}{1-q_2^0} &= \frac{q_2}{1-q_2} \frac{1-P_{e,1,[2]}^{\mathrm{I}}}{P_{e,1,[2]}^{\mathrm{II}}} \\ \implies q_2^0 &= \frac{q_2(1-P_{e,1,[2]}^{\mathrm{I}})}{q_2(1-P_{e,1,[2]}^{\mathrm{I}}) + (1-q_2)P_{e,1,[2]}^{\mathrm{II}}}, \end{aligned} \tag{4.22}$$

his decision threshold is $\lambda_2^0 := \lambda_2(q_2^0)$, and the probabilities of error are

$$\begin{aligned} P_{e,2}^{\mathrm{I}_0} &= p_{\widehat{H}_2|\widehat{H}_1,H}(1|0,0) = \int_{\lambda_2^0}^{\infty} f_{Y_2|H}(y|0)dy, \\ P_{e,2}^{\mathrm{II}_0} &= p_{\widehat{H}_2|\widehat{H}_1,H}(0|0,1) = \int_{-\infty}^{\lambda_2^0} f_{Y_2|H}(y|1)dy. \end{aligned}$$

Likewise, when Alexis decides $\widehat{H}_1 = 1$, Blake updates his belief q_2 to the posterior q_2^1:

$$\begin{aligned} \frac{q_2^1}{1-q_2^1} &= \frac{q_2}{1-q_2} \frac{P_{e,1,[2]}^{\mathrm{I}}}{1-P_{e,1,[2]}^{\mathrm{II}}} \\ \implies q_2^1 &= \frac{q_2 P_{e,1,[2]}^{\mathrm{I}}}{q_2 P_{e,1,[2]}^{\mathrm{I}} + (1-q_2)(1-P_{e,1,[2]}^{\mathrm{II}})}, \end{aligned} \tag{4.23}$$

his decision threshold is $\lambda_2^1 := \lambda_2(q_2^1)$, and the probabilities of error are

$$\begin{aligned} P_{e,2}^{\mathrm{I}_1} &= p_{\widehat{H}_2|\widehat{H}_1,H}(1|1,0) = \int_{\lambda_2^1}^{\infty} f_{Y_2|H}(y|0)dy, \\ P_{e,2}^{\mathrm{II}_1} &= p_{\widehat{H}_2|\widehat{H}_1,H}(0|1,1) = \int_{-\infty}^{\lambda_2^1} f_{Y_2|H}(y|1)dy. \end{aligned}$$

Now we compute the system's Bayes risk (or Blake's Bayes risk) R_2:

$$\begin{aligned} R_2 &= c_{10} p_{\widehat{H}_2,H}(1,0) + c_{01} p_{\widehat{H}_2,H}(0,1) \\ &= c_{10} \sum_{\widehat{h}_1 \in \{0,1\}} p_{\widehat{H}_2|\widehat{H}_1,H}(1|\widehat{h}_1,0) p_{\widehat{H}_1|H}(\widehat{h}_1|0) p_H(0) \\ &\quad + c_{01} \sum_{\widehat{h}_1 \in \{0,1\}} p_{\widehat{H}_2|\widehat{H}_1,H}(0|\widehat{h}_1,1) p_{\widehat{H}_1|H}(\widehat{h}_1|1) p_H(1) \\ &= c_{10} \left[P_{e,2}^{\mathrm{I}_0}(1-P_{e,1}^{\mathrm{I}}) + P_{e,2}^{\mathrm{I}_1} P_{e,1}^{\mathrm{I}} \right] p_0 \\ &\quad + c_{01} \left[P_{e,2}^{\mathrm{II}_0} P_{e,1}^{\mathrm{II}} + P_{e,2}^{\mathrm{II}_1}(1-P_{e,1}^{\mathrm{II}}) \right] (1-p_0). \end{aligned} \tag{4.24}$$

Note that the Bayes risk R_2 in (4.24) is a function of q_1 and q_2. One might think that R_2 is minimum at $q_1 = q_2 = p_0$ as Alexis makes the best decision for the true prior and Blake does not misunderstand her decision. Surprisingly, however, this turns out to be untrue. We prove this by studying Alexis's optimal belief q_1^* that minimizes R_2.

Theorem 14. *Alexis's and Blake's optimal beliefs q_1^*, q_2^* that minimize R_2 satisfy*

$$\frac{q_1^*}{1-q_1^*} = \frac{p_0(P_{e,2}^{I_1} - P_{e,2}^{I_0})}{(1-p_0)(P_{e,2}^{II_0} - P_{e,2}^{II_1})}. \tag{4.25}$$

Before proceeding to the proof, note that error probability terms in the right-side are dependent on q_2, but not on q_1. Furthermore, the value of $(P_{e,2}^{\mathrm{I}_1} - P_{e,2}^{\mathrm{I}_0})/(P_{e,2}^{\mathrm{II}_0} - P_{e,2}^{\mathrm{II}_1})$ is generally not 1, i.e., in general $q_1 = q_2 = p_0$ is not the optimal belief. For example, for the additive Gaussian noise model considered in the next section, the ratio is not equal to 1 except when $p_0 = c_{01}/(c_{10} + c_{01})$.

Proof of Thm. 14. Let us consider the first derivative of (4.24) with respect to q_1:

$$\begin{aligned}\frac{\partial R_2}{\partial q_1} &= c_{10}p_0(P_{e,2}^{\mathrm{I}_1} - P_{e,2}^{\mathrm{I}_0})\frac{\partial P_{e,1}^{\mathrm{I}}}{\partial q_1} \\ &\quad + c_{01}(1-p_0)(P_{e,2}^{\mathrm{II}_0} - P_{e,2}^{\mathrm{II}_1})\frac{\partial P_{e,1}^{\mathrm{II}}}{\partial q_1}.\end{aligned}$$

We want to find q_1 that minimizes R_2, i.e., q_1 makes the first derivative zero. Using

$$\begin{aligned}\frac{dP_{e,1}^{\mathrm{I}}}{dq_1} &= \frac{dP_{e,1}^{\mathrm{I}}}{d\lambda_1}\frac{d\lambda_1}{dq_1} = -f_{Y_1|H}(\lambda_1|0)\frac{d\lambda_1}{dq_1}, \\ \frac{dP_{e,1}^{\mathrm{II}}}{dq_1} &= \frac{dP_{e,1}^{\mathrm{II}}}{d\lambda_1}\frac{d\lambda_1}{dq_1} = f_{Y_1|H}(\lambda_1|1)\frac{d\lambda_1}{dq_1};\end{aligned}$$

this occurs when

$$\begin{aligned}&c_{10}p_0(P_{e,2}^{\mathrm{I}_1} - P_{e,2}^{\mathrm{I}_0})f_{Y_1|H}(\lambda_1|0) \\ &= c_{01}(1-p_0)(P_{e,2}^{\mathrm{II}_0} - P_{e,2}^{\mathrm{II}_1})f_{Y_1|H}(\lambda_1|1) \\ &\Leftrightarrow \frac{f_{Y_1|H}(\lambda_1|1)}{f_{Y_1|H}(\lambda_1|0)} = \frac{c_{10}p_0(P_{e,2}^{\mathrm{I}_1} - P_{e,2}^{\mathrm{I}_0})}{c_{01}(1-p_0)(P_{e,2}^{\mathrm{II}_0} - P_{e,2}^{\mathrm{II}_1})}.\end{aligned} \tag{4.26}$$

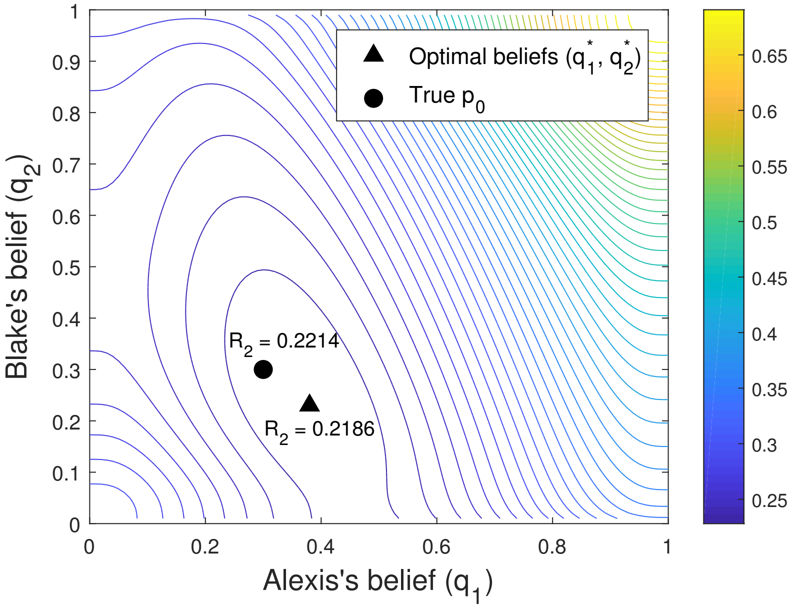


Figure 4.3: The Bayes risk for $q_1, q_2 \in (0, 1)$ with $p_0 = 0.3$, $c_{10} = c_{01} = 1$, and additive standard Gaussian noise. The pair of optimal beliefs (▲) yields $R_2 = 0.2186$, while the true prior (●) yields $R_2 = 0.2214$.

Note that $\lambda_1 = \lambda_1(q_1)$ is the solution to (4.4),

$$\frac{f_{Y_1|H}(\lambda_1|1)}{f_{Y_1|H}(\lambda_1|0)} = \frac{c_{10}q_1}{c_{01}(1-q_1)}. \tag{4.27}$$

Equating (4.26) and (4.27) completes the proof. □

The theorem considers general continuous likelihoods $\{f_{Y_n|H}\}$ with the monotonicity assumption on $\lambda(q)$. It is interesting to evaluate the optimal beliefs in the case of Gaussian likelihoods (i.e., additive Gaussian noise) and obtain insights into optimality in the sequential decision-making problem.

4.4 Gaussian Likelihoods

We now focus on Gaussian likelihoods and study their optimal beliefs in this section. Suppose the nth agent receives the signal $Y_n = H + Z_n$, where Z_n is an independent additive Gaussian noise with zero mean and variance $\sigma_n^2 > 0$. Thus, the received signal probability densities for $H = h$ are

$$f_{Y_n|H}(y_n|h) = \phi(y_n; h, \sigma_n^2).$$

For a belief q_n, the decision threshold is then determined by the likelihood ratio test,

$$\mathcal{L}_n(y_n) = \frac{f_{Y_n|H}(y_n|1)}{f_{Y_n|H}(y_n|0)} \underset{\widehat{H}_1=0}{\overset{\widehat{H}_1=1}{\gtrless}} \frac{c_{10}q_n}{c_{01}(1-q_n)},$$

that simplifies to the following simple threshold condition for Gaussian likelihoods:

$$y_n \underset{\widehat{H}_1=0}{\overset{\widehat{H}_1=1}{\gtrless}} \lambda_n(q_n) = \frac{1}{2} + \sigma_n^2 \log\left(\frac{c_{10}q_n}{c_{01}(1-q_n)}\right). \tag{4.28}$$

Here the index n represents the nth agent in the system, as the belief and variance of the agent varies along the chain.

Using the recursive update in Sec. 4.2 and decision threshold (4.28), it is possible to obtain the Bayes risk of Blake (i.e., $N = 2$) for given beliefs q_1, q_2. Fig. 4.3 depicts Blake's Bayes risk for $q_1, q_2 \in (0, 1)$, and explicitly shows that knowing true prior probability is not optimal. The social learning problem with Bayes costs $c_{10} = c_{01} = 1$, prior $p_0 = 0.3$, and additive Gaussian noise with zero mean and unit variance results in a Bayes risk that is minimum when Alexis's belief is 0.38 and Blake's belief is 0.23 (triangle), as shown in the figure where it is also compared to the true prior (circle).

Figs. 4.4 and 4.5 show the trend of optimal belief pair that minimizes the last agent's Bayes risk, when all agents have the same noise levels for the case of two and three agents respectively. We can observe several common characteristics. First, the non-terminal agents (i.e., Alexis for $N = 2$ and Alexis and Blake for $N = 3$) overweight their beliefs if p_0 is small and underweight it if p_0 is large. We call this *open-minded* behavior as it enhances less likely events. Second, the last agent (i.e., Blake for $N = 2$ and Chuck for $N = 3$) underweights the belief if p_0 is small and overweights it if p_0 is large, implicitly compensating for the biases of the preceding agents. Such behavior is referred to as being *closed-minded* as it represents a cautious outlook to the decision-making problem. Lastly, there is a unique, non-trivial prior, $p_0 \in (0, 1)$, where all agents' optimal beliefs are identical to the true prior.

However, the case of nonidentical noise variances of agents results in a very different behavior of optimal beliefs, especially when the last agent has smaller noise. The optimal beliefs for $N = 2$ and the case of the preceding agent having smaller noise, and that of

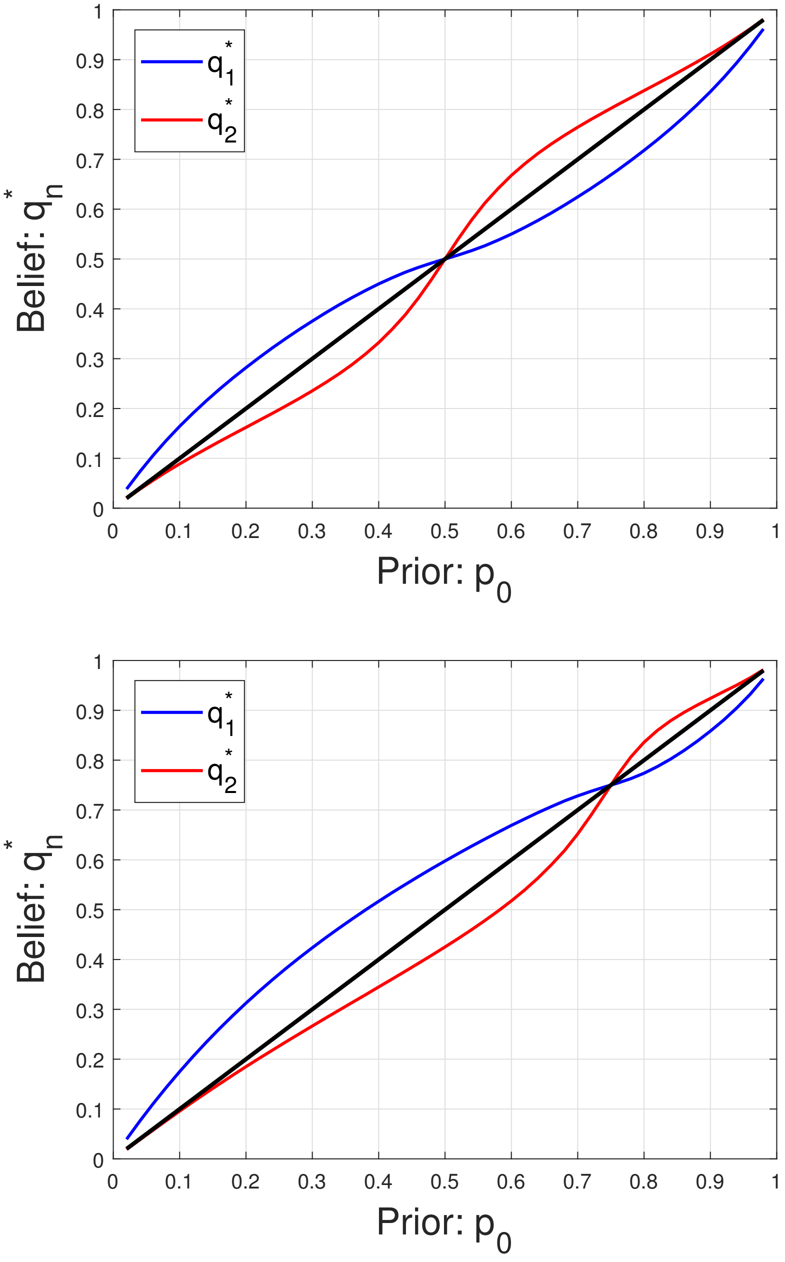


Figure 4.4: The trend of the optimal beliefs for $N = 2$ (Alexis, Blake). Z_1, Z_2 are standard Gaussian. Top panel: $c_{10} = c_{01} = 1$. Bottom panel: $c_{10} = 1, c_{10} = 3$.

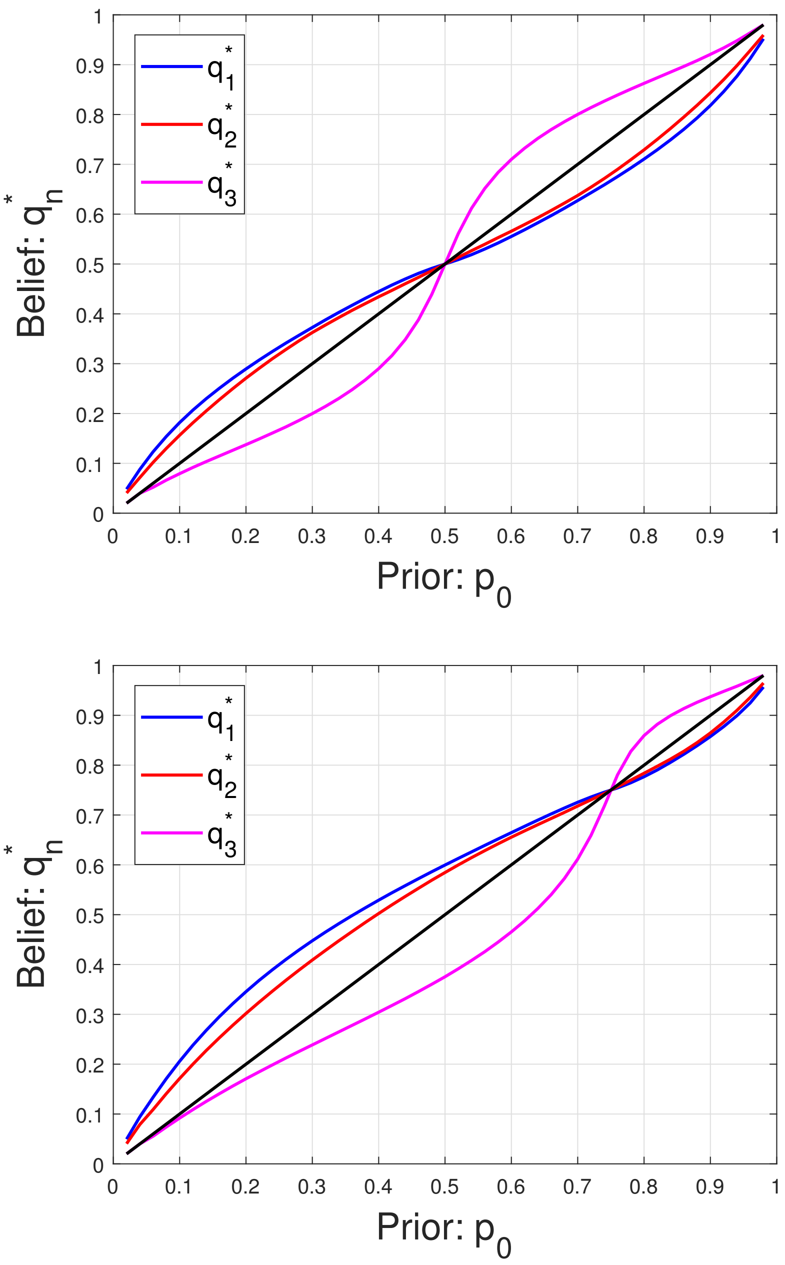


Figure 4.5: The trend of the optimal beliefs for $N = 3$ (Alexis, Blake, and Chuck). Z_1, Z_2, Z_3 are standard Gaussian. Top panel: $c_{10} = c_{01} = 1$. Bottom panel: $c_{10} = 1, c_{10} = 3$.

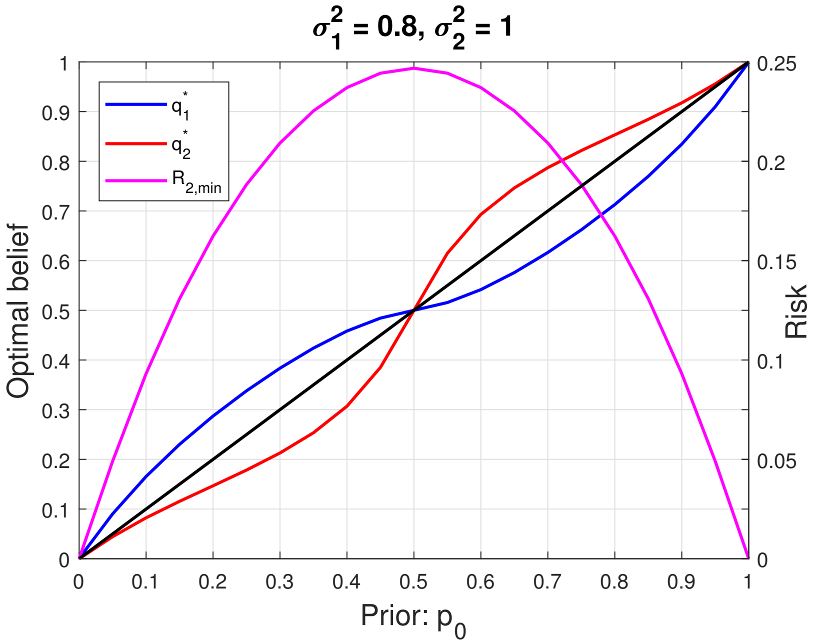


Figure 4.6: Optimal beliefs when the preceding agent has smaller noise, where $\sigma_1^2 = 0.8$ and $\sigma_2^2 = 1$.

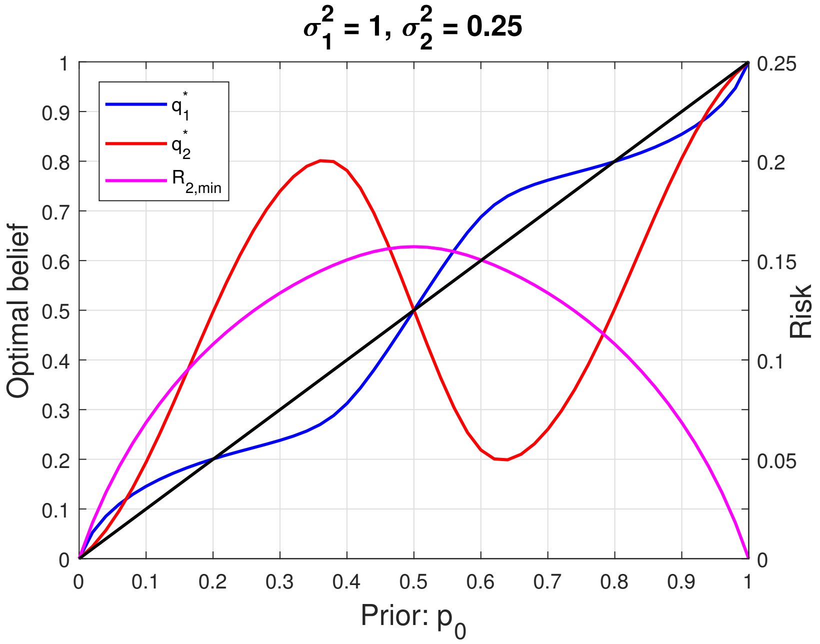


Figure 4.7: Optimal beliefs when the later-acting agent has smaller noise, where $\sigma_1^2 = 1$ and $\sigma_2^2 = 0.25$.

the last agent having smaller noise are shown in Figs. 4.6 and 4.7, respectively. As can be observed, the optimal belief curves are markedly different when the last agent has smaller noise, and we now derive some analytical properties of q_1^*, q_2^*.

Theorem 15. *For any σ_1^2 and σ_2^2, q_1^* and q_2^* satisfy:*

1. *for $p_0 \in (0,1)$, $q_1^* \leq p_0$ if and only if $q_2^* \geq \frac{c_{01}}{c_{01}+c_{10}}$, with equality for $q_2^* = \frac{c_{01}}{c_{01}+c_{10}}$.*
2. *$p_0 = q_1^* = q_2^*$ if and only if $p_0 \in \left\{0, \frac{c_{01}}{c_{01}+c_{10}}, 1\right\}$.*

Proof. Given in App. C.1. □

Thm. 15 highlights the fact that if the last agent believes the null hypothesis is more likely, then the ideal predecessor underweights the prior, and vice versa. Additionally, for p_0 near zero (near one) the optimal predecessor overweights (underweights) the prior.

In particular, let us consider two cases separately. First, let the predecessor have smaller noise. Then the curves for optimal beliefs and the corresponding Bayes risk are as shown in Fig. 4.6. The behavior here is similar to the case with equal noise, indicating that the reducing noise of the predecessor does not alter the overall behaviors of beliefs, as the last agent is unaware of this improved signal quality.

On the other hand, when the last agent has smaller noise, we notice that the nature of curves changes, as shown in Fig. 4.7. The behavior of the ideal agents indicates that when the predecessor has significantly larger noise than the last agent, the last agent stays open-minded. In addition, q_1^* has multiple crossings with p_0, but q_2^* has a single crossing at $q_2^* = c_{01}/(c_{01}+c_{10})$.

As expected, the ideal predecessor is open-minded for near-deterministic priors (p_0 close to zero or one). However, when the prior uncertainty in the hypotheses is high (p_0 near 1/2), we note that the ideal last agent with less noise favors the less likely hypothesis. This can be attributed to the fact that the last agent stays open-minded to the less likely hypothesis when the predecessor with larger noise is more likely to make errors. To further understand the nature of such a predecessor, we characterize the crossings of the optimal belief curve with the prior $q_1^* = p_0$.

Theorem 16. *The set of all p_0 such that $q_1^* = p_0$, $q_2^* = \frac{c_{01}}{c_{01}+c_{10}}$ is given by the solutions to*

$$e^x = \frac{1 - \beta Q(-\alpha + \sigma_1 x)}{1 - \beta Q(-\alpha - \sigma_1 x)}, \tag{4.29}$$

where

$$x = \log\left(\frac{c_{10}p_0}{c_{01}(1-p_0)}\right),\ \alpha = \frac{1}{2\sigma_1},\ \beta = 1 - \frac{Q\left(1/2\sigma_2\right)}{Q\left(-1/2\sigma_2\right)}.$$

Proof. Given in App. C.2. □

We note that $p^* = \frac{c_{01}}{c_{01}+c_{10}}$ is always a solution to (4.29). The case of multiple solutions to (4.29) is of particular interest and a sufficient condition is given in the following corollary.

Corollary 11. *If*

$$\frac{2\beta\sigma_1\phi(\alpha)}{1-\beta Q(-\alpha)} > 1, \tag{4.30}$$

then (4.29) *has at least* 3 *solutions in* $(0,1)$.

Proof. Since x is a monotonic function of p_0, it is sufficient to show that (4.29) has at least 3 solutions in x. From the symmetry in (4.29), since $x = 0$ is always a root, it suffices to show the existence of at least one more root in $x > 0$. First note the ranges of variables, $x \in (-\infty,\infty), \alpha \in (0,\infty), \beta \in (0,1)$.

Letting $r(x)$ be the right side of (4.29), since $0 \leq Q(\cdot) \leq 1$, we have

$$1-\beta \leq r(x) := \frac{1-\beta Q(-\alpha+\sigma_1 x)}{1-\beta Q(-\alpha-\sigma_1 x)} \leq \frac{1}{1-\beta},$$

indicating that $r(x) \in [1-\beta, \frac{1}{1-\beta}]$. However, note that e^x monotonically increases in $(1,\infty)$ for $x > 0$. Since $e^x, r(x)$ coincide at $x = 0$, it follows that they cross at least once on $(0,\infty)$ and also on $(-\infty,0)$, if $r'(x) > \frac{d}{dx}e^x$ at $x = 0$ by the intermediate value theorem. Thus, the sufficient condition follows:

$$r'(0) = \frac{2\sigma_1\beta\phi(\alpha;0,1)}{1-\beta Q(-\alpha)} > 1 = \left.\frac{d}{dx}e^x\right|_{x=0}.$$

□

Cor. 11 provides a sufficient condition on the noise level of agents under which there exist multiple crossings of the curves $q_1^*(p_0)$ and p_0. The range of standard deviations of the additive Gaussian noise of the preceding and last agents that satisfy the sufficient condition of Cor. 11 is shown in Fig. 4.8. Note from the figure that the area below the red dotted contour in Fig. 4.8 has multiple solutions to $q_1^* = p_0$, i.e., when the last agent has comparatively smaller than the preceding agent.

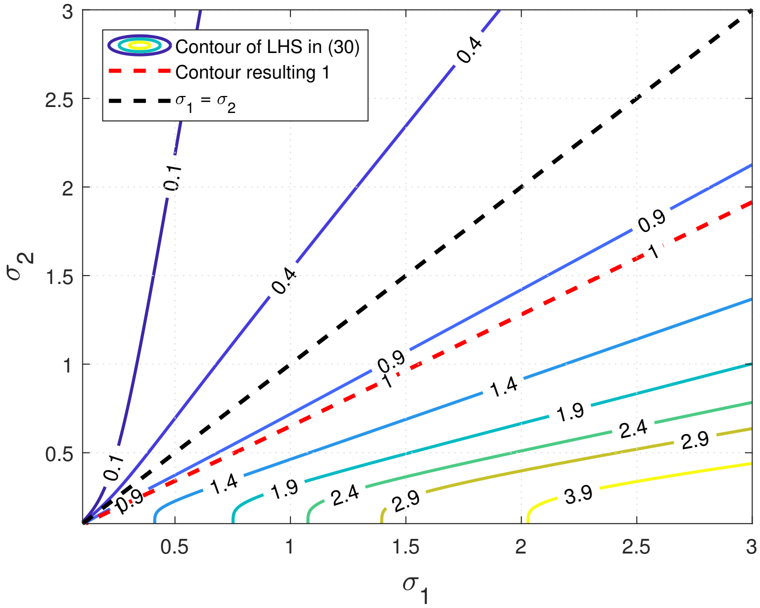


Figure 4.8: Contour plot of (4.30) with values for various σ_1, σ_2. The red dotted contour shows the contour that results in 1 so that the area below it satisfies (4.30) and therefore has multiple solutions to (4.29).

This is important as the crossings indicate a change in the perceived bias of the predecessor and also indicate the regions in which the last agent overweights the unlikely hypothesis as in Fig. 4.7.

4.5 Team Construction Criterion

Having studied the mathematical conditions for optimal reweighting of initial beliefs, we now investigate team selection for social learning. Naturally, a social planner who is aware of the context p_0 can pick the optimal agent pairs to minimize Bayes risk. However, it is not clear if agents are capable of organizing themselves into ideal teams in the absence of contextual knowledge. Thus, we now identify the criterion for the last agent to identify the optimal predecessors among a set of given predecessors.

Theorem 17. *Consider two predecessors with $q_1 < q_{1'}$. Let $\lambda_1, \lambda_{1'}$ be the decision thresholds of the respective predecessors. Then, the predecessor with belief q_1 is the optimal choice if* *and only if*

$$\frac{\mathbb{P}_1\left[Y_1 \in [\lambda_1, \lambda_{1'}], Y_2 \in [\lambda_2^1, \lambda_2^0]\right]}{\mathbb{P}_0\left[Y_1 \in [\lambda_1, \lambda_{1'}], Y_2 \in [\lambda_2^1, \lambda_2^0]\right]} \geq \frac{c_{10} p_0}{c_{01}(1-p_0)}. \tag{4.31}$$

Proof. Given in App. C.3. □

In other words, by rewriting (4.31) in a likelihood ratio form, we observe that the criterion for picking the predecessor with a smaller belief is given by the likelihood ratio test

$$\mathcal{L}\left[\widehat{H}_1 = \widehat{H}_2 = 1, \widehat{H}_{1'} = \widehat{H}_{2'} = 0\right] \geq \frac{c_{10} p_0}{c_{01}(1-p_0)},$$

where $\widehat{H}_{2'}$ is the decision made by the last agent following the decision of the predecessor with belief $q_{1'}$.

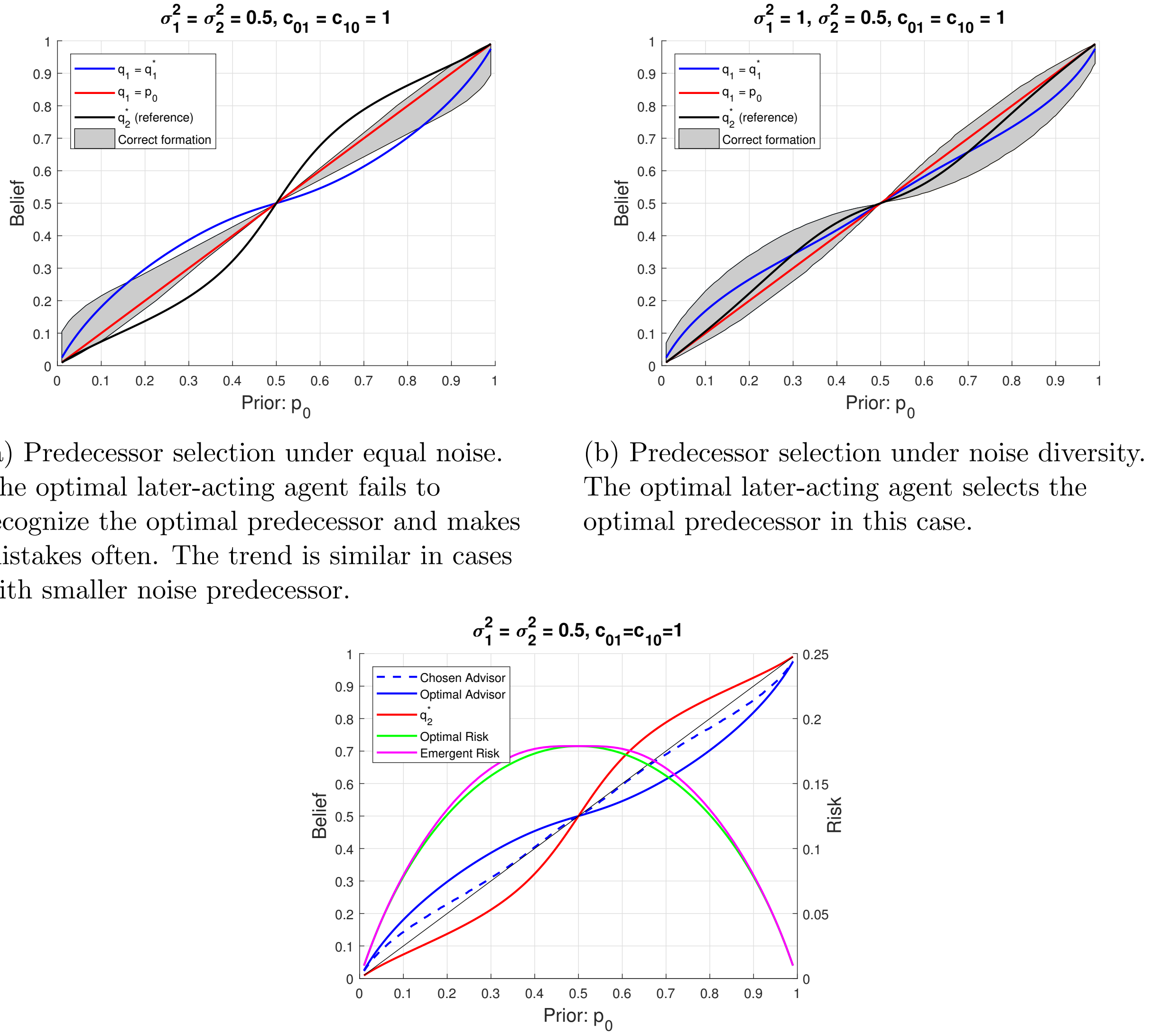
Thus selecting an ideal predecessor requires a social planner who is aware of the context p_0. Without this, the last agent selects a predecessor according to his personal belief q_2. That is, the last agent verifies condition (4.31) by replacing p_0 by q_2. Such a choice of predecessor might not always conform to the optimal choice when the belief of the last agent deviates significantly from the prior. To illustrate, we consider the problem of choosing between two predecessors with beliefs $q_1(p_0) = q_1^*(p_0)$ and $q_{1'}(p_0) = p_0$. Let $q(p_0, q_2)$ be the belief of the optimal predecessor choice for a given pair (p_0, q_2). We identify the region of correct selection by shading, $\mathcal{S} = \{(p_0, q_2) : q(p_0, q_2) = q(q_2, q_2)\}$.

First, when noise levels are equal, the region in which the last agent picks the correct preceding agent is shown in Fig. 4.9a. We note that the correct region is relatively small and does not include q_2^*. In particular, the last agent with optimal belief chooses the wrong predecessor always, whereas a suboptimal last agent with beliefs in the shaded region picks the correct one.

On the other hand, when the last agent has smaller noise than the predecessor, the corresponding region is as shown in Fig. 4.9b. Here we note that the last agent with optimal belief picks the correct preceding agent always.

Thus, we note that knowledge of the mathematically optimal beliefs does not guarantee selection of the right preceding agent. Further, we also observe that the diversity of noise levels may increase the feasibility of selecting the right preceding agent when the last agent has optimal belief.

We also explore the optimal choice of predecessor for the given optimal last agent in the absence of knowledge of the prior probability. From (4.25), the belief of the optimal preceding



(a) Predecessor selection under equal noise. The optimal later-acting agent fails to recognize the optimal predecessor and makes mistakes often. The trend is similar in cases with smaller noise predecessor.

(b) Predecessor selection under noise diversity. The optimal later-acting agent selects the optimal predecessor in this case.

(c) Context-unaware predecessor selection. The last agent chooses predecessor using (4.32) without p_0 and the Bayes risk increases as a result.

Figure 4.9: Context-unaware team selection.

agent, $\tilde{q}_1$ chosen by a last agent, in the absence of context (prior probability p_0) satisfies

$$\frac{\tilde{q}_1}{1-\tilde{q}_1} = \frac{p_0}{1-p_0} \frac{P_{e,2}^{\mathrm{I}_1} - P_{e,2}^{\mathrm{I}_0}}{P_{e,2}^{\mathrm{II}_0} - P_{e,2}^{\mathrm{II}_1}}. \tag{4.32}$$

The last agent's behavior with belief q_2^* is as shown in Fig. 4.9c. We note that the preceding agent chosen by the last agent differs from the optimal choice. Further, it is also evident that this choice consequently results in an increased Bayes risk. Such behavior in team selection highlights the significance of context, and thus of a social planner, for identifying the right team.

4.6 Human-AI Collaboration Systems

In this section, we use mathematical results from previous sections to study the engineering design problem of constructing human-AI collaborative systems. To do so, we make the following assumptions from the behavioral sciences: Human agents perform Bayesian decision-making [85–88] and their perceptions follow the Prelec reweighting function [55]. In addition, agents experience varying observational noise which is additive and Gaussian (as it is a common model in human signal perception [89, 90]). As usual in sequential social learning setup, all agents make selfish decisions [29, 33].

4.6.1 Approximation by Prelec Family

To design human-AI collaborative systems, we first determine whether optimal belief functions from previous sections are close to human behavior as modeled by cumulative prospect theory [52, 55].[4]

We approximate the optimal belief curves q_n^* by the Prelec function and study the resulting increase in the Bayes risk. We restrict to the Prelec family whose fixed point is identical

[4] Bounded rationality models have been categorized into two main classes—costly bounded rationality and truly bounded rationality [48]. Costly rationality considers the emergence of boundedly irrational behavior as optimization under some costs of decision-making such as computation and communication. On the other hand, truly bounded rationality is not based on an optimization framework. Though not the focus of the present chapter, one might wonder whether people are (approximately) naturally optimal for social learning. That is: Since the optimal belief curves result from limitations in computation (selfish decision-making) and communication (public signal quantization), do cumulative prospect-theoretic models emerge from a costly rationality framework for social learning?

to $p^* = \frac{c_{10}}{c_{01}+c_{10}}$, and then find best parameters (α_n, β_n) in the minimax absolute error sense, i.e.,

$$(\alpha_n, \beta_n) = \operatorname*{arg\,min}_{\alpha,\beta: w(p^*;\alpha,\beta)=p^*} \|q_n^*(\cdot) - w(\cdot;\alpha,\beta)\|_\infty.$$

Let the Prelec function approximations be $(q_{1,\mathrm{Pre}}, q_{2,\mathrm{Pre}})$.

The Prelec approximations for the two-agent case are shown in dotted curves in Fig. 4.10. When the preceding agent has smaller noise as in the top panel of Fig. 4.10, the Prelec function approximates the optimal beliefs well and the Bayes risk does not increase by much. To evaluate the loss from the approximation, consider the set of correct beliefs $q_1 = q_2 = p_0$, that result in a Bayes risk of $R_{2,\mathrm{corr}}$. The maximal loss in terms of Bayes risk from using the correct beliefs is $\max_{p_0}(R_{2,\mathrm{corr}} - R_{2,\mathrm{min}}) \approx 0.0039$. On the other hand, the maximal loss from the best Prelec approximation is ≈ 0.0009. This indicates that the natural cognitive biases of humans (i.e., Prelec reweighting) are effective for social learning when the preceding agent has smaller noise.

On the other hand, when the last agent has smaller noise as in the bottom panel of Fig. 4.10, the Prelec approximation does not accurately mimic the optimal behavior of agents. Recall that the Prelec function is always increasing and has only one crossing with unit slope line in $(0, 1)$. Therefore, the Prelec function fails to account for all the variations in the optimal belief. Moreover, while the additional loss of Bayes risk by the Prelec fitting is ≈ 0.0187, the loss from using the correct beliefs, $p_0 = q_1 = q_2$, is ≈ 0.0060. This indicates that even though the Prelec weighting functions serve as good approximations with predecessors having less noisy observations, they do not model the optimal behavior in the case of predecessors having noisier observations. These results suggest that human agents following cumulative prospect theory models [52] yield small Bayes risk when predecessors have smaller noise.

4.6.2 Human-AI Teams

The previous subsection informs the design of AI-human collaboration structures [56]. In many human-AI joint teams, a human agent makes the final decision based on the advice of an AI component as depicted in Fig. 4.11a, but the opposite structure of Fig. 4.11b is also possible. It is thus important to identify the best team configuration [91]. Indeed, D. Kahneman recently stated that "You can combine humans and machines, provided the

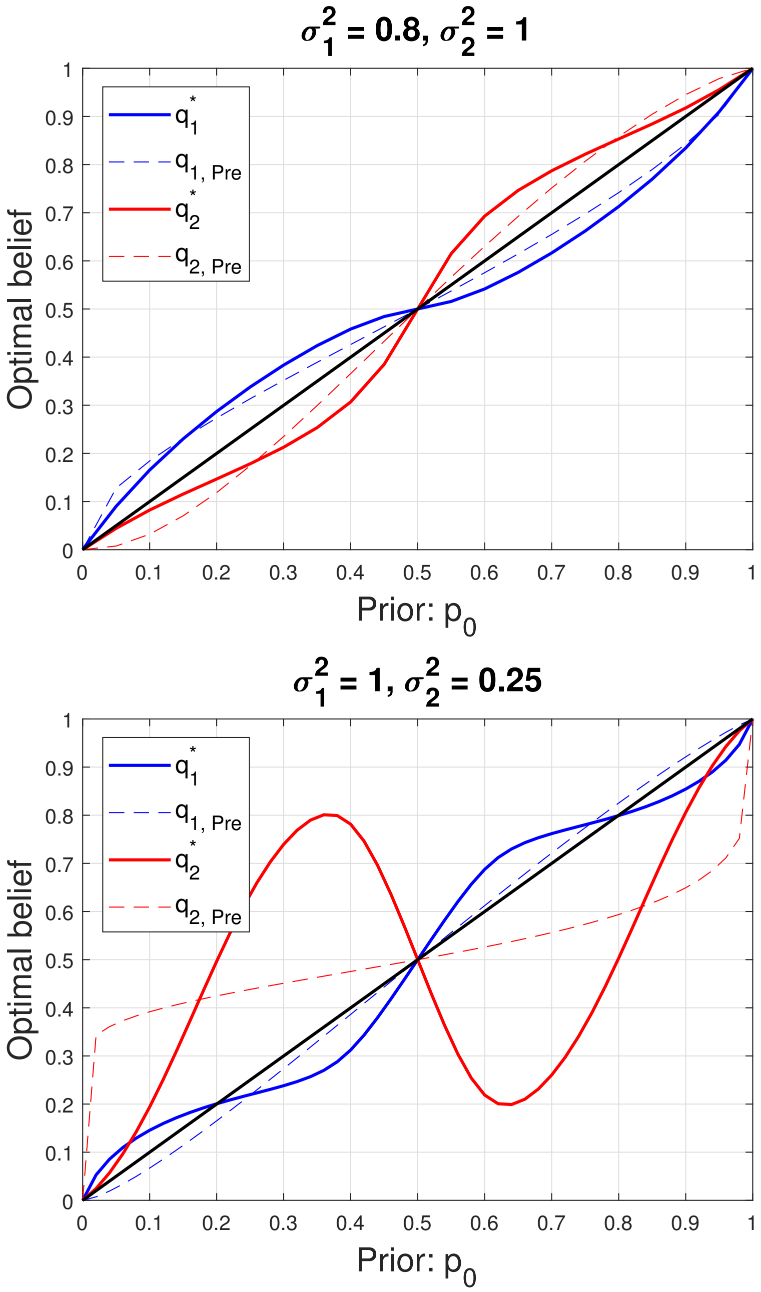
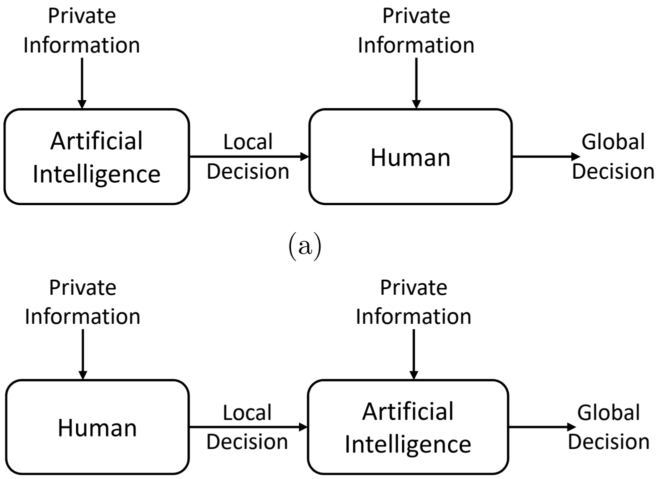


Figure 4.10: Optimal beliefs as compared to Prelec-weighted beliefs. Top panel: When the preceding agent has smaller noise. Bottom panel: When the later-acting agent has smaller noise.



(a)

(b)

Figure 4.11: Models of AI-human collaboration, where a machine provides input for human judgement or vice versa.

machine has the last word" [92].

Our results indicate that an AI assistant with smaller noise could be an effective predecessor to the human decision-maker. In particular, an open-minded AI predecessor and a closed-minded human final decision-maker with appropriate Prelec reweighted beliefs work well together, as in Fig. 4.6. However, an AI component with greater noise might not be a good predecessor to the human last agent who does not have beliefs that mimic the optimal behavior in Fig. 4.7 and so, perhaps counterintuitively, the architecture of Fig. 4.11b should be adopted, with the AI agent having larger noise making the global decision.

Additionally, these results along with those of Sec. 4.5 provide some insight into human-AI teams when the human agent picks an AI predecessor, given a choice among different agents. In particular, consider the AI-human team where the human, who has a Prelec-weighted belief, chooses one of two possible AI predecessors—one that has the optimal belief q_1^* and the other that is aware of the true prior p_0. In case the human agent has larger noise, and a closed-minded Prelec belief as in Fig. 4.9a, she unfortunately picks the AI predecessor with $q_1 = p_0$ and the team becomes suboptimal. However, if the human agent has smaller noise, and an open-minded Prelec belief, she picks the optimal AI component $q_1 = q_1^*$ and therefore can make the optimal decision as in Fig. 4.9b. Thus it is evident that optimal team organization is feasible when the human has smaller noise and the appropriate open-minded belief.

4.7 Chapter Summary

We discuss the sequential social learning problem with individual biased beliefs. Unlike previous works on herding, we focus on the Bayes risk of the last-acting agent. We first derive the optimal belief update rule for general likelihoods and evaluated for Gaussian likelihoods. Counterintuitively, optimal beliefs that yield minimum Bayes risk are in general different from the true prior. Under equal expertise levels, we observe that optimal advisors have open-minded beliefs, that is, they overweight small priors and underweight large priors, while the optimal advisee has closed-minded belief. However, the trend may change depending on varying expertise levels such that, especially when the advisee has much more expertise, optimal belief of the advisee is inverted as she becomes open-minded.

We also show that the Prelec reweighting function from cumulative prospect theory approximates the behavior of the optimal beliefs under specific levels of expertise; however, when the advisee has much more expertise, it fails to capture all the behavioral traits of the optimal beliefs.

Finally, we consider the ability of agents to organize themselves into optimal teams and show that in the absence of a social planner, the advisee can get paired with the wrong advisor when the individual belief deviates significantly from the underlying prior value. The setup arises from the consideration of AI and it tells us that, without knowing the true prior, our human-machine team construction could be misorganized.

Chapter 5

The CEO Problem with rth Power of Difference and Logarithmic Distortions

In this chapter, we explore two CEO problems that differ from the prior works listed in Sec. 1.1 in that the models not only have a non-Gaussian source-observation pair, but also have general rth power difference distortion $d(x, \hat{x}) = |x - \hat{x}|^r$ or logarithmic distortion. The models and our contributions are briefly summarized here.

- (Sec. 5.2, *regular* model) Continuous source and observation supported on $\mathbb{R}$ satisfying some regularity conditions, including the jointly Gaussian CEO problem [57, 58, 60], but with $|x - \hat{x}|^r$ distortion: The distortion scales as $R_{\mathsf{sum}}^{-r/2}$. Achievability is by the Berger-Tung scheme [93] and median estimator and converse is by the Shannon lower bound [94].

- (Sec. 5.3, *non-regular* model) Bounded source and observation such that estimation-theoretic regularity conditions do not hold, including copula [63] or additive uniform noise model with $|x - \hat{x}|^r$ distortion: The distortion scales as R_{sum}^{-r}. Achievability is by the Berger-Tung scheme and midrange estimator [95] and converse is by the Chazan-Ziv-Zakai bound [96, 97].

- (Sec. 5.4, equivalence) The regular model as in Sec. 5.2: If test channels satisfy some conditions, quadratic (i.e, $r = 2$) and logarithmic distortions are asymptotically equivalent as $L \to \infty$, bridged by entropy power relation $D_{\mathsf{Q}} = \frac{1}{2\pi e} 2^{D_{\mathsf{Log}}}$, where $D_{\mathsf{Q}}, D_{\mathsf{Log}}$ are quadratic and logarithmic distortions, respectively. It also implies logarithmic distortion decays as $-\log R_{\mathsf{sum}}$.

With the results of [57, 58, 60, 61], our results suggest that the Berger-Tung achievable scheme might be asymptotically optimal even for various types of models, not listed here. Furthermore, noting that the jointly Gaussian model, a special case of regular models, is the worst model among additive noises [65], we can conclude that other regular models are not much easier to estimate since they all have $R_{\mathsf{sum}}^{-r/2}$ asymptotics. It is possible to further argue that regular models are essentially the worst model among all variance-bounded additive

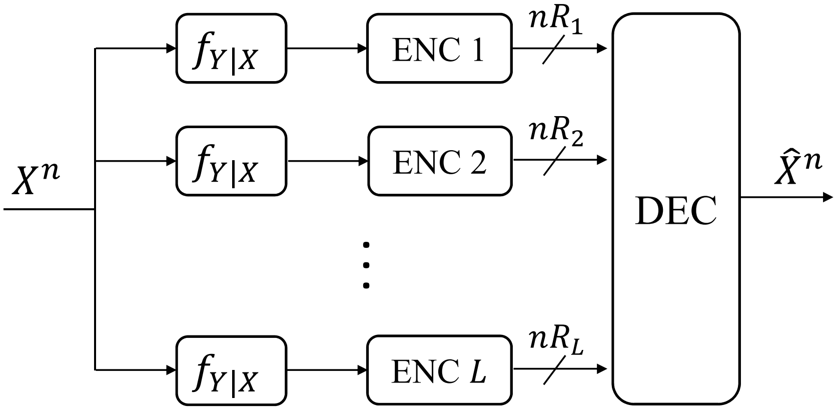


Figure 5.1: The CEO problem model with L agents.

noise models (not necessarily regular) in the sense of sum rate asymptotics by the argument of [65]. In contrast, non-regular models that have R_{sum}^{-r} are easier to estimate than regular models. The equivalence of the two distortions is interesting since we already know the entropy power inequality $\mathsf{Var}(X|Z) \geq \frac{1}{2\pi e} e^{h(X|Z)}$ [70], where the left and right sides are interpreted as quadratic and logarithmic distortions respectively, but the equivalence shows asymptotically equality.

5.1 CEO Problem Formulation

We consider the CEO problem as in [39], but with real-valued alphabets, i.e., $\mathcal{X}, \mathcal{Y} \subset \mathbb{R}$. The source $\{X(t)\}_{i=1}^{\infty}$ that the CEO is interested in is independent and identically distributed (i.i.d.) from a density function $f_X(x)$. There are L agents who collect the source information, but the ith agent is only given a noisy version $\{Y_i(t)\}_{t=1}^{\infty}$, i.i.d. drawn from a common observation distribution $f_{Y|X}$. The agents encode observations separately into messages of rate $\{R_i\}_{i=1}^{L}$; more precisely, the ith agent encodes a length n block of observations into a codeword C_i from codebook $\mathcal{C}_i$ of rate R_i and proceeds to send the codeword index. Sum rate of the link to the CEO is limited to $R_{\mathsf{sum}} = \sum_{i=1}^{L} R_i$. Upon receiving codewords from agents, the CEO wishes to estimate $\{\hat{X}(t)\}_{t=1}^{n}$ that minimizes the expected distortion of length n,

$$D^n(X^n, \hat{X}^n) := \frac{1}{n}\mathbb{E}\left[\sum_{t=1}^{n} |X(t) - \hat{X}(t)|^r\right] = \frac{1}{n}\sum_{t=1}^{n} \mathbb{E}\left[|X(t) - \hat{X}(t)|^r\right],$$

where $\hat{X} \in \mathcal{X}$, if the distortion is rth power of difference distortion.

The other distortion measure in this chapter is logarithmic distortion, which commonly arises in machine learning literature and also recently in information theory [61],

$$D^n(X^n, \hat{X}^n) := \frac{1}{n}\mathbb{E}\left[\sum_{t=1}^{n} -\log \hat{X}(X;t)\right] = \frac{1}{n}\sum_{t=1}^{n} \mathbb{E}\left[-\log \hat{X}(X;t)\right],$$

where $\hat{X}$ is a probability distribution over $\mathcal{X}$, i.e., $\hat{X} \in \mathcal{P}(\mathcal{X})$, where $\mathcal{P}(\mathcal{X})$ denotes the probability distribution space over $\mathcal{X}$. The problem model is illustrated in Fig. 5.1.

In this work, we are interested in the asymptotic tradeoff between R_{sum} and $D^n(X^n, \hat{X}^n)$. To see this, define

$$\begin{aligned} D^n(L, R_{\mathsf{sum}}) &:= \min_{\{\mathcal{C}_i\}_{i=1}^{L}:\sum_{i=1}^{L} R_i \le R_{\mathsf{sum}}} D^n(X^n, \hat{X}^n), \\ D(L, R_{\mathsf{sum}}) &:= \lim_{n\to\infty} D^n(L, R_{\mathsf{sum}}). \end{aligned}$$

As we will see, $D(L, R_{\mathsf{sum}})$ asymptotically vanishes as L, R_{sum} grow without bound, but keeping the average individual rate R_{sum}/L unchanged. So we investigate the following quantities:

$$\begin{aligned} \beta_{\mathsf{reg}} &:= \lim_{L, R_{\mathsf{sum}}\to\infty} R_{\mathsf{sum}}^{r/2} D(L, R_{\mathsf{sum}}) \quad \text{in Sec. 5.2}, \\ \beta_{\mathsf{n\text{-}reg}} &:= \lim_{L, R_{\mathsf{sum}}\to\infty} R_{\mathsf{sum}}^{r} D(L, R_{\mathsf{sum}}) \quad \text{in Sec. 5.3}. \end{aligned}$$

Hence, if β_{reg} and $\beta_{\mathsf{n\text{-}reg}}$ are constant, it tells us that the speeds of distortion decay are $R_{\mathsf{sum}}^{-r/2}$ and R_{sum}^{-r} for regular and non-regular models, respectively.

Before proceeding with formal definitions of regular and non-regular models in the following sections, recall one of the regularity conditions of the Fisher information (and thus the Cramer-Rao lower bound) [98, Sec. 2.5]:

> The support of $f_{Y|X}$ is common for all x, i.e., the set $\{y : f_{Y|X}(y|x) > 0\}$ is independent of x.

In this context, a model is called *regular* if it satisfies the above condition as well as other conditions in Sec. 5.2, whereas it is *non-regular* if the above does not hold, but conditions in Sec. 5.3 hold. Note that these two definitions do not form a disjoint partition, and there are examples that are neither regular nor non-regular

In the sequel, f, p denote continuous and discrete probability densities, respectively. We will use the natural logarithm so that the unit of information rate is nats. Hat notation $\hat{V}$ is for estimated values and tilde notation $\widetilde{V}$ is for quantized values. The function $q(\cdot)$ also stands for the quantization function so $q(V)$ and $\widetilde{V}$ are interchangeable. The round bracket subscript $V_{(i)}$ denotes ith order statistics, that is, reordered sequence from $\{V_i\}_{i=1}^L$ in increasing order $V_{(1)} \leq V_{(2)} \leq \cdots \leq V_{(L)}$. When $L = 2m+1, m \in \mathbb{Z}_+$, $V_{(m+1)}$ is the sample median and it is often denoted by $\mathsf{med}(\{V_i\}_{i=1}^L)$. Also the true median of f_V is denoted by $\mathsf{med}(V)$ with abuse of notation.

5.2 Regular CEO Problem

5.2.1 Model and Result

We consider unbounded source and observation alphabets $\mathcal{X} = \mathcal{Y} = \mathbb{R}$ and impose regularity conditions on probability distributions that enable us to characterize β_{reg} explicitly. Let us first state source and observation conditions (A1)–(A4).

(A1) The source has a finite absolute moment of order r.

(A2) The density f_X is continuous and positive almost everywhere in $\mathbb{R}$ and the density $f_{Y|X}$ is twice continuously differentiable with respect to x for almost every $x \in \mathbb{R}$ and almost every $y \in \mathbb{R}$.

(A3) For almost every $x \in \mathbb{R}$,

$$\mathbb{E}_{Y|x}\left[\left|\frac{\partial}{\partial x}\log f_{Y|X}(Y|x)\right|^2\right] < \infty \quad \text{and} \quad \mathbb{E}_{Y|x}\left[\left|\frac{\partial^2}{\partial x^2}\log f_{Y|X}(Y|x)\right|^2\right] < \infty,$$

and the Fisher information $I_Y(x) := \mathbb{E}_{Y|x}[(\frac{\partial}{\partial x}\log f_{Y|X}(Y|x))^2]$ is well-defined, finite, and positive for almost every $x \in \mathbb{R}$.

(A4) The posterior distribution of x given Y^n asymptotically concentrates on the true value sufficiently fast for every $x \in \mathbb{R}$. Formally speaking, for any $\delta > 0, x \in \mathbb{R}$

$$\mathbb{P}[f_{X|Y^n}[N_x^c] > \delta] = o(1/\log n),$$

for every open set N_x containing the true x.

Condition (A1) is necessary not only for technical evaluation, but also for the rate-distortion formulation as in [99]. Conditions (A2)–(A4) are smoothness conditions that enable us to characterize asymptotics explicitly, especially (A4) leads to a simple expression of Lem. 14.

Next we impose some conditions for the existence of an auxiliary random variable U that satisfies some properties. Recall that $\mathsf{med}(U|x)$ is the median of $f_{U|X=x}$, i.e.,

$$\int_{-\infty}^{\mathsf{med}(U|x)} f_{U|X}(u|x)du = \frac{1}{2}.$$

(A5) The Markov chain $X - Y - U$ holds and U has a finite absolute moments of order r.

(A6) Medians of $f_{U|x_0}, f_{U|x_1}$ are distinct when $x_0 \neq x_1$. In addition, the function $u = \ell(x) := \mathsf{med}(U|x)$ that maps x to the median of $f_{U|x}$ is bi-Lipschitz continuous, i.e., $\ell(\cdot)$ and $\ell^{-1}(\cdot)$ are both Lipschitz. Suppose ℓ^{-1} has a Lipschitz constant $K > 0$.

(A7) For some positive constant c, it holds that $\alpha := \inf_{x\in\mathbb{R}} f_{U|x}(\mathsf{med}(U|x)|x) > c$.

Define $\mathcal{S}_{\mathsf{reg}}$ to be the set of Us that satisfy (A5)–(A7). Condition (A5) enables forward test channels in compression step. Also since we will use median estimation, conditions (A6) and (A7) are technically required because upon obtaining the exact median of U conditioned on x, one should be able to recover x from it. The Lipschitz property also guarantees that when error in estimating the median of U is small, error in X is small as well up to the Lipschitz constant factor. If one adopts another estimation scheme such as mean estimation or maximum likelihood estimation, different conditions will be required. It is however remarkable that (A1)–(A7) all hold for the Gaussian CEO problem with additive Gaussian test channel as in [57], where sample mean is used.

As mentioned, the distortion measure we will consider is the rth power of difference, i.e.,

$$d(x, \hat{x}) = |x - \hat{x}|^r,$$

under which our main result of this section is the following.

Theorem 18 (Regular CEO problem)**.** *Suppose conditions (A1)–(A4) hold for source and observation model and suppose there exists U such that (A5)–(A7) hold. Then, for distortion measure $d(x, \hat{x}) = |x - \hat{x}|^r$,*

$$C_1 \left(\min_{U:X-Y-U} I(Y;U|X) \right)^{r/2} \leq \beta_{reg} \leq C_2 \left(\min_{U\in\mathcal{S}_{reg}} I(Y;U|X) \right)^{r/2},$$

where

$$C_1 = \frac{1}{re}\left(V_1\Gamma\left(1+\frac{1}{r}\right)\frac{e^{-\frac{1}{2}\mathbb{E}[\log\det I_Y(X)]}}{\sqrt{2\pi e}}\right)^{-r},$$
$$C_2 = \left(\frac{K}{2\alpha}\right)^r 2^{3r/2}\frac{\Gamma(\frac{r+1}{2})}{\sqrt{\pi}},$$

and the minimum of the lower bound is taken over non-trivial random variables to ensure that the mutual information is non-zero.

5.2.2 Direct Coding Theorem

We will make use of standard achievable scheme in [57]. That is, first finely quantize continuous alphabets and apply Berger-Tung encoding and decoding over incurred discrete alphabets, and then, estimate the source. Our estimation is based on sample median estimator, which is the best for absolute distortion, i.e., $|x-\hat{x}|$. Suppose the number of agents is odd, i.e., $L = 2m+1, m \in \mathbb{Z}_+$ to simplify notation. Before proceeding, note that random variables $\{U_i\}_{i=1}^L$ are all generated through an identical test channel $f_{U|Y}$ that satisfies (A5)–(A7). This assumption does not lose optimality.[1]

Quantization

Quantizing the real line enables agents to use subsequent codes and Slepian-Wolf compression in discrete domain. Let $\widetilde{X}, \widetilde{Y}, \widetilde{U}$ denote the quantized versions of X, Y, U. We suppose our fine quantization ensures that the loss due to quantization is negligible. Formally, we take a quantization scheme that satisfies the following conditions: for some small $\delta_i > 0, i \in \{0,1,2\}$,

$$\mathbb{E}\left[|U_{(m+1)} - q(U_{(m+1)})|^r\right] \le \delta_0 \ \text{ and } \ \mathbb{E}\left[|q(U_{(m+1)}) - \mathsf{med}(\{q(U_i)\}_{i=1}^L)|^r\right] \le \delta_0 \tag{5.1}$$

$$|I(Y;U) - I(\widetilde{Y};\widetilde{U})| \le \delta_1, \tag{5.2}$$
$$|I(X;U) - I(\widetilde{X};\widetilde{U})| \le \delta_2. \tag{5.3}$$

[1] Suppose that nonidentical test channels achieve a smaller distortion D. As agents are symmetric, the distortion must be invariant under permutation. Time-sharing argument that averages nonidentical channels shows that identical test channels also achieve D, which yields a contradiction.

It is easy to see that there exists a quantization scheme with finite cardinality that satisfies (5.1) from the finite moment condition, as well as (5.2) and (5.3) from the definition of mutual information for arbitrary ensembles [100]; hence, a common refinement of quantization schemes satisfies all three conditions. This quantization also induces discrete probability distributions for $\widetilde{X}, \widetilde{Y}, \widetilde{U}$:

$$\begin{aligned}
p_{\widetilde{Y},\widetilde{U}}(\widetilde{y},\widetilde{u}) &= \int_{\{(y,u):q(y)=\widetilde{y},q(u)=\widetilde{u}\}} f_{Y,U}(y,u)dydu, \\
p_{\widetilde{X},\widetilde{U}}(\widetilde{x},\widetilde{u}) &= \int_{\{(x,u):q(x)=\widetilde{x},q(u)=\widetilde{u}\}} f_{X,U}(x,u)dxdu, \\
p_{\widetilde{Y}|X}(\widetilde{y}|x) &= \int_{\{y:q(y)=\widetilde{y}\}} f_{Y|X}(y|x)dy, \\
p_{\widetilde{U}|\widetilde{Y}}(\widetilde{u}|\widetilde{y}) &= \frac{p_{\widetilde{Y},\widetilde{U}}(\widetilde{y},\widetilde{u})}{p_{\widetilde{Y}}(\widetilde{y})}.
\end{aligned}$$

Spaces of $\widetilde{X}, \widetilde{Y}, \widetilde{U}$ are denoted by $\widetilde{\mathcal{X}}, \widetilde{\mathcal{Y}}, \widetilde{\mathcal{U}}$.

Codes Approximating Test Channel

Each agent takes block length n_0 and encodes quantized observation $\widetilde{Y}^{n_0}$ into a codeword, instead of Y^{n_0}. Let $\varphi : \widetilde{\mathcal{Y}}^{n_0} \mapsto \widetilde{\mathcal{U}}^{n_0}$ be the (possibly stochastic) block code encoder, common for all agents. This mapping induces the following empirical distributions:

$$\begin{aligned}
\hat{p}_{\widetilde{Y}^{n_0},\widetilde{U}^{n_0}}(\widetilde{y}^{n_0},\widetilde{u}^{n_0}) &= p_{\widetilde{Y}^{n_0}}(\widetilde{y}^{n_0})\mathbb{1}_{\{\varphi(\widetilde{y}^{n_0})=\widetilde{u}^{n_0}\}}, \\
\hat{p}_{\widetilde{Y},\widetilde{U}}(\widetilde{Y}(t)=\widetilde{y},\widetilde{U}=\widetilde{u}) &= \mathbb{E}_{p_{\widetilde{Y}^n}}\left[\mathbb{1}_{\{\widetilde{U}(t)=\widetilde{u},\widetilde{Y}(t)=\widetilde{y}\}}\right], \\
\hat{p}_{\widetilde{U}|\widetilde{Y}}(\widetilde{U}(t)=\widetilde{u}|\widetilde{Y}(t)=\widetilde{y}) &= \frac{\hat{p}_{\widetilde{Y},\widetilde{U}}(\widetilde{Y}(t)=\widetilde{y},\widetilde{U}(t)=\widetilde{u})}{p_{\widetilde{Y}}(\widetilde{Y}(t)=y)},
\end{aligned}$$

where $\mathbb{1}_{\{.\}}$ is the indicator function. Then the existence of a block code that approximates the true test channel $f_{U|Y}$ follows from [57, Prop. 3.1].

Proposition 12 ([57])**.** *For every $\epsilon, \delta > 0$, there exists a deterministic mapping $\varphi : \widetilde{Y}^{n_0} \mapsto \widetilde{U}^{n_0}$ with the range cardinality M such that*

$$\frac{1}{n_0}\log M \leq I(Y;U) + \epsilon$$

and

$$\sum_{\widetilde{u}\in\widetilde{\mathcal{U}}} |\hat{p}_{\widetilde{U}|X}(\widetilde{U}(t)=\widetilde{u}|x) - p_{\widetilde{U}|X}(\widetilde{U}(t)=\widetilde{u}|x)| \leq \frac{\epsilon}{|\widetilde{\mathcal{X}}|}$$

for all $t \in [1:n_0]$ and all $x \in \mathbb{R}$.

Encoding and Decoding

The overall encoding scheme is two-step as [39, 57]: in the first stage, each agent encodes $\widetilde{Y}_i^{n_0}$ into $\widetilde{U}_i^{n_0}$ by common $\varphi(\cdot)$. Note that $\{\widetilde{U}_i^{n_0}\}_{i=1}^L$ are correlated; the second stage performs Slepian-Wolf (or SW, for short) encoding to remove the correlation. Let $W_i \in \mathcal{W}$ be the index of the codeword $\widetilde{U}_i^{n_0}$. Formally speaking, the SW encoder at the ith agent is the mapping $\xi_i : \mathcal{W}^n \to \{0, 1, \ldots, N_i - 1\}$. Individual and sum rates are therefore defined to be

$$\begin{aligned} R_i &= \frac{1}{nn_0} \log N_i, \\ R_{\mathsf{sum}} &= \sum_{i=1}^{L} R_i = \frac{1}{nn_0} \sum_{i=1}^{L} \log N_i. \end{aligned}$$

The complete encoder of ith agent is given by

$$Z_i := \xi_i \circ \varphi^{n_0}(\widetilde{y}^{nn_0}) \in \{0, 1, \ldots, N_i - 1\}.$$

The CEO performs decoding in reverse: it recovers $\{\hat{U}_i^{nn_0}\}_{i=1}^L$ from $\{Z_i\}_{i=1}^L$, and then estimates X from $\{\hat{U}_i^{nn_0}\}_{i=1}^L$.

The next proposition ([57, Prop. 3.2 and Sec. III.D]) specifies the average individual rate upper bound in multi- and single-letter mutual information forms, and its error probability.

Proposition 13 ([57]). *For every $\epsilon, \lambda > 0$ and $\epsilon' > \epsilon$, there exists sufficiently large L, n and index encoders $\{\xi_i\}_{i=1}^L$ such that*

$$\begin{aligned} &\frac{R_{\mathsf{sum}}}{L} \leq \frac{1}{n_0} H(\widetilde{U}^{n_0}|\widetilde{X}^{n_0}) + \epsilon \leq I(Y;U|X) + \epsilon', \\ &\mathbb{P}[\mathcal{B}] \leq \lambda, \end{aligned}$$

where $\mathcal{B} := \{(\hat{U}_1^{n_0}, \ldots, \hat{U}_L^{n_0}) \neq (\widetilde{U}_1^{n_0}, \ldots, \widetilde{U}_L^{n_0})\}$ is the error event.

Estimation Upper Bound

If the CEO has the true $U_{(m+1)} = \mathsf{med}(\{U_i\}_{i=1}^L)$, the median of $\{U_1, \dots, U_L\}$, then she can uniquely determine X by mapping ℓ^{-1}. Therefore our goal is to estimate $U_{(m+1)}$ as accurately as possible from decoded $\{\hat{U}_i\}_{i=1}^L$. Note that for a given $X = x$, the true median of U is $\mathsf{med}(U|x) = \ell(x)$.

Lemma 10 (Median Estimator [101])**.** *Let F, f be the cumulative distribution and density function of V. Then, the sample median of $L = 2m+1$ samples follows the density function*

$$\mathbb{P}[V_{(m+1)} = v] = \frac{(2m+1)!}{m!m!}(F(v))^m(1-F(v))^m f(v) = \frac{(F(v))^m(1-F(v))^m}{B(m+1, m+1)} dF(v),$$

where $B(\cdot,\cdot)$ is the Beta function, so it is the $\mathsf{Beta}(m+1, m+1)$ distribution scaled by $F(v)$. Furthermore, $V_{(m+1)}$ is approximately Gaussian $\mathcal{N}\left(\mathsf{med}(V), \frac{1}{4Lf^2(\mathsf{med}(V))}\right)$ provided that L is large.

Lemma 11. *Under the notations of Lem. 10, the following holds when L is large:*

$$\mathbb{E}[|V_{(m+1)} - \mathsf{med}(V)|^r] \le \left(\frac{2}{Lf^2(\mathsf{med}(V))}\right)^{r/2} \frac{\Gamma(\frac{r+1}{2})}{\sqrt{\pi}} + \epsilon.$$

Proof. See App. D.1. □

Now we can derive the distortion asymptotics in terms of R_{sum}.

Theorem 19 (Achievability of Regular CEO Problem)**.**

$$\beta_{\mathsf{reg}} \le 2^{3r/2} \left(\frac{K}{\alpha}\right)^r \frac{\Gamma(\frac{r+1}{2})}{\sqrt{\pi}} \left(\min_{U \in \mathcal{S}_{\mathsf{reg}}} I(Y; U|X)\right)^{r/2}.$$

Proof. Given t, instantaneous error is bounded as follows. Since $\ell(\cdot)$ is the function that maps x to $\mathsf{med}(U|x) \in \mathcal{U}$ and $X = \ell^{-1}(\mathsf{med}(U|X))$, our estimation is $\hat{X} = \ell^{-1}(\hat{U}_{(m+1)}(t))$.

$$
\begin{aligned}
&\mathbb{E}\left[|X(t)-\hat{X}(t)|^r\right] = \mathbb{E}\left[|X(t)-\ell^{-1}(\hat{U}_{(m+1)}(t))|^r\right] \\
&\overset{(a)}{\leq} K^r \mathbb{E}\left[|\mathsf{med}(U|X)-\hat{U}_{(m+1)}(t)|^r\right] \\
&= K^r \mathbb{E}\left[|\mathsf{med}(U|X)-U_{(m+1)}(t)+U_{(m+1)}(t)-\hat{U}_{(m+1)}(t)|^r\right] \\
&\overset{(b)}{\leq} (2K)^r \mathbb{E}\left[|\mathsf{med}(U|X)-U_{(m+1)}(t)|^r|\right] + (2K)^r \mathbb{E}\left[|U_{(m+1)}(t)-\hat{U}_{(m+1)}(t)|^j\right] \\
&\overset{(c)}{\leq} (2K)^r \mathbb{E}\left[|\mathsf{med}(U|X)-U_{(m+1)}(t)|^r\right] + \epsilon_1,
\end{aligned}
$$

where (a) follows from the Lipschitz property of ℓ^{-1}; (b) follows from the triangle inequality and Prop. 18 in App. D.4; and (c) is proven as Prop. 15 in App. D.2.

Regarding the first term, since the median estimator is approximately Gaussian $\mathcal{N}\left(\mathsf{med}(U|X), \frac{1}{4Lf^2(\mathsf{med}(U|X))}\right)$ distributed,

$$
\begin{aligned}
&(2K)^r \mathbb{E}\left[|\mathsf{med}(U|X)-U_{(m+1)}(t)|^r\right] \\
&= (2K)^r \mathbb{E}_X \mathbb{E}_{U|X}\left[|\mathsf{med}(U|X)-U_{(m+1)}(t)|^r|X\right] \\
&\leq (2K)^r \mathbb{E}_X\left[\left(\frac{2}{Lf^2(\mathsf{med}(U|X))}\right)^{r/2} \frac{\Gamma(\frac{r+1}{2})}{\sqrt{\pi}} + \epsilon_2|X\right] \\
&= 2^{3r/2}\left(\frac{K}{\alpha}\right)^r \frac{1}{L^{r/2}} \frac{\Gamma(\frac{r+1}{2})}{\sqrt{\pi}} + \epsilon_2,
\end{aligned}
$$

where the inequality follows from Lem. 11.

Summing over all $t \in [1:n]$, we have

$$
\begin{aligned}
D^n(X^n,\hat{X}^n) = \frac{1}{n}\sum_{t=1}^{n} \mathbb{E}\left[|X(t)-\hat{X}(t)|^r\right] \leq 2^{3r/2}\left(\frac{K}{\alpha}\right)^r \frac{1}{L^{r/2}} \frac{\Gamma(\frac{r+1}{2})}{\sqrt{\pi}} + \epsilon \\
\implies D(L, R_{\mathsf{sum}}) \leq 2^{3r/2}\left(\frac{K}{\alpha}\right)^r \frac{1}{L^{r/2}} \frac{\Gamma(\frac{r+1}{2})}{\sqrt{\pi}} + \epsilon_2.
\end{aligned}
$$

From Prop. 13, we have $\frac{R_{\mathsf{sum}}}{L} \leq I(Y;U|X)$, therefore,

$$
\begin{aligned}
\beta_{\mathsf{reg}} &= \lim_{L, R_{\mathsf{sum}}\to\infty} R_{\mathsf{sum}}^{r/2} D(L, R_{\mathsf{sum}}) \\
&\leq 2^{3r/2}\left(\frac{K}{\alpha}\right)^r \frac{\Gamma(\frac{r+1}{2})}{\sqrt{\pi}} I(Y;U|X)^{r/2}.
\end{aligned}
$$

Taking infimum over $\mathcal{S}_{\mathsf{reg}}$ completes the proof. □

5.2.3 Converse Coding Theorem

A key feature of the converse is the Shannon lower bound [99, 102], which for real-valued sources with difference normed distortion is given in [94]. It is one of the few tools that evaluates the rate distortion function as a closed-form expression and is known for asymptotically tightness when distortion goes to zero [103, 104]. Combining Lem. 14 stemming from [105, 106], we can show the matching converse. As we will see, the Shannon lower bound is essentially an uncoded lower bound, that is, the bound is for estimation from $\{Y_i\}_{i=1}^L$ rather than from received codewords. It therefore shows a lower bound only due to intrinsic observational noise, yet is sufficient to show the matching asymptotics. Converse argument regarding coding rate also follows standard argument in [57], but we state it for completeness.

Coding Rate Lower Bound

Let us first derive coding rate lower bound.

$$\begin{aligned}
nR_i &= \log |\mathcal{C}_i^n| \\
&\geq I(Y_i^n; C_i | X_n) = \sum_{t=1}^{n} I(Y_i(t); C_i | Y_i^{t-1}, X_n) \\
&= \sum_{t=1}^{n} \left[h(Y_i(t) | Y_i^{t-1}, X^n) - h(Y_i(t) | C_i, Y_i^{t-1}, X^n) \right] \\
&= \sum_{t=1}^{n} \left[h(Y_i(t) | X^n) - h(Y_i(t) | C_i, Y_i^{t-1}, X^n) \right] \\
&\geq \sum_{t=1}^{n} \left[h(Y_i(t) | X^n) - h(Y_i(t) | C_i, X^n) \right] \\
&= \sum_{t=1}^{n} I(Y_i(t); C_i | X^n).
\end{aligned}$$

The sum rate lower bound is therefore given by

$$R_{\mathsf{sum}} \geq \frac{1}{n}\sum_{t=1}^{n}\sum_{i=1}^{L} I(Y_i(t); C_i|X^n).$$

Define $\breve{X}_t := (X(1), \ldots, X(t-1), X(t+1), \ldots, X(n))$ and let $U_i(t, \breve{x}_t)$ be a random variable whose joint distribution with $X(t)$ and $Y_i(t)$ is

$$\begin{aligned}
&\mathbb{P}[x \leq X(t) \leq x + dx, y \leq Y_i(t) \leq y + dy, U_i(t, \breve{x}_t) = c] \\
&= f_X(x) f_{Y|X}(y|x)\mathbb{P}[C_i = c|Y_i(t) = y, X(t) = x, \breve{X}_t = \breve{x}_t]dxdy \\
&= f_X(x) f_{Y|X}(y|x)\mathbb{P}[C_i = c|Y_i(t) = y, \breve{X}_t = \breve{x}_t]dxdy,
\end{aligned}$$

since the codeword C_i depends on $X(t)$ only through $Y_i(t)$. Hence, the Markov chain $X(t) - Y_i(t) - U_i(t, \breve{x}_t)$ holds for each i and given $\breve{x}_t$, which gives the following lower bound in expectation form.

$$R_{\mathsf{sum}} \geq \frac{1}{n}\sum_{t=1}^{n}\sum_{i=1}^{L} \mathbb{E}_{\breve{X}_t}[I(Y_i(t); U_i(t, \breve{X}_t)|X(t))].$$

Note that $\hat{X}(t) = g(C_1, \ldots, C_L) = g'(U_1(t), \ldots, U_L(t))$ for some functions g, g'.

Estimation Lower Bound

An estimate of the CEO problem is $\hat{X}^n(C_1, C_2, \cdots, C_L)$; however, it is obvious that there is an estimate $\hat{X}'^n = \hat{X}'^n(\{Y_i^n\}_{i=1}^L)$ based on $\{Y_i^n\}_{i=1}^L$ yielding a better estimate than $\hat{X}^n$. We will derive the performance lower bound for $\hat{X}'^n$ using the Shannon lower bound and it turns out that this lower bound for $\hat{X}'^n$ is sufficient to show the asymptotics.

Lemma 12. *Let $\hat{X}$ be an arbitrary estimate from $\{Y_i\}_{i=1}^L$. Then,*

$$I(X^n; \hat{X}^n) \leq nI(X; \{Y_i\}_{i=1}^L).$$

Proof.

$$
\begin{aligned}
& I(X^n; \hat{X}^n) \\
& \overset{(a)}{\leq} I(X^n; \{Y_i^n\}_{i=1}^L) \\
& = h(\{Y_i^n\}_{i=1}^L) - h(\{Y_i^n\}_{i=1}^L | X^n) \\
& = \sum_{t=1}^n h(\{Y_i(t)\}_{i=1}^L | \{Y_i^{t-1}\}_{i=1}^L) - \sum_{t=1}^n h(\{Y_i^n(t)\}_{i=1}^L | X^n, \{Y_i^{t-1}\}_{i=1}^L) \\
& \overset{(b)}{\leq} \sum_{t=1}^n h(\{Y_i(t)\}_{i=1}^L) - \sum_{t=1}^n h(\{Y_i(t)\}_{i=1}^L | X^n, \{Y_i^{t-1}\}_{i=1}^L) \\
& \overset{(c)}{=} \sum_{t=1}^n h(\{Y_i(t)\}_{i=1}^L) - \sum_{t=1}^n h(\{Y_i(t)\}_{i=1}^L | X(t)) \\
& = \sum_{t=1}^n I(X(t); \{Y_i(t)\}_{i=1}^L) \\
& \overset{(d)}{=} nI(X; Y^L),
\end{aligned}
$$

where (a) follows from the data processing inequality for $X^n - \{Y_i^n\}_{i=1}^L - \hat{X}^n$; (b) follows from the fact that removing conditions only increases entropy; (c) follows since $\{Y_i(t)\}_{i=1}^L$ depends only on $X(t)$; and (d) follows since $X(t), \{Y_i(t)\}_{i=1}^L$ are i.i.d. over time. □

Lemma 13 (Shannon lower bound [94]). *Suppose $X, \hat{X}$ are d-dimensional vectors in $\mathbb{R}^d$ and consider any norm $\|X - \hat{X}\|$. Define the standard rate distortion function*

$$R(D) := \inf_{P_{\hat{X}|X} : \mathbb{E}[\|X - \hat{X}\|^r] \leq D} I(X; \hat{X}).$$

Then, the Shannon lower bound is given by

$$R(D) \geq R_{\mathsf{SLB}}(D) := h(X) - \frac{d}{r} \log \left(\frac{rD}{d} (V_d \Gamma(1 + d/r))^{r/d} e \right),$$

where V_d is the volume of d-dimensional unit ball such that $\{x : \|x\| \leq 1, x \in \mathbb{R}^d\}$ and $\Gamma(\cdot)$ is the Gamma function.

Lemma 14 ([105, 106]). *Suppose $X \in \mathbb{R}^d$ and conditions (C2)–(C4) hold. Then,*

$$I(X; \{Y_i\}_{i=1}^L) = \frac{d}{2} \log \frac{L}{2\pi e} + h(X) + \frac{1}{2} \mathbb{E}[\log \det I_Y(X)] + o(1).$$

Combining all of the above, we can prove the converse.

Theorem 20 (Converse of Regular CEO Problem)**.**

$$\beta_{\textsf{\textit{n-reg}}} \geq C_1 \left(\min_{U:X-Y-U} I(Y;U|X) \right)^{r/2},$$

where

$$C_1 = \frac{1}{re} \left(V_1 \Gamma \left(1 + \frac{1}{r} \right) \frac{e^{-\frac{1}{2}\mathbb{E}[\log \det I_Y(X)]}}{\sqrt{2\pi e}} \right)^{-r}.$$

Proof. In particular, suppose that $\hat{X}'(t)$ in Lem. 12 is an estimate achieving distortion $D' := \frac{1}{n}\sum_{t=1}^{n} \mathbb{E}[|X(t) - \hat{X}'(t)|^r]$ from $\{Y_i^n\}_{i=1}^{L}$ with $D' \leq D$. Then, combining all lemmas we have the following chain of inequalities:

$$\begin{aligned}
& h(X) - \frac{d}{r}\log\left(\frac{rD'}{d}(V_d\Gamma(1+d/r))^{r/d}e\right) \\
& \overset{(a)}{\leq} R_{\mathsf{SLB}}(D') \\
& \overset{(b)}{\leq} \inf_{P_{\hat{X}|X}:\mathbb{E}[\|X-\hat{X}\|^r]\leq D'} I(X;\hat{X}) \\
& \leq \frac{1}{n} I(X^n; \hat{X}'^n) \\
& \overset{(c)}{\leq} I(X;Y^L) \\
& \overset{(d)}{=} \frac{1}{2}\log\frac{L}{2\pi e} + h(X) + \frac{1}{2}\mathbb{E}[\log\det I_Y(X)] + o(1),
\end{aligned}$$

where (a), (b) are from the Shannon lower bound Lem. 13; (b) is from Lem. 12; and (c) is from Lem. 14.

With $d = 1$, we have the following inequality:

$$h(X) - \frac{1}{r}\log\left(rD'(V_1\Gamma(1+1/r))^r e\right) \leq \frac{1}{2}\log\frac{L}{2\pi e} + h(X) + \frac{1}{2}\mathbb{E}[\log\det I_Y(X)] + o(1).$$

Arranging terms, we obtain

$$D \geq D' \geq \left(\frac{1}{\sqrt{L}}\right)^r \frac{1}{re}\left(V_1\Gamma\left(1+\frac{1}{r}\right)\frac{e^{-\frac{1}{2}\mathbb{E}[\log\det I_Y(X)]}}{\sqrt{2\pi e}}\right)^{-r} =: \frac{C_1}{L^{r/2}},$$

where

$$C_1 = \frac{1}{re}\left(V_1\Gamma\left(1+\frac{1}{r}\right)\frac{e^{-\frac{1}{2}\mathbb{E}[\log\det I_Y(X)]}}{\sqrt{2\pi e}}\right)^{-r}.$$

It is easy to see $D(L, R_{\mathsf{sum}}) \geq C_1 L^{-r/2}$.

Multiplying $D(L, R_{\mathsf{sum}})$ by $R_{\mathsf{sum}}^{r/2}$,

$$\begin{aligned}
\beta_{\mathsf{reg}} &\geq \lim_{L\to\infty}\left(\frac{1}{n}\sum_{t=1}^{n}\sum_{i=1}^{L}\mathbb{E}_{\breve{X}_t}[I(Y_i(t);U_i(t,\breve{X}_t)|X(t))]\right)^{r/2}\cdot\frac{C_1}{L^{r/2}}\\
&= C_1\lim_{L\to\infty}\left(\frac{1}{nL}\sum_{t=1}^{n}\sum_{i=1}^{L}\mathbb{E}_{\breve{X}_t}[I(Y_i(t);U_i(t,\breve{X}_t)|X(t))]\right)^{r/2}\\
&\geq C_1\lim_{L\to\infty}\left(\frac{1}{nL}\sum_{t=1}^{n}\sum_{i=1}^{L}\min_{t,i,\breve{X}_t} I(Y_i(t);U_i(t,\breve{X}_t)|X(t))\right)^{r/2}\\
&\geq C_1\left(\min_{U:X-Y-U} I(Y;U|X)\right)^{r/2}.
\end{aligned}$$

So the lower bound has been proved. □

Discussion

It is interesting to evaluate the Shannon lower bound for jointly Gaussian CEO problem. When the model is jointly Gaussian as in [57], but with general rth power of difference, it is possible to exactly evaluate the right side of the chain of inequalities without resorting to Lem. 14. Note that when X, Y^L are jointly Gaussian, once receiving y^L the posterior distribution $\mathbb{P}(X|Y^L = y^L)$ is also Gaussian. Let $X, Z \sim \mathcal{N}(0, \sigma_X^2), \mathcal{N}(0, \sigma_Z^2)$, respectively, and $Y_i(t) = X(t) + Z_i(t)$.

Letting $\bar{y}$ be the sample mean, $\bar{y} := \frac{1}{L}\sum_{\ell=1}^{L} y_i$,

$$\begin{aligned}
\mathbb{P}(X|Y^L = y^L) &\sim \mathcal{N}(\mathbb{E}[X|Y^L = y^L], \mathsf{Var}[X|Y^L = y^L])\\
&= \mathcal{N}\left(\frac{\sigma_X^2}{\sigma_X^2 + \frac{\sigma_Z^2}{L}}\bar{y}, \frac{\sigma_X^2}{1 + \frac{\sigma_X^2}{\sigma_Z^2}L}\right).
\end{aligned}$$

This results in the mutual information as follows:

$$\begin{aligned}
I(X;Y^L) &= h(X) - h(X|Y^L)\\
&= \frac{1}{2}\log(2\pi e\sigma_X^2) - h(X|Y^L)\\
&= \frac{1}{2}\log(2\pi e\sigma_X^2) - \int p(y^L)h(X|Y^L = y^L)dy^L\\
&= \frac{1}{2}\log(2\pi e\sigma_X^2) - \int p(y^L)\frac{1}{2}\log\left(2\pi e\frac{\sigma_X^2}{1+\frac{\sigma_X^2}{\sigma_Z^2}L}\right)dy^L\\
&= \frac{1}{2}\log(2\pi e\sigma_X^2) - \frac{1}{2}\log\left(2\pi e\frac{\sigma_X^2}{1+\frac{\sigma_X^2}{\sigma_Z^2}L}\right).
\end{aligned}$$

It is immediately apparent that $R_{\mathsf{SLB}}(D) \leq I(X;Y^L)$ gives the same asymptotics $R_{\mathsf{sum}}^{-r/2}$ (up to a different constant factor). This verifies our aforementioned conclusion that non-Gaussian regular models do not perform much better than the Gaussian model in the sense of sum-rate asymptotics, although Gaussianity is the worst compressible model [65].

It should also be noted that the median estimator is neither unique nor the best, but achieves the correct sum rate asymptotics. For instance, the (scaled version of) sample mean estimator in [57] turns out to be the best estimator for the quadratic Gaussian CEO problem even in non-asymptotic regime [58, 60] because the minimum mean-squared error estimator (MMSE) is in fact a linear summation of codewords for the additive Gaussian test channel. To illustrate pros and cons of those estimators, consider a simple estimation problem of X from observation $Y_i = X + Z_i, i \in [1 : L]$, where Y_i is given observation, Z_i is additive and i.i.d. drawn from some f_Z with zero mean and σ_Z^2 variance. In this case, sample mean estimator is distributed approximately $\mathcal{N}(0, \sigma_Z^2/L)$ by the central limit theorem and so yields approximately σ_Z^2/L quadratic distortion. However, the quadratic distortion induced by the median estimator is $(4Lf_Z^2(0))^{-1}$ according to Lem. 10. Since the performance of the median estimator is independent of variance, the median estimator is more efficient when Z is sufficiently heavy-tailed. Also the Gaussianity of Lem. 10 suggests a further extension to a broader class of estimators called *consistent and asymptotic normal (CAN)* estimators; for example, the maximum likelihood estimator (MLE) is also CAN and furthermore asymptotically *efficient* [107]. The asymptotic normality of MLE by the Bernstein-von Mises theorem will be a stepping stone to the equivalence of quadratic and logarithmic distortions in Sec. 5.4.

5.3 Non-regular Model

5.3.1 Model and Result

This section considers the bounded source and observation in [63], where the source-observation model is assumed to be *non-regular* in the sense of regularity conditions of the Cramer-Rao lower bound [98, 108]. A special case of such non-regular model is known as a copula[2] that models dependency between two (or multiple) uniform random variables and is widely used in quantitative finance: the CEO wishes to estimate some economic event or financial risk such as bankruptcy of a firm, but only related indicators governed by the copula model are observable. The formal definition of the non-regular model is as follows.

(B1) The source and the observation are finitely supported, that is, $\mathcal{X}, \mathcal{Y} \subset \mathbb{R}$ are finite intervals of the real line. Without loss of generality, we suppose $\mathcal{X} = \mathcal{Y} = [0, 1]$. In addition, $f_{Y|x}$ is discontinuous at both end points of support conditioned on x, i.e., let $\mathcal{Y}_x := [e_\ell(x), e_u(x)]$ be the support of $f_{Y|x}$, then, $f_{Y|x}(e_\ell(x)), f_{Y|x}(e_u(x)) > 0$.

(B2) There exists a random variable $U \in \mathcal{U}$ satisfying the following: 1) the Markov chain $X - Y - U$ holds; 2)$f_{U|x}$ has bounded support $[a(x), b(x)]$ for which $(a+b)(x)$ is invertible; 3) the inverse function $\ell^{-1} = (a+b)^{-1}$ exists and is Lipschitz with constant $K > 0$; and 4) $f_{U|x}(u|x)$ does not vanish at either end point $a(x), b(x)$, i.e., $f_{U|x}(a(x)|x), f_{U|x}(b(x)|x) > \delta$ for some positive δ that does not depend on x. Without loss of generality, we assume $\mathcal{U} = [0, 1]$.

As illustration, a simple example of (X, Y) that satisfies (B1) is a copula [109]. Another example is a uniform source with independent additive uniform observational noise, i.e., $Y_i = X + Z_i$ where $X, Z_i \sim \mathsf{unif}[0, 1]$. Also $X \sim \mathsf{unif}[0, 1]$ with $Y \sim \mathsf{unif}[0, X]$ is an example that satisfies (B1). Verifying (B1) for the copula example is given in [63], and for the uniform examples is immediate. Also it should be noted that one of regularity conditions of the Cramer-Rao lower bound [98, Sec. 2.5], namely that the support of $f_{Y|X}$ is the same for all $x \in \mathcal{X}$, is violated in (B1) as well as in all examples above, so that the model is called non-regular.

Let $\mathcal{S}_{\textsf{n-reg}}$ be the set of Us that satisfy (B2). Applying a copula test channel (e.g., Clayton copula) $f_{U|Y}$ to Y satisfies (B2) so that $\mathcal{S}_{\textsf{n-reg}}$ is nonempty.

[2] A copula is a multivariate distribution that has the uniform distribution for each marginal [109].

Theorem 21 (Non-regular CEO problem)**.** *Suppose condition (B1) holds for source and observation model and there exists U such that (B2) holds. Then, for distortion measure $d(x, \hat{x}) = |x - \hat{x}|^r$,*

$$C_3 \left(\min_{U:X-Y-U} I(Y;U|X) \right)^r \leq \beta_{\textit{n-reg}} \leq C_4 \left(\min_{U \in \mathcal{S}_{\textit{n-reg}}} I(Y;U|X) \right)^r,$$

where

$$\begin{aligned} C_3 &= r2^{-r} \int_{\widetilde{h}=0}^{1} \widetilde{h}^{r-1} \int_{x=0}^{1} f_X(x) e^{-\widetilde{h}g(x)} dx d\widetilde{h}, \\ C_4 &= \frac{r!2^{r+1}K^r}{\delta^r}, \end{aligned}$$

with

$$g(x) = \frac{d}{d\Delta} \left(- \min_{s \in [0,1]} \log \left(\int f^s_{Y|X}(y|x) f^{1-s}_{Y|X}(y|x+\Delta) dy \right) \right) \Bigg|_{\Delta=0},$$

and the minimum of the lower bound is taken over non-trivial random variables to ensure that the mutual information is non-zero.

Before proceeding, it should be noted that proofs in the sequel repeat parts of standard achievability and converse proofs in Sec. 5.2.2 and Sec. 5.2.3 and so are omitted.

5.3.2 Direct Coding Theorem

Like Sec. 5.2.2, we repeat quantization, Berger-Tung compression-decompression, and then estimation of the source X. Conditions for the quantization are the following:

$$\begin{aligned} \mathbb{E}[|U - \widetilde{U}|^r] &\leq \delta_0, \\ |I(Y;U) - I(\widetilde{Y};\widetilde{U})| &\leq \delta_1, \\ |I(X;U) - I(\widetilde{X};\widetilde{U})| &\leq \delta_2. \end{aligned} \tag{5.4}$$

The remaining steps are the same as Sec. 5.2.2 except for the estimation step. Midrange estimator will be used to estimate the source since it is optimal in several cases with bounded support [95,110,111]. Furthermore, it is more efficient than sample mean in many cases such as the cosine, parabolic, rectangular, and inverted parabolic distributions [95].

Theorem 22 (Achievability of Non-regular CEO Problem)**.**

$$\beta_{\textsf{n-reg}} \leq \frac{r!2^{r+1}K^r}{\delta^r}\left(\min_{U\in\mathcal{S}_{\textsf{n-reg}}} I(Y;U|X)\right)^r,$$

where $\delta > 0$ is given in the condition (B2).

Proof. As mentioned, the CEO estimates by sample midrange estimator, i.e.,

$$\hat{X}(t) = \ell^{-1}\left(\frac{\hat{U}_{(1)}(t)+\hat{U}_{(L)}(t)}{2}\right).$$

Then we have the following distortion upper bound:

$$\begin{aligned}
\mathbb{E}\left[|X(t)-\hat{X}(t)|^r\right] &= \mathbb{E}\left[\left|X(t)-\ell^{-1}\left(\frac{\hat{U}_{(1)}(t)+\hat{U}_{(L)}(t)}{2}\right)\right|^r\right]\\
&= \mathbb{E}\left[\left|\ell^{-1}\left(\frac{a(X(t))+b(X(t))}{2}\right)-\ell^{-1}\left(\frac{\hat{U}_{(1)}(t)+\hat{U}_{(L)}(t)}{2}\right)\right|^r\right]\\
&\leq K^r\mathbb{E}\left[\left|\frac{a(X(t))+b(X(t))}{2}-\frac{\hat{U}_{(1)}(t)+\hat{U}_{(L)}(t)}{2}\right|^r\right]
\end{aligned}$$

since ℓ^{-1} is Lipschitz with constant K. For notational simplicity, let us denote $a_X = a(X(t)), b_X = b(X(t))$ and omit '(t)'.

$$\begin{aligned}
&K^r\mathbb{E}\left[\left|\frac{a_X+b_X}{2}-\frac{\hat{U}_{(1)}+\hat{U}_{(L)}}{2}\right|^r\right]\\
&\overset{(a)}{\leq} K^r\mathbb{E}\left[\left(\left|\frac{a_X+b_X}{2}-\frac{U_{(1)}+U_{(L)}}{2}\right|+\left|\frac{U_{(1)}+U_{(L)}}{2}-\frac{\hat{U}_{(1)}+\hat{U}_{(L)}}{2}\right|\right)^r\right]\\
&\overset{(b)}{\leq} (2K)^r\mathbb{E}\left[\left|\frac{a_X+b_X}{2}-\frac{U_{(1)}+U_{(L)}}{2}\right|^r\right]+(2K)^r\mathbb{E}\left[\left|\frac{U_{(1)}+U_{(L)}}{2}-\frac{\hat{U}_{(1)}+\hat{U}_{(L)}}{2}\right|^r\right]\\
&= K^r\mathbb{E}\left[\left|a_X+b_X-U_{(1)}-U_{(L)}\right|^r\right]+K^r\mathbb{E}\left[\left|U_{(1)}+U_{(L)}-\hat{U}_{(1)}-\hat{U}_{(L)}\right|^r\right]\\
&\overset{(c)}{\leq} K^r\mathbb{E}\left[\left|a_X+b_X-U_{(1)}-U_{(L)}\right|^r\right]+\epsilon,
\end{aligned}$$

where (a) follows from the triangle inequality; (b) follows from Prop. 18 in App. D.4; and (c) is proven by Lem. 16 in App. D.2.

Recall that $f_{U|X}$ does not vanish at either end point, a_X and b_X. Define the set $\mathcal{I} :=$

$\{u_{(1)} > a_X + \epsilon \text{ or } u_{(L)} < b_X - \epsilon_1\}$ so that $\mathcal{I}^c = \{U_{(1)} \leq a_X + \epsilon_1 \text{ and } U_{(L)} \geq b_X - \epsilon_1\}$.

$$\mathbb{E}\left[\left|a_X + b_X - U_{(1)} - U_{(L)}\right|^r\right] = \mathbb{E}_X \mathbb{E}_{U|X}\left[\left|a_X + b_X - U_{(1)} - U_{(L)}\right|^r |X\right].$$

The conditional expectation is

$$\begin{aligned}
&\mathbb{E}_{U|X}\left[\left|a_X + b_X - U_{(1)} - U_{(L)}\right|^r |X\right] \\
&= \int_{\mathcal{I}} \left|a_X + b_X - U_{(1)} - U_{(L)}\right|^r f_{U_{(1)},U_{(L)}|X}(u_{(1)}, u_{(L)}|x) du_{(1)} du_{(L)} \\
&\quad + \int_{\mathcal{I}^c} \left|a_X + b_X - U_{(1)} - U_{(L)}\right|^r f_{U_{(1)},U_{(L)}|X}(u_{(1)}, u_{(L)}|x) du_{(1)} du_{(L)} \\
&\leq \mathit{const} \cdot \mathbb{P}[\mathcal{I}|X] + \int_{\mathcal{I}^c} \left|a_X + b_X - U_{(1)} - U_{(L)}\right|^r f_{U_{(1)},U_{(L)}|X}(u_{(1)}, u_{(L)}|x) du_{(1)} du_{(L)}. \qquad (5.5)
\end{aligned}$$

Let us separately evaluate each term. First, since $\{U_i\}_{i=1}^L$ are independent when conditioned on X,

$$\begin{aligned}
\mathbb{P}[\mathcal{I}|X] &\leq \mathbb{P}[U_{(1)} > a_X + \epsilon_1 |X] + \mathbb{P}[U_{(L)} < b_X - \epsilon_1 |X] \\
&= \prod_{i=1}^L \mathbb{P}[U_i > a_X + \epsilon_1 |X] + \prod_{i=1}^L \mathbb{P}[U_i < b_X - \epsilon_1 |X] \\
&= \prod_{i=1}^L (1 - \mathbb{P}[U_i \leq a_X + \epsilon_1 |X]) + \prod_{i=1}^L (1 - \mathbb{P}[U_i \geq b_X - \epsilon_1 |X]) \\
&= (1 - \mathbb{P}[U \leq a_X + \epsilon_1 |X])^L + (1 - \mathbb{P}[U \geq b_X - \epsilon_1 |X])^L,
\end{aligned}$$

where the last equality follows since agents are i.i.d. Since $f_{U|X}$ is continuous and does not vanish at a_X, b_X:

$$\begin{aligned}
&\lim_{u \to a_X \text{ or } b_X} f_{U|X}(u|x) \geq \delta \\
&\mathbb{P}[U \leq a_X + \epsilon_1 |X] \geq \delta\epsilon_1 \quad \text{and} \quad \mathbb{P}[U \geq b_X - \epsilon_1 |X] \geq \delta\epsilon_1.
\end{aligned}$$

Therefore

$$\mathbb{P}[\mathcal{I}|X] \leq (1 - \mathbb{P}[U \leq a_X + \epsilon_1 |X])^L + (1 - \mathbb{P}[U \geq b_X - \epsilon_1 |X])^L \leq 2(1 - \delta\epsilon_1)^L,$$

so the first term vanishes exponentially fast as L.

Let us consider the second term of (5.5). Take random variables

$$\eta := L\int_{a_X}^{U_{(1)}} f_{U|X}(u)du \geq \delta L(U_{(1)} - a_X),$$
$$\xi := L\int_{U_{(L)}}^{b_X} f_{U|X}(u)du \geq \delta L(b_X - U_{(L)}),$$

where $a_X \leq U_{(1)} \leq a_X + \epsilon_1, b_X - \epsilon_1 \leq U_{(L)} \leq b_X$ with marginal and joint distributions [112]

$$f_\xi(s) = f_\eta(s) = \left(1 - \frac{s}{L}\right)^{L-1} \text{ and}$$
$$f_{\xi,\eta}(s_1, s_2) = \frac{L-1}{L}\left(1 - \frac{s_1 + s_2}{L}\right)^{L-2},$$

where $s_1, s_2 \geq 0$ and $s_1 + s_2 \leq L$. Also note that as $L \to \infty$, ξ and η are asymptotically independent and $f_\xi(s), f_\eta(s) \to e^{-s}$. From the definition of ξ, η,

$$|a_X + b_X - (U_{(1)} + U_{(L)})|^r = \big((U_{(1)} - a_X) + (b_X - U_{(L)})\big)^r \leq \frac{2^r(\xi^r + \eta^r)}{(L\delta)^r},$$

where the last inequality follows from Prop. 18 and the definitions of η and ξ. Therefore, when L is large the second term is

$$\int_{\mathcal{I}^c} \left|a_X + b_X - U_{(1)} + U_{(L)}\right|^r f_{u_{(1)},u_{(L)}|X}(u_{(1)}, u_{(L)}|x)du_{(1)}du_{(L)}$$
$$\leq \frac{2^r}{(L\delta)^r}\int_0^{L(1-F_{U|X}(b_X-\epsilon_1)}\int_0^{LF_{U|X}(a_X+\epsilon_1)}(s_1^r + s_2^r)f_{\xi,\eta}(s_1, s_2)ds_1ds_2.$$

Combining all of the above,

$$R_{\mathsf{sum}}^r D(L, R)$$
$$\leq R_{\mathsf{sum}}^r K^r \mathbb{E}_X\left[\frac{2^r}{(L\delta)^r}\int_0^{L(1-F_{U|X}(b_X-\epsilon_1)}\int_0^{LF_{U|X}(a_X+\epsilon_1)}(s_1 + s_2)f_{\xi,\eta}(s_1, s_2)ds_1ds_2 + \epsilon\right].$$

As $\int_0^\infty s^r e^{-s}ds = r!$,

$$\lim_{L\to\infty} R_{\mathsf{sum}}^r D(L, R_{\mathsf{sum}}) \leq \frac{(2K)^r I(Y;U|X)^r}{\delta^r}\left(2\int_0^\infty s^r e^{-s}ds\right) = \frac{r!2^{r+1}K^r I(Y;U|X)^r}{\delta^r}.$$

Taking infimum over $\mathcal{S}_{\textsf{n-reg}}$ gives us the achievability. □

5.3.3 Converse Coding Theorem

To show the converse, we will use the generalized Chazan-Ziv-Zakai bound since it still holds for the non-regularity conditions (B1) and (B2) unlike the Cramer-Rao lower bound. The next lemma is a generalized version of the Chazan-Ziv-Zakai bound. Proof is an easy extension of special case $r = 2$ [96, 113], but for the sake of completeness we include it in App. D.3. Note that P_{min} in the next theorem is a function of $f_{Y|X}$ so it is also an uncoded lower bound like the Shannon lower bound in Sec. 5.2.3. However, it gives a matching asymptotic lower bound up to a constant.

Lemma 15 (Chazan-Ziv-Zakai Bound for $r \in \mathbb{N}$)**.** *Suppose $X \in [0,1]$. Then,*

$$\mathbb{E}\left[|X-\hat{X}|^r\right] \geq \int_{h=0}^{1} r2^{-r}h^{r-1}\int_{x=0}^{1-h}\frac{f_X(x)+f_X(x+h)}{2}P_{\mathit{min}}[x,x+h]dxdh,$$

where P_{min} is the minimum probability of error of binary hypothesis testing with $H_0 : Y \sim f_{Y|x}$ and $H_1 : Y \sim f_{Y|x+h}$.

Proof. See App. D.3. □

Recall that the same argument in Sec. 5.2.3 gives the sum rate lower bound

$$R_{\mathsf{sum}} \geq \frac{1}{n}\sum_{t=1}^{n}\sum_{i=1}^{L}\mathbb{E}_{\breve{X}_t}[I(Y_i(t);U_i(t,\breve{X}_t)|X(t)].$$

Theorem 23 (Converse for Non-regular CEO Problem)**.**

$$\beta_{\mathit{n\text{-}reg}} \geq r2^{-r}\left(\min_{U:X-Y-U} I(Y;X|U)\right)\int_{\widetilde{h}=0}^{1}\widetilde{h}^{r-1}\int_{x=0}^{1} f_X(x)e^{-\widetilde{h}g(x)}dxd\widetilde{h},$$

where

$$g(x) = \frac{d}{d\Delta}\left(-\min_{s\in[0,1]}\log\left(\int f_{Y|X}^{s}(y|x)f_{Y|X}^{1-s}(y|x+\Delta)dy\right)\right)\Bigg|_{\Delta=0}.$$

Proof. It is obvious that the estimate from $\{Y_i^n\}_{i=1}^{L}$ performs better than an estimate from codewords. Let $\hat{X}'$ be the uncoded estimate, i.e., $\hat{X}' = \hat{X}'(\{Y_i^n\}_{i=1}^L)$. Then,

$$
\begin{aligned}
&D^n(X^n, \hat{X}^n) \\
&\geq D^n(X^n, (\hat{X}')^n) \\
&= \frac{1}{n}\sum_{t=1}^{n} \mathbb{E}\left[|X(t) - \hat{X}(t)|\right] \\
&\overset{(a)}{\geq} \frac{1}{n}\frac{r2^{-r}}{2}\sum_{t=1}^{n}\int_{h=0}^{1} h^{r-1}\int_{x=0}^{1-h} (f_X(x) + f_X(x+h))P_{\mathsf{min}}(x, x+h)dxdh \\
&= \frac{1}{n}\frac{r2^{-r}}{2L^r}\sum_{t=1}^{n}\int_{h=0}^{1} (Lh)^{r-1}\int_{x=0}^{1-h} (f_X(x) + f_X(x+h))P_{\mathsf{min}}(x, x+h)dxd(hL) \\
&\overset{(b)}{=} \frac{1}{n}\frac{r2^{-r}}{2L^r}\sum_{t=1}^{n}\int_{\widetilde{h}=0}^{1} \widetilde{h}^{r-1}\int_{x=0}^{1-\frac{\widetilde{h}}{L}} \left(f_X(x) + f_X(x+\frac{\widetilde{h}}{L})\right) P_{\mathsf{min}}\left(x, x+\frac{\widetilde{h}}{L}\right) dxd\widetilde{h} \\
&\overset{(c)}{\geq} \frac{r2^{-r}}{2L^r}\frac{1}{\frac{1}{n}\sum_{t=1}^{n}\left[\int_{\widetilde{h}=0}^{1}\int_{x=0}^{1-\frac{\widetilde{h}}{L}} \left(f_X(x) + f_X(x+\frac{\widetilde{h}}{L})\right) P_{\mathsf{min}}\left(x, x+\frac{\widetilde{h}}{L}\right) dxd\widetilde{h}\right]^{-1}},
\end{aligned}
$$

where (a) follows from the Chazan-Ziv-Zakai bound in Lem. 15; (b) is obtained by letting $\widetilde{h} = hL$; and (c) follows from the arithmetic-harmonic (AM-HM) inequality. In addition,

$$
\begin{aligned}
&R^r_{\mathsf{sum}} D^n(X^n, \hat{X}^n) \\
&\geq \frac{r2^{-r}}{2L^r} \frac{\left(\frac{1}{n}\sum_{t=1}^n \sum_{i=1}^L \mathbb{E}_{\breve{X}_t}[I(Y_i(t); U_i(t, \breve{X}_t)|X(t)]\right)^r}{\frac{1}{n}\sum_{t=1}^n \left[\int_{\widetilde{h}=0}^1 \widetilde{h}^{r-1} \int_{x=0}^{1-\frac{\widetilde{h}}{L}} \left(f_X(x) + f_X(x+\frac{\widetilde{h}}{L})\right) P_{\mathsf{min}} \left(x, x+\frac{\widetilde{h}}{L}\right) dx d\widetilde{h}\right]^{-1}} \\
&\overset{(a)}{\geq} \frac{r2^{-r}}{2L^r} \frac{\frac{1}{n}\sum_{t=1}^n \left(\sum_{i=1}^L \mathbb{E}_{\breve{X}_t}[I(Y_i(t); U_i(t, \breve{X}_t)|X(t)]\right)^r}{\frac{1}{n}\sum_{t=1}^n \left[\int_{\widetilde{h}=0}^1 \widetilde{h}^{r-1} \int_{x=0}^{1-\frac{\widetilde{h}}{L}} \left(f_X(x) + f_X(x+\frac{\widetilde{h}}{L})\right) P_{\mathsf{min}} \left(x, x+\frac{\widetilde{h}}{L}\right) dx d\widetilde{h}\right]^{-1}} \\
&= \frac{r2^{-r}}{2L^r} \frac{\sum_{t=1}^n \left(\sum_{i=1}^L \mathbb{E}_{\breve{X}_t}[I(Y_i(t); U_i(t, \breve{X}_t)|X(t)]\right)^r}{\sum_{t=1}^n \left[\int_{\widetilde{h}=0}^1 \widetilde{h}^{r-1} \int_{x=0}^{1-\frac{\widetilde{h}}{L}} \left(f_X(x) + f_X(x+\frac{\widetilde{h}}{L})\right) P_{\mathsf{min}} \left(x, x+\frac{\widetilde{h}}{L}\right) dx d\widetilde{h}\right]^{-1}} \\
&\overset{(b)}{\geq} \frac{r2^{-r}}{2L^r} \min_t \frac{\left(\sum_{i=1}^L \mathbb{E}_{\breve{X}_t}[I(Y_i(t); U_i(t, \breve{X}_t)|X(t)]\right)^r}{\left[\int_{\widetilde{h}=0}^1 \widetilde{h}^{r-1} \int_{x=0}^{1-\frac{\widetilde{h}}{L}} \left(f_X(x) + f_X(x+\frac{\widetilde{h}}{L})\right) P_{\mathsf{min}} \left(x, x+\frac{\widetilde{h}}{L}\right) dx d\widetilde{h}\right]^{-1}} \\
&\geq \frac{r2^{-r}}{2} \min_{t,i} \frac{\left(\mathbb{E}_{\breve{X}_t}[I(Y_i(t); U_i(t, \breve{X}_t)|X(t)]\right)^r}{\left[\int_{\widetilde{h}=0}^1 \widetilde{h}^{r-1} \int_{x=0}^{1-\frac{\widetilde{h}}{L}} \left(f_X(x) + f_X(x+\frac{\widetilde{h}}{L})\right) P_{\mathsf{min}} \left(x, x+\frac{\widetilde{h}}{L}\right) dx d\widetilde{h}\right]^{-1}} \\
&= \frac{r2^{-r}}{2} \min_{t,i} \left(\mathbb{E}_{\breve{X}_t}[I(Y_i(t); U_i(t, \breve{X}_t)|X(t)]\right)^r \\
&\qquad \times \int_{\widetilde{h}=0}^1 \widetilde{h}^{r-1} \int_{x=0}^{1-\frac{\widetilde{h}}{L}} \left(f_X(x) + f_X(x+\frac{\widetilde{h}}{L})\right) P_{\mathsf{min}} \left(x, x+\frac{\widetilde{h}}{L}\right) dx d\widetilde{h},
\end{aligned}
$$

where (a) follows after applying the Jensen's inequality on the numerator; and (b) follows from Prop. 17 in App. D.4. Also the Chernoff-Stein lemma [70] gives

$$P_{\mathsf{min}} \left(x, x+\frac{\widetilde{h}}{L}\right) = e^{-LC(x, x+\frac{\widetilde{h}}{L})},$$

where $C\left(x, x+\frac{\widetilde{h}}{L}\right)$ is the Chernoff information between two conditional densities of y given x and $x+\frac{\widetilde{h}}{L}$. Since $L \to \infty$, the quantity $G_x\left(\frac{\widetilde{h}}{L}\right) := C\left(x, x+\frac{\widetilde{h}}{L}\right)$ can be approximated by the Maclaurin expansion

$$G_x(\Delta) = G_x(0) + \Delta \cdot G_x'(0) + O(\Delta^2).$$

As $G_x(0) = 0$, we have

$$P_{\mathsf{min}}\left(x, x + \frac{\widetilde{h}}{L}\right) = e^{-LC(x,x+\frac{\widetilde{h}}{L})} = e^{-\widetilde{h}G_x'(0)+O(L^{-1})}.$$

Therefore, for large L,

$$\begin{aligned}
&R_{\mathsf{sum}}^r D^n(X^n, \hat{X}^n) \\
&\geq \frac{r2^{-r}}{2}\left(\min_{t,i} \mathbb{E}_{\breve{X}_t}[I(Y_i(t); U_i(t, \breve{X}_t)|X(t)]\right)^r \\
&\qquad \times \int_{\widetilde{h}=0}^{1} \widetilde{h}^{r-1} \int_{x=0}^{1-\frac{\widetilde{h}}{L}} \left(f_X(x) + f_X(x + \frac{\widetilde{h}}{L})\right) e^{-\widetilde{h}G_x'(0)+O(L^{-1})} dx d\widetilde{h} \\
&\overset{(a)}{=} \frac{r2^{-r}}{2}\left(\min_{t,i} \mathbb{E}_{\breve{X}_t}[I(Y_i(t); U_i(t, \breve{X}_t)|X(t)]\right)^r \int_{\widetilde{h}=0}^{1} \widetilde{h}^{r-1} \int_{x=0}^{1} 2f_X(x)e^{-\widetilde{h}g(x)} dx d\widetilde{h} \\
&= r2^{-r}\left(\min_{t,i} \mathbb{E}_{\breve{X}_t}[I(Y_i(t); U_i(t, \breve{X}_t)|X(t)]\right)^r \int_{\widetilde{h}=0}^{1} \widetilde{h}^{r-1} \int_{x=0}^{1} f_X(x)e^{-\widetilde{h}g(x)} dx d\widetilde{h} \\
&\geq r2^{-r}\left(\min_{t,i,\breve{X}_t} I(Y_i(t); U_i(t, \breve{X}_t)|X(t)\right)^r \int_{\widetilde{h}=0}^{1} \widetilde{h}^{r-1} \int_{x=0}^{1} f_X(x)e^{-\widetilde{h}g(x)} dx d\widetilde{h},
\end{aligned}$$

where (a) follows from the fact that L is sufficiently large; and $g(x)$ is the first derivative of $G_x'(0)$, i.e.,

$$g(x) = \frac{d}{d\Delta}\left(-\min_{s\in[0,1]} \log\left(\int f_{Y|X}^s(y|x) f_{Y|X}^{1-s}(y|x+\Delta)dy\right)\right)\Bigg|_{\Delta=0}.$$

Taking limits concludes the proof that

$$\begin{aligned}
\beta_{\mathsf{n\text{-}reg}} &= \lim_{L,R_{\mathsf{sum}}\to\infty} \lim_{n\to\infty} R_{\mathsf{sum}}^r D^n(X^n, \hat{X}^n) \\
&\geq r2^{-r}\left(\min_{U:X-Y-U} I(Y;U|X)\right)^r \int_{\widetilde{h}=0}^{1} \widetilde{h}^{r-1} \int_{x=0}^{1} f_X(x)e^{-\widetilde{h}g(x)} dx d\widetilde{h}.
\end{aligned}$$

□

5.4 Equivalence of Quadratic and Logarithmic Distortions

In this section, we will show that quadratic distortion D_{Q} and logarithmic distortion D_{Log} [61] are in fact asymptotically equivalent under some conditions. Those two are in general related by the entropy power inequality [70], that is, when Z is a set of received messages,

$$\mathsf{Var}(X|Z) \geq \tfrac{1}{2\pi e} e^{h(X|Z)} \implies D_{\mathsf{Q}} \geq \tfrac{1}{2\pi e} 2^{D_{\mathsf{Log}}}.$$

We previously showed equality in the case of the jointly Gaussian CEO problem with finite number of agents [62] due to entropy maximization property of Gaussians. Here we extend it to our regular CEO problem and provide conditions for which D_{Q} and D_{Log} are equivalent under the entropy power conversion $D_{\mathsf{Q}} = \frac{1}{2\pi e} 2^{D_{\mathsf{Log}}}$, as $L \to \infty$. Regarding such universality of logarithmic distortion, it is known that logarithmic distortion is equivalent to *any* distortion measure in a direct source coding problem [114], but note that it is not true for remote source coding problems.

To argue the equivalence, we state again that each agent's test channel $f_{U_i|Y_i}$ is identical to $f_{U|Y}$ as assumed in the previous sections since it does not lose optimality.[3] Also beyond the regular model in Sec. 5.2, we further suppose the following conditions on test channel for logarithmic optimal codewords [115, 116]:

(C1) For all $x \in \mathcal{X}$, it holds that $\int_{\mathcal{U}} \frac{\partial^2}{\partial x^2} f_{U|X}(u|x) du = 0$. Also the Fisher information is finite and positive, i.e.,

$$0 < I_U(x) := \mathbb{E}_{U|x}[(\tfrac{\partial}{\partial x} \log f_{U|X}(U|x))^2] < \infty$$

for all $x \in \mathcal{X}$.

(C2) Let x_0 denote the true source. Then, there exists $k(u)$ such that $|\frac{\partial^2}{\partial x^2} f_{U|x}(u|x)| \leq k(u)$ on small neighborhood of x_0 and such that $\mathbb{E}_{x_0}[k(U)]$ is finite.

Define $\mathcal{S}'_{\mathsf{reg}}$ to be the set of Us that satisfy (C1) and (C2) as well as (A5)–(A7). Note that although $\mathcal{S}'_{\mathsf{reg}} \subset \mathcal{S}_{\mathsf{reg}}$, it only affects a constant factor in Thm. 18.

Theorem 24. *Given L, R_{sum}, suppose optimal codebook for logarithmic distortion is gener-*

[3] Note that individual rates need not be identical; however, the sum rate that agents must satisfy is unchanged by the Slepian-Wolf coding regardless of individual rate allocation.

ated from a member of $\mathcal{S}'_{\mathsf{reg}}$. Then, $D_{\mathsf{Q}}(L, R_{\mathsf{sum}})$ and $D_{\mathsf{Log}}(L, R_{\mathsf{sum}})$ asymptotically satisfy

$$D_{\mathsf{Q}}(L, R_{\mathsf{sum}}) - \frac{1}{2}\log\left(2\pi e D_{\mathsf{Log}}(L, R_{\mathsf{sum}})\right) \to 0 \quad \textit{as } L, R_{\mathsf{sum}} \to \infty.$$

It is easy to anticipate that the logarithmic distortion decays as $-\log L$ (so that $-\log R_{\mathsf{sum}}$) since the minimum logarithmic distortion is always $h(X|U^L)$ by declaring posterior distribution [61, Lem. 1] and $h(X|U^L)$ decreases as $-\log L$ from Lem. 14. The above theorem not only validates such intuition, but also shows its asymptotic equivalence to quadratic distortion with entropy power relation. Before proceeding to the proof, let us state the Bernstein-von Mises theorem which is often referred to as *asymptotic normality of posterior*, without the prior having an effect.

Lemma 16 (Bernstein-von Mises [115,116])**.** *Suppose (C1) and (C2) as well as (A1)–(A7) hold. Then, for any $x_0 \in \mathcal{X}$,*

$$\|f(X|U^L) - \mathcal{N}(\hat{X}_{\mathsf{MLE}}, (LI_U(x_0))^{-1})\|_{\mathsf{TV}} \to 0 \textit{ as } L \to \infty \textit{ with } f_{Y|x_0}\textit{-probability } 1,$$

where $\hat{X}_{\mathsf{MLE}}$ and $\|\cdot\|_{\mathsf{TV}}$ denote the maximum likelihood estimator and total variation distance, respectively.

Now we can prove the equivalence, which relies on the Bernstein-von Mises theorem.

Proof of Thm. 24. Let us consider the quadratic optimal codebook and fix some codewords $(w_1, w_2, \ldots, w_L)$. Then, incurred quadratic distortion is

$$\begin{aligned}
&\left(D^n_{\mathsf{Q}}(L, R_{\mathsf{sum}}|\{w_i\}_{i=1}^L)\right)^n \\
&= \left(\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[|X_i - \hat{X}_i|^2|\{w_i\}_{i=1}^L\right]\right)^n = \left(\frac{1}{n}\sum_{i=1}^n \mathsf{Var}(X_i|\{w_i\}_{i=1}^L)\right)^n \\
&\overset{(a)}{\geq} \prod_{i=1}^n \mathsf{Var}(X_i|\{w_i\}_{i=1}^L) \\
&\overset{(b)}{\geq} \prod_{i=1}^n \frac{1}{2\pi e} e^{2h(X_i|\{w_i\}_{i=1}^L)} = \frac{1}{(2\pi e)^n} e^{2\sum h(X_i|\{w_i\}_{i=1}^L)} \\
&\overset{(c)}{=} \frac{1}{(2\pi e)^n} 2^{2nD^n_{\mathsf{Log}}(L, R_{\mathsf{sum}}|\{w_i\}_{i=1}^L)} = \left(\frac{1}{2\pi e} 2^{2D^n_{\mathsf{Log}}(L, R_{\mathsf{sum}}|\{w_i\}_{i=1}^L)}\right)^n,
\end{aligned}$$

where (a) follows from the arithmetic-geometric inequality; (b) follows from the fact that Gaussian maximizes differential entropy for a given variance; and (c) follows by declaring the true posterior distribution [61, Lem. 1]. Hence, taking expectation over all codewords,

$$\begin{aligned} D^n_{\mathsf{Q}}(L, R_{\mathsf{sum}}) &= \mathbb{E}\left[D^n_{\mathsf{Q}}(L, R_{\mathsf{sum}}|\{W_i\}_{i=1}^L)\right] \geq \mathbb{E}\left[\frac{1}{2\pi e} 2^{2D^n_{\mathsf{Log}}(L, R_{\mathsf{sum}}|\{W_i\}_{i=1}^L)}\right] \\ &\overset{(d)}{\geq} \frac{1}{2\pi e} 2^{2\mathbb{E}[D^n_{\mathsf{Log}}(L, R_{\mathsf{sum}}|\{W_i\}_{i=1}^L)]} = \frac{1}{2\pi e} 2^{2\widetilde{D}^n_{\mathsf{Log}}(L, R_{\mathsf{sum}})}, \end{aligned}$$

where (d) follows from the Jensen's inequality and $\widetilde{D}^n_{\mathsf{Log}}$ is the logarithmic distortion incurred by quadratic optimal codebook. It is therefore obvious that the logarithmic optimal codebook achieves a smaller distortion. It shows one direction

$$D^n_{\mathsf{Q}}(L, R_{\mathsf{sum}}) \geq \frac{1}{2\pi e} 2^{2D^n_{\mathsf{Log}}(L, R_{\mathsf{sum}})}.$$

To show the other direction, consider the logarithmic optimal codebook.

$$D^n_{\mathsf{Log}}(L, R_{\mathsf{sum}}) = h(X|U^L) \overset{(a)}{=} \frac{1}{2}\log\left(2\pi e \mathsf{Var}(X|U^L)\right) \overset{(b)}{\geq} \frac{1}{2}\log\left(2\pi e D^n_{\mathsf{Q}}(L, R_{\mathsf{sum}})\right),$$

where (a) is in fact '$\geq$', but the equality holds asymptotically by the Bernstein-von Mises theorem; and (b) follows since the logarithmic optimal codebook is suboptimal for quadratic distortion. The theorem is proved. □

5.5 Chapter Summary

In this chapter, we study two continuous alphabet CEO problems—regular and non-regular—and find their matching sum rate asymptotics $R_{\mathsf{sum}}^{-r/2}$ and R_{sum}^{-r}, respectively, for $|x - \hat{x}|^r$ distortion. We also propose practical estimators, sample median and midrange estimators, unlike usual MLE or MMSE [39, 57] that are computationally expensive.

Inspired by the Bernstein-von Mises theorem, we also provide a condition for the regular model, under which quadratic and logarithmic distortions are asymptotically equivalent by entropy power relation as the number of agents grows.

Chapter 6

Conclusion and Future Research

In this dissertation, we have discussed three examples of human-machine information processing systems. As noted, the main reason for the difference from machine-only systems is the bounded rationality of humans. We aim at separate examples of bounded rationality in each chapter.

We model the bounded rationality as workload-dependent (in Chaps. 2 and 3) information processing quality and study Shannon's capacity theorem. The purpose of this research is to understand how systems that involve humans behave and how we can optimally design them. We only consider one aspect of the queueing metric, namely queue length seen by departure; we believe this is the most reasonable metric at least in a crowdsourcing application. However, moving to another queueing metric, we can also consider the case that service quality relies on the queue-length seen by arrivals. This is especially the case when customers in a hurry are the source of errors. The nature of Chaps. 2 and 3, however, remains unchanged since, for the single-user case, distributions of queue-length seen by arrivals and departures are identical. The distributions seen by arrivals and departures are nonidentical in general. Waiting-time-dependent service quality is interesting and a good proxy for quantum bit processing [117].

Social learning in sequential decision-making (Chap. 4) is a new formulation in the sense that people in sequential social learning are not aware of others' beliefs based on which previous decisions are made. This study brought to light two main attributes. Firstly, having perfect prior information does not guarantee that the team can function well together. In fact, complementary sets of traits constitute good teams, therein informing us of what pairs of agents can function well together. In addition, comparing the optimal beliefs with the Prelec reweighting functions, we observe that humans might be fundamentally predisposed to functioning well in teams if the advisor is open-minded and has more expertise than the closed-minded learner. The social learning study allows us to deduce optimal interactions between human and machine depending on expertise and bias on an inference problem. With the growth of machine learning, designing machines for humans or systems that interact with humans, for example crowdsourcing systems or AI-assisted physicians, is crucial. Therefore, developing machines involving humans in the loop is a problem of great interest. From cumulative prospect theory, we can assume the Prelec reweighting function is a good analytic model for humans, which leads us to a further machine-human interacting system design.

Lastly, we formulate collaborative decision-making as a CEO problem and then extend the classical CEO problem to two continuous alphabet settings, called regular and non-regular CEO problems, with general rth power of difference and logarithmic distortions, and study matching asymptotics of distortions. Noting that the conditions for regular models (A1)–(A7) and for non-regular models (B1)–(B2) do not form a disjoint partition, there are many other models that do not belong to either of the two. For example, when the observational noise is additive triangular, it does not satisfy non-vanishing probability density in (B1) so that the midrange estimator does not gives tight asymptotics with the Chazan-Ziv-Zakai based converse. So it would be an interesting future direction to find a generalization of the source-observation condition and its matching estimation scheme in (non-)asymptotic regime. As mentioned, the jointly Gaussian model is the worst model among all finite variance models as shown in [65] and all regular models have the same asymptotics. Therefore, regular models belong to the class of the slowest distortion decay $R_{\mathsf{sum}}^{-r/2}$, and our non-regular models are another class of decay R_{sum}^{-r}. In this context, it is interesting to classify various models by distortion decay.

Appendix A

Proofs for Chapter 2

A.1 Proof of Lemma 1

We need separate approaches for Type I and Type II arrival processes, as the nature of $\{Q_i\}$ process depends on the type.

First consider Type II processes. For these, $\lambda < 1$ implies $\sum_k km_A(k) < 1$. If $m_A(0) = 0$, the mean arrival rate must be equal or greater than 1, contradicting the assumption. Hence, $m_A(0) > 0$. Also from the assumption in Sec. 2.1, $m_A(1) > 0$.

Under the assumptions, we show $\{Q_i\}$ is an irreducible and aperiodic Markov chain by proving $\mathbb{P}(Q_{i+1} = Q_i + 1|Q_i)$, $\mathbb{P}(Q_{i+1} = \max(Q_i - 1, 0)|Q_i)$, $\mathbb{P}(Q_{i+1} = Q_i|Q_i) > 0$ for all Q_i. If this is true then any state can be reached from any other state, since states are in $\mathbb{Z}_+$. Notice the enumerated probabilities are probabilities of the events corresponding to two, one, and no arrivals, respectively, during a service time.

By the above result and assumption, $m_A(0), m_A(1) > 0$ and there exists an $s > 1$ such that $P_S(s) > 0$. Note the probability of exactly two arrivals in a service time is lower bounded by

$$(m_A(1))^2(m_A(0))^{s-2}P_S(s)$$

for any $s > 1$. As there exists an $s > 1$ such that $P_S(s) > 0$ and $m_A(a), m_A(1) > 0$, this bound is strictly positive. Probability of exactly one arrival in a service time is lower bounded by $m_A(1)(m_A(0))^{s-1}P_S(s)$ for any $s > 0$, which again is strictly positive. Probability of no arrival is lower bounded by $P_S(s)(m_A(0))^s$, which is also strictly positive.

Note that as $\mathbb{P}(Q_{i+1} = Q_i|Q_i) > 0$, this Markov chain is also aperiodic. Due to the self-loop if $\mathbb{P}(Q_{i+k} = q'|Q_i = q)$ is positive, then so is $\mathbb{P}(Q_{i+k+1} = q'|Q_i = q)$.

Positive recurrence follows by considering queue-length to be the Lyapunov function, because $\lambda < \mu$. Hence, the result follows for Type II processes due to the existence of a unique stationary distribution for an irreducible and aperiodic positive recurrent Markov chain.

Hence, $\{Q_i\}$ is ergodic.

For Type I, $\{Q_i\}$ is not a Markov chain, and we take a different approach. First note that as $\mu < 1$, $\sum_s sP_S(s) > 1$. This implies there exists an $s > 1$ such that $P_S(s) > 0$. Note that by assumption $\lambda < \mu$. Then, for Type I arrival processes this implies there exists an $a > 1$ such that $P_A(a) > 0$.

Consider the process $\{W_i\}$, the sojourn time for jobs. We first claim that under the assumption, this is an irreducible, aperiodic, and positive recurrent Markov chain. It is known in queuing theory that for i.i.d. inter-arrival and service times, $\{W_i\}$ is a Markov chain. Next, we show irreducibility and aperiodicity by showing that $\mathbb{P}(W_{i+1} = W_i + 1 | W_i)$, $\mathbb{P}(W_{i+1} = W_i | W_i)$, and $\mathbb{P}(W_{i+1} = \max(W_i - 1, 0), W_i) > 0$.

First, we consider the case when P_A has a support that spans $\mathbb{Z}_+$. As $\mu < 1$, there exists an $s > 1$ such that $P_S(s) > 0$. Consider a possible path from W_i to $W_{i+1} = \max(W_i + b, 0), b \in \{0, \pm 1\}$. This can happen as follows: the $(i+1)$th job brings a service time requirement of s, and it reaches the system $s - b$ time after the ith job. As the service times and inter-arrival times are independent, probability of this sample path event is exactly $P_S(s)P_A(s-b)$, which is strictly positive.

Next, we consider the case when P_S has a support spanning $\mathbb{Z}_+$. As $\lambda < 1$, there exists an $a > 1$ such that $P_A(a) > 0$. Then a possible path for the events is as follows: the $(i+1)$th job comes a time after ith job and brings with it a service requirement of $a + b$. The rest follows by evaluating the probability of this event.

Note that $\{W_i\}$ is an irreducible and aperiodic Markov chain. Note that given W_i, Q_i is independent of anything else because given W_i, it only depends on the number of arrivals in the time W_i:

$$\mathbb{P}(Q_i = q) = \mathbb{P}\left(\sum_{i=1}^{q} A_i \leq W_i < \sum_{i=1}^{q+1} A_i\right).$$

As the A_i are i.i.d., this also implies that given a distribution of W_i, the distribution of Q_i is fixed.

It follows from queuing theory that $\{W_i\}$ is positive recurrent for $\lambda < \mu$. Hence, $\{W_i\}$ converges in distribution to a stationary distribution, and by the above argument, so does $\{Q_i\}$. Ergodicity of Q_i follows from the ergodicity of W_i. ■

A.2 Proof of Theorem 2

Let $\{\widehat{Q}_i\}$ be queue-lengths seen by the arrivals, then the stationary distribution of $\widehat{Q}_i$ is the same as that of Q_i. Note that there is only one arrival and one departure at a time. Since the queue-length is stable, the fraction of time the queue-length increases by 1 from a value q is the same as the fraction of time the queue-length decreases by 1 from q, for all q. Since increase corresponds to arrival and decrease corresponds to departure, the fractions of arrivals and departures that see a queue-length q are the same. Thus it is sufficient to show that the stationary distribution of $\{\widehat{Q}\}$ is $\pi_k = (1-\sigma)\sigma^k$, where σ solves $x = \sum_{n=0}^{\infty} P_A(n)(1-\mu+x\mu)^n$ in $(0,1)$.

We shall first show the uniqueness of the stationary distribution from the fact that $\{\widehat{Q}_i\}$ is an irreducible Markov chain, and then derive the stationary distribution.

Consider the transition probability

$$\mathbb{P}(\widehat{Q}_{i+1} = q' | \widehat{Q}_i = q, \widehat{Q}_{i-1}, \ldots).$$

As at most one arrival is possible, the probability is 0 for $q' - q > 1$. For $q' - q \leq 1$,

$$\begin{aligned}
&\mathbb{P}(\widehat{Q}_{i+1} = q' | \widehat{Q}_i = q, \widehat{Q}_{i-1}, \ldots) \\
&= \mathbb{P}(\text{there are } q - q' + 1 \text{ departures} \\
&\quad\quad \text{between } i \text{ and } i+1 \text{ arrival} \mid \widehat{Q}_i = q, \widehat{Q}_{i-1}, \ldots).
\end{aligned}$$

As service time is geometric with mean $\frac{1}{\mu}$ and hence memoryless, starting from any time, the time to the next departure is geometric with the same mean, if there is a job in the queue. After any arrival, there is always at least one job in the queue, and hence, time to the next departure is geometric. Thus the probability that there are $q - q' + 1$ departures given the past is nothing but the probability that the sum of $q - q' + 1$ geometric random variables is less than a realization of P_A. Thus,

$$\begin{aligned}
&\mathbb{P}(\widehat{Q}_{i+1} = q' | \widehat{Q}_i = q, \widehat{Q}_{i-1}, \ldots) \\
&\quad = \sum_{t=0}^{\infty} P_A(t) \mathbb{P}\left(\sum_{i=1}^{q-q'+1} S_i \leq t \leq \sum_{i=1}^{q-q'} S_i \right) \\
&\quad = \sum_{t=0}^{\infty} P_A(t) \mathbb{P}(\mathrm{Bin}(t, \mu) = q - q' + 1) \qquad \text{(A.1)} \\
&\quad = \sum_{t=0}^{\infty} P_A(t) \binom{t}{q - q' + 1} (1-\mu)^{t-q+q'-1} \mu^{q-q'+1}.
\end{aligned}$$

Eq. (A.1) follows because the service times are geometric, meaning each time a job in service gets completed according to a Bernoulli random variable, and the sum of Bernoulli random variables is binomial. This derivation implies the transition depends only on q and q', further implying the process is Markov.

Thus the probability of the $q \to 0$ transition is

$$\sum_{t=0}^{\infty} P_A(t) \binom{t}{q+1} (1-\mu)^{t-q-1} \mu^{q+1}.$$

Note that the transitions can be written as the amount of change in the queue-length, meaning a $q \to q'$ transition is a $q - q'$ change, and is nothing but the probability of having $q - q' + 1$ departures before an arrival.

For $k \geq 0$, let β_k denote the probability that the sum of k geometric random variables is less than the time between two arrivals. Then, for $q' > 0$,

$$\mathbb{P}(\widehat{Q}_{i+1} = q' | \widehat{Q}_i = q) = \beta_{q-q'+1},$$

and for $q' = 0$,

$$\mathbb{P}(\widehat{Q}_{i+1} = 0 | \widehat{Q}_i = q) = 1 - \sum_{k=0}^{q} \beta_k.$$

Also, as $\beta_0, \beta_1, \beta_2 > 0$, the Markov chain is irreducible and aperiodic. Thus there exists a unique stationary distribution π which solves $\pi = \pi[P]$, where $[P]$ is the probability transition matrix. The transition matrix $[P]$ is written as a matrix whose first column is $(1 - \beta_0, 1 - \sum_{k=0}^{1} \beta_k, \ldots)^T$ and other columns are $(0, \ldots, \beta_0, \beta_1, \beta_2, \ldots)^T$, where β_0 is the $(i, i+1)$th entry.

From $\pi = \pi[P]$ it follows that

$$\pi_0 = \sum_{i=0}^{\infty} \left(1 - \sum_{k=0}^{i} \beta_k\right) \pi_i,$$
$$\pi_k = \sum_{i=0}^{\infty} \pi_{k-1+i}\beta_i \quad \text{for } k > 0.$$

Like in the analysis of GI/M/1 queue [72], we guess a solution $\pi_k = \pi_0\sigma^k$ for some $\sigma < 1$. Next, we check if this solution satisfies $\pi = \pi[P]$ for a unique $\sigma < 1$.

It follows from $\pi = \pi[P]$, as above, that σ must satisfy

$$\begin{aligned}\sigma &= \sum_{i=0}^{\infty} \sigma^i \beta_i \\ &= \sum_{i=0}^{\infty} \sigma^i \sum_{t=0}^{\infty} P_A(t) \binom{t}{i} (1-\mu)^{t-i}\mu^i \\ &= \sum_{t=0}^{\infty} P_A(t) \sum_{i=0}^{t} \binom{t}{i} (1-\mu)^{t-i}(\sigma\mu)^i \qquad \text{(A.2)} \\ &= \sum_{t=0}^{\infty} P_A(t)(1-\mu+\sigma\mu)^t. \qquad \text{(A.3)}\end{aligned}$$

Eq. (A.2) follows by interchanging the two sums, as per the Fubini-Tonelli theorem since terms are non-negative. Eq. (A.3) follows using the binomial theorem.

To show that the distribution π is unique, we show that $x = \sum_{t=0}^{\infty} P_A(t)(1-\mu+x\mu)^t$ has a unique solution in $0 < x < 1$. Towards this we characterize $\sum_{t=0}^{\infty} P_A(t)(1-\mu+x\mu)^t$, in Lems. 17 and 18 given below, which complete the proof. ■

Lemma 17. *For any P_A on $\mathbb{Z}_+$ and $\mu \in (0,1)$, $\sum_{t=0}^{\infty} P_A(t)(1-\mu+x\mu)^t$ is an increasing function of x in $(0,1)$, and strictly convex in $(0,1)$.*

Proof. Define

$$f(x) = \sum_{t=0}^{\infty} P_A(t)(1-\mu+x\mu)^t.$$

It is sufficient to show that $f'(x), f''(x)$ are both strictly positive in $x \in (0,1)$.

Let the partial sum up to T in $f(x)$ be $f_T(x)$, i.e.,

$$f_T(x) = \sum_{t=0}^{T} P_A(t)(1-\mu+x\mu)^t,$$

and then

$$f_T'(x) = \sum_{t=0}^{T} \mu t P_A(t)(1-\mu+x\mu)^t.$$

It is easy to see that $f_T'(x)$ is increasing as $0 < 1-\mu+x\mu < 1$. In addition, $f_T'(x)$ is bounded since

$$\begin{aligned} f_T'(x) &= \sum_{t=0}^{T} \mu t P_A(t)(1-\mu+x\mu)^t \\ &\leq \mu \sum_{t=0}^{T} t P_A(t) < \infty. \end{aligned}$$

Since $f_T'(x)$ is increasing and bounded, $\lim_{T\to\infty} f_T'(x)$ exists for all $x \in (0,1)$.

Next, note that for any $x \in (0,1)$, the difference between $f_T'(x)$ and $f'(x)$ is

$$\begin{aligned} f'(x) - f_T'(x) &= \sum_{t=T+1}^{\infty} \mu t P_A(t)(1-\mu+x\mu)^{t-1} \\ &\leq \mu \sum_{t=T+1}^{\infty} t P_A(t) \to 0 \end{aligned}$$

as $T \to \infty$, where the inequality follows from $0 < 1-\mu+x\mu < 1$ and the limit follows from the condition of fixed mean. Then, $\lim_{T\to\infty} f_T'(x) = f'(x)$ uniformly in $(0,1)$. That $f'(x) > 0$ follows from

$$0 < 1-\mu+x\mu < 1.$$

Similarly, we can show the existence and strict positivity of $f''(x)$, which completes the proof. □

Lemma 18. *The equation* $x = \sum_{t=0}^{\infty} P_A(t)(1-\mu+x\mu)^t$ *has a unique solution in* $(0,1)$.

Proof. Note that $x = 1$ is a solution to this fixed-point equation. First, we show that there is at least one fixed point in $(0,1)$.

Again, $f(x) = \sum_{t=0}^{\infty} P_A(t)(1-\mu+x\mu)^t > 0$ for $x = 0$. Hence, if there is no fixed point in $(0,1)$ this implies that $f(x)$ is strictly greater than x in $(0,1)$.

Now, consider the derivative of $f(x)$ at $1 - \frac{\delta}{\mu}$, which is $\mu \sum_t t(1-\delta)^t P_A(t) = \mu \widehat{A}(1-\delta)$, where $\widehat{A}(\alpha) = \mathbb{E}_{P_A} \alpha^A$. We know that generating function $\widehat{A}$ is continuous around 1. Hence, as $\delta \to 0$, $\widehat{A}(1-\delta) \to \frac{1}{\lambda}$. As $\frac{\mu}{\lambda} > 1$, there exists $\delta > 0$ such that $\mu \sum_t t(1-\delta)^t P_A(t) > 1$. This means that the derivative of $f(x)$ at $x = 1-\delta$ is > 1.

If $f(x) > x$ for all $x \in (0,1)$, then the following is true. From convexity of f,

$$\begin{aligned} f(1) &\geq f(1-\delta) + \delta f'(1-\delta) \\ &> 1 - \delta + \delta f'(1-\delta) \\ &> 1 - \delta + \delta = 1. \end{aligned}$$

This is a contradiction. So, there exists a fixed point in $(0,1)$.

Let us assume there is more than one fixed point in $(0,1)$. By Lem. 17, $f(x)$ is convex in $(0,1)$. A convex function can intersect a line at most twice. As $f(x)$ crosses $y = x$ at $x = 1$, there can be only one fixed point in $[0,1)$, but 0 is not a fixed point. □

A.3 Proof of Lemma 2

Note that for a geometric random variable A_i with mean $1/\lambda_i$ and letting $\alpha = (1-\mu-\sigma\mu) \in (0,1)$,

$$\begin{aligned} \widetilde{A}(P_{A_i}, \sigma) &= \sum_{t=1}^{\infty} \alpha^t P_{A_i}(t) \\ &= \sum_{t=1}^{\infty} \alpha^t (1-\lambda_i)^{t-1} \lambda_i = \alpha\lambda_i \sum_{t=0}^{\infty} (\alpha(1-\lambda_i))^t \\ &= \frac{\alpha\lambda_i}{(1-\alpha) + \alpha\lambda_i} = \frac{\frac{\alpha}{1-\alpha}}{\frac{1}{\lambda_i} + \frac{\alpha}{1-\alpha}}. \end{aligned} \tag{A.4}$$

Consider the sum-of-geometric random variables $\mathcal{A}^s$ first. Then for any sum-of-geometric random variable $P_{A^s} \in \mathcal{A}^s$,

$$\begin{aligned}
\widetilde{A}(P_{A^s}, \sigma) &= \sum_{t=1}^{\infty} \alpha^t P_{A^s}(t) \\
&= \sum_{t_i=1,\ 1\le i\le I}^{\infty} \alpha^{t_1+t_2+\cdots+t_I} P_{A_1}(t_1)\cdots P_{A_I}(t_I) \\
&= \prod_{i=1}^{I} \sum_{t_i=1}^{\infty} \alpha^{t_i} P_{A_i}(t_i) = \prod_{i=1}^{I} \frac{\frac{\alpha}{1-\alpha}}{\frac{1}{\lambda_i} + \frac{\alpha}{1-\alpha}}.
\end{aligned}$$

The last equality follows from (A.4). Note that the inequality

$$\prod_i (1 + x_i) \ge 1 + \sum_i x_i$$

holds for any $x_i > 0$. Hence, inverting both sides of this inequality and scaling both numerator and denominator by $\frac{\alpha}{1-\alpha}$,

$$\begin{aligned}
\widetilde{A}(p_{A^s}, \sigma) &= \prod_{i=1}^{I} \frac{\frac{\alpha}{1-\alpha}}{\frac{1}{\lambda_i} + \frac{\alpha}{1-\alpha}} \\
&\le \frac{\frac{\alpha}{1-\alpha}}{\sum_{i=1}^{I} \frac{1}{\lambda_i} + \frac{\alpha}{1-\alpha}} = \frac{\frac{\alpha}{1-\alpha}}{\frac{1}{\lambda} + \frac{\alpha}{1-\alpha}} \\
&= \widetilde{A}(\mathsf{geo}, \sigma).
\end{aligned}$$

Next since A^m is mixed, for any $P_{A^m} \in \mathcal{A}^m$,

$$\begin{aligned}
\widetilde{A}(P_{A^m}, \sigma) &= \sum_{t=1}^{\infty} \alpha^t P_{A^m}(t) = \sum_{i=1}^{I} c_i \sum_{t_i=1}^{\infty} \alpha^{t_i} P_{A_i}(t_i) \\
&= \sum_{i=1}^{I} c_i \frac{\frac{\alpha}{1-\alpha}}{\frac{1}{\lambda_i} + \frac{\alpha}{1-\alpha}}.
\end{aligned}$$

The last expression is convex in $1/\lambda_i$. Hence by Jensen's inequality

$$\begin{aligned}
\widetilde{A}(P_{A^m}, \sigma) &= \sum_{i=1}^{I} c_i \frac{\frac{\alpha}{1-\alpha}}{\frac{1}{\lambda_i} + \frac{\alpha}{1-\alpha}} \\
&\geq \frac{\frac{\alpha}{1-\alpha}}{\sum_{i=1}^{I} c_i \frac{1}{\lambda_i} + \frac{\alpha}{1-\alpha}} = \frac{\frac{\alpha}{1-\alpha}}{\frac{1}{\lambda} + \frac{\alpha}{1-\alpha}} \\
&= \widetilde{A}(\mathsf{geo}, \sigma).
\end{aligned}$$

A.4 Proof of Theorem 3

Consider the following transition probability for $q > 0$:

$$\begin{aligned}
&\mathbb{P}(Q_{i+1} = q' | Q_i = q, Q_{i-1}, \ldots) \\
&= \mathbb{P}(\text{there are } q' - q + 1 \text{ arrivals} \\
&\qquad \text{between departures } i-1 \text{ and } i \mid Q_i = q, Q_{i-1}, \ldots) \\
&= \mathbb{P}(\text{sum of } q' - q + 1 \text{ geometric times} \\
&\qquad \leq \text{interdeparture time between } i-1 \text{ and } i) \qquad \text{(A.5)} \\
&= \sum_{t=0}^{\infty} P_S(t) \mathbb{P}(\text{Bin}(t, \lambda) = q' - q + 1) \qquad \text{(A.6)} \\
&= \sum_{t=0}^{\infty} P_S(t) \binom{t}{q' - q + 1} (1-\mu)^{t - q' + q - 1} \mu^{q' - q + 1} \\
&= k_{q'-q+1}.
\end{aligned}$$

Eq. (A.5) follows because geometric random variables are memoryless. Geometric interarrival is the same as Bernoulli arrival per time slot, and the sum of Bernoulli variables is binomial, which leads to (A.6).

When $Q_i = 0$, note that just before the $(i+1)$th arrival, the queue-length is 0, and it is 1 just after the $(i+1)$th arrival. Then the probability that $Q_{i+1} = q'$ is equal to the probability that there are exactly q' arrivals during the service time of the $(i+1)$th job. From above, this is equal to $k_{q'}$.

This proves $\{Q_i\}$ is Markov; irreducibility and aperiodicity follow since $\mathbb{P}(Q_{i+1} = Q_i + \delta | Q_i) > 0$ for $\delta \in \{0, \pm 1\}$.

From $\pi = \pi[P]$ for this Markov chain it follows that

$$\begin{aligned} \pi_0 k_0 + \pi_1 k_0 &= \pi_0 \\ \pi_0 k_1 + \pi_1 k_1 + \pi_2 k_0 &= \pi_1 \\ &\vdots \end{aligned}$$

Multiplying the first equation by z^0, the second by z, the third by z^2, and so on, and then summing all of them we get

$$\pi_0 K(z) + K(z)(\pi_1 + \pi_2 z + \cdots) = \Pi(z),$$

which, after some algebra, gives

$$\Pi(z) = \frac{\pi_0(z-1)K(z)}{z - K(z)}.$$

We know that $\Pi(1) = 1$, so the left side must also be 1 for $z = 1$. But it is $\frac{0}{0}$ when evaluated at $z = 1$, as $K(1) = \sum_j k_j = 1$. Thus using l'Hôpital's rule we get

$$\pi_0 = \frac{1 - K'(1)}{K(1)}.$$

Note that $K'(z) = \sum_j j k_j z^j$ which gives $K'(1) = \sum_j j k_j$, i.e., $K'(1)$ is the expected number of arrivals in a time distributed as P_S. As arrivals are Bernoulli and are independent from service times, from Wald's lemma we get

$$K'(1) = \frac{\lambda}{\mu},$$

which in turn gives $\pi_0 = 1 - \frac{\lambda}{\mu}$.

From $\Pi(z)$ we can obtain π_1 by evaluating $\frac{\Pi(z)-\pi_0}{z}$ as $z \to 0$. By repeating the procedure we can obtain π_k by evaluating the limit of $\frac{\Pi(z)-\sum_{j=0}^{k-1}\pi_j z^j}{z^k}$ as $z \to 0$.

Appendix B

Proofs for Chapter 3

B.1 Proof of Theorem 7 (GI/M/1 queues)

In the case of GI/M/1 queues, it is easier to derive the queue-length distribution seen by ith arrival (i.e., just prior to arrivals), say $\widehat{Q}_i$, than Q_i because of the memoryless property of the server. From the same argument as in the proof of Lem. 3, we know that generic random variable $\widehat{Q} \overset{d}{=} Q$ when it is stationary, so we will consider $\widehat{Q}$ instead of Q.

Notice that $\widehat{Q}_{n+1} = (\widehat{Q}_n - \beta_n + 1)_+$, where β_n is the number of jobs completed during the inter-arrival time A_{n+1}. As $\{A_n\}$ is i.i.d., it does not depend on the past history of the queue and neither does β_n. Therefore, $\widehat{Q}_n$ forms a discrete-time Markov chain.

Define ℓ_q to be the probability of q job completions between two consecutive arrivals, i.e.,

$$\ell_q := \mathbb{P}[\beta_n = q | \widehat{Q}_n \geq q] = \int_0^\infty P^A(t) \frac{e^{-\mu t}(\mu t)^q}{q!} dt. \tag{B.1}$$

Then, the transition matrix $[P]$ is given by[1]

$$[P] = \begin{bmatrix} 1-\ell_0 & \ell_0 & 0 & 0 & \cdots \\ 1-\ell_0-\ell_1 & \ell_1 & \ell_0 & 0 & \cdots \\ 1-\ell_0-\ell_1-\ell_2 & \ell_2 & \ell_1 & \ell_0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix},$$

[1] Such a matrix is called a lower Hessenberg matrix.

and the stationary distribution relationship $\widehat{\pi} = \widehat{\pi}[P]$ yields

$$\widehat{\pi}(0) = \sum_{q=0}^{\infty} \widehat{\pi}(q) \left(1 - \sum_{i=0}^{q} \ell_i\right),$$
$$\widehat{\pi}(i) = \sum_{q=0}^{\infty} \ell_q \widehat{\pi}(i + q - 1) \text{ for } i > 1. \tag{B.2}$$

As the stationary distribution is unique, it suffices to show that $\widehat{\pi}(q) = \widehat{\pi}(0)\sigma^q$ for some $\sigma < 1$. Substituting $\widehat{\pi}(q) = \widehat{\pi}(0)\sigma^q$ into (B.2), we have

$$\sigma = \sum_{q=0}^{\infty} \ell_q \sigma^q =: B(\sigma). \tag{B.3}$$

Note that $B(0) = \ell_0 > 0$, $B(1) = 1$, and $B(\sigma)$ is convex over $\sigma \in [0, 1]$ since $B'(\sigma), B''(\sigma) \geq 0$. There are two possible cases: no fixed point in $(0, 1)$ or a unique fixed point in $(0, 1)$. Recall $B(\sigma)$ is a probability generating function of ℓ_q and thus, $B'(1) = \frac{\mu}{\lambda} = \rho^{-1} > 1$ since it is the number of job completions normalized by inter-arrivals. Therefore, the latter is the only possibility and the fixed point in $(0, 1)$ is unique. Let σ^* denote the solution.

On the other hand, substituting (B.1) into (B.3),

$$\begin{aligned}
\sigma = \sum_{q=0}^{\infty} \ell_q \sigma^q &= \sum_{q=0}^{\infty} \left(\int_0^{\infty} P^A(t) \frac{e^{-\mu t}(\mu t)^q}{q!} dt \right) \sigma^q \\
&= \int_0^{\infty} P^A(t) \sum_{q=0}^{\infty} \frac{e^{-\mu t}}{q!} (\mu t \sigma)^q dt \\
&= \int_0^{\infty} P^A(t) e^{-\mu t} \sum_{q=0}^{\infty} \frac{(\mu t \sigma)^q}{q!} dt \\
&= \int_0^{\infty} P^A(t) e^{-\mu t(1-\sigma)} dt \\
&= A^*(\mu(1 - \sigma)),
\end{aligned}$$

where $A^*(\cdot)$ is the Laplace-Stieltjes transform of $P^A(t)$. Hence, the fixed point solution σ^* is the unique root of

$$\sigma = A^*(\mu(1 - \sigma)).$$

As $\sum_q \widehat{\pi}(q) = 1$, it is easy to see $\widehat{\pi}(0) = 1 - \sigma^*$. Therefore,

$$\pi(q) = \widehat{\pi}(q) = (1 - \sigma^*)(\sigma^*)^q.$$

B.2 Proof of Theorem 8 ($\mathsf{M}/\mathsf{GI}/1$ queues)

To derive $\pi(Q)$ in closed form, we will first show that $\{Q_i\}$ forms a Markov chain, and then represent the stationary distribution in terms of P^S.

Let Q_{n+1} be the queue-length seen by $(n+1)$th departure. Then, we observe that

$$Q_{n+1} = \begin{cases} Q_n + \alpha_{n+1} - 1 & \text{if } Q_n \geq 1, \\ \alpha_{n+1} & \text{if } Q_n = 0, \end{cases}$$

where α_{n+1} is the number of jobs arriving during the service time of $(n+1)$th job. Since α_{n+1} is independent of past history $\{Q_n, Q_{n-1}, \cdots, Q_1\}$, we know that $\{Q_n\}$ forms a discrete-time Markov chain. Furthermore, it is time-homogeneous as inter-arrivals and services are i.i.d.

Denote the transition probability of the Markov chain by $p_{ij} := \mathbb{P}[Q_{n+1} = j | Q_n = i]$. Then,

$$p_{ij} = \begin{cases} \mathbb{P}[j - i + 1 \text{ arrivals during service}] & \text{if } i \geq 1, \\ \mathbb{P}[j \text{ arrivals during service}] & \text{if } i = 0. \end{cases}$$

We obtain p_{ij} by marginalizing joint probability. Since the number of arrivals is Poisson,

$$\begin{aligned} &\mathbb{P}[q \text{ arrivals during service}] \\ &= \int_0^\infty \mathbb{P}[S = t \text{ and } q \text{ arrivals}] dt \\ &= \int_0^\infty P^S(t) \mathbb{P}[q \text{ arrivals} | S = t] dt \\ &= \int_0^\infty P^S(t) \frac{e^{-\lambda t}(\lambda t)^q}{q!} dt. \end{aligned}$$

Letting $k_q := \mathbb{P}[q \text{ arrivals during service}]$ for brevity, the transition matrix is given as follows.[2]

$$[P] = \begin{bmatrix} k_0 & k_1 & k_2 & k_3 & \cdots \\ k_0 & k_1 & k_2 & k_3 & \cdots \\ 0 & k_0 & k_1 & k_2 & \cdots \\ 0 & 0 & k_0 & k_1 & \cdots \\ 0 & 0 & 0 & k_0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

Solving the stationary distribution identity $\pi = \pi[P]$,

$$\pi(q) = \pi(0)k_q + \sum_{j=1}^{q+1} \pi(j)k_{q-j+1} \quad \text{for } q = 0, 1, \cdots .$$

Multiplying the equations of each q by z^q and summing over $q = 0, 1, \cdots$, we have

$$\Pi(z) = \frac{\pi(0)(1-z)K(z)}{K(z) - z},$$

where $\Pi(z), K(z)$ are probability generating functions of $\pi(q), k_q$,

$$\Pi(z) = \sum_{q=0}^{\infty} \pi(q)z^q \text{ and } K(z) = \sum_{q=0}^{\infty} k_q z^q.$$

Note that $K(1) = \sum_q k_q = 1$. By l'Hôpital's rule at $z = 1$, we have $\pi(0) = 1 - K'(1)$. Since k_q is the normalized number of arrivals during service time, the first moment $K'(1) = \rho = \frac{\lambda}{\mu}$, which implies $\pi(0) = 1 - \rho$. Therefore,

$$\Pi(z) = \frac{(1-\rho)(1-z)K(z)}{K(z) - z}.$$

B.3 Proof of Lemma 6

To prove the 'seen by departures' result, we start from continuous-time ergodicity in [118]. We first take a continuous-time piecewise-deterministic Markov process [119]. Then, since it is strong Markov, the stopped process at user k departures forms a stationary and ergodic

[2] Such a matrix is called an upper Hessenberg matrix.

discrete-time Markov chain. Suppose that once job processing is completed and the job departs at time t, the next job enters the server at time t^+.

Let us take a continuous-time Markov process $Z(t) := (\mathbf{L}(t), \mathbf{A}(t), \mathbf{S}(t)) \in \mathcal{Z}$, where

- $\mathbf{L}(t)$ is the vector of transmitter jobs in order of their arrivals including the job in the server. If the system is empty, $\mathbf{L}(t) = \emptyset$. Otherwise, $\mathbf{L}(t) = (\ell_0, \ell_1, \ell_2, \cdots) \in [1 : K]^{Q(t)+1}$, where $Q(t)$ is the queue-length at time t.

- $\mathbf{A}(t) \in \mathbb{R}_+^K$ is the residual arrival time vector whose component $A_k(t)$ indicates the remaining time until the next arrival of kth user.

- $\mathbf{S}(t) \in (\mathbb{R}_+ \cup \infty)^K$ is the residual service time vector whose component $S_k(t)$ indicates residual service time if user k's job is being served, infinite otherwise.

Under condition (3.7), this is Harris recurrent so that there exists the stationary distribution $\widehat{\pi}$ and the following holds [118, Thm. 6.4]: For any $g : \mathcal{Z} \mapsto \mathbb{R}_+$,

$$\lim_{t\to\infty} \frac{1}{t} \int_0^t g(Z(s))ds = \mathbb{E}_{\widehat{\pi}}[g(Z)] \text{ almost surely.} \tag{B.4}$$

Fix a user k and take a sequence of stopping times $(t_1, t_2, \cdots)$ such that $t_n := \min\{t > t_{n-1} : S_k(t-) > 0, S_k(t) = 0\}$ (assume $t_0 < 0$ for simplicity), i.e., the sequence of hitting times at which user kth job departs. Take a small $\Delta > 0$ and $g_1 := \mathbf{1}_{\{S_k(t) \leq \Delta\}}, g_2 := \mathbf{1}_{\{|\mathbf{L}(t)|=q+1, S_k(t)\leq\Delta\}}$. Since either inter-arrival time distributions or service time distribution are continuous, we know that $\widehat{\pi}$ is also continuous. Therefore, (B.4) implies

$$\lim_{n\to\infty} \frac{1}{t_n} \int_0^{t_n} g_1(Z(s))ds \approx \Delta \cdot \widehat{\pi}\{Z(t) : S_k(t) = 0\},$$
$$\lim_{n\to\infty} \frac{1}{t_n} \int_0^{t_n} g_2(Z(s))ds \approx \Delta \cdot \widehat{\pi}\{Z(t) : Q(t) = q, S_k(t) = 0\}.$$

Taking $\Delta \to 0$ and using the fact that the queue-length is a deterministic function of $\mathbf{L}(t)$, it follows that the stationary distribution exists and

$$\pi_{Kk}(q) := \frac{\widehat{\pi}\{Z(t) : |\mathbf{L}(t)| = q + 1, S_k(t) = 0\}}{\widehat{\pi}\{Z(t) : S_k(t) = 0\}}. \tag{B.5}$$

Next show the ergodicity. Define samplings

$$\begin{aligned} h_1(Z(t)) &:= \mathbf{1}_{\{S_k(t)\leq\Delta\}}, \\ h_2(Z(t)) &:= \mathbf{1}_{\{S_k(t)\leq\Delta\}} f(q(t)), \end{aligned}$$

and note that

$$\lim_{n\to\infty}\frac{1}{t_n}\int_0^{t_n} h_1(Z(s))ds = \lim_{n\to\infty}\frac{n\Delta}{t_n} = \lambda_{Kk}\Delta$$

and

$$\begin{aligned} \lim_{n\to\infty}\frac{1}{t_n}\int_0^{t_n} h_2(Z(s))ds &= \lim_{n\to\infty}\frac{1}{t_n}\sum_{i=1}^{n} f(q(t_j))\Delta \\ = \lim_{n\to\infty}\frac{n}{t_n}\frac{1}{n}\sum_{i=1}^{n} f(q(t_i))\Delta &= \lambda_{Kk}\Delta\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} f(q(t_i)), \end{aligned}$$

where $\lim_n \frac{n}{t_n}\to\lambda_{Kk}$ is used due to the system stability. Then,

$$\frac{\lim_{n\to\infty}\frac{1}{t_n}\int_0^{t_n} h_2(Z(s))ds}{\lim_{n\to\infty}\frac{1}{t_n}\int_0^{t_n} h_1(Z(s))ds} = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} f(q(t_i)). \tag{B.6}$$

Also letting $\Delta\to 0$ and applying (B.4) to the left side of (B.6),

$$\begin{aligned} \frac{\lim_{n\to\infty}\frac{1}{t_n}\int_0^{t_n} h_2(Z(s))ds}{\lim_{n\to\infty}\frac{1}{t_n}\int_0^{t_n} h_1(Z(s))ds} &= \frac{\mathbb{E}_{\widehat{\pi}}[h_2(Z)]}{\mathbb{E}_{\widehat{\pi}}[h_1(Z)]} \\ = \frac{\sum_{q=0}^{\infty} f(q)\widehat{\pi}\{Z(t):S_k(t)=0,|\mathbf{L}(t)|=q+1\}}{\widehat{\pi}\{Z(t):S_k(t)=0\}} \\ = \sum_{q=0}^{\infty} f(q)\frac{\widehat{\pi}\{S_k(t)=0,|\mathbf{L}(t)|=q+1\}}{\widehat{\pi}\{S_k(t)=0\}} \\ = \sum_{q=0}^{\infty} f(q)\pi_{Kk}(q) = \mathbb{E}_{\pi_{Kk}}[f(Q)]. \end{aligned}$$

Since $Q_i = Q(t_i)$, the following holds:

$$\frac{1}{n}\sum_{i=1}^{n} f(q_i) = \frac{1}{n}\sum_{i=1}^{n} f(q(t_i)) = \mathbb{E}_{\pi_{Kk}}[f(Q)]$$

almost surely.

B.4 Proof of Lemma 7

We restricted to PPs over bounded B so Φ_K, Φ_K^* both have no events outside of B. Therefore it is sufficient to show that for all $B' \in \mathfrak{B}$ such that $B' \subset B$,

$$d_{\mathsf{TV}}(N_K(B'), N_K^*(B')) \to 0 \text{ as } K \to \infty.$$

Note that Poisson processes are infinitely divisible, so we can split into K independent Poisson PPs $\{\Phi_{Kk}^*\}_{k\in[1:K]}$ with intensity λ_{Kk}. Let N_{Kk}^* be the counting measure of Φ_{Kk}^*. From the Poisson distribution and its Taylor expansion when $|B|\lambda_{Kk}$ is small:

$$\begin{aligned}&\mathbb{P}[N_{Kk}^*(B) = 1] = |B|\lambda_{Kk} + O(|B|^2\lambda_{Kk}^2),\\&\mathbb{P}[N_{Kk}^*(B) \geq 2] = O(|B|^2\lambda_{Kk}^2).\end{aligned}$$

Hence, the total variational distance between individual PPs is computed as follows, where argument B is omitted for simplicity.

$$\begin{aligned}
&2d_{\mathsf{TV}}(N_{Kk}, N^*_{Kk}) \\
= &\sum_{j \in \mathbb{Z}_+} |\mathbb{P}[N_{Kk} = j] - \mathbb{P}[N^*_{Kk} = j]| \\
= &|(1 - \mathbb{P}[N_{Kk} \geq 1]) - (1 - \mathbb{P}[N^*_{Kk} \geq 1])| \\
&\quad + \sum_{j \geq 1} |\mathbb{P}[N_{Kk} = j] - \mathbb{P}[N^*_{Kk} = j]| \\
= &|\mathbb{P}[N^*_{Kk} = 1] + \mathbb{P}[N^*_{Kk} \geq 2] - \mathbb{P}[N_{Kk} = 1] \\
&\quad - \mathbb{P}[N_{Kk} \geq 2]| + \sum_{j \geq 1} |\mathbb{P}[N_{Kk} = j] - \mathbb{P}[N^*_{Kk} = j]| \\
\overset{(a)}{\leq} &|\mathbb{P}[N^*_{Kk} = 1] - \lambda_{Kk}| + O(|B|^2 \lambda^2_{Kk}) + \mathbb{P}[N_{Kk} \geq 2] \\
&\quad + \sum_{j \geq 1} |\mathbb{P}[N_{Kk} = j] - \mathbb{P}[N^*_{Kk} = j]| \\
\overset{(b)}{\leq} &O(|B|^2 \lambda^2_{Kk}) + \mathbb{P}[N_{Kk} \geq 2] \\
&\quad + \sum_{j \geq 1} |\mathbb{P}[N_{Kk} = j] - \mathbb{P}[N^*_{Kk} = j]| \\
\overset{(c)}{\leq} &O(|B|^2 \lambda^2_{Kk}) + \mathbb{P}[N_{Kk} \geq 2] + \mathbb{P}[N_{Kk} \geq 2] + \mathbb{P}[N^*_{Kk} \geq 2] \\
= &O(|B|^2 \lambda^2_{Kk}) + 2\mathbb{P}[N_{Kk} \geq 2],
\end{aligned}$$

where (a) follows from the triangle inequality, (3.9), and the Taylor expansion; (b) follows from the Taylor expansion; and (c) follows from the triangle inequality and the Taylor expansion.

Now we bound the total variation between two sums of independent random variables as follows.

$$\begin{aligned}
d_{\mathsf{TV}}(N_K, N_K^*) &\overset{(a)}{\leq} \sum_{k\in[1:K]} d_{\mathsf{TV}}(N_{Kk}, N_{Kk}^*) \\
&\overset{(b)}{\leq} \sum_{k\in[1:K]} O\left(|B|^2\lambda_{Kk}^2\right) + \sum_{k\in[1:K]} \mathbb{P}[N_{Kk} \geq 2] \\
&\leq c|B|^2 \cdot \sum_{k\in[1:K]} \lambda_{Kk} \left(\max_{k\in[1:K]} \lambda_{Kk}\right) + \sum_{k\in[1:K]} \mathbb{P}[N_{Kk} \geq 2] \\
&= c|B|^2 \cdot \lambda_K^* \cdot g_2(K) + \sum_{k\in[1:K]} \mathbb{P}[N_{Kk} \geq 2],
\end{aligned}$$

where (a) follows from the total variation inequality for product measures, and (b) follows from the above derivation.

Therefore, the first term vanishes at speed $O(|B|^2 g_2(K))$, the second term $\sum_k P[N_{Kk} \geq 2] \to 0$ at speed $O(g_1(K,B))$. So the overall speed of convergence is given by $O(g(K,B))$, where $g(K,B) := \max\{g_1(K,B), |B|^2 g_2(K)\}$.

Finally, for all subsets $B' \subset B$ with $B' \in \mathfrak{B}$, we can repeat the above argument, but the speed of convergence still holds since $g_1(K,B') \leq g_1(K,B)$ and $|B'|g_2(K) \leq |B|g_2(K)$.

B.5 Proof of Lemma 8

We will first restrict the superposed RMPP on B, and then apply the data processing inequality (also known as monotone theorem in some literature [120]) to show $Q_i^{(K)} \overset{\mathsf{TV}}{\to} Q_i^*$. Without loss of generality, we only consider some arbitrary ith symbol whose arrival was at $t_i > 0$.

Let us introduce *empty points* [75]. When $\phi(t)$ is a specific realization of $\Phi(t)$, an arrival time instance $e_j(\phi)$ at which there is no job in the system (in the queue and in the server both) is called an empty point.[3] List $e_j(\phi)$ in order

$$\cdots < e_{-1}(\phi) < e_0(\phi) \leq 0 < e_1(\phi) < \cdots .$$

The jth empty point implies that the queue state after $t = e_j(\phi)$ is completely determined only by arrivals after $e_j(\phi)$. Then, we know that $e_0(\Phi_K) \overset{\mathsf{TV}}{\to} e_0(\Phi^*)$ with speed $O(g(K,B))$

[3] This is different from the regenerative cycles, introduced in Sec. 3.2. Since we are considering arbitrary superposition process Φ that is not renewal in general, $e_j(\Phi)$ is not regenerative.

by data processing inequality and thus, $e_j(\Phi_K) \xrightarrow{\mathsf{TV}} e_j(\Phi^*)$ for any j by stationarity.

Take a set of PP realizations $A_{u_1} := \{\phi : -u_1 < e_0(\phi) \leq 0\}$. Since $e_0(\Phi_K) \xrightarrow{\mathsf{TV}} e_0(\Phi^*)$, for arbitrary $\epsilon_1 > 0$ it is possible to take u_1, K_0 such that for all $K > K_0$,

$$P_K[A_{u_1}] > 1 - \epsilon_1 \text{ and } P^*[A_{u_1}] > 1 - \epsilon_1.$$

Also, take a set $A_{u_2} := \{\phi : 0 < t_i(\phi) < u_2\}$. Thus it is immediate that for arbitrary $\epsilon_2 > 0$ we can take $u_2 > 0$ such that $P_K[A_{u_2}] > 1 - \epsilon_2$ and $P^*[A_{u_2}] > 1 - \epsilon_2$.

Let $q(i, \phi)$ be the queue-length seen by ith departure of ϕ, and $u := \max(u_1, u_2)$, $\epsilon := \epsilon_1 + \epsilon_2$. By the property of the empty point and A_{u_1}, A_{u_2},

$$\begin{aligned} P^*[\phi : q(i, \phi) = q(i, \mathbf{1}_{[-u,u)}\phi)] &\geq P^*[A_{u_1} \cap A_{u_2}] > 1 - \epsilon, \\ P_K[\phi : q(i, \phi) = q(i, \mathbf{1}_{[-u,u)}\phi)] &\geq P_K[A_{u_1} \cap A_{u_2}] > 1 - \epsilon. \end{aligned}$$

Setting $B = [-u, u)$, we can bound total variation as follows:

$$\begin{aligned} & d_{\mathsf{TV}}(Q_i(\Phi_K), Q_i(\Phi^*)) \\ \overset{(a)}{\leq} \; & d_{\mathsf{TV}}(Q_i(\Phi_K), Q_i(\mathbf{1}_B\Phi_K)) + d_{\mathsf{TV}}(Q_i(\mathbf{1}_B\Phi_K), Q_i(\mathbf{1}_B\Phi^*)) \\ & \quad + d_{\mathsf{TV}}(Q_i(\mathbf{1}_B\Phi^*), Q_i(\Phi^*)) \\ \overset{(b)}{\leq} \; & 2\epsilon + d_{\mathsf{TV}}(Q_i(\mathbf{1}_B\Phi_K), Q_i(\mathbf{1}_B\Phi^*)) \\ \overset{(c)}{\leq} \; & 2\epsilon + d_{\mathsf{TV}}(\mathbf{1}_B\Phi_K, \mathbf{1}_B\Phi^*) \leq 2\epsilon + O(g(K, B)), \end{aligned}$$

where (a) follows from the triangle inequality, (b) follows from the property of empty point, and (c) follows from the data processing inequality since $Q_i(\cdot)$ is a function of a PP. Since ϵ_1, ϵ_2 are arbitrary, the statement is proved.

Appendix C

Proofs for Chapter 4

C.1 Proof of Theorem 15

Let us prove Thm. 15 starting with the premise that $q_1^* \geq p_0$. First, from (4.25), we have

$$q_1^* \geq p_0 \Leftrightarrow \frac{P_{e,2}^{\mathrm{II}_1} - P_{e,2}^{\mathrm{II}_0}}{P_{e,2}^{\mathrm{I}_1} - P_{e,2}^{\mathrm{I}_0}} \geq -1. \tag{C.1}$$

To study the ratio in (C.1), consider the Type I vs. Type II error curve for binary hypothesis testing under additive Gaussian noise.[1] This is shown in Fig. C.1, and as seen here is a convex function [108]. Note that on the curve, the Type I and Type II error probabilities, $(P_e^{\mathrm{I}}, P_e^{\mathrm{II}})$, are the points on the curve that have tangents with slope matching $-\left(\frac{c_{10}q}{c_{01}(1-q)}\right)$, where q is the corresponding prior probability, and σ^2 is the variance of the additive Gaussian noise.

First, from Thm. 12, we know that $q_2^0 \geq q_2^1$ which in turn implies that $\lambda_2^0 \geq \lambda_2^1$. This in turn indicates that

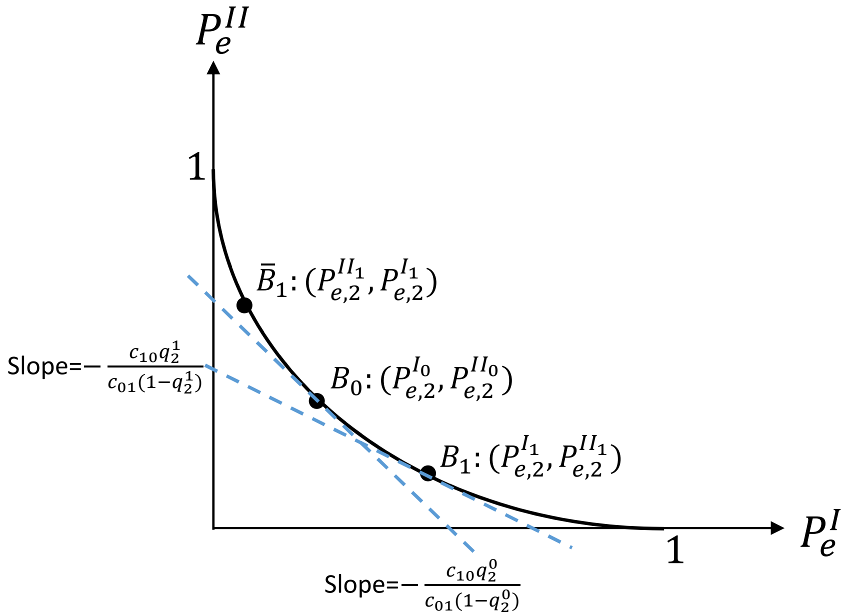
$$P_{e,2}^{\mathrm{I}_0} = Q\left(\tfrac{\lambda_2^0}{\sigma_2}\right) \leq Q\left(\tfrac{\lambda_2^1}{\sigma_2}\right) = P_{e,2}^{\mathrm{I}_1}.$$

Similarly, $P_{e,2}^{\mathrm{II}_0} \geq P_{e,2}^{\mathrm{II}_1}$, and thus, as shown in the figure, the point $B_0 = \left(P_{e,2}^{\mathrm{I}_0}, P_{e,2}^{\mathrm{II}_0}\right)$ lies to the left of $B_1 = \left(P_{e,2}^{\mathrm{I}_1}, P_{e,2}^{\mathrm{II}_1}\right)$.

Further, since B_1 lies on the curve, so does the point $\bar{B}_1 = \left(P_{e,2}^{\mathrm{II}_1}, P_{e,2}^{\mathrm{I}_1}\right)$ as it caters to the error probabilities corresponding to the probability of the null hypothesis $\mathbb{P}\left[H=0\right] = 1 - q_2^1$. Thus, the line $\overline{B_1\bar{B}_1}$ has a slope of -1.

Note that condition (C.1) translates to the slope of the line $\overline{B_0B_1}$ being greater than -1. Observe that if $\bar{B}_1$ lies to the right of B_1 then it implies that the slope of $\overline{B_0B_1}$ is less than -1, violating (C.1). Similarly, if B_0 lies to the left of $\bar{B}_1$, then again the (C.1) is violated.

[1] It is also called receiver operating characteristic (ROC) curve [108, 121] when the curve is vertically inverted.



P_e^{II}
1
$\bar{B}_1: (P_{e,2}^{II_1}, P_{e,2}^{I_1})$
Slope= $-\frac{c_{10}q_2^1}{c_{01}(1-q_2^1)}$
$B_0: (P_{e,2}^{I_0}, P_{e,2}^{II_0})$
$B_1: (P_{e,2}^{I_1}, P_{e,2}^{II_1})$
1
P_e^I
Slope= $-\frac{c_{10}q_2^0}{c_{01}(1-q_2^0)}$

Figure C.1: The point B_0 always exists between points B_1 and $\bar{B}_1$.

On the other hand, if B_0 lies between $\bar{B}_1$ and B_1, then we know that the slope of $\overline{B_0 B_1}$ is greater than that of $\overline{B_1 \bar{B}_1}$, therein satisfying (C.1). Thus, (C.1) is true if and only if the point B_0 lies between the two points B_1 and $\bar{B}_1$.

From the convexity of the curve and comparing coordinates of B_0 and $\bar{B}_1$, we have

$$\begin{aligned} q_1^* \geq p_0 &\Leftrightarrow P_{e,2}^{\mathrm{I}_0} \geq P_{e,2}^{\mathrm{II}_1} \text{ and } P_{e,2}^{\mathrm{II}_0} \leq P_{e,2}^{\mathrm{I}_1} \\ &\overset{(a)}{\Leftrightarrow} Q\left(\frac{\lambda_2^0}{\sigma_2}\right) \geq 1 - Q\left(\frac{\lambda_2^1 - 1}{\sigma_2}\right) \text{ and } Q\left(\frac{\lambda_2^1}{\sigma_2}\right) \geq 1 - Q\left(\frac{\lambda_2^0 - 1}{\sigma_2}\right) \\ &\overset{(b)}{\Leftrightarrow} \lambda_2^0 + \lambda_2^1 \leq 1 \\ &\overset{(c)}{\Leftrightarrow} 2\lambda_{1,[2]} + \sigma_2^2 \log\left(\frac{P_{e,1,[2]}^{\mathrm{I}}\left(1 - P_{e,1,[2]}^{\mathrm{I}}\right)}{P_{e,1,[2]}^{\mathrm{II}}\left(1 - P_{e,1,[2]}^{\mathrm{II}}\right)}\right) \leq 1, \end{aligned} \qquad \text{(C.2)}$$

where (a) follows from the false alarm and missed detection probabilities in terms of the Q-function of the standard Gaussian random variable; (b) follows from the fact that the Q-function is monotonically decreasing and that $1 - Q(x) = Q(-x)$; and (c) follows from (4.22), (4.23), and $\lambda_{1,[2]} = \lambda_2(q_2)$.

From (4.28), we have

$$\lambda_{1,[2]} = \frac{1}{2} + \sigma_2^2 \log\left(\frac{c_{10}q_2^*}{c_{01}(1 - q_2^*)}\right).$$

Substituting in (C.2), we have

$$q_1^* \geq p_0 \Leftrightarrow 2\log\left(\frac{c_{10}q_2^*}{c_{01}(1-q_2^*)}\right) \leq \log\left(\frac{P_{e,1,[2]}^{\mathrm{II}}\left(1-P_{e,1,[2]}^{\mathrm{II}}\right)}{P_{e,1,[2]}^{\mathrm{I}}\left(1-P_{e,1,[2]}^{\mathrm{I}}\right)}\right).$$

Letting $x := \log\left(\frac{c_{10}q_2^*}{c_{01}(1-q_2^*)}\right) = \frac{1}{\sigma_2^2}\left(\lambda_2 - \frac{1}{2}\right)$ and using $Q(\cdot)$ representation of error probabilities, we have

$$q_1^* \geq p_0 \Leftrightarrow 2x \leq \log\left(\frac{Q\left(\sigma_2 x - \frac{1}{2\sigma_2}\right)Q\left(-\sigma_2 x + \frac{1}{2\sigma_2}\right)}{Q\left(\sigma_2 x + \frac{1}{2\sigma_2}\right)Q\left(-\sigma_2 x - \frac{1}{2\sigma_2}\right)}\right). \tag{C.3}$$

From Cor. 10, we know that the function

$$\widetilde{g}(x) = x + \log\left(\frac{Q\left(\sigma x + \frac{1}{2\sigma}\right)}{Q\left(\sigma x - \frac{1}{2\sigma}\right)}\right)$$

is an increasing function of x. Thus, reformulating (C.3) using $\widetilde{g}(\cdot)$,

$$\begin{aligned} q_1^* \geq p_0 &\Leftrightarrow \widetilde{g}(x) \leq \widetilde{g}(-x) \\ &\Leftrightarrow x \leq 0 \Leftrightarrow q_2^* \leq \frac{c_{01}}{c_{01}+c_{10}}. \end{aligned}$$

The condition for equality follows from observing the condition for equality at all the inequalities, proving the first part of the result.

The second part follows directly from the first, taking into account the trivial cases of $p_0 \in \{0,1\}$.

C.2 Proof of Theorem 16

We will consider the case of $c_{01} = c_{10} = 1$ for convenience. The proof extends directly by a simple scaling argument.

The optimal belief of worker two satisfies $\frac{\partial R_2}{\partial q_2} = 0$. Thus, differentiating (4.24) with respect to q_2 and rearranging,

$$p_0 \left[(1 - P^{\mathrm{I}}_{e,1}) f_{Y_2|H}(\lambda^0_2|0) \frac{\partial \lambda^0_2}{\partial q_2} + P^{\mathrm{I}}_{e,1} f_{Y_2|H}(\lambda^1_2|0) \frac{\partial \lambda^1_2}{\partial q_2} \right] =$$
$$(1 - p_0) \left[P^{\mathrm{II}}_{e,1} f_{Y_2|H}(\lambda^0_2|1) \frac{\partial \lambda^0_2}{\partial q_2} + (1 - P^{\mathrm{II}}_{e,1}) f_{Y_2|H}(\lambda^1_2|1) \frac{\partial \lambda^1_2}{\partial q_2} \right].$$

Let $x = \log\left(\frac{p_0}{1-p_0}\right)$. For $q^*_2 = 1/2$ and $q^*_1 = p_0$, we have

$$\lambda_1 = \frac{1}{2} + \sigma^2_1 x \text{ and } \lambda_{1,[2]} = \frac{1}{2}.$$

It implies $P^{\mathrm{I}}_{e,1,[2]} = P^{\mathrm{II}}_{e,1,[2]} = Q(1/2\sigma_2)$. Then,

$$\begin{aligned} \mathcal{L}(\lambda^0_2) &= \frac{f_{Y_2|H}(\lambda^0_2|1)}{f_{Y_2|H}(\lambda^0_2|0)} = \frac{q_2}{1-q_2} \frac{(1 - P^{\mathrm{I}}_{e,1,[2]})}{P^{\mathrm{II}}_{e,1,[2]}} = \frac{Q(-1/2\sigma_2)}{Q(1/2\sigma_2)} =: \frac{1}{c}, \\ \mathcal{L}(\lambda^1_2) &= \frac{f_{Y_2|H}(\lambda^1_2|1)}{f_{Y_2|H}(\lambda^1_2|0)} = \frac{q_2}{1-q_2} \frac{P^{\mathrm{I}}_{e,1,[2]}}{(1 - P^{\mathrm{II}}_{e,1,[2]})} = \frac{Q(1/2\sigma_2)}{Q(-1/2\sigma_2)} = c. \end{aligned}$$

Equivalently, this implies that

$$\lambda^0_2 = \frac{1}{2} + \sigma^2 \log\left(\frac{1}{c}\right), \quad \lambda^1_2 = \frac{1}{2} - \sigma^2 \log\left(\frac{1}{c}\right).$$

Thus, $\lambda^0_2 + \lambda^1_2 = 1$, and so

$$\begin{aligned} f_{Y_2|H}(\lambda^1_2|1) &= \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left(-\tfrac{(\lambda^1_2 - 1)^2}{2\sigma^2_2}\right) \\ &= \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left(-\tfrac{(\lambda^0_2)^2}{2\sigma^2_2}\right) = f_{Y_2|H}(\lambda^0_2|0). \end{aligned}$$

Similarly, we also have

$$f_{Y_2|H}(\lambda^1_2|0) = f_{Y_2|H}(\lambda^0_2|1).$$

Further, from (4.22) and (4.23), we have

$$\frac{d\lambda_2^0}{dq_2} = \frac{d\lambda_2^0}{d\lambda_{1,[2]}}\frac{d\lambda_{1,[2]}}{dq_2} = \left[1 + \frac{\sigma_2^2\phi\left(\frac{\lambda_{1,[2]}}{\sigma_2}\right)}{1 - P^{\mathrm{I}}_{e,1,[2]}} - \frac{\sigma_2^2\phi\left(\frac{\lambda_{1,[2]}-1}{\sigma_2}\right)}{P^{\mathrm{II}}_{e,1,[2]}}\right]\frac{d\lambda_{1,[2]}}{dq_2},$$

$$\frac{d\lambda_2^1}{dq_2} = \frac{d\lambda_2^1}{d\lambda_{1,[2]}}\frac{d\lambda_{1,[2]}}{dq_2} = \left[1 - \frac{\sigma_2^2\phi\left(\frac{\lambda_{1,[2]}}{\sigma_2}\right)}{P^{\mathrm{I}}_{e,1,[2]}} + \frac{\sigma_2^2\phi\left(\frac{\lambda_{1,[2]}-1}{\sigma_2}\right)}{1 - P^{\mathrm{II}}_{e,1,[2]}}\right]\frac{d\lambda_{1,[2]}}{dq_2}.$$

When $\lambda_{1,[2]} = \frac{1}{2}$, $P^{\mathrm{I}}_{e,1,[2]} = P^{\mathrm{II}}_{e,1,[2]} = Q\left(\frac{1}{2\sigma_2}\right)$, and $\phi\left(\frac{\lambda_{1,[2]}}{\sigma_2}\right) = \phi\left(\frac{\lambda_{1,[2]}-1}{\sigma_2}\right)$. Thus, $\frac{d\lambda_2^0}{dq_2} = \frac{d\lambda_2^1}{dq_2}$.

Using these, the values of prior for which $q_1^* = p_0, q_2^* = 1/2$ are given by

$$\frac{p_0}{1-p_0} = \frac{Q\left(\frac{-1}{2\sigma_2}\right)Q\left(\frac{-1}{2\sigma_1} - \sigma_1 x\right) + Q\left(\frac{1}{2\sigma_2}\right)Q\left(\frac{1}{2\sigma_1} + \sigma_1 x\right)}{Q\left(\frac{-1}{2\sigma_2}\right)Q\left(\frac{-1}{2\sigma_2} + \sigma_1 x\right) + Q\left(\frac{1}{2\sigma_2}\right)Q\left(\frac{1}{2\sigma_1} - \sigma_1 x\right)}. \tag{C.4}$$

Using the definitions of x, α, β in (C.4), and the fact that $Q(-y) = 1 - Q(y)$, the result follows.

C.3 Proof of Theorem 17

From (4.1), we note that the Bayes risk for social learning with beliefs (q_1, q_2) is

$$\begin{aligned} &R_2(q_1, q_2) \\ &\quad = c_{10}p_0\left[P^{\mathrm{I}_0}_{e,2}(1 - P^{\mathrm{I}}_{e,1}) + P^{\mathrm{I}_1}_{e,2}P^{\mathrm{I}}_{e,1}\right] + c_{01}(1-p_0)\left[P^{\mathrm{II}_0}_{e,2}P^{\mathrm{II}}_{e,1} + P^{\mathrm{II}_1}_{e,2}(1 - P^{\mathrm{II}}_{e,1})\right].\end{aligned}$$

Then, the difference in Bayes risk between the two choices of advisors is given by

$$\begin{aligned}\Delta R_2 &= R_2(q_1, q_2) - R_2(q_{1'}, q_2) \\ &= c_{10}p_0(P^{\mathrm{I}}_{e,1} - P^{\mathrm{I}}_{e,1'})(P^{\mathrm{I}_1}_{e,2} - P^{\mathrm{I}_0}_{e,2}) + c_{01}(1-p_0)(P^{\mathrm{II}}_{e,1} - P^{\mathrm{II}}_{e,1'})(P^{\mathrm{II}_0}_{e,2} - P^{\mathrm{II}_1}_{e,2}).\end{aligned} \tag{C.5}$$

Since $q_1 < q_{1'}$, the decision thresholds satisfy $\lambda_1 < \lambda_{1'}$. Thus, from (C.5) and independence of Y_1, Y_2 given H, we see that $\Delta R_2 \leq 0$ if and only if (4.31) holds.

Appendix D

Proofs for Chapter 5

D.1 Proof of Lemma 10

Let us introduce notations first. Let W be a Gaussian random variable with distribution $\mathcal{N}\left(\mathsf{med}(V), \frac{1}{4Lf^2(\mathsf{med}(V))}\right)$. Let $\xi_m := \mathbb{E}[V_{(m+1)}]$ and note that $\xi_m \neq \mathsf{med}(V)$ in general since $V_{(m+1)}$ is a biased estimator in general. Also let γ_r, γ'_r be absolute central moments of $V_{(m+1)}$ and W, i.e.,

$$\begin{aligned}\gamma_r &:= \mathbb{E}[|V_{(m+1)} - \xi_m|^r],\\ \gamma'_r &:= \mathbb{E}[|W - \mathsf{med}(V)|^r],\end{aligned}$$

and ρ_r, ρ'_r be central moments of $V_{(m+1)}$ and W, i.e.,

$$\begin{aligned}\rho_r &:= \mathbb{E}[(V_{(m+1)} - \xi_m)^r],\\ \rho'_r &:= \mathbb{E}[(W - \mathsf{med}(V))^r].\end{aligned}$$

Then, our proof is based on the following result.

Proposition 14 ([122])**.** $\lim_{m\to\infty} \rho_r = \rho'_r$ *for all* $r \geq 2$*, and* $\lim_{m\to\infty} \xi_m = \mathsf{med}(V)$*.*

Due to the triangle inequality and Prop. 18,

$$\begin{aligned}&\mathbb{E}[|V_{(m+1)} - \mathsf{med}(V)|^r]\\ &= \mathbb{E}[|V_{(m+1)} - \xi_m + \xi_m - \mathsf{med}(V)|^r]\\ &\leq 2^r\mathbb{E}[|V_{(m+1)} - \xi_m|^r] + 2^r\mathbb{E}[|\xi_m - \mathsf{med}(V)|^r]\\ &= 2^r\mathbb{E}[|V_{(m+1)} - \xi_m|^r] + 2^r|\xi_m - \mathsf{med}(V)|^r,\end{aligned}$$

where the last equality follows since $\xi_m, \mathsf{med}(V)$ are deterministic quantities. Furthermore, due to Prop. 14, we can take large m for any positive δ such that $|\xi_m - \mathsf{med}(V)|^r \leq \delta$.

Consider the first term. Since $\rho_r \to \rho'_r$, we know that γ_r, γ'_r are bounded. Letting $A := V_{(m+1)} - \xi_m$ and $B := W - \mathsf{med}(V)$ for brevity, we can take large $p \in \mathbb{N}$ such that

$$\Big|\mathbb{E}[|A|^r] - \mathbb{E}[|A|^r \wedge p]\Big| \leq \delta \ \text{ and } \ \Big|\mathbb{E}[|B|^r] - \mathbb{E}[|B|^r \wedge p]\Big| \leq \delta.$$

Then,

$$\begin{aligned}\Big|\mathbb{E}[|A|^r] - \mathbb{E}[|B|^r]\Big| &= \Big|\mathbb{E}[|A|^r - |A|^r \wedge p + |A|^r \wedge p] - \mathbb{E}[|B|^r - |B|^r \wedge p + |B|^r \wedge p]\Big| \\ &\leq 2\delta + \Big|\mathbb{E}[|A|^r \wedge p] - \mathbb{E}[|B|^r \wedge p]\Big|.\end{aligned}$$

Note that $|\cdot|^r \wedge p$ is a bounded continuous function and $A \to B$ in distribution as $m \to \infty$ by Lem. 10. Therefore we can take large m such that $\Big|\mathbb{E}[|A|^r \wedge p] - \mathbb{E}[|B|^r \wedge p]\Big| \leq \delta$ by the continuous mapping theorem, which leads us to

$$\Big|\mathbb{E}[|A|^r] - \mathbb{E}[|B|^r]\Big| \leq 3\delta.$$

Hence, we have

$$\mathbb{E}[|V_{(m+1)} - \mathsf{med}(V)|^r] \leq 2^r \mathbb{E}[|B|^r] + 2^r 3\delta + 2^r \delta.$$

Note that $\mathbb{E}[|B|^r]$ is the rth absolute central moment of Gaussian, so

$$\mathbb{E}[|B|^r] = \left(\frac{1}{2Lf^2(\mathsf{med}(V))}\right)^{r/2} \frac{\Gamma(\frac{r+1}{2})}{\sqrt{\pi}}.$$

Since δ is arbitrary, the proof is completed.

D.2 Distortion Bounds in Achievability

The next proposition is a part of the proof of the regular model achievability.

Proposition 15 (Regular CEO Problem)**.**

$$(2K)^r \mathbb{E}\left[|U_{(m+1)}(t) - \widehat{U}_{(m+1)}(t)|^r\right] \leq \epsilon.$$

Proof. For the sake of notational brevity, we omit '(t)' so

$$\begin{aligned}
&(2K)^r \mathbb{E}\left[|U_{(m+1)}(t) - \widehat{U}_{(m+1)}(t)|^r\right] \qquad\qquad (\text{D.1})\\
&= (2K)^r \mathbb{E}\left[|U_{(m+1)} - \widehat{U}_{(m+1)}|^r\right]\\
&\le 2^{2r} K^r \mathbb{E}\left[|U_{(m+1)} - q(U_{(m+1)})|^r\right] + 2^{2r} K^r \mathbb{E}\left[|q(U_{(m+1)}) - \widehat{U}_{(m+1)}|^r\right]\\
&\le 2^{2r} K^r \mathbb{E}\left[|U_{(m+1)} - q(U_{(m+1)})|^r\right] + 2^{3r} K^r \mathbb{E}\left[|q(U_{(m+1)}) - \mathsf{med}(\{\widetilde{U}_i\}_{i=1}^L)|^r\right]\\
&\quad + 2^{3r} K^r \mathbb{E}\left[|\mathsf{med}(\{\widetilde{U}_i\}_{i=1}^L) - \widehat{U}_{(m+1)}|^r\right],
\end{aligned}$$

where both inequalities are due to the triangle inequality and Prop. 18.

Now the first and the second terms are small enough because of (5.1), i.e.,

$$2^{2r} K^r \mathbb{E}\left[|U_{(m+1)} - q(U_{(m+1)})|^r\right] \le 2^{2r} K^r \delta_0^r,$$
$$2^{3r} K^r \mathbb{E}\left[|q(U_{(m+1)}) - \mathsf{med}(\{\widetilde{U}_i\}_{i=1}^L)|^r\right] \le 2^{3r} K^r \delta_0^r.$$

The last term is positive only when there is a Slepian-Wolf decoding error $\mathcal{B}$ defined in Prop. 13, so

$$2^{3r} K^r \mathbb{E}\left[|\mathsf{med}(\{\widetilde{U}_i\}_{i=1}^L) - \widehat{U}_{(m+1)}|^r\right] \le 2^{3r} K^r (2\widetilde{u}_{\mathsf{max}})^r \mathbb{P}[\mathcal{B}] \le 2^{3r} K^r (2\widetilde{u}_{\mathsf{max}})^r \lambda,$$

and $\lambda \to 0$ as $n \to \infty$. Therefore (D.1) can be bounded by ϵ if we choose small δ and large n appropriately. □

The next proposition is a part of the proof of the non-regular model achievability.

Proposition 16 (Non-regular CEO Problem)**.** *For any $\epsilon > 0$, there exist a quantization scheme and block length n such that*

$$K^r \mathbb{E}\left[\left|U_{(1)} + U_{(L)} - \widehat{U}_{(1)} - \widehat{U}_{(L)}\right|^r\right] \le \epsilon.$$

Proof. Using the triangle inequality and Prop. 18, we have

$$\begin{aligned}
&K^r \mathbb{E}\left[\left|U_{(1)} + U_{(L)} - \widehat{U}_{(1)} - \widehat{U}_{(L)}\right|^r\right]\\
&\le (2K)^r \mathbb{E}\left[\left|U_{(1)} + U_{(L)} - \widetilde{U}_{(1)} - \widetilde{U}_{(L)}\right|^r\right] + (2K)^r \mathbb{E}\left[\left|\widetilde{U}_{(1)} + \widetilde{U}_{(L)} - \widehat{U}_{(1)} - \widehat{U}_{(L)}\right|^r\right]. \qquad (\text{D.2})
\end{aligned}$$

The first term is decomposed into two terms by Prop. 18 and we take sufficiently fine quantization points (5.4),

$$
\begin{aligned}
&(2K)^r\mathbb{E}\left[\left|U_{(1)}+U_{(L)}-\widetilde{U}_{(1)}-\widetilde{U}_{(L)}\right|^r\right]\\
&\le 2^{2r}K^r\mathbb{E}\left[\left|U_{(1)}+U_{(L)}\right|^r\right]+2^{2r}K^r\mathbb{E}\left[\left|\widetilde{U}_{(1)}-\widetilde{U}_{(L)}\right|^r\right]\\
&\le 2^{2r+1}K^r\delta_0^r.
\end{aligned}
$$

Next, to bound the second term recall the decoding error probability $\mathbb{P}[\mathcal{B}]\le\lambda$ given in Prop. 13. Then,

$$
\begin{aligned}
&(2K)^r\mathbb{E}\left[\left|\widetilde{U}_{(1)}+\widetilde{U}_{(L)}-\widehat{U}_{(1)}+\widehat{U}_{(L)}\right|^r\right]\\
&\le 2^{2r}K^r\left(\mathbb{E}[|\widetilde{U}_{(1)}-\widehat{U}_{(1)}|^r]+\mathbb{E}[|\widetilde{U}_{(L)}-\widehat{U}_{(L)}|^r]\right)\\
&\le 2^{2r}K^r2(2\widetilde{u}_{\mathsf{max}})^r\mathbb{P}[\mathcal{B}]\le 2^{2r}K^r2(2\widetilde{u}_{\mathsf{max}})^r\lambda,
\end{aligned}
$$

where $\widetilde{u}_{\mathsf{max}}:=\max\{|\widetilde{u}|:\widetilde{u}\in\mathcal{U}\}<1$ as $\mathcal{U}=[0,1]$.

Hence, taking sufficiently fine quantization and taking sufficiently large n, we can bound (D.2) for any $\epsilon>0$. □

D.3 Proof of Lemma 15

Let us start with the following identity for a non-negative random variable Z:

$$
\mathbb{E}[Z]=\int_0^\infty\mathbb{P}[Z\ge t]dt.
$$

Letting $Z=|\widehat{X}-X|^r$ and $t=\left(\frac{h}{2}\right)^r$, we have the following identity by change of variable.

$$
\begin{aligned}
\mathbb{E}[|X-\widehat{X}|^r]&=\int_0^\infty\mathbb{P}[|X-\widehat{X}|^r\ge t]dt\\
&=\int_0^\infty r2^{-r}h^{r-1}\mathbb{P}\left[|X-\widehat{X}|^r\ge\left(\frac{h}{2}\right)^r\right]dh\\
&=\int_0^\infty r2^{-r}h^{r-1}\mathbb{P}\left[|X-\widehat{X}|\ge\frac{h}{2}\right]dh.
\end{aligned}
$$

Let us derive a lower bound of $\mathbb{P}\left[|\widehat{X}-X| \geq \frac{h}{2}\right]$.

$$\begin{aligned}
&\mathbb{P}\left[|\widehat{X}-X| \geq \frac{h}{2}\right] \\
&= \mathbb{P}\left[\widehat{X}-X \geq \frac{h}{2}\right] + \mathbb{P}\left[\widehat{X}-X < -\frac{h}{2}\right] \\
&= \int_0^1 f_X(x)\mathbb{P}\left[\widehat{X}-X \geq \frac{h}{2}\Big|X=x\right]dx + \int_0^1 f_X(x)\mathbb{P}\left[\widehat{X}-X < -\frac{h}{2}\Big|X=x\right]dx.
\end{aligned}$$

By change of variable $x = t+h$ in the second integration, we have

$$\begin{aligned}
&\mathbb{P}\left[|\widehat{X}-X| \geq \frac{h}{2}\right] \\
&= \int_0^1 f_X(x)\mathbb{P}\left[\widehat{X}-X \geq \frac{h}{2}\Big|X=x\right]dx + \int_{-h}^{1-h} f_X(x)\mathbb{P}\left[\widehat{X}-X < -\frac{h}{2}\Big|X=t+h\right]dt \\
&= \int_0^1 f_X(x)\mathbb{P}\left[\widehat{X}-x \geq \frac{h}{2}\Big|X=x\right]dx + \int_{-h}^{1-h} f_X(t+h)\mathbb{P}\left[\widehat{X}-t < \frac{h}{2}\Big|X=t+h\right]dt \\
&\geq \int_0^{1-h} f_X(x)\mathbb{P}\left[\widehat{X}-x \geq \frac{h}{2}\Big|X=x\right]dx + \int_{0}^{1-h} f_X(t+h)\mathbb{P}\left[\widehat{X}-t < \frac{h}{2}\Big|X=t+h\right]dt \\
&= \int_0^{1-h} f_X(x)\mathbb{P}\left[\widehat{X}-x \geq \frac{h}{2}\Big|X=x\right] + f_X(x+h)\mathbb{P}\left[\widehat{X}-x < \frac{h}{2}\Big|X=x+h\right]dx \\
&= \int_0^{1-h} (f_X(x)+f_X(x+h))\Bigg\{\frac{f_X(x)}{f_X(x)+f_X(x+h)}\mathbb{P}\left[\widehat{X}-x \geq \frac{h}{2}\Big|X=x\right] \\
&\qquad\qquad + \frac{f_X(x+h)}{f_X(x)+f_X(x+h)}\mathbb{P}\left[\widehat{X}-x < \frac{h}{2}\Big|X=x+h\right]\Bigg\}dx.
\end{aligned}$$

So the quantity in the curly bracket implies the error probability of a decision rule

$$\widehat{X}-x \underset{X=x}{\overset{X=x+h}{\gtrless}} \frac{h}{2}$$

and then it is further bounded by the optimal error probability $P_{\mathsf{min}}(x, x+h)$. Then, we have the final lower bound as follows:

$$\begin{aligned}
\mathbb{P}\left[|\widehat{X}-X| \geq \frac{h}{2}\right] &\geq \int_0^{1-h} (f_X(x)+f_X(x+h))P_{\mathsf{min}}(x,x+h)dx \\
\implies \mathbb{E}[|X-\widehat{X}|^r] &\geq \int_0^{\infty} r2^{-r}h^{r-1}\int_0^{1-h}\frac{f_X(x)+f_X(x+h)}{2}P_{\mathsf{min}}(x,x+h)dxdh \\
&= \int_0^{1} r2^{-r}h^{r-1}\int_0^{1-h}\frac{f_X(x)+f_X(x+h)}{2}P_{\mathsf{min}}(x,x+h)dxdh.
\end{aligned}$$

The proof is completed.

D.4 Inequalities

Proposition 17. *Suppose $a_i, b_i > 0$ for all $i \in [1:n]$. Then,*

$$\frac{\sum_{i=1}^n a_i}{\sum_{i=1}^n b_i} \geq \min_{i\in[1:n]}\left(\frac{a_i}{b_i}\right).$$

Proof. Let $m := \min_i\left(\frac{a_i}{b_i}\right)$. Then,

$$\begin{aligned}
& a_i \geq mb_i \quad \forall i \in [1:n], \\
\implies & \sum_{i=1}^n a_i \geq m\sum_{i=1}^n b_i, \\
\implies & \frac{\sum_{i=1}^n a_i}{\sum_{i=1}^n b_i} \geq m = \min_{i\in[1:n]}\left(\frac{a_i}{b_i}\right).
\end{aligned}$$

□

Proposition 18. *For $a, b \geq 0$ and $r \in \mathbb{N}$,*

$$(a+b)^r \leq 2^r(a^r+b^r).$$

Proof. By the binomial expansion theorem,

$$\begin{aligned}(a+b)^r &= \sum_{i=0}^{r} \binom{r}{i} a^i b^{r-i} \\ &\leq \sum_{i=0}^{r} \binom{r}{i} \left(\max(a,b)\right)^r \\ &= 2^r \left(\max(a,b)\right)^r \\ &\leq 2^r \left(a^r + b^r\right).\end{aligned}$$

□

References

[1] C. E. Shannon, “A mathematical theory of communication,” *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, 623–656, July/Oct. 1948.

[2] R. R. Kline, *The Cybernetics Moment: Or Why we Call Our Age the Information Age.* Baltimore: John Hopkins University Press, 2015.

[3] C. E. Shannon, “The bandwagon,” *IRE Trans. Inf. Theory*, vol. IT-2, no. 1, p. 3, 1956.

[4] D. Slepian, “Information theory in the fifties,” *IEEE Trans. Inf. Theory*, vol. IT-19, no. 2, pp. 145–148, Mar. 1973.

[5] R. Kaplan and D. Saccuzzo, *Psychological Testing: Principles, Applications, and Issues.* Belmont, CA: Wadsworth, 2009.

[6] B. Schwartz, “Queues, priorities, and social process,” *Soc. Psychol.*, vol. 41, no. 1, pp. 3–12, Mar. 1978.

[7] M. Jamal, “Job stress and job performance controversy revisited: An empirical examination in two countries,” *Int. J. Stress Management*, vol. 14, no. 2, pp. 175–187, May 2007.

[8] R. W. Derlet and J. R. Richards, “Overcrowding in the nation’s emergency departments: Complex causes and disturbing effects,” *Ann. Emerg. Med.*, vol. 35, no. 1, pp. 63–68, Jan. 2000.

[9] D. C. Dugdale, R. Epstein, and S. Z. Pantilat, “Time and the patient-physician relationship,” *J. Gen. Intern. Med.*, vol. 14, no. S1, pp. S34–S40, Jan. 1999.

[10] S. Branson, G. Van Horn, C. Wah, P. Perona, and S. Belongie, “The ignorant led by the blind: A hybrid human-machine vision system for fine-grained categorization,” *Int. J. Comput. Vis.*, vol. 108, no. 1-2, pp. 3–29, May 2014.

[11] M. Borokhovich, A. Chatterjee, J. Rogers, L. R. Varshney, and S. Vishwanath, “Improving impact sourcing via efficient global service delivery,” in *Proc. Data for Good Exchange (D4GX)*, Sep. 2015.

[12] A. Vempaty, L. R. Varshney, and P. K. Varshney, "Reliable crowdsourcing for multi-class labeling using coding theory," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 4, pp. 667–679, Aug. 2014.

[13] K. Sriram and D. M. Lucantoni, "Traffic smoothing effects of bit dropping in a packet voice multiplexer," *IEEE Trans. Commun.*, vol. 37, no. 7, pp. 703–712, Jul. 1989.

[14] S. C. Draper, M. D. Trott, and G. W. Wornell, "A universal approach to queuing with distortion control," *IEEE Trans. Autom. Control*, vol. 50, no. 4, pp. 532–537, Apr. 2005.

[15] V. K. Goyal, "Multiple description coding: Compression meets the network," *IEEE Signal Process. Mag.*, vol. 18, no. 5, pp. 74–93, Sep. 2001.

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

[17] A. S. Bedekar and M. Azizoğlu, "The information-theoretic capacity of discrete-time queues," *IEEE Trans. Inf. Theory*, vol. 44, no. 2, pp. 446–461, Mar. 1998.

[18] B. Prabhakar and R. Gallager, "Entropy and the timing capacity of discrete queues," *IEEE Trans. Inf. Theory*, vol. 49, no. 2, pp. 357–370, Feb. 2003.

[19] R. Sundaresan and S. Verdú, "Sequential decoding for the exponential server timing channel," *IEEE Trans. Inf. Theory*, vol. 46, no. 2, pp. 705–709, Mar. 2000.

[20] M. Costa, M. Codreanu, and A. Ephremides, "On the age of information in status update systems with packet management," *IEEE Trans. Inf. Theory*, vol. 62, no. 4, pp. 1897–1910, Apr. 2016.

[21] İ. E. Telatar, "Multi-access communications with decision feedback decoding," Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA, May 1992.

[22] İ. E. Telatar and R. G. Gallager, "Combining queueing theory with information theory for multiaccess," *IEEE J. Sel. Areas Commun.*, vol. 13, no. 6, pp. 963–969, Aug. 1995.

[23] S. Raj, E. Telatar, and D. Tse, "Job scheduling and multiple access," in *Advances in Network Information Theory*, P. Gupta, G. Kramer, and A. J. van Wijngaarden, Eds. Providence: DIMACS, American Mathematical Society, 2004, pp. 127–137.

[24] S. Musy and E. Telatar, "On the transmission of bursty sources," in *Proc. 2006 IEEE Int. Symp. Inf. Theory*, Jul. 2006, pp. 2899–2903.

[25] N. Michelusi, J. Boedicker, M. Y. El-Naggar, and U. Mitra, "Queuing models for abstracting interactions in bacterial communities," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 3, pp. 584–599, Mar. 2016.

[26] G. Ellison and D. Fudenberg, "Rules of thumb for social learning," *J. Polit. Econ.*, vol. 101, no. 4, pp. 612–643, Aug. 1993.

[27] V. Krishnamurthy and H. V. Poor, "Social learning and Bayesian games in multiagent signal processing: How do local and global decision makers interact?" *IEEE Signal Process. Mag.*, vol. 30, no. 3, pp. 43–57, May 2013.

[28] A. V. Banerjee, "A simple model of herd behavior," *Quart. J. Econ.*, vol. 107, no. 3, pp. 797–817, Aug. 1992.

[29] S. Bikhchandani, D. Hirshleifer, and I. Welch, "Learning from the behavior of others: Conformity, fads, and informational cascades," *J. Econ. Perspect.*, vol. 12, no. 3, pp. 151–170, 1998.

[30] V. Bala and S. Goyal, "Conformism and diversity under social learning," *Econ. Theor.*, vol. 17, no. 1, pp. 101–120, Jan. 2001.

[31] L. Smith and P. Sørensen, "Pathological outcomes of observational learning," *Econometrica*, vol. 68, no. 2, pp. 371–398, Mar. 2000.

[32] B. Çelen and S. Kariv, "Observational learning under imperfect information," *Games Econ. Behav.*, vol. 47, no. 1, pp. 72–86, Apr. 2004.

[33] D. Gale and S. Kariv, "Bayesian learning in social networks," *Games Econ. Behav.*, vol. 45, no. 2, pp. 329–346, Nov. 2003.

[34] D. Acemoglu, M. A. Dahleh, I. Lobel, and A. Ozdaglar, "Bayesian learning in social networks," *Rev. Econ. Stud.*, vol. 78, no. 4, pp. 1201–1236, Oct. 2011.

[35] J. B. Rhim, L. R. Varshney, and V. K. Goyal, "Quantization of prior probabilities for collaborative distributed hypothesis testing," *IEEE Trans. Signal Process.*, vol. 60, no. 9, pp. 4537–4550, Sep. 2012.

[36] J. B. Rhim and V. K. Goyal, "Distributed hypothesis testing with social learning and symmetric fusion," *IEEE Trans. Signal Process.*, vol. 62, no. 23, Dec. 2014.

[37] V. V. Veeravalli, T. Başar, and H. V. Poor, "Decentralized sequential detection with a fusion center performing the sequential test," *IEEE Trans. Inf. Theory*, vol. 39, no. 2, pp. 433–442, Mar. 1993.

[38] R. Viswanathan and P. K. Varshney, "Distributed detection with multiple sensors: Part I—fundamentals," *Proc. IEEE*, vol. 85, no. 1, pp. 54–63, Jan. 1997.

[39] T. Berger, Z. Zhang, and H. Viswanathan, "The CEO problem," *IEEE Trans. Inf. Theory*, vol. 42, no. 3, pp. 887–902, May 1996.

[40] V. Saligrama, M. Alanyali, and O. Savas, “Distributed detection in sensor networks with packet losses and finite capacity links,” *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4118–4132, Nov. 2006.

[41] S. Kar and J. M. F. Moura, “Distributed consensus algorithms in sensor networks with imperfect communication: Link failures and channel noise,” *IEEE Trans. Signal Process.*, vol. 57, pp. 355–369, Jan. 2009.

[42] ——, “Distributed consensus algorithms in sensor networks: Quantized data and random link failures,” *IEEE Trans. Signal Process.*, vol. 58, pp. 1383–1400, Mar. 2010.

[43] M. Alanyali, S. Venkatesh, O. Savas, and S. Aeron, “Distributed Bayesian hypothesis testing in sensor networks,” in *Proc. Am. Contr. Conf. (ACC 2004)*, vol. 6, June-July 2004, pp. 5369–5374.

[44] K. R. Rad and A. Tahbaz-Salehi, “Distributed parameter estimation in networks,” in *Proc. 49th IEEE Conf. Decision Control*, Dec. 2010, pp. 5050–5055.

[45] A. Jadbabaie, P. Molavi, A. Sandroni, and A. Tahbaz-Salehi, “Non-Bayesian social learning,” *Games Econ. Behav.*, vol. 76, no. 1, pp. 210–225, Sep. 2012.

[46] S. Shahrampour, A. Rakhlin, and A. Jadbabaie, “Distributed detection: Finite-time analysis and impact of network topology,” *IEEE Trans. Autom. Control*, vol. 61, no. 11, pp. 3256–3268, Nov. 2016.

[47] A. Nedić, A. Olshevsky, and C. A. Uribe, “Nonasymptotic convergence rates for cooperative learning over time-varying directed graphs,” in *Proc. Am. Contr. Conf. (ACC 2015)*, Jul. 2015, pp. 5884–5889.

[48] A. K. Sahu and S. Kar, “Distributed sequential detection for gaussian shift-in-mean hypothesis testing,” *IEEE Trans. Signal Process.*, vol. 64, no. 1, pp. 89–103, Jan. 2016.

[49] A. Lalitha, T. Javidi, and A. D. Sarwate, “Social learning and distributed hypothesis testing,” *IEEE Trans. Inf. Theory*, vol. 64, no. 9, pp. 6161–6179, Sep. 2018.

[50] J. B. Rhim and V. K. Goyal, “Social teaching: Being informative vs. being right in sequential decision making,” in *Proc. 2013 IEEE Int. Symp. Inf. Theory*, Jul. 2013, pp. 2602–2606.

[51] P. K. Varshney, *Distributed Detection and Data Fusion.* New York: Springer-Verlag, 1997.

[52] A. Tversky and D. Kahneman, “Advances in prospect theory: Cumulative representation of uncertainty,” *J. Risk Uncertainty*, vol. 5, no. 4, pp. 297–323, Oct. 1992.

[53] V. S. S. Nadendla, S. Brahma, and P. K. Varshney, “Towards the design of prospect-theory based human decision rules for hypothesis testing,” in *Proc. 54th Annu. Allerton Conf. Commun. Control Comput.*, Sep. 2016, pp. 766–773.

[54] V. S. S. Nadendla, E. Akyol, C. Langbort, and T. Başar, "Strategic communication between prospect theoretic agents over a Gaussian test channel," in *Proc. Mil. Commun. Conf. (MILCOM 2017)*, Oct. 2017, pp. 109–114.

[55] D. Prelec, "The probability weighting function," *Econometrica*, vol. 66, no. 3, pp. 497–527, May 1998.

[56] A. McAfee and E. Brynjolfsson, *Machine, Platform, Crowd: Harnessing Our Digital Future.* WW Norton & Company, 2017.

[57] H. Viswanathan and T. Berger, "The quadratic Gaussian CEO problem," *IEEE Trans. Inf. Theory*, vol. 43, no. 5, pp. 1549–1559, Sep. 1997.

[58] Y. Oohama, "The rate-distortion function for the quadratic Gaussian CEO problem," *IEEE Trans. Inf. Theory*, vol. 44, no. 3, pp. 1057–1070, May 1998.

[59] ——, "Rate-distortion theory for Gaussian multiterminal source coding systems with several side informations at the decoder," *IEEE Trans. Inf. Theory*, vol. 51, no. 7, pp. 2577–2593, Jul. 2005.

[60] V. Prabhakaran, D. Tse, and K. Ramachandran, "Rate region of the quadratic Gaussian CEO problem," in *Proc. 2004 IEEE Int. Symp. Inf. Theory*, June-July 2004, p. 117.

[61] T. A. Courtade and T. Weissman, "Multiterminal source coding under logarithmic loss," *IEEE Trans. Inf. Theory*, vol. 60, no. 1, pp. 740–761, Jan. 2014.

[62] D. Seo and L. R. Varshney, "Information-theoretic limits of algorithmic noise tolerance," in *IEEE Int. Conf. Reboot. Comput. (ICRC)*, Nov. 2016, pp. 1–4.

[63] A. Vempaty and L. R. Varshney, "The non-regular CEO problem," *IEEE Trans. Inf. Theory*, vol. 61, no. 5, pp. 2764–2775, May 2015.

[64] K. Eswaran and M. Gastpar, "Remote source coding under Gaussian noise: Dueling roles of power and entropy power," arXiv:1805.06515 [cs.IT]., May 2018.

[65] H. Asnani, I. Shomorony, A. S. Avestimehr, and T. Weissman, "Network compression: Worst case analysis," *IEEE Trans. Inf. Theory*, vol. 61, no. 7, pp. 3980–3995, Jul. 2015.

[66] R. Zamir and T. Berger, "Multiterminal source coding with high resolution," *IEEE Trans. Inf. Theory*, vol. 45, no. 1, pp. 106–117, Jan. 1999.

[67] M. S. Pinsker, *Information and Information Stability of Random Variables and Processes.* San Francisco: Holden-Day, 1964.

[68] S. Verdú and T. S. Han, "A general formula for channel capacity," *IEEE Trans. Inf. Theory*, vol. 40, no. 4, pp. 1147–1157, Jul. 1994.

[69] T. S. Han, *Information-Spectrum Methods in Information Theory.* Berlin: Springer, 2003.

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

[71] G. Caire and S. Shamai (Shitz), "On the capacity of some channels with channel state information," *IEEE Trans. Inf. Theory*, vol. 45, no. 6, pp. 2007–2019, Sep. 1999.

[72] L. Kleinrock, *Queuing Systems, Volume I: Theory.* John Wiley & Sons, Inc., 1975.

[73] A. Chatterjee, L. R. Varshney, and S. Vishwananth, "Work capacity of freelance markets: Fundamental limits and decentralized schemes," in *Proc. 2015 IEEE INFOCOM*, Apr. 2015, pp. 1769–1777.

[74] S. Higginbotham, "Autonomous trucks need people," *IEEE Spectr.*, vol. 56, no. 3, p. 21, Mar. 2019.

[75] P. Franken, D. König, U. Arndt, and V. Schmidt, *Queues and Point Processes.* New York: John Wiley & Sons, 1982.

[76] D. J. Daley and D. Vere-Jones, *An Introduction to the Theory of Point Processes.* Berlin, Germany: Springer-Verlag, 1998.

[77] O. Kallenberg, *Random Measures, Theory and Applications.* Cham, Switzerland: Springer, 2017.

[78] A. Chatterjee, D. Seo, and L. R. Varshney, "Capacity of systems with queue-length dependent service quality," *IEEE Trans. Inf. Theory*, vol. 63, no. 6, pp. 3950–3963, Jun. 2017.

[79] S. Asmussen, *Applied Probability and Queues, 2nd ed.* New York, USA: Springer-Verlag, 2003.

[80] S. M. Samuels, "A characterization of the Poisson process," *J. Appl. Probab.*, no. 1, pp. 72–85, Mar. 1974.

[81] R. Radner, "Team decision problems," *Ann. Math. Stat.*, vol. 33, no. 3, pp. 857–881, Sep. 1962.

[82] A. al-Nowaihi and S. Dhami, "Probability weighting functions," *Wiley Encyclopedia of Operations Research and Management Science*, Feb. 2011.

[83] J. P. Mills, "Table of the ratio: area to bounding ordinate, for any portion of normal curve," *Biometrika*, no. 3/4, pp. 395–400, Nov. 1926.

[84] M. R. Sampford, "Some inequalities on Mill's ratio and related functions," *Ann. Math. Stat.*, vol. 24, no. 1, pp. 130–132, Mar. 1953.

[85] J. A. Swets, W. P. Tanner, Jr., and T. G. Birdsall, “Decision processes in perception,” *Psychol. Rev.*, vol. 68, no. 5, pp. 301–340, Sep. 1961.

[86] W. K. Viscusi, “Are individuals Bayesian decision makers?” *Am. Econ. Rev.*, vol. 75, no. 2, pp. 381–385, May 1985.

[87] G. L. Brase, L. Cosmides, and J. Tooby, “Individuation, counting, and statistical inference: The role of frequency and whole-object representations in judgment under uncertainty,” *J. Exp. Psychol. Gen.*, vol. 127, no. 1, pp. 3–21, Mar. 1998.

[88] M. Glanzer, A. Hilford, and L. T. Maloney, “Likelihood ratio decisions in memory: Three implied regularities,” *Psychon. Bull. Rev.*, vol. 16, no. 3, pp. 431–455, Jun. 2009.

[89] D. C. Knill and W. Richards, *Perception as Bayesian Inference.* Cambridge: Cambridge University Press, 1996.

[90] W. A. Yost, A. N. Popper, and R. R. Fay, *Human Psychophysics.* New York, USA: Springer-Verlag, 1993.

[91] M. H. Jarrahi, “Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making,” *Business Horizons*, vol. 61, no. 4, pp. 577 – 586, 2018.

[92] MIT IDE, “Where humans meet machines: Intuition, expertise and learning,” https://medium.com/mit-initiative-on-the-digital-economy/where-humans-meet-machines-intuition-expertise-and-learning-be639f00bade, May 2018.

[93] S.-Y. Tung, “Multiterminal source coding,” Ph.D. dissertation, Cornell University, Ithaca, NY, May 1978.

[94] Y. Yamada, S. Tazakia, and R. M. Gray, “Asymptotic performance of block quantizers with difference distortion measures,” *IEEE Trans. Inf. Theory*, vol. IT-26, no. 1, pp. 6–14, Jan. 1980.

[95] P. R. Rider, “The midrange of a sample as an estimator of the population midrange,” *J. Am. Stat. Assoc.*, vol. 52, no. 280, pp. 537–542, Dec. 1957.

[96] D. Chazan, M. Zakai, and J. Ziv, “Improved lower bounds on signal parameter estimation,” *IEEE Trans. Inf. Theory*, vol. IT-21, no. 1, pp. 90–93, Jan. 1975.

[97] K. L. Bell, “Performance bounds in parameter estimation with application to bearing estimation,” Ph.D. thesis, George Mason University, Fairfax, VA, 1995.

[98] E. L. Lehmann and G. Casella, *Theory of Point Estimation, 2nd ed.* New York, USA: Springer, 2006.

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

[100] R. L. Dobrushin, "A general formulation of the fundamental theorem of Shannon in the theory of information," *Uspekhi Mat. Nauk*, vol. 14, no. 6, pp. 3–104, 1959.

[101] H. A. David and H. N. Nagaraja, *Order Statistics*, 3rd ed. Hoboken, NJ: Wiley-Interscience, 2003.

[102] C. E. Shannon, "Coding theorems for a discrete source with a fidelity criterion," in *IRE Nat. Conv. Rec., Part 4*, Mar. 1959, pp. 142–163.

[103] T. Linder and R. Zamir, "On the asymptotic tightness of the Shannon lower bound," *IEEE Trans. Inf. Theory*, vol. 40, no. 6, pp. 2026–2031, Nov. 1994.

[104] T. Koch, "The Shannon lower bound is asymptotically tight," *IEEE Trans. Inf. Theory*, vol. 62, no. 11, pp. 6155–6161, Nov. 2016.

[105] B. S. Clarke and A. R. Barron, "Information-theoretic asymptotics of Bayes methods," *IEEE Trans. Inf. Theory*, vol. 36, no. 3, pp. 453–471, May 1990.

[106] ——, "Jeffreys' prior is asymptotically least favorable under entropy risk," *J. Stat. Planning Inference*, vol. 41, no. 1, pp. 37–60, Aug. 1994.

[107] L. Wasserman, *All of Statistics: A Concise Course in Statistical Inference.* New York, USA: Springer, 2013.

[108] H. L. Van Trees, *Detection, Estimation, and Modulation Theory.* John Wiley & Sons, 1968.

[109] R. B. Nelsen, *An Introduction to Copulas.* New York: Springer, 2006.

[110] J. Neyman and E. S. Pearson, "On the use and interpretation of certain test criteria for purposes of statistical inference: Part I," *Biometrika*, vol. 20A, no. 1/2, pp. 175–240, Jul. 1928.

[111] G. R. Arce and S. A. Fontana, "On the midrange estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 36, no. 6, pp. 920–922, Jun. 1988.

[112] H. Akçay, H. Hjalmarsson, and L. Ljung, "On the choice of norms in system identification," *IEEE Trans. Autom. Control*, vol. 41, no. 9, pp. 1367–1372, Sep. 1996.

[113] K. L. Bell, Y. Steinberg, Y. Ephraim, and H. L. Van Trees, "Extended Ziv-Zakai lower bound for vector parameter estimation," *IEEE Trans. Inf. Theory*, vol. 43, no. 2, pp. 624–637, Mar. 1997.

[114] A. No and T. Weissman, "Universality of logarithmic loss in lossy compression," in *Proc. 2015 IEEE Int. Symp. Inf. Theory*, Jun. 2015, pp. 2166–2170.

[115] A. W. van der Vaart, *Asymptotic Statistics.* Cambridge: Cambridge University Press, 1998.

[116] A. DasGupta, *Asymptotic Theory of Statistics and Probability.* New York, USA: Springer, 2008.

[117] K. Jagannathan, A. Chatterjee, and P. Mandayam, "Qubits through queues: The capacity of channels with waiting time dependent errors," in *Proc. 25th National Conf. Commun. (NCC'19)*, Feb. 2019.

[118] J. Dai and S. Meyn, "Stability and convergence of moments for multiclass queueing networks via fluid limit models," *IEEE Trans. Autom. Control*, vol. 40, no. 11, pp. 1889–1904, Nov. 1995.

[119] M. H. A. Davis, "Piecewise-deterministic Markov processes: A general class of non-diffusion stochastic models," *J. Roy. Stat. Soc. Ser. B*, vol. 46, no. 3, pp. 353–388, 1984.

[120] R.-D. Reiss, *A Course on Point Processes.* New York, USA: Springer–Verlag, 1993.

[121] H. V. Poor, *An Introduction to Signal Detection and Estimation.* Springer Science & Business Media, 1988.

[122] J. T. Chu and H. Hotelling, "The moments of the sample median," *Ann. Math. Stat.*, vol. 26, no. 4, pp. 593–606, 1955.