UNIVERSITY OF CALIFORNIA SAN DIEGO

Belief Refinement Approaches to Communication and Inference Problems

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Electrical Engineering (Communication Theory and Systems)

by

Anusha Lalitha

Committee in charge:

Professor Tara Javidi, Chair
Professor Sanjoy Dasgupta
Professor Massimo Franceschetti
Professor Young-Han Kim
Professor Piya Pal

2019

The dissertation of Anusha Lalitha is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____

Chair

University of California San Diego

2019

DEDICATION

To my lovely parents

Padmaja Betanabhotla and Subramanya Prasad Betanabhotla

whose support, encouragement, and endless love made this dissertation possible.

EPIGRAPH

को अद्धा वेद क इह प्र वोचत्कुत आजाता कुत इयं विसृष्टिः ।
अर्वाग्देवा अस्य विसर्जनेनाथा को वेद यत आबभूव ॥

इयं विसृष्टिर्यत आबभूव यदि वा दधे यदि वा न ।
यो अस्याध्यक्षः परमे व्योमन्त्सो अङ्ग वेद यदि वा न वेद ॥

But, after all, who knows, and who can say
Whence it all came, and how creation happened?
the gods themselves are later than creation,
so who knows truly whence it has arisen?

Whence all creation had its origin,
the creator, whether he fashioned it or whether he did not,
the creator, who surveys it all from highest heaven,
he knows — or maybe even he does not know.

—Nasadiya Sukta, Rigveda 10.129-6,7

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisor, Professor Tara Javidi, for her mentorship and support. As an advisor, she constantly challenged me, and patiently helped me confront my weaknesses. I am deeply grateful to Tara for believing in me and for propelling and empowering me to reach far beyond what I thought I was ever capable of.

I wish to express my thankfulness and appreciation to my dissertation committee members, Professor Young-Han Kim, Professor Piya Pal, Professor Massimo Franceschetti and Professor Sanjoy Dasgupta. I thank my lab-mates Sung-En Chiu, Yongxi Lu, Nancy Ronquillo, Chang-Heng Wang, and Xinghan Wang. I would also like to thank my collaborators Professor Anand Sarwate and Professor Anatoly Khina for many fruitful research interactions. I have numerous people to thank that have invested their time into me, which has helped me reach where I am today, for which I am very grateful. There are far too many people to name individually, and to those people I leave out I apologize.

I thank my husband Kishore for always being a source of strength for me and for unwavering support over the course of these years. Your desire for me to excel exceeded even mine and it has allowed me to follow my dreams. Thanks for always accompanying me and never letting me feel lonely. I also thank my in-laws, Bharati aunty and Rathinavel uncle for their unconditional support to my academic career.

Finally, this thesis would not have been possible without the constant support and encouragement of my parents and my brother Sameer. Thank you Sameer for taking care of me especially the past six months, being my truest friend and instilling in me the ability to appreciate Carnatic music. I thank my grandparents for their blessings, discipline and timeless teachings which are more valuable than a doctorate. There are no words to express my gratitude to my parents for performing the selfless penance of raising me, always believing in me, and smiling through the pain of separation while I chase my dreams. Thank you *amma* and *nanna* for being there in every step of this long journey. I can never repay for the sacrifices you have made for me.

But as the smallest token of my appreciation, I dedicate this thesis to you.

Chapter 2, in full, is a reprint of the material as it appears in the paper: Anusha Lalitha, Nancy Ronquillo and Tara Javidi, “Improved Target Acquisition Rates With Feedback Codes", in *IEEE Journal of Selected Topics in Signal Processing*, vol. 12, no. 5, pp. 871-885, Oct. 2018. The dissertation author was the primary investigator and author of this paper.

Chapter 3, in part, is a reprint of the material as it appears in the paper: Anusha Lalitha and Tara Javidi, “Reliability of sequential hypothesis testing can be achieved by an almost-fixed-length test", in *IEEE International Symposium on Information Theory*, Barcelona, pp. 1710-1714, 2016. The dissertation author was the primary investigator and author of this paper.

Chapter 3, in full, is currently being prepared for submission for publication as: Anusha Lalitha and Tara Javidi, “Almost-fixed-length strategies for Channel Coding and Hypothesis Testing". The dissertation author was the primary investigator and author of this paper.

Chapter 4, in part, is a reprint of the material as it appears in the paper: Anusha Lalitha, Anatoly Khina, Tara Javidi, and Victoria Kostina, “Real-time binary posterior matching", in *IEEE International Symposium on Information Theory*, 2019. The dissertation author was the primary investigator and author of this paper.

Chapter 5, in full, is a reprint of the material as it appears in the paper: Anusha Lalitha, Tara Javidi and Anand D. Sarwate, “Social Learning and Distributed Hypothesis Testing", in *IEEE Transactions on Information Theory*, vol. 64, no. 9, pp. 6161-6179, Sept. 2018. The dissertation author was the primary investigator and author of this paper.

Chapter 6, in full, has been submitted for publication as: Anusha Lalitha, Xinghan Wang, Cihan Kilinc, Yongxi Lu, Tara, Javidi, and Farinaz Koushanfar, “Decentralized Bayesian Learning over Graphs”, available on *arXiv preprint arXiv:1905.10466*.The dissertation author was the primary investigator and author of this paper.

VITA

2012	B. Tech in Electrical Engineering, Indian Institute of Technology, Gandhinagar, India
2015	M. S. in Electrical Engineering (Communication Theory and Systems), University of California, San Diego
2019	Ph. D. in Electrical Engineering (Communication Theory and Systems), University of California, San Diego

PUBLICATIONS

A. Lalitha and T. Javidi, Almost-fixed-length strategies for channel coding and hypothesis tests *in Preparation.*

A. Lalitha, X. Wang, C. Kilinc, Y. Lu, T. Javidi, and F. Koushanfar, Farinaz, "Decentralized Bayesian Learning over Graphs", *arXiv preprint arXiv:1905.10466* under submission.

A. Lalitha, A. Khina, T. Javidi, and V. Kostina, "Real-time binary posterior matching", in *IEEE International Symposium on Information Theory*, 2019.

A. Lalitha, N. Ronquillo and T. Javidi, "Improved Target Acquisition Rates With Feedback Codes", in *IEEE Journal of Selected Topics in Signal Processing*, vol. 12, no. 5, pp. 871-885, Oct. 2018.

A. Lalitha, T. Javidi and A. D. Sarwate, "Social Learning and Distributed Hypothesis Testing", in *IEEE Transactions on Information Theory*, vol. 64, no. 9, pp. 6161-6179, Sept. 2018.

ABSTRACT OF THE DISSERTATION

Belief Refinement Approaches to Communication and Inference Problems

by

Anusha Lalitha

Doctor of Philosophy in Electrical Engineering (Communication Theory and Systems)

University of California San Diego, 2019

Professor Tara Javidi, Chair

This dissertation considers a problem where a single agent or a group of agents aim to estimate/learn unknown (possibly time-varying) parameters of interest despite making noisy observations. The agents take a Bayesian-like approach by maintaining a posterior probability distribution or “belief" over a parameter space conditioned on past observations. The agents aim to iteratively refine their belief over the parameter space as new information is acquired from their private observations or through collaboration with other agents. In particular, the agents aim to ensure that sufficient belief is assigned in neighborhoods centered around the true parameter with high probability or “reliability". In the context of communication problems considered in this dissertation, the agents may be *active*, i.e., agents may additionally take actions which

provide new observations. Furthermore, agents may employ an *adaptive* strategy, i.e., using their past actions and the resulting observations, agents can adaptively choose actions to control the concentration of the belief. When the agents are active, we propose and analyze *adaptive* belief refinement approaches to obtain belief concentration on the unknown parameter with high reliability. In a different context, namely that of decentralized inference, we consider passive agents. Here, agents face an additional challenge due to the statistical insufficiency of their private observations to learn the unknown parameter. While individual agents' observations are not informative enough, we assume that the agents' observations are *collectively* informative to learn the unknown parameter. Here, we propose and analyze *decentralized* belief refining strategies to collaboratively obtain belief concentration on the unknown parameter.

In the first part of this dissertation, we consider active strategies that are extensions of the posterior matching strategy (PM) introduced by Horstein, which is a generalization of the well-known binary search algorithm. We propose and analyze PM based strategies in the context of modern communication systems, namely the problem of establishing initial access in mm-Wave communication and spectrum sensing for Cognitive Radio. We propose and analyze channel coding strategies for real-time streaming and control applications. The second part of the dissertation investigates the belief refinement approaches for decentralized learning. In particular, it focusing on developing and analyzing a decentralized learning rule for statistical hypothesis testing and its application to decentralized machine learning.

Chapter 1

Introduction

This dissertation considers a problem where a single agent or a group of agents aim to estimate/learn unknown (possibly time-varying) parameters of interest despite making noisy observations. The agents take a Bayesian-like approach by maintaining a posterior probability distribution or “belief" over a parameter space conditioned on past observations. The agents aim to iteratively refine their belief over the parameter space as new information is acquired from their private observations or through collaboration with other agents. In particular, the agents aim to ensure that sufficient belief is assigned in neighborhoods centered around the true parameter with high probability or “reliability". In the context of communication problems considered in this dissertation, the agents may be *active*, i.e., agents may additionally take actions which provide new observations. Furthermore, agents may employ an *adaptive* strategy, i.e., using their past actions and the resulting observations, agents can adaptively choose actions to control the concentration of the belief. When the agents are active, we propose and analyze *adaptive* belief refinement approaches to obtain belief concentration on the unknown parameter with high reliability. In a different context, namely that of decentralized inference, we consider passive agents. Here, agents face an additional challenge due to the statistical insufficiency of their private observations to learn the unknown parameter. While individual agents' observations are not

informative enough, we assume that the agents' observations are *collectively* informative to learn the unknown parameter. Here, we propose and analyze *decentralized* belief refining strategies to collaboratively obtain belief concentration on the unknown parameter.

In the first part of this dissertation, we consider active strategies that are extensions of the posterior matching strategy (PM) introduced by Horstein in [1], which is a generalization of the well-known binary search algorithm. We briefly describe the strategy next in the context of a search problem but note that a search strategy can double as a channel coding strategy with feedback and vice versa, which will be discussed later (in chapter 2). Here, the goal of the agent is to estimate the unknown location of a target and the agent's action is to choose a subset of the search space to query. If the target lies in the search subset, then the query outcome is one, otherwise outcome is zero. If the agent always observes the search outcome without any noise, then it is known that the binary search algorithm which always searches half the search space at each time instant, can be used to estimate the target location efficiently. If the agent observes one, it updates the belief over the search subset to one and assigns zero belief to rest the search space. Hence, the agent zooms into the region where it observes one, thus reducing the search space by half after every query. However, when the search outcomes are corrupted by noise, a more sophisticated approach such as PM strategy is necessary. Here, the agent always chooses a search subset to query whose belief is equal to half. After observing the search outcome, the agent updates its belief using the Bayes rule. Intuitively, as indicated by the noisy observation, the belief increases in the region where the target is believed to be, and decreases elsewhere. Shayevitz and Feder show in [2] that the agent's belief concentrates to an exponentially small neighborhood around the true location (as will be shown later the exponent here denotes the rate of information acquisition) with high probability. More recently, Waeber et al. in [3] show that belief concentrates on the true parameter with probability error vanishing exponentially fast as the number of observations grows. Several discretized versions [4–8] of the PM strategy exists where the algorithm splits the search space into a finite number of intervals and aims to locate

the interval that contains the target. It is known that using the discretized PM strategies posterior concentrates with probability error vanishing exponentially fast as the number of observations grows. We devise strategies based on the PM strategy to address various problems at the core of communications which we will describe next.

Chapter 2 considers a discretized version of the above target search problem, where an agent is searching for an unknown target location among a finite number of locations. Here, agent's action consists of simultaneously probing a group of locations. The resulting observation consists of a sum of an indicator of the target's presence in the probed region, and a zero mean Gaussian noise term whose variance is a function of the measurement vector. This problem formulation captures two engineering problems in the context of modern communication systems, namely the problem of establishing initial access in mm-Wave communication and spectrum sensing for Cognitive Radio. We propose a simple two-phase strategy adaptive target search strategy based on PM strategy actively shapes the belief while reducing the noise encountered by the agent to achieve higher reliability. We provide a non-asymptotic upper bound on the expected number of measurements collected under the proposed two-phase strategy. Using a non-asymptotic lower bound of the expected number of measurements for optimal non-adaptive search strategy, a non-trivial lower bound to the adaptivity gain in using an adaptive strategy over a non-adaptive strategy. Our non-asymptotic analysis of adaptivity gain reveals two qualitatively different asymptotic regimes depending on how the number of locations grow.

Note that the adaptive two-phase strategy proposed in chapter 2 has a stopping time i.e., the number of observations it collects, is random variable with bounded expectation. Such a constraint is called average-length constraint. However, in many practical applications using strategies which satisfy only average-length constraint has major limitations since it does not prohibit stopping time from being occasionally very long. Moreover, an average-length constraint does not limit the variability of the stopping time around its expected value. This leads us to the question: is it essential to allow some variability in the stopping time to achieve higher reliability?

In chapter 3 we demonstrate that this flexibility need not be significant. More specifically, we defined a new class of strategies for channel coding and statistical hypothesis testing (two important problems at the core of communications and statistics) which are bounded almost surely and exceed the given test length constraint with an exponentially small probability. We show that reliability achieved by the fixed-length strategies can be significantly by the almost-fixed-length strategies. Furthermore, it is possible to achieve optimal reliability of variable-length strategies in an almost fixed length manner.

In the context of channel coding with feedback, the classical PM strategy and its discretized versions, assume that the entire information (possibly infinite bit) sequence to be transmitted is available non-causally to the transmitter, prior to the beginning of transmission. Consequently, the non-causal knowledge assumption precludes the use of the classical PM strategy for real-time streaming applications and in control scenarios, in which the data to be transmitted is determined in a causal fashion. In Chapter 4 we propose a horizon-free causal posterior matching extends PM strategy to transmit bits which arrive one-by-one at random times. For deterministic inter-bit arrival times we provide characterize the error exponents in horizon-free manner by extending Burnashev-Zigangirov's analysis in [4] such that the older bits have higher reliability than latter bits. For random inter-bit arrival times, the error exponent is obtained for two regimes: a high rate regime where all the bits that are available at the encoder are decoded and a low rate regime where the minimum number of bits are decoded but with a higher exponent. This strategy is then used to estimate a scalar linear stochastic process with unknown initial condition and additive bounded disturbances over BSC.

The second part of the dissertation investigates the belief refinement approaches for decentralized learning. In particular, the second part of the dissertation consists of two chapters focusing on developing and analyzing a decentralized learning rule for statistical hypothesis testing and its application to decentralized machine learning. Recall that agents' private observations may be statistical insufficiency to learn the unknown parameter. While individual agents' observations

are not informative enough, we assume that the agents' observations are collectively informative to learn the unknown parameter.

Chapter 5 considers a problem of decentralized hypothesis testing over a network. Individual agents in a network receive noisy local (private) observations whose distribution is parameterized by a discrete parameter (hypothesis). The marginals of the joint observation distribution conditioned on each hypothesis are known locally at the agents, but the true parameter/hypothesis is not known. An belief merging rule is analyzed in which agents first perform a Bayesian update of their belief (distribution estimate) of each hypothesis based on their local observations, communicate these updates to their neighbors, and then perform a "non-Bayesian" linear consensus using the log-beliefs of their neighbors. Under mild assumptions, we show that the belief of any agent on a wrong hypothesis converges to zero exponentially fast. We characterize the exponential rate of learning, in terms of the agents' influence of the network and the divergences between the observations' distributions. For a broad class of observation statistics which includes distributions with unbounded support such as Gaussian mixtures, we show that rate of rejection of wrong hypothesis satisfies a large deviation principle, i.e., the probability of sample paths on which the rate of rejection of wrong hypothesis deviates from the mean rate vanishes exponentially fast and we characterize the rate function in terms of the agents' influence of the network and the local observation models.

Chapter 6 extends the decentralized belief merge rule to parametric learning and it provides strong analytic guarantees on convergence as well as a closed form characterization of the rate of convergence. We also show that our methodology can be combined with efficient Bayesian inference techniques such as Variational Inference to train Bayesian neural networks in a decentralized manner. By empirical studies we show that our theoretical analysis can guide the design of network/social interactions and data partitioning to achieve convergence. When the training dataset is divided across the network, we demonstrate that we can train a fully connected as well as a convolutional neural network at each agent in a decentralized manner such that it can

successfully perform image classification on the global dataset.

Notation: Vectors are denoted by boldface letters $\mathbf{A}$ and $\mathbf{A}(j)$ is the j^{th} element of a vector. Let $|\mathbf{A}|$ denote the L_1 norm of vector $\mathbf{A}$. Matrices are denoted by overlined boldface letters. $\mathbb{N}$ denotes the set of natural numbers. For $k,t \in \mathbb{N}$, $k < t$, the sequence $\{s_k, s_{k+1}, \ldots, s_t\}$ is denoted as s_k^t. For a set S and scalar $a \in \mathbb{R}$, $a+S$ denotes the set $\{x+a : x \in S\}$ and aS denotes the set $\{ax : x \in S\}$. For sets S_1, S_2, $S_1 \times S_2$ denotes the set $\{(x_1,x_2) : x_1 \in S_1, x_2 \in S_2\}$. For $M \in \mathbb{N}$, the sequence of integers $\{1,2,\ldots,M\}$ is denoted $[M]$. $\mathrm{Ber}(p)$ denotes the Bernoulli distribution with parameter p. The binary entropy of probability p is denoted by $h(p) = -p\log p - \bar{p}\log\bar{p}$ with $\bar{p} := 1-p$; all logarithms in this work are to the base 2. Let $G(x;\mu,\sigma^2)$ denote the pdf of Gaussian random variable with mean μ and variance σ^2 at x. For any probability mass function (pmf) p, let $p^{\otimes i}$ denote p convolved with itself i times. The Kullback–Leibler (KL) divergence between two probability density functions $\mathsf{P}_1(\cdot)$ and $\mathsf{P}_2(\cdot)$ on space $\mathcal{X}$ is defined as $D(\mathsf{P}_1 \| \mathsf{P}_2) = \sum_{\mathcal{X}} \mathsf{P}_1(x) \log \frac{\mathsf{P}_1(x)}{\mathsf{P}_2(x)}$, with the convention $0 \log \frac{a}{0} = 0$ and $b \log \frac{b}{0} = \infty$ for $a,b \in [0,1]$ with $b \neq 0$. Let $\mathcal{U}_M$ denote the set $\{\mathbf{u} \in \mathbb{R}^M : u(j) \in \{0,1\}\}$. Let $[g]_a = g$ if $g \geq a$ otherwise $[g]_a = 0$. For any $M \in \mathbb{N}$, let $\mathsf{l}_M := \{\frac{1}{M}, \frac{2}{M}, \ldots, 1\}$ and let $[M] = \{1,2,\ldots,M\}$. For vectors $\mathbf{x},\mathbf{y} \in \mathbb{R}^d$, let $\mathbf{x} \leq \mathbf{y}$ denote $x_i \leq y_i$ for each i-th element of vector $\mathbf{x}$ and $\mathbf{y}$ and let $\langle \mathbf{x},\mathbf{y} \rangle$ denote $\sum_{i=1}^d x_i y_i$. Let $\mathbf{1}$ denote the vector of where each element is 1. For any subset $F \subset \mathbb{R}^{M-1}$, let F^o be the interior of F and $\bar{F}$ the closure. For $\varepsilon > 0$ let $F_{\varepsilon^+} = \{\mathbf{x} + \delta\mathbf{1}, \forall 0 < \delta \leq \varepsilon \text{ and } \mathbf{x} \in F\}$, $F_{\varepsilon^-} = \{\mathbf{x} - \delta\mathbf{1}, \forall 0 < \delta \leq \varepsilon \text{ and } \mathbf{x} \in F\}$.

Chapter 2

Improved Target Acquisition Rates with Feedback Codes

2.1 Introduction

Consider a single target acquisition over a search region of width B with a resolution up to width δ. Mathematically, this is the problem of estimating a unit vector $\mathbf{W} \in \{0,1\}^{\frac{B}{\delta}}$ via a sequence of noisy linear measurements

$$Y_n = \langle \mathbf{S}_n, \mathbf{W} + \Xi_n \rangle, \quad n = 1,2,\ldots \tag{2.1}$$

where a binary measurement vector $\mathbf{S}_n \in \{0,1\}^{\frac{B}{\delta}}$ denotes the locations inspected and the vector $\Xi_n \in \mathbb{R}^{\frac{B}{\delta}}$ denotes the additive measurement noise per location. More generally, the observation Y_n at time n can be written as

$$Y_n = \langle \mathbf{S}_n, \mathbf{W} \rangle + Z_n(\mathbf{S}_n), \tag{2.2}$$

where $Z_n(\mathbf{S}_n)$ is a noise term whose statistics are a function of the measurement vector $\mathbf{S}_n$. The goal is to design a sequence of measurement vectors $\{\mathbf{S}_n\}_{n=1}^{\tau}$, such that the target location $\mathbf{W}$ is

estimated with high reliability, while keeping the (expected) number of measurements τ as low as possible.

In this chapter, we first consider the linear model (2.1) when the elements of Ξ_n are i.i.d Gaussian with zero mean and variance $\delta\sigma^2$. This means that $Z_n(\mathbf{S}_n)$ in (2.2) are distributed as $\mathcal{N}(0, |\mathbf{S}_n|\delta\sigma^2)$. For this case we show that the problem of searching for a target under measurement dependent Gaussian noise $Z_n(\mathbf{S}_n)$ is equivalent to channel coding over a binary additive white Gaussian noise (BAWGN) channel with state and feedback (in Section 4.6 [9]). This allows us not only to retrofit the known channel coding schemes based on sorted Posterior Matching (sortPM) [10] as adaptive search strategies, but also to obtain information theoretic converses to characterize fundamental limits on the target acquisition rate under both adaptive and non-adaptive strategies. Furthermore, by providing a non-asymptotic analysis of our two stage sorted Posterior-Matching-based adaptive strategy and our converse for non-adaptive strategy, we obtain a lower bound on the adaptivity gain.

2.1.1 Our Contributions

Our main results are inspired by the analogy between target acquisition under measurement dependent noise and channel coding with state and feedback. This connection was utilized in [11] under a Bernoulli noise model. In this chapter, in Proposition 1, we formalize the connection between our target acquisition problem with Gaussian measurement dependent noise and channel coding over a BAWGN channel with state. Here, the channel state denotes the measurement vector. The channel transition depends on the channel state as $\mathcal{N}(X_n, |\mathbf{S}_n|\delta\sigma^2)$ for input codeword $X_n \in \{0, 1\}$. Adapting the codeword to the past channel outputs, i.e. using feedback codes is known to increase the capacity of a channel with state and feedback. This motivates us to use adaptivity when searching, i.e., to utilize past observations $\{Y_i\}_{i=1}^{n-1}$ when selecting the next measurement vector $\mathbf{S}_n$. Furthermore, this information theoretic perspective allows us to quantify the increase in the adaptive target acquisition rate. Our analysis of improvement in the target

acquisition rate as well as the adaptivity gain, measured as the reduction in expected number of measurements, while using an adaptive strategy over a non-adaptive strategy has two components. Firstly, we utilize information theoretic converse for an optimal non-adaptive search strategy to obtain a non-asymptotic lower bound on the minimum expected number of measurements required while maintaining a desired reliability. As a consequence, this provides the best non-adaptive target acquisition rate. Secondly, we utilize a feedback code based on sorted Posterior Matching as a two-stage adaptive search strategy and obtain a non-asymptotic upper bound on the expected number of measurements while achieving a desired reliability. These two components of our analysis allow us to characterize a lower bound on the adaptivity gain.

Our non-asymptotic analysis of adaptivity gain reveals two qualitatively different asymptotic regimes. In particular, we show that adaptivity gain depends on the manner in which the number of locations grow. We show that the adaptivity grows logarithmically with the number of locations $\frac{B}{\delta}$, i.e., $O\left(\log \frac{B}{\delta}\right)$ when refining the search resolution δ (δ going to 0) and while keeping total search width B fixed. On the other hand, we show that as the search width B expands (B goes to ∞) while keeping search resolution δ fixed, the adaptivity gain grows with the number of locations as $O\left(\frac{B}{\delta}\log \frac{B}{\delta}\right)$.

The problem of searching for a target under noisy observations has roots in [1]. Our problem setup is closely related to the problem of sequential estimation of the target location via the noisy 20 questions game studied by Jedynak et al. in [12] and collaborative 20 questions for target localization by Tsiligkaridis et al. in [13]. However, unlike these problem setups which focus on measurement independent noise we consider measurement dependent noise and hence focus on prior work considering the same. The problem of searching for a target under a binary measurement dependent noise, whose crossover probability increases with the weight of the measurement vector was studied by [11] and analyzed under sort PM strategy in [10]. In particular, [11] and [10] provide asymptotic analysis of the adaptivity gain for the case where $B = 1$ and δ approaches zero. Our prior work [14] by utilizing a (suboptimal) hard decoding of

Gaussian observation Y_n, strengthens [11] and [10] by also accounting for the regime in which B grows. While the analysis in [14] strengthens the non-asymptotic bounds in [10] with Bernoulli noise it failed to provide tight analysis for our problem with Gaussian observations. In this chapter, by strengthening our analysis in [14] we further extend the prior work. We provide the following detailed list of our main contributions in this chapter:

1. We consider the problem of searching for a target (vacant) narrow band of width δ over a total bandwidth B via linear binary measurements subject to measurement dependent Gaussian noise (this model is referred to as noise folding in some literature). We establish the equivalence of this problem to the problem of binary-input channel coding over an additive Gaussian channel with state and feedback. This allows us to consider information theoretic techniques to characterize the fundamental limits of searching as well as construct feedback codes as effective search strategies (see Proposition 1 and Corollary 1).

2. We propose a simple intuitive two stage adaptive target search strategy inspired by known feedback codes. This strategy allows us to provide a tight non-asymptotic upper bound on the expected number of measurements needed by an optimal adaptive search strategy. The only known upper bound on the expected number of measurements needed by an optimal adaptive search strategy in the use of Binary Symmetric Channel (BSC) [10, 11] provides an upper bound that is very loose in general and particularly loose for Gaussian observations. In this sense our upper bound significantly tightens the prior analysis (see Remark 3 and Figures 2.6 and 2.7).

3. Obtaining tight non-asymptotic upper bounds on the expected number of measurements needed by an optimal adaptive search strategy allows us to provide better bound on adaptivity gain (see Theorem 1).

4. Our result extends and significantly improves the prior work in the asymptotic regime of B goes to ∞:

- Our setup specializes to two practically relevant asymptotic problems given by Example 1 where resolution δ shrinks while search space width B remains fixed (noise variance shrinking to zero) and Example 2 where resolution δ is fixed but search space width B grows (half bandwidth noise variance linearly growing).
- Our two-stage strategy is shown to be asymptotically optimal in the regime where δ goes to zero and B is fixed. In this regime, our results extends prior work to on BSC [10, 11] to the Gaussian additive noise case with noise folding. Here we note that the BSC work in [10, 11] can be viewed as a pessimistic analysis of the Gaussian measurements with hard-decoding.
- In the asymptotic regime where δ is fixed and B grows, our result significantly improves prior work [11] in incorporating an optimization of the proposed strategy in terms of a parameter α. Note that without such optimization, even if one accepts the hard decoding approximation, all known schemes and analysis fail to provide a non-trivial bound in the asymptotic regime (see Figure 2.6).

2.1.2 Applications

Our problem formulation addresses two challenging engineering problems which arise in the context of modern communication systems. We will discuss the two problems in the following examples and then provide the details of the state of art.

Example 1 (Establishing initial access in mm-Wave communication)**.** Consider the problem of detecting the direction of arrival for initial access in millimeter wave (mmWave) Communications. Prior to data transmission the base station and user equipment are tasked with aligning the transmitter and receiver antennas in the angular space. In other words, each sequential beam can be viewed as a measurement vector $\mathbf{S}_n$ searching the angular space $B \subset (0, 360^\circ)$. Furthermore, beam forming with varying beam widths as we consider in this chapter is associated with effects in

perceived SNR. That is, while a narrow beam exploits antenna gains, a wide beam with the same allocated power inherently achieves lower gains [15, 16](where the gain is inversely proportional to the beam width). This effect results in lower perceived SNR for wider beams, and thus higher probability of error in detecting a synchronization signal. This effect of beam width on the SNR, has motivated our model of measurement dependent noise, where $Z_n(\mathbf{S}_n)$. For initial access search, the angular space B is bounded to 360°, however, the resolution of the search δ is dependent on the number of antennas used, where finer and finer beams can be resolved as more and more antennas are used.

Example 2 (Spectrum Sensing for Cognitive Radio)**.** Consider the problem of opportunistically searching for a vacant subband of bandwidth δ over a total bandwidth of B. In this problem secondary user desires to locate the single stationary vacant subband quickly and reliably, by making measurements $\mathbf{S}_n$ at every time n. We consider the energy based detection [17] where joint multi-band detection is employed. Specifically, we are inspired by the group testing based techniques for cognitive radio presented by [18] where a signal occupancy measurement is acquired by jointly deciding the occupancy of a group of subbands. Sequential measurements are made with fixed sampling rate, i.e. a fixed power consumption. Due to noise folding effects caused by sub-Nyquist sampling [19] at each time instant n, the noise intensity depends on the number of subbands probed as dictated by a measurement vector $\mathbf{S}_n$. Thus, the noisy observation Y_n is a function of measurement dependent noise $Z_n(\mathbf{S}_n)$. The resolution of the search δ can be limited by energy detection technology, while the searchable bandwidth space B is subject to change depending on the needs of the secondary user and is potentially unbounded.

Giordani et al. [15] compare the exhaustive search like the Sequential Beamspace Scanning considered by Barati et al. [20], where the base station sequentially searches through all angular sectors, against a two stage iterative hierarchical search strategy. In the first stage an exhaustive search identifies a coarse sector by repeatedly probing each coarse region for a predetermined SNR to be achieved. In the second stage an exhaustive search over all locations identifies the

target. Giordani et al. show that in general the adaptive iterative strategy reduces the number of measurements over exhaustive search except when desired SNR is too high, forcing the number of measurements required at each stage to get too large. We observe this in through our simulations in Section 2.6-A. In fact, as confirmed by our simulations random-coding-based non-adaptive strategies including the Agile-Link protocol [21], outperform the repetition based adaptive strategies.

Past literature on spectrum sensing for cognitive radio [22–24] and support vector recovery [25, 26] have focused on the problem where $\mathbf{S}_n$ can be real or complex, with measurement independent noise applying both exhaustive search and multiple adaptive search strategies. In contrast, our work considers a simple binary model, $\mathbf{S}_n \in \{0,1\}^{\frac{B}{\delta}}$, but captures the implications of measurement dependence of the noise, which is known in the spectrum sensing literature as noise folding. The problem of measurement dependent noise (known as noise folding) has been investigated in [19] where non-adaptive design of complex measurements matrix satisfying RIP condition has been investigated. Our work compliments this study by characterizing the gain associated with adaptively addressing the measurement dependent noise (noise folding), albeit for the simpler case of binary measurements. We note that the case of adaptively finding a subset of a sufficiently large vacant bandwidth with noise folding is considered in [18], where ideas from group testing and noisy binary search have been utilized. The solutions however depend strongly on the availability of sufficiently large consecutive vacant band and does not apply to our setting.

2.2 Problem Setup

In this section, we describe the mathematical formulation of the target acquisition problem followed by the performance criteria.

2.2.1 Problem Formulation

We consider a search agent interested in quickly and reliably finding the true location of a single stationary target by making measurements over time about the target's presence. In particular, we consider a total search region of width B that contains the target in a location of width δ. In other words, the search agent is searching for the target's location among $\frac{B}{\delta}$ total locations. Let $\mathbf{W} \in \mathcal{U}_{\frac{B}{\delta}}$ denote the true location of the target, where $\mathbf{W}(j) = 1$ if and only if the target is located at location j. The target location $\mathbf{W}$ can take $\frac{B}{\delta}$ possible values uniformly at random whose value remains fixed during the search. A measurement at time n is given by a vector $\mathbf{S}_n \in \mathcal{U}_{\frac{B}{\delta}}$, where $\mathbf{S}_n(j) = 1$ if and only if location j is probed. Each measurement can be imagined to result in a clean observation $X_n = \mathbf{W}^\intercal \mathbf{S}_n \in \{0,1\}$ indicating of the presence of the target in the measurement vector $\mathbf{S}_n$. However, only a noisy version of the clean observation X_n is available to the agent. The resulting noisy observation $Y_n \in \mathbb{R}$ is given by the following linear model with additive measurement dependent noise

$$Y_n = X_n + Z_n(\mathbf{S}_n). \tag{2.3}$$

Here, we assume $Z_n \sim \mathcal{N}(0, |\mathbf{S}_n|\delta\sigma^2)$ which corresponds to the case of i.i.d white Gaussian noise with σ^2 denotes the noise variance per unit width. Conditioned on the measurement vector $\mathbf{S}_n$, the noise Z_n is independent over time. Also define $\sigma^2_{\text{Total}} := \frac{B\sigma^2}{2}$, which denotes the noise intensity when the agent searches half of the search region $\frac{B}{2}$.

A search consisting of τ measurements can be represented by a sequence of measurement vectors $\mathbf{S}_1^\tau = \{\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_\tau\}$ which yields a sequence of observations $Y_1^\tau = \{Y_1, Y_2, \ldots, Y_\tau\}$. At any time instant $n \in [\tau]$, the agent selects the measurement vector in general as a function of the past observations and measurements. Mathematically,

$$\mathbf{S}_n = g_n\left(Y_1^{n-1}, \mathbf{S}_1^{n-1}\right), \tag{2.4}$$

for some causal (possibly random) function $g_n : \mathbb{R}^{n-1} \times \mathcal{U}_{\frac{B}{\delta}}^{n-1} \to \mathcal{U}_{\frac{B}{\delta}}$. After observing the noisy observations Y_1^τ and the sequence of measurement vectors $\mathbf{S}_1^\tau$, the agent estimates the target location $\mathbf{W}$ as follows

$$\hat{\mathbf{W}} = d\left(Y_1^\tau, \mathbf{S}_1^\tau\right), \tag{2.5}$$

for some decision function $d : \mathbb{R}^\tau \times \mathcal{U}_{\frac{B}{\delta}}^\tau \to \mathcal{U}_{\frac{B}{\delta}}$. The probability of error for a search is given by $\mathsf{P}_e = \mathsf{P}(\hat{\mathbf{W}} \neq \mathbf{W} | \mathbf{Y}, \overline{\mathbf{S}})$ and the average probability of error is given by $\overline{\mathsf{P}}_e = \mathsf{P}(\hat{\mathbf{W}} \neq \mathbf{W})$. Now we define the measurement strategy:

Definition 1 (ε-Reliable Search Strategy $\mathfrak{c}_\varepsilon$)**.** For some $\varepsilon \in (0,1)$, an ε-*reliable search strategy*, denoted by $\mathfrak{c}_\varepsilon$, is defined as a sequence of τ (possibly random) number of causal functions $\{g_1, g_2, \ldots, g_\tau\}$, according to which the measurement matrix $\mathbf{S}_1^\tau$ is selected, and a decision function d which provides an estimate $\hat{\mathbf{W}}$ of $\mathbf{W}$, such that the average probability of error $\overline{\mathsf{P}}_e$ is at most ε.

2.2.2 Types of Search Strategies

Every measurement vector $\mathbf{S}_n$ and the number of total measurements τ can be selected either based on the past observations Y_1^{n-1}, or independent of them. Based on these two choices, strategies can be divided into four types i) having fixed length versus variable length of sequence of measurement vectors $\mathbf{S}_1^\tau$, and ii) being adaptive versus non-adaptive. A *fixed length* ε-*reliable strategy* $\mathfrak{c}_\varepsilon$ uses a fixed number of measurements τ predetermined offline independent of the observations, to obtain estimate $\hat{\mathbf{W}}$. On the other hand, a *variable length* ε-*reliable strategy* $\mathfrak{c}_\varepsilon$ uses a random number of measurements τ (possibly determined as a function of the observation sequence Y_1^τ) to obtain an estimate $\hat{\mathbf{W}}$. For example, τ can be selected such that agent achieves $\mathsf{P}_e \leq \varepsilon$ in every search and hence τ is a random variable which is a function of the past noisy observations. Under an *adaptive strategy* $\mathfrak{c}_\varepsilon$ the agent designs the measurement vector $\mathbf{S}_n$ as a

function of the past observations Y_1^{n-1}, i.e., g_n is a function of both $\mathbf{S}_1^{n-1}$ and Y_1^{n-1}.

Definition 2. Let $\mathcal{C}_\varepsilon^A$ be a class of all ε-reliable adaptive strategies.

Under a *non-adaptive strategy*, the agent designs the measurement vector $\mathbf{S}_n$ offline independent of past observations, i.e., g_n does not depend on $\mathbf{S}_1^{n-1}$ or Y_1^{n-1}.

Definition 3. Let $\mathcal{C}_\varepsilon^{NA}$ be a class of all ε-reliable non-adaptive strategies.

2.3 Preliminaries

In this section, we review fundamentals of channel coding with state and feedback and related the information theoretic concepts. The aim is to connect the problem of searching under measurement dependent Gaussian noise to the problem of channel coding with state and feedback. We then formulate an equivalent model of channel coding with state and feedback for comparison to (2.3).

2.3.1 Channel Coding with State and Feedback

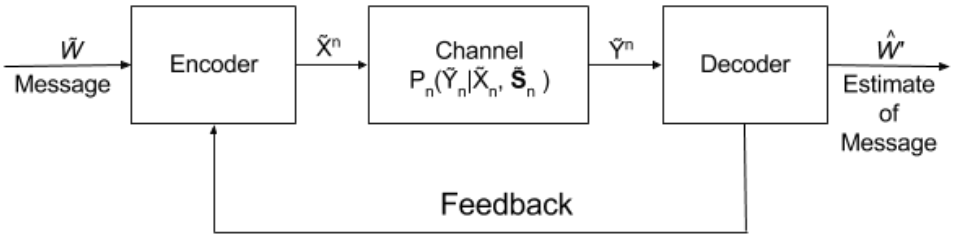


Figure 2.1: Transmission over a communication channel with state and feedback

A communication channel is specified by a set of inputs $\tilde{X} \in \tilde{\mathcal{X}}$, a set of outputs $\tilde{Y} \in \tilde{\mathcal{Y}}$, and a channel transition probability measure $\mathsf{P}(\tilde{y}|\tilde{x})$ for every $\tilde{x} \in \tilde{\mathcal{X}}$ and $\tilde{y} \in \tilde{\mathcal{Y}}$ that expresses the probability of observing a certain output $\tilde{y}$ given that an input $\tilde{x}$ was transmitted [27]. Throughout this work, we will concentrate on coding over a channel with state and feedback (section 4.6

in [9]). Formally, at time n the channel state, $\tilde{\mathbf{S}}_n$ belongs to a discrete and finite set $\tilde{\mathcal{A}}$. We assume that the channel state is known at both the encoder and the decoder. The transition probability at time n is specified by the conditional probability assignment

$$\mathsf{P}_n\left(\tilde{Y}_n\tilde{S}_{n+1}|\tilde{Y}_1^{n-1},\tilde{X}_1^n,\tilde{\mathbf{S}}_1^n\right)=\mathsf{P}_n\left(\tilde{S}_{n+1}|\tilde{Y}_1^n,\tilde{X}_1^n,\tilde{\mathbf{S}}_1^n\right)\mathsf{P}\left(\tilde{Y}_n|\tilde{X}_n,\tilde{S}_n\right). \tag{2.6}$$

Transmission over such a channel is shown in Figure 2.1. In general, the channel state $\tilde{\mathbf{S}}_n$ at time n evolves as a function of all past outputs, inputs, and states,

$$\tilde{\mathbf{S}}_n=\tilde{g}_n(\tilde{Y}_1^{n-1},\tilde{X}_1^{n-1},\tilde{\mathbf{S}}_1^{n-1}). \tag{2.7}$$

The goal is to encode and transmit a uniformly distributed message $\tilde{\mathbf{W}}\in[M]$ over the channel. The encoding function ϕ_n at any time n depends on the message to be transmitted $\tilde{\mathbf{W}}$, all past states, and past outputs. Thus the next symbol to be transmitted is given by

$$\tilde{X}_n=\phi_n(\tilde{Y}_1^{n-1},\tilde{\mathbf{S}}_1^{n-1},\tilde{\mathbf{W}}). \tag{2.8}$$

The encoder obtains the past outputs from the decoder due to the availability of a noiseless feedback channel from decoder to encoder. In this chapter, we assume that both encoder and decoder know the evolution of the channel state, i.e., the sequence $\tilde{\mathbf{S}}_1^n$. After τ channel uses, the decoder uses the noisy observations $\tilde{Y}_1^\tau$ and state information $\tilde{\mathbf{S}}_1^\tau$ to find the best estimate $\tilde{\mathbf{W}}'$, of the message $\tilde{\mathbf{W}}$. The probability of error at the end of message transmission is given by $\mathsf{P}_e=\mathsf{P}(\tilde{\mathbf{W}}'\neq\tilde{\mathbf{W}}|\tilde{Y}_1^\tau,\tilde{\mathbf{S}}_1^\tau)$ and the average probability of error is given by $\overline{\mathsf{P}}_e=\mathsf{P}(\tilde{\mathbf{W}}'\neq\tilde{\mathbf{W}})$.

Example 3 (Binary Additive White Gaussian Noise channel with State and feedback)**.** Consider a Binary Additive White Gaussian Noise (BAWGN) channel with noisy output $\tilde{Y}_n$ given as the sum of input $\tilde{X}_n\in\{0,1\}$ and Gaussian random variable $\tilde{Z}_n\in\mathbb{R}$ whose distribution is a function of the channel state $\tilde{\mathbf{S}}_n$. Specifically, $\tilde{Z}_n$ is a Gaussian random variable with state dependent noise

variance $|\tilde{\mathbf{S}}_n|\delta\sigma^2$ for some $\delta > 0$. In other words, we have

$$\tilde{Y}_n = \tilde{X}_n + \tilde{Z}_n(\tilde{\mathbf{S}}_n), \tag{2.9}$$

where $\tilde{Z}_n \sim \mathcal{N}(0, |\tilde{\mathbf{S}}_n|\delta\sigma^2)$, and the state evolves as $\tilde{\mathbf{S}}_n = \tilde{g}_n(\tilde{Y}_1^{n-1}, \tilde{X}_1^{n-1}\tilde{\mathbf{S}}_1^{n-1})$. Transmission over a BAWGN channel is illustrated in Figure 2.2.

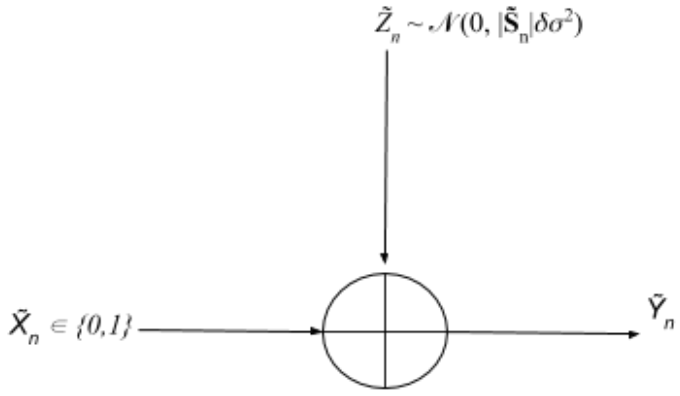


Figure 2.2: Transmission over a BAWGN channel with binary input $\tilde{X}_n$ and Gaussian noise $\tilde{Z}_n$.

Proposition 1. *The problem of searching under measurement dependent Gaussian noise can be cast as a problem of channel coding over a BAWGN channel with state and feedback. Specifically,*

- *The true location vector* $\mathbf{W}$ *can be cast as a message* $\tilde{\mathbf{W}}$ *to be transmitted over the BAWGN. Therefore, by setting* $\tilde{\mathbf{W}} = \mathbf{W}$ *there are* $\frac{B}{\delta}$ *possible messages.*

- *The measurement vector fixes the probability* $\mathsf{P}(\tilde{Y}_n|\tilde{x}_n, \tilde{\mathbf{S}}_n) = \mathcal{N}(\tilde{x}_n, |\boldsymbol{S}_n|\delta\sigma^2)$ *since noise distribution is* $\tilde{Z}_n \sim \mathcal{N}(0, |\mathbf{S}_n|\delta\sigma^2)$ *for* $\tilde{x}_n \in \{0,1\}$. *In other words, by setting* $\tilde{\mathbf{S}}_n = \mathbf{S}_n$ *the measurement vector acts as the channel state and fixes the channel transition probability.*

- *An* ε*-reliable search strategy* $\mathfrak{c}_\varepsilon$ *provides a sequence of* $\{g_1, g_2, \ldots, g_\tau\}$ *such that* $\mathsf{P}(\mathbf{W}' \neq \mathbf{W}) \leq \varepsilon$. *Hence, by setting* $\tilde{g}_i = g_i$ *for all* $i \in \{1, 2, \ldots, \tau\}$, *the search strategy dictates the evolution of channel states* $\tilde{\mathbf{S}}_n$.

- *The sequence of measurement vectors $\mathbf{S}_1^n$ can be used as the codebook. Specifically, the codewords and the encoding strategy are obtained by setting*

$$\tilde{X}_n = \phi_n(\tilde{Y}_1^{n-1}, \tilde{\mathbf{S}}_1^{n-1}, \tilde{\mathbf{W}}) = \mathbf{W}^\intercal \mathbf{S}_n. \tag{2.10}$$

In other words, ε-reliable search strategy $\mathfrak{c}_\varepsilon$ provides an encoding strategy with at most ε probability of error in decoding the true message.

Corollary 1. *The problem of coding over BAWGN channel with codebook dependent state and feedback can be cast a problem of searching under measurement dependent Gaussian noise when the codebook dependent state is given as*

$$\tilde{\mathbf{S}}_n = [\phi_n(\tilde{Y}_1^{n-1}, \tilde{\mathbf{S}}_1^{n-1}, \tilde{\mathbf{W}}(1)), \phi_n(\tilde{Y}_1^{n-1}, \tilde{\mathbf{S}}_1^{n-1}, \tilde{\mathbf{W}}(2)), \ldots, \phi_n(\tilde{Y}_1^{n-1}, \tilde{\mathbf{S}}_1^{n-1}, \tilde{\mathbf{W}}(M))]^\intercal. \tag{2.11}$$

The measurement vector is obtained by setting $\mathbf{S}_n = \tilde{\mathbf{S}}_n$. Therefore, a channel coding strategy with $\mathsf{P}(\tilde{W}' \neq \tilde{W}) \leq \varepsilon$ provides an ε-reliable search strategy over M locations.

The equivalence of problem of searching under measurement dependent noise and the problem of channel coding with state and feedback, implied by Proposition 1 and Corollary 1, provides an efficient way to design and compare non-adaptive and adaptive search strategies. Furthermore, it is known that feedback can improve the capacity of a channel with state [9]. In other words, adaptive coding strategies provide a gain in rate of transmission over non-adaptive coding strategies. This motivates our analysis of the gains to be seen when using an adaptive search strategy over a non-adaptive search strategy. Next we define appropriate figures merit to characterize the gain in using adaptive strategies for the problem of searching under measurement dependent noise.

2.3.2 Target Acquisition Rate and Adaptivity Gain

For any ε-reliable strategy $\mathfrak{c}_\varepsilon$, the performance is measured by the expected number of measurements $\mathbb{E}_{\mathfrak{c}_\varepsilon}[\tau]$. The following definition captures the growth of expected number of measurements as number of search locations grow.

Definition 4 (Achievable Target Acquisition Rate)**.** A target acquisition rate R is said to be ε-*achievable*, if for any $\xi > 0$ there exists an $n(\xi)$ such that for all $n \geq n(\xi)$ there is an ε-reliable search strategy $\mathfrak{c}_\varepsilon$ which satisfies the following

$$\mathbb{E}_{\mathfrak{c}_\varepsilon}[\tau] \leq n, \tag{2.12}$$

$$\frac{B}{\delta} \geq 2^{n(R-\xi)}. \tag{2.13}$$

A targeting rate R is said to be *achievable target acquisition rate* if it is ε-achievable for all $\varepsilon \in (0,1)$.

The above definition is motivated by information theoretic notion of transmission rate over a communication channel, which captures the exponential rate at which the number of messages grows with the number of channel uses while the receiver can decode with a small average error probability. Similarly, the target acquisition rate captures the exponential rate at which the number of target locations grow with the number of measurement vectors while a search strategy can still locate the target with a diminishing average error probability.

Definition 5. The BAWGN capacity with input distribution $\mathsf{Bern}(q)$ and noise variance σ^2 is defined as

$$C\left(q,\sigma^2\right) := -\int_{-\infty}^{\infty} \left((1-q)G(y;0,\sigma^2) + qG(y;1,\sigma^2)\right) \log\left((1-q)G(y;0,\sigma^2) + qG(y;1,\sigma^2)\right) dy - \frac{1}{2}\log(2\pi e\sigma^2). \tag{2.14}$$

Corollary 2. *From channel coding over a BAWGN channel with state and feedback, we obtain that for any small $\xi > 0$ and n large enough, there exists an ε-reliable search strategy $\mathfrak{c}_\varepsilon$ such the*

following holds

$$\mathbb{E}_{\mathfrak{c}_\varepsilon}[\tau] \leq n, \tag{2.15}$$

$$2^{n(C(\frac{1}{2},\sigma^2_{Total})-\xi)} \overset{(a)}{\leq} \frac{B}{\delta} \overset{(b)}{<} 2^{nC(\frac{1}{2},\delta\sigma^2)}, \tag{2.16}$$

where (a) *follows by combining our Corollary 1 with Theorem 4.6.1 in [9], and* (b) *follows by combining our Corollary 1 with the converse of the noisy channel coding theorem [27] and using the fact that the best channel is obtained when noise variance is the least, i.e.,* $\delta\sigma^2$.

Definition 6 (Target Acquisition Capacity under $\sigma^2_{\text{Total}} \geq \rho$)**.** The supremum of achievable target acquisition rates R under $\sigma^2_{\text{Total}} \geq \rho$ is called the target acquisition capacity C_ρ under $\sigma^2_{\text{Total}} \geq \rho$.

Remark 1. Let C_ρ^{NA} and C_ρ^{A} denote the target acquisition capacity under total noise variance $\sigma^2_{\text{Total}} \geq \rho$ over the class of non-adaptive and adaptive strategies respectively. From [10, 11], the gain in the target acquisition capacity when using adaptive strategies is given as

$$C_\rho^{\text{A}} - C_\rho^{\text{NA}} = 1 - \sup_q C(q, 2q\rho) > 0. \tag{2.17}$$

When the width of the search region B grows the noise intensity σ^2_{Total} grows unboundedly and the achievable rate goes to zero. Hence, we first characterize the following notion of adaptivity gain before we characterize the the improvement in target acquisition rate when using adaptive strategies over non-adaptive strategies.

Definition 7 (Adaptivity Gain)**.** The adaptivity gain is defined as the best reduction in the expected number of measurements when searching with an ε-reliable adaptive strategy $\mathfrak{c}'_\varepsilon \in \mathcal{C}_\varepsilon^A$, over an ε-reliable non-adaptive strategy $\mathfrak{c}_\varepsilon \in \mathcal{C}_\varepsilon^{NA}$. Mathematically, it is given as

$$\min_{\mathfrak{c}_\varepsilon \in \mathcal{C}_\varepsilon^{NA}} \mathbb{E}[\tau] - \min_{\mathfrak{c}'_\varepsilon \in \mathcal{C}_\varepsilon^{A}} \mathbb{E}[\tau']. \tag{2.18}$$

2.4 Main Results

In this section, we characterize a lower bound on the adaptivity gain $\min_{\mathfrak{c}_\varepsilon \in \mathcal{C}_\varepsilon^{NA}} \mathbb{E}[\tau] - \min_{\mathfrak{c}'_\varepsilon \in \mathcal{C}_\varepsilon^{A}} \mathbb{E}[\tau']$; the performance improvement measured in terms of reduction in the expected number of measurements for searching over a width B among $\frac{B}{\delta}$ locations under measurement dependent Gaussian noise. First we characterize a lower bound on $\min_{\mathfrak{c}_\varepsilon \in \mathcal{C}_\varepsilon^{NA}} \mathbb{E}[\tau]$.

2.4.1 Non-adaptive Strategies

Lemma 1. *The minimum expected number of measurements required for any ε-reliable non-adaptive search strategy can be lower bounded as*

$$\min_{\mathfrak{c}_\varepsilon \in \mathcal{C}_\varepsilon^{NA}} \mathbb{E}_{\mathfrak{c}_\varepsilon}[\tau] \geq \frac{(1-\varepsilon)\log\left(\frac{B}{\delta}\right) - h(\varepsilon)}{C\left(q^*, q^* B\sigma^2\right)}.$$

Proof of the Lemma 1 is provided in Appendix-A. The proof follows from the fact that clean signal X_i and noise Z_i are independent over time and independent of past observations for $i \in [n]$, due to the non-adaptive nature of the search strategy. In the absence of information from past observation outcomes, the agent tries to maximize the mutual information $I(X_i, Y_i)$ at every measurement. Since $X_i \sim \mathsf{Bern}(q_i)$ and $Z_i \sim \mathcal{N}(0, q_i B\sigma^2)$, the mutual information $I(X_i, Y_i) = C\left(q_i, q_i B\sigma^2\right)$ is maximized at $q_i = q^*$.

Figure 2.3 shows the behavior of the maximum mutual information for a non-adaptive strategy $C\left(q^*, q^* B\sigma^2\right)$ as number of locations grow. When $B = 1$ and δ goes to zero, $C\left(q^*, q^* B\sigma^2\right)$ remains same for a given σ^2. On the other hand, when $\delta = 1$ and B grows, $C\left(q^*, q^* B\sigma^2\right)$ goes to zero. Furthermore, $C\left(q^*, q^* B\sigma^2\right)$ goes to zero faster when $\sigma^2 = 0.05$ than when $\sigma^2 = 0.001$. This implies non-adaptive strategies need a growing number of measurements as B grows. On the other hand, an adaptive strategy can reduce the number measurements as follows. Whenever the agent narrows down the target's location to some coarse fraction of the total search region (say a section of width αB) with high confidence, an adaptive strategy can zoom in and search only

within αB section. This reduces the noise intensity in the measurements unlike a non adaptive strategy which still searches regions of width q^*B. Hence, non-adaptive strategies perform poorly in comparison to adaptive strategies that rapidly zoom in to smaller regions especially when $C\left(q^*, q^*B\sigma^2\right)$ close to zero (as shown in Figure 2.3 for $\sigma^2 = 0.05$).

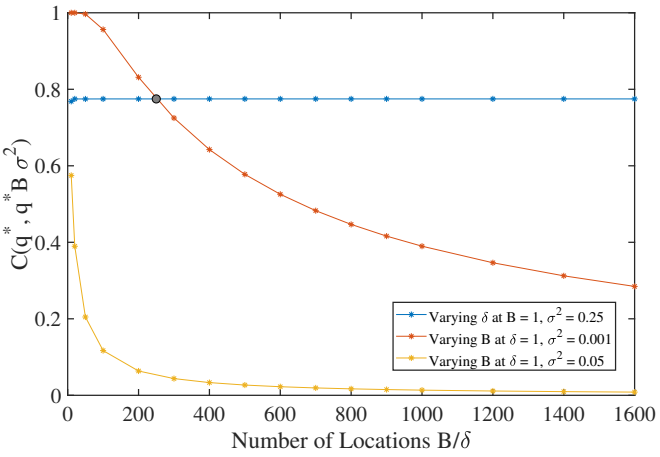


Figure 2.3: Non-adaptive capacity as number of locations grow for various values of σ^2.

2.4.2 Lower Bound on Adaptivity Gain

The expected number of measurements required to zoom in to a region of width αB with high confidence is larger when α is small. On the other hand, noise intensity reduces more significantly after zooming in to a region of width αB, for small α than for large α. This reduces the expected number of measurements needed to locate the target within the region αB upto a resolution δ with high confidence. For any adaptive strategy, there is a trade-off between how rapidly an adaptive strategy can zoom in to and the width of the region to which it zooms in. This trade-off is controlled by the value of parameter α. Since adaptive strategies observe less noisy measurements than non-adaptive strategies after zooming, parameter α also controls the adaptivity gain. This intuition is formalized by the following theorem.

Theorem 1. *Let* $\varepsilon \in (0,1)$. *For any* ε*-reliable non-adaptive strategy* $\mathfrak{c}_\varepsilon \in \mathcal{C}_\varepsilon^{NA}$ *searching over*

a search region of width B among $\frac{B}{\delta}$ *locations with* τ *number of measurements, there exists an* ε*-reliable adaptive strategy* $\mathfrak{c}'_\varepsilon \in \mathcal{C}_\varepsilon^A$ *with* τ' *number of measurements, such that for any* $\eta > 0$ *the following holds*

$$\min_{\mathfrak{c}_\varepsilon \in \mathcal{C}_\varepsilon^{NA}} \mathbb{E}_{\mathfrak{c}_\varepsilon}[\tau] - \min_{\mathfrak{c}'_\varepsilon \in \mathcal{C}_\varepsilon^{A}} \mathbb{E}_{\mathfrak{c}'_\varepsilon}[\tau'] \geq \max_{\alpha \in \mathsf{I}_{\frac{B}{\delta}}} \left\{ \frac{\log \frac{1}{\alpha}}{g^{(1)}_{\varepsilon,\eta}(q^*, B\sigma^2)} + \frac{\log \frac{\alpha B}{\delta}}{g^{(2)}_{\varepsilon,\eta}(\alpha, q^*, B\sigma^2)} - h_{\varepsilon,\eta}(\delta, \alpha, B\sigma^2) \right\},$$

where

$$g^{(1)}_{\varepsilon,\eta}(q^*, B\sigma^2) = \left(\frac{(1-\varepsilon)}{C(q^*, q^* B\sigma^2)} - \frac{1}{C\left(q^*, q^* B\sigma^2\right) - \eta} \right)^{-1},$$
$$g^{(2)}_{\varepsilon,\eta}(\alpha, q^*, B\sigma^2) = \left(\frac{(1-\varepsilon)}{C(q^*, q^* B\sigma^2)} - \frac{1}{C\left(\frac{1}{2}, \frac{\alpha B\sigma^2}{2}\right) - \eta} \right)^{-1},$$

and

$$h_{\varepsilon,\eta}(\delta, \alpha, B\sigma^2) = \frac{\log\left(\frac{2}{\varepsilon}\right) + \log\log\left(\frac{1}{\alpha}\right) + a_\eta}{C\left(q^*, q^* B\sigma^2\right) - \eta} + \frac{\log\left(\frac{2}{\varepsilon}\right) + \log\log\left(\frac{\alpha B}{\delta}\right) + a_\eta}{C\left(\frac{1}{2}, \frac{\alpha B\sigma^2}{2}\right) - \eta} + \frac{h(\varepsilon)}{C(q^*, q^* B\sigma^2)},$$

$q^* = \text{argmax}_{q \in \mathsf{I}_{\frac{B}{\delta}}} C(q, qB\sigma^2)$, *and* a_η *is the solution of the following equation*

$$\eta = \frac{a}{a-3} \max_{q \in \mathsf{I}_{\frac{B}{\delta}}} \int_{-\infty}^{\infty} \frac{e^{-\frac{y^2}{2Bq\sigma^2}}}{\sqrt{2\pi q B\sigma^2}} \left[\frac{2y-1}{2qB\sigma^2} \right]_{(a-3)} dy. \tag{2.19}$$

Proof of Theorem 1 is obtained by combining Lemma 1 and Lemma 2. Theorem 1 provides a non-asymptotic lower bound on adaptivity gain.The first two terms in the above lower bound can be viewed as corresponding to two stages. Intuitively, the first part corresponds to the initial stage of the search, where the agent narrows down the target's location to some coarse α

fraction of the total search region, i.e., narrows to a section of width αB with high confidence. The second stage corresponds to refined the search within one of the coarse sections αB obtained from initial stage. For the first stage, our bound predicts negligible gains in using an adaptive strategy over a non-adaptive strategy and it is captured by the term $\frac{1}{g^{(1)}_{\varepsilon,\eta}(q^*,B\sigma^2)}$. However, in the second stage where an adaptive strategy zooms in to a width of αB, a significant gain can be seen especially as B grows and it is captured by the term $\frac{1}{g^{(2)}_{\varepsilon,\eta}(\alpha,q^*,B\sigma^2)}$. The following corollary characterizes the adaptivity gain in the two asymptotic regimes δ going to zero and B growing.

Corollary 3. *Let $\varepsilon \in (0,1)$. For any ε-reliable non-adaptive strategy $\mathfrak{c}_\varepsilon \in \mathcal{C}^{NA}_\varepsilon$ searching over a search region of width B among $\frac{B}{\delta}$ with τ number of measurements, there exists an ε-reliable adaptive strategy $\mathfrak{c}'_\varepsilon \in \mathcal{C}^{A}_\varepsilon$ with τ' number of measurements, such that for a fixed B the asymptotic adaptivity gain grows logarithmically with the total number of locations,*

$$\lim_{\delta\to 0} \frac{\mathbb{E}_{\mathfrak{c}_\varepsilon}[\tau] - \mathbb{E}_{\mathfrak{c}'_\varepsilon}[\tau']}{\log \frac{B}{\delta}} \geq \frac{1-\varepsilon}{C(q^*, q^*B\sigma^2)} - 1. \tag{2.20}$$

For a fixed δ, the asymptotic adaptivity gain grows at least linearly with total number of locations,

$$\lim_{B\to\infty} \frac{\mathbb{E}_{\mathfrak{c}_\varepsilon}[\tau] - \mathbb{E}_{\mathfrak{c}'_\varepsilon}[\tau']}{\frac{B}{\delta}\log \frac{B}{\delta}} \geq \frac{(1-\varepsilon)\delta\sigma^2}{\log e}. \tag{2.21}$$

Furthermore, we have

$$\lim_{B\to\infty} \frac{\min_{\mathfrak{c}_\varepsilon \in \mathcal{C}^{NA}_\varepsilon} \mathbb{E}_{\mathfrak{c}_\varepsilon}[\tau]}{\frac{B}{\delta}\log \frac{B}{\delta}} \geq \frac{(1-\varepsilon)\delta\sigma^2}{\log e}, \tag{2.22}$$

and

$$\lim_{B\to\infty} \frac{\min_{\mathfrak{c}_\varepsilon \in \mathcal{C}^{A}_\varepsilon} \mathbb{E}_{\mathfrak{c}'_\varepsilon}[\tau']}{\frac{B}{\delta}} \leq 16\delta\sigma^2. \tag{2.23}$$

The proof of the above corollary is provided in Appendix-C.

Remark 2. The above corollary characterizes the two qualitatively different regimes discussed previously. For fixed B, as δ goes to zero the asymptotic adaptivity gain scales as only $\log \frac{B}{\delta}$, whereas for fixed δ, as B increases the asymptotic adaptivity gain scales as $\frac{B}{\delta} \log \frac{B}{\delta}$. In other words, adaptivity provides a larger reduction in the expected number of measurements for the regime where the total search width is growing than in the case where we fix the total width and shrink the location widths. In Section 2.6 we related this phenomenon to the diminishing capacity of BAWGN channel when the total noise σ^2_{Total} grows.

Next we provide the main technical components of the proof of Theorem 1.

2.4.3 Adaptive Search Strategies

Consider the following two stage search strategy.

First Stage (Fixed Composition Strategy $\mathfrak{c}_{\frac{\varepsilon}{2}}^{\frac{1}{\alpha}}$)

We group the $\frac{B}{\delta}$ locations of width δ into $\frac{1}{\alpha}$ sections of width αB. Let $\mathbf{W}'$ denote the true location of the target among the sections of width αB. Now, we use a non-adaptive strategy to search for the target location among $\frac{1}{\alpha}$ sections of width αB. In particular, we use a fixed composition strategy where at every time instant n, the fraction of total locations probed is fixed to be q^*. In other words, the measurement vector $\mathbf{S}'_n$ at every instant n is picked uniformly randomly from the set of measurement vectors $\{\mathbf{S}' \in \mathcal{U}_{\frac{1}{\alpha}} : |\mathbf{S}'| = \lfloor \frac{q^*}{\alpha} \rfloor\}$. For the ease of exposition, we assume that $\frac{q^*}{\alpha}$ is an integer. Hence, for this strategy, at every n, $X_n \sim \mathsf{Bern}(q^*)$ and $Z_n \sim \mathcal{N}(0, q^* B \sigma^2)$. For all $i \in \{1, 2, \ldots, \frac{1}{\alpha}\}$, let $\rho'_n(i)$ be the posterior probability of the estimate $\hat{\mathbf{W}}'(i) = 1$ after reception of $\mathbf{Y}^{n-1}$, i.e., $\rho'_n(i) := \mathsf{P}\left(\hat{\mathbf{W}}'(i) = 1 | Y^{n-1}\right)$ and let $\rho'_n := \left\{\rho'_n(1), \rho'_n(2), \ldots, \rho'_n\left(\frac{1}{\alpha}\right)\right\}$. Assume that agent begins with a uniform probability over the $\frac{1}{\alpha}$ sections, i.e., $\rho'_0 = \{\alpha, \alpha, \ldots, \alpha\}$. The posterior probability $\rho'_{n+1}(i)$ at time $n+1$ when

$Y_n = y$ is obtained by the following Bayesian update:

$$\rho'_{n+1}(i) = \begin{cases} \frac{\rho'_n(i)G(y;1,q^*B\sigma^2)}{\mathcal{D}'_n} & \text{if } S'_n(i) = 1, \\ \frac{\rho'_n(i)G(y;0,q^*B\sigma^2)}{\mathcal{D}'_n} & \text{if } S'_n(i) = 0, \end{cases} \tag{2.24}$$

where

$$\mathcal{D}'_n = \sum_{j:\mathbf{1}_{\{S_n(j)=1\}}} \rho'_n(j)G(y;1,q^*B\sigma^2) + \sum_{j:\mathbf{1}_{\{S_n(j)=0\}}} \rho'_n(j)G(y;0,q^*B\sigma^2). \tag{2.25}$$

Let $\tau^1 := \inf\left\{n : \max_i \rho'_n(i) \geq 1 - \frac{\varepsilon}{2}\right\}$ be the number of measurements used under stage 1. Note that τ^1 is a random variable. Hence, first stage is a non-adaptive variable length strategy. Now, the expected stopping time $\mathbb{E}_{\mathfrak{c}^1_{\frac{\varepsilon}{2}}}[\tau^1]$ can be upper bounded using Lemma 3 from Appendix-B.

Second Stage (Sorted Posterior Matching Strategy $\mathfrak{c}^2_{\frac{\varepsilon}{2}}$)

In the second stage, the agent zooms into the αB width section obtained from the first stage and uses an adaptive strategy to search only within this αB section. The agent searches for the target location of width δ among the remaining $\frac{\alpha B}{\delta}$ locations. In particular, we use the sorted posterior matching strategy proposed in [10] which we describe next. Let $\mathbf{W}''$ denote the true target location of width δ. For all $i \in \{1,2,\ldots,\frac{\alpha B}{\delta}\}$, let $\rho''_n(i)$ be the posterior probability of the estimate $\hat{\mathbf{W}}''(i) = 1$ after reception of $\mathbf{Y}^{n-1}$, i.e., $\rho'_n(i) := \mathsf{P}\left(\hat{\mathbf{W}}''(i) = 1 | \mathbf{Y}^{n-1}\right)$ and let $\rho''(n) := \{\rho''_n(1), \rho''_n(2), \ldots, \rho''_n\left(\frac{\alpha B}{\delta}\right)\}$. Assume the agent begins with a uniform probability over the $\frac{\alpha B}{\delta}$ sections, i.e., $\rho''_0 = \left\{\frac{\delta}{\alpha B}, \frac{\delta}{\alpha B}, \ldots, \frac{\delta}{\alpha B}\right\}$. At every time instant n, we sort the posterior values in descending order to obtain the sorted posterior vector $\rho''^{\downarrow}_n$. Let vector I_n denote the corresponding ordering of the location indices in the new sorted posterior. Define

$$k^*_n := \underset{i}{\operatorname{argmin}} \left| \sum_{j=1}^{i} \rho''^{\downarrow}_n(j) - \frac{1}{2} \right|. \tag{2.26}$$

We choose the measurement vector $\mathbf{S}_n''$ such that $\mathbf{S}_n''(j) = 1$ if and only if $j \in \{I_n(1), \ldots, I_n(k_n^*)\}$. Note that for this strategy, at every n, the noise is $Z_n \sim \mathcal{N}(0, |\mathbf{S}_n''|\delta\sigma^2)$ and the worst noise intensity is $\mathcal{N}(0, \frac{\alpha B\sigma^2}{2})$. The posterior probability $\rho_{n+1}''(i)$ at time $n+1$ when $Y_n = y$ is obtained by the following Bayesian update:

$$\rho_{n+1}''(i) = \begin{cases} \frac{\rho_n''(i)G(y;1,|\mathbf{S}_n''|\delta\sigma^2)}{\mathcal{D}_n''} & \text{if } S_n''(i) = 1, \\ \frac{\rho_n''(i)G(y;0,|\mathbf{S}_n''|\delta\sigma^2)}{\mathcal{D}_n''} & \text{if } S_n''(i) = 0, \end{cases} \tag{2.27}$$

where

$$\mathcal{D}_n'' = \sum_{j:\mathbf{1}_{\{S_n(j)=1\}}} \rho_n''(j)G\left(y;1,|\mathbf{S}_n''|\delta\sigma^2\right) + \sum_{j:\mathbf{1}_{\{S_n(j)=0\}}} \rho_n''(j)G\left(y;0,|\mathbf{S}_n''|\delta\sigma^2\right). \tag{2.28}$$

Let $\tau^2 := \inf\left\{n : \max_i \rho_n^2(i) \geq 1 - \frac{\varepsilon}{2}\right\}$ be the number of measurements used under stage 2. Note that τ^2 is a random variable. Hence, the second stage is an adaptive variable length strategy. The expected number of measurements $\mathbb{E}_{\mathfrak{c}^2_{\frac{\varepsilon}{2}}}[\tau'']$ can be upper bounded using Lemma 6 from Appendix-B.

Noting that the total probability of error of the two stage search strategy is less than ε and that the expected stopping time is $\mathbb{E}_{\mathfrak{c}'_\varepsilon}[\tau'] = \mathbb{E}_{\mathfrak{c}^1_{\frac{\varepsilon}{2}}}[\tau^1] + \mathbb{E}_{\mathfrak{c}^2_{\frac{\varepsilon}{2}}}[\tau^2]$, we have the assertion of the following lemma.

Lemma 2. *The minimum expected number of measurements required for the above* ε*-reliable adaptive search strategy* $\mathfrak{c}'_\varepsilon$ *can be upper bounded as*

$$\mathbb{E}_{\mathfrak{c}'_\varepsilon}[\tau'] \leq \min_{\alpha \in \mathsf{I}_{\frac{B}{\delta}}} \left\{ \frac{\log\frac{1}{\alpha} + \log\frac{2}{\varepsilon} + \log\log\frac{1}{\alpha} + a_\eta}{C\left(q^*, q^*B\sigma^2\right) - \eta} + \frac{\log\frac{\alpha B}{\delta} + \log\frac{2}{\varepsilon} + \log\log\frac{\alpha B}{\delta} + a_\eta}{C\left(\frac{1}{2}, \frac{\alpha B\sigma^2}{2}\right) - \eta} \right\}. \tag{2.29}$$

Remark 3. Recall that $\min_{\mathfrak{c}'_\varepsilon \in \mathcal{C}_\varepsilon^A} \mathbb{E}_{\mathfrak{c}'_\varepsilon}[\tau']$ denotes the minimum expected number of measurements required by the optimal adaptive strategy non-asymptotically. Lemma 2 provides an upper bound

on $\min_{\mathfrak{c}'_\varepsilon \in \mathcal{C}_\varepsilon^A} \mathbb{E}_{\mathfrak{c}'_\varepsilon}[\tau']$ using the two stage adaptive strategy. The sorted posterior matching strategy proposed in [10] provides another upper bound for $\min_{\mathfrak{c}'_\varepsilon \in \mathcal{C}_\varepsilon^A} \mathbb{E}_{\mathfrak{c}'_\varepsilon}[\tau']$. However, this bound is very loose. In fact, sorted posterior matching startegy empirically performs significantly better than the bound predicted by the analysis in [10]. Lemma 2 using a possibly sub-optimal strategy than sorted posterior matching provides a significantly tighter bound on $\min_{\mathfrak{c}'_\varepsilon \in \mathcal{C}_\varepsilon^A} \mathbb{E}_{\mathfrak{c}'_\varepsilon}[\tau']$ as illustrated in Figures 2.6 and 2.7.

Remark 4. In the regime of fixed B and diminishing δ, Lemma 2 together with Corollary 2 establishes the asymptotic optimality of our proposed algorithm.

2.5 Extensions and Generalizations

2.5.1 Generalization to other noise models

The main results presented in this chapter consider the setup where the noise Z_n is distributed as $\mathcal{N}(0, |\mathbf{S}_n|\delta\sigma^2)$. In other words, the variance of the noise given by $(|\mathbf{S}_n|\delta\sigma^2)$ is a linear function of the size of a measurement vector $|\mathbf{S}_n|$. This model assumption holds when each target location adds noise equally and independently of other locations when probed together. In general, due to correlation across locations the additive noise variance can be assumed to scale as a non-decreasing function $f(\cdot)$ of the measurement vector $|\mathbf{S}_n|$. In this section, we extend our model to a general formulation for the noise $Z_n \sim \mathcal{N}(0, f(|\mathbf{S}_n|)\delta\sigma^2)$, where $f(\cdot)$ is a non-decreasing function of $|\mathbf{S}_n|$. For example, $f(\mathbf{S}_n) = |\mathbf{S}_n|^\gamma$ for some $\gamma > 0$. Figure 2.4 shows that the effect of the noise function $f(|\mathbf{S}_n|)$ on the capacity. The following theorem is an extension of Theorem 1 to the general formulation of noise. We provide the theorem without a proof since it closely follows the proof of Theorem 1.

Theorem 2. *Let $\varepsilon \in (0,1)$ and let $f(\cdot)$ be a non-decreasing function. For any ε-reliable non-adaptive strategy $\mathfrak{c}_\varepsilon \in \mathcal{C}_\varepsilon^{NA}$ searching over a search region of width B among $\frac{B}{\delta}$ locations with τ*

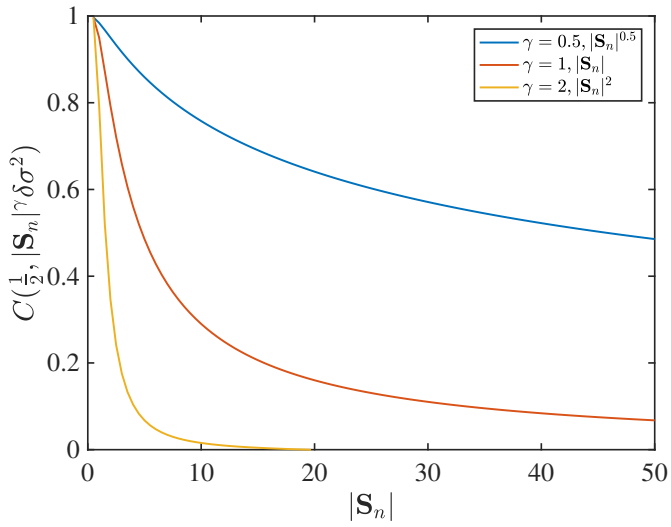


Figure 2.4: Behavior of capacity of BAWGN channel with $\sigma^2 = 0.25$ over a total search region of width $B = 10$, location width $\delta = 0.1$, as a function of the size of a measurement $|\mathbf{S}_n|$.

number of measurements, there exists an ε*-reliable adaptive strategy* $\mathfrak{c}'_\varepsilon \in \mathcal{C}_\varepsilon^A$ *with* τ' *number of measurements, such that for any constant* $\eta > 0$ *the following holds*

$$\min_{\mathfrak{c}_\varepsilon \in \mathcal{C}_\varepsilon^{NA}} \mathbb{E}_{\mathfrak{c}_\varepsilon}[\tau] - \min_{\mathfrak{c}'_\varepsilon \in \mathcal{C}_\varepsilon^{A}} \mathbb{E}_{\mathfrak{c}'_\varepsilon}[\tau'] \geq \max_{\alpha \in \mathsf{I}_{\frac{B}{\delta}}} \left\{ \frac{\log \frac{1}{\alpha}}{g^{(1)}_{\varepsilon,\eta}(q^*, B\sigma^2)} \frac{\log \frac{\alpha B}{\delta}}{g^{(2)}_{\varepsilon,\eta}(\alpha, q^*, B\sigma^2)} \right\} (1 + o(1)),$$

where

$$g^{(1)}_{\varepsilon,\eta}(q^*, B\sigma^2) = \left(\frac{(1-\varepsilon)}{C(q^*, f(\frac{q^* B}{\delta})\delta\sigma^2)} - \frac{1}{C\left(q^*, f(\frac{q^* B}{\delta})\delta\sigma^2\right) - \eta} \right)^{-1},$$

$$g^{(2)}_{\varepsilon,\eta}(\alpha, q^*, B\sigma^2) = \left(\frac{(1-\varepsilon)}{C(q^*, f(\frac{q^* B}{\delta})\delta\sigma^2)} - \frac{1}{C\left(\frac{1}{2}, f(\frac{\alpha B}{2\delta})\sigma^2\right) - \eta} \right)^{-1},$$

and $q^* = \operatorname{argmax}_{q \in \mathsf{I}_{\frac{B}{\delta}}} C(q, f(\frac{qB}{\delta})\delta\sigma^2)$, *and* $o(1)$ *goes to 0 as* $\frac{B}{\delta} \to \infty$.

2.5.2 Multiple Targets

The problem formulation and the main results of this chapter consider the special case when there exists a single stationary target. Suppose instead the agent aims to find the true location of r unique targets quickly and reliably. Our problem formulation is easily extended to the general case where there may exist multiple targets. In our generalization to multiple targets under the linear noise model (2.3), the clean signal indicates the the number of targets present in the measurement vector $\mathbf{S}_n$. In particular, let $\mathbf{W}^{(i)} \in \mathcal{U}_{\frac{B}{\delta}}$ be such that $\mathbf{W}^{(i)}(j) = 1$ if and only if j-th location contains the i-th target. Then, the noisy observation is given as

$$Y_n = \sum_{i=1}^{r} (\mathbf{W}^{(i)})^{\intercal} \mathbf{S}_n + Z_n, \tag{2.30}$$

where $Z_n \sim \mathcal{N}(0, |\mathbf{S}_n| \delta \sigma^2)$. Setting $X_n^{(i)} = (\mathbf{W}^{(i)})^{\intercal} \mathbf{S}_n$ for $i \in [r]$, we have

$$Y_n = \sum_{i=1}^{r} X_n^{(i)} + Z_n. \tag{2.31}$$

The problem of searching for multiple targets is equivalent to the problem of channel coding over a Multiple Access Channel (MAC) with state and feedback [28]. In other words, we can extend the Proposition 1, to channel coding over a MAC with state and feedback with the following constraints: (i) $\mathbf{W}^{(i)}$ can be viewed as the message to be transmitted by the i-th transmitter, (ii) the measurement matrix $\overline{\mathbf{S}}_n$ can be viewed as the common codebook shared by all the transmitters, and (iii) a search strategy dictates the evolution of the MAC state. The channel transition is then fixed by the channel state which is measurement dependent.

Example 1′ (Establishing initial access in mm-Wave communications). In the deployment of mm-Wave links into a cellular or 802.11 network, the base station needs to to quickly switch between users and accommodate multiple mobile clients. In a dense network as such, a received signal at the base station will be corrupted by measurement dependent noise (due to channel

properties) and by neighboring interfering users. Thus, in our model at time n the noisy observation, Y_n of eq.2.31, is a function of inputs from multiple users in the network, in addition to a measurement dependent noise.

Example $2'$ (Spectrum Sensing for Cognitive Radio). In spectrum sensing for cognitive radio, an agent is tasked with opportunistically searching for r vacant subbands of bandwidth δ over a total bandwidth of B. In this problem we desire to locate r stationary vacant subbands quickly and reliably, by making measurements over time. Here again the noise intensity depends on the number of subbands probed, $\mathbf{S}_n$, at each time instant n.

Searching for multiple targets with measurement dependent noise is a significantly harder problem compared to a single target case and achievability strategies for this problem even in the absence of noise are far more complex [29, 30].

2.6 Numerical Results

In this section we provide numerical analysis.

2.6.1 Comparing Search Strategies

In this section, we empirically compare the performance in expected number of measurements $\mathbb{E}_{\mathfrak{c}_\varepsilon}[\tau]$ required by four ε-reliable strategies proposed in the literature. In addition to the sortPM strategy $\mathfrak{c}^2_\varepsilon$, and the optimal variable length non-adaptive strategy i.e., the fixed composition strategy $\mathfrak{c}^1_\varepsilon$, we also consider two noisy variants of the binary search strategy. The noisy binary search applied to our search model selects the locations to be searched at time n, i.e. the search region $\mathbf{S}_n$, to be half the width of the previous search region $\mathbf{S}_{n-1}$. In particular, it zooms in to the half region of $\mathbf{S}_{n-1}$ which has accumulated higher posterior probability.

The first variant we consider is the fixed length noisy binary search, which resembles the adaptive iterative hierarchical search strategy [15]. In this strategy, each measurement is repeated

$\alpha_\varepsilon(\mathbf{S}_n)|\mathbf{S}_n|$ number of times, where $\alpha_\varepsilon(\mathbf{S}_n)$ is a number chosen as a function of $\mathbf{S}_n$ such that all combined measurements result in an ε-reliable search strategy. That is, each measurement vector $\mathbf{S}_n$ is used $\alpha_\varepsilon(\mathbf{S}_n)|\mathbf{S}_n|$ number of times, before the strategy zooms into a region of half the size. The second variant is a variable length version of the similar noisy binary search where each measurement vector $\mathbf{S}_n$ is used until we obtain error probability less than $\varepsilon_p := \frac{\varepsilon}{\log B/\delta}$ either inside or outside of $\mathbf{S}_n$. Table I provides a quick summary of the search strategies. Note that Table I also includes a short summary of our two-stage strategy, although this strategy is studied in the next section (Section VI-B).

Table 2.1: Candidate Search Strategies

Strategies $\mathfrak{c}_\varepsilon \in \mathcal{C}_\varepsilon$	Description of $\mathbf{S}_n$ selection
Optimal non-adaptive	• Select $\mathbf{S}_n$ s.t. $\lvert\mathbf{S}_n\rvert = \frac{q^*B}{\delta}$ as dictated by strategy $\mathfrak{c}_\varepsilon^1$
Fixed Length Noisy Binary	• Select $\mathbf{S}_n$ as dictated by binary search strategy • Repeat $\alpha_\varepsilon(\mathbf{S}_n)\lvert\mathbf{S}_n\rvert$ times
Variable Length Noisy Binary	• Select $\mathbf{S}_n$ as dictated by binary search strategy • Repeat τ times s.t. $\tau = \min\{n : \lVert\rho_n\rVert_\infty \geq 1-\varepsilon_p\}$
Sorted Posterior Matching	• Select $\mathbf{S}_n$ as dictated by Sort PM strategy $\mathfrak{c}_\varepsilon^2$
Two-stage Strategy	• Phase 1: Search among $(\frac{1}{\alpha})$ large subsets. Select $\mathbf{S}_n$ fixed composition s.t. $\lvert\mathbf{S}_n\rvert = \frac{q^*B}{\delta}$ • Phase 2: Zoom into region of size αB, and select $\mathbf{S}_n$ with Sort PM strategy $\mathfrak{c}_\varepsilon^2$

Figure 2.5, shows the performance of each ε-reliable search strategy, when considering fixed parameters B, δ, and ε. We note that the fixed length noisy binary strategy performs poorly in comparison to the optimal non-adaptive strategy. This shows that randomized non-adaptive

search strategies, similar to the one considered in [21] perform better than both the exhaustive search (as shown in [21]) and the iterative hierarchical search strategy. In particular, it performs better than the variable length noisy binary search since when the noise variance parameter σ^2 is large, this higher noise intensity requires that each measurement is repeated far too many times in order to be ε-reliable. The performance of the optimal fully adaptive variable length strategies sort PM [10] is superior to all strategies even in the non-asymptotic regime.

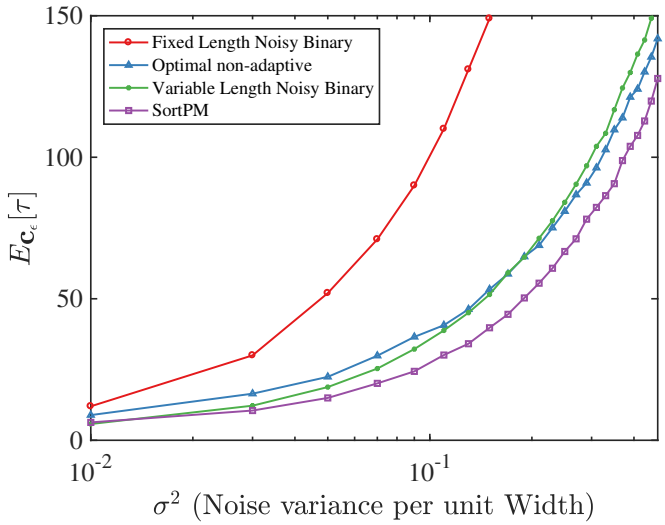


Figure 2.5: $\mathbb{E}_{\mathfrak{c}_\varepsilon}[\tau]$ with $\varepsilon = 10^{-4}$, $B = 16$, and $\delta = 1$, as a function of σ^2 for various strategies.

2.6.2 Two Distinct Regimes of Operation

In this section, for a fixed σ^2 we are interested in the expected number of measurements required $\mathbb{E}_{\mathfrak{c}_\varepsilon}[\tau]$ by an ε-reliable strategy $\mathfrak{c}_\varepsilon$, in the following two regimes: (1) varying δ while keeping B fixed, and (2) varying B while keeping δ fixed. Figures 2.6 and 2.7 show the simulation results of $\mathbb{E}_{\mathfrak{c}_\varepsilon}[\tau]$ as a function of width B, i.e. regime (1) and resolution δ, i.e. regime (2), respectively. The empirical performance is studied for the following: for the fixed composition non adaptive strategy $\mathfrak{c}_\varepsilon \in \mathcal{C}_\varepsilon^{NA}$, for the sort PM adaptive strategy $\mathfrak{c}_\varepsilon \in \mathcal{C}_\varepsilon^{A}$ and its respective upper bound (obtained from the analysis of [10]), for our proposed two-stage strategy along with

dominant terms of the lower bound of Lemma 1, and the upper bound of Lemma 2.. In both regimes, we observe better performance using the sortPM strategy over our two-stage strategy. However, the upper bound of the sortPM strategy is extremely loose and fails to guarantee any adaptivity gain. In fact, in Figure 2.7 the sortPM upper bound is approximately 4 times larger in $\mathbb{E}_{c_\varepsilon}[\tau]$ than the sortPM strategy. On the other hand, under both regimes of operation, our tighter bounds empirically show positive adaptivity gain, albeit in distinctly different manners for each regime. For both regimes, we see that the adaptivity gain grows as the total number of locations increases; however in distinctly different manner as seen in Corollary 3.

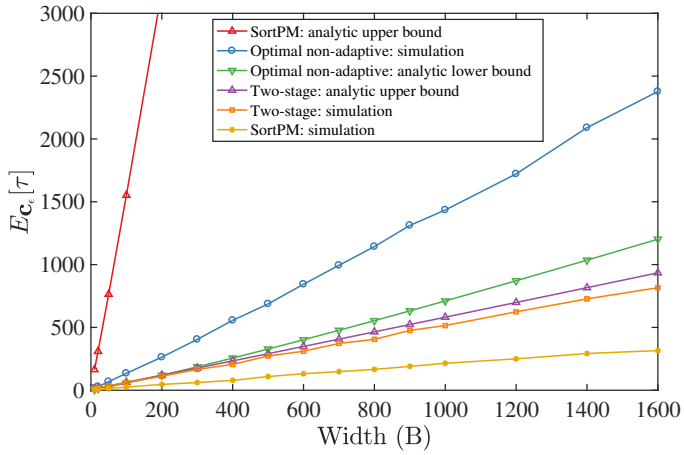


Figure 2.6: $\mathbb{E}_{c_\varepsilon}[\tau]$ with $\varepsilon = 10^{-4}$, $\sigma^2 = 0.05$, and $\delta = 1$, as a function of B.

2.6.3 Relating the Regimes of Operation to Capacity

In this section, we attempt to relate these two regimes of operation to the manner in which the capacity of a BAWGN channel varies. Let noise parameter $Z_n \sim \mathcal{N}(0, 2q\sigma^2_{\text{Total}})$, where $q = \frac{|\mathbf{S}_n|\delta}{B}$ is the fraction of the search region measured and $\sigma^2_{\text{Total}} = \frac{B\sigma^2}{2}$ is the half bandwidth variance. Figure 2.9 shows the effects of the half bandwidth variance on the capacity of a search as a function of q. Intuitively, the target acquisition rate of the adaptive strategy relates to the time spent searching sets of size q as q varies from $\frac{1}{2}$ to $\frac{\delta}{B}$. This means for sufficiently small

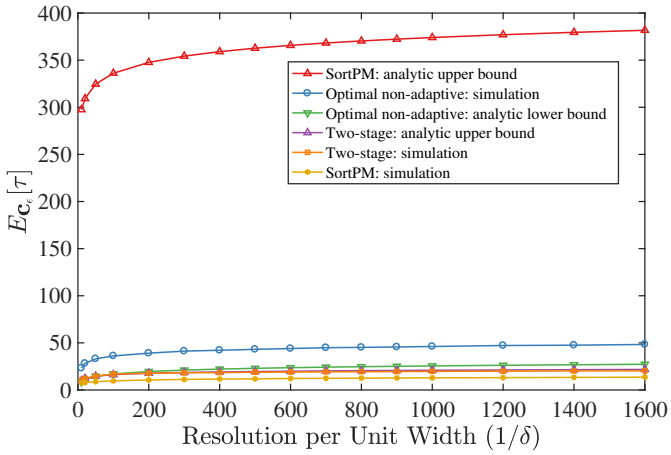


Figure 2.7: $\mathbb{E}_{\mathbf{c}_\epsilon}[\tau]$ with $\varepsilon = 10^{-4}$, $\sigma^2 = 1$ and $B = 1$, as a function of δ.

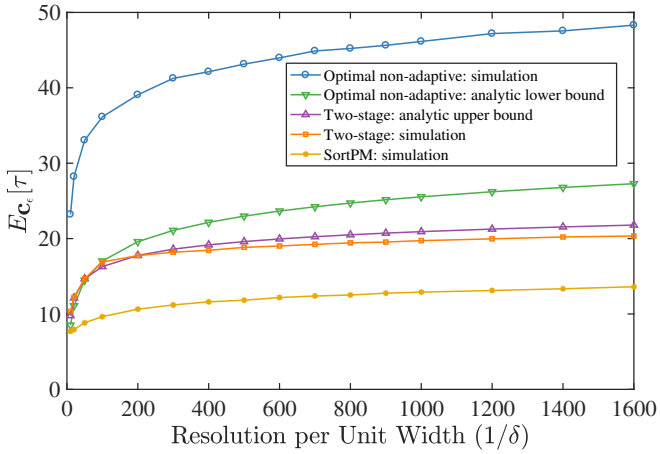


Figure 2.8: Close up of $\mathbb{E}_{\mathbf{c}_\epsilon}[\tau]$ with $\varepsilon = 10^{-4}$, $\sigma^2 = 1$ and $B = 1$, as a function of δ.

σ^2_{Total} (≤ 0.025 in this example), the adaptivity gain is negligible since $C(\frac{1}{2}, 2q\sigma^2_{\text{Total}})$ is about 1 for all q. For medium range σ^2_{Total} (for e.g., 0.25 in this example), the adaptivity effects the target acquisition rate from $C(\frac{1}{2}, 2q^*\sigma^2_{\text{Total}})$ to $C(\frac{1}{2}, 2\frac{\delta}{B}\sigma^2_{\text{Total}})$. When σ^2_{Total} grows significantly, however, the capacity drops rather quickly to zero, forcing the non-adaptive strategies to operate close to exhaustive search, whose measurement time increases linearly in $\frac{B}{\delta}$. This is the regime with most significant adaptivity gain as predicted by Corollary 3.

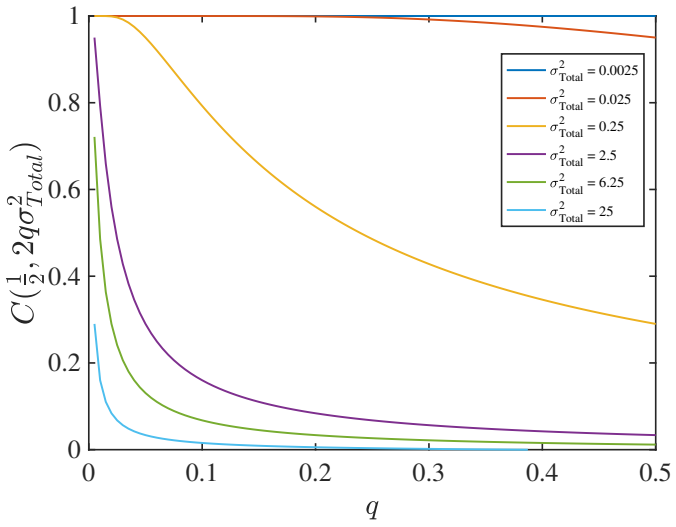


Figure 2.9: For arbitrary B and δ, and with $\varepsilon = 10^{-4}$, $C(\frac{1}{2}, 2q\sigma^2_{Total})$ as a function of q for different values of total noise variance (σ^2_{Total})

2.6.4 Beyond a Linear Noise Model

In this section, we analyze $\mathbb{E}_{c_\varepsilon}[\tau]$ under a general noise model, as presented in section (V-A). Recall, $Y_n \sim \mathcal{N}(X_n, f(|\mathbf{S}_n|)\delta\sigma^2)$, where f is a non-decreasing function of the measurement vector $|\mathbf{S}_n|$. Figure 2.4 shows that the behavior of the capacity range of a search with fixed parameters B, δ, $\mathbf{S}_n$ can be significantly affected by the function $f(\cdot)$. Let us consider the noise function $f(\cdot)$ to be of the form $|\mathbf{S}_n|^\gamma$. Figure 2.10 shows the plot of dominant terms of the lower bound of Lemma 1, and the upper bound of Lemma 2 as a function of σ^2 for the values of $\gamma \in \{0.5, 1, 2\}$. The adaptivity gain is clearly more significant for larger values of gamma and hence, validates the need for generalizing the noise function.

2.7 Conclusion and Future Work

We considered the problem of searching for a target's unknown location under measurement dependent Gaussian noise. We showed that this problem is equivalent to channel coding over a BAWGN channel with state and feedback. We used this connection to utilize feedback code

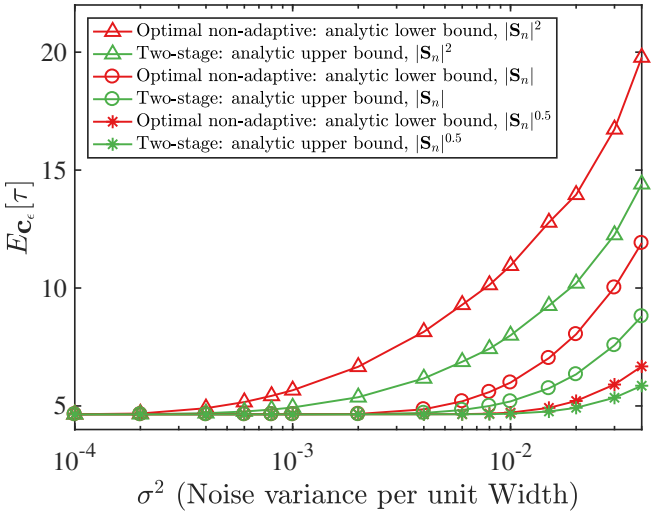


Figure 2.10: $\mathbb{E}_{\mathbf{c}_\varepsilon}[\tau]$ with $\varepsilon = 10^{-4}$, $\sigma^2 = 0.25$ and $B = 25$, $\delta = 1$, as a function of γ when $Z_n \sim \mathcal{N}(0, |\mathbf{S}_n|^\gamma \delta \sigma^2)$.

based adaptive search strategies. We obtained information theoretic converses to characterize the fundamental limits on the target acquisition rate under both adaptive and non-adaptive strategies. As a corollary, we obtained a lower bound on the adaptivity gain. We identified two asymptotic regimes with practical applications where our analysis shows that adaptive strategies are far more critical when either noise intensity or the total search width is large. In contrast, in scenarios where neither the total width nor noise intensity is large, non-adaptive strategies might perform quite well. The immediate step is the extension of this work to a model with $r > 1$ target locations, where the problem has been shown to be equivalent to MAC encoding with feedback [28].

2.8 Appendix

2.8.1 Proof of Lemma 1

Applying Fano's inequality [27] to any non-adaptive search strategy that locates the target among $\frac{B}{\delta}$ locations with $P_e \leq \varepsilon$, we have

$$\begin{aligned}
\log\left(\frac{B}{\delta}\right) &\overset{(a)}{\leq} \frac{1}{1-\varepsilon}\sup_{X^n}\sum_{i=1}^{n} I(X_i,Y_i) + \frac{h(\varepsilon)}{1-\varepsilon} \\
&\overset{(b)}{\leq} \frac{1}{1-\varepsilon}\sum_{i=1}^{n} C\left(q_i, q_i B\sigma^2\right) + \frac{h(\varepsilon)}{1-\varepsilon} \\
&\leq \frac{n}{1-\varepsilon}\max_{q\in \mathbf{I}_{\frac{B}{\delta}}} C(q, qB\sigma^2) + \frac{h(\varepsilon)}{1-\varepsilon},
\end{aligned} \tag{2.32}$$

where (a) follows from the fact that X_i and Z_i for $i = 1,2,\ldots,n$ are independent over time and independent of past observations due to the non-adaptive nature of the search strategy. Since $X_i \sim \mathsf{Bern}(q_i)$ and $Z_i \sim \mathcal{N}(0, q_i B\sigma^2)$, (b) follows from the fact that $I(X_i,Y_i) = C\left(q_i, q_i B\sigma^2\right)$. Rearranging the above equation, we have the assertion of the lemma.

2.8.2 Proof of Lemma 2

Before we provide the proof of Lemma 2, we define quantities required in the proof. For any $q \in \mathsf{I}_{\frac{B}{\delta}}$ and under any measurement vector $\mathbf{S}_n \in \mathcal{U}_{\frac{B}{\delta}}$ such that $|\mathbf{S}_n| = \frac{qB}{\delta}$ we have the following

$$\left|\log\frac{\mathsf{P}(y|\mathbf{S}_n, \mathbf{W}(i)=1)}{\mathsf{P}(y|\mathbf{S}_n, \mathbf{W}(j)=1)}\right| = \begin{cases} 0 & \text{if } \mathbf{S}_n(i)=1 \text{ and } \mathbf{S}_n(j)=1, \\ 0 & \text{if } \mathbf{S}_n(i)\neq 1 \text{ and } \mathbf{S}_n(j)\neq 1, \\ \left|\frac{2y-1}{2qB\sigma^2}\right| & \text{Otherwise.} \end{cases} \tag{2.33}$$

Hence, we have

$$\max_{i,j\in[\frac{B}{\delta}]}\max_{\mathbf{S}_n\in\mathcal{U}_{\frac{B}{\delta}}}\int_{-\infty}^{\infty}\mathsf{P}(y|\mathbf{S}_n,\mathbf{W}(i)=1)\left|\log\frac{\mathsf{P}(y|\mathbf{S}_n,\mathbf{W}(i)=1)}{\mathsf{P}(y|\mathbf{S}_n,\mathbf{W}(j)=1)}\right|^{1+\gamma}dy$$
$$=\max_{q\in\mathsf{I}_{\frac{B}{\delta}}}\left\{\int_{-\infty}^{\infty}\frac{e^{-\frac{y^2}{2qB\sigma^2}}}{\sqrt{2\pi qB\sigma^2}}\left|\frac{2y-1}{2qB\sigma^2}\right|^{1+\gamma}dy\right\}. \tag{2.34}$$

Therefore, there exists $\xi_{\frac{B}{\delta}}<\infty$ and $\gamma>0$ such that

$$\max_{i,j\in[\frac{B}{\delta}]}\max_{\mathbf{S}_n\in\mathcal{U}_{\frac{B}{\delta}}}\int_{-\infty}^{\infty}\mathsf{P}(y|\mathbf{S}_n,\mathbf{W}(i)=1)\left|\log\frac{\mathsf{P}(y|\mathbf{S}_n,\mathbf{W}(i)=1)}{\mathsf{P}(y|\mathbf{S}_n,\mathbf{W}(j)=1)}\right|^{1+\gamma}dy\leq\xi_{\frac{B}{\delta}}. \tag{2.35}$$

Define

$$\psi_{\frac{B}{\delta}}(a):=\max_{q\in\mathsf{I}_{\frac{B}{\delta}}}\left\{\int_{-\infty}^{\infty}\frac{e^{-\frac{y^2}{2Bq\sigma^2}}}{\sqrt{2\pi qB\sigma^2}}\left[\frac{2y-1}{2qB\sigma^2}\right]_a dy\right\}, \tag{2.36}$$

and recall that $[g]_a=g$ if $g\geq a$ otherwise $[g]_a=0$. The quantity a controls the maximum jump in the log-likelihood ratio of the Gaussian observations under all possible search sets determined by the values of $q\in\mathsf{I}_{\frac{B}{\delta}}$ and the quantity $\psi_{\frac{B}{\delta}}(a)$ controls the tail probability of log-likelihood ratios whose value is greater than a. Furthermore, we have $\psi_{\frac{B}{\delta}}(a)$ is non-increasing in a, and we have $\psi_{\frac{B}{\delta}}(a)\leq a^{-\gamma}\xi_{\frac{B}{\delta}}$. Therefore, the tail probability goes to 0, i.e, $\psi_{\frac{B}{\delta}}(a)\to 0$ as $a\to\infty$. Now we are ready to provide the proof for Stage I of our two stage strategy.

Stage I

Lemma 3. *Under the fixed composition search strategy while searching over a search region of width B among $\frac{1}{\alpha}$ locations such that $|\mathbf{S}'_n|\alpha=q^*$ for $n\geq 1$, the following holds true for all $n\geq 1$*

$$\mathbb{E}\left[U(\rho'_{n+1})-U(\rho'_n)|\mathcal{F}_n,\mathbf{S}'_n\right]\geq C\left(q^*,q^*B\sigma^2\right), \tag{2.37}$$

where define $U(\rho'_n) := \sum_{i=1}^{\frac{1}{\alpha}} \rho'_n(i) \log \frac{\rho'_n(i)}{1-\rho'_n(i)}$.

Proof. The proof follows closely the proof of inequality (9) in [31]. There are $\frac{1}{\alpha}$ locations of length αB and hence query vector $\mathbf{S}'_n \in \mathcal{U}_{\frac{1}{\alpha}}$. At every time instant under the fixed composition strategy $K^* = |\mathbf{S}_n| = \frac{q^*}{\alpha}$ number of locations are searched. i.e., a region of length q^*B is searched. Let $\mathcal{P}_{K^*}$ denote the collection of all partitions p of search locations 1 to $\frac{1}{\alpha}$ into sets A_n^0 and A_n^1 such that $|A_n^1| = K^*$. The probability of picking a partition $p \in \mathcal{P}_{K^*}$ is $\lambda_p = \binom{\frac{1}{\alpha}}{K^*}^{-1}$. For simplicity of exposition let $M = \frac{1}{\alpha}$. Also, we have $\sum_{p\in\mathcal{P}_{K^*}} \lambda_p \mathbf{1}_{\{i\in A_n^0\}} = \pi_0^* := \frac{M-K^*}{M}$, and $\sum_{p\in\mathcal{P}_{K^*}} \lambda_p \mathbf{1}_{\{i\in A_n^1\}} = \pi_1^* := \frac{K^*}{M}$.

Since a region of q^*B is searched at every time instant, the noise variance is fixed at $q^*B\sigma^2$. Hence, let $\mathsf{P}_k = \mathsf{P}(Y|X=k, |A_n^1| = K^*) = \mathcal{N}(k, q^*B\sigma^2)$ for $k \in \{0,1\}$. Consider

$$
\begin{aligned}
&\mathbb{E}\left[U(\rho'_{n+1}) - U(\rho'_n) | \mathcal{F}_n, \mathbf{S}_n\right] \\
&= \sum_{p\in\mathcal{P}_{K^*}} \lambda_p \sum_{i=1}^{M} \sum_{k=0}^{1} \rho'_n(i) \mathbf{1}_{\{i\in A_n^k\}} D\left(\mathsf{P}_k \middle\| \sum_{j\neq i} \sum_{l=1}^{1} \frac{\rho'_n(j)}{1-\rho'_n(i)} \mathbf{1}_{\{i\in A_n^l\}} \mathsf{P}_l\right) \\
&= \sum_{i=1}^{M} \rho'_n(i) \sum_{k=0}^{1} \pi_k^* \sum_{p\in\mathcal{P}_{K^*}} \frac{\lambda_p}{\pi_k^*} \mathbf{1}_{\{i\in A_n^k\}} D\left(\mathsf{P}_k \middle\| \sum_{j\neq i} \sum_{l=1}^{1} \frac{\rho'_n(j)}{1-\rho'_n(i)} \mathbf{1}_{\{i\in A_n^l\}} \mathsf{P}_l\right) \\
&\overset{(a)}{\geq} \sum_{i=1}^{M} \rho'_n(i) \sum_{k=0}^{1} \pi_k^* D\left(\mathsf{P}_k \middle\| \sum_{j\neq i} \sum_{l=1}^{1} \frac{\rho'_n(j)}{1-\rho'_n(i)} \sum_{p\in\mathcal{P}_{K^*}} \frac{\lambda_p}{\pi_k^*} \mathbf{1}_{\{i\in A_t^k\}} \mathbf{1}_{\{i\in A_t^l\}} \mathsf{P}_l\right) \\
&\overset{(b)}{=} \sum_{i=1}^{M} \rho'_n(i) \left(\pi_1^* D\left(\mathsf{P}_1 \middle\| \frac{K^*-1}{M-1}\mathsf{P}_1 + \frac{M-K^*}{M-1} P_0\right) \pi_0^* D\left(\mathsf{P}_0 \middle\| \frac{M-K^*-1}{M-1}\mathsf{P}_0 + \frac{K^*}{M-1} P_1\right)\right] \\
&\geq \sum_{i=1}^{M} \rho'_n(i) \left(\pi_1^* D\left(\mathsf{P}_1 \middle\| \frac{K^*}{M}\mathsf{P}_1 + \frac{M-K^*}{M} P_0\right) \pi_0^* D\left(\mathsf{P}_0 \middle\| \frac{M-K^*}{M}\mathsf{P}_0 + \frac{K^*}{M} P_1\right)\right) \\
&\overset{(c)}{=} C\left(q^*, q^*B\sigma^2\right),
\end{aligned}
$$

where (a) follows from Jensen's inequality (b) follows from the definition of π_0^*, π_1^* and

$$\sum_{p\in\mathcal{P}_{K^*}} \lambda_p \mathbf{1}_{\{i\in A_n^0\}}\mathbf{1}_{\{j\in A_n^1\}} = \sum_{p\in\mathcal{P}_{K^*}} \lambda_p \mathbf{1}_{\{i\in A_n^1\}}\mathbf{1}_{\{j\in A_n^0\}} = \frac{K^*(M-K^*)}{M(M-1)},$$
$$\sum_{p\in\mathcal{P}_{K^*}} \lambda_p \mathbf{1}_{\{i\in A_n^0\}}\mathbf{1}_{\{j\in A_n^0\}} = \frac{\pi_0^*(M-K^*-1)}{M-1},$$
$$\sum_{p\in\mathcal{P}_{K^*}} \lambda_p \mathbf{1}_{\{i\in A_n^1\}}\mathbf{1}_{\{j\in A_n^1\}} = \frac{\pi_1^*(K^*-1)}{M-1},$$

and (c) is the definition of non-adaptive BAWGN channel capacity with input distribution $\mathrm{Ber}(q^*)$ and noise variance $q^*B\sigma^2$. □

Lemma 4. *Under the fixed composition search strategy while searching over a search region of width B among $\frac{1}{\alpha}$ locations such that $|\mathbf{S}_n'|\alpha = q^*$ for $n \geq 1$, the following holds true for the expected number of queries while searching with $\mathsf{P}_e \leq \frac{\varepsilon}{2}$*

$$\mathbb{E}_{\mathfrak{c}_\varepsilon^1}[\tau^1] \leq \frac{\log\frac{1}{\alpha} + \log\frac{2}{\varepsilon} + \log\log\frac{B}{\delta} + a_\eta}{C(q^*, q^*B\sigma^2) - \eta}. \tag{2.38}$$

Proof is similar to the proof of Lemma 6.

Stage II

Note that BAWGN capacity for all $q \in \mathsf{I}_{\frac{B}{\delta}}$ with capacity achieving input is

$$\begin{aligned} C\left(\frac{1}{2}, qB\sigma^2\right) &= D\left(\mathcal{N}(0,qB\sigma^2)\middle\|\frac{1}{2}\mathcal{N}(0,qB\sigma^2)+\frac{1}{2}\mathcal{N}(1,qB\sigma^2)\right) \\ &= D\left(\mathcal{N}(1,qB\sigma^2)\middle\|\frac{1}{2}\mathcal{N}(0,qB\sigma^2)+\frac{1}{2}\mathcal{N}(1,qB\sigma^2)\right). \end{aligned} \tag{2.39}$$

Hence, the following Lemma follows from Proposition 3 in [32].

Lemma 5. *Under the sortPM search strategy while searching over a search region of width αB*

among $\frac{\alpha B}{\delta}$ *locations, the following holds true for all* $n \geq 1$

$$\mathbb{E}\left[U(\rho''_{n+1}) - U(\rho''_n)|\mathcal{F}_n, \mathbf{S}_n\right] \geq C\left(\frac{1}{2}, \frac{\alpha B\sigma^2}{2}\right), \tag{2.40}$$

where define $U(\rho''_n) := \sum_{i=1}^{\frac{\alpha B}{\delta}} \rho''_n(i) \log \frac{\rho''_n(i)}{1-\rho''_n(i)}$.

Lemma 6. *Under the sortPM search strategy, the following holds true for the expected number of queries while searching over the search width* αB *among* $\frac{\alpha B}{\delta}$ *locations with* $\mathsf{P}_e \leq \frac{\varepsilon}{2}$

$$\mathbb{E}_{\zeta_\varepsilon^2}[\tau^2] \leq \frac{\log\frac{\alpha B}{\delta} + \log\frac{2}{\varepsilon} + \log\log\frac{\alpha B}{\delta} + a_\eta}{C\left(\frac{1}{2}, \frac{\alpha B\sigma^2}{2}\right) - \eta}, \tag{2.41}$$

where a_η *is the solution of the following equation*

$$\eta = \frac{a}{a-3}\psi_{\frac{B}{\delta}}(a-3). \tag{2.42}$$

Proof. Fix some $a > 0$ to be chosen later. Let $M = \frac{\alpha B}{\delta}$. Let $\tilde{\rho}' = 1 - \frac{1}{1+\max\{\log M, \frac{2}{\varepsilon}\}}$. Now, define $U'(\rho''_0) = U(\rho''_0) - \log\frac{\tilde{\rho}'}{1-\tilde{\rho}'}$ and define $U'(\rho''_n)$ as follows: if $U'(\rho''_n) < 0$, then

$$U'(\rho''_{n+1}) = \begin{cases} U(\rho''_{n+1}) - U(\rho''_n) + U'(\rho''_n) \\ \text{if } U(\rho''_{n+1}) - U(\rho''_n) < a - U'(\rho''_n), \\ a \\ \text{if } U(\rho''_{n+1}) - U(\rho''_n) \geq a - U'(\rho''_n), \end{cases} \tag{2.43}$$

and if $U'(\rho''_n) \geq 0$, then

$$U'(\rho''_{n+1}) = \begin{cases} U(\rho''_{n+1}) - U(\rho''_n) + U'(\rho''_n) \\ \text{if } U(\rho''_{n+1}) - U(\rho''_n) < a, \\ a + U'(\rho''_n) \\ \text{if } U(\rho''_{n+1}) - U(\rho''_n) \geq a. \end{cases} \tag{2.44}$$

By induction we can show that

$$\log \frac{\tilde{\rho}'}{1-\tilde{\rho}'} \leq U(\rho''_n) - U'(\rho''_n). \tag{2.45}$$

We have

$$\begin{aligned} \mathbb{E}\left[U'(\rho''_{n+1}) - U'(\rho''_n)|\mathcal{F}_n\right] &= \mathbb{E}\left[U(\rho''_{n+1}) - U(\rho''_n)|\mathcal{F}_n\right] \\ &\quad + \mathbb{E}\left[\left[-b - U(\rho''_{n+1}) + U(\rho''_n) - U'(\rho''_n)\mathbf{1}_{\{U'(\rho''_n)<0\}}\right]^+ |\mathcal{F}_n\right] \\ &\overset{(a)}{\geq} \mathbb{E}\left[U(\rho''_{n+1}) - U(\rho''_n)|\mathcal{F}_n\right] - \frac{a}{a-3}\psi_{\frac{B}{\delta}}(a-3) \\ &\overset{(b)}{\geq} C\left(\frac{1}{2}, \frac{\alpha B\sigma^2}{2}\right) - \frac{a}{a-3}\psi_{\frac{B}{\delta}}(a-3), \end{aligned} \tag{2.46}$$

where (a) follows from [33] equation (4.140) and (b) follows Lemma 5. Let $\tau' = \min\{n : U'(\rho''_n) \geq 0\}$ and $\tau_{\frac{\varepsilon}{\varepsilon}} = \min\{n : U(\rho''_n) \geq \log \frac{\tilde{\rho}}{1-\tilde{\rho}}\}$ where $\tilde{\rho} = 1 - \frac{2}{\varepsilon}$. From equation (2.45) and since $\tilde{\rho}' > \tilde{\rho}$, we have

$$\mathbb{E}_{\zeta^2_\varepsilon}[\tau_{\frac{\varepsilon}{2}}] \leq \mathbb{E}_{\zeta^2_\varepsilon}|\tilde{\tau}']. \tag{2.47}$$

The sequence $\frac{U'(\rho''_n)}{C\left(\frac{1}{2}, \frac{\alpha B\sigma^2}{2}\right) - \frac{a}{a-3}\psi_{\frac{B}{\delta}}(a-3)} - n$ forms a submartingale with respect to filtration $\mathcal{F}_n$. Now

by Doob's Stopping Theorem we have

$$\frac{U'(\rho_0'')}{C\left(\frac{1}{2},\frac{\alpha B\sigma^2}{2}\right)-\frac{a}{a-3}\psi_{\frac{B}{\delta}}(a-3)} \leq \mathbb{E}\left[\frac{U'(\rho_{\tilde{\tau}'}'')}{C\left(\frac{1}{2},\frac{\alpha B\sigma^2}{2}\right)-\frac{a}{a-3}\psi_{\frac{B}{\delta}}(a-3)}-\tilde{\tau}'\right]. \tag{2.48}$$

Hence, we have

$$\begin{aligned}
\mathbb{E}_{\zeta_\varepsilon^2}[\tilde{\tau}'] &\leq \frac{-U'(\rho_0'')+\mathbb{E}[U'(\rho_{\tilde{\tau}'}'')]}{C\left(\frac{1}{2},\frac{\alpha B\sigma^2}{2}\right)-\frac{a}{a-3}\psi_{\frac{B}{\delta}}(a-3)} \\
&= \frac{-U(\rho_0'')+\log\frac{\tilde{\rho}'}{1-\tilde{\rho}'}+\mathbb{E}[U'(\rho_{\tilde{\tau}'}'')]}{C\left(\frac{1}{2},\frac{\alpha B\sigma^2}{2}\right)-\frac{a}{a-3}\psi_{\frac{B}{\delta}}(a-3)} \\
&\overset{(a)}{\leq} \frac{\log\frac{\alpha B}{\delta}+\log\log\frac{\alpha B}{\delta}+\log\frac{2}{\varepsilon}+\mathbb{E}[U'(\rho_{\tilde{\tau}'}'')]}{C\left(\frac{1}{2},\frac{\alpha B\sigma^2}{2}\right)-\frac{a}{a-3}\psi_{\frac{B}{\delta}}(a-3)} \\
&\overset{(b)}{\leq} \frac{\log\frac{\alpha B}{\delta}+\log\log\frac{\alpha B}{\delta}+\log\frac{2}{\varepsilon}+a}{C\left(\frac{1}{2},\frac{\alpha B\sigma^2}{2}\right)-\frac{a}{a-3}\psi_{\frac{B}{\delta}}(a-3)},
\end{aligned} \tag{2.49}$$

where (a) follows from the fact that $U(\rho_0'') = -\log(\frac{B}{\delta}-1)$ and (b) follows from the fact that for all $n < \tau'$, $U'(\rho_n'') < 0$ and hence from equation (2.43) we have $U'(\rho_{\tilde{\tau}'}'') < a$. Let $\eta > 0$ such that $\eta \ll C\left(\frac{1}{2},\frac{\alpha B\sigma^2}{2}\right)$. Choose a to be the solution of the following

$$\eta = \frac{a}{a-3}\psi_{\frac{B}{\delta}}(a-3), \tag{2.50}$$

i.e., choose $a = a_\eta$. We have the assertion of the lemma by combining above equation with equations (2.47) and (2.49). Note that we control the maximum jump in the one-step evolution of average $U(\rho_n'')$ by truncating the log likelihood ratio of the Gaussian observations under all possible search sets by the term a_η. However, truncating the log-likelihood results in a cutback in our capacity by an amount η, i.e., we obtain $C(\frac{1}{2},\frac{\alpha B\sigma^2}{2})-\eta$. □

2.8.3 Proof of Corollary 2

We make the following sub-optimal choices of a_η and $\alpha(\frac{B}{\delta})$ to obtain asymptotic bounds. Choose $a_\eta = \log\log\frac{B}{\delta}$ so that η goes to zero as $\frac{B}{\delta} \to \infty$, and choose $\alpha(\frac{B}{\delta}) = \frac{1}{\log\frac{B}{\delta}}$. Note that $\alpha(\frac{B}{\delta})$ goes to 0 slower than δ goes to 0. Combining this with Theorem 1 and using the fact $\lim_{\delta\to 0} C\left(\frac{1}{2}, \frac{1}{2}\alpha\left(\frac{B}{\delta}\right)B\sigma^2\right) = 1$, we have equation (2.20). Similarly, note that $\alpha(\frac{B}{\delta})$ goes to 0 slower than B goes to ∞. Using loose approximations $C(q^*, q^*B\sigma^2) \leq \frac{\log e}{B\sigma^2}$ and $C\left(\frac{1}{2}, \alpha(\frac{B}{\delta})B\sigma^2\right) \geq \frac{\log(\frac{B}{\delta})}{16B\sigma^2}\left(1 - \frac{\log(\frac{B}{\delta})}{16B\sigma^2}\right)$ with Theorem 1 we have equations (2.22–2.21).

Chapter 2, in full, is a reprint of the material as it appears in the paper: Anusha Lalitha, Nancy Ronquillo and Tara Javidi, “Improved Target Acquisition Rates With Feedback Codes", in *IEEE Journal of Selected Topics in Signal Processing*, vol. 12, no. 5, pp. 871-885, Oct. 2018. The dissertation author was the primary investigator and author of this paper.

Chapter 3

Almost-Fixed-Length Strategies for Channel Coding and Hypothesis Testing

3.1 Introduction

Channel coding and statistical hypothesis testing which are at the core of information theory and statistics have been traditionally studied under two settings, namely under the *average-length constraint* and the *fixed-length constraint*. Under the average-length constraint, strategies are designed such that their expected stopping time is bounded whereas under the fixed-length constraint, strategies are designed such that their stopping time is strictly bounded. Furthermore, the class of strategies which satisfy an average-length constraint provide various appealing improvements compared to those that satisfy fixed-length constraint. For instance, channel coding strategies under average-length constraint significantly improve the rate-reliability trade-off [5] and similarly, hypothesis testing under average-length constraint improve the type-I and type-II error exponent trade-off [34], over the fixed-length strategies. However, in many practical applications using strategies which satisfy only average-length constraint has major limitations since it does not prohibit stopping time from being occasionally very long. Moreover, unlike the

fixed-length constraint, an average-length constraint does not limit the variability of the stopping time around the average value. This suggests that allowing some variability in the stopping time is essential for achieving better error exponents. The main contribution of this chapter is to demonstrate that this flexibility need not be significant.

This chapter introduces a new class of coding and hypothesis testing strategies referred to as *almost-fixed-length channel codes* and *almost-fixed-length hypothesis tests* in which the stopping time is kept fixed ($\leq n$) for almost all sample paths except for an exponentially rare set for which the stopping time is allowed exceed n but remains bounded by Kn where K is a constant. More specifically, the probability of stopping time exceeding n approaches zero exponentially fast with an exponent $\gamma > 0$. In other words, variance of the stopping time of almost-fixed-length tests approaches zero as n grows. We also note that the proposed class of strategies does not require a full sequential computation, and hence, is not as computationally cumbersome as the variable-length strategies under average-length constraint. In this chapter, we show that it is possible to achieve optimal performance of variable-length strategies using almost-fixed-length strategies. Hence, neither growing variability nor the computational complexity are essential to obtaining the optimal error exponents. The performance achieved by the two settings, namely average-length constraint and fixed-length constraint, that were thought to be very distinct are in fact the extremities of a continuum of performance curves achieved by almost-fixed-length strategies parametrized by γ.

It is known that feedback can significantly improve the rate-reliability trade-off of fixed length feedback codes provided that variable-length codes are allowed. Burnashev [5] established that the error exponent improves in this setting and the reliability function for any rate R below capacity is given by

$$E(R) = C_1 \left(1 - \frac{R}{C}\right), \tag{3.1}$$

where C denotes the capacity of the DMC and C_1 denotes the maximal KL-divergence between the conditional output distributions given any two inputs. On the other hand, the fixed-length codes are bounded above and below by the Haroutunian bound established in [35] and the random coding exponent demonstrated in [36], which are both strictly less than Burnashev's optimal reliability function. The following example shows that significant gap in the performance of the two settings.

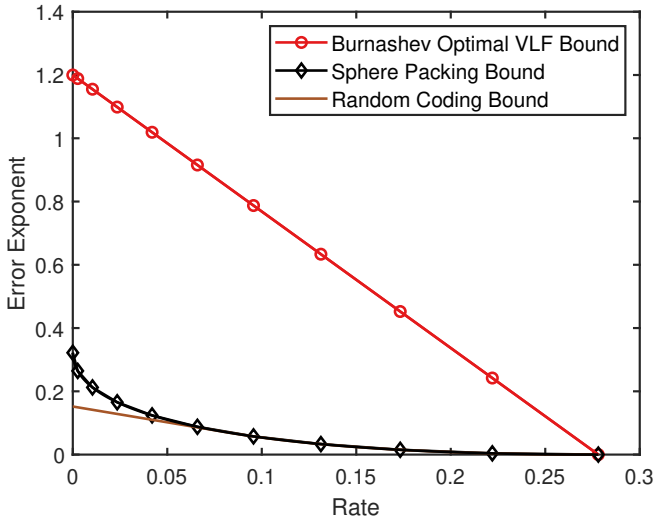


Figure 3.1: Figure shows the optimal error exponents of variable-length feedback codes shown by Burnashev along with upper bound and lower bound on fixed-length feedback codes i.e., sphere packing bound and random coding bound for a BSC with cross-over probability $p = 0.2$.

Example 4. Consider a Binary Symmetric Channel (BSC) with cross-over probability $p = 0.2$. Figure 3.1 shows Burnashev's optimal reliability function, the random coding bound which is a lower bound for fixed-length codes and the sphere packing bound which is an upper bound for fixed-length codes[1]. We can see that the variable-length codes significantly improve the error exponents achieved by the class of fixed-length codes even in the presence of feedback. We shall return to this example to illustrate how one can go from the fixed-length error exponent curve to Burnashev's optimal error exponent curve in an almost-fixed-length manner.

[1] Haroutunian bound coincides with the sphere packing bound in the case of a BSC.

There is a large body of literature on the asymptotic analysis of type-I and type-II errors as the (expected) number of samples n grows large. More specifically, the error exponents in both variants of hypothesis testing is well-known and understood [27, 34, 37–39]. In the fixed-length regime, the error exponents of the two types of errors can only be traded-off against each other, the sequential hypothesis tests can achieve both exponents simultaneously. In other words, by allowing the stopping time to be a random number with bounded expected value, the sequential hypothesis test resolves the trade-off between error-types.

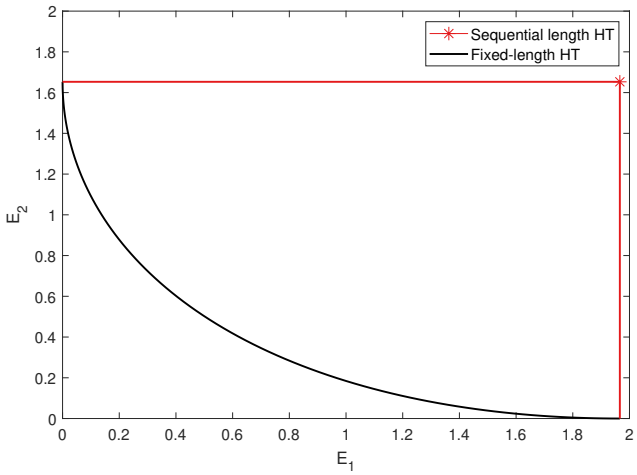


Figure 3.2: Figure shows the optimal error exponents of fixed-length hypothesis test and sequential hypothesis test for Bernoulli samples with parameters given by $p_1 = 0.9$ under H_1 and $p_2 = 0.2$ under H_2.

Example 5. Consider $H_1 : X \sim \mathsf{Bern}(0.9)$ and $H_2 : X \sim \mathsf{Bern}(0.2)$. Figure 3.2 shows the optimal error exponents in both fixed-length and sequential setting. We can see that the sequential hypothesis test provides a significant improvement over the fixed-length hypothesis testing. We shall return to this example to illustrate how one can go from the fixed-length curve to the sequential curve.

3.1.1 Related Work

Channel codes with probabilistic delay constraints were considered by Altug et al. in [40], where they show that if the constraint on expected stopping time is replaced by a probabilistic one then the first order gain in rate ceases to exist. However, these works fail to notice the improvement in the error exponent achieved by codes under probabilistic delay constraints such as the class of almost-fixed-length codes over the class of fixed-length codes. Our results show that it is possible to achieve Burnashev's optimal error exponent with finite and bounded block-length as γ approaches zero. As shown in Figure 3.1, almost-fixed-length codes indeed bridge a significant gap between in an almost-fixed-length manner.

For achievability, we propose a simple construction for a two-phase almost-fixed-length feedback channel code which builds upon fixed-length feedback codes with error-erasure decoder considered by Forney in [41], Telatar and Gallger in [42] and more recently by Nakiboglu and Zheng in [43]. In the first phase we utilize an error-erasure code and whenever an erasure is declared we proceed to the second phase which consists of a fixed-length channel code. Leveraging the bounds on error-erasure exponents, we provide upper and lower bounds on the optimal error exponents achievable in an almost-fixed-length manner.

We propose a simple two-phase hypothesis test using which the overall reliability is increased significantly and the trade-off between type-I and type-II error exponents is relaxed. Our converse proof closely follows a pair of papers by Grigoryan et. al. [44] and Sason [45] on hypothesis testing with rejection.

3.2 Types of Channel Codes

Consider a discrete memoryless channel (DMC) with input alphabet $\mathcal{X}$, output alphabet $\mathcal{Y}$ and a sequence of conditional output distributions $\{\mathsf{P}_{Y_n|X_1^n Y_1^{n-1}}\}_{n\geq 1}^{\infty}$ which satisfy the following

$$\mathsf{P}_{Y_n|X_1^n Y_1^{n-1}}(y_n|x_1^n y_1^{n-1}) = \mathsf{P}_{Y|X}(y_n|x_n) \quad \forall n \in \mathbb{N}. \tag{3.2}$$

We assume that the feedback channel is of infinite capacity, noiseless and delay free i.e., the input of the feedback channel is observed at the transmitter before transmission of X_n at each time $n \in \mathbb{N}$.

3.2.1 Fixed Length Feedback Channel Codes

Definition 8. An (ℓ, M, ε) fixed-length feedback (FLF) code, where $\ell, M \in \mathbb{N}$, and $\varepsilon \in (0,1)$, is defined by:

(i) A common randomness $U \in \mathcal{U}$, with a probability distribution P_U, whose realization is used to initialize the encoder and the decoder before the start of transmission.

(ii) A sequence of encoders $f_n : \mathcal{U} \times \{1,\ldots,M\} \times \mathcal{Y}^{n-1} \to \mathcal{X}$ for $n \in \mathbb{N}$ defining the channel inputs

$$X_n = f_n(U, W, Y^{n-1}), \tag{3.3}$$

where $W \in \{1,\ldots,M\}$ is the equiprobable message.

(iii) A sequence of decoders $g_n : \mathcal{U} \times \mathcal{Y}^n \to \{1,\ldots,M\}$ for $n \in \mathbb{N}$ providing an estimate of W at each time n.

(iv) A stopping time $\tau \in \mathbb{N}$ which satisfies

$$\tau = \ell \quad \text{a.s.} \tag{3.4}$$

The final estimate is computed at time τ and it is given by

$$\hat{W}_\tau := g_\tau(U, Y^\tau) \tag{3.5}$$

such that the error probability satisfies

$$\mathsf{P}(\hat{W}_\tau \neq W) \leq \varepsilon. \tag{3.6}$$

Definition 9. A rate-reliability pair (R, E) is said to be achievable in *fixed-length manner with feedback* if for any $\delta > 0$ there exists an $\ell(\delta) \in \mathbb{N}$ such that for all $\ell \geq \ell(\delta)$ there is a $(\ell, M_\ell, \varepsilon_\ell)$ FLF code which satisfies

$$M_\ell \geq 2^{\ell R}, \tag{3.7}$$

$$\varepsilon_\ell \leq 2^{-\ell(E-\delta)}. \tag{3.8}$$

For a given rate R below capacity, the reliability function $E_{\text{FLF}}(R)$ is defined as the best achievable error exponent at rate R in a fixed-length manner with feedback.

The following fact characterizes an upper bound and lower bound on the optimal achievable reliability in a fixed-length manner with feedback.

Fact 1. *For the class of FLF codes, the optimal achievable reliability as a function of rate R can bounded above and below as follows*

$$E_r(R) \leq E_{FLF}(R) \leq E_H(R), \tag{3.9}$$

where $E_H(R)$ is an upper established by Haroutunian in [35] and $E_r(R)$ is the random coding exponent demonstrated in [36]. The upper bound $E_H(R)$ is strictly larger than the sphere packing exponent $E_{sp}(R)$ (demonstrated in [36]) for the class of non-symmetric channels and coincides with $E_{sp}(R)$ for the class of symmetric channels including the Binary Symmetric Channel (BSC) [46].

3.2.2 Variable Length Feedback Channel Codes

Definition 10. An (ℓ, M, ε) variable-length feedback (VLF) code, where $\ell, M \in \mathbb{N}$, and $\varepsilon \in (0,1)$, is defined similarly to FLF codes with an exception that condition (iv) in Definition 8 is replaced by:

(iv)$'$ A random stopping time $\tau \in \mathbb{N}$ which satisfies

$$\mathbb{E}[\tau] \leq \ell \quad \text{a.s.} \tag{3.10}$$

The final estimate is computed at time τ and it is given by

$$\hat{W}_\tau := g_\tau(U, Y^\tau) \tag{3.11}$$

such that the error probability satisfies

$$\mathsf{P}(\hat{W}_\tau \neq W) \leq \varepsilon. \tag{3.12}$$

Definition 11. A rate-reliability pair (R, E) is said to be achievable in *variable-length manner with feedback* if for any $\delta > 0$ there exists an $\ell(\delta) \in \mathbb{N}$ such that for all $\ell \geq \ell(\delta)$ there is a

$(\ell, M_\ell, \varepsilon_\ell)$ VLF code which satisfies

$$M_\ell \geq 2^{\ell R}, \tag{3.13}$$

$$\varepsilon_\ell \leq 2^{-\ell(E-\delta)}. \tag{3.14}$$

For a given rate R below capacity, the reliability function $E_{\mathrm{VLF}}(R)$ is defined as the best achievable error exponent at rate R in a variable-length manner with feedback.

The following fact characterizes the optimal achievable reliability in a variable-length manner with feedback.

Fact 2. *Burnashev in [5] established that for the class of VLF codes, the optimal achievable reliability as a function of rate R is given by*

$$E_{VLF}(R) = C_1 \left(1 - \frac{R}{C}\right), \tag{3.15}$$

where C denotes the capacity of the DMC and C_1 denotes the maximal KL-divergence between the conditional output distributions given any two inputs i.e.,

$$C_1 := \max_{x,x' \in \mathcal{X}} D(\mathsf{P}_{Y|X}(\cdot|x) || \mathsf{P}_{Y|X}(\cdot|x')). \tag{3.16}$$

3.2.3 Almost Fixed Length Feedback Channel Codes

We introduce a new class of channel codes for which the number of channel uses are bounded but have some variability in terms of stopping time. By construction, this new class of $(\ell, M, \gamma, K, \varepsilon)$ almost-fixed-length channel codes are given an exponentially small flexibility for the stopping time τ to be larger than ℓ, while keeping the maximum length of any test to be bounded by a constant times ℓ.

Definition 12. An $(\ell, M, \gamma, K, \varepsilon)$ almost-fixed-length feedback (AFLF) code, where $\ell, M, K \in \mathbb{N}$,

$\gamma \geq 0$, and $\varepsilon \in (0,1)$, is defined similarly to FLF codes with an exception that condition (iv) in Definition 8 is replaced by:

(iv)$''$ A random stopping time $\tau \in \mathbb{N}$ which satisfies

$$\mathsf{P}(\tau > \ell) \leq 2^{-\gamma \ell}, \tag{3.17}$$

$$\tau \leq K\ell \quad \text{a.s.} \tag{3.18}$$

The final estimate is computed at time τ is given by

$$\hat{W}_\tau := g_\tau(U, Y^\tau) \tag{3.19}$$

such that the error probability satisfies

$$\mathsf{P}(\hat{W}_\tau \neq W) \leq \varepsilon. \tag{3.20}$$

Remark 5. The definition of AFLF code implies that:

(i) Fixed Length Feedback (FLF) codes are a special case of AFLF codes where $\gamma = \infty$ and $K = 1$ and hence $\mathsf{P}(\tau > \ell) = 0$.

(ii) AFLF codes are a special case of Variable Length Feedback (VLF) codes, where the condition on the stopping time

$$\mathbb{E}[\tau] \leq \ell \tag{3.21}$$

is replaced by more stringent conditions given in equations (3.17) and (3.18). In other words, AFLF codes not only require the stopping time τ to be bounded in expectation but also require the probability that τ exceeds ℓ to be exponentially small.

(iii) AFLF codes are a special case of VLF* codes considered by Altug et al. in [40], where the condition on the stopping time

$$\min\{n \in \mathbb{N} : \mathsf{P}(\tau > n) \leq \varepsilon_d\} \leq \ell \tag{3.22}$$

for some $\varepsilon_d \in (0,1)$ is replaced by more stringent conditions given in equations (3.17) and (3.18). In other words, AFLF codes not only require the probability that τ exceeds ℓ to be less than a fixed threshold ε_d but also require the threshold to decay exponentially in ℓ.

Definition 13. A rate-reliability pair (R,E) is said to be achievable in an *almost-fixed-length manner with feedback* if for any $\delta > 0$ there exists an $\ell(\delta) \in \mathbb{N}$ such that for all $\ell \geq \ell(\delta)$ there is a $(\ell, M_\ell, \gamma, K, \varepsilon_\ell)$ AFLF code which satisfies

$$M_\ell \geq 2^{\ell R}, \tag{3.23}$$

$$\varepsilon_\ell \leq 2^{-\ell(E-\delta)}. \tag{3.24}$$

For a given rate R below capacity, the reliability function $E_{\text{AFLF}}(R,\gamma,K)$ is defined as the best achievable error exponent at rate R in an almost-fixed length manner with feedback.

The following fact characterizes an upper bound and lower bound on the optimal achievable reliability in an almost-fixed-length manner with feedback.

Corollary 4. *For the class of AFLF codes, for all $\gamma \geq 0$ and $K \in \mathbb{N}$ the optimal achievable reliability as a function of rate R can be bounded as follows*

$$E_{FLF}(R) = E_{AFLF}(R,\infty,1) \leq E_{AFLF}(R,\gamma,K),$$
$$E_{AFLF}(R,\gamma,K) \leq E_{AFLF}(R,0,\infty) = E_{VLF}(R).$$

Hence, we have

$$E_r(R) \leq E_{AFLF}(R,\infty,1) \leq E_H(R), \tag{3.25}$$

$$E_r(R) \leq E_{AFLF}(R,\gamma,K) \leq C_1\left(1-\frac{R}{C}\right). \tag{3.26}$$

The above corollary is be obtained by combining Remark 5 (i) with Fact 1 and by combining Remark 5 (ii) with Fact 2.

3.3 Rate-Reliability of Almost-Fixed-Length Feedback Codes

3.3.1 Achievability: Construction of Almost-Fixed-Length Codes

In this section we show that AFLF codes can be easily constructed from existing channel coding strategies. More specifically, we build upon error-erasure codes considered by Forney in [41], Telatar and Gallger in [42] and more recently by Nakiboglu and Zheng in [43]. Next we provide some definitions for error-erasure codes.

Definition 14. An $(\ell,M,\varepsilon,\varepsilon_X)$ fixed length feedback code with error-erasure decoding, where $\ell,M \in \mathbb{N}$, and $\varepsilon,\varepsilon_X \in (0,1)$, is defined by:

(i) A common randomness $U \in \mathcal{U}$, with a probability distribution P_U, whose realization is used to initialize the encoder and the decoder before the start of transmission.

(ii) A sequence of encoders $f_n : \mathcal{U} \times \{1,\ldots,M\} \times \mathcal{Y}^{n-1} \to \mathcal{X}$ for $n \in \mathbb{N}$ defining the channel inputs

$$X_n = f_n(U,W,Y^{n-1}), \tag{3.27}$$

where $W \in \{1,\ldots,M\}$ is the equiprobable message.

(iii) A sequence of decoders $g_n : \mathcal{U} \times \mathcal{Y}^n \to \{1,\ldots,M\} \cup \{e\}$ for $n \in \mathbb{N}$ providing the best estimate of W in $\{1,\ldots,M\}$ or declare an erasure e at time n.

(iv) A stopping time $\tau \in \mathbb{N}$ which satisfies

$$\tau = \ell \quad \text{a.s.} \tag{3.28}$$

The final estimate is computed at time τ and it is given by

$$\hat{W}_\tau := g_\tau(U, Y^\tau) \tag{3.29}$$

such that the error probability satisfies

$$\mathsf{P}(\hat{W}_\tau \neq W \mid \hat{W}_\tau \neq e) \leq \varepsilon, \tag{3.30}$$

and the erasure probability satisfies

$$\mathsf{P}(\hat{W}_\tau = e) \leq \varepsilon_X. \tag{3.31}$$

Definition 15. A rate-reliability pair (R, E, E_X) is said to be achievable in *fixed-length manner with feedback under error-erasure decoding* if for any $\delta > 0$ there exists an $\ell(\delta) \in \mathbb{N}$ such that for all $\ell \geq \ell(\delta)$ there is a $(\ell, M_\ell, \varepsilon_\ell, \varepsilon_{\ell,X})$ fixed-length error-erasure code which satisfies

$$M_\ell \geq 2^{\ell R}, \tag{3.32}$$

$$\varepsilon_\ell \leq 2^{-\ell(E-\delta)}, \tag{3.33}$$

$$\varepsilon_{\ell,X} \leq 2^{-\ell(E_X-\delta)}. \tag{3.34}$$

For a given rate R below capacity and $E_X \geq 0$, the error reliability function $E_{\mathrm{ee}}(R, E_X)$ is defined

as the best achievable error exponent at rate R and at erasure exponent E_X in a fixed-length manner with feedback under error-erasure decoding.

Next we provide a simple and intuitive construction of AFLF codes using the class of error-erasure codes and FLF codes.

Proposition 2. *Consider a $(\ell, M, \varepsilon_1, \varepsilon_X)$ fixed-length feedback code under error-erasure decoding and a $((K-1)\ell, M, \varepsilon_2)$ fixed-length feedback code, applied sequentially in a two phase strategy as shown below:*

(i) Phase I: For all time instants $n \leq \ell$, use the encoding functions of $(\ell, M, \varepsilon_1, \varepsilon_X)$ fixed-length feedback code to obtain the next channel input. At $n = \ell$, use the decoding function of the $(\ell, M, \varepsilon_1, \varepsilon_X)$ code to obtain an estimate $\hat{W}_\ell$ of the message. If $\hat{W}_\ell \neq e$ then stop and if $\hat{W}_\ell = e$, then proceed to Phase II.

(ii) Phase II: For all time instants $\ell < n \leq K\ell$, discard the previous channel observations and use the encoding functions of $((K-1)\ell, M, \varepsilon_2)$ FLF code to obtain the next channel input. At $n = K\ell$, use the decoding function of the $((K-1)\ell, M, \varepsilon_2)$ FLF code to obtain an estimate $\hat{W}_{K\ell}$ of the message.

The resulting two-phase strategy is an (ℓ, M) almost-fixed-length feedback code which satisfies

$$\mathsf{P}(\tau > \ell) \leq \varepsilon_X, \tag{3.35}$$

$$\mathsf{P}(\hat{W}_\tau \neq W) \leq \varepsilon_1 + \varepsilon_2. \tag{3.36}$$

We obtain the following corollary as a consequence of the above proposition.

Corollary 5. *For the class of AFLF codes, for all $\gamma \geq 0$ and $K \in \mathbb{N}$, the optimal achievable reliability as a function of rate R can be lower bounded as*

$$E_{AFLF}(R, \gamma, K) \geq \min\left\{ E'_{ee}(R, \gamma), (K-1) E'_{FLF}\left(\frac{R}{(K-1)} \right) \right\}, \tag{3.37}$$

if the error exponent $E'_{ee}(R,\gamma)$ is achievable in a fixed-length manner with feedback under error-erasure decoding where the erasure exponent is γ and if $E'_{FLF}\left(\frac{R}{(K-1)}\right)$ is achievable in a fixed-length manner with feedback.

Next two theorems provide a lower bound to the optimal achievable reliability in an almost-fixed-length manner with feedback.

Theorem 3 (Special Case: $\gamma = 0$)**.** *For the class of AFLF codes, for $\gamma = 0$ and $K \in \mathbb{N}$, the optimal achievable reliability as a function of rate R can be lower bounded as*

$$E_{AFLF}(R,0,K) \geq \min\left\{C_1\left(1-\frac{R}{C}\right),(K-1)E_r\left(\frac{R}{(K-1)}\right)\right\}. \tag{3.38}$$

Furthermore, for all $K \geq 1+\frac{C_1}{E_r(0)}$ we have

$$E_{AFLF}(R,0,K) = C_1\left(1-\frac{R}{C}\right). \tag{3.39}$$

Proof of the above theorem is provided in Appendix 3.7.1 and is based on a truncated version of the Yamamoto-Itoh strategy [47]. Theorem 3 shows that the optimal Burnashev bound $E_{\mathrm{VLF}}(R)$ can be achieved for any rate R below capacity with bounded number of channel uses. The class of VLF codes also achieve the optimal Burnashev bound, however under bounded number of average channel uses, where unlike AFLF codes occasionally very large number of channel uses are required.

Remark 6. Polyanskiy et al. in [48] show that feedback improves the first order rate. However, for channel codes with more stringent constraints on stopping time designed to reduce the variability in number channel uses provide no such improvement even in the presence of feedback. Altug et al. in [40] show that if the constraint on expected number of channel uses in the class of VLF codes is replaced by a probabilistic one, given by equation (3.22), as seen in the class of VLF* codes then the first order gain in rate ceases to exist. However, these works fail to notice the

improvement in the error exponent achieved by codes under probabilistic delay constraints such as the class of AFLF codes over the class of FLF codes. Theorem 3 and 4 show that the class of AFLF codes can improve the error exponent significantly. In particular, it is possible to achieve Burnashev's optimal error exponent with finite and bounded block-length as γ approaches zero.

Remark 7. The strategies considered in Proposition 2 and in the proof of Theorem 3, we use a decoder which discards the past channel outputs in the case of an erasure. Instead the decoder can use all the past channel outputs to jointly decode the transmitted message at $\tau = K\ell$. Such a strategy was considered by Gopala et al. in [49] where the error exponent satisfies

$$E_{\mathrm{AFLF}}(R,0,K) \geq \min\left\{E_f(R), KE_r^*\left(\frac{R}{K}\right)\right\}, \tag{3.40}$$

where E_{r}^* is the error exponent achieved by a random code whose input probability distribution is optimal for Forney's error-erasure decoder considered in [41] where erasure exponent tends to zero and $E_f(R)$ is the Forney's decision feedback exponent defined as the largest possible error exponent while the erasure exponent is positive [41]. For a BSC, if $K \geq \frac{E_f(0)}{E_r(0)}$, where $E_r(R)$ is the BSC random coding exponent, then for all $0 \leq R \leq C$, we have $E_{\mathrm{AFLF}}(R,0,K) \geq E_f(R)$. As seen in Figure 3.3, while joint decoding of channel outputs reduces the number of channel uses in the worst case, it comes at the cost of a smaller error exponent guarantee.

Theorem 4 (Strictly Positive $\gamma > 0$)**.** *For any rate R below capacity, let*

$$\alpha^*(R,\gamma) := \frac{R}{g^{-1}(\frac{\gamma}{R})}, \tag{3.41}$$

where $g(a) = \frac{E_r(a)}{a}$*. For the class of AFLF codes, for* $0 < \gamma < E_r(R)$ *and* $K \in \mathbb{N}$*, the optimal achievable reliability as a function of rate R can be lower bounded as*

$$E_{AFLF}(R,\gamma,K) \geq \min\left\{E'_{ee}(R,\gamma), (K-1)E_r\left(\frac{R}{K-1}\right)\right\}, \tag{3.42}$$

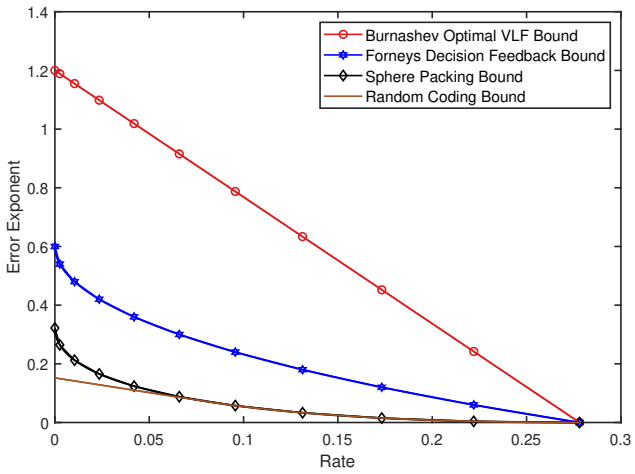


Figure 3.3: Figure shows Forney's decision feedback bound with the optimal error exponents of VLF codes shown by Burnashev and with upper bound and lower bound on FLF codes i.e., sphere packing bound and random coding bound for a BSC with cross-over probability $p = 0.2$.

where we define

$$E'_{ee}(R,\gamma) := \max_{\alpha\in[\alpha^*(R,\gamma),1]} \max_{\substack{\lambda\in[0,1] \\ (1-\alpha)D(\mathsf{P}^{(\lambda)}||\mathsf{P}_{Y|X}(\cdot|x))\geq\gamma}} E''_{ee}(\alpha,\lambda,R), \tag{3.43}$$

and define

$$E''_{ee}(\alpha,\lambda,R) := \alpha E_r\left(\frac{R}{\alpha}\right) + (1-\alpha)\, D(\mathsf{P}^{(\lambda)}||\mathsf{P}_{Y|X}(\cdot|x')), \tag{3.44}$$

where for any $\lambda \in [0,1]$, *the* λ*-tilted distribution* $\mathsf{P}^{(\lambda)}$ *is given by*

$$\mathsf{P}^{(\lambda)}(y) := \frac{\mathsf{P}^{1-\lambda}_{Y|X}(y|x)\mathsf{P}^{\lambda}_{Y|X}(y|x')}{\sum_{a\in\mathcal{Y}}\mathsf{P}^{1-\lambda}_{Y|X}(a|x)\mathsf{P}^{\lambda}_{Y|X}(a|x')}, \quad \forall y \in \mathcal{Y}. \tag{3.45}$$

Additionally, for $\gamma > E_r(R)$ *the lower bound on optimal achievable reliability as a function of rate* R *given by equation* (3.42) *reduces to* $E_{AFLF}(R,\gamma,K) \geq E_r(R)$.

Proof of the above theorem is provided in Appendix 3.7.1 and is based on a truncated

version of the Yamamoto-Itoh strategy [47].

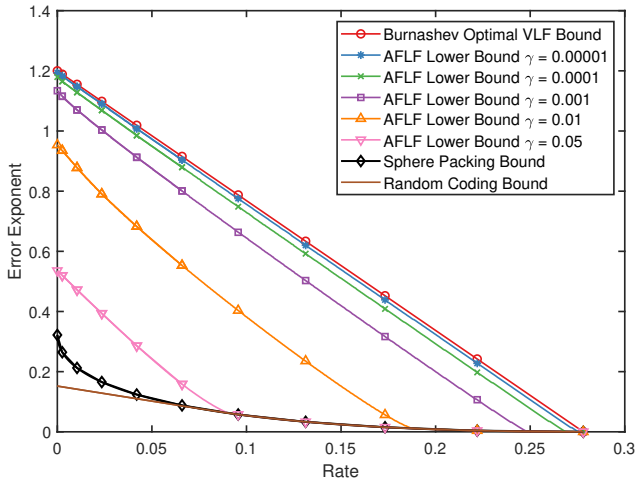


Figure 3.4: Figure shows the error exponents achieved by AFLF codes where $K \geq K^* = 7$ for values of γ approaching zero. The AFLF bounds are compared with the optimal error exponents of VLF codes shown by Burnashev along with upper bound and lower bound on FLF codes i.e., sphere packing bound and random coding bound for a BSC with cross-over probability $p = 0.2$.

Example 1 (Revisited). For setup considered in Example 4, Figure 3.4 shows the lower bound described in Theorem 4 equation (3.42) for the optimal AFLF error exponent $E_{\text{AFLF}}(R, \gamma, K)$. As γ decreases, the rate-reliability curve improves as predicted by the lower bound in equation (3.42). In particular, for any rate R below capacity it is possible to achieve error exponent arbitrarily close to Burnashev's optimal error exponent $E_{\text{VLF}}(R)$ in an almost-fixed-length manner by selecting γ close zero.

3.3.2 Converse for Almost-Fixed-Length Feedback Codes

Our converse bounds the performance of a $(\ell, M, \gamma, K, \varepsilon)$ AFLF code with that of a fixed-length error-erasure where the probability of erasure approaches zero exponentially fast with an exponent at most γ. More specifically, given an $(\ell, M, \gamma, K, \varepsilon)$ AFLF code we can construct a fixed-length error-erasure code by declaring an erasure whenever $\tau > \ell$. Such a code is a $(\ell, M, 2^{-\gamma\ell}, \varepsilon)$ fixed-length feedback code with error-erasure decoder. As a consequence we obtain

the following result.

Proposition 3 (Converse)**.** *For the class of AFLF codes, for $\gamma \geq 0$ and $K \in \mathbb{N}$, the optimal achievable reliability as a function of rate R can be upper bounded as*

$$E_{AFLF}(R,\gamma,K) \leq \min\left\{E_{ee}(R,\gamma), KE_H\left(\frac{R}{K}\right)\right\}, \tag{3.46}$$

where recall that $E_{ee}(R,\gamma)$ is the optimal achievable error exponent under error-erasure decoding where the erasure exponent is γ and $E_H(R)$ denotes the upper bound establised by Haroutunian in [35] for the class of FLF codes.

Proof of the above proposition is provided in the Appendix 3.7.2.

3.4 Types of Hypothesis Tests

In this section, we extend the notion of almost-fixed-length strategies to hypothesis testing and provide matching upper and lower bounds for error exponents achieved by binary hypothesis tests in an almost-fixed-length manner. Consider two hypotheses H_1 and H_2 which correspond to the two possible underlying distributions, P_1 and P_2, governing the samples. In other words, we have

$$H_1 : X \sim \mathsf{P}_1, \quad \text{and} \quad H_2 : X \sim \mathsf{P}_2,$$

where X takes values in a finite set $\mathcal{X}$. Consider collecting τ number of i.i.d samples, where τ is a random stopping time with respect to the underlying filtration given by $\sigma(X_1,\ldots,X_n)$. The expectation under hypothesis H_i, for $i \in \{1,2\}$, is denoted by $\mathbb{E}_i[\cdot]$.

A general *hypothesis test* decides between H_1 and H_2, for any given τ samples by dividing the sample space $\mathcal{X}^\tau$ into two sets or two "decision regions". A decision region, denoted by A_i^τ, is

a collection of samples $X^\tau \in \mathcal{X}^\tau$ for which the test chooses H_i, for $i \in \{1,2\}$. The type-I error is defined as an error event that occurs when the test accepts hypothesis H_2 when hypothesis H_1 is true and its probability is given by $\mathsf{P}_1\left(A_2^\tau\right)$. Similarly, type-II error is defined as an error event when the test accepts hypothesis H_1 when hypothesis H_2 is true and its probability is given by $\mathsf{P}_2\left(A_1^\tau\right)$. It is known that growing the number of samples results in an exponential reduction in these probabilities of error. This fact is characterized by two classical asymptotic results depending on the manner in which τ grows. First, we review two classical regimes of hypothesis tests based on the growth of τ.

3.4.1 Fixed-Length Hypothesis Tests

In this setting τ is assumed to be a bounded integer i.e., it satisfies $\tau \leq n$, where $n \in \mathbb{N}$.

Definition 16. The error exponents (E_1, E_2) are said to be achievable in a *fixed-length* manner, if for every $\delta > 0$ there exists an $N(\delta) \in \mathbb{N}$ such that for all $n \geq N(\delta)$ there exists a hypothesis test satisfying the following constraints

$$\tau \leq n \quad \mathsf{P}_i - \text{a.s. for } i \in \{1,2\}, \tag{3.47}$$

$$\mathsf{P}_1\left(A_2^\tau\right) \leq e^{-(E_1-\delta)n}, \tag{3.48}$$

$$\mathsf{P}_2\left(A_1^\tau\right) \leq e^{-(E_2-\delta)n}. \tag{3.49}$$

Definition 17. For any $\lambda \in [0,1]$, the λ-tilted distribution $\mathsf{P}^{(\lambda)}$ with respect to P_1 and P_2 is given by

$$\mathsf{P}^{(\lambda)}(x) := \frac{\mathsf{P}_1^{1-\lambda}(x)\mathsf{P}_2^{\lambda}(x)}{\sum_{a\in\mathcal{X}} \mathsf{P}_1^{1-\lambda}(a)\mathsf{P}_2^{\lambda}(a)}, \quad \forall x \in \mathcal{X}. \tag{3.50}$$

The following fact characterizes the set of all error exponents, denoted by $\mathcal{R}_{\mathrm{FL}}$, achievable in a fixed-length manner.

Fact 3 (Theorem 11.7.1 in [27]). *The set of error exponents feasible for the class of fixed-length hypothesis tests is given by*

$$\mathcal{R}_{FL} = \left\{ (E_1, E_2) : E_i \leq D\left(\mathsf{P}^{(\lambda)} \| \mathsf{P}_i\right), i \in \{1,2\}, \textit{ for some } \lambda \in [0,1] \right\}. \tag{3.51}$$

Furthermore, the following fixed-length test achieves the optimal error exponents on the boundary of $\mathcal{R}_{FL}$. If

$$\begin{aligned} \tfrac{1}{n}\textstyle\sum_{i=1}^{n} \log \frac{\mathsf{P}_1(X_i)}{\mathsf{P}_2(X_i)} \geq \alpha \quad & \textit{stop and choose } H_1, \\ \tfrac{1}{n}\textstyle\sum_{i=1}^{n} \log \frac{\mathsf{P}_1(X_i)}{\mathsf{P}_2(X_i)} < \alpha \quad & \textit{stop and choose } H_2, \end{aligned} \tag{3.52}$$

where α is given by

$$\alpha = D\left(\mathsf{P}^{(\lambda)} \| \mathsf{P}_2\right) - D\left(\mathsf{P}^{(\lambda)} \| \mathsf{P}_1\right), \lambda \in [0,1]. \tag{3.53}$$

Definition 18. Let λ^* be such that

$$D\left(\mathsf{P}^{(\lambda^*)} \| \mathsf{P}_1\right) = D\left(\mathsf{P}^{(\lambda^*)} \| \mathsf{P}_2\right). \tag{3.54}$$

Then, the Chernoff exponent D^* is defined as

$$D^* := D\left(\mathsf{P}^{(\lambda^*)} \| \mathsf{P}_1\right), \tag{3.55}$$

and it characterizes the optimal reliability of Bayesian tests. In other words, D^* denotes the optimal exponent that can be achieved simultaneously by both type-I and type-II errors in a fixed-length manner.

3.4.2 Sequential Hypothesis Tests

In this setting, τ is allowed to be a random variable (potentially unbounded) such that $\max\{\mathbb{E}_1[\tau], \mathbb{E}_2[\tau]\} \leq n$, where $n \in \mathbb{N}$.

Definition 19. The error exponents (E_1, E_2) are said to be sequentially achievable, if for every $\delta > 0$ there exists an $N(\delta) \in \mathbb{N}$ such that for all $n \geq N(\delta)$ there exists a hypothesis test satisfying the following constraints

$$\max\{\mathbb{E}_1[\tau], \mathbb{E}_2[\tau]\} \leq n, \tag{3.56}$$

$$\mathsf{P}_1(A_2^\tau) \leq e^{-(E_1-\delta)n}, \tag{3.57}$$

$$\mathsf{P}_2(A_1^\tau) \leq e^{-(E_2-\delta)n}. \tag{3.58}$$

The following fact characterizes the set of all error exponents, denoted by $\mathcal{R}_{\text{seq}}$, achievable in sequential manner.

Fact 4 (Wald and Wolfowitz [34])**.** *The set of error exponents feasible for the class of sequential hypothesis test are given by*

$$\mathcal{R}_{seq} = \{E_1 : E_1 \leq D(\mathsf{P}_2 \| \mathsf{P}_1)\} \times \{E_2 : E_2 \leq D(\mathsf{P}_1 \| \mathsf{P}_2)\}. \tag{3.59}$$

Furthermore, the following sequential hypothesis test achieves the above optimal error exponents $(D(\mathsf{P}_2 \| \mathsf{P}_1), D(\mathsf{P}_1 \| \mathsf{P}_2))$*. At any instant* $k \in \mathbb{N}$*,*

$$\begin{array}{ll} \sum_{i=1}^k \log \frac{\mathsf{P}_1(X_i)}{\mathsf{P}_2(X_i)} \geq \alpha & \textit{stop and choose } H_1, \\ \sum_{i=1}^k \log \frac{\mathsf{P}_1(X_i)}{\mathsf{P}_2(X_i)} \leq \beta & \textit{stop and choose } H_2, \\ \beta < \sum_{i=1}^k \log \frac{\mathsf{P}_1(X_i)}{\mathsf{P}_2(X_i)} < \alpha & \textit{take an extra sample} \\ & \textit{and repeat for } k+1, \end{array} \tag{3.60}$$

where $\alpha = (D(\mathsf{P}_1 \| \mathsf{P}_2) - \delta)n$ *and* $\beta = -(D(\mathsf{P}_2 \| \mathsf{P}_1) - \delta)n$*.*

Remark 8. Our definition of sequentially achievable error exponents, given by equations (3.56), (3.57), and (3.58), coincides with the achievable error exponents defined in [50]. Alternatively, the definition can be modified such that for $n_1, n_2 \in \mathbb{N}$ large enough the hypothesis test satisfies

$$\mathbb{E}_1[\tau] \leq n_1, \quad \mathbb{E}_2[\tau] \leq n_2, \tag{3.61}$$

$$\mathsf{P}_1(A_2^\tau) \leq e^{-(E_1-\delta)n_1}, \tag{3.62}$$

$$\mathsf{P}_2(A_1^\tau) \leq e^{-(E_2-\delta)n_2}, \tag{3.63}$$

as considered in [51]. In contrast, only the case where $n_1 = n_2$ is considered in [50]. Our definition which includes the definition of [50] as a special case is more stringent than the definition considered in [51]. For instance this definition does not admit sequential tests that increase the error exponent E_1 arbitrarily under H_1 by taking arbitrarily large number of samples under H_1 than under H_2, i.e., by making $\frac{n_1}{n_2}$ arbitrarily large.

In summary, an optimal fixed-length hypothesis test can only achieve the maximum error exponent in one type of error if the probability of the other error-type is kept fixed. In contrast, a sequential hypothesis test achieves both optimal error exponents simultaneously. The following example with Figure 1 illustrates this.

3.4.3 Almost-Fixed-Length Hypothesis Tests

We introduce a new class of hypothesis tests in the same spirit as the class of AFLF codes.

Definition 20. The error exponents (E_1, E_2) are said to be achievable in a (γ, K)-almost-fixed-length manner for some $\gamma \geq 0$ and $K \in \mathbb{N}$, if for every $\delta > 0$ there exists an $N(\delta) \in \mathbb{N}$ such that

for all $n \geq N(\delta)$ there exists a hypothesis test satisfying the following

$$\mathsf{P}_i(\tau > n) \leq e^{-\gamma n} \quad i \in \{1,2\}, \tag{3.64}$$

$$\tau \leq Kn \quad \mathsf{P}_i - \text{a.s. for } i \in \{1,2\}, \tag{3.65}$$

$$\mathsf{P}_1(A_2^\tau) \leq e^{-(E_1-\delta)n}, \tag{3.66}$$

$$\mathsf{P}_2(A_1^\tau) \leq e^{-(E_2-\delta)n}. \tag{3.67}$$

Let $\mathcal{R}_{\text{AFL}}^{(\gamma,K)}$ denote the region of all feasible error exponents of the class of (γ,K)-almost-fixed-length tests.

Remark 9. Note that as γ tends to ∞, this class of tests recover the class of fixed-length hypothesis tests, hence $\mathcal{R}_{\text{FL}} \subset \mathcal{R}_{\text{AFL}}^{(\gamma,K)}$, for every $\gamma \geq 0$ and $K \in \mathbb{N}$. Similarly, for all $\varepsilon > 0$ and n large enough, we have that $\mathbb{E}_i[\tau] \leq n + \varepsilon$, for $i \in \{1,2\}$. This implies that $\mathcal{R}_{\text{AFL}}^{(\gamma,K)} \subset \mathcal{R}_{\text{seq}}$.

3.5 Exponents of Almost-Fixed-Length Hypothesis Tests

Next we provide our main result for almost-fixed-length hypothesis tests.

Theorem 5. *For any $\gamma \geq 0$, define*

$$\mathcal{R}_\gamma := \mathcal{R}_{FL} \cup \{E_1 : E_1 \leq E_1(\gamma)\} \times \{E_2 : E_2 \leq E_2(\gamma)\}, \tag{3.68}$$

where define

$$E_1(\gamma) := \max_{\lambda \in [0,1]} \left\{ D\left(\mathsf{P}^{(\lambda)} \| \mathsf{P}_1\right) : D\left(\mathsf{P}^{(\lambda)} \| \mathsf{P}_2\right) \geq \gamma \right\}, \tag{3.69}$$

and

$$E_2(\gamma) := \max_{\lambda \in [0,1]} \left\{ D\left(\mathsf{P}^{(\lambda)} \| \mathsf{P}_2\right) : D\left(\mathsf{P}^{(\lambda)} \| \mathsf{P}_1\right) \geq \gamma \right\}. \tag{3.70}$$

For any $\gamma \geq 0$ *and* $K \in \mathbb{N}$*, we have*

$$\mathcal{R}_\gamma \cap K\mathcal{R}_{FL} \subset \mathcal{R}_{AFL}^{(\gamma,K)}. \tag{3.71}$$

and conversely, we have

$$\mathcal{R}_{AFL}^{(\gamma,K)} \subset \mathcal{R}_\gamma \cap K\mathcal{R}_{FL}. \tag{3.72}$$

The proof of the above theorem follows by from combining Proposition 4 with Proposition 5. The following corollary is an immediate consequence of the above theorem.

Corollary 6. *For* $\gamma > D^*$*, and for all* $K \in \mathbb{N}$*, we have*

$$\mathcal{R}_{AFL}^{(\gamma,K)} = \mathcal{R}_{FL} \tag{3.73}$$

since $\{E_1 : E_1 \leq E_1(\gamma)\} \times \{E_2 : E_2 \leq E_2(\gamma)\} \subset \mathcal{R}_{FL}$.

3.5.1 Achievability: A Two Phase Hypothesis Test

For $\gamma > D^*$, the achievability of $\mathcal{R}_\gamma$ coincides with that of the class of fixed-length hypothesis tests, $\mathcal{R}_{FL}$, ($\mathsf{P}_i(\tau > n) = 0$ and since $\mathcal{R}_\gamma = \mathcal{R}_{\mathrm{FL}}$), so any fixed-length hypothesis test achieves $\mathcal{R}_\gamma$. Let us consider $\gamma \leq D^*$ and $K \in \mathbb{N}$. We propose a hypothesis test that decides between the hypotheses at two evaluation points, one at n and the other at Kn. Formally the two phase hypothesis test is described as follows for $\gamma \leq D^*$.

(i) Phase-I: We collect n samples and choose whether to stop and decide between the hypotheses

or to continue to collect extra samples. Define

$$\lambda_1(\gamma) := \underset{\lambda\in[0,1]}{\operatorname{argmax}} \left\{ D\left(\mathsf{P}^{(\lambda)} \| \mathsf{P}_1\right) : D\left(\mathsf{P}^{(\lambda)} \| \mathsf{P}_2\right) \geq \gamma \right\}, \tag{3.74}$$

and

$$\lambda_2(\gamma) := \underset{\lambda\in[0,1]}{\operatorname{argmax}} \left\{ D\left(\mathsf{P}^{(\lambda)} \| \mathsf{P}_2\right) : D\left(\mathsf{P}^{(\lambda)} \| \mathsf{P}_1\right) \geq \gamma \right\}. \tag{3.75}$$

Note that $\lambda_1(\gamma)$ and $\lambda_2(\gamma)$ denote $\lambda \in [0,1]$ which achieves in the maximum in equation (3.69) and (3.70) respectively. For the ease of exposition we will denote $\lambda_i(\gamma)$ as λ_i where $i \in \{1,2\}$. Furthermore, let

$$\alpha_1 = D\left(\mathsf{P}^{(\lambda_2)} \| \mathsf{P}_2\right) - D\left(\mathsf{P}^{(\lambda_2)} \| \mathsf{P}_1\right), \tag{3.76}$$

$$\beta_1 = D\left(\mathsf{P}^{(\lambda_1)} \| \mathsf{P}_2\right) - D\left(\mathsf{P}^{(\lambda_1)} \| \mathsf{P}_1\right). \tag{3.77}$$

Phase-I of the strategy is conducted as follows: if

$$\begin{array}{ll} \frac{1}{n}\sum_{i=1}^{n} \log \frac{\mathsf{P}_1(X_i)}{\mathsf{P}_2(X_i)} \geq \alpha_1 & \text{stop and choose 1,} \\ \frac{1}{n}\sum_{i=1}^{n} \log \frac{\mathsf{P}_1(X_i)}{\mathsf{P}_2(X_i)} \leq \beta_1 & \text{stop and choose 2,} \\ \beta_1 < \frac{1}{n}\sum_{i=1}^{n} \log \frac{\mathsf{P}_1(X_i)}{\mathsf{P}_2(X_i)} < \alpha_1 & \text{proceed to Phase-II.} \end{array} \tag{3.78}$$

(ii) Phase-II: In the second phase, $(K-1)n$ extra samples are obtained. Fix some $\lambda \in [0,1]$ and let

$$\alpha = D\left(\mathsf{P}^{(\lambda)} \| \mathsf{P}_2\right) - D\left(\mathsf{P}^{(\lambda)} \| \mathsf{P}_1\right). \tag{3.79}$$

Phase-II of the strategy is conducted as follows: if

$$\begin{aligned} \tfrac{1}{Kn}\textstyle\sum_{i=1}^{Kn} \log \tfrac{\mathsf{P}_1(X_i)}{\mathsf{P}_2(X_i)} \geq \alpha \quad \text{stop and choose 1,} \\ \tfrac{1}{Kn}\textstyle\sum_{i=1}^{Kn} \log \tfrac{\mathsf{P}_1(X_i)}{\mathsf{P}_2(X_i)} < \alpha \quad \text{stop and choose 2.} \end{aligned} \tag{3.80}$$

Proposition 4. *Let $\gamma \leq D^*$ and $K \in \mathbb{N}$. The two phase hypothesis test as given by equations* (3.78) *and* (3.80) *is an almost-fixed-length hypothesis test and the set of error exponents achieved is given by*

$$\mathcal{R}_\gamma \cap K\mathcal{R}_{FL}.$$

Furthermore, for all $K \geq K^*$ and for $\alpha = 0$ in Phase-II in equation (3.80)*, the two phase hypothesis test achieves any $(E_1, E_2) \in \mathcal{R}_\gamma$ where*

$$K^* := \max\left\{ \frac{D(\mathsf{P}_2 \,\|\mathsf{P}_1)}{D^*}, \frac{D(\mathsf{P}_1 \,\|\mathsf{P}_2)}{D^*} \right\}.$$

The proof of the above proposition is provided in Appendix 3.7.3.

Example 2 (Revisited). Figure 3.5 shows the region of error exponents $\mathcal{R}_{\text{ALF}}^{(\gamma,K)}$ described in Theorem 5 at different values of γ for $K \geq K^* = 4$. As γ decreases, the trade-off between error exponents (E_1, E_2) improves. In particular, it shows that it is possible to achieve the error exponents that are arbitrarily close to optimal error exponents of sequential hypothesis tests, i.e. $(D(\mathsf{P}_2 \,\|\mathsf{P}_1), D(\mathsf{P}_1 \,\|\mathsf{P}_2))$, by selecting γ arbitrarily close to zero. Figure 3.6 shows $\mathcal{R}_{\text{ALF}}^{(\gamma,K)}$ for $K = 2$ which is strictly less than $K^* = 4$. We see that for smaller values of γ the feasible region of error exponents is bounded by $K\mathcal{R}_{\text{FL}}$.

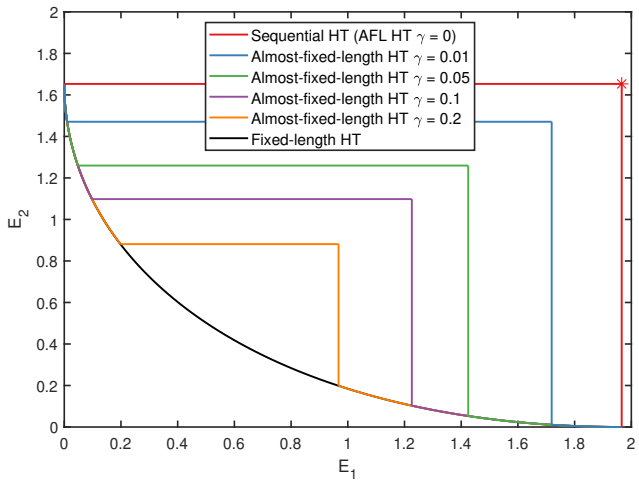


Figure 3.5: This figure shows the region $\mathcal{R}_{\text{AFL}}^{(\gamma,K)}$ for various values of γ when $K \geq K^* = 4$ when the samples are Bernoulli with parameters $p_1 = 0.9$ under H_1 and $p_2 = 0.2$ under H_2. As γ decreases the trade-off between the error exponents gets better and the test achieves the optimal sequential exponents $(D(\mathsf{P}_2 \| \mathsf{P}_1), D(\mathsf{P}_1 \| \mathsf{P}_2))$.

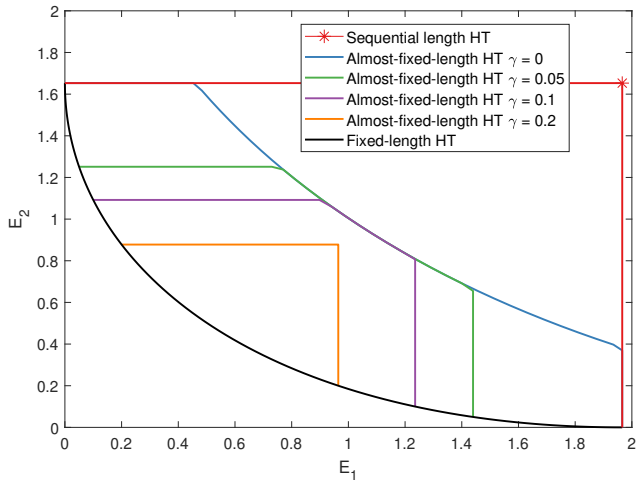


Figure 3.6: This figure shows the achievable region of the two phase hypothesis test as γ increases for $K = 2$ ($K^* = 4$), when the samples are Bernoulli with parameters $p_1 = 0.9$ under H_1 and $p_2 = 0.2$ under H_2.

3.5.2 Converse: Hypothesis Testing with Rejection Option

Our converse bounds the performance of a γ-almost-fixed-length hypothesis test with that of a fixed-length hypothesis test with rejection option where the probability of rejection

approaches zero exponentially fast with an exponent at most γ. A hypothesis test with rejection option, at end of τ samples divides the sample space $\mathcal{X}^\tau$ into three sets or decision regions, given by A_i^τ for $i \in \{1,2\}$ where the test accepts H_i, and A_Ω^τ which denotes the region for which the test rejects both hypotheses H_1 and H_2. Given a (γ, K)-almost-fixed-length hypothesis test we can construct a hypothesis test with rejection option by rejecting to choose either of the hypotheses whenever $\tau > n$.

Definition 21. The exponents (E_1, E_2, E_Ω) are said be achievable, if for every $\delta > 0$ there exists an $N(\delta) \in \mathbb{N}$ such that for all $n \geq N(\delta)$ there exists a hypothesis test with rejection option that satisfies the following

$$\tau \leq n, \tag{3.81}$$

$$\mathsf{P}_1(A_2^\tau) \leq e^{-(E_1-\delta)n}, \quad \mathsf{P}_2(A_1^\tau) \leq e^{-(E_2-\delta)n}, \tag{3.82}$$

$$\mathsf{P}_1(A_\Omega^\tau) + \mathsf{P}_2(A_\Omega^\tau) \leq e^{-(E_\Omega-\delta)n}. \tag{3.83}$$

Lemma 7. *For any $\gamma \geq 0$, let $\overline{\mathcal{R}}_\gamma$ denote the region of all feasible error exponents for the class of hypothesis tests with rejection option, then we have*

$$\mathcal{R}_\gamma \times \{E_\Omega = \gamma\} \subset \overline{\mathcal{R}}_\gamma.$$

Conversely, for every $\gamma \geq 0$ we have

$$\overline{\mathcal{R}}_\gamma \subset \mathcal{R}_\gamma \times \{E_\Omega = \gamma\}.$$

A variant of above the lemma has been studied under the class of hypothesis tests with rejection option considered by Grigoryan et al. in [44], Sason in [45] and Gutman in [52]. The following proposition builds on the converse of hypothesis tests with rejection option to provide a converse for almost-fixed-length hypothesis tests.

Proposition 5. *Let $\gamma \geq 0$ and $K \in \mathbb{N}$. The region of all feasible error exponents of the class of (γ, K)-almost-fixed-length tests satisfies*

$$\mathcal{R}_{AFL}^{(\gamma,K)} \subset \mathcal{R}_{\gamma} \cap K\mathcal{R}_{FL}. \tag{3.84}$$

The proof of the above proposition is provided in Appendix 3.7.4.

3.6 Conclusion and Future Work

We looked at a new class of strategies for channel coding and hypothesis tests that have a slight flexibility over fixed-length strategies by allowing a slightly large stopping time in exponentially small fraction of sample paths. We show that when stopping times are allowed to be slightly large in only exponentially small cases, the overall reliability is increased significantly. We showed that it is possible to achieve optimal performance of variable-length strategies using almost-fixed-length strategies. Since, for any $n \geq 1$, for all $(\ell, M, \gamma, K, \varepsilon)$ AFLF code we have that

$$\lim_{\ell \to \infty} \mathbb{E}\left[\left(\frac{\tau_\ell}{\ell}\right)^n\right] = 1, \quad \lim_{\ell \to \infty} \mathrm{Var}(\tau_\ell) = 0.$$

Similarly, the above equation holds for the stopping time of any (γ, K) AFL hypothesis test. This means that the class of strategies for which the variance of the stopping time is required be to zero, is no more restrictive than the class of strategies that satisfy only an average-length constraint, in terms of reliability. Hence, neither growing variability nor the computational complexity are essential to obtaining the optimal error exponents. Similar statements can be made for constraining higher moments of the stopping time.

3.7 Appendix

3.7.1 Two-Phase AFLF Code based on Truncated Yamamoto-Itoh Strategy

The following achievability strategy is similar to the strategy considered in [43], however for completeness we provide a simpler and intuitive proof which is more natural to our problem. Our two-phase strategy is described as follows:

1. Phase-I (Truncated Yamamoto-Itoh Strategy): Fix some $\alpha \in [0,1]$. For the first phase we consider the Yamamoto-Itoh strategy [47] with block-length ℓ and $2^{\ell R}$ number of messages, where we divide the first phase into two parts of length $\alpha\ell$ and $(1-\alpha)\ell$. In the first part, we transmit ℓR bits over $\alpha\ell$ channel uses using a random code [36]. Hence, the probability of making an error in the first part is given by

$$\mathsf{P}_{1\mathrm{e}} \leq 2^{-\alpha\ell E_{\mathrm{r}}\left(\frac{R}{\alpha}\right)}. \tag{3.85}$$

Now fix $x, x' \in \mathcal{X}$ such that $D(P_{Y|X}(\cdot|x)||\mathsf{P}_{Y|X}(\cdot|x'))$ is maximized, i.e.,

$$C_1 = D(P_{Y|X}(\cdot|x)||\mathsf{P}_{Y|X}(\cdot|x')).$$

If the received message has been decoded correctly we send ACKs, i.e., transmit input x otherwise we send NACKs, i.e., transmit input x' for the remaining length $(1-\alpha)\ell$. Now construct a fixed length hypothesis test as shown in Fact 3, for some $\lambda \in [0,1]$, to distinguish between the ACK and NACK symbols received such that

$$\mathsf{P}_{2\mathrm{ec}} \leq 2^{-(1-\alpha)\ell D(\mathsf{P}^{(\lambda)}||\mathsf{P}_{Y|X}(\cdot|x'))}, \tag{3.86}$$

and

$$\mathsf{P}_{2\text{ce}} \leq 2^{-(1-\alpha)\ell D(\mathsf{P}^{(\lambda)}||\mathsf{P}_{Y|X}(\cdot|x))}, \tag{3.87}$$

where $\mathsf{P}_{2\text{ec}}$ denotes the probability of receiving an ACK when NACK was transmitted and $\mathsf{P}_{2\text{ce}}$ denotes the probability of receiving a NACK when ACK was transmitted. If the hypothesis test declares a NACK, proceed to Phase-II otherwise accept the decoded message.

2. Phase-II (Random Code):In the second phase send ℓR bits using random channel coding with blocklength $(K-1)\ell$ at rate $\frac{R}{K-1}$ so that $\tau \leq K\ell$.

Clearly the above strategy is an AFLF code whose γ and error exponent are determined as follows. For the above strategy the probability of entering Phase-II is given by

$$\begin{aligned}
\mathsf{P}(\tau > \ell) &= \mathsf{P}_{1e}(1-\mathsf{P}_{2\text{ec}}) + (1-\mathsf{P}_{1e})\mathsf{P}_{2\text{ce}} && (3.88)\\
&\leq \mathsf{P}_{1\text{e}} + \mathsf{P}_{2\text{ce}} && (3.89)\\
&\leq 2^{-\alpha\ell E_{\text{r}}\left(\frac{R}{\alpha}\right)} + 2^{-(1-\alpha)\ell D(\mathsf{P}^{(\lambda)}||\mathsf{P}_{Y|X}(\cdot|x))}. && (3.90)
\end{aligned}$$

Furthermore, let $\varepsilon_{(K-1)\ell}$ denote the probability of error in Phase-II then the total probability of error of the AFLF code is given by

$$\begin{aligned}
\varepsilon_\ell &= \mathsf{P}_{1\text{e}}\mathsf{P}_{2\text{ec}} + \mathsf{P}(\tau > \ell)\varepsilon_{(K-1)\ell}\\
&\leq 2^{-\alpha\ell E_{\text{r}}\left(\frac{R}{\alpha}\right)} 2^{-(1-\alpha)\ell D(\mathsf{P}^{(\lambda)}||\mathsf{P}_{Y|X}(\cdot|x'))} + 2^{-(K-1)\ell E_{\text{r}}\left(\frac{R}{K-1}\right)}. && (3.91)
\end{aligned}$$

Proof of Theorem 3

Fix $\varepsilon > 0$ and let $\alpha = \frac{R}{C-\varepsilon}$. We construct the hypothesis test such that the exponent of P_{2ce} error is 0 and the exponent of P_{2ec} is $D(\mathsf{P}_{Y|X}(\cdot|x)||\mathsf{P}_{Y|X}(\cdot|x'))$, i.e., C_1 of the channel. Hence,

as ε goes to 0 the probability of entering Phase-II goes to 0, and the error exponent of the strategy satisfies

$$E_{\text{AFLF}}(R,0,K) \geq \liminf_{\ell\to\infty} -\frac{1}{\ell}\log \varepsilon_\ell \tag{3.92}$$

$$= \min\left\{C_1\left(1-\frac{R}{C}\right), (K-1)E_{\text{r}}\left(\frac{R}{K-1}\right)\right\}. \tag{3.93}$$

To obtain the value of K for which the optimal $E_{\text{VLF}}(R)$ is achieved using the above two phase strategy as an AFLF code, we extend the argument used in [49] to any general DMC. Note that

$$(K-1)E_{\text{r}}\left(\frac{C}{K-1}\right) \geq E_{\text{VLF}}(C) = 0 \tag{3.94}$$

and for $K \geq 1+\frac{C_1}{E_{\text{R}}(0)}$ we have

$$(K-1)E_{\text{r}}(0) \geq E_{\text{VLF}}(0) = C_1. \tag{3.95}$$

Since $C_1 > C$ we have

$$\frac{\partial E_{\text{VLF}}(R)}{\partial R} = -\frac{C_1}{C} \leq -1. \tag{3.96}$$

Furthermore, it is known that [36]

$$\frac{\partial E_{\text{r}}(R)}{\partial R} \geq -1. \tag{3.97}$$

Hence, for all $K \geq 1+\frac{C_1}{E_{\text{R}}(0)}$ it follows that

$$(K-1)E_{\text{R}}\left(\frac{R}{K-1}\right) \geq E_{\text{VLF}}(R) \tag{3.98}$$

for all $0 \leq R \leq C$. Combining with Corollary 4 we have the assertion of the theorem.

Proof of Theorem 4

Fix some $\alpha \in [0,1]$. Now note that

$$\liminf_{\ell\to\infty} -\frac{1}{\ell}\log \mathsf{P}(\tau > \ell) \geq \min\left\{\alpha E_{\mathrm{r}}\left(\frac{R}{\alpha}\right), (1-\alpha)\, D(\mathsf{P}^{(\lambda)}||\mathsf{P}_{Y|X}(\cdot|x))\right\}, \tag{3.99}$$

and similarly we have

$$E_{\mathrm{AFLF}}(R,\gamma,K) \geq \min\left\{\alpha E_{\mathrm{r}}\left(\frac{R}{\alpha}\right) + (1-\alpha)\, D(\mathsf{P}^{(\lambda)}||\mathsf{P}_{Y|X}(\cdot|x')), (K-1)E_{\mathrm{r}}\left(\frac{R}{K-1}\right)\right\} \tag{3.100}$$

Case 1: Consider the case where $0 < \gamma < E(R)$. Let $\alpha^*(R,\gamma)$ denote α which satisfies

$$\alpha E_{\mathrm{r}}\left(\frac{R}{\alpha}\right) = \gamma, \tag{3.101}$$

and hence

$$\alpha^*(R,\gamma) = \frac{R}{g^{-1}(\frac{\gamma}{R})}, \tag{3.102}$$

where $g(a) = \frac{E_{\mathrm{r}}(a)}{a}$. Then, we have

$$E_{\mathrm{AFLF}}(R,\gamma,K) \geq \min\left\{E'_{\mathrm{ee}}(R,\gamma), (K-1)E_{\mathrm{r}}\left(\frac{R}{K-1}\right)\right\}, \tag{3.103}$$

where we define

$$E'_{\mathrm{ee}}(R,\gamma) := \max_{\alpha\in[\alpha^*(R,\gamma),1]} \max_{\substack{\lambda\in[0,1] \\ (1-\alpha)D(\mathsf{P}^{(\lambda)}||\mathsf{P}_{Y|X}(\cdot|x))\geq\gamma}} E''_{\mathrm{ee}}(\alpha,\lambda,R), \tag{3.104}$$

and define

$$E''_{\mathrm{ee}}(\alpha,\lambda,R) := \alpha E_{\mathrm{r}}\left(\frac{R}{\alpha}\right) + (1-\alpha)\, D(\mathsf{P}^{(\lambda)}||\mathsf{P}_{Y|X}(\cdot|x')). \tag{3.105}$$

Case 2: Consider the case where $\gamma > E_{\mathrm{r}}(R)$. Since for all $\alpha \in [0,1]$ we have $\alpha E_{\mathrm{r}}\left(\frac{R}{\alpha}\right) \leq E_{\mathrm{r}}(R)$, choose $\alpha = 1$ so that the decoder never declares NACK and hence $\tau = \ell$ a.s. Furthermore we have $E_{\mathrm{AFLF}}(R,\gamma,K) \geq E_r(R)$ for all $0 \leq R \leq C$.

3.7.2 Proof of Proposition 3

For every $(\ell, M, \gamma, K, \varepsilon)$ AFLF code we obtain an $(\ell, M, 2^{-\gamma\ell}, \varepsilon)$ error-erasure code. Since the optimal error exponent of an error-erasure code with erasure exponent γ is given by $E_{\mathrm{ee}}(R,\gamma)$, the probability of error of $(\ell, M, 2^{-\gamma\ell}, \varepsilon)$ error-erasure code satisfies

$$\varepsilon \geq 2^{-\ell E_{\mathrm{ee}}(R,\gamma)}. \tag{3.106}$$

Therefore, the probability of error of $(\ell, M, \gamma, K, \varepsilon)$ AFLF code also satisfies $\varepsilon \geq 2^{-\ell E_{\mathrm{ee}}(R,\gamma)}$. Furthermore, the AFLF code uses $K\ell$ channel uses which implies

$$\varepsilon \geq 2^{-E_{\mathrm{H}}(R)}. \tag{3.107}$$

In other words, the probability of error of $(\ell, M, \gamma, K, \varepsilon)$ AFLF code

$$\varepsilon \geq \max\left\{2^{-\ell E_{\mathrm{ee}}(R,\gamma)}, 2^{-E_{\mathrm{H}}(R)}\right\}. \tag{3.108}$$

Hence, the assertion of the proposition follows.

3.7.3 Proof of Proposition 4

It is straightforward to check that equations (3.74) and (3.75) imply

$$D\left(\mathsf{P}^{(\lambda_1)}||\mathsf{P}_1\right) = E_1(\gamma), \text{ and } D\left(\mathsf{P}^{(\lambda_1)}||\mathsf{P}_2\right) = \gamma, \tag{3.109}$$

$$D\left(\mathsf{P}^{(\lambda_2)}||\mathsf{P}_2\right) = E_2(\gamma), \text{ and } D\left(\mathsf{P}^{(\lambda_2)}||\mathsf{P}_1\right) = \gamma. \tag{3.110}$$

Note that by construction, $\tau \leq Kn$ a.s. Additionally, we must show that the probability of the sample paths where $\tau > n$ is exponentially small with an exponent γ. Consider

$$\begin{aligned}\mathsf{P}_1(\tau > n) &= \mathsf{P}_1\left(\beta_1 < \frac{1}{n}\sum_{i=1}^{n}\log\frac{\mathsf{P}_1(X_i)}{\mathsf{P}_2(X_i)} < \alpha_1\right) \\ &\leq \mathsf{P}_1\left(\frac{1}{n}\sum_{i=1}^{n}\log\frac{\mathsf{P}_1(X_i)}{\mathsf{P}_2(X_i)} < \alpha_1\right).\end{aligned}$$

Hence, for any $\delta > 0$ there exists an $N(\delta)$ such that for all $n \geq N(\delta)$ we have

$$\mathsf{P}_1(\tau > n) \overset{(a)}{\leq} e^{-(D\left(\mathsf{P}^{(\lambda_2)}\|\mathsf{P}_1\right)-\delta)n} \overset{(b)}{\leq} e^{-(\gamma-\delta)n},$$

where (a) is obtaining using Sanov's Theorem (Theorem 11.4.1 in [27]) and equation (3.77), and (b) comes from equation (3.110). Similarly, for all $n \geq N(\delta)$ we also have $\mathsf{P}_2(\tau > n) \leq e^{-(\gamma-\delta)n}$ using Sanov's Theorem and equations (3.76) and (3.109). Hence, this test belongs to the class of (γ, K)-almost-fixed-length hypothesis test.

The error of type-I is given as follows,

$$\begin{aligned}\mathsf{P}_1\left(A_2^{\tau}\right) = \mathsf{P}_1&\left(\frac{1}{n}\sum_{i=1}^{n}\log\frac{\mathsf{P}_1(X_i)}{\mathsf{P}_2(X_i)} \leq \beta_1\right) + \\ &\mathsf{P}_1\left(\left\{\beta_1 < \frac{1}{n}\sum_{i=1}^{n}\log\frac{\mathsf{P}_1(X_i)}{\mathsf{P}_2(X_i)} < \alpha_1\right\} \cap \left\{\frac{1}{Kn}\sum_{i=1}^{Kn}\log\frac{\mathsf{P}_1(X_i)}{\mathsf{P}_2(X_i)} < \alpha\right\}\right).\end{aligned}$$

Fix $\lambda \in [0,1]$. Using Sanov's Theorem and from the definition of α_1 and β_1, for any $\delta > 0$ there

exists an $N_0(\delta)$ such that for all $n \geq N_0(\delta)$ we have

$$\mathsf{P}_1\left(A_2^{\tau}\right) \leq e^{-\left(D\left(\mathsf{P}^{(\lambda_1)}\|\mathsf{P}_1\right)-\delta\right)n} + e^{-\left(KD\left(\mathsf{P}^{(\lambda)}\|\mathsf{P}_1\right)-\delta\right)n}.$$

Now, taking limit we obtain

$$\lim_{n\to\infty} \frac{1}{n} - \log \mathsf{P}_1\left(A_2^{\tau}\right) \geq \min\left\{D\left(\mathsf{P}^{(\lambda_1)}\|\mathsf{P}_1\right), KD\left(\mathsf{P}^{(\lambda)}\|\mathsf{P}_1\right)\right\}.$$

Similarly, we obtain

$$\lim_{n\to\infty} \frac{1}{n} - \log \mathsf{P}_2\left(A_1^{\tau}\right) \geq \min\left\{D\left(\mathsf{P}^{(\lambda_2)}\|\mathsf{P}_2\right), KD\left(\mathsf{P}^{(\lambda)}\|\mathsf{P}_2\right)\right\}.$$

Now if we set $\alpha = 0$ in Phase-II, we have $D\left(\mathsf{P}^{(\lambda)}\|\mathsf{P}_1\right) = D\left(\mathsf{P}^{(\lambda)}\|\mathsf{P}_2\right) = D^*$. Hence, we have the assertion of the proposition.

3.7.4 Proof of Proposition 5

From the definition of (γ, K)-almost-fixed-length tests we have

$$\mathcal{R}_{\mathrm{AFL}}^{(\gamma,K)} \subset \mathcal{R}_{\mathrm{AFL}}^{(0,K)} \subset K\mathcal{R}_{\mathrm{FL}}. \tag{3.111}$$

From the converse of hypothesis test with rejection we have

$$\mathcal{R}_{\mathrm{AFL}}^{(\gamma,K)} \times \{E_\Omega = \gamma\} \subset \mathcal{R}_\gamma \times \{E_\Omega = \gamma\}. \tag{3.112}$$

Therefore, for every $\gamma \geq 0$ and $K \in \mathbb{Z}^+$ we have

$$\mathcal{R}_{\mathrm{AFL}}^{(\gamma,K)} \subset \mathcal{R}_\gamma \cap K\mathcal{R}_{\mathrm{FL}}. \tag{3.113}$$

Hence, we have the converse for Theorem 5.

Chapter 3, in part, is a reprint of the material as it appears in the paper: Anusha Lalitha and Tara Javidi, “Reliability of sequential hypothesis testing can be achieved by an almost-fixed-length test", in *IEEE International Symposium on Information Theory*, Barcelona, pp. 1710-1714, 2016. The dissertation author was the primary investigator and author of this paper.

Chapter 3, in full, is currently being prepared for submission for publication as: Anusha Lalitha and Tara Javidi, “Almost-fixed-length strategies for Channel Coding and Hypothesis Testing". The dissertation author was the primary investigator and author of this paper.

Chapter 4

Real-time Binary Posterior Matching

4.1 Introduction

While feedback cannot increase the capacity of memoryless channels [27, Ch. 7.12], it can dramatically reduce the probability of error and the complexity of the communication schemes that achieve them. For the binary symmetric channel (BSC), a *horizon-free* sequential scheme was proposed by Horstein [1]; it was rigorously proved to attain capacity by Shayevitz and Feder [2] for this and other channels, via its generalization—the *posterior matching* (PM) scheme. Exponential error-probability guarantees, for the *finite-horizon* setting, were constructed in [4, 6, 7, 53]. An exponential bound on the error probability in the horizon-free case has been devised by Waeber et al. [3], although this bound becomes trivial for rates much below the capacity.

The availability of instantaneous noiseless feedback obviates the need of transmitting long error-correcting codes across long epochs, and enables instead the use of sequential communication schemes, by providing full knowledge of the receiver's state to the transmitter. A class of problems where this may have powerful implications is that of stabilizing an unstable control plant over a noisy channel. In particular, in the presence of feedback, the structure of the

horizon-free PM decoder seems to match the structure of anytime reliable decoders (proposed for stabilizing unstable linear plants over noisy channel [54–56]).

However, the classical PM schemes assume that the entire information (possibly infinite bit) sequence is available essentially non-causally to the transmitter, prior to the beginning of transmission. That is, they are sequential with respect to the transmitted sequence (codeword) but not with respect to the information sequence. Consequently, the non-causal knowledge assumption precludes the use of the classical PM scheme for real-time and control scenarios, in which the data to be transmitted is determined in a causal fashion.

In the current work, we consider a real-time setting, described in detail in 5.2, in which the bits arrive to the transmitter one-by-one at random times, under the assumption that the inter-arrival times (time-arrival differences) have a known finite support. We construct, in 4.3, a causal (horizon-free) PM scheme for this setting, i.e., a scheme that is sequential with respect to both the information and the transmitted sequences. We provide exponential guarantees for the error probability akin to those of [3], in 4.4.

We apply the proposed scheme, in 4.5, for control over a BSC with feedback and compare its analytic and empirical stabilization performance with those of the anytime-reliable codes of Sahai and Mitter [54] that use no feedback but are computationally demanding, as well as with those of Simsek et al. [57],[1] in 4.5.1. We conclude the chapter with a discussion, in 4.6.

4.2 Problem Formulation

The transmitter wishes to transmit an infinite stream of bits over a BSC with cross over probability $p \in (0, 1/2)$. Let $s_i \in \{0,1\}$ denote the i-th bit in the infinite bit stream where $s_i \sim \mathsf{Bern}(1/2)$. Notationally, we consider the infinite bit sequence as the binary expansion

[1] Analytic guarantees for the scheme of [57] exist only for the case in which the entire information sequence is known in advance, which corresponds, to the case of stabilizing an unstable linear system with possibly unknown initial conditions but with no system disturbance.

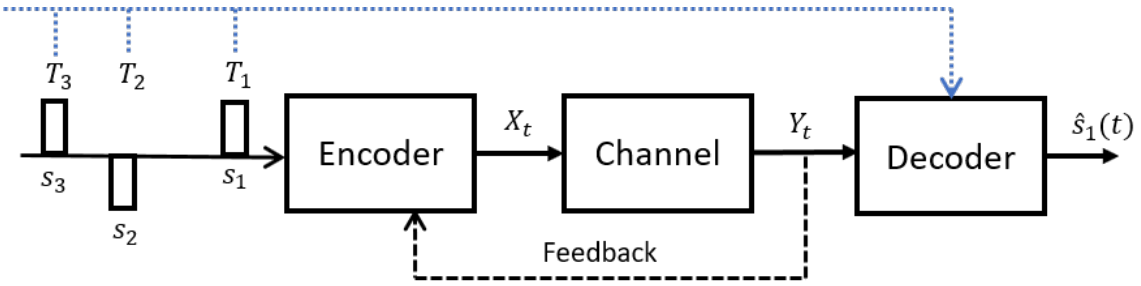


Figure 4.1: Figure shows the transmission of a stream of bits which arrive at the encoder at random times T_i. The arrival times are available at both at the encoder and decoder.

of a single message point Θ uniformly distributed over the unit interval i.e., $\Theta \sim \text{Unif}[0,1)$. We assume the bits of the message point Θ are revealed to the transmitter causally at arbitrary (possibly random) times, as follows. Let $\{N_i\}_{i\geq 1}$ be an i.i.d. random process where each N_i has a pmf p_N and $N_i \in [n_{\min}, n_{\max}]$, where $n_{\min} < n_{\max}$, and $n_{\min}, n_{\max} \in \mathbb{N}$. Furthermore, the i-th bit arrives at time $T_i := \sum_{j=1}^{i-1} N_j + 1$ for all $i \geq 2$ with $T_1 = 1$. For all time instants $t \in \mathbb{N}$, define the following random variable

$$b(t) := \max\{i \in \mathbb{N} : T_i \leq t\}. \tag{4.1}$$

In other words, $b(t)$ denotes the number of bits that have arrived by time t. Assuming that the first bit is available at the beginning (i.e., $T_1 = 1$), at time t at most first $\lceil \frac{t}{n_{\min}} \rceil$ bits have arrived with a non-zero probability and similarly the first $\lceil \frac{t}{n_{\max}} \rceil$ bits have arrived with probability 1, which implies

$$\mathsf{P}\left(\left\lceil \frac{t}{n_{\max}} \right\rceil \leq b(t) \leq \left\lceil \frac{t}{n_{\min}} \right\rceil \right) = 1. \tag{4.2}$$

Remark 10 (Periodic arrival times)**.** An important special instance of this framework is the case of deterministic and periodic arrival times, in which a new information bit is revealed every fixed $n \in \mathbb{N}$ time steps.

We now define the feedback communication scheme of an information bit sequence that

is made available causally to the encoder with random inter bit-arrival times, depicted in 4.1. We assume that the times at which the bits are revealed to the transmitter are known at the receiver.

The encoder $\mathcal{E}$ is described by a sequence of (causal) functions $\{\mathcal{E}_t\}_{t\geq 1}$. A causal encoder with feedback emits a channel input symbol $x_t \in \{0,1\}$ as a function of number of bits available $s_1^{b(t)}$ and past channel outputs y_1^{t-1}:

$$x_t = \mathcal{E}_t\left(s_1^{b(t)}, y_1^{t-1}\right). \tag{4.3}$$

The decoder $\mathcal{D}$ is described by the sequence of functions $\{\mathcal{D}_t\}_{t\geq 1}$. After observing t channel outputs, the decoder outputs a vector of estimates of all the bits available at the encoder thus far, $\hat{s}_1^{b(t)}(t) = [\hat{s}_1(t), \hat{s}_2(t), \ldots, \hat{s}_{b(t)}(t)] \in \{0,1\}^{b(t)}$:

$$\hat{s}_1^{b(t)}(t) = \mathcal{D}_t\left(y_1^t\right). \tag{4.4}$$

At any time instant t, we want to analyze the probability of error in decoding the first j bits $\mathsf{P}\left(\hat{s}_1^j(t) \neq s_1^j\right)$ for $1 \leq j \leq b(t)$. Since the bits that arrive early get encoded for longer duration it is natural to expect that the probability of error in decoding the older bits is smaller than that in decoding the newer bits.

4.3 Causal Posterior Matching Strategy

In this section, we propose a causal PM based encoding and decoding strategy to transmit a causally available message where the inter bit-arrival times are random.

First, we provide an overview of the strategy. At time t, suppose only the first i bits are available to the encoder i.e., consider the event $b(t) = i$. Consider a unit interval $[0,1]$ and divide it into bins of equal length 2^{-i}. The message point Θ is located on the unit interval, whose first i bits s_1^i provide the index of the bin containing Θ, where the index takes values in the set

$\{0,1,\ldots,2^i-1\}$. The encoder and decoder maintain a posterior probability of the message point Θ belonging to each bin after observing the past channel outputs. For the next N_i channel uses, we use causal posterior matching (described in detail in Sections 4.3.2 and 4.3.3 below) to encode the first i bits. After observing each channel output the decoder and encoder (using feedback) perform a Bayesian update to the posterior probability of Θ. After these N_i channel uses, a new bit arrives. To accommodate the new bit, we divide each bin from the previous 2^i bins into 2 bins, resulting in 2^{i+1} bins in total. Furthermore, we divide the posterior probability equally into the newly created bins. Now, the first $i+1$ bits provide the index of the bin containing Θ on a grid with $2^{(i+1)}$ bins. This process of dividing the existing bins and the posterior probability to accommodate a new bit continues in a horizon-free manner. At any time t, the binary expansion of the index of the bin that contains the median of the posterior distribution are declared as estimates of the bits available at the encoder.

4.3.1 Preliminaries

Let BSC(p) denote a BSC with cross-over probability $p \in (0,1/2)$ with input $X \in \{0,1\}$, output $Y \in \{0,1\}$:

$$\mathsf{P}(Y=y|X=x) = \begin{cases} p & \text{if } y \neq x, \\ \bar{p} & \text{if } y = x. \end{cases} \tag{4.5}$$

Let $C(p) := 1-h(p)$ denote the capacity of BSC(p).

Suppose after t channel uses, the encoder has access to only the first i bits, i.e., consider the event $b(t)=i$. Furthermore, the decoder maintains a posterior distribution of Θ after observing t channel outputs y_1^t, i.e., $\mathsf{P}_{\Theta|Y_1^t}\left(\Theta \in [(k-1)2^{-i}, k2^{-i}) \middle| y_1^t\right)$ for all $k \in \{0,\ldots,2^i-1\}$. Let $F_{\Theta|Y_1^t}$ denote the cumulative distribution function (CDF) of posterior probability distribution. Due to the presence of feedback, the posterior distribution maintained by the decoder is available to

the encoder as well. We refer to the point $F^{-1}_{\Theta|Y_1^t}\left(1/2\middle|y_1^t\right) \in [0,1]$ as the median of the posterior probability distribution at time t.

The following definitions will be useful, as we shall see, in describing the causal PM strategy.

- For every $n \in \mathbb{N}$, let $\beta(n)$ denote the solution of the following equation

$$\beta = \psi^*(\beta) - \frac{1}{n}, \tag{4.6}$$

where

$$\psi(\lambda) := -\log\left\{(2p)^\lambda + (2\bar{p})^\lambda\right\} + 1, \tag{4.7}$$

and define $\psi^*(\beta)$ as the Legendre–Fenchel transform of $\psi(\lambda)$:

$$\psi^*(\beta) := \sup_{\lambda>0}\left(\psi(\lambda) - \lambda\beta\right). \tag{4.8}$$

Further denote by $\lambda^*(n) \in [0,1]$ the λ that achieves the supremum in (4.8) when $\psi^*(\beta)$ satisfies (4.6). In other

- For all $i,t \in \mathbb{N}$, let $k_i^{(t)} \in \{0,\ldots,2^i-1\}$ denote the index of the bin containing the median $F^{-1}_{\Theta|Y_1^t}(1/2)$ in the grid with resolution 2^{-i} over the unit interval, i.e,

$$k_i^{(t)} 2^{-i} \le F^{-1}_{\Theta|Y_1^t}\left(1/2\middle|y_1^t\right) < \left(k_i^{(t)}+1\right)2^{-i}. \tag{4.9}$$

We are now ready to describe the causal PM strategy in detail.

4.3.2 Encoder

Fix a parameter $\lambda \in \{\lambda^*(n_{\min}), \lambda^*(n_{\max})\}$. After t channel uses and the next channel input at time instant $t+1$ is given as follows. Recall that $b(t+1)$ denotes the number of bits available to the encode before transmission at time instant $t+1$. Let $d_1^{(t)}$ and $d_2^{(t)}$ denote the values of the probability to the left and to the right of the median in the bin $k_{b(t+1)}^{(t)}$, respectively:

$$d_1^{(t)} := 1/2 - F_{\Theta|Y_1^t}\left(k_{b(t+1)}^{(t)} 2^{-b(t+1)}\right), \tag{4.10}$$

$$d_2^{(t)} := F_{\Theta|Y_1^t}\left(\left(k_{b(t+1)}^{(t)} + 1\right) 2^{-b(t+1)}\right) - 1/2. \tag{4.11}$$

Define further, for any $\lambda \in [0,1]$,

$$\pi_1^{(t+1)}(\lambda) := \frac{h(\lambda, d_2^{(t)})}{h(\lambda, d_1^{(t)}) + h(\lambda, d_2^{(t)})}, \tag{4.12}$$

$$\pi_2^{(t+1)}(\lambda) := 1 - \pi_1^{(t+1)}(\lambda), \tag{4.13}$$

where

$$h(\lambda, d) := (1 - 2(\bar{p} - p)d)^{-\lambda} - (1 + 2(\bar{p} - p)d)^{-\lambda}. \tag{4.14}$$

The next channel input at time $t+1$, conditioned on the past observations y_1^t, with probability $\pi_1^{(t+1)}(\lambda)$ is given by

$$X_{t+1} = \begin{cases} 0 & \text{if } 0.s_1^{b(t+1)} \leq k_{b(t+1)}^{(t)} 2^{-b(t+1)}, \\ 1 & \text{if } 0.s_1^{b(t+1)} > k_{b(t+1)}^{(t)} 2^{-b(t+1)}. \end{cases} \tag{4.15}$$

and with probability $\pi_2^{(t+1)}(\lambda)$ is given by

$$X_{t+1} = \begin{cases} 0 & \text{if } 0.s_1^{b(t+1)} \leq \left(k_{b(t+1)}^{(t)} + 1\right) 2^{-b(t+1)}, \\ 1 & \text{if } 0.s_1^{b(t+1)} > \left(k_{b(t+1)}^{(t)} + 1\right) 2^{-b(t+1)}. \end{cases} \tag{4.16}$$

Note that there two cases for the encoding operation:

(i) No new bit arrives at $t+1$: In this case we have $b(t+1) = b(t)$. Hence, the resolution of the grid $2^{-b(t)}$ remains changed.

(ii) A new bit arrives at $t+1$: In this case we have $b(t+1) = b(t)+1$. Hence, the resolution of the grid decreases from $2^{-b(t)}$ to $2^{-b(t+1)}$. Therefore, the encoder divides each of the previous bins into two equal-length bins with equal posterior probabilities in each bin. In particular, for any $i \in \mathbb{N}$, after t channel uses under the event $b(t) = i$, $b(t+1) = i+1$, the encoder sets

$$\mathsf{P}_{\Theta|Y_1^t}\left(\Theta \in \left[(2k)2^{-i-1}, (2k+1)2^{-i-1}\right) \middle| y_1^t\right) \tag{4.17}$$

$$= \mathsf{P}_{\Theta|Y_1^t}\left(\Theta \in \left[(2k+1)2^{-i-1}, (2k+2)2^{-i-1}\right) \middle| y_1^t\right) \tag{4.18}$$

$$= \frac{1}{2}\mathsf{P}_{\Theta|Y_1^t}\left(\Theta \in \left[k2^{-i}, (k+1)2^{-i}\right) \middle| y_1^t\right), \tag{4.19}$$

for all $k \in \{0, \ldots, 2^i - 1\}$.

4.3.3 Decoder

Upon receiving the channel output at time instant $t+1$, the decoder performs a Bayesian update to the posterior of Θ as follows.

Lemma 8. *For $t \in \mathbb{N}$, consider $i \in \{1, \ldots, b(t+1)\}$. For all $k \leq k_i^{(t)}$, since $F_{\Theta|Y_1^t}\left(k2^{-i} \middle| y_1^t\right) \leq 1/2$,*

the Bayesian update after observing the channel output at time $t+1$ is given as follows:

$$\frac{F_{\Theta|Y_1^{t+1}}\left(k2^{-i}\middle|y_1^{t+1}\right)}{F_{\Theta|Y_1^{t}}\left(k2^{-i}\middle|y_1^{t}\right)} = \begin{cases} \frac{p}{\frac{1}{2}+(\bar{p}-p)d_1^{(t)}} & \text{if } Y_{t+1}=1, \\ \frac{\bar{p}}{\frac{1}{2}-(\bar{p}-p)d_1^{(t)}} & \text{if } Y_{t+1}=0, \end{cases} \tag{4.20}$$

with probability $\pi_1^{(t+1)}(\lambda)$ and

$$\frac{F_{\Theta|Y_1^{t+1}}\left(k2^{-i}\middle|y_1^{t+1}\right)}{F_{\Theta|Y_1^{t}}\left(k2^{-i}\middle|y_1^{t}\right)} = \begin{cases} \frac{p}{\frac{1}{2}-(\bar{p}-p)d_2^{(t)}} & \text{if } Y_{t+1}=1, \\ \frac{\bar{p}}{\frac{1}{2}+(\bar{p}-p)d_2^{(t)}} & \text{if } Y_{t+1}=0, \end{cases} \tag{4.21}$$

with probability $\pi_2^{(t+1)}(\lambda)$. For $k > k_i^{(t)}$, since $1-F_{\Theta|Y^t}(k2^{-i}) < 1/2$, the Bayesian update for $1-F_{\Theta|Y^t}(k2^{-i}) < 1/2$ can be specified similarly.

The proof of Lemma 8 is given in Appendix 4.7.1.

At any time instant t, the decoder generates an estimate $\widehat{\Theta}_t = F^{-1}_{\Theta|Y^t}\left(1/2\middle|y_1^t\right)$ of Θ. The estimates $\hat{s}_1^{b(t)}(t)$ of the first $b(t)$ bits are the binary expansion of the index of the bin containing the median $k_{b(t)}^{(t)}$. Furthermore, when a new bit arrives, similar to the encoder, the decoder divides each bin into two equal-length bins and equally divides the posterior probability.

Remark 11. In the special case where after t channel uses the median coincides with the (left) end point of a bin $k_i^{(t)}2^{-i}$ for some $i \in \{1,\ldots,b(t+1)\}$, at time $t+1$ the encoder transmits 1 if $s_1^{b(t+1)}$ bits are to the right of the median and—0 otherwise. Furthermore, the decoder's update reduces to the update of non-causal PM considered by Horstein in [1] and Shayevitz and Feder in [2], where each $F_{\Theta|Y^t}(k2^{-i})$, $k \le k_i^{(t)}$ and similarly $1-F_{\Theta|Y^t}(k2^{-i})$, $k > k_i^{(t)}$, expands by $2\bar{p}$ or shrinks by $2p$.

Remark 12. For any $i \ge 1$, for all $t \ge T_i$, the encoder has access to the first i bits and hence the number of bins is at least 2^i. In other words, from time T_i to t, the causal PM strategy operates on

a grid whose resolution is finer than 2^{-i} and hence updates $F_{\Theta|Y^t}(k2^{-i})$ for all $k \in \{0,\dots,k_i^{(t)}\}$ and $1-F_{\Theta|Y^t}(k2^{-i})$ for all $k \in \{k_i^{(t)}+1,\dots,2^i-1\}$. This implies that, for all $t \geq T_i$, we always encode the first i bits along with the newly available bits. Furthermore, although we assume bits arrive one at a time, the strategy and our analysis can be extended to the case where any $k \in \mathbb{N}$ bits arrive at a time.

4.4 Main Results

In this section, we provide our main result on the error exponent attained by the causal PM strategy.

Theorem 6. *Consider the causal PM strategy with parameter λ over a BSC(p). The i-th bit arrives at the encoder at a random time T_i, whose pmf is $p_N^{\otimes i}$, where the inter-arrival times lie in the set $[n_{\min},\dots,n_{\max}]$ and let*

$$b(t) = \max\{i \in \mathbb{N} : T_i \leq t\}.$$

Then,

(i) For $\lambda = \lambda^*(n_{\min})$, the probability of error in decoding the first $i \in \{1,\dots,\lceil t/n_{\min}\rceil\}$ message bits after t channel uses is bounded by

$$\mathsf{P}\left(\hat{s}_1^i(t) \neq s_1^i \middle| b(t) > i\right) \leq \kappa \mathbb{E}\left[2^{-\beta(n_{\min})(t-T_i)} \middle| b(t) > i\right], \tag{4.22}$$

where $\beta(n_{\min})$ is the solution of (4.6) for $n = n_{\min}$ and where κ is a finite positive constant.

(ii) For $\lambda = \lambda^*(n_{\max})$, the probability of error in decoding the first $i \in \{1,\dots,\lfloor t/n_{\max}\rfloor\}$ message

bits after t channel uses is bounded by

$$\mathsf{P}\left(\hat{s}_1^i(t) \neq s_1^i\right) \leq \kappa 2^{-\beta(n_{\max})(t-n_{\max}(i-1))}, \tag{4.23}$$

where $\beta(n_{\max})$ is the solution of (4.6) *for $n = n_{\max}$ and where κ is a finite positive constant.*

Theorem 6 shows that the causal PM strategy can operate in two regimes based on how the randomization probabilities $\pi_1^{(t)}$ and $\pi_2^{(t)}$ are chosen given the past observations and the number of bits available at the encoder, i.e., by setting the parameter λ appropriately. In the setting (i), causal PM can be thought of as operating in a "*high-rate regime*", since it decodes all the arrived information bits, but with a lower error exponent of (4.22), corresponding to $\beta(n_{\min})$. In contrast, in the setting (ii), causal PM can be thought of as operating in a "*low-rate regime*", as it decodes only the first $\lfloor t/n_{\max} \rfloor$ bits, but with a higher error exponent of (4.23), corresponding to $\beta(n_{\min})$.

The proof above theorem is provided in Appendix 4.7.2 and 4.7.3. The proof relies on the analyzing the tails of the posterior probability distribution given by $\min\{F_{\Theta|Y_1^t}\left(\theta\middle|y_1^t\right), 1 - F_{\Theta|Y_1^t}\left(\theta\middle|y_1^t\right)\}$ for $\theta \in (0,1)$, which is inspired by the analysis in [3]. However, the analysis of the expected value of decay of the tails is based on the analysis of Burnashev and Zigangirov in [4].

Corollary 7 (Periodic arrival times)**.** *Consider the causal PM strategy with parameter λ over a BSC(p). The i-th bit arrives at the encoder at time $T_i = n(i-1)+1$, i.e., the inter-arrival time is constant $n \in \mathbb{N}$. Then, for $\lambda = \lambda^*(n)$, the probability of error in decoding the first $j \in [\lfloor t/n \rfloor]$ bits of a message after t channel uses is bounded by*

$$\mathsf{P}\left(\hat{s}_1^j(t) \neq s_1^j\right) \leq \kappa\left(2^{-\beta(n)(t-n(j-1))}\right), \tag{4.24}$$

where $\beta(n)$ is the solution of (4.6) *for n, and where $0 \leq \kappa < \infty$.*

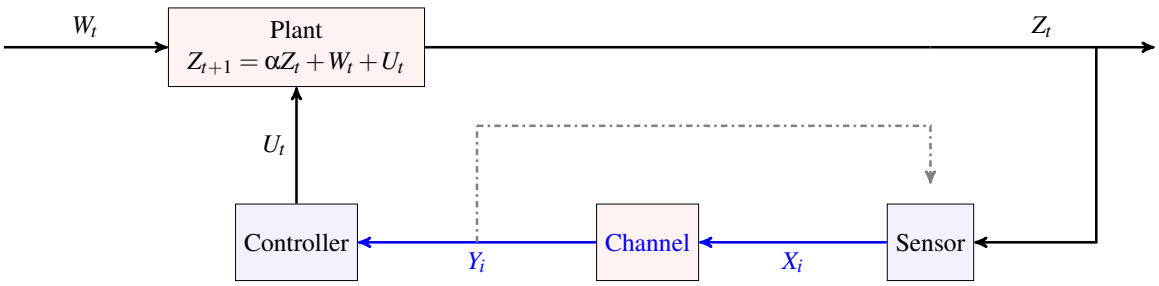


Figure 4.2: A scalar linear plant that is controlled over a noisy channel. The Sensor transmits to the controller over a noisy channel with feedback; n channel uses $\{X_i\}$ per control sample Z_t are assumed.

4.5 Application to Control over Noisy Channels

Consider the problem of stabilizing an unstable scalar plant,

$$Z_{t+1} = \alpha Z_t + W_t + U_t, \tag{4.25}$$

where $\alpha > 1$, the initial state is a random variable $Z_0 \in [-\Delta, \Delta]$, the disturbances $\{W_t\}_{t\geq 0}$ are i.i.d. with a bounded support $W_t \in [-W, W]$ and U_t is a control signal applied by the controller at time t. The controller, that generates U_t, is separated from the sensor that measures Z_t by a BSC(p) with feedback, i.e., n channel uses per each control sample Z_t are available. For $\eta \geq 1$, we want to stabilize the η-th moment, i.e., $\sup_t \mathbb{E}[|Z_t|^\eta] < \infty$. To that end, suppose the observer quantizes the plant measurements into 1 bit, which implies a new bit arrives after every n channel uses. This is a special case of our strategy where the inter-arrival time of the bits is fixed (recall 10). This model is depicted in 4.2.

Remark 13. For the ease of exposition, we consider a 1 bit quantizer but the strategy can be extended to a k-bit quantizer for any $1 \leq k \leq n$.

To stabilize the plant it suffices to apply a control signal $U_t = -\alpha \hat{Z}_t$, where $\{\hat{Z}_t\}_{t\geq 1}$ satisfies $\sup_t \mathbb{E}[|Z_t - \hat{Z}_t|^\eta] < \infty$. The following corollary provides the values α for which the plant can be stabilized.

Corollary 8. *Consider the plant of* (4.25) *for* $\alpha > 1$ *observed through a BSC*(p) *with feedback with a budget of n channel uses. Then, for all* $\eta \geq 1$*, the plant is* η*-stabilizable, i.e.,* $\sup_t \mathbb{E}\left[\left|Z_t - \hat{Z}_t\right|^{\eta}\right] < \infty$*, for*

$$\log \alpha \leq \min\left\{1/n, \beta(n)/\eta\right\}, \tag{4.26}$$

where $\beta(n)$ *is the solution of* (4.6)*.*

Proof. We use the causal PM strategy to transmit the quantized plant measurements over a BSC(p) with feedback. This is a special case of our causal PM strategy where the inter-arrival time of the bits is a constant n, hence we set $\lambda = \lambda^*(n)$. For each step of the plant evolution we convey one bit over n channel uses. Corollary 7 provides the following guarantees on the estimates generated by the causal PM strategy

$$\mathsf{P}\left(\hat{s}_1^j(nt) \neq s_1^j\right) \leq \kappa\left(2^{-\beta(n)n(t-j)}\right),$$

for all $j \in [t]$, where $\beta(n)$ is the solution of (4.6). Hence, using [54, Theorem 4.1] we have that the plant is η-stabilizable if (4.26) holds. □

Remark 14. The constraint $\log \alpha < 1/n \leq 1$ is due to a 1-bit quantization requirement that we implicitly impose by assuming that a single bit arrives at a time. This requirement can be lifted by allowing higher quantization rates, along with the appropriate adaptation of the proposed scheme, at the price of reducing the error exponent β. In other words, two conflicting effects can be seen in the problem of stabilizing an unstable plant over a noisy channel: (i) Source quantization: we wish to maximize the quantization resolution to allow for finer source approximation, however this results in higher channel-coding rate since more bits have to be sent over a given channel budget n (ii) Channel coding: we wish to minimize the channel-coding rate to minimize the error due to decoding, i.e., to maximize the error exponent. These two effects are manifested by the

two minimands in (4.26).

Remark 15. As a consequence of 8 , for a given $\eta \geq 1$ and $p \in (0, 1/2)$, we obtain a lower bound $R(p)$ on the maximum rate (i.e., minimum channel budget $\lceil 1/R(p) \rceil$) at which the communication channel BSC(p) can be operated such that the plant (4.25) is η-stabilizable for some $\alpha > 1$. Using (4.6), note that we have

$$\min\left\{\frac{1}{n}, \frac{\beta(n)}{\eta}\right\} \geq \max_{\beta>0} \min\left\{\frac{\beta}{\eta}, \frac{1}{\eta}\left(\psi^*(\beta) - \frac{1}{n}\right), \frac{1}{n}\right\}$$

Hence, using 8, this implies that $R(p)$ is the largest $R > 0$ that satisfies the following equation:

$$\psi^*(\eta R) = (\eta + 1)R. \tag{4.27}$$

In other words, we obtain that $2^{R(p)}$ is a lower bound on the largest α for which the plant (4.25) can be η-stabilized over a BSC(p) for any channel budget $n > 1$.

4.5.1 Simulations for Control over Noisy Channels

We compare the performance of the proposed causal PM strategy with previously proposed upper and lower bounds for the maximal value of α for which the plant (4.25) can be stabilized. 4.3 compares the stabilizability of a system as a function of the crossover probability of a BSC. The empirical as well as the theoretical performance of both the causal PM-based strategy and a strategy proposed by Simsek et al. [57] (albeit for the interference-free case: $W_t \equiv 0$), as well as the Sahai–Mitter lower bound without feedback (anytime-reliable tree codes) of [54] and the capacity upper bound are illustrated. From 4.3 we see that the bound $2^{R(p)}$ on α, provided by our analysis of the causal PM-based scheme, is rather conservative in comparison to its empirical performance. The latter clearly outperforms the Simsek et al. strategy [57] (for which the analysis is rather tight) and exceeds the Sahai–Mitter lower bound. This demonstrates that the causal

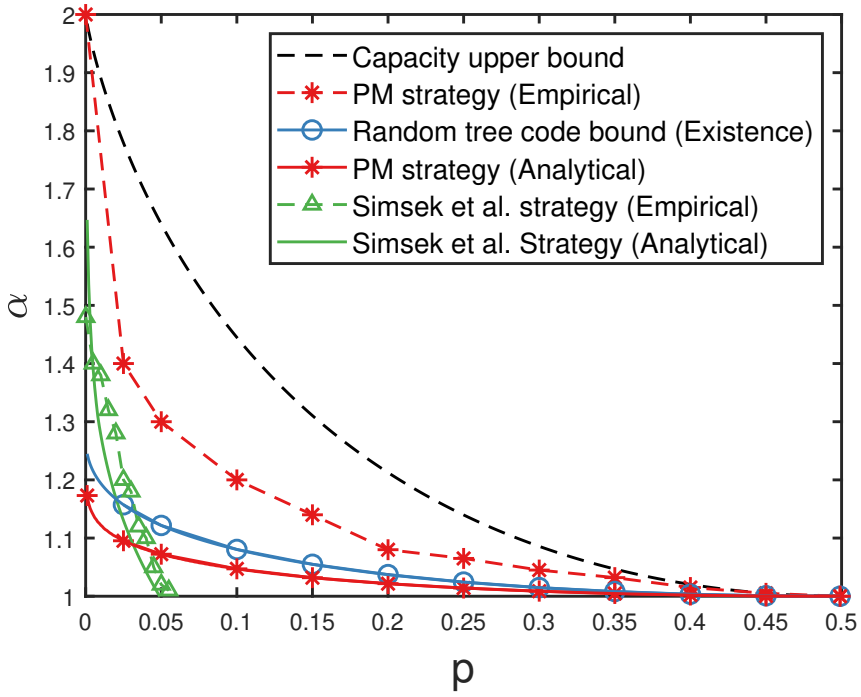


Figure 4.3: The maximum eigenvalue α of a plant that is stabilizable over a BSC(p) as a function of p using: the causal PM strategy (analytically and empirically), the Simsek et al.strategy [57] (analytically and empirically), the Sahai–Mitter tree-code lower bound [54], and the capacity upper bound.

PM-based strategy provides better performance both in terms of stability and complexity. We further note that the causal PM-based scheme can stabilize the plant for α values that are strictly greater than one for all crossover probabilities $p \in [0, 1/2)$, even under the provided conservative analysis. This is in stark contrast to the strategy of Simsek et al., which can stabilize unstable plants only below a certain threshold crossover probability.

4.6 Conclusions and Future Work

We considered the problem of transmitting an infinite stream of bits over a BSC where the bits are revealed to the transmitter causally and the inter bit-arrival time may be random. We proposed a causal PM strategy and provided guarantees for the error exponent of the decoded bits

using this strategy. The causal PM is parameterized by $\lambda(n)$ which decides the randomization of the encoding functions. Hence, it implicitly decides the number of bits decoded and their error exponent. We derived explicit results for two extremes of $\lambda(n)$. An interesting area of future work would be to extend our analysis to any λ between these two extremes. Another important future direction is to extend our analysis to the case where the bit arrival times are unknown at the receiver.

Furthermore, we applied our strategy to the problem of stabilizing a control plant over a BSC. We provided analytical guarantees on the maximal plant eigenvalue for which the plant can be stabilized using causal posterior matching. Closing the gap between our analysis and the empirical performance is an important area of future.

4.7 Appendix

4.7.1 Preliminaries

Proof of Lemma 8

For $k \in \{0,\dots,2^i-1\}$, note that $F_{\Theta|Y_1^t}(k2^{-i}|y_1^t) = \mathsf{P}_{\Theta|Y_1^t}(\Theta \in [0,k2^{-i}) \mid y_1^t)$. For any $y \in \{0,1\}$ after observing $Y_{t+1} = y$, the decoder updates the posterior distribution of Θ using the Bayes rule as follows

$$F_{\Theta|Y_1^{t+1}}(k2^{-i}|y_1^t, Y_{t+1} = y) = \frac{F_{\Theta|Y_1^t}(k2^{-i}|y_1^t)\mathsf{P}(Y_{t+1} = y \mid \Theta \in [0,k2^{-i}], y_1^t)}{\mathsf{P}(Y_{t+1} = y \mid y_1^t)}. \tag{4.28}$$

Let $Y_{t+1} = 1$ and the case where $Y_{t+1} = 0$ can be obtained similarly. First consider the case where median is approximated as the left end point of the $k_{b(t+1)}^{(t)}$th i.e., $k_{b(t+1)}^{(t)} 2^{-b(t+1)}$. Then recall that the encoding function is given by equation (4.15). For all $i \in \{1,\dots,b(t+1)\}$ since

$$k_i^{(t)} 2^{-i} \le k_{b(t+1)}^{(t)} 2^{-b(t+1)} \tag{4.29}$$

we have $\mathsf{P}(Y_{t+1}=1 \mid \Theta \in [0,k2^{-i}], y_1^t) = p$ for all $k \in \{0,\ldots,k_i^{(t)}\}$. Furthermore, we have

$$\mathsf{P}(Y_{t+1}=1 \mid y_1^t) = \mathsf{P}(Y_{t+1}=1 \mid \Theta \in [0,k_{b(t+1)}^{(t)}2^{-b(t+1)}), y_1^t) F_{\Theta|Y_1^t}(k_{b(t+1)}^{(t)}2^{-b(t+1)} \mid y_1^t) \quad (4.30)$$

$$+ \mathsf{P}(Y_{t+1}=1 \mid \Theta \in [k_{b(t+1)}^{(t)}2^{-b(t+1)}, 1), y_1^t)(1 - F_{\Theta|Y_1^t}(k_{b(t+1)}^{(t)}2^{-b(t+1)} \mid y_1^t)) \quad (4.31)$$

$$= p\left(\frac{1}{2} - d_1^{(t)}\right) + \bar{p}\left(\frac{1}{2} + d_1^{(t)}\right) \quad (4.32)$$

$$= \frac{1}{2} + (\bar{p}-p)d_1^{(t)}. \quad (4.33)$$

Hence, for all $k \in \{0,\ldots,k_i^{(t)}\}$ we obtain

$$F_{\Theta|Y_1^{t+1}}(k2^{-i}|y_1^{t+1}) = \frac{F_{\Theta|Y_1^t}(k2^{-i}|y_1^t)p}{\frac{1}{2} + (\bar{p}-p)d_1^{(t)}}. \quad (4.34)$$

Similarly we can obtain the Bayes update for $1 - F_{\Theta|Y_1^{t+1}}(k2^{-i}|y_1^{t+1})$ when $k \in \{k_i^{(t)}+1,\ldots,2^i-1\}$. Now consider the case where median is approximated as the right end point of the $k_{b(t+1)}^{(t)}$th i.e., $(k_{b(t+1)}^{(t)}+1)2^{-b(t+1)}$. Then recall that the encoding function is given by equation (4.16). Again noting that

$$k_i^{(t)}2^{-i} \leq k_{b(t+1)}^{(t)}2^{-b(t+1)} \quad (4.35)$$

$$(k_{b(t+1)}^{(t)}+1)2^{-b(t+1)} \leq (k_i^{(t)}+1)2^{-i} \quad (4.36)$$

we have $\mathsf{P}(Y_{t+1}=1 \mid \Theta \in [0,k2^{-i}], y_1^t) = p$ for all $k \in \{0,\ldots,k_i^{(t)}\}$ and $\mathsf{P}(Y_{t+1}=1 \mid y_1^t) = \frac{1}{2} - (\bar{p}-p)d_2^{(t)}$. The Bayesian update for the rest of the all cases can be obtained similarly. Note that the proof relies on the fact that $i \leq b(t+1)$, hence for bits which have not arrived by $t+1$ the update of the posterior probabilities may not be dictated by the assertion of the lemma.

4.7.2 Proof of Theorem 6 Part (i)

Consider the sample path where the inter-bit arrival realizations $\{N_j\}_{j\geq 1}$ are $\{n_j\}_{j\geq 1}$. Let $t_i := \sum_{j=1}^{i-1} n_j + 1$. Fix $i \in \{1,\ldots,\frac{t}{n_{\min}}\}$ and consider the event where $T_i = t_i < t$. This is the case where first i bits arrive by time instant t with non-zero probability. Recall that $k_i^{(t)}$ denotes the index of the bin containing the median $F^{-1}_{\Theta|Y^t}\left(\frac{1}{2}\right)$ in grid with resolution 2^{-i} i.e.,

$$k_i^{(t)} 2^{-i} \leq F^{-1}_{\Theta|Y^t}\left(\frac{1}{2}\right) < (k_i^{(t)}+1)2^{-i}. \tag{4.37}$$

We bound the probability of error in decoding first i bits conditioned on the event $\{N_1^{i-1} = n_1^{i-1}\} \cap \{T_i < t\}$

$$\begin{aligned}
&\mathsf{P}\left(\hat{s}_1^i(t) \neq s_1^i \middle| N_1^{i-1} = n_1^{i-1}, T_i < t\right) && (4.38)\\
&= \mathsf{P}\left(\Theta \notin \left(k_i^{(t)}2^{-i}, (k_i^{(t)}+1)2^{-i}\right] \middle| N_1^{i-1} = n_1^{i-1}, T_i < t\right) && (4.39)\\
&= \mathbb{E}\left[F_{\Theta|Y^t}\left(k_i^{(t)}2^{-i}\right) + 1 - F_{\Theta|Y^t}\left((k_i^{(t)}+1)2^{-i}\right) \middle| N_1^{i-1} = n_1^{i-1}, T_i < t\right]. && (4.40)
\end{aligned}$$

For any $\tau \geq 1$, define

$$\xi_{\Theta|Y^\tau}(x) := \min\left\{F_{\Theta|Y^\tau}(x), 1 - F_{\Theta|Y^\tau}(x)\right\} \tag{4.41}$$

for $x = k2^{-i}$ where $k \in \{0,\ldots,2^i-1\}$. Since $F_{\Theta|Y^t}\left(k_i^{(t)}2^{-i}\right) \leq \frac{1}{2}$ and $1 - F_{\Theta|Y^t}\left((k_i^{(t)}+1)2^{-i}\right) \leq \frac{1}{2}$, observe that

$$\xi_{\Theta|Y^t}\left(k_i^{(t)}2^{-i}\right) = F_{\Theta|Y^t}\left(k_i^{(t)}2^{-i}\right)$$

and that

$$\xi_{\Theta|Y^t}\left((k_i^{(t)}+1)2^{-i}\right) = 1 - F_{\Theta|Y^t}\left((k_i^{(t)}+1)2^{-i}\right).$$

Hence, for every $\beta > 0$ we can write

$$\mathsf{P}\left(\hat{s}_1^i(t) \neq s_1^i \middle| N_1^{i-1} = n_1^{i-1}, T_i < t\right) \tag{4.42}$$

$$= \mathbb{E}\left[\xi_{\Theta|Y^t}\left(k_i^{(t)}2^{-i}\right) + \xi_{\Theta|Y^t}\left((k_i^{(t)}+1)2^{-i}\right) \middle| N_1^{i-1} = n_1^{i-1}, T_i < t\right] \tag{4.43}$$

$$\leq 2^{-\beta(t-t_i)} + \mathsf{P}\left(\xi_{\Theta|Y^t}\left((k_i^{(t)}+1)2^{-i}\right) + \xi_{\Theta|Y^t}\left(k_i^{(t)}2^{-i}\right) \geq 2^{-\beta(t-t_i)} \middle| N_1^{i-1} = n_1^{i-1}, T_i < t\right). \tag{4.44}$$

The second term in the above equation can be bounded using Markov's inequality for any $\lambda > 0$ as follows

$$\mathsf{P}\left(\xi_{\Theta|Y^t}\left((k_i^{(t)}+1)2^{-i}\right) + \xi_{\Theta|Y^t}\left(k_i^{(t)}2^{-i}\right) \geq 2^{-\beta(t-t_i)} \middle| N_1^{i-1} = n_1^{i-1}, T_i < t\right) \tag{4.45}$$

$$\leq 2^{\lambda\beta(t-t_i)}\mathbb{E}\left[\left(\xi_{\Theta|Y^t}\left(k_i^{(t)}2^{-i}\right) + \xi_{\Theta|Y^t}\left((k_i^{(t)}+1)2^{-i}\right)\right)^{\lambda} \middle| N_1^{i-1} = n_1^{i-1}, T_i < t\right]. \tag{4.46}$$

Furthermore, note that

$$\mathbb{E}\left[\left(\xi_{\Theta|Y^t}\left(k_i^{(t)}2^{-i}\right) + \xi_{\Theta|Y^t}\left((k_i^{(t)}+1)2^{-i}\right)\right)^{\lambda} \middle| N_1^{i-1} = n_1^{i-1}, T_i < t\right] \tag{4.47}$$

$$\leq 2^{\lambda}\mathbb{E}\left[\max_k \xi^{\lambda}_{\Theta|Y^t}\left(k2^{-i}\right) \middle| N_1^{i-1} = n_1^{i-1}, T_i < t\right]$$

$$\leq 2^{\lambda}\mathbb{E}\left[\sum_{k=0}^{2^i-1} \xi^{\lambda}_{\Theta|Y^t}\left(k2^{-i}\right) \middle| N_1^{i-1} = n_1^{i-1}, T_i < t\right]$$

$$= 2^{\lambda}\sum_{k=0}^{2^i-1} \mathbb{E}\left[\xi^{\lambda}_{\Theta|Y^t}\left(k2^{-i}\right) \middle| N_1^{i-1} = n_1^{i-1}, T_i < t\right]. \tag{4.48}$$

Fix some $k \in \{0,\dots,2^i-1\}$. Then, for any $t_i \le \tau \le t$ consider

$$\mathbb{E}\left[\xi^{\lambda}_{\Theta|Y^{\tau}}(k2^{-i})\middle| N_1^{i-1}=n_1^{i-1}, T_i<t\right] \tag{4.49}$$

$$=\mathbb{E}\left[2^{\lambda\log\xi_{\Theta|Y^{\tau}}(k2^{-i})}\middle| N_1^{i-1}=n_1^{i-1}, T_i<t\right] \tag{4.50}$$

$$=\mathbb{E}\left[2^{\lambda\log\xi_{\Theta|Y^{\tau-1}}(k2^{-i})}\,\mathbb{E}\left[2^{\lambda\log\frac{\xi_{\Theta|Y^{\tau}}(k2^{-i})}{\xi_{\Theta|Y^{\tau-1}}(k2^{-i})}}\middle| Y^{\tau-1}\right]\middle| N_1^{i-1}=n_1^{i-1}, T_i<t\right] \tag{4.51}$$

$$\overset{(a)}{\le} 2^{-\psi(\lambda)}\mathbb{E}\left[2^{\lambda\log\xi_{\Theta|Y^{\tau-1}}(k2^{-i})}\middle| N_1^{i-1}=n_1^{i-1}, T_i<t\right], \tag{4.52}$$

where $\psi(\lambda)=-\log\left(\frac{(2p)^{\lambda}+(2\bar{p})^{\lambda}}{2}\right)$ and (a) obtained by applying Lemma 9.

Conditioned on the event $\{N_1^{i-1}=n_1^{i-1}\}\cap\{T_i<t\}$, for all $\tau \ge t_i$ the causal PM strategy operates on a grid whose resolution is finer than 2^{-i}. As discussed in Remark 12 this implies that $\xi_{\Theta|Y^{\tau}}(k2^{-i})$ is updated for all $t_i \le \tau \le t$. Therefore, applying Lemma 9 repeatedly for all time instants $t_i \le \tau \le t$, we obtain the following for every $k \in \{0,\dots,2^i-1\}$

$$\mathbb{E}\left[\xi^{\lambda}_{\Theta|Y^{t}}(k2^{-i})\middle| N_1^{i-1}=n_1^{i-1}, T_i<t\right] \le 2^{-\psi(\lambda)(t-t_i)}\mathbb{E}\left[\log\xi^{\lambda}_{\Theta|Y^{t_{i-1}}}(k2^{-i})\middle| N_1^{i-1}=n_1^{i-1}, T_i<t\right]. \tag{4.53}$$

Substituting equation (4.53) in equation (4.47) we have

$$\mathbb{E}\left[\left(\xi_{\Theta|Y^{t}}\left(k_i^{(t)}2^{-i}\right)+\xi_{\Theta|Y^{t}}\left((k_i^{(t)}+1)2^{-i}\right)\right)^{\lambda}\middle| N_1^{i-1}=n_1^{i-1}, T_i<t\right] \tag{4.54}$$

$$\le 2^{-\psi(\lambda)(t-t_i)}\sum_{k=0}^{2^i-1}\mathbb{E}\left[\xi^{\lambda}_{\Theta|Y^{t_{i-1}}}(k2^{-i})\middle| N_1^{i-1}=n_1^{i-1}, T_i<t\right]. \tag{4.55}$$

Now, applying Lemma 10 for λ such that $\psi(\lambda) - \frac{1}{n_{\min}} > 0$ we have

$$\mathbb{E}\left[\left(\xi_{\Theta|Y^t}\left(k_i^{(t)}2^{-i}\right) + \xi_{\Theta|Y^t}\left((k_i^{(t)}+1)2^{-i}\right)\right)^{\lambda}\middle| N_1^{i-1} = n_1^{i-1}, T_i < t\right] \tag{4.56}$$

$$\leq 2^{-\psi(\lambda)(t-t_i)} \frac{1}{1-2^{-\left(\psi(\lambda)-\frac{1}{n_{\min}}\right)n_{\min}}}. \tag{4.57}$$

Substituting the above inequality in equation (4.42), the probability of error in decoding first i bits conditioned on the event $\{N_1^{i-1} = n_1^{i-1}\} \cap \{T_i < t\}$ for all $\lambda > 0$ such that $\psi(\lambda) - \frac{1}{n_{\min}} > 0$ is given by

$$\mathsf{P}\left(\hat{s}_1^i(t) \neq s_1^i \middle| N_1^{i-1} = n_1^{i-1}, T_i < t\right) \leq 2^{-\beta(t-t_i)} + \frac{2^{-(\psi(\lambda)-\lambda\beta)(t-t_i)}}{1-2^{-\left(\psi(\lambda)-\frac{1}{n_{\min}}\right)n_{\min}}}. \tag{4.58}$$

For any $\beta > 0$, if there exists a $\lambda > 0$ such that $\psi(\lambda) - \lambda\beta - \frac{1}{n_{\min}} > 0$, then we have $\psi(\lambda) - \frac{1}{n_{\min}} > 0$ and $\psi(\lambda) - \lambda\beta > 0$. Also, if $\sup_{\lambda>0}(\psi(\lambda) - \lambda\beta) - \frac{1}{n_{\min}} > 0$ then the supremum achieving λ also satisfies $\psi(\lambda) - \frac{1}{n_{\min}} > 0$. Therefore, we fix $\lambda = \lambda^*(n_{\min})$. Hence, we obtain the following

$$\mathsf{P}\left(\hat{s}_1^i(t) \neq s_1^i \middle| N_1^{i-1} = n_1^{i-1}, T_i < t\right) \tag{4.59}$$

$$\leq 2^{-\beta(t-t_i)} + \kappa' 2^{-\left(\psi^*(\beta)-\frac{1}{n_{\min}}\right)(t-t_i)} \tag{4.60}$$

$$\leq \kappa 2^{-\max_{\beta>0}\min\left\{\beta,\psi^*(\beta)-\frac{1}{n_{\min}}\right\}(t-t_i)}, \tag{4.61}$$

$$\leq \kappa 2^{-\beta(n_{\min})(t-t_i)}. \tag{4.62}$$

for some positive constants $\kappa', \kappa < \infty$ independent of t. Now, note that the event $\{N_1^{i-1} = n_1^{i-1}\} \cap \{T_i < t\}$ is equivalent to the event $\{N_1^{i-1} = n_1^{i-1}\} \cap \{b(t) > i\}$. Taking the expectation with respect to $p_N^{\otimes i}$ we have

$$\mathsf{P}\left(\hat{s}_1^i(t) \neq s_1^i \middle| b(t) > i\right) \leq \kappa \mathbb{E}\left[2^{-\beta(n_{\min})(t-T_i)}\middle| b(t) > i\right], \tag{4.63}$$

for all $i \in \left[\left\lceil \frac{t}{n_{\min}} \right\rceil\right]$.

4.7.3 Proof of Theorem 6 Part (ii)

Now following the steps as in the proof of Theorem 6 part(i) we obtain the following

$$\mathbb{E}\left[\xi_{\Theta|Y^t}\left(k_i^{(t)}2^{-i}\right)+\xi_{\Theta|Y^t}\left((k_i^{(t)}+1)2^{-i}\right)\right]^{\lambda} \leq 2^{\lambda}2^{-\psi(\lambda)(t-t_i)}\sum_{k=0}^{2^i-1}\mathbb{E}\left[\log \xi^{\lambda}_{\Theta|Y^{t_i-1}}(k2^{-i})\right]. \tag{4.64}$$

Now, applying Lemma 11 for λ such that $\psi(\lambda)-\frac{1}{n_{\max}}>0$ we have

$$\mathbb{E}\left[\xi_{\Theta|Y^t}\left(k_i^{(t)}2^{-i}\right)+\xi_{\Theta|Y^t}\left((k_i^{(t)}+1)2^{-i}\right)\right]^{\lambda} \leq 2^{-\psi(\lambda)(t-t_i)}\frac{1}{1-2^{-\left(\psi(\lambda)-\frac{1}{n_{\max}}\right)n_{\max}}}. \tag{4.65}$$

Following the same steps as in the proof of Theorem 6 part (i) and setting $\lambda=\lambda^*(n_{\max})$, for $i\in\{1,\ldots,\lceil\frac{t}{n_{\max}}\rceil\}$ we obtain the following

$$\mathsf{P}\left(\hat{s}_1^i(t)\neq s_1^i\right) \leq \kappa 2^{-\max_{\beta>0}\min\{\beta,\psi^*(\beta)-\frac{1}{n_{\max}}\}(t-n_{\max}(i-1))} \tag{4.66}$$

$$= \kappa 2^{-\beta(n_{\max})(t-n_{\max}(i-1))}, \tag{4.67}$$

some positive constant $\kappa<\infty$ independent of t.

4.7.4 Technical Background

In this appendix, we provide some preliminary lemmata which are technical and only helpful in proving the main results of the chapter.

Lemma 9. *For $t\geq 1$ consider $i\in\left[\left\lceil\frac{t}{n_{\min}}\right\rceil\right]$. Consider a sample path where $N_1^{i-1}=n_1^{i-1}$ and $b(t)>i$. Then for all τ such $t_i-1=\sum_{j=1}^{i-1}n_j\leq\tau\leq t$, for any point $x=k2^{-i}$, where $k\in\{0,\ldots,2^i-1\}$,*

and $0 < \lambda \leq 1$ *the following holds true*

$$\mathbb{E}\left[\left(\frac{\xi_{\Theta|Y^{\tau+1}}(x)}{\xi_{\Theta|Y^{\tau}}(x)}\right)^{\lambda}\middle| Y^{\tau}, N_1^{i-1} = n_1^{i-1}, T_i < t\right] \leq \frac{(2p)^{\lambda} + (2\bar{p})^{\lambda}}{2}, \tag{4.68}$$

when randomization probabilities are chosen as follows

$$\pi_1^{(\tau+1)}(\lambda) = \frac{h(\lambda, d_2^{(\tau)})}{h(\lambda, d_1^{(\tau)}) + h(\lambda, d_2^{(\tau)})}, \tag{4.69}$$

$$\pi_2^{(\tau+1)}(\lambda) = 1 - \pi_1^{(\tau+1)}(\lambda), \tag{4.70}$$

where $d_1^{(\tau)}$ *and* $d_2^{(\tau)}$ *are as defined in equation* (4.12) *respectively and* (4.13) *and* $h(\lambda, d)$ *is defined in equation* (4.14).

Proof. The proof of this lemma is based on the analysis of moment generated function provided by Burnashev and Zigangirov in [4].

Case 1: Suppose the median does not cut any bin i.e., median coincides with the end point of some bin.

Case 1a: Using Lemma 8, the update of $\xi_{\Theta|Y^{\tau}}(x)$ when $F_{\Theta|Y^{\tau}}(x) \leq \frac{1}{2}$ is given as follows

$$\frac{\xi_{\Theta|Y^{\tau+1}}(x)}{\xi_{\Theta|Y^{\tau}}(x)} \leq \begin{cases} 2p & \text{if } Y_{\tau+1} = 1 \quad \text{w.p. } \frac{1}{2}, \\ 2\bar{p} & \text{if } Y_{\tau+1} = 0 \quad \text{w.p. } \frac{1}{2}. \end{cases} \tag{4.71}$$

Case 1b: Using Lemma 8, the update of $\xi_{\Theta|Y^{\tau}}(x)$ when $1 - F_{\Theta|Y^{\tau}}(x) < \frac{1}{2}$ is given as follows

$$\frac{\xi_{\Theta|Y^{\tau+1}}(x)}{\xi_{\Theta|Y^{\tau}}(x)} \leq \begin{cases} 2\bar{p} & \text{if } Y_{\tau+1} = 1 \quad \text{w.p. } \frac{1}{2}, \\ 2p & \text{if } Y_{\tau+1} = 0 \quad \text{w.p. } \frac{1}{2}. \end{cases} \tag{4.72}$$

Hence, for all $x = k2^{-i}$ where $k \in \{0,\ldots,2^i-1\}$ under Case 1 we have

$$\mathbb{E}\left[\left(\frac{\xi_{\Theta|Y^\tau}(x)}{\xi_{\Theta|Y^{\tau-1}}(x)}\right)^\lambda \middle| Y^\tau, N_1^{i-1} = n_1^{i-1}\right] \leq f_1(\lambda) := \frac{(2p)^\lambda + (2\bar{p})^\lambda}{2}. \tag{4.73}$$

Case 2: When median lies inside some bin we randomize the encoding.

Case 2a: Consider $x = k2^{-i}$ where $k \in \{0,\ldots,2^i-1\}$ such that $\xi_{\Theta|Y^\tau}(x) = F_{\Theta|Y^\tau}(x) \leq \frac{1}{2}$. Using Lemma 8 with probability $\pi_1^{(\tau+1)}$ the update is given as

$$\frac{\xi_{\Theta|Y^{\tau+1}}(x)}{\xi_{\Theta|Y^\tau}(x)} \leq \begin{cases} \frac{p}{\frac{1}{2}+(\bar{p}-p)d_1^{(\tau)}} & \text{if } Y_{\tau+1}=1, \\ \frac{\bar{p}}{\frac{1}{2}-(\bar{p}-p)d_1^{(\tau)}} & \text{if } Y_{\tau+1}=0 \end{cases} \tag{4.74}$$

where the probability of $\mathsf{P}(Y_{\tau+1}=1 \mid y_1^\tau) = \frac{1}{2}+(\bar{p}-p)d_1^{(\tau)}$ and $\mathsf{P}(Y_{\tau+1}=0 \mid y_1^\tau) = \frac{1}{2}-(\bar{p}-p)d_1^{(\tau)}$. Similarly with probability $\pi_2^{(\tau+1)}$ the update is given as

$$\frac{\xi_{\Theta|Y^{\tau+1}}(x)}{\xi_{\Theta|Y^\tau}(x)} \leq \begin{cases} \frac{p}{\frac{1}{2}-(\bar{p}-p)d_2^{(\tau)}} & \text{if } Y_{\tau+1}=1, \\ \frac{\bar{p}}{\frac{1}{2}+(\bar{p}-p)d_2^{(\tau)}} & \text{if } Y_{\tau+1}=0, \end{cases} \tag{4.75}$$

where $\mathsf{P}(Y_{\tau+1}=1 \mid y_1^\tau) = \frac{1}{2}-(\bar{p}-p)d_2^{(\tau)}$ and $\mathsf{P}(Y_{\tau+1}=0 \mid y_1^\tau) = \frac{1}{2}+(\bar{p}-p)d_2^{(\tau)}$. Hence, for all $x = k2^{-i}$ where $k \in \{0,\ldots,2^i-1\}$ such that $\xi_{\Theta|Y^\tau}(x) = F_{\Theta|Y^\tau}(x) \leq \frac{1}{2}$, we have

$$\mathbb{E}\left[\left(\frac{\xi_{\Theta|Y^{\tau+1}}(x)}{\xi_{\Theta|Y^\tau}(x)}\right)^\lambda \middle| Y^\tau, N_1^{i-1} = n_1^{i-1}\right] \leq f_2(\lambda) := \pi_1^{(\tau+1)} g_\lambda\left(d_1^{(\tau)}\right) + \pi_2^{(\tau+1)} g_\lambda\left(-d_2^{(\tau)}\right), \tag{4.76}$$

where we define

$$g_\lambda(d) := \frac{p^\lambda}{\left(\frac{1}{2}+(\bar{p}-p)d\right)^{\lambda-1}} + \frac{\bar{p}^\lambda}{\left(\frac{1}{2}-(\bar{p}-p)d\right)^{\lambda-1}}. \tag{4.77}$$

Case 2b: Consider $x = k2^{-i}$ where $k \in \{0,\ldots,2^i-1\}$ such that $\xi_{\Theta|Y^\tau}(x) = 1 - F_{\Theta|Y^\tau}(x) \leq \frac{1}{2}$. Using Lemma 8 with probability $\pi_1^{(\tau+1)}$ the update is given as

$$\frac{\xi_{\Theta|Y^{\tau+1}}(x)}{\xi_{\Theta|Y^\tau}(x)} \leq \begin{cases} \frac{\bar{p}}{\frac{1}{2}+(\bar{p}-p)d_1^{(\tau)}} & \text{if } Y_{\tau+1}=1, \\ \frac{p}{\frac{1}{2}-(\bar{p}-p)d_1^{(\tau)}} & \text{if } Y_{\tau+1}=0, \end{cases} \tag{4.78}$$

where $\mathsf{P}(Y_{\tau+1}=1 \mid y_1^\tau) = \frac{1}{2}+(\bar{p}-p)d_1^{(\tau)}$ and $\mathsf{P}(Y_{\tau+1}=0 \mid y_1^\tau) = \frac{1}{2}-(\bar{p}-p)d_1^{(\tau)}$. Similarly, with probability $\pi_2^{(\tau+1)}$ the update is given as

$$\frac{\xi_{\Theta|Y^{\tau+1}}(x)}{\xi_{\Theta|Y^\tau}(x)} \leq \begin{cases} \frac{\bar{p}}{\frac{1}{2}-(\bar{p}-p)d_2^{(\tau)}} & \text{if } Y_{\tau}=1, \\ \frac{p}{\frac{1}{2}+(\bar{p}-p)d_2^{(\tau)}} & \text{if } Y_{\tau}=0, \end{cases} \tag{4.79}$$

where $\mathsf{P}(Y_{\tau+1}=1 \mid y_1^\tau) = \frac{1}{2}-(\bar{p}-p)d_1^{(\tau)}$ and $\mathsf{P}(Y_{\tau+1}=0 \mid y_1^\tau) = \frac{1}{2}+(\bar{p}-p)d_1^{(\tau)}$. Hence, $x = k2^{-i}$ where $k \in \{0,\ldots,2^i-1\}$ such that $\xi_{\Theta|Y^\tau}(x) = 1 - F_{\Theta|Y^\tau}(x) \leq \frac{1}{2}$, we have

$$\mathbb{E}\left[\left(\frac{\xi_{\Theta|Y^{\tau+1}}(x)}{\xi_{\Theta|Y^\tau}(x)}\right)^{\lambda} \middle| Y^\tau, N_1^{i-1} = n_1^{i-1}\right] \leq f_3(\lambda) := \pi_1^{(\tau+1)} g_\lambda(-d_1^{(\tau)}) + \pi_2^{(\tau+1)} g_\lambda(d_2^{(\tau)}). \tag{4.80}$$

Now we want $\pi_1^{(\tau+1)}$ and $\pi_2^{(\tau+1)}$ which minimize the $\max\{f_2(\lambda), f_3(\lambda)\}$. Hence we choose $\pi_1^{(\tau+1)}$ and $\pi_2^{(\tau+1)}$ for which $f_2(\lambda) = f_3(\lambda)$ and obtain

$$\pi_1^{(\tau+1)} = \frac{g_\lambda\left(d_2^{(\tau)}\right) - g_\lambda\left(-d_2^{(\tau)}\right)}{g_\lambda\left(d_1^{(\tau)}\right) - g_\lambda\left(-d_1^{(\tau)}\right) + g_\lambda\left(d_2^{(\tau)}\right) - g_\lambda\left(-d_2^{(\tau)}\right)}. \tag{4.81}$$

Note that for the above choice of $\pi_1^{(\tau)}$ and $\pi_2^{(\tau)}$ we have

$$\max_{d_1^{(\tau)},d_2^{(\tau)}} \max\{f_2(\lambda), f_3(\lambda)\} \leq \max_{d_1^{(\tau)},d_2^{(\tau)}} \left\{ \frac{f_2(\lambda)+f_3(\lambda)}{2} \right\} \tag{4.82}$$

$$= \frac{1}{2}\pi_1^{(\tau+1)} \max_{d_1^{(\tau)}} \left\{ g_\lambda(d_1^{(\tau)}) + g_\lambda(-d_1^{(\tau)}) \right\} \tag{4.83}$$

$$+ \frac{1}{2}\pi_2^{(\tau+1)} \max_{d_2^{(\tau)}} \left\{ g_\lambda(d_2^{(\tau)}) + g_\lambda(-d_2^{(\tau)}) \right\} \tag{4.84}$$

$$= \max_d \frac{1}{2} \left\{ g_\lambda(d) + g_\lambda(-d) \right\}. \tag{4.85}$$

Furthermore, we have

$$\frac{g_\lambda(d)+g_\lambda(-d)}{2} = \left(\frac{(2p)^\lambda + (2\bar{p})^\lambda}{4} \right) \Big((1+2(\bar{p}-p)d)^{1-\lambda} + \tag{4.86}$$

$$+ (1-2(\bar{p}-p)d)^{1-\lambda} \Big). \tag{4.87}$$

Since $d_1, d_2 \leq \frac{1}{2}$ we have $2(\bar{p}-p)d_1, 2(\bar{p}-p)d_2 \leq 1$. If $0 < \lambda \leq 1$, then using the inequality $(1-x)^\lambda + (1+x)^\lambda \leq 2$, we have

$$\max_d \{ g_\lambda(d) + g_\lambda(-d) \} \leq \frac{(2p)^\lambda + (2\bar{p})^\lambda}{2}. \tag{4.88}$$

Therefore, for all three cases we can upper bound as follows

$$\max_{d_1^{(\tau)},d_2^{(\tau)}} \max\{f_1(\lambda), f_2(\lambda), f_3(\lambda)\} \leq \frac{(2p)^\lambda + (2\bar{p})^\lambda}{2}. \tag{4.89}$$

Hence, we have the assertion of the lemma. □

Lemma 10. *For $t \geq 1$ consider $i \in \{1, \ldots, \lceil \frac{t}{n_{\min}} \rceil\}$. Consider a sample path where $N_1^{i-1} = n_1^{i-1}$ and $b(t) > i$. Then, using causal posterior matching strategy the following holds true for all*

$\lambda > 0$ *such that* $\psi(\lambda) - \frac{1}{n_{\min}} > 0$*:*

$$\sum_{k=1}^{2^i-1} \mathbb{E}\left[\xi^{\lambda}_{\Theta|Y^{t_i-1}}(k2^{-i})\middle| N_1^{i-1} = n_1^{i-1}, T_i < t\right] \leq \frac{1}{1-2^{-\left(\psi(\lambda)-\frac{1}{n_{\min}}\right)n_{\min}}}. \tag{4.90}$$

Proof. The first bit has been encoded for all times instants from 1 to $t_i - 1$, hence for $k = 2^{i-1}$ we have

$$\mathbb{E}\left[\xi^{\lambda}_{\Theta|Y^{t_{i-1}}}(k2^{-i})\middle| N_1^{i-1} = n_1^{i-1}, T_i < t\right] \leq 2^{-\psi(\lambda)(t_i-1)}.$$

Note that there are 2^{i-1} tails which satisfy the above equation. Similarly, for $1 \leq \ell \leq i-1$ the ℓth bit has been encoded since time t_ℓ to $t_i - 1$ we have that there are $2^{i-\ell}$ tails of the remaining ones which are less than $2^{-\psi(\lambda)(t_i-t_\ell)}$. Therefore we have

$$\begin{aligned}
\sum_{k=0}^{2^i-1} \mathbb{E}\left[\xi^{\lambda}_{\Theta|Y^{t_i-1}}(k2^{-i})\middle| N_1^{i-1} = n_1^{i-1}, T_i < t\right] &\leq \sum_{\ell=0}^{i} 2^{i-\ell} 2^{-\psi(\lambda)(t_i-t_\ell)} && (4.91)\\
&\overset{(a)}{\leq} \sum_{\ell=0}^{i} 2^{i-\ell} 2^{-\psi(\lambda) n_{\min}(i-\ell)} && (4.92)\\
&= \sum_{\ell=0}^{i} 2^{-\left(\psi(\lambda)-\frac{1}{n_{\min}}\right)n_{\min}(i-\ell)} && (4.93)\\
&\leq \sum_{\ell=0}^{\infty} 2^{-\left(\psi(\lambda)-\frac{1}{n_{\min}}\right)n_{\min}\ell} && (4.94)\\
&\overset{(b)}{=} \frac{1}{1-2^{-\left(\psi(\lambda)-\frac{1}{n_{\min}}\right)n_{\min}}}, && (4.95)
\end{aligned}$$

where (a) follows from $t_i - t_\ell \geq n_{\min}(i-\ell)$ and (b) follows from that fact that $\psi(\lambda) - \frac{1}{n_{\min}} > 0$. □

Lemma 11. *For $t \geq 1$ consider $i \in \{1, \ldots, \lceil \frac{t}{n_{\max}} \rceil\}$. Let $t_i := n_{\max}(i-1)+1$. Then, using causal posterior matching strategy the following holds true for all $\lambda > 0$ such that $\psi(\lambda) - \frac{1}{n_{\max}} > 0$:*

$$\sum_{k=0}^{2^i-1} \mathbb{E}\left[\xi^{\lambda}_{\Theta|Y^{t_i-1}}(k2^{-i})\right] \leq \frac{1}{1-2^{-\left(\psi(\lambda)-\frac{1}{n_{\max}}\right)n_{\max}}}. \tag{4.96}$$

Proof. The first bit has been encoded for all times instants from 1 to $t_i - 1$, hence for $k = 2^{i-1}$ we have

$$\mathbb{E}\left[\xi^{\lambda}_{\Theta|Y^{T_{i-1}}}(k2^{-i})\right] \leq = 2^{-\psi(\lambda)(t_i-1)}. \tag{4.97}$$

Note that there are 2^{i-1} tails which satisfy the above equation. Similarly, for $1 \leq \ell \leq i-1$ the ℓth bit has been encoded since time t_ℓ to $t_i - 1$ we have that there are $2^{i-\ell}$ tails of the remaining ones which are less than $2^{-\psi(\lambda)(t_i - t_\ell)}$. Noting that $t_i - t_\ell = n_{\max}(i-\ell)$ and following the proof of Lemma 10 we have the assertion of the lemma. □

Chapter 4, in part, is a reprint of the material as it appears in the paper: Anusha Lalitha, Anatoly Khina, Tara Javidi, and Victoria Kostina, “Real-time binary posterior matching", in *IEEE International Symposium on Information Theory*, 2019. The dissertation author was the primary investigator and author of this paper.

Chapter 5

Social Learning and Distributed Hypothesis Testing

5.1 Introduction

Learning in distributed settings is more than a phenomenon of social networks; it is also an engineering challenge for networked system designers. For instance, in today's data networks, many applications need estimates of certain parameters: file-sharing systems need to know the distribution of (unique) documents shared by their users, internet-scale information retrieval systems need to deduce the criticality of various data items, and monitoring networks need to compute aggregates in a duplicate-insensitive manner. Finding scalable, efficient, and accurate methods for computing such metrics (e.g. number of documents in the network, sizes of database relations, distributions of data values) is of critical value in a wide array of network applications.

We consider a network of nodes that sample local observations (over time) governed by an unknown true hypothesis θ^* taking values in a finite discrete set Θ. We model the i-th node's distribution (or local channel, or likelihood function) of the observations conditioned on the true hypothesis by $f_i(\cdot;\theta^*)$ from a collection $\{f_i(\cdot;\theta) : \theta \in \Theta\}$. Nodes neither have access to each

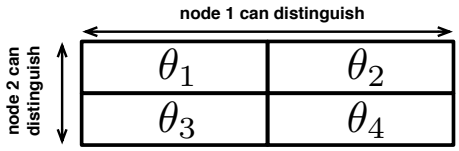


Figure 5.1: Example of a parameter space in which no node can identify the true parameter. There are 4 parameters, $\{\theta_1, \theta_2, \theta_3, \theta_4\}$, and 2 nodes. The node 1 has $f_1(\cdot;\theta_1) = f_1(\cdot;\theta_3)$ and $f_1(\cdot;\theta_2) = f_1(\cdot;\theta_4)$, and the node 2 has $f_2(\cdot;\theta_1) = f_2(\cdot;\theta_2)$ and $f_2(\cdot;\theta_3) = f_2(\cdot;\theta_4)$.

others' observations nor the joint distribution of observations across all nodes in the network. Every node in the network aims to learn the unknown true hypothesis θ^*. A simple two-node example is illustrated in Figure 5.1 – one node can only learn the column in which the true hypothesis lies, and the other can only learn the row. In this example, the local observations of a given node are not sufficient to recover the underlying hypothesis in isolation. In this chapter we study a learning rule that enables the nodes to learn the unknown true hypothesis based on message passing between one hop neighbors (local communication) in the network. In particular, each node performs a local Bayesian update and send its belief vectors (message) to its neighbors. After receiving the messages from the neighbors each node performs a consensus averaging on a reweighting of the *log beliefs*. Our result shows that under our learning rule each node can reject the wrong hypothesis exponentially fast.

We show that the rate of rejection of wrong hypothesis is the weighted sum of Kullback-Leibler (KL) divergences between likelihood function of the true parameter and the likelihood function of the wrong hypothesis, where the sum is over the nodes in the network and the weights are the nodes' influences as dictated by the learning rule. Furthermore, we show that the probability of sample paths on which the rate of rejection deviates from the mean rate vanishes exponentially fast. For any strongly connected network and bounded ratios of log-likelihood functions, we obtain a lower bound on this exponential rate. Furthermore, for any aperiodic network we characterize the exact exponent with which probability of sample paths on which the rate of rejection deviates from the mean rate vanishes (i.e., obtain a large deviation principle) for a broader class of observation statistics which includes distributions with unbounded support

such as Gaussian mixtures and Gamma distribution. The large deviation rate function is shown to be a function of observation model and the nodes' influences on the network as dictated by the learning rule.

Outline of the chapter. The rest of the chapter is organized as follows. We provide the model in Section 5.2 which defines the nodes' observation model and network. This section also contains the learning rule and assumptions on model. We then provide results on rate of convergence and their proofs in Section 5.3. We apply our learning rule to various examples in Section 5.4 and discuss some practical issues in Section 5.4.3. We conclude with a summary in Section 5.5.

5.1.1 Related Work

The literature on distributed learning, estimation and detection can divided into two broad sets. One set deals with the fusion of information observed by a group nodes at a fusion center where the communication links (between the nodes and fusion center) are either rate limited [58–66] or subject to channel imperfections such as fading and packet drops [67–69]. Our work belongs to the second set, which models the communication network as a directed graph whose vertices/nodes are agents and an edge from node i to j indicates that i may send a message to j with perfect fidelity (the link is a noiseless channel of infinite capacity). These "protocol" models study how message passing in a network can be used to achieve a pre-specified computational task such as distributed learning [70, 71], general function evaluation [72],or stochastic approximations [73]. Message passing protocols may be synchronous or asynchronous (such as the "gossip" model [74–78]). This graphical model of the communication, instead of assuming a detailed physical-layer formalization, implicitly assumes a PHY/MAC-layer abstraction where sufficiently high data rates are available to send the belief vectors with desired precision when nodes are within each others' communication range. A missing edge indicates the corresponding link has zero capacity.

Due to the large body of work in distributed detection, estimation and merging of opinions,

we provide a long yet detailed summary of all the related works and their relation to our setup. Readers familiar with these works can skip to Section 5.2 without loss of continuity.

Several works [79–83] consider an update rule which uses local Bayesian updating combined with a linear consensus strategy on the beliefs [84] that enables all nodes in the network identify the true hypothesis. Jadbabaie et al. [79] characterize the "learning rate" of the algorithm in terms of the total variational error across the network and provide an almost sure upper bound on this quantity in terms of the KL-divergences and influence vector of agents. In Corollary 10 we analytically show that the proposed learning rule in this chapter provides a strict improvement over linear consensus strategies [79]. Simultaneous and independent works by Shahrampour et al. [85] and Nedić et al. [86] consider a similar learning rule (with a change of order in the update steps). They obtain similar convergence and concentration results under the assumption of bounded ratios of likelihood functions. Nedić et al. [86] analyze the learning rule for time-varying graphs. Theorem 9 strengthens these results for static networks by providing a large deviation analysis for a broader class of likelihood functions which includes Gaussian mixtures.

Rad and Tahbaz-Salehi [82] study distributed parameter estimation using a Bayesian update rule and average consensus on the log-likelihoods similar to (6.2)–(5.3). They show that the maximum of each node's belief distribution converges in probability to the true parameter under certain analytic assumptions (such as log-concavity) on the likelihood functions of the observations. Our results show almost sure convergence and concentration of the nodes' beliefs when the parameter space is discrete and the log-likelihood function is concave. Kar et al. in [87] consider the problem of distributed estimation of an unknown underlying parameter where the nodes make noisy observations that are non-linear functions of an unknown global parameter. They form local estimates using a quantized message-passing scheme over randomly-failing communication links, and show the local estimators are consistent and asymptotically normal. Note that for any general likelihood model and static strongly connected network, our Theorem 7 strengthens the results of distributed estimation (where the error vanishes inversely with the

square root of total number of observations) by showing exponentially fast convergence of the beliefs. Furthermore, Theorem 8 and 9 strengthen this by characterizing the rate of convergence.

Similar non-Bayesian update rules have been in the context of one-shot merging of opinions [83] and beliefs in [88] and [89]. Olfati-Saber et al. [83] studied an algorithm for distributed one-shot hypothesis testing using belief propagation (BP), where nodes perform average consensus on the log-likelihoods under a single observation per node. The nodes can achieve a consensus on the product of their local likelihoods. A benefit of our approach is that nodes do not need to know each other's likelihood functions or indeed even the space from which their observations are drawn. Saligrama et al. [88] and Alanyali et al. [89], consider a similar setup of belief propagation (after observing single event) for the problem of distributed identification of the MAP estimate (which coincides with the true hypothesis for sufficiently large number of observations) for certain balanced graphs. Each node passes messages which are composed by taking a product of the recent messages then taking a weighted average over all hypotheses. Alanyali et al. [89] propose modified BP algorithms that achieve MAP consensus for arbitrary graphs. Though the structure of the message composition of the BP algorithm based message passing is similar to our proposed learning rule, we consider a dynamic setting in which observations are made infinitely often. Our rule incorporates new observation every time a node updates its belief to learn the true hypothesis. Other works study collective MAP estimation when nodes communicate discrete decisions based on Bayesian updates [90, 91] Harel et el. in [90] study a two-node model where agents exchange decisions rather than beliefs and show that unidirectional transmission increases the speed of convergence over bidirectional exchange of local decisions. Mueller-Frank [91] generalized this result to a setting in which nodes similarly exchange local strategies and local actions to make inferences.

Several recently-proposed models study distributed sequential binary hypothesis testing detecting between different means with Gaussian [92] and non-Gaussian observation models [93]. Jakovetic et al. [93] consider a distributed hypothesis test for i.i.d observations over time and

across nodes where nodes exchange weighted sum of a local estimate from previous time instant and ratio of likelihood functions of the latest local observation with the neighbors. When the network is densely connected (for instance, a doubly stochastic weight matrix), after sufficiently long time nodes gather all the observations throughout network. By appropriately choosing a local threshold for local Neyman-Pearson test, they show that the performance of centralized Neyman-Pearson test can be achieved locally. In contrast, our M-ary learning rule applies for observations that are correlated across nodes and exchanges more compact messages i.e., the beliefs (two finite precision real values for binary hypothesis test) as opposed to messages composed of the raw observations (in the case of $\mathbb{R}^d$ Gaussian observations with $d \gg 2$, d finite precision real values for binary hypothesis test). Sahu and Kar [92] consider a variant of this test for the special case of Gaussians with shifted mean and show that it minimizes the expected stopping times under each hypothesis for given detection errors.

5.2 The Model

5.2.1 Nodes' Observation Model

Consider a group of n individual nodes. Let $\Theta = \{\theta_1, \theta_2, \ldots, \theta_M\}$ denote a finite set of M parameters which we call *hypotheses*: each θ_i denotes a hypothesis. At each time instant t, every node $i \in [n]$ makes an observation $X_i^{(t)} \in \mathcal{X}_i$, where $\mathcal{X}_i$ denotes the observation space of node i. The joint observation profile at any time t across the network, $\{X_1^{(t)}, X_2^{(t)}, \ldots, X_n^{(t)}\}$, is denoted by $\mathbf{X^{(t)}} \in \mathcal{X}$, where $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \ldots \times \mathcal{X}_n$. The joint likelihood function for all $X \in \mathcal{X}$ given θ_k is the true hypothesis is denoted as $f(X; \theta_k)$. We assume that the observations are statistically governed by a fixed global "true hypothesis" $\theta^* \in \Theta$ which is unknown to the nodes. Without loss of generality we assume that $\theta^* = \theta_M$. Furthermore, we assume that no node in network knows the joint likelihood functions $\{f(\cdot; \theta_k)\}_{k=1}^M$ but every node $i \in [n]$ knows the *local likelihood functions* $\{f_i(\cdot; \theta_k)\}_{k=1}^M$, where $f_i(\cdot; \theta_k)$ denotes the i-th marginal of $f(\cdot; \theta_k)$. Each

node's observation sequence (in time) is conditionally independent and identically distributed (i.i.d) but the observations might be correlated across the nodes at any given time.

In this setting, nodes attempt to learn the "true hypothesis" θ_M using their knowledge of $\{f_i(\cdot;\theta_k)\}_{k=1}^{M}$. In isolation, if $f_i(\cdot;\theta_k) \neq f_i(\cdot;\theta_M)$ for some $k \in [M-1]$, node i can rule out hypothesis θ_k in favor of θ_M exponentially fast with an exponent which is equal to $D\left(f_i(\cdot;\theta_M) \,\|\, f_i(\cdot;\theta_k)\right)$ [94, Section 11.7]. Hence, for a given node the KL-divergence between the distribution of the observations conditioned over the hypotheses is a useful measure of the distinguishability of the hypotheses. Now, define

$$\begin{aligned}\bar{\Theta}_i &= \{k \in [M] : f_i(\cdot;\theta_k) = f_i(\cdot;\theta_M)\} \\ &= \{k \in [M] : D\left(f_i(\cdot;\theta_M) \,\|\, f_i(\cdot;\theta_k)\right) \neq 0\}.\end{aligned}$$

In other words, let $\bar{\Theta}_i$ be the set of all hypotheses that are *locally indistinguishable* to node i. In this work, we are interested in the case where $|\bar{\Theta}_i| > 1$ for some node i, but the true hypothesis θ_M is *globally identifiable* (see (5.1)).

Assumption 1. *For every pair $k \neq j$, there is at least one node $i \in [n]$ for which the KL-divergence $D\left(f_i(\cdot;\theta_k) \,\|\, f_i(\cdot;\theta_j)\right)$ is strictly positive.*

In this case, we ask whether nodes can collectively go beyond the limitations of their local observations and learn θ_M. Since

$$\{\theta_M\} = \bar{\Theta}_1 \cap \bar{\Theta}_2 \cap \ldots \cap \bar{\Theta}_n, \tag{5.1}$$

it is straightforward to see that Assumption 1 is a sufficient condition for the global identifiability of θ_M when only marginal distributions are known at the nodes. Also, note that this assumption does not require the existence of a single node that can distinguish θ_M from all other hypotheses θ_k, where $k \in [M-1]$. We only require that for every pair $k \neq j$, there is at least one node $i \in [n]$

for which $f_i(\cdot;\theta_k) \neq f_i(\cdot;\theta_j)$.

Finally, we define a probability triple $\left(\Omega, \mathcal{F}, \mathsf{P}^{\theta_M}\right)$, where

$$\Omega = \{\omega : \omega = (\mathbf{X}^{(\mathbf{0})}, \mathbf{X}^{(\mathbf{1})}, \ldots), \forall \mathbf{X}^{(\mathbf{t})} \in \mathcal{X}, \forall t\}$$

, $\mathcal{F}$ is the $\sigma-$ algebra generated by the observations and P^{θ_M} is the probability measure induced by paths in Ω, i.e., $\mathsf{P}^{\theta_M} = \prod_{t=0}^{\infty} f(\cdot;\theta_M)$. We use $\mathbb{E}^{\theta_M}[\cdot]$ to denote the expectation operator associated with measure P^{θ_M}. For simplicity we drop θ_M to denote P^{θ_M} by P and denote $\mathbb{E}^{\theta_M}[\cdot]$ by $\mathbb{E}[\cdot]$.

5.2.2 Network

We model the communication network between nodes via a directed graph with vertex set $[n]$. We define the neighborhood of node i, denoted by $\mathcal{N}(i)$, as the set of all nodes which have an edge starting from themselves to node i. This means if node $j \in \mathcal{N}(i)$, it can send the information to node i along this edge. In other words, the neighborhood of node i denotes the set of all sources of information available to it. Moreover, we assume that the nodes have knowledge of their neighbors $\mathcal{N}(i)$ only and they have no knowledge of the rest of the network [95].

Assumption 2. *The underlying graph of the network is strongly connected, i.e. for every $i, j \in [n]$ there exists a directed path starting from node i and ending at node j.*

We consider the case where the nodes are connected to every other node in the network by at least one multi-hop path, i.e. a strongly connected graph allows the information gathered to be disseminated at every node throughout the network. Such a network enables learning even when some nodes in the network may not be able to distinguish the true hypothesis on their own, i.e. the case where $|\bar{\Theta}_i| > 1$ for some nodes.

5.2.3 The Learning Rule

In this section we provide a learning rule for the nodes to learn θ_M by collaborating with each other through the local communication alone.

We begin by defining the variables required in order to define the learning rule. At every time instant t each node i maintains a private belief vector $\mathbf{q_i^{(t)}} \in \mathcal{P}(\Theta)$ and a public belief vector $\mathbf{b_i^{(t)}} \in \mathcal{P}(\Theta)$, which are probability distributions on Θ. The social interaction of the nodes is characterized by a stochastic matrix W. More specifically, weight $W_{ij} \in [0,1]$ is assigned to the edge from node j to node i such that $W_{ij} > 0$ if and only if $j \in \mathcal{N}(i)$ and $W_{ii} = 1 - \sum_{j=1}^{n} W_{ij}$. The weight W_{ij} denotes the (relative) confidence node i has on the information it receives from node j.

The steps of learning are given below. Suppose each node i starts with an initial private belief vector $\mathbf{q_i^{(0)}}$. At each time $t = 1, 2, \ldots$ the following events happen:

1. Each node i draws a conditionally i.i.d observation $X_i^{(t)} \sim f_i(\cdot; \theta_M)$.

2. Each node i performs a local Bayesian update on $\mathbf{q_i^{(t-1)}}$ to form $\mathbf{b_i^{(t)}}$ using the following rule. For each $k \in [M]$,

$$b_i^{(t)}(\theta_k) = \frac{f_i\left(X_i^{(t)}; \theta_k\right) q_i^{(t-1)}(\theta_k)}{\sum_{a \in [M]} f_i\left(X_i^{(t)}; \theta_a\right) q_i^{(t-1)}(\theta_a)}. \tag{5.2}$$

3. Each node i sends the message $\mathbf{Y_i^{(t)}} = \mathbf{b_i^{(t)}}$ to all nodes j for which $i \in \mathcal{N}(j)$. Similarly receives messages from its neighbors $\mathcal{N}(i)$.

4. Each node i updates its private belief of every θ_k, by averaging the log beliefs it received from its neighbors. For each $k \in [M]$,

$$q_i^{(t)}(\theta_k) = \frac{\exp\left(\sum_{j=1}^{n} W_{ij} \log b_j^{(t)}(\theta_k)\right)}{\sum_{a \in [M]} \exp\left(\sum_{j=1}^{n} W_{ij} \log b_j^{(t)}(\theta_a)\right)}. \tag{5.3}$$

Note that the private belief vector $\mathbf{q_i^{(t)}}$ remain locally with the nodes while their public belief vectors $\mathbf{b_i^{(t)}}$ are exchanged with the neighbors. The objective of learning rule is to ensure that the private belief vector $\mathbf{q_i^{(t)}}$ of each node $i \in [n]$ converges to $\mathbf{1}_M(\cdot)$.

Given the weight matrix W, the network can be thought of as a weighted strongly connected network. Assumption 2, implies that weight matrix W is irreducible. In this context we recall the following fact.

Fact 5 (Section 2.5 of Hoel et. al. [96])**.** *Let W be the transition matrix of a Markov chain. If W is irreducible then the stationary distribution of the Markov chain denoted by $\mathbf{v} = [v_1, v_2, \ldots, v_n]$ is the normalized left eigenvector of W associated with eigenvalue 1 and it is given as*

$$v_i = \sum_{j=1}^{n} v_j W_{ji}. \tag{5.4}$$

Furthermore, all components of $\mathbf{v}$ are strictly positive. If the Markov chain is aperiodic, then

$$\lim_{t\to\infty} W^t(i,j) = v_j, \quad i,j \in [n]. \tag{5.5}$$

If the chain is periodic with period d, then for each pair of states $i, j \in [n]$, there exists an integer $r \in [d]$, such that $W^t(i,j) = 0$ unless $t = md + r$ for some nonnegative integer m, and

$$\lim_{m\to\infty} W^{md+r}(i,j) = v_j d. \tag{5.6}$$

In the social learning literature, the eigenvector $\mathbf{v}$ also known as the eigenvector centrality; it is a measure of social influence of a node in the network. In particular we will see that v_i determines the contribution of node i in the collective network learning rate.

Definition 22 (Network Divergence)**.** For all $k \in [M-1]$, the network divergence between θ_M

and θ_k, denoted by $K(\theta_M, \theta_k)$, is defined as

$$K(\theta_M, \theta_k) := \sum_{i=1}^{n} v_i D\left(f_i(\cdot;\theta_M) \,\|\, f_i(\cdot;\theta_k)\right), \tag{5.7}$$

where $\mathbf{v} = [v_1, v_2, \ldots, v_n]$ is the normalized left eigenvector of W associated with eigenvalue 1.

Fact 5 together with Assumption 1 guarantees that $K(\theta_M, \theta_k)$ is strictly positive for every $k \in [M-1]$.

Due to the form of our learning rule, if the initial belief of any $\theta_k, k \in [M]$, for some node is zero then beliefs of that θ_k remain zero in subsequent time intervals. Hence, we require the following assumption.

Assumption 3. *For all $i \in [n]$, the initial private belief $q_i^{(0)}(\theta_k) > 0$ for every $k \in [M]$.*

5.3 Main Results

5.3.1 The Criteria for Learning

Before we present our main results, we discuss the metrics we use to evaluate the performance of a learning rule in the given distributed setup.

Definition 23 (Rate of Rejection of Wrong Hypothesis)**.** For any node $i \in [n]$ and $k \in [M-1]$, define the following

$$\rho_i^{(t)}(\theta_k) := -\frac{1}{t} \log q_i^{(t)}(\theta_k). \tag{5.8}$$

The rate of rejection of θ_k in favor of θ_M at node i is defined as

$$\rho_i(\theta_k) := \liminf_{t\to\infty} \rho_i^{(t)}(\theta_k). \tag{5.9}$$

Now, let

$$\tilde{\mathbf{q}}_{\mathbf{i}}^{(\mathbf{t})} := \left[q_i^{(t)}(\theta_1), q_i^{(t)}(\theta_2), \ldots, q_i^{(t)}(\theta_{M-1})\right]^T. \tag{5.10}$$

Then

$$\rho_i^{(t)} := -\frac{1}{t}\log \tilde{\mathbf{q}}_{\mathbf{i}}^{(\mathbf{t})} \tag{5.11}$$

and the rate of rejection at node i is defined as

$$\rho_i := \liminf_{t\to\infty} \rho_i^{(t)}. \tag{5.12}$$

If $\rho_i(\theta_k) > 0$ for all $k \in [M-1]$, under a given learning rule the belief vector of node i not only converges to the true hypothesis, it converges exponentially fast. Another way to measure the performance of a learning rule is the rate at which the belief of true hypothesis converges to one.

Definition 24 (Rate of Convergence to True Hypothesis)**.** For any $i \in [n]$ and $k \in [M-1]$, define the rate of convergence μ_i to θ_M by

$$\mu_i := \liminf_{t\to\infty} -\frac{1}{t}\log(1 - q_i^{(t)}(\theta_M)). \tag{5.13}$$

Definition 25 (Rate of Social Learning)**.** The total variational error across the network when the underlying true hypothesis is θ_k (where we allow the true hypothesis to vary, i.e. $\theta^* = \theta_k$ for any $k \in [M]$ instead of assuming that it is fixed at $\theta^* = \theta_M$) is given as

$$e^{(t)}(k) = \frac{1}{2}\sum_{i=1}^{n} ||q_i^{(t)}(\cdot) - \mathbf{1}_k(\cdot)|| = \sum_{i=1}^{n}\sum_{j\neq k} q_i^{(t)}(\theta_j). \tag{5.14}$$

This equals the total probability that all nodes in the network assign to "wrong hypotheses". Now,

define

$$e^{(t)} := \max_{k\in[M]} e^{(t)}(k). \tag{5.15}$$

The rate of social learning is defined as the rate at which total variational error, $e^{(t)}$, converges to zero and mathematically it is defined as

$$\rho_L := \liminf_{t\to\infty} -\frac{1}{t}\log e^{(t)}. \tag{5.16}$$

This measure of performance for the learning rule has been used in the social learning literature [81]. For a given network and a given observation model for nodes, ρ_L gives the least rate of learning guaranteed in the network and therefore provides a worst case guarantee. It is straightforward to see that with a characterization for $\rho_i(\theta_k)$ for all $k \in [M-1]$ we obtain a lower bound on rate of convergence to true hypothesis, μ_i, and on the rate of social learning, ρ_L, under a given learning rule.

5.3.2 Learning: Convergence to True Hypothesis

Theorem 7 (Rate of Rejecting Wrong Hypotheses, ρ_i)**.** *Let θ_M be the true hypothesis. Under the Assumptions 1–3, for every node in the network, the private belief (and hence the public belief) under the proposed learning rule converges to true hypothesis exponentially fast with probability one. Furthermore, the rate of rejecting hypothesis θ_k in favor of θ_M is given by the network divergence between θ_M and θ_k. Specifically, we have*

$$\lim_{t\to\infty} \mathbf{q_i^{(t)}} = \mathbf{1}_M \quad \mathsf{P}\text{-}a.s. \tag{5.17}$$

and

$$\rho_i = -\lim_{t\to\infty} \frac{1}{t}\log \tilde{\mathbf{q}}_\mathbf{i}^{(\mathbf{t})} = \mathbf{K} \quad \mathsf{P}\text{-}a.s. \tag{5.18}$$

where

$$\mathbf{K} = [K(\theta_M, \theta_1), K(\theta_M, \theta_2), \ldots, K(\theta_M, \theta_{M-1})]^T . \tag{5.19}$$

The proof of Theorem 7 is provided in Appendix A. Theorem 7 establishes that the beliefs of wrong hypotheses, θ_k for $k \in [M-1]$, vanish exponentially fast and it characterizes the exponent with which a node rejects θ_k in favor of θ_M. The rate of rejection is a function of the node's ability to distinguish between the hypotheses, which is given by the KL-divergences and structure of the weighted network, weighted by the eigenvector centrality of the nodes. Hence, every node influences the rate in two ways. Firstly, if the node has higher eigenvector centrality (i.e. the node is centrality located), it has larger influence over the beliefs of other nodes as a result has a greater influence over the rate of exponential decay as well. Secondly, if the node has high KL-divergence (i.e highly informative observations that can distinguish between θ_k and θ_M), then again it increases the rate. If an influential node has highly informative observations then it boosts the rate of rejecting θ_k by improving the rate. We will illustrate this through numerical examples in Section 5.4.1.

We obtain lower bound on the rate of convergence to the true hypothesis and rate of learning as corollaries to Theorem 7.

Corollary 9 (Lower Bound on Rate of Convergence to θ_M)**.** *Let θ_M be the true hypothesis. Under the Assumptions 1–3, for every $i \in [n]$, the rate of convergence to θ_M can be lower-bounded as*

$$\mu_i \geq \min_{k \in [M-1]} K(\theta_M, \theta_k) \quad \mathsf{P}\text{-}a.s. \tag{5.20}$$

Corollary 10 (Lower Bound on Rate of Learning)**.** *Let θ_M be the true hypothesis. Under the*

Assumptions 1–3, the rate of learning ρ_L across the network is lower-bounded by,

$$\rho_L \geq \min_{i,j\in[M]} K(\theta_i, \theta_j) \quad \mathsf{P}\text{-}a.s.$$

Remark 16. Jadbabaie et. al. proposed a learning rule in [79], which differs from the proposed rule at the private belief vector $\mathbf{q_i^{(t)}}$ formation step. Instead of averaging the log beliefs, nodes average the beliefs received as messages from their neighbors. In [81], Jadbabaie et. al. provide an upper bound on the rate of learning ρ_L obtained using their algorithm. They show

$$\rho_L \leq \alpha \min_{i,j\in[M]} K(\theta_i, \theta_j) \quad \mathsf{P}\text{-a.s.} \tag{5.21}$$

where α is a constant strictly less than one. Corollary 10 shows that lower bound on ρ_L using the proposed algorithm is greater than the upper bound provided in (5.21).

5.3.3 Concentration under Bounded Log-likelihood ratios

Under mild assumptions, Theorem 7 shows that the belief about a wrong hypothesis θ_k for $k \in [M-1]$ converges to zero exponentially fast at rate equal to the network divergence, $K(\theta_M, \theta_k)$, between θ_M and θ_k with probability one. We strength this result for periodic networks with period d under the following assumption.

Assumption 4. *There exists a positive constant L such that*

$$\max_{i\in[n]} \max_{j,k\in[M]} \sup_{X\in\mathcal{X}_i} \left| \log \frac{f_i\left(X;\theta_j\right)}{f_i\left(X;\theta_k\right)} \right| \leq L. \tag{5.22}$$

Theorem 8 (Concentration of Rate of Rejecting Wrong Hypotheses, $\rho_i^{(t)}(\theta_k)$)**.** *Let θ_M be the true hypothesis. Under Assumptions 1–4, for periodic networks with period d, for every node $i \in [n]$,*

$k \in [M-1]$, *and for all* $\varepsilon > 0$ *we have*

$$\lim_{t\to\infty} \frac{1}{t} \log \mathsf{P}\left(\rho_i^{(t)}(\theta_k) \leq K(\theta_M, \theta_k) - \varepsilon\right) \leq -\frac{\varepsilon^2}{2L^2 d}. \tag{5.23}$$

For $0 < \varepsilon \leq L - K(\theta_M, \theta_k)$*, we have*

$$\lim_{t\to\infty} \frac{1}{t} \log \mathsf{P}\left(\rho_i^{(t)}(\theta_k) \geq K(\theta_M, \theta_k) + \varepsilon\right) \leq -\frac{1}{2L^2 d} \min\left\{\varepsilon^2, \min_{j\in[M-1]} K(\theta_M, \theta_j)^2\right\}. \tag{5.24}$$

For $\varepsilon \geq L - K(\theta_M, \theta_k)$ *we have*

$$\lim_{t\to\infty} \frac{1}{t} \log \mathsf{P}\left(\rho_i^{(t)}(\theta_k) \geq K(\theta_M, \theta_k) + \varepsilon\right) \leq -\min_{k\in[M-1]} \left\{\frac{K(\theta_M, \theta_k)^2}{2L^2 d}\right\}. \tag{5.25}$$

Corollary 11 (Rate of convergence to True Hypothesis)**.** *Let* θ_M *be the true hypothesis. Under Assumptions 1–4, for every* $i \in [n]$*, we have*

$$\mu_i = \min_{k\in[M-1]} K(\theta_M, \theta_k) \quad \mathsf{P}\text{-}a.s.$$

Proofs of Theorem 8 and Corollary 11 are provided in Appendix B. From Theorem 7 we know that $\rho_i^{(t)}(\theta_k)$ converges to $K(\theta_M, \theta_k)$ almost surely. Theorem 8 strengthens Theorem 7 by showing that the probability of sample paths where $\rho_i^{(t)}(\theta_k)$ deviates by some fixed ε from $K(\theta_M, \theta_k)$ vanishes exponentially fast. This implies that $\rho_i^{(t)}(\theta_k)$ converges to $K(\theta_M, \theta_k)$ exponentially fast in probability. Theorem 8 also characterizes a lower bound on the exponent when the probability of such events vanishes and shows that periodicity of the network reduces the exponent.

5.3.4 Large Deviation Analysis

We require a technical assumption that relaxes the assumption of bounded ratios of the likelihood functions in prior work [85, 97–99].

Assumption 5. *For every pair $\theta_i \neq \theta_j$ and every node $k \in [n]$, the random variable $\left|\log \frac{f_k(X_k;\theta_i)}{f_k(X_k;\theta_j)}\right|$ has finite log moment generating function under distribution $f_k\left(\cdot;\theta_j\right)$.*

Next, we give examples of families of distributions which satisfy Assumption 5 but violate Assumption 4.

Remark 17. Distributions $f(X;\theta_i)$ and $f(X;\theta_j)$ for $i \neq j$ which the following properties for some positive constants C and β, satisfy Assumption 5

$$\mathsf{P}_i\left(\frac{f(X;\theta_j)}{f(X;\theta_i)} \geq x\right) \leq \frac{C}{x^\beta}, \quad \mathsf{P}_i\left(\frac{f(X;\theta_i)}{f(X;\theta_j)} \geq x\right) \leq \frac{C}{x^\beta}. \tag{5.26}$$

Note that (5.26) is a sufficient condition but not a necessary condition. Examples 6–7 below do not satisfy (5.26) yet satisfy Assumption 5.

Example 6 (Gaussian Mixtures)**.** Let $f(X;\theta_1) = \mathcal{N}(\mu_1,\sigma)$ and $f(X;\theta_2) = \mathcal{N}(\mu_2,\sigma)$. Then

$$g_1(x) := \left|\log \frac{f(x;\theta_1)}{f(x;\theta_2)}\right| \leq c_1|x| + c_2, \tag{5.27}$$

where $c_1 = \left|\frac{\mu_1-\mu_2}{\sigma^2}\right|$ and $c_2 = \left|\frac{\mu_1^2-\mu_2^2}{2\sigma^2}\right|$. Hence, for $i \in \{1,2\}$ and for $\lambda \geq 0$ we have

$$\mathbb{E}_i\left[e^{\lambda g_1(X)}\right] \leq e^{c_2\lambda}\mathbb{E}_i\left[e^{c_1\lambda|X|}\right] < \infty. \tag{5.28}$$

More generally for $i \in \{1,2\}$, and $p \in [0,1]$, let

$$f(x;\theta_i) = \frac{p}{\sigma\sqrt{2\pi}}\exp\left(\frac{-(x-\alpha_i)^2}{2\sigma^2}\right) + \frac{1-p}{\sigma\sqrt{2\pi}}\exp\left(\frac{-(x-\beta_i)^2}{2\sigma^2}\right). \tag{5.29}$$

Then the log moment generating function of $\left|\log \frac{f(X;\theta_1)}{f(X;\theta_2)}\right|$ is finite for all $\lambda \geq 0$.

Example 7 (Gamma distribution)**.** Let

$$f(X;\theta_1) = \frac{\beta^{\alpha_1}}{\Gamma(\alpha_1)} x^{\alpha_1 - 1} e^{-\beta x}$$

and

$$f(X;\theta_2) = \frac{\beta^{\alpha_2}}{\Gamma(\alpha_2)} x^{\alpha_2 - 1} e^{-\beta x},$$

then

$$g_2(x) := \left|\log \frac{f(x;\theta_1)}{f(x;\theta_2)}\right| \leq c_1 |\log x| + c_2, \tag{5.30}$$

where $c_1 = |\alpha_1 - \alpha_2|$ and $c_2 = \left|(\alpha_1 - \alpha_2)\log\beta + \log\frac{\Gamma(\alpha_2)}{\Gamma(\alpha_1)}\right|$. Hence, for $i \in \{1,2\}$ and for $\lambda \geq 0$ we have

$$\mathbb{E}_i\left[e^{\lambda g_2(X)}\right] \leq e^{c_2\lambda}\mathbb{E}_i\left[e^{c_1\lambda|\log X|}\right] < \infty. \tag{5.31}$$

The above examples show that Assumption 5 is satisfied for distributions which have unbounded support. In order to analyze the concentration of $\rho_i^{(t)}$ under Assumption 5 we replace Assumption 2 with the following assumption.

Assumption 2′. The underlying graph of the network is strongly connected and aperiodic.

Next we provide few more definitions. Let

$$\mathbf{Y}^{(t)}(\theta_k) := \langle \mathbf{v}, \mathbf{L}^{(t)}(\theta_k)\rangle, \tag{5.32}$$

where $\mathbf{L}^{(t)}(\theta_k)$ is the vector of log likelihood ratios given by

$$\mathbf{L}^{(t)}(\theta_k) = \left[\log \frac{f_1\left(X_1^{(t)};\theta_k\right)}{f_1\left(X_1^{(t)};\theta_M\right)}, \ldots, \log \frac{f_n\left(X_n^{(t)};\theta_k\right)}{f_n\left(X_n^{(t)};\theta_M\right)} \right]^T. \tag{5.33}$$

Definition 26 (Moment Generating Function)**.** For every $\lambda_k \in \mathbb{R}$, let $\Lambda_k(\lambda_k)$ denote the log moment generating function of $\mathbf{Y}^{(t)}(\theta_k)$ by

$$\Lambda_k(\lambda_k) := \log \mathbb{E}[e^{\lambda_k \mathbf{Y}^{(t)}(\theta_k)}] = \log \mathbb{E}[e^{\lambda_k \langle \mathbf{v}, \mathbf{L}(\theta_k)\rangle}] \tag{5.34}$$

For every $\lambda \in \mathbb{R}^{M-1}$, let $\Lambda(\lambda)$ denote the log moment generating function of $\mathbf{Y}$ by

$$\Lambda(\lambda) := \log \mathbb{E}[e^{\langle \lambda, \mathbf{Y}\rangle}]. \tag{5.35}$$

Note that each entry of vector $\mathbf{Y}^{(t)}$ is a function of joint observation vector $\mathbf{X}^{(t)}$ whose distribution is governed by $f(\cdot;\theta_M)$.

Definition 27 (Large Deviation Rate Function)**.** For all $x \in \mathbb{R}$, let $I_k(x)$ denote the Fenchel-Legendre transform of $\Lambda_k(\cdot)$

$$I_k(x) := \sup_{\lambda_k \in \mathbb{R}} \{\lambda x - \Lambda_k(\lambda_k)\}. \tag{5.36}$$

For all $\mathbf{x} \in \mathbb{R}^{M-1}$, let $I(\mathbf{x})$ denote the Fenchel-Legendre transform of $\Lambda(\cdot)$

$$I(\mathbf{x}) := \sup_{\lambda \in \mathbb{R}^{M-1}} \{\langle \lambda, \mathbf{x}\rangle - \Lambda(\lambda)\}. \tag{5.37}$$

Theorem 9 (Large Deviations of $\rho_i^{(t)}$)**.** *Let θ_M be the true hypothesis. Under Assumptions 1, 2′, 3, 5, the rate of rejection $\rho_i^{(t)}$ satisfies an Large Deviation Principle with rate function $J(\cdot)$, i.e.,*

for any set $F \subset \mathbb{R}^{M-1}$ we have

$$\liminf_{t\to\infty} \frac{1}{t} \log \mathsf{P}\left(\rho_i^{(t)} \in F\right) \geq - \inf_{\mathbf{y}\in F^o} J(\mathbf{y}), \tag{5.38}$$

and

$$\limsup_{t\to\infty} \frac{1}{t} \log \mathsf{P}\left(\rho_i^{(t)} \in F\right) \leq - \inf_{\mathbf{y}\in \bar{F}} J(\mathbf{y}), \tag{5.39}$$

where large deviation rate function $J(\cdot)$ is defined as

$$J(\mathbf{y}) := \inf_{\mathbf{x}\in\mathbb{R}^{M-1}: g(\mathbf{x})=\mathbf{y}} I(\mathbf{x}),\ \forall \mathbf{y} \in \mathbb{R}^{M-1}, \tag{5.40}$$

where $g : \mathbb{R}^{M-1} \to \mathbb{R}^{M-1}$ is a continuous mapping given by

$$g(\mathbf{x}) := [g_1(\mathbf{x}), g_2(\mathbf{x}), \ldots, g_{M-1}(\mathbf{x})]^T, \tag{5.41}$$

and

$$g_k(\mathbf{x}) := x_k - \max\{0, x_1, x_2, \ldots, x_{M-1}\}. \tag{5.42}$$

The proof of Theorem 9 is provided in Appendix C. Theorem 9 characterizes the asymptotic rate of concentration of $\rho_i^{(t)}$ in any set $F \subset \mathbb{R}^{M-1}$. In other words, it characterizes the rate at which the probability of deviations in each $\rho_i^{(t)}(\theta_k)$ from the rate of rejection $K(\theta_M, \theta_k)$ for every $\theta_k \neq \theta_M$ vanish simultaneously. It characterizes the asymptotic rate as a function of the observation model of each node (not just the bound L on the ratios of log-likelihood function) and as a function of eigenvector centrality $\mathbf{v}$. The following corollary specializes this result to obtain the individual rate of rejecting a wrong hypothesis at every node. It can be obtained by repeating the proof of Theorem 9 for each hypothesis alone.

Corollary 12. *Let θ_M be the true hypothesis. Under Assumptions 1, 2′, 3, 5, for $0 < \varepsilon \leq K(\theta_M, \theta_k)$, $k \in [M-1]$, we have*

$$\lim_{t\to\infty} \frac{1}{t} \log \mathsf{P}\left(\rho_i^{(t)}(\theta_k) \leq K(\theta_M, \theta_k) - \varepsilon\right) = -I_k\left(K(\theta_M, \theta_k) - \varepsilon\right). \tag{5.43}$$

For $\varepsilon > 0$, we have

$$\lim_{t\to\infty} \frac{1}{t} \log \mathsf{P}\left(\rho_i^{(t)}(\theta_k) \geq K(\theta_M, \theta_k) + \varepsilon\right) = -I_k\left(K(\theta_M, \theta_k) + \varepsilon\right). \tag{5.44}$$

Using Theorem 9 and Hoeffding's Lemma, we obtain the following corollary.

Corollary 13. *Suppose Assumption 4 is satisfied for some finite $L \in \mathbb{R}$. For ε as specified in Theorem 8, we recover the exponents of Theorem 8 under aperiodic networks, given by*

$$\lim_{t\to\infty} \frac{1}{t} \log \mathsf{P}\left(\rho_i^{(t)}(\theta_k) \geq K(\theta_M, \theta_k) + \varepsilon\right) \leq -\frac{\varepsilon^2}{2L^2}, \tag{5.45}$$

and

$$\lim_{t\to\infty} \frac{1}{t} \log \mathsf{P}\left(\rho_i^{(t)}(\theta_k) \leq K(\theta_M, \theta_k) - \varepsilon\right) \leq -\frac{\varepsilon^2}{2L^2}. \tag{5.46}$$

Remark 18. Under Assumption 4, Corollary 13 shows that lower bound on the asymptotic rate of concentration of $\rho_i^{(t)}$ as characterized by Theorem 8 is loose in comparision to that obtained from Theorem 9. Nedic et al. [86] and Shahrampour et al. [85] provide non-asymptotic lower bounds on the rate of concentration of $\rho_i^{(t)}$ whose asymptotic form coincides with the lower bound on rate characterized by Theorem 8 for aperiodic networks. This implies that under Assumption 4 Theorem 9 provides a tighter asymptotic rate than their results in [85, 86]. Hence, Theorem 9 strengthens Theorem 8 by extending the large deviation to larger class of distributions and providing a tighter bound that captures the complete effect of nodes' influence in the network

and the local observation statistics.

5.4 Examples

In this section through numerical examples we illustrate how nodes learn using the proposed learning rule and examine the factors which affect the rate of rejection of wrong hypotheses and its rate of concentration.

5.4.1 Factors influencing Convergence

Example 8. Consider a group of two nodes as shown in Figure 5.1, where the set of hypotheses is $\Theta = \{\theta_1, \theta_2, \theta_3, \theta_4\}$ and true hypothesis $\theta^* = \theta_4$. Observations at each node at time t, $X_i^{(t)}$, take values in $\mathbb{R}^{100}$ and have a Gaussian distribution. For node 1, $f_1(\cdot;\theta_1) = f_1(\cdot;\theta_3) = \mathcal{N}(\mu_{11}, \Sigma)$ and $f_1(\cdot;\theta_2) = f_1(\cdot;\theta_4) = \mathcal{N}(\mu_{12}, \Sigma)$, and for node 2, $f_2(\cdot;\theta_1) = f_2(\cdot;\theta_2) = \mathcal{N}(\mu_{21}, \Sigma)$ and $f_2(\cdot;\theta_3) = f_2(\cdot;\theta_4) = \mathcal{N}(\mu_{22}, \Sigma)$, where $\mu_{11}, \mu_{12}, \mu_{21}, \mu_{22} \in \mathbb{R}^{100}$ and Σ is a positive semi-definite matrix of size 100-by-100. Here, node 1 can identify the column containing θ_4, and node 2 can identify the row. In other words, $\bar{\Theta}_1 = \{\theta_2, \theta_4\}$ and $\bar{\Theta}_2 = \{\theta_3, \theta_4\}$. Also, $\theta_4 = \bar{\Theta}_1 \cap \bar{\Theta}_2$, hence θ_4 is globally identifiable.

Strong Connectivity

Nodes are connected to each other in a network and the weight matrix is given by

$$W = \begin{pmatrix} 0.9 & 0.1 \\ 0.4 & 0.6 \end{pmatrix}. \tag{5.47}$$

Figure 5.2 shows the evolution of beliefs with time for node 2 on a single sample path. We see that using the proposed learning rule, belief of θ_4 goes to one while the beliefs of wrong hypotheses go to zero. This example shows that each node through collaboration is able to learn

θ_4. Figure 5.3 shows the rate of rejection of wrong hypotheses. We see that the rate of rejection θ_k for $k \in \{1,2,3\}$ closely follows the asymptotic rate $K(\theta_4, \theta_k)$.

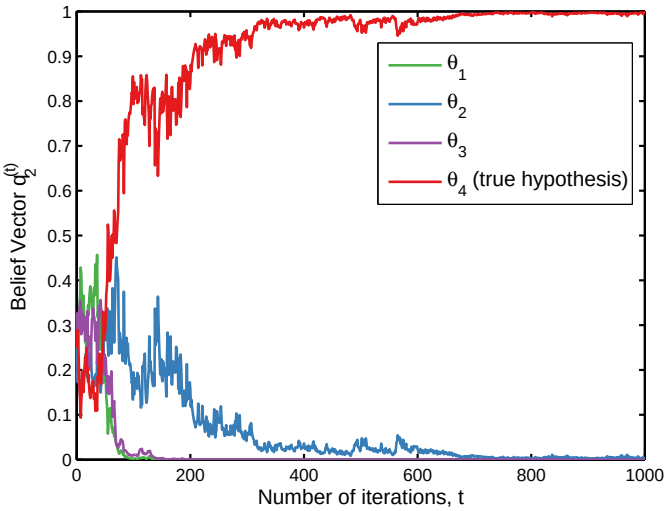


Figure 5.2: For the set of nodes described in Figure 5.1, this figure shows the evolution of beliefs for one instance using the proposed learning rule. Belief of the true hypothesis θ_4 of node 2 converges to 1 and beliefs of all other hypotheses go to zero.

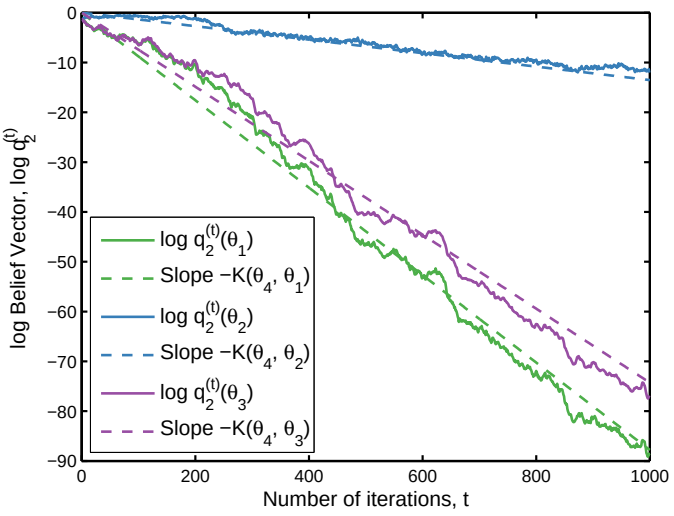


Figure 5.3: Figure shows the exponential decay of beliefs of θ_1, θ_2 and θ_3 of node 2 using the learning rule.

Suppose the nodes are connected to each other in a network whose weight matrix is given

by

$$W = \begin{pmatrix} 1 & 0 \\ 0.5 & 0.5 \end{pmatrix}. \tag{5.48}$$

Since there is no path from node 2 to node 1, the network is not strongly connected. Node 2 as seen in Figure 5.4 does not converge to θ_4. Even though node 1 cannot distinguish the elements of $\bar{\Theta}_1$ from θ_4, it rejects the hypotheses in $\{\theta_1, \theta_3\}$ in favor of θ_4. This forces node 2 also to reject the set $\{\theta_1, \theta_3\}$. For node 1, θ_2 and θ_4 are observationally equivalent, hence their respective beliefs equal half. But node 2 oscillates between θ_2 and θ_4 and is unable to learn θ_4. Hence, when the network is not strongly connected both nodes fail to learn.

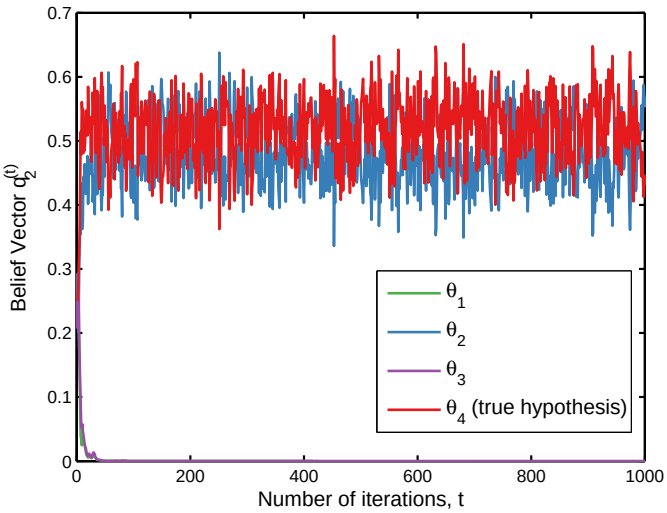


Figure 5.4: Figure shows the beliefs of node 2 shown in Figure 5.1. When the network is not strongly connected node 2 cannot learn θ_4.

In this setup we apply the learning rule considered in [79], where in the consensus step public beliefs are updated by averaging the beliefs received from the neighbors instead of averaging the logarithm of the beliefs. As seen in Figure 5.5, rate of rejecting learning using the proposed learning rule is greater than the upper bound on learning rule in [79]. Note that the precision of the belief vectors in the simulations is 8 bytes (64 bits) per hypothesis. This implies

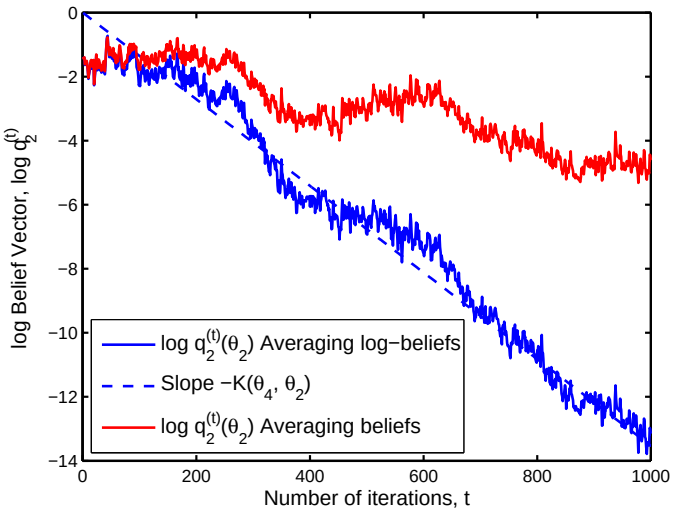


Figure 5.5: Figure shows that the rate of rejection of θ_2 using the proposed learning rule (averaging the log beliefs) is greater than the rate of rejection of θ_2 obtained using the learning rule in [79] (averaging the beliefs).

the nodes each send 32 bytes per unit time, which is less than the case when nodes exchange local Gaussian observations which may require data rate as high as 800 bytes per observation.

Periodicity

Now suppose the nodes are connected to each other in periodic network with period 2 and the weight matrix given by

$$W = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}. \tag{5.49}$$

From Figure 5.6, we see that the belief on wrong hypotheses converges to zero but beliefs oscillate significantly about the expected value of rate of rejection as compared to the case of an aperiodic network considered in (5.47).

Even though nodes do not have a positive self-weight (W_{ii}), the new information (through observations) entering at every node reaches its neighbors and gets dispersed in throughout the

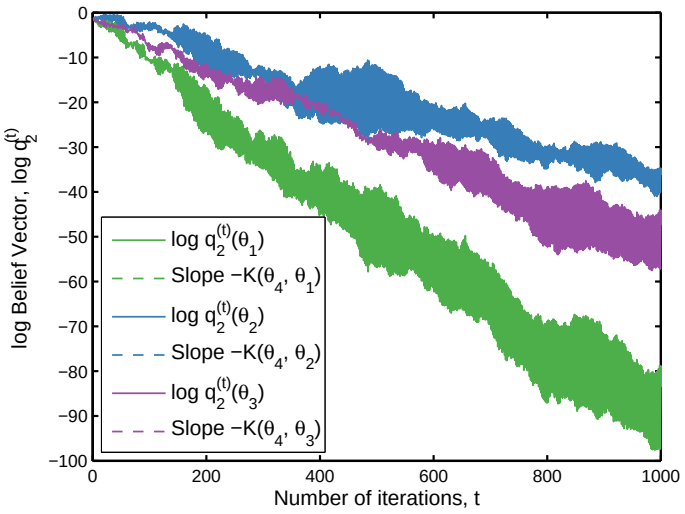


Figure 5.6: Figure shows the exponential decay of beliefs of θ_1, θ_2, and θ_3 of node 2 connected to node 1 in a periodic network with period 2.

network; eventually reaches every node. Hence, nodes learn even when the network is periodic as long as it remains strongly connected.

Eigenvector Centrality and Extent of distinguishability

From Theorem 7, we know that a larger weighted sum of the KL divergences, *i.e.* a larger network divergence, $K(\theta_M, \theta_k)$, yields a better rate of rejecting hypothesis θ_k. We look at a numerical example to show this.

Example 9. Let $\Theta = \{\theta_1, \theta_2, \theta_3, \theta_4, \theta_5\}$ and $\theta^* = \theta_4$. Consider a set of 25 nodes which are arranged in 5×5 array to form a grid. We obtain a grid network by connecting every node to its adjacent nodes. We define the weight matrix as,

$$W_{ij} = \begin{cases} \frac{1}{|\mathcal{N}(i)|}, & \text{if } j \in \mathcal{N}(i) \\ 0, & \text{otherwise} \end{cases} \tag{5.50}$$

Consider an extreme scenario where only one node can distinguish true hypothesis θ_1 from the rest and to the remaining nodes in the network all hypotheses are observationally equivalent

i.e. $\bar{\Theta}_i = \Theta$ for 24 nodes and $\bar{\Theta}_i = \{\theta_1\}$ for only one node. We call that one node which can distinguish the true hypothesis from other hypotheses as the "informed node" and the rest of the nodes called the "non-informed nodes".

For the weight matrix in (5.50), the eigenvector centrality of node i is proportional to $\mathcal{N}(i)$, which means in this case, more number of neighbors implies higher social influence. This implies that the corner nodes (namely node 1, node 5, node 20 and node 25 at the four corners of the grid) have least eigenvector centrality among all nodes. Hence, they are least influential. The nodes on four edges have a greater influence than the corner nodes. Most influential nodes are the ones with four connections, such as node 13 which is located in third row and third column of the grid. It is also the central location of the grid.

Figure 5.7 shows the variation in the rate of rejection of θ_2 of node 5 as the location of informed node changes. We see that if the informed node is at the center of the grid then the rate of rejection is fastest and the rate is slowest when the informed node is placed at a corner. In other words, rate of convergence is highest when the most influential node in the network has high distinguishability.

5.4.2 Factors influencing Concentration

Now to examine the results from Theorem 8 and Theorem 9, we go back to Example 8, where two nodes are in a strongly connected aperiodic network given by (5.47). Observation model for each node is defined as follows. For node 1, $f_1(\cdot;\theta_1) = f_1(\cdot;\theta_3) \sim \mathsf{Ber}(\frac{4}{5})$ and $f_1(\cdot;\theta_2) = f_1(\cdot;\theta_4) \sim \mathsf{Ber}(\frac{1}{4})$, and for node 2, $f_2(\cdot;\theta_1) = f_2(\cdot;\theta_2) \sim \mathsf{Ber}(\frac{1}{3})$ and $f_2(\cdot;\theta_3) = f_2(\cdot;\theta_4) \sim \mathsf{Ber}(\frac{1}{4})$. Figure 5.8 shows the exponential decay of θ_1 for 25 instances. We see that the number of sample paths that deviate more than $\varepsilon = 0.1$ from $K(\theta_4,\theta_1)$ decrease with number of iterations. Theorem 8 characterizes the asymptotic rate at which the probability of such sample paths vanishes when the log-likelihoods are bounded. This asymptotic rate is given as a function of L and period of the network. From Corollary 13 the rate given by Theorem 8 is loose for

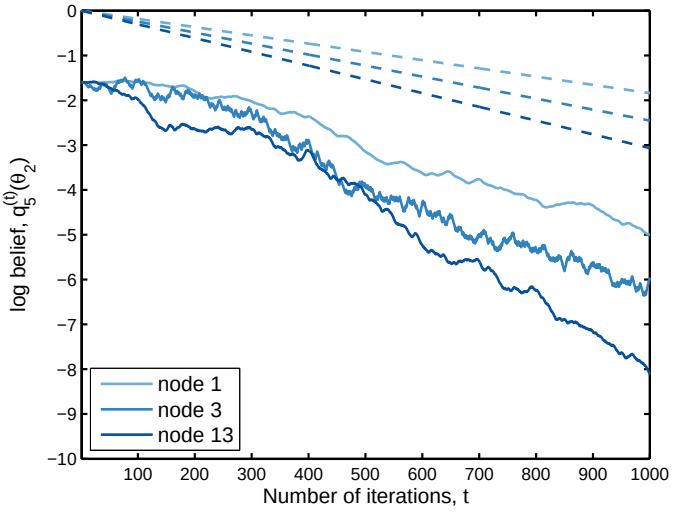


Figure 5.7: Figure illustrates the manner in which rate of rejection of θ_2 at node 5 is influenced by varying the location of an informed node. As seen here when the informed node is more central *i.e.* at node 13, rate of rejection is fastest and when the informed node is at the corner node 1, rate of rejection is slowest.

aperiodic networks. A tighter bound which utilizes the complete observation model is given by Theorem 9. Figure 5.9 shows the gap between the rates.

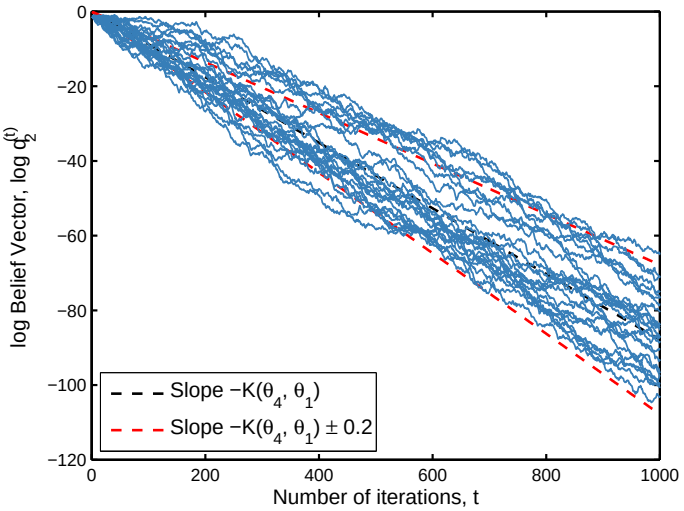


Figure 5.8: Figure shows the decay of belief of θ_1 (wrong hypothesis) of node 2 for 25 instances. We see that the number of sample paths on which the rate of rejecting θ_1 deviates more than $\eta = 0.1$ reduces as the number of iterations increase.

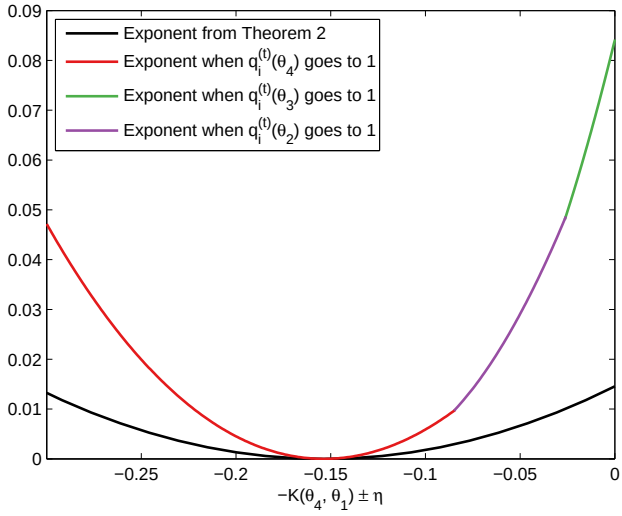


Figure 5.9: Figure shows the asymptotic exponent with which the probability of events where rate of rejecting θ_1 deviates by η from $K(\theta_4, \theta_1)$; θ_4 is the true hypothesis. The black curve shows the asymptotic exponent as characterized by Theorem 8. The colored curve shows the exact asymptotic exponent as characterized by Theorem 9, where the exponent depends on the hypothesis to which the learning rule is converging. This shows that small deviations from $K(\theta_4, \theta_1)$ occur when the learning rule is converging to θ_4 and larger deviations occur when the learning rule is converging to a wrong hypothesis.

Figure 5.9 in the context of Example 8 shows the rate at which the probability of sample paths deviating from rate of rejection can be thought of as operating in three different regimes. Here, each regime denotes the hypothesis to which the learning rule is converging. In order to see this consider the rate function of θ_1, i.e. $J_1(\cdot)$ from Corollary 12;

$$J_1(\mathbf{y}) = \inf_{x \in \mathbb{R}^3 : g(\mathbf{x}) = y} I(\mathbf{x}), \forall y \in \mathbb{R}.$$

The behavior of the rate function $J_1(\cdot)$ depends on the function $g_1(\mathbf{x}) = x_1 - \max\{0, x_1, x_2, x_3\}$. Whenever $g_1(\mathbf{x}) = x_1$, the rate function is $I_1(\cdot)$. This shows that whenever there is a deviation of $x - k(\theta_4, \theta_1)$ from the rate of rejection of θ_1, the sample paths that vanish with slowest exponents are those for which $\frac{1}{t} \log \frac{q_i^{(t)}(\theta_1)}{q_i^{(t)}(\theta_4)} < 0$ as $t \to \infty$. In other words, small deviations occur when the learning rule is converging to true hypothesis θ_4 and they depend on $I_1(\cdot)$ (and hence θ_1) alone. Whereas large deviations occur when the learning rule is mistakenly converging to a wrong

hypothesis and hence, the rate function depends on θ_1 and the wrong hypothesis to which the learning rule is converging. Hence, we have three different regimes corresponding to the three wrong hypotheses.

5.4.3 Learning with Communication Constraints

Now, we consider a variant of our learning rule where the communication between the nodes is quantized to belong to a predefined finite set. Each node i starts with an initial private belief vector $\mathbf{q_i^{(0)}}$ and at each time $t = 1, 2, \ldots$ the following events happen:

1. Each node i draws a conditionally i.i.d observation $X_i^{(t)} \sim f_i(\cdot; \theta_M)$.

2. Each node i performs a local Bayesian update on $\mathbf{q_i^{(t-1)}}$ to form $\mathbf{b_i^{(t)}}$ using the following rule. For each $k \in [M]$,

$$b_i^{(t)}(\theta_k) = \frac{f_i\left(X_i^{(t)}; \theta_k\right) q_i^{(t-1)}(\theta_k)}{\sum_{a \in [M]} f_i\left(X_i^{(t)}; \theta_a\right) q_i^{(t-1)}(\theta_a)}. \tag{5.51}$$

3. Each node i sends the message $Y_i^{(t)}(\theta_k) = \left[D b_i^{(t)}(\theta_k)\right]$, for all $k \in [M]$, to all nodes j for which $i \in \mathcal{N}(j)$, where $D \in \mathbb{Z}^+$ and

$$[x] = \begin{cases} \lfloor x \rfloor + 1, & \text{if } x > \lfloor x \rfloor + 0.5, \\ \lfloor x \rfloor, & \text{if } x \leq \lfloor x \rfloor + 0.5, \end{cases} \tag{5.52}$$

   where $\lfloor x \rfloor$ denotes the largest integer less than x.

4. Each node i normalizes the beliefs received from the neighbors $\mathcal{N}(i)$ as

$$\tilde{Y}_i^{(t)}(\theta_k) = \frac{Y_i^{(t)}(\theta_k)}{\sum_{a \in [M]} Y_i^{(t)}(\theta_a)} \tag{5.53}$$

and updates its private belief of θ_k, for each $k \in [M]$,

$$q_i^{(t)}(\theta_k) = \frac{\exp\left(\sum_{j=1}^n W_{ij} \log \tilde{Y}_i^{(t)}(\theta_k)\right)}{\sum_{a\in[M]} \exp\left(\sum_{j=1}^n W_{ij} \tilde{Y}_i^{(t)}(\theta_a)\right)}. \tag{5.54}$$

In the above learning rule, the belief on each hypothesis belongs to a set of size $D+1$. Hence transmitting the entire belief vector, i.e., transmitting the entire message requires $M\log(D+1)$ bits.

Note that all of our simulations so far, we have used 64-bit precision to represent the belief on each hypothesis, meaning our simulations can be interpreted as limiting the communication links to support 64 bits, or equivalently 8 bytes, per hypothesis per unit of time. Our previous numerical results show a close match with the analysis using this level of quantization. Next we show the impact of a coarser quantization.

Example 10. Consider a network of radars or ultrasound sensors whose aim is to find the location of a target. Each sensor can sense the target's location along one dimension only, whereas the target location is a point in three-dimensional space. Consider the configuration in Figure 5.10: there are two nodes along each of the three coordinate axes at locations $[\pm 2, 0, 0]$, $[0, \pm 2, 0]$, and $[0, 0, \pm 2]$. The communication links are given by the directed arrows. Nodes located on the x-axis can sense whether x-coordinate of the target lies in the interval $(-2, -1]$ or in the interval $(-1, 0)$ or in the interval $[0, 1)$ or in the interval $[1, 2)$. If a target is located in the interval $(-\infty, -2] \cup [2, \infty)$ on the x-axis then no node can detect it. Similarly nodes on y-axis and z-axis can each distinguish between 4 distinct non-intersecting intervals on the y-axis and the z-axis respectively. Therefore, the total number of hypotheses is $M = 4^3 = 64$.

The sensors receive signals which are three dimensional Gaussian vectors whose mean is altered in the presence of a target. In the absence of a target, the ambient signals have a Gaussian distribution with mean $[0, 0, 0]$. For the sensor node along x-axis located at $[2, 0, 0]$, if the target has x-coordinate $\theta_x \in (-2, 2)$, the mean of the sensor's observation is $[\lfloor 3 + \theta_x \rfloor, 0, 0]$. If a target is

located in $(-\infty,-2]\cup[2,\infty)$ on the x-axis, then the mean of the Gaussian observations is $[0,0,0]$. Local marginals of the nodes along y-axis and z-axis are described similarly, i.e., as the target moves away from the node by one unit the signal mean strength goes by one unit. For targets located at a distance four units and beyond the sensor cannot detect the target. In this example, suppose θ_1 is the true hypothesis.

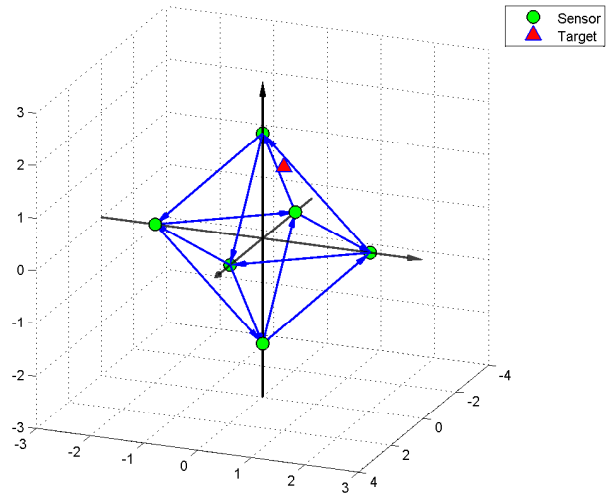


Figure 5.10: Figure shows a sensor network where each node is a low cost radar that can sense along the axis it is placed and not the other. The directed edges indicate the directed communication between the nodes. Through cooperative effort the nodes aim to learn location of the target in 3 dimensions.

Consider $D = 2^{12}-1$ which implies that belief on each hypothesis is of size 12 bits or equivalently 1.5 bytes. Figure 5.11 shows evolution of log beliefs of node 3 for hypotheses for θ_2, θ_5 and θ_6 for 500 instances when the link rate is limited to 1.5 bytes per hypothesis per unit time. We see that the learning rule converges to the true hypotheses on all 500 instances. Similarly, Figure 5.12 shows the evolution of beliefs of node 3 for hypotheses θ_2, θ_5 and θ_6 when the link rate is limited to 1 byte per hypothesis per unit time, i.e., when $D = 2^8-1$. We see that the learning rule converges to a wrong hypothesis θ_2. However, on the same sample path in Figure 5.13 we see that if the link rate is 1.5 bytes per hypothesis per unit time, the learning

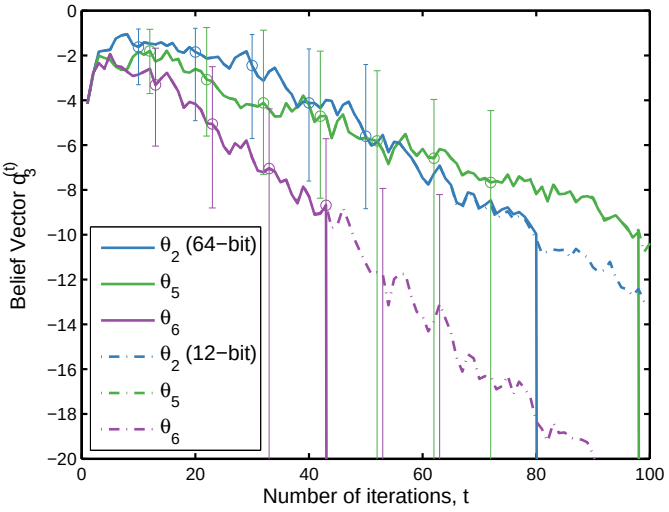


Figure 5.11: The solid lines in figure show the evolution of the log beliefs of node 3 with time for hypotheses θ_2, θ_5 and θ_6 when links support a maximum of 12 bits per hypothesis per unit time. This is compared with the evolution of the log beliefs with no rate restriction case (dotted lines) which translates a maximum of 64 bits per hypothesis per unit time. Figure also shows the confidence intervals (one standard deviation above and below) around log beliefs over 500 instances of learning rule with 12 bits per hypothesis. We see the learning rule with link rate 12 bits per hypothesis converges in all the instances.

rule converges to true hypothesis. This happens because on every sample path our learning rule has an initial transient phase where beliefs may have large fluctuations during which the belief on true hypothesis may get close to zero. For low link rates (small D), even when the belief on true hypothesis is strictly positive but less than $\frac{1}{2D}$, it gets quantized to zero. Recall that for our learning rule, when a belief goes to zero, propagates the zero belief to all subsequent time instants. This shows that as we increase link rate (increase value of D), the quantized learning rule is more robust to the initial fluctuations. Moreover, we observe that for both Examples 8 and 10, when link rates are greater than or equal to 1.5 bytes per hypothesis per unit time the learning rule converges for all instances and its performance coincides with the prediction of our the analysis under the assumption of perfect links.

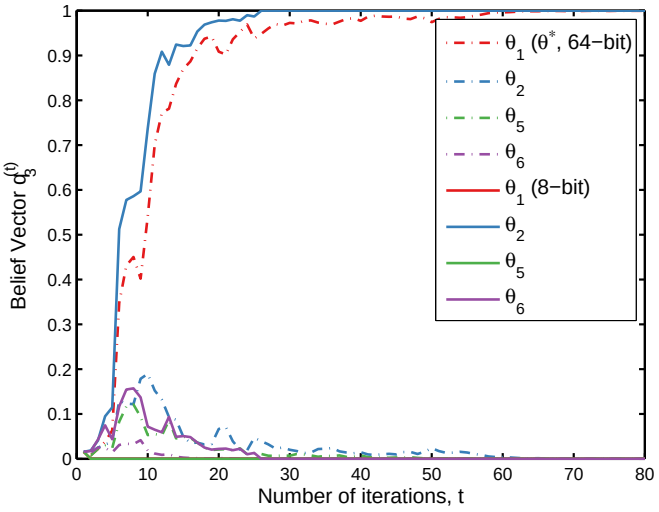


Figure 5.12: The solid lines in the figure show the evolution of the log beliefs of node 3 with time for hypotheses θ_2, θ_5 and θ_6 when links support a maximum of 8 bits per hypothesis per unit time. This is compared with the evolution of the log beliefs with no rate restriction case (dotted lines) which translates a maximum of 64 bits per hypothesis per unit time. For this sample path, we see that learning rule converges to a wrong hypothesis θ_5 when the communication is restricted to 8 bits per hypothesis.

5.5 Discussion

In this chapter we study a learning rule through which a network of nodes make observations and communicate in order to collectively learn an unknown fixed global hypothesis that statistically governs the distribution of their observations. Our learning rule performs local Bayesian updating followed by averaging log-beliefs. We showed that our rule guarantees exponentially fast convergence to the true hypothesis almost surely. We showed the rate of rejection of any wrong hypothesis has an explicit characterization in terms of the local divergences and network topology. Furthermore, under the (mild technical) Assumption 5 on the tail of the log-likelihood ratios of observations, we provide an asymptotically tight characterization of rate of concentration for the rate of rejection of wrong hypotheses. This assumption admits a broad class of distributions with unbounded support such as Gaussian mixtures. In the next subsections we address two important aspects of our algorithm construction and network model.

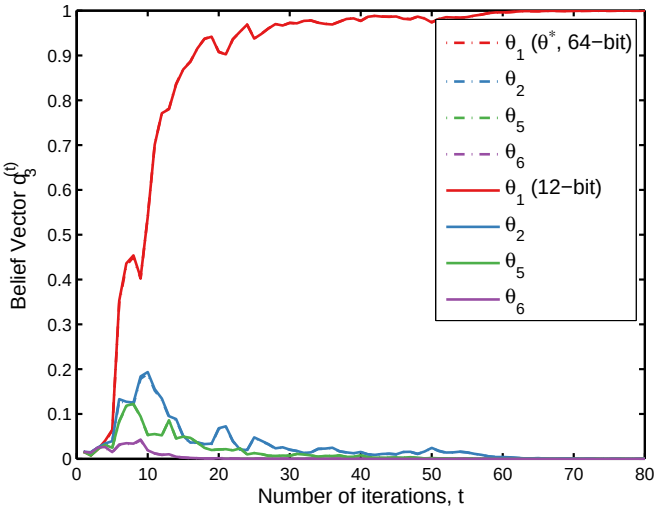


Figure 5.13: The solid lines in figure show the evolution of the beliefs of node 3 with time for hypotheses θ_2, θ_5 and θ_6 when links support a maximum of 12 bits per hypothesis per unit time. This is compared with the evolution of the beliefs with no rate restriction case (dotted lines) which in our simulations translates to the case when the links support a maximum of 64 bits per hypothesis per unit time. On the same sample path in Figure 5.12, we see that learning rule converges to true hypothesis when the communication is restricted to 12 bits per hypothesis.

5.5.1 Lack of Knowledge of Joint Observation Distribution

Our algorithm does not require that the the nodes in the network (a) have knowledge of the full joint distribution of the observations nor (b) share their raw local observations. These two properties of our algorithm are highly desirable in many social network settings due to privacy considerations. The performance of our algorithm seems to be overtly pessimistic compared to the performance of a fully cooperative network with identically distributed and independent observations across the nodes (where the rate of rejecting the wrong hypothesis is n times our rate $K(\theta^*, \theta)$). However interestingly, in the case of fully correlated identical observations across the network, our algorithm performs as well as a centralized aggregator would perform. In short, our work can be viewed as a first step towards addressing these questions in settings where nodes keep their local observations and marginal distributions and completely prioritizing local privacy. Nonetheless, we acknowledge that many non-trivial questions remain: (i) what is the trade-off

between privacy preservation and learning rate and (ii) what are the cost/benefits of learning the joint distribution in order to optimally combine the local observations.

5.5.2 Availability of Perfect Communication Links

In this work, we have assumed that communicating public beliefs among the neighbors can occur with an infinite precision. Although this is a hard assumption to justify in resource-constrained settings, we believe that it is a reasonable abstraction for a practical “protocol-level" model of communication constraints, in which sufficiently high data rates are available to send messages when nodes are within each others’ communication range, whereas no communication is possible for physically distant nodes. In Section 5.4.3, we have provided detailed simulations to show that the gap between the true model and the idealized protocol model is not of significant practical consequence. In particular, Examples 3 and 5 show the impact of quantizing the beliefs before exchanging them is negligible at even low link rates. However, from a theoretical perspective, a study of distributed hypothesis testing with constraints on communication is a major topic of ongoing research [64, 100].

Furthermore, through Example 5, we have also highlighted the practical gains, in terms of communication, associated with communicating the beliefs instead of the raw local observations where the observations are in a high dimensional space. In other words, the nodes that rely on our learning rule do not need to keep track of their neighbors’ reported observations, but only the beliefs.

5.6 Appendix

5.6.1 Proof of Theorem 7

We begin with the following recursion for each node i and $k \in [M-1]$:

$$\begin{aligned}\log \frac{q_i^{(t)}(\theta_M)}{q_i^{(t)}(\theta_k)} &= \sum_{j=1}^{n} W_{ij} \log \frac{b_j^{(t)}(\theta_M)}{b_j^{(t)}(\theta_k)} \\ &= \sum_{j=1}^{n} W_{ij} \left(\log \frac{f_j\left(X_j^{(t)};\theta_M\right)}{f_j\left(X_j^{(t)};\theta_k\right)} + \log \frac{q_j^{(t-1)}(\theta_M)}{q_j^{(t-1)}(\theta_k)} \right), \end{aligned} \tag{5.55}$$

where the first and the second equalities follow from (5.3) and (6.2), respectively. Now for each node j we rewrite $\log \frac{q_j^{(\cdot)}(\theta_M)}{q_j^{(\cdot)}(\theta_k)}$ in terms of node j's neighbors and their samples at the previous instants. We can expand in this way until we express everything in terms of the samples collected and the initial estimates. Noting that $W^t(i,j) = \sum_{i_{t-1}=1}^{n} \dots \sum_{i_1=1}^{n} W_{ii_1} \dots W_{i_{t-1}j}$, it is easy to check that (5.55) can be further expanded to obtain the following:

$$\begin{aligned}\lim_{t\to\infty} \frac{1}{t} \log \frac{q_i^{(t)}(\theta_M)}{q_i^{(t)}(\theta_k)} &= \lim_{t\to\infty} \frac{1}{t} \sum_{j=1}^{n} \sum_{\tau=1}^{t} W^{\tau}(i,j) \log \frac{f_j\left(X_j^{(t-\tau+1)};\theta_M\right)}{f_j\left(X_j^{(t-\tau+1)};\theta_k\right)} \\ &\quad + \lim_{t\to\infty} \frac{1}{t} \sum_{j=1}^{n} W^t(i,j) \log \frac{q_j^{(0)}(\theta_M)}{q_j^{(0)}(\theta_k)}. \end{aligned} \tag{5.56}$$

From Assumption 3, the prior $q_j^{(0)}(\theta_k)$ is strictly positive for every node j and every $k \in [M]$. Since $W^t(i,j) \leq 1$, we have

$$\lim_{t\to\infty} \frac{1}{t} \left\{ \sum_{j=1}^{n} W^t(i,j) \log \frac{q_j^{(0)}(\theta_M)}{q_j^{(0)}(\theta_k)} \right\} = 0. \tag{5.57}$$

Let W be periodic with period d. If W is aperiodic, then the same proof still holds by

putting $d = 1$. Now, we fix node i as a reference node and for every $r \in [d]$, define

$$A_r = \{j \in [n] : W^{md+r}(i,j) > 0 \text{ for some } m \in \mathbb{N}\}.$$

In particular, $(A_1, A_2, \ldots, A_d)$ is a partition of $[n]$; these sets form cyclic classes of the Markov chain. Fact 5 implies that for every $\delta > 0$, there exists an integer N which is function of δ alone, such that for all $m \geq N$, for some fixed $r \in [d]$, if $j \in A_r$, then

$$\left| W^{md+r}(i,j) - v_j d \right| \leq \delta \tag{5.58}$$

and if $j \notin A_r$

$$0 \leq W^{md+r}(i,j) \leq \delta. \tag{5.59}$$

Using this the first term in (5.56) can be decomposed as follows

$$\begin{aligned}
&\lim_{t\to\infty} \frac{1}{t} \sum_{j=1}^{n} \sum_{\tau=1}^{t} W^{\tau}(i,j) \log \frac{f_j\left(X_j^{(t-\tau+1)}; \theta_M\right)}{f_j\left(X_j^{(t-\tau+1)}; \theta_k\right)} \\
&= \lim_{t\to\infty} \frac{1}{t} \sum_{j=1}^{n} \sum_{\tau=1}^{Nd-1} W^{\tau}(i,j) \log \frac{f_j\left(X_j^{(t-\tau+1)}; \theta_M\right)}{f_j\left(X_j^{(t-\tau+1)}; \theta_k\right)} \\
&\quad + \lim_{t\to\infty} \frac{1}{t} \sum_{j=1}^{n} \sum_{\tau=Nd}^{t} W^{\tau}(i,j) \log \frac{f_j\left(X_j^{(t-\tau+1)}; \theta_M\right)}{f_j\left(X_j^{(t-\tau+1)}; \theta_k\right)}.
\end{aligned} \tag{5.60}$$

Using the triangle inequality and the fact that $W^{\tau}(i,j) \leq 1$ for every $\tau \in \mathbb{N}$ we have

$$\left| \lim_{t\to\infty} \frac{1}{t} \sum_{\tau=1}^{Nd-1} W^{\tau}(i,j) \log \frac{f_j\left(X_j^{(t-\tau)}; \theta_M\right)}{f_j\left(X_j^{(t-\tau)}; \theta_k\right)} \right| \leq \lim_{t\to\infty} \frac{1}{t} \sum_{\tau=1}^{Nd-1} \left| \log \frac{f_j\left(X_j^{(t-\tau)}; \theta_M\right)}{f_j\left(X_j^{(t-\tau)}; \theta_k\right)} \right|.$$

For every $j \in [n]$, $\log \frac{f_j(X_j;\theta_M)}{f_j(X_j;\theta_k)}$ is integrable, implying $\left| \log \frac{f_j(X_j;\theta_M)}{f_j(X_j;\theta_k)} \right|$ is almost surely finite. This

implies that

$$\lim_{t\to\infty}\frac{1}{t}\sum_{\tau=1}^{Nd-1}W^{\tau}(i,j)\log\frac{f_j\left(X_j^{(t-\tau)};\theta_M\right)}{f_j\left(X_j^{(t-\tau)};\theta_k\right)}=0\quad\text{P-a.s.}\tag{5.61}$$

Using (5.57) and (5.61), (5.60) becomes

$$\lim_{t\to\infty}\frac{1}{t}\log\frac{q_i^{(t)}(\theta_M)}{q_i^{(t)}(\theta_k)}=\lim_{t\to\infty}\frac{1}{t}\sum_{j=1}^{n}\sum_{\tau=Nd}^{t}W^{\tau}(i,j)\log\frac{f_j\left(X_j^{(t-\tau+1)};\theta_M\right)}{f_j\left(X_j^{(t-\tau+1)};\theta_k\right)}$$

with probability one. It is straightforward to see that the above equation can be rewritten as

$$\lim_{t\to\infty}\frac{1}{t}\log\frac{q_i^{(t)}(\theta_M)}{q_i^{(t)}(\theta_k)}=\lim_{T\to\infty}\frac{1}{Td}\sum_{j=1}^{n}\sum_{m=N}^{T-1}\left\{\sum_{r=1}^{d}W^{md+r}(i,j)\log\frac{f_j\left(X_j^{(Td-md-r+1)};\theta_M\right)}{f_j\left(X_j^{(Td-md-r+1)};\theta_k\right)}\right\}$$

with probability one. For every $\delta>0$ and N such that for all $m\in N$ equations (5.58) and (5.59) hold true, using Lemma 12 we get that

$$\lim_{t\to\infty}\frac{1}{t}\log\frac{q_i^{(t)}(\theta_M)}{q_i^{(t)}(\theta_k)}$$

with probability one lies in the interval with end points

$$K(\theta_M,\theta_k)-\frac{\delta}{d}\sum_{j=1}^{n}\mathbb{E}\left[\left|\log\frac{f_j\left(X_j;\theta_M\right)}{f_j\left(X_j;\theta_k\right)}\right|\right]$$

and

$$K(\theta_M,\theta_k)+\frac{\delta}{d}\sum_{j=1}^{n}\mathbb{E}\left[\left|\log\frac{f_j\left(X_j;\theta_M\right)}{f_j\left(X_j;\theta_k\right)}\right|\right].$$

Since this holds for any $\delta > 0$, we have

$$\lim_{t\to\infty} \frac{1}{t} \log \frac{q_i^{(t)}(\theta_M)}{q_i^{(t)}(\theta_k)} = K(\theta_M, \theta_k) \quad \mathsf{P}\text{-a.s.}$$

Hence, with probability one, for every $\varepsilon > 0$ there exists a time T' such that $\forall t \geq T'$, $\forall k \in [M-1]$ we have

$$\left| \frac{1}{t} \log \frac{q_i^{(t)}(\theta_M)}{q_i^{(t)}(\theta_k)} - K(\theta_M, \theta_k) \right| \leq \varepsilon,$$

which implies

$$\frac{1}{1 + \sum_{k\in[M-1]} e^{-K(\theta_M,\theta_k)t + \varepsilon t}} \leq q_i^{(t)}(\theta_M) \leq 1.$$

Hence we have the assertion of the theorem.

Lemma 12. *For a given $\delta > 0$ and for some $N \in \mathbb{N}$ for which* (5.58) *and* (5.59) *hold true for all $m \geq N$, the following expression*

$$\lim_{T\to\infty} \frac{1}{Td} \sum_{j=1}^{n} \sum_{m=N}^{T-1} \left\{ \sum_{r=1}^{d} W^{md+r}(i,j) \log \frac{f_j\left(X_j^{(Td-md-r+1)}; \theta_M\right)}{f_j\left(X_j^{(Td-md-r+1)}; \theta_k\right)} \right\}$$

with probability one lies in an interval with end points

$$K(\theta_M, \theta_k) - \frac{\delta}{d} \sum_{j=1}^{n} \mathbb{E}\left[\left| \log \frac{f_j\left(X_j; \theta_M\right)}{f_j\left(X_j; \theta_k\right)} \right| \right],$$

and

$$K(\theta_M, \theta_k) + \frac{\delta}{d} \sum_{j=1}^{n} \mathbb{E}\left[\left| \log \frac{f_j\left(X_j; \theta_M\right)}{f_j\left(X_j; \theta_k\right)} \right| \right].$$

Proof. To the given expression we add and subtract $v_j d$ from $W^{md+r}(i,j)$ for all $j \in A_r$ to obtain

$$\begin{aligned}
&\lim_{T\to\infty} \frac{1}{Td} \sum_{j=1}^{n} \sum_{m=N}^{T-1} \left\{ \sum_{r=1}^{d} W^{md+r}(i,j) \log \frac{f_j\left(X_j^{(Td-md-r+1)};\theta_M\right)}{f_j\left(X_j^{(Td-md-r+1)};\theta_k\right)} \right\} \\
&= \sum_{r=1}^{d} \sum_{j\notin A_r} \left\{ \lim_{T\to\infty} \frac{1}{Td} \sum_{m=N}^{T-1} W^{md+r}(i,j) \log \frac{f_j\left(X_j^{(Td-md-r+1)};\theta_M\right)}{f_j\left(X_j^{(Td-md-r+1)};\theta_k\right)} \right\} \\
&\quad + \sum_{r=1}^{d} \sum_{j\in A_r} \left\{ \lim_{T\to\infty} \frac{1}{Td} \sum_{m=N}^{T-1} \left(W^{md+r}(i,j) - v_j d\right) \log \frac{f_j\left(X_j^{(Td-md-r+1)};\theta_M\right)}{f_j\left(X_j^{(Td-md-r+1)};\theta_k\right)} \right\} \\
&\quad + \sum_{r=1}^{d} \sum_{j\in A_r} \left\{ \lim_{T\to\infty} \frac{1}{Td} \sum_{m=N}^{T-1} v_j d \log \frac{f_j\left(X_j^{(Td-md-r+1)};\theta_M\right)}{f_j\left(X_j^{(Td-md-r+1)};\theta_k\right)} \right\}. \qquad (5.62)
\end{aligned}$$

For each r and some $j \in A_r$, using (5.58) and the strong law of large numbers we have

$$\begin{aligned}
&\left| \lim_{T\to\infty} \frac{1}{Td} \left\{ \sum_{m=N}^{T-1} \left(W^{md+r}(i,j) - v_j d\right) \log \frac{f_j\left(X_j^{(Td-md-r+1)};\theta_M\right)}{f_j\left(X_j^{(Td-md-r+1)};\theta_k\right)} \right\} \right| \\
&\leq \frac{\delta}{d} \mathbb{E}\left[\left| \log \frac{f_j\left(X_j;\theta_M\right)}{f_j\left(X_j;\theta_k\right)} \right| \right] \quad \text{P-a.s..}
\end{aligned}$$

Similarly for $j \notin A_r$, using (5.59) we have

$$\left| \lim_{T\to\infty} \frac{1}{Td} \sum_{m=N}^{T-1} W^{md+r}(i,j) \log \frac{f_j\left(X_j^{(Td-md-r+1)};\theta_M\right)}{f_j\left(X_j^{(Td-md-r+1)};\theta_k\right)} \right| \leq \frac{\delta}{d} \mathbb{E}\left[\left| \log \frac{f_j\left(X_j;\theta_M\right)}{f_j\left(X_j;\theta_k\right)} \right| \right] \quad \text{P-a.s..}$$

Again, by the strong law of large numbers we have

$$\begin{aligned}
\sum_{r=1}^{d} \sum_{j\in A_r} v_j \left\{ \lim_{T\to\infty} \frac{1}{T} \sum_{m=N}^{T-1} \log \frac{f_j\left(X_j^{(Td-md-r+1)};\theta_M\right)}{f_j\left(X_j^{(Td-md-r+1)};\theta_k\right)} \right\} &= \sum_{r=1}^{d} \sum_{j\in A_r} v_j \mathbb{E}\left[\log \frac{f_j\left(X_j;\theta_M\right)}{f_j\left(X_j;\theta_k\right)} \right] \\
&= K(\theta_M,\theta_k) \quad \text{P-a.s..}
\end{aligned}$$

Now combining this with (5.62) we have the assertion of the lemma.

□

5.6.2 Proof of Theorem 8

Recall the following equation:

$$\lim_{t\to\infty}\frac{1}{t}\log\frac{q_i^{(t)}(\theta_M)}{q_i^{(t)}(\theta_k)} = \lim_{t\to\infty}\frac{1}{t}\sum_{j=1}^{n}\sum_{\tau=1}^{Nd-1}W^{\tau}(i,j)\log\frac{f_j\left(X_j^{(t-\tau+1)};\theta_M\right)}{f_j\left(X_j^{(t-\tau+1)};\theta_k\right)} + \lim_{t\to\infty}\frac{1}{t}\sum_{j=1}^{n}\sum_{\tau=Nd}^{t}W^{\tau}(i,j)\log\frac{f_j\left(X_j^{(t-\tau+1)};\theta_M\right)}{f_j\left(X_j^{(t-\tau+1)};\theta_k\right)}, \tag{5.63}$$

where N is such that for all $m \geq N, m \in \mathbb{N}$ equations (5.58) and (5.59) are satisfied. For any fixed t, using Assumption 4, the first term in the summation on the right hand side of (5.63) can be bounded as

$$\left|\frac{1}{t}\sum_{j=1}^{n}\sum_{\tau=1}^{Nd-1}W^{\tau}(i,j)\log\frac{f_j\left(X_j^{(t-\tau+1)};\theta_M\right)}{f_j\left(X_j^{(t-\tau+1)};\theta_k\right)}\right| \leq \frac{nNdL}{t}.$$

Also, the second term in the summation on the right hand side of (5.63) can be bounded as

$$\left|\frac{1}{t}\sum_{j=1}^{n}\sum_{\tau=Nd}^{t}W^{\tau}(i,j)\log\frac{f_j\left(X_j^{(t-\tau+1)};\theta_M\right)}{f_j\left(X_j^{(t-\tau+1)};\theta_k\right)} - \sum_{r=1}^{d}\sum_{j\in A_r}\frac{v_j}{Td}\sum_{m=0}^{T-1}\log\frac{f_j\left(X_j^{(Td-md-r+1)};\theta_M\right)}{f_j\left(X_j^{(Td-md-r+1)};\theta_k\right)}\right|$$
$$\leq \delta\frac{1}{Td}\sum_{m=0}^{T-1}\left|\log\frac{f_j\left(X_j^{(Td-md-r+1)};\theta_M\right)}{f_j\left(X_j^{(Td-md-r+1)};\theta_k\right)}\right|.$$

Using Assumption 4 we have

$$\frac{1}{Td}\sum_{m=0}^{T-1}\left|\log\frac{f_j\left(X_j^{(Td-md-r+1)};\theta_M\right)}{f_j\left(X_j^{(Td-md-r+1)};\theta_k\right)}\right|\leq\frac{L}{d}.$$

Therefore, we have

$$\left|\frac{1}{t}\log\frac{q_i^{(t)}(\theta_M)}{q_i^{(t)}(\theta_k)}-\sum_{r=1}^{d}\sum_{j\in A_r}\frac{v_j}{Td}\sum_{m=0}^{T-1}\log\frac{f_j\left(X_j^{(Td-md-r+1)};\theta_M\right)}{f_j\left(X_j^{(Td-md-r+1)};\theta_k\right)}\right|\leq\frac{\delta nL}{d}.$$

Applying Hoeffding's inequality (Theorem 2 of [101]), for every $0<\varepsilon\leq K(\theta_M,\theta_k)$, we can write (5.63) for $t\geq Nd$ as

$$\frac{1}{t}\log\frac{q_i^{(t)}(\theta_M)}{q_i^{(t)}(\theta_k)}\leq K(\theta_M,\theta_k)-\varepsilon+o\left(\frac{1}{t},\delta\right),$$

with probability at most $\exp\left(-\frac{\varepsilon^2T}{2L^2}\right)$ where $o\left(\frac{1}{t},\delta\right)=\frac{\delta nL}{d}+\frac{nNdL}{t}$. Similarly, for $0<\varepsilon\leq L-K(\theta_M,\theta_k)$ we have

$$\frac{1}{t}\log\frac{q_i^{(t)}(\theta_M)}{q_i^{(t)}(\theta_k)}\geq K(\theta_M,\theta_k)+\varepsilon+o\left(\frac{1}{t},\delta\right),$$

with probability at most $\exp\left(-\frac{\varepsilon^2T}{2L^2}\right)$ and for $\varepsilon>L-K(\theta_M,\theta_k)$ we have

$$\frac{1}{t}\log\frac{q_i^{(t)}(\theta_M)}{q_i^{(t)}(\theta_k)}\geq K(\theta_M,\theta_k)+\varepsilon+o\left(\frac{1}{t},\delta\right),$$

with probability 0. Now, taking limit and letting δ go to zero, for $0<\varepsilon\leq K(\theta_M,\theta_k)$ we have

$$\lim_{t\to\infty}\frac{1}{t}\log\mathsf{P}\left(\rho_i^{(t)}(\theta_k)-\rho_i^{(t)}(\theta_M)\leq K(\theta_M,\theta_k)-\varepsilon\right)\leq-\frac{\varepsilon^2}{2L^2d},$$

for $0 < \varepsilon \leq L - K(\theta_M, \theta_k)$ we have

$$\lim_{t\to\infty} \frac{1}{t} \log \mathsf{P}\left(\rho_i^{(t)}(\theta_k) - \rho_i^{(t)}(\theta_M) \geq K(\theta_M, \theta_k) + \varepsilon\right) \leq -\frac{\varepsilon^2}{2L^2 d},$$

and for $\varepsilon > L - K(\theta_M, \theta_k)$ we have

$$\lim_{t\to\infty} \frac{1}{t} \log \mathsf{P}\left(\rho_i^{(t)}(\theta_k) - \rho_i^{(t)}(\theta_M) \geq K(\theta_M, \theta_k) + \varepsilon\right) = -\infty.$$

Since $q_i^{(t)}(\theta_M) \leq 1$, all the events ω which lie in the set $\{\omega : \rho_i^{(t)}(\theta_k) \leq K(\theta_M, \theta_k) - \varepsilon\}$ also lie in the set $\{\omega : \rho_i^{(t)}(\theta_k) \leq K(\theta_M, \theta_k) - \varepsilon + \rho_i^{(t)}(\theta_M)\}$. Hence, for every $0 < \varepsilon \leq K(\theta_M, \theta_k)$ we have

$$\lim_{t\to\infty} \frac{1}{t} \log \mathsf{P}\left(\rho_i^{(t)}(\theta_k) \leq K(\theta_M, \theta_k) - \varepsilon\right) \leq -\frac{\varepsilon^2}{2L^2 d}. \tag{5.64}$$

For $k \in [M-1]$ and any $\alpha \geq 0$, the set

$$\left\{\rho_i^{(t)}(\theta_k) \geq K(\theta_M, \theta_k) + \varepsilon\right\}$$

lies in the complement of the following set:

$$\left\{\rho_i^{(t)}(\theta_k) - \rho_i^{(t)}(\theta_M) < K(\theta_M, \theta_k) + \varepsilon - \alpha\right\} \cap \left\{\rho_i^{(t)}(\theta_M) < \alpha\right\}.$$

This implies that

$$\begin{aligned} &\mathsf{P}\left(\rho_i^{(t)}(\theta_k) \geq K(\theta_M, \theta_k) + \varepsilon\right) \\ &\leq \mathsf{P}\left(\rho_i^{(t)}(\theta_k) - \rho_i^{(t)}(\theta_M) \geq K(\theta_M, \theta_k) + \varepsilon - \alpha\right) + \mathsf{P}\left(\rho_i^{(t)}(\theta_M) \geq \alpha\right). \end{aligned} \tag{5.65}$$

Using Lemma 13 we have that for every $\delta > 0$ there exists a T such that for all $t \geq T$

$$\mathsf{P}\left(\rho_i^{(t)}(\theta_k) \geq K(\theta_M, \theta_k) + \varepsilon\right) \tag{5.66}$$

$$\leq \exp\left(-\frac{(\varepsilon-\alpha)^2}{2L^2 d}t + \delta t\right) + \exp\left(-\min_{k\in[M-1]}\left\{\frac{K(\theta_M,\theta_k)^2}{2L^2 d}\right\}t + \delta t\right). \tag{5.67}$$

Taking the limit as $\alpha \to 0^+$ for $0 < \varepsilon \leq L - K(\theta_M, \theta_k)$ we have

$$\lim_{t\to\infty}\frac{1}{t}\log \mathsf{P}\left(\rho_i^{(t)}(\theta_k) \geq K(\theta_M,\theta_k)+\varepsilon\right) \leq -\frac{1}{2L^2 d}\min\left\{\varepsilon^2, \min_{j\in[M-1]} K^2(\theta_M,\theta_j)\right\}. \tag{5.68}$$

For $\varepsilon \geq L - K(\theta_M, \theta_k)$ we have

$$\lim_{t\to\infty}\frac{1}{t}\log \mathsf{P}\left(\rho_i^{(t)}(\theta_k) \geq K(\theta_M,\theta_k)+\varepsilon\right) \leq -\min_{k\in[M-1]}\left\{\frac{K(\theta_M,\theta_k)^2}{2L^2 d}\right\}. \tag{5.69}$$

Lemma 13. *For all $\alpha > 0$, we have the following for the sequence $q_i^{(t)}(\theta_M)$*

$$\lim_{t\to\infty}\frac{1}{t}\log \mathsf{P}\left(\rho_i^{(t)}(\theta_M) \geq \alpha\right) \leq -\min_{k\in[M-1]}\left\{\frac{K(\theta_M,\theta_k)^2}{2L^2 d}\right\}. \tag{5.70}$$

Proof. For any $\alpha > 0$, consider

$$\mathsf{P}\left(\rho_i^{(t)}(\theta_M) \geq \alpha\right) \leq \sum_{k\in[M-1]} \mathsf{P}\left(\frac{1}{M-1}\left(1-e^{-\alpha t}\right) \leq q_i^{(t)}(\theta_k)\right)$$

$$= \sum_{k\in[M-1]} \mathsf{P}\left(\rho_i^{(t)}(\theta_k) \leq K(\theta_M,\theta_k) - \eta_t(\theta_k)\right), \tag{5.71}$$

where $\eta_t(\theta_k) = K(\theta_M,\theta_k) - \frac{1}{t}\log(M-1) + \frac{1}{t}\log\left(1-e^{-\alpha t}\right)$. For every $\varepsilon > 0$, there exists $T(\varepsilon)$

such that for all $t \geq T(\varepsilon)$ we have

$$\begin{aligned}\mathsf{P}\left(\rho_i^{(t)}(\theta_k) \geq \alpha\right) &\leq \sum_{k\in[M-1]} \mathsf{P}\left(\rho_i^{(t)}(\theta_k) \leq K(\theta_M,\theta_k) - K(\theta_M,\theta_k) + \varepsilon\right) \\ &= \sum_{k\in[M-1]} \mathsf{P}\left(\rho_i^{(t)}(\theta_k) \leq \varepsilon\right).\end{aligned}$$

Therefore, for every $\varepsilon > 0$, $\delta > 0$, there exists $T = \max\{T(\varepsilon), T(\delta)\}$ such that for all $t \geq T$ we have

$$\mathsf{P}\left(\rho_i^{(t)}(\theta_M) \geq \alpha\right) \leq (M-1) \max_{k\in[M-1]} \exp\left\{-\frac{(K(\theta_M,\theta_k)-\varepsilon)^2}{2L^2 d} t + \delta t\right\}.$$

By taking the limit and making ε arbitrarily small, we have

$$\lim_{t\to\infty} \frac{1}{t} \log \mathsf{P}\left(\rho_i^{(t)}(\theta_M) \geq \alpha\right) \leq -\min_{k\in[M-1]} \left\{\frac{K(\theta_M,\theta_k)^2}{2L^2 d}\right\}.$$

□

Proof of Corollary 11

From Theorem 8, we have

$$\lim_{t\to\infty} \frac{1}{t} \log \mathsf{P}\left(\mu_i \geq \min_{k\in[M-1]} K(\theta_M,\theta_k) + \varepsilon\right) \leq -\frac{1}{2L^2 d} \min\left\{\varepsilon^2, \min_{k\in[M-1]} K(\theta_M,\theta_k)^2\right\}.$$

Now, applying the Borel-Cantelli Lemma to the above equation we have

$$\mu_i \leq \min_{k\in[M-1]} K(\theta_M,\theta_k) \quad \mathsf{P}\text{-a.s.}$$

Letting $\varepsilon \to 0$ and by combining this with Corollary 9 we have

$$\mu_i = \min_{k \in [M-1]} K(\theta_M, \theta_k) \quad \mathsf{P}\text{-a.s.}$$

5.6.3 Proof of Theorem 9

Fact 6 (Cramer's Theorem, Theorem 3.8 [102])**.** *Consider a sequence of d-dimensional i.i.d random vectors* $\{\mathbf{X}_n\}_{n=1}^{\infty}$. *Let* $\mathbf{S}_n = \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i$. *Then, the sequence of* $\mathbf{S}_n$ *satisfies a large deviation principle with rate function* $\Lambda^*(\cdot)$, *namely: For any set* $F \subset \mathbb{R}^d$,

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathsf{P}(\mathbf{S}_n \in F) \geq -\inf_{\mathbf{x} \in F^o}, \tag{5.72}$$

and

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathsf{P}(\mathbf{S}_n \in F) \leq -\inf_{\mathbf{x} \in \bar{F}}, \tag{5.73}$$

where $\Lambda^*(\cdot)$ *is given by*

$$\Lambda^*(\mathbf{x}) := \sup_{\lambda \in \mathbb{R}^d} \{\langle \lambda, \mathbf{x} \rangle - \Lambda(\lambda)\}. \tag{5.74}$$

and $\Lambda(\cdot)$ *is the log moment generating function of* $\mathbf{S}_n$ *which is given by*

$$\Lambda(\lambda) := \log \mathbb{E}[e^{\langle \lambda, \mathbf{Y} \rangle}]. \tag{5.75}$$

Fact 7 (Contraction Principle, Theorem 3.20 [102])**.** *Let* $\{\mathsf{P}_t\}$ *be a sequence of probability*

measures on a Polish space $\mathcal{X}$ that satisfies LDP with rate function I. Let

$$\begin{cases} \mathcal{Y} & \text{be a Polish space} \\ T : \mathcal{X} \to \mathcal{Y} & \text{a continuous map} \\ \mathsf{Q}_t = \mathsf{P}_t \circ T^{-1} & \text{an image probability measure.} \end{cases} \tag{5.76}$$

Then $\{\mathsf{Q}_t\}$ satisfies the LDP on $\mathcal{Y}$ with rate function J given by

$$J(y) = \inf_{x \in \mathcal{X}: T(x)=y} I(x). \tag{5.77}$$

To prove that $\frac{1}{t}\log \tilde{\mathbf{q}}_{\mathbf{i}}^{(\mathbf{t})}$ satisfies the LDP, first we establish the LDP satisfied by the following vector:

$$\mathbf{Q}_i^{(t)} = \left[\frac{q_i^{(t)}(\theta_1)}{q_i^{(t)}(\theta_M)}, \frac{q_i^{(t)}(\theta_2)}{q_i^{(t)}(\theta_M)}, \ldots, \frac{q_i^{(t)}(\theta_{M-1})}{q_i^{(t)}(\theta_M)} \right]^T. \tag{5.78}$$

Note that $\mathbf{Q}_i^{(t)} = \frac{\tilde{\mathbf{q}}_{\mathbf{i}}^{(\mathbf{t})}}{q_i^{(t)}(\theta_M)}$. From Lemma 14, we obtain that $\frac{1}{t}\log \mathbf{Q}_i^{(t)}$ satisfies the LDP with rate function $I(\cdot)$, as given by (5.37). Now we apply the Contraction Principle (Fact 7), for

$$\begin{aligned} &\mathcal{X} = \mathbb{R}^{M-1}, \quad \mathcal{Y} = \mathbb{R}^{M-1}, \\ &T(\mathbf{x}) = g(\mathbf{x}), \quad \forall \mathbf{x} \in \mathbb{R}^{M-1}, \\ &\mathsf{P}_t = \mathsf{P}\left(\frac{1}{t}\log \mathbf{Q}_i^{(t)} \in \cdot \right), \\ &\mathsf{Q}_t = \mathsf{P}\left(g\left(\frac{1}{t}\log \mathbf{Q}_i^{(t)} \right) \in \cdot \right), \end{aligned}$$

and we get that $g\left(\frac{1}{t}\log \mathbf{Q}_i^{(t)}\right)$ satisfies an LDP with a rate function $J(\cdot)$, i.e., for every $F \subset \mathbb{R}^{M-1}$

we have

$$\liminf_{t\to\infty} \frac{1}{t} \log \mathsf{P}\left(g\left(\frac{1}{t} \log \mathbf{Q}_i^{(t)} \right) \in F \right) \geq - \inf_{\mathbf{y}\in F^o} J(\mathbf{y}), \tag{5.79}$$

and

$$\limsup_{t\to\infty} \frac{1}{t} \log \mathsf{P}\left(g\left(\frac{1}{t} \log \mathbf{Q}_i^{(t)} \right) \in F \right) \leq - \inf_{\mathbf{y}\in \bar{F}} J(\mathbf{y}). \tag{5.80}$$

Combining Lemma 15 with (5.79) and (5.80), we obtain that $\frac{1}{t} \log \tilde{\mathbf{q}}_{\mathbf{i}}^{(\mathbf{t})}$ satisfies the LDP with rate function $J(\cdot)$ as well. Hence, we have the assertion of the theorem.

Lemma 14. *The random vector $\frac{1}{t} \log \mathbf{Q}_i^{(t)}$ satisfies the LDP with rate function given by $I(\cdot)$ in* (5.36). *That is, for any set $F \subset \mathbb{R}^{M-1}$ with interior F^o and closure $\bar{F}$, we have*

$$\liminf_{t\to\infty} \frac{1}{t} \log \mathsf{P}\left(\frac{1}{t} \log \mathbf{Q}_i^{(t)} \in F \right) \geq - \inf_{\mathbf{x}\in F^o} I(\mathbf{x}), \tag{5.81}$$

and

$$\limsup_{t\to\infty} \frac{1}{t} \log \mathsf{P}\left(\frac{1}{t} \log \mathbf{Q}_i^{(t)} \in F \right) \leq - \inf_{\mathbf{x}\in \bar{F}} I(\mathbf{x}). \tag{5.82}$$

Proof. Using the learning rule we have

$$\begin{aligned} \frac{1}{t} \log \mathbf{Q}_i^{(t)} &= \frac{1}{t} \sum_{\tau=1}^{t} \sum_{j=1}^{n} W^{\tau}(i,j) \mathbf{L}_j^{(t-\tau+1)} \\ &= \frac{1}{t} \sum_{\tau=1}^{t} \sum_{j=1}^{n} \left(W^{\tau}(i,j) - v_j \right) \mathbf{L}_j^{(t-\tau+1)} + \frac{1}{t} \sum_{\tau=1}^{t} \mathbf{Y}^{(\tau)}, \end{aligned} \tag{5.83}$$

where $\mathbf{L}$ is given by (5.33) and $\mathbf{Y}$ by (5.32). Using Cramer's Theorem (Fact 6) in $\mathbb{R}^{M-1}$, for any

set $F \subset \mathbb{R}^{M-1}$, we have

$$\liminf_{t\to\infty} \frac{1}{t} \log \mathsf{P}\left(\frac{1}{t}\sum_{\tau=1}^{t} \mathbf{Y}^{(\tau)} \in F\right) \geq -\inf_{\mathbf{x}\in F^o} I(\mathbf{x}), \tag{5.84}$$

and

$$\limsup_{t\to\infty} \frac{1}{t} \log \mathsf{P}\left(\frac{1}{t}\sum_{\tau=1}^{t} \mathbf{Y}^{(\tau)} \in F\right) \leq -\inf_{\mathbf{x}\in \bar{F}} I(\mathbf{x}). \tag{5.85}$$

Consider

$$\left|\frac{1}{t}\sum_{\tau=1}^{t}\sum_{j=1}^{n} \left(W^{\tau}(i,j) - v_j\right) \mathbf{L}_j^{(t-\tau+1)}\right| \leq \frac{n}{t}\sum_{\tau=1}^{t} |\lambda_{\max}^{\tau}(W)| \left(\sum_{j=1}^{n} \left|\mathbf{L}_j^{(t-\tau+1)}\right|\right). \tag{5.86}$$

From Assumption 5, we have that $\Lambda(\lambda)$ is finite for $\lambda \in \mathbb{R}^n$. Now, using Lemma 16, we have

$$\lim_{t\to\infty} \frac{1}{t} \log \mathsf{P}\left(\left|\frac{1}{t}\sum_{\tau=1}^{t}\sum_{j=1}^{n} \left(W^{\tau}(i,j) - v_j\right) \mathbf{L}_j^{(t-\tau+1)}\right| \geq \delta\right) = -\infty. \tag{5.87}$$

Using Lemma 17 on $\frac{1}{t}\log \mathbf{Q}_i^{(t)}$, we have the assertion of the theorem. □

Lemma 15. *For every set $F \subset \mathbb{R}^{M-1}$ and for all $i \in [n]$, we have*

$$\liminf_{t\to\infty} \frac{1}{t} \log \mathsf{P}\left(\frac{1}{t}\log \tilde{\mathbf{q}}_{\mathbf{i}}^{(\mathbf{t})} \in F\right) \geq \liminf_{t\to\infty} \frac{1}{t} \log \mathsf{P}\left(g\left(\frac{1}{t}\log \mathbf{Q}_i^{(t)}\right) \in F\right), \tag{5.88}$$

and

$$\limsup_{t\to\infty} \frac{1}{t} \log \mathsf{P}\left(\frac{1}{t}\log \tilde{\mathbf{q}}_{\mathbf{i}}^{(\mathbf{t})} \in F\right) \leq \limsup_{t\to\infty} \frac{1}{t} \log \mathsf{P}\left(g\left(\frac{1}{t}\log \mathbf{Q}_i^{(t)}\right) \in F\right). \tag{5.89}$$

Proof. For all $t \geq 0$, we have

$$\frac{1}{t}\log \tilde{\mathbf{q}}_{\mathbf{i}}^{(\mathbf{t})} = g\left(\frac{1}{t}\log \mathbf{Q}_i^{(t)}\right) - \frac{1}{t}\log\left(e^{-C^{(t)}t} + \sum_{j=1}^{M-1} e^{g_j\left(\frac{1}{t}\log \mathbf{Q}_i^{(t)}\right)t}\right)\mathbf{1}, \tag{5.90}$$

where

$$C^{(t)} = \max\left\{0, \frac{1}{t}\log\frac{q_i^{(t)}(\theta_1)}{q_i^{(t)}(\theta_M)}, \frac{1}{t}\log\frac{q_i^{(t)}(\theta_2)}{q_i^{(t)}(\theta_M)}, \ldots, \frac{1}{t}\log\frac{q_i^{(t)}(\theta_{M-1})}{q_i^{(t)}(\theta_M)}\right\}.$$

Also for all $t \geq 0$, we have

$$1 \leq e^{-C^{(t)}t} + \sum_{j=1}^{M-1} e^{g_j\left(\frac{1}{t}\log \mathbf{Q}_i^{(t)}\right)t} \leq M.$$

Hence for all $\varepsilon > 0$, there exists $T(\varepsilon)$ such that for all $t \geq T(\varepsilon)$ we have

$$g\left(\frac{1}{t}\log \mathbf{Q}_i^{(t)}\right) - \varepsilon\mathbf{1} \leq \frac{1}{t}\log \tilde{\mathbf{q}}_{\mathbf{i}}^{(\mathbf{t})} \leq g\left(\frac{1}{t}\log \mathbf{Q}_i^{(t)}\right). \tag{5.91}$$

For any $F \subset \mathbb{R}^{M-1}$, let $F_{\varepsilon^+} = \{\mathbf{x} + \delta\mathbf{1}, \forall 0 < \delta \leq \varepsilon \text{ and } \mathbf{x} \in F\}$, $F_{\varepsilon^-} = \{\mathbf{x} - \delta\mathbf{1}, \forall 0 < \delta \leq \varepsilon \text{ and } \mathbf{x} \in F\}$. Therefore, for every $\varepsilon > 0$ we have

$$\liminf_{t\to\infty} \frac{1}{t}\log \mathsf{P}\left(g\left(\frac{1}{t}\log \mathbf{Q}_i^{(t)}\right) \in F\right) \leq \liminf_{t\to\infty} \frac{1}{t}\log \mathsf{P}\left(\frac{1}{t}\log \tilde{\mathbf{q}}_{\mathbf{i}}^{(\mathbf{t})} \in F_{\varepsilon^-}\right). \tag{5.92}$$

Making ε arbitrarily small, $F_{\varepsilon^-} \to F$, and by monotonicity and continuity of probability measure we have

$$\liminf_{t\to\infty} \frac{1}{t}\log \mathsf{P}\left(g\left(\frac{1}{t}\log \mathbf{Q}_i^{(t)}\right) \in F\right) \leq \liminf_{t\to\infty} \frac{1}{t}\log \mathsf{P}\left(\frac{1}{t}\log \tilde{\mathbf{q}}_{\mathbf{i}}^{(\mathbf{t})} \in F\right). \tag{5.93}$$

For $t \geq T(\varepsilon)$ we also have

$$\frac{1}{t}\log \tilde{\mathbf{q}}_{\mathbf{i}}^{(\mathbf{t})} \leq g\left(\frac{1}{t}\log \mathbf{Q}_i^{(t)}\right) \leq \frac{1}{t}\log \tilde{\mathbf{q}}_{\mathbf{i}}^{(\mathbf{t})} + \varepsilon\mathbf{1}. \tag{5.94}$$

This implies for every $\varepsilon > 0$ we have

$$\limsup_{t\to\infty} \frac{1}{t}\log \mathsf{P}\left(\frac{1}{t}\log \tilde{\mathbf{q}}_{\mathbf{i}}^{(\mathbf{t})} \in F\right) \leq \limsup_{t\to\infty} \frac{1}{t}\log \mathsf{P}\left(g\left(\frac{1}{t}\log \mathbf{Q}_i^{(t)}\right) \in F_{\varepsilon^+}\right). \tag{5.95}$$

Again, by making ε arbitrarily small we have

$$\limsup_{t\to\infty} \frac{1}{t}\log \mathsf{P}\left(\frac{1}{t}\log \tilde{\mathbf{q}}_{\mathbf{i}}^{(\mathbf{t})} \in F\right) \leq \limsup_{t\to\infty} \frac{1}{t}\log \mathsf{P}\left(g\left(\frac{1}{t}\log \mathbf{Q}_i^{(t)}\right) \in F\right). \tag{5.96}$$

Hence, we have the assertion of the lemma. □

5.6.4 Proof of the Lemmas

Lemma 16. *Let q be a real number such that $q \in (0,1)$. Let $\mathbf{X}_i$ be a sequence of non-negative i.i.d random vectors in $\mathbb{R}^n$, distributed as $\mathbf{X}$ and let $\Lambda(\lambda)$ denote its log moment generating function which is finite for $\lambda \in \mathbb{R}^n$, then for every $\delta > 0$, we have*

$$\lim_{t\to\infty} \frac{1}{t}\log \mathsf{P}\left(\frac{1}{t}\sum_{i=1}^{t}(q)^i \mathbf{X}_i \geq \delta\mathbf{1}\right) = -\infty. \tag{5.97}$$

Proof. Applying Chebychev's inequality and using the definition of log moment generating function, for $\lambda \in \mathbb{R}^n$, we have

$$\mathsf{P}\left(\frac{1}{t}\sum_{i=1}^{t}(q)^i \mathbf{X}_i \geq \delta\mathbf{1}\right) \leq e^{-t\left(\langle\lambda,\delta\mathbf{1}\rangle - \frac{1}{t}\sum_{i=1}^{t}\Lambda((q)^i\lambda)\right)}. \tag{5.98}$$

From convexity of Λ, we have $\sum_{i=1}^{t}\Lambda((q)^i\lambda) \leq \Lambda(\lambda)\sum_{i=1}^{t}(q)^i$. Since $\Lambda(\lambda)$ is finite and $\sum_{i=1}^{\infty}(q)^i <$

∞, for all $\delta > 0$ we have

$$\lim_{t\to\infty} \frac{1}{t} \log \mathsf{P}\left(\frac{1}{t}\sum_{i=1}^{t} (q)^i \mathbf{X}_i \geq \delta \mathbf{1}\right) \leq -\langle \lambda, \delta \mathbf{1} \rangle. \tag{5.99}$$

Since, the above equation is true for all $\lambda \in \mathbb{R}^n$, we have the assertion of the lemma. □

Lemma 17. *Consider a sequence $\{\mathbf{Z}^{(t)}\}_{t=0}^{\infty}$ where $\mathbf{Z}^{(t)} \in \mathbb{R}^d$ such that*

$$\mathbf{Z}^{(t)} = \mathbf{X}^{(t)} + \mathbf{Y}^{(t)}, \tag{5.100}$$

where sequences $\{\mathbf{X}^{(t)}\}_{t=0}^{\infty}$ and $\{\mathbf{Y}^{(t)}\}_{t=0}^{\infty}$ have the following properties:

1. *The sequence $\{\mathbf{X}^{(t)}\}_{t=0}^{\infty}$ satisfies*

$$\liminf_{t\to\infty} \frac{1}{t} \log \mathsf{P}\left(\mathbf{X}^{(t)} \in F\right) \geq -\inf_{\mathbf{x}\in F^o} I_X(\mathbf{x}), \tag{5.101}$$

$$\limsup_{t\to\infty} \frac{1}{t} \log \mathsf{P}\left(\mathbf{X}^{(t)} \in F\right) \leq -\inf_{\mathbf{x}\in \bar{F}} I_X(\mathbf{x}), \tag{5.102}$$

   where $I_X : \mathbb{R}^d \to \mathbb{R}$ is a well-defined LDP rate function.

2. *For every $\varepsilon > 0$, sequence $\{\mathbf{Y}^{(t)}\}_{t=0}^{\infty}$ satisfies*

$$\lim_{t\to\infty} \frac{1}{t} \log \mathsf{P}(|\mathbf{Y}^{(t)}| \geq \varepsilon \mathbf{1}) = -\infty. \tag{5.103}$$

   Then $\{\mathbf{Z}^{(t)}\}_{t=0}^{\infty}$ satisfies

$$\liminf_{t\to\infty} \frac{1}{t} \log \mathsf{P}(\mathbf{Z}^{(t)} \in F) \geq -\inf_{\mathbf{x}\in F^o} I_X(\mathbf{x}), \tag{5.104}$$

$$\limsup_{t\to\infty} \frac{1}{t} \log \mathsf{P}(\mathbf{Z}^{(t)} \in F) \leq -\inf_{\mathbf{x}\in \bar{F}} I_X(\mathbf{x}). \tag{5.105}$$

Proof. For every $t \geq 0$, we have

$$\begin{aligned}\mathsf{P}\left(\mathbf{Z}^{(t)} \in F_{\varepsilon^+} \cup F_{\varepsilon^-}\right) &\geq \mathsf{P}\left(\{\mathbf{X}^{(t)} \in F\} \cap \{|\mathbf{Y}^{(t)}| \leq \varepsilon \mathbf{1}\}\right) \\ &\geq \mathsf{P}\left(\mathbf{X}^{(t)} \in F\right) - \mathsf{P}\left(|\mathbf{Y}^{(t)}| > \varepsilon \mathbf{1}\right).\end{aligned}$$

For all $\delta > 0$, there exists a $T(\delta)$ such that for all $t \geq T(\delta)$ we have

$$\mathsf{P}\left(\mathbf{X}^{(t)} \in F\right) \geq e^{-\inf_{\mathbf{x} \in F^o} I_X(\mathbf{x})t - \delta t}.$$

For all $B > 0$, there exists a $T(B)$ such that for all $t \geq T(B)$ we have

$$\mathsf{P}\left(|\mathbf{Y}^{(t)}| > \varepsilon \mathbf{1}\right) \geq e^{-Bt}.$$

Now choose $B > \inf_{\mathbf{x} \in F^o} I_X(\mathbf{x}) + \delta$ and $t \geq \max\{T(\delta), T(B)\}$, then we have

$$\mathsf{P}\left(\mathbf{Z}^{(t)} \in F_{\varepsilon^+} \cup F_{\varepsilon^-}\right) \geq e^{-\inf_{\mathbf{x} \in F^o} I_X(\mathbf{x})t - \delta t}\left(1 - e^{-Bt + \inf_{\mathbf{x} \in F^o} I_X(\mathbf{x})t + \delta t}\right).$$

Sending ε to zero and taking the limit we have

$$\liminf_{t \to \infty} \frac{1}{t} \log \mathsf{P}\left(\mathbf{Z}^{(t)} \in F\right) \geq -\inf_{\mathbf{x} \in F^o} I_X(\mathbf{x}).$$

Similarly, using the fact that $\mathsf{P}(\{\mathbf{Z}^{(t)} \in F\} \cap \{|\mathbf{Y}^{(t)}| \leq \varepsilon \mathbf{1}\}) \leq \mathsf{P}\left(\mathbf{X}^{(t)} \in F_{\varepsilon^+}\right)$ we have the other LDP bound. □

Chapter 5, in full, is a reprint of the material as it appears in the paper: Anusha Lalitha, Tara Javidi and Anand D. Sarwate, "Social Learning and Distributed Hypothesis Testing", in *IEEE Transactions on Information Theory*, vol. 64, no. 9, pp. 6161-6179, Sept. 2018. The dissertation author was the primary investigator and author of this paper.

Chapter 6

Decentralized Bayesian Learning over Graphs

6.1 Introduction

Personal edge devices can often use their locally observed data to learn machine learning models that improve the user experience on the device as well as on other devices. However, the use of local data for learning globally rich machine learning models has to address two important challenges. Firstly, this type of localized data, in isolation from the data collected by other devices, is unlikely to be statistically sufficient to learn a global model. Secondly, there might be severe restrictions on sharing raw forms of personal/local data due to privacy and communication cost concerns. In light of these challenges and restrictions, an alternative approach has emerged which leaves the training data distributed on the edge devices while enabling the decentralized learning of a shared model. This alternative, known as *Federated Learning*, is based on edge devices' periodic communication with a central (cloud-based) server responsible for iterative model aggregation. While addressing the privacy constraints on raw data sharing, and significantly reducing the communication overload as compared to synchronized stochastic gradient descent

(SGD), this approach falls short in fully decentralizing the training procedure. Many practical peer to peer networks are dynamic and a regular access to a fixed central server, which coordinates the learning across devices, is not always possible. Existing methods based on federated learning cannot handle such general networks where central server is absent. To summarize, some of the major challenges encountered in a fully decentralized learning paradigm are: (i)*Statistical Insufficiency:* The local and individually observed data distributions are likely to be less rich than the global training set. For example, a subset of features associated with the global model may be missing locally. (ii)*Restriction on Data Exchange:* Due to privacy concerns, agents do not share their raw training data with the neighbors. Furthermore, model parameter sharing has been shown to reduce the communication requirements significantly. (iii) *Lack of Synchronization:* There may not be a single agent with whom every agent communicates which can synchronize the learning periodically. (iv) *Localized Information Exchange:* Agents are likely to limit their interactions and information exchange to a small group of their peers which can be viewed as the 1-hop neighbors on the social network graph. Furthermore, information obtained from different peers might be viewed differently, requiring a heterogeneous model aggregation strategy.

Contributions: We consider a fully decentralized learning paradigm where agents iteratively update their models using local data and aggregate information from their neighbors to their local models. In particular, we consider a learning rule where agents take a Bayesian-like approach via the introduction of a posterior distribution over a parameter space characterizing the unknown global model. Our theoretical and conceptual contributions are as follows: (i) Our decentralized learning rule generalizes a learning rule considered in the social literature [103–105] by restricting the posterior distribution to a predetermined family of distributions for computational tractability. (ii) We provide theoretical guarantee that each agent will eventually learn the true parameters associated with global model under mild assumptions. (iii) We provide analytical characterization of the rate of convergence of the posterior probability at each agent in the network as a function of network structure and local learning capacity. (iv) Unlike prior work, we allow a

fully general network structure as long as it is strongly connected. As a consequence, our work provides first known theoretical guarantees on convergence for a Bayesian variant of federated learning.

In addition to our theoretical results we show that our methodology can be combined with efficient Bayesian inference techniques to train Bayesian neural networks in a decentralized manner. By empirical studies we show that our theoretical analysis can guide the design of network/social interaction and data partition to achieve convergence. We also show the scalability of our method by training over 100 neural networks on asynchronous time-varying networks. Our Bayesian approach has the added advantage of obtaining confidence value over agents' predictions and can directly benefit from Bayesian learning literature which shows that these models offer robustness to over-fitting, regularization of the weights, uncertainty/confidence estimation, and can easily learn from small datasets [106, 107]. In this regard, our work bridges the gap between decentralized training methodologies and Bayesian neural networks.

Related Work: Our fully decentralized training methodology extends federated learning [108–110] to general graphs in a Bayesian setting and does away with the need of having a centralized controller. In particular, our learning rule also generalizes various Bayesian inference techniques such as [107,111–113] and variational continual learning techniques such as [111,113]. Lastly, our work can be viewed as a Bayesian variant of communication-efficient methods based on SGD [114–116] which also allow the agents to make several local computations and then periodically average the local models. This is unlike decentralized optimization and SGD based methods [117–123] where local (stochastic) gradients are computed for each instance of data and communication happens at a rate comparable to number of local updates. For a detailed overview on the communication-efficient SGD methods contrasted with decentralized optimization methods refer to [124].

6.2 Problem Formulation

The Model: Let $\mathcal{X}$ denote the global input space and let $\mathcal{Y}$ denote the set of all possible labels. The global dataset has input-label pairs belonging to $(\mathcal{X},\mathcal{Y})$ which are distributed as $\mathcal{D} = \mathsf{P}_X \times \mathsf{P}_{Y|X}$. Consider a group of N individual agents, where each agent i has access to input-label pairs taken from a subset $(\mathcal{X}_i,\mathcal{Y})$ such that $\cup_{i=1}^N \mathcal{X}_i \subset \mathcal{X}$. The samples $\left\{X_i^{(1)}, X_i^{(2)}, \ldots\right\}$ are independent and identically distributed (i.i.d), and are generated according to the distribution $\mathsf{P}_i \in \mathcal{P}(\mathcal{X})$. Furthermore, we assume that each agent has a set of *candidate local likelihood functions* over the label space which are parametrized by $\theta \in \Theta$ and given by $\{\ell_i(y \mid \theta, x) : y \in \mathcal{Y}, \theta \in \Theta, x \in \mathcal{X}\}$. Each agent i is aiming to learn a distribution over Θ which achieves the following

$$\inf_{\pi \in \mathcal{P}(\Theta)} \mathbb{E}_{\mathsf{P}_X}\left[D_{\mathrm{KL}}\left(\mathsf{P}_{Y|X}(\cdot|X) \,\middle\|\, \int_\Theta \ell_i(\cdot|\theta,X)\pi(\theta)d\theta \right) \right]. \tag{6.1}$$

Note that for any input $x \sim \mathsf{P}_X$, the distribution $\int_\Theta \ell_i(\cdot|\theta,x)\pi(\theta)d\theta$ denotes predictive distribution over the label space $\mathcal{Y}$. Minimizing the objective in equation (6.1) ensures that each agent makes statistically similar predictions as the true labelling function over the global dataset.

Definition 28. A social learning model is said to be *realizable* if there exists a $\theta^* \in \Theta$ such that $\ell_i(\cdot \mid \theta^*, x) = \mathsf{P}_{Y|X}(\cdot \mid x)$ for $i \in [N]$.

We note that, in the realizable case, the minimizer of equation (6.1) is the trivial distribution which takes value one at θ^* and zero elsewhere. In other words, in the realizable case, each agent's goal is to learn the *true model parameter* θ^*.

Definition 29. If $\mathsf{P}_i = \mathsf{P}_X$ for all agents $i \in [N]$, then all agents have identically distributed observations across the network. We refer to this as the *IID data distribution* setting. In contrast, we call the local data to have *non-IID data distribution* when there exists $i \in [N]$ for which $\mathsf{P}_i \neq \mathsf{P}_X$.

Example 11 (Decentralized Linear Regression with non-IID Data Distribution)**.** Let $d \geq 2$ and $\Theta = \mathbb{R}^d$. Consider a linear realizable model where there exists a $\theta^* = [\theta_0^*, \ldots, \theta_{d-1}^*] \in \Theta$, data input $\mathbf{x} \in \mathbb{R}^d$, the label $y \in \mathbb{R}$ given as $y = \theta^{*T}\phi(\mathbf{x}) + \eta$, where the basis function $\phi : \mathbb{R}^d :\to \mathbb{R}$ provides the feature vector $\phi(\mathbf{x}) = [\phi_0(\mathbf{x}), \ldots, \phi_{d-1}(\mathbf{x})]^T$ and η denotes the additive Gaussian noise $\eta \sim N(0, \alpha^2)$. This implies true probabilistic model generating the labels as well as local likelihood function at any agent i, given an input $\mathbf{x}$ is given by $\mathsf{P}_{Y|X}(y \mid \mathbf{x}) = \ell_i(y \mid \theta, \mathbf{x}) = G(y, \theta^{*T}\phi(\mathbf{x}), \alpha^2)$. Now we consider a non-IID data distribution: Fix some $0 < m < d$ and let

$$\mathcal{X}_1 = \left\{ \mathbf{x} \in \mathbb{R}^d \mid \phi(\mathbf{x}) = [\phi_0(\mathbf{x}), \ldots, \phi_{m-1}(\mathbf{x}), 0, \ldots, 0]^T \right\}$$

and

$$\mathcal{X}_2 = \left\{ \mathbf{x} \in \mathbb{R}^d \mid \phi(\mathbf{x}) = [0, \ldots, 0, \phi_m(\mathbf{x}), \ldots, \phi_{d-1}(\mathbf{x})]^T \right\}.$$

Suppose that agent 1 make observations in $\mathcal{X}_1$ or can access only m features locally. Similarly, agent 2 observations lie in $\mathcal{X}_2$, i.e. the remaining $d - m$ features locally. It is clear that the local features at each agent is such that the true parameter θ^* cannot be locally learned and there is a need for communication and model aggregation.

Example 12 (Decentralized Image Classification using Deep Neural Networks)**.** Consider the problem of learning a neural network which can approximate the input-label probabilistic model with distribution $\mathsf{P}_{Y|X}(\cdot \mid \mathbf{x})$ over the label space for each input image $\mathbf{x} \in \mathcal{X}$. In this setting, the local likelihood function at any agent i, given an image $\mathbf{x} \in \mathcal{X}$ was observed, conditioned on the DNN weights θ is obtained as follows $\ell_i(y \mid \theta, \mathbf{x}) = \text{Softmax}(y, \mathbf{f}_\theta(\mathbf{x})) := \frac{\exp(\mathbf{f}_\theta^y(\mathbf{x}))}{\sum_{y' \in \mathcal{Y}} \exp(\mathbf{f}_\theta^{y'}(\mathbf{x}))}$, where $\mathbf{f}_\theta^y(\cdot)$ denotes the value of the output layer of the neural network at label y.

The Communication Network: We model the communication network between agents via a directed graph with vertex set $[N]$. We define the neighborhood of agent i, denoted by $\mathcal{N}(i)$, as the set of all agents j who have an edge going from j to i. We assume $i \in \mathcal{N}(i)$. Furthermore, if

agent $j \in \mathcal{N}(i)$, agent i receives information from agent j. The social interaction of the agents is characterized by a stochastic matrix W. The weight $W_{ij} \in [0,1]$ is strictly positive if and only if $j \in \mathcal{N}(i)$ and $\sum_{j=1}^{N} W_{ij} = 1$. The weight W_{ij} denotes the confidence agent i has on the information it receives from agent j.

6.2.1 Decentralized Learning Rule

We introduce a decentralized learning rule which generalizes a learning rule considered in the social learning literature [103–105]. However, we restrict local posterior distributions to a predetermined family of distributions. This allows us to implement the decentralized algorithm in a computationally tractable manner. Let $\mathcal{Q} \subset \mathcal{P}(\Theta)$ a family of posterior distributions. Start with $\mathbf{q}_i^{(0)} \in \mathcal{P}(\Theta)$ with $\mathbf{q}_i^{(0)}(\theta) > 0$ for all $\theta \in \Theta$ and $i \in [N]$. At each time step $n = 1, 2, \ldots$ the following events happen at every agent $i \in [N]$:

1. Draw a batch of M i.i.d samples $\left(\mathbf{X}_i^{(n)}, \mathbf{Y}_i^{(n)}\right) \sim \mathsf{P}_{Y|X}(\mathbf{Y}_i^{(n)}|\mathbf{X}_i^{(n)})\mathsf{P}_i^M(\mathbf{X}_i^{(n)})$.

2. **Local Bayesian Update of Posterior:** Perform a local Bayesian update on $\mathbf{q}_i^{(n-1)}$ to form the public posterior vector $\mathbf{b}_i^{(n)}$ using the following rule. For each $\theta \in \Theta$,

$$b_i^{(n)}(\theta) = \frac{\ell_i\left(\mathbf{Y}_i^{(n)} \mid \theta, \mathbf{X}_i^{(n)}\right) q_i^{(n-1)}(\theta)}{\int_\Theta \ell_i\left(\mathbf{Y}_i^{(n)} \mid \phi, \mathbf{X}_i^{(n)}\right) q_i^{(n-1)}(\phi) d\phi}. \tag{6.2}$$

3. **Projection onto Allowed Family of Posteriors:** Project onto an allowed family of posterior distributions $\mathcal{Q}$ by employing KL-divergence minimization,

$$\Pi_{\mathcal{Q}}(\mathbf{b}_i^{(n)}) = \operatorname*{argmin}_{\pi \in \mathcal{Q}} D\left(\pi \middle\| \mathbf{b}_i^{(n)}\right). \tag{6.3}$$

4. **Communication Step:** Agent i sends $\Pi_{\mathcal{Q}}(\mathbf{b}_i^{(n)})$ to agent j if $i \in \mathcal{N}(j)$ and receives $\Pi_{\mathcal{Q}}(\mathbf{b}_j^{(n)})$ from neighbors $j \in \mathcal{N}(i)$.

5. **Consensus Step:** Update private posterior distribution by averaging the log posterior distributions received from neighbors, i.e., for each $\theta \in \Theta$,

$$q_i^{(n)}(\theta) = \frac{\exp\left(\sum_{j=1}^N W_{ij} \log \Pi_Q(\mathbf{b}_j^{(n)})(\theta)\right)}{\int_\Theta \exp\left(\sum_{j=1}^N W_{ij} \log \Pi_Q(\mathbf{b}_j^{(n)})(\phi)\right) d\phi}. \tag{6.4}$$

Remark 19 (Variational Inference)**.** In above learning rule, local Bayesian update of the posterior step (6.2) can be combined with the projection onto allowed family of distributions (6.3) as follows

$$\mathbf{b}_i^{(n)} = \operatorname*{argmin}_{\pi \in Q} D\left(\pi \left\| \frac{1}{Z_i^{(n)}} \ell_i\left(\mathbf{Y}_i^{(n)} \mid \cdot, \mathbf{X}_i^{(n)}\right) b_i^{(n-1)}(\cdot)\right.\right) \tag{6.5}$$

$$= \operatorname*{argmin}_{\pi \in Q} D\left(\pi \left\| \mathbf{q}_i^{(n-1)}\right.\right) + \mathbb{E}_\pi\left[-\log \ell_i\left(\mathbf{Y}_i^{(n)} \mid \cdot, \mathbf{X}_i^{(n)}\right)\right], \tag{6.6}$$

where $Z_i^{(n)}$ is the possibly intractable normalization constant. Minimization performed in Equation (6.6) is referred to as Variational Inference (VI) and the minimand is referred to as the variational free energy [106, 107, 112, 125].

Remark 20 (Gaussian Approximate Posterior)**.** Gaussian approximate posterior can be obtained in an computationally efficient manner via VI techniques [107, 112]. More specifically, let Q denote the family of Gaussian posterior distributions with pdf given by $G(\theta, \mu, \Sigma)$. Let $(\mu_i^{(n)}, \Sigma_i^{(n)})$ denote the mean and the covariance matrix of $\mathbf{b}_i^{(n)}$ at agent at i obtained using equation (6.6). Then we can show that the posterior distribution $\mathbf{q}_i^{(n)}$ obtained after the consensus step also belongs to Q for all $i \in [N]$. Furthermore, the mean and covariance matrix $(\widetilde{\mu}_i^{(n)}, \widetilde{\Sigma}_i^{(n)})$ of $\mathbf{q}_i^{(n)}$ is given as follows

$$\widetilde{\Sigma}_i^{(n)^{-1}} = \sum_{j=1}^N W_{ij} \Sigma_j^{(n)^{-1}}, \quad \widetilde{\mu}_i^{(n)} = \widetilde{\Sigma}_i^{(n)} \sum_{j=1}^N W_{ij} \Sigma_j^{(n)^{-1}} \mu_j^{(n)}. \tag{6.7}$$

Hence, the family of Gaussian distributions not only makes the algorithm tractable, it simplifies

the consensus step by eliminating the normalization involved in equation (6.4) by reducing to updates on the mean and covariance matrix. Derivation is provided in the supplementary.

6.3 Analytic Results: Rate of Convergence

Assumption 6. *The network is a connected aperiodic graph. Specifically, W is an aperiodic and irreducible stochastic matrix.*

Assumption 7. *Let Θ be a finite set and let $\overline{\Theta}_i := \operatorname{argmin}_{\theta\in\Theta} \mathbb{E}_{\mathsf{P}_i}\left[D_{KL}\left(\mathsf{P}_{Y|X}(\cdot|X_i)||\ell_i(\cdot|\theta,X_i)\right)\right]$ and $\Theta^* := \cap_{i=1}^N \overline{\Theta}_i$. There exists a parameter $\theta^* \in \Theta$ that is globally learnable, i.e, $\cap_{i=1}^N \overline{\Theta}_i \neq \varnothing$.*

Assumption 8. *For all agents $i \in [N]$, assume:(i) The prior posterior $b_i^{(0)}(\theta) > 0$ for all $\theta \in \Theta$. (ii) There exists an $\alpha > 0, L > 0$ such that $\alpha < \ell_i(y \mid \theta, x) < L$, for all $y \in \mathcal{Y}$, $\theta \in \Theta$ and $x \in \mathcal{X}$.*

These assumptions are natural. Assumption 6 states that one can always restrict attention to the connected components of the social network where the information gathered locally by the agents can disseminated within the component. Assumption 7 ensures the combined observation of the agents across the network is statistically sufficient to learn the global model. Finally, Assumption 8 prevents the degenerate case where a zero Bayesian prior prohibits learning.

Theorem 10. *Let Θ be a finite set and let $\mathcal{Q} = \mathcal{P}(\Theta)$. Under assumptions 6, 7 and 8, using the decentralized learning algorithm described in 6.2.1 for any given confidence parameter $\delta \in (0,1)$ and any arbitrarily small $\varepsilon > 0$, we have*

$$\max_{i\in[N]} \max_{\theta\notin\Theta^*} b_i^{(n)}(\theta) < e^{-n(K(\Theta)-\varepsilon)} \tag{6.8}$$

when the number of samples satisfies $n \geq \frac{8C\log\frac{N|\Theta|}{\delta}}{\varepsilon^2(1-\lambda_{max}(W))}$, where we define the rate of convergence

of the posterior distribution as follows

$$K(\Theta) := \min_{\theta^* \in \Theta^*, \theta \in \Theta \setminus \Theta^*} \sum_{j=1}^{N} v_j I_j(\theta^*, \theta), \tag{6.9}$$

and $I_j(\theta^*, \theta) := \mathbb{E}_{\mathsf{P}_j^M}\left[D_{KL}\left(\mathsf{P}_{Y|X}(\cdot|\mathbf{X}_j)||\ell_j(\cdot \mid \theta, \mathbf{X}_j)\right) - D_{KL}\left(\mathsf{P}_{Y|X}(\cdot|\mathbf{X}_j)||\ell_j(\cdot \mid \theta^*, \mathbf{X}_j)\right)\right]$, *where eigenvector centrality* $\mathbf{v} = [v_1, v_2, \ldots, v_N]$ *is the unique stationary distribution of* W *with strictly positive components, furthermore define* $\lambda_{max}(W) := \max_{1 \leq i \leq N-1} \lambda_i(W)$, *where* $\lambda_i(W)$ *denotes i-th eigenvalue of* W *counted with algebraic multiplicity and* $\lambda_0(W) = 1$, *and* $C := \left|\log \frac{L}{\alpha}\right|$.

Proof of the theorem and additional comments on the rate of convergence are provided in the supplementary material.

Remark 21. The rate of convergence characterized by (6.9) is a function of the agent's ability to distinguish between the parameters given by the KL-divergences and structure of the weighted network which is captured by the eigenvector centrality $\mathbf{v}$ of the agents. Hence, every agent influences the rate in two ways. Firstly, if the agent has higher eigenvector centrality (i.e. the agent is centrality located), it has larger influence over the posterior distributions of other agents as a result has a greater influence over the rate of exponential decay as well. Secondly, if the agent has high KL-divergence (i.e highly informative local observations that can distinguish between parameters), then again it increases the rate. If an influential agent has highly informative observations then it boosts the rate of convergence. We will illustrate this through extensive simulations in 6.4.

6.4 Experiments

6.4.1 Decentralized Bayesian Linear Regression

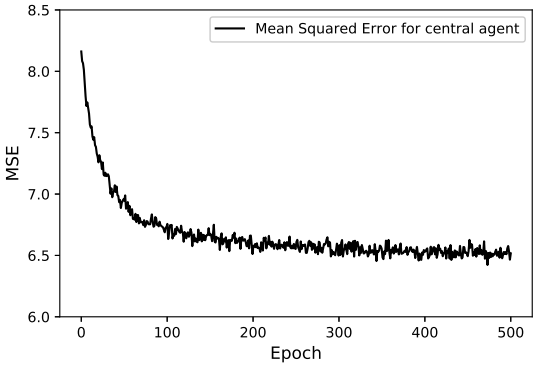
To illustrate our approach, we construct an example of Bayesian linear regression (Example 11) in the realizable setting over the network with 4 agents. We show that our proposed

social learning framework enables a fully decentralized and fast learning of a global model even when the local data is severely deficient. More specifically, we assume that each agent makes observations along only one coordinate of $\mathbf{x}$ even though the global test set consists of observations belonging to any $\mathbf{x}$ (further details of the experimental setup are provided in the supplementary). Note that this is a case of extreme non-IID data partition across the agents. 6.1c shows that the MSE of both agents, when trained using the learning rule, matches that of a central agent implying that the agents converge to the true θ^* as our theory predicts.

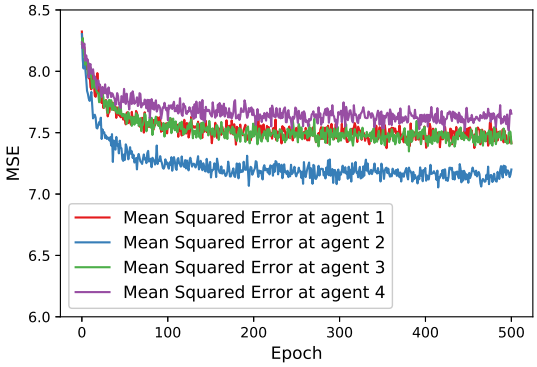
Remark 22. Note that Gaussian likelihood functions considered in Example 11 violate the bounded likelihood functions assumption. Furthermore, the parameters belong to a continuous parameter set Θ. This example and those that follow demonstrate that our analytical assumptions on the likelihood functions and the parameter set are sufficient but not necessary for convergence of our decentralized learning rule.

6.4.2 Decentralized Image Classification

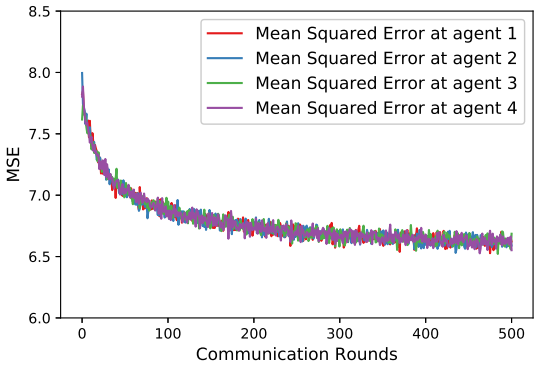
To illustrate the performance of our learning rule on real datasets we consider the problem of decentralized training of Bayesian neural networks for an image classification task on the MNIST digits dataset [126] and the Fashion-MNIST (FMNIST) dataset [127]. For all our experiments we consider a fully connected NN with the same architecture considered in the context of federated learning in [110]. Additional details regarding the implementation are provided in the supplementary. At each time step n, we sample $\theta_k \sim \mathbf{b}_i^{(n)}$ for $k \in [L]$ and for each test set image $\mathbf{x}$, we employ Monte Carlo to obtain the prediction and confidence in the prediction as $y = \text{argmax}_{y' \in \mathcal{Y}} \frac{1}{L} \sum_{k=1}^{L} \text{Softmax}(y', \mathbf{f}_{\theta_k}(\mathbf{x}))$, and $\mathsf{P}(y) = \frac{1}{L} \sum_{k=1}^{L} \text{Softmax}(y, \mathbf{f}_{\theta_k}(\mathbf{x}))$ respectively. The posterior probability $\mathsf{P}(y)$ in Bayesian Deep Learning literature [106, 128, 129], is interpreted as the *confidence* of agent i in predicting y as the true label. In our experiments, we divide the training dataset into subsets with non-overlapping label subsets. Hence, agents must learn $\mathbf{b}_i^{(n)}$ such that the resulting predictive distribution can perform well over the global dataset without



(a) Central agent



(b) Learning without cooperation



(c) Learning with cooperation

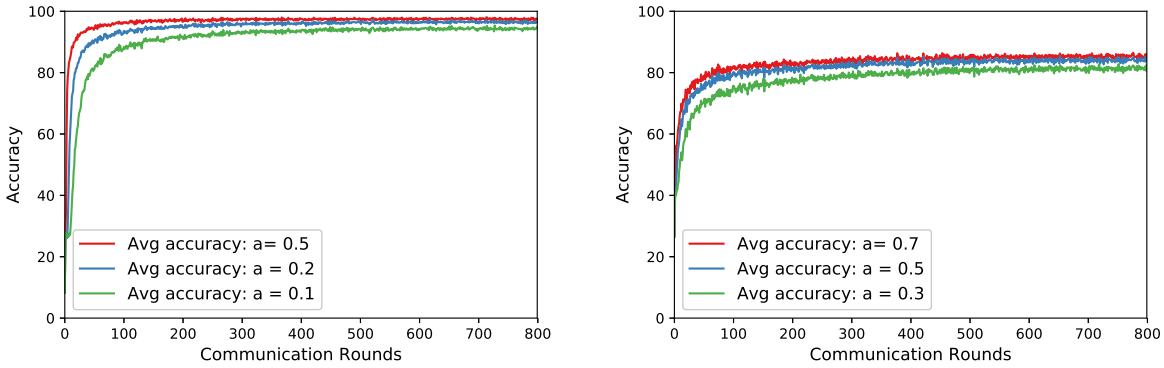
Figure 6.1: Figure compares the Mean Squared Error (MSE) of the predictions over a test dataset under three cases: (i) a benchmark scenario where all training data is shared with a central (cloud) agent, (ii) another benchmark case in which local agents, despite the severe deficiency of their observations, learn without cooperation using local training data only, and (iii) our learning paradigm where agents learn using the proposed decentralized learning rule.

sharing the local data and hence not having seen input example associated with the labels that are missing locally. In other words, our agents, at test time, will produce labels of items that they might have never encountered during the training phase. To make the distinction, we refer to a label agent i produces as an in-domain (ID) label if training data corresponding to that label is available locally otherwise they are referred to as out-of-domain (OOD) labels. We now describe our empirical studies.

Design of Social Interaction Matrix W

In this section, we investigate how the social interaction matrix W should be designed for a given network structure and a given data partition such that we maximize the rate of convergence in decentralized training. We examine this on a network with a star topology, where a central agent is connected to 8 other edge agents. Let the social interaction weights for the central agent be $\mathbf{W}_1 = [\frac{1}{9}, \ldots, \frac{1}{9}]$. For $a \in (0,1)$, we assume that an edge agent i puts a confidence $\mathbf{W}_{i1} = a$ on the central agent, $\mathbf{W}_{ii} = 1 - a$ on itself and zero on others. Note that as the confidence a which the edge agents put on the central agent increases, the eigenvector centrality of the central agent v_1 increases i.e., central agent becomes more influential over the network. For both MNIST and FMNIST, we partition the dataset such that the central agent has more informative local observations. Hence, using equation (6.9) we know that placing more confidence a on the central agent increases the rate of convergence to the true parameter and increases rate of convergence of the test dataset accuracy. This is demonstrated in 6.2a and 6.2b where both accuracy and the rate of convergence improve as a increases. In other words, rate of convergence and the average accuracy is the highest when the agent with most informative local observations has most influence on the network. Furthermore, on star topology we also demonstrate the scalability of our method through asynchronous implementathion over time-varying networks with 25 agents and 100 nodes where we achieve 96.5% and 92.3% accuracy respectively (6.6.4 in supplementary).

We focus on star topology since federated learning methods [108–110] (only) consider



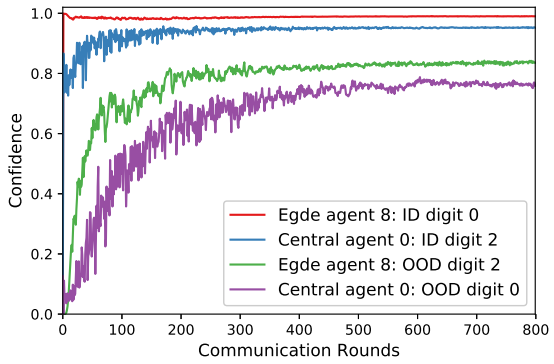
(a) On MNIST dataset

(b) On FMNIST dataset

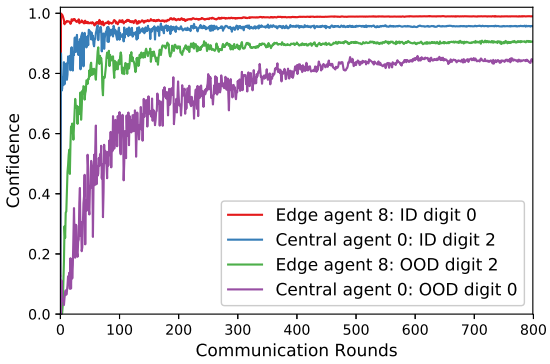
Figure 6.2: Figure shows the variation in the average accuracy over a star network topology as the eigenvector centrality of the central agent is changed.

networks with this structure. We compare the performance of our learning rule to the best reported results. On MNIST for $a = 0.5$ the average accuracy we obtain is 97.55% which is comparable to the federated learning method [110] FedAvg obtaining 98% for the same architecture and data partition. Similarly, on FMNIST for $a = 0.7$ the average accuracy we obtain is 84.21% slightly inferior to the federated learning method FedAvg [110] which obtains 87.33% accuracy in similar setting. For asynchronous time-varying networks, when we increase the number of agents in the network from 25 to 100, we again see a drop in the accuracy from 96.5% to 92.3% (6.6.4 in the supplementary). We believe the lack of periodic global synchronization results in this difference and for detailed discussion refer to Remark 25 in the supplementary. An important area of future work is to overcome this challenge.

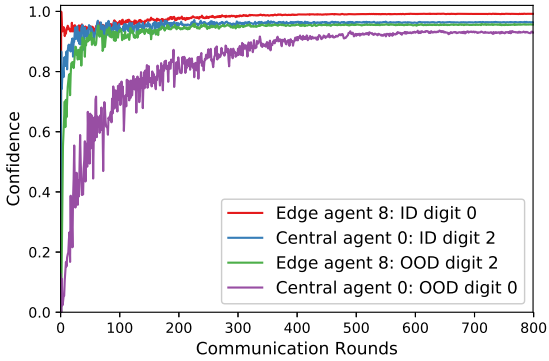
Effect on confidence over predictions: In addition to accuracy, Bayesian neural networks provide confidence estimates for each agent's predictions. Hence, we investigate the effect of network structure on confidence. 6.3 shows the confidence on digits 0 and 2 at both central and edge agents as a is varied. In all cases, we observe that both central agent and edge agents learn to predict their ID labels with higher confidence than the OOD labels. Furthermore, 6.3a, 6.3b and 6.3c show that as eigenvector centrality of central agent (most informative agent) increases,



(a) a = 0.1



(b) a = 0.2



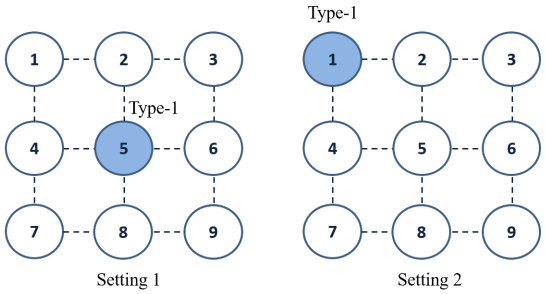
(c) a = 0.5

Figure 6.3: Figure shows the increase in the confidence on ID digit and OOD digit at the central agent and an edge agent over communication rounds. Agents are connected in a network with star topology.

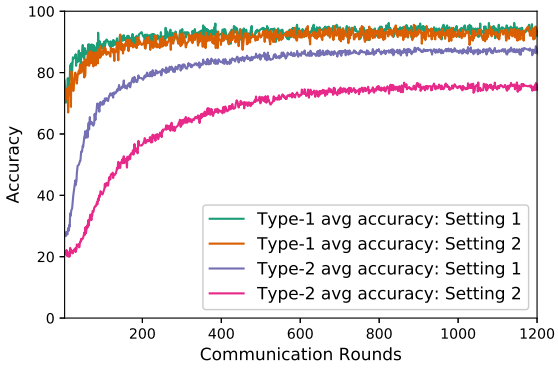
the confidence on OOD label at agent edge increase as expected.

Effect of Data Partition Over the Network

Effect of the agent placements: In this section, we investigate the appropriate placement of a locally informative agent in the network in a manner that maximizes the rate of convergence. We examine this on a 3×3 grid network obtained by connecting every agent to its adjacent agents as shown in 6.4a. The social interaction weights are defined as $W_{ij} = \frac{1}{|\mathcal{N}(i)|}$ if $j \in \mathcal{N}(i)$ and zero otherwise. In this network, the eigenvector centrality of agent i is proportional to its degree $|\mathcal{N}(i)|$; hence, more number of neighbors implies higher social influence. We divide the data such



(a) Settings of Grid Topology



(b) Average accuracy

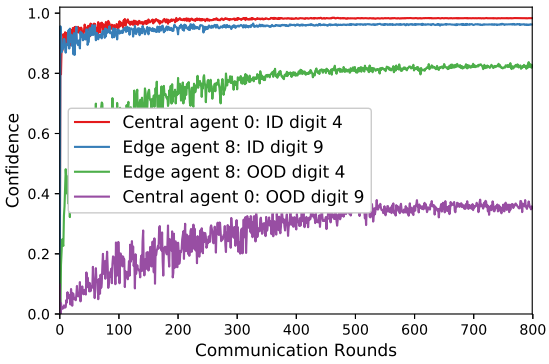
Figure 6.4: Figure shows average accuracy over 9 agents in a network with grid topology.

that the local training set for one of our agents (the Type-1 agent) is statistically informative than the local training set for all other (Type-2) agents. Now, we consider two possible placements of the Type-1 agent in the network (shown in 6.4a): (i) *Setting 1:* Type-1 agent is placed at the center (position 5) of the network and (ii) *Setting 2:* Type-1 agent is placed in a corner location (position 1) in the network. Using equation (6.9) we can predict that setting 1 has a higher rate of convergence to the true parameter and a higher rate of convergence of the test dataset accuracy compared to setting 2 which is demonstrated in 6.9a. In other words, rate of convergence is highest when the most influential agent in the network has access to an informative training dataset.

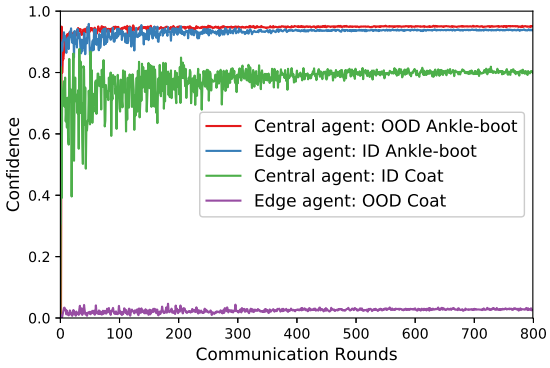
Effect of the type of data partition: Theorem 1 establishes the convergence of our learning rule under Assumption 7. Theoretical implication of this result is that all agents eventually learn the labeling function that best fits the global data if every wrong parameter labeling function

can be eliminated by some agent in the network. In the case where the agents use neural networks, local learning can only learn features discriminative to in-domain labels. Our theoretical result suggests that agents are guaranteed to converge to the correct labeling function only when every pair of OOD labels is distinguished by some agents in the network. This also suggests that some non-IID data partitioning of the labels can lead to convergence to an ambiguous set of labeling functions. This has been also shown to lead to poor accuracy empirically in the federated learning literature [130]. Unlike federated learning, our analytic Bayesian framework allows us to theoretically predict the issue.

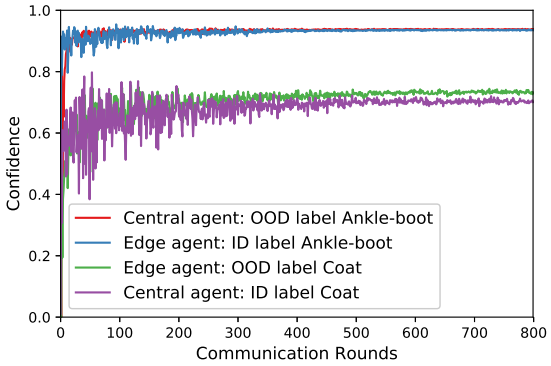
In order to understand the practical implications of Assumption 7, we construct an example where violating assumption leads to poor accuracy. Consider a star network with $a = 0.5$ where the central agent has access labels $\{0,\ldots,7\}$ and edge agents have access to labels $\{8,9\}$. Given that $\{4,9\}$ share many common features and since no agent in the network has access to both digits, our analytic results fall short to ensure learning features that can directly distinguish $\{4,9\}$. Indeed, 6.5a the confidence on OOD digit 9 at the central agent and on OOD digit 4 at an edge agent remains low. The effect of data partition described above is more pronounced in the case of FMNIST dataset. Let central agent have access to labels {`t-shirt`, `trouser`, `dress`, `coat`, `shirt`, `bag`} and edge agents has access to labels {`pullover`, `sandal`, `sneaker` `ankle-boot`}. Agents do not learn to distinguish label `pullover` at edge agents from the labels at the central agent. 6.5b shows that the confidence on OOD label `coat` at the edge agents is significantly low for this data partition and the average accuracy drops to 69.7%. Contrast this with the less ambiguous and less severe data partition of FMNIST data considered for 6.2b where all the labels with shirt-like features, are assigned to a single type, both accuracy and confidence improve as seen in 6.5c.



(a) MNIST dataset with ambiguous partitioning



(b) FMNIST dataset with ambiguous partitioning



(c) FMNIST dataset with non-ambiguous partitioning

Figure 6.5: Figures shows confidence at central and edge agents in a star network over various partition of the MNIST and FMNIST datasets.

6.5 Conclusion

In this chapter, we considered the problem of decentralized learning over a network of agents with no central server. We considered a peer-to-peer learning algorithm in which agents iterate and aggregate the beliefs of their one-hop neighbors and collaboratively estimate the global optimal parameter. We obtained high probability bounds on convergence and a full characterization of the rate of convergence across the network. We illustrated the effectiveness of algorithm for learning neural networks in computationally tractable manner while achieving high accuracies. Our experimental illustrate the predictive power of analysis of the algorithm. An important area of future work includes extensive empirical studies on various deep neural network architectures.

6.6 Appendix

6.6.1 Comments on Rate of Convergence

Remark 23 (Positivity of $K(\Theta)$)**.** We make a few comments on the quantity $K(\Theta)$. Note that in the realizable setting, for any $\theta^* \in \Theta^*$ and $\theta \in \Theta \setminus \Theta^*$ we get

$$I_j(\theta^*, \theta) = \mathbb{E}_{\mathsf{P}_j^M}\left[D_{\mathrm{KL}}\left(\ell_j(\cdot|\theta^*, \mathbf{X}_j)||\ell_j(\cdot \mid \theta, \mathbf{X}_j)\right)\right]$$

which is non-negative. The KL-divergence between the likelihood functions conditioned on the input captures the extent of distinguishability of parameter θ^* from θ. For a wrong parameter $\theta \in \Theta \setminus \Theta^*$, if $I_j(\theta^*, \theta)$ is very small then we say that the local observations at agent j are not informative enough to distinguish between θ^* and θ. Similarly for the non-realizable setting, for $\theta^* \in \Theta^*$ and $\theta \in \Theta \setminus \Theta^*$ by definition we have $\mathbb{E}_{\mathsf{P}_i^M}[D_{\mathrm{KL}}\left(\mathsf{P}_{Y|X}(\cdot|\mathbf{X}_j)||\ell_j(\cdot \mid \theta^*, \mathbf{X}_j)\right)] < \mathbb{E}_{\mathsf{P}_i^M}[D_{\mathrm{KL}}\left(\mathsf{P}_{Y|X}(\cdot|\mathbf{X}_j)||\ell_j(\cdot \mid \theta, \mathbf{X}_j)\right)]$ for all j. Hence, $K(\Theta)$ is always positive. In the social

learning literature, eigenvector centrality $\mathbf{v}$ is a measure of social influence of an agent in the network, since each v_i determines the contribution of agent i in the collective network learning rate $K(\Theta)$.

6.6.2 Consensus Step on Gaussian distributions

Let $(\mu_i^{(n)}, \Sigma_i^{(n)})$ denote the mean and the covariance matrix of $\mathbf{b}_i^{(n)}$ at agent at i obtained using equation (6.6). Using equation (6.4), we have

$$\sum_{j=1}^{N} W_{ij} \ln G(\theta, \mu_j^{(n)}, \Sigma_j^{(n)}) \tag{6.10}$$

$$= -\frac{1}{2} \sum_{j=1}^{N} W_{ij} \left((\theta - \mu_j^{(n)})^{\mathrm{T}} \Sigma_j^{(n)^{-1}} (\theta - \mu_j^{(n)}) \right) - \frac{1}{2} \sum_{i=1}^{N} W_{ij} \ln(2\pi)^k |\Sigma_j^{(n)}| \tag{6.11}$$

$$= -\frac{1}{2} \left(\theta^T \sum_{j=1}^{N} W_{ij} \Sigma_j^{(n)^{-1}} \theta + \sum_{j=1}^{N} \mu_j^{(n)^T} W_{ij} \Sigma_j^{(n)^{-1}} \mu_j^{(n)} \right) \tag{6.12}$$

$$+ \frac{1}{2} \left(\sum_{j=1}^{N} \mu_j^{(n)^T} W_{ij} \Sigma_j^{(n)^{-1}} \theta + \theta^T \sum_{j=1}^{N} W_{ij} \Sigma_j^{(n)^{-1}} \mu_i^{(n)} \right) - \frac{1}{2} \sum_{j=1}^{N} W_{ij} \ln(2\pi)^k |\Sigma_j^{(n)}|. \tag{6.13}$$

By completing the squares we obtain $\mathbf{q}_i^{(n)}$ is Gaussian distribution and we have

$$\widetilde{\Sigma}_i^{(n)^{-1}} = \sum_{j=1}^{N} W_{ij} \Sigma_j^{(n)^{-1}}, \tag{6.14}$$

and

$$\widetilde{\Sigma}^{(n)^{-1}}_{i} \widetilde{\mu}_i^{(n)} = \sum_{j=1}^{N} W_{ij} \Sigma^{(n)^{-1}}_{j} \mu_j^{(n)} \implies \widetilde{\mu}_i^{(n)} = \widetilde{\Sigma}_i^{(n)} \sum_{j=1}^{N} W_{ij} \Sigma_j^{(n)^{-1}} \mu_j^{(n)}. \tag{6.15}$$

6.6.3 Details on Bayesian Linear Regression Experiment

Let $\theta^* = [-0.3, 0.5, 0.5, 0.1, 0.2]^T$ and let noise be distributed as $\eta \sim \mathcal{N}(0, \alpha^2)$ where $\alpha = 0.8$. Agent i makes observations $(\mathbf{x}, y)$, where $\mathbf{x} = [0 \ldots, 0, x_i, 0, \ldots, 0]^T$ and x_i is sampled from

Unif$[-1,1]$ for $i=1$, Unif$[-1.5,1.5]$ for $i=2$, Unif$[-1.25,1.25]$ for $i=3$, and Unif$[-0.75,0.75]$ for $i=4$. We assume each agent starts with a Gaussian prior over θ with zero mean vector and covariance matrix given by diag$[0.5,0.5,0.5,0.5]$, where diag$(\mathbf{x})$ denotes a diagonal matrix with diagonal elements given by vector $\mathbf{x}$. The social interaction weights are given as $\mathbf{W}_1 = [0.5,0.5,0,0]$, $\mathbf{W}_2 = [0.3,0.1,0.3,0.3]$, $\mathbf{W}_3 = [0,0.5,0.5,0]$ and $\mathbf{W}_4 = [0,0.5,0,0.5]$. We assume each agent starts with a Gaussian prior over Θ and hence the posterior distribution after a Bayesian update remains Gaussian. This implies $\mathcal{Q}$ remains fixed as the family of Gaussian distributions and the consensus step reduces to equation (6.7).

6.6.4 Details on Bayesian Deep Learning Experiments on Image Classification

We consider two datasets: (i) the MNIST digits dataset [126] where each image is assigned a label in $\mathcal{Y} = [0,\ldots,9]$ and (ii) the Fashion-MNIST (FMNIST) dataset [127] where each image is assigned a label in $\mathcal{Y} = [$ `t-shirt`, `trouser`, `pullover`, `dress`, `coat`, `sandal`, `shirt`, `sneaker`, `bag`, `ankle-boot`$]$. Both datasets consist of 60,000 training images and 10,000 testing images of size 28 by 28. For all our experiments we consider a fully connected NN with 2-hidden layers with 200 units each using ReLU activations which is same as the architecture considered in the context of federated learning in [110].

For all the experiments we choose $\mathcal{Q}$ to be the family of Gaussian mean-field approximate posterior distributions with pdf given by $G(\theta,\mu,\Sigma)$, where Σ is a strictly diagonal matrix [111,112]. As discussed in 19 this corresponds to performing variational inference to obtain a Gaussian approximation of the local posterior distribution, i.e., minimizing the variational free energy given in equation (6.6) over $\mathcal{Q}$. While we compute the KL divergence in (6.6) in a closed form, we employ simple Monte Carlo to compute the gradients using Bayes by Backprop [107, 112].

Remark 24 (Prediction on Test Dataset)**.** In the absence of cooperation among the agents,

each agent i using the Bayes rule only learns the local posterior distribution $\mathsf{P}(\theta|\mathbf{X}_i^n, \mathbf{Y}_i^n)$ and makes predictions on the test dataset input $\mathbf{x}$ using the predictive distribution $\mathsf{P}(y|\mathbf{x}) = \int_\Theta \ell_i(y \mid \theta, \mathbf{x})\mathsf{P}(\theta|\mathbf{X}_i^n, \mathbf{Y}_i^n)d\theta$. However at any time step n, using the decentralized learning rule each agent i learns a posterior distribution $\mathbf{b}_i^{(n)}$ and makes predictions on the test dataset input $\mathbf{x}$ using a predictive distribution $\int_\Theta \ell_i(y \mid \theta, \mathbf{x})b_i^{(n)}(\theta)d\theta$. Applying 10 we see that as the local posterior $\mathbf{b}_i^{(n)}$ converges to θ^* for each agent i, it can locally predict as if was trained on global dataset.

Remark 25. Federated learning paradigm, unlike our fully decentralized setup, requires a centralized controller to aggregate the local models from each agent. Furthermore, after each round of communication with the central controller, every agent before training initializes its local model with the global model obtained from the central controller. The periodic shared initialization using a global model across the network, while it is a stringent constraint, is required to prevent the averaging performed at the central controller from producing an arbitrarily bad model [110]. Without modelling the correlation between the weights and bias of the agents across the network, different random initialization at each agent can lead to different local minima and result in diverging local models at the agents [131]. However, modelling of correlation between the weights and bias of the agents across the network is computationally prohibitive. We overcome this challenge by using shared initialization when the local models are trained for the first time at each agent, however we do not perform this after each communication round. Our method overcomes the need for shared initialization after each communication round by incorporating the global information (on the weights and bias across the agents) in the local training by using the prior $\mathbf{q}_i^{(n)}$, obtained locally via the consensus step (6.4) at each agent i, in the minimization of variational free energy (6.6). It would be interesting to investigate other shared initialization suitable for decentralized training which addresses the gap in the performance.

Design of Social Interaction Matrix W

For experiments in 6.4.2, we use a network with a star topology, where there is one central agent and 8 edge agents. We vary confidence a which the edge agents put on the central agent over $[0.1, 0.2, 0.3, 0.5, 0.7]$, the eigenvector centrality of the central agent v_1 increases as $[0.1, 0.18, 0.25, 0.36, 0.44]$. We partition the MNIST dataset into two subsets so that the central agent dataset has all images of labels $[2, \ldots, 9]$ and edge agents has all images of labels $[0, 1]$. To ensure all the edge agents has equal number of images, we shuffle the images with labels $[0, 1]$ and partition them into 8 non-overlapping subsets. We call this partition `MNIST-Setup1`.

Similarly, for Fashion-MNIST (FMNIST) dataset, we first partition into two subsets so that central agent has access to labels [`t-shirt`, `pullover`, `dress`, `coat`, `shirt`, `bag`] and edge agents have access to labels [`trouser`, `sandal`, `sneaker`, `ankle-boot`]. We shuffle the images with labels [`trouser`, `sandal`, `sneaker`, `ankle-boot`] and partition them into 8 non-overlapping subsets. We call this partition `FMNIST-Setup1`.

We ensure that all agents has same number of local updates u per communication round, which is equal to $(\lfloor n_{edge}/B \rfloor)E$. For the central agent, this means that for each local epoch, the central agent is trained on a random subset of its local dataset, whereas the edge agents use all the local dataset. For all agents, we use Adam optimizer [132] with initial learning rate of 0.001 and learning rate decay of 0.99 per communication round.

Table 6.1: Settings for Star Topology Network Experiment: E is number of local epochs, B is the local minibatch size, u is the number of local updates per communication round, η is the initial learning rate for all agents, ε is the learning rate decay rate, n_{center} is the dataset size of the central agent, n_{edge} is the dataset size of each of the edge agent.

Experiment	E	B	u	η	ε	n_{center}	n_{edge}	comm rounds
`MNIST-Setup1`	5	50	155	0.001	0.99	47335	1583	800
`MNIST-Setup2`	5	50	145	0.001	0.99	48200	1475	800
`MNIST-Setup3`	5	50	145	0.001	0.99	48209	1473	800
`FMNIST-Setup1`	5	100	150	0.001	0.99	36000	3000	800
`FMNIST-Setup2`	5	100	150	0.001	0.99	36000	3000	800

Effect of Data Partition over the Network

Effect of the agent placements: We use a 3 by 3 grid network illustrated by 6.4a in 6.4.2. We assign MNIST images with labels $[2,\ldots,9]$ to an agent of Type-1 and divide images with labels $[0,1]$ among 8 agents of Type-2. In `Center` setting, we place Type-1 agent at the central location. In `Corner` setting, we place Type-1 agent in a corner location. Similar to 6.6.4, We ensure that all agents has same number of local updates u per communication round, which is equal to $(\lfloor n_{Type2}/B \rfloor)E$. Again, we use Adam optimizer for all agents.

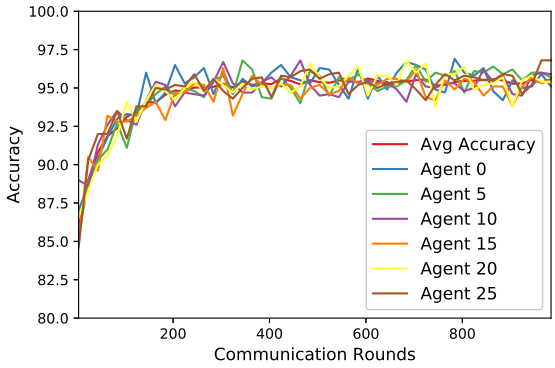
Table 6.2: Settings for Grid Topology Network Experiment: E is number of local epochs, B is the local minibatch size, u is the number of local updates per communication round, η is the initial learning rate for all agents, ε is the learning rate decay rate, n_{Type1} is the dataset size of the Type-1 agent, n_{Type2} is the dataset size of each of the Type-2 agent.

Experiment	E	B	u	η	ε	n_{Type1}	n_{Type2}	comm rounds
`Corner`	5	50	155	0.001	0.99	47335	1583	1200
`Center`	5	50	155	0.001	0.99	47335	1583	1200

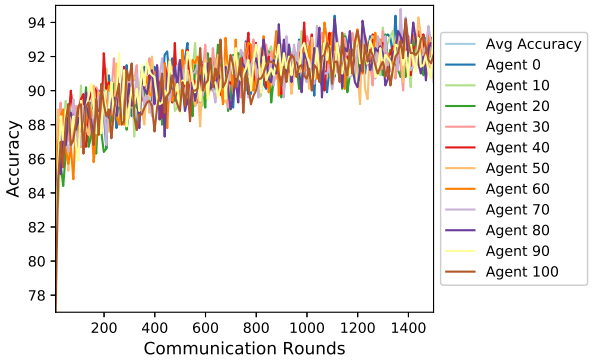
Effect of the type data partition: In ablation study, we again use a star network and consider two other ways of partitioning the MNIST dataset: (1) the central agent dataset has all images of labels $[0,\ldots,7]$ and edge agents has all images of labels $[8,9]$, we call this `MNIST-Setup2`, and (2) the edge agents has all images of labels [4,9] and the central agent other labels, we call this `MNIST-Setup3`. For FMNIST dataset, central agent has access to images with labels [`t-shirt`, `trouser`, `dress`, `coat`, `shirt`, `bag`] and edge agents have access to images with labels [`pullover`, `sandal`, `sneaker` `ankle-boot`], we call this `FMNIST-Setup2`.

Asynchronous Decentralized Learning on Time-varying Networks Experiment

Now we implement our learning rule on time-varying networks which model practical peer-to-peer networks where synchronous updates are not easy or very costly to implement.



(a) Average accuracy over all 26 nodes.



(b) Accuracy over all 100 nodes.

Figure 6.6: Figure shows the accuracies of agents in a time-varying network.

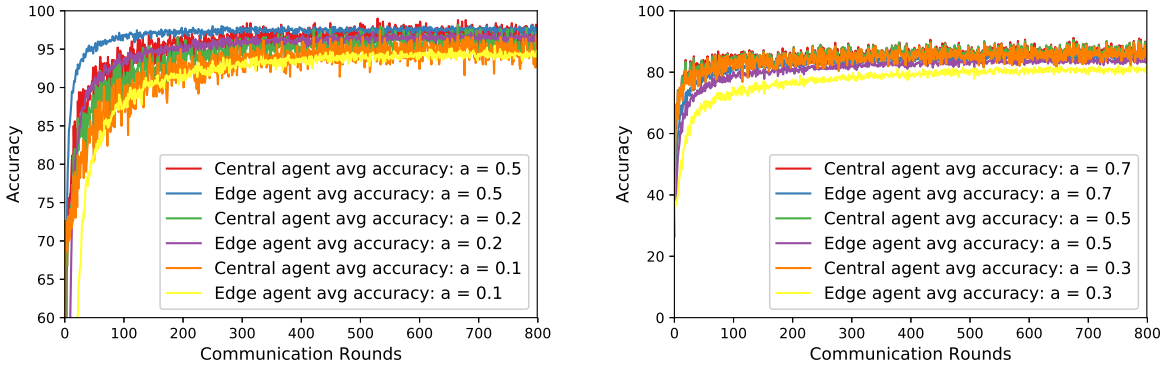
We consider a time-varying network of $N+1$ agents numbered as $\{0,1,\ldots,N\}$. At any give time, only N_0 agents are connected to agent 0 in a star topology. For $k \in [\frac{N}{N_0}]$, let $\mathcal{G}_k$ denote a graph with a star topology where the central agent 0 is connected to edge agents whose indices belong to $\{N_0(k-1)+1,\ldots,N_0k\}$. This implies at any given time only a small fraction of agents $\frac{N_0}{N}$ are training over their local data. Note that $\cup_{k=1}^{\frac{N}{N_0}} \mathcal{G}_k$ is strongly connected network over all $N+1$ agents. The social interaction weights for the central agent are $\mathbf{W}_0 = [\frac{1}{N_0+1},\ldots,\frac{1}{N_0+1}]$. Let $a = 0.5$. An edge agent $i \in \mathcal{G}_k$ puts a confidence $\mathbf{W}_{i0} = a$ on the central agent 0, $\mathbf{W}_{ii} = 1-a$ on itself and zero on others. The MNIST dataset is divided in an i.i.d manner, i.e., data is shuffled and each agent is randomly assigned approximately $(\frac{60{,}000}{N+1})$ samples. For $N = 25, N_0 = 5$, we obtain an average accuracy of 95.6% over all agents and 95.1% accuracy at the central agent

and for $N = 100, N_0 = 10$, we obtain an average accuracy of 92.3% over all agents and 93.1% accuracy at the central agent. This also demonstrates that decentralized learning can be achieved with as few as 600 samples locally.

Table 6.3: Settings for Time-varying Network Experiment: E is number of local epochs, B is the local minibatch size, u is the number of local updates per communication round, η is the initial learning rate for all agents, ε is the learning rate decay rate, n is the dataset size of any agent. Since all agents have same number of samples, they automatically have equal number of local updates per communication round. Adam optimizer is used for all agents.

Experiment	E	B	u	η	ε	n	comm rounds
`N = 25`	1	50	47	0.001	0.99	2307	1000
`N = 100`	2	10	120	0.001	0.998	594	1000

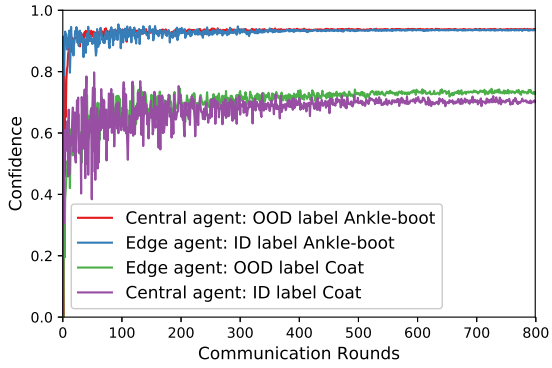
6.6.5 Additional Figures



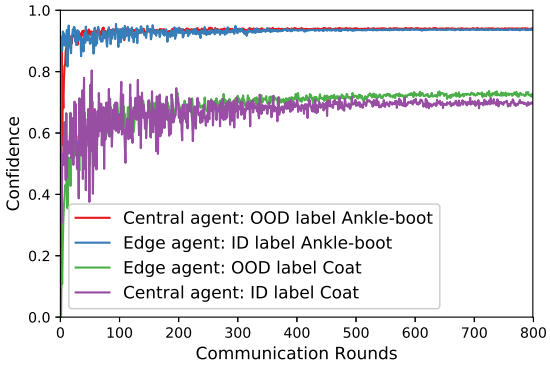
(a) On MNIST dataset

(b) On Fashion-MNIST dataset

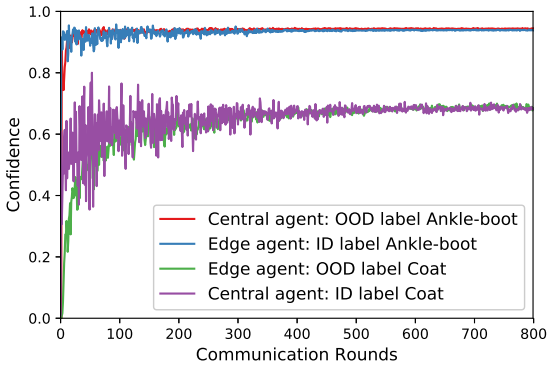
Figure 6.7: Figures shows average accuracy of 9 agents connected in a network with star topology.



(a) a = 0.7

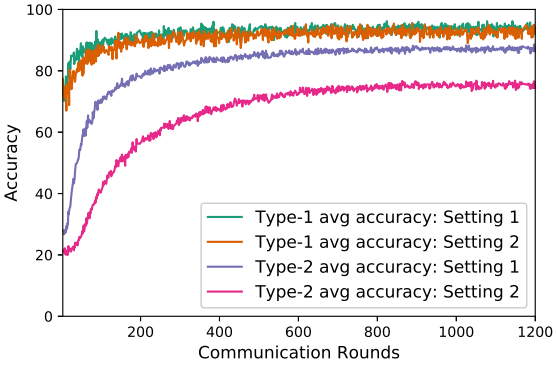


(b) a = 0.5

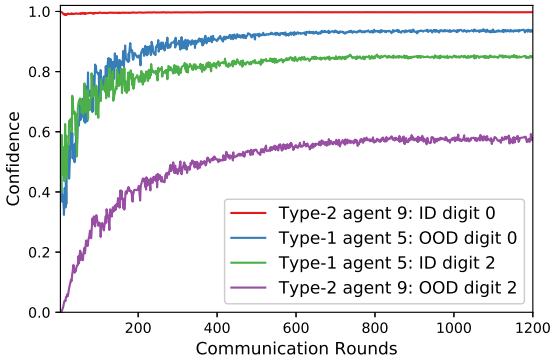


(c) a = 0.3

Figure 6.8: Figures shows the increase in the confidence on an ID label and OOD label at the central and edge agents over communication rounds for FMNIST dataset. Agents are connected in a network with star topology and the value of a varies over $[0.7, 0.5, 0.3]$.



(a) Average accuracy



(b) Confidence values in setting 1

Figure 6.9: Figure shows average accuracy over 9 agents in a network with grid topology.

6.6.6 Proof of Theorem 1

The proof of Theorem 1 is based the proof provided in [103–105]. For the ease of exposition, let $b_i^{(0)}(\theta) = \frac{1}{|\Theta|}$ for all $\theta \in \Theta$. Fix a $\theta^* \in \Theta^*$. We begin with the following recursion for each node $i \in [N]$ and for any $\theta \notin \Theta^*$,

$$\frac{1}{n}\log\frac{b_i^{(n)}(\theta^*)}{b_i^{(n)}(\theta)} = \frac{1}{n}\sum_{j=1}^{N}\sum_{k=1}^{n} W_{ij}^k z_j^{(n-k+1)}(\theta^*,\theta), \tag{6.16}$$

where

$$z_j^{(k)}(\theta^*,\theta) = \log\frac{l_j\left(X_j^{(k)} \mid \theta^*, X_i^{(k)}\right)}{l_j\left(X_j^{(k)} \mid \theta, X_i^{(k)}\right)}. \tag{6.17}$$

From the above recursion we have

$$\frac{1}{n}\log\frac{b_i^{(n)}(\theta^*)}{b_i^{(n)}(\theta)} \geq \frac{1}{n}\sum_{j=1}^{N} v_j\left(\sum_{k=1}^{n} z_j^{(k)}(\theta^*,\theta)\right) - \frac{1}{n}\sum_{j=1}^{N}\sum_{k=1}^{n}\left|W_{ij}^k - v_j\right|\left|z_j^{(k)}(\theta^*,\theta)\right| \tag{6.18}$$

$$\overset{(a)}{\geq} \frac{1}{n}\sum_{j=1}^{N} v_j\left(\sum_{k=1}^{n} z_j^{(k)}(\theta^*,\theta)\right) - \frac{4C\log N}{n(1-\lambda_{\max}(W))}, \tag{6.19}$$

where (a) follows from Lemma 18 and the boundedness assumption of log-likelihood ratios. Now fix $n \geq \frac{8C\log N}{\varepsilon(1-\lambda_{\max}(W))}$, since $b_i^{(n)}(\theta^*) \leq 1$ we have

$$-\frac{1}{n}\log b_i^{(n)}(\theta) \geq -\frac{\varepsilon}{2} + \frac{1}{n}\sum_{j=1}^{N} v_j\left(\sum_{k=1}^{n} z_j^{(k)}(\theta^*,\theta)\right).$$

Furthermore, we have

$$\mathsf{P}\left(-\frac{1}{n}\log b_i^{(n)}(\theta) \leq \sum_{j=1}^{N} v_j I_j(\theta^*,\theta) - \varepsilon\right) \leq \mathsf{P}\left(\frac{1}{n}\sum_{j=1}^{N} v_j \sum_{k=1}^{n} z_j^{(k)}(\theta^*,\theta) \leq \sum_{j=1}^{N} v_j I_j(\theta^*,\theta) - \frac{\varepsilon}{2}\right).$$

Now for any $j \in [N]$ note that

$$\sum_{j=1}^{N} v_j \sum_{k=1}^{n} z_j^{(k)}(\theta^*,\theta) - n\sum_{j=1}^{N} v_j I_j(\theta^*,\theta) = \sum_{k=1}^{n}\left(\sum_{j=1}^{N} v_j z_j^{(k)}(\theta^*,\theta) - \sum_{j=1}^{N} v_j \mathbb{E}[z_j^{(k)}(\theta^*,\theta)]\right).$$

For any $\theta \notin \Theta^*$, applying McDiarmid's inequality for all $\varepsilon > 0$ and for all $n \geq 1$ we have

$$\mathsf{P}\left(\sum_{k=1}^{n}\left(\sum_{j=1}^{N} v_j z_j^{(k)}(\theta^*,\theta) - \sum_{j=1}^{N} v_j \mathbb{E}[z_j^{(k)}(\theta^*,\theta)]\right) \leq -\frac{\varepsilon n}{2}\right) \leq e^{-\frac{\varepsilon^2 n}{2C}}.$$

Hence, for all $\theta \notin \Theta^*$, for $n \geq \frac{8C\log N}{\varepsilon(1-\lambda_{\max}(W))}$ we have

$$\mathsf{P}\left(\frac{-1}{n}\log b_i^{(n)}(\theta) \leq \sum_{j=1}^{N} v_j I_j(\theta^*,\theta) - \varepsilon\right) \leq e^{-\frac{\varepsilon^2 n}{4C}}, \tag{6.20}$$

which implies

$$\mathsf{P}\left(b_i^{(n)}(\theta) \geq e^{-n(\sum_{j=1}^{N} v_j I_j(\theta^*,\theta)-\varepsilon)}\right) \leq e^{-\frac{\varepsilon^2 n}{4C}}. \tag{6.21}$$

Using this we obtain a bound on the worst case error over all θ and across the entire network as follows

$$\mathsf{P}\left(\max_{i\in[N]}\max_{\theta\notin\Theta^*} b_i^{(n)}(\theta) \geq e^{-n(K(\Theta)-\varepsilon)}\right) \leq N|\Theta|e^{-\frac{\varepsilon^2 n}{4C}}, \tag{6.22}$$

where $K(\Theta) := \min_{\theta\in\Theta^*,\psi\in\Theta\setminus\Theta^*}\sum_{j=1}^{N} v_j I_j(\theta,\psi)$. From Assumption 7 and Lemma 18 we have that $K(\Theta) > 0$. Then, with probability $1-\delta$ we have

$$\max_{i\in[N]}\max_{\theta\notin\Theta^*} b_i^{(n)}(\theta) < e^{-n(K(\Theta)-\varepsilon)}, \tag{6.23}$$

when the number of samples satisfies

$$n \geq \frac{8C\log\frac{N|\Theta|}{\delta}}{\varepsilon^2(1-\lambda_{\max}(W))}. \tag{6.24}$$

Lemma 18 ([104]). *For an irreducible and aperiodic stochastic matrix W, the stationary distribution $\mathbf{v} = [v_1, v_2, \ldots, v_N]$ is unique and has strictly positive components and satisfies $v_i = \sum_{j=1}^{n} v_j W_{ji}$. Furthermore, for any $i \in [N]$ the weight matrix satisfies*

$$\sum_{k=1}^{n} \sum_{j=1}^{N} \left| W_{ij}^{k} - v_j \right| \leq \frac{4 \log N}{1 - \lambda_{max}(W)},$$

where $\lambda_{max}(W) = \max_{i \in [N-1]} \lambda_i(W)$, and $\lambda_i(W)$ denotes eigenvalue of W counted with algebraic multiplicity and $\lambda_0(W) = 1$.

Chapter 6, in full, has been submitted for publication as: Anusha Lalitha, Xinghan Wang, Cihan Kilinc, Yongxi Lu, Tara, Javidi, and Farinaz Koushanfar, "Decentralized Bayesian Learning over Graphs", available on *arXiv preprint arXiv:1905.10466*. The dissertation author was the primary investigator and author of this paper.

Bibliography

[1] M. Horstein. Sequential transmission using noiseless feedback. *IEEE Transactions of Information Theory*, 9(3):136–143, Jul. 1963.

[2] O. Shayevitz and M. Feder. Optimal feedback communication via posterior matching. *IEEE Transactions of Information Theory*, 57(3):1186–1222, Mar. 2011.

[3] Rolf Waeber, Peter I Frazier, and Shane G Henderson. Bisection search with noisy responses. *SIAM Journal on Control and Optimization*, 51(3):2261–2279, May 2013.

[4] Marat Valievich Burnashev and Kamil'Shamil'evich Zigangirov. An interval estimation problem for controlled observations. *Problemy Peredachi Informatsii*, 10(3):51–61, 1974.

[5] M. V. Burnashev. Data transmission over a discrete channel with feedback. Random transmission time. *Problemy Peredachi Informatsii*, 12(4):10–30, 1976.

[6] Cheuk Ting Li and Abbas El Gamal. An efficient feedback coding scheme with low error probability for discrete memoryless channels. *IEEE Transactions of Information Theory*, 61(6):2953–2963, Jun. 2015.

[7] Mohammad Naghshvar, Tara Javidi, and Michele Wigger. Extrinsic Jensen–Shannon divergence: Applications to variable-length coding. *IEEE Transactions of Information Theory*, 61(4):2148–2164, Apr. 2015.

[8] Sung-En Chiu and T. Javidi. Sequential measurement-dependent noisy search. In *Proceedings of IEEE Information Theory Workshop (ITW)*, pages 221–225, Sep. 2016.

[9] R. G. Gallager. *Information Theory and Reliable Communication*. John Wiley & Sons, Inc., 1968.

[10] S. Chiu and T. Javidi. Sequential measurement-dependent noisy search. In *IEEE Information Theory Workshop (ITW)*, Sept. 2016.

[11] Yonatan Kaspi, Ofer Shayevitz, and Tara Javidi. Searching with measurement dependent noise. *IEEE Transactions on Information Theory*, 64(4):2690–2705, April 2018.

[12] Bruno Jedynak, Peter I. Frazier, and Raphael Sznitman. Twenty questions with noise: Bayes optimal policies for entropy loss. *J. Appl. Probab.*, 49(1):114–136, 03 2012.

[13] T. Tsiligkaridis, B. M. Sadler, and A. O. Hero. Collaborative 20 questions for target localization. *IEEE Transactions on Information Theory*, 60(4):2233–2252, April 2014.

[14] A. Lalitha, N. Ronquillo, and T. Javidi. Measurement dependent noisy search: The gaussian case. In *Proceedings of the 2017 IEEE International Symposium on Information Theory (ISIT)*, pages 3090–3094, June 2017.

[15] M. Giordani, M. Mezzavilla, C. N. Barati, S. Rangan, and M. Zorzi. Comparative analysis of initial access techniques in 5g mmwave cellular networks. In *Proceedings of the 2016 Annual Conference on Information Science and Systems (CISS)*, pages 268–273, March 2016.

[16] V. Va and R. W. Heath. Performance analysis of beam sweeping in millimeter wave assuming noise and imperfect antenna patterns. In *IEEE 84th Vehicular Technology Conference*, volume 63, September 2016.

[17] E. Axell, G. Leus, E. G. Larsson, and H. V. Poor. Spectrum sensing for cognitive radio : State-of-the-art and recent advances. *IEEE Signal Processing Magazine*, 29(3):1053–5888, 05 2012.

[18] A. Sharma and C.R. Murthy. Group testing-based spectrum hole search for cognitive radios. *IEEE Transactions on Vehicular Technology*, 63(8):3794–3805, October 2014.

[19] J. Treichler, M. Davenport, and R. Baraniuk. Application of compressive sensing to the design of wideband signal acquisition receivers. In *Proceedings of the 6th US/Australia Joint Work. Defense Applications of Signal Processing (DASP)*, pages 1–10, September 2009.

[20] C. N. Barati, S. A. Hosseini, M. Mezzavilla, P. Amiri-Eliasi, S. Rangan, T. Korakis, S. S. Panwar, and M. Zorzi. Directional initial access for millimeter wave cellular systems. In *Proceedings of the 49th Asilomar Conference on Signals, Systems and Computers*, pages 307–311, Nov 2015.

[21] O. Abari, H. Hassanieh, and D. Katabi. Millimeter wave communications: From point-to-point links to agile network connections. In *Proceedings of the 15th ACM Workshop on Hot Topics in Networks*, pages 169–175, November 2016.

[22] Z. Quan, S.Cui, A. H. Sayed, and H. V. Poor. Optimal Multiband Joint Detection for Spectrum Sensing in Cognitive Radio Networks. *IEEE Transactions Signal Process*, 57(3):1128–1140, March 2009.

[23] Y. Feng and X. Wang. Adaptive Multiband Spectrum Sensing. *IEEE Wireless Communications Letters*, 1(2):121–124, April 2012.

[24] A. Tajer, R. Castro, and X. Wang. Adaptive Spectrum Sensing for Agile Cognitive Radios. In *Proceedings of the 2010 IEEE Conference on Acoustics Speech and Signal Processing (ICASSP)*, March 2010.

[25] M. L. Malloy and R. D. Nowak. Near-optimal adaptive compressed sensing. *IEEE Transactions on Information Theory*, 60(7):4001–4012, May 2014.

[26] Y. Jin, Y. Kim, and B. Rao. Limits on support recovery of sparse signals via multiple-access communication techniques. *IEEE Transactions on Information Theory*, 57(12):7877–7892, December 2011.

[27] T. M. Cover and J. A. Thomas. *Elements of information theory (2nd ed)*. John Wiley & Sons, Inc., New York, NY, USA, 2006.

[28] N. Ronquillo and T. Javidi. Multi-band Noisy Spectrum Sensing with Codebooks. In *Proceedings of the 50th Asilomar Conference on Signals, Systems, and Computers*, pages 1687–1691, November 2016.

[29] N. Bshouty. Optimal algorithms for the coin weighing problem with a spring scale. In *Proceedings of the 22nd Conference on Learning Theory (COLT)*, pages 1–10, June 2009.

[30] S. C. Chang and E. Weldon. Coding for t-user multiple-access channels. *IEEE Transactions on Information Theory*, 25(6):684–691, November 1979.

[31] M. Naghshvar and T. Javidi. Optimal reliability over a DMC with feedback via deterministic sequential coding. In *Proceedings of the 2012 IEEE International Symposium on Information Theory and its Applications (ISITA)*, pages 51–55, Oct 2012.

[32] M. Naghshvar and T. Javidi. Extrinsic Jensen–Shannon Divergence: Applications to Variable-Length Coding. In *IEEE Transactions on Information Theory*, pages 2191–2195, July 2012.

[33] Mohammad Naghshvar. *Active Learning and Hypothesis Testing*. PhD thesis, University of California San Diego, 2013.

[34] A. Wald and J. Wolfowitz. Optimum character of the sequential probability ratio tests. *The Annals of Mathematical Statistics*, 19(3):326–339, 1948.

[35] E. A. Haroutunian. A lower bound on the probability of error for channels with feedback. 13(2):36–44, 1977.

[36] Robert G. Gallager. *Information Theory and Reliable Communication*. John Wiley & Sons, Inc., New York, NY, USA, 1968.

[37] R.E. Blahut. Hypothesis testing and information theory. *IEEE Transactions on Information Theory*, 20(4):405–417, Jul 1974.

[38] E. Tuncel. Extensions of error exponent analysis in hypothesis testing. In *Proceedings of International Symposium on Information Theory (ISIT)*, pages 835–839, Sept 2005.

[39] H. Chernoff. Sequential design of experiments. *The Annals of Mathematical Statistics*, 30:755–770, 1959.

[40] Y. Altuğ, H. V. Poor, and S. Verdú. Variable-length channel codes with probabilistic delay guarantees. In *Proceedings of 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 642–649, Sep. 2015.

[41] G. Forney. Exponential error bounds for erasure, list, and decision feedback schemes. *IEEE Transactions on Information Theory*, 14(2):206–220, Mar 1968.

[42] I. E. Telatar and R. G. Gallager. New exponential upper bounds to error and erasure probabilities. In *Proceedings of IEEE International Symposium on Information Theory (ISIT)*, pages 379–, Jun 1994.

[43] B. Nakiboglu and L. Zheng. Errors-and-erasures decoding for block codes with feedback. *IEEE Transactions on Information Theory*, 58(1):24–49, Jan 2012.

[44] N. Grigoryan, A. Harutyunyan, S. Voloshynovskiy, and O. Koval. On multiple hypothesis testing with rejection option. In *Proceedings of IEEE Information Theory Workshop (ITW)*, pages 75–79, Oct 2011.

[45] I. Sason. Moderate deviations analysis of binary hypothesis testing. In *Proceedings of IEEE International Symposium on Information Theory (ISIT)*, pages 821–825, July 2012.

[46] B. Nakiboğlu and L. Zheng. Upper bounds to error probability with feedback. In *Proceedings of 47th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 865–871, Sep. 2009.

[47] H. Yamamoto and K. Itoh. Asymptotic performance of a modified schalkwijk-barron scheme for channels with noiseless feedback (corresp.). *IEEE Transactions on Information Theory*, 25(6):729–733, Nov 1979.

[48] Y. Polyanskiy, H. V. Poor, and S. Verdu. Feedback in the non-asymptotic regime. *IEEE Transactions on Information Theory*, 57(8):4903–4925, Aug 2011.

[49] P. K. Gopala, Y. H. Nam, and H. El Gamal. On the error exponents of arq channels with deadlines. *IEEE Transactions on Information Theory*, 53(11):4265–4273, Nov 2007.

[50] I. Csiszar and P. C. Shields. Information theory and statistics: a tutorial. *Foundations and Trends in Communications and Information Theory*, 1(4):417–528, December 2004.

[51] Y. Polyanskiy and S. Verdu. Binary hypothesis testing with feedback. In *Information Theory and Applications Workshop (ITA),*, 2011.

[52] M. Gutman. Asymptotically optimal classification for multiple tests with empirically observed statistics. *IEEE Transactions on Information Theory*, 35(2):401–408, March 1989.

[53] Arkadii Georgievich D'yachkov. Upper bounds on the error probability for discrete memoryless channels with feedback. *Problemy Peredachi Informatsii*, 11(4):13–28, 1975.

[54] A. Sahai and S. K. Mitter. The necessity and sufficiency of anytime capacity for stabilization of a linear system over a noisy communication link—part I: Scalar systems. *IEEE Transactions of Information Theory*, 52(8):3369–3395, Aug. 2006.

[55] R. T. Sukhavasi and B. Hassibi. Linear time-invariant anytime codes for control over noisy channels. *IEEE Transactions of Automatic Control*, 61(12):3826–3841, Dec. 2016.

[56] A. Khina, W. Halbawi, and B. Hassibi. (Almost) practical tree codes. In *Proceedings of IEEE Intenational Symposuim in Information Theory (ISIT)*, pages 2404–2408, Barcelona, Spain, Jul. 2016.

[57] Tunc Simsek, Rahul Jain, and Pravin Varaiya. Scalar estimation and control with noisy binary observations. *IEEE Transactions on Automatic Control*, 49(9):1598–1603, 2004.

[58] R. Ahlswede and I. Csiszar. Hypothesis testing with communication constraints. *IEEE Transactions on Information Theory*, 32(4):533–542, July 1986.

[59] Te Han. Hypothesis testing with multiterminal data compression. *IEEE Transactions on Information Theory*, 33(6):759–772, November 1987.

[60] M. Longo, T. D. Lookabaugth, and R. M. Gray. Quantization for decentralized hypothesis testing under communication constraints. *IEEE Transactions on Information Theory*, 36(2):241–255, March 1990.

[61] V. V. Veeravalli, T. Basar, and H. V. Poor. Decemberentralized sequential detection with a fusion center performing the sequential test. *IEEE Transactions on Information Theory*, 39(2):433–442, March 1993.

[62] H. Shimokawa, Te Sun Han, and S. Amari. Error bound of hypothesis testing with data compression. In *Proceedings of the 1994 IEEE International Symposium on Information Theory*, pages 114–119, June 1994.

[63] Te Sun Han and S. Amari. Statistical inference under multiterminal data compression. *IEEE Transactions on Information Theory*, 44(6):2300–2324, October 1998.

[64] Y. Xiang and Y. H. Kim. Interactive hypothesis testing against independence. In *Proceedings of the 2013 IEEE International Symposium on Information Theory Proceedings*, pages 2840–2844, July 2013.

[65] Y. Mei. Asymptotic optimality theory for decemberentralized sequential hypothesis testing in sensor networks. *IEEE Transactions on Information Theory*, 54(5):2072–2089, May 2008.

[66] M. S. Rahman and A. B. Wagner. On the optimality of binning for distributed hypothesis testing. *IEEE Transactions on Information Theory*, 58(10):6282–6303, October 2012.

[67] Biao Chen, Ruixiang Jiang, T. Kasetkasem, and P. K. Varshney. Channel aware decemberision fusion in wireless sensor networks. *IEEE Transactions on Signal Processing*, 52(12):3454–3458, December 2004.

[68] Biao Chen and P. K. Willett. On the optimality of the likelihood-ratio test for local sensor decemberision rules in the presence of nonideal channels. *IEEE Transactions on Information Theory*, 51(2):693–699, February 2005.

[69] A. Anandkumar and L. Tong. Distributed statistical inference using type based random access over multi-access fading channels. In *Proceedings of the 40th Annual Conference on Information Sciences and Systems*, pages 38–43, March 2006.

[70] J. B. Predd, S. R. Kulkarni, and H. V. Poor. A collaborative training algorithm for distributed learning. *IEEE Transactions on Information Theory*, 55(4):1856–1871, April 2009.

[71] S. Kar, J. M. F. Moura, and H. V. Poor. $\mathcal{QD}$-learning: A collaborative distributed strategy for multi-agent reinforcement learning through consensus + innovations. *IEEE Transactions on Signal Processing*, 61(7):1848–1862, April 2013.

[72] D. Mosk-Aoyama and D. Shah. Fast distributed algorithms for computing separable functions. *IEEE Transactions on Information Theory*, 54(7):2997–3007, July 2008.

[73] P. Bianchi, G. Fort, and W. Hachem. Performance of a distributed stochastic approximation algorithm. *IEEE Transactions on Information Theory*, 59(11):7405–7418, November 2013.

[74] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah. Randomized gossip algorithms. *IEEE Transactions on Information Theory*, 52(6):2508–2530, June 2006.

[75] F. Benezit, A. G. Dimakis, P. Thiran, and M. Vetterli. Order-optimal consensus through randomized path averaging. *IEEE Transactions on Information Theory*, 56(10):5150–5167, October 2010.

[76] T. C. Aysal and K. E. Barner. Convergence of consensus models with stochastic disturbances. *IEEE Transactions on Information Theory*, 56(8):4101–4113, August 2010.

[77] Y. Yang and R. S. Blum. Broadcast-based consensus with non-zero-mean stochastic perturbations. *IEEE Transactions on Information Theory*, 59(6):3971–3989, June 2013.

[78] S. Kar and J. M. F. Moura. Distributed consensus algorithms in sensor networks: Quantized data and random link failures. *IEEE Transactions on Signal Processing*, 58(3):1383–1400, March 2010.

[79] Ali Jadbabaie, Pooya Molavi, Alvaro Sandroni, and Alireza Tahbaz-Salehi. Non-Bayesian social learning. *Games and Economic Behavior*, 76(1):210–225, 2012.

[80] S. Shahrampour and A. Jadbabaie. Exponentially fast parameter estimation in networks using distributed dual averaging. In *Proceedings of the 52nd Annual IEEE Conference on Decision and Control (CDC), 2013*, pages 6196–6201, December 2013.

[81] Ali Jadbabaie, Pooya Molavi, and Alireza Tahbaz-salehi. Information Heterogeneity and the Speed of Learning in Social Networks. *Columbia Business School Research Paper*, (13-28), May 2013.

[82] K. Rahnama Rad and A. Tahbaz-Salehi. Distributed parameter estimation in networks. In *Proceedings of the 49th IEEE Conference on Decision and Control*, pages 5050–5055, December 2010.

[83] Reza Olfati-Saber, Elisa Franco, Emilio Frazzoli, and Jeff S. Shamma. Belief Consensus and Distributed Hypothesis Testing in Sensor Networks. In *Workshop on Network Embedded Sensing and Control*, Notre Dame University, South Bend, IN, October 2005.

[84] Morris H. DeGroot. Reaching a consensus. *Journal of the American Statistical Association*, 69(345):118–121, 1974.

[85] S. Shahrampour, A. Rakhlin, and A. Jadbabaie. Distributed detection: Finite-time analysis and impact of network topology. *IEEE Transactions on Automatic Control*, 61(11):3256–3268, November 2016.

[86] A. Nedić, A. Olshevsky, and C. A. Uribe. Nonasymptotic convergence rates for cooperative learning over time-varying directed graphs. In *Proceedings of the 2015 American Control Conference (ACC)*, pages 5884–5889, July 2015.

[87] S. Kar, J. M. F. Moura, and K. Ramanan. Distributed parameter estimation in sensor networks: Nonlinear observation models and imperfect communication. *IEEE Transactions on Information Theory*, 58(6):3575–3605, June 2012.

[88] V. Saligrama, M. Alanyali, and O. Savas. Distributed detection in sensor networks with packet losses and finite capacity links. *IEEE Transactions on Signal Processing*, 54(11):4118–4132, November 2006.

[89] M. Alanyali, S. Venkatesh, O. Savas, and S. Aeron. Distributed bayesian hypothesis testing in sensor networks. In *Proceedings of the 2004 American Control Conference*, volume 6, pages 5369–5374 vol.6, June 2004.

[90] Matan Harel, Elchanan Mossel, Philipp Strack, and Omer Tamuz. On the speed of social learning. *CoRR*, abs/1412.7172, 2014.

[91] Manuel Mueller-Frank. A general framework for rational learning in social networks. *Theoretical Economics*, 8(1):1–40, 2013.

[92] A. K. Sahu and S. Kar. Distributed sequential detection for gaussian shift-in-mean hypothesis testing. *IEEE Transactions on Signal Processing*, 64(1):89–103, January 2016.

[93] D. Bajovic, D. Jakovetic, J. M. F. Moura, J. Xavier, and B. Sinopoli. Large deviations performance of consensus+innovations distributed detection with non-gaussian observations. *IEEE Transactions on Signal Processing*, 60(11):5987–6002, November 2012.

[94] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, New York, NY, USA, 2 edition, 1991.

[95] A. G. Dimakis, S. Kar, J. M.F. Moura, M. G. Rabbat, and A. Scaglione. Gossip algorithms for distributed signal processing. *Proceedings of the IEEE*, 98(11):1847–1864, November 2010.

[96] Charles J. Stone Paul G. Hoel, Sidney C. Port. *Introduction to Stochastic Processes*. Waveland Press, 1972.

[97] A. Lalitha, A. Sarwate, and T. Javidi. Social learning and distributed hypothesis testing. In *Proceedings of the 2014 IEEE International Symposium on Information Theory*, pages 551–555, June 2014.

[98] A. Lalitha and T. Javidi. On the rate of learning in distributed hypothesis testing. In *Proceedings of the 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1–8, September 2015.

[99] A. Nedić, A. Olshevsky, and C. A. Uribe. Fast Convergence Rates for Distributed Non-Bayesian Learning. *IEEE Transactions on Automatic Control*, PP(99):1–1, 2017.

[100] Sadaf Salehkalaibar, Michele A. Wigger, and Ligong Wang. Hypothesis testing in multi-hop networks. volume abs/1708.05198, 2017.

[101] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):pp. 13–30, 1963.

[102] F. den Hollander. *Large Deviations*, volume 14 of *Fields Institute Monographs*. American Mathematical Society, 2000.

[103] A. Nedić, A. Olshevsky, and C. A. Uribe. Non-asymptotic Convergence Rates for Cooperative Learning over Time-varying Directed Graphs. In *2015 American Control Conference (ACC)*, pages 5884–5889, July 2015.

[104] S. Shahrampour, A. Rakhlin, and A. Jadbabaie. Distributed Detection: Finite-Time Analysis and Impact of Network Topology. *IEEE Transactions on Automatic Control*, 61(11):3256–3268, Nov 2016.

[105] A. Lalitha, T. Javidi, and A. D. Sarwate. Social Learning and Distributed Hypothesis Testing. *IEEE Transactions on Information Theory*, 64(9):6161–6179, Sept 2018.

[106] Yarin Gal. Uncertainty in Deep Learning. *University of Cambridge*, 2016.

[107] Durk P Kingma, Tim Salimans, and Max Welling. Variational dropout and the local reparameterization trick. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 2575–2583. Curran Associates, Inc., 2015.

[108] Jakub Konecný, H. Brendan McMahan, Daniel Ramage, and Peter Richtárik. Federated optimization: Distributed machine learning for on-device intelligence. *CoRR*, abs/1610.02527, 2016.

[109] Jakub Konečnỳ, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated Learning: Strategies for Improving Communication Efficiency. *arXiv preprint arXiv:1610.05492*, 2016.

[110] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Artificial Intelligence and Statistics*, pages 1273–1282, 2017.

[111] Cuong V. Nguyen, Yingzhen Li, Thang D. Bui, and Richard E. Turner. Variational continual learning. In *International Conference on Learning Representations*, 2018.

[112] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight Uncertainty in Neural Networks. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 1613–1622. JMLR.org, 2015.

[113] Tamara Broderick, Nicholas Boyd, Andre Wibisono, Ashia C Wilson, and Michael I Jordan. Streaming variational bayes. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 1727–1735. Curran Associates, Inc., 2013.

[114] Virginia Smith, Simone Forte, Chenxin Ma, Martin Takáč, Michael I. Jordan, and Martin Jaggi. CoCoA: A general framework for communication-efficient distributed optimization. *Journal of Machine Learning Research*, 18(230):1–49, 2018.

[115] Pratik Chaudhari, Carlo Baldassi, Riccardo Zecchina, Stefano Soatto, and Ameet Talwalkar. Parle: parallelizing stochastic gradient descent. *CoRR*, abs/1707.00424, 2017.

[116] Tao Lin, Sebastian U. Stich, and Martin Jaggi. Don't use large mini-batches, use local SGD. *CoRR*, abs/1808.07217, 2018.

[117] John C. Duchi, Alekh Agarwal, and Martin J. Wainwright. Dual averaging for distributed optimization: Convergence analysis and network scaling. *IEEE Transactions on Automatic Control*, 57:592–606, 2012.

[118] E. Wei and A. Ozdaglar. Distributed alternating direction method of multipliers. In *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*, pages 5445–5450, Dec 2012.

[119] Hanlin Tang, Xiangru Lian, Ming Yan, Ce Zhang, and Ji Liu. d^2: Decentralized training over decentralized data. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4848–4856, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

[120] L. Bottou, F. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

[121] Xiangru Lian, Ce Zhang, Huan Zhang, Cho-Jui Hsieh, Wei Zhang, and Ji Liu. Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5330–5340. Curran Associates, Inc., 2017.

[122] Zhanhong Jiang, Aditya Balu, Chinmay Hegde, and Soumik Sarkar. Collaborative deep learning in fixed topology networks. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5904–5914. Curran Associates, Inc., 2017.

[123] Peter H. Jin, Qiaochu Yuan, Forrest N. Iandola, and Kurt Keutzer. How to scale distributed deep learning? *CoRR*, abs/1611.04581, 2016.

[124] Jianyu Wang and Gauri Joshi. Cooperative SGD: A unified framework for the design and analysis of communication-efficient SGD algorithms. *CoRR*, abs/1808.07576, 2018.

[125] Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012.

[126] Yann LeCun and Corinna Cortes. MNIST handwritten digit database. 2010.

[127] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. 2017.

[128] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1050–1059, New York, New York, USA, 20–22 Jun 2016. PMLR.

[129] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5574–5584. Curran Associates, Inc., 2017.

[130] Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Damon Civin, and Vikas Chandra. Federated learning with non-iid data. *CoRR*, abs/1806.00582, 2018.

[131] Ian Goodfellow, Oriol Vinyals, and Andrew Saxe. Qualitatively characterizing neural network optimization problems. In *International Conference on Learning Representations*, 2015.

[132] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2015.