© 2016 Aolin Xu

INFORMATION-THEORETIC LIMITATIONS OF DISTRIBUTED INFORMATION PROCESSING

BY

AOLIN XU

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Electrical and Computer Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2016

Urbana, Illinois

Doctoral Committee:

Assistant Professor Maxim Raginsky, Chair
Professor Bruce Hajek
Professor Olgica Milenkovic
Professor Rayadurgam Srikant
Adjunct Assistant Professor Yihong Wu

Abstract

In a generic distributed information processing system, a number of agents connected by communication channels aim to accomplish a task collectively through local communications. The fundamental limits of distributed information processing problems depend not only on the intrinsic difficulty of the task, but also on the communication constraints due to the distributedness. In this thesis, we reveal these dependencies quantitatively under information-theoretic frameworks.

We consider three typical distributed information processing problems: decentralized parameter estimation, distributed function computation, and statistical learning under adaptive composition. For the first two problems, we derive converse results on the Bayes risk and the computation time, respectively. For the last problem, we first study the relationship between the generalization capability of a learning algorithm and its stability property measured by the mutual information between its input and output, and then derive achievability results on the generalization error of adaptively composed learning algorithms. In all cases, we obtain general results on the fundamental limits with respect to a general model of the problem, so that the results can be applied to various specific scenarios. Our information-theoretic analyses also provide general approaches to inferring global properties of a distributed information processing system from local properties of its components.

To Haohao

Acknowledgments

First of all, I would like to thank my advisor Prof. Maxim Raginsky. This thesis would not have been possible without his patient guidance, insightful advice, and tremendous intellectual support. I have also benefited from his broad knowledge spanning the fields of information theory, probability theory, and statistical learning, which I believe comes from his immense amount of reading. Prof. Raginsky not only reads and digests for himself; he also reads for everybody, which is manifested by his accessible lecture notes and blog posts. Often I have a feeling that he translates essays written in classical Chinese by some probabilists/statisticians into modern Chinese, or interprets strange dialects spoken by some computer scientists into Mandarin, so that more people can get an idea of something interesting but difficult to comprehend. To me, this characterizes an exemplary researcher and teacher.

I am grateful to Profs. Bruce Hajek, Olgica Milenkovic, Rayadurgam Srikant, and Yihong Wu, for kindly serving on my doctoral committee and giving me valuable comments and feedback. I would like to thank Profs. Pierre Moulin and Yihong Wu for teaching me the knowledge of information theory and statistics that I have been using throughout the doctoral study. I am also thankful to my MS advisor Prof. Kewu Peng, for encouraging me to pursue a doctoral degree abroad, and my formal advisor Prof. Naresh Shanbhag for his kind understanding of my interest in the new direction.

My gratitude also goes to my friends and colleagues at UIUC. I want to thank Peter Kairouz, for helping me both in course study and research, and for our fruitful discussions on many topics. I also want to thank the excellent group members, Peng Guan, Jaeho Lee, and Ehsan Shafieepoorfard, for learning new things together. A big thanks to Gong Zhang, Yingyan Lin, and Sai Zhang, who helped me tremendously in daily life especially when things got tough. I would like to thank all of my college friends for their constant encouragement. A special thanks to Lijie Huang and Huacheng

Zeng, who also went abroad to pursue PhD degrees, for sharing with me the joys and hard times throughout the years.

I am deeply indebted to my parents, for their unwavering love and support and always being ready to help. Finally, I would like to thank my wife, Haohao, for having been with me all the time and trying her best to help me focus on the doctoral study. I dedicate this thesis to her.

Table of Contents

Chapter 1

Introduction

In this thesis we use information-theoretic tools to study the fundamental limits of distributed information processing. In a generic distributed information processing system, a number of agents connected by communication channels aim to accomplish a task collectively through local communications. The communication among the agents can be either static, where certain agents passively receive messages sent by the other agents; or dynamic, where the agents interact with each other so that the messages sent by each agent depend on the messages received from the other agents; or sequential, where each agent receives messages from the upstream agents, and sends messages to downstream agents. The fundamental limits of distributed information processing thus depend not only on the intrinsic difficulty of the task, but also on the communication constraints due to the distributedness. Our objective is to reveal these dependencies quantitatively under information-theoretic frameworks. The methods of analysis we develop also provide general approaches to inferring the global properties of a distributed information processing system from the local properties of its components.

We consider three typical distributed information processing problems: decentralized parameter estimation, as an example of non-interactive processing; distributed function computation, as an example of interactive processing; and statistical learning under adaptive composition, as an example of sequential processing. For each problem, our goal is to obtain general results on the fundamental limits with respect to a general model of the problem, so that the results can be applied to various specific scenarios. For the first two problems, we derive converse results on the Bayes risk and the computation time, respectively. For the last problem, we first study the relationship between the generalization capability of a learning algorithm and its stability property measured by the mutual information between its input and output, and then derive achievability results on the generalization error of learning

algorithms obtained from adaptive composition.

1.1 Decentralized Estimation

In Chapter 2, we study lower bounds for the Bayes risk in decentralized parameter estimation. In decentralized estimation, the estimator does not have direct access to the samples generated according to the parameter of interest, but only to the quantized and possibly noise-contaminated data received from a local processor that observes the samples. The estimation performance is therefore not only subject to the statistical relationship between the parameter and the samples, but also subject to the communication constraints caused by the separation of the local processor from the estimator. When there are more than one local processor, the estimation performance further degrades because of the distribution of the samples and the communication resources to multiple processors.

We start with deriving Bayes risk lower bounds for the centralized estimation, where we find ways to relate the Bayes risk to various information-theoretic quantities, such as the small ball probability of the parameter and the mutual information between the parameter and the samples. The lower bounds reflect how the estimation performance is limited by the statistical constraint, which exists even in the absence of the communication constraints, and is determined by the joint distribution of the parameter and the samples as well as the distortion function. The lower bounds for the centralized estimation also serve as the basis for deriving lower bounds for the decentralized estimation.

When the estimation is decentralized, due to the quantization of the samples and the transmission over noisy channels, the mutual information between the parameter and the data received by the estimator is a contraction of the mutual information between the parameter and the original samples. Therefore, when applying the lower bounds derived for centralized estimation to the decentralized estimation problems, it is crucial to sharply quantify this contraction of information according to the communication constraints. We use a powerful tool called the strong data processing inequality for this purpose. The strong data processing inequality also couples the communication constraint and the statistical constraint together in the lower bounds, which

allows us to obtain much tighter Bayes risk lower bounds than one could get from the ordinary data processing inequality.

For the situation with multiple local processors, we consider the cases where the processors observe independent or dependent sample sets conditional on the parameter, and where the communication protocol is non-interactive or interactive. Our general results can recover and improve the existing Bayes risk and minimax risk lower bounds for specific decentralized estimation problems with noiseless channels, and also capture the effect of noisy channels on the estimation performance. Moreover, our lower bounds provide a general way to quantify the degradation of estimation performance caused by distributing samples and communication resources to multiple processors, which is only discussed for specific examples in existing works.

1.2 Distributed Function Computation

In Chapter 3, we study the problem of distributed function computation. We consider a general model where the computing agents, or nodes, are connected by point-to-point discrete memoryless channels, and each node aims to compute a common function of the observations of all the nodes in the network through local communication and computation. We are interested in obtaining lower bounds for the fundamental limit on the computation time, i.e., the minimum time needed by any algorithm to achieve a given accuracy with a given confidence on the computation result at each node.

The quantity that plays a key role in our derivation is the conditional mutual information between the function value and the computation result of an arbitrary node, given the observations in a subset of nodes that contains this node. Any requirement on the accuracy and confidence of the computation results translates into a lower bound on this conditional mutual information. On the other hand, the computation time of the algorithm, as well as the communication constraints imposed by the network topology and by the channel noise, place an upper bound on this conditional mutual information. Our main objective is to establish tight lower and upper bounds on this conditional mutual information according to the performance requirement and the communication and time constraints, which in turn will provide us with lower bounds on the computation time.

For the lower bound on the conditional mutual information of interest, we propose a bound via the small ball probability, which captures the dependence of the computation time on the joint distribution of the observations at the nodes, the structure of the function, and the accuracy requirement. For linear functions, the small ball probability can be expressed in terms of Lévy concentration functions of sums of independent random variables, which lead to strict improvements over existing results.

For upper bounds on the conditional mutual information of interest, we propose a bound based on the strong data processing inequality, which complements and strengthens the cutset-capacity upper bound in the literature. In addition, to address the limitation of the single-cutset analysis in the literature, we propose a multi-cutset analysis, which quantifies the dissipation of information as it flows across a sequence of cutsets in the network. This analysis is based on reducing a general network to a bidirected chain, and the results highlight the dependence of the computation time on the diameter of the network, a fundamental parameter that is missing from most of the existing results.

1.3 Stability and Generalization of Statistical Learning Algorithms

In Chapter 4, we study the information-theoretic stability and generalization capabilities of statistical learning algorithms. A statistical learning algorithm can be viewed as a (possibly randomized) transformation that maps the training dataset to a hypothesis. We say that such an algorithm is *stable* if its output does not depend too much on any individual training instance. Since stability is closely connected to generalization capabilities of learning algorithms, it is of theoretical and practical interest to obtain sharp quantitative estimates on the generalization error of learning algorithms in terms of their stability properties.

We propose an information-theoretic notion of stability based on the mutual information between a learning algorithm's input dataset and its output hypothesis. This notion of stability naturally follows from the idea that stability imposes limits on the amount of information the algorithm can extract from the observed data. We derive upper bounds on the expected general-

ization error and the absolute generalization error of a learning algorithm in terms of its input-output mutual information. These upper bounds formalize and quantify the intuition that the less information the output of a learning algorithm contains about the input dataset, the better the algorithm generalizes, or the less it overfits. We also discuss how to design learning algorithms with controlled input-output mutual information, and show that regularizing the empirical risk minimization algorithm with the input-output mutual information leads to the well-known Gibbs algorithm.

Another benefit of relating the generalization error of a learning algorithm to its input-output mutual information is the ease of analyzing the generalization capability of learning algorithms obtained from adaptive composition of constituent algorithms. Adaptive composition can be realized in a decentralized manner by multiple processors sharing the same dataset and sequentially executing the constituent algorithms: each processor takes the dataset as well as the outputs of the executed algorithms as its input, executes its own algorithm, and sends the output to the downstream processors. By bounding the input-output mutual information of each constituent algorithm, we can upper-bound the generalization error of the final output of the composed algorithm. The information-theoretic analysis also helps to capture how the communication constraints due to separation of the processors influence the generalization capability of the composed algorithm: although the effective hypothesis space may be reduced by the communication constraint, the composed algorithm can generalize better because of it. The same techniques can be applied to adaptive data analytics, where the analyst chooses queries by interacting with the dataset in multiple rounds, a topic that has become popular in recent years.

Chapter 2

Lower Bounds for Bayes Risk in Decentralized Estimation

2.1 Introduction

2.1.1 Decentralized Estimation

In decentralized estimation, the estimator does not have direct access to the samples generated according to the parameter of interest, but only to the data received from local processors that observe the samples. In this chapter, we consider a general model of decentralized estimation, where each local processor observes a set of samples generated according to a common random parameter W, quantizes the samples to a fixed-length binary message, and then encodes and sends the message to the estimator over an independent and possibly noisy communication channel. When the communication channels are noiseless and feedback from the estimator to the local processors is available, the processors can operate in an interactive protocol by taking turns to send messages, where the message sent by each processor can depend on the previous messages sent by the other processors. An estimate $\widehat{W}$ is then computed based on the messages received from the local processors. The estimation performance is measured by the expected distortion between W and $\widehat{W}$, with respect to some distortion function. The minimum possible expected distortion is defined as the Bayes risk. We derive lower bounds on the Bayes risk for this estimation problem, and gain insight into the fundamental limits of decentralized estimation.

There are three types of constraints inherent in decentralized estimation. The first, and the most fundamental one, is the *statistical constraint*, determined by the joint distribution of the parameter and the samples. The statistical constraint exists even in the centralized estimation, where the estimator can directly observe the samples. To study how the estimation

performance is limited by the statistical constraint, we start with deriving lower bounds on the Bayes risk for centralized estimation in Sec. 2.2. The results obtained in Sec. 2.2 apply to the decentralized estimation as well, but, more importantly, they also serve as the basis for the refined lower bounds for the decentralized estimation in Sec. 2.4 and Sec. 2.5.

The second is the *communication constraint*, due to the separation between the local processors and the estimator. The communication constraint arises even when there is only one local processor. It can be caused by the finite precision of analog-to-digital conversion, limitations on the storage of intermediate results, limited transmission blocklength, channel noise, etc. In Sec. 2.4, we present a detailed study of decentralized estimation with a single processor and reveal the influence of the communication constraint on the estimation performance. Section. 2.3 contains background information on *strong data processing inequalities*, the major tool used in our analysis of the communication constraint.

The third constraint appears when there are more than one local processors. It is the *penalty of decentralization*, caused by distributing the samples and communication resources to multiple processors. We study decentralized estimation with multiple processors in Sec. 2.5, where we show that, regardless of whether or not the sample sets seen by different local processors are conditionally independent given the parameter, the degradation of estimation performance becomes more pronounced when the resources are distributed to more processors. We also provide lower bounds on the Bayes risk for interactive protocols, where the processors take turns to send their messages, and each processor sends one message based on its sample set and the previous messages sent by other processors.

2.1.2 Method of Analysis

Our method of analysis is information-theoretic in nature. The major quantity we examine is the conditional mutual information $I(W;\widehat{W}|U)$ with a judiciously chosen auxiliary random variable U.

We first lower-bound this quantity according to the estimation performance, such as the probability of excess distortion or the expected distortion. The lower bounds will also depend on the *a priori* uncertainty about

W, measured either by its small ball probability or by its differential entropy. Any such lower bound can be viewed as a generalization of Fano's inequality, which indicates the least amount of information about W that must be contained in $\widehat{W}$ in order to achieve a certain estimation performance. We also analyze the probability of excess distortion and the expected distortion via the distribution of the conditional information density $i(W;\widehat{W}|U)$.

On the other hand, various constraints inherent in decentralized estimation impose upper bounds on $I(W;\widehat{W}|U)$. According to the statistical constraint, $I(W;\widehat{W}|U)$ is upper-bounded by the conditional mutual information between W and the samples. The communication constraint further implies that the amount of information about W contained in the estimator's indirect observation of the samples will be a contraction of the amount contained in the samples. We use strong data processing inequalities to quantify this contraction of information and to couple the communication constraint and the statistical constraint together in the upper bounds on $I(W;\widehat{W}|U)$. When there are multiple processors, strong data processing inequalities also give an upper bound that decreases as the samples and communication resources are distributed to more processors, which reflects the penalty of decentralization. In addition, we rely on a cutset analysis that chooses the conditioning random variable U to consist of all the samples seen by only a subset of the processors; this choice is useful for analyzing the situation where the processors observe sample sets that are dependent conditional on W.

Finally, by combining the upper and lower bounds on $I(W;\widehat{W}|U)$, we obtain lower bounds on the Bayes risk.

2.1.3 Related Work

Early work on the fundamental limits of decentralized estimation mainly focused on the asymptotic setting, e.g., determining the error exponent in multiterminal hypothesis testing with fixed quantization rates. That work is surveyed by Han and Amari [1]. In recent years, the focus has shifted towards determining explicit dependence of the estimation performance on the communication constraint (see, e.g., [2–6] and references therein). For instance, Zhang et al. [2] and Duchi et al. [3] derived lower bounds on the minimax risk of several decentralized estimation problems with noiseless communica-

tion channels. Their results also provide lower bounds on the number of bits needed in quantization to achieve the same minimax rate as in the centralized estimation. Garg et al. [4] extended the lower bound for interactive protocols in [2], which centered on the one-dimensional Gaussian location model, to the setting of high-dimensional Gaussian location models. Braverman et al. [5] presented lower bounds for decentralized estimation of a sparse multivariate Gaussian mean. Their derivation is based on a "distributed data processing inequality," which quantifies the information loss in decentralized binary hypothesis testing under the Gaussian location model. Shamir [6] showed that the analysis of several decentralized estimation and online learning problems can be reduced to a certain meta-problem involving discrete parameter estimation with interactive protocols, and derived minimax lower bounds for this meta-problem.

The main idea underlying all of the above works is that one has to quantify the contraction of information due to the communication constraint; however, this is often done in a case-by-case manner for each particular problem, and the resulting contraction coefficients are generally not sharp. Additionally, these works only consider the situation where the sample sets are conditionally independent given the parameter and where the communication channels connecting the processors to the estimator are noiseless.

By contrast, we derive general lower bounds on the Bayes risk, which automatically serve as lower bounds on the minimax risk. We use strong data processing inequalities as a unifiying general method for quantifying the contraction of mutual information in decentralized estimation. Our results apply to general priors, sample generating models, and distortion functions. When particularized to the examples in the existing works, our results can lead to sharper lower bounds on both the Bayes and the minimax risk. For example, we improve the lower bound for the mean estimation on the unit cube studied in [2], as well as the lower bound for the meta-problem of Shamir [6]. Moreover, we consider the situations where the sample sets are conditionally dependent and where the communication channels are noisy. We also provide a general way to quantify the degradation of estimation performance caused by distributing resources to multiple processors, which is only discussed for specific examples in existing work.

2.2 Bayes Risk Lower Bounds for Centralized Estimation

In the standard Bayesian estimation framework, $\mathcal{P} = \{P_{X|W=w} : w \in \mathsf{W}\}$ is a family of distributions on an observation space X, where the parameter space W is endowed with a prior distribution P_W. Given $W = w$, a sample X is generated from $P_{X|W=w}$. In centralized estimation, the unknown random parameter $W \sim P_W$ is estimated from X as $\widehat{W} = \psi(X)$, via an estimator $\psi : \mathsf{X} \to \mathsf{W}$. Given a non-negative distortion function $\ell : \mathsf{W} \times \mathsf{W} \to \mathbb{R}^+$, define the Bayes risk for estimating W from X with respect to ℓ as

$$R_{\mathrm{B}} = \inf_{\psi} \mathbb{E}[\ell(W, \psi(X))]. \tag{2.1}$$

In this section, we derive lower bounds on the Bayes risk in the context of centralized estimation. These bounds serve as lower bounds for the decentralized setting as well, but they can also be used to derive refined lower bounds for decentralized estimation, as shown in Secs. 2.4 and 2.5. We first present lower bounds on the Bayes risk based on small ball probability, mutual information, and information density in Secs. 2.2.1 and 2.2.2. These lower bounds apply to estimation problems with an arbitrary joint distribution $P_{W,X}$ and an arbitrary distortion function ℓ, and also provide generalizations of Fano's inequality, as discussed in Sec. 2.2.3. Next, in Sec. 2.2.4, we present a lower bound based on mutual information and differential entropy, which applies to parameter estimation problems in $\mathbb{R}^d$, with distortion functions of the form $\ell(w, \widehat{w}) = \|w - \widehat{w}\|^r$ for some norm $\|\cdot\|$ in $\mathbb{R}^d$ and some $r \geq 1$.

2.2.1 Lower Bounds Based on Mutual Information and Small Ball Probability

The *small ball probability* of W with respect to distortion function ℓ is defined as

$$\mathcal{L}_W(\rho) = \sup_{w \in \mathsf{W}} \mathbb{P}[\ell(W, w) \leq \rho]. \tag{2.2}$$

Given another random variable U jointly distributed with W, the conditional small ball probability of W given $U = u$ is defined as

$$\mathcal{L}_{W|U}(u, \rho) = \sup_{w \in \mathsf{W}} \mathbb{P}[\ell(W, w) \le \rho | U = u]. \tag{2.3}$$

Each of these two quantities measures how well the distribution P_W or $P_{W|U=u}$ concentrates in a small region of size ρ as measured by $\ell(\cdot, \cdot)$. The larger the small ball probability, the more concentrated the corresponding distribution is. We give a lower bound on the probability of excess distortion in terms of conditional mutual information and conditional small ball probability:

Lemma 2.1. *For any estimate $\widehat{W}$ of W, any $\rho \ge 0$, and any auxiliary random variable U,*

$$\mathbb{P}[\ell(W, \widehat{W}) > \rho] \ge 1 - \frac{I(W; \widehat{W}|U) + 1}{\log\left(1/\mathbb{E}[\mathcal{L}_{W|U}(U, \rho)]\right)}. \tag{2.4}$$

Proof. The inequality (2.4) is a direct consequence of the lower bound on the conditional mutual information: whenever $\mathbb{P}[\ell(W, \widehat{W}) > \rho] \le \delta$, it holds that

$$I(W; \widehat{W}|U) \ge (1 - \delta) \log \frac{1}{\mathbb{E}[\mathcal{L}_{W|U}(U, \rho)]} - 1,$$

which follows from the proof of Lemma 3.1 in Chapter 3. In Sec. 2.7.1, we present an alternative unified proof of Lemmas 2.1 and 2.2 using properties of the Neyman–Pearson function. □

Our first lower bound on the Bayes risk for centralized estimation is an immediate consequence of Lemma 2.1:

Theorem 2.1. *The Bayes risk for estimating the parameter W based on the sample X with respect to the distortion function ℓ satisfies*

$$R_{\mathrm{B}} \ge \sup_{P_{U|W,X}} \sup_{\rho > 0} \rho \left(1 - \frac{I(W; X|U) + 1}{\log(1/\mathbb{E}[\mathcal{L}_{W|U}(U, \rho)])}\right). \tag{2.5}$$

In particular,

$$R_{\mathrm{B}} \ge \sup_{\rho > 0} \rho \left(1 - \frac{I(W; X) + 1}{\log(1/\mathcal{L}_W(\rho))}\right). \tag{2.6}$$

Proof. For an arbitrary estimator $\psi : \mathsf{X} \to \mathsf{W}$,

$$I(W;\widehat{W}|U) \le I(W;X|U) \tag{2.7}$$

by the data processing inequality. It follows from Lemma 2.1 that

$$\mathbb{P}[\ell(W,\widehat{W}) > \rho] \ge 1 - \frac{I(W;X|U)+1}{\log\left(1/\mathbb{E}[\mathcal{L}_{W|U}(U,\rho)]\right)}, \qquad \rho \ge 0. \tag{2.8}$$

Theorem 2.1 follows from Markov's inequality $\mathbb{E}[\ell(W,\widehat{W})] \ge \rho\,\mathbb{P}[\ell(W,\widehat{W}) > \rho]$ and from the arbitrariness of ψ, $P_{U|W,X}$, and $\rho \ge 0$. □

Remark 2.1. Precise evaluation of the expected conditional small ball probability $\mathbb{E}[\mathcal{L}_{W|U}(U,\rho)]$ in Theorem 2.1 can be difficult. The following technique may sometimes be useful: Suppose we can upper-bound $\mathbb{E}[\mathcal{L}_{W|U}(U,\rho)]$ by some increasing function $g(\rho)$, which has an inverse function $g^{-1}(p) = \sup\{\rho > 0 : g(\rho) \le p\}$. Given some $s \in (0,1)$, choosing a suitable $\rho > 0$ such that

$$g(\rho) \le 2^{-(I(W;X|U)+1)/(1-s)} \tag{2.9}$$

guarantees

$$1 - \frac{I(W;X|U)+1}{\log\left(1/\mathbb{E}[\mathcal{L}_{W|U}(U,\rho)]\right)} \ge s. \tag{2.10}$$

It then follows from Theorem 2.1 that

$$R_{\mathrm{B}} \ge \sup_{P_{U|W,X}} \sup_{0<s<1} s g^{-1}\left(2^{-(I(W;X|U)+1)/(1-s)}\right). \tag{2.11}$$

A similar methodology for deriving lower bounds on the Bayes risk has been recently proposed by Chen et al. [7], who obtained unconditional lower bounds similar to (2.6) in terms of general f-informativities [8] and a quantity essentially the same as the small ball probability. However, as will be shown later, the conditional lower bound (2.5) can lead to tighter results compared to the unconditional version (2.6), and is also useful in the context of decentralized estimation problems.

For the problem of estimating W based on n samples $X_1, \ldots, X_n$ conditionally i.i.d. given W, we can choose the conditioning random variable U in (2.5)

to be an independent copy of $X^n = (X_1, \ldots, X_n)$ conditional on W, denoted as X'^n — that is, $P_{W,X^n,X'^n} = P_W \otimes P_{X^n|W} \otimes P_{X'^n|W}$ and $P_{X'^n|W} = P_{X^n|W}$. This choice leads to

$$R_{\mathrm{B}} \geq \sup_{\rho>0} \rho \left(1 - \frac{I(W; X^n|X'^n) + 1}{\log\left(1/\mathbb{E}[\mathcal{L}_{W|X^n}(X^n, \rho)]\right)} \right). \tag{2.12}$$

We then need to evaluate or upper-bound $I(W; X^n|X'^n)$ and $\mathbb{E}[\mathcal{L}_{W|X^n}(X^n, \rho)]$. For example, in the smooth parametric case when $\mathcal{P}$ is a subset of a finite-dimensional exponential family and W has a density supported on a compact subset of $\mathbb{R}^d$, it was shown by Clarke and Barron [9, 10] that

$$I(W; X^n) = \frac{d}{2} \log \frac{n}{2\pi e} + h(W) + \frac{1}{2}\mathbb{E}[\log \det J_{X|W}(W)] + o(1) \qquad \text{as } n \to \infty \tag{2.13}$$

where $h(W)$ is the differential entropy of W, and $J_{X|W}(w)$ is the Fisher information matrix about w contained in X. When (2.13) holds, we have

$$I(W; X^n|X'^n) = I(W; X^n, X'^n) - I(W; X'^n) \tag{2.14}$$

$$\to \frac{d}{2} \qquad \text{as } n \to \infty, \tag{2.15}$$

meaning that $I(W; X^n|X'^n)$ in (2.12) is asymptotically independent of n. Upper-bounding $\mathbb{E}[\mathcal{L}_{W|X^n}(X^n, \rho)]$ is more problem-specific. We give two examples below, in both of which we consider the absolute distortion $\ell(w, \widehat{w}) = |w - \widehat{w}|$, such that the Bayes risk gives the minimum mean absolute error (MMAE). A benefit of lower-bounding MMAE is that the square of the resulting lower bound also serves as a lower bound for the minimum mean squared error (MMSE).

Example 2.1 (Estimating Gaussian mean with Gaussian prior)**.** *Consider the case where the parameter $W \sim N(0, \sigma_W^2)$, the samples are $X_i = W + Z_i$ with $Z_i \sim N(0, \sigma^2)$ independent of W for $i = 1, \ldots, n$, and $\ell(w, \widehat{w}) = |w - \widehat{w}|$.*

From the conditional lower bound (2.12), we get the following lower bound for Example 2.1:

Corollary 2.1. *In Example 2.1, the Bayes risk is lower bounded by*

$$R_{\mathrm{B}} \geq \frac{1}{16}\sqrt{\frac{\pi\sigma_W^2}{2(1+n\sigma_W^2/\sigma^2)}}. \tag{2.16}$$

Proof. Section 2.7.2. □

Note that the MMAE in Example 2.1 is upper bounded by

$$R_{\mathrm{B}} \leq \sqrt{\frac{\sigma_W^2}{1+n\sigma_W^2/\sigma^2}}, \tag{2.17}$$

which is achieved by $\widehat{W} = \mathbb{E}[W|X^n]$. Thus the non-asymptotic lower bound on the Bayes risk in (2.16) captures the correct dependence on n, and is off from the true Bayes risk by a constant factor. If we apply the unconditional lower bound (2.6) to Example 2.1, we can only get an asymptotic lower bound

$$R_{\mathrm{B}} \gtrsim \frac{1}{4\log\left(1+n\sigma_W^2/\sigma^2\right)}\sqrt{\frac{\pi\sigma_W^2}{1+n\sigma_W^2/\sigma^2}} \qquad \text{as } n \to \infty, \tag{2.18}$$

which differs from the upper bound by a logarithmic factor in n. This example shows that the conditional lower bound (2.5) can provide tighter results than its unconditional counterpart (2.6).

Example 2.2 (Estimating Bernoulli bias with uniform prior)**.** *Consider the example where the parameter $W \sim U[0,1]$, the samples $X_i \sim \mathrm{Bern}(w)$ conditional on $W = w$ for $i = 1, \ldots, n$, and $\ell(w, \widehat{w}) = |w - \widehat{w}|$.*

Corollary 2.2. *In Example 2.2, the Bayes risk is lower bounded by*

$$R_{\mathrm{B}} \gtrsim \frac{1}{16\sqrt{2\pi n}} \qquad \text{as } n \to \infty. \tag{2.19}$$

Proof. Section 2.7.2. □

Note that the MMAE in Example 2.2 is upper bounded by

$$R_{\mathrm{B}} \leq \frac{1}{\sqrt{6n}}, \tag{2.20}$$

which is achieved by the sample mean estimator $\widehat{W} = \frac{1}{n}\sum_{i=1}^n X_i$. Thus, the

lower bound in (2.19) asymptotically captures the correct dependence on n, and is off from the true Bayes risk by a constant factor.

2.2.2 Lower Bounds Based on Information Density and Small Ball Probability

For a joint distribution $P_{U,W,X}$ on $\mathsf{U} \times \mathsf{W} \times \mathsf{X}$, define the conditional information density as

$$i(w;x|u) = \log \frac{\mathrm{d}P_{W|U=u,X=x}}{\mathrm{d}P_{W|U=u}}(w). \tag{2.21}$$

We give a lower bound on the probability of excess distortion in terms of conditional information density and conditional small ball probability:

Lemma 2.2. *For any estimate $\widehat{W}$ of W based on the sample X, any $\rho \geq 0$, $\gamma > 0$, and any auxiliary random variable U,*

$$\begin{aligned}\mathbb{P}[\ell(W,\widehat{W}) > \rho] \geq &\mathbb{P}[i(W;X|U) < \log\gamma] - \gamma\mathbb{E}[\mathcal{L}_{W|U}(U,\rho)] + \\ &\gamma \inf_{u,w,x} \frac{\mathrm{d}P_{W|U=u}}{\mathrm{d}P_{W|U=u,X=x}}(w)\, \mathbb{P}[i(W;X|U) \geq \log\gamma].\end{aligned} \tag{2.22}$$

Proof. The proof, inspired by the metaconverse technique from [11], is given in Sec. 2.7.1. □

Our second Bayes risk lower bound for centralized estimation is a consequence of Lemma 2.2:

Theorem 2.2. *The Bayes risk for estimating the parameter W based on the sample X with respect to the distortion function ℓ satisfies*

$$R_{\mathrm{B}} \geq \sup_{P_{U|W,X}} \sup_{\rho,\gamma>0} \rho\big(\mathbb{P}[i(W;X|U) < \log\gamma] - \gamma\mathbb{E}[\mathcal{L}_{W|U}(U,\rho)]\big). \tag{2.23}$$

In particular,

$$R_{\mathrm{B}} \geq \sup_{\rho,\gamma>0} \rho\big(\mathbb{P}[i(W;X) < \log\gamma] - \gamma\mathcal{L}_W(\rho)\big). \tag{2.24}$$

Proof. With the aid of Markov's inequality, (2.22) leads to the inequality

$$R_{\mathrm{B}} \geq \sup_{P_{U|W,X}} \sup_{\rho,\gamma>0} \rho\Bigg(\mathbb{P}[i(W;X|U) < \log\gamma] - \gamma\mathbb{E}[\mathcal{L}_{W|U}(U,\rho)] + \gamma \inf_{u,w,x} \frac{\mathrm{d}P_{W|U=u}}{\mathrm{d}P_{W|U=u,X=x}}(w)\mathbb{P}[i(W;X|U) \geq \log\gamma]\Bigg). \quad (2.25)$$

The lower bound in (2.23) follows by replacing $\inf_{u,w,x} \frac{\mathrm{d}P_{W|U=u}}{\mathrm{d}P_{W|U=u,X=x}}(w)$ with zero. □

We give a high-dimensional example to illustrate the usefulness of Theorem 2.2:

Example 2.3 (Estimating d-dimensional Gaussian mean with uniform prior on d-ball)**.** *Consider the case where the parameter $W \in \mathbb{R}^d$ is distributed uniformly on the ball $\mathsf{W} = \{w \in \mathbb{R}^d : \|w\|_2 \leq a\}$, the samples are $X_i = W + Z_i$ with $Z_i \sim N(0, \sigma^2\mathbf{I}_d)$ independent of W for $i = 1, \ldots, n$, and $\ell(w, \widehat{w}) = \|w - \widehat{w}\|_2$.*

Corollary 2.3. *In Example 2.3, for any $a > 0$, $\sigma^2 > 0$, and $d \geq 1$, the Bayes risk is lower bounded by*

$$R_{\mathrm{B}} \gtrsim \frac{1}{20}\sqrt{\frac{2\pi\sigma^2 d}{n}} \qquad \text{as } n \to \infty. \quad (2.26)$$

Proof. Section 2.7.3. □

Note that the Bayes risk in Example 2.3 is upper bounded by

$$R_{\mathrm{B}} \leq \sqrt{\frac{\sigma^2 d}{n}}, \quad (2.27)$$

which is achieved by the sample mean estimator $\widehat{W} = \frac{1}{n}\sum_{i=1}^n X_i$. Thus, the lower bound in (2.26) captures the correct dependence on n (asymptotically) and d (non-asymptotically), and is off from the true Bayes risk by a constant factor. Moreover, by squaring (2.26), we get a lower bound on the MMSE that also captures the correct dependence on n and d.

2.2.3 Generalizations of Fano's Inequality

The lower bounds on the probability of excess distortion in Lemmas 2.1 and 2.2 can be viewed as generalizations of Fano's inequality.

When W takes values on $\{1, \ldots, M\}$ and $\ell(w, \widehat{w}) = \mathbf{1}\{w \neq \widehat{w}\}$, setting $\rho = 0$ in (2.4) without conditioning on U recovers the following generalization of Fano's inequality due to Han and Verdú [12]:

$$\mathbb{P}[\widehat{W} \neq W] \geq 1 - \frac{I(W;X) + 1}{\log(1/\max_{w \in [M]} P_W(w))}. \tag{2.28}$$

Similarly, setting $\rho = 0$ in (2.22) without conditioning on U, we get

$$\begin{aligned}\mathbb{P}[\widehat{W} \neq W] \geq \sup_{\gamma > 0} \mathbb{P}[i(W;X) < \log \gamma] - \gamma \max_{w \in [M]} P_W(w) + \\ \gamma \inf_{w,x} \frac{\mathrm{d}P_W}{\mathrm{d}P_{W|X=x}}(w) \mathbb{P}[i(W;X) \geq \log \gamma].\end{aligned} \tag{2.29}$$

When W is uniformly distributed on $\{1, \ldots, M\}$, (2.28) reduces to the usual Fano's inequality

$$\mathbb{P}[\widehat{W} \neq W] \geq 1 - \frac{I(W;X) + 1}{\log M}, \tag{2.30}$$

while (2.29) reduces to the Poor–Verdú bound [13]

$$\mathbb{P}[\widehat{W} \neq W] \geq \sup_{\gamma > 0} \left(1 - \frac{\gamma}{M}\right) \mathbb{P}[i(W;X) < \log \gamma]. \tag{2.31}$$

When W is continuous, (2.4) and (2.22) provide continuum generalizations of Fano's inequality. For example, when $\mathsf{W} \subset \mathbb{R}^d$ and $\ell(w, \widehat{w}) = \|w - \widehat{w}\|_2$, (2.4) leads to

$$\mathbb{P}\big[\|\widehat{W} - W\|_2 > \rho\big] \geq 1 - \frac{I(W;X) + 1}{\log(1/\sup_{w \in \mathsf{W}} \mathbb{P}[\|W - w\|_2 \leq \rho])}, \tag{2.32}$$

which is also obtained by Chen et al. [7], and generalizes the result of Duchi and Wainwright [14]. Similarly, (2.22) leads to

$$\mathbb{P}[\|\widehat{W} - W\|_2 > \rho] \geq \sup_{\gamma > 0} \left(\mathbb{P}[i(W;X) < \log \gamma] - \gamma \sup_{w \in \mathsf{W}} \mathbb{P}[\|W - w\|_2 \leq \rho] \right). \tag{2.33}$$

2.2.4 Lower Bounds Based on Mutual Information and Differential Entropy

For the problem of estimating a real-valued parameter W with respect to the quadratic distortion $\ell(w, \widehat{w}) = (w - \widehat{w})^2$, it can be shown that (see, e.g. Lemma 3.3 in Sec. 3.2.1), if $\mathbb{E}(W - \widehat{W})^2 \leq \alpha$, then

$$I(W; \widehat{W}|U) \geq h(W|U) - \frac{1}{2}\log(2\pi e\alpha). \tag{2.34}$$

Upper-bounding $I(W; \widehat{W}|U)$ by $I(W; X|U)$, we obtain a lower bound on the MMSE

$$\inf_{\psi} \mathbb{E}(W - \widehat{W})^2 \geq \sup_{P_{U|W,X}} \frac{1}{2\pi e} 2^{-2(I(W;X|U) - h(W|U))}. \tag{2.35}$$

More generally, for the problem of estimating a parameter W taking values in $\mathbb{R}^d$, the Shannon lower bound on the rate-distortion function (see, e.g., [15, Chap. 4.8]) can be used to show that, if $\mathbb{E}\|W - \widehat{W}\|^r \leq \alpha$ with an arbitrary norm $\|\cdot\|$ in $\mathbb{R}^d$ and an arbitrary $r \geq 1$, then

$$I(W; \widehat{W}) \geq h(W) - \log\left(V_d\Big(\frac{\alpha r e}{d}\Big)^{d/r} \Gamma\Big(1 + \frac{d}{r}\Big)\right), \tag{2.36}$$

where V_d is the volume of the unit ball in $(\mathbb{R}^d, \|\cdot\|)$ and $\Gamma(\cdot)$ is the gamma function. For example, this method can be used to recover the lower bounds of Seidler [16] for the problem of estimating a parameter in $\mathbb{R}^d$ with respect to squared weighted ℓ_2 norms, and gives tight lower bounds on the Bayes risk and the minimax risk in high-dimensional estimation problems [17, Lec. 13]. A simple extension of (2.36) via an auxiliary random variable U gives

$$I(W; \widehat{W}|U) \geq h(W|U) - \log\left(V_d\Big(\frac{\alpha r e}{d}\Big)^{d/r} \Gamma\Big(1 + \frac{d}{r}\Big)\right). \tag{2.37}$$

As a consequence, we obtain a lower bound on the Bayes risk in terms of conditional mutual information and conditional differential entropy:

Theorem 2.3. *For an arbitrary norm $\|\cdot\|$ in $\mathbb{R}^d$ and any $r \geq 1$, the Bayes risk for estimating the parameter $W \in \mathbb{R}^d$ based on the sample X with respect*

to the distortion function $\ell(w, \widehat{w}) = \|w - \widehat{w}\|^r$ *satisfies*

$$R_{\mathrm{B}} \geq \sup_{P_{U|W,X}} \frac{d}{re} \left(V_d \Gamma\Big(1 + \frac{d}{r}\Big) \right)^{-r/d} 2^{-(I(W;X|U) - h(W|U))r/d}. \tag{2.38}$$

In particular, for estimating a real-valued W *with respect to* $\ell(w, \widehat{w}) = |w - \widehat{w}|$,

$$R_{\mathrm{B}} \geq \sup_{P_{U|W,X}} \frac{1}{2e} 2^{-(I(W;X|U) - h(W|U))}. \tag{2.39}$$

One advantage of Theorem 2.3 is that its unconditional version can yield tighter Bayes risk lower bounds than the unconditional version of Theorem 2.1. For example, consider the case where W is uniformly distributed on $[0, 1]$, and is estimated based on X with respect to the absolute distortion. Setting $g(\rho) = 2\rho$ in Remark 2.1 and optimizing s in (2.11), the unconditional version of Theorem 2.1 yields an asymptotic lower bound

$$R_{\mathrm{B}} \gtrsim \frac{1}{8I(W;X)} 2^{-I(W;X)} \qquad \text{as } I(W;X) \to \infty. \tag{2.40}$$

By contrast, the unconditional version of Theorem 2.3 yields a tighter and non-asymptotic lower bound

$$R_{\mathrm{B}} \geq \frac{1}{2e} 2^{-I(W;X)}. \tag{2.41}$$

2.3 Mutual Information Contraction via SDPI

While the results in Sec. 2.2 all apply to general estimation problems, either centralized or decentralized, the results in terms of mutual information (Theorems 2.1 and 2.3) are particularly amenable to tightening in the context of the decentralized estimation. For example, Theorem 2.1 reveals two sources of the difficulty of estimating W: the spread of the prior distribution P_W or its conditional counterpart $P_{W|U}$, captured by $\mathcal{L}_{\mathcal{W}}$ or $\mathcal{L}_{W|U}$, and the amount of information about W contained in the sample X, captured by $I(W;X)$ or $I(W;X|U)$. When an estimator does not have direct access to X, but can only receive information about it from one or more local processors, the amount of information about W contained in the estimator's indirect observations will contract relative to $I(W;X)$ or $I(W;X|U)$. The contraction is

caused by the communication constraints between the local processors and the estimator, such as finite precision of analog-to-digital conversion, storage limitations of intermediate results, limited transmission blocklength, channel noise, etc.

We will quantify this contraction of mutual information through strong data processing inequalities, or SDPI's, for the relative entropy (see [18] and references therein). Given a stochastic kernel (or channel) K with input alphabet X and output alphabet Y, and a reference input distribution μ on X, we say that K satisfies an SDPI at μ with constant $c \in [0,1)$ if $D(\nu K \| \mu K) \leq cD(\nu\|\mu)$ for any other input distribution ν on X. Here, μK denotes the marginal distribution of the channel output when the input has distribution μ. The *SDPI constant of K at input distribution μ* is defined as

$$\eta(\mu, K) \triangleq \sup_{\nu:\, \nu \neq \mu} \frac{D(\nu K \| \mu K)}{D(\nu\|\mu)}, \tag{2.42}$$

while the *SDPI constant of K* is defined as

$$\eta(K) \triangleq \sup_{\mu} \eta(\mu, K). \tag{2.43}$$

It is shown in [19] that the SDPI constants are also the maximum contraction ratios of mutual information in a Markov chain: for a Markov chain

$$W - X - Y,$$

we have

$$\sup_{P_{W|X}} \frac{I(W;Y)}{I(W;X)} = \eta(P_X, P_{Y|X}) \tag{2.44}$$

if the joint distribution $P_{X,Y}$ is fixed, and

$$\sup_{P_{W,X}} \frac{I(W;Y)}{I(W;X)} = \eta(P_{Y|X}) \tag{2.45}$$

if only the channel $P_{Y|X}$ is fixed. This fact leads to the following chain of inequalities for mutual information:

$$I(W;Y) \leq I(W;X)\eta(P_X, P_{Y|X}) \leq I(W;X)\eta(P_{Y|X}). \tag{2.46}$$

This is a stronger result than the ordinary data processing inequality for mutual information, as it quantitatively captures the amount by which the mutual information contracts after passing through a channel.

It is generally hard to compute the SDPI constant for an arbitrary pair of μ and K, except for some special cases:

- For the binary symmetric channel, $\eta(\mathrm{Bern}(\frac{1}{2}), \mathrm{BSC}(\varepsilon)) = \eta(\mathrm{BSC}(\varepsilon)) = (1-2\varepsilon)^2$ [20].
- For the binary erasure channel, $\eta(\mathrm{Bern}(\frac{1}{2}), \mathrm{BEC}(\varepsilon)) = \eta(\mathrm{BEC}(\varepsilon)) = 1-\varepsilon$.
- If X and Y are jointly Gaussian with correlation coefficient $\rho_{X,Y}$, then [21]

$$\eta(P_X, P_{Y|X}) = \rho_{X,Y}^2. \tag{2.47}$$

In the remainder of this section, we collect a few upper bounds and properties of the SDPI constants, which will be used in the following sections. The first upper bound is due to Cohen et al. [22]:

Lemma 2.3. *Define the* Dobrushin contraction coefficient *of a stochastic kernel $P_{X|W}$ by*

$$\vartheta(P_{X|W}) = \max_{w,w'} \|P_{X|W=w} - P_{X|W=w'}\|_{\mathrm{TV}}. \tag{2.48}$$

Then

$$\eta(P_{X|W}) \le \vartheta(P_{X|W}). \tag{2.49}$$

The next upper bound is proved in [18, Remark 3.2] for arbitrary f-divergences:

Lemma 2.4. *Suppose there exist a constant $\alpha \in (0,1]$ and a distribution Q_X, such that*[1]

$$\frac{\mathrm{d}P_{X|W=w}}{\mathrm{d}Q_X}(x) \ge \alpha \qquad \text{for all } x \in \mathsf{X} \text{ and } w \in \mathsf{W}. \tag{2.50}$$

[1] In Markov chain theory, this is known as a Doeblin minorization condition.

Then

$$\eta(P_{X|W}) \le 1-\alpha. \tag{2.51}$$

Lemma 2.4 leads to the following property:

Lemma 2.5. *For a joint distribution $P_{W,X}$, suppose there is a constant $\alpha \in (0,1]$ such that the forward channel $P_{X|W}$ satisfies*

$$\frac{\mathrm{d}P_{X|W=w}}{\mathrm{d}P_{X|W=w'}}(x) \ge \alpha \qquad \text{for all } x \in \mathsf{X} \text{ and } w, w' \in \mathsf{W}. \tag{2.52}$$

Then the SDPI constants of the forward channel $P_{X|W}$ and the backward channel $P_{W|X}$ satisfy

$$\eta(P_{X|W}) \le 1-\alpha \quad \text{and} \quad \eta(P_{W|X}) \le 1-\alpha. \tag{2.53}$$

Proof. To prove the claim for the forward channel, pick any $w' \in \mathsf{W}$ and let $Q_X = P_{X|W=w'}$. Then the condition in Lemma 2.4 is satisfied with this Q_X. To prove the claim for the backward channel, consider any $x \in \mathsf{X}$ and $w \in \mathsf{W}$. Then

$$\begin{aligned}
\frac{\mathrm{d}P_{W|X=x}}{\mathrm{d}P_W}(w) &= \frac{\mathrm{d}P_{X|W=w}}{\mathrm{d}\int P_{X|W=w'}P_W(\mathrm{d}w')}(x) && (2.54)\\
&= \frac{1}{\int \frac{\mathrm{d}P_{X|W=w'}}{\mathrm{d}P_{X|W=w}}(x) P_W(\mathrm{d}w')} && (2.55)\\
&\ge \frac{1}{\frac{1}{\alpha}\int P_W(\mathrm{d}w')} && (2.56)\\
&= \alpha, && (2.57)
\end{aligned}$$

where (2.56) uses the fact that $\frac{\mathrm{d}P_{X|W=w'}}{\mathrm{d}P_{X|W=w}}(x) \le 1/\alpha$, due to the assumption in (2.52). Using Lemma 2.4 with $Q_W = P_W$, we get the result. □

In decentralized estimation, we will encounter the SDPI constant $\eta(P_{X^n}, P_{W|X^n})$. The following lemma gives an upper bound for this SDPI constant, which is often easier to compute:

Lemma 2.6. *If $W - Z - X^n$ form a Markov chain, then*

$$\eta(P_{X^n}, P_{W|X^n}) \le \eta(P_Z, P_{W|Z}). \tag{2.58}$$

In particular, Z can be any sufficient statistic of X^n for estimating W.

Proof. It suffices to show that for any Y such that $W - X^n - Y$ form a Markov chain,

$$I(W;Y) \le \eta(P_Z, P_{W|Z}) I(X^n;Y). \tag{2.59}$$

Indeed, by the definition of $\eta(P_Z, P_{W|Z})$ and the fact that $W - Z - X^n - Y$ form a Markov chain,

$$\begin{aligned} I(W;Y) &\le \eta(P_Z, P_{W|Z}) I(Z;Y) && (2.60)\\ &\le \eta(P_Z, P_{W|Z}) I(X^n;Y), && (2.61) \end{aligned}$$

which proves (2.59) and the lemma. □

We will often need a conditional version of the SDPI:

Lemma 2.7. *For any Markov chain $W, V \to X \to Y$ with $P_{Y|X} = K$,*

$$I(W;Y|V) \le \eta(K) I(W;X|V).$$

Proof. For binary channels, this result was first proved by Evans and Schulman [23, Corollary 1]. Here, we give a general proof. For each $v \in \mathsf{V}$ we have the Markov chain $W \to X \to Y$ conditional on $V = v$, hence

$$\begin{aligned} I(W;Y|V=v) &\le \eta(P_{Y|X,V=v}) I(W;X|V=v) \\ &= \eta(K) I(W;X|V=v). \end{aligned}$$

Taking expectation with respect to V on both sides, we get $I(W;Y|V) \le \eta(K) I(W;X|V)$. □

For product input distributions and product channels, the SDPI constant tensorizes [19] (see [18] for a more general result for other f-divergences):

Lemma 2.8. *For distributions $\mu_1, \ldots, \mu_m$ on X and channels $K_1, \ldots, K_m$ with input alphabet X,*

$$\eta(\mu_1 \otimes \ldots \otimes \mu_m, K_1 \otimes \ldots \otimes K_m) = \max_{1 \le i \le m} \eta(\mu_i, K_i).$$

Finally, the following lemma that gives an SDPI for multiple uses of a channel. It is a special case of Polyanskiy and Wu [24, Corollary 2], obtained using the method of Evans and Schulman [23]. We give the proof, since we adapt the underlying technique at several points in this chapter as well as in Chapter 3.

Lemma 2.9. *Consider sending a message Y through T uses of a memoryless channel $P_{V|U}$ with feedback, where $U_t = \varphi(Y, V^{t-1}, t)$ with some encoder φ for $t = 1, \dots, T$. Then for any random variable W such that $W - Y - U^T, V^T$ form a Markov chain,*

$$I(W; V^T) \le I(W; Y)\big(1 - (1 - \eta(P_{V|U}))^T\big). \tag{2.62}$$

In particular, the result holds when the channel is used T times without feedback.

Proof. Let $\eta = \eta(P_{V|U})$. Then

$$\begin{aligned} I(W; V^T) &= I(W; V^{T-1}) + I(W; V_T | V^{T-1}) && (2.63)\\ &\le I(W; V^{T-1}) + \eta I(W; U_T | V^{T-1}) && (2.64)\\ &= (1-\eta) I(W; V^{T-1}) + \eta I(W; V^{T-1}, U_T) && (2.65)\\ &\le (1-\eta) I(W; V^{T-1}) + \eta I(W; Y), && (2.66) \end{aligned}$$

where (2.64) follows from the Markov chain $W, V^{T-1} - U_T - V_T$ and the conditional version of SDPI (Lemma 2.7); (2.66) follows from the Markov chain $W - Y - V^{T-1}, U_T$. Unrolling the above recursive upper bound on $I(W; V^T)$ and noting that $I(W; V_1) \le \eta I(W; Y)$, we get (2.62). □

Using the same proof technique, we can obtain an upper bound for the SDPI constant of a product channel:

Lemma 2.10. *For a product channel $K = \bigotimes_{i=1}^m K_i$, if the constituent channels satisfy $\eta(K_i) \le \eta$ for $i \in \{1, \dots, m\}$, then*

$$\eta(K) \le 1 - (1 - \eta)^m. \tag{2.67}$$

2.4 Decentralized Estimation: Single Processor

We start the discussion of decentralized estimation with the single-processor setup. Consider the following decentralized estimation problem with one local processor, shown schematically in Fig. 2.1:

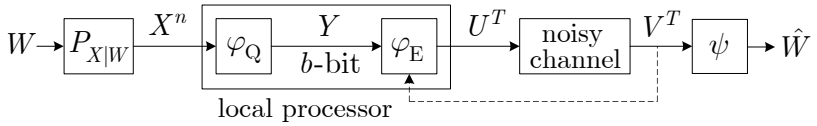


Figure 2.1: Model of decentralized estimation (single processor).

- W is an unknown parameter (discrete or continuous, scalar or vector) with prior distribution P_W.
- Conditional on $W = w$, n samples $X^n = (X_1, \ldots, X_n)$ are independently drawn from the distribution $P_{X|W=w}$.
- The local processor observes X^n and maps it to a b-bit message $Y = \varphi_{\mathrm{Q}}(X^n)$.
- The encoder maps Y to a codeword $U^T = \varphi_{\mathrm{E}}(Y)$ with blocklength T, and transmits U^T over a discrete memoryless channel (DMC) $P_{V|U}$. We allow the possibility of feedback from the estimator to the processor, in which case $U_t = \varphi_{\mathrm{E}}(Y, V^{t-1}, t)$, $t = 1, \ldots, T$.
- The estimator computes $\widehat{W} = \psi(V^T)$ as an estimate of W, based on the received codeword V^T.

The Bayes risk in the single processor setup is defined as

$$R_{\mathrm{B}} = \inf_{\varphi_{\mathrm{Q}}, \varphi_{\mathrm{E}}, \psi} \mathbb{E}\big[\ell(W, \psi(V^T))\big], \tag{2.68}$$

which depends on the problem specification including $P_{W,X}$, ℓ, n, b, T, and $P_{V|U}$. We can use the unconditional versions of Theorems 2.1 and 2.3 to obtain lower bounds for R_{B}, by replacing $I(W;X)$ with $I(W;V^T)$. To reveal the dependence of R_{B} on various problem specifications, we need an upper bound on $I(W;V^T)$ which is independent of φ_{Q} and φ_{E}:

Theorem 2.4. *In decentralized estimation with a single processor, for any choice of φ_{Q} and φ_{E},*

$$I(W;V^T) \le \min\Big\{ I(W;X^n)\eta_T,\ \ \eta(P_{X^n},P_{W|X^n})\,(H(X^n)\wedge b)\,\eta_T, \\ \eta(P_{X^n},P_{W|X^n})CT\Big\}, \tag{2.69}$$

where C is the Shannon capacity of the channel $P_{V|U}$, and

$$\eta_T = \begin{cases} 1-(1-\eta(P_{V|U}))^T & \textit{with feedback} \\ \eta(P_{V|U}^{\otimes T}) & \textit{without feedback} \end{cases}. \tag{2.70}$$

Proof. When the channel is used with feedback, the problem setup gives rise to the Markov chain $W - X^n - Y - U^T, V^T$. With $\eta_T = 1-(1-\eta(P_{V|U}))^T$, as a consequence of Lemma 2.9, we have

$$I(W;V^T) \le I(W;Y)\eta_T \le I(W;X^n)\eta_T. \tag{2.71}$$

Alternatively,

$$\begin{aligned} I(W;V^T) &\le I(W;Y)\eta_T && (2.72)\\ &\le \eta(P_{X^n},P_{W|X^n})I(X^n;Y)\eta_T && (2.73)\\ &\le \eta(P_{X^n},P_{W|X^n})(H(X^n)\wedge b)\eta_T, && (2.74) \end{aligned}$$

where (2.73) is from the SDPI in (2.46); (2.74) is because $I(X^n;Y) \le \min\{H(X^n),H(Y)\}$ and $Y \in [2^b]$. Lastly, from the SDPI and following the proof that feedback does not increase the capacity of a discrete memoryless channel [25],

$$I(W;V^T) \le \eta(P_{X^n},P_{W|X^n})I(Y;V^T) \le \eta(P_{X^n},P_{W|X^n})CT.$$

We complete the proof for the case with feedback by taking the minimum of the three resulting estimates to get the tightest bound on $I(W;V^T)$.

When the channel is used without feedback, we have the Markov chain $W - X^n - Y - U^T - V^T$. In this case, (2.71) holds with $\eta_T = \eta(P_{V|U}^{\otimes T})$ as a consequence of the SDPI. The rest of the proof for this case is the same as the case with feedback. □

Note that, with the ordinary data processing equality, we can only get the upper bound

$$I(W;V^T) \le \min\Big\{I(W;X^n),\, H(X^n)\wedge b,\, CT\Big\}, \tag{2.75}$$

where the first term reflects the statistical constraint due to the finite number of samples, the second term reflects the communication constraint due to the quantization, and the third term reflects the communication constraint due to the noisy channel. All of these terms are tightened in (2.69) via the multiplication by various contraction coefficients. Thus, using the SDPI, we can tighten the results of Theorems 2.1 and 2.3 in the setting of decentralized estimation by quantifying the communication constraint, and by coupling the statistical constraint and the communication constraint together.

Next we study a few examples of this problem setup to illustrate the effectiveness of using Theorem 2.4 to derive lower bounds on the Bayes risk.

2.4.1 Transmitting a Bit over a BSC

Example 2.4. *Consider the case where the parameter takes values* 0 *and* 1 *with equal probabilities, the local processor directly observes* W*, and communicates the value of* W *to the estimator through* T *uses of the channel* $\mathrm{BSC}(\varepsilon)$*. Formally,* W *is* $\mathrm{Bern}(\frac{1}{2})$*,* $W = X^n = Y$*, and* $P_{V|U} = \mathrm{BSC}(\varepsilon)$*. The Bayes risk is defined as* $R_{\mathrm{B}} = \inf_{\varphi_{\mathrm{E}},\psi} \mathbb{P}[W \neq \widehat{W}]$*.*

In this simple example, there is no statistical constraint since W can be directly observed by the local processor, while the communication constraint is imposed by the T uses of a BSC. Using Theorem 2.4, we can derive lower bounds on R_{B} and obtain upper bounds on the error exponent when the channel is used with or without feedback:

Corollary 2.4. *In Example 2.4, if the channel is used without feedback, then*

$$R_{\mathrm{B}} \ge h_2^{-1}\left(\frac{1}{\sqrt{2T}}(4\varepsilon(1-\varepsilon))^{\frac{T}{2}}\right), \tag{2.76}$$

and

$$\limsup_{T\to\infty} -\frac{1}{T}\log R_{\mathrm{B}} \le \frac{1}{2}\log\frac{1}{4\varepsilon(1-\varepsilon)}. \tag{2.77}$$

If the channel is used with feedback, then

$$R_{\mathrm{B}} \geq h_2^{-1}\left((4\varepsilon(1-\varepsilon))^T\right), \tag{2.78}$$

and

$$\limsup_{T\to\infty} -\frac{1}{T}\log R_{\mathrm{B}} \leq \log\frac{1}{4\varepsilon(1-\varepsilon)}. \tag{2.79}$$

Proof. Choose the φ_{E} and ψ that attain R_{B}. In this case, we can bypass Theorem 2.1 by using the binary-alphabet version of Fano's inequality:

$$1 - h_2(\mathbb{P}[\widehat{W} \neq W]) \leq I(W;V^T). \tag{2.80}$$

If the channel is used without feedback, it follows from Theorem 2.4 and Lemma 2.3 that

$$\begin{aligned} I(W;V^T) &\leq I(W;X^n)\eta\big(\mathrm{BSC}(\varepsilon)^{\otimes T}\big) \\ &\leq \vartheta\big(\mathrm{BSC}(\varepsilon)^{\otimes T}\big) \\ &\leq 1 - \frac{1}{\sqrt{2T}}(4\varepsilon(1-\varepsilon))^{T/2}, \end{aligned} \tag{2.81}$$

where the upper bound on $\vartheta\big(\mathrm{BSC}(\varepsilon)^{\otimes T}\big)$ is evaluated in [24]. Combining (2.80) and (2.81), and using the fact that [26, Theorem 2.2]

$$h_2^{-1}(x) \geq \frac{x}{2\log(6/x)} \qquad \text{for } x \in [0,1], \tag{2.82}$$

we obtain (2.76) and (2.77).

If the channel is used with feedback, Theorem 2.4 gives

$$I(W;V^T) \leq I(W;X^n)\big(1-(1-\eta(\mathrm{BSC}(\varepsilon))^T\big) \leq 1-(4\varepsilon(1-\varepsilon))^T, \tag{2.83}$$

where we used the fact that $\eta(\mathrm{BSC}(\varepsilon)) = (1-2\varepsilon)^2$. Combining (2.80), (2.82), and (2.83), we obtain (2.78) and (2.79). □

Using the Chernoff bound, it can be shown that a blocklength-T repetition code without feedback can achieve $\mathbb{P}[\widehat{W} \neq W] \leq (4\varepsilon(1-\varepsilon))^{-T/2}$ [27]. Thus,

when the channel is used without feedback,

$$\liminf_{T\to\infty} -\frac{1}{T}\log R_{\mathrm{B}} \geq \frac{1}{2}\log\frac{1}{4\varepsilon(1-\varepsilon)}, \tag{2.84}$$

which matches the upper bound on the error exponent given by (2.77). Therefore, Theorem 2.4 can effectively capture the communication constraint in this example.

2.4.2 Estimating a Discrete Parameter

Example 2.5. *Consider the case where W is uniformly distributed on $\{-1,1\}^d$. The sample $X \in \{-1,1\}^d$ is generated conditionally on W as follows. For $j = 1,\ldots,d$, given $W_j = w_j$, the jth coordinate of X, denoted by X_j, is independently drawn from the distribution $P_{X_j|W_j=w_j}(x_j) = (1 + x_j w_j \delta)/2$ for some $\delta \in [0,1]$. In other words, $P_{X_j|W_j}$ is $\mathrm{BSC}(\frac{1-\delta}{2})$. It follows that X_j is uniformly distributed on $\{-1,1\}$, and $P_{W_j|X_j}$ is $\mathrm{BSC}(\frac{1-\delta}{2})$ as well. The communication channel $P_{V|U}$ is assumed to be an arbitrary DMC.*

Theorem 2.4 gives the following upper bound on $I(W;V^T)$ for this example:

Corollary 2.5. *In Example 2.5,*

$$I(W;V^T) \leq \min\Big\{d\big(1-h_2\big(\tfrac{1-\delta}{2}\big)\big)\eta_T,\ \delta^2 b\eta_T,\ \delta^2 CT\Big\}. \tag{2.85}$$

Proof. Since $(W_1,X_1),\ldots,(W_d,X_d)$ are independent in this case, we can apply the tensorization property of the SDPI constant (Lemma 2.8), which states that

$$\eta(P_X, P_{W|X}) = \max_{1\leq j\leq d} \eta(P_{X_j}, P_{W_j|X_j}). \tag{2.86}$$

Due to the fact that X_j is uniform on $\{-1,1\}$ and $P_{W_j|X_j} = \mathrm{BSC}(\frac{1-\delta}{2})$, we have the exact SDPI constant

$$\eta(P_{X_j}, P_{W_j|X_j}) = \delta^2. \tag{2.87}$$

We also have $I(W;X) = d\big(1-h_2\big(\frac{1-\delta}{2}\big)\big)$. The results then follow from Theorem 2.4. □

The same problem with a noiseless communication channel was considered in [2]. The result in [2, Lemma 3], proved in a much more complicated way, shows that

$$I(W;Y) \le \frac{32\delta^2(d \wedge b)}{(1-\delta)^4} \tag{2.88}$$

where the contraction coefficient is less than 1 only when $\delta < 0.133$. By contrast, the contraction coefficient in (2.85) never exceeds 1. Moreover, since $1 - h_2\big(\frac{1-\delta}{2}\big) \le \delta^2$, the upper bound in (2.85) is a considerable improvement on the one in (2.88) over all $\delta \in [0,1]$, especially for large δ, under the same noiseless channel assumption ($\eta_T = 1$). Corollary 2.5 can also be used to derive lower bounds on the minimax risk of estimating the mean of an arbitrary probability distribution on the cube $[-1,1]^d$. We discuss this application in Sec. 2.5.1, in the multi-processor setup.

From another point of view, Example 2.5 is essentially a problem of noisy lossy source coding [28] of an i.i.d. Bern($\frac{1}{2}$) source of length d observed through a BSC($\frac{1-\delta}{2}$), with an additional challenge of sending the quantized message over T uses of another noisy channel. Using Corollary 2.5, we can obtain lower bounds on the average bit error probability for estimating the source W and on the quantization rate of the sample X:

Corollary 2.6. *In Example 2.5, let* $\ell(w,\widehat{w}) = \frac{1}{d}\sum_{j=1}^{d} \mathbf{1}\{w_j \neq \widehat{w}_j\}$. *Then,*

$$R_{\mathrm{B}} \ge h_2^{-1}\left(1 - \frac{1}{d}\min\left\{d\big(1 - h_2\big(\tfrac{1-\delta}{2}\big)\big)\eta_T,\ \delta^2 b\eta_T,\ \delta^2 CT\right\}\right), \tag{2.89}$$

provided b, d, *and* T *are such that the argument of* $h_2^{-1}(\cdot)$ *lies in* $[0,1]$. *Moreover, to achieve* $R_{\mathrm{B}} \le p$, *it is necessary that*

$$\frac{b}{d} \ge \frac{1 - h_2(p)}{\delta^2 \eta_T}, \tag{2.90}$$

where $\eta_T = 1 - (1 - \eta(P_{V|U}))^T$.

Proof. Choose the φ_{Q}, φ_{E}, and ψ that attain R_{B}. In this case, we can again bypass Theorem 2.1 by using the following chain of inequalities to relate the

average bit error probability with $I(W;V^T)$:

$$1 - h_2(R_{\mathrm{B}}) = d_2(R_{\mathrm{B}} \| \tfrac{1}{2}) \tag{2.91}$$

$$\leq \frac{1}{d} \sum_{j=1}^{d} d_2\big(\mathbb{P}[W_j \neq \widehat{W}_j] \| \tfrac{1}{2}\big) \tag{2.92}$$

$$\leq \frac{1}{d} \sum_{j=1}^{d} I(W_j; \widehat{W}_j) \tag{2.93}$$

$$\leq \frac{1}{d} I(W; V^T) \tag{2.94}$$

$$\leq \frac{1}{d} \min \Big\{ d\big(1 - h_2\big(\tfrac{1-\delta}{2}\big)\big)\eta_T,\, \delta^2 b \eta_T,\, \delta^2 CT \Big\}, \tag{2.95}$$

where (2.92) uses the fact that $R_{\mathrm{B}} = \frac{1}{d}\sum_{j=1}^{d} \mathbb{P}[W_j \neq \widehat{W}_j]$ and the convexity of divergence; (2.93) uses the fact that W_j is uniform on $\{-1, 1\}$ and the data processing inequality for divergence; (2.94) uses the fact that W_j's are independent; (2.95) follows from Corollary 2.5. Applying h_2^{-1} to both sides, we get (2.89). The lower bound (2.90) is a consequence of (2.95). □

The asymptotic rate limit of noisy lossy coding of an i.i.d. Bern($\frac{1}{2}$) source observed through a BSC($\frac{1-\delta}{2}$) with distortion p is given by

$$\tilde{R}(p) = 1 - h_2\left(\frac{2p + \delta - 1}{2\delta}\right), \qquad 0 \leq \frac{1-\delta}{2} \leq p \leq \frac{1}{2}. \tag{2.96}$$

In Fig. 2.2, the lower bounds on the quantization rate given by (2.90) with different values of $\eta(P_{V|U})$ are compared with $\tilde{R}(p)$. The lower bounds are also compared with the rate-distortion function of an i.i.d. Bern($\frac{1}{2}$) source, given by

$$R(p) = 1 - h_2(p), \qquad 0 \leq p \leq \frac{1}{2}. \tag{2.97}$$

We can see that with $\eta(P_{V|U}) = 1$, the lower bound well matches the asymptotically achievable rate given by (2.96) for large δ. With $\eta(P_{V|U}) < 1$, the elevated lower bounds capture the need to increase the quantization rate for sending the quantized message through another noisy channel.

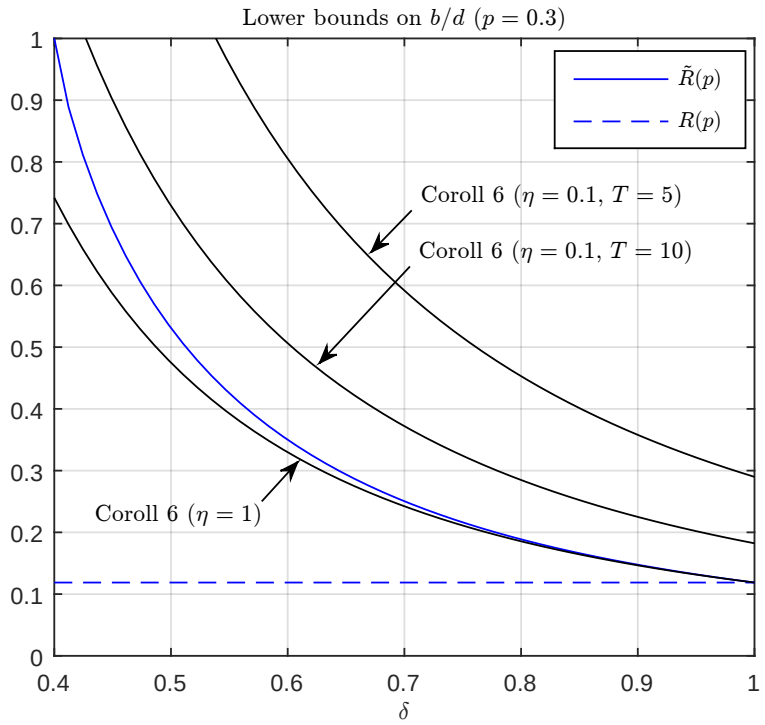


Figure 2.2: Comparison of lower bounds on b/d, where $p = 0.3$ and $\eta = \eta(P_{V|U})$.

2.4.3 Estimating a Continuous Parameter

Example 2.6. *Consider the problem of estimating the bias of a Bernoulli random variable through a BSC. In this case, W is assumed to be uniformly distributed on $[0, 1]$, $P_{X|W=w}$ is $\mathrm{Bern}(w)$, and $P_{V|U}$ is $\mathrm{BSC}(\varepsilon)$. We are interested in lower-bounding the Bayes risk with respect to the absolute distortion $\ell(w, \widehat{w}) = |w - \widehat{w}|$.*

Define $I^* = \sup_{\varphi_{\mathrm{Q}}, \varphi_{\mathrm{E}}} I(W; V^T)$. Replacing $I(W; X)$ with I^* in (2.41), we obtain the following lower bound on the Bayes risk for this example as a consequence of Theorem 2.3:

$$R_{\mathrm{B}} \geq \frac{1}{2e} 2^{-I^*}. \tag{2.98}$$

Now we only need to upper-bound I^*:

Corollary 2.7. *In Example 2.6, for any choice of φ_{Q} and φ_{E},*

$$I(W; V^T) \leq \min\left\{\left(\frac{1}{2}\log n + \gamma_n\right)\eta_T,\ (1 - 2^{-n})b\eta_T,\ (1 - 2^{-n})(1 - h_2(\varepsilon))T\right\},$$

where γ_n is some sequence such that $\lim_{n\to\infty}\gamma_n = -0.6$, and $\eta_T = 1-(4\varepsilon(1-\varepsilon))^T$.

Proof. From (2.13),

$$I(W;X^n) = \frac{1}{2}\log\frac{n}{2\pi e} + h(W) + \frac{1}{2}\mathbb{E}\left[\log\frac{1}{W(1-W)}\right] + o(1) \tag{2.99}$$

$$= \frac{1}{2}\log n - 0.6 + o(1) \qquad \text{as } n\to\infty. \tag{2.100}$$

Moreover, from Lemma 2.3,

$$\eta(P_{W|X^n}) \le \vartheta(P_{W|X^n}) = 1 - 2^{-n}, \tag{2.101}$$

where the Dobrushin coefficient is evaluated in Sec. 2.7.4. In addition, $\eta(\mathrm{BSC}(\varepsilon)) = (1-\varepsilon)^2$. With these facts, the result follows from Theorem 2.4. □

Now we apply the above results to two special cases.

Case 1: $\varepsilon = 0$, $T \ge b$. In this case, the communication constraint only comes from the quantization of the samples, since the quantized message can be perfectly received by the estimator. Setting $b = \frac{1}{2}\log n$, the lower bound in (2.98) together with Corollary 2.7 imply that

$$R_{\mathrm{B}} \ge \frac{1}{2e}2^{-(1-2^{-n})b} \ge \frac{1}{2e\sqrt{n}}. \tag{2.102}$$

To obtain an upper bound on R_{B}, consider the scheme where the local processor computes the sample mean $\bar{X} = \frac{1}{n}\sum_{j=1}^n X_j$, which is uniformly distributed on $\{0, 1/n, \dots, 1\}$, and quantizes $\bar{X}$ into $\tilde{X}$ using a uniform b-bit quantization of $[0,1]$. The estimator sets $\widehat{W} = \tilde{X}$. By the triangle inequality,

$$\mathbb{E}|W - \widehat{W}| \le \mathbb{E}|W - \bar{X}| + \mathbb{E}|\bar{X} - \tilde{X}| \le \sqrt{\mathbb{E}[\mathrm{Var}(\bar{X}|W)]} + 2^{-b} \tag{2.103}$$

$$= \frac{1}{\sqrt{6n}} + 2^{-b}. \tag{2.104}$$

Thus for $b = \frac{1}{2}\log n$,

$$R_{\mathrm{B}} \le \frac{1.41}{\sqrt{n}}, \tag{2.105}$$

which differs from the lower bound only by a constant factor.

Case 2: $\varepsilon > 0$, $b \geq \log(n+1)$. In this case, the communication constraint only comes from the noisy channel, since $\log(n+1)$ bits are enough to perfectly represent the sample mean $\bar{X}$, which is a sufficient statistic of X^n for estimating W and can take only $n+1$ values. From (2.98) and Corollary 2.7,

$$R_{\mathrm{B}} \geq \max\left\{\frac{1}{2en^{\eta_T/2}}2^{-\gamma_n\eta_T},\ \frac{1}{2e}2^{-(1-2^{-n})(1-h_2(\varepsilon))T}\right\}. \tag{2.106}$$

To obtain an upper bound on R_{B}, consider the scheme where the local processor first uses $\log(n+1)$ bits to represent the sample mean $\bar{X}$ as a message uniformly distributed on $\{0, 1/n, \ldots, 1\}$, then transmits the message over the channel using an optimal blocklength-T code. The estimator decodes $\bar{X}$ as $\widehat{X}$, and sets $\widehat{W} = \widehat{X}$. Then

$$\begin{aligned}
\mathbb{E}|W - \widehat{W}| &\leq \mathbb{E}|W - \bar{X}| + \mathbb{E}|\bar{X} - \widehat{X}| && (2.107)\\
&\leq \frac{1}{\sqrt{6n}} + \mathbb{P}[\bar{X} \neq \widehat{X}] && (2.108)\\
&\leq \frac{1}{\sqrt{6n}} + 2^{-E_{\mathrm{r}}\left(\frac{1}{T}\log(n+1)\right)T}, && (2.109)
\end{aligned}$$

where $E_{\mathrm{r}}(\cdot)$ is the random coding error exponent of BSC(ε) [27, p. 146]. For $\frac{1}{T}\log(n+1) \leq 1 - h_2\left(\frac{\sqrt{\varepsilon}}{\sqrt{\varepsilon}+\sqrt{1-\varepsilon}}\right)$,

$$E_{\mathrm{r}}\left(\tfrac{1}{T}\log(n+1)\right) = 1 - \log(1+\sqrt{4\varepsilon(1-\varepsilon)}) - \tfrac{1}{T}\log(n+1). \tag{2.110}$$

If the channel is used with feedback, then $E_{\mathrm{r}}(\cdot)$ in (2.109) can be replaced by $E_{\mathrm{f}}(\cdot)$, the best attainable error exponent on BSC using block codes with feedback. In particular [29, Problem 10.36],

$$\lim_{R\to 0} E_{\mathrm{f}}(R) = E_{\mathrm{f}}(0) = -\log\left(\varepsilon^{1/3}(1-\varepsilon)^{2/3} + \varepsilon^{2/3}(1-\varepsilon)^{1/3}\right) > E_{\mathrm{r}}(0). \tag{2.111}$$

From the lower bound in (2.106) and the upper bound in (2.109), we know that the Bayes risk in this case decays polynomially in n and exponentially

in T. Moreover,

$$1 \leq \frac{1-h_2(\varepsilon)}{-\log\left(\varepsilon^{1/3}(1-\varepsilon)^{2/3}+\varepsilon^{2/3}(1-\varepsilon)^{1/3}\right)} \leq \frac{9}{8} \qquad \text{for } \varepsilon \in \left(\frac{2}{9}, \frac{1}{2}\right), \quad (2.112)$$

which implies that the error exponent with respect to T in the lower bound can closely match that in the upper bound when transmission rate is low and ε is relatively large.

2.5 Decentralized Estimation: Multiple Processors

We now consider the problem setup with m local processors. The ith processor, $i = 1, \ldots, m$, observes n samples $X_{(i)}^n$ generated from a common random parameter W. Given $W = w$, the joint distribution of the $m \times n$ array of samples is $P_{X_{(1)},\ldots,X_{(m)}|W=w}^{\otimes n}$. In other words, the samples across different processors can be dependent conditional on W, but, at each processor i, the samples are i.i.d. draws from $P_{X_{(i)}|W=w}$. As in the single-processor setup, the ith processor maps its samples to a b-bit message $Y_{(i)} = \varphi_{\mathrm{Q},i}(X_{(i)}^n)$, then maps the message to a blocklength-T codeword $U_{(i)}^T = \varphi_{\mathrm{E},i}(Y_{(i)})$, and sends it to the estimator via T uses of a discrete memoryless channel. The estimator computes $\widehat{W} = \psi(V^{m\times T})$ based on the received codewords $V^{m\times T} = (V_{(1)}^T, \ldots, V_{(m)}^T)$. Here we assume that the channels between the processors and the estimator are independent and have the same probability transition law $P_{V|U}$.[2] The Bayes risk in this multi-processor setup is defined as

$$R_{\mathrm{B}} = \inf_{\varphi_{\mathrm{Q}}^m, \varphi_{\mathrm{E}}^m, \psi} \mathbb{E}\big[\ell(W, \psi(V^{m\times T}))\big]. \quad (2.113)$$

Compared with the single processor setup, the multi-processor setup gives rise to some new problems:

- The sample sets observed by different processors can be either independent or dependent conditionally on W, depending on the joint distribution $P_{X_{(1)},\ldots,X_{(m)}|W=w}$. In Sec. 2.5.1, we derive lower bounds for the case where $X_{(1)}, \ldots, X_{(m)}$ are conditionally independent given W; in Sec. 2.5.2, we study the case where $X_{(1)}, \ldots, X_{(m)}$ are dependent

[2] The results can be straightforwardly generalized to the case where the parameters n, b, T, and the channels are different across the processors.

conditionally on W. We will see that the Bayes risk can behave quite differently in these two cases.

- Suppose the $m \times n$ array of samples $(X_{(1)}^n, \ldots, X_{(m)}^n)$ can be observed by a single processor, which can map the samples to an mb-bit message and use the channel mT times to send the message, and the estimation is based on the received codeword of blocklength mT. How will the estimation performance degrade once these resources are distributed into m processors in the multi-processor setup? We examine this performance degradation through the Bayes risk lower bounds, for both cases where the sample sets are conditionally independent and dependent.

- When the channels are noiseless and feedback is available from the estimator to the local processors, each processor can observe the messages sent by the other processors. This allows for interactive protocols, as studied in [2, 4, 5]. We will mainly focus on the case where the communication from local processors to the estimator is carried out without feedback, except for Sec. 2.5.3, where we consider the case where feedback is available and derive lower bounds on the Bayes risk for interactive protocols.

Before delving into various special cases, we give two general lower bounds for Bayes risk in the multi-processor setup, which are immediate consequences of Theorems 2.1 and 2.3, respectively:

Theorem 2.5. *In the multi-processor setup, the Bayes risk satisfies*

$$R_{\mathrm{B}} \geq \inf_{\varphi_{\mathrm{Q}}^m, \varphi_{\mathrm{E}}^m} \sup_{\mathcal{S} \subset [m],\, \rho > 0} \rho \left(1 - \frac{I(W; V^{m \times T} | X_{\mathcal{S}}^n) + 1}{\log(1/\mathbb{E}[\mathcal{L}_W(X_{\mathcal{S}}^n, \rho)])} \right), \tag{2.114}$$

where $X_{\mathcal{S}}^n = (X_{(i)}^n)_{i \in \mathcal{S}}$. *When* $W \in \mathbb{R}^d$ *and* $\ell(w, \widehat{w}) = \|w - \widehat{w}\|^r$ *for any norm* $\|\cdot\|$ *in* $\mathbb{R}^d$ *and any* $r \geq 1$,

$$R_{\mathrm{B}} \geq \inf_{\varphi_{\mathrm{Q}}^m, \varphi_{\mathrm{E}}^m} \sup_{\mathcal{S} \subset [m]} \frac{d}{re} \left(V_d \Gamma\left(1 + \frac{d}{r}\right) \right)^{-r/d} 2^{-(I(W; V^{m \times T} | X_{\mathcal{S}}^n) - h(W | X_{\mathcal{S}}^n))r/d}. \tag{2.115}$$

The proof of Theorem 2.5 is inspired by the proof of the Slepian-Wolf converse for distributed almost-lossless source coding using the cutset argument [25,

Chap. 15.4]: choose the auxiliary random variable $U = X_{\mathcal{S}}^n$ in Theorems 2.1 and 2.3, then optimize over $\mathcal{S}$.

2.5.1 Sample Sets Conditionally Independent Given W

We first study the case where the sets of samples observed by the processors are conditionally independent given the parameter W. In this case, we can simply choose $\mathcal{S} = \varnothing$ in Theorem 2.5 to obtain lower bounds on the Bayes risk. To that end, we need an upper bound on $I(W; V^{m\times T})$ which is independent of φ_{Q}^m and φ_{E}^m:

Theorem 2.6. *In the multi-processor setup, where the samples observed by the processors are conditionally i.i.d. given W, for any choice of φ_{Q}^m and φ_{E}^m,*

$$I(W; V^{m\times T}) \le \min\Big\{ I(W; X^{m\times n})\eta_{mT},\ \ \eta(P_{X^n}, P_{W|X^n})mb\eta_T, \\ \eta(P_{X^n}, P_{W|X^n})mCT\Big\}, \tag{2.116}$$

where $\eta_T = \eta(P_{V|U}^{\otimes T})$. The first upper bound can be replaced by $mI(W; X^n)\eta_T$.

Proof. Applying SDPI to the Markov chain $W - X^{m\times n} - U^{m\times T} - V^{m\times T}$, we get the first upper bound in (2.116). Due to the independence assumption, the codewords $V_{(1)}^T, \ldots, V_{(m)}^T$ received by the estimator are conditionally independent given W. This implies that (see, e.g., [3, Lemma 4])

$$I(W; V^{m\times T}) \le \sum_{i=1}^m I(W; V_{(i)}^T). \tag{2.117}$$

Using Theorem 2.4 to upper-bound each term, we obtain the second and the third upper bound in (2.116), as well as an alternative $mI(W; X^n)\eta_T$ to the first upper bound. □

To capture the penalty of decentralization, consider the situation where a total number of N conditionally i.i.d. samples are allocated to a single processor, which maps them to a B-bit message and uses the channel L times to send the message. In this situation, Theorem 2.4 gives the upper

bound

$$I(W;V^L) \le \min\Big\{I(W;X^N)\eta_L,\ \eta(P_{X^N},P_{W|X^N})B\eta_L,\ \eta(P_{X^N},P_{W|X^N})CL\Big\}. \tag{2.118}$$

Once these resources are evenly distributed to m processors, so that each processor observes N/m samples, maps then to a B/m-bit message, and uses the channel L/m times to send the message, Theorem 2.6 implies that

$$\begin{aligned} I(W;V^{m\times\frac{L}{m}}) \le \min\Big\{&I(W;X^N)\eta_L,\ \eta(P_{X^{N/m}},P_{W|X^{N/m}})B\eta_{L/m},\\ &\eta(P_{X^{N/m}},P_{W|X^{N/m}})CL\Big\}, \end{aligned} \tag{2.119}$$

where the first upper bound can be replaced by $mI(W;X^{N/m})\eta_{L/m}$. Comparing (2.119) with (2.118), we see that the differences are in the SDPI constants $\eta(P_{X^{N/m}},P_{W|X^{N/m}})$ and $\eta_{L/m}$. Since $W-X^n-X^k$ form a Markov chain whenever $k\le n$, Lemma 2.6 implies that $\eta(P_{X^{N/m}},P_{W|X^{N/m}})$ is decreasing in m. For example, when $W\sim N(0,\sigma_W^2)$ and $X_i=W+Z_i$ with Z_i drawn i.i.d. from $N(0,\sigma^2)$ for $i=1,\dots,n$, we have $\eta(P_{\bar X},P_{W|\bar X})=n\sigma_W^2/(n\sigma_W^2+\sigma^2)$ by (2.47). Then by Lemma 2.6

$$\eta(P_{X^{N/m}},P_{W|X^{N/m}}) \le \frac{\sigma_W^2 N/m}{\sigma_W^2 N/m+\sigma^2} \tag{2.120}$$

$$\approx \frac{N}{m}\frac{\sigma_W^2}{\sigma^2} \qquad \text{when } \frac{\sigma_W^2}{\sigma^2} \text{ is small.} \tag{2.121}$$

Moreover, from (2.67) we know that $\eta_{L/m}$ is decreasing in m as well, and

$$\eta_{L/m} \approx \frac{L}{m}\eta(P_{V|U}) \qquad \text{when } \eta(P_{V|U}) \text{ is small.} \tag{2.122}$$

Thus, when the processors observe sample sets that are conditionally independent given the parameter, the penalty of decentralization can be captured by the reduced SDPI constants. The resulting upper bound on $I(W;V^{m\times\frac{L}{m}})$ decreases as the resources are distributed to more processors.

To illustrate the effectiveness of Theorem 2.6, we first show an example of mean estimation in the d-dimensional Gaussian location model with a Gaussian prior:

Example 2.7. *Consider the decentralized estimation of* $W\sim N(0,\sigma_W^2\mathbf{I}_d)$

with m processors, where the samples are i.i.d. draws from $N(w, \sigma^2 \mathbf{I}_d)$ given $W = w$. The distortion function is $\ell(w, \widehat{w}) = \|w - \widehat{w}\|_2^2$. Suppose there are N samples in total, a budget of B bits for quantization, and L available uses of the channels. These resources are evenly distributed to the m processors.

Combining (2.119) from Theorem 2.6 and (2.115) in Theorem 2.5, we get the following Bayes risk lower bound for Example 2.7:

Corollary 2.8. *In Example 2.7, the Bayes risk satisfies*

$$R_{\mathrm{B}} \geq d\sigma_W^2 \max\left\{\left(1 + \frac{N\sigma_W^2}{\sigma^2}\right)^{-\eta_L},\ \exp\left(-\frac{N\sigma_W^2 \ln 4}{N\sigma_W^2 + m\sigma^2}\frac{(B\eta_{L/m} \wedge CL)}{d}\right)\right\}, \tag{2.123}$$

where $\eta_L = \eta(P_{V|U}^{\otimes L})$.

The first lower bound captures the increase of the Bayes risk due to the noisy communication channels, as compared to the Bayes risk $\frac{d\sigma_W^2}{1+N\sigma_W^2/\sigma^2}$ of the centralized estimation. From the second lower bound, we can see the order increase of the Bayes risk when the samples and the communication resources are distributed to more processors. When the communication channels are noiseless, the lower bound in Corollary 2.8 reduces to

$$R_{\mathrm{B}} \geq \max\left\{\frac{d\sigma_W^2}{1 + N\sigma_W^2/\sigma^2},\ d\sigma_W^2 \exp\left(-\frac{N\sigma_W^2 \ln 4}{N\sigma_W^2 + m\sigma^2}\frac{B}{d}\right)\right\}. \tag{2.124}$$

It shows that, with noiseless communication channels, in order to achieve the same performance as in the centralized scenario, the total number of bits allocated for quantization needs to be at least

$$B \geq \left(1 + \frac{m\sigma^2}{N\sigma_W^2}\right)\frac{d}{2}\log\left(1 + \frac{N\sigma_W^2}{\sigma^2}\right). \tag{2.125}$$

Note that it is necessary to have $N \geq m$, since each processor should observe at least one sample. Whether the lower bound in (2.125) is a sufficient condition for achieving the Bayes rate of centralized estimation is an open problem.

As a second example, we use Theorem 2.6 to derive lower bounds on the minimax risk for a nonparametric estimation problem studied in [2]. Here we assume that the communication channels are noisy:

Example 2.8. *Consider the decentralized estimation of the mean of an unknown distribution P on $\mathsf{X} = [-1,1]^d$, where each processor $i \in [m]$ only observes a single independent sample $X_{(i)}$ drawn from P. We use $\mathcal{P}$ to denote the family of probability distributions on $[-1,1]^d$, and define $\theta(P) = \mathbb{E}_P[X]$ for a distribution $P \in \mathcal{P}$. The minimax risk of this example is defined as*

$$R_{\mathrm{M}} = \inf_{\varphi_{\mathrm{Q}}^m, \varphi_{\mathrm{E}}^m, \psi} \sup_{P \in \mathcal{P}} \mathbb{E}_P \|\theta(P) - \psi(V^{m\times T})\|_2^2, \tag{2.126}$$

where ψ is an estimator of $\theta \in [-1,1]^d$.

Corollary 2.9. *In Example 2.8, the minimax risk satisfies*

$$R_{\mathrm{M}} > \frac{d}{5} \min\left\{1, \frac{d}{m \min\{d\eta_T,\, b\eta_T,\, CT\}}\right\}, \tag{2.127}$$

where $\eta_T = \eta(P_{V|U}^{\otimes T})$.

Proof. At a high level, the proof strategy follows that in [2] by reducing the minimax estimation problem to the Bayes estimation problem in Example 2.5 of Sec. 2.4.2. However, here we use the result of Corollary 2.6 instead of the distance-based Fano's inequality used in [2] to obtain a tighter lower bound. The lower bound will also be able to capture the influence of noisy channels between the processors and the estimator.

Let W, δ, and $P_{X_j|W_j}$ be defined as in Example 2.5. Conditional on $W = w$, each processor observes an independent copy of X, whose coordinates are drawn according to $P_{X_j|W_j=w_j}$ for $i = 1, \ldots, d$. Hence $P_{X|W=w} \in \mathcal{P}$ for all $w \in \{-1,1\}^d$. Let $\theta_w \triangleq \theta(P_{X|W=w}) = \delta w$, then

$$\|\theta_w - \theta_{w'}\|^2 = 4\delta^2 \ell_{\mathrm{H}}(w, w'), \tag{2.128}$$

where ℓ_{H} denotes the Hamming distance. Define

$$R_{\mathrm{B}} = \inf_{\varphi_1^m, \varphi_2^m} \inf_{\psi} \mathbb{E}[\ell_{\mathrm{H}}(W, \widehat{W})], \tag{2.129}$$

where the second infimum is over all estimators of $W \sim \mathrm{Unif}(\{-1,+1\}^d)$. Then, for $0 \le \delta \le 1$,

$$R_{\mathrm{M}} \ge 4\delta^2 R_{\mathrm{B}}. \tag{2.130}$$

From the proof of Corollary 2.6 and Theorem 2.6, we have

$$1 - h_2(R_{\mathrm{B}}/d) \le \frac{1}{d} I(W; V^{m\times T}) \le \frac{\delta^2 m}{d} \min\Big\{d\eta_T,\, b\eta_T,\, CT\Big\}, \tag{2.131}$$

where we have replaced the first upper bound in Theorem 2.6 with $mI(W;X)\eta_T$, and used the fact that $1 - h_2((1-\delta)/2) \le \delta^2$. Thus,

$$R_{\mathrm{M}} \ge 4\delta^2 d\, h_2^{-1}\left(1 - \frac{\delta^2 m}{d} \min\{d\eta_T,\, b\eta_T,\, CT\}\right). \tag{2.132}$$

With $\delta^2 = \min\{1, d/(2m\min\{d\eta_T,\, b\eta_T,\, CT\})\}$, the quantity in the parentheses is at least $1/2$, and since $h_2^{-1}(1/2) > 1/10$, we obtain the desired result. □

When the communication channels are noiseless, Corollary 2.9 reduces to

$$R_{\mathrm{M}} > \frac{d}{5} \min\Big\{1,\, \frac{d}{m(d \wedge b)}\Big\}, \tag{2.133}$$

which recovers the lower bound in [2, Proposition 2] and improves the multiplicative constant. The lower bound can be achieved within a constant factor when $b = d$, using a method described in [2].

As the last example of this section, we apply Theorem 2.6 to the case where the parameter is a vector of length n, and each component of the sample set is generated according to the corresponding component of the parameter.

Example 2.9 (CEO problem with noisy channels)**.** *Suppose the unknown parameter now is a random sequence W^n, consisting of n i.i.d. draws from some prior distribution P_W on $\mathbb{R}^d$. $X_{(1)}, \ldots, X_{(m)}$ are assumed to be independent, but not necessarily identically distributed, conditional on W. Given $W^n = w^n$, the ith processor observes the sample set $X_{(i)}^n$, whose jth component is independently drawn from $P_{X_{(i)}|W=w_j}$, for $j = 1, \ldots, n$. The ith processor then maps $X_{(i)}^n$ to a b_i-bit message and encodes it for transmission via T uses of a noisy channel $P_{V|U}$. The estimator computes $\widehat{W}^n$ from the m received codewords as an estimate of W^n. The distortion is measured by $\frac{1}{n}\sum_{j=1}^n \|w_j - \widehat{w}_j\|^r$ with some norm $\|\cdot\|$ on $\mathbb{R}^d$ and some $r \ge 1$.*

When the channels between the processors and the estimator are noiseless, Example 2.9 coincides with the chief estimation officer (CEO) problem [30]. Courtade [31] worked out a lower bound on the sum rate of the CEO problem

using SDPI. The following result is an extension of the result in [31] to the case where the channels between the processors and the estimator are noisy:

Corollary 2.10. *For the CEO problem with noisy channels in Example 2.9, if* $\frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\|W_j-\widehat{W}_j\|^r\le\alpha$, *then the quantization rates* b_i/n, $i=1,\ldots,m$, *need to satisfy*

$$\sum_{i=1}^{m}\frac{b_i}{n}\eta(P_{X_{(i)}},P_{W|X_{(i)}})\eta_T\ge h(W)-\log\left(V_d\Big(\frac{\alpha re}{d}\Big)^{d/r}\Gamma\Big(1+\frac{d}{r}\Big)\right),\quad(2.134)$$

where $\eta_T=\eta(P_{V|U}^{\otimes T})$.

Proof. Since $X_{(1)}^n,\ldots,X_{(m)}^n$ are conditionally independent given W^n, Theorem 2.6 gives

$$I(W^n;\widehat{W}^n)\le\sum_{i=1}^{m}b_i\eta(P_{X_{(i)}^n},P_{W^n|X_{(i)}^n})\eta_T\quad(2.135)$$

$$=\sum_{i=1}^{m}b_i\eta(P_{X_{(i)}},P_{W|X_{(i)}})\eta_T,\quad(2.136)$$

where the second step follows from the independence among $(W_j,X_{(i),j})$'s for each fixed $i=1,\ldots,m$, and the tensorization property of the SDPI constant (Lemma 2.8).

Now define

$$R_W(\alpha)=\inf_{P_{\widehat{W}|W}:\,\mathbb{E}\|W-\widehat{W}\|^r\le\alpha}I(W;\widehat{W})$$

and

$$R_{W^n}(\alpha)=\inf_{P_{\widehat{W}^n|W^n}:\,\frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\|W_j-\widehat{W}_j\|^r\le\alpha}I(W^n;\widehat{W}^n)$$

to be the rate-distortion functions of W and W^n, respectively. We have

$$I(W^n;\widehat{W}^n)\ge R_{W^n}(\alpha)\quad(2.137)$$

$$=nR_W(\alpha)\quad(2.138)$$

$$\ge n\left(h(W)-\log\left(V_d\Big(\frac{\alpha re}{d}\Big)^{d/r}\Gamma\Big(1+\frac{d}{r}\Big)\right)\right),\quad(2.139)$$

where (2.137) is because of the assumption that $\frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\|W_j-\widehat{W}_j\|^r\le\alpha$; (2.138) uses the additivity property of the rate-distortion function under additive distortions; and (2.139) is a consequence of (2.36). The proof of (2.134)

is completed by combining the upper and lower bounds on $I(W^n; \widehat{W}^n)$. $\square$

2.5.2 Sample Sets Conditionally Dependent Given W

Now we consider the situation where the processors observe dependent sample sets conditional on the parameter. To obtain tight Bayes risk lower bounds, we need to choose a suitable conditioning subset $\mathcal{S}$ in Theorem 2.5. Once $\mathcal{S}$ is chosen, we need to evaluate or upper-bound the expected conditional small ball probability $\mathbb{E}[\mathcal{L}_W(X_{\mathcal{S}}^n, \rho)]$ or the conditional differential entropy $h(W|X_{\mathcal{S}}^n)$. We also need to upper-bound $I(W; V^{m\times T}|X_{\mathcal{S}}^n)$ regardless of the choice of φ_{Q}^m and φ_{E}^m . Here we give a general upper bound on $I(W; V^{m\times T}|X_{\mathcal{S}}^n)$, which holds regardless of whether or not the sample sets are conditionally independent given W:

Theorem 2.7. *In the multi-processor setup, for any choice of φ_{Q}^m and φ_{E}^m, and for any $\mathcal{S} \subset [m]$,*

$$I(W; V^{m\times T}|X_{\mathcal{S}}^n) \le \min\Big\{I(W; X_{\mathcal{S}^c}^n|X_{\mathcal{S}}^n)\eta_{|\mathcal{S}^c|T},\ \eta(\mathcal{S})|\mathcal{S}^c|b\eta_{|\mathcal{S}^c|T},\ \eta(\mathcal{S})|\mathcal{S}^c|CT\Big\}, \tag{2.140}$$

where $\mathcal{S}^c = [m] \setminus \mathcal{S}$, $\eta_{|\mathcal{S}^c|T} = \eta\big(P_{V|U}^{\otimes|\mathcal{S}^c|T}\big)$, and

$$\eta(\mathcal{S}) = \sup_{x_{\mathcal{S}}^n} \eta\big(P_{X_{\mathcal{S}^c}^n|X_{\mathcal{S}}^n=x_{\mathcal{S}}^n}, P_{W|X_{\mathcal{S}^c}^n, X_{\mathcal{S}}^n=x_{\mathcal{S}}^n}\big). \tag{2.141}$$

In particular, when the channels are noiseless, we have

$$I(W; V^{m\times T}|X_{\mathcal{S}}^n) \le \min\Big\{I(W; X_{\mathcal{S}^c}^n|X_{\mathcal{S}}^n),\ \eta(\mathcal{S})|\mathcal{S}^c|b\Big\}. \tag{2.142}$$

Proof. Section 2.7.5. $\square$

Theorem 2.7 can be used to capture the penalty of decentralization when the sample sets are conditionally dependent. Consider the situation where all of the m sample sets $X_{(1)}^n, \ldots, X_{(m)}^n$ are observed by a single processor, which maps them to an mb-bit message and uses the channel mT times to

send the message. In this situation, we have the upper bound

$$I(W;V^{mT}|X_{\mathcal{S}}^n) \le \min\left\{I(W;X_{\mathcal{S}^c}^n|X_{\mathcal{S}}^n)\eta_{mT},\ \eta(\mathcal{S})mb\eta_{mT},\ \eta(\mathcal{S})mCT\right\} \tag{2.143}$$

(see Section 2.7.5 for the proof). In particular, when the channels are noiseless, we have

$$I(W;V^{mT}|X_{\mathcal{S}}^n) \le \min\left\{I(W;X_{\mathcal{S}^c}^n|X_{\mathcal{S}}^n),\ \eta(\mathcal{S})mb\right\}. \tag{2.144}$$

Comparing (2.140) with (2.143), we can see that, when the sample sets are dependent conditionally on W, the penalty of decentralization can still be captured by the reduced upper bound on $I(W;V^{mT}|X_{\mathcal{S}}^n)$. In particular, when the channels are noiseless, for a fixed $\mathcal{S}$, the second upper bound in (2.142) is only a $\frac{m-|\mathcal{S}|}{m}$ fraction of the second upper bound in (2.144). However, this does not mean that choosing $\mathcal{S}$ as large as possible leads to the tightest lower bound on the Bayes risk. The reason is that a larger $\mathcal{S}$ generally corresponds to a larger $\mathbb{E}[\mathcal{L}_W(X_{\mathcal{S}}^n,\rho)]$ or a smaller $h(W|X_{\mathcal{S}}^n)$, which may offset the decrease of the upper bound on $I(W;V^{mT}|X_{\mathcal{S}}^n)$. The optimal $\mathcal{S}$ to choose thus depends on the specific problem.

We study two examples to illustrate the effectiveness of combining the upper bound on $I(W;V^{mT}|X_{\mathcal{S}}^n)$ in Theorem 2.7 with the lower bounds in Theorem 2.5. For simplicity, we focus on the case where the communication channels are noiseless.

Example 2.10. *Consider a two-processor case, where* $W \sim U[0,1]$ *and* $X_1, X_2 \in \{0,1\}$. *The conditional distribution* $P_{X_{(1)},X_{(2)}|W=w}$ *is specified as* $P_{X_{(1)},X_{(2)}|W=w}(0,0) = P_{X_{(1)},X_{(2)}|W=w}(1,1) = (1-w)/2$, *and* $P_{X_{(1)},X_{(2)}|W=w}(0,1) = P_{X_{(1)},X_{(2)}|W=w}(1,0) = w/2$. *Note that* X_1 *and* X_2 *are marginally independent of* W, *but are jointly dependent on* W. *In the decentralized estimation, processor* i *observes* $X_{(i)}^n$ *and maps the samples to a* b*-bit message. The estimator computes* $\widehat{W}$ *based on the noiselessly received messages. The distortion function is* $\ell(w,\widehat{w}) = |w-\widehat{w}|$.

For this example, we can choose $\mathcal{S} = \{2\}$, then use (2.115) in Theorem 2.5 and (2.142) in Theorem 2.7 to obtain the following lower bound on the Bayes risk:

Corollary 2.11. *In Example 2.10, the Bayes risk satisfies*

$$R_B \geq \frac{1}{2e} 2^{-(1-2^{-n})b}. \tag{2.145}$$

Proof. Since $X_{(2)}^n$ is independent of W, $h(W|X_{(2)}^n) = h(W) = 0$. Moreover, since $X_{(1)}^n$ and $X_{(2)}^n$ are independent, and $Z^n = X_{(1)}^n \oplus X_{(2)}^n$ is a sufficient statistic of $X_{(1)}^n$ and $X_{(2)}^n$ for W,

$$\eta(P_{X_{(1)}^n|X_{(2)}^n=x_{(2)}^n}, P_{W|X_{(1)}^n,X_{(2)}^n=x_{(2)}^n}) = \eta(P_{Z^n}, P_{W|Z^n}) \qquad \text{for all } x_{(2)}^n, \tag{2.146}$$

where Z_i's are i.i.d. Bern(1/2) and $P_{Z_i|W=w} = \text{Bern}(w)$. As shown in Sec. 2.7.4, $\vartheta(P_{W|Z^n}) = 1 - 2^{-n}$. Thus

$$\sup_{x_{(2)}^n} \eta(P_{X_{(1)}^n|X_{(2)}^n=x_{(2)}^n}, P_{W|X_{(1)}^n,X_{(2)}^n=x_{(2)}^n}) \leq 1 - 2^{-n}. \tag{2.147}$$

Combining (2.115) in Theorem 2.5 and (2.142) in Theorem 2.7, we get

$$R_B \geq \frac{1}{2e} 2^{-I(W;Y_{(1)},Y_{(2)}|X_{(2)}^n)+h(W|X_{(2)}^n)} \tag{2.148}$$

$$\geq \frac{1}{2e} 2^{-(1-2^{-n})b}, \tag{2.149}$$

which proves the claim. □

In the extremal case when Processor 1 does not send anything to the estimator, no matter how many bits Processor 2 can send to the estimator, the Bayes risk is lower-bounded by

$$R_B \geq \frac{1}{2e}, \tag{2.150}$$

which follows from (2.145) by setting $b = 0$. This conforms to the fact that $X_{(2)}^n$ is independent of W. It shows that the communication constraint can have much more severe effects on the estimation performance when the sample sets are dependent conditionally on the parameter, as compared to the case where the processors can observe samples that are conditionally i.i.d. given the parameter.

The lower bound in (2.145) may not be tight in general. Setting $b = \frac{1}{2}\log n$,

(2.145) implies that

$$R_{\mathrm{B}} \geq \frac{1}{2e\sqrt{n}}. \tag{2.151}$$

This lower bound would be achievable up to a constant factor when Processor 1 could observe both $X_{(1)}^n$ and $X_{(2)}^n$, in which case the problem is reduced to Example 2.6 with noiseless channel. But it is unlikely to be achievable when the sample sets are distributed to the two processors. A recent paper of El Gamal and Lai [32] studies the problem of decentralized minimum-variance unbiased estimation of W based on observations quantized at the rate of b/n. It is shown that Slepian–Wolf rates are not necessary to achieve the centralized estimation performance, but in their protocol b needs to be proportional to n. The optimal rate region for this decentralized estimation problem is still unknown.

Now we examine the penalty of decentralization. First consider the situation where a single processor can observe both $X_{(1)}^n$ and $X_{(2)}^n$ and map them to a $2b$-bit message. In this situation, (2.115) in Theorem 2.5 together with (2.144) lead to

$$R_B \geq \frac{1}{2e} 2^{-(1-2^{-n})2b}. \tag{2.152}$$

Choosing $2b = \frac{1}{2}\log n$, we have

$$R_B \geq \frac{1}{2e\sqrt{n}}. \tag{2.153}$$

For achievability, the processor can compute the sufficient statistic $Z^n = X_{(1)}^n \oplus X_{(2)}^n$, where Z_i's are i.i.d. Bern(w) given $W = w$, and use $\frac{1}{2}\log n$ bits to uniformly quantize the sample mean of Z^n over $[0, 1]$. Following the same analysis as in Case 1 of Example 2.6, we obtain

$$R_B \leq \frac{1.41}{\sqrt{n}}. \tag{2.154}$$

Thus the lower bound (2.153) is tight up to a constant factor in this situation. Once the sample sets and the $2b = \frac{1}{2}\log n$ bits are distributed to the two

processors, it follows from (2.145) that

$$R_B \geq \frac{1}{2en^{1/4}}. \tag{2.155}$$

Compared with (2.153), we can see the order increase of the lower bound. Therefore, although the Bayes risk lower bound given by (2.145) may be conservative, it can already reflect the penalty of distributing the sample sets and the communication resources to two processors.

Example 2.10 can be extended to the m-processor case:

Example 2.11. *Consider the following conditional distribution of a length-m binary vector $(X_{(1)}, \ldots, X_{(m)})$ given W:*

$$P_{X_{(1)},\ldots,X_{(m)}|W=w}(x_{(1)}, \ldots, x_{(m)}) = \begin{cases} (1-w)2^{-(m-1)}, & \text{if } x_{(1)} \oplus \ldots \oplus x_{(m)} = 0 \\ w2^{-(m-1)}, & \text{if } x_{(1)} \oplus \ldots \oplus x_{(m)} = 1 \end{cases}. \tag{2.156}$$

The vector $(X_{(1)}, \ldots, X_{(m)})$ has the property that any $m-1$ or fewer of its coordinates are independent of W, while the entire vector is dependent on W. Moreover, $Z = X_{(1)} \oplus \ldots \oplus X_{(m)}$ is Bern(w) *conditional on $W = w$, and Z is a sufficient statistic of $(X_{(1)}, \ldots, X_{(m)})$ for estimating W. In decentralized estimation, the ith processor observes $X_{(i)}^n$, $i = 1, \ldots, m$, and maps its samples to a b-bit message. The estimator computes $\widehat{W}$ based on the noiselessly received messages. The distortion function is $\ell(w, \widehat{w}) = |w - \widehat{w}|$.*

With $\mathcal{S} = \{2, \ldots, m\}$, following a similar analysis as in Example 2.10, we can show that

$$h(W|X_{\mathcal{S}}^n) = h(W) = 0, \tag{2.157}$$

and

$$\sup_{x_{\mathcal{S}}^n} \eta(P_{X_{(1)}^n|X_{\mathcal{S}}^n=x_{\mathcal{S}}^n}, P_{W|X_{(1)}^n,X_{\mathcal{S}}^n=x_{\mathcal{S}}^n}) \leq 1 - 2^{-n}. \tag{2.158}$$

Thus combining (2.115) in Theorem 2.5 with Theorem 2.7, we get a lower

bound on the Bayes risk in Example 2.11:

$$R_B \geq \frac{1}{2e} 2^{-(1-2^{-n})b}. \tag{2.159}$$

Again, we can examine the penalty of decentralization. In the situation where a single processor can observe $(X_{(1)}^n, \ldots, X_{(m)}^n)$ and map them to a mb-bit message, it follows from (2.114) in Theorem 2.5 and (2.144) that

$$R_B \geq \frac{1}{2e} 2^{-(1-2^{-n})mb}. \tag{2.160}$$

Choosing $mb = \frac{1}{2} \log n$, we have

$$R_B \geq \frac{1}{2e\sqrt{n}}, \tag{2.161}$$

which is tight up to a constant factor. Once the sample sets and the $mb = \frac{1}{2} \log n$ bits are distributed to the m processors, it follows from (2.159) that

$$R_B \geq \frac{1}{2en^{1/(2m)}}. \tag{2.162}$$

Compared with (2.161), we can see the order increase of the lower bound as m increases, which reflects the penalty of distributing the sample sets and the communication resources to more processors.

2.5.3 Interactive Protocols

When the communications channels are noiseless and feedback is available from the estimator to the processors, each processor can observe the messages sent by the other processors. This allows for the interactive protocols, as studied in [2, 4, 5]. Here we consider a case where the processors take turns to send messages to the estimator, and each processor transmits only once. The message sent by a processor can depend on the previous messages sent by other processors, and is noiselessly received by the estimator. This serial interactive setup has also been considered by Shamir [6].

Theorem 2.8. *Consider the multi-processor setup, where the processors observe sample sets $X_{(1)}^n, \ldots, X_{(m)}^n$ that are conditionally i.i.d. given W, and*

where the message sent by the ith processor is given by

$$Y_{(i)} = \varphi_i(X_{(i)}^n, Y^{i-1}), \qquad i = 1, \ldots, m. \tag{2.163}$$

If the backward channel $P_{X|W}$ satisfies

$$\frac{\mathrm{d}P_{X|W=w}}{\mathrm{d}P_{X|W=w'}}(x) \geq \alpha, \qquad \text{for all } x \in \mathsf{X} \text{ and } w, w' \in \mathsf{W} \tag{2.164}$$

for some constant $\alpha \in (0, 1]$, then, for any choice of φ^m and ψ,

$$I(W; Y^m) \leq \min\Big\{I(W; X^{m\times n}),\, (1-\alpha^n)mb\Big\}. \tag{2.165}$$

In particular, the above upper bound holds in the non-interactive case as well.

Proof. Section 2.7.6. □

We can apply Theorem 2.8 to the "hide-and-seek" problem formulated by Shamir [6] as a generic model for a number of decentralized estimation problems and online learning problems:

Example 2.12. *Consider a family of distributions $\mathcal{P} = \{P_w : w = 1, \ldots, d\}$ on $\{0,1\}^d$. Under P_w, the wth coordinate of the random vector $X \in \{0,1\}^d$ has bias $\frac{1}{2}+\rho$, while the other coordinates of X are independently drawn from* $\mathrm{Bern}(\frac{1}{2})$. *For $i = 1, \ldots, m$, the ith processor observes n samples $X_{(i)}^n$ drawn independently from P_w, and sends a b-bit message $Y_{(i)} = \varphi_i(X_{(i)}^n, Y^{i-1})$ to the estimator. The estimator computes $\widehat{W}$ from the received messages Y^m. The minimax risk of this example is defined as*

$$R_{\mathrm{M}} = \inf_{\varphi^m, \psi} \max_{w \in [d]} \mathbb{P}[\widehat{W} \neq w]. \tag{2.166}$$

The minimax lower bound for this problem obtained in [6] is

$$R_{\mathrm{M}} \geq 1 - \left(\frac{3}{d} + 5\sqrt{\min\left\{\frac{10\rho nmb}{d}, mn\rho^2\right\}}\right) \qquad \text{for } 0 \leq \rho \leq \frac{1}{4n}. \tag{2.167}$$

The question was left open whether this lower bound can be improved. The following result gives an affirmative answer.

Corollary 2.12. *In Example 2.12, the minimax risk is lower bounded by*

$$R_{\mathrm{M}} \geq 1 - \frac{1}{\log d} \min \left\{ \left[1 - \left(\frac{1-2\rho}{1+2\rho} \right)^n \right] mb + 1,\ (4mn\rho^2 \wedge \log d) + 1 \right\} \tag{2.168}$$

for $0 \leq \rho \leq \frac{1}{2}$.

Proof. Let W be uniformly distributed on $\{1, \ldots, d\}$. Then we can use the techniques developed so far to derive lower bounds on the average error probability $\mathbb{P}[\widehat{W} \neq W]$, which will provide lower bounds on the minimax risk. Using the fact that

$$\frac{P_{X|W=w}(x)}{P_{X|W=w'}(x)} \geq \frac{\frac{1}{2}-\rho}{\frac{1}{2}+\rho} \qquad \text{for all } x \in \mathsf{X} \text{ and } w, w' \in \mathsf{W}, \tag{2.169}$$

Theorem 2.8 gives

$$I(W; Y^m) \leq \left[1 - \left(\frac{1-2\rho}{1+2\rho} \right)^n \right] mb \qquad \text{for } 0 \leq \rho \leq \frac{1}{2}. \tag{2.170}$$

In addition, since the entries in $X^{m\times n}$ are i.i.d. conditional on $W = w$, defining Q as the uniform distribution on $\{0,1\}^d$, we have

$$\begin{aligned} I(W; X^{m\times n}) &\leq mnD(P_{X|W} \| P_X | P_W) && (2.171)\\ &\leq mnD(P_{X|W} \| Q | P_W) && (2.172)\\ &= mn\big(1 - h_2(\tfrac{1}{2}+\rho)\big) && (2.173)\\ &\leq 4mn\rho^2, && (2.174) \end{aligned}$$

where (2.172) follows from the identity

$$D(P_{X|W} \| P_X | P_W) = D(P_{X|W} \| Q | P_W) - D(P_X \| Q),$$

and in the last step we have used the fact that $h_2(p) \geq 4p(1-p)$. We also know that $I(W; X^{m\times n}) \leq H(W) = \log d$. Therefore, for $0 \leq \rho \leq \frac{1}{2}$,

$$I(W; Y^m) \leq \min \left\{ \left[1 - \left(\frac{1-2\rho}{1+2\rho} \right)^n \right] mb,\ (4mn\rho^2 \wedge \log d) \right\}. \tag{2.175}$$

Moreover, the lower bound (2.114) in Theorem 2.5 with the choice $\mathcal{S} = \varnothing$

and the distortion function $\ell(w, \widehat{w}) = \mathbf{1}\{\widehat{w} \neq w\}$ becomes the usual Fano's inequality

$$\mathbb{P}[\widehat{W} \neq W] \geq 1 - \frac{I(W; Y^m) + 1}{\log d}. \tag{2.176}$$

Plugging in the upper bound (2.175), we get the result. □

Now we compare the result of Corollary 2.12 and the lower bound in (2.167). Note that the lower bound in (2.167) holds only for $0 \leq \rho \leq \frac{1}{4n}$, whereas the lower bound given in Corollary 2.12 holds for all $0 \leq \rho \leq \frac{1}{2}$. We compare them in two cases. In the first case we set $\rho = \frac{1}{4n}$, and in the second case we set $\rho = 0.01$ for all n. In both cases we set $m = 10$, $d = 512$, and $b = 3d$, as [6] considers the situation where $b = O(d)$. With n varying from 1 to 1000, we plot the lower bounds for the two cases in Fig. 2.3 and Fig. 2.4, respectively. We can see that the lower bound given by Corollary 2.12 is tighter in the plotted range of n in both cases.

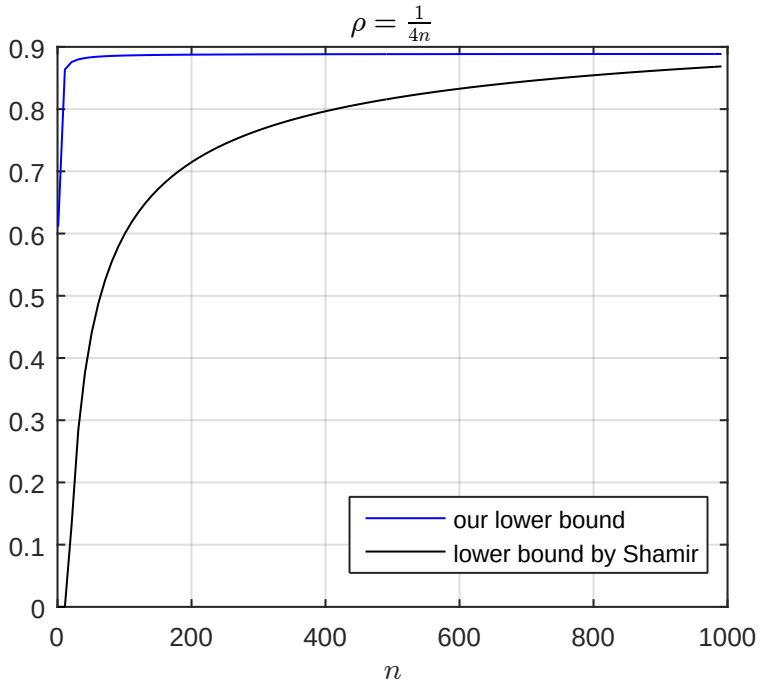


Figure 2.3: Comparison of minimax lower bounds given by Corollary 2.12 and by [6], where $m = 10$, $d = 512$, $b = 3d$, and $\rho = \frac{1}{4n}$.

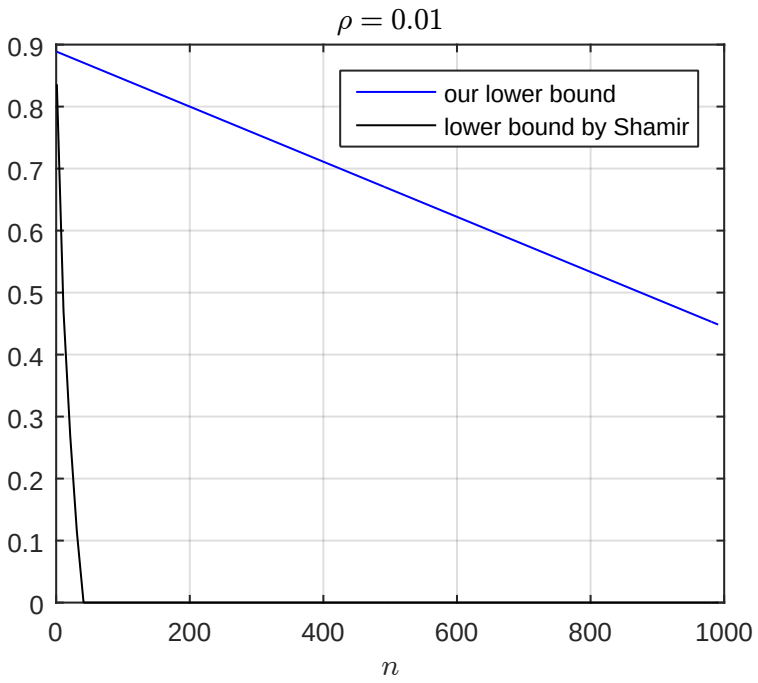


Figure 2.4: Comparison of minimax lower bounds given by Corollary 2.12 and by [6], where $m = 10$, $d = 512$, $b = 3d$, and $\rho = 0.01$ (the lower bound in [6] is set to 0 when $n > 1/4p$).

2.6 Conclusion and Future Research Directions

We have proposed an information-theoretic framework for deriving general lower bounds on the Bayes risk in a systematic way, with applications to decentralized estimation. The main contributions are summarized below.

- Starting in the context of centralized estimation, we have derived lower bounds on the Bayes risk in terms of mutual information (Theorem 2.1) and information density (Theorem 2.2). Both lower bounds involve the small ball probability. They are proved by lower-bounding the probability of excess distortion using properties of the Neyman-Pearson function, and then converting these bounds into lower bounds on the expected distortion using Markov's inequality. The lower bounds in Theorem 2.1 and Theorem 2.2 apply to general parameter spaces, prior distributions, sample generating models, and distortion functions.

- Theorem 2.3 gives a lower bound on the Bayes risk in terms of mutual information and differential entropy. The proof does not involve a detour to bounding the probability of excess distortion, and instead relies

on the Shannon lower bound for the rate-distortion function, which directly relates the mutual information to the expected distortion. Its unconditional version can yield tighter lower bounds than that of Theorem 2.1. However, it only applies when the parameter space is $\mathbb{R}^d$ and the distortion is measured by some norm.

- All of our lower bounds on the Bayes risk for centralized estimation involve an auxiliary conditioning random variable U. A proper choice of U can lead to tighter lower bounds than the ones without conditioning. Moreover, when applied to decentralized estimation, choosing U as a subcollection of sample sets enables us to handle the case where the processors observe conditionally dependent sample sets (Theorem 2.5).

- In the context of decentralized estimation, the general results are refinements of the lower bounds on the Bayes risk based on mutual information (Theorem 2.1 and Theorem 2.3). We have used strong data processing inequalities (SDPIs) as a unified method to quantify the contraction of mutual information caused by communication constraints. The essence of this method is exhibited already in the upper bounds on the mutual information for the single-processor setup (Theorem 2.4). For the multi-processor setup, we have discussed two cases depending on whether the sample sets are conditionally independent or not (Theorem 2.6 and Theorem 2.7). The resulting lower bounds on the Bayes risk (Theorem 2.5) provide us with a systematic way to quantify the penalty of decentralization.

- Finally, we have obtained upper bounds on the mutual information (Theorem 2.8) for interactive communication protocols, where the processors take turns to send their messages, and each processor transmits only once. Deriving general upper bounds on the mutual information using SDPIs for multi-round interactive protocols is an interesting direction for future research.

2.7 Additional Proofs for Chapter 2

2.7.1 Proofs of Lemma 2.1 and Lemma 2.2

The proof relies on the properties of the Neyman–Pearson function, which arises in the context of binary hypothesis testing, and is defined as follows: Given two probability measures P and Q on a common measurable space Z, for any $\alpha \in [0,1]$ let

$$\beta_\alpha(P,Q) = \inf_{f\colon \mathsf{Z}\to[0,1]} \left\{ \int_{\mathsf{Z}} f \,\mathrm{d}Q : \int_{\mathsf{Z}} f \,\mathrm{d}P \geq \alpha \right\}. \tag{2.177}$$

We will need the following properties of $\beta_\alpha(P,Q)$:

- Data processing inequality: For any Markov kernel K from Z into another measurable space Y,

  $$\beta_\alpha(PK, QK) \geq \beta_\alpha(P,Q), \tag{2.178}$$

  where PK and QK are the images of P and Q under K [33].

- Weak converse: For any $\alpha \in [0,1]$,

  $$d_2(\alpha\|\beta_\alpha) \leq D(P\|Q), \tag{2.179}$$

  where $d_2(p\|q) \triangleq p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q}$ is the binary relative entropy [34].

- Strong converse: For any $\alpha \in [0,1]$,

  $$\alpha - \gamma\beta_\alpha \leq \left(1 - \gamma \inf_z \tfrac{\mathrm{d}Q}{\mathrm{d}P}(z)\right) P\left[\tfrac{\mathrm{d}P}{\mathrm{d}Q}(Z) \geq \gamma\right] \qquad \forall \gamma > 0. \tag{2.180}$$

  (see [35, Lemma 35]).

Now we proceed to the proof. Fixing an arbitrary $P_{U|W,X}$, define $\mathbb{P} = P_{U,W,X}$ and $\mathbb{Q} = P_U \otimes P_{W|U} \otimes P_{X|U}$. For any estimator $\psi : \mathsf{X} \to \mathsf{W}$ and any $\rho \geq 0$, consider the function $f(w,x) = \mathbf{1}\{\ell(w,\widehat{w}) \leq \rho\}$. Then $\int f \,\mathrm{d}\mathbb{P} =$

$\mathbb{P}[\ell(W,\widehat{W}) \le \rho]$ and $\int f \,\mathrm{d}\mathbb{Q} = \mathbb{Q}[\ell(W,\widehat{W}) \le \rho]$. On the one hand,

$$\mathbb{Q}[\ell(W,\widehat{W}) \le \rho] = \int_{\mathsf{U}} \int_{\mathsf{W}} \int_{\mathsf{W}} \mathbf{1}\{\ell(w,\widehat{w}) \le \rho\} P_{W|U}(\mathrm{d}w|u) P_{\widehat{W}|U}(\mathrm{d}\widehat{w}|u) P_U(\mathrm{d}u) \tag{2.181}$$

$$= \int_{\mathsf{U}} \int_{\mathsf{W}} \mathbb{P}[\ell(W,\widehat{w}) \le \rho | U = u] P_{\widehat{W}|U}(\mathrm{d}\widehat{w}|u) P_U(\mathrm{d}u) \tag{2.182}$$

$$\le \int_{\mathsf{U}} \sup_{\widehat{w} \in \mathsf{W}} \mathbb{P}[\ell(W,\widehat{w}) \le \rho | U = u] P_U(\mathrm{d}u) \tag{2.183}$$

$$= \mathbb{E}[\mathcal{L}_{W|U}(U,\rho)]. \tag{2.184}$$

On the other hand, by the definition of β_α and by the data processing inequality (2.178),

$$\mathbb{Q}[\ell(W,\widehat{W}) \le \rho] \ge \beta_{\mathbb{P}[\ell(W,\widehat{W}) \le \rho]}\big(\mathbb{P}_{W,\widehat{W}}, \mathbb{Q}_{W,\widehat{W}}\big) \tag{2.185}$$

$$\ge \beta_{\mathbb{P}[\ell(W,\widehat{W}) \le \rho]}(\mathbb{P},\mathbb{Q}). \tag{2.186}$$

Combining (2.184), (2.185), and (2.179), and using the fact that $d_2(\alpha\|\beta) \ge \alpha \log \frac{1}{\beta} - h_2(\alpha)$, we obtain a lower bound on the excess distortion probability

$$\mathbb{P}[\ell(W,\widehat{W}) > \rho] \ge 1 - \frac{I(W;\widehat{W}|U) + 1}{\log\big(1/\mathbb{E}[\mathcal{L}_{W|U}(U,\rho)]\big)}, \tag{2.187}$$

which proves Lemma 2.1.

Combining (2.184), (2.186), and (2.180), we obtain another lower bound on the excess distortion probability

$$\mathbb{P}[\ell(W,\widehat{W}) > \rho] \ge \mathbb{P}[i(W;X|U) < \log\gamma] - \gamma \mathbb{E}[\mathcal{L}_{W|U}(U,\rho)] + \gamma \inf_{u,w,x} \frac{\mathrm{d}P_{W|U=u}}{\mathrm{d}P_{W|U=u,X=x}}(w) \mathbb{P}[i(W;X|U) \ge \log\gamma] \qquad \forall \gamma > 0, \tag{2.188}$$

which proves Lemma 2.2.

2.7.2 Proofs of Corollary 2.1 and Corollary 2.2

Proof of Corollary 2.1

We prove this result using Theorem 2.1, by choosing U as an conditionally independent copy of X^n given W. In Example 2.1, we have the conditional pdf

$$p_{W|X^n=x^n} = N\big(\mathbb{E}[W|X^n=x^n], \mathrm{Var}[W|X^n=x^n]\big), \tag{2.189}$$

where

$$\mathbb{E}[W|X^n=x^n] = \frac{\sigma_W^2}{\sigma_W^2+\sigma^2/n}\bar{x}, \qquad \mathrm{Var}[W|X^n=x^n] = \frac{\sigma_W^2}{1+n\sigma_W^2/\sigma^2}, \tag{2.190}$$

and $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$. Thus,

$$\big\|p_{W|X^n=x^n}\big\|_\infty = \sup_w |p_{W|X^n=x^n}(w)| = \sqrt{\frac{1}{2\pi}\left(\frac{1}{\sigma_W^2}+\frac{n}{\sigma^2}\right)}, \tag{2.191}$$

and therefore

$$\begin{aligned}
\mathcal{L}_{W|X^n}(x^n,\rho) &= \sup_{w\in\mathbb{R}} \mathbb{P}[|W-w|\le\rho|X^n=x^n] && (2.192)\\
&= \sup_{w\in\mathbb{R}} \int_{w-\rho}^{w+\rho} p_{W|X^n=x^n}(w')\mathrm{d}w' && (2.193)\\
&\le 2\rho\big\|p_{W|X^n=x^n}\big\|_\infty && (2.194)\\
&= \rho\sqrt{\frac{2}{\pi}\left(\frac{1}{\sigma_W^2}+\frac{n}{\sigma^2}\right)}. && (2.195)
\end{aligned}$$

In addition,

$$I(W;X^n|X'^n) = I(W;X^n,X'^n) - I(W;X'^n) = \frac{1}{2}\log\frac{1+2n\sigma_W^2/\sigma^2}{1+n\sigma_W^2/\sigma^2}. \tag{2.196}$$

From (2.11),

$$R_{\mathrm{B}} \geq \sup_{0<s<1} \sqrt{\frac{\pi\sigma_W^2}{2(1+n\sigma_W^2/\sigma^2)}} s 2^{-(I(W;X^n|X'^n)+1)/(1-s)} \tag{2.197}$$

$$\geq \frac{1+\sigma^2/(n\sigma_W^2)}{8(2+\sigma^2/(n\sigma_W^2))} \sqrt{\frac{\pi\sigma_W^2}{2(1+n\sigma_W^2/\sigma^2)}} \tag{2.198}$$

$$\geq \frac{1}{16} \sqrt{\frac{\pi\sigma_W^2}{2(1+n\sigma_W^2/\sigma^2)}}, \tag{2.199}$$

where the second line follows by setting $s = 1/2$.

Proof of Corollary 2.2

Again, we use Theorem 2.1 by choosing U as an conditionally independent copy of X^n given W. In Example 2.2, we have the conditional pdf

$$p_{W|X^n}(w|x^n) = (n+1)\binom{n}{k}(1-w)^{n-k}w^k \mathbf{1}\{0 \leq w \leq 1\}, \tag{2.200}$$

where $k = \sum_{i=1}^n x_i$. Since the maximum of the function

$$w \mapsto (1-w)^{n-k}w^k \mathbf{1}\{0 \leq w \leq 1\}$$

is achieved at $w^* = k/n$, we have

$$\left\| p_{W|X^n=x^n} \right\|_\infty = (n+1)\binom{n}{k}\left(1-\frac{k}{n}\right)^{n-k}\left(\frac{k}{n}\right)^k, \tag{2.201}$$

and therefore

$$\mathcal{L}_{W|X^n}(x^n, \rho) \leq 2\rho \left\| p_{W|X^n=x^n} \right\|_\infty = 2\rho(n+1)\binom{n}{k}\left(1-\frac{k}{n}\right)^{n-k}\left(\frac{k}{n}\right)^k. \tag{2.202}$$

Since the marginal distribution of $K = \sum_{i=1}^n X_i$ is uniform over $\{0, \ldots, n\}$,

$$\mathbb{E}[\mathcal{L}_{W|X^n}(X^n, \rho)] \leq 2\rho \sum_{k=0}^n \binom{n}{k}\left(1-\frac{k}{n}\right)^{n-k}\left(\frac{k}{n}\right)^k, \tag{2.203}$$

and, using Stirling's approximation [36, p. 54], we have the estimate

$$\binom{n}{k}\left(1-\frac{k}{n}\right)^{n-k}\left(\frac{k}{n}\right)^{k} \le \sqrt{\frac{n}{2\pi k(n-k)}}, \qquad k=1,\dots,n-1. \tag{2.204}$$

With these upper bounds, we have

$$\mathbb{E}[\mathcal{L}_{W|X^n}(X^n,\rho)] \le 2\rho\left(2+\sum_{k=1}^{n-1}\sqrt{\frac{n}{2\pi k(n-k)}}\right) \le 2\rho\big(2+\sqrt{\pi n/2}\big). \tag{2.205}$$

In addition, from (2.15),

$$I(W;X^n|X'^n) \to \frac{1}{2} \qquad \text{as } n\to\infty. \tag{2.206}$$

Therefore, using Eq. (2.11), we find

$$\begin{aligned} R_{\mathrm{B}} &\ge \sup_{0<s<1} \frac{s}{2(2+\sqrt{\pi n/2})} 2^{-(I(W;X^n|X'^n)+1)/(1-s)} && (2.207)\\ &\ge \frac{1}{4(2+\sqrt{\pi n/2})} 2^{-2(I(W;X^n|X'^n)+1)} && (2.208)\\ &\sim \frac{1}{16\sqrt{2\pi n}} \qquad \text{as } n\to\infty, && (2.209) \end{aligned}$$

where the second line follows by setting $s=1/2$.

2.7.3 Proof of Corollary 2.3

We use the lower bound in (2.24) to prove this result. In Example 2.3, the conditional pdf $p_{W|X^n=x^n}$ is a truncated Gaussian distribution

$$p_{W|X^n}(w|x^n) = \frac{\mathbf{1}\{\|w\|_2 \le a\}}{c_n(\bar{x})(2\pi\sigma^2/n)^{d/2}} \exp\left(-\frac{n}{2\sigma^2}\|\bar{x}-w\|_2^2\right), \tag{2.210}$$

where $\bar{x}=\frac{1}{n}\sum_{i=1}^n x_i \in \mathbb{R}^d$, and the normalizing factor is

$$\begin{aligned} c_n(\bar{x}) &= \int_{\mathbb{R}^d} \frac{\mathbf{1}\{\|w\|_2 \le a\}}{(2\pi\sigma^2/n)^{d/2}} \exp\left(-\frac{n}{2\sigma^2}\|\bar{x}-w\|_2^2\right) \mathrm{d}w && (2.211)\\ &= \mathbb{P}[\|\bar{X}+U_n\|_2 \le a | \bar{X}=\bar{x}] && (2.212) \end{aligned}$$

with $U_n \sim N(0, \frac{\sigma^2}{n}\mathbf{I}_d)$ independent of $\bar{X}$. We can show that[3]

$$c_n(\bar{X}) \xrightarrow{P} 1 \qquad \text{as } n \to \infty. \tag{2.213}$$

Indeed, since $\bar{X} \xrightarrow{P} W$ and $U_n \xrightarrow{d} 0$, we have $\bar{X} + U_n \xrightarrow{d} W$ [37, Lemma 4.5 and Corollary 4.7], hence

$$\mathbb{E}[|c_n(\bar{X}) - 1|] = 1 - \mathbb{E}[c_n(\bar{X})] = \mathbb{P}[\|\bar{X} + U_n\|_2 > a] \to \mathbb{P}[\|W\|_2 > a] = 0 \tag{2.214}$$

as $n \to \infty$, and thus $c_n(\bar{X}) \xrightarrow{L^1} 1$ as $n \to \infty$. Since $Z_n \xrightarrow{P} Z$ is equivalent to $\mathbb{E}[|Z_n - Z| \wedge 1] \to 0$ as $n \to \infty$, we arrive at (2.213). From (2.210),

$$\left\|p_{W|X^n=x^n}\right\|_\infty = \begin{cases} \dfrac{1}{c_n(\bar{x})}\left(\dfrac{n}{2\pi\sigma^2}\right)^{d/2}, & \|\bar{x}\|_2 \le a \\ \dfrac{1}{c_n(\bar{x})}\left(\dfrac{n}{2\pi\sigma^2}\right)^{d/2} \exp\left(-\dfrac{n(\|\bar{x}\|_2 - a)^2}{2\sigma^2}\right), & \|\bar{x}\|_2 > a \end{cases}. \tag{2.215}$$

Let V_d denote the volume of the unit ball in $(\mathbb{R}^d, \|\cdot\|_2)$. Then, for all x^n and $\|w\|_2 \le a$,

$$\frac{p_{W|X^n=x^n}(w)}{p_W(w)} \le V_d a^d \left\|p_{W|X^n=x^n}\right\|_\infty \le \frac{V_d a^d}{c_n(\bar{x})}\left(\frac{n}{2\pi\sigma^2}\right)^{d/2}. \tag{2.216}$$

Choosing $\gamma = (1+\delta)V_d a^d \left(\frac{n}{2\pi\sigma^2}\right)^{d/2}$ (for an arbitrary $\delta > 0$) and $\rho = a(2\gamma)^{-1/d}$ in (2.24), we get

$$\begin{aligned} R_{\mathrm{B}} &\ge \rho\left(\mathbb{P}\Big[i(W;X^n) < \log\gamma\Big] - \gamma\mathcal{L}_W(\rho)\right) && (2.217)\\ &\ge \rho\left(\mathbb{P}\left[\frac{V_d a^d}{c_n(\bar{X})}\left(\frac{n}{2\pi\sigma^2}\right)^{d/2} < \gamma\right] - \gamma\left(\frac{\rho}{a}\right)^d\right) && (2.218)\\ &\ge \left(\frac{1}{2(1+\delta)}\right)^{1/d} V_d^{-1/d}\sqrt{\frac{2\pi\sigma^2}{n}}\left(\mathbb{P}\left[\frac{1}{c_n(\bar{X})} < 1+\delta\right] - \frac{1}{2}\right) && (2.219)\\ &\gtrsim \frac{1}{20}\sqrt{\frac{2\pi\sigma^2 d}{n}} \qquad \text{as } n \to \infty, && (2.220) \end{aligned}$$

[3]Given a sequence of real-valued random variables $\{Z_n\}$, we write $Z_n \xrightarrow{L^1} Z$, $Z_n \xrightarrow{P} Z$, and $Z_n \xrightarrow{d} Z$ to indicate the convergence in L^1, in probability, and in distribution, respectively.

where the last step follows from the fact that $c_n(\bar{X}) \xrightarrow{P} 1$ (hence $1/c_n(\bar{X}) \xrightarrow{P} 1$), $(1/2)^{1/d} \geq 1/2$ for all $d \geq 1$, $V_d^{1/d} \leq 5/\sqrt{d}$ for all $d \geq 1$, and the fact that $\delta > 0$ is arbitrary. We thus obtain a lower bound that is asymptotic in n and non-asymptotic in a, σ^2, and d.

2.7.4 Proof of Equation (2.101)

We have $p_W(w) = 1$ for $w \in [0,1]$, and $P_{X^n|W}(x^n|w) = w^s(1-w)^{n-s}$, where s is the Hamming weight (the number of 1's) of x^n. Thus,

$$P_{X^n}(x^n) = \int_0^1 w^s(1-w)^{n-s}\mathrm{d}w = \frac{1}{(n+1)\binom{n}{s}}$$

and

$$P_{W|X^n}(w|x^n) = w^s(1-w)^{n-s}(n+1)\binom{n}{s}.$$

This gives

$$\|P_{W|X^n=x^n} - P_{W|X^n=\tilde{x}^n}\|_{\mathrm{TV}} = \frac{n+1}{2} \int_0^1 \left| w^s(1-w)^{n-s}\binom{n}{s} - w^{\tilde{s}}(1-w)^{n-\tilde{s}}\binom{n}{\tilde{s}} \right| \mathrm{d}w,$$

which is maximized by choosing x^n and $\tilde{x}^n$ such that $s = 0$ and $\tilde{s} = n$. Hence

$$\vartheta(P_{W|X^n}) = \frac{n+1}{2}\int_0^1 \left|(1-w)^n - w^n\right|\mathrm{d}w = 1 - 2^{-n}.$$

2.7.5 Proofs of Theorem 2.7 and Equation (2.143)

Proof of Theorem 2.7

The first upper bound follows from

$$I(W;V^{m\times T}|X_{\mathcal{S}}^n) = I(W;V_{\mathcal{S}^c}^T|X_{\mathcal{S}}^n) \tag{2.221}$$
$$\leq \eta(P_{V_{\mathcal{S}^c}^T|U_{\mathcal{S}^c}^T}) I(W;U_{\mathcal{S}^c}^T|X_{\mathcal{S}}^n) \tag{2.222}$$
$$\leq \eta_{|\mathcal{S}^c|T} I(W;Y_{\mathcal{S}^c}|X_{\mathcal{S}}^n) \tag{2.223}$$
$$\leq \eta_{|\mathcal{S}^c|T} I(W;X_{\mathcal{S}^c}^n|X_{\mathcal{S}}^n), \tag{2.224}$$

where (2.221) follows from the Markov chain $W, V_{\mathcal{S}^c}^T - X_{\mathcal{S}}^n - V_{\mathcal{S}}^T$, and (2.222) follows from the Markov chain $W, X_{\mathcal{S}}^n - U_{\mathcal{S}^c}^T - V_{\mathcal{S}^c}^T$ and the conditional version of SDPI (Lemma 2.7).

Alternatively, we can upper-bound $I(W;Y_{\mathcal{S}^c}|X_{\mathcal{S}}^n)$ in (2.223) with the following chain of inequalities:

$$I(W;V^{m\times T}|X_{\mathcal{S}}^n) \leq \eta_{|\mathcal{S}^c|T} I(W;Y_{\mathcal{S}^c}|X_{\mathcal{S}}^n) \tag{2.225}$$
$$= \eta_{|\mathcal{S}^c|T} \int I(W;Y_{\mathcal{S}^c}|X_{\mathcal{S}}^n = x_{\mathcal{S}}^n) P_{X_{\mathcal{S}}^n}(\mathrm{d}x_{\mathcal{S}}^n) \tag{2.226}$$
$$\leq \eta_{|\mathcal{S}^c|T} \int I(X_{\mathcal{S}^c}^n;Y_{\mathcal{S}^c}|X_{\mathcal{S}}^n = x_{\mathcal{S}}^n) \eta\big(P_{X_{\mathcal{S}^c}^n|X_{\mathcal{S}}^n=x_{\mathcal{S}}^n}, P_{W|X_{\mathcal{S}^c}^n,X_{\mathcal{S}}^n=x_{\mathcal{S}}^n}\big) P_{X_{\mathcal{S}}^n}(\mathrm{d}x_{\mathcal{S}}^n) \tag{2.227}$$
$$\leq \eta_{|\mathcal{S}^c|T} \sup_{x_{\mathcal{S}}^n} \eta\big(P_{X_{\mathcal{S}^c}^n|X_{\mathcal{S}}^n=x_{\mathcal{S}}^n}, P_{W|X_{\mathcal{S}^c}^n,X_{\mathcal{S}}^n=x_{\mathcal{S}}^n}\big) |\mathcal{S}^c| b, \tag{2.228}$$

where (2.227) is from the Markov chain $W - X_{\mathcal{S}^c}^n - Y_{\mathcal{S}^c}$ conditional on $X_{\mathcal{S}}^n = x_{\mathcal{S}}^n$ and the SDPI, and (2.228) is because $I(X_{\mathcal{S}^c}^n;Y_{\mathcal{S}^c}|X_{\mathcal{S}}^n) \leq H(Y_{\mathcal{S}^c}) \leq |\mathcal{S}^c| b$.

Lastly, from the Markov chain $W - X_{\mathcal{S}^c}^n - V_{\mathcal{S}^c}^T$ conditional on $X_{\mathcal{S}}^n = x_{\mathcal{S}}^n$ and the SDPI,

$$I(W;V^{m\times T}|X_{\mathcal{S}}^n) = I(W;V_{\mathcal{S}^c}^T|X_{\mathcal{S}}^n) \tag{2.229}$$
$$\leq I(X_{\mathcal{S}^c}^n;V_{\mathcal{S}^c}^T|X_{\mathcal{S}}^n) \sup_{x_{\mathcal{S}}^n} \eta\big(P_{X_{\mathcal{S}^c}^n|X_{\mathcal{S}}^n=x_{\mathcal{S}}^n}, P_{W|X_{\mathcal{S}^c}^n,X_{\mathcal{S}}^n=x_{\mathcal{S}}^n}\big) \tag{2.230}$$
$$\leq |\mathcal{S}^c| C T \sup_{x_{\mathcal{S}}^n} \eta\big(P_{X_{\mathcal{S}^c}^n|X_{\mathcal{S}}^n=x_{\mathcal{S}}^n}, P_{W|X_{\mathcal{S}^c}^n,X_{\mathcal{S}}^n=x_{\mathcal{S}}^n}\big), \tag{2.231}$$

where the last step follows from $I(X_{\mathcal{S}^c}^n;V_{\mathcal{S}^c}^T|X_{\mathcal{S}}^n) \leq I(U_{\mathcal{S}^c}^T;V_{\mathcal{S}^c}^T|X_{\mathcal{S}}^n) \leq I(U_{\mathcal{S}^c}^T;V_{\mathcal{S}^c}^T)$,

because of the Markov chain $X_{\mathcal{S}}^n - U_{\mathcal{S}^c}^T - V_{\mathcal{S}^c}^T$.

Proof of Equation (2.143)

The proof parallels that of Theorem 2.7. For the first upper bound in (2.143),

$$I(W;V^{mT}|X_{\mathcal{S}}^n) \le \eta(P_{V^{mT}|U^{mT}}) I(W;U^{mT}|X_{\mathcal{S}}^n) \tag{2.232}$$
$$\le \eta_{mT} I(W;Y|X_{\mathcal{S}}^n) \tag{2.233}$$
$$\le \eta_{mT} I(W;X_{\mathcal{S}^c}^n|X_{\mathcal{S}}^n), \tag{2.234}$$

where (2.232) is from the Markov chain $W, X_{\mathcal{S}}^n - U^{mT} - V^{mT}$.

Alternatively, we can upper-bound $I(W;Y|X_{\mathcal{S}}^n)$ in (2.233) with the following chain of inequalities:

$$I(W;V^{mT}|X_{\mathcal{S}}^n) \le \eta_{mT} I(W;Y|X_{\mathcal{S}}^n) \tag{2.235}$$
$$= \eta_{mT} \int I(W;Y|X_{\mathcal{S}}^n = x_{\mathcal{S}}^n) P_{X_{\mathcal{S}}^n}(\mathrm{d}x_{\mathcal{S}}^n) \tag{2.236}$$
$$\le \eta_{mT} \int I(X_{\mathcal{S}^c}^n;Y|X_{\mathcal{S}}^n = x_{\mathcal{S}}^n) \eta\big(P_{X_{\mathcal{S}^c}^n|X_{\mathcal{S}}^n=x_{\mathcal{S}}^n}, P_{W|X_{\mathcal{S}^c}^n,X_{\mathcal{S}}^n=x_{\mathcal{S}}^n}\big) P_{X_{\mathcal{S}}^n}(\mathrm{d}x_{\mathcal{S}}^n) \tag{2.237}$$
$$\le \eta_{mT} \sup_{x_{\mathcal{S}}^n} \eta\big(P_{X_{\mathcal{S}^c}^n|X_{\mathcal{S}}^n=x_{\mathcal{S}}^n}, P_{W|X_{\mathcal{S}^c}^n,X_{\mathcal{S}}^n=x_{\mathcal{S}}^n}\big) mb, \tag{2.238}$$

where (2.237) is from the Markov chain $W - X_{\mathcal{S}^c}^n - Y$ conditional on $X_{\mathcal{S}}^n = x_{\mathcal{S}}^n$ and the SDPI; (2.238) is because $I(X_{\mathcal{S}^c}^n;Y|X_{\mathcal{S}}^n) \le H(Y) \le mb$.

Lastly, from the Markov chain $W - X_{\mathcal{S}^c}^n - V^{mT}$ conditional on $X_{\mathcal{S}}^n = x_{\mathcal{S}}^n$ and the SDPI,

$$I(W;V^{mT}|X_{\mathcal{S}}^n) \le I(X_{\mathcal{S}^c}^n;V^{mT}|X_{\mathcal{S}}^n) \sup_{x_{\mathcal{S}}^n} \eta\big(P_{X_{\mathcal{S}^c}^n|X_{\mathcal{S}}^n=x_{\mathcal{S}}^n}, P_{W|X_{\mathcal{S}^c}^n,X_{\mathcal{S}}^n=x_{\mathcal{S}}^n}\big) \tag{2.239}$$
$$\le mCT \sup_{x_{\mathcal{S}}^n} \eta\big(P_{X_{\mathcal{S}^c}^n|X_{\mathcal{S}}^n=x_{\mathcal{S}}^n}, P_{W|X_{\mathcal{S}^c}^n,X_{\mathcal{S}}^n=x_{\mathcal{S}}^n}\big), \tag{2.240}$$

where the last step follows from $I(X_{\mathcal{S}^c}^n;V^{mT}|X_{\mathcal{S}}^n) \le I(U^{mT};V^{mT}|X_{\mathcal{S}}^n) \le I(U^{mT};V^{mT})$, because of the Markov chain $X_{\mathcal{S}}^n - U^{mT} - V^{mT}$.

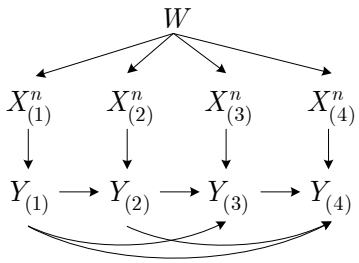


Figure 2.5: Bayesian network of $(W, X^{m\times n}, Y^m)$ in the interactive case $(m = 4)$.

2.7.6 Proof of Theorem 2.8

The first upper bound in (2.165) follows from the Markov chain $W - X^{m\times n} - Y^m$.

To prove the second upper bound in (2.165), we use the chain rule to decompose $I(W; Y^m)$ as

$$I(W; Y^m) = \sum_{i=1}^{m} I(W : Y_{(i)}|Y^{i-1}), \tag{2.241}$$

and then apply SDPI to each term. Since $Y_{(i)} = \varphi_i(X^n_{(i)}, Y^{i-1})$, we know that $W - X^n_{(i)} - Y_{(i)}$ form a Markov chain given $Y^{i-1} = y^{i-1}$. Thus the SDPI gives

$$I(W; Y_{(i)}|Y^{i-1} = y^{i-1}) \le \eta(P_{W|X^n_{(i)},Y^{i-1}=y^{i-1}}) I(X^n_{(i)}; Y_{(i)}|Y^{i-1} = y^{i-1}). \tag{2.242}$$

Now the goal is to upper-bound $\eta(P_{W|X^n_{(i)},Y^{i-1}=y^{i-1}})$. We can view $P_{W|X^n_{(i)},Y^{i-1}=y^{i-1}}$ as the backward channel and $P_{X^n_{(i)}|W,Y^{i-1}=y^{i-1}}$ as the forward channel. Since we assume that each processor sends its message only once, $X^n_{(i)}$ and Y^{i-1} are conditionally independent given W, which can be seen from the Bayesian network in Fig. 2.5. Therefore,

$$\frac{\mathrm{d}P_{X^n_{(i)}|W=w,Y^{i-1}=y^{i-1}}}{\mathrm{d}P_{X^n_{(i)}|W=w',Y^{i-1}=y^{i-1}}}(x^n_{(i)}) = \frac{\mathrm{d}P_{X^n_{(i)}|W=w}}{\mathrm{d}P_{X^n_{(i)}|W=w'}}(x^n_{(i)}) \tag{2.243}$$

$$\ge \alpha^n \quad \text{for all } x^n_{(i)}, w, \text{ and } w', \tag{2.244}$$

where (2.244) follows from the condition in (2.164) and the assumption that the samples in $X^n_{(i)}$ are conditionally i.i.d. given W. Then by Lemma 2.5, the

SDPI constant of the backward channel satisfies

$$\eta(P_{W|X_{(i)}^n, Y^{i-1}=y^{i-1}}) \leq 1 - \alpha^n. \tag{2.245}$$

Since the above inequalities hold for any y^{i-1}, we have

$$\begin{aligned} I(W; Y_{(i)}|Y^{i-1}) &\leq (1-\alpha^n) I(X_{(i)}^n; Y_{(i)}|Y^{i-1}) && (2.246)\\ &\leq (1-\alpha^n) I(X^{m\times n}; Y_{(i)}|Y^{i-1}). && (2.247) \end{aligned}$$

It follows that

$$\begin{aligned} I(W; Y^m) &\leq (1-\alpha^n) I(X^{m\times n}; Y^m) && (2.248)\\ &\leq (1-\alpha^n) mb. && (2.249) \end{aligned}$$

Chapter 3

Lower Bounds for Distributed Function Computation

3.1 Introduction and Preview of Results

3.1.1 Model and Problem Formulation

The problem of distributed function computation arises in such applications as inference and learning in networks, and consensus or coordination of multiple agents. Each node of the network has an initial random observation and aims to compute a common function of the observations of all the nodes by exchanging messages with its neighbors over discrete memoryless point-to-point channels and by performing local computations. A problem of theoretical and practical interest is to determine the fundamental limits on the *computation time*, i.e., the minimum number of steps needed by any distributed computation algorithm to guarantee that, when the algorithm terminates, each node has an accurate estimate of the function value with high probability.

Formally, a network consisting of nodes connected by point-to-point channels is represented by a directed graph $G = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is a finite set of nodes and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is a set of edges. Node u can send messages to node v only if $(u, v) \in \mathcal{E}$. Accordingly, to each edge $e \in \mathcal{E}$ we associate a discrete memoryless channel with finite input alphabet X_e, finite output alphabet Y_e, and stochastic transition law K_e that specifies the transition probabilities $K_e(y_e|x_e)$ for all $(x_e, y_e) \in \mathsf{X}_e \times \mathsf{Y}_e$. The channels corresponding to different edges are assumed to be independent. Initially, each node v has access to an observation given by a random variable (r.v.) W_v taking values in some space W_v. We assume that the joint probability law $\mathbb{P}_W$ of $W \triangleq (W_v)_{v\in\mathcal{V}}$ is known to all the nodes. Given a function $f : \prod_{v\in\mathcal{V}} \mathsf{W}_v \to \mathsf{Z}$, each node aims to estimate the value $Z = f(W)$ via local communication and computation. For example, when f is given by the identity mapping $Z = W$, the goal of

each node is to estimate the observations of all other nodes in the network.

The operation of the network is synchronized, and takes place in discrete time. A *T-step algorithm* $\mathcal{A}$ is a collection of deterministic encoders $(\varphi_{v,t})$ and estimators (ψ_v), for all $v \in \mathcal{V}$ and $t \in \{1, \ldots, T\}$, given by mappings

$$\varphi_{v,t} : \mathsf{W}_v \times \mathsf{Y}_{v\leftarrow}^{t-1} \to \mathsf{X}_{v\rightarrow}, \quad \psi_v : \mathsf{W}_v \times \mathsf{Y}_{v\leftarrow}^{T} \to \mathsf{Z},$$

where $\mathsf{X}_{v\rightarrow} = \prod_{u\in\mathcal{N}_{v\rightarrow}} \mathsf{X}_{(v,u)}$ and $\mathsf{Y}_{v\leftarrow} = \prod_{u\in\mathcal{N}_{v\leftarrow}} \mathsf{Y}_{(u,v)}$. Here, $\mathcal{N}_{v\leftarrow} \triangleq \{u \in \mathcal{V} : (u,v) \in \mathcal{E}\}$ and $\mathcal{N}_{v\rightarrow} \triangleq \{u \in \mathcal{V} : (v,u) \in \mathcal{E}\}$ are, respectively, the in-neighborhood and the out-neighborhood of node v. The algorithm operates as follows: at each step t, each node v computes $X_{v,t} \triangleq (X_{(v,u),t})_{u\in\mathcal{N}_{v\rightarrow}} = \varphi_{v,t}\big(W_v, Y_v^{t-1}\big) \in \mathsf{X}_{v\rightarrow}$, and then transmits each message $X_{(v,u),t}$ along the edge $(v,u) \in \mathcal{E}$. For each $(u,v) \in \mathcal{E}$, the received message $Y_{(u,v),t}$ at each t is related to the transmitted message $X_{(u,v),t}$ via the stochastic transition law $K_{(u,v)}$. At step T, each node v computes $\widehat{Z}_v = \psi_v(W_v, Y_v^T)$ as an estimate of Z, where $Y_{v,t} \triangleq (Y_{(u,v),t})_{u\in\mathcal{N}_{v\leftarrow}} \in \mathsf{Y}_{v\leftarrow}$ for $t \in \{1, \ldots, T\}$.

Given a nonnegative *distortion function* $\ell : \mathsf{Z} \times \mathsf{Z} \to \mathbb{R}^+$, we use the excess distortion probability $\mathbb{P}\big[\ell(Z, \widehat{Z}_v) > \varepsilon\big]$ to quantify the computation fidelity of the algorithm at node v. A key fundamental limit of distributed function computation is the *(ε,δ)-computation time*:

$$T(\varepsilon,\delta) \triangleq \inf\Big\{T \in \mathbb{N} : \exists \text{ a } T\text{-step algorithm } \mathcal{A} \text{ such that } \max_{v\in\mathcal{V}} \mathbb{P}\big[\ell(Z, \widehat{Z}_v) > \varepsilon\big] \le \delta\Big\}. \tag{3.1}$$

If an algorithm $\mathcal{A}$ has the property that

$$\max_{v\in\mathcal{V}} \mathbb{P}\big[\ell(Z, \widehat{Z}_v) > \varepsilon\big] \le \delta,$$

then we say that it *achieves accuracy* ε with *confidence* $1-\delta$. Thus, $T(\varepsilon,\delta)$ is the minimum number of time steps needed by any algorithm to achieve accuracy ε with confidence $1-\delta$. The objective of this chapter is to derive general lower bounds on $T(\varepsilon,\delta)$ for arbitrary network topologies, discrete memoryless channel models, continuous or discrete observations, and functions f.

Previously, this problem (for real-valued functions and quadratic distortion) has been studied by Ayaso et al. [38] and by Como and Dahleh [39] using

information-theoretic techniques. This problem is also related to the study of communication complexity of distributed computing over noisy channels. In that context, Goyal et al. [40] studied the problem of computing Boolean functions in complete graphs, where each pair of nodes communicates over a pair of independent binary symmetric channels (BSCs), and obtained tight lower bounds on the number of serial broadcasts using an approach tailored to that special problem. The technique used in [40] has been extended to random planar networks by Dutta et al. [41]. Other related, but differently formulated, problems include communication complexity and information complexity in distributed computing over noiseless channels, surveyed in [42]; minimum communication rates for distributed computing [43–45], compression, or estimation based on infinite sequences of observations, surveyed in [46, Chap. 21]; and distributed computing in wireless networks, surveyed in [47]. Some achievability results for specific distributed function computation problems can be found in [38, 48–55].

3.1.2 Method of Analysis and Summary of Main Results

Our analysis builds upon the information-theoretic framework proposed by Ayaso et al. [38] and Como and Dahleh [39]. The underlying idea is rather natural and exploits a fundamental trade-off between the minimal amount of information any good algorithm must *necessarily* extract about the function value Z when it terminates and the maximal amount of information any algorithm is able to obtain due to time and communication constraints. To be more precise, given any set of nodes $\mathcal{S} \subseteq \mathcal{V}$, let $W_{\mathcal{S}} \triangleq (W_v)_{v\in\mathcal{S}}$ denote the vector of observations at all the nodes in $\mathcal{S}$. The quantity that plays a key role in the analysis is the conditional mutual information $I(Z;\widehat{Z}_v|W_{\mathcal{S}})$ between the actual function value Z and the estimate $\widehat{Z}_v$ at an arbitrary node v, given the observations in an arbitrary subset of nodes $\mathcal{S}$ containing node v.

Consider an arbitrary T-step algorithm $\mathcal{A}$ that achieves accuracy ε with confidence $1-\delta$. Then, as we show in Lemma 3.1 of Sec. 3.2.1, this mutual information can be lower-bounded by

$$I(Z;\widehat{Z}_v|W_{\mathcal{S}}) \geq (1-\delta)\log\frac{1}{\mathbb{E}[\mathcal{L}_{Z|W_{\mathcal{S}}}(W_{\mathcal{S}},\varepsilon)]} - h_2(\delta), \tag{3.2}$$

where $h_2(\delta) = -\delta \log \delta - (1-\delta)\log(1-\delta)$ is the binary entropy function, and

$$\begin{aligned}\mathcal{L}_{Z|W_{\mathcal{S}}}(w_{\mathcal{S}}, \varepsilon) &= \sup_{z \in \mathsf{Z}} \mathbb{P}[\ell(Z, z) \leq \varepsilon | W_{\mathcal{S}} = w_{\mathcal{S}}] \\ &= \sup_{z \in \mathsf{Z}} \mathbb{P}[\ell(f(W), z) \leq \varepsilon | W_{\mathcal{S}} = w_{\mathcal{S}}]\end{aligned}$$

is the conditional small ball probability of $Z = f(W)$ given $W_{\mathcal{S}} = w_{\mathcal{S}}$ as defined in (2.3). The conditional small ball probability quantifies the difficulty of localizing the value of $Z = f(W)$ in a "distortion ball" of size ε given partial knowledge about the value of W, namely $W_{\mathcal{S}} = w_{\mathcal{S}}$. For example, as discussed in Sec. 3.4, when f is a linear function of the observations W, the conditional small ball probability can be expressed in terms of so-called *Lévy concentration functions* [56], for which tight estimates are available under various regularity conditions.

On the other hand, if $\mathcal{A}$ is a T-step algorithm, then the amount of information any node v has about Z once $\mathcal{A}$ terminates can be upper bounded by a quantity that increases with T and also depends on the network topology and on the information transmission capabilities of the channels connecting the nodes. To quantify this amount of information, we consider a *cut* of the network, i.e., a partition of the set of nodes $\mathcal{V}$ into two disjoint subsets $\mathcal{S}$ and $\mathcal{S}^c \triangleq \mathcal{V} \backslash \mathcal{S}$, such that $v \in \mathcal{S}$. The underlying intuition is that any information that nodes in $\mathcal{S}$ receive about $W_{\mathcal{S}^c}$ must flow across the edges from nodes in $\mathcal{S}^c$ to nodes in $\mathcal{S}$. The set of these edges, denoted by $\mathcal{E}_{\mathcal{S}}$, is referred to as the *cutset* induced by $\mathcal{S}$. Figure 3.1 illustrates these concepts on a simple four-node network. We then have the following upper bound [38,39] (see also

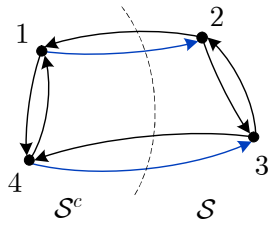


Figure 3.1: A four-node network with a cut defined by $\mathcal{S} = \{2, 3\}$ and $\mathcal{S}^c = \{1, 4\}$. The cutset $\mathcal{E}_{\mathcal{S}}$ consists of edges $(1, 2)$ and $(4, 3)$, marked in blue.

Lemma 3.4 in Sec. 3.2.2):

$$I(Z;\widehat{Z}_v|W_{\mathcal{S}}) \leq TC_{\mathcal{S}}. \tag{3.3}$$

The quantity $C_{\mathcal{S}}$, referred to as the *cutset capacity*, is the sum of the Shannon capacities of all the channels located on the edges in the cutset $\mathcal{E}_{\mathcal{S}}$. Thus, if there exists a cut $(\mathcal{S},\mathcal{S}^c)$ with a small value of $C_{\mathcal{S}}$, then the amount of information gained by the nodes in $\mathcal{S}$ about Z will also be small. Note that the cutset upper bound grows linearly with T. However, when the initial observations W are discrete, we also know that

$$I(Z;\widehat{Z}_v|W_{\mathcal{S}}) \leq I(W_{\mathcal{S}^c};\widehat{Z}_v|W_{\mathcal{S}}) \leq H(W_{\mathcal{S}^c}|W_{\mathcal{S}}),$$

where $H(W_{\mathcal{S}^c}|W_{\mathcal{S}})$ is the conditional entropy of $W_{\mathcal{S}^c}$ given $W_{\mathcal{S}}$, which does not depend on T. In fact, we sharpen this bound by showing in Lemma 3.5 in Sec. 3.2.3 that

$$I(Z;\widehat{Z}_v|W_{\mathcal{S}}) \leq \big(1-(1-\eta_v)^T\big)H(W_{\mathcal{S}^c}|W_{\mathcal{S}}). \tag{3.4}$$

Here, η_v is defined as

$$\eta_v = \sup \frac{I(U;Y_v)}{I(U;X_v)},$$

where the supremum is over all triples (U,X_v,Y_v) of r.v.'s, such that U takes values in an arbitrary alphabet, $U \to X_v \to Y_v$ is a Markov chain, X_v takes values in $\mathsf{X}_{v\leftarrow}$, Y_v takes values in $\mathsf{Y}_{v\leftarrow}$, and the conditional probability law $\mathbb{P}_{Y_v|X_v}$ is equal to the product of all the channels entering v. As discussed in detail in Sec. 2.3, this constant is related to so-called strong data processing inequalities, and quantifies the information transmission capabilities of the channels entering v. When $\eta_v < 1$, the upper bound (3.4) is *strictly smaller* than $H(W_{\mathcal{S}^c}|W_{\mathcal{S}})$. With the upper bound (3.4), we can strengthen the cutset bound to the following:

$$I(Z;\widehat{Z}_v|W_{\mathcal{S}}) \leq \min\big\{TC_{\mathcal{S}}, \big(1-(1-\eta_v)^T\big)H(W_{\mathcal{S}^c}|W_{\mathcal{S}})\big\}. \tag{3.5}$$

Combining the bounds in (3.2) and (3.5), we conclude that, if there exists a

T-step algorithm $\mathcal{A}$ that achieves accuracy ε with confidence $1-\delta$, then

$$T \geq \max\left\{ \frac{1}{C_{\mathcal{S}}} \left((1-\delta) \log \frac{1}{\mathbb{E}[\mathcal{L}_{Z|W_{\mathcal{S}}}(W_{\mathcal{S}}, \varepsilon)]} - h_2(\delta) \right), \right. \\ \left. \frac{\log\left(1 - \frac{1}{H(W_{\mathcal{S}^c}|W_{\mathcal{S}})} \left((1-\delta) \log \frac{1}{\mathbb{E}[\mathcal{L}_{Z|W_{\mathcal{S}}}(W_{\mathcal{S}}, \varepsilon)]} - h_2(\delta) \right)\right)}{\log(1-\eta_v)} \right\}; \tag{3.6}$$

moreover, this inequality holds for all choices of $\mathcal{S} \subset \mathcal{V}$ and $v \in \mathcal{S}$. The precise statements of the resulting lower bounds on $T(\varepsilon, \delta)$ are given in Theorem 3.1 and Theorem 3.3.

The lower bound in (3.6) accounts for the difficulty of estimating the value of $Z = f(W)$ given only a subset of observations $W_{\mathcal{S}}$ through the small ball probability $\mathcal{L}_{Z|W_{\mathcal{S}}}(W_{\mathcal{S}}, \varepsilon)$, and for the communication bottlenecks in the network through the cutset capacity $C_{\mathcal{S}}$ and the constants η_v. The presence of $\mathcal{L}_{Z|W_{\mathcal{S}}}(W_{\mathcal{S}}, \varepsilon)$ in the bound ensures the correct scaling of $T(\varepsilon, \delta)$ in the high-accuracy limit $\varepsilon \to 0$. In particular, when the function f is real-valued and the probability distribution of $Z = f(W)$ has a density, it is not hard to see that $\mathcal{L}_{Z|W_{\mathcal{S}}}(W_{\mathcal{S}}, \varepsilon) = O(\varepsilon)$, and therefore $T(\varepsilon, \delta)$ grows without bound at the rate of $\log(1/\varepsilon)$ as $\varepsilon \to 0$. By contrast, the bounds of Ayaso et al. [38] saturate at a finite constant even when no computation error is allowed, i.e., when $\varepsilon = 0$. Detailed comparison with existing bounds is given in Sec. 3.4, where we particularize our lower bounds to the computation of linear functions. Moreover, in certain cases our lower bound on $T(\varepsilon, \delta)$ tends to infinity in the high-confidence regime $\delta \to 0$. By contrast, existing lower bounds that rely on cutset capacity estimates remain bounded regardless of how small we make δ.

Throughout this chapter, we provide several concrete examples that illustrate the tightness of the general lower bound in (3.6). In particular, Example 3.1 in Sec. 3.2.4 concerns the problem of computing the mod-2 sum of two independent Bern($\frac{1}{2}$) random variables in a network of two nodes communicating over binary symmetric channels (BSCs). For that problem, we obtain a lower bound on $T(0, \delta)$ that matches an achievable upper bound within a factor of 2. In Example 3.2 in Sec. 3.2.4, we consider the case where the nodes aim to distribute their discrete observations to all other nodes, and obtain a lower bound on $T(0, \delta)$ that captures the *conductance* of the

network, which plays a prominent role in the previously published bounds of Ayaso et al. [38]. In Sec. 3.5, we study two more examples: computing a sum of independent Rademacher random variables in a dumbbell network of BSCs, and distributed averaging of real-valued observations in an arbitrary network of binary erasure channels (BECs). Our lower bound for the former example precisely captures the dependence of the computation time on the number of nodes in the network, while for the latter example it captures the correct dependence of the computation time on the accuracy parameter ε.

A significant limitation of the analysis based on a single cut $(\mathcal{S}, \mathcal{S}^c)$ of the network is that it only captures the flow of information across the cutset $\mathcal{E}_{\mathcal{S}}$, but does not account for the time it takes the algorithm to disseminate this information to all the nodes in $\mathcal{S}$. We address this limitation in Sec. 3.3 through a multi-cutset analysis. The main idea is to partition the set of nodes $\mathcal{V}$ into *several* subsets $\mathcal{S}_1, \ldots, \mathcal{S}_n$, such that, for all $\mathcal{P}_i \triangleq \mathcal{S}_1 \cup \ldots \cup \mathcal{S}_i$, the cutsets $\mathcal{E}_{\mathcal{P}_1}, \ldots, \mathcal{E}_{\mathcal{P}_{n-1}}, \mathcal{E}_{\mathcal{P}_1^c}, \ldots, \mathcal{E}_{\mathcal{P}_{n-1}^c}$ are disjoint, and to analyze the flow of information across this sequence of cutsets. Once such a partition is selected, the analysis is based on a network reduction argument (Lemma 3.7), which lumps all the nodes in each $\mathcal{S}_i$ into a single virtual "supernode." The construction of the partition ensures that each supernode i only communicates with supernodes $i-1$ and $i+1$, and can also send noisy messages to itself (this is needed to simulate noisy communication among the nodes within $\mathcal{S}_i$ in the original network). Thus, the reduced network takes the form of a chain with n nodes communicating with their nearest neighbors over bidirectional noisy links and, in addition, sending noisy messages to themselves. We refer to this network as a *bidirected chain* of length $n-1$. Figure 3.2a shows the partition of a six-node network, and the bidirected chain reduced from this network is shown in Fig. 3.2b.

Once this reduction is carried out, we can convert any T-step algorithm $\mathcal{A}$ running on the original network into a *randomized* T-step algorithm $\mathcal{A}'$ running on the reduced network with the same accuracy and confidence guarantees as $\mathcal{A}$. Consequently, it suffices to analyze distributed function computation in bidirected chains. The key quantitative statement that emerges from this analysis can be informally stated as follows: For any bidirected chain with $n > 3$ nodes, there exists a constant $\eta \in [0, 1]$ that plays the same role as η_v in (3.4) and quantifies the information transmission capabilities of the channels in the chain, such that, for any algorithm $\mathcal{A}$ that runs on

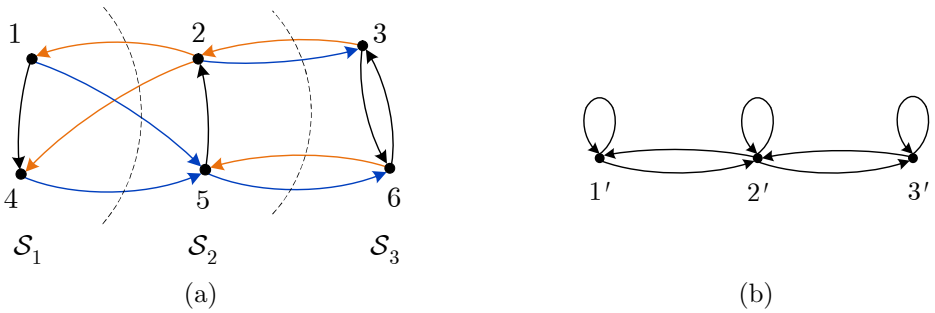


Figure 3.2: A six-node network partitioned into three sets, $\mathcal{S}_1 = \{1, 4\}$, $\mathcal{S}_2 = \{2, 5\}$, and $\mathcal{S}_3 = \{3, 6\}$. Here, $\mathcal{P}_1 = \{1, 4\}$, $\mathcal{P}_2 = \{1, 2, 4, 5\}$, and the cutsets $\mathcal{E}_{\mathcal{P}_1} = \{(2, 1), (2, 4)\}$, $\mathcal{E}_{\mathcal{P}_2} = \{(3, 2), (6, 5)\}$, $\mathcal{E}_{\mathcal{P}_1^c} = \{(1, 5), (4, 5)\}$, and $\mathcal{E}_{\mathcal{P}_2^c} = \{(2, 3), (5, 6)\}$ are disjoint. Observe that nodes in $\mathcal{S}_1$ communicate only with nodes in $\mathcal{S}_2$ and $\mathcal{S}_1$, and nodes in $\mathcal{S}_3$ communicate only with nodes in $\mathcal{S}_2$ and $\mathcal{S}_3$. The bidirected chain reduced from the network is shown on the right.

this chain and takes time $T = O(n/\eta)$, the conditional mutual information between the function value Z and its estimate $\widehat{Z}_n$ at the rightmost node n given the observations of nodes 2 through n is upper bounded by

$$I(Z; \widehat{Z}_n | W_{2:n}) = O\left(\frac{C_{(1,2)} n^2}{\eta} e^{-2n\eta^2}\right), \tag{3.7}$$

where $C_{(1,2)}$ is the Shannon capacity of the channel from node 1 to node 2. The precise statement is given in Lemma 3.8 in Sec. 3.3.1. Intuitively, this shows that, unless the algorithm uses $\Omega(n/\eta)$ steps, the information about W_1 will *dissipate* at an exponential rate by the time it propagates through the chain from node 1 to node n. Combining (3.7) with the lower bound on $I(Z; \widehat{Z}_n | W_{2:n})$ based on small ball probabilities, we can obtain lower bounds on the computation time $T(\varepsilon, \delta)$. The precise statement is given in Theorem 3.4. Moreover, as we show, it is always possible to reduce an arbitrary network with bidirectional point-to-point channels between the nodes to a bidirected chain whose length is equal to the *diameter* of the original network, which implies that, for networks with sufficiently large diameter, and for sufficiently small values of ε, δ,

$$T(\varepsilon, \delta) = \Omega\left(\frac{\operatorname{diam}(G)}{\eta}\right), \tag{3.8}$$

where $\operatorname{diam}(G)$ denotes the diameter. This dependence on $\operatorname{diam}(G)$, which

cannot be captured by the single-cutset analysis, is missing in almost all of the existing lower bounds on computation time. An exception is the paper by Rajagopalan and Schulman [50] that gives an asymptotic lower bound on the time required to broadcast a single bit over a chain of unidirectional BSCs. Our multi-cutset analysis applies to both discrete and continuous observations, and to general network topologies. It can be straightforwardly particularized to specific networks, such as bidirected chains, rings, trees, and grids, as discussed in Sec. 3.3.2. We note that techniques involving multiple (though not necessarily disjoint) cutsets have also been proposed in the study of multi-party communication complexity by Tiwari [57] and more recently by Chattopadhyay et al. [58], while our concern is the influence of network topology and channel noise on the computation time.

3.2 Single-cutset Analysis

We start by deriving information-theoretic lower bounds on the computation time $T(\varepsilon, \delta)$ based on a single cutset in the network. Recall that a *cutset* associated to a partition of $\mathcal{V}$ into two disjoint sets $\mathcal{S}$ and $\mathcal{S}^c \triangleq \mathcal{V} \setminus \mathcal{S}$ consists of all edges that connect a node in $\mathcal{S}^c$ to a node in $\mathcal{S}$:

$$\mathcal{E}_{\mathcal{S}} \triangleq \big\{(u,v) \in \mathcal{E} : u \in \mathcal{S}^c, v \in \mathcal{S}\big\} \equiv (\mathcal{S}^c \times \mathcal{S}) \cap \mathcal{E}.$$

When $\mathcal{S}$ is a singleton, i.e., $\mathcal{S} = \{v\}$, we will write $\mathcal{E}_v$ instead of the more clunky $\mathcal{E}_{\{v\}}$. As the discussion in Sec. 3.1.2 indicates, our analysis revolves around the conditional mutual information $I(Z; \widehat{Z}_v | W_{\mathcal{S}})$ for an arbitrary set of nodes $\mathcal{S} \subset \mathcal{V}$ and for an arbitrary node $v \in \mathcal{S}$. The lower bound on $I(Z; \widehat{Z}_v | W_{\mathcal{S}})$ expresses quantitatively the intuition that any algorithm that achieves

$$\max_{v \in \mathcal{V}} \mathbb{P}\big[\ell(Z, \widehat{Z}_v) > \varepsilon\big] \leq \delta$$

must necessarily extract a sufficient amount of information about the value of $Z = f(W) = f(W_{\mathcal{S}}, W_{\mathcal{S}^c})$. On the other hand, the upper bounds on $I(Z; \widehat{Z}_v | W_{\mathcal{S}})$ formalize the idea that this amount cannot be too large, since any information that nodes in $\mathcal{S}$ receive about $W_{\mathcal{S}^c}$ must flow across the edges in the cutset $\mathcal{E}_{\mathcal{S}}$ (cf. [25, Sec. 15.10] for a typical illustration of this type of

cutset arguments). We capture this information limitation in two ways: via channel capacity and via SDPI constants.

The remainder of this section is organized as follows. We first present conditional mutual information lower bounds in Sec. 3.2.1. Then we state the upper bound based on cutset capacity in Sec. 3.2.2 and the upper bounds based on SDPI in Sec. 3.2.3. Finally, we combine the lower and upper bounds to derive lower bounds on $T(\varepsilon, \delta)$ in Sec. 3.2.4.

3.2.1 Lower Bounds on $I(Z; \widehat{Z}_v | W_{\mathcal{S}})$

Lower Bound via Small Ball Probability

For any $\varepsilon \geq 0$, $\mathcal{S} \subset \mathcal{V}$, and $w_{\mathcal{S}} \in \prod_{v \in \mathcal{S}} \mathsf{W}_v$, according to the definition in (2.3), the conditional small ball probability of Z given $W_{\mathcal{S}} = w_{\mathcal{S}}$ is

$$\mathcal{L}_{Z|W_{\mathcal{S}}}(w_{\mathcal{S}}, \varepsilon) = \sup_{z \in \mathsf{Z}} \mathbb{P}[\ell(Z, z) \leq \varepsilon | W_{\mathcal{S}} = w_{\mathcal{S}}]. \tag{3.9}$$

This quantity measures how well the conditional distribution of Z given $W_{\mathcal{S}} = w_{\mathcal{S}}$ concentrates in a small region of size ε as measured by $\ell(\cdot, \cdot)$. A useful fact about the conditional small ball probability is the monotonicity of the set function $\mathcal{S} \mapsto \mathbb{E}[\mathcal{L}_{Z|W_{\mathcal{S}}}(W_{\mathcal{S}}, \varepsilon)]$: if $\mathcal{S} \subseteq \mathcal{S}' \subset \mathcal{V}$, then

$$\mathbb{E}[\mathcal{L}_{Z|W_{\mathcal{S}}}(W_{\mathcal{S}}, \varepsilon)] \leq \mathbb{E}[\mathcal{L}_Z(W_{\mathcal{S}'}, \varepsilon)]. \tag{3.10}$$

Indeed, by the law of iterated expectation, for any $\mathcal{S} \subseteq \mathcal{S}' \subset \mathcal{V}$ we have

$$\begin{aligned}
\mathcal{L}_{Z|W_{\mathcal{S}}}(w_{\mathcal{S}}, \varepsilon) &= \sup_{z \in \mathsf{Z}} \mathbb{E}\big[\mathbb{E}[\mathbf{1}\{\ell(Z, z) \leq \varepsilon\} | W_{\mathcal{S}'}] \big| W_{\mathcal{S}} = w_{\mathcal{S}}\big] \\
&\leq \mathbb{E}\Big[\sup_{z \in \mathsf{Z}} \mathbb{E}[\mathbf{1}\{\ell(Z, z) \leq \varepsilon\} | W_{\mathcal{S}'}] \Big| W_{\mathcal{S}} = w_{\mathcal{S}}\Big] \\
&= \mathbb{E}\big[\mathcal{L}_Z(W_{\mathcal{S}'}, \varepsilon) \big| W_{\mathcal{S}} = w_{\mathcal{S}}\big].
\end{aligned}$$

Integrating over $w_{\mathcal{S}}$, we obtain (3.10).

The following lower bound on $I(Z; \widehat{Z}_v | W_{\mathcal{S}})$ in terms of the conditional small ball probability is essential for proving lower bounds on $T(\varepsilon, \delta)$.

Lemma 3.1. *If an algorithm $\mathcal{A}$ achieves*

$$\max_{v\in\mathcal{V}} \mathbb{P}\big[\ell(Z,\widehat{Z}_v) > \varepsilon\big] \le \delta \le 1/2, \tag{3.11}$$

then for any set $\mathcal{S}\subset\mathcal{V}$ and any node $v\in\mathcal{S}$,

$$I(Z;\widehat{Z}_v|W_{\mathcal{S}}) \ge (1-\delta)\log\frac{1}{\mathbb{E}[\mathcal{L}_{Z|W_{\mathcal{S}}}(W_{\mathcal{S}},\varepsilon)]} - h_2(\delta), \tag{3.12}$$

where $h_2(\delta) = -\delta\log\delta - (1-\delta)\log(1-\delta)$ is the binary entropy function.

Proof. Fix an arbitrary $\mathcal{S}\subset\mathcal{V}$ and an arbitrary $v\in\mathcal{S}$. Consider the probability distributions $\mathbb{P} = \mathbb{P}_{W_{\mathcal{S}},Z,\widehat{Z}_v}$ and $\mathbb{Q} = \mathbb{P}_{W_{\mathcal{S}}}\otimes\mathbb{P}_{Z|W_{\mathcal{S}}}\otimes\mathbb{P}_{\widehat{Z}_v|W_{\mathcal{S}}}$. Define the indicator random variable $\Upsilon \triangleq \mathbf{1}\big\{\ell(Z,\widehat{Z}_v)\le\varepsilon\big\}$. Then from (3.11) it follows that $\mathbb{P}[\Upsilon=1]\ge 1-\delta$. On the other hand, since $Z\to W_{\mathcal{S}}\to\widehat{Z}_v$ form a Markov chain under $\mathbb{Q}$, by Fubini's theorem,

$$\begin{aligned}
\mathbb{Q}[\Upsilon=1] &= \int_{\mathsf{W}_{\mathcal{S}}}\int_{\mathsf{Z}}\int_{\mathsf{Z}} \mathbf{1}\big\{\ell(z,\widehat{z}_v)\le\varepsilon\big\}\mathbb{P}(\mathrm{d}z|w_{\mathcal{S}})\mathbb{P}(\mathrm{d}\widehat{z}_v|w_{\mathcal{S}})\mathbb{P}(\mathrm{d}w_{\mathcal{S}}) \\
&= \int_{\mathsf{W}_{\mathcal{S}}}\int_{\mathsf{Z}} \mathbb{P}\big[\ell(Z,\widehat{z}_v)\le\varepsilon\big|W_{\mathcal{S}}=w_{\mathcal{S}}\big]\mathbb{P}(\mathrm{d}\widehat{z}_v|w_{\mathcal{S}})\mathbb{P}(\mathrm{d}w_{\mathcal{S}}) \\
&\le \int_{\mathsf{W}_{\mathcal{S}}} \sup_{\widehat{z}_v\in\mathsf{Z}} \mathbb{P}\big[\ell(Z,\widehat{z}_v)\le\varepsilon\big|W_{\mathcal{S}}=w_{\mathcal{S}}\big]\mathbb{P}(\mathrm{d}w_{\mathcal{S}}) \\
&= \mathbb{E}[\mathcal{L}_{Z|W_{\mathcal{S}}}(W_{\mathcal{S}},\varepsilon)].
\end{aligned} \tag{3.13}$$

Consequently,

$$\begin{aligned}
I(Z;\widehat{Z}_v|W_{\mathcal{S}}) &= D(\mathbb{P}\|\mathbb{Q}) \\
&\overset{\text{(a)}}{\ge} d_2(\mathbb{P}[\Upsilon=1]\|\mathbb{Q}[\Upsilon=1]) \\
&\overset{\text{(b)}}{\ge} \mathbb{P}[\Upsilon=1]\log\frac{1}{\mathbb{Q}[\Upsilon=1]} - h_2(\mathbb{P}[\Upsilon=1]) \\
&\overset{\text{(c)}}{\ge} (1-\delta)\log\frac{1}{\mathbb{E}[\mathcal{L}_{Z|W_{\mathcal{S}}}(W_{\mathcal{S}},\varepsilon)]} - h_2(\delta),
\end{aligned}$$

where

(a) follows from the data processing inequality for divergence, where $d_2(p\|q)\triangleq p\log(p/q)+(1-p)\log((1-p)/(1-q))$ is the binary divergence function;

(b) follows from the fact that $d_2(p\|q)\ge p\log(1/q)-h_2(p)$; and

(c) follows from the fact that $\mathbb{P}[\Upsilon = 1] \geq 1 - \delta \geq 1/2$ by (3.11), and $\mathbb{Q}[\Upsilon = 1] \leq \mathbb{E}[\mathcal{L}_{Z|W_{\mathcal{S}}}(W_{\mathcal{S}}, \varepsilon)]$ by (3.13).

□

For a fixed ε, Lemma 3.1 captures the intuition that, the more spread the conditional distribution $\mathbb{P}_{Z|W_{\mathcal{S}}}$ is, the more information we need about Z to achieve the required accuracy; similarly, for a fixed $\mathbb{P}_{Z|W_{\mathcal{S}}}$, the smaller the accuracy parameter ε, the more information is necessary. In Sec. 3.4, we provide explicit expressions and upper bounds for the conditional small ball probability $\mathcal{L}_{Z|W_{\mathcal{S}}}(w_{\mathcal{S}}, \varepsilon)$ in the context of computing linear functions of real-valued r.v.'s with absolutely continuous probability distributions. We show that, in such cases, $\mathcal{L}_{Z|W_{\mathcal{S}}}(w_{\mathcal{S}}, \varepsilon) = O(\varepsilon)$, which implies that the lower bound of Lemma 3.1 grows at least as fast as $\log(1/\varepsilon)$ in the high-accuracy limit $\varepsilon \to 0$.

Lower Bound via Rate-distortion Functions

In Sec. 3.1.1, we have defined the (ε, δ)-computation time where the excess distortion probability is used as a measure of the computation fidelity. Alternatively, we can use the expected distortion $\mathbb{E}\big[\ell(Z, \widehat{Z}_v)\big]$ to quantify the computation fidelity of node v, and define the ε-computation time as

$$T(\varepsilon) \triangleq \inf \Big\{ T \in \mathbb{N} : \exists \mathcal{A} \in \mathfrak{A}(T) \text{ such that } \max_{v \in \mathcal{V}} \mathbb{E}\big[\ell(Z, \widehat{Z}_v)\big] \leq \varepsilon \Big\}. \quad (3.14)$$

Lemma 3.2 summarizes the relationship between the two definitions of computation time.

Lemma 3.2.

1. *For all $\varepsilon \geq 0$ and $\delta \in (0, 1)$,*

$$T(\varepsilon/\delta, \delta) \leq T(\varepsilon).$$

2. *If the distortion function is bounded, i.e., $d_{\max} \triangleq \max_{z, \widehat{z} \in \mathsf{Z}} \ell(z, \widehat{z}) < \infty$, then for all $\varepsilon \in [0, d_{\max}]$ and $\delta \in [0, 1)$,*

$$T(\varepsilon + \delta d_{\max}) \leq T(\varepsilon, \delta).$$

Proof. 1. If $\mathbb{E}\big[\ell(\widehat{Z}_v, Z)\big] \le \varepsilon$, then due to the non-negativity of the distortion function and Markov's inequality,

$$\mathbb{P}\big[\ell(\widehat{Z}_v, Z) > \varepsilon/\delta\big] \le \frac{\mathbb{E}\big[\ell(\widehat{Z}_v, Z)\big]}{\varepsilon/\delta} \le \delta.$$

2. If $\mathbb{P}\big[\ell(\widehat{Z}_v, Z) > \varepsilon\big] \le \delta$, then by the assumption that $0 \le \ell(z, \widehat{z}) \le d_{\max}$,

$$\begin{aligned}\mathbb{E}\big[\ell(\widehat{Z}_v, Z)\big] &= \int_0^\infty \mathbb{P}\big[\ell(\widehat{Z}_v, Z) > u\big]\mathrm{d}u \\ &= \int_0^\varepsilon \mathbb{P}\big[\ell(\widehat{Z}_v, Z) > u\big]\mathrm{d}u + \int_\varepsilon^{d_{\max}} \mathbb{P}\big[\ell(\widehat{Z}_v, Z) > u\big]\mathrm{d}u \\ &\le \varepsilon + \delta d_{\max}.\end{aligned}$$

□

To obtain lower bounds for the ε-computation time defined with respect to the expected distortion, we can make use of the conditional rate-distortion function of Z given $W_{\mathcal{S}}$ [59], defined as

$$R_{Z|W_{\mathcal{S}}}(\varepsilon) = \inf_{\mathbb{P}_{\widehat{Z}|Z,W_{\mathcal{S}}}:\, \mathbb{E}[\ell(Z,\widehat{Z})]\le\varepsilon} I(Z; \widehat{Z}|W_{\mathcal{S}}). \tag{3.15}$$

The operational meaning of $R_{Z|W_{\mathcal{S}}}(\varepsilon)$ is the minimum asymptotic rate to encode the source Z within expected distortion ε when both the encoder and the decoder are provided with the side information $W_{\mathcal{S}}$. Also denote the usual rate distortion function of Z by

$$R_Z(\varepsilon) = \inf_{\mathbb{P}_{\widehat{Z}|Z}:\, \mathbb{E}[\ell(Z,\widehat{Z})]\le\varepsilon} I(Z; \widehat{Z}).$$

Under the expected distortion criterion, $I(Z; \widehat{Z}_v|W_{\mathcal{S}})$ can be lower-bounded in terms of $R_{Z|W_{\mathcal{S}}}(\varepsilon)$ or $R_Z(\varepsilon)$. Moreover, when Z is continuous and the distortion function is quadratic, the lower bound only involves the conditional differential entropy of Z given $W_{\mathcal{S}}$, as stated in the following lemma.

Lemma 3.3. *If an algorithm achieves* $\max_{v\in\mathcal{V}} \mathbb{E}\big[\ell(Z, \widehat{Z}_v)\big] \le \varepsilon$*, then for any set* $\mathcal{S} \subset \mathcal{V}$ *and any node* $v \in \mathcal{S}$*,*

$$I(Z; \widehat{Z}_v|W_{\mathcal{S}}) \ge R_{Z|W_{\mathcal{S}}}(\varepsilon) \ge R_Z(\varepsilon) - I(Z; W_{\mathcal{S}}).$$

If Z is continuous and $\ell(z,\widehat{z}) = (z-\widehat{z})^2$, then under the same condition,

$$I(Z;\widehat{Z}_v|W_{\mathcal{S}}) \geq h(Z|W_{\mathcal{S}}) + \frac{1}{2}\log\frac{1}{2\pi e\varepsilon}.$$

Proof. For any set $\mathcal{S} \subset \mathcal{V}$ and any node $v \in \mathcal{S}$, an algorithm that achieves $\max_{v\in\mathcal{V}} \mathbb{E}\big[\ell(Z,\widehat{Z}_v)\big] \leq \varepsilon$ yields a conditional distribution $\mathbb{P}_{\widehat{Z}_v|Z,W_{\mathcal{S}}}$ that lies in the feasible set for the infimization in (3.15). Thus,

$$I(Z;\widehat{Z}_v|W_{\mathcal{S}}) \geq R_{Z|W_{\mathcal{S}}}(\varepsilon).$$

The conditional rate distortion function can be further lower bounded in terms of $R_Z(\varepsilon)$ [59],

$$R_{Z|W_{\mathcal{S}}}(\varepsilon) \geq R_Z(\varepsilon) - I(Z;W_{\mathcal{S}}).$$

If Z is continuous and $\ell(z,\widehat{z}) = (z-\widehat{z})^2$, then

$$\begin{aligned}
I(Z;\widehat{Z}_v|W_{\mathcal{S}}) &= h(Z|W_{\mathcal{S}}) - h(Z|\widehat{Z}_v, W_{\mathcal{S}}) \\
&= h(Z|W_{\mathcal{S}}) - h(Z-\widehat{Z}_v|\widehat{Z}_v, W_{\mathcal{S}}) \\
&\geq h(Z|W_{\mathcal{S}}) - h(Z-\widehat{Z}_v) \\
&\geq h(Z|W_{\mathcal{S}}) - \frac{1}{2}\log\left(2\pi e\mathbb{E}[(Z-\widehat{Z}_v)^2]\right) \\
&\geq h(Z|W_{\mathcal{S}}) + \frac{1}{2}\log\frac{1}{2\pi e\varepsilon},
\end{aligned}$$

where we have used the fact that conditioning reduces differential entropy and $h(Z-\widehat{Z}_v) \leq \frac{1}{2}\log\left(2\pi e\mathbb{E}[(Z-\widehat{Z}_v)^2]\right)$. This completes the proof of Lemma 3.3. □

3.2.2 Upper Bound on $I(Z;\widehat{Z}_v|W_{\mathcal{S}})$ via Cutset Capacity

Our first upper bound involves the *cutset capacity* $C_{\mathcal{S}}$, defined as

$$C_{\mathcal{S}} \triangleq \sum_{e\in\mathcal{E}_{\mathcal{S}}} C_e.$$

Here, C_e denotes the Shannon capacity of the channel K_e.

Lemma 3.4. *For any set $\mathcal{S} \subset \mathcal{V}$, let $\widehat{Z}_{\mathcal{S}} \triangleq (\widehat{Z}_v)_{v\in\mathcal{S}}$. Then, for any T-step*

algorithm $\mathcal{A}$ and for any $v \in \mathcal{S}$,

$$I(Z; \widehat{Z}_v | W_{\mathcal{S}}) \leq I(Z; \widehat{Z}_{\mathcal{S}} | W_{\mathcal{S}}) \leq TC_{\mathcal{S}}.$$

Proof. The first inequality follows from the data processing lemma for mutual information. The second inequality has been obtained in [38] and [39] as well, but the proof in [38] relies heavily on differential entropy. Our proof is more general, as it only uses the properties of mutual information.

For a set of nodes $\mathcal{S} \subset \mathcal{V}$, let $X_{\mathcal{S},t} \triangleq (X_{v,t})_{v \in \mathcal{S}}$ and $Y_{\mathcal{S},t} \triangleq (Y_{v,t})_{v \in \mathcal{S}}$. For two subsets $\mathcal{S}_1$ and $\mathcal{S}_2$ of $\mathcal{V}$, define $X_{(\mathcal{S}_1,\mathcal{S}_2),t} \triangleq \big(X_{(u,v),t} : u \in \mathcal{S}_1, v \in \mathcal{S}_2, (v,u) \in \mathcal{E}\big)$ as the messages sent from nodes in $\mathcal{S}_1$ to nodes in $\mathcal{S}_2$ at step t, and $Y_{(\mathcal{S}_1,\mathcal{S}_2),t} \triangleq \big(Y_{(u,v),t} : u \in \mathcal{S}_1, v \in \mathcal{S}_2, (u,v) \in \mathcal{E}\big)$ as the messages received by nodes in $\mathcal{S}_2$ from nodes in $\mathcal{S}_1$ at step t. We will be using this notation in the proofs that follow, as well.

If $T = 0$, then for any $v \in \mathcal{S}$, $\widehat{Z}_v = \psi_v(W_v)$, hence $I(Z; \widehat{Z}_{\mathcal{S}} | W_{\mathcal{S}}) \leq I(Z; W_{\mathcal{S}} | W_{\mathcal{S}}) = 0$. For $T \geq 1$, we start with the following chain of inequalities:

$$\begin{aligned}
I(Z; \widehat{Z}_{\mathcal{S}} | W_{\mathcal{S}}) &\overset{(a)}{\leq} I(W_{\mathcal{S}}, W_{\mathcal{S}^c}; W_{\mathcal{S}}, Y_{\mathcal{S}}^T | W_{\mathcal{S}}) \\
&= I(W_{\mathcal{S}^c}; Y_{\mathcal{S}}^T | W_{\mathcal{S}}) \\
&= \sum_{t=1}^{T} I(W_{\mathcal{S}^c}; Y_{\mathcal{S},t} | W_{\mathcal{S}}, Y_{\mathcal{S}}^{t-1}) \\
&\overset{(b)}{=} \sum_{t=1}^{T} I(W_{\mathcal{S}^c}; Y_{\mathcal{S},t} | W_{\mathcal{S}}, Y_{\mathcal{S}}^{t-1}, X_{\mathcal{S},t}) \\
&\leq \sum_{t=1}^{T} I(W_{\mathcal{S}^c}, X_{\mathcal{S}^c,t}; Y_{\mathcal{S},t} | W_{\mathcal{S}}, Y_{\mathcal{S}}^{t-1}, X_{\mathcal{S},t}) \\
&= \sum_{t=1}^{T} I(X_{\mathcal{S}^c,t}; Y_{\mathcal{S},t} | W_{\mathcal{S}}, Y_{\mathcal{S}}^{t-1}, X_{\mathcal{S},t}) + I(W_{\mathcal{S}^c}; Y_{\mathcal{S},t} | W_{\mathcal{S}}, Y_{\mathcal{S}}^{t-1}, X_{\mathcal{S},t}, X_{\mathcal{S}^c,t}) \\
&\overset{(c)}{=} \sum_{t=1}^{T} I(X_{\mathcal{S}^c,t}; Y_{\mathcal{S},t} | W_{\mathcal{S}}, Y_{\mathcal{S}}^{t-1}, X_{\mathcal{S},t}) \\
&\overset{(d)}{\leq} \sum_{t=1}^{T} I(X_{\mathcal{S}^c,t}; Y_{\mathcal{S},t} | X_{\mathcal{S},t}), \qquad (3.16)
\end{aligned}$$

where

(a) follows from data processing inequality, and the fact that $Z = f(W_{\mathcal{S}}, W_{\mathcal{S}^c})$ and $\widehat{Z}_v = \psi_v(W_v, Y_v^T)$;

(b) follows from the fact that $X_{v,t} = \varphi_{v,t}(W_v, Y_v^{t-1})$;

(c) follows from the memorylessness of the channels, hence the Markov chain $W_{\mathcal{S}^c}, W_{\mathcal{S}}, Y_{\mathcal{S}}^{t-1} \to X_{\mathcal{S},t}, X_{\mathcal{S}^c,t} \to Y_{\mathcal{S},t}$, and the weak union property of conditional independence [60, p. 25]; and

(d) follows from the Markov chain $W_{\mathcal{S}}, Y_{\mathcal{S}}^{t-1} \to X_{\mathcal{S},t}, X_{\mathcal{S}^c,t} \to Y_{\mathcal{S},t}$, together with the fact that, if $X \to A, B \to C$ form a Markov chain, then

$$I(A; C|X, B) \leq I(A; C|B).$$

To prove this, we expand $I\left(A, X; C|\, B\right)$ in two ways to get

$$\begin{aligned} I\left(A, X; C|B\right) &= I(X; C|B) + I\left(A; C|\, X, B\right) \\ &= I(A; C|B) + I\left(X; C|\, A, B\right). \end{aligned}$$

The claim follows from the fact that $I\left(X; C|\, A, B\right) = 0$.

From now on we drop the step index t and denote $X_{(\mathcal{S}_1,\mathcal{S}_2),t}$ as $X_{\mathcal{S}_1\mathcal{S}_2}$ to simplify the notation. Note that $X_{\mathcal{S}} = (X_{\mathcal{S}\mathcal{S}}, X_{\mathcal{S}\mathcal{S}^c})$ and $Y_{\mathcal{S}} = (Y_{\mathcal{S}\mathcal{S}}, Y_{\mathcal{S}^c\mathcal{S}})$. We have

$$\begin{aligned} I(X_{\mathcal{S}^c}; Y_{\mathcal{S}}|X_{\mathcal{S}}) &= I(X_{\mathcal{S}^c}; Y_{\mathcal{S}^c\mathcal{S}}, Y_{\mathcal{S}\mathcal{S}}|X_{\mathcal{S}}) \\ &= I(X_{\mathcal{S}^c}; Y_{\mathcal{S}^c\mathcal{S}}|X_{\mathcal{S}}) + I(X_{\mathcal{S}^c}; Y_{\mathcal{S}\mathcal{S}}|X_{\mathcal{S}}, Y_{\mathcal{S}^c\mathcal{S}}) \\ &\overset{\text{(a)}}{=} I(X_{\mathcal{S}^c\mathcal{S}}, X_{\mathcal{S}^c\mathcal{S}^c}; Y_{\mathcal{S}^c\mathcal{S}}|X_{\mathcal{S}}) \\ &= I(X_{\mathcal{S}^c\mathcal{S}}; Y_{\mathcal{S}^c\mathcal{S}}|X_{\mathcal{S}}) + I(X_{\mathcal{S}^c\mathcal{S}^c}; Y_{\mathcal{S}^c\mathcal{S}}|X_{\mathcal{S}}, X_{\mathcal{S}^c\mathcal{S}}) \\ &\overset{\text{(b)}}{\leq} I(X_{\mathcal{S}^c\mathcal{S}}; Y_{\mathcal{S}^c\mathcal{S}}) \\ &\overset{\text{(c)}}{\leq} \sum_{e \in \mathcal{E}_{\mathcal{S}}} C_e, \end{aligned} \tag{3.17}$$

where

(a) follows from the Markov chain $X_{\mathcal{S}^c}, Y_{\mathcal{S}^c\mathcal{S}} \to X_{\mathcal{S}} \to Y_{\mathcal{S}\mathcal{S}}$ and the weak union property of conditional independence;

(b) follows from the Markov chains $X_{\mathcal{S}} \to X_{\mathcal{S}^c\mathcal{S}} \to Y_{\mathcal{S}^c\mathcal{S}}$ and $X_{\mathcal{S}^c\mathcal{S}^c}, X_{\mathcal{S}} \to X_{\mathcal{S}^c\mathcal{S}} \to Y_{\mathcal{S}^c\mathcal{S}}$, and the weak union property of conditional independence; and

(c) follows from the fact that the channels associated with $\mathcal{E}_{\mathcal{S}}$ are independent, and the fact that the capacity of a product channel is at most the sum of the capacities of the constituent channels [34].

Then the statement of Lemma 3.4 follows from (3.16) and (3.17). □

3.2.3 Upper Bound on $I(Z; \widehat{Z}_v|W_{\mathcal{S}})$ via SDPI

In Sec. 2.3, we have introduced the necessary background on strong data processing inequalities (SDPIs). We now state our upper bounds on $I(Z; \widehat{Z}_v|W_{\mathcal{S}})$ based on SDPI constants. Let $K_v \triangleq \bigotimes_{e\in\mathcal{E}_v} K_e$ be the overall transition law of the channels across the cutset $\mathcal{E}_v$. Define

$$\eta_v \triangleq \eta(K_v)$$

as the SDPI constant of K_v according to (2.43), and

$$\eta_v^* \triangleq \max_{e\in\mathcal{E}_v} \eta(K_e)$$

as the largest SDPI constant among all the channels across $\mathcal{E}_v$. Our second upper bound on $I(Z; \widehat{Z}_v|W_{\mathcal{S}})$ involves these SDPI constants, and the conditional entropy of $W_{\mathcal{S}^c}$ given $W_{\mathcal{S}}$.

Lemma 3.5. *For any set $\mathcal{S} \subset \mathcal{V}$, any node $v \in \mathcal{S}$, and any T-step algorithm $\mathcal{A}$,*

$$\begin{aligned} I(Z; \widehat{Z}_v|W_{\mathcal{S}}) &\le (1 - (1-\eta_v)^T) H(W_{\mathcal{S}^c}|W_{\mathcal{S}}) \\ &\le (1 - (1-\eta_v^*)^{|\mathcal{E}_v|T}) H(W_{\mathcal{S}^c}|W_{\mathcal{S}}). \end{aligned}$$

Proof. We adapt the proof of Lemma 2.10. For any v and t, define the shorthand $X_{v\leftarrow,t} \triangleq X_{(\mathcal{N}_{v\leftarrow},v),t}$. If $T = 0$, then for any $v \in \mathcal{S}$, $\widehat{Z}_v = \psi_v(W_v)$;

hence $I(Z;\widehat{Z}_v|W_{\mathcal{S}}) \le I(Z;W_v|W_{\mathcal{S}}) = 0$. If $T \ge 1$, then for any $v \in \mathcal{S}$,

$$\begin{aligned} I(Z;\widehat{Z}_v|W_{\mathcal{S}}) &\le I(W_{\mathcal{S}}, W_{\mathcal{S}^c}; W_v, Y_v^T|W_{\mathcal{S}}) \\ &= I(W_{\mathcal{S}^c}; Y_v^T|W_{\mathcal{S}}) \\ &= I(W_{\mathcal{S}^c}; Y_v^{T-1}|W_{\mathcal{S}}) + I(W_{\mathcal{S}^c}; Y_{v,T}|W_{\mathcal{S}}, Y_v^{T-1}) \\ &\le I(W_{\mathcal{S}^c}; Y_v^{T-1}|W_{\mathcal{S}}) + \eta_v I(W_{\mathcal{S}^c}; X_{v\leftarrow,T}|W_{\mathcal{S}}, Y_v^{T-1}) \\ &= (1-\eta_v) I(W_{\mathcal{S}^c}; Y_v^{T-1}|W_{\mathcal{S}}) + \eta_v I(W_{\mathcal{S}^c}; Y_v^{T-1}, X_{v\leftarrow,T}|W_{\mathcal{S}}), \end{aligned} \tag{3.18}$$

where the fourth line follows from the conditional SDPI (Lemma 2.7) and the fact that $W_{\mathcal{S}^c}, W_{\mathcal{S}}, Y_v^{t-1} \to X_{v\leftarrow,t} \to Y_{v,t}$ form a Markov chain for $t \in \{1,\dots,T\}$. Unrolling the recursive upper bound (3.18) on $I(W_{\mathcal{S}^c}; Y_v^T|W_{\mathcal{S}})$, and noting that $I(W_{\mathcal{S}^c}; Y_{v,1}|W_{\mathcal{S}}) \le \eta_v I(W_{\mathcal{S}^c}; X_{v\leftarrow,1}|W_{\mathcal{S}})$, we get

$$\begin{aligned} I(W_{\mathcal{S}^c}; Y_v^T|W_{\mathcal{S}}) &\le (1-\eta_v)^{T-1}\eta_v I(W_{\mathcal{S}^c}; X_{v\leftarrow,1}|W_{\mathcal{S}}) + \dots + \\ &\quad (1-\eta_v)\eta_v I(W_{\mathcal{S}^c}; Y_v^{T-2}, X_{v\leftarrow,T-1}|W_{\mathcal{S}}) + \eta_v I(W_{\mathcal{S}^c}; Y_v^{T-1}, X_{v\leftarrow,T}|W_{\mathcal{S}}) \\ &\le \big((1-\eta_v)^{T-1} + \dots + (1-\eta_v) + 1\big)\eta_v H(W_{\mathcal{S}^c}|W_{\mathcal{S}}) \\ &= \big(1-(1-\eta_v)^T\big) H(W_{\mathcal{S}^c}|W_{\mathcal{S}}). \end{aligned}$$

The weakened upper bound follows from the fact that $\eta_v \le 1-(1-\eta_v^*)^{|\mathcal{E}_v|}$, due to Lemma 2.10. This completes the proof of Lemma 3.5. □

Comparing Lemma 3.4 and Lemma 3.5, we note that the upper bound in Lemma 3.4 captures the communication constraints through the cutset capacity alone, in accordance with the fact that the communication constraints do not depend on W or Z. The bound applies when W is either discrete or continuous; however, it grows linearly with T. By contrast, the upper bound in Lemma 3.5 builds on the fact that $I(Z;\widehat{Z}_v|W_{\mathcal{S}})$ is upper bounded by $H(W_{\mathcal{S}^c}|W_{\mathcal{S}})$, and goes a step further by capturing the communication constraint through a multiplicative contraction of $H(W_{\mathcal{S}^c}|W_{\mathcal{S}})$. It never exceeds $H(W_{\mathcal{S}^c}|W_{\mathcal{S}})$ as T increases. However, it is useful only when the conditional entropy $H(W_{\mathcal{S}^c}|W_{\mathcal{S}})$ is well-defined and finite (e.g., when W is discrete). We give an explicit comparison of Lemma 3.4 and Lemma 3.5 in the following example:

Example 3.1. *Consider a two-node network, where the nodes are connected by BSCs. The problem is for the two nodes to compute the mod-2 sum of*

their one-bit observations. Formally, we have $G = (\mathcal{V}, \mathcal{E})$ *with* $\mathcal{V} = \{1, 2\}$, $\mathcal{E} = \{(1, 2), (2, 1)\}$, $K_{(1,2)} = K_{(2,1)} = \mathrm{BSC}(p)$, W_1 *and* W_2 *are independent* $\mathrm{Bern}(\frac{1}{2})$ *r.v.'s,* $Z = W_1 \oplus W_2$, *and* $\ell(z, \widehat{z}) = \mathbf{1}\{z \neq \widehat{z}\}$.

Choosing $\mathcal{S} = \{2\}$, Lemma 3.4 gives

$$I(Z; \widehat{Z}_2 | W_2) \leq (1 - h_2(p))T, \tag{3.19}$$

whereas Lemma 3.5, together with the fact that $\eta(\mathrm{BSC}(p)) = (1-2p)^2$, gives

$$I(Z; \widehat{Z}_2 | W_2) \leq 1 - (4p\bar{p})^T, \tag{3.20}$$

where, for $p \in [0, 1]$, $\bar{p} \triangleq 1 - p$. For this example, the cutset-capacity upper bound is always tighter for small T, as

$$\frac{\partial\big(1 - (4p\bar{p})^T\big)}{\partial T}\Big|_{T=0} = \log \frac{1}{4p\bar{p}} \geq 1 - h_2(p), \quad p \in [0, 1].$$

Fig. 3.3 shows the two upper bounds with $p = 0.3$: the cutset-capacity upper bound is tighter when $T < 5$.

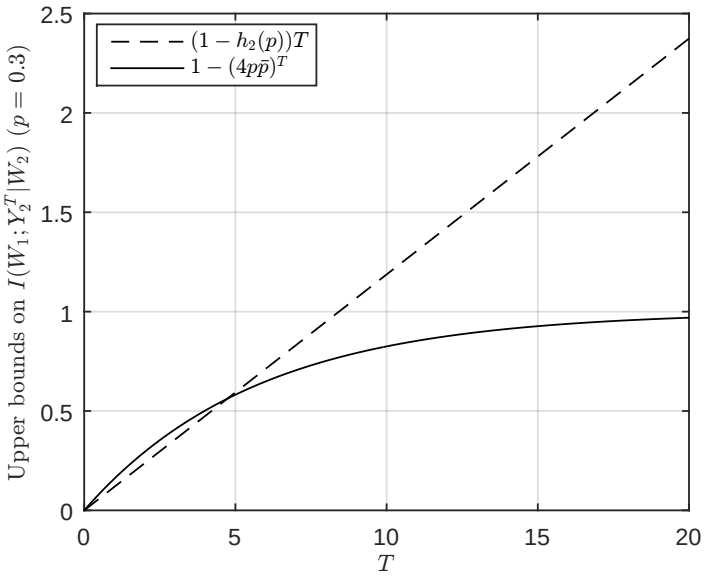


Figure 3.3: Comparison of upper bounds in Lemma 3.4 and Lemma 3.5 for computing mod-2 sum in a two-node network.

3.2.4 Lower Bounds on Computation Time

We now proceed to derive lower bounds on the computation time $T(\varepsilon, \delta)$ and $T(\varepsilon)$ based on the previously derived lower and upper bounds on the conditional mutual information $I(Z; \widehat{Z}_v | W_{\mathcal{S}})$. Define the shorthand notation

$$\mathcal{I}(\mathcal{S}, \varepsilon, \delta) \triangleq (1-\delta) \log \frac{1}{\mathbb{E}[\mathcal{L}_{Z|W_{\mathcal{S}}}(W_{\mathcal{S}}, \varepsilon)]} - h_2(\delta),$$

which is the lower bound on $I(Z; \widehat{Z}_v | W_{\mathcal{S}})$ in Lemma 3.1.

Cutset-capacity Bounds

Combined with the conditional small ball probability lower bound in Lemma 3.1, the cutset-capacity upper bound in Lemma 3.4 leads to a lower bound on $T(\varepsilon, \delta)$:

Theorem 3.1. *For an arbitrary network, for any $\varepsilon \geq 0$ and $\delta \in [0, 1/2]$,*

$$T(\varepsilon, \delta) \geq \max_{\mathcal{S} \subset \mathcal{V}} \frac{\mathcal{I}(\mathcal{S}, \varepsilon, \delta)}{C_{\mathcal{S}}}.$$

Combined with the rate-distortion lower bound in Lemma 3.3, the cutset-capacity upper bound leads to a lower bound on $T(\varepsilon)$:

Theorem 3.2. *For an arbitrary network,*

$$T(\varepsilon) \geq \max_{\mathcal{S} \subset \mathcal{V}} \frac{R_Z(\varepsilon) - I(Z; W_{\mathcal{S}})}{C_{\mathcal{S}}}.$$

If Z is continuous and $\ell(z, \widehat{z}) = (z - \widehat{z})^2$, then

$$T(\varepsilon) \geq \max_{\mathcal{S} \subset \mathcal{V}} \frac{1}{C_{\mathcal{S}}} \left(h(Z|W_{\mathcal{S}}) + \frac{1}{2} \log \frac{1}{2\pi e \varepsilon} \right).$$

From an operational point of view, the lower bound of Theorem 3.1 reflects the fact that the problem of distributed function computation is, in a certain sense, a joint source-channel coding (JSCC) problem with possibly noisy feedback. In particular, the lower bound on $I(Z; \widehat{Z}_v | W_{\mathcal{S}})$ from Lemma 3.1, which is used to prove Theorem 3.1, can be interpreted in terms of a reduction of JSCC to generalized list decoding [61, Sec. III.B]. Given

any algorithm $\mathcal{A}$ and any node $v \in \mathcal{V}$, we may construct a "list decoder" as follows: given the estimate $\widehat{Z}_v$, we generate a "list" $\{z \in \mathsf{Z} : \ell(z, \widehat{Z}_v) \leq \varepsilon\}$. If we fix a set $\mathcal{S} \subset \mathcal{V}$ and allow all the nodes in $\mathcal{S}$ to share their observations $W_{\mathcal{S}}$, then $\mathbb{E}[\mathcal{L}_{Z|W_{\mathcal{S}}}(W_{\mathcal{S}}, \varepsilon)]$ is an upper bound on the $\mathbb{P}_W$-measure of the list of any node $v \in \mathcal{S}$. Therefore, $\mathcal{I}(\mathcal{S}, \varepsilon, \delta)$ is a lower bound on the total amount of information that is necessary for the JSCC problem. The complementary cutset upper bound on $I(Z; \widehat{Z}_v|W_{\mathcal{S}})$ bounds the amount of information that can be accumulated with each channel use. The lower bound on $T(\varepsilon, \delta)$ can thus be interpreted as a lower bound on the blocklength of the JSCC problem. Similarly, the lower bound from Lemma 3.3 involves the (conditional) rate-distortion function, which quantifies the asymptotic fundamental limits of lossy source coding. The complementary upper bounds on $I(Z; \widehat{Z}_v|W_{\mathcal{S}})$ involve channel capacity, which quantifies the asymptotic fundamental limits of channel coding.

As we will demonstrate in Sec. 3.4, based on Theorem 3.1, it is possible to exploit structural properties of the function f (such as linearity) and of the probability law $\mathbb{P}_W$ (such as log-concavity) to derive lower bounds on the computation time that are often tighter than existing bounds.

SDPI Bounds

Combining the lower bound of Lemma 3.1 with the SDPI upper bound of Lemma 3.5, we get the following:

Theorem 3.3. *For an arbitrary network, for any $\varepsilon \geq 0$ and $\delta \in [0, 1/2]$,*

$$T(\varepsilon, \delta) \geq \max_{\mathcal{S} \subset \mathcal{V}} \max_{v \in \mathcal{S}} \frac{\log\left(1 - \frac{\mathcal{I}(\mathcal{S}, \varepsilon, \delta)}{H(W_{\mathcal{S}^c}|W_{\mathcal{S}})}\right)^{-1}}{|\mathcal{E}_v| \log(1 - \eta_v^*)^{-1}}, \tag{3.21}$$

where $\eta_v^* \triangleq \max_{e \in \mathcal{E}_v} \eta(K_e)$.

We can obtain a lower bound on $T(\varepsilon)$ of the same form by replacing $\mathcal{I}(\mathcal{S}, \varepsilon, \delta)$ in (3.21) with the lower bounds on $I(Z; \widehat{Z}_v|W_{\mathcal{S}})$ in Lemma 3.3.

The lower bounds in Theorem 3.1 and Theorem 3.3 for $T(\varepsilon, \delta)$ can behave quite differently. To illustrate this, we compare them in two cases:

When $H(W_{\mathcal{S}^c}|W_{\mathcal{S}}) \gg \log \frac{1}{\mathbb{E}[\mathcal{L}_{Z|W_{\mathcal{S}}}(W_{\mathcal{S}},\varepsilon)]}$, Theorem 3.3 gives

$$\begin{aligned} T(\varepsilon,\delta) &\geq \max_{\mathcal{S}\subset\mathcal{V}} \max_{v\in\mathcal{S}} \frac{\log\left(1-\frac{\mathcal{I}(\mathcal{S},\varepsilon,\delta)}{H(W_{\mathcal{S}^c}|W_{\mathcal{S}})}\right)^{-1}}{|\mathcal{E}_v|\log(1-\eta_v^*)^{-1}} \\ &\approx \max_{\mathcal{S}\subset\mathcal{V}} \max_{v\in\mathcal{S}} \frac{\mathcal{I}(\mathcal{S},\varepsilon,\delta)\log e}{H(W_{\mathcal{S}^c}|W_{\mathcal{S}})|\mathcal{E}_v|\log(1-\eta_v^*)^{-1}}, \end{aligned}$$

which has essentially the same dependence on $\mathcal{I}(\mathcal{S},\varepsilon,\delta)$ as the lower bound given by Theorem 3.1. In this case, Theorem 3.1 gives more useful lower bounds as long as $C_{\mathcal{S}} \ll H(W_{\mathcal{S}^c}|W_{\mathcal{S}})$, especially when W is continuous.

When $H(W_{\mathcal{S}^c}|W_{\mathcal{S}}) \approx \log \frac{1}{\mathbb{E}[\mathcal{L}_{Z|W_{\mathcal{S}}}(W_{\mathcal{S}},\varepsilon)]}$ and δ is small, $H(W_{\mathcal{S}^c}|W_{\mathcal{S}})$ serves as a sharp proxy of $\mathcal{I}(\mathcal{S},\varepsilon,\delta)$. Theorem 3.1 in this case gives

$$T(\varepsilon,\delta) \geq \max_{\mathcal{S}\subset\mathcal{V}} \frac{\mathcal{I}(\mathcal{S},\varepsilon,\delta)}{C_{\mathcal{S}}} \approx \max_{\mathcal{S}\subset\mathcal{V}} \frac{H(W_{\mathcal{S}^c}|W_{\mathcal{S}})}{C_{\mathcal{S}}},$$

while Theorem 3.3 gives

$$\begin{aligned} T(\varepsilon,\delta) &\geq \max_{\mathcal{S}\subset\mathcal{V}} \max_{v\in\mathcal{S}} \frac{\log\left(1-\frac{\mathcal{I}(\mathcal{S},\varepsilon,\delta)}{H(W_{\mathcal{S}^c}|W_{\mathcal{S}})}\right)^{-1}}{|\mathcal{E}_v|\log(1-\eta_v^*)^{-1}} \\ &\approx \max_{\mathcal{S}\subset\mathcal{V}} \max_{v\in\mathcal{S}} \frac{\log H(W_{\mathcal{S}^c}|W_{\mathcal{S}}) + \log\frac{1}{h_2(\delta)}}{|\mathcal{E}_v|\log(1-\eta_v^*)^{-1}}, \end{aligned}$$

where in the last step we have used the fact that $\log\left(\delta + \frac{h_2(\delta)}{H(W_{\mathcal{S}^c}|W_{\mathcal{S}})}\right) \sim \log\left(\frac{h_2(\delta)}{H(W_{\mathcal{S}^c}|W_{\mathcal{S}})}\right)$ as $\delta \to 0$. Theorem 3.1 in this case is sharper in capturing the dependence of $T(\varepsilon,\delta)$ on the amount of information contained in Z, in that the lower bound is proportional to $H(W_{\mathcal{S}^c}|W_{\mathcal{S}})$, whereas the lower bound given by Theorem 3.3 depends on $H(W_{\mathcal{S}^c}|W_{\mathcal{S}})$ only through $\log H(W_{\mathcal{S}^c}|W_{\mathcal{S}})$. On the other hand, Theorem 3.3 in this case is much sharper in capturing the dependence of $T(\varepsilon,\delta)$ on the confidence parameter δ, since $\log h_2(\delta)$ grows without bound as $\delta \to 0$, while the lower bound given by Theorem 3.1 remains bounded. We consider two examples for this case.

The first is Example 3.1 in Sec. 3.2.3, for the two-node mod-2 sum problem. We have $\mathcal{L}_{Z|W_2}(w_2,0) = \max_{z\in\{0,1\}} \mathbb{P}[W_1 \oplus W_2 = z|W_2 = w_2] = \frac{1}{2}$, and $\mathcal{I}(\mathcal{S},0,\delta) = 1-\delta-h_2(\delta)$. Theorems 3.1 and 3.3 imply the following:

Corollary 3.1. *For the problem in Example 3.1, for $\delta \in [0,1/2]$, the $(0,\delta)$-*

computation time satisfies

$$T(0,\delta) \geq \max\left\{\frac{1-\delta-h_2(\delta)}{1-h_2(p)}, \frac{\log(\delta+h_2(\delta))^{-1}}{\log(4p\bar{p})^{-1}}\right\}, \tag{3.22}$$

where the first lower bound is given by Theorem 3.1, and the second one is given by Theorem 3.3.

To obtain an achievable upper bound on $T(0,\delta)$ in Example 3.1, we consider the algorithm where each node uses a length-T repetition code to send its one-bit observation to the other node. Using the Chernoff bound, as in [27], it can be shown that the probability of decoding error at each node is upper-bounded by $(4p\bar{p})^{T/2}$, and therefore this algorithm achieves accuracy $\varepsilon = 0$ with confidence parameter $\delta \leq (4p\bar{p})^{T/2}$. This gives the upper bound

$$T(0,\delta) \leq \frac{2\log(\delta)^{-1}}{\log(4p\bar{p})^{-1}}. \tag{3.23}$$

Comparing (3.23) with the second lower bound in (3.22), we see that they asymptotically differ only by a factor of 2 as $\delta \to 0$, as $\lim_{\delta\to 0} \log(\delta + h_2(\delta))/\log(\delta) = 1$. Thus, for the problem in Example 3.1, the converse lower bound on $T(0,\delta)$ obtained from the SDPI closely matches the achievable upper bound on $T(0,\delta)$.

The second example concerns the problem of disseminating all of the observations through an arbitrary network:

Example 3.2. *Consider the problem where W_v's are i.i.d. samples from the uniformly distribution over $\{1,\ldots,M\}$, $Z = W$, and $\ell(z,\widehat{z}) = \mathbf{1}\{z \neq \widehat{z}\}$. In other words, the goal of the nodes is to distribute their observations to all other nodes.*

In this example, $H(W_{\mathcal{S}^c}|W_{\mathcal{S}}) = |\mathcal{S}^c|\log M$, and $\mathcal{I}(\mathcal{S},0,\delta) = (1-\delta)|\mathcal{S}^c|\log M - h_2(\delta)$. Following Ayaso et al. [38, Def. III.4], we define the *conductance* of the network G as

$$\Phi(G) \triangleq \min_{\mathcal{S}\in\mathcal{V}:|\mathcal{V}|/2<|\mathcal{S}|<|\mathcal{V}|} \frac{C_{\mathcal{S}}}{|\mathcal{S}^c|}.$$

Then we have the following corollary:

Corollary 3.2. *For the problem in Example 3.2, Theorem 3.1 gives*

$$T(0,\delta) \geq \max_{\mathcal{S}\subset\mathcal{V}} \frac{(1-\delta)|\mathcal{S}^c|\log M - h_2(\delta)}{C_{\mathcal{S}}} \tag{3.24}$$

$$\gtrsim \frac{\log M}{\Phi(G)} \qquad \text{as } \delta \to 0, \tag{3.25}$$

whereas Theorem 3.3 gives

$$T(0,\delta) \gtrsim \max_{\mathcal{S}\subset\mathcal{V}} \max_{v\in\mathcal{S}} \frac{\log\left(|\mathcal{S}^c|\log M\right) + \log h_2(\delta)^{-1}}{|\mathcal{E}_v|\log(1-\eta_v^*)^{-1}} \qquad \text{as } \delta \to 0. \tag{3.26}$$

Again, we see that the lower bound obtained from SDPI is much sharper for capturing the dependence of $T(0,\delta)$ on δ, since $\log h_2(\delta)^{-1} \to +\infty$ as $\delta \to 0$. On the other hand, the lower bound obtained from the cutset capacity upper bound is tighter in its dependence on M, and can also capture the dependence on the conductance of the network.

Finally, we point out that Theorem 3.1 gives the correct lower bound $T(\varepsilon,\delta) = +\infty$ when the network graph G is disconnected (assuming f depends on the observations of all nodes): If $\mathcal{V}$ consists of two disconnected components $\mathcal{S}$ and $\mathcal{S}^c$, then $C_{\mathcal{S}} = 0$, which results in $T(\varepsilon,\delta) = +\infty$. Despite the sharp dependence of the lower bounds of Theorems 3.1 and 3.3 on ε and δ, they have the same limitation as all previously known bounds obtained via single-cutset arguments: they examine only the flow of information across a cutset $\mathcal{E}_{\mathcal{S}}$, but not within $\mathcal{S}$; hence they cannot capture the dependence of computation time on the diameter of the network. We address this limitation in the following section.

3.3 Multi-cutset Analysis

We now extend the techniques of Sec. 3.2 to a multi-cutset analysis, to address the limitation of the results obtained from the single-cutset analysis. In particular, the new results are able to quantify the dissipation of information as it flows across a succession of cutsets in the network. As briefly sketched in Sec. 3.1.2, we accomplish this by partitioning a general network using multiple disjoint cutsets, such that the operation of any algorithm on the network can be simulated by another algorithm running on a chain of bidirectional

noisy links. We then derive tight mutual information upper bounds for such chains, which in turn can be used to lower-bound the computation time for the original network.

3.3.1 Network Reduction

Consider an arbitrary network $G = (\mathcal{V}, \mathcal{E})$. If there exists a collection of nested subsets $\mathcal{P}_1 \subset \ldots \subset \mathcal{P}_{n-1}$ of $\mathcal{V}$, such that the associated cutsets $\mathcal{E}_{\mathcal{P}_1}, \ldots, \mathcal{E}_{\mathcal{P}_{n-1}}$ are disjoint, and the cutsets $\mathcal{E}_{\mathcal{P}_1^c}, \ldots, \mathcal{E}_{\mathcal{P}_{n-1}^c}$ are also disjoint, then we say that G is *successively partitioned* according to $\mathcal{P}_1, \ldots, \mathcal{P}_{n-1}$ into n subsets $\mathcal{S}_1, \ldots, \mathcal{S}_n$, where $\mathcal{S}_i = \mathcal{P}_i \setminus \mathcal{P}_{i-1}$, with $\mathcal{P}_0 \triangleq \varnothing$ and $\mathcal{P}_n \triangleq \mathcal{V}$. For $i \in \{2, \ldots, n\}$, a node in $\mathcal{S}_i$ is called a left-bound node of $\mathcal{S}_i$ if there is an edge from it to a node in $\mathcal{S}_{i-1}$. The set of left-bound nodes of $\mathcal{S}_i$ is denoted by $\overleftarrow{\partial}\mathcal{S}_i$. For $\mathcal{S}_1$, define $\overleftarrow{\partial}\mathcal{S}_1 = \{v\}$ for an arbitrary $v \in \mathcal{S}_1$. In addition, for $i \in \{2, \ldots, n\}$, let

$$d_i \triangleq |\mathcal{E}_{\mathcal{P}_{i-1}^c}| + |\mathcal{E}_{\mathcal{P}_i}| + |\{\mathcal{E} \cap (\mathcal{S}_i \times \overleftarrow{\partial}\mathcal{S}_i)\}| \tag{3.27}$$

be the number of edges entering $\mathcal{S}_i$ from its neighbors $\mathcal{S}_{i-1}$ and $\mathcal{S}_{i+1}$, plus the number of edges entering $\overleftarrow{\partial}\mathcal{S}_i$ from $\mathcal{S}_i$ itself. For example, Fig. 3.2a in Sec. 3.1.2 illustrates a successive partition of a six-node network into three subsets $\mathcal{S}_1 = \{1, 4\}$, $\mathcal{S}_2 = \{2, 5\}$, and $\mathcal{S}_3 = \{3, 6\}$, with $\overleftarrow{\partial}\mathcal{S}_1 = \{4\}$, $\overleftarrow{\partial}\mathcal{S}_2 = \{2\}$, and $\overleftarrow{\partial}\mathcal{S}_3 = \{3, 6\}$. In addition, $d_2 = 5$ and $d_3 = 4$. As another example, the network in Fig. 3.4a, where each undirected edge represents a pair of channels with opposite directions, can be successively partitioned into $\mathcal{S}_1 = \{1\}$, $\mathcal{S}_2 = \{2, 7\}$, $\mathcal{S}_3 = \{3, 6, 8, 9\}$, $\mathcal{S}_4 = \{4, 10\}$, and $\mathcal{S}_5 = \{5\}$, with $\overleftarrow{\partial}\mathcal{S}_1 = \{1\}$, $\overleftarrow{\partial}\mathcal{S}_1 = \{2, 7\}$, $\overleftarrow{\partial}\mathcal{S}_3 = \{3, 8\}$, $\overleftarrow{\partial}\mathcal{S}_4 = \{4, 10\}$, and $\overleftarrow{\partial}\mathcal{S}_i = \{5\}$. In addition, $d_2 = 6$, $d_3 = 7$, $d_4 = 6$, and $d_5 = 2$.

Formally, a network G has *bidirectional links* if, for any pair of nodes $u, v \in \mathcal{V}$, $(u, v) \in \mathcal{E}$ if and only if $(v, u) \in \mathcal{E}$. A *path* between u and v is a sequence of edges $\{(v_i, v_{i+1})\}_{i=1}^{k-1}$, such that $v_1 = u$ and $v_k = v$ (if G is connected, there is at least one path between any pair of nodes). The *graph distance* between u and v, denoted by $d_G(u, v)$, is the length of a shortest path between u and v (shortest paths are not necessarily unique). The *diameter*

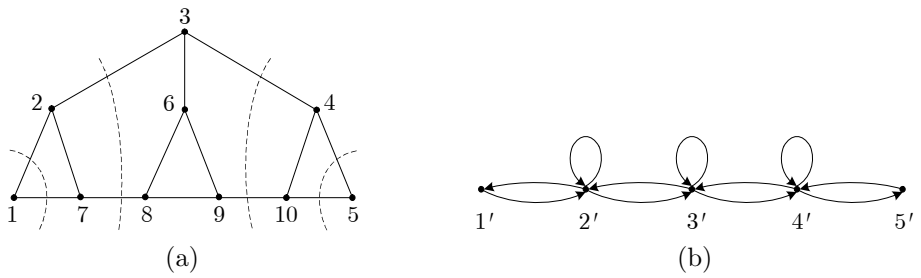


Figure 3.4: A successive partition of a network and the chain reduced according to it.

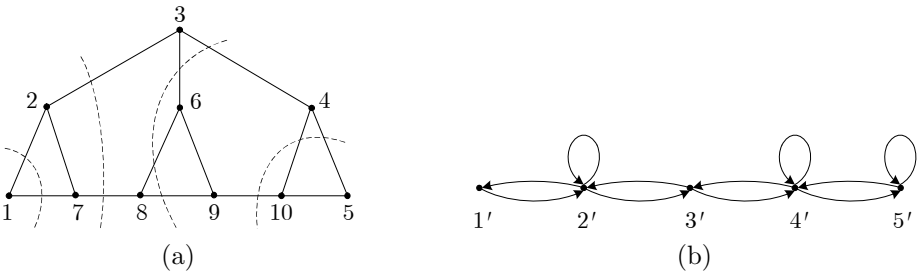


Figure 3.5: Another successive partition (using the construction in the proof of Lemma 3.6) and the chain reduced according to it.

of G is then defined by

$$\operatorname{diam}(G) \triangleq \max_{u\in\mathcal{V}} \max_{v\in\mathcal{V}} d_G(u,v).$$

The following lemma states that any such network G can be successively partitioned into $n = \operatorname{diam}(G) + 1$ subsets:

Lemma 3.6. *Any network $G = (\mathcal{V}, \mathcal{E})$ with bidirectional links (i.e., $(u,v) \in \mathcal{E}$ if and only if $(v,u) \in \mathcal{E}$) admits a successive partition into subsets $\mathcal{S}_1, \ldots, \mathcal{S}_n$ with $n = \operatorname{diam}(G) + 1$.*

Proof. For any $v \in \mathcal{V}$ and any $r \in \{0 : \operatorname{diam}(G)\}$, we define the sets

$$\mathbb{B}_G(v,r) \triangleq \{u \in \mathcal{V} : d_G(v,u) \le r\}$$

and

$$\mathbb{S}_G(v,r) \triangleq \{u \in \mathcal{V} : d_G(v,u) = r\},$$

i.e., the ball and the sphere of radius r centered at v. In particular, $\mathbb{B}_G(v,r) = \mathbb{B}_G(v,r-1) \cup \mathbb{S}_G(v,r)$.

We now construct the desired successive partition. Let $n = \text{diam}(G) + 1$, and pick any pair of nodes $v_0, v_1 \in \mathcal{V}$ that achieve the maximum in the definition of $\text{diam}(G)$. With this, we take

$$\mathcal{P}_i = \mathbb{B}_G(v_0, i-1), \qquad i = 1, \ldots, n.$$

Clearly, $\mathcal{P}_1 = \{v_0\} \subset \mathcal{P}_2 \subset \ldots \subset \mathcal{P}_n = \mathcal{V}$, and moreover

$$\mathcal{S}_i = \mathbb{S}_G(v_0, i-1), \qquad i = 1, \ldots, n.$$

From this construction, we see that

$$\mathcal{E}_{\mathcal{P}_i} = \{(u, v) \in \mathcal{E} : u \in \mathcal{S}_{i+1},\, v \in \mathcal{S}_i\}$$

and

$$\mathcal{E}_{\mathcal{P}_i^c} = \{(u, v) \in \mathcal{V} : u \in \mathcal{S}_i,\, v \in \mathcal{S}_{i+1}\}\,.$$

The pairwise disjointness of the cutsets $\mathcal{E}_{\mathcal{P}_i}$, as well as of the cutsets $\mathcal{E}_{\mathcal{P}_i^c}$, is immediate. □

Remarks:

- Using the construction underlying the proof, we can also show that, for any two nodes in G, we can successively partition G into $n = d_G(u, v)+1$ subsets.
- For the successive partition constructed in the proof, all nodes in $\mathcal{S}_i$ are left-bound nodes, and d_i is the sum of the in-degrees of the nodes in $\mathcal{S}_i$.

As an example, Fig. 3.5a shows the successive partition of the network in Fig. 3.4a using the construction in the proof, where $\mathcal{S}_1 = \{1\}$, $\mathcal{S}_2 = \{2, 7\}$, $\mathcal{S}_3 = \{3, 8\}$, $\mathcal{S}_4 = \{4, 6, 9\}$, $\mathcal{S}_5 = \{5, 10\}$, with $\overleftarrow{\partial}\mathcal{S}_i = \mathcal{S}_i$, $i \in \{1, \ldots, 5\}$, and $d_2 = 6$, $d_3 = 6$, $d_4 = 9$, and $d_5 = 5$.

The successive partition of G ensures that nodes in $\mathcal{S}_i$ only communicate with nodes in $\mathcal{S}_{i-1}$ and $\mathcal{S}_{i+1}$, as well as among themselves. Indeed, suppose that the network graph G includes an edge $e = (u, v) \in \mathcal{E}$ with $u \in \mathcal{S}_i$ and $v \in \mathcal{S}_j$, where $i > j + 1$. By construction of the successive partition, $u \in \mathcal{P}_{j+1}^c \subset \mathcal{P}_j^c$ and $v \in \mathcal{P}_j \subset \mathcal{P}_{j+1}$. Therefore, e belongs to both $\mathcal{E}_{\mathcal{P}_j}$ and

$\mathcal{E}_{\mathcal{P}_{j+1}}$. However, the cutsets $\mathcal{E}_{\mathcal{P}_j}$ and $\mathcal{E}_{\mathcal{P}_{j+1}}$ are disjoint, so we arrive at a contradiction. Likewise, we can use the disjointness of the cutsets $\mathcal{E}_{\mathcal{P}_i^c}$ and $\mathcal{E}_{\mathcal{P}_j^c}$ to show that the network graph contains no edges (u, v) with $u \in \mathcal{S}_i$, $v \in \mathcal{S}_j$, and $j > i + 1$.

In view of this, we can associate to the partition $\{\mathcal{S}_i\}$ a *bidirected chain* $G' = (\mathcal{V}', \mathcal{E}')$, i.e., a network with vertex set $\mathcal{V}' = \{1', \ldots, n'\}$, edge set

$$\mathcal{E}' = \big\{(i', (i-1)')\big\}_{i=2}^{n} \cup \big\{(i', (i+1)')\big\}_{i=1}^{n-1} \cup \big\{(i', i')\big\}_{i=1}^{n},$$

and channel transition laws

$$K_{(i',(i-1)')} = \bigotimes_{(u,v)\in\mathcal{E}: u\in\mathcal{S}_i, v\in\mathcal{S}_{i-1}} K_{(u,v)} \tag{3.28}$$

$$K_{(i',(i+1)')} = \bigotimes_{(u,v)\in\mathcal{E}: u\in\mathcal{S}_i, v\in\mathcal{S}_{i+1}} K_{(u,v)} \tag{3.29}$$

$$K_{(i',i')} = \bigotimes_{(u,v)\in\mathcal{E}: u\in\mathcal{S}_i, v\in\overleftarrow{\partial}\mathcal{S}_i} K_{(u,v)}, \tag{3.30}$$

where node i' in G' observes

$$W_{i'} = W_{\mathcal{S}_i}.$$

In other words, the subset $\mathcal{S}_i$ in G is reduced to node i' in G'; the channels across the subsets in G are reduced to the channels between the nodes in G'; and the channels from $\mathcal{S}_i$ to $\overleftarrow{\partial}\mathcal{S}_i$ in G are reduced to a self-loop at node i' in G'. The channels from $\mathcal{S}_i$ to $\mathcal{S}_i \setminus \overleftarrow{\partial}\mathcal{S}_i$ in G are not included in G', and will be simulated by node i' using private randomness. For the network in Fig. 3.2a in Sec. 3.1.2, according to the illustrated partition, it can be reduced to a 3-node bidirected chain in Fig. 3.2b, with $K_{(1',1')} = K_{(1,4)}$, $K_{(2',2')} = K_{(5,2)}$, and $K_{(3',3')} = K_{(3,6)} \otimes K_{(6,3)}$. For the network in Fig. 3.4a, according to the illustrated partition, it can be reduced to a 5-node bidirected chain in Fig. 3.4b, with $K_{(2',2')} = K_{(2,7)} \otimes K_{(7,2)}$, $K_{(3',3')} = K_{(6,3)} \otimes K_{(6,8)} \otimes K_{(9,8)}$, and $K_{(4',4')} = K_{(4,10)} \otimes K_{(10,4)}$. According to the partition in Fig. 3.5a, the same network can be reduced to a 5-node bidirected chain in Fig. 3.5b, with $K_{(2',2')} = K_{(2,7)} \otimes K_{(7,2)}$, $K_{(4',4')} = K_{(6,9)} \otimes K_{(9,6)}$, and $K_{(5',5')} = K_{(5,10)} \otimes K_{(10,5)}$.

For the bidirected chain G' reduced from G, we consider a class of *randomized* T-step algorithms that run on G' and are of a more general form

compared to the deterministic algorithms considered so far. Such a randomized algorithm operates as follows: at step $t \in \{1, \ldots, T\}$, node i' computes the outgoing messages $X_{(i',(i-1)'),t} = \overleftarrow{\varphi}_{i',t}(W_{i'}, Y_{i'}^{t-1})$, $X_{(i',(i+1)'),t} = \vec{\varphi}_{i',t}(W_{i'}, Y_{i'}^{t-1}, U_{i'}^{t-1})$, and $X_{(i',i'),t} = \mathring{\varphi}_{i',t}(W_{i'}, Y_{i'}^{t-1}, U_{i'}^{t-1})$, and computes the private message $U_{i',t} = \vartheta_{i',t}(W_{i'}, Y_{i'}^{t-1}, U_{i'}^{t-1}, R_{i',t})$, where $R_{i',t}$ is the private randomness held by node i', uniformly distributed on $[0,1]$ and independent across $i' \in \mathcal{V}'$ and $t \in \{1, \ldots, T\}$. At step T, node i' computes the final estimate $\widehat{Z}_{i'} = \psi_{i'}(W_{i'}, Y_{i'}^T)$ of Z. These randomized algorithms have the feature that the message sent to the node on the left and the final estimate of a node are computed solely based on the node's initial observation and received messages, whereas the messages sent to the node on the right and to itself are computed based on the node's initial observation, received messages, as well as private messages, and the computation of the private messages involves the node's private randomness. Define

$$T'(\varepsilon, \delta) = \inf \Big\{ T \in \mathbb{N} : \exists \text{ a randomized } T\text{-step algorithm } \mathcal{A}' \text{ such that } \max_{i' \in \mathcal{V}'} \mathbb{P}[\ell(Z, \widehat{Z}_{i'}) > \varepsilon] \leq \delta \Big\} \tag{3.31}$$

as the (ε, δ)-computation time for Z on G' using the randomized algorithms described above. The following lemma indicates that we can obtain lower bounds on $T(\varepsilon, \delta)$ by lower-bounding $T'(\varepsilon, \delta)$.

Lemma 3.7. *Consider an arbitrary network G that can be successively partitioned into $\mathcal{S}_1, \ldots, \mathcal{S}_n$, such that $\overleftarrow{\partial}\mathcal{S}_i$'s are all nonempty. Let $G' = (\mathcal{V}', \mathcal{E}')$ be the bidirected chain constructed from G according to the partition. Then, given any T-step algorithm on G that achieves $\max_{v \in \mathcal{V}} \mathbb{P}[\ell(Z, \widehat{Z}_v) > \varepsilon] \leq \delta$, we can construct a randomized T-step algorithm $\mathcal{A}'$ on G', such that $\max_{i' \in \mathcal{V}'} \mathbb{P}[\ell(Z, \widehat{Z}_{i'}) > \varepsilon] \leq \delta$. Consequently, $T(\varepsilon, \delta)$ for computing Z on G is lower bounded by $T'(\varepsilon, \delta)$ defined in (3.31).*

Proof. Section 3.7.1. □

Remark: In the network reduction, we can alternatively map all the channels from $\mathcal{S}_i$ to $\mathcal{S}_i$ (instead of only mapping the channels from $\mathcal{S}_i$ to $\overleftarrow{\partial}\mathcal{S}_i$) in the original network G to the self-loop at node i' of the reduced chain G'. By doing so, to simulate the operation of an algorithm $\mathcal{A}$ that runs on G, the algorithm $\mathcal{A}'$ that runs on G' no longer needs to generate private messages using the nodes' private randomness, since all the channels in G are

preserved in G'. In other words, under this alternative reduction, any T-step algorithm $\mathcal{A}$ in that runs on G can be simulated by a T-step algorithm $\mathcal{A}'$ of the same deterministic type as $\mathcal{A}$ that runs on G'. However, this alternative reduction increases the information transmission capability of the self-loops in G', and will result in a looser lower bound on $T(\varepsilon, \delta)$, as will be discussed in the remark following Theorem 3.4.

In light of Lemma 3.7, in order to lower-bound $T(\varepsilon, \delta)$ for computing Z on G, we just need to lower-bound $T'(\varepsilon, \delta)$ defined in (3.31). To this end, we derive upper bounds on the conditional mutual information for bidirected chains by extending the techniques behind Lemma 3.4 and Lemma 3.5:

Lemma 3.8. *Consider an n-node bidirected chain with vertex set $\mathcal{V} = \{1, \ldots, n\}$ and edge set*

$$\mathcal{E} = \big\{(i, i-1)\big\}_{i=2}^{n} \cup \big\{(i, i+1)\big\}_{i=1}^{n-1} \cup \big\{(i, i)\big\}_{i=1}^{n},$$

and an arbitrary randomized T-step algorithm $\mathcal{A}'$ that runs on this chain. Let $\eta_i \triangleq \eta(K_i)$ denote the SDPI constant of the channel $K_i \triangleq \bigotimes_{j:\,(j,i)\in\mathcal{E}} K_{(j,i)}$, and let $\eta \triangleq \max_{i=1,\ldots,n} \eta_i$. If $T \leq n-2$, then

$$I(Z; \widehat{Z}_n | W_{2:n}) = 0.$$

If $T \geq n-1$, then

$$I(Z; \widehat{Z}_n | W_{2:n}) \leq \begin{cases} H(W_1|W_{2:n})\eta \displaystyle\sum_{i=1}^{T-n+2} \mathcal{B}(T-i, n-2, \eta), n \geq 2 & \text{(3.32a)} \\ C_{(1,2)}\eta \displaystyle\sum_{i=1}^{T-n+2} \mathcal{B}(T-i-1, n-3, \eta) i, \quad n \geq 3 & \text{(3.32b)} \end{cases}$$

with $\mathcal{B}(m, k, p) \triangleq \binom{m}{k} p^k (1-p)^{m-k}$. For $n \geq 2$, the above upper bounds can be weakened to

$$I(Z; \widehat{Z}_n | W_{2:n}) \leq \begin{cases} H(W_1|W_{2:n}) \big(1 - (1-\eta)^{T-n+2}\big)^{n-1}, & \text{(3.33a)} \\ C_{(1,2)}(T-n+2) \big(1 - (1-\tilde{\eta})^{T-n+2}\big)^{n-2} & \text{(3.33b)} \end{cases}$$

Moreover, if $n \geq 4$ *and* $n-1 \leq T \leq 2+(n-3)\gamma/\eta$ *for some* $\gamma \in (0,1)$*, then*

$$I(Z;\widehat{Z}_n|W_{2:n}) \leq C_{(1,2)}\frac{(n-3)^2\gamma^2}{\eta}\exp\left(-2\left(\frac{\eta}{\gamma}-\eta\right)^2(n-3)\right). \tag{3.34}$$

Proof. Section 3.7.2. □

Equation (3.32a) is reminiscent of a result of Rajagopalan and Schulman [50] on the evolution of mutual information in broadcasting a bit over a unidirectional chain of BSCs. The result in [50] is obtained by solving a system of recursive inequalities on the mutual information involving suboptimal SDPI constants. Our results apply to chains of general bidirectional links and to the computation of general functions. We arrive at a system of inequalities similar to the one in [50], which can be solved in a similar manner and gives (3.32a) and (3.32b). We also obtain weakened upper bounds in (3.33a) and (3.33b), which show that, for a fixed T, the conditional mutual information decays at least exponentially fast in n. The upper bound in (3.34) provides another weakening of (3.32a) and (3.32b), and shows explicitly the dependence of the upper bound on n.

Assuming for simplicity that $H(W_1|W_{2:n}) = 1$, Fig. 3.6 compares (3.32a) with the weakened upper bound in (3.33a). We can see that the gap can be large when n is large and T is much larger than n. Nevertheless, the weakened upper bounds in (3.33a) and (3.33b) allow us to derive lower bounds on computation time that are non-asymptotic in n, and explicit in ε, δ, and channel properties.

3.3.2 Lower Bounds on Computation Time

We now build on the results presented above to obtain lower bounds on the $T(\varepsilon,\delta)$ by reducing the original problem to function computation over bidirected chains. We first provide the result for an arbitrary network, and then particularize it to several specific topologies (namely, chains, rings, grids, and trees).

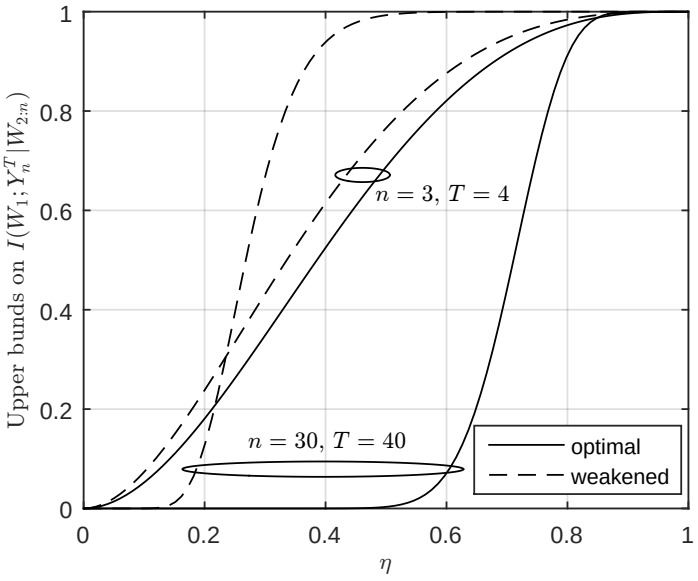


Figure 3.6: Upper bound in (3.32a) (solid line) vs. the weakened one in (3.33a) (dashed line) for chains.

Lower Bound for an Arbitrary Network

Theorem 3.4 below contains general lower bounds on computation time for an arbitrary network. The statement of the theorem is somewhat lengthy, but can be parsed as follows: Given an arbitrary connected network with bidirectional links, any reduction of that network to a bidirected chain gives rise to a system of inequalities that must be satisfied by the computation time $T(\varepsilon, \delta)$. These inequalities, presented in (3.35), are nonasymptotic in nature and involve explicitly computable parameters of the network, but cannot be solved in closed form. The first inequality follows from an SDPI-based analysis analogous to Theorem 3.3, while the second inequality is a cutset bound in the spirit of Theorem 3.1. Explicit but weaker expressions that lower-bound $T(\varepsilon, \delta)$ in terms of network parameters appear below as (3.36) and (3.38), together with asymptotic expressions for large n (the size of the reduced bidirected chain). Both of these bounds state that $T(\varepsilon, \delta)$ is lower bounded by the size of the bidirected chain plus a correction term that accounts for the effect of channel noise (via channel capacities and SDPI constants). Finally, (3.39) and (3.40) provide the precise version of the bound in (3.8): asymptotically, the computation time $T(\varepsilon, \delta)$ scales as $\Omega(n/\tilde{\eta})$, where $\tilde{\eta}$ is the worst-case SDPI constant of the reduced network. By Lemma 3.6, it is always possible to reduce the network to a bidirected chain of length

$\mathrm{diam}(G)+1$, so the main message of Theorem 3.4 is that the computation time $T(\varepsilon,\delta)$ scales at least linearly in the network diameter. Thus, the main advantage of the multi-cutset analysis over the usual single-cutset analysis is that it can capture this dependence on the network diameter.

Theorem 3.4. *Assume the following:*

- *The network graph $G=(\mathcal{V},\mathcal{E})$ is connected, the capacities of all edge links are upper bounded by C, and the SDPI constants of edge links are upper bounded by η.*
- *G admits a successive partition into $\mathcal{S}_1,\ldots,\mathcal{S}_n$, such that $\overleftarrow{\partial}\mathcal{S}_i$'s are all nonempty.*

Let

$$\Delta \triangleq \max_{i\in\{2:n\}} d_i$$

where

$$d_i = |\mathcal{E}_{\mathcal{P}_{i-1}^c}| + |\mathcal{E}_{\mathcal{P}_i}| + |\{\mathcal{E}\cap(\mathcal{S}_i\times\overleftarrow{\partial}\mathcal{S}_i)\}|$$

as defined in (3.27), and let

$$\tilde{\eta} = 1-(1-\eta)^{\Delta}.$$

Then for $\varepsilon\geq 0$ and $\delta\in(0,1/2]$, the (ε,δ)-computation time $T(\varepsilon,\delta)$ must satisfy the inequalities

$$\mathcal{I}(\mathcal{S}_1^c,\varepsilon,\delta) \leq \begin{cases} H(W_{\mathcal{S}_1}|W_{\mathcal{S}_1^c})\tilde{\eta}\sum_{i=1}^{T(\varepsilon,\delta)-n+2}\mathcal{B}(T(\varepsilon,\delta)-i,n-2,\tilde{\eta}), & n\geq 2\\ C_{\mathcal{S}_1^c}\tilde{\eta}\sum_{i=1}^{T(\varepsilon,\delta)-n+2}\mathcal{B}(T(\varepsilon,\delta)-i-1,n-3,\tilde{\eta})i, & n\geq 3. \end{cases} \tag{3.35}$$

The above results can be weakened to

$$T(\varepsilon,\delta) \geq \frac{\log\left(1-\left(\frac{\mathcal{I}(\mathcal{S}_1^c,\varepsilon,\delta)}{H(W_{\mathcal{S}_1}|W_{\mathcal{S}_1^c})}\right)^{\frac{1}{n-1}}\right)^{-1}}{\Delta\log(1-\eta)^{-1}}+n-2 \tag{3.36}$$

$$\sim \frac{\log(n-1)+\log\left(1-\frac{\mathcal{I}(\mathcal{S}_1^c,\varepsilon,\delta)}{H(W_{\mathcal{S}_1}|W_{\mathcal{S}_1^c})}\right)^{-1}}{\Delta\log(1-\eta)^{-1}}+n-2 \quad \text{as } n\to\infty, \tag{3.37}$$

and

$$T(\varepsilon,\delta) \geq \frac{\mathcal{I}\left(\mathcal{S}_1^c,\varepsilon,\delta\right)}{C_{\mathcal{S}_1^c}} + n - 2. \tag{3.38}$$

Moreover, if the partition size n is large enough, so that $n \geq 4$ and

$$\frac{C|\mathcal{V}|^2(n-3)^2}{4\eta}\exp\left(-2\eta^2(n-3)\right) < \mathcal{I}(\mathcal{S}_1^c,\varepsilon,\delta), \tag{3.39}$$

then

$$T(\varepsilon,\delta) > 2 + \frac{n-3}{2\tilde{\eta}} = \Omega\left(\frac{n}{\tilde{\eta}}\right). \tag{3.40}$$

Proof. In light of Lemma 3.7, it suffices to show that the lower bounds in Theorem 3.4 need to be satisfied by $T'(\varepsilon,\delta)$ for the bidirected chain G', to which G reduces according to the partition $\{\mathcal{S}_i\}$.

Consider any randomized T-step algorithm $\mathcal{A}'$ that achieves

$$\max_{i'\in\mathcal{V}'} \mathbb{P}[\ell(Z,\widehat{Z}_{i'}) > \varepsilon] \leq \delta$$

on G'. From Lemma 3.1,

$$I(Z;\widehat{Z}_{n'}|W_{2':n'}) \geq \mathcal{I}(\{2':n'\},\varepsilon,\delta).$$

Then from Lemma 3.8 and the fact that

$$\eta_{i'} = \eta(K_{((i-1)',i')} \otimes K_{((i+1)',i')} \otimes K_{i',i'}) \leq 1-(1-\eta)^{d_i} \leq 1-(1-\eta)^{\Delta}, \tag{3.41}$$

we have

$$\mathcal{I}(\{2':n'\},\varepsilon,\delta) \leq \begin{cases} H(W_{1'}|W_{2':n'})\tilde{\eta}\sum_{i=1}^{T-n+2}\mathcal{B}(T-i,n-2,\tilde{\eta}), & n\geq 2 \\ C_{(1',2')}\tilde{\eta}\sum_{i=1}^{T-n+2}\mathcal{B}(T-i-1,n-3,\tilde{\eta})i, & n\geq 3 \end{cases}$$

and

$$\mathcal{I}(\{2':n'\},\varepsilon,\delta) \le \begin{cases} H(W_{1'}|W_{2':n'}) \prod_{i=2}^{n} \left(1-(1-\eta)^{d_i(T-n+2)}\right) \\ C_{(1',2')}(T-n+2) \prod_{i=3}^{n} \left(1-(1-\eta)^{d_i(T-n+2)}\right) \end{cases}, n \ge 2. \tag{3.42}$$

Since $\mathcal{I}(\{2':n'\},\varepsilon,\delta) = \mathcal{I}(\mathcal{S}_1^c,\varepsilon,\delta)$, $H(W_{1'}|W_{2':n'}) = H(W_{\mathcal{S}_1}|W_{\mathcal{S}_1^c})$, and $C_{(1',2')} = C_{\mathcal{S}_1^c}$, we see that $T'(\varepsilon,\delta)$ must satisfy (3.35) in Theorem 3.4.

Using (3.41), (3.42) can be weakened to

$$\mathcal{I}(\mathcal{S}_1^c,\varepsilon,\delta) \le \begin{cases} H(W_{\mathcal{S}_1}|W_{\mathcal{S}_1^c})\left(1-(1-\eta)^{\Delta(T-n+2)}\right)^{n-1} \\ C_{\mathcal{S}_1^c}(T-n+2)\left(1-(1-\eta)^{\Delta(T-n+2)}\right)^{n-2} \end{cases}. \tag{3.43}$$

The first line of (3.43) leads to

$$\begin{aligned} T'(\varepsilon,\delta) &\ge \frac{\log\left(1-\left(\frac{\mathcal{I}(\mathcal{S}_1^c,\varepsilon,\delta)}{H(W_{\mathcal{S}_1}|W_{\mathcal{S}_1^c})}\right)^{\frac{1}{n-1}}\right)^{-1}}{\Delta\log(1-\eta)^{-1}} + n - 2 \\ &\sim \frac{\log(n-1) + \log\left(1-\frac{\mathcal{I}(\mathcal{S}_1^c,\varepsilon,\delta)}{H(W_{\mathcal{S}_1}|W_{\mathcal{S}_1^c})}\right)^{-1}}{\Delta\log(1-\eta)^{-1}} + n - 2 \quad \text{as } n\to\infty, \end{aligned}$$

where the last step follows from the fact that $\log\left(1-p^{\frac{1}{n}}\right)^{-1} \sim \log\frac{n}{1-p}$ as $n\to\infty$ for $p\in(0,1)$. The second line of (3.43) leads to

$$T'(\varepsilon,\delta) \ge \frac{\mathcal{I}\left(\mathcal{S}_1^c,\varepsilon,\delta\right)}{C_{\mathcal{S}_1^c}} + n - 2.$$

Finally, we prove that $T'(\varepsilon,\delta) = \Omega(n/\tilde{\eta})$ under the assumption that (3.39) holds. Suppose that $T'(\varepsilon,\delta) \le 2 + (n-3)/2\tilde{\eta}$. Then, from (3.34) in Lemma 3.8, we have

$$\mathcal{I}(\mathcal{S}_1^c,\varepsilon,\delta) \le C_{\mathcal{S}_1^c}\frac{(n-3)^2}{4\tilde{\eta}}\exp\left(-2\tilde{\eta}^2(n-3)\right) \quad \text{if } n\ge 4.$$

Note that $\Delta \ge 1$ by the assumption that G is connected, thus $\tilde{\eta} = 1-(1-\eta)^{\Delta} \ge \eta$. Moreover, $C_{\mathcal{S}_1^c} \le C|\mathcal{E}| \le C|\mathcal{V}|^2$. As a result,

$$\mathcal{I}(\mathcal{S}_1^c,\varepsilon,\delta) \le \frac{C|\mathcal{V}|^2(n-3)^2}{4\eta}\exp\left(-2\eta^2(n-3)\right) \quad \text{if } n\ge 4,$$

which contradicts the assumption that (3.39) holds. Thus,

$$T'(\varepsilon, \delta) > 2 + \frac{n-3}{2\tilde{\eta}} = \Omega\left(\frac{n}{\tilde{\eta}}\right).$$

Theorem 3.4 then follows from Lemma 3.7. □

Remarks:

- We call a node in $\mathcal{S}_i$ a *boundary node* if there is an edge (either inward or outward) between it and a node in $\mathcal{S}_{i-1}$ or $\mathcal{S}_{i+1}$. Denote the set of boundary nodes of $\mathcal{S}_i$ by $\partial\mathcal{S}_i$. The results in Theorem 3.4 can be weakened by replacing d_i with

  $$\partial d_i = \sum_{v \in \partial\mathcal{S}_i} |\mathcal{E}_v|,$$

  namely the summation of the in-degrees of boundary nodes of $\mathcal{S}_i$, since $d_i \leq \partial d_i$ for $i \in \{2, \ldots, n\}$.

- As discussed in the remark following Lemma 3.7, an alternative network reduction is to map all the channels from $\mathcal{S}_i$ to $\mathcal{S}_i$ (instead of only mapping the channels from $\mathcal{S}_i$ to $\overleftarrow{\partial}\mathcal{S}_i$) in the original network G to the self-loop at node i' of the reduced chain G'. Using the same proof strategy with this alternative reduction, we can obtain lower bounds on $T(\varepsilon, \delta)$ of the same form as the results in Theorem 3.4, but with d_i's replaced by

  $$\tilde{d}_i \triangleq |\mathcal{E}_{\mathcal{P}_{i-1}^c}| + |\mathcal{E}_{\mathcal{P}_i}| + |\{\mathcal{E} \cap (\mathcal{S}_i \times \mathcal{S}_i)\}|.$$

  Since $d_i \leq \partial d_i \leq \tilde{d}_i$ for $i \in \{2, \ldots, n\}$, the lower bounds on $T(\varepsilon, \delta)$ obtained by this alternative network reduction are weaker than the results in Theorem 3.4, and are even weaker than the results obtained by replacing d_i's with ∂d_is.

- Due to Lemma 3.6, for a network G of bidirectional links, we can always find a successive partition of G such that n in Theorem 3.4 is equal to the diam$(G) + 1$. By contrast, the diameter cannot be captured in general by the theorems in Sec. 3.2.

- Choosing a successive partition of G with $n = 2$ is equivalent to choosing a single cutset. In that case, we see that (3.38) recovers Theorem 3.1, while (3.36) recovers a weakened version of Theorem 3.3 (in (3.36), $\Delta = d_2$ is at least the sum of the in-degrees of the left-bound nodes of $\mathcal{S}_2$, while Theorem 3.3 involves the in-degree of only one node in $\mathcal{S}_2$).

We now apply Theorem 3.4 to networks with specific topologies. We assume that nodes communicate via bidirectional links. Thus, any such network will be represented by an undirected graph, where each undirected edge represents a pair of channels with opposite directions.

Chains

For chains, the proof of Theorem 3.4 already contains lower bounds on $T'(\varepsilon, \delta)$. These lower bounds apply to $T(\varepsilon, \delta)$ as well, since the class of T-step algorithms on a chain is a subcollection of randomized T-step algorithms on the same chain. We thus have the following corollary.

Corollary 3.3. *Consider an n-node bidirected chain without self-loops, where the SDPI constants of all channels are upper bounded by η. Then for $\varepsilon \geq 0$ and $\delta \in (0, 1/2]$, $T(\varepsilon, \delta)$ must satisfy the inequalities in Theorem 3.4 with $\mathcal{S}_1 = \{1\}$ and $d_i = 2$ for all $i \in \{1, \ldots, n\}$. In particular, if all channels are* BSC(p)*, then*

$$T(\varepsilon, \delta) \geq \max \left\{ \frac{\mathcal{I}(\mathcal{V} \setminus \{1\}, \varepsilon, \delta)}{1 - h_2(p)}, \frac{\log(n-1) + \log\left(1 - \frac{\mathcal{I}(\mathcal{V}\setminus\{1\}, \varepsilon, \delta)}{H(W_1 | W_{\mathcal{V}\setminus\{1\}})}\right)^{-1}}{2\log(4p\bar{p})^{-1}} \right\} + n - 2,$$

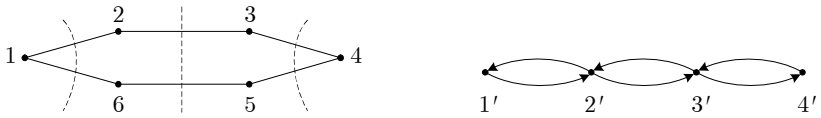
for all sufficiently large n.

Here and below, the estimates for a network of bidirectional BSCs are obtained using the bounds (3.19) and (3.20).

Rings

Consider a ring with $2n - 2$ nodes, where the nodes are labeled clockwise from 1 to $2n-2$. The diameter is equal to $n-1$. According to the successive partition in the proof of Lemma 3.6, this ring can be partitioned into $\mathcal{S}_1 =$

$\{1\}$, $\mathcal{S}_i = \{i, 2n-i\}$, $i \in \{2, \ldots, n-1\}$, and $\mathcal{S}_n = \{n\}$. As an example, Fig. 3.7a shows a 6-node ring and Fig. 3.7b shows the chain reduced from it. With this partition, we can apply Theorem 3.4 and get the following



(a) (b)

Figure 3.7: A ring network and the chain reduced from it.

corollary.

Corollary 3.4. *Consider a $(2n-2)$-node ring, where the SDPI constants of all channels are upper bounded by η. Then for $\varepsilon \geq 0$ and $\delta \in (0, 1/2]$, $T(\varepsilon, \delta)$ must satisfy the inequalities in Theorem 3.4 with $\mathcal{S}_1 = \{1\}$ and $d_i = 4$ for all $i \in \{1, \ldots, n\}$. In particular, if all channels are* BSC(p)*, then*

$$T(\varepsilon, \delta) = \max\left\{ \frac{\mathcal{I}\big(\mathcal{V} \setminus \{1\}, \varepsilon, \delta\big)}{2(1 - h_2(p))}, \frac{\log(n-1) + \log\left(1 - \frac{\mathcal{I}(\mathcal{V}\setminus\{1\}, \varepsilon, \delta)}{H(W_1 | W_{\mathcal{V}\setminus\{1\}})}\right)^{-1}}{4\log(4p\bar{p})^{-1}} \right\} + n - 2,$$

for all sufficiently large n.

Grids

Consider an $\frac{n+1}{2} \times \frac{n+1}{2}$ grid (where we assume n is odd), which has diameter $n-1$. Figure 3.8a shows a successive partition of a $\frac{n+1}{2} \times \frac{n+1}{2}$ grid into $\frac{n+1}{2}$ subsets, with $\Delta = \max_{i \in \{2:n\}} d_i = 2n$. Figure 3.8b shows the successive partition in the proof of Lemma 3.6, which partitions the network into n subsets, with $\Delta = \max_{i \in \{2:n\}} d_i = 2(n-1)$, thus resulting in strictly tighter lower bounds on computation time compared to the ones obtained from the partition in Fig. 3.8a. With the latter partition, we get the following corollary.

Corollary 3.5. *Consider an $\frac{n+1}{2} \times \frac{n+1}{2}$ grid, where $1 - \ldots - n$ is one of the longest paths. Assume that the SDPI constants of all channels are upper bounded by η. Then for $\varepsilon \geq 0$ and $\delta \in (0, 1/2]$, $T(\varepsilon, \delta)$ must satisfy the inequalities in Theorem 3.4 with $\mathcal{S}_1 = \{1\}$, $d_i = d_{n+1-i} = 4(i-2) + 6$,*

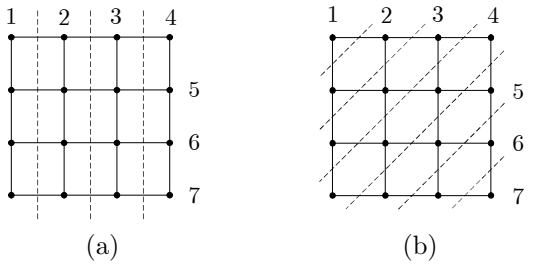


Figure 3.8: Successive partitions of a 4×4 ($n = 7$) grid network. The length of the labeled path is the diameter of the network.

$i \in \{1, \ldots, \frac{n-1}{2}\}$, *and* $d_{(n+1)/2} = 2(n-1)$. *In particular, if all channels are* BSC(p), *then*

$$T(\varepsilon, \delta) \geq \max \left\{ \frac{\mathcal{I}\big(\mathcal{V} \setminus \{1\}, \varepsilon, \delta\big)}{2(1 - h_2(p))}, \frac{\log(n-1) + \log\left(1 - \frac{\mathcal{I}(\mathcal{V}\setminus\{1\}, \varepsilon, \delta)}{H(W_1 | W_{\mathcal{V}\setminus\{1\}})}\right)^{-1}}{2(n-1)\log(4p\bar{p})^{-1}} \right\} + n - 2,$$

for all sufficiently large n.

Trees

Consider a tree, whose nodes are numbered in such a way that $1 - \ldots - n$ is one of the longest paths. Then the diameter of the tree is $n - 1$, and nodes 1 and n are necessarily leaf nodes. The tree can be viewed as being rooted at node 1. Let $\mathcal{D}_i$ be the union of node i and its descendants in the rooted tree, and let $\mathcal{S}_i = \mathcal{D}_i \setminus \mathcal{D}_{i+1}$, $i \in \{1, \ldots, n\}$. The tree can then be successively partitioned into $\mathcal{S}_1, \ldots, \mathcal{S}_n$. In the n-node bidirected chain reduced according to this partition, the edges between nodes i' and $(i+1)'$ are the pair of channels between nodes i and $i+1$ in the tree, and the self-loop of node i', $i \in \{2, \ldots, n-1\}$, is the channel from $\mathcal{S}_i \setminus \{i\}$ to node i in the tree. As an example, Fig. 3.9a shows this partition of a tree network, where the chain reduced from it has the same form as the one in Fig. 3.4b. With this partition, we get the following corollary.

Corollary 3.6. *Consider a d-regular tree network where $1 - \ldots - n$ is one of the longest paths. Assume that the SDPI constants of all channels are upper bounded by η. Then for $\varepsilon \geq 0$ and $\delta \in (0, 1/2]$, $T(\varepsilon, \delta)$ must satisfy the inequalities in Theorem 3.4 with $\mathcal{S}_1 = \{1\}$ and $d_i = d$ for all $i \in \{1, \ldots, n\}$.*

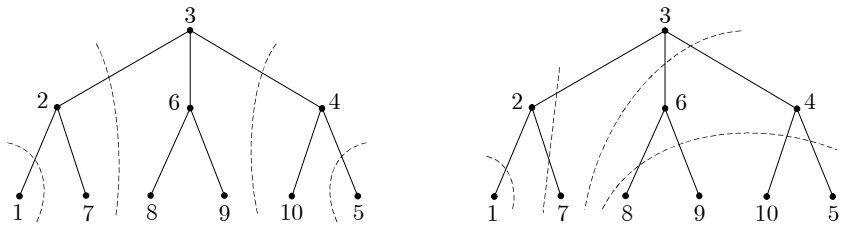


Figure 3.9: Successive partitions of a tree network.

In particular, if all channels are BSC(p)*, then*

$$T(\varepsilon,\delta) \geq \max\left\{\frac{\mathcal{I}\big(\mathcal{V}\setminus\{1\},\varepsilon,\delta\big)}{1-h_2(p)}, \frac{\log(n-1)+\log\left(1-\frac{\mathcal{I}(\mathcal{V}\setminus\{1\},\varepsilon,\delta)}{H(W_1|W_{\mathcal{V}\setminus\{1\}})}\right)^{-1}}{d\log(4p\bar{p})^{-1}}\right\}+n-2,$$

for all sufficiently large n*.*

If we use the successive partition in the proof of Lemma 3.6 on a d-regular tree with diameter $n-1$, then the tree will be reduced to an n-node bidirected chain without self-loops. Figure 3.9b shows such an example. However, with this partition, $\Delta = \max_{i\in\{2:n\}} d_i$ increases with n, which renders the resulting lower bound on computation time looser than the one in Corollary 3.6. It means that, although the partition in the proof of Lemma 3.6 always captures the diameter of a network, it may not always give the best lower bound on computation time among all possible successive partitions.

3.4 Small Ball Probability Estimates for Computation of Linear Functions

The bounds stated in the preceding sections involve the conditional small ball probability, defined in (3.9). In this section, we provide estimates for this quantity in the context of a distributed computation problem of wide interest — the computation of linear functions. Specifically, we assume that the observations $W_v, v \in \mathcal{V}$, are independent real-valued random variables,

and the objective is to compute a linear function

$$Z = f(W) = \sum_{v \in \mathcal{V}} a_v W_v \tag{3.44}$$

for a fixed vector of coefficients $(a_v)_{v \in \mathcal{V}} \in \mathbb{R}^{|\mathcal{V}|}$, subject to the absolute error criterion $\ell(z, \widehat{z}) = |z - \widehat{z}|$. We will use the following shorthand notation: for any set $\mathcal{S} \subset \mathcal{V}$, let $a_{\mathcal{S}} = (a_v)_{v \in \mathcal{S}}$ and $\langle a_{\mathcal{S}}, W_{\mathcal{S}} \rangle = \sum_{v \in \mathcal{S}} a_v W_v$.

The independence of the W_v's and the additive structure of f allow us to express the conditional small ball probability $\mathcal{L}_{Z|W_{\mathcal{S}}}(w_{\mathcal{S}}, \varepsilon)$ defined in (3.9) in terms of so-called *Lévy concentration functions* of random sums [56]. The Lévy concentration function of a real-valued r.v. U (also known as the "small ball probability") is defined as

$$L(U, \rho) = \sup_{u \in \mathbb{R}} \mathbb{P}\left[|U - u| \leq \rho\right], \qquad \rho > 0.$$

If we fix a subset $\mathcal{S} \subset \mathcal{V}$, and consider a specific realization $W_{\mathcal{S}} = w_{\mathcal{S}}$ of the observations of the nodes in $\mathcal{S}$, then

$$\begin{aligned}
\mathcal{L}_{Z|W_{\mathcal{S}}}(w_{\mathcal{S}}, \varepsilon) &= \sup_{z \in \mathbb{R}} \mathbb{P}\left[\Big| \sum_{v \in \mathcal{V}} a_v W_v - z \Big| \leq \varepsilon \,\Big|\, W_{\mathcal{S}} = w_{\mathcal{S}}\right] \\
&= \sup_{z \in \mathbb{R}} \mathbb{P}\left[\Big| \sum_{v \in \mathcal{S}^c} a_v W_v + \sum_{v \in \mathcal{S}} a_v w_v - z \Big| \leq \varepsilon\right] \\
&= \sup_{z \in \mathbb{R}} \mathbb{P}\left[\Big| \sum_{v \in \mathcal{S}^c} a_v W_v - z \Big| \leq \varepsilon\right] \\
&= L\left(\langle a_{\mathcal{S}^c}, W_{\mathcal{S}^c} \rangle, \varepsilon\right),
\end{aligned} \tag{3.45}$$

where in the second line we have used the fact that the W_v's are independent r.v.'s, while in the third line we have used the fact that for any function $g : \mathbb{R} \to \mathbb{R}$ and any $a \in \mathbb{R}$, $\sup_z g(z) = \sup_z g(z + a)$. In other words, for a fixed $\mathcal{S}$, the quantity $\mathcal{L}_{Z|W_{\mathcal{S}}}(w_{\mathcal{S}}, \varepsilon)$ is independent of the boundary condition $w_{\mathcal{S}}$, and is controlled by the probability law of the random sum $\langle a_{\mathcal{S}^c}, W_{\mathcal{S}^c} \rangle$, i.e., the part of the function f that depends on the observations of the nodes in $\mathcal{S}^c$.

The problem of estimating Lévy concentration functions of sums of independent random variables has a long history in the theory of probability —

for random variables with densities, some of the first results go back at least to Kolmogorov [62], while for discrete random variables it is closely related to the so-called Littlewood–Offord problem [63]. We provide a few examples to illustrate how one can exploit available estimates for Lévy concentration functions under various regularity conditions to obtain tight lower bounds on the computation time for linear functions. The examples are illustrated through Theorem 3.1, as it tightly captures the dependence of computation time on $\mathcal{I}(\mathcal{S}, \varepsilon, \delta)$. (However, since the results of Theorems 3.3 and 3.4 also involve the quantity $\mathcal{I}(\mathcal{S}, \varepsilon, \delta)$, the estimates for Lévy concentration functions can be applied there as well.)

3.4.1 Computing Linear Functions of Continuous Observations

Gaussian Sums

Suppose that the local observations W_v, $v \in \mathcal{V}$, are i.i.d. standard Gaussian random variables. Then, for any $\mathcal{S} \subseteq \mathcal{V}$, $\langle a_{\mathcal{S}}, W_{\mathcal{S}} \rangle$ is a zero-mean Gaussian r.v. with variance $\|a_{\mathcal{S}}\|_2^2 = \sum_{v \in \mathcal{S}} a_v^2$ (here, $\|\cdot\|_2$ is the usual Euclidean ℓ_2 norm). A simple calculation shows that

$$\mathcal{L}_{Z|W_{\mathcal{S}}}(w_{\mathcal{S}}, \varepsilon) = L\left(N\left(0, \|a_{\mathcal{S}^c}\|_2^2\right), \varepsilon\right) \leq \sqrt{\frac{2}{\pi}} \frac{\varepsilon}{\|a_{\mathcal{S}^c}\|_2}.$$

Using this in Theorem 3.1, we get the following result.

Corollary 3.7. *For the problem of computing a linear function in (3.44), where* $(W_v) \overset{\text{i.i.d.}}{\sim} N(0,1)$*, suppose that the coefficients* a_v *are all nonzero. Then for* $\varepsilon \geq 0$ *and* $\delta \in (0, 1/2]$*,*

$$T(\varepsilon, \delta) \geq \max_{\mathcal{S} \subset \mathcal{V}} \frac{1}{C_{\mathcal{S}}} \left(\frac{1-\delta}{2} \log \frac{\pi \|a_{\mathcal{S}^c}\|_2^2}{2\varepsilon^2} - h_2(\delta) \right).$$

Thus, the lower bound on the computation time for (3.44) depends on the vector of coefficients a only through its ℓ_2 norm.

Sums of Independent r.v.'s with Log-concave Distributions

Another instance in which sharp bounds on the Lévy concentration function are available is when the observations of the nodes are independent random variables with log-concave distributions (we recall that a real-valued r.v. U is said to have a log-concave distribution if it has a density of the form $p_U(u) = e^{-F(u)}$, where $F : \mathbb{R} \to (-\infty, +\infty]$ is a convex function; this includes Gaussian, Laplace, uniform, etc.). The following result was obtained recently by Bobkov and Chistyakov [64, Theorem 1.1]: Let $U_1, \ldots, U_k$ be independent random variables with log-concave distributions, and let $S_k = U_1 + \ldots + U_k$. Then, for any $\rho \geq 0$,

$$\frac{1}{\sqrt{3}} \frac{\rho}{\sqrt{\mathrm{Var}(S_k) + \rho^2/3}} \leq L(S_k, \rho) \leq \frac{2\rho}{\sqrt{\mathrm{Var}(S_k) + \rho^2/3}}. \tag{3.46}$$

Corollary 3.8. *For the problem of computing a linear function in (3.44), where the W_v's are independent random variables with log-concave distributions and with variances at least σ^2, suppose that the coefficients a_v are all nonzero. Then for $\varepsilon \geq 0$ and $\delta \in (0, 1/2]$,*

$$T(\varepsilon, \delta) \geq \max_{\mathcal{S} \subset \mathcal{V}} \frac{1}{C_{\mathcal{S}}} \left(\frac{1-\delta}{2} \log \left(\frac{\sigma^2 \|a_{\mathcal{S}^c}\|_2^2}{4\varepsilon^2} + \frac{1}{12} \right) - h_2(\delta) \right).$$

Proof. For each $v \in \mathcal{V}$, $a_v W_v$ also has a log-concave distribution, and, for any $\mathcal{S} \subset \mathcal{V}$,

$$\mathrm{Var}(\langle a_{\mathcal{S}^c}, W_{\mathcal{S}^c} \rangle) = \sum_{v \in \mathcal{S}^c} |a_v|^2 \mathrm{Var}(W_v) \geq \|a_{\mathcal{S}^c}\|_2^2 \sigma^2.$$

The lower bound follows from Theorem 3.1 and from (3.46). □

Sums of Independent r.v.'s with Bounded Third Moments

It is known that random variables with log-concave distributions have bounded moments of any order. Under a much weaker assumption that the local observations W_v, $v \in \mathcal{V}$ have bounded third moments, we can prove the following result.

Corollary 3.9. *Consider the problem of computing the linear function in (3.44), where the W_v's are independent zero-mean r.v.'s with variances at*

least 1 *and with third moments bounded by* B*, and the coefficients* a_v *satisfy the constraint* $K_1 \leq |a_v| \leq K_2$ *for some* $K_1, K_2 > 0$*. Then for* $\varepsilon \geq 0$ *and* $\delta \in (0, 1/2]$*,*

$$T(\varepsilon, \delta) \geq \max_{\mathcal{S} \subset \mathcal{V}} \frac{1}{C_{\mathcal{S}}} \left(\frac{1-\delta}{2} \log \frac{|\mathcal{V} \setminus \mathcal{S}|}{M^2(\varepsilon)} - h_2(\delta) \right),$$

where $M(\varepsilon) \triangleq c\big(\varepsilon/K_1 + B(K_2/K_1)^3\big)$ *with some absolute constant* c*.*

Proof. Under the conditions of the theorem, a small ball estimate due to Rudelson and Vershynin [65, Corollary 2.10] can be used to show that, for any $\mathcal{S} \subset \mathcal{V}$,

$$L(\langle a_{\mathcal{S}}, W_{\mathcal{S}} \rangle, \varepsilon) \leq \frac{M(\varepsilon)}{\sqrt{|\mathcal{S}|}}.$$

The desired conclusion follows immediately. □

3.4.2 Linear Vector-valued Functions

Similar to the Lévy concentration function of a real-valued random variable, the Lévy concentration function of a random vector U taking values in $\mathbb{R}^n$ can be defined as

$$L(U, \rho) = \sup_{u \in \mathbb{R}^n} \mathbb{P}\left[\|U - u\|_2 \leq \rho\right], \qquad \rho > 0.$$

Consider the case where each node observes an independent real-valued random variable W_v, and the observations form a $|\mathcal{V}| \times 1$ vector $W_{\mathcal{V}}$. Suppose the nodes wish to compute a linear transform of $W_{\mathcal{V}}$,

$$Z = AW_{\mathcal{V}} \tag{3.47}$$

with some fixed $n \times |\mathcal{V}|$ matrix A, subject to the Euclidean-norm distortion criterion $\ell(z, \widehat{z}) = \|z - \widehat{z}\|_2$. In this case

$$\begin{aligned}
\mathcal{L}_{Z|W_{\mathcal{S}}}(w_{\mathcal{S}}, \varepsilon) &= \sup_{z \in \mathbb{R}^n} \mathbb{P}[\|AW_{\mathcal{V}} - z\|_2 \leq \varepsilon | W_{\mathcal{S}} = w_{\mathcal{S}}] \\
&= \sup_{z \in \mathbb{R}^n} \mathbb{P}[\|A_{\mathcal{S}^c} W_{\mathcal{S}^c} + A_{\mathcal{S}} w_{\mathcal{S}} - z\|_2 \leq \varepsilon] \\
&= \sup_{z \in \mathbb{R}^n} \mathbb{P}[\|A_{\mathcal{S}^c} W_{\mathcal{S}^c} - z\|_2 \leq \varepsilon] \\
&= L(A_{\mathcal{S}^c} W_{\mathcal{S}^c}, \varepsilon),
\end{aligned}$$

where $A_{\mathcal{S}^c}$ is the submatrix formed by the columns of A with indices in $\mathcal{S}^c$. We will need the following result, due to Rudelson and Vershynin [66]. Let $s_j(A_{\mathcal{S}^c})$, $j = 1, \ldots, \min\{n, |\mathcal{S}^c|\}$, denote the singular values of $A_{\mathcal{S}^c}$ arranged in non-increasing order, and define the stable rank of $A_{\mathcal{S}^c}$ by

$$r(A_{\mathcal{S}^c}) = \left\lfloor \frac{\|A_{\mathcal{S}^c}\|_{\mathrm{HS}}^2}{\|A_{\mathcal{S}^c}\|^2} \right\rfloor,$$

where $\|A_{\mathcal{S}^c}\|_{\mathrm{HS}} = \left(\sum_{j=1}^{\min\{n, |\mathcal{S}^c|\}} s_j(A_{\mathcal{S}^c})^2 \right)^{1/2}$ is the Hilbert-Schmidt norm of $A_{\mathcal{S}^c}$, and $\|A_{\mathcal{S}^c}\| = s_1(A_{\mathcal{S}^c})$ is the spectral norm of $A_{\mathcal{S}^c}$. (Note that for any non-zero matrix $A_{\mathcal{S}^c}$, $1 \leq r(A_{\mathcal{S}^c}) \leq \mathrm{rank}(A_{\mathcal{S}^c})$.) Then, provided

$$L(W_v, \varepsilon / \|A_{\mathcal{S}^c}\|_{\mathrm{HS}}) \leq p$$

for all $v \in \mathcal{S}^c$, we will have

$$L(A_{\mathcal{S}^c} W_{\mathcal{S}^c}, \varepsilon) \leq (cp)^{0.9 r(A_{\mathcal{S}^c})},$$

where c is an absolute constant [66, Theorem 1.4]. This result relates the Lévy concentration function of the linear transform of a vector to the Lévy concentration function of each coordinate of the vector. Applying this result in Theorem 3.1, we get a lower bound on $T(\varepsilon, \delta)$ for computing linear vector-valued functions.

Corollary 3.10. *For the problem of computing a linear transform of the observations defined in (3.47), where W_v's are independent real-valued r.v.'s,*

suppose the rows of A are nonzero vectors. Then for $\varepsilon \geq 0$ and $\delta \in (0, 1/2]$,

$$T(\varepsilon, \delta) \geq \max_{\mathcal{S} \subset \mathcal{V}} \frac{1}{C_{\mathcal{S}}} \left(0.9(1-\delta) r(A_{\mathcal{S}^c}) \log \frac{1}{c \max_{v \in \mathcal{S}^c} L(W_v, \varepsilon / \|A_{\mathcal{S}^c}\|_{\mathrm{HS}})} - h_2(\delta) \right)$$

for some absolute constant c.

3.4.3 Linear Function of Discrete Observations

Finally, we consider a case when the local observations W_v have discrete distributions. Specifically, let the W_v's be i.i.d. Rademacher random variables; i.e., each W_v takes values ± 1 with equal probability. We still use the absolute distortion function $\ell(z, \widehat{z}) = |z - \widehat{z}|$ to quantify the estimation error. In this case, the Lévy concentration function $L(\langle a_{\mathcal{S}}, W_{\mathcal{S}} \rangle, \varepsilon)$ will be highly sensitive to the *direction* of the vector $a_{\mathcal{S}}$, rather than just its norm. For example, consider the extreme case when $a_v = |\mathcal{V}|$ for a single node $v \in \mathcal{S}$, and all other coefficients are zero. Then $L(\langle a_{\mathcal{S}}, W_{\mathcal{S}} \rangle, 0) = L(|\mathcal{V}| W_v, 0) = 1/2$. On the other hand, if $a_v = 1$ for all $v \in \mathcal{V}$ and $|\mathcal{S}|$ is even, then

$$L(\langle a_{\mathcal{S}}, W_{\mathcal{S}} \rangle, 0) = 2^{-|\mathcal{S}|} \binom{|\mathcal{S}|}{|\mathcal{S}|/2} \sim \sqrt{\frac{2}{\pi |\mathcal{S}|}} \qquad \text{as } |\mathcal{S}| \to \infty,$$

where the last step is due to Stirling's approximation. Moreover, a celebrated result due to Littlewood and Offord, improved later by Erdős [67], says that, if $|a_v| \geq 1$ for all v, then

$$L(\langle a_{\mathcal{S}}, W_{\mathcal{S}} \rangle, 1) \leq 2^{-|\mathcal{S}|} \binom{|\mathcal{S}|}{\lfloor |\mathcal{S}|/2 \rfloor} \sim \sqrt{\frac{2}{\pi |\mathcal{S}|}} \qquad \text{as } |\mathcal{S}| \to \infty,$$

which translates into a lower bound on the $(1, \delta)$-computation time which is of the same order as the lower bound on the zero-error computation time.

Corollary 3.11. *For the problem of computing the linear function in (3.44), where the W_v's are independent Rademacher random variables, suppose that $|a_v| \geq 1$ for all v, and $\delta < 1/2$. Then*

$$T(0, \delta) \geq T(1, \delta) \gtrsim \max_{\mathcal{S} \subset \mathcal{V}} \frac{1}{C_{\mathcal{S}}} \left(\frac{1-\delta}{2} \log \frac{\pi |\mathcal{V} \setminus \mathcal{S}|}{2} - h_2(\delta) \right) \qquad \text{as } |\mathcal{S}| \to \infty.$$

3.4.4 Comparison with Existing Results

We illustrate the utility of the above bounds through comparison with some existing results. For example, Ayaso et al. [38] derive lower bounds on a related quantity

$$\tilde{T}(\varepsilon,\delta) \triangleq \inf\Big\{T \in \mathbb{N} : \exists \text{ a } T\text{-step algorithm } \mathcal{A} \text{ such that } \max_{v\in\mathcal{V}} \mathbb{P}\big[\widehat{Z}_v \notin [(1-\varepsilon)Z, (1+\varepsilon)Z]\,\big] < \delta\Big\}.$$

One of their results is as follows: if $Z = f(W)$ is a linear function of the form (3.44) and $(W_v) \overset{\text{i.i.d.}}{\sim} \text{Uniform}([1, 1+B])$ for some $B > 0$, then

$$\tilde{T}(\varepsilon,\delta) \geq \max_{\mathcal{S}\subset\mathcal{V}} \frac{|\mathcal{S}|}{2C_{\mathcal{S}}} \log \frac{1}{B\varepsilon^2 + \kappa\delta + (1/B)^{2/|\mathcal{V}|}} \tag{3.48}$$

for all sufficiently small $\varepsilon, \delta > 0$, where $\kappa > 0$ is a fixed constant [38, Theorem III.5]. Let us compare (3.48) with what we can obtain using our techniques. It is not hard to show that

$$\tilde{T}(\varepsilon,\delta) \geq T\big(\|a\|_1(1+B)\varepsilon, \delta\big), \tag{3.49}$$

where $\|a\|_1 = \sum_{v\in\mathcal{V}} |a_v|$ is the ℓ_1 norm of a. Moreover, since any r.v. uniformly distributed on a bounded interval of the real line has a log-concave distribution, we can use Corollary 3.8 to lower-bound the right-hand side of (3.49). This gives

$$\tilde{T}(\varepsilon,\delta) \geq \max_{\mathcal{S}\subset\mathcal{V}} \frac{1}{C_{\mathcal{S}}} \left(\frac{1-\delta}{2} \log \frac{B^2\|a_{\mathcal{S}^c}\|_2^2}{48(B+1)^2\|a\|_1^2\varepsilon^2} - h_2(\delta)\right) \tag{3.50}$$

for all sufficiently small $\varepsilon, \delta > 0$. We immediately see that this bound is tighter than the one in (3.48). In particular, the right-hand side of (3.48) remains bounded for vanishingly small ε and δ, and in the limit of $\varepsilon, \delta \to 0$ tends to

$$\max_{\mathcal{S}\subset\mathcal{V}} \frac{|\mathcal{S}|}{C_{\mathcal{S}}} \frac{\log B}{|\mathcal{V}|} \leq \frac{\log B}{\min_{\mathcal{S}\subset\mathcal{V}} C_{\mathcal{S}}}.$$

By contrast, as $\varepsilon, \delta \to 0$, the right-hand side of (3.50) grows without bound as $\log(1/\varepsilon)$.

Another lower bound on the (ε, δ)-computation time $T(\varepsilon, \delta)$ was obtained by Como and Dahleh [39]. Their starting point is the following continuum generalization of Fano's inequality [39, Lemma 2] in terms of conditional differential entropy: if $Z, \widehat{Z}$ are two jointly distributed real-valued r.v.'s, such that $\mathbb{E}Z^2 < \infty$, then, for any $\varepsilon > 0$,

$$h(Z|\widehat{Z}) \leq \mathbb{P}\big[|Z - \widehat{Z}| \leq \varepsilon\big] \log \varepsilon + \frac{1}{2} \log \big(16\pi e \mathbb{E}Z^2\big). \tag{3.51}$$

If we use (3.51) instead of Lemma 3.1 to lower-bound $I(Z; \widehat{Z}_v|W_{\mathcal{S}})$, then we get

$$T(\varepsilon, \delta) \geq \max_{\mathcal{S} \subset \mathcal{V}} \frac{1}{C_{\mathcal{S}}} \left(\frac{1-\delta}{2} \log \frac{1}{\varepsilon^2} + h(Z|W_{\mathcal{S}}) - \frac{1}{2} \log \big(16\pi e \mathbb{E}Z^2\big) \right). \tag{3.52}$$

Again, let us consider the case when $Z = f(W)$ is a linear function of the form (3.44) with all a_v nonzero and with $(W_v) \overset{\text{i.i.d.}}{\sim} N(0,1)$. Then (3.52) becomes

$$T(\varepsilon, \delta) \geq \max_{\mathcal{S} \subset \mathcal{V}} \frac{1}{C_{\mathcal{S}}} \left(\frac{1-\delta}{2} \log \frac{1}{\varepsilon^2} + \frac{1}{2} \log \frac{\|a_{\mathcal{S}^c}\|_2^2}{8\|a\|_2^2} \right). \tag{3.53}$$

The lower bound of our Corollary 3.7 will be tighter than (3.53) for all $\varepsilon > 0$ as long as

$$\frac{1-\delta}{2} \log \frac{\pi \|a_{\mathcal{S}^c}\|_2^2}{2} - h_2(\delta) \geq \frac{1}{2} \log \frac{\|a_{\mathcal{S}^c}\|_2^2}{8\|a\|_2^2}, \quad \forall \mathcal{S} \subset \mathcal{V}. \tag{3.54}$$

Note that the quantity on the right-hand side is nonpositive. More generally, for observations with log-concave distributions, the result of Lemma 3.1 can be weakened to get a lower bound involving the conditional differential entropy $h(Z|W_{\mathcal{S}})$, which is tighter than similar results obtained in [39].

Corollary 3.12. *If the observations W_v, $v \in \mathcal{V}$, have log-concave distributions, then for computing the sum $Z = \sum_{v \in \mathcal{V}} W_v$ subject to the absolute error criterion $\ell(z, \widehat{z}) = |z - \widehat{z}|$, for $\varepsilon \geq 0$ and $\delta \in (0, 1/2]$,*

$$T(\varepsilon, \delta) \geq \max_{\mathcal{S} \subset \mathcal{V}} \frac{1}{C_{\mathcal{S}}} \left((1-\delta) \left(h(Z|W_{\mathcal{S}}) + \log \frac{1}{2e\varepsilon} \right) - h_2(\delta) \right).$$

Proof. Let $p_{\mathcal{S}}(z)$ denote the probability density of $\sum_{v\in\mathcal{S}^c} W_v$. From (3.45),

$$\mathcal{L}_{Z|W_{\mathcal{S}}}(w_{\mathcal{S}},\varepsilon) = \sup_{z\in\mathbb{R}} \int_{z-\varepsilon}^{z+\varepsilon} p_{\mathcal{S}}(z)\mathrm{d}z \le 2\varepsilon\|p_{\mathcal{S}}\|_\infty, \quad \forall w_{\mathcal{S}} \in \prod_{v\in\mathcal{S}} \mathsf{W}_v, \tag{3.55}$$

where $\|p_{\mathcal{S}}\|_\infty$ is the sup norm of $p_{\mathcal{S}}$. By a result of Bobkov and Madiman [68, Proposition I.2], if U is a real-valued r.v. with a log-concave density p, then the differential entropy $h(U)$ is upper bounded by $\log e + \log \|p\|_\infty^{-1}$. Using this fact together with (3.55), the log-concavity of $p_{\mathcal{S}}$, and the fact that the W_v's are mutually independent, we can write

$$\begin{aligned}
\log \frac{1}{\mathbb{E}[\mathcal{L}_{Z|W_{\mathcal{S}}}(W_{\mathcal{S}},\varepsilon)]} &\ge \log\frac{1}{2\varepsilon} + \log\frac{1}{\|p_{\mathcal{S}}\|_\infty} \\
&\ge \log\frac{1}{2e\varepsilon} + h\Big(\sum_{v\in\mathcal{S}^c} W_v\Big) \\
&= \log\frac{1}{2e\varepsilon} + h(Z|W_{\mathcal{S}}).
\end{aligned}$$

Using this estimate in Theorem 3.1, we get the desired lower bound on $T(\varepsilon,\delta)$. □

3.5 Comparison with Upper Bounds on Computation Time

For the two-node mod-2 sum problem in Example 3.1, we have shown in Corollary 3.1 that the lower bound on computation given by Theorem 3.3 can tightly match the upper bound. In this section, we provide two more examples in which our lower bounds on computation time are tight. In the first example, our lower bound precisely captures the dependence of computation time on the number of nodes in the network. In the second example, our lower bound tightly captures the dependence of computation time on the accuracy parameter ε.

3.5.1 Rademacher Sum over a Dumbbell Network

Example 3.3. *Consider a dumbbell network of bidirectional BSCs with the same crossover probability. Formally, suppose $|\mathcal{V}|$ is even, and let the nodes be*

indexed from 1 *to* $|\mathcal{V}|$. *Nodes* 1 *to* $|\mathcal{V}|/2$ *form a clique (i.e., each pair of nodes is connected by a pair of BSCs), while nodes* $|\mathcal{V}|/2+1$ *to* $|\mathcal{V}|$ *form another clique. The two cliques are connected by a pair of BSCs between nodes* $|\mathcal{V}|/2$ *and* $|\mathcal{V}|/2+1$. *Each node initially observes a* $\mathrm{Bern}(\frac{1}{2})$ *(or Rademacher) r.v. The goal is for the nodes to compute the sum of the observations of all nodes. The distortion function is* $\ell(z,\widehat{z}) = |z-\widehat{z}|$.

By choosing the cutset as the pair of BSCs that joins the two cliques, our lower bound for random Rademacher sums in Corollary 3.11 gives the following lower bound on computation time.

Corollary 3.13. *For the problem of in Example 3.3, for* $\delta \in (0, 1/2)$,

$$T(0,\delta) \gtrsim \frac{1}{C}\left(\frac{1-\delta}{2}\log\frac{\pi|\mathcal{V}|}{4} - h_2(\delta)\right) \quad \text{as } |\mathcal{V}| \to \infty,$$

which implies

$$T(0,\delta) = \Omega\left(\log|\mathcal{V}|\right).$$

Now we show that the above lower bound matches the upper bound on the computation time, which turns out to be

$$T(0,\delta) = O\left(\log|\mathcal{V}|\right).$$

As shown by Gallager [48], for a fixed success probability, nodes $|\mathcal{V}|/2$ and $|\mathcal{V}|/2+1$ can learn the partial sum of the observations in their respective cliques in $O\big(\log\log|\mathcal{V}|\big)$ steps. These two nodes then exchange their partial sum estimates using binary block codes. Each partial sum can take $|\mathcal{V}|/2+1$ values, and can be encoded losslessly with $\log(|\mathcal{V}|/2+1)$ bits. The blocklength needed for transmission of the encoded partial sums is thus $O\big(\log(|\mathcal{V}|/2+1)\big)$, where the hidden factor depends on the required success probability and the channel crossover probability, but not on $|\mathcal{V}|$. Having learned the partial sum of the other clique, nodes $|\mathcal{V}|/2$ and $|\mathcal{V}|/2+1$ continue to broadcast this partial sum to other nodes in their own clique. This takes another $O\big(\log(|\mathcal{V}|/2+1)\big)$ step. In total, the computation can be done in $O\big(\log\log|\mathcal{V}|\big) + 2O\big(\log(|\mathcal{V}|/2+1)\big) = O(\log|\mathcal{V}|)$ steps, to have all nodes learn the sum of all observations, for any prescribed success probability. This shows that $T(0,\delta) = O\left(\log|\mathcal{V}|\right)$.

3.5.2 Distributed Averaging over Discrete Noisy Channels

Example 3.4. *Consider a network where the nodes are connected by binary erasure channels with the same erasure probability. Each node initially observes a log-concave r.v. The goal is for the nodes to compute the average of the observations of all nodes.*

For this example, Carli et al. [51] define the computation time as

$$\tilde{T}(\varepsilon) \triangleq \inf \left\{ T \in \mathbb{N} : \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \mathbb{E}\big[(Z - \widehat{Z}_v(t))^2\big] \leq \varepsilon, \, \forall t \geq T \right\}$$

and show that

$$\tilde{T}(\varepsilon) \leq c_1 + c_2 \frac{\log^3 \varepsilon^{-1}}{\log^2 \rho^{-1}}, \tag{3.56}$$

where ρ is the second largest singular value of the consensus matrix adapted to the network, and c_1 and c_2 are positive constants depending only on channel erasure probability. It can be shown that the above upper bound still holds (with different constants) when channels are BSCs.

We use Corollary 3.12 to derive the following lower bound on $\tilde{T}(\varepsilon)$.

Corollary 3.14. *For the problem in Example 3.4,*

$$\tilde{T}(\varepsilon) \geq \max_{S \subset \mathcal{V}} \frac{1}{C_S} \left(\frac{1}{2} \left(h(Z|W_S) + \log \frac{1}{4e|\mathcal{V}|} + \frac{1}{2} \log \frac{1}{\varepsilon} \right) - 1 \right). \tag{3.57}$$

Proof. Using Jensen's inequality twice, we can write

$$\begin{aligned} \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \mathbb{E}\big[(Z - \widehat{Z}_v(T))^2\big] &\geq \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \big(\mathbb{E}|Z - \widehat{Z}_v(T)|\big)^2 \\ &\geq \left(\frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \mathbb{E}|Z - \widehat{Z}_v(T)| \right)^2. \end{aligned}$$

Therefore, $|\mathcal{V}|^{-1} \sum_{v \in \mathcal{V}} \mathbb{E}\big[(Z - \widehat{Z}_v(T))^2\big] \leq \varepsilon$ implies that $\mathbb{E}|Z - \widehat{Z}_v(T)| \leq |\mathcal{V}|\sqrt{\varepsilon}$ for all $v \in |\mathcal{V}|$, and

$$\mathbb{P}\left[|Z - \widehat{Z}_v(T)| \geq \frac{|\mathcal{V}|\sqrt{\varepsilon}}{\delta} \right] \leq \delta, \qquad \forall v \in \mathcal{V}, \delta \in (0, 1/2]$$

by Markov's inequality. Then by Corollary 3.12,

$$\begin{aligned}\tilde{T}(\varepsilon) &\geq T\left(\frac{|\mathcal{V}|\sqrt{\varepsilon}}{\delta}, \delta\right) \\ &\geq \max_{\mathcal{S}\subset\mathcal{V}} \frac{1}{C_{\mathcal{S}}}\left((1-\delta)\left(h(Z|W_{\mathcal{S}}) + \log\frac{\delta}{2e|\mathcal{V}|\sqrt{\varepsilon}}\right) - h_2(\delta)\right).\end{aligned}$$

Choosing $\delta = 1/2$, we obtain (3.57). □

Alternatively, we can use the lower bound on $T(\varepsilon)$ to obtain a lower bound on $\tilde{T}(\varepsilon)$. Noting that $\tilde{T}(\varepsilon)$ can be lower bounded by $T(|\mathcal{V}|\varepsilon)$ with $\ell(z, \widehat{z}) = (z - \widehat{z})^2$. The lower bound for $T(\varepsilon)$ in Theorem 3.2 leads to

$$\tilde{T}(\varepsilon) \geq \max_{\mathcal{S}\subset\mathcal{V}} \frac{1}{C_{\mathcal{S}}}\left(h(Z|W_{\mathcal{S}}) + \frac{1}{2}\log\frac{1}{2\pi e|\mathcal{V}|\varepsilon}\right).$$

The above lower bounds imply that $\tilde{T}(\varepsilon)$ is necessarily logarithmic in ε^{-1}, which tightly matches the poly-logarithmic dependence on ε^{-1} in the upper bound given by (3.56). As pointed out in Carli et al. [69], it is possible to prove that a computation time logarithmic in ε^{-1} is achievable by embedding a quantized consensus algorithm for noiseless networks into the simulation framework developed by Rajagopalan and Schulman [50] for noisy networks.

3.6 Conclusion and Future Research Directions

We have studied the fundamental time limits of distributed function computation from an information-theoretic perspective. The computation time depends on the amount of information about the function value needed by each node and the rate for the nodes to accumulate such an amount of information. The small ball probability lower bound on conditional mutual information reveals how much information is necessary, while the cutset-capacity upper bound and the SDPI upper bound capture the bottleneck on the rate for the information to be accumulated. The multi-cutset analysis provides a more refined characterization of the information dissipation in a network.

Here are some questions that are worthwhile to consider in the future:

- In the multi-cutset analysis, the purpose of introducing self-loops when

reducing the network to a chain is to establish necessary Markov relations for proving upper bounds on $I(Z;\widehat{Z}_n|W_{\mathcal{S}})$ in bidirected chains, and the reason for considering left-bound nodes is to improve the lower bounds on computation time. We could have included all channels from $\mathcal{S}_i$ to $\mathcal{S}_i$ into the self-loop at node i' in G', but this would result in looser lower bounds on computation time (cf. the remark after Theorem 3.4). However, there might be other network reduction methods, e.g., different ways to construct the bidirected chain, that will yield even tighter lower bounds on computation time than our proposed method.

- In the first step of the derivation of Lemma 3.4 and Lemma 3.5, we have upper-bounded $I(Z;\widehat{Z}_v|W_{\mathcal{S}})$ using the ordinary data processing inequality as

$$I(Z;\widehat{Z}_v|W_{\mathcal{S}}) \leq I(W_{\mathcal{S}^c};\widehat{Z}_v|W_{\mathcal{S}}).$$

  One may wonder whether we can tighten this step by a judicious use of SDPIs. The answer is negative. It can be shown that

$$I(Z;\widehat{Z}_v|W_{\mathcal{S}}) \leq I(W_{\mathcal{S}^c};\widehat{Z}_v|W_{\mathcal{S}}) \sup_{w_{\mathcal{S}}\in\prod_{v\in\mathcal{S}}\mathsf{W}_v} \eta(\mathbb{P}_{W_{\mathcal{S}^c}|W_{\mathcal{S}}=w_{\mathcal{S}}}, \mathbb{P}_{Z|W_{\mathcal{S}^c},W_{\mathcal{S}}=w_{\mathcal{S}}}),$$

  where the contraction coefficient depends on the joint distribution of the observations $\mathbb{P}_W$ and the function $Z = f(W)$. However,

$$\eta(\mathbb{P}_{W_{\mathcal{S}^c}|W_{\mathcal{S}}=w_{\mathcal{S}}}, \mathbb{P}_{Z|W_{\mathcal{S}^c},W_{\mathcal{S}}=w_{\mathcal{S}}}) = 1$$

  for both discrete and continuous observations. For discrete observations, this is a consequence of the fact that [20]

$$\eta(\mathbb{P}_X, \mathbb{P}_{Y|X}) < 1 \Longleftrightarrow \text{graph } \big\{(x,y) : \mathbb{P}_X(x) > 0, \mathbb{P}_{Y|X}(y|x) > 0\big\} \text{ is connected,}$$

  and the fact that, for any $\mathbb{P}_{Y|X}$ induced by a deterministic function $f : \mathsf{X} \to \mathsf{Y}$, this graph is always disconnected. This condition can be extended to continuous alphabets [70]. It would be interesting to see whether *nonlinear* SDPIs, e.g., of the sort recently introduced by Polyanskiy and Wu [71], can be somehow applied here to tighten the

upper bounds.

- If the function to be computed is the identity mapping, i.e., $Z = W$, then the goal of the nodes is to distribute their observations to all other nodes in the network. In this case, our results on the computation time can provide non-asymptotic lower bounds on the blocklength of the codes for the source-channel coding problems in multi-terminal networks. In Example 3.2, we have considered one such case with discrete observations, and obtained lower bounds in Corollary 3.2 based on the single cutset analysis. It would be interesting to apply the multi-cutset analysis to the source-channel coding problems in multi-terminal, multi-hop networks.

3.7 Additional Proofs for Chapter 3

3.7.1 Proof of Lemma 3.7

The goal of this proof is to show that, given any T-step algorithm $\mathcal{A}$ running on G, we can construct a randomized T-step algorithm $\mathcal{A}'$ running on G' that simulates $\mathcal{A}$. Fix any T-step algorithm $\mathcal{A}$ that runs on G. For each t, we can factor the conditional distribution of the messages $X_t \triangleq (X_{v,t})_{v\in\mathcal{V}}$ given W, X^{t-1}, Y^{t-1} as follows:

$$\begin{aligned}\mathbb{P}_{X_t|W,X^{t-1},Y^{t-1}}(x_t|w,x^{t-1},y^{t-1}) &= \prod_{v\in\mathcal{V}} \mathbb{P}_{X_{v,t}|W_v,Y_v^{t-1}}(x_{v,t}|w_v,y_v^{t-1}) \\ &= \prod_{i=1}^{n}\prod_{v\in\mathcal{S}_i} \mathbb{P}_{X_{v,t}|W_v,Y_v^{t-1}}\left(x_{v,t}\middle|w_v,y_v^{t-1}\right) \\ &= \prod_{i=1}^{n} \mathbb{P}_{X_{\mathcal{S}_i,t}|W_{\mathcal{S}_i},Y_{\mathcal{S}_i}^{t-1}}\left(x_{\mathcal{S}_i,t}\middle|w_{\mathcal{S}_i},y_{\mathcal{S}_i}^{t-1}\right). \quad (3.58)\end{aligned}$$

Likewise, the conditional distribution of the received messages $Y_t \triangleq (Y_{v,t})_{v\in\mathcal{V}}$

given W, X^t, Y^{t-1} can be factored as

$$\begin{aligned}\mathbb{P}_{Y_t|W,X^t,Y^{t-1}}(y_t|w,x^t,y^{t-1}) &= \prod_{e\in\mathcal{E}} \mathbb{P}_{Y_{e,t}|X_{e,t}}(y_{e,t}|x_{e,t}) \\ &= \prod_{e\in\mathcal{E}} K_e(y_{e,t}|x_{e,t}) \\ &= \prod_{i=1}^{n} \prod_{u\in\mathcal{S}_i} \prod_{v\in\mathcal{V}:\,(u,v)\in\mathcal{E}} K_{(u,v)}(y_{(u,v),t}|x_{(u,v),t}). \quad (3.59)\end{aligned}$$

Since the successive partition of G ensures that nodes in $\mathcal{S}_i$ can communicate with nodes in $\mathcal{S}_j$ only if $|i-j| \leq 1$, the messages originating from $\mathcal{S}_i$ at step t can be decomposed as

$$\begin{aligned}X_{\mathcal{S}_i,t} &= (X_{(\mathcal{S}_i,\mathcal{S}_{i-1}),t}, X_{(\mathcal{S}_i,\mathcal{S}_{i+1}),t}, X_{(\mathcal{S}_i,\mathcal{S}_i),t}) \\ &= (X_{(\mathcal{S}_i,\mathcal{S}_{i-1}),t}, X_{(\mathcal{S}_i,\mathcal{S}_{i+1}),t}, X_{(\mathcal{S}_i,\overleftarrow{\partial}\mathcal{S}_i),t}, X_{(\mathcal{S}_i,\mathcal{S}_i\setminus\overleftarrow{\partial}\mathcal{S}_i),t}),\end{aligned}$$

and the messages received by nodes in $\mathcal{S}_i$ at step t can be decomposed as

$$\begin{aligned}Y_{\mathcal{S}_i,t} &= (Y_{(\mathcal{S}_{i-1},\mathcal{S}_i),t}, Y_{(\mathcal{S}_{i+1},\mathcal{S}_i),t}, Y_{(\mathcal{S}_i,\mathcal{S}_i),t}) \\ &= (Y_{(\mathcal{S}_{i-1},\mathcal{S}_i),t}, Y_{(\mathcal{S}_{i+1},\mathcal{S}_i),t}, Y_{(\mathcal{S}_i,\overleftarrow{\partial}\mathcal{S}_i),t}, Y_{(\mathcal{S}_i,\mathcal{S}_i\setminus\overleftarrow{\partial}\mathcal{S}_i),t}). \quad (3.60)\end{aligned}$$

According to the operation of algorithm $\mathcal{A}$, for each $(u,v) \in \mathcal{E}$ there exists a mapping $\varphi_{(u,v),t}$, such that $X_{(u,v),t} = \varphi_{(u,v),t}(W_u, Y_u^{t-1})$. By the definition of $\overleftarrow{\partial}\mathcal{S}_i$, we can write

$$X_{(\mathcal{S}_i,\mathcal{S}_{i-1}),t} = \big(\varphi_{(u,v),t}(W_u, Y_u^{t-1}) : (u,v) \in \mathcal{E}, u \in \overleftarrow{\partial}\mathcal{S}_i, v \in \mathcal{S}_{i-1}\big).$$

Thus, there exists a mapping $\overleftarrow{\varphi}_{\mathcal{S}_i,t}$, such that

$$X_{(\mathcal{S}_i,\mathcal{S}_{i-1}),t} = \overleftarrow{\varphi}_{\mathcal{S}_i,t}(W_{\overleftarrow{\partial}\mathcal{S}_i}, Y_{\overleftarrow{\partial}\mathcal{S}_i}^{t-1}), \quad (3.61)$$

where

$$Y_{\overleftarrow{\partial}\mathcal{S}_i,t} = \big(Y_{(\mathcal{S}_{i-1},\overleftarrow{\partial}\mathcal{S}_i),t}, Y_{(\mathcal{S}_{i+1},\overleftarrow{\partial}\mathcal{S}_i),t}, Y_{(\mathcal{S}_i,\overleftarrow{\partial}\mathcal{S}_i),t}\big). \quad (3.62)$$

By the same token, there exist mappings $\vec{\varphi}_{\mathcal{S}_i,t}$, $\mathring{\varphi}_{\mathcal{S}_i,t}$ and $\bar{\varphi}_{\mathcal{S}_i,t}$, such that

$$X_{(\mathcal{S}_i,\mathcal{S}_{i+1}),t} = \vec{\varphi}_{\mathcal{S}_i,t}(W_{\mathcal{S}_i}, Y_{\mathcal{S}_i}^{t-1}), \tag{3.63}$$

$$X_{(\mathcal{S}_i,\overleftarrow{\partial}\mathcal{S}_i),t} = \mathring{\varphi}_{\mathcal{S}_i,t}(W_{\mathcal{S}_i}, Y_{\mathcal{S}_i}^{t-1}), \tag{3.64}$$

$$X_{(\mathcal{S}_i,\mathcal{S}_i\setminus\overleftarrow{\partial}\mathcal{S}_i),t} = \bar{\varphi}_{\mathcal{S}_i,t}(W_{\mathcal{S}_i}, Y_{\mathcal{S}_i}^{t-1}). \tag{3.65}$$

Define the random variables

$$\begin{aligned}
W_i &\triangleq W_{\mathcal{S}_i},\\
X_{i,t} &= (X_{(i,i-1),t}, X_{(i,i+1),t}, X_{(i,i),t})\\
&\triangleq (X_{(\mathcal{S}_i,\mathcal{S}_{i-1}),t}, X_{(\mathcal{S}_i,\mathcal{S}_{i+1}),t}, X_{(\mathcal{S}_i,\overleftarrow{\partial}\mathcal{S}_i),t}),\\
Y_{i,t} &= (Y_{(i-1,i),t}, Y_{(i+1,i),t}, Y_{(i,i),t})\\
&\triangleq (Y_{(\mathcal{S}_{i-1},\mathcal{S}_i),t}, Y_{(\mathcal{S}_{i+1},\mathcal{S}_i),t}, Y_{(\mathcal{S}_i,\overleftarrow{\partial}\mathcal{S}_i),t}),\\
U_{i,t} &\triangleq (X_{(\mathcal{S}_i,\mathcal{S}_i\setminus\overleftarrow{\partial}\mathcal{S}_i),t}, Y_{(\mathcal{S}_i,\mathcal{S}_i\setminus\overleftarrow{\partial}\mathcal{S}_i),t}).
\end{aligned}$$

From the decomposition of $Y_{\mathcal{S}_i,t}$ in (3.60), we know that (Y_i^{t-1}, U_i^{t-1}) contains $Y_{\mathcal{S}_i}^{t-1}$; while from the decomposition of $Y_{\overleftarrow{\partial}\mathcal{S}_i,t}$ in (3.62), we know that Y_i^{t-1} contains $Y_{\overleftarrow{\partial}\mathcal{S}_i}^{t-1}$. Therefore, from (3.61) and (3.63)-(3.65), we deduce the existence of mappings $\overleftarrow{\varphi}_{i,t}$, $\vec{\varphi}_{i,t}$, $\mathring{\varphi}_{i,t}$, and $\bar{\varphi}_{i,t}$, such that the messages transmitted by nodes in $\mathcal{S}_i$ at time t can be generated as

$$X_{(i,i-1),t} = \overleftarrow{\varphi}_{i,t}(W_i, Y_i^{t-1}), \tag{3.66}$$

$$X_{(i,i+1),t} = \vec{\varphi}_{i,t}(W_i, Y_i^{t-1}, U_i^{t-1}), \tag{3.67}$$

$$X_{(i,i),t} = \mathring{\varphi}_{i,t}(W_i, Y_i^{t-1}, U_i^{t-1}), \tag{3.68}$$

$$X_{(\mathcal{S}_i,\mathcal{S}_i\setminus\overleftarrow{\partial}\mathcal{S}_i),t} = \bar{\varphi}_{i,t}(W_i, Y_i^{t-1}, U_i^{t-1}). \tag{3.69}$$

Note that the computation of $X_{(i,i-1),t}$ does not involve U_i^{t-1}. Next, the messages received by nodes in $\mathcal{S}_i$ at step t are related to the transmitted messages as

$$\begin{aligned}
X_{(i-1,i),t} &\xrightarrow{K_{(i-1,i)}} Y_{(i-1,i),t},\\
X_{(i+1,i),t} &\xrightarrow{K_{(i+1,i)}} Y_{(i+1,i),t},\\
X_{(i,i),t} &\xrightarrow{K_{(i,i)}} Y_{(i,i),t},
\end{aligned}$$

where the stochastic transition laws have the same form as those in (3.28) to

(3.30). In addition, since $X_{(\mathcal{S}_i,\mathcal{S}_i\setminus\overleftarrow{\partial}\mathcal{S}_i),t}$ and $Y_{(\mathcal{S}_i,\mathcal{S}_i\setminus\overleftarrow{\partial}\mathcal{S}_i),t}$ are related through the channels from $\mathcal{S}_i$ to $\mathcal{S}_i \setminus \overleftarrow{\partial}\mathcal{S}_i$, there exists a mapping $\kappa_{i,t}$ such that $Y_{(\mathcal{S}_i,\mathcal{S}_i\setminus\overleftarrow{\partial}\mathcal{S}_i),t}$ can be realized as

$$Y_{(\mathcal{S}_i,\mathcal{S}_i\setminus\overleftarrow{\partial}\mathcal{S}_i),t} = \kappa_{i,t}(X_{(\mathcal{S}_i,\mathcal{S}_i\setminus\overleftarrow{\partial}\mathcal{S}_i),t}, R_{i,t}), \tag{3.70}$$

where $R_{i,t}$ can be taken as a random variable uniformly distributed over $[0,1]$ and independent of everything else. From (3.69) and (3.70), we know that $U_{i,t}$ can be realized by a mapping $\vartheta_{i,t}$ as

$$U_{i,t} = \vartheta_{i,t}(W_i, Y_i^{t-1}, U_i^{t-1}, R_{i,t}). \tag{3.71}$$

Taking all of this into account, we can rewrite the factorization (3.58) as follows:

$$\begin{aligned}
&\mathbb{P}_{X_t|W,X^{t-1},Y^{t-1}}(x_t|w,x^{t-1},y^{t-1}) \\
&= \prod_{i=1}^{n} \mathbf{1}\big\{x_{(i-1,i),t} = \overleftarrow{\varphi}_{i,t}(w_i, y_i^{t-1})\big\} \cdot \mathbf{1}\big\{x_{(i,i+1),t} = \vec{\varphi}_{i,t}(w_i, y_i^{t-1}, u_i^{t-1})\big\} \\
&\qquad \cdot \mathbf{1}\big\{x_{(i,i),t} = \mathring{\varphi}_{i,t}(w_i, y_i^{t-1}, u_i^{t-1})\big\} \cdot \mathbf{1}\big\{x_{(\mathcal{S}_i,\mathcal{S}_i\setminus\overleftarrow{\partial}\mathcal{S}_i),t} = \bar{\varphi}_{i,t}(w_i, y_i^{t-1}, u_i^{t-1})\big\},
\end{aligned} \tag{3.72}$$

and we can rewrite the factorization (3.59) as

$$\begin{aligned}
&\mathbb{P}_{Y_t|W,X^t,Y^{t-1}}(y_t|w,x^t,y^{t-1}) \\
&= \prod_{i=1}^{n} K_{(i-1,i)}(y_{(i-1,i),t}|x_{(i-1,i),t}) \cdot K_{(i+1,i)}(y_{(i+1,i),t}|x_{(i+1,i),t}) \cdot K_{(i,i)}(y_{(i,i),t}|x_{(i,i),t}) \\
&\qquad \cdot \bigotimes_{(u,v)\in\mathcal{E}:u\in\mathcal{S}_i,v\in\mathcal{S}_i\setminus\overleftarrow{\partial}\mathcal{S}_i} K_{(u,v)}(y_{(\mathcal{S}_i,\mathcal{S}_i\setminus\overleftarrow{\partial}\mathcal{S}_i),t}|x_{(\mathcal{S}_i,\mathcal{S}_i\setminus\overleftarrow{\partial}\mathcal{S}_i),t}),
\end{aligned} \tag{3.73}$$

where the channel $\bigotimes_{(u,v)\in\mathcal{E}:u\in\mathcal{S}_i,v\in\mathcal{S}_i\setminus\overleftarrow{\partial}\mathcal{S}_i} K_{(u,v)}$ can be realized by the mapping $\kappa_{i,t}$ with the r.v. $R_{i,t}$.

To summarize: the mappings defined in (3.66) to (3.68) and (3.71) specify a randomized T-step algorithm $\mathcal{A}'$ that runs on G' and simulates the T-step algorithm $\mathcal{A}$ that runs on G. Specifically, using these mappings, each node i' in G' can generate all the transmitted and received messages of $\mathcal{S}_i$ in $\mathcal{A}$ as $(X_{i'}^T, Y_{i'}^T, U_{i'}^T)$. Moreover, from (3.72) and (3.73) we see that the random

objects

$$\big(W_{\mathcal{S}_i}, X^T_{\mathcal{S}_i}, Y^T_{\mathcal{S}_i} : i \in \{1,\dots,n\}\big) \qquad \text{and} \qquad \big(W_{i'}, X^T_{i'}, Y^T_{i'}, U^T_{i'} : i' \in \{1,\dots,n\}\big)$$

have the same joint distribution.

Finally, as we have assumed that $\overleftarrow{\partial}\mathcal{S}_i$'s are all nonempty, we can define

$$\widehat{Z}_i \triangleq \widehat{Z}_v = \psi_v(W_v, Y_v^T)$$

with an arbitrary $v \in \overleftarrow{\partial}\mathcal{S}_i$. From the definition of $Y_{i,t}$ and the fact that Y_i^T contains Y_v^T, it follows that there exists a mapping ψ_i such that

$$\widehat{Z}_i = \psi_i(W_i, Y_i^T).$$

Using this mapping, node i' in G' can generate the final estimate of the chosen $v \in \overleftarrow{\partial}\mathcal{S}_i$ in $\mathcal{A}$ as $\widehat{Z}_{i'}$, such that $(Z, \widehat{Z}_i : i \in \{1,\dots,n\})$ and $(Z, \widehat{Z}_{i'} : i \in \{1,\dots,n\})$ have the same joint distribution. This guarantees that

$$\begin{aligned}\max_{i' \in \mathcal{V}'} \mathbb{P}[\ell(Z, \widehat{Z}_{i'}) > \varepsilon] &= \max_{i \in \{1:n\}} \mathbb{P}[\ell(Z, \widehat{Z}_i) > \varepsilon] \\ &\le \max_{v \in \mathcal{V}} \mathbb{P}[\ell(Z, \widehat{Z}_v) > \varepsilon] \\ &\le \delta.\end{aligned}$$

The claim that $T(\varepsilon, \delta)$ for computing Z on G is lower bounded by $T'(\varepsilon, \delta)$ for computing Z on G' then follows from the definition of $T'(\varepsilon, \delta)$ in (3.31). This proves Lemma 3.7.

3.7.2 Proof of Lemma 3.8

Recall that, for any randomized T-step algorithm $\mathcal{A}'$, at step $t \in \{1,\dots,T\}$, node $i \in \{1,\dots,n\}$ computes the outgoing messages $X_{(i,i-1),t} = \overleftarrow{\varphi}_{i,t}(W_i, Y_i^{t-1})$, $X_{(i,i+1),t} = \vec{\varphi}_{i,t}(W_i, Y_i^{t-1}, U_i^{t-1})$, and $X_{(i,i),t} = \mathring{\varphi}_{i,t}(W_i, Y_i^{t-1}, U_i^{t-1})$, and the private message $U_{i,t} = \vartheta_{i,t}(W_i, Y_i^{t-1}, U_i^{t-1}, R_{i,t})$, where $R_{i,t}$ is the private randomness of node i. At step T, node i computes $\widehat{Z}_i = \psi_i(W_i, Y_i^T)$. We will use the Bayesian network formed by all the relevant variables and the d-separation criterion [60, Theorem 3.3] to find conditional independences among these variables. To simplify the Bayesian network, we merge some of

the variables by defining

$$\tilde{U}_{i,t} \triangleq (X_{(i,i),t}, X_{(i,i+1),t}, U_{i,t})$$

and

$$\tilde{Y}_{i,t} \triangleq (Y_{(i,i),t}, Y_{(i+1,i),t})$$

for $i \in \{1, \ldots, n\}$. The joint distribution of the variables can then be factored as

$$\begin{aligned}
&\mathbb{P}_{W,X^T,U^T,Y^T}(w, x^T, u^T, y^T) \\
&\quad = \mathbb{P}_W(w) \prod_{t=1}^{T} \prod_{i=1}^{n} \mathbf{1}\{x_{(i,i-1),t} = \overleftarrow{\varphi}_{i,t}(w_i, y_i^{t-1})\} \mathbb{P}_{\tilde{U}_{i,t}|W_i,Y_i^{t-1},\tilde{U}_i^{t-1}}(\tilde{u}_{i,t}|w_i, y_i^{t-1}, \tilde{u}_i^{t-1}) \\
&\qquad \times \prod_{i=1}^{n} \mathbb{P}_{Y_{(i-1,i),t}|\tilde{U}_{i-1,t}}(y_{(i-1,i),t}|\tilde{u}_{i-1,t}) \mathbb{P}_{\tilde{Y}_{i,t}|\tilde{U}_{i,t},X_{(i+1,i),t}}(\tilde{y}_{i,t}|\tilde{u}_{i,t}, x_{(i+1,i),t}).
\end{aligned} \tag{3.74}$$

The Bayesian network corresponding to this factorization for $n = 4$ and $T = 4$ is shown in Fig. 3.10 at the end of this chapter.

If $T = 0$, then $\widehat{Z}_n = \psi(W_n)$, hence $I(Z; \widehat{Z}_n|W_{2:n}) \leq I(Z; W_n|W_{2:n}) = 0$. For $T \geq 1$, we prove the upper bounds in the following steps, where we assume $n \geq 4$. The case $n = 3$ can be proved by skipping Step 2, and the case $n = 2$ can be proved by skipping Step 1 and Step 2.

Step 1:

For any i and t, define the shorthand $X_{i\leftarrow,t} \triangleq X_{(\mathcal{N}_{i\leftarrow},i),t}$, where $\mathcal{N}_{i\leftarrow}$ is the in-neighborhood of node i. From the Markov chain $W, Y_n^{T-1} \to X_{n\leftarrow,T} \to Y_{n,T}$ and Lemma 2.7, we follow the same argument as the one used for proving Lemma 3.5 to show that

$$\begin{aligned}
I(Z; \widehat{Z}_n|W_{2:n}) &\leq I(W_1; Y_n^T|W_{2:n}) \\
&\leq (1 - \eta_n) I(W_1; Y_n^{T-1}|W_{2:n}) + \eta_n I(W_1; Y_n^{T-1}, X_{n\leftarrow,T}|W_{2:n}).
\end{aligned}$$

Applying the d-separation criterion to the Bayesian network corresponding to (3.74) (see Fig. 3.10 at the end of this chapter for an illustration), we can

read off the Markov chain

$$W_1 \to W_{2:n}, Y_{n-1}^{t-1} \to Y_n^{t-1}, \tilde{U}_{n-1,t}, \tilde{U}_{n,t}$$

for $t \in \{1, \ldots, T\}$, since all trails from W_1 to $(Y_n^{t-1}, \tilde{U}_{n-1,t}, \tilde{U}_{n,t})$ are blocked by $(W_{2:n}, Y_{(n-2,n-1)}^{t-1})$, and all trails from $(Y_n^{t-1}, \tilde{U}_{n-1,t}, \tilde{U}_{n,t})$ to W_1 are blocked by $(W_{2:n}, \tilde{Y}_{n-1}^{t-1})$. This implies the Markov chain

$$W_1 \to W_{2:n}, Y_{n-1}^{T-1} \to Y_n^{T-1}, X_{n\leftarrow,T},$$

since $X_{(n-1,n),T}$ is included in $\tilde{U}_{n-1,T}$ and $X_{(n,n),T}$ is included in $\tilde{U}_{n,T}$. Consequently,[1]

$$I(W_1; Y_n^T | W_{2:n}) \leq (1-\eta_n) I(W_1; Y_n^{T-1} | W_{2:n}) + \eta_n I(W_1; Y_{n-1}^{T-1} | W_{2:n}). \quad (3.75)$$

Note that $I(W_1; Y_{n,1} | W_{2:n}) \leq I(W_1; X_{n\leftarrow,1} | W_{2:n}) \leq I(W_1; W_{\mathcal{N}_{n\leftarrow}} | W_{2:n}) = 0$.

Step 2:
For $i \in \{1, \ldots, n-3\}$, from the Markov chain $W, Y_{n-i}^{T-i-1} \to X_{(n-i)\leftarrow,T-i} \to Y_{n-i,T-i}$ and Lemma 2.7,

$$\begin{aligned} I(W_1; Y_{n-i}^{T-i} | W_{2:n}) \leq & (1-\eta_{n-i}) I(W_1; Y_{n-i}^{T-i-1} | W_{2:n}) + \\ & \eta_{n-i} I(W_1; Y_{n-i}^{T-i-1}, X_{(n-i)\leftarrow,T-i} | W_{2:n}) \end{aligned}$$

From the Bayesian network corresponding to (3.74), we can read off the Markov chain

$$W_1 \to W_{2:n}, Y_{n-i-1}^{t-1} \to Y_{n-i}^{t-1}, \tilde{U}_{n-i-1,t}, \tilde{U}_{n-i,t}, X_{(n-i+1,n-i),t}$$

for $t = 1, \ldots, T-i$, since all trails from W_1 to $(Y_{n-i}^{t-1}, \tilde{U}_{n-i-1,t}, \tilde{U}_{n-i,t}, X_{(n-i+1,n-i),t})$ are blocked by $(W_{2:n}, Y_{(n-i-2,n-i-1)}^{t-1})$, and all trails from

$$(Y_{n-i}^{t-1}, \tilde{U}_{n-i-1,t}, \tilde{U}_{n-i,t}, X_{(n-i+1,n-i),t})$$

to W_1 are blocked by $(W_{2:n}, \tilde{Y}_{n-i-1}^{t-1})$. This implies the Markov chain $W_1 \to W_{2:n}, Y_{n-i-1}^{T-i-1} \to Y_{n-i}^{T-i-1}, X_{(n-i)\leftarrow,T-i}$, since $X_{(n-i-1,n-i),T-i}$ is included in

[1] This follows from the ordinary DPI and from the fact that, if $X \to A, B \to C$ is a Markov chain, then $X \to B \to C$ is a Markov chain conditioned on $A = a$.

$\tilde{U}_{n-i-1,T-i}$ and $X_{(n-i,n-i),T-i}$ is included in $\tilde{U}_{n-i,T-i}$. Therefore,

$$I(W_1; Y_{n-i}^{T-i}|W_{2:n}) \leq (1-\eta_{n-i})I(W_1; Y_{n-i}^{T-i-1}|W_{2:n}) + \eta_{n-i}I(W_1; Y_{n-i-1}^{T-i-1}|W_{2:n}) \tag{3.76}$$

for $i \in \{1, \ldots, n-3\}$. Also note that

$$I(W_1; Y_{n-i,1}|W_{2:n}) \leq I(W_1; X_{(n-i)\leftarrow,1}|W_{2:n}) \leq I(W_1; W_{\mathcal{N}_{(n-i)\leftarrow}}|W_{2:n}) = 0.$$

Step 3:
Finally, we upper-bound $I(W_1; Y_2^{T-n+2}|W_{2:n})$ for $T \geq n-1$. From the Markov chain $W, Y_2^{t-1} \to X_{2\leftarrow,t} \to Y_{2,t}$ and Lemma 2.7,

$$I(W_1; Y_2^{T-n+2}|W_{2:n}) \leq (1-\eta_2)I(W_1; Y_2^{T-n+1}|W_{2:n}) + \eta_2 H(W_1|W_{2:n}). \tag{3.77}$$

This upper bound is useful only when $H(W_1|W_{2:n})$ is finite. If the observations are continuous r.v.'s, we can upper-bound $I(W_1; Y_2^{T-n+2}|W_{2:n})$ in terms of the channel capacity $C_{(1,2)}$:

$$\begin{aligned}
I(W_1; Y_2^{T-n+2}|W_{2:n}) &= \sum_{t=1}^{T-n+2} I(W_1; Y_{2,t}|W_{2:n}, Y_2^{t-1}) \\
&= \sum_{t=1}^{T-n+2} I(W_1; Y_{(1,2),t}|W_{2:n}, Y_2^{t-1}) + I(W_1; \tilde{Y}_{2,t}|W_{2:n}, Y_2^{t-1}, Y_{(1,2),t}) \\
&\leq \sum_{t=1}^{T-n+2} I(X_{(1,2),t}; Y_{(1,2),t}|W_{2:n}, Y_2^{t-1}) \\
&\leq \sum_{t=1}^{T-n+2} I(X_{(1,2),t}; Y_{(1,2),t}) \\
&\leq C_{(1,2)}(T-n+2),
\end{aligned} \tag{3.78}$$

where we have used the Markov chain $W_1 \to W_{2:n}, Y_2^{t-1}, Y_{(1,2),t} \to \tilde{Y}_{2,t}$ for $t \in \{1, \ldots, T-n+2\}$, which follows by applying the d-separation criterion to the Bayesian network corresponding to the factorization in (3.74), so that the second term in the second line is zero; the Markov chain $W, Y_2^{t-1} \to X_{(1,2),t} \to Y_{(1,2),t}$, which also implies the Markov chain $W_1 \to X_{(1,2),t}, W_{2:n}, Y_2^{t-1} \to Y_{(1,2),t}$ by the weak union property of conditional independence, hence the

third line and the fourth line; and the fact that $I(X_{(1,2),t}; Y_{(1,2),t}) \le C_{(1,2)}$.

Step 4:
Define $I_{i,t} = I(W_1; Y_i^t | W_{2:i})$ for $i \ge 2$ and $t \ge 1$. From (3.75), (3.76), (3.77), and (3.78), we can write, for $n \ge 3$ and $T \ge n-1$,

$$I_{n-i,T-i} \le \bar{\eta}_{n-i} I_{n-i,T-i-1} + \eta_{n-i} I_{n-i-1,T-i-1}, \quad i \in \{0, \dots, n-3\} \tag{3.79}$$

where $\bar{\eta}_{n-i} = 1 - \eta_{n-i}$, and $I_{n-i,1} = 0$. In addition, for $T \ge n-1$,

$$I_{2,T-n+2} \le \begin{cases} \bar{\eta}_2 I_{2,T-n+1} + \eta_2 H(W_1|W_{2:n}) \\ C_{(1,2)}(T-n+2) \end{cases}, \tag{3.80}$$

and $I_{2,0} = 0$.

An upper bound on $I(W_1; Y_n^T | W_{2:n})$ can be obtained by solving this set of recursive inequalities with the specified boundary conditions. It can be checked by induction that $I(W_1; Y_n^T | W_{2:n}) = 0$ if $T \le n-2$. For $T \ge n-1$, if $\eta_i \le \tilde{\eta}$ for all $i \in \{1, \dots, n\}$, then the above inequalities continue to hold with η_i's replaced with $\tilde{\eta}$. The resulting set of inequalities is similar to the one obtained by Rajagopalan and Schulman [50] for the evolution of mutual information in broadcasting a bit over a unidirectional chain of BSCs. With $\mathcal{B}(m,k,p) \triangleq \binom{m}{k} p^k (1-p)^{m-k}$, the exact solution is given by

$$\begin{aligned} I(W_1; Y_n^T | W_{2:n}) &\le H(W_1|W_{2:n}) \tilde{\eta} \sum_{i=1}^{T-n+2} \tilde{\eta}^{n-2} (1-\tilde{\eta})^{T-i-n+2} \binom{T-i}{n-2} \\ &= H(W_1|W_{2:n}) \eta \sum_{i=1}^{T-n+2} \mathcal{B}(T-i, n-2, \eta) \end{aligned}$$

for $n \ge 2$, and

$$\begin{aligned} I(W_1; Y_n^T | W_{2:n}) &\le C_{(1,2)} \tilde{\eta} \sum_{i=1}^{T-n+2} \tilde{\eta}^{n-3} (1-\tilde{\eta})^{T-i-n+2} \binom{T-i-1}{n-3} i \\ &= C_{(1,2)} \eta \sum_{i=1}^{T-n+2} \mathcal{B}(T-i-1, n-3, \eta) i \end{aligned}$$

for $n \ge 3$. This proves (3.32a) and (3.32b).

For general η_i's, we obtain a suboptimal upper bound by unrolling the first term in (3.79) for each i and using the fact that $I_{n-i,t} = 0$ for $t \le n-i-2$,

getting

$$\begin{aligned} I_{n-i,T-i} &\le \bar{\eta}_{n-i}^{T-n+1}\eta_{n-i}I_{n-i-1,n-i-2} + \ldots + \bar{\eta}_{n-i}\eta_{n-i}I_{n-i-1,T-i-2} + \eta_{n-i}I_{n-i-1,T-i-1} \\ &\le \big(\bar{\eta}_{n-i}^{T-n+1} + \ldots + \bar{\eta}_{n-i} + 1\big)\eta_{n-i}I_{n-i-1,T-i-1} \\ &= \big(1 - \bar{\eta}_{n-i}^{T-n+2}\big)I_{n-i-1,T-i-1}. \end{aligned}$$

Iterating over i, and noting that $I_{2,T-n+2} \le \min\big\{H(W_1|W_{2:n})(1-\bar{\eta}_2^{T-n+2}), C_{(1,2)}(T-n+2)\big\}$, we get for $n \ge 2$ and $T \ge n-1$,

$$I(W_1; Y_n^T|W_{2:n}) \le \begin{cases} H(W_1|W_{2:n})\prod_{i=2}^n \big(1-(1-\eta_i)^{T-n+2}\big) \\ C_{(1,2)}(T-n+2)\prod_{i=3}^n \big(1-(1-\eta_i)^{T-n+2}\big) \end{cases}. \quad (3.81)$$

The weakened upper bounds in (3.33a) and (3.33b) are obtained by replacing η_i in (3.81) with $\eta \triangleq \max_{i=1,\ldots,n}\eta_i$.

Finally, we show (3.34) using an argument similar to the one in [50]. If $n \ge 4$ and $T \le 2 + (n-3)\gamma/\eta$ for some $\gamma \in (0,1)$, then

$$\eta < \frac{\eta}{\gamma} \le \frac{n-3}{T-2} \le \frac{n-2}{T-1} \le 1,$$

where the last inequality follows from the assumption that $T \ge n-1$, since otherwise $I(Z;\widehat{Z}_n|W_{2:n}) = 0$. The upper bounds in (3.32a) and (3.32b) can be weakened to

$$\begin{aligned} I(Z;\widehat{Z}_n|W_{2:n}) &\le \begin{cases} H(W_1|W_{2:n})\eta(T-n+2)\mathcal{B}(T-1,n-2,\eta) \\ C_{(1,2)}\eta(T-n+2)^2\mathcal{B}(T-2,n-3,\eta) \end{cases} \\ &\le \min\{H(W_1|W_{2:n}), C_{(1,2)}\}\eta(T-n+2)^2\mathcal{B}(T-2,n-3,\eta) \\ &\le C_{(1,2)}\eta(T-n+2)^2\exp\left(-2\left(\frac{n-3}{T-2}-\eta\right)^2(T-2)\right) \\ &\le C_{(1,2)}\frac{(n-3)^2\gamma^2}{\eta}\exp\left(-2\left(\frac{\eta}{\gamma}-\eta\right)^2(n-3)\right), \end{aligned}$$

where the first and second lines follow from monotonicity properties of the binomial distribution; the third line follows from the Chernoff–Hoeffding bound; and the fourth line follows from the assumption that $n \ge 4$ and $n-1 \le T \le 2+(n-3)\gamma/\eta$.

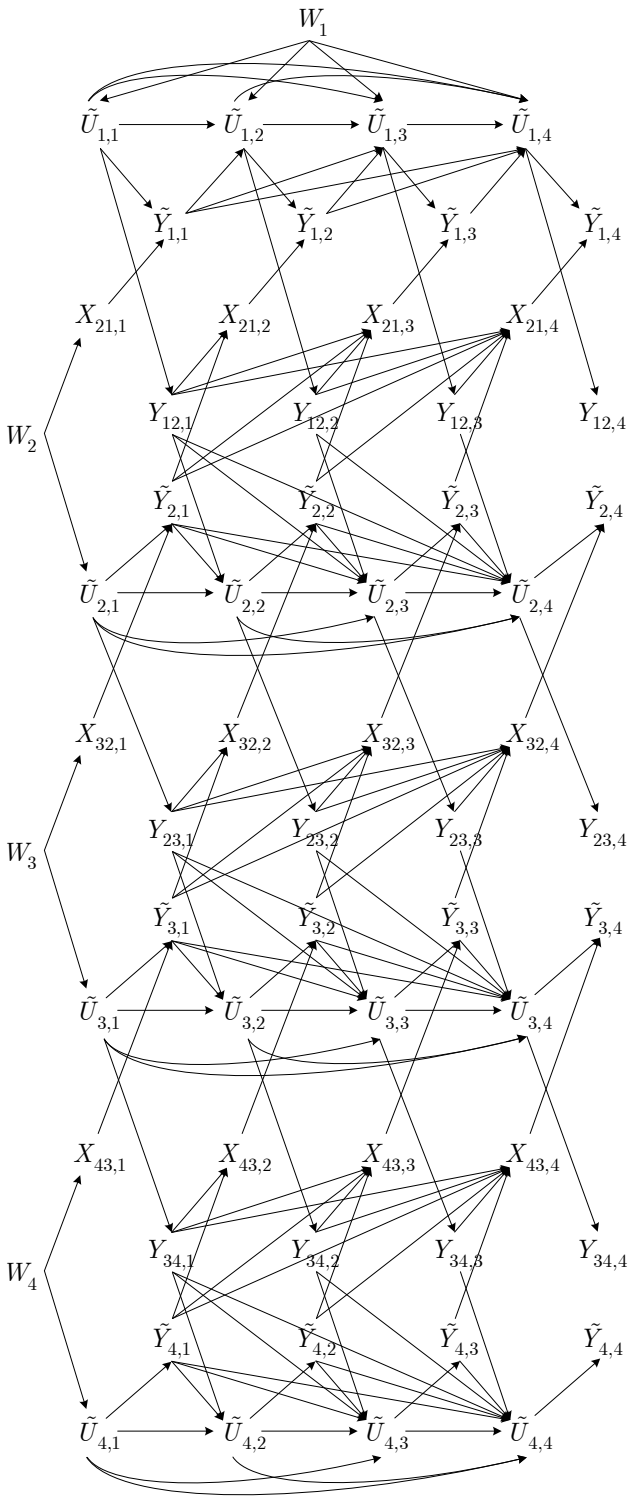


Figure 3.10: Bayesian network of (W, X^T, U^T, Y^T) for the randomized algorithm $\mathcal{A}'$ on a 4-node bidirected chain with $T = 4$. ($W_{1:4}$ are arbitrarily correlated, and not all edges emanating from $W_{2:4}$ are shown.)

Chapter 4

Upper Bounds for Generalization Error of Statistical Learning Algorithms

4.1 Introduction

Machine learning algorithms can be viewed as stochastic transformations that map training data to hypotheses. The performance of a learning algorithm is assessed by the true risk of its output hypothesis: based on a random training dataset, a good learning algorithm should generate a hypothesis with small true risk either in expectation or with high probability. The generalization error of a hypothesis is defined as the difference between its true risk and its empirical risk on the training data. A small generalization error means that the true risk of a hypothesis can be accurately estimated by its empirical risk. A hypothesis will have a small true risk if both its empirical risk and its generalization error are small, meaning that it both can fit the training data and is able to generalize, or in other words, does not overfit. It turns out that the generalization capability of a learning algorithm is determined by its stability properties, which pertain to sensitivity of the learning algorithm's output to local modifications of the input dataset. Algorithmic stability was introduced in the 1970's by Devroye and Wagner [72] and Rogers and Wagner [73] as a tool for estimating the generalization error, and studied more recently by Kearns and Ron [74], Bousquet and Elisseeff [75], Poggio et al. [76], and Shalev-Shwartz et al. [77] to establish sufficient and necessary conditions for learnability.

In recent years, the interest in stability was renewed through the work on differential privacy [78], which quantifies the sensitivity of the *distribution* of the algorithm's output to the dataset, and can therefore be viewed as a form of information-theoretic stability. Once the connection to generalization error bounds was established, it was used to study adaptive data analytics, where the analyst chooses queries by interacting with the dataset

in multiple rounds [79, 80]. Differential privacy also behaves nicely under composition of algorithms [81, 82], which makes it particularly amenable to information-theoretic analysis. Based on the idea that the distribution of the output of a stable learning algorithm cannot depend too much on any particular instance in the input dataset, a number of new information-theoretic notions of stability, e.g., stability in erasure mutual information and stability in Wasserstein distance, have been proposed recently by Raginsky et al. [83]. The notion of stability in erasure mutual information is weaker (i.e., less restrictive) than differential privacy; whereas stability in Wasserstein distance is a stronger notion, which is based on the theory of optimal transportation, and can be related to other information-theoretic stability notions via transportation-information inequalities.

In this chapter, we define a new notion of stability through the mutual information between the input dataset and the output hypothesis of a learning algorithm, and call it stability in input-output mutual information. This notion of stability naturally captures the idea that stability imposes limits on the amount of information the algorithm can glean from the observed data. We derive an upper bound on the expected generalization error for learning algorithms that are stable in input-output mutual information (Theorem 4.2). Our generalization error bound is similar to the result obtained by Russo and Zou [84], which is in terms of the mutual information between the output hypothesis and the collection of empirical risks. For learning algorithms that generate the output hypothesis solely based on the empirical risks, these two upper bounds are equivalent. However, the proof in [84] requires the hypothesis space to be a finite set, whereas our formulation allows uncountably infinite hypothesis spaces. Moreover, for algorithms that are stable in input-output mutual information, we derive a high-probability bound for the absolute generalization error that decays exponentially in the size of the dataset (Theorem 4.4). In addition, we discuss several approaches to designing learning algorithms with input-output mutual information stability, and show that the popular Gibbs algorithm [85] can be viewed as a regularized variant of empirical risk minimization, where the regularization controls the input-output mutual information.

We also discuss the input-output mutual information stability in the adaptive composition of learning algorithms, where a number of learning algorithms are sequentially executed, and the output of each algorithm may de-

pend on the dataset as well as on the outputs of the previously executed learning algorithms. Adaptive composition can be used as a way of obtaining complex learning algorithms by combining simple constituent algorithms, and can be realized in a decentralized fashion by multiple processors sharing the same dataset and running the constituent algorithms. By analyzing the input-output mutual information of each constituent algorithm, we can upper-bound the generalization error of the final output of the composed algorithm. The information-theoretic analysis also helps to capture the influence of communication constraints on the generalization capability of the algorithm obtained from adaptive composition. Finally, we apply the relationship between input-output mutual information stability and generalization error to analyzing bias and accuracy in adaptive data analytics.

4.2 Preliminaries

4.2.1 Formulation of General Statistical Learning Problem

In the standard framework of statistical learning theory [77, 86], we have an *instance space* Z, a *hypothesis space* W, and a nonnegative loss function $\ell : \mathsf{W} \times \mathsf{Z} \to \mathbb{R}^+$. The learning algorithm is given a *dataset* of size n, i.e., an n-tuple

$$S = (Z_1, \ldots, Z_n)$$

of i.i.d. random elements of Z with distribution μ, serving as training samples. The distribution μ of the training samples is unknown to the learning algorithm. A possibly randomized learning algorithm is a Markov kernel $P_{W|S}$ that maps a dataset S to a random element W of the hypothesis space W. The randomness in W comes from two sources: the randomness of the dataset S, and the private randomness utilized by the learning algorithm if it is randomized. The *true risk* of a hypothesis $w \in \mathsf{W}$ on μ is given by

$$L_\mu(w) \triangleq \mathbb{E}[\ell(w, Z)] = \int_{\mathsf{Z}} \ell(w, z)\mu(\mathrm{d}z). \tag{4.1}$$

The goal of a learning algorithm is to output a hypothesis W based on the input dataset S such that the true risk of W is small either in expectation or

with high probability. Had the learning algorithm known the data-generating distribution μ, it could pick a hypothesis from W that minimizes $L_\mu(w)$ and attain the minimum true risk over W, which is $\inf_{w\in\mathsf{W}} L_\mu(w)$. The *excess risk* of a learning algorithm is the difference $L_\mu(W) - \inf_{w\in\mathsf{W}} L_\mu(w)$, which is always nonnegative. For a learning algorithm characterized by $P_{W|S}$, the expected excess risk on μ is given by

$$R_{\text{excess}}(\mu, P_{W|S}) \triangleq \mathbb{E}[L_\mu(W)] - \inf_{w\in\mathsf{W}} L_\mu(w)$$

where the expectation is taken over the marginal distribution of W. A good learning algorithm should have a small expected excess risk for any data-generating distribution μ. Since μ is unknown, the learning algorithm cannot directly compute $L_\mu(w)$ for any $w \in \mathsf{W}$, but can instead compute the *empirical risk* of w

$$L_S(w) \triangleq \frac{1}{n}\sum_{i=1}^{n} \ell(w, Z_i) \tag{4.2}$$

as a proxy of $L_\mu(w)$. For a learning algorithm characterized by $P_{W|S}$, the *generalization error* on μ is the difference $L_\mu(W) - L_S(W)$, and we are interested in its expected value

$$\text{gen}(\mu, P_{W|S}) \triangleq \mathbb{E}[L_\mu(W) - L_S(W)],$$

where the expectation is taken with respect to the joint distribution $P_{S,W} = \mu^{\otimes n} \otimes P_{W|S}$. When a learning algorithm has a small empirical risk, it means that its output fits the training samples well. When a learning algorithm has a small generalization error, it means that the difference between the true risk of its output and the empirical risk of its output is small; in other words, the algorithm does not overfit. As will be discussed in the next subsection, if a learning algorithm has a small expected empirical risk, and at the same time has a small expected generalization error, then it will have a small expected excess risk.

The above framework of statistical learning applies to both supervised and unsupervised learning. As an example of supervised learning, consider the problem of learning a neural network for image classification. In this case, each training sample $Z_i = (X_i, Y_i)$, where $X_i \in \mathsf{X}$ is an image and $Y_i \in \mathsf{Y}$ is

the label of that image. The hypothesis space W is the set of predictors of the form $w : \mathsf{X} \to \mathsf{Y}$ that can be implemented by the neural network. For a given network structure, each $w \in \mathsf{W}$ is determined by a configuration of the set of weights over the edges in the network. For any $w \in \mathsf{W}$ and any instance $z = (x, y)$, the loss function takes the form $\ell(w, z) = \mathbf{1}\{w(x) \neq y\}$. A learning algorithm for a neural network takes a dataset S consisting of n training samples as input, and outputs a configuration of the set of weights in the network, hence a predictor $W \in \mathsf{W}$.

As an example of unsupervised learning, consider the problem of k-means clustering in $\mathbb{R}^d$. In this case, $\mathsf{Z} = \mathbb{R}^d$, W is the collection of all subsets of $\mathbb{R}^d$ of size k, and $\ell(w, z) = \min_{c \in w} \|z - c\|^2$. Here, each $w \in \mathsf{W}$ represents a set of k centroids, and the loss function measures the squared Euclidean distance between an instance z and its nearest centroid, according to the hypothesis w. A learning algorithm for k-means clustering takes as input a dataset S consisting of n i.i.d. training samples drawn from μ and outputs a set W of k centroids.

4.2.2 Trade-off Between Empirical Risk and Generalization Error

A learning algorithm aims to output a hypothesis $W \in \mathsf{W}$ with a small true risk $L_\mu(W)$ either in expectation or with high probability. The expected true risk can be expressed as

$$\begin{aligned} \mathbb{E}[L_\mu(W)] &= \mathbb{E}[L_S(W)] + \mathbb{E}[L_\mu(W)] - \mathbb{E}[L_S(W)] \\ &= \mathbb{E}[L_S(W)] + \mathrm{gen}(\mu, P_{W|S}). \end{aligned} \tag{4.3}$$

The first term in (4.3) is the expected empirical risk of W, which reflects how well the output hypothesis fits the training samples, while the second term in (4.3) is the expected generalization error, which reflects how well the output hypothesis generalizes. To minimize the expected true risk of the algorithm we need both terms to be small.

A learning algorithm is called an empirical risk minimization (ERM) algorithm if it always outputs a hypothesis $W_{\mathrm{ERM}} \in \mathsf{W}$ that minimizes the

empirical risk, i.e.,

$$L_S(W_{\mathrm{ERM}}) = \inf_{w \in \mathsf{W}} L_S(w).$$

Note that the expected minimum empirical risk is less than the minimum true risk, as

$$\mathbb{E}[L_S(W_{\mathrm{ERM}})] = \mathbb{E}\Big[\inf_{w \in \mathsf{W}} L_S(w)\Big] \leq \inf_{w \in \mathsf{W}} \mathbb{E}[L_S(w)] = \inf_{w \in \mathsf{W}} L_\mu(w). \tag{4.4}$$

A less restrictive requirement for a learning algorithm to have small empirical risk is *asymptotic empirical risk minimization* (AERM) [77], where the output hypothesis W is required to satisfy

$$\sup_{\mu} \mathbb{E}[L_S(W) - L_S(W_{\mathrm{ERM}})] \xrightarrow{n \to \infty} 0.$$

We say that a learning algorithm $P_{W|S}$ *generalizes on average* if

$$\sup_{\mu} \big|\mathrm{gen}(\mu, P_{W|S})\big| \xrightarrow{n \to \infty} 0.$$

We say that a learning algorithm $P_{W|S}$ is *consistent* if

$$\sup_{\mu} R_{\mathrm{excess}}(\mu, P_{W|S}) \xrightarrow{n \to \infty} 0.$$

The following theorem formalizes the intuition that if a learning algorithm has both small expected empirical risk and small expected generalization error, then it has small expected excess risk.

Theorem 4.1 (Shalev-Shwartz et al. [77])**.** *If a learning algorithm is AERM and generalizes on average, then it is consistent.*

Proof. For any data-generating distribution μ,

$$\begin{aligned}
R_{\mathrm{excess}}(\mu, P_{W|S}) &= \mathbb{E}[L_\mu(W)] - \inf_{w \in \mathsf{W}} L_\mu(w) \\
&= \mathbb{E}[L_\mu(W)] - \mathbb{E}[L_S(W)] + \mathbb{E}[L_S(W)] - \mathbb{E}[L_S(W_{\mathrm{ERM}})] + \\
&\quad \mathbb{E}[L_S(W_{\mathrm{ERM}})] - \inf_{w \in \mathsf{W}} L_\mu(w) \\
&\leq \mathrm{gen}(\mu, P_{W|S}) + (\mathbb{E}[L_S(W)] - \mathbb{E}[L_S(W_{\mathrm{ERM}})]),
\end{aligned}$$

where in the last step we have used (4.4). From the definition of AERM and

generalization on-average, we know that

$$\lim_{n\to\infty} \sup_{\mu} R_{\text{excess}}(\mu, P_{W|S}) \leq 0.$$

The claim follows because $R_{\text{excess}}(\mu, P_{W|S})$ is nonnegative for all n. □

However, it is generally impossible to minimize the expected empirical risk and the expected generalization error simultaneously: on one hand, if W contains a hypothesis that perfectly fits the training samples, then choosing this hypothesis will result in zero empirical risk, but at the same time lead to overfitting, such that the hypothesis would fail on fresh instances and result in large generalization error; on the other hand, by ignoring the training samples, the learning algorithm can output a hypothesis with zero expected generalization error (which will be shown in Sec. 4.4.2), but clearly this will lead to large empirical risk. Therefore, any learning algorithm faces a trade-off between the empirical risk and the generalization error. In this chapter, we focus on information-theoretic analysis of the generalization error. We will mainly show that we can control the generalization error of a learning algorithm by controlling the mutual information between the input and output of the algorithm. We will use the analytical results to design algorithms that balance the empirical risk and the generalization error, and also apply the results to adaptive composition of learning algorithms and adaptive data analytics.

4.2.3 Adaptive Composition of Learning Algorithms

Adaptive composition is a way of obtaining complex learning algorithms by combining simple constituent algorithms. It can be realized in a decentralized fashion by multiple processors (machines) sharing the same dataset and running the constituent algorithms. Under a k-fold adaptive composition, the dataset S is shared by k processors. The jth processor runs a learning algorithm $A_j = P_{W_j|S,W^{j-1}}$, which outputs a random element $W_j \in \mathsf{W}$ based on the dataset S and the outputs W^{j-1} of the algorithms $A_1, \ldots, A_{j-1}$ run by the first $j-1$ processors. Figure 4.1 shows the dependence among the dataset and the algorithm outputs under a 4-fold adaptive composition.

Suppose each of the constituent algorithms $A_1, \ldots, A_k$ satisfies certain gen-

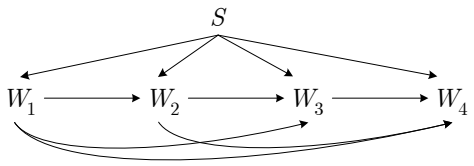


Figure 4.1: Dependence of the dataset and the algorithm outputs in a 4-fold composition.

eralization guarantees conditional on the outputs of the previous algorithms. We would like to find out what generalization performance the overall algorithm can achieve. Our information-theoretic analysis provides us with the right tool to tackle such a problem, so that we can upper-bound the generalization error of the composed learning algorithm using the knowledge of local generalization guarantees of the constituent algorithms. The same information-theoretic analysis can be applied to upper-bounding the bias in the adaptive data analytics [79,80], a topic that has become popular in recent years.

4.3 Stability and Generalization of Learning Algorithms

As discussed in Sec. 4.2.2, having a small generalization error is crucial for a learning algorithm to produce an output hypothesis with a small true risk. It turns out that the generalization error of a learning algorithm is determined by its stability properties. Roughly speaking, a learning algorithm is stable if a small change of the input to the algorithm does not change the output of the algorithm much. In this section, we first review the traditional notions of stability that quantify the variability of the output W relative to local changes of the dataset S. Then we introduce an information-theoretic notion of stability that measures the statistical dependence between the input and output of a learning algorithm based on the mutual information $I(S;W)$. We also review some other information-theoretic notions of stability and discuss their relationships.

4.3.1 Traditional Notions of Stability

On-average Stability

The first notion of stability we present is the on-average stability defined in [77], which is equivalent to generalization on average. Let $S' = (Z'_1, \ldots, Z'_n)$ be an i.i.d. copy of the dataset S, and let

$$S_{(i)} = (Z_1, \ldots, Z_{i-1}, Z'_i, Z_{i+1}, \ldots, Z_n).$$

In other words, $S_{(i)}$ is obtained by replacing the ith sample in S with the ith sample in S'. For the same learning algorithm $P_{W|S}$, let W be its output when the input dataset is S, and let $W_{(i)}$ be its output when the input dataset is $S_{(i)}$. We say an algorithm is (ε, μ)-*stable on average* if, under the data-generating distribution μ,

$$\left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[\ell(W_{(i)}, Z_i)] - \mathbb{E}[\ell(W, Z_i)] \right| \leq \varepsilon,$$

where the expectations are taken over the random tuples $(S_{(i)}, W_{(i)}, Z_i)$ and (S, W), respectively. A learning algorithm is said to be *stable on average* if

$$\sup_{\mu} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[\ell(W_{(i)}, Z_i)] - \mathbb{E}[\ell(W, Z_i)] \right| \xrightarrow{n \to \infty} 0.$$

It is straightforward to show that

$$\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[\ell(W_{(i)}, Z_i)] - \mathbb{E}[\ell(W, Z_i)] = \mathrm{gen}(\mu, P_{W|S})$$

by noting that $W_{(i)}$ is independent of Z_i. Therefore, a learning algorithm generalizes on average if and only if it is stable on average.

Uniform Stability

A stronger stability notion is uniform stability [75]. We say that an algorithm is ε-*uniformly stable* if, for all datasets $s, s' \in \mathsf{Z}^n$ differing in at most one

instance, i.e., their Hamming distance $d_{\mathrm{H}}(s, s') \leq 1$,

$$\sup_{z \in \mathsf{Z}} \left| \mathbb{E}[\ell(W, z)|S = s] - \mathbb{E}[\ell(W, z)|S = s'] \right| \leq \varepsilon.$$

Note that if an algorithm is ε-uniformly stable, then for all $i = 1, \ldots, n$,

$$\begin{aligned}
&\mathbb{E}[\ell(W_{(i)}, Z_i)] - \mathbb{E}[\ell(W, Z_i)] \\
&= \int_{\mathsf{Z}^n} \mu^{\otimes n}(\mathrm{d}s_{(i)}) \int_{\mathsf{Z}} \mu(\mathrm{d}z_i) \mathbb{E}[\ell(W_{(i)}, z_i)|S_{(i)} = s_{(i)}] - \int_{\mathsf{Z}^n} \mu^{\otimes n}(\mathrm{d}s) \mathbb{E}[\ell(W, z_i)|S = s] \\
&= \int_{\mathsf{Z}^{n+1}} \mu^{\otimes n}(\mathrm{d}s) \mu(\mathrm{d}z_i') \big(\mathbb{E}[\ell(W, z_i)|S = s_{(i)}] - \mathbb{E}[\ell(W, z_i)|S = s] \big) \\
&\leq \varepsilon;
\end{aligned}$$

hence the algorithm is also (ε, μ)-stable on average, and

$$\left| \mathrm{gen}(\mu, P_{W|S}) \right| \leq \varepsilon$$

for any μ. Moreover, the notion of uniform stability is strong enough, such that for uniformly stable deterministic algorithms, we can derive high-probability bounds for the absolute generalization error $|L_\mu(W) - L_S(W)|$. Specifically, it is shown by Bousquet and Elisseeff [75] that if a deterministic learning algorithm is ε-uniformly stable, and the loss function ℓ takes values in $[0, 1]$, then

$$\mathbb{P}[|L_\mu(W) - L_S(W)| > \alpha + 2\varepsilon] \leq 2e^{-2n\alpha^2/(4n\varepsilon+1)^2}.$$

Thus, if $\varepsilon = O(1/n)$, then the probability for $|L_\mu(W) - L_S(W)|$ being large decays exponentially in n.

4.3.2 Information-theoretic Notions of Stability

Stability in Input-output Mutual Information

The traditional notions of stability suggest that the generalization capability of a learning algorithm hinges on how sensitive the output of the algorithm is to local modifications of the input dataset. It suggests that the more independent the output hypothesis W is of the input dataset S, the better

the learning algorithm generalizes. The dependence between S and W can be naturally measured by the mutual information between them, which prompts the following definition: a learning algorithm is (ε, μ)-*stable in input-output mutual information* if, under the data-generating distribution μ,

$$I(S; W) \leq \varepsilon.$$

This information-theoretic definition of stability says that the less information the output of a learning algorithm can provide about its input dataset, the more stable it is. We mainly focus on studying the consequences of this notion of stability in this chapter. In the next section, we will show that a learning algorithm that is (ε, μ)-stable in input-output mutual information has strong generalization guarantees. But before doing that, we review some other information-theoretic notions of stability, and discuss some of their relationships.

Stability in Erasure Mutual Information

As proposed by Raginsky et al. [83], a learning algorithm is (ε, μ)-stable in mutual information if, under the data-generating distribution μ,

$$\frac{1}{n} \sum_{i=1}^{n} I(Z_i; W | S^{-i}) \leq \varepsilon$$

where $S^{-i} = (Z_1, \ldots, Z_{i-1}, Z_{i+1}, \ldots, Z_n)$. To distinguish this notion of stability from the preceding definition of stability in input-output mutual information, we call it *stability in erasure mutual information* in this chapter, as the quantity $\sum_{i=1}^{n} I(Z_i; W | S^{-i})$ coincides with the erasure mutual information [87, Def. 6] between S and W. Since

$$\frac{1}{n} \sum_{i=1}^{n} I(W; Z_i | S^{-i}) = I(W; S) - \frac{1}{n} \sum_{i=1}^{n} I(W; S^{-i}),$$

we see that if an algorithm is (ε, μ)-stable in input-output mutual information, then it is (ε, μ)-stable in erasure mutual information.

On-average KL-stability

As a slight modification of the definition proposed by Wang et al. [88], we say a learning algorithm is (ε, μ)-*KL-stable on average* if, under the data-generating distribution μ,

$$\frac{1}{n}\sum_{i=1}^{n}\int \mu^{\otimes n}(\mathrm{d}s)\mu(\mathrm{d}z_i')D(P_{W|S=s}\|P_{W|S=s_{(i)}}) \leq \varepsilon, \tag{4.5}$$

where $s_{(i)} = (z_1, \ldots, z_{i-1}, z_i', z_i, \ldots, z_n)$. Using the i.i.d. assumption on S and the convexity of relative entropy, we can show that

$$I(Z_i; W|S^{-i}) \leq \int \mu^{\otimes n}(\mathrm{d}s)\mu(\mathrm{d}z_i')D(P_{W|S=s}\|P_{W|S=s_{(i)}}).$$

Therefore, if an algorithm is (ε, μ)-KL-stable on average, then it is (ε, μ)-stable in erasure mutual information.

KL-stability and TV-stability

The notions of KL-stability and TV-stability are introduced recently by Bassily et al. [80]. A learning algorithm is ε-*KL-stable* if

$$\sup_{s,s'\in\mathsf{Z}^n:\, d_{\mathrm{H}}(s,s')\leq 1} D(P_{W|S=s}\|P_{W|S=s'}) \leq \varepsilon$$

and is ε-*TV-stable* if

$$\sup_{s,s'\in\mathsf{Z}^n:\, d_{\mathrm{H}}(s,s')\leq 1} \|P_{W|S=s} - P_{W|S=s'}\|_{\mathrm{TV}} \leq \varepsilon.$$

It is clear that if a learning algorithm is ε-KL-stable, then it is (ε, μ)-on-average KL-stable for any μ, and hence (ε, μ)-stable in erasure mutual information for any μ. Moreover, from the variational representation of the total-variation distance

$$\|P - Q\|_{\mathrm{TV}} = \sup_{f:\mathsf{X}\to[0,1]} \mathbb{E}_P f(X) - \mathbb{E}_Q f(X) \tag{4.6}$$

we see that, if the loss function ℓ takes values in $[0,1]$, then for any $s, s' \in \mathsf{Z}^n$

$$\sup_{z \in \mathsf{Z}} \left| \mathbb{E}[\ell(W,z)|S=s] - \mathbb{E}[\ell(W,z)|S=s'] \right| \leq \|P_{W|S=s} - P_{W|S=s'}\|_{\mathrm{TV}}.$$

Therefore, if a learning algorithm is ε-TV-stable and $\ell(\cdot,\cdot) \in [0,1]$, then it is ε-uniformly stable. In addition, from Pinsker's inequality [1]

$$\|P_{W|S=s} - P_{W|S=s'}\|_{\mathrm{TV}} \leq \sqrt{\frac{1}{2} D(P_{W|S=s} \| P_{W|S=s'})}$$

it follows that, if a learning algorithm is ε-KL-stable and $\ell(\cdot,\cdot) \in [0,1]$, then it is $\sqrt{\varepsilon/2}$-uniformly stable.

(ε, δ)-Differential Privacy

Recently the concept of (ε, δ)-differential privacy, introduced in the context of statistical processing of databases, has been adopted as a notion of algorithmic stability for statistical learning and adaptive data analytics [79, 80]. A learning algorithm is said to be (ε, δ)-*differentially private* for some $\varepsilon \geq 0$ and $\delta \in [0,1]$ if, for any two datasets $s, s' \in \mathsf{Z}^n$ with $d_{\mathrm{H}}(s, s') \leq 1$ and for any measurable set $F \subseteq \mathsf{W}$,

$$P_{W|S=s}(F) \leq e^{\varepsilon} P_{W|S=s'}(F) + \delta.$$

The relationship between (ε,δ)-differential privacy and other information-theoretic stability notions has been a popular topic in the literature, and some recent results are obtained in [83] and [89].

Stability in Max-information

Dwork et al. [79] propose to measure the stability of a learning algorithm using the *max-information* between S and W, defined as

$$I_\infty(S;W) \triangleq \sup_{s \in \mathsf{Z}^n, w \in \mathsf{W}} \log \frac{\mathrm{d}P_{S,W}}{\mathrm{d}(P_S \otimes P_W)}(s,w).$$

[1] All logarithms are natural in this chapter.

A learning algorithm is *ε-stable in max-information* if

$$I_\infty(S;W) \le \varepsilon.$$

It can be shown that

$$I_\infty(S;W) = \sup_{s,s' \in \mathsf{Z}^n} \sup_{w \in \mathsf{W}} \log \frac{\mathrm{d}P_{W|S=s}}{\mathrm{d}P_{W|S=s'}}(w) \ge I(S;W).$$

Therefore, if a learning algorithm is ε-stable in max-information, then it is (ε, μ)-stable in input-output mutual information for any μ. Moreover, if a learning algorithm is $(\varepsilon, 0)$-differentially private, then it is ε-stable in max-information.

Stability in Wasserstein Distance

Suppose that W is a complete separable metric space with metric d. For $p \ge 1$, the p-Wasserstein distance between two probability measures P and Q on W is defined as [90]

$$\mathbb{W}_p(P,Q) \triangleq \left(\inf_{W \sim P, W' \sim Q} \mathbb{E}[d^p(W,W')] \right)^{1/p},$$

where the infimum is over all *couplings* of P and Q, i.e., random couples (W, W') taking values in the product space $\mathsf{W} \times \mathsf{W}$, such that the marginal distribution of W (respectively, W') is equal to P (respectively, Q).

As proposed by Raginsky et al. [83], a learning algorithm $P_{W|S}$ is ε-stable in p-Wasserstein distance if, for any two $s, s' \in \mathsf{Z}^n$ with $d_{\mathrm{H}}(s,s') \le 1$,

$$\mathbb{W}_p(P_{W|S=s}, P_{W|S=s'}) \le \varepsilon.$$

It is shown in [83] that if the function $w \mapsto \ell(w,z)$ is ρ-Lipschitz for any $z \in \mathsf{Z}$, i.e., $|\ell(w,z) - \ell(w',z)| \le \rho d(w,w')$, then a learning algorithm that is ε-stable in 1-Wasserstein distance implies that the algorithm is $\rho\varepsilon$-uniformly stable. To see this, fix s and s' with $d_{\mathrm{H}}(s,s') = 1$, and let $\Pi \in \mathcal{P}(\mathsf{W} \times \mathsf{W})$ be the optimal coupling of A_s and $A_{s'}$, i.e., the one that achieves $\mathbb{W}_1$. Then,

for any $z \in \mathsf{Z}$,

$$\begin{aligned}
&\left|\mathbb{E}[\ell(W,z)|S=s] - \mathbb{E}[\ell(W,z)|S=s']\right| \\
&= \left|\int_{\mathsf{W}} \ell(w,z) A_s(\mathrm{d}w) - \int_{\mathsf{W}} \ell(w,z) A_{s'}(\mathrm{d}w)\right| \\
&= \left|\int_{\mathsf{W}\times\mathsf{W}} (\ell(w,z) - \ell(w',z))\, \Pi(\mathrm{d}w, \mathrm{d}w')\right| \\
&\leq \rho \int_{\mathsf{W}} d(w,w') \Pi(\mathrm{d}w, \mathrm{d}w') \\
&= \rho \mathbb{W}_1(A_s, A_{s'}) \\
&\leq \rho\varepsilon.
\end{aligned}$$

4.4 Upper-bounding Generalization Error via $I(S;W)$

We have defined a learning algorithm to be (ε, μ)-stable in input-output mutual information, if under the data-generating distribution μ,

$$I(S;W) \leq \varepsilon.$$

Now we turn to deriving generalization guarantees for learning algorithms with this property.

4.4.1 A Decoupling Estimate

We start with a digression from the statistical learning problem to a more general problem. Suppose there is a pair of random variables S and W with joint distribution $P_{S,W}$. Let $\bar{S}$ be an independent copy of S, and $\bar{W}$ an independent copy of W, such that $P_{\bar{S},\bar{W}} = P_S \otimes P_W$. Consider an arbitrary real-valued function $f : \mathsf{S} \times \mathsf{W} \to \mathbb{R}$ of S and W. The problem is to upper-bound the absolute difference between $\mathbb{E}[f(S,W)]$ and $\mathbb{E}[f(\bar{S},\bar{W})]$. There are two cases where we can obtain information-theoretic upper bounds on this quantity.

The first case is when the function f takes values in the unit interval.[2] From a straightforward application of the variational representation of the

[2] Actually any bounded function will work; restricting the function to be bounded in $[0,1]$ is just for convenience.

total-variation distance in (4.6), we can show that

Lemma 4.1. *If $f(s,w) \in [0,1]$ for all $s \in \mathsf{S}$ and $w \in \mathsf{W}$, then*

$$\big|\mathbb{E}[f(S,W)] - \mathbb{E}[f(\bar{S},\bar{W})]\big| \le \|P_{S,W} - P_S \otimes P_W\|_{\mathrm{TV}}$$

where the right-hand side of the above inequality coincides with the so-called T-information between S and W [71].

The second case is when the random variable $f(\bar{S},\bar{W})$ is σ-subgaussian. Recall that a random variable X is σ-subgaussian if $\log \mathbb{E}[e^{\lambda(X-\mathbb{E}X)}] \le \lambda^2\sigma^2/2$ for all $\lambda \in \mathbb{R}$ [91].

Lemma 4.2. *If $f(\bar{S},\bar{W})$ is σ-subgaussian under $P_{\bar{S},\bar{W}} = P_S \otimes P_W$, then*

$$\big|\mathbb{E}[f(S,W)] - \mathbb{E}[f(\bar{S},\bar{W})]\big| \le \sqrt{2\sigma^2 I(S;W)}.$$

Proof. For any $s \in \mathsf{S}$ and $w \in \mathsf{W}$, let

$$F(s,w) \triangleq f(s,w) - \mathbb{E}[f(\bar{S},\bar{W})].$$

By the subgaussian assumption,

$$\log \mathbb{E}\big[e^{\lambda F(\bar{S},\bar{W})}\big] = \log \mathbb{E}\big[e^{\lambda(f(\bar{S},\bar{W}) - \mathbb{E}[f(\bar{S},\bar{W})])}\big] \le \frac{\lambda^2\sigma^2}{2} \qquad \forall \lambda \in \mathbb{R}. \tag{4.7}$$

Just like Russo and Zou [84], we exploit the Donsker–Varadhan variational representation of the relative entropy [92, Cor. 4.15]: for any two probability measures π, ρ on a common measurable space $(\Omega, \mathcal{F})$,

$$D(\pi\|\rho) = \sup_F \left\{ \int_\Omega F \,\mathrm{d}\pi - \log \int_\Omega e^F \mathrm{d}\rho \right\}, \tag{4.8}$$

where the supremum is over all measurable functions $F : \Omega \to \mathbb{R}$, such that $e^F \in L^1(\rho)$. From (4.8) and (4.7), we know that for any $\lambda \in \mathbb{R}$,

$$\begin{aligned} D(P_{S,W}\|P_S \otimes P_W) &\ge \mathbb{E}[\lambda F(S,W)] - \log \mathbb{E}\big[e^{\lambda F(\bar{S},\bar{W})}\big] \\ &\ge \lambda\big(\mathbb{E}[f(S,W)] - \mathbb{E}[f(\bar{S},\bar{W})]\big) - \frac{\lambda^2\sigma^2}{2}. \end{aligned} \tag{4.9}$$

The above inequality gives a nonnegative parabola in λ, whose discriminant

must be nonpositive, which implies

$$\left|\mathbb{E}[f(S,W)] - \mathbb{E}[f(\bar{S},\bar{W})]\right| \leq \sqrt{2\sigma^2 D(P_{S,W}\|P_S \otimes P_W)}.$$

The result follows by noting that $I(S;W) = D(P_{S,W}\|P_S \otimes P_W)$. □

Generally, the function $f(s,w)$ need not have an additive structure for $f(S,w)$ to be subgaussian. As an example, when $S = (Z_1,\ldots,Z_n)$ where Z_i's are i.i.d. standard Gaussian, if $f(s,w)$ is ρ-Lipchitz in s, i.e.,

$$|f(s,w) - f(s',w)| \leq \rho\|s - s'\|_2$$

then $f(S,w)$ is ρ-subgaussian [91].

4.4.2 Upper Bound on Expected Generalization Error

Upper-bounding the generalization error of a learning algorithm $P_{W|S}$ is a special case of the general problem considered in the preceding subsection. In this case,

$$S = (Z_1,\ldots,Z_n) \sim \mu^{\otimes n}$$

is the input dataset, and W is the output hypothesis of the learning algorithm. For some loss function $\ell : \mathsf{W} \times \mathsf{Z} \to \mathbb{R}^+$, let the function f take the form

$$f(s,w) = \frac{1}{n}\sum_{i=1}^{n} \ell(w,z_i).$$

For an arbitrary $w \in \mathsf{W}$, recall that the empirical risk is

$$L_S(w) \triangleq \frac{1}{n}\sum_{i=1}^{n} \ell(w,Z_i) = f(S,w)$$

and the true risk is

$$L_\mu(w) \triangleq \mathbb{E}[\ell(w,Z)] = \mathbb{E}[f(S,w)].$$

Also recall the expected generalization error of the learning algorithm $P_{W|S}$, which can be written as

$$\begin{aligned}\mathrm{gen}(\mu, P_{W|S}) &\triangleq \mathbb{E}[L_\mu(W) - L_S(W)] \\ &= \mathbb{E}[f(\bar{S}, \bar{W})] - \mathbb{E}[f(S, W)],\end{aligned}$$

where the joint distribution of S and W is $P_{S,W} = \mu^{\otimes n} \otimes P_{W|S}$.

Theorem 4.2. *Suppose $\ell(w, Z)$ is σ-subgaussian under μ for all $w \in \mathsf{W}$, i.e.,*

$$\mathbb{E}\big[e^{\lambda(\ell(w,Z) - \mathbb{E}[\ell(w,Z)])}\big] \le e^{\lambda^2\sigma^2/2} \qquad \forall \lambda \in \mathbb{R}. \tag{4.10}$$

Then when $S \sim \mu^{\otimes n}$,

$$\big|\mathrm{gen}(\mu, P_{W|S})\big| \le \sqrt{\frac{2\sigma^2}{n} I(S; W)}. \tag{4.11}$$

Proof. In this case

$$f(S, w) = \frac{1}{n} \sum_{i=1}^{n} \ell(w, Z_i). \tag{4.12}$$

Since $\ell(w, Z)$ is assumed to be σ-subgaussian and Z_i's are i.i.d. random variables, we have

$$\begin{aligned}\mathbb{E}[\exp\{\lambda(f(S, w) - \mathbb{E}f(S, w))\}] &= \mathbb{E}\Big[\exp\Big\{\sum_{i=1}^{n} \frac{\lambda}{n}\big(\ell(w, Z_i) - L_\mu(w)\big)\Big\}\Big] \\ &= \prod_{i=1}^{n} \mathbb{E}\Big[\exp\Big\{\frac{\lambda}{n}\big(\ell(w, Z_i) - L_\mu(w)\big)\Big\}\Big] \\ &\le \exp\Big\{\frac{\lambda^2\sigma^2}{2n}\Big\},\end{aligned}$$

which means that $f(S, w)$ is $\sigma/\sqrt{n}$-subgaussian for all $w \in \mathsf{W}$, hence $f(\bar{S}, \bar{W})$ is $\sigma/\sqrt{n}$-subgaussian. The claim then follows from Lemma 4.2. □

Theorem 4.2 implies that if a learning algorithm is (ε, μ)-stable in input-output mutual information and if $\ell(w, Z)$ is σ-subgaussian under μ, then

$$\big|\mathrm{gen}(\mu, P_{W|S})\big| \le \sqrt{\frac{2\sigma^2\varepsilon}{n}}.$$

It suggests that, by controlling the mutual information between the input and the output of a learning algorithm, we can control the learning algorithm's generalization error.

Comparison with the Result by Russo and Zou

Russo and Zou [84] considered the case where the hypothesis space W is a finite set and showed that, if $\ell(w, Z)$ is σ-subgaussian for all $w \in \mathsf{W}$, then

$$\left|\mathrm{gen}(\mu, P_{W|S})\right| \le \sqrt{\frac{2\sigma^2}{n} I(\Lambda_{\mathsf{W}}(S); W)}, \tag{4.13}$$

where

$$\Lambda_{\mathsf{W}}(S) \triangleq \big(L_S(w)\big)_{w \in \mathsf{W}}$$

is the collection of empirical risks of the hypotheses in W. Since for each $w \in \mathsf{W}$, $L_S(w)$ is a deterministic functions of S, we always have the Markov chain

$$\Lambda_{\mathsf{W}}(S) - S - W,$$

hence

$$I(\Lambda_{\mathsf{W}}(S); W) \le I(S; W).$$

Thus, in the case considered in [84], the result of Theorem 4.2 can be obtained as a consequence of (4.13). On the other hand, if the output W of the learning algorithm depends on S only through the empirical risks $\Lambda_{\mathsf{W}}(S)$, in other words, when the Markov chain

$$S - \Lambda_{\mathsf{W}}(S) - W$$

holds, then

$$I(\Lambda_{\mathsf{W}}(S); W) = I(S; W)$$

and the result of Theorem 4.2 implies (4.13). The advantage of Theorem 4.2 is that it does not put any restriction on W, which is allowed to be an uncountably infinite set.

Comparison with the Upper bound via Erasure Mutual Information

It is shown by Raginsky et al. [83] that, under certain regularity conditions on the loss function, if a learning algorithm is (ε, μ)-stable in erasure mutual information, then the expected generalization error is upper bounded by $\sqrt{2\sigma^2\varepsilon}$:

Theorem 4.3 ([83, Theorem 2]). *If for any $s \in \mathsf{Z}^n$ and any $i \in [n]$, $\ell(W, z_i)$ is σ-subgaussin with respect to $P_{W|S^{-i}=s^{-i}}$, i.e.*

$$\log \mathbb{E}\Big[\exp\left\{\lambda(\ell(W, z_i) - \mathbb{E}[\ell(W, z_i)|S^{-i} = s^{-i}])\right\} \Big| S^{-i} = s^{-i}\Big] \leq \frac{\lambda^2\sigma^2}{2} \quad (4.14)$$

for all $\lambda \in \mathbb{R}$, where

$$S^{-i} = (S_1, \ldots, S_{i-1}, S_{i+1}, \ldots, S_n),$$

then

$$\big|\mathrm{gen}(\mu, P_{W|S})\big| \leq \sqrt{\frac{2\sigma^2}{n}\sum_{i=1}^{n} I(W; Z_i|S^{-i})}. \quad (4.15)$$

For completeness, we include the proof of Theorem 4.3 in Sec. 4.8.1. The results in Theorem 4.2 and Theorem 4.3 are complementary to each other. Since

$$\frac{1}{n}\sum_{i=1}^{n} I(W; Z_i|S^{-i}) = I(W; S) - \frac{1}{n}\sum_{i=1}^{n} I(W; S^{-i}),$$

requiring an algorithm to have a small $I(S; W)$ for a given μ is more restrictive than to have a small $\frac{1}{n}\sum_{i=1}^{n} I(W; Z_i|S^{-i})$. However, having a small $I(S; W)$ implies that the absolute generalization error $|L_\mu(W) - L_S(W)|$ is small with high probability, as discussed in the next subsection.

4.4.3 High-probability Bound on $|L_\mu(W) - L_S(W)|$

We now turn to the study of guarantees on the absolute generalization error $|L_\mu(W) - L_S(W)|$ of a learning algorithm which is (ε, μ)-stable in input-output mutual information. The objective is to derive a high-probability bound for $|L_\mu(W) - L_S(W)|$ to be small. First of all, for any fixed $w \in \mathsf{W}$,

if the loss function takes values in $[0,1]$, then from the Chernoff-Hoeffding bound,

$$\mathbb{P}[|L_\mu(w) - L_S(w)| > \alpha] \leq 2e^{-2n\alpha^2}.$$

It means that for each $w \in \mathsf{W}$, to make

$$\mathbb{P}[|L_\mu(w) - L_S(w)| > \alpha] \leq \beta$$

it is sufficient to have a sample complexity

$$n \geq \frac{1}{2\alpha^2} \log \frac{2}{\beta},$$

which is polynomial in $1/\alpha$ and logarithmic in $1/\beta$. The following results (cf. Corollary 4.2) show that, even if W is a random element from W that is dependent on S, as long as the mutual information $I(S;W)$ is small, a sample complexity polynomial in $1/\alpha$ and logarithmic in $1/\beta$ still suffices to guarantee $\mathbb{P}[|L_\mu(W) - L_S(W)| > \alpha] \leq \beta$, where the probability now is taken with respect to the joint distribution of S and W.

Theorem 4.4. *Suppose the loss function ℓ takes values in $[0,1]$, and the dataset S has the distribution $\mu^{\otimes n}$. For any $0 < \alpha, \beta \leq 1$, if an algorithm $P_{W|S}$ is (ε, μ)-stable in input-output mutual information and $\varepsilon \geq \frac{8\beta}{\alpha} \log \frac{2}{\beta}$, then choosing $n \geq \frac{16\varepsilon}{\alpha^2\beta}$ guarantees that $\mathbb{P}[|L_S(W) - L_\mu(W)| > \alpha] \leq \beta$.*

Choosing $\varepsilon = \frac{8\beta}{\alpha} \log \frac{2}{\beta}$ in Theorem 4.4, we get the following result.

Corollary 4.1. *Suppose the loss function ℓ takes values in $[0,1]$, and the dataset S has the distribution $\mu^{\otimes n}$. For any $0 < \alpha, \beta \leq 1$, if an algorithm $P_{W|S}$ satisfies*

$$I(S;W) \leq \frac{8\beta}{\alpha} \log \frac{2}{\beta},$$

then choosing

$$n \geq \frac{128}{\alpha^3} \log \frac{2}{\beta}$$

guarantees that $\mathbb{P}[|L_S(W) - L_\mu(W)| > \alpha] \leq \beta$.

Another corollary of Theorem 4.4 can be stated as follows.

Corollary 4.2. *Suppose the loss function ℓ takes values in $[0,1]$, and the*

dataset S has the distribution $\mu^{\otimes n}$. For any $0 < \alpha, \beta \leq 1$, if

$$n \geq \frac{128}{\alpha^3} \log \frac{2}{\beta}$$

and if the algorithm $P_{W|S}$ satisfies

$$I(S;W) \leq \frac{\alpha^2 \beta}{16} n,$$

then $\mathbb{P}[|L_S(W) - L_\mu(W)| > \alpha] \leq \beta$.

Note that the above results are independent of the size of the hypothesis space W, which is allowed to be an uncountable set. To prove Theorem 4.4, we need the following lemmas.

Lemma 4.3. *Consider the parallel execution of m independent copies of $P_{W|S}$ on independent datasets $S_1, \ldots, S_m$: for $t = 1, \ldots, m$, an independent copy of $P_{W|S}$ takes $S_t \sim \mu^{\otimes n}$ as input and outputs W_t. Define $\tilde{S} \triangleq (S_1, \ldots, S_m)$. If $P_{W|S}$ is (ε, μ)-stable in input-output mutual information, then the overall algorithm $P_{W^m|\tilde{S}}$ satisfies $I(\tilde{S}; W^m) \leq m\varepsilon$.*

Proof. The proof is based on the independence among (S_t, W_t), $t = 1, \ldots, m$, and the chain rule of mutual information. □

Lemma 4.4. *Let $\tilde{S} \triangleq (S_1, \ldots, S_m)$, where $S_t \sim \mu^{\otimes n}$. If an algorithm $P_{W,T,R|\tilde{S}} : \mathsf{Z}^{m \times n} \to \mathsf{W} \times [m] \times \{\pm 1\}$ satisfies $I(\tilde{S}; W, T, R) \leq \varepsilon$, and if $\ell(w, Z)$ is σ-subgaussian for all $w \in \mathsf{W}$, then*

$$\mathbb{E}\big[R(L_{S_T}(W) - L_\mu(W))\big] \leq \sqrt{\frac{2\sigma^2 \varepsilon}{n}}.$$

Proof. For any $\tilde{s} \in \mathsf{Z}^{m \times n}$, $w \in \mathsf{W}$, $t \in [m]$ and $r \in \{\pm 1\}$, let

$$u(\tilde{s}, w, t, r) \triangleq r(L_{s_t}(w) - L_\mu(w)).$$

If $\ell(w, Z)$ is σ-subgaussian under $Z \sim \mu$ for all $w \in \mathsf{W}$, then $\frac{r}{n} \sum_{i=1}^n \ell(w, Z_{t,i})$ is $\sigma/\sqrt{n}$-subgaussian under $S_t \sim \mu^{\otimes n}$ for all $w \in \mathsf{W}$, $t \in [m]$ and $r \in \{\pm 1\}$,

hence

$$\log \mathbb{E}\big[e^{\lambda u(\tilde{S},w,r,t)}\big] = \log \mathbb{E}\Big[\exp\Big\{\lambda\Big(\frac{r}{n}\sum_{i=1}^{n}\ell(w,Z_{t,i}) - \frac{r}{n}\sum_{i=1}^{n}\mathbb{E}[\ell(w,Z_{t,i})]\Big)\Big\}\Big] \\ \leq \frac{\lambda^2\sigma^2}{2n} \qquad \text{for all } \lambda \in \mathbb{R}.$$

From the Donsker-Varadhan variational representation of the relative entropy (4.8),

$$D(P_{\tilde{S}|W=w,T=t,R=r}\|P_{\tilde{S}}) \geq \mathbb{E}[\lambda u(\tilde{S},w,t,r)|W=w,T=t,R=r] - \log \mathbb{E}\big[e^{\lambda u(\tilde{S},w,r,t)}\big] \\ \geq \lambda\big(\mathbb{E}[rL_{S_t}(w)|W=w,T=t,R=r] - rL_\mu(w)\big) - \frac{\lambda^2\sigma^2}{2n}.$$

Averaging both sides with respect to $P_{W,T,R}$, we get

$$I(\tilde{S};W,T,R) \geq \lambda\mathbb{E}\big[R(L_{S_T}(W) - L_\mu(W))\big] - \frac{\lambda^2\sigma^2}{2n} \qquad \text{for all } \lambda \in \mathbb{R},$$

which implies that

$$\mathbb{E}\big[R(L_{S_T}(W) - L_\mu(W))\big] \leq \sqrt{\frac{2\sigma^2 I(\tilde{S};W,T,R)}{n}}$$

and proves the claim. □

Note that the upper bound in Lemma 4.4 does not depend on m. The following lemma pertains to the accuracy of the so-called exponential mechanism introduced by McSherry and Talwar [93] in the context of differential privacy.

Lemma 4.5 (Bassily et al. [80, Lemma 7.1])**.** *Let F be a finite set, f be a function $F \to \mathbb{R}$, and $\eta > 0$. If a random variable X on F has the distribution*

$$P_X(x) \propto e^{\eta f(x)}, \qquad x \in F \tag{4.16}$$

then

$$\mathbb{E}f(X) \geq \max_{x\in F} f(x) - \frac{1}{\eta}\log|F|. \tag{4.17}$$

With these lemmas, we can prove Theorem 4.4.

Proof of Theorem 4.4. The proof is an adaptation of a "monitor technique" proposed by Bassily et al. [80]. First, let $P_{W^m|\tilde{S}}$ be the parallel execution of m independent copies of $P_{W|S}$: for $t = 1, \dots, m$, an independent copy of $P_{W|S}$ takes an independent $S_t \sim \mu^{\otimes n}$ as input and outputs W_t. Given the outputs w^m, define the set

$$F = \{t = 1, \dots, m : (w_t, t, 1), (w_t, t, -1)\} \tag{4.18}$$

with cardinality $2m$. Then, let the output of the "monitor" be a sample (W^*, T^*, R^*) drawn from F according to the distribution

$$P_{W^*,T^*,R^*|\tilde{S}=\tilde{s},W^m=w^m}(w^*, t^*, r^*) \propto \exp\left(\frac{\gamma n r^*}{2}\big(L_\mu(w^*) - L_{s_{t^*}}(w^*)\big)\right) \tag{4.19}$$

for $(w^*, t^*, r^*) \in F$, with some $\gamma > 0$. Note that given $\tilde{s}$ and w^m, the output (W^*, T^*, R^*) is essentially obtained from an exponential mechanism [93] applied to $\tilde{s}$ with respect to the function

$$u\big((w^*, t^*, r^*), \tilde{s}\big) = r^*\big(L_\mu(w^*) - L_{s_{t^*}}(w^*)\big), \qquad (w^*, t^*, r^*) \in F. \tag{4.20}$$

It can be shown that the above exponential mechanism with $\ell \in [0, 1]$ has the following property: for two datasets $\tilde{s}, \tilde{s}' \in \mathsf{Z}^{m\times n}$ such that $d_H(\tilde{s}, \tilde{s}') \le 1$,

$$e^{-\gamma} \le \frac{P_{W^*,T^*,R^*|\tilde{S}=\tilde{s},W^m=w^m}(w^*, t^*, r^*)}{P_{W^*,T^*,R^*|\tilde{S}=\tilde{s}',W^m=w^m}(w^*, t^*, r^*)} \le e^{\gamma}, \quad \forall (w^*, t^*, r^*) \in F \tag{4.21}$$

which means that the algorithm $P_{W^*,T^*,R^*|\tilde{S},W^m=w^m}$ is $(\gamma, 0)$-differentially private for all w^m. By the group privacy property of $(\gamma, 0)$-differentially private algorithms [81, Theorem 2.2],

$$\begin{aligned} I(\tilde{S}; W^*, T^*, R^*|W^m) &\le \sup_{w^m} \sup_{\tilde{s},\tilde{s}'} D(P_{W^*,T^*,R^*|\tilde{S}=\tilde{s},W^m=w^m} \| P_{W^*,T^*,R^*|\tilde{S}=\tilde{s}',W^m=w^m}) \\ &\le n\gamma. \end{aligned} \tag{4.22}$$

In addition, since $P_{W|S}$ satisfies $I(S; W) \le \varepsilon$, Lemma 4.3 implies that

$$I(\tilde{S}; W^m) \le m\varepsilon. \tag{4.23}$$

Therefore, by the chain rule of mutual information and the data processing

inequality, we have

$$\begin{aligned} I(\tilde{S}; W^*, T^*, R^*) &\le I(\tilde{S}; W^m, W^*, T^*, R^*) \\ &\le m\varepsilon + n\gamma. \end{aligned} \tag{4.24}$$

By Lemma 4.4 and the assumption that $\ell(\cdot,\cdot) \in [0,1]$ (hence $\sigma^2 = 1/4$ in Lemma 4.4),

$$\mathbb{E}\big[R^*\big(L_{S_{T^*}}(W^*) - L_\mu(W^*)\big)\big] \le \sqrt{\frac{m\varepsilon + n\gamma}{2n}}. \tag{4.25}$$

On the other hand, in view of (4.54), we can apply Lemma 4.5 with the set F, the function $f(w^*, t^*, r^*) = u\big((w^*, t^*, r^*), \tilde{s}\big)$, and $\eta = \gamma n/2$ to get

$$\begin{aligned} &\mathbb{E}\big[R^*\big(L_{S_{T^*}}(W^*) - L_\mu(W^*)\big)\big|\tilde{S} = \tilde{s}, W^m = w^m\big] \\ &\ge \max_{(w^*,t^*,r^*)\in F} u\big((w^*, t^*, r^*), \tilde{s}\big) - \frac{2}{\gamma n}\log|F| \\ &= \max_{t\in[m]} \big|L_{s_t}(w_t) - L_\mu(w_t)\big| - \frac{2}{\gamma n}\log(2m), \end{aligned}$$

which implies

$$\mathbb{E}\big[R^*\big(L_{S_{T^*}}(W^*) - L_\mu(W^*)\big)\big] \ge \mathbb{E}\Big[\max_{t\in[m]} \big|L_{S_t}(W_t) - L_\mu(W_t)\big|\Big] - \frac{2}{\gamma n}\log(2m). \tag{4.26}$$

Combining (4.60) and (4.61) gives

$$\mathbb{E}\Big[\max_{t\in[m]} \big|L_{S_t}(W_t) - L_\mu(W_t)\big|\Big] \le \frac{2}{\gamma n}\log(2m) + \sqrt{\frac{m\varepsilon + n\gamma}{2n}}. \tag{4.27}$$

The rest of the proof is by contradiction. Choose $m = \lfloor 1/\beta \rfloor$, and $\gamma = \varepsilon/\beta n$. Suppose the algorithm $P_{W|S}$ does not satisfy the claimed generalization property, namely,

$$\mathbb{P}\big[\big|L_S(W) - L_\mu(W)\big| > \alpha\big] > \beta. \tag{4.28}$$

Then by the independence among the pairs (S_t, W_t), $t = 1, \ldots, m$,

$$\mathbb{P}\Big[\max_{t\in[m]} \big|L_{S_t}(W_t) - L_\mu(W_t)\big| > \alpha\Big] > 1 - \Big(1-\beta\Big)^{\lfloor 1/\beta \rfloor} > \frac{1}{2}.$$

Thus

$$\mathbb{E}\Big[\max_{t\in[m]} \big|L_{S_t}(W_t) - L_\mu(W_t)\big|\Big] > \frac{\alpha}{2}. \tag{4.29}$$

Combining (4.62) and (4.63) gives

$$\frac{\alpha}{2} < \frac{2\beta}{\varepsilon}\log\frac{2}{\beta} + \sqrt{\frac{\varepsilon}{\beta n}}. \tag{4.30}$$

Since it is assumed that $\varepsilon \geq \frac{8\beta}{\alpha}\log\frac{2}{\beta}$, the above inequality implies that

$$n < \frac{\varepsilon}{\beta(\frac{\alpha}{2} - \frac{2\beta}{\varepsilon}\log\frac{2}{\beta})^2} \leq \frac{16\varepsilon}{\alpha^2\beta},$$

which contradicts the assumption that $n \geq \frac{16\varepsilon}{\alpha^2\beta}$, and hence completes the proof. □

Using the same monitor technique, we can also obtain a high-probability bound on $|L_\mu(W) - L_S(W)|$ for algorithms stable in erasure mutual information, stated in the following theorem. Since stability in erasure mutual information is much weaker than stability in input-output mutual information, the resulting sample complexity is polynomial in both $1/\alpha$ and $1/\beta$.

Theorem 4.5. *Suppose $P_{W|S}$ is (ε, μ)-stable in erasure mutual information, and $\ell : \mathsf{W} \times \mathsf{Z} \to [0,1]$. If $\varepsilon < \alpha^2\beta^2/4$ and*

$$n \geq \frac{2}{\varepsilon\big(\frac{\alpha}{2} - \frac{1}{\beta}\sqrt{\varepsilon}\big)}\log\frac{2}{\beta} \tag{4.31}$$

for some $0 < \alpha, \beta < 1$, then

$$\mathbb{P}\big[\big|L_\mu(W) - L_S(W)\big| > \alpha\big] \leq \beta.$$

Consequently, if $P_{W|S}$ is $(\alpha^2\beta^2/16, \mu)$-stable in erasure mutual information,

then

$$n \geq \frac{128}{\alpha^3\beta^2} \log \frac{2}{\beta} \tag{4.32}$$

guarantees $\mathbb{P}\big[\big|L_\mu(W) - L_S(W)\big| > \alpha\big] \leq \beta$.

Proof. Section 4.8.2. □

Note that, if we simply use the Markov inequality instead of the monitor technique, then it can be shown that, if $P_{W|S}$ is $(\alpha^2\beta^2/4, \mu)$-stable in erasure mutual information, then

$$n \geq \frac{16 \log 2}{\alpha^3\beta^3} \tag{4.33}$$

guarantees $\mathbb{P}[|L_\mu(W) - L_S(W)| > \alpha] \leq \beta$. This is a worse bound on the sample complexity because it depends on β through $\frac{1}{\beta^3}$, whereas the sample complexity bound in (4.32) depends on β through $\frac{1}{\beta^2}\log\frac{1}{\beta}$.

4.4.4 Upper Bound on $\mathbb{E}|L_\mu(W) - L_S(W)|$

The proof of Theorem 4.4 also yields an upper bound for the expected absolute generalization error.

Theorem 4.6. *Suppose the loss function ℓ takes values in $[0,1]$, and the dataset S has the distribution $\mu^{\otimes n}$. If a learning algorithm $P_{W|S}$ is (ε, μ)-stable in input-output mutual information, then*

$$\mathbb{E}\big|L_S(W) - L_\mu(W)\big| \leq \inf_{\gamma > 0} \left(\frac{\log 2}{\gamma n} + \sqrt{\frac{\varepsilon}{2n} + \gamma} \right).$$

Proof. Choose $m = 1$ in the proof of Theorem 4.4. Combining (4.60) and (4.61) gives

$$\mathbb{E}\big|L_S(W) - L_\mu(W)\big| \leq \frac{2\log 2}{\gamma n} + \sqrt{\frac{\varepsilon + n\gamma}{2n}}. \tag{4.34}$$

The claim holds because $\gamma > 0$ is arbitrary. □

For the more general case where $\ell(w, Z)$ is σ-subgausian instead of bounded in $[0,1]$, Russo and Zou [84, Proposition 4] derived an upper bound for

$\mathbb{E}\big|L_S(W) - L_\mu(W)\big|$ stated as

$$\mathbb{E}\big|L_S(W) - L_\mu(W)\big| \leq \frac{\sigma}{\sqrt{n}} + c\sigma\sqrt{\frac{2I(\Lambda_{\mathsf{W}}(S);W)}{n}},$$

where $c < 36$ is a constant. This result relies on a proof that, if a random variable U is subgaussian, then $|U|$ is also subgaussian. We will compare these two bounds later when we discuss adaptive data analysis.

4.5 Learning Algorithms with Input-output Mutual Information Stability

In this section, we analyze several learning algorithms from the viewpoint of input-output mutual information stability. The first two algorithms we consider, namely the Gibbs algorithm and noisy empirical risk minimization, are in the paradigm of approximate ERM algorithms, where the output hypothesis approximates the ERM hypothesis in a certain sense. In this paradigm, the output W of the learning algorithm depends on the dataset S only through the collection of empirical risks $\Lambda_{\mathsf{W}}(S)$; thus we have the Markov chain

$$S - \Lambda_{\mathsf{W}}(S) - W$$

and

$$I(S;W) = I(\Lambda_{\mathsf{W}}(S);W),$$

as the Markov chain $\Lambda_{\mathsf{W}}(S) - S - W$ always holds. We also briefly discuss the method of inducing input-output mutual information stability by preprocessing of the dataset, where strong data processing inequalities can play a role in the analysis. Finally, we analyze the input-output mutual information stability of learning algorithms obtained from adaptive composition of constituent algorithms.

4.5.1 Gibbs Algorithm

As discussed in Sec. 4.2.2, the expected true risk of the output hypothesis W can be decomposed as

$$\mathbb{E}[L_\mu(W)] = \mathbb{E}[L_S(W)] + \mathrm{gen}(\mu, P_{W|S}).$$

This decomposition suggests that, to obtain a learning algorithm with small true risk, the output should, on one hand, have small empirical risk (fit the dataset), and, on the other hand, have small generalization error (not overfit). Since Theorem 4.2 shows that the generalization error can be upper bounded in terms of the mutual information $I(S;W)$, it is natural to consider an algorithm that minimizes the empirical risk regularized by $I(S;W)$:

$$P^\star_{W|S} = \operatorname*{arg\,inf}_{P_{W|S}} \left(\mathbb{E}[L_S(W)] + \frac{1}{\beta} I(S;W) \right), \tag{4.35}$$

where $\beta > 0$ is a parameter that balances fitting and generalization. To deal with the issue that μ is unknown to the learning algorithm, we can replace the mutual information term with an upper bound $D(P_{W|S}\|Q|P_S)$ on it that does not depend on μ, where Q is an arbitrary distribution on W. From

$$\begin{aligned} P^*_{W|S} &= \operatorname*{arg\,inf}_{P_{W|S}} \left(\mathbb{E}[L_S(W)] + \frac{1}{\beta} D(P_{W|S}\|Q|P_S) \right) \qquad (4.36) \\ &= \operatorname*{arg\,inf}_{P_{W|S}} \int_{\mathsf{Z}^n} \mu^{\otimes n}(\mathrm{d}s) \left(\mathbb{E}[L_s(W)|S=s] + \frac{1}{\beta} D(P_{W|S=s}\|Q) \right) \\ &= \int_{\mathsf{Z}^n} \mu^{\otimes n}(\mathrm{d}s) \operatorname*{arg\,inf}_{P_{W|S=s}} \left(\mathbb{E}[L_s(W)|S=s] + \frac{1}{\beta} D(P_{W|S=s}\|Q) \right), \end{aligned}$$

it follows that for each $s \in \mathsf{Z}^n$, the algorithm $P^*_{W|S}$ that minimizes (4.36) satisfies

$$P^*_{W|S=s} = \operatorname*{arg\,inf}_{P_{W|S=s}} \left(\mathbb{E}[L_s(W)|S=s] + \frac{1}{\beta} D(P_{W|S=s}\|Q) \right). \tag{4.37}$$

In the minimization in (4.37), the term $\frac{1}{\beta} D(P_{W|S=s}\|Q)$ can be viewed as a stabilizer for the ERM algorithm, which is added to improve the generalization capability of the algorithm. The closer Q is to P_W in relative entropy,

the closer $D(P_{W|S}\|Q|P_S)$ is to $I(S;W)$, as

$$I(S;W) = D(P_{W|S}\|Q|P_S) - D(P_W\|Q)$$

and the better $P^*_{W|S}$ approximates $P^\star_{W|S}$ that minimizes (4.35). It turns out that the algorithm $P^*_{W|S}$ that satisfies (4.37) for each $s \in \mathsf{Z}^n$ is the Gibbs algorithm [85], a randomized algorithm that outputs a hypothesis with smaller empirical risk with exponentially larger probability, which satisfies

$$P^*_{W|S=s}(\mathrm{d}w) = \frac{e^{-\beta L_s(w)}Q(\mathrm{d}w)}{\mathbb{E}_Q[e^{-\beta L_s(W)}]}$$

for each $s \in \mathsf{Z}^n$. The parameter $\beta > 0$ controls how well the Gibbs algorithm approximates the ERM algorithm. The Gibbs algorithm can thus be interpreted as a way to stabilize the ERM algorithm by controlling the input-output mutual information $I(S;W)$.

If the loss function ℓ takes value, in $[0,1]$, then

$$e^{-2\beta/n} \le \frac{\mathrm{d}P^*_{W|S=s}}{\mathrm{d}P^*_{W|S=s'}} \le e^{2\beta/n}$$

for all $s, s' \in \mathsf{Z}^n$ such that $d_{\mathrm{H}}(s,s') \le 1$. This implies that the Gibbs algorithm with $\ell(\cdot,\cdot) \in [0,1]$ is $(2\beta/n, 0)$-differentially private, $2\beta/n$-KL-stable, and $(2\beta/n, \mu)$-stable in erasure mutual information for any μ. We can also upper-bound the mutual information $I(S;W)$ for the Gibbs algorithm. From the group privacy property of $(2\beta/n, 0)$-differentially private mechanisms [81, Theorem 2.2], we know that, if the loss function $\ell(\cdot,\cdot) \in [0,1]$, then

$$e^{-2\beta} \le \frac{\mathrm{d}P^*_{W|S=s}}{\mathrm{d}P^*_{W|S=s'}} \le e^{2\beta} \qquad \forall s, s' \in \mathsf{Z}^n,$$

which implies that for any μ

$$I(S;W) \le \sup_{s,s' \in \mathsf{Z}^n} D(P^*_{W|S=s}\|P^*_{W|S=s'}) \le 2\beta.$$

By Theorem 4.2, the generalization error of the Gibbs algorithm with ℓ taking

values in $[0,1]$ satisfies

$$\left|\operatorname{gen}(\mu, P^*_{W|S})\right| \le \sqrt{\frac{\beta}{n}}.$$

This estimate is the same as the one given by the guarantee that the algorithm is $2\beta/n$-stable in erasure mutual information [83]. In addition, from Hoeffding's lemma and the fact that the Gibbs algorithm is $(1-e^{-2\beta/n})$-TV stable, it is shown in [83] that

$$\left|\operatorname{gen}(\mu, P^*_{W|S})\right| \le (1-e^{-2\beta/n}) \wedge \frac{\beta}{4n} \wedge \sqrt{\frac{\beta}{n}}. \tag{4.38}$$

Moreover, it is shown by Wang et al. [88, Theorem 4] that, for the Gibbs algorithm,

$$\operatorname{gen}(\mu, P^*_{W|S}) = \frac{1}{\beta}\sum_{i=1}^{n}\int \mu^{\otimes n}(\mathrm{d}s)\mu(\mathrm{d}z_i')D(P_{W|S=s}\|P_{W|S=s_{(i)}}). \tag{4.39}$$

Recalling the notion of on-average KL-stability defined in (4.5), the above identity implies that the Gibbs algorithm is (ε,μ)-KL-stable on average if and only if

$$\operatorname{gen}(\mu, P^*_{W|S}) \le \frac{n\varepsilon}{\beta}.$$

In Sec. 4.8.3, we give a more readable proof of (4.39).

We can also analyze the excess risk of the Gibbs algorithm when W is a finite set and ℓ takes values in $[0,1]$. Suppose W has cardinality k. Using a proof similar to that of Lemma 4.5, we can show that, for any dataset s, the empirical risk of the Gibbs algorithm (with Q chosen as the uniform distribution on W) satisfies

$$\mathbb{E}[L_s(W)|S=s] \le \min_{w\in\mathsf{W}} L_s(w) + \frac{1}{\beta}\log k. \tag{4.40}$$

Choosing $\beta = 2\sqrt{n\log k}$, we have

$$\begin{aligned}\mathbb{E}[L_S(W)] &\le \mathbb{E}[L_S(W_{\mathrm{ERM}})] + \frac{1}{2}\sqrt{\frac{\log k}{n}}\\ &\le \min_{w\in\mathsf{W}} L_\mu(w) + \frac{1}{2}\sqrt{\frac{\log k}{n}}\end{aligned}$$

where the last step is due to (4.4). Combining with the upper bound (4.38) on generalization error, the true risk can be upper bounded by

$$\mathbb{E}[L_\mu(W)] \leq \min_{w \in \mathsf{W}} L_\mu(w) + \frac{1}{2}\sqrt{\frac{\log k}{n}} + \frac{1}{2}\sqrt{\frac{\log k}{n}}.$$

Therefore, the expected excess risk of the Gibbs algorithm in this case satisfies

$$R_{\text{excess}}(\mu, P^*_{W|S}) \leq \sqrt{\frac{\log k}{n}}.$$

4.5.2 Noisy Empirical Risk Minimization

The second algorithm we consider is the noisy empirical risk minimization algorithm. The algorithm adds independent noise to the empirical risk of each hypothesis, and then outputs the hypothesis that minimizes the noisy empirical risks. Suppose the hypothesis space $\mathsf{W} = \{w_1, \ldots, w_k\}$ is a finite set with cardinality k, and the output of the algorithm is $W = w_{J^*}$, with

$$J^* = \arg\min_{j \in [k]} \big(L_S(w_j) + N_j\big).$$

We first consider the case where N_j's are are i.i.d. Gaussian with zero-mean and variance σ_N^2. In this case, if $\ell(w, Z)$ is σ-subgaussian, then

$$\begin{aligned}
I(\Lambda_{\mathsf{W}}(S); W) &\leq I\big((L_S(w_i))_{i\in[k]}; (L_S(w_i) + N_i)_{i\in[k]}\big) \\
&\leq \sum_{j=1}^{k} I(L_S(w_j); L_S(w_j) + N_j) \\
&\leq \max_{j\in[k]} \frac{k}{2} \log\left(1 + \frac{\mathrm{Var}[L_S(w_j)]}{\sigma_N^2}\right) \\
&\leq \frac{k}{2}\log\left(1 + \frac{\sigma^2}{n\sigma_N^2}\right) \\
&\leq \frac{k\sigma^2}{2n\sigma_N^2},
\end{aligned}$$

where we have used the data processing inequality for mutual information; the fact that for product channels, the mutual information between the overall input and output is upper bounded by the sum of the input-output mutual information of individual channels [34]; the formula for the capacity of

input-power constrained Gaussian channel; the fact that $L_S(w_j)$'s are $\sigma/\sqrt{n}$ subgaussian hence $\mathrm{Var}[L_S(w_j)] \le \sigma^2/n$; and the fact that $\log(1+x) \le x$. Also note that in this case

$$I(\Lambda_{\mathsf{W}}(S); W) \le H(W) \le \log k.$$

Choosing a small noise variance σ_N^2, we obtain an approximate ERM algorithm, with

$$\left|\mathrm{gen}(\mu, P_{W|S})\right| \le \frac{\sigma^2\sqrt{k}}{n\sigma_N} \wedge \sqrt{\frac{2\sigma^2 \log k}{n}}$$

by Theorem 4.2. For example, by choosing $\sigma_N^2 = 1/n$, we have

$$\left|\mathrm{gen}(\mu, P_{W|S})\right| \le \sigma^2\sqrt{\frac{k}{n}} \wedge \sqrt{\frac{2\sigma^2 \log k}{n}}.$$

Now we consider the case where N_j's are are i.i.d. exponential random variables with mean b. In this case, if $\ell(\cdot,\cdot) \in [0,1]$, then

$$\begin{aligned} I(\Lambda_{\mathsf{W}}(S); W) &\le \sum_{j=1}^{k} I(L_S(w_j); L_S(w_j) + N_j) \\ &\le \max_{j\in[k]} k \log\left(1 + \frac{\mathbb{E}[L_S(w_j)]}{b}\right) \\ &\le k \log\left(1 + \frac{1}{b}\right), \end{aligned}$$

where we have used the fact that, for any nonnegative random variable X with mean a and an exponential random variable N independent of X with mean b [94],

$$I(X; X+N) \le \log\left(1 + \frac{a}{b}\right).$$

Choosing a small noise mean b, we obtain an approximate ERM algorithm, with

$$\left|\mathrm{gen}(\mu, P_{W|S})\right| \le \sqrt{\frac{k}{2n}\log\left(1+\frac{1}{b}\right)} \wedge \sqrt{\frac{\log k}{2n}}$$

by Theorem 4.2.

4.5.3 Preprocessing of the Dataset

Another method of inducing input-output mutual information stability is to preprocess the dataset S to obtain $\tilde{S}$, and then run a learning algorithm on the preprocessed dataset $\tilde{S}$. The preprocessing can consists of adding noise to the data or erasing some of the instances in the dataset, etc. In any case, we have the Markov chain

$$S - \tilde{S} - W.$$

The strong data processing inequality introduced in Sec. 2.3 implies that

$$I(S;W) \leq \min\left\{I(S;\tilde{S}),\, I(\tilde{S};W)\eta(P_{\tilde{S}}, P_{S|\tilde{S}})\right\}.$$

If $\tilde{Z}_i$ is generated from Z_i independently of everything else, then by the tensorization property of the SDPI constant (Lemma 2.8),

$$\eta(P_{\tilde{S}}, P_{S|\tilde{S}}) = \max_{i\in[n]} \eta(P_{\tilde{Z}_i}, P_{Z_i|\tilde{Z}_i}) \leq \max_{i\in[n]} \eta(P_{Z_i|\tilde{Z}_i}).$$

As an example, if Z_i and $\tilde{Z}_i$ are jointly Gaussian with correlation coefficient ρ, then

$$\eta(P_{\tilde{Z}_i}, P_{Z_i|\tilde{Z}_i}) = \rho^2.$$

It would be interesting to find preprocessing methods such that we can evaluate $\eta(P_{\tilde{Z}_i}, P_{Z_i|\tilde{Z}_i})$ or $\eta(P_{Z_i|\tilde{Z}_i})$, so that we can sharply bound the input-output mutual information $I(S;W)$.

4.5.4 Adaptive Composition

Consider the situation where k learning algorithms are sequentially executed. The output of the jth algorithm may depend on the dataset S, as well as on the outputs W^{j-1} of the executed learning algorithms. The output at the final step can be viewed as obtained from an adaptive composition of the k constituent learning algorithms. From the chain rule of mutual information,

$$I(S;W_k) \leq I(S;W^k) = \sum_{j=1}^{k} I(S;W_j|W^{j-1}).$$

It suggests that we can control the generalization error of the final output by controlling the conditional mutual information $I(S;W_j|W^{j-1})$ at each step of the composition.

When the constituent algorithms are run on different processors sharing the same dataset, the output hypothesis of each processor needs to be communicated to other processors to serve as an input. In this case, communication constraints may occur due to the quantization of the output hypothesis, the finite blocklength for transmission, and the noisy channels connecting the processors. These communication constraints will limit the effective hypothesis space of each constituent algorithm and the accuracy for reconstructing the hypotheses at each processor. At the same time, since the mutual information terms $I(S;W_j|W^{j-1})$ are limited by the communication constraints as well, the generalization capability of the composed algorithm may be improved. For example, if each output of the first $k-1$ algorithms is constrained to be represented by b bits, then each of the first $k-1$ hypothesis spaces is effectively confined to a set of cardinality 2^b, and the final input-output mutual information can be bounded by

$$I(S;W_k) \leq (k-1)b\log 2 + I(S;W_k|W^{k-1}).$$

The methods developed in Chapter 2 and 3 can also be used to upper bound the mutual information $I(S;W_j|W^{j-1})$, $j=1,\ldots,k$, according to the communication constraints; then Theorem 4.2 can be invoked to upper-bound the generalization error.

It is the chain rule of mutual information that makes the adaptive composition easy to analyze under the notion of input-output mutual information stability. We can also apply the chain rule of mutual information or relative entropy to analyze the adaptive composition for other information-theoretic stability notions, e.g., stability in erasure mutual information and KL-stability. Specifically, if the algorithms $P_{W_j|S,W^{j-1}}$, $j=1,\ldots,k$, satisfy

$$\frac{1}{n}\sum_{i=1}^{n} I(Z_i;W_j|S^{-i},W^{j-1}) \leq \varepsilon_j$$

for some μ, then

$$\frac{1}{n}\sum_{i=1}^{n} I(Z_i; W_1, \ldots, W_k | S^{-i}) \le \sum_{j=1}^{k} \varepsilon_j$$

for the same μ; moreover, if the algorithms $P_{W_j|S,W^{j-1}}$, $j = 1, \ldots, k$, satisfy

$$\sup_{s,s' \in \mathsf{Z}^n : d_{\mathrm{H}}(s,s') \le 1} \sup_{w^{j-1}} D(P_{W_j|S=s,W^{j-1}=w^{j-1}} \| P_{W_j|S=s',W^{j-1}=w^{j-1}}) \le \varepsilon_j,$$

then

$$\sup_{s,s' \in \mathsf{Z}^n : d_{\mathrm{H}}(s,s') \le 1} D(P_{W^k|S=s} \| P_{W^k|S=s'}) \le \sum_{j=1}^{k} \varepsilon_j.$$

For other notions of information-theoretic stability, the adaptive composition is not as easy to analyze. For example, for the (ε, δ)-differential privacy, one may need to use results in binary hypothesis testing to characterize the privacy degradation under adaptive composition [82].

4.6 Application to Adaptive Data Analytics

4.6.1 Non-adaptive and Adaptive Data Analytics

In non-adaptive data analytics, there is an unknown distribution μ on Z, and a random dataset $S \in \mathsf{Z}^n$ drawn from $\mu^{\otimes n}$. Given a query space W and a function $\ell : \mathsf{W} \times \mathsf{Z} \to \mathbb{R}^+$, the data analyst picks some query $w \in \mathsf{W}$ and wishes to evaluate the expectation of $\ell(w, Z)$ under $Z \sim \mu$, denoted as

$$L_\mu(w) = \mathbb{E}[\ell(w, Z)].$$

Although the distribution μ is unknown, there is an answer-generating mechanism holding the dataset S, which accepts the query w and returns the empirical mean

$$L_S(w) = \frac{1}{n}\sum_{i=1}^{n} \ell(w, Z_i)$$

to the data analyst. The query picked by the data analyst could also be a random element W in W, which is independent of the dataset S. By the

law of large numbers (assuming the function ℓ is bounded), we know that $L_S(W) - L_\mu(W)$ converges to zero both in L^1 and in probability uniformly for all μ. Therefore, in non-adaptive data analytics, due to the independence between the query W and the dataset S, there is a strong guarantee that the answer given as $L_S(W)$ can well approximate the true expectation $L_\mu(W)$ for any μ. As an example, consider the problem of performance evaluation in the general statistical learning framework: W is a hypothesis generated by some learning algorithm based on a training dataset S'; the learner (playing the role of the data analyst) would like to evaluate the true risk of W, $L_\mu(W)$, under μ; the tester (playing the role of the answer-generating mechanism) holds a testing dataset S independent of S' and provides the empirical risk of W on S, $L_S(W)$, to the learner as a proxy of $L_\mu(W)$. Due to the independence between W and S, the answer $L_S(W)$ is an accurate estimate of the true risk $L_\mu(W)$ of the hypothesis W.

In practice, data analytics is often performed in multiple rounds in an adaptive manner: in the jth round, the data analyst issues a query W_j based on the previously issued queries W^{j-1} as well as the answers Y^{j-1} received so far; a new answer Y_j is then generated based on the dataset S and the query W_j. In this case, the queries W_j for $j \geq 2$ are no longer independent of the dataset S; hence, the empirical mean $L_S(W_j)$ can severely deviate from the true expectation $L_\mu(W_j)$. The difference $L_\mu(W_j) - L_S(W_j)$ is called the *bias* of W_j on S. An important problem in adaptive data analytics is to design answer-generating mechanisms such that the answers Y_j are close to the true expectations $L_\mu(W_j)$ under multiple rounds of adaptive analysis. Continuing the preceding example of performance evaluation in statistical learning: once the learner receives $Y_1 = L_S(W_1)$ from the tester, he can modify the learning algorithm based on it, and come up with a new hypothesis W_2 to see whether it can reduce the empirical risk on the testing dataset. If the tester naively returns the empirical risks $Y_j = L_S(W_j)$ all the time, the learner may gradually find a hypothesis that overfits the testing dataset such that empirical risk on S is small, but the true risk is large. The answer-generating mechanism thus has to be carefully designed to prevent the learner from overfitting the testing dataset.

Recently, ideas in differential privacy have been brought to designing the answer-generating mechanisms in adaptive data analytics [79,80,95]. In these works, the major concern is the bias analysis, which is based on deriving

generalization guarantees of differentially private algorithms. Once the bias is upper bounded, the accuracy of the answers can be analyzed by combining the upper bound on the bias and the accuracy guarantees of various privacy-inducing mechanisms. In this section, we use information-theoretic methods developed earlier in this chapter to analyze the bias and accuracy in adaptive data analytics. Our analyses are simpler than those based on differential privacy, and provide information-theoretic insights about how to design good answer-generating mechanisms that reduce bias and improve accuracy.

4.6.2 Analyzing Bias and Accuracy Using $I(S;W)$

We consider the k-round adaptive analysis, where both the queries and answers can be drawn from randomized mechanisms. At the jth round, the data analyst issues a query W_j drawn according to the kernel $P_{W_j|W^{j-1},Y^{j-1}}$, and receives an answer Y_j to the query drawn according to the kernel $P_{Y_j|S,W_j}$. The Bayesian network of the query-answer pairs is shown in Fig. 4.2 for $k = 4$.

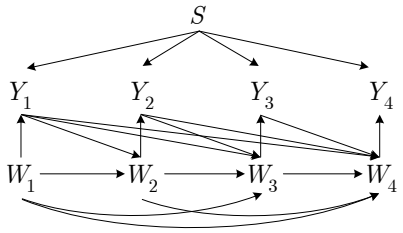


Figure 4.2: Bayesian network of adaptive data analysis, $k = 4$.

Note that the bias $L_\mu(W_j) - L_S(W_j)$ is equivalent to the generalization error discussed earlier in this chapter. Consequently, we can upper-bound the bias of the jth query W_j in terms of the mutual information $I(S;W_j)$ using the results obtained in Sec. 4.4. From the chain rule of mutual information, we

have the following chain of inequalities: for $j = 1, \ldots, k$,

$$\begin{aligned} I(S; W_j) &\le I(S; W^{j-1}, Y^{j-1}) \\ &= \sum_{i=1}^{j-1} I(S; W_i, Y_i | W^{i-1}, Y^{i-1}) \\ &= \sum_{i=1}^{j-1} I(S; W_i | W^{i-1}, Y^{i-1}) + I(S; Y_i | W_i, W^{i-1}, Y^{i-1}) \\ &\le \sum_{i=1}^{j-1} I(S; Y_i | W_i), \end{aligned} \tag{4.41}$$

where the last step uses the fact that $I(S; W_i | W^{i-1}, Y^{i-1}) = 0$ because of the Marjov chain $S \to W^{i-1}, Y^{i-1} \to W_i$, and the fact that

$$I(S; Y_i | W_i, W^{i-1}, Y^{i-1}) \le I(W^{i-1}, Y^{i-1}, S; Y_i | W_i) = I(S; Y_i | W_i)$$

because $W^{i-1}, Y^{i-1} \to S \to Y_i$ form a Markov chain conditioned on W_i.

Russo and Zou considered the same problem in [84]. However, their assumption is that (1) the query space W is a finite set, and (2) the answer Y_j is generated by adding noise on $L_S(W_j)$. Their result [84, Lemma 1] shows that

$$I(\Lambda_{\mathsf{W}}(S); W_j) \le \sum_{i=1}^{j-1} I(L_S(W_i); Y_i | W^{i-1}, Y^{i-1}, W_i),$$

where $\Lambda_{\mathsf{W}}(S) = (L_S(w))_{w \in \mathsf{W}}$. Our result is more general, as neither assumption is needed. When the answer Y_j is generated by adding noise on $L_S(W_j)$, we can obtain the same upper bound on $I(\Lambda_{\mathsf{W}}(S); W_j)$:

$$\begin{aligned} I(\Lambda_{\mathsf{W}}(S); W_j) &\le I(S; W_j) \\ &\le \sum_{i=1}^{j-1} I(S; Y_i | W_i, W^{i-1}, Y^{i-1}) \\ &\le \sum_{i=1}^{j-1} I(L_S(W_i); Y_i | W_i, W^{i-1}, Y^{i-1}) \end{aligned}$$

and we allow the query space W to be uncountably infinite.

Gaussian Noise-adding

Now we evaluate the upper bound in the special case where Y_j is generated by adding Gaussian noise to $L_S(W_j)$, i.e., for $j = 1, \ldots, k$,

$$Y_j = L_S(W_j) + N_j,$$

where N_j's are i.i.d. zero-mean Gaussian with variance σ_j^2. If $\ell(w, Z)$ is σ-subgaussian for all $w \in \mathsf{W}$, then from (4.41),

$$\begin{aligned} I(S; W_j) &\le \sum_{i=1}^{j-1} I(S; Y_i | W_i) \\ &\le \sum_{i=1}^{j-1} I(L_S(W_i); L_S(W_i) + N_i | W_i) \\ &\le \sum_{i=1}^{j-1} \frac{1}{2} \log \left(1 + \frac{\sup_{w \in \mathsf{W}} \mathrm{Var}[L_S(w)]}{\sigma_i^2} \right) \\ &\le \sum_{i=1}^{j-1} \frac{\sigma^2}{2n\sigma_i^2}. \end{aligned}$$

By Theorem 4.2, the generalization error (bias) of the jth query in this special case satisfies

$$\left| \mathbb{E}[L_\mu(W_j)] - \mathbb{E}[L_S(W_j)] \right| \le \frac{\sigma^2}{n} \sqrt{\sum_{i=1}^{j-1} \frac{1}{\sigma_i^2}}, \qquad j = 1, \ldots, k.$$

Moreover, we can study the accuracy of the jth answer measured by the absolute error using Theorem 4.6. Suppose $\ell(\cdot, \cdot) \in [0, 1]$, then for any $\gamma > 0$,

$$\begin{aligned} \mathbb{E}|Y_j - L_\mu(W_j)| &\le \mathbb{E}|Y_j - L_S(W_j)| + \mathbb{E}|L_S(W_j) - L_\mu(W_j)| \\ &\le \sigma_j + \frac{\log 2}{\gamma n} + \sqrt{\frac{I(S; W_j)}{2n} + \gamma} \\ &\le \sigma_j + \frac{\log 2}{\gamma n} + \sqrt{\frac{1}{16n^2} \sum_{i=1}^{j-1} \frac{1}{\sigma_i^2} + \gamma}. \end{aligned}$$

Choosing $\sigma_j^2 = \sqrt{j}/2n$, $\gamma = \sqrt{j}/4n$, and using the fact that $\sum_{i=1}^{j} 1/\sqrt{i} \leq 2\sqrt{j}$, we get

$$\mathbb{E}|Y_j - L_\mu(W_j)| \leq \frac{\sqrt{2}j^{1/4}}{\sqrt{n}} + \frac{4\log 2}{\sqrt{j}}, \qquad j = 1, \ldots, k. \tag{4.42}$$

Note that the choice of the noise variance $\sigma_j^2 = \sqrt{j}/2n$ does not depend on k, meaning that the answer-generating mechanism does not need to know the total number of queries that will be issued by the analyst in advance. From (4.42), we see that for a sufficiently large k such that $k = \Omega(n^{2/3})$, we have

$$\max_{j \in [k]} \mathbb{E}|Y_j - L_\mu(W_j)| \lesssim \frac{\sqrt{2}k^{1/4}}{\sqrt{n}}. \tag{4.43}$$

This upper bound is on the same order as the result obtained by Russo and Zou [84, Proposition 9] (under the assumption that the query space W is a finite set), which states that

$$\max_{j \in [k]} \mathbb{E}|Y_j - L_\mu(W_j)| \leq \frac{ck^{1/4}}{\sqrt{n}}$$

with some constant c. Under the assumption that the collection of empirical risks $(L_S(w))_{w \in \mathsf{W}}$ is a Gaussian process with variance σ^2/n, and under a richness assumption on the query space, Wang et al. [96] obtained the minimax rate of the mean squared error for the k-fold adaptive data analytics:

$$\inf_{\{P_{Y_j|S,W_j}\}_{j=1}^k} \sup_{\{P_{W_j|W^{j-1},Y^{j-1}}\}_{j=1}^k} \max_{j \in [k]} \mathbb{E}(Y_j - L_\mu(W_j))^2 = O\Big(\frac{\sqrt{k}\sigma^2}{n}\Big).$$

For the more general problem setup that we have considered above, whether the upper bound in (4.43) can be improved is an open problem. It would also be interesting to consider other answer-generating mechanisms beyond the noise-adding method, and analyze the corresponding bias and accuracy via $I(S; Y_j|W_j)$, $j = 1, \ldots, k$.

4.7 Conclusion and Future Research Directions

In this chapter, we mainly analyzed the generalization error of a learning algorithm via the mutual information between its input and output. We derived an upper bound on the expected generalization error and a high-probability bound on the absolute generalization error for algorithms that are stable in input-output mutual information. We also discussed how to design learning algorithms with input-output mutual information stability, and showed that the Gibbs algorithm can be viewed as an input-output mutual information regularized ERM algorithm. In addition, we discussed the input-output mutual information stability in adaptive composition, which is useful for analyzing the generalization performance when the constituent algorithms are run on multiple processors in a decentralized setting with communication constraints. The results have also been applied to analyzing the bias and accuracy in adaptive data analytics. There are a few problems worthwhile for future study.

- The notion of input-output mutual information stability is a somewhat strong condition for stability. It is not satisfied by classical learning algorithms such as the ERM algorithm and the stochastic gradient descent (SGD) algorithm. The notion of Wasserstein stability is less restrictive, and can be used to analyze the generalization error of ERM algorithm and algorithms with random incremental updates [83]. An interesting research topic is to use Wasserstein stability to analyze the generalization performance of the SGD algorithm.

- Our upper bound on the expected generalization error (Theorem 4.2) only requires that the loss function $\ell(w, z)$ is subgaussian in z for any $w \in \mathsf{W}$. However, the high-probability bound (Theorem 4.4) and the expectation bound for the absolute generalization error $|L_\mu(W) - L_S(W)|$ (Theorem 4.6) both require the loss function to be bounded. This is due to the need for upper-bounding the input-output mutual information of the exponential mechanism used in the proof of these results, which requires a bounded loss function. Are there better ways to prove these results which do not require boundedness of the loss function?

- We have derived information-theoretic upper bounds on the generaliza-

tion error, but did not have any discussion on how to lower-bound this quantity. For the Gibbs algorithm, Wang et al. [88] provide an exact information-theoretic characterization of the generalization error, as stated in (4.39). It would be interesting to study information-theoretic lower bounds on the generalization error for general algorithms.

- We have been focusing on the generalization error. However, as discussed in Sec. 4.2.2, having a small generalization error alone is not sufficient to have a small true risk. Can we provide information-theoretic conditions for a learning algorithm to be asymptotic ERM, so that we can characterize the consistency of a learning algorithm in an information-theoretic manner?

4.8 Additional Proofs for Chapter 4

4.8.1 Proof of Theorem 4.3

Given an index $i \in [n]$ and a sample $s = (z_1, \ldots, z_n) \in \mathsf{Z}^n$, let $\pi = P_{W|S=s}$, $\rho = P_{W|S^{-i}=s^{-i}}$, and $F(w) = -\lambda \ell(w, z_i)$, where $\lambda \in \mathbb{R}$ is an arbitrary parameter. Then from the Donsker-Varadhan variational representation of relative entropy (4.8), we have

$$D\big(P_{W|S=s}\big\|P_{W|S^{-i}=s^{-i}}\big) \geq -\lambda \mathbb{E}\left[\ell(W, z_i)\big|S = s\right] - \log \mathbb{E}\big[e^{-\lambda \ell(W, z_i)}\big|S^{-i} = s^{-i}\big]. \tag{4.44}$$

By the subgaussianity assumption (4.14), we can write

$$\log \mathbb{E}\big[e^{-\lambda \ell(W, z_i)}\big|S^{-i} = s^{-i}\big] \leq \lambda \mathbb{E}\big[\ell(W, z_i)\big|S^{-i} = s^{-i}\big] + \frac{\lambda^2 \sigma^2}{2}. \tag{4.45}$$

Using (4.45) in (4.44), we obtain

$$D\big(P_{W|S=s}\big\|P_{W|S^{-i}=s^{-i}}\big) \geq \lambda \left(\mathbb{E}\big[\ell(W, z_i)\big|S^{-i} = s^{-i}\big] - \mathbb{E}\big[\ell(W, z_i)\big|S = s\big]\right) - \frac{\lambda^2 \sigma^2}{2}. \tag{4.46}$$

Let $S' = (Z'_1, \ldots, Z'_n)$ be an n-tuple of i.i.d. draws from μ, independent of (S, W), and let $W_{(i)}$ denote the output of the learning algorithm A operating

on $S^{i,Z_i'} \triangleq (Z_1, \ldots, Z_{i-1}, Z_i', Z_{i+1}, \ldots, Z_n)$. Using the fact that S and S' are two independent samples of size n from μ, we can write

$$\begin{aligned}\int \mu^{\otimes n}(\mathrm{d}s)\mathbb{E}\big[\ell(W, z_i)\big|S^{-i} = s^{-i}\big] &= \int \mu^{\otimes n}(\mathrm{d}s) \int P_{W,Z_i'|S^{-i}=s^{-i}}(\mathrm{d}w, \mathrm{d}z_i')\ell(w, z_i) \\ &= \int \mu^{\otimes n}(\mathrm{d}s)\mu^{\otimes n}(\mathrm{d}s')P_{W|S=s_{(i)}}(\mathrm{d}w)\ell(w, z_i) \\ &= \int \mu^{\otimes n}(\mathrm{d}s)\mu^{\otimes n}(\mathrm{d}s')P_{W|S=s}(\mathrm{d}w)\ell(w, z_i') \\ &= \mathbb{E}[\ell(W, Z_i')]\end{aligned}$$

and

$$\begin{aligned}\int \mu^{\otimes n}(\mathrm{d}s)\mathbb{E}\big[\ell(W, z_i)\big|S = s\big] &= \int \mu^{\otimes n}(\mathrm{d}s)\mu^{\otimes n}(\mathrm{d}s')P_{W|S=s}(\mathrm{d}w)\ell(w, z_i) \\ &= \int \mu^{\otimes n}(\mathrm{d}s)\mu^{\otimes n}(\mathrm{d}s')P_{W|S=s_{(i)}}\ell(w, z_i') \\ &= \mathbb{E}[\ell(W_{(i)}, Z_i')].\end{aligned}$$

Therefore, taking expectations of both sides of (4.46) with respect to Z^n, we have

$$I(W; Z_i|S^{-i}) \geq -\frac{\lambda^2\sigma^2}{2} + \lambda\left(\mathbb{E}[\ell(W, Z_i')] - \mathbb{E}[\ell(W_{(i)}, Z_i')]\right). \tag{4.47}$$

Summing (4.47) over i gives

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n I(W; Z_i|S^{-i}) &\geq -\frac{\lambda^2\sigma^2}{2} + \frac{\lambda}{n}\sum_{i=1}^n \left(\mathbb{E}[\ell(W, Z_i')] - \mathbb{E}[\ell(W_{(i)}, Z_i')]\right) \\ &= -\frac{\lambda^2\sigma^2}{2} + \lambda\mathbb{E}\left[L(W) - L_S(W)\right].\end{aligned}$$

Maximizing the right-hand side with respect to $\lambda \in \mathbb{R}$, we obtain (4.15).

4.8.2 Proof of Theorem 4.5

Similar to the proof for Theorem 4.4, we need the following two lemmas to prove Theorem 4.5. As before, define $\tilde{S} = (S_1, \ldots, S_m)$, where $S_t = (Z_{t,1}, \ldots, Z_{t,n})$, $t = 1, \ldots, m$, are i.i.d. datasets drawn from $\mu^{\otimes n}$.

Lemma 4.6. *Consider the parallel execution of m independent copies of*

$P_{W|S}$: for $t = 1, \ldots, m$, an independent copy of $P_{W|S}$ takes $\tilde{S}_t$ as input, and outputs W_t. If $P_{W|S}$ is (ε, μ)-stable in erasure mutual information:

$$\frac{1}{n}\sum_{i=1}^{n} I(W; Z_i|S^{-i}) \le \varepsilon \tag{4.48}$$

for some μ, then the overall algorithm $P_{W_1,\ldots,W_m|\tilde{S}}$ is also (ε, μ)-stable in erasure mutual information:

$$\frac{1}{mn}\sum_{t=1}^{m}\sum_{i=1}^{n} I(W_1, \ldots, W_m; Z_{t,i}|\tilde{S}^{-t,i}) \le \varepsilon. \tag{4.49}$$

Proof. The proof is based on the independence among (S_t, W_t), $t = 1, \ldots, m$. □

Lemma 4.7. *Suppose an algorithm $P_{W,T,R|\tilde{S}} : \mathcal{Z}^{m\times n} \to \mathcal{W} \times [m] \times \{-1, 1\}$ is (ε, μ)-stable in erasure mutual information, i.e.,*

$$\frac{1}{mn}\sum_{t=1}^{m}\sum_{i=1}^{n} I(W, T, R; Z_{t,i}|\tilde{S}^{-t,i}) \le \varepsilon, \tag{4.50}$$

and for all $\tilde{s}$, $t \in [m]$ and $i \in [n]$, the random variable $\mathbf{1}\{T = t\}\ell(W, z_{t,i})R$ with (W, T, R) distributed according to $P_{W,T,R|\tilde{S}^{-t,i}=\tilde{s}^{-t,i}}$ is σ-subgaussian, i.e., for all $\lambda \in \mathbb{R}$

$$\log \mathbb{E}\Big[\exp\Big\{\lambda\Big(\mathbf{1}\{T = t\}\ell(W, z_{t,i})R - \mathbb{E}\big[\mathbf{1}\{T = t\}\ell(W, z_{t,i})R\big|\tilde{S}^{-t,i} = \tilde{s}^{-t,i}\big]\Big)\Big\}\Big|\tilde{S}^{-t,i} = \tilde{s}^{-t,i}\Big] \le \frac{\lambda^2\sigma^2}{2}. \tag{4.51}$$

Then

$$\mathbb{E}\big[R\big(L_\mu(W) - L_{S_T}(W)\big)\big] \le m\sqrt{2\sigma^2\varepsilon}.$$

Proof. Let $\tilde{S}'$ be an independent copy of $\tilde{S}$. First of all, we have

$$\mathbb{E}[L_\mu(W)R] - \mathbb{E}[L_{S_T}(W)R] = \frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{m}\Big(\mathbb{E}\big[\mathbf{1}\{T = t\}\ell(W, Z'_{t,i})R\big] - \mathbb{E}\big[\mathbf{1}\{T = t\}\ell(W, Z_{t,i})R\big]\Big). \tag{4.52}$$

Then, from the Donsker-Varadhan variational representation of the relative

entropy and the subgausisan assumption in (4.51), we have

$$D(P_{W,T,R|\tilde{S}=\tilde{s}}\|P_{W,T,R|\tilde{S}^{-t,i}=s^{-t,i}}) \geq \lambda\Big(\mathbb{E}\big[\mathbf{1}\{T=t\}\ell(W,z_{t,i})R\big|\tilde{S}^{-t,i}=\tilde{s}^{-t,i}\big]- \mathbb{E}\big[\mathbf{1}\{T=t\}\ell(W,z_{t,i})R\big|\tilde{S}=\tilde{s}\big]\Big)-\frac{\lambda^2\sigma^2}{2}.$$

Integrating with $\mu^{\otimes mn}(\mathrm{d}\tilde{s})$, and using the fact that

$$\mathbb{E}\big[\mathbf{1}\{T=t\}\ell(W,Z'_{t,i})R\big]=\int\mu^{\otimes mn}(\mathrm{d}\tilde{s})\mathbb{E}\big[\mathbf{1}\{T=t\}\ell(W,z_{t,i})R\big|\tilde{S}^{-t,i}=\tilde{s}^{-t,i}\big]$$

and

$$\mathbb{E}\big[\mathbf{1}\{T=t\}\ell(W,Z_{t,i})R\big]=\int\mu^{\otimes mn}(\mathrm{d}\tilde{s})\mathbb{E}\big[\mathbf{1}\{T=t\}\ell(W,z_{t,i})R\big|\tilde{S}=\tilde{s}\big],$$

we have

$$I(W,T,R;Z_{t,i}|S^{-t,i})\geq-\frac{\lambda^2\sigma^2}{2}+\lambda\Big(\mathbb{E}\big[\mathbf{1}\{T=t\}\ell(W,Z'_{t,i})R\big]- \mathbb{E}\big[\mathbf{1}\{T=t\}\ell(W,Z_{t,i})R\big]\Big).$$

Therefore, by (4.52),

$$\mathbb{E}[L_\mu(W)R]-\mathbb{E}[L_{S_T}(W)R]\leq\frac{m\lambda\sigma^2}{2}+\frac{m}{\lambda}\frac{1}{mn}\sum_{i=1}^{n}\sum_{t=1}^{m}I(W,T,R;Z_{t,i}|\tilde{S}^{-t,i})$$

for all $\lambda\in R$. Optimizing over λ, we get

$$\mathbb{E}[L_\mu(W)R]-\mathbb{E}[L_{S_T}(W)R]\leq m\sqrt{\frac{2\sigma^2}{mn}\sum_{i=1}^{n}\sum_{t=1}^{m}I(W,T,R;Z_{t,i}|\tilde{S}^{-t,i})}$$

which proves the claim. □

Proof of Theorem 4.5

The proof is based on the monitor technique proposed in [80], and parallels with the proof for Theorem 4.4. First, let $P_{W_1,\ldots,W_m|\tilde{S}}$ be the parallel execution of $m=\lfloor 1/\beta\rfloor$ independent copies of $P_{W|S}$: for $t=1,\ldots,m$, an independent

copy of $P_{W|S}$ takes S_t as input and outputs W_t. Define the set

$$F = \{t = 1, \ldots, m : (W_t, t, 1), (W_t, t, -1)\}. \tag{4.53}$$

Then, let the output of the monitor be a sample (W^*, T^*, R^*) from F according to the distribution

$$P_{W^*,T^*,R^*|\tilde{S}=\tilde{s},W^m=w^m}(w^*, t^*, r^*) \propto \exp\Big(\frac{\varepsilon n r^*}{2}\big(L_\mu(w^*) - L_{s_{t^*}}(w^*)\big)\Big) \tag{4.54}$$

for $(w^*, t^*, r^*) \in F$. Note that given $\tilde{s}$ and $w_1, \ldots, w_m$, the output (W^*, T^*, R^*) is essentially obtained from an exponential mechanism applied to $\tilde{s}$ with respect to the function

$$u\big((w^*, t^*, r^*), \tilde{s}\big) = r^*\big(L_\mu(w^*) - L_{s_{t^*}}(w^*)\big), \qquad (w^*, t^*, r^*) \in F. \tag{4.55}$$

It can be shown that the above exponential mechanism with $\ell \in [0, 1]$ satisfies that, for two datasets $\tilde{s}$ and $\tilde{s}'$ such that $d_H(\tilde{s}, \tilde{s}') = 1$,

$$e^{-\varepsilon} \leq \frac{P_{W^*,T^*,R^*|\tilde{S}=\tilde{s},W^m=w^m}(w^*, t^*, r^*)}{P_{W^*,T^*,R^*|\tilde{S}=\tilde{s}',W^m=w^m}(w^*, t^*, r^*)} \leq e^{\varepsilon}, \quad \forall (w^*, t^*, r^*) \in F; \tag{4.56}$$

hence,

$$\sup_{\tilde{s},\tilde{s}':d_H(\tilde{s},\tilde{s}')=1} \sup_{w^m} D(P_{W^*,T^*,R^*|\tilde{S}=\tilde{s},W^m=w^m} \| P_{W^*,T^*,R^*|\tilde{S}=\tilde{s}',W^m=w^m}) \leq \varepsilon \tag{4.57}$$

and

$$\frac{1}{mn}\sum_{t=1}^{m}\sum_{i=1}^{n} I(W^*, T^*, R^*; Z_{t,i}|\tilde{S}^{-t,i}, W_1, \ldots, W_m) \leq \varepsilon. \tag{4.58}$$

In addition, by Lemma 4.6, we know that $P_{W_1,\ldots,W_m|\tilde{S}}$ satisfies

$$\frac{1}{mn}\sum_{t=1}^{m}\sum_{i=1}^{n} I(W_1, \ldots, W_m; Z_{t,i}|\tilde{S}^{-t,i}) \leq \varepsilon. \tag{4.59}$$

Therefore, by the chain rule of mutual information,

$$\begin{aligned}
&\frac{1}{mn}\sum_{t=1}^{m}\sum_{i=1}^{n} I(W^*, T^*, R^*; Z_{t,i}|\tilde{S}^{-t,i}) \\
&\le \frac{1}{mn}\sum_{t=1}^{m}\sum_{i=1}^{n} I(W_1, \ldots, W_m, W^*, T^*, R^*; Z_{t,i}|\tilde{S}^{-t,i}) \\
&\le 2\varepsilon.
\end{aligned}$$

By Lemma 4.7 and the assumption that $\ell \in [0,1]$ (hence $\sigma^2 = 1/4$ in Lemma 4.7),

$$\mathbb{E}\big[R^*\big(L_\mu(W^*) - L_{\tilde{S}_{T^*}}(W^*)\big)\big] \le m\sqrt{\varepsilon}. \tag{4.60}$$

On the other hand, in view of (4.54), we can apply Lemma 4.5 with the set F, the function $f(w^*, t^*, r^*) = u\big((w^*, t^*, r^*), \tilde{s}\big)$, and $\eta = \varepsilon n/2$ to get

$$\begin{aligned}
&\mathbb{E}\big[R^*\big(L_\mu(W^*) - L_{\ S_{T^*}}(W^*)\big)\big|\tilde{S} = \tilde{s}, W^m = w^m\big] \\
&\ge \max_{(w^*,t^*,r^*)\in F} u\big((w^*, t^*, r^*), \tilde{s}\big) - \frac{2}{\varepsilon n}\log *|F| \\
&= \max_{t\in[m]} \big|L_\mu(w_t) - L_{s_t}(w_t)\big| - \frac{2}{\varepsilon n}\log(2m),
\end{aligned}$$

which implies

$$\mathbb{E}\big[R^*\big(L_\mu(W^*) - L_{S_{T^*}}(W^*)\big)\big] \ge \mathbb{E}\Big[\max_{t\in[m]} \big|L_\mu(W_t) - L_{S_t}(W_t)\big|\Big] - \frac{2}{\varepsilon n}\log(2m). \tag{4.61}$$

Combining (4.60) and (4.61) gives

$$\mathbb{E}\Big[\max_{t\in[m]} \big|L_\mu(W_t) - L_{S_t}(W_t)\big|\Big] < \frac{2}{\varepsilon n}\log(2m) + m\sqrt{\varepsilon}. \tag{4.62}$$

The rest of the proof is by contradiction. Suppose the algorithm $P_{W|S}$ does not satisfy the claimed generalization property, namely,

$$\mathbb{P}\big[\big|L_\mu(W) - L_S(W)\big| > \alpha\big] > \beta.$$

Then by the independence among the pairs $(\tilde{S}_t, W_t)$, $t = 1, \ldots, m$,

$$\mathbb{P}\Big[\max_{t\in[m]} \big|L_\mu(W_t) - L_{S_t}(W_t)\big| \geq \alpha\Big] > 1 - \Big(1-\beta\Big)^m > \frac{1}{2}.$$

Thus

$$\mathbb{E}\Big[\max_{t\in[m]} \big|L_\mu(W_t) - L_{S_t}(W_t)\big|\Big] > \alpha/2. \tag{4.63}$$

Combining (4.62) and (4.63) gives

$$n < \frac{2}{\varepsilon\big(\frac{\alpha}{2} - \frac{1}{\beta}\sqrt{\varepsilon}\big)} \log\frac{2}{\beta},$$

which contradicts the assumption that $n \geq \frac{2}{\varepsilon\left(\frac{\alpha}{2} - \frac{1}{\beta}\sqrt{\varepsilon}\right)} \log\frac{2}{\beta}$, and hence completes the proof.

4.8.3 Proof of Equation (4.39)

Recall that for a Gibbs algorithm $P_{W|S}$, $\mathrm{d}P_{W|S=s}/\mathrm{dQ}$ is given for all $s \in \mathsf{Z}^n$. Thus, for any two datasets $s, s' \in \mathsf{Z}^n$,

$$\log\frac{\mathrm{d}P_{W|S=s}}{\mathrm{d}P_{W|S=s}} = \log \mathrm{d}P_{W|S=s} - \log \mathrm{d}P_{W|S=s'}.$$

According to the definition of the Gibbs algorithm, for any w, s, and z_i',

$$\log \mathrm{d}P_{W|S=s}(w) = -\frac{\beta}{n}\sum_{j\neq i}\ell(w, z_j) - \frac{\beta}{n}\ell(w, z_i) + \log \mathrm{d}Q(w) - g(s)$$

and

$$\log \mathrm{d}P_{W|S=s_{(i)}}(w) = -\frac{\beta}{n}\sum_{j\neq i}\ell(w, z_j) - \frac{\beta}{n}\ell(w, z_i') + \log \mathrm{d}Q(w) - g(s_{(i)}),$$

where

$$g(s) \triangleq \log \mathbb{E}_Q\Big[\exp\Big\{-\frac{\beta}{n}\sum_{j\neq i}\ell(\bar{W}, z_j) - \frac{\beta}{n}\ell(\bar{W}, z_i)\Big\}\Big]$$

and

$$g(s_{(i)}) \triangleq \log \mathbb{E}_Q\Big[\exp\Big\{-\frac{\beta}{n}\sum_{j\neq i}\ell(\bar{W}, z_j) - \frac{\beta}{n}\ell(\bar{W}, z_i')\Big\}\Big].$$

Note that for each $i \in [n]$,

$$\int \mu^{\otimes n}(\mathrm{d}s)\mu(\mathrm{d}z_i')g(s) = \int \mu^{\otimes n}(\mathrm{d}s)\mu(\mathrm{d}z_i')g(s_{(i)}).$$

Therefore,

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\int \mu^{\otimes n}(\mathrm{d}s)\mu(\mathrm{d}z_i')D(P_{W|S=s}\|P_{W|S=s_{(i)}}) \\
&\quad= \frac{1}{n}\sum_{i=1}^{n}\int \mu^{\otimes n}(\mathrm{d}s)\mu(\mathrm{d}z_i')P_{W|S=s}(\mathrm{d}w)\big(\log \mathrm{d}P_{W|S=s}(w) - \log \mathrm{d}P_{W|S=s_{(i)}}(w)\big) \\
&\quad= \frac{\beta}{n^2}\sum_{i=1}^{n}\int \mu^{\otimes n}(\mathrm{d}s)\mu(\mathrm{d}z_i')P_{W|S=s}(\mathrm{d}w)\big(\ell(w, z_i') - \ell(w, z_i)\big) + \\
&\qquad\frac{1}{n}\sum_{i=1}^{n}\int \mu^{\otimes n}(\mathrm{d}s)\mu(\mathrm{d}z_i')\big(g(s_{(i)}) - g(s)\big) \\
&\quad= \frac{\beta}{n^2}\sum_{i=1}^{n}\Big(\mathbb{E}[L_\mu(W)] - \int \mu^{\otimes n}(\mathrm{d}s)P_{W|S=s}(\mathrm{d}w)\ell(w, z_i)\Big) \\
&\quad= \frac{\beta}{n}\big(\mathbb{E}[L_\mu(W)] - \mathbb{E}[L_S(W)]\big),
\end{aligned}$$

which proves (4.39).

References

[1] T. Han and S. Amari, "Statistical inference under multiterminal data compression," *IEEE Trans. Inform. Theory*, vol. 44, no. 6, pp. 2300–2324, 1998.

[2] Y. Zhang, J. Duchi, M. Jordan, and M. Wainwright, "Information-theoretic lower bounds for distributed statistical estimation with communication constraints," in *27th Annual Conference on Neural Information Processing Systems (NIPS)*, 2013.

[3] J. Duchi, M. Jordan, M. Wainwright, and Y. Zhang, "Optimality guarantees for distributed statistical estimation," *arXiv preprint*, 2014. [Online]. Available: http://arxiv.org/abs/1405.0782

[4] A. Garg, T. Ma, and H. L. Nguyen, "On communication cost of distributed statistical estimation and dimensionality," in *28th Annual Conference on Neural Information Processing Systems (NIPS)*, 2014.

[5] M. Braverman, A. Garg, T. Ma, H. L. Nguyen, and D. P. Woodruff, "Communication lower bounds for statistical estimation problems via a distributed data processing inequality," in *Proceedings of 48th Annual ACM SIGACT Symposium on Theory of Computing (STOC)*, 2016.

[6] O. Shamir, "Fundamental limits of online and distributed algorithms for statistical learning and estimation," in *28th Annual Conference on Neural Information Processing Systems (NIPS)*, 2014.

[7] X. Chen, A. Guntuboyina, and Y. Zhang, "On bayes risk lower bounds," *arXiv preprint*, 2014. [Online]. Available: http://arxiv.org/abs/1410.0503

[8] I. Csiszár, "A class of measures of informativity of observation channels," *Periodica Math. Hungar.*, vol. 2, no. 1–4, pp. 191–213, 1972.

[9] B. S. Clarke and A. R. Barron, "Jeffreys' prior is asymptotically least favorable under entropy risk," *Journal of Statistical Planning and Inference*, vol. 41, no. 1, pp. 37–60, 1994.

[10] B. S. Clarke and A. R. Barron, "Information-theoretic asymptotics of Bayes methods," *IEEE Trans. Inform. Theory*, vol. 36, no. 3, pp. 453–471, 1990.

[11] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inform. Theory*, vol. 56, no. 5, pp. 2307–2359, 2010.

[12] T. S. Han and S. Verdú, "Generalizing the Fano inequality," *IEEE Trans. Inform. Theory*, vol. 40, no. 4, pp. 1247–1251, 1994.

[13] H. V. Poor and S. Verd, "A lower bound on the error probability of multihypothesis testing," *IEEE Trans. Inform. Theory*, vol. 41, no. 6, pp. 1992–1993, 1995.

[14] J. Duchi and M. Wainwright, "Distance-based and continuum Fano inequalities with applications to statistical estimation," *Technical report, UC Berkeley*, 2013.

[15] R. Gray, *Source Coding Theory*. Kluwer Academic Publishers, 1990.

[16] J. Seidler, "Bounds on the mean-square error and the quality of domain decisions based on mutual information," *IEEE Trans. Inform. Theory*, vol. 17, no. 6, pp. 655–665, 1971.

[17] Y. Wu, *Lecture notes for ECE598 (UIUC): Information-theoretic methods in high-dimensional statistics*, 2016. [Online]. Available: http://www.ifp.illinois.edu/~yihongwu/teaching/598

[18] M. Raginsky, "Strong data processing inequalities and Φ-Sobolev inequalities for discrete channels," *IEEE Trans. Inform. Theory*, vol. 62, no. 6, pp. 3355–3389, 2016.

[19] V. Anantharam, A. Gohari, S. Kamath, and C. Nair, "On maximal correlation, hypercontractivity, and the data processing inequality studied by Erkip and Cover," *arXiv preprint*, 2013. [Online]. Available: http://arxiv.org/abs/1304.6133

[20] R. Ahlswede and P. Gács, "Spreading of sets in product spaces and hypercontraction of the Markov operator," *Ann. Probab.*, vol. 4, no. 6, pp. 925–939, 1976.

[21] E. Erkip and T. Cover, "The efficiency of investment information," *IEEE Trans. Inform. Theory*, vol. 44, no. 3, pp. 1026–1040, 1998.

[22] J. E. Cohen, Y. Iwasa, G. Rautu, M. B. Ruskai, E. Seneta, and G. Zbăganu, "Relative entropy under mappings by stochastic matrices," *Lin. Algebra Appl.*, vol. 179, pp. 211–235, 1993.

[23] W. Evans and L. Schulman, "Signal propagation and noisy circuits," *IEEE Trans. Inform. Theory*, vol. 45, no. 7, pp. 2367–2373, 1999.

[24] Y. Polyanskiy and Y. Wu, "Strong data-processing inequalities for channels and Bayesian networks," *arXiv preprint*, 2015. [Online]. Available: http://arxiv.org/abs/1508.06025

[25] T. Cover and J. Thomas, *Elements of Information Theory*, 2nd ed. New York: Wiley.

[26] C. Calabro, "The exponential complexity of satisfiability problems," Ph.D. dissertation, University of California, San Diego, 2009.

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

[28] R. Dobrushin and B. Tsybakov, "Information transmission with additional noise," *IRE Trans. on Inform. Theory*, vol. 8, no. 5, pp. 293–304, Sep 1962.

[29] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, 2nd ed. Cambridge University Press, 2011.

[30] T. Berger, Z. Zhang, and H. Viswanathan, "The CEO problem," *IEEE Trans. Inform. Theory*, vol. 42, no. 3, pp. 887–902, 1996.

[31] T. A. Courtade, "Outer bounds for multiterminal source coding via a strong data processing inequality," in *IEEE International Symposium on Information Theory (ISIT)*, 2013.

[32] M. El Gamal and L. Lai, "Are Slepian-Wolf rates necessary for distributed parameter estimation?" in *Proc. 53th Annu. Allerton Conf. on Commun., Control, and Comput.*, 2015.

[33] Y. Polyanskiy and S. Verdú, "Arimoto channel coding converse and Rényi divergence," in *Proc. 48th Annu. Allerton Conf. on Commun., Control, and Comput.*, 2010, pp. 1327–1333.

[34] Y. Polyanskiy and Y. Wu, "Lecture Notes on Information Theory," Lecture Notes for ECE563 (UIUC) and 6.441 (MIT), 2012-2016. [Online]. Available: http://people.lids.mit.edu/yp/homepage/data/itlectures_v4.pdf

[35] Y. Polyanskiy, "Channel coding: Non-asymptotic fundamental limits," Ph.D. dissertation, Princeton University, 2010.

[36] W. Feller, *An Introduction to Probability Theory and Its Applications*, 3rd ed. Wiley, 1968, vol. 1.

[37] O. Kallenberg, *Foundations of Modern Probability*, 2nd ed. Springer, 2002.

[38] O. Ayaso, D. Shah, and M. Dahleh, "Information-theoretic bounds for distributed computation over networks of point-to-point channels," *IEEE Trans. Inform. Theory*, vol. 56, no. 12, pp. 6020–6039, 2010.

[39] G. Como and M. Dahleh, "Lower bounds on the estimation error in problems of distributed computation," in *Proc. Inform. Theory and Applications Workshop*, 2009, pp. 70–76.

[40] N. Goyal, G. Kindler, and M. Saks, "Lower bounds for the noisy broadcast problem," *SIAM Journal on Computing*, vol. 37, no. 6, pp. 1806–1841, 2008.

[41] C. Dutta, Y. Kanoria, D. Manjunath, and J. Radhakrishnan, "A tight lower bound for parity in noisy communication networks," in *Proc. ACM Symposium on Discrete Algorithms (SODA)*, 2014, pp. 1056–1065.

[42] M. Braverman, "Interactive information and coding theory," in *Proc. Int. Congress Math.*, 2014.

[43] A. Orlitsky and J. Roche, "Coding for computing," *IEEE Trans. Inform. Theory*, vol. 47, no. 3, pp. 903–917, 2001.

[44] J. Körner and K. Marton, "How to encode the modulo-two sum of binary sources," *IEEE Trans. Inform. Theory*, vol. 25, no. 2, pp. 219–221, 1979.

[45] A. B. Wagner, S. Tavildar, and P. Viswanath, "Rate region of the quadratic Gaussian two-encoder source-coding problem," *IEEE Trans. Inform. Theory*, vol. 54, no. 5, pp. 1938–1961, 2008.

[46] A. El Gamal and Y.-H. Kim, *Network Information Theory*. Cambridge Univ. Press, 2011.

[47] A. Giridhar and P. Kumar, "Toward a theory of in-network computation in wireless sensor networks," *IEEE Communications Magazine*, vol. 44, no. 4, pp. 98–107, April 2006.

[48] R. Gallager, "Finding parity in a simple broadcast network," *IEEE Trans. Inform. Theory*, vol. 34, no. 2, pp. 176–180, 1988.

[49] L. Schulman, "Coding for interactive communication," *IEEE Trans. Inform. Theory*, vol. 42, no. 6, pp. 1745–1756, 1996.

[50] S. Rajagopalan and L. Schulman, "A coding theorem for distributed computation," in *ACM Symposium on Theory of Computing*, 1994.

[51] R. Carli, G. Como, P. Frasca, and F. Garin, "Distributed averaging on digital erasure networks," *Automatica*, vol. 47, no. 115-121, 2011.

[52] S. Kar and J. Moura, "Distributed consensus algorithms in sensor networks with imperfect communication: Link failures and channel noise," *IEEE Trans. Signal Process.*, vol. 57, no. 1, pp. 355–369, 2009.

[53] N. Noorshams and M. Wainwright, "Non-asymptotic analysis of an optimal algorithm for network-constrained averaging with noisy links," *IEEE J. Sel. Top. Sign. Proces.*, vol. 5, no. 4, pp. 833–844, 2011.

[54] L. Ying, R. Srikant, and G. Dullerud, "Distributed symmetric function computation in noisy wireless sensor networks with binary data," in *International Symposium on Modeling and Optimization in Mobile, Ad-Hoc and Wireless networks (WiOpt)*, 2006.

[55] S. Deb, M. Medard, and C. Choute, "Algebraic gossip: a network coding approach to optimal multiple rumor mongering," *IEEE Trans. Inform. Theory*, vol. 52, no. 6, pp. 2486–2507, 2006.

[56] V. V. Petrov, *Sums of Independent Random Variables*. Berlin: Springer-Verlag, 1975.

[57] P. Tiwari, "Lower bounds on communication complexity in distributed computer networks," *J. ACM*, vol. 34, no. 4, pp. 921–938, Oct. 1987.

[58] A. Chattopadhyay, J. Radhakrishnan, and A. Rudra, "Topology matters in communication," in *Proc. IEEE Annu. Symp. on Foundations of Comp. Sci. (FOCS)*, Oct 2014, pp. 631–640.

[59] R. Gray, "A new class of lower bounds to information rates of stationary sources via conditional rate-distortion functions," *IEEE Trans. Inform. Theory*, vol. 19, no. 4, pp. 480–489, 1973.

[60] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

[61] V. Kostina and S. Verdú, "Lossy joint source-channel coding in the finite blocklength regime," *IEEE Trans. Inform. Theory*, vol. 59, no. 5, pp. 2545–2575, 2013.

[62] A. Kolmogorov, "Sur les propriétés des fonctions de concentrations de M. P. Lévy," *Ann. Inst. H. Poincaré*, vol. 16, pp. 27–34, 1958.

[63] H. H. Nguyen and V. H. Vu, "Small ball probability, inverse theorems, and applications," in *Erdős Centennial*, ser. Bolyai Society Mathematical Studies. Springer, 2013, vol. 25. [Online]. Available: http://arxiv.org/abs/1301.0019

[64] S. G. Bobkov and G. P. Chistyakov, "On concentration functions of random variables," *J. Theor. Probab.*, vol. 28, no. 3, pp. 976–988, 2015, published online.

[65] M. Rudelson and R. Vershynin, "The Littlewood–Offord problem and invertibility of random matrices," *Adv. Math.*, vol. 218, pp. 600–633, 2008.

[66] M. Rudelson and R. Vershynin, "Small ball probabilities for linear images of high dimensional distributions," *arXiv1402.4492R*, Feb. 2014. [Online]. Available: https://arxiv.org/abs/1402.4492

[67] P. Erdős, "On a lemma of Littlewood and Offord," *Bull. Amer. Math. Soc.*, vol. 51, pp. 898–902, 1945.

[68] S. Bobkov and M. Madiman, "The entropy per coordinate of a random vector is highly constrained under convexity conditions," *IEEE Trans. Inform. Theory*, vol. 57, no. 8, pp. 4940–4954, 2011.

[69] R. Carli, G. Como, P. Frasca, and F. Garin, "Average consensus on digital noisy networks," *1st IFAC Workshop on Estimation and Control of Networked Systems*, 2009.

[70] H. S. Witsenhausen, "On sequences of pairs of dependent random variables," *SIAM J. Appl. Math.*, vol. 28, no. 1, pp. 100–113, Jan. 1975.

[71] Y. Polyanskiy and Y. Wu, "Dissipation of information in channels with input constraints," *IEEE Trans. Inform. Theory*, vol. 62, no. 1, pp. 35–55, 2016.

[72] L. Devroye and T. Wagner, "Distribution-free performance bounds for potential function rules," *IEEE Transactions on Information Theory*, vol. 25, no. 5, pp. 601–604, 1979.

[73] W. H. Rogers and T. J. Wagner, "A finite sample distribution-free performance bound for local discrimination rules," *The Annals of Statistics*, vol. 6(3), pp. 506–514, 1978.

[74] D. Ron and M. Kearns, "Algorithmic stability and sanity-check bounds for leave-one-out crossvaildation," *Neural Computation*, vol. 11, no. 6, pp. 1427–1453, 1999.

[75] O. Bousquet and A. Elisseeff, "Stability and generalization," *J. Machine Learning Res.*, vol. 2, pp. 499–526, 2002.

[76] T. Poggio, R. Rifkin, S. Mukherjee, and P. Niyogi, "General conditions for predictivity in learning theory," *Nature*, vol. 428, no. 6981, pp. 419–422, 2004.

[77] S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan, "Learnability, stability and uniform convergence," *J. Mach. Learn. Res.*, vol. 11, pp. 2635–2670, 2010.

[78] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Theory of Cryptography Conference*. Springer Berlin Heidelberg, 2006, pp. 265–284.

[79] C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. Roth, "Generalization in adaptive data analysis and holdout reuse," in *28th Annual Conference on Neural Information Processing Systems (NIPS)*, 2015.

[80] R. Bassily, K. Nissim, A. Smith, T. Steinke, U. Stemmer, and J. Ullman, "Algorithmic stability for adaptive data analysis," in *Proceedings of The 48th Annual ACM SIGACT Symposium on Theory of Computing*, 2016.

[81] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Foundations and Trends in Theoretical Computer Science*, vol. 9, no. 3-4, 2014.

[82] P. Kairouz, S. Oh, and P. Viswanath, "The composition theorem for differential privacy," in *Proc. 32nd Int. Conf. on Machine Learning (ICML)*, 2015.

[83] M. Raginsky, A. Rakhlin, M. Tsao, Y. Wu, and A. Xu, "Information-theoretic analysis of stability and bias of learning algorithms," in *Proceedings of IEEE Information Theory Workshop*, 2016.

[84] D. Russo and J. Zou, "Controlling bias in adaptive data analysis using information theory," in *Proceedings of The 19th International Conference on Artificial Intelligence and Statistics*, 2016.

[85] T. Zhang, "Information-theoretic upper and lower bounds for statistical estimation," *IEEE Trans. Inform. Theory*, vol. 52, no. 4, pp. 1307 – 1321, 2006.

[86] M. Raginsky, "Lecture Notes on Statistical Learning Theory," Lecture Notes for ECE598MR (UIUC), 2013-2015. [Online]. Available: http://maxim.ece.illinois.edu/teaching/fall15b/index.html

[87] S. Verdú and T. Weissman, "The information lost in erasures," *IEEE Trans. Inform. Theory*, vol. 54, no. 11, pp. 5030–5058, 2008.

[88] Y.-X. Wang, J. Lei, and S. E. Fienberg, "On-average kl-privacy and its equivalence to generalization for max-entropy mechanisms," in *Proceedings of the International Conference on Privacy in Statistical Databases*, 2016.

[89] P. Cuff and L. Yu, "Differential privacy as a mutual information constraint," in *Proceedings of 23rd ACM Conference on Computer and Communications Security*, 2016.

[90] C. Villani, *Topics in Optimal Transportation*, ser. Graduate Studies in Mathematics. Providence, RI: Amer. Math. Soc., 2003, vol. 58.

[91] M. Raginsky and I. Sason, *Concentration of Measure Inequalities in Information Theory, Communications, and Coding*, 2nd ed. Now Publishers, 2014.

[92] S. Boucheron, G. Lugosi, and P. Massart, *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford Univ. Press, 2013.

[93] F. McSherry and K. Talwar, "Mechanism design via differential privacy," in *Proceedings of 48th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, 2007.

[94] S. Verdú, "The exponential distribution in information theory," *Problems of Information Transmission*, vol. 32, no. 1, pp. 86–95, 1996.

[95] C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. Roth, "Preserving statistical validity in adaptive data analysis," in *Proc. of 47th ACM Symposium on Theory of Computing (STOC)*, 2015.

[96] Y.-X. Wang, J. Lei, and S. E. Fienberg, "A minimax theory for adaptive data analysis," *arXiv:1602.04287*, 2016. [Online]. Available: https://arxiv.org/abs/1602.04287