Information Theory and Noisy Computation

William S. Evans[1]

TR-94-057
November, 1994

Abstract

The information carried by a signal unavoidably decays when the signal is corrupted by random noise. This occurs when a noisy channel transmits a message as well as when a noisy component performs computation. We first study this signal decay in the context of communication and obtain a tight bound on the decay of the information carried by a signal as it crosses a noisy channel. We then use this information theoretic result to obtain depth lower bounds in the noisy circuit model of computation defined by von Neumann. In this model, each component fails (produces 1 instead of 0 or vice-versa) independently with a fixed probability, and yet the output of the circuit should be correct with high probability. Von Neumann showed how to construct circuits in this model that reliably compute a function and are no more than a constant factor deeper than noiseless circuits for the function. Our result implies that such a multiplicative increase in depth is necessary for reliable computation. The result also indicates that above a certain level of component noise, reliable computation is impossible.

We use a similar technique to lower bound the size of reliable circuits in terms of the noise and complexity of their components, and the sensitivity of the function they compute. Our

[1] The author was partially supported by National Science Foundation grant CCR92-01092.

bound is asymptotically equivalent to previous bounds as a function of sensitivity, but unlike previous bounds, its dependence on component noise implies that as this noise increases to 1/2, the size of reliable circuits must increase unboundedly. In all cases, the bound is strictly stronger than previous results.

Using different techniques, we obtain the exact threshold for component noise, above which noisy formulas cannot reliably compute all functions. We obtained an upper bound on this threshold in studying the depth of noisy circuits. The fact that this bound is only slightly larger than the true threshold indicates the high precision of our information theoretic techniques.

Contents

List of Figures **iii**

1 Introduction **1**

1.1 Information in Noisy Computation . . . 2

1.2 Model of Noisy Computation . . . 2

1.3 Outline of Results . . . 4

1.4 Information Theory . . . 5

2 Signal Decay **7**

2.1 Reduction to Weak Signal Case . . . 8

2.2 Reduction to Divergence Ratio . . . 11

2.3 Divergence as Square of L_2 Distance . . . 12

2.4 Signal Decay Theorem . . . 13

2.5 Bounds for 2-input, m-output Channels . . . 15

3 Noisy Circuit Depth **18**

3.1 Intuition . . . 19

3.2 Circuit Anomalies . . . 20

3.3 Noisy Circuit Depth Lower Bound . . . 22

4 Noisy Circuit Size **27**

4.1 Noisy Input Wires . . . 28

4.2 Lower Bound on Noisy Circuit Size . . . 30

4.2.1 Sensitivity . . . 30

4.2.2 Block Sensitivity 34
4.2.3 Using Mutual Information 35

5 Threshold for Noisy Computation 36
5.1 Threshold Value 37
5.2 Negative Result 40
5.3 Contraction of $\mathbf{S}(\mathbf{a})$ 42
5.4 Reduction lemmas 43
5.5 Positive Result 44

6 Future Work 47
6.1 Information Theory 47
6.2 Depth and Size Bounds 48
6.3 Threshold Bounds 48
6.4 Markov Chains 48
6.5 Quadratic Dynamical Systems 49

A Information Theory 51
A.1 Definitions 51
A.2 Properties of Entropy and Information 52

B L_c Distance 57

C Bounding Noisy Circuit Size via Information 61

Bibliography 64

List of Figures

2.1 A binary channel and its corresponding row-stochastic matrix. 8

2.2 Visualization of $I(X;Y)$ and $I(X;Z)$ 9

2.3 The map σ applied to two infinitesimally separated distributions. 12

3.1 Upper bounds on $I(X;Z)/I(X;Y)$ for symmetric channels with noise $(1-\xi)/2$. 20

3.2 Lower bounds on factor of increase in circuit depth using 3-input, ϵ-noisy gates. . 25

4.1 Lower bounds on factor of $s \log s$ for noisy circuit size using 3-input gates. 34

5.1 Bounds on the threshold for noisy computation. 38

5.2 Error of ϵ-noisy, 3-input majority gate as a function of input error for $\epsilon = \beta_3$ and for $\epsilon < \beta_3$. 39

5.3 Contraction of $S(a)$ to $S(m_{\epsilon,k}(a))$ caused by one noisy gate. 42

Acknowledgments

Many people contributed to my enjoyable education at Berkeley. My advisor, Manuel Blum, viewed every research problem I presented to him with an enthusiasm that was infectious. I thank him for his support and for inspiring a joy for research and teaching that saved me from early discouragement.

Leonard Schulman also served as my “advisor” during the past two years. Leonard was instrumental in the research presented in this thesis, but his contribution goes beyond that. I learned more from him about how to do research than from anyone. I thank him for his patience, accessibility, and clarity of thought that made daunting problems approachable. It truly has been a pleasure to work with him.

Berkeley has a tremendous theory faculty. I thank Umesh Vazirani for showing me that beautiful problems lie just below the surface. I thank both Umesh and Richard Karp for teaching the best classes I’ve ever had. Mike Luby and Alistair Sinclair I thank for their friendship and encouragement. I also thank Mike and Manuel for making me feel welcome at ICSI (the International Computer Science Institute), where I wrote most of this thesis.

The theory students at Berkeley are what make it such a terrific place to learn. Nina Amenta encouraged me to have more fun than I usually thought advisable, for which I give my warmest thanks. I thank Dana Randall and Diane Hernek for their great friendship, sense of humor, and advice. For coffee and the cryptic as well as his friendship and a welcome perspective, I thank Ashu Rege. I thank Sandy Irani and Ronitt Rubinfeld for their advice and inspiration.

I cannot imagine Berkeley without Dan Jurafsky. I loved his dinners, his projects, and his company. Thanks to Nervous for Nigel and the Haste Street Players for terrific performances and great camaraderie. Many, many other people helped make life at Berkeley exciting, interesting, and diverse. I especially thank Carolyn Helmke, Becky Gross, Jackie Caplan-Auerbach, and Mary Lammert.

I save a special thanks for John Hartman, Steve Lucco, Ken Shirriff, Ramón Cáceres, Mike Hohmeyer, Tristan Barrientos, and Randi Weinstein. Living in the Hillegass House was like being a part of a big, fun family. It’s hard to leave such a great bunch of people.

Finally, I thank my parents David and Dorothy and my brother Hal for their constant love and support.

Chapter 1

Introduction

> *Our present treatment of error is unsatisfactory and ad hoc. It is the author's conviction, voiced over many years, that error should be treated by thermodynamical methods, and be the subject of a thermodynamical theory, as information has been, by the work of L. Szilard and C.E. Shannon.*
>
> J. von Neumann 1952

The decay of an information signal as it propagates through a medium is an unavoidable phenomenon, familiar in almost every form of communication: sound, wire, radio and so on.

The problem of signal decay is not restricted to communication: that it plagues long computations, as well, was all too apparent to the first users of electronic computers, and was for example the spur for Hamming's interest in coding theory [12].

Von Neumann recognized that, rather than being technological and passing, this signal decay was an essential difficulty for large-scale computations, which by their nature rely on the propagation of long chains of events [28]. Von Neumann's goal was to subject noisy computation to the same thermodynamical treatment as communication had received in the contemporary work of Shannon [23]. Surprisingly, it took over thirty-five years before the tools developed by Shannon to study information and communication were successfully applied to the problem of noisy computation, in the work of Pippenger [16]. During that time, Shannon's methods proved so useful for communication that an entire area devoted to the study of information developed.

In this thesis, we present new results in the area of information theory and use these results, along with other techniques, to obtain lower bounds on the complexity of noisy computation. A primary goal of this research is to indicate the applicability of information theoretic tools to the study of computation.

1.1 Information in Noisy Computation

Imagine computation as a sequence of individual steps. Each step produces an output which carries some information about the original input; and that output is the input to the next step in the path. The computation forms a chain in which each intermediate result depends only on its predecessor. The final output of the computation must carry a large amount of information about the input (assuming the function being computed is non-trivial). If the steps are subject to random noise then the information about the input carried by successive outputs along one path decreases. We can obtain the required amount of information at the output by increasing the number of "parallel" computation paths. Thus, the decay in information caused by each noisy step combined with the required accuracy of the final output leads to complexity bounds on the structure of the computation.

To make the reasoning more formal, we use Shannon's mutual information as a measure of this intuitive notion of information. Let X, the outcome of a probabilistic experiment, be the input to the computation. Let Y represent the output of some step on the computation path. The random variable Y may be thought of as a noisy signal reporting on the outcome of the experiment. That is, each experimental outcome will give rise to a different conditional probability distribution on the random variable Y. The mutual information $I(X;Y)$ measures how statistically distinguishable these conditional distributions are. If the distributions are different then the information Y carries about X is large. If, on the other hand, the conditional distributions on Y given X are all the same (i.e. Y is independent of X) then Y carries zero information about X.

Now suppose Z is the output of the computation step with input Y in the computation path. Since Z depends on X only through Y, it should be the case that the information Z carries about X is no more than the information carried by Y. This is precisely the data processing inequality; that $I(X;Z)$ is no more than $I(X;Y)$ (see section 1.4 and appendix A). The statement reflects the fact that the conditional distributions on Z are less distinguishable than those on Y.

In this simple view of computation, a noisy computation step can be viewed as a noisy communication channel, and our interest is in the amount of information that such a channel preserves. The first result in this thesis is a tight upper bound on the fraction of information about X that is preserved in crossing a noisy binary channel (i.e. $I(X;Z)/I(X;Y)$). It is worth emphasizing that the bound holds regardless of the distribution on X and Y, and is a property of the channel alone.

1.2 Model of Noisy Computation

Our picture of computation as a single chain of steps illustrates the applicability of information theory but is rather imprecise. To make quite clear the mechanism of computation, we adopt a model of noisy computation, inspired by the work of Turing [25] and McCulloch and

Pitts [14], which was first proposed by von Neumann in 1952 [28].

The model of computation is the *noisy circuit*. A circuit takes n Boolean values as input and produces one Boolean output. It is composed of a collection of individual components called gates. Each gate in the circuit is one of a finite set of allowable gates called the *basis*. A gate takes a fixed number of Boolean inputs and produces a Boolean output. The inputs to a gate in the circuit may be the outputs of other gates in the circuit, inputs to the circuit, or constants 0 or 1. The output of the circuit is the output of one of the gates.

We assume that the interconnection pattern does not allow “feedback”. That is, the interconnection structure of the circuit forms a directed, acyclic graph. Vertices of the graph correspond to gates and a directed edge from u to v corresponds to gate v taking as input the output of gate u. If, in addition, the graph forms a tree (each vertex has one outgoing edge) then we call the circuit a *formula*.

The *depth* of a circuit is the length of (i.e. number of vertices on) the longest directed path in this graph. Depth is a measure of latency under the assumption that each gate causes the same delay in computation. The *size* of a circuit is the number of vertices in the graph (i.e. the number of gates in the circuit). In addition, there is the question of the complexity of the components used to construct the circuit (i.e. the basis). For our purposes, the *complexity of the basis* will be the maximum number of inputs to any one gate in the basis.

In von Neumann’s noisy circuit, each gate fails (outputs a 0 instead of a 1 or vice-versa) independently with probability ϵ. We may assume $\epsilon \leq 1/2$, since any gate which fails with probability greater than $1/2$ is acting like a $(1-\epsilon)$-noisy version of its complement.

Since each gate (including the final gate) is subject to this noise, the output of a noisy circuit for computing a function will be incorrect with probability at least ϵ. Thus a noisy circuit cannot “compute” a function in the normal sense of the word. Instead of requiring the circuit to be always correct, we ask that it be usually correct or *reliable*. A noisy circuit $(1-\delta)$-*reliably computes* a Boolean function f taking n Boolean inputs if, for all inputs $\boldsymbol{x} \in \{0,1\}^n$, the circuit outputs $f(\boldsymbol{x})$ on input $\boldsymbol{x}$ with probability at least $1-\delta$.

This is a very simple model of noise, and one which is computationally powerful. What makes this model perhaps unreasonably powerful is the assumption that gates fail with probability precisely ϵ. It is possible that a noisy circuit in this model can $(1-\delta)$-reliably compute a function, but will fail to compute the function when its gates are noiseless. In effect, precise noise permits the construction of “random bits”, allowing one to implement randomized algorithms. This objection has been addressed by designing models which are more “realistic”. (See [18] for examples and [19] for a survey of results in the theory of reliable computation.)

Most of the results in this thesis are negative results or lower bounds, saying that reliable, noisy computation cannot be done under various conditions. If we show a negative result assuming a particular model of computation, then the result holds for all less computationally powerful models as well. So assuming the most powerful model of computation leads to the widest applicability. Thus von Neumann’s noise model is an appropriate choice as our model of noisy computation. It has been used to obtain virtually all lower bounds on noisy circuit complexity [2, 16, 7, 11, 9, 22, 6, 8].

1.3 Outline of Results

Chapter 2 presents a new information theoretic result which bounds the fraction of information that can cross a noisy channel. The data processing inequality establishes an upper bound of 1 on this fraction. In his paper which pioneered the use of information for studying noisy computation, Pippenger [16] determined an upper bound of ξ on the fraction for symmetric binary channels which have probability $\epsilon = (1 - \xi)/2$ of complementing their input. We determine an exact upper bound on the fraction for any binary channel. For the symmetric channels considered by Pippenger, our bound is ξ^2. This result can be seen as a quantified version of the data processing inequality for binary valued random variables. It will be an essential part of the lower bound in chapter 3.

The remaining chapters of the thesis are devoted to presenting several lower bounds on reliable, noisy circuit complexity.

Chapter 3 concerns the *depth* of noisy circuits. In 1952, Von Neumann [28] provided the first upper bound on reliable circuit depth by showing, for any function, how to construct a reliable circuit for the function which is no more than a constant factor deeper than its noiseless circuit. The construction works as long as ϵ (the noise at each gate) is less than some threshold. In 1988, Pippenger [16] proved the first lower bound on reliable circuit depth by using information theoretic techniques to prove that there exist functions whose reliable formula depth is at least $1/\log_k(k\xi)$ times their noiseless formula depth where k is the number of inputs to each gate and $\xi = 1 - 2\epsilon$. Pippenger also showed that for $\epsilon \geq 1/2 - 1/2k$, reliable computation[1] using noisy formulas is impossible. In 1989, Feder [7] extended Pippenger's result to general circuits by abandoning the information theoretic approach for a strictly probabilistic argument.

In chapter 3, we use more precise information theoretic techniques than those used by Pippenger, and show that the factor of increase in reliable circuit depth is at least $1/\log_k(k\xi^2)$. The result also implies that reliable computation is impossible for $\epsilon \geq 1/2 - 1/2\sqrt{k}$. The result relies heavily on the "signal decay" theorem of chapter 2.

In chapter 4, we lower bound the *size* of reliable circuits. Von Neumann's 1952 work implies that, for any function, there exists a reliable circuit for the function of size $O(c \log c)$ where c is the size of a noiseless circuit for the function. This implication was made explicit in the work of Dobrushin and Ortyukov [3] and Pippenger [15].

In 1977, Dobrushin and Ortyukov [2] claimed a lower bound of $\Omega(s \log s)$ on the reliable circuit size of any function with sensitivity[2] s. In 1991, Pippenger, Stamoulis, and Tsitsiklis [20]

[1] Reliable computation refers to the ability to $(1 - \delta)$-reliably compute all Boolean functions for some fixed $\delta < 1/2$.

[2] A function's sensitivity is the maximum (over all input vectors) of the number of bits which change the function

showed that this work contains serious flaws. By a different argument, they were able to prove an $\Omega(n \log n)$ bound on the size of reliable circuits computing the parity of n inputs. In 1991, Gál [9] and Reischuk and Schmeltz [22] successfully reproved Dobrushin and Ortyukov's original claim, that functions with sensitivity s require reliable circuits of size $\Omega(s \log s)$.

Our lower bound is asymptotically equivalent to these previous bounds as a function of the sensitivity of the function being computed. The difference is in the bound's dependence on the noise and complexity of the individual components. In particular, as the component noise increases to $1/2$, our lower bound increases to infinity (whereas previous bounds do not). In all cases, our bound is strictly stronger than previous results.

In chapter 5, we establish a *threshold* on the noise of individual components above which reliable computation of all Boolean functions is impossible. In other words, if ϵ is above some threshold then for any value $\delta < 1/2$ there exist functions which cannot be $(1-\delta)$-reliably computed using ϵ-noisy components. The threshold depends only on the complexity of the individual components (i.e. the maximum number of inputs to a component). In the case of formulas built from components with an odd number of inputs, we show that this bound is tight; that reliable computation can be achieved if and only if ϵ is less than the threshold. Strangely, our result requires the component noise to be precise. In other words, that each component must fail independently with precisely probability ϵ. This extends work done by Hajek and Weller [11] who prove a tight threshold for computation by formulas using 3-input gates.

1.4 Information Theory

In order to set the stage for our discussion, we must introduce the tools from information theory which we will be using. Suppose the result X of a probabilistic experiment has a known distribution $\boldsymbol{p}_X$ (i.e. $X = i$ with probability $\boldsymbol{p}_X(i)$). How much information do we gain by performing the experiment X? If $\boldsymbol{p}_X(i) = 1$ for some i, we gain no information since the outcome is pre-determined. If $\boldsymbol{p}_X(0) = 1/2$ and $\boldsymbol{p}_X(1) = 1/2$ then we do gain some information. The question is how to measure the information gained.

Shannon's answer to this question is an information measure called the *entropy* or *self-information* of the random variable X. It is defined as

$$H(X) = -\sum_{i=1}^{n} \boldsymbol{p}_X(i) \log \boldsymbol{p}_X(i).$$

We will assume here and throughout the thesis that logarithms are base 2. One way to view the entropy of X is as the average number of bits needed to describe X. For example, if X is equally likely to be any one of n possibilities then we require $\log n$ bits to describe X.

Just a quick word about notation. We use boldface letters (e.g. $\boldsymbol{p}$, $\boldsymbol{a}$, and $\boldsymbol{\epsilon}$) to denote vectors. Since all distributions in this thesis are over discrete sample spaces, we specify

value when flipped individually (see the definition in section 4.2.1).

distributions as vectors with non-negative elements summing to 1. We often use $\boldsymbol{p}_X$ to denote a probability distribution on random variable X, and similarly, $\boldsymbol{p}_{X|Y=y}$ or $\boldsymbol{p}_{X|y}$ to denote a probability distribution on random variable X conditioned on $Y = y$.

Now suppose we could perform a second experiment with outcome Y related to X. Given the value Y, the information conveyed by X is called the *conditional entropy of X given Y* and is defined as

$$H(X|Y) = -\sum_{x,y} \boldsymbol{p}_{X,Y}(x,y) \log \boldsymbol{p}_{X|Y=y}(x).$$

Intuitively, the information gained from X should decrease (or at most remain the same) given the value of Y. In other words, Y tells us something about X. The amount of information Y conveys about X is the *mutual information between X and Y*. It is defined as

$$I(X;Y) = H(X) - H(X|Y).$$

In words, the information Y conveys about X plus the information X conveys given Y equals the total information conveyed by X. It is easy to verify that mutual information is symmetric $I(X;Y) = I(Y;X)$, and thus the information Y conveys about X is the same as the information X conveys about Y. Also $I(X;X) = H(X)$ which justifies our use of the term self-information for entropy.

Originally, mutual information was used to measure the rate of transmission across noisy channels. For example, the capacity of a noisy channel, which Shannon showed to be the maximum rate at which information can cross a channel with arbitrarily low error, is defined as the maximum mutual information between an input and output of the channel. (In an attempt to translate this result to noisy computation, Elias studied the “capacity” of noisy computation, but concluded that arbitrarily low error requires an arbitrarily low “rate” of computation [5].)

In this thesis, we view information as a measure of correlation. One of the fundamental results which support this view is the data processing inequality. Informally, it states that a function of Y cannot carry more information about X than Y itself. More precisely, if X, Y, and Z are random variables such that Z is independent of X given Y then $I(X;Z)/I(X;Y) \leq 1$. The following chapter is devoted to showing a more precise version of this inequality; one which depends on the relation between Y and Z.

Appendix A contains further definitions and results from information theory which are used in this thesis.

Chapter 2

Signal Decay

In this chapter we investigate the propagation of information signals in noisy media. We study a basic question which is relevant to any such propagation, whether in communication or in computation. To set the framework we first recall the well known “data processing inequality” for information. Let X be a random variable which is the input to a communication channel, and let Y be the output of that channel. Let Y in turn be input to another communication channel, and let Z be the output of that channel. (Thus Z depends on X solely through Y.) The mutual information $I(X;Y)$ is a nonnegative real number measuring the information available about X after the first channel; likewise $I(X;Z)$ measures the information available after the second channel. The data processing inequality states that no matter what the properties of the second channel, $I(X;Z) \leq I(X;Y)$.

$$\underbrace{\overbrace{X \to Y}^{I(X;Y)} \to Z}_{I(X;Z)}$$

If the second channel is noisy then one may expect that this inequality is strict, and further, that the signal decay affect the capabilities of the communication or computation system.

Our objective is therefore to obtain, as a function of the $Y \to Z$ channel alone, a tight upper bound on the ratio $I(X;Z)/I(X;Y)$. Thus the bound is required to hold for every distribution on X and for every form of dependence of Y on X. The desire for an inequality which is true under such a stringent requirement is motivated by the intended application of the inequality: namely inferring the global properties of communication or computation systems from the local properties of their components.

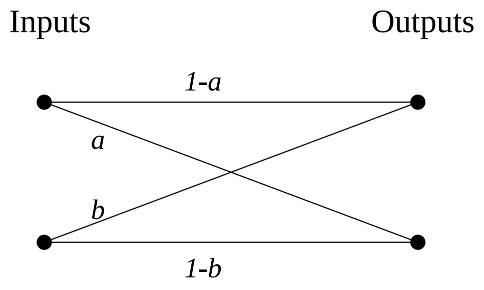
The first inequality of this type on the ratio $I(X;Z)/I(X;Y)$ was derived by Pippenger (for binary symmetric channels) as a key step in his method for showing a lower bound on the depth, and an upper bound on the maximum tolerable component noise, of noisy formulas [16].

In this chapter, we obtain an exact upper bound on the maximum achievable "signal strength ratio" $I(X;Z)/I(X;Y)$, for every binary channel (two input values and two output values). We also obtain bounds on the signal strength ratio when the $Y \to Z$ channel takes two input values to m output values.

Noisy Channel

Before discussing this bound, we more formally define a noisy channel. An *n-input, m-output noisy channel* takes one of n values as input and produces one of m values as output according to a probability distribution which depends on the input value. If the input to a channel is a random variable then the channel will produce a random variable whose distribution (viewed as a vector) is the product of the distribution of the input random variable with a matrix describing the channel. The i,jth entry of the matrix is the probability input i is transformed into output j by the channel. Thus the entries of the matrix are non-negative and the rows sum to 1.

If $n = m = 2$ then the channel is a *binary channel*. See figure 2.1 for an illustration of a binary channel and its corresponding matrix. If the input to the channel is a random variable X with distribution $\boldsymbol{p}_X$ then the channel outputs a random variable Y whose distribution is $\boldsymbol{p}_Y = \boldsymbol{p}_X \cdot A$. In particular, $\boldsymbol{p}_Y(0) = \boldsymbol{p}_X(0)(1-a) + \boldsymbol{p}_X(1)b$ and $\boldsymbol{p}_Y(1) = \boldsymbol{p}_X(0)a + \boldsymbol{p}_X(1)(1-b)$.



$$A = \begin{bmatrix} 1-a & a \\ b & 1-b \end{bmatrix}$$

Figure 2.1: A binary channel and its corresponding row-stochastic matrix.

2.1 Reduction to Weak Signal Case

Our first step in obtaining an upper bound on $I(X;Z)/I(X;Y)$ relies upon a geometric interpretation of mutual information. From the definition of information (see appendix A), we have:

$$I(X;Y) \quad = \quad H(Y) - H(Y|X)$$

$$
\begin{aligned}
&= H(Y) - \sum_x \boldsymbol{p}_X(x) H(Y|X=x) \\
&= h
\end{aligned}
$$

Here h is the altitude marked in figure 2.2; $\boldsymbol{p}_{Y|0}$ and $\boldsymbol{p}_{Y|1}$ are the conditional distributions on Y given $X = 0$ and 1; and $\boldsymbol{p}_Y$ is the unconditioned distribution on Y (i.e. the average of the conditional distributions with the weights $\boldsymbol{p}_X(0)$ and $\boldsymbol{p}_X(1)$). Thus mutual information can be interpreted as a discrete second derivative of the entropy function H.

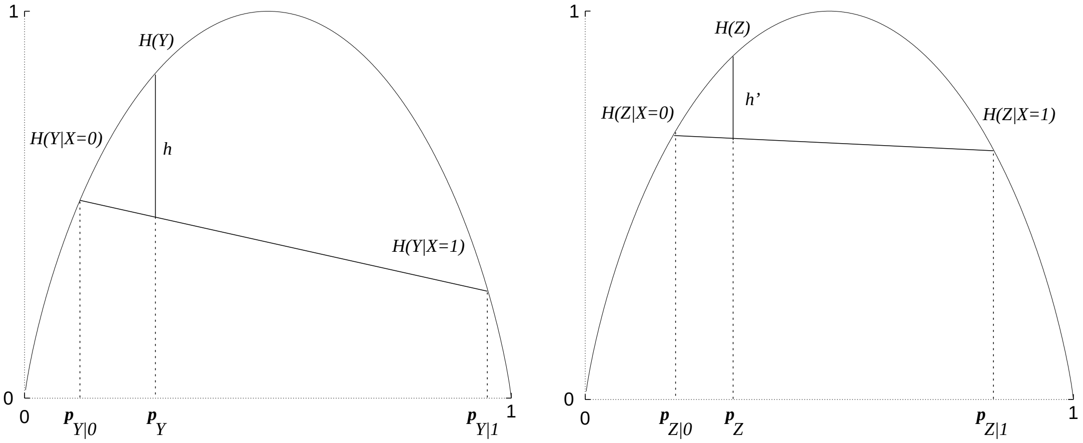


Figure 2.2: Visualization of $I(X;Y)$ and $I(X;Z)$

Now suppose we pass the random variable Y through a channel A and obtain the output Z. For each $x = 0, 1$, the distribution $\boldsymbol{p}_{Z|x}$ equals $\boldsymbol{p}_{Y|x} \cdot A$. Just as for Y, the mutual information $I(X;Z)$ is the discrete second derivative among the points $H(Z|X=0)$, $H(Z)$ and $H(Z|X=1)$. Thus, $I(X;Z)$ is the altitude h' in figure 2.2. Recall that we wish to obtain an upper bound, as a function of the channel A, on the ratio $I(X;Z)/I(X;Y)$. This is equivalent to determining the maximum over all $\boldsymbol{p}_{Y|0}, \boldsymbol{p}_{Y|1}$ and all weights $\boldsymbol{p}_X$, of the ratio h'/h.

We will find the maximum ratio h'/h by explicitly identifying parameters for which it is attained. Our first step in determining these parameters relies on a very general fact about maximizing the ratio between two discrete second derivatives.

For any function f, any two values x, y in the domain of f, and any $p \in [0,1]$, let $f_2(x,y,p) = f(px + (1-p)y) - pf(x) - (1-p)f(y)$ denote the discrete second derivative of f.

Lemma 2.1.1 *For any strictly concave functions f, g on closed and bounded intervals and any $p \in [0,1]$, the ratio*

$$r(x,y) = g_2(x,y,p)/f_2(x,y,p)$$

is maximized in the limit $|x - y| \to 0$.

Proof: Let x^* and y^* be a closest pair of points which achieve the maximum ratio r (say $x^* < y^*$). We obtain a contradiction by finding a closer pair x,y which achieve at least ratio r.

The function g is bounded since it is a continuous function on a closed and bounded interval. Since f and g are strictly concave it follows that $0 < r < \infty$. By suitable affine linear transformations of f and g we can reduce to the case in which the functions are equal at the endpoints (i.e. $f(x^*) = g(x^*)$ and $f(y^*) = g(y^*)$); and we can scale the maximum ratio r to 1 (thus $f(px^* + (1-p)y^*) = g(px^* + (1-p)y^*)$).

We now produce a pair x, y with $|x - y| < |x^* - y^*|$ and $r(x, y) \geq r(x^*, y^*)$. There are two cases. If $g(z) > f(z)$ for some $z \in [x^*, y^*]$ then let x be the greatest value less than z, for which $f(x) = g(x)$; and let y be the least value greater than z, for which $f(y) = g(y)$. Observe that $x \neq y$, and also that at least one of x, y is not at an endpoint x^* or y^*, since by assumption $f(px^* + (1-p)y^*) = g(px^* + (1-p)y^*)$. Further observe that $g(z') > f(z')$ for all $z' \in [x, y]$, and in particular for $z' = px + (1-p)y$. Hence x, y are as desired.

In the other case $g(z) \leq f(z)$ for all $z \in [x^*, y^*]$. Then any pair x, y such that $px + (1-p)y = px^* + (1-p)y^*$ and such that $x^* < x < y < y^*$, will complete the proof. □

We now reexamine the ratio of signal strengths, $I(X;Z)/I(X;Y)$. We find that the fraction of information about X which is preserved going through channel A is maximized for a pair of distributions $\boldsymbol{p}_{Y|X=0}$ and $\boldsymbol{p}_{Y|X=1}$ which are almost indistinguishable:

Corollary 2.1.1 *The ratio $I(X;Z)/I(X;Y)$ is maximized in the limit $|\boldsymbol{p}_{Y|X=0} - \boldsymbol{p}_{Y|X=1}| \rightarrow 0$.*

Recall that $\boldsymbol{p}_{Y|X=0}$ and $\boldsymbol{p}_{Y|X=1}$ correspond to points on the unit interval. The distance function is induced from the interval.

Proof: Fix any weights $\boldsymbol{p}_X(0)$ and $\boldsymbol{p}_X(1)$. Then $I(X;Y)$ and $I(X;Z)$ are the discrete second derivatives of the strictly concave entropy function on $[0,1]$. (See [1] for proof of concavity of entropy.) □

Observe also that unless the channel is either perfectly noiseless or perfectly noisy, that is unless the entries of A are all 0's and 1's, the corollary will hold strictly; which is to say that the maximum ratio is achieved only in the limit of very close distributions. Thus only when it is carrying a very weak signal can a (nontrivial) noisy channel perform at its peak efficiency.

For example suppose we transmit one bit of information over a long cable; and suppose that each meter of the cable introduces some random noise which is symmetric in the sense that it affects 0's and 1's with the same frequency. We will later see that in this symmetric case, the signal strength ratio is maximized when each of the distributions $\boldsymbol{p}_{Y|X=0}$ and $\boldsymbol{p}_{Y|X=1}$ are asymptotically close to the uniform distribution (in which 0's and 1's are equally likely). This is also the distribution each signal eventually approaches as it travels along this cable. Hence the corollary implies that the greatest information loss occurs in the first part of the cable.

For a homogeneous cable this observation could be more simply made by examining powers of the matrix describing the properties of a meter of cable. Our result shows that this is actually a general phenomenon regarding transmission over noisy channels, rather than being a property of multiplication of stochastic matrices.

Another lesson which is suggested by the corollary is that if several signals carry information about an event, one may wish to propagate each signal separately rather than combine the information into a single, clearer signal. Of course the corollary must be applied with care since not every weak signal achieves the minimum loss.

2.2 Reduction to Divergence Ratio

Shannon's mutual information may be written in several ways (see appendix A). One of the most useful is,

$$I(X;Y) = \sum_x \boldsymbol{p}_X(x) \sum_y \boldsymbol{p}_{Y|X=x}(y) \log \frac{\boldsymbol{p}_{Y|X=x}(y)}{\boldsymbol{p}_Y(y)}.$$

The inner sum is known as the *Kullback-Leibler divergence* or simply the *divergence* of $\boldsymbol{p}_{Y|X=x}$ from $\boldsymbol{p}_Y$, and is defined, in general, for distributions $\boldsymbol{p}$ and $\boldsymbol{q}$ as

$$D(\boldsymbol{p}||\boldsymbol{q}) \equiv \sum_x \boldsymbol{p}(x) \log \frac{\boldsymbol{p}(x)}{\boldsymbol{q}(x)}.$$

Divergence is always non-negative and is zero if and only if $\boldsymbol{p}$ and $\boldsymbol{q}$ are identical (see A.2.3). In that sense, it behaves like a distance measure between distributions. However, it is not a true distance measure since it is not symmetric and does not obey the triangle inequality. Nonetheless, we show in the following section that divergence may be interpreted as the square of a distance when the distributions $\boldsymbol{p}$ and $\boldsymbol{q}$ are extremely close.

Our interest in the divergence of infinitesimally separated distributions, comes from our goal of establishing an upper bound on the ratio $r = I(X;Z)/I(X;Y)$. In section 2.1, we showed that r is maximized when the distributions on Y given $X = 0$ and $X = 1$ are infinitesimally separated. This greatly simplifies the task of identifying the distributions which maximize r since instead of having to consider two parameters (specifying the distributions on Y), we can range over just one parameter (specifying one of the distributions), and express r as a series expansion in terms of the distance between the two distributions.

Rather than trying to solve this maximization problem directly, we first show that it is equivalent to maximizing a ratio of divergences. We then show how to interpret these divergences in order to obtain a simple expression for their ratio.

In general, since Z is the output of channel A on input Y,

$$\frac{I(X;Z)}{I(X;Y)} = \frac{\boldsymbol{p}_X(0) D(\boldsymbol{p}_{Y|0} \cdot A||\boldsymbol{p}_Y \cdot A) + \boldsymbol{p}_X(1) D(\boldsymbol{p}_{Y|1} \cdot A||\boldsymbol{p}_Y \cdot A)}{\boldsymbol{p}_X(0) D(\boldsymbol{p}_{Y|0}||\boldsymbol{p}_Y) + \boldsymbol{p}_X(1) D(\boldsymbol{p}_{Y|1}||\boldsymbol{p}_Y)}.$$

If the conditional distributions on Y are infinitesimally separated, we can write $\boldsymbol{p}_Y$ as $\boldsymbol{p}$ and $\boldsymbol{p}_{Y|0}$ as $\boldsymbol{p} + \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} = (\epsilon, -\epsilon)$. (This forces $\boldsymbol{p}_{Y|1} = \boldsymbol{p} - \frac{\boldsymbol{p}_X(0)}{\boldsymbol{p}_X(1)}\boldsymbol{\epsilon}$.) For fixed $\boldsymbol{p}$,

$$\frac{I(X;Z)}{I(X;Y)} = \frac{D((\boldsymbol{p}+\boldsymbol{\epsilon}) \cdot A || \boldsymbol{p} \cdot A)}{D(\boldsymbol{p}+\boldsymbol{\epsilon}||\boldsymbol{p})} + O(\epsilon).$$

In particular, the weights $\boldsymbol{p}_X(0), \boldsymbol{p}_X(1)$ vanish from the problem. Our task reduces to maximizing the constant term in the expansion of $D((\boldsymbol{p}+\boldsymbol{\epsilon}) \cdot A||\boldsymbol{p} \cdot A)/D(\boldsymbol{p}+\boldsymbol{\epsilon}||\boldsymbol{p})$ over all distributions $\boldsymbol{p}$ on $\{0, 1\}$.

2.3 Divergence as Square of L_2 Distance

There is a parameter space in which this maximization problem is addressed most simply, and in which the locus of maximization and value of the maximum, are expressed most naturally. We now give this parameterization and then solve for the maximum.

Let $\boldsymbol{p}$ be a probability distribution on $\{0, 1\}$. Define

$$\boldsymbol{\sigma}(\boldsymbol{p}) = (\sqrt{\boldsymbol{p}(0)}, \sqrt{\boldsymbol{p}(1)}).$$

Geometrically $\boldsymbol{\sigma}$ maps the segment between $(1, 0)$ and $(0, 1)$ in $\Re^2$ (the standard parameterization of the probability distributions) to the quarter circle, centered at the origin, between $(1, 0)$ and $(0, 1)$. See figure 2.3.

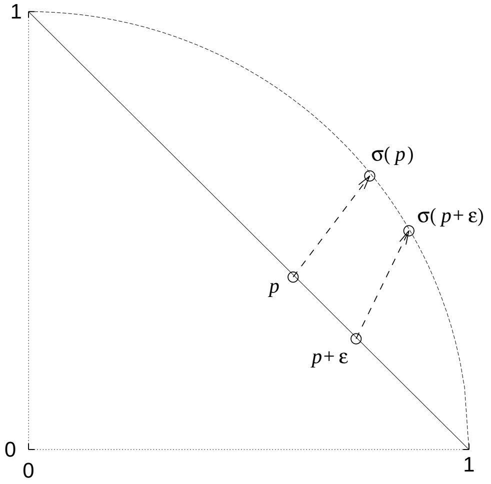


Figure 2.3: The map σ applied to two infinitesimally separated distributions.

Let $\boldsymbol{\epsilon} = (\epsilon, -\epsilon)$ so that both $\boldsymbol{p}$ and $\boldsymbol{p} + \boldsymbol{\epsilon}$ are probability distributions. For sufficiently small ϵ, $D(\boldsymbol{p} + \boldsymbol{\epsilon}||\boldsymbol{p})$ is approximated by a power series expansion in ϵ. The coefficients of this expansion depend on $\boldsymbol{p}$, the probability distribution about which the expansion is being taken. However, after re-parameterization by $\boldsymbol{\sigma}$, the first term of the power series expansion no longer has a dependence on $\boldsymbol{p}$. In fact, the first term is simply proportional to the square of the L_2 distance between the re-parameterized distributions:

Lemma 2.3.1

$$D(\boldsymbol{p}+\boldsymbol{\epsilon}||\boldsymbol{p}) = 2||\boldsymbol{\sigma}(\boldsymbol{p}+\boldsymbol{\epsilon}) - \boldsymbol{\sigma}(\boldsymbol{p})||_2^2 + O(\epsilon^3)$$

Proof: By series expansion. □

There is some intuition for this re-parameterization. It is well known that the divergence $D(\boldsymbol{p}+\boldsymbol{\epsilon}||\boldsymbol{p})$ measures how statistically distinguishable the two distributions $\boldsymbol{p}+\boldsymbol{\epsilon}$ and $\boldsymbol{p}$ are. (E.g. how many coin-tossing trials are required to distinguish a coin with distribution $\boldsymbol{p}+\boldsymbol{\epsilon}$ from one with distribution $\boldsymbol{p}$.) A fixed $\boldsymbol{\epsilon}$ is more significant for a highly biased distribution $\boldsymbol{p}$, than for $\boldsymbol{p}$ near $(1/2, 1/2)$. This is clearest when considering $\boldsymbol{p} = (0, 1)$ and $\boldsymbol{p}+\boldsymbol{\epsilon} = (\epsilon, 1-\epsilon)$, in which case a coin with distribution $\boldsymbol{p}$ will *never* be mistaken for a coin with distribution $\boldsymbol{p}+\boldsymbol{\epsilon}$. The map $\boldsymbol{\sigma}$ stretches the ends of the segment to capture this dependence on $\boldsymbol{p}$ exactly, so that the statistical distinguishability of two nearby distributions is simply captured by their L_2 distance on the circle.

2.4 Signal Decay Theorem

Beyond simplifying the form taken by the divergence, the parameterization of distributions by points on the circle is especially natural for expressing the following theorem.

Theorem 2.4.1 *Let X and Y be Boolean random variables. Let the channel A be*

$$A = \begin{bmatrix} \sin^2\alpha & \cos^2\alpha \\ \sin^2\beta & \cos^2\beta \end{bmatrix}$$

Let Z be the Boolean random variable output by the channel A on input Y. Then

$$\frac{I(X;Z)}{I(X;Y)} \leq \sin^2(\alpha-\beta).$$

Note, $\alpha-\beta$ is the angle (at the origin) between the points on the quarter circle which are the images under the square root map of the most extreme possible distributions on Z.

Proof: As discussed, it suffices to show for any distribution $\boldsymbol{p}$ on Y that

$$\frac{D((\boldsymbol{p}+\boldsymbol{\epsilon})\cdot A||\boldsymbol{p}\cdot A)}{D(\boldsymbol{p}+\boldsymbol{\epsilon}||\boldsymbol{p})} \leq \sin^2(\alpha-\beta) \tag{2.1}$$

where $\boldsymbol{\epsilon} = (\epsilon, -\epsilon)$ and ϵ infinitesimal. The resulting distribution on Z is $\boldsymbol{p}_Z(0) = \boldsymbol{p}(0)\sin^2\alpha + (1-\boldsymbol{p}(0))\sin^2\beta, \boldsymbol{p}_Z(1) = 1 - \boldsymbol{p}(0)\sin^2\alpha - (1-\boldsymbol{p}(0))\sin^2\beta$. Substituting A into the ratio (2.1), and expanding in terms of ϵ, we find that

$$\frac{D((\boldsymbol{p}+\boldsymbol{\epsilon})\cdot A||\boldsymbol{p}\cdot A)}{D(\boldsymbol{p}+\boldsymbol{\epsilon}||\boldsymbol{p})} = (\sin^2\alpha - \sin^2\beta)^2 \frac{\boldsymbol{p}(0)\boldsymbol{p}(1)}{\boldsymbol{p}_Z(0)\boldsymbol{p}_Z(1)} + O(\epsilon).$$

By differentiation one can determine that this expression is maximized for the distribution $\boldsymbol{p}$ specified by

$$\boldsymbol{p} = \left(\frac{\cos\beta\sin\beta}{\cos\beta\sin\beta + \cos\alpha\sin\alpha}, \frac{\cos\alpha\sin\alpha}{\cos\beta\sin\beta + \cos\alpha\sin\alpha}\right).$$

The value of the ratio for this distribution is $\sin^2(\alpha - \beta)$. □

Theorem 2.4.1 also holds under conditioning by certain random variables. In particular, if Q is a random variable such that Z (the output of the channel A) is independent of (Q, X) given Y (the input to the channel) then theorem 2.4.1 holds under conditioning by Q. The requirement that Z be independent of (Q, X) given Y implies that $\boldsymbol{p}_{Z|QXY} = \boldsymbol{p}_{Z|Y}$. In other words, the channel A, which is an expression of $\boldsymbol{p}_{Z|Y}$, remains fixed regardless of the value of Q. This would not be true if, for example, Q equals Z, since in that case, Z is fixed once Q is given, rather than being the output of channel A on input Y.

Corollary 2.4.1 *Let X and Y be Boolean random variables. Let the channel A be*

$$A = \begin{bmatrix} \sin^2\alpha & \cos^2\alpha \\ \sin^2\beta & \cos^2\beta \end{bmatrix}$$

Let Z be the Boolean random variable output by the channel A on input Y. Let Q be a (not necessarily Boolean) random variable such that Z is independent of (Q, X) given Y. Then

$$\frac{I(X;Z|Q)}{I(X;Y|Q)} \leq \sin^2(\alpha - \beta).$$

Proof: Since Z is independent of (Q, X) given Y, $\boldsymbol{p}_{Z|QXY} = \boldsymbol{p}_{Z|Y}$ and thus $\boldsymbol{p}_{Z|Q=q,X=x} = \boldsymbol{p}_{Y|Q=q,X=x} \cdot A$ for all values q and x taken by the random variables Q and X respectively. Therefore the distributions on X, Y, and Z given $Q = q$ satisfy the conditions on the distributions of X, Y, and Z in theorem 2.4.1. It follows from the theorem that

$$\frac{I(X;Z|Q=q)}{I(X;Y|Q=q)} \leq \sin^2(\alpha - \beta)$$

The corollary follows since

$$\frac{I(X;Z|Q)}{I(X;Y|Q)} = \frac{\sum_q \boldsymbol{p}_Q(q) I(X;Z|Q=q)}{\sum_q \boldsymbol{p}_Q(q) I(X;Y|Q=q)} \leq \max_q \frac{I(X;Z|Q=q)}{I(X;Y|Q=q)} \leq \sin^2(\alpha - \beta).$$

□

2.5 Bounds for 2-input, m-output Channels

The previous discussion was limited to binary channels: channels which take two input values and produce two output values. Similar results may be obtained for the case of 2-input to m-output channels. The basic outline of the proof is the same as in the case of binary channels. We prove that the conditional distributions on Y which maximize the ratio $r = I(X;Z)/I(X;Y)$ for a fixed channel are infinitesimally separated. We then express the ratio r as a ratio of divergences. Unfortunately, finding the distribution which maximizes this ratio is not as simple as in the binary channel case, and we do not know a closed form expression for it. However, we do obtain an upper bound on the ratio r which is exact for some channels. For example, for $m = 2$, the bound is identical to the exact bound given in theorem 2.4.1.

Let the channel A be described by the 2 × m matrix,

$$A = \begin{bmatrix} a_1 & a_2 & \cdots & a_m \\ b_1 & b_2 & \cdots & b_m \end{bmatrix}.$$

Let $\boldsymbol{a} = (a_1, \ldots, a_m)$ and $\boldsymbol{b} = (b_1, \ldots, b_m)$. In the case $m = 2$, theorem 2.4.1 upper bounds the ratio r by $\sin^2$ of the angle between $\boldsymbol{\sigma}(\boldsymbol{a})$ and $\boldsymbol{\sigma}(\boldsymbol{b})$ (the images of the rows of A under the square root map). Recall that $\boldsymbol{\sigma}$ denoted the square root map for distributions on $\{0, 1\}$.

It is easy to extend $\boldsymbol{\sigma}$ to cover discrete distributions on m elements. For $\boldsymbol{a} = (a_1, \ldots, a_m)$ a distribution on m elements, let

$$\boldsymbol{\sigma}(\boldsymbol{a}) = (\sqrt{a_1}, \ldots, \sqrt{a_m}).$$

In this section, we show that, for all $m \geq 2$, $\sin^2$ of the angle between $\boldsymbol{\sigma}(\boldsymbol{a})$ and $\boldsymbol{\sigma}(\boldsymbol{b})$ is an upper bound on the ratio r. We will not be able to make the further claim that for any channel, the bound can be achieved in the limit. Such a statement is true in the case $m = 2$ since we explicitly determine the distributions which maximize the ratio. For $m > 2$, the proof does not determine these distributions.

Theorem 2.5.1 *Let X and Y be Boolean random variables. Let the channel A be*

$$A = \begin{bmatrix} a_1 & a_2 & \cdots & a_m \\ b_1 & b_2 & \cdots & b_m \end{bmatrix}$$

Let Z be the Boolean random variable output by the channel A on input Y. Then

$$\frac{I(X;Z)}{I(X;Y)} \leq 1 - \left(\sum_{i=1}^{m} \sqrt{a_i b_i}\right)^2.$$

Note that $1 - (\sum_{i=1}^{m} \sqrt{a_i b_i})^2$ is $\sin^2$ of the angle between $\boldsymbol{\sigma}(\boldsymbol{a})$ and $\boldsymbol{\sigma}(\boldsymbol{b})$.

Proof: The first step of the proof is to show that the conditional distributions on Y which maximize $r = I(X;Z)/I(X;Y)$ are infinitesimally separated. In particular, we desire an extension of corollary 2.1.1 to the case when Z is an m valued random variable. As in the proof of that corollary, for fixed weights $\boldsymbol{p}_X(0)$ and $\boldsymbol{p}_X(1)$, $I(X;Y)$ is the discrete second derivative of the entropy function on $[0,1]$. Since Z is the output of channel A on input Y, the distributions $\boldsymbol{p}_{Z|0}$ and $\boldsymbol{p}_{Z|1}$ lie on the line in $\Re^m$ between $(0,1)\cdot A$ and $(1,0)\cdot A$. Thus $I(X;Z)$ is the discrete second derivative of the entropy function of m valued distributions restricted to this line. Since entropy is a strictly concave function, this restriction is strictly concave and the extension follows from lemma 2.1.1.

Restriction to the case of infinitesimally separated distributions allows us to write the ratio r for fixed $\boldsymbol{p}$ as,

$$\frac{I(X;Z)}{I(X;Y)} = \frac{D((\boldsymbol{p}+\boldsymbol{\epsilon})\cdot A||\boldsymbol{p}\cdot A)}{D(\boldsymbol{p}+\boldsymbol{\epsilon}||\boldsymbol{p})} + O(\epsilon).$$

By series expansion,

$$\frac{D((\boldsymbol{p}+\boldsymbol{\epsilon})\cdot A||\boldsymbol{p}\cdot A)}{D(\boldsymbol{p}+\boldsymbol{\epsilon}||\boldsymbol{p})} = \boldsymbol{p}(0)\boldsymbol{p}(1)\sum_{i=1}^{m}\frac{(a_i-b_i)^2}{\boldsymbol{p}(0)a_i+\boldsymbol{p}(1)b_i} + O(\epsilon).$$

It remains to prove that

$$\boldsymbol{p}(0)\boldsymbol{p}(1)\sum_{i=1}^{m}\frac{(a_i-b_i)^2}{\boldsymbol{p}(0)a_i+\boldsymbol{p}(1)b_i} \le 1 - \left(\sum_{i=1}^{m}\sqrt{a_ib_i}\right)^2.$$

This follows from the concavity of the square root function[1]. After some simple manipulation, proving the inequality reduces to proving

$$\sum_{i=1}^{m}\sqrt{a_ib_i} \le \left(\sum_{i=1}^{m}\frac{a_ib_i}{\boldsymbol{p}(0)a_i+\boldsymbol{p}(1)b_i}\right)^{1/2}.$$

Let $\lambda_i = \boldsymbol{p}(0)a_i + \boldsymbol{p}(1)b_i$ and $x_i = a_ib_i/\lambda_i^2$. Note that $\sum_i \lambda_i = 1$ and $0 \le x_i < \infty$. Since, square root is a concave function,

$$\sum_{i=1}^{m}\lambda_i\sqrt{x_i} \le \sqrt{\sum_{i=1}^{m}\lambda_i x_i}$$

and the proof is complete. □

As in the case of the binary channel, this result also holds under conditioning.

Corollary 2.5.1 *Let X and Y be Boolean random variables. Let the channel A be*

$$A = \begin{bmatrix} a_1 & a_2 & \cdots & a_m \\ b_1 & b_2 & \cdots & b_m \end{bmatrix}$$

[1] Thanks to Simon C. Harris, Trinity College, Cambridge for pointing out this fact.

Let Z be the Boolean random variable output by the channel A on input Y. Let Q be a (not necessarily Boolean) random variable such that Z is independent of (Q, X) given Y. Then

$$\frac{I(X;Z|Q)}{I(X;Y|Q)} \leq 1 - \left(\sum_{i=1}^{m} \sqrt{a_i b_i} \right)^2 .$$

Proof: Use the same argument as in the proof of corollary 2.4.1. □

Chapter 3

Noisy Circuit Depth

Among the fundamental concerns in computation are the depth and size of circuits required to compute Boolean functions. The depth of circuits, in particular, measures latency of computation. This is of critical importance in circuits for real-time computation (e.g. the FFT); and it is central to the study of parallel complexity classes.

In view of the limitations of physical circuits, von Neumann asked whether circuits with noisy components can compute the same functions as circuits with reliable gates; and if so, at what cost in latency (depth)? He provided the following positive, but qualified, response to this question: Every circuit with noiseless gates can be reliably simulated by a circuit with noisy gates, whose depth is at most a constant[1] times the depth of the original circuit, provided that the probability of error in each component of the circuit is no more than some threshold $< 1/2$.

This answer has two especially interesting features. The first is the threshold on component failure, above which the construction fails. The second is that the construction requires a slow-down (i.e. increase in depth) by a factor strictly greater than 1. For a long time it was not known whether these features were necessary, or were artifacts of von Neumann's construction. Finally, Pippenger showed through an elegant information-theoretic argument that both features were necessary, at least in noisy formulas (circuits in which the output of each gate is the input to at most one other gate) [16]. Shortly afterward Feder extended Pippenger's bound to general noisy circuits [7].

Surprisingly, Feder's extension to circuits (which obtained the same threshold and factor of depth increase bounds as Pippenger) did not argue in terms of mutual information. Thus, Pippenger's use of information theoretic techniques to attack this problem seemed extravagant. In fact, an argument based on the L_1 distance between conditional distributions is sufficiently precise to obtain Pippenger's result. We discuss the use of the L_1 distance in greater detail in appendix B since some of the results obtained therein are used in later chapters.

[1] Von Neumann's original construction implies a depth increase by a factor of 2 for $\epsilon < 0.0073$ when computation is by 3-input majority gates.

The first indication that information theoretic techniques were not extravagant and could provide extremely precise bounds on noisy computation, came in our work on lower bounding the depth of noisy formula [6]. In this chapter, we discuss our previous result and extend it to general circuits. We increase the lower bound on the factor of depth increase; and decrease the upper bound on the noise threshold. Thus we improve on the results of both Pippenger and Feder, and offer some justification for von Neumann's desire to use information theoretic techniques in the study of noisy computation.

The key to our improved depth lower bound is the signal decay theorem (theorem 2.4.1) and its conditional version (corollary 2.4.1). These results give precise bounds on the fraction of information which can cross a noisy channel. We first give an informal outline of the proof and explain the role of the results from chapter 2. We then state and prove the main theorem of this chapter.

3.1 Intuition

At a high level, the application of our information theoretic analysis to the lower bound on circuit depth follows the outline of Pippenger's argument which, very briefly, has the following structure.

For each input bit X upon which the function depends there is a setting of the other inputs so that the function is X (or $\overline{X}$, the complement of X). A reliable circuit for the function, with this setting of the inputs, must output a value which is highly correlated with X. In other words, if X is a random variable then the mutual information carried by the output about X must be high.

On the other hand, the amount of information the input X can "send" to the output is restricted by the structure of the intervening noisy circuit. In particular, the amount of information is bounded by the sum over all paths from X to the output of a quantity that is exponentially small in the length of the path. To establish this for formulas, Pippenger first showed that the total information sent is bounded by the sum of the information sent over each path from X to the output. This supports the view of information as a kind of fluid which flows from the input X to the output along the wires of the formula. At each gate, several paths combine, but the fluid flowing out of the gate is no more than the sum of the fluid flowing in. In a formula, this fact follows from the fact that the inputs to the gate are mutually independent given X (see the sub-additivity of information in appendix A). In circuits, information may not be sub-additive at a gate, which ruins the interpretation of information as a fluid. Thus we cannot decompose a circuit into a set of disjoint paths as Pippenger did in the case of formulas. In the next section, we present a new technique which overcomes this difficulty.

Once the formula was decomposed into a set of disjoint paths, Pippenger bounded the information carried by each path by an amount which is exponentially small in the path length. Let Y be the pre-noise output of some gate on a path from X to the circuit's output and let Z be the output after the random noise. (Z equals Y with probability $1 - \epsilon = (1 + \xi)/2$, and $\overline{Y}$ with probability $\epsilon = (1 - \xi)/2$.) The information carried by the random variable Y about the input X drops by a certain factor when Y is affected by random noise. We need a bound on this factor. In particular, we need the ratio $I(X;Z)/I(X;Y)$ to be bounded by a function of the noise bias ξ. Pippenger shows that $I(X;Z)/I(X;Y) \leq \xi$. We use the results from chapter 2 to show that this ratio is bounded by ξ^2.

The improved bound on $I(X;Z)/I(X;Y)$ follows from theorem 2.4.1. We model the noisy dependence of Z on Y by a binary communication channel which outputs Z upon input Y. For von Neumann's model of noisy components, the channel is symmetric with noise $(1 - \xi)/2$ (i.e. $\cos^2 \alpha = \sin^2 \beta = (1 - \xi)/2$). Theorem 2.4.1 implies that $I(X;Z)/I(X;Y) \leq \xi^2$. This improves strictly on Pippenger's corresponding bound of ξ for all channels (except, of course, that it is identical for perfect or totally noisy channels). See figure 3.1. The distinction has its greatest significance for relatively noisy channels. For, Pippenger's bound grows linearly as the channel is perturbed away from a totally noisy channel ($\xi = 0$); whereas our bound has a quadratic basin about that point. Such a result should be anticipated on qualitative grounds, since the ratio we are studying is a smooth, nonnegative function of ξ and is equal to zero for the totally noisy channel.

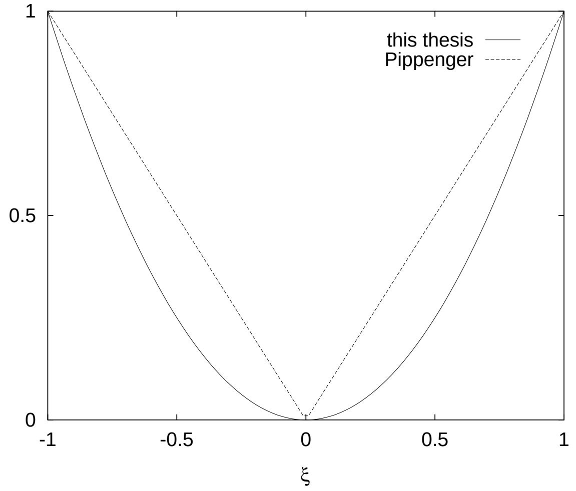


Figure 3.1: Upper bounds on $I(X;Z)/I(X;Y)$ for symmetric channels with noise $(1 - \xi)/2$.

3.2 Circuit Anomalies

In a circuit, it is not necessarily true that the total information sent by an input to the output is bounded by the sum of the information sent over the paths. Such a statement requires two inequalities to hold. One of the inequalities is the data processing inequality which states that $I(Y;X) \leq I(Y_1, \ldots, Y_k; X)$ where $Y_1, \ldots, Y_k$ are the inputs to a gate with pre-noise output Y. This is also true for circuits. The other is that $I(Y_1, \ldots, Y_k; X) \leq \sum_i I(Y_i; X)$. This is true if the Y_i are mutually independent given X, but in a circuit Y_i and Y_j $(i \neq j)$ may be dependent given X (if they share some common noise source; e.g. if they are the output of the same noisy gate). In fact, in a circuit, $I(Y_1, \ldots, Y_k; X)$ may be greater than $\sum_i I(Y_i; X)$. Thus the method of decomposing the circuit into a set of disjoint paths while not decreasing the information between input and output, which works in the case of formulas, seems unlikely to succeed.

Rather than going through the intermediate step of decomposing the circuit into disjoint paths, we directly bound the information between any set of wires (taken together) and the input X by the sum of $\xi^{2|P|}$ over all paths P from X to these wires. This establishes that the total information sent by an input to the output is bounded by the sum over all paths P from that input to the output of $\xi^{2|P|}$. Unfortunately, it makes the argument for the ξ^2 drop in information at every noisy gate more complicated. We use the conditioned version of the signal decay theorem 2.4.1 to handle this added complexity.

We first prove this lemma and then show how to use it to obtain a lower bound on circuit depth.

Lemma 3.2.1 *Let G be a circuit composed of $(1-\xi)/2$-noisy gates. Suppose each input to G is X (a Boolean random variable) or a constant. Let W be the vector of random values carried by a set of wires in G. Then*

$$I(W; X) \leq \sum_{P \text{ from } X \text{ to } W} \xi^{2|P|}$$

where the sum is over paths P in G from input X to wires in W, and $|P|$ is the number of gates on the path P.

Proof: View G as a directed acyclic graph whose vertices are gates or the inputs to the circuit (X and constants 0 and 1), and whose edges are wires. Direct a wire (edge) from vertex h to g if the output of h is the input of g. The wire which is the output of the circuit is a special edge which is directed from its one endpoint. Number the gate vertices distinctly from 1 to the number of gates in G and number the input vertices 0 so that each wire starts from its smaller numbered endpoint. Such a numbering is possible since G is acyclic. Number the wires with the number of their smaller numbered endpoint.

The proof is by induction on the number of the highest numbered wire in W. If the highest numbered wire has number 0 then the edges in W carry a combination of constant values and X. If W contains a wire with value X then $I(W; X) = 1$ and there is at least one wire which originates at X, i.e. one path of length 0 from X to wires in W. If all the wires in W are constant then $I(W; X) = 0$ and there are no paths from X to wires in W. In either case,

$$I(W; X) \leq \sum_{P \text{ from } X \text{ to } W} \xi^{2|P|}.$$

Assume the lemma holds for all W which contain wires numbered $\leq t$. Consider W which contains wires numbered $\leq t + 1$. Let Z be the binary random value carried by the wires numbered $t+1$ in W. Since each gate vertex has a distinct number and noise occurs at the gate, Z is well defined. Let $W_1, \ldots, W_m$ be the wires in W numbered $\leq t$. Since all wires numbered $t+1$ carry the same value, $I(W;X) = I(Z, W_1, \ldots, W_m; X)$. From the definition of information,

$$I(Z, W_1, \ldots, W_m; X) = I(Z; X | W_1, \ldots, W_m) + I(W_1, \ldots, W_m; X).$$

Let Y be the pre-noise output of gate $t+1$. The output Z of gate $t+1$ is the result of passing Y through a symmetric channel with noise $(1-\xi)/2$. The input X and the values $W_1, \ldots, W_m$, since they are the output of gates numbered $\leq t$, are independent of Z given Y. Thus corollary 2.4.1 implies

$$I(Z; X | W_1, \ldots, W_m) \leq \xi^2 I(Y; X | W_1, \ldots, W_m).$$

Let $Y_1, \ldots, Y_k$ be the inputs to gate $t+1$. By the data processing inequality,

$$I(Y; X | W_1, \ldots, W_m) \leq I(Y_1, \ldots, Y_k; X | W_1, \ldots, W_m).$$

Therefore,

$$\begin{aligned} I(Z, W_1, \ldots, W_m; X) &\leq \xi^2 I(Y_1, \ldots, Y_k; X | W_1, \ldots, W_m) + I(W_1, \ldots, W_m; X) \\ &= \xi^2 I(Y_1, \ldots, Y_k, W_1, \ldots, W_m; X) + (1-\xi^2) I(W_1, \ldots, W_m; X) \end{aligned}$$

Since $Y_1, \ldots, Y_k$ are inputs to gate $t+1$, they are wires with numbers $\leq t$. Thus we can apply the inductive hypothesis to both terms to obtain,

$$\begin{aligned} I(Z, W_1, \ldots, W_m; X) &\leq \xi^2 \sum_{\substack{P \text{ from } X \text{ to} \\ \{Y_1, \ldots, Y_k, W_1, \ldots, W_m\}}} \xi^{2|P|} + (1-\xi^2) \sum_{\substack{P \text{ from } X \text{ to} \\ \{W_1, \ldots, W_m\}}} \xi^{2|P|} \\ &\leq \sum_{\substack{P \text{ from } X \text{ to} \\ \{Z, W_1, \ldots, W_m\}}} \xi^{2|P|} \\ &\leq \sum_{P \text{ from } X \text{ to } W} \xi^{2|P|} \end{aligned}$$

□

3.3 Noisy Circuit Depth Lower Bound

Let f be a Boolean function which depends on n arguments.[2] We show that any circuit C which computes the function f with high probability using noisy k-input gates must have depth at least $R \log_k n$ with a certain $R > 1$ depending on k and the noise level. This implies a lower

[2] A function depends on an argument x if there exists a setting of the other arguments such that the function restricted to that setting is not a constant.

bound on the factor by which the depth of a circuit must increase when going from the perfect to the noisy gate model. In particular, suppose there exists a gate which computes a function g that depends on k inputs, and no gate that depends on more than k inputs. The function f which is the d-fold composition of g depends on $n = k^d$ inputs and can be computed by a depth d circuit in the perfect gate model. Feder's result implies that in the noisy circuit model, the depth must be at least Rd where $R = 1/\log_k(k\xi)$. Our result implies $R = 1/\log_k(k\xi^2)$.

Theorem 3.3.1 *Let f be a function which depends on n inputs. Let C be a circuit of depth c using gates with at most k inputs, where each gate fails independently with probability $(1-\xi)/2$. Suppose C $(1-\delta)$-reliably computes the function f where $\delta < 1/2$. Let $\Delta = 1+\delta\log\delta+(1-\delta)\log(1-\delta)$.*

- *If $\xi^2 > 1/k$ then $c \geq \frac{\log(n\Delta)}{\log(k\xi^2)}$*
- *If $\xi^2 \leq 1/k$ then $n \leq 1/\Delta$*

Proof: Let $x_1, \ldots, x_n$ be the inputs to the function f. Since f depends on all inputs, for each input x_i there exists a setting of the other $n-1$ inputs so that f is either the function x_i or $\overline{x_i}$. Let C_i be the circuit C with this setting for the $n-1$ inputs other than x_i. Let X be a Boolean random variable uniformly distributed over $\{0,1\}$. Let $C_i(X)$ be the random variable which is the output of C_i when $x_i = X$. By Fano's inequality (see theorem A.2.6),

$$I(C_i(X); X) \geq \Delta. \tag{3.1}$$

In other words, since $C_i(X)$ and X are correlated, the mutual information between them is large.

We apply lemma 3.2.1 with $G = C_i$ and $W = C_i(X)$ to obtain the upper bound

$$I(C_i(X); X) \leq \sum_{P \in C_i} \xi^{2|P|} \tag{3.2}$$

where the sum is over paths in C from x_i to C's output.

Combining the bounds (3.1) and (3.2) and summing over all C_i gives

$$n\Delta \leq \sum_{P \in C} \xi^{2|P|}. \tag{3.3}$$

The first result of the theorem follows easily from the following lemma.

Lemma 3.3.1 *For all circuits C of depth c that are composed of k-input gates, if $\xi^2 > 1/k$ then*

$$\sum_{P \in C} \xi^{2|P|} \leq k^c \xi^{2c}$$

where the sum is over paths in C from C's inputs to C's output.

Proof: It suffices to show that when $\xi^2 > 1/k$, the expression $\sum_{P \in C} \xi^{2|P|}$ is maximized for C equal to the complete k-ary tree of depth c, since this tree has $\sum_P \xi^{2|P|} = k^c \xi^{2c}$.

If C is not a tree then by duplicating any gate with multiple outputs, we can change C into a tree without affecting the number or length of paths. We can thus assume that C is a tree. If C is not complete then some vertex v at depth $l < c$ has fewer than k children. If v is not a leaf then adding a child to v increases the sum over paths by $\xi^{2(l+1)}$. If v is a leaf then adding k children to v increases the sum by $k\xi^{2(l+1)} - \xi^{2l}$ which is strictly positive since $\xi^2 > 1/k$. □

Combining the result of lemma 3.3.1 with (3.3), we obtain

$$n\Delta \leq k^c \xi^{2c}$$

which implies the first result of the theorem.

For the second result, notice that every gate increases the number of paths from inputs to output. However, it also increases the distance (path length) from its inputs to the output. If the gate is too noisy, the additional paths it provides will not compensate for the loss in signal clarity. Eventually, the output will bear little relation to the inputs. Thus, there is a threshold on the noisiness of the gates. Above this threshold, gates are too noisy to allow reliable computation and we cannot compute functions of an arbitrary number of inputs.

In order to calculate this threshold, we first claim that there exists $1 \leq i \leq n$ such that

$$\sum_{P \in C_i} 1/k^{|P|} \leq 1/n$$

where the sum is over paths in C from x_i to C's output. The claim follows by an averaging argument and the fact that $\sum_{P \in C} 1/k^{|P|} \leq 1$ (the Kraft inequality, which can be proven by induction).

Combining (3.1) and (3.2) with the above claim, for $\xi^2 \leq 1/k$, we obtain

$$\Delta \leq \sum_{P \in C_i} \xi^{2|P|} \leq \sum_{P \in C_i} 1/k^{|P|} \leq 1/n$$

which implies the second result of the theorem. □

Our result improves on the results of Pippenger and Feder in two ways. First, we increase the lower bound on the threshold below which computation in the noisy gate model is infeasible. This will be discussed further in chapter 5.

Second, we increase the factor by which the depth of the reliable circuit must increase. To compute a function which depends on n inputs, Feder shows that a reliable circuit must have depth greater than $\log n$ by at least a factor $1/\log(k\xi)$. Our result is that this factor must be at least $1/\log(k\xi^2)$. Since there exist functions which can be computed in $\log_k n$ depth using k-input gates, our result implies that the reliable circuit depth must be at least a factor $1/\log_k(k\xi^2)$ greater than the noiseless circuit depth. See figure 3.2.

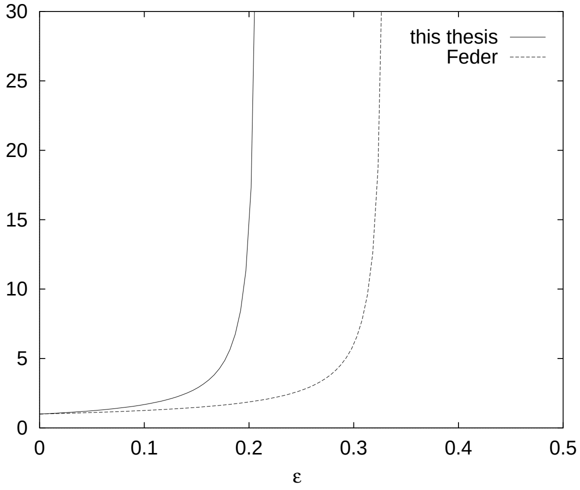


Figure 3.2: Lower bounds on factor of increase in circuit depth using 3-input, ϵ-noisy gates.

Von Neumann's original construction implies a depth increase by a factor of 2 for $\epsilon < 0.0073$ when computation is performed by 3-input majority gates. Pippenger [16] provides a more careful analysis of von Neumann's method and shows that, for computation by 3-input parity gates, as $\epsilon \to 0$, the factor is asymptotic to $1/(1 - 2/\log_3(1/\epsilon))$. Our result implies that this factor is at least $1/(1 - 2\log_3(1/\xi))$ which is asymptotic to $4\epsilon/\ln 3$.

Von Neumann's construction uses majority voting to reduce error. For a thorough discussion of error correction by majority voting, see Pippenger [17].

Our depth bounds can be easily extended to the case of asymmetric noise in which a component fails with probability that depends on its pre-noise output. The model allows the possibility that a gate may fail with a different probability when producing a 0 than when producing a 1. In this case, if Y is the pre-noise output of the gate then the noisy output Z of the gate is the output of an arbitrary binary channel on input Y. Let A denote the binary channel between Y and Z. In this model, a noisy gate for computing a function g calculates g correctly and then transmits the result across the noisy channel A to produce an output. Theorem 2.4.1 bounds the fraction of information preserved in crossing this more general channel. All that is needed is to replace the bound ξ^2 with this new bound to obtain the generalization.

Theorem 3.3.2 *Let f be a function which depends on n inputs. Let C be a circuit of depth c using gates with at most k inputs, where each gate fails independently with asymmetric noise described by the channel*

$$A = \begin{bmatrix} \sin^2\alpha & \cos^2\alpha \\ \sin^2\beta & \cos^2\beta \end{bmatrix}$$

Suppose C $(1-\delta)$-reliably computes the function f where $\delta < 1/2$. Let $\Delta = 1 + \delta\log\delta + (1-\delta)\log(1-\delta)$.

- *If* $\sin^2(\alpha - \beta) > 1/k$ *then* $c \geq \frac{\log(n\Delta)}{\log(k \sin^2(\alpha-\beta))}$
- *If* $\sin^2(\alpha - \beta) \leq 1/k$ *then* $n \leq 1/\Delta$

Proof: The proof is identical to the proof of theorem 3.3.1 with the bound $\sin^2(\alpha - \beta)$ (from theorem 2.4.1) replacing ξ^2. □

Chapter 4

Noisy Circuit Size

The size of a circuit, i.e. the number of components in it, is roughly a measure of the size or cost of it's implementation. In order to construct a circuit to compute a particular function, one would expect to be forced to use more noisy components than noiseless ones. We will show that this is indeed the case for certain functions.

Noisy circuit size has received somewhat more attention than noisy circuit depth. Von Neumann planted the seeds of this investigation by heuristically bounding the redundancy required for computation using noisy components [28]. In 1977, Dobrushin and Ortyukov refined von Neumann's method to prove that for all functions there exist noisy circuits whose size is $O(c \log c)$ where c is the size of a noiseless circuit for the function [3]. Pippenger strengthened this result by exhibiting an explicit construction of the noisy circuit and extending the result to less powerful models of noisy computation [15].

Accompanying these positive upper bounds are lower bounds of the same general form. In particular, Dobrushin and Ortyukov claimed that $\Omega(s \log s)$ noisy circuit size is necessary to compute functions reliably where s is the sensitivity[1] of the function. Pippenger *et. al.* showed that this work contains serious flaws [20]. By a different argument, they were able to prove that noisy circuits computing the parity of n bits require $\Omega(n \log n)$ size. Gál [9] and Reischuk and Schmeltz [22] successfully reproved Dobrushin and Ortyukov's original claim, that all functions with sensitivity s require $\Omega(s \log s)$ size noisy circuits. This statement treats as constants the noise of each component, the reliability requirement of the noisy circuit, and the complexity of the basis. In this chapter, we improve the dependence of this lower bound on these factors, and obtain a bound which grows to infinity as the noise ϵ of the components approaches $1/2$.

In order to lower bound the size of the circuit, we lower bound the number of wires which lead to the circuit from the inputs. Since each gate has a bounded number of inputs, this implies a bound on the number of gates. Intuitively, because the gates are noisy, the circuit must

[1] A function's sensitivity is the maximum (over all input vectors) of the number of bits which change the function value when flipped individually.

have several wires to each input bit in order to gain enough knowledge about the input to compute the function reliably. From the previous section, it would seem that an ideal measure of knowledge is Shannon's mutual information. However, in this case a simpler proof and a slightly better result can be obtained using L_1 distance. We present the proof based on L_1 distance in this chapter and include the argument based on mutual information in appendix C.

4.1 Noisy Input Wires

We typically view an ϵ-noisy gate as computing a Boolean function g (correctly) and then complementing the result with probability ϵ. Noise, in this view, is a single random event which is independent of the input to the gate. This is the view taken by von Neumann when describing his noisy gate model. It captures the defining property of an ϵ-noisy, k-input gate that computes g:

Property 4.1.1 *For each input* $\boldsymbol{x} \in \{0,1\}^k$, *the gate outputs* $g(\boldsymbol{x})$ *with probability* $1-\epsilon$.

We found it convenient to adopt this view of a noisy gate when proving bounds on noisy circuit depth. However, to bound noisy circuit size, we adopt a different view first proposed by Dobrushin and Ortyukov in [2]. This is only a conceptual aid. The model is still that of the ϵ-noisy gate defined by property 4.1.1.

In this new view, rather than the output wire failing with probability ϵ, each input wire fails independently with probability ω. The output of the gate is then computed based on the noisy version of the input. In particular, an input $\boldsymbol{x}$ becomes $\boldsymbol{y}$ with probability

$$A(\boldsymbol{x},\boldsymbol{y}) = \omega^d(1-\omega)^{k-d}$$

where d is the Hamming distance between $\boldsymbol{x}$ and $\boldsymbol{y}$.[2] The gate then outputs 0 with probability $\boldsymbol{\psi}(\boldsymbol{y})$ and 1 with probability $1-\boldsymbol{\psi}(\boldsymbol{y})$ where $\boldsymbol{\psi}$ is a function characterizing the gate. It is worth emphasizing that given $\boldsymbol{y}$, the output of the gate is independent of $\boldsymbol{x}$.

The function $\boldsymbol{\psi}$ must be chosen so that the gate obeys property 4.1.1. If ω is too large, it may not be possible to find such a $\boldsymbol{\psi}$ and in this case the ϵ-noisy gate cannot be viewed as having ω-noisy input wires.

To calculate the error on a particular input, treat the function $\boldsymbol{\psi}$ as a 2^k dimensional column vector in $[0,1]^{2^k}$, and $A(\boldsymbol{x},\boldsymbol{y})$ as a $2^k \times 2^k$ matrix. The error on input $\boldsymbol{x}$ is the $\boldsymbol{x}$th component of the matrix, vector product $A \cdot \boldsymbol{\psi}$ which we denote $(A \cdot \boldsymbol{\psi})(\boldsymbol{x})$. With this notation, the condition that the gate be ϵ-noisy is that, for all $\boldsymbol{x} \in \{0,1\}^k$,

$$\boldsymbol{\psi}(\boldsymbol{x}) \in [0,1] \quad \text{and} \quad (A \cdot \boldsymbol{\psi})(\boldsymbol{x}) = \begin{cases} 1-\epsilon & \text{if } g(\boldsymbol{x}) = 0 \\ \epsilon & \text{if } g(\boldsymbol{x}) = 1 \end{cases} \tag{4.1}$$

[2] The Hamming distance between x and y is the number of bit positions in which x and y differ.

Our goal is to find the values of ω for which there is a $\boldsymbol{\psi}$ that satisfies 4.1. In other words, how noisy can we view the input wires of an ϵ-noisy gate?

The following lemma answers this question. It improves lemma 3.1 from [2] which required $\omega \in [0, \epsilon/k]$. As in chapter 3, let ξ be twice the bias of ϵ from 1/2 (i.e. $\xi = 1 - 2\epsilon$).

Lemma 4.1.1 *For every Boolean function* $g : \{0,1\}^k \mapsto \{0,1\}$, $\epsilon \in [0, 1/2]$, *and* $\omega \in [0, \frac{1-\sqrt[k]{\xi}}{2}]$ *there exists an* ϵ*-noisy gate with* ω*-noisy input wires that computes* g.

Proof: Fix k and $\epsilon \in [0, 1/2]$. We show that for any $\omega \in [0, \frac{1-\sqrt[k]{\xi}}{2}]$, and any Boolean function g on k inputs, there exists a function $\boldsymbol{\psi}$ satisfying condition 4.1. If $\epsilon = 1/2$ then the output of the gate is a fair coin flip, independent of the input, and the function $\boldsymbol{\psi} = 1/2$ satisfies condition 4.1 for any $\omega \in [0, 1/2]$.

If $\epsilon < 1/2$, we show the existence of the function $\boldsymbol{\psi}$ by treating condition 4.1 as a system of linear equations. We show that $\boldsymbol{\psi}$ is uniquely defined by this system and that for $\omega \in [0, \frac{1-\sqrt[k]{\xi}}{2}]$, $\boldsymbol{\psi}(\boldsymbol{x})$ is a probability (i.e. lies in [0, 1]) for all $\boldsymbol{x}$.

The matrix A can be written as the tensor product of a 2 × 2 matrix with itself k times.

$$A = \begin{pmatrix} 1-\omega & \omega \\ \omega & 1-\omega \end{pmatrix}^{\otimes k}.$$

The inverse of A exists and is easy to calculate from its tensor form,

$$A^{-1} = \frac{1}{(1-2\omega)^k} \begin{pmatrix} 1-\omega & -\omega \\ -\omega & 1-\omega \end{pmatrix}^{\otimes k}.$$

From condition 4.1, the function $\boldsymbol{\psi}$ (treated as a vector) is $A^{-1} \cdot (\overline{\frac{1}{2}} + \boldsymbol{\phi})$ where $\overline{\frac{1}{2}}$ is the 2^k dimensional column vector of all $1/2$'s and $\boldsymbol{\phi} \in \{+\xi/2, -\xi/2\}^{2^k}$ (for $\xi = 1 - 2\epsilon$) corresponds to the Boolean function g:

$$\boldsymbol{\phi}(\boldsymbol{x}) = \begin{cases} +\xi/2 & \text{if } g(\boldsymbol{x}) = 0 \\ -\xi/2 & \text{if } g(\boldsymbol{x}) = 1 \end{cases}$$

To complete the proof, we must show that for all Boolean functions g (i.e. all vectors $\boldsymbol{\phi}$), the value $\boldsymbol{\psi}(\boldsymbol{x})$ lies in $[0, 1]$ for all $\boldsymbol{x}$. Since $\overline{\frac{1}{2}}$ is an eigenvector of A^{-1} with eigenvalue 1, $A^{-1} \cdot (\overline{\frac{1}{2}} + \boldsymbol{\phi}) = \overline{\frac{1}{2}} + A^{-1} \cdot \boldsymbol{\phi}$. Thus it remains to prove that

$$\max_{\boldsymbol{\phi} \in \{+\xi/2, -\xi/2\}^{2^k}} \max_{\boldsymbol{x} \in \{0,1\}^k} |(A^{-1} \cdot \boldsymbol{\phi})(\boldsymbol{x})| \leq 1/2.$$

A row of A^{-1} is a permutation of the 2^k terms in the binomial expansion $((1-\omega) - \omega)^k$ each weighted by $(1 - 2\omega)^{-k}$. The sum of the absolute value of the entries in a row is $(1 -$

$2\omega)^{-k}((1-\omega)+\omega)^k = (1-2\omega)^{-k}$. Since the magnitude of the entries in $\boldsymbol{\phi}$ is $\xi/2$, the maximum of $|(A^{-1}\cdot\boldsymbol{\phi})(\boldsymbol{x})|$ is $\xi(1-2\omega)^{-k}/2$. Thus $\boldsymbol{\psi}(\boldsymbol{x}) \in [0,1]$ if $\xi(1-2\omega)^{-k}/2 \leq 1/2$ or equivalently

$$\omega \leq \frac{1-\xi^{1/k}}{2}$$

□

4.2 Lower Bound on Noisy Circuit Size

Assume that k is an upper bound on the number of inputs to a gate in the basis. We lower bound the size of a $(1-\delta)$-reliable circuit by lower bounding the number of wires from the external inputs into the circuit. Since each gate in the circuit has at most k inputs, a lower bound on the number of wires translates directly into a lower bound on the number of gates. The result of the previous section allows us to think of the wires as failing independently with probability ω. Thus, we can view the noisy circuit as a probabilistic function which obtains a noisy version of the true input. Intuitively, the circuit must obtain several samples of each input bit in order to correctly compute the function. If the circuit fails to take sufficiently many samples of an input bit, it will be unable to distinguish the input in which the bit is set to 1 from that in which the bit is 0. If, in addition, the function is different on these two inputs then the circuit fails to correctly compute the function on at least one of the two inputs.

4.2.1 Sensitivity

To make this intuition precise, we define the sensitivity of a Boolean function f which takes input $\boldsymbol{x} = x_1, x_2, ..., x_n$. Let $\boldsymbol{e}_i$ denote the n-bit vector which is all 0's except for a 1 in the ith position. Let $\boldsymbol{e}_0$ denote the n-bit vector of 0's.

A function f is *sensitive on input $\boldsymbol{x}$ to an input bit x_i* if $f(\boldsymbol{x}) \neq f(\boldsymbol{x}\oplus\boldsymbol{e}_i)$ (i.e. the value of f changes when x_i is flipped). The *sensitivity of f on input $\boldsymbol{x}$* is

$$|\{i : f(\boldsymbol{x}) \neq f(\boldsymbol{x}\oplus\boldsymbol{e}_i)\}|.$$

In other words, it is the number of input bits to which f is sensitive on input $\boldsymbol{x}$. The *sensitivity of f* is the maximum over all inputs $\boldsymbol{x}$ of the sensitivity of f on $\boldsymbol{x}$.

Theorem 4.2.1 *For $\epsilon \in (0,1/2]$ and $\delta \in [0,1/2)$, if a Boolean function f is $(1-\delta)$-reliably computed by a circuit with ϵ-noisy, k-input gates then the number of gates in the circuit is at least*

$$\frac{s\log s + 2s\log(2(1-2\delta))}{k\log t}$$

where s is the sensitivity of the function f, $t = \frac{\omega^3+(1-\omega)^3}{\omega(1-\omega)}$, and $\omega = \frac{1-\sqrt[k]{\xi}}{2}$.

Proof: Without loss of generality, we assume that the function f has maximum sensitivity on the input $\boldsymbol{e}_0$; that $f(\boldsymbol{e}_0) = 0$; and that on input $\boldsymbol{e}_0$, f is sensitive to $x_1, ..., x_s$ (i.e. $f(\boldsymbol{e}_i) = 1$ for $i = 1, ..., s$).

Let m_i be the number of wires from input x_i. By lemma 4.1.1, we may assume that these input wires fail with probability $\omega = \frac{1-\sqrt[k]{\xi}}{2}$. Thus an input to the circuit causes a probability distribution on the values transmitted by the input wires.

Consider the input wires from the inputs $x_1, ..., x_s$. Let $p(\boldsymbol{y}|\boldsymbol{e}_i)$ be the probability that these wires transmit a particular vector $\boldsymbol{y}$ of values when the input to the circuit is $\boldsymbol{e}_i$. The vector $\boldsymbol{y}$ is a $\sum_{i=1}^{s} m_i$ dimensional vector of 0's and 1's.

Since the circuit is reliable, it outputs 0 with probability $\geq 1 - \delta$ on input $\boldsymbol{e}_0$ and 1 with probability $\geq 1 - \delta$ on input $\boldsymbol{e}_i$ for $i = 1, ..., s$. In order to distinguish input $\boldsymbol{e}_0$ from $\boldsymbol{e}_i$, the distributions $p(\cdot|\boldsymbol{e}_0)$ and $p(\cdot|\boldsymbol{e}_i)$ must be different. Let X be a Boolean random variable which is 0 if the circuit's input is $\boldsymbol{e}_0$ and 1 if the circuit's input is $\boldsymbol{e}_i$. Let Z be the output of the circuit. The reliability of the circuit implies that $||\boldsymbol{p}_{Z|X=0} - \boldsymbol{p}_{Z|X=1}||_1 \geq 2(1 - 2\delta)$. In addition, Z is independent of X given the values transmitted by the input wires. By the data processing inequality for L_1 distance (lemma B.0.2), the L_1 distance between $p(\cdot|\boldsymbol{e}_0)$ and $p(\cdot|\boldsymbol{e}_i)$ must be at least $2(1 - 2\delta)$. That is,

$$\sum_{y} |p(\boldsymbol{y}|\boldsymbol{e}_0) - p(\boldsymbol{y}|\boldsymbol{e}_i)| \geq ||\boldsymbol{p}_{Z|X=0} - \boldsymbol{p}_{Z|X=1}||_1 \geq 2(1 - 2\delta).$$

This implies that the L_1 distance between $p(\cdot|\boldsymbol{e}_0)$ and any average of the $p(\cdot|\boldsymbol{e}_i)$ is at least $2(1-2\delta)$, i.e.

$$D \equiv \sum_{y} \left| p(\boldsymbol{y}|\boldsymbol{e}_0) - \sum_{i=1}^{s} \alpha_i p(\boldsymbol{y}|\boldsymbol{e}_i) \right| \geq 2(1 - 2\delta) \tag{4.2}$$

with $\alpha_i \geq 0$ and $\sum \alpha_i = 1$.

Let us now derive an upper bound on this L_1 distance. Our first step is to rewrite the distance as an expectation.

$$D = \mathbf{E} \left| \sum_{i=1}^{s} \alpha_i \left(1 - \frac{p(\boldsymbol{y}|\boldsymbol{e}_i)}{p(\boldsymbol{y}|\boldsymbol{e}_0)} \right) \right|$$

where the expectation is with respect to the distribution $p(\cdot|\boldsymbol{e}_0)$. Since $\epsilon \neq 0$, we have $\omega \neq 0$ and $p(\boldsymbol{y}|\boldsymbol{e}_0) \neq 0$ for all $\boldsymbol{y}$.

After rewriting $|x|$ as $\sqrt{x^2}$, we use the Cauchy-Schwarz inequality to obtain,

$$D^2 \leq \mathbf{E} \left[\left(\sum_{i=1}^{s} \alpha_i \left(1 - \frac{p(\boldsymbol{y}|\boldsymbol{e}_i)}{p(\boldsymbol{y}|\boldsymbol{e}_0)} \right) \right)^2 \right]$$

Since $\boldsymbol{e}_i$ and $\boldsymbol{e}_0$ differ only in the ith input bit, the ratio

$$\frac{p(\boldsymbol{y}|\boldsymbol{e}_i)}{p(\boldsymbol{y}|\boldsymbol{e}_0)} = \left(\frac{\omega}{1 - \omega} \right)^{Z_i} \left(\frac{1 - \omega}{\omega} \right)^{m_i - Z_i}$$

where Z_i is the number of 0's transmitted by the m_i wires from the input x_i.[3] Under distribution $p(\cdot|\boldsymbol{e}_0)$, Z_i is independent of Z_j for $j \neq i$. Also,

$$\mathbf{E}\left[1 - \frac{p(\boldsymbol{y}|\boldsymbol{e}_i)}{p(\boldsymbol{y}|\boldsymbol{e}_0)}\right] = \sum_y p(\boldsymbol{y}|\boldsymbol{e}_0) - p(\boldsymbol{y}|\boldsymbol{e}_i) = 0$$

Thus the $\alpha_i(1 - p(\boldsymbol{y}|\boldsymbol{e}_i)/p(\boldsymbol{y}|\boldsymbol{e}_0))$ for $i = 1, \ldots, s$ are mutually independent and have mean zero. If $X_1, \ldots, X_m$ are independent, mean zero random variables then $\mathbf{E}[(\sum_i X_i)^2] = \mathbf{E}[\sum_i X_i^2]$. It follows that

$$\begin{aligned}\mathbf{E}\left[\left(\sum_{i=1}^{s} \alpha_i \left(1 - \frac{p(\boldsymbol{y}|\boldsymbol{e}_i)}{p(\boldsymbol{y}|\boldsymbol{e}_0)}\right)\right)^2\right] &= \sum_{i=1}^{s} \mathbf{E}\left[\left(\alpha_i \left(1 - \frac{p(\boldsymbol{y}|\boldsymbol{e}_i)}{p(\boldsymbol{y}|\boldsymbol{e}_0)}\right)\right)^2\right] \\ &= \sum_{i=1}^{s} \alpha_i^2 \mathbf{E}\left[\left(\frac{p(\boldsymbol{y}|\boldsymbol{e}_i)}{p(\boldsymbol{y}|\boldsymbol{e}_0)}\right)^2 - 1\right]\end{aligned}$$

Our next step is to calculate the expectation within the preceding sum. Since the wires fail independently with probability ω, $p(\boldsymbol{y}|\boldsymbol{e}_i)/p(\boldsymbol{y}|\boldsymbol{e}_0)$ is the product of m_i independent and identically distributed random variables. Each random variable is $\omega/(1-\omega)$ with probability $1-\omega$ and $(1-\omega)/\omega$ with probability ω, and so,

$$\mathbf{E}\left[\left(\frac{p(\boldsymbol{y}|\boldsymbol{e}_i)}{p(\boldsymbol{y}|\boldsymbol{e}_0)}\right)^2\right] = t^{m_i}$$

where $t = \frac{\omega^3 + (1-\omega)^3}{\omega(1-\omega)}$. Thus,

$$\mathbf{E}\left[\left(\sum_{i=1}^{s} \alpha_i \left(1 - \frac{p(\boldsymbol{y}|\boldsymbol{e}_i)}{p(\boldsymbol{y}|\boldsymbol{e}_0)}\right)\right)^2\right] = \sum_{i=1}^{s} \alpha_i^2 (t^{m_i} - 1)$$

Up to this point, the α_i could be any set of non-negative weights whose sum is one. We now choose particular values for α_i to minimize the right hand side of the preceding equation. This constrained optimization problem can be solved using the theory of Lagrange multipliers. The sum is minimized for

$$\alpha_i = \frac{1}{N(t^{m_i} - 1)}$$

[3] When y is distributed according to $p(\cdot|\epsilon_0)$, Z_i is a binomially distributed random variable:

$$\mathbf{P}[Z_i = a] = \binom{m_i}{a}(1-\omega)^a \omega^{m_i - a}$$

where N is the normalizing constant $N = \sum_i 1/(t^{m_i} - 1)$. With this choice of α_i,

$$D^2 \leq \left(\sum_{i=1}^{s} \frac{1}{t^{m_i} - 1} \right)^{-1} \leq \left(\sum_{i=1}^{s} \frac{1}{t^{m_i}} \right)^{-1}$$

Combining this with our lower bound (4.2) gives,

$$(2(1 - 2\delta))^2 \leq \left(\sum_{i=1}^{s} \frac{1}{t^{m_i}} \right)^{-1} \tag{4.3}$$

Using the inequality of arithmetic and geometric means, yields

$$(2(1 - 2\delta))^2 \leq \frac{1}{s} \left(\prod_{i=1}^{s} t^{m_i} \right)^{1/s}$$

which after taking logarithms implies the theorem. □

The proof of the theorem provides more information about the m_i than simply that their sum is large. Since $t \geq 1$, the bound 4.3 implies that the number of inputs with fewer than m input wires is at most

$$\frac{t^m}{(2(1 - 2\delta))^2}.$$

In other words, the circuit cannot sample a few inputs many times to make up for neglecting most of the inputs.

For fixed k, ϵ, and δ, theorem 4.2.1 implies that the size of a noisy circuit to compute a function with sensitivity s is $\Omega(s \log s)$. In 1991, this asymptotic behavior was proved by both Gál (for $\delta < 1/4$) [9] and independently by Reischuk and Schmeltz [22]. Recently, Gács and Gál extended Gál's result to all values $\delta < 1/2$ [8]. The improvement presented in this thesis is in the dependence of the bound on the noise ϵ introduced by each gate. In particular, we show that as ϵ approaches $1/2$, the size of the circuit increases unboundedly. See figure 4.1 for a comparison of the multiplicative factors of $s \log s$ as a function of ϵ.

Theorem 4.2.1 is fundamentally a statement about computation based on noisy samples. The fact that the computational model is a circuit of k-input gates allows us to transform a lower bound on the number of samples into a lower bound on circuit size. However, in bounding the number of samples, the computation can be any probabilistic function. The only requirements are that access to the input is exclusively through the noisy samples (in order to apply the L_1 data processing inequality) and that the sampling pattern is fixed, independent of the values obtained in sampling. Thus, for example, our result extends to the Static Noisy Decision Tree model just as in [22].

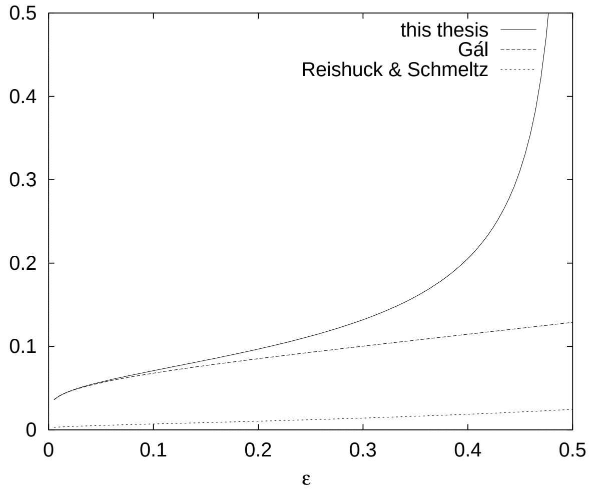


Figure 4.1: Lower bounds on factor of $s \log s$ for noisy circuit size using 3-input gates.

4.2.2 Block Sensitivity

As in [9], theorem 4.2.1 also holds when the sensitivity of f is replaced by the block sensitivity of f.

For any set $B \subseteq \{1, \ldots, n\}$, let $\boldsymbol{e}_B$ be the n-bit vector which is all 0's except that for every $i \in B$, $\boldsymbol{e}_B(i) = 1$. ($\boldsymbol{e}_B$ is the characteristic vector for the set B.) A function f has block sensitivity b if b is the largest number such that there exists disjoint sets $B_1, \ldots, B_b \subseteq \{1, \ldots, n\}$ and an input $\boldsymbol{x} = x_1, \ldots, x_n$ with $f(\boldsymbol{x}) \neq f(\boldsymbol{x} \oplus \boldsymbol{e}_{B_i})$ for $i = 1, \ldots, b$. The blocks B_i are sensitive blocks of f on input $\boldsymbol{x}$. Obviously, the block sensitivity of f is at least the sensitivity of f.

Theorem 4.2.2 *For $\epsilon \in (0, 1/2]$ and $\delta \in [0, 1/2)$, if a Boolean function f is $(1-\delta)$-reliably computed by a circuit with ϵ-noisy, k-input gates then the number of gates in the circuit is at least*

$$\frac{b \log b + 2b \log(2(1-2\delta))}{k \log t}$$

where b is the block sensitivity of the function f, $t = \frac{\omega^3 + (1-\omega)^3}{\omega(1-\omega)}$, and $\omega = \frac{1-\sqrt[k]{\xi}}{2}$.

Proof: As in the proof of theorem 4.2.1, assume without loss of generality that the function f has maximum block sensitivity on the input $\boldsymbol{e}_0$; that $f(\boldsymbol{e}_0) = 0$; and that B_i is the ith sensitive block of input bits (i.e. $f(\boldsymbol{e}_{B_i}) = 1$ for $i = 1, ..., b$). Redefine m_i to be the number of input wires from inputs in B_i.

The proof follows that given for theorem 4.2.1 with $\boldsymbol{e}_{B_i}$ replacing $\boldsymbol{e}_i$. □

4.2.3 Using Mutual Information

The proof given in this chapter is based upon lower and upper bounding the L_1 distance between conditional distributions. The distributions in question are on the input wires to the circuit; in one case, given that $\boldsymbol{e}_0$ is the input, and in the other, given that a randomly chosen $\boldsymbol{e}_i$ is the input. The obvious question in the context of this thesis is: Does using mutual information in place of the L_1 distance yield a better bound? We obtain a bound using mutual information that is essentially the same as that obtained using L_1 distance (the additive constant is smaller). In fact, the upper bound on the mutual information comes from the L_1 distance upper bound in the proof of theorem 4.2.1. The most interesting feature of the proof is its use of the signal decay theorem for 2-input, m-output channels. Improving the signal decay theorem for the particular distributions required by the proof, may lead to improved bounds on the size of noisy circuits. See appendix C for details of the information theoretic proof.

Chapter 5

Threshold for Noisy Computation

If the components (gates) of a circuit are too noisy then the circuit cannot be used to compute functions with any degree of reliability. In other words, for some values of ϵ, reliable computation[1] using ϵ-noisy, k-input gates is impossible. For instance, if $\epsilon = 1/2$ then each gate (including the output gate of the circuit) produces a fair coin flip. A result of chapter 3 (theorem 3.3.1) shows that if $\epsilon \geq 1/2 - 1/2\sqrt{k}$ then functions which depend on a sufficiently large number of inputs cannot be reliably computed. The basic question addressed in this chapter is: what is the exact threshold above which reliable computation is impossible using ϵ-noisy, k-input gates?

To make the notion of a threshold more precise, let $\mathcal{C}_k$ be the unique value in $[0, 1/2]$ such that if $\epsilon < \mathcal{C}_k$ then reliable computation is possible, but if $\epsilon \geq \mathcal{C}_k$ then reliable computation is impossible. Let $\mathcal{F}_k$ be the analogous quantity when computation is restricted to formulas. The values $\mathcal{C}_k$ and $\mathcal{F}_k$ are the noise thresholds for reliable computation by circuits and formulas using noisy, k-input gates. Since a formula is a circuit, reliable computation by formula implies reliable computation by circuit. Thus $\mathcal{F}_k \leq \mathcal{C}_k$.

The threshold phenomenon was first noticed by von Neumann, who showed that for $\epsilon < 0.0073$, reliable computation is possible using ϵ-noisy, 3-input majority gates. His method was to interleave "computation levels" of the circuit, i.e. levels which correspond to levels of the original (noiseless) circuit, with "error-correction levels", in which 3-input majority gates combine the output of three separate copies of each computation, in order to obtain an output which is more likely to be correct than any single copy.

As von Neumann noted, this idea cannot lead to reliable computation if $\epsilon \geq 1/6$. If each input to a 3-input majority gate is incorrect independently with probability a, then an ϵ-noisy, 3-input majority gate will be incorrect with probability

$$(1-\epsilon)(a^3 + 3a^2(1-a)) + \epsilon((1-a)^3 + 3a(1-a)^2) \tag{5.1}$$

[1] Reliable computation refers to the ability to $(1-\delta)$-reliably compute all Boolean functions for some fixed $\delta < 1/2$.

If $\epsilon < 1/6$ then this value can be smaller than a. Such error-correction is necessary for von Neumann's argument to work. However, if $\epsilon \geq 1/6$, then for every $a < 1/2$, the error probability of the output is greater than a, i.e. the output of the majority gate is less reliable than its inputs; and von Neumann's method fails.

This suggested to von Neumann that perhaps reliable computation is not possible by ϵ-noisy, 3-input gates if $\epsilon \geq 1/6$. The first proof of the impossibility of reliable computation by noisy components came in 1988 from Pippenger's work on formula depth bounds [16]. He showed that if $\epsilon \geq 1/2 - 1/2k$ then reliable computation by formula is impossible using ϵ-noisy, k-input gates. This implies that $\mathcal{F}_k \leq 1/2 - 1/2k$ (e.g. $\mathcal{F}_3 \leq 1/3$). Soon after, Feder extended this result to general circuits, showing $\mathcal{C}_k \leq 1/2 - 1/2k$ [7]. In chapter 3, we used more precise bounds on the information decay caused by noisy components to show that $\mathcal{C}_k \leq 1/2 - 1/2\sqrt{k}$.

In 1991, Hajek and Weller used a completely different technique to prove a tight threshold for reliable computation by formulas with noisy 3-input gates [11], showing that $\mathcal{F}_3 = 1/6$. In this chapter, we extend the work of Hajek and Weller to prove a tight threshold for reliable computation by formulas using noisy, k-input gates (k odd). The main result of the chapter is summarized in the following

Theorem 5.0.3 *For k odd and*

$$\beta_k = \frac{1}{2} - \frac{2^{k-2}}{k\binom{k-1}{\frac{k-1}{2}}},$$

there exists $\delta < 1/2$ such that all Boolean functions can be $(1-\delta)$-reliably computed by noisy formulas if and only if $\epsilon < \beta_k$.

Observe that the theorem implies that $\mathcal{F}_k = \beta_k$ and $\mathcal{C}_k \geq \beta_k$.

Using Stirling's approximation, $\beta_k \approx 1/2 - \sqrt{\pi}/2\sqrt{2k}$. This is only slightly smaller than the upper bound on $\mathcal{C}_k$ we obtained using information theoretic tools in chapter 3. The fact that the bounds are so close reflects the precision of our information theoretic bounds. See figure 5.1 for a comparison of these threshold bounds.

5.1 Threshold Value

To calculate the threshold for reliable computation, we consider a generalization of von Neumann's expression for the error of a 3-input majority gate (5.1). Let

$$m_{\epsilon,k}(a) = (1-\epsilon)\phi_k(a) + \epsilon(1-\phi_k(a)) \tag{5.2}$$

where

$$\phi_k(a) = \sum_{i=0}^{\lfloor k/2 \rfloor} \binom{k}{i} (1-a)^i a^{k-i}$$

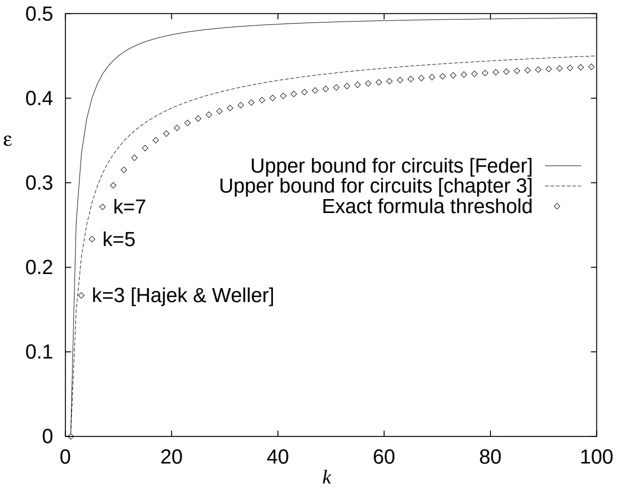


Figure 5.1: Bounds on the threshold for noisy computation.

Equation 5.2 represents the probability that an ϵ-noisy, k-input majority gate is incorrect given that its inputs are incorrect independently with probability a.

For $k = 3$, if $\epsilon \geq 1/6$ then $m_{\epsilon,3}(a) > a$ for all $a \in [0, 1/2)$. This is von Neumann's observation that the output of a noisy, 3-input majority gate is less reliable than its inputs, if ϵ is large. In this section, we generalize von Neumann's observation to noisy, k-input gates.

Lemma 5.1.1 *For k odd,*

1. *if $\epsilon \geq \beta_k$ then $m_{\epsilon,k}(a) > a$ for all $a \in [0, 1/2)$*
2. *if $\epsilon < \beta_k$ then there exists $\nu_{\epsilon,k} \in [0, 1/2)$ such that $m_{\epsilon,k}(\nu_{\epsilon,k}) = \nu_{\epsilon,k}$ and*
   - *if $a < \nu_{\epsilon,k}$ then $m_{\epsilon,k}(a) > a$*
   - *if $a > \nu_{\epsilon,k}$ then $m_{\epsilon,k}(a) < a$*

See figure 5.2 for an example of $m_{\epsilon,k}$ when $\epsilon = \beta_k$ and $\epsilon < \beta_k$.

Proof: Rewrite a as $(1 - \alpha)/2$ (i.e. α is twice the bias of a from 1/2) and let

$$f_k(\epsilon, \alpha) = m_{\epsilon,k}((1 - \alpha)/2) - \frac{1 - \alpha}{2}$$

(I.e. $f_k(\epsilon, \alpha)$ is the difference between the unreliability of the output and the unreliability of the inputs.)

We first show that $m_{0,k}(a) \leq a$ and $m_{\beta_k,k}(a) > a$ for $a \in [0, 1/2)$. This and the linearity of $m_{\epsilon,k}(a)$ in ϵ prove the first statement in the lemma.

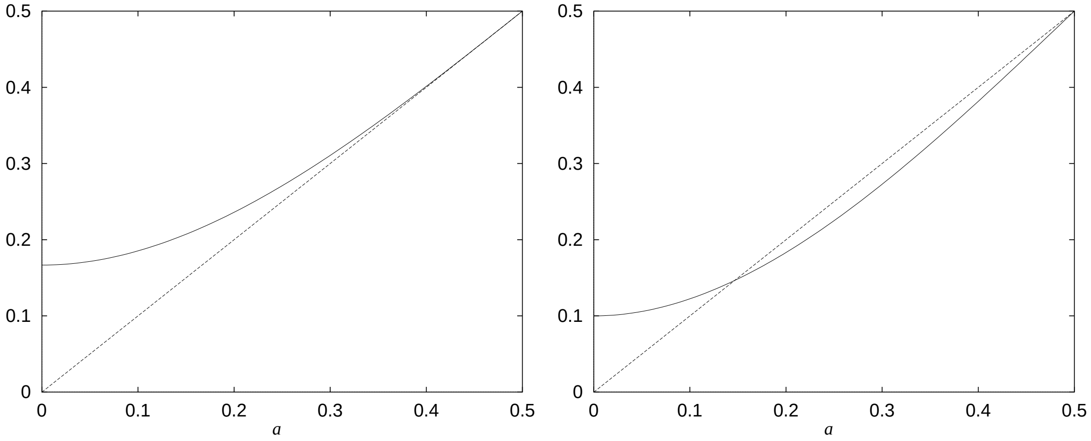


Figure 5.2: Error of ϵ-noisy, 3-input majority gate as a function of input error for $\epsilon = \beta_3$ and for $\epsilon < \beta_3$.

$m_{0,k}(a)$ is the probability that the noiseless majority of k inputs is incorrect given that each input is incorrect with probability a. To show $m_{0,k}(a) \leq a$, we show

$$m_{0,k}(a) = \sum_{i=0}^{\lfloor k/2 \rfloor} \binom{k}{i} (1-a)^i a^{k-i} \leq a \sum_{i=0}^{k} \binom{k}{i} (1-a)^i a^{k-i} = a$$

which for k odd is equivalent to

$$\sum_{i=0}^{\lfloor k/2 \rfloor} \binom{k}{i} (1-a)^i a^{k-i} \leq a \sum_{i=0}^{\lfloor k/2 \rfloor} \binom{k}{i} ((1-a)^i a^{k-i} + (1-a)^{k-i} a^i)$$

This inequality holds term by term.

Since $f_k(\beta_k, 0) = 0$, to show $m_{\beta_k,k}(a) > a$ for $a \in [0, 1/2)$ it suffices to show $f_k(\beta_k, \alpha)$ is an increasing function of $\alpha \in [0, 1]$ (i.e. $df_k(\beta_k, \alpha)/d\alpha \geq 0$),

$$\frac{df_k(\epsilon, \alpha)}{d\alpha} = 1/2 + (1 - 2\epsilon) \frac{d}{d\alpha} \phi_k \tag{5.3}$$

Since,

$$\frac{d}{d\alpha} \phi_k = -\frac{k}{2^k} \binom{k-1}{\frac{k-1}{2}} (1 - \alpha^2)^{\frac{k-1}{2}}$$

substituting into (5.3) yields,

$$\frac{df_k(\epsilon, \alpha)}{d\alpha} = 1/2 - (1 - 2\epsilon) \frac{k}{2^k} \binom{k-1}{\frac{k-1}{2}} (1 - \alpha^2)^{\frac{k-1}{2}}$$

For $\epsilon = \beta_k$,

$$\frac{df_k(\epsilon,\alpha)}{d\alpha} = \frac{1}{2} - \frac{(1-\alpha^2)^{\frac{k-1}{2}}}{2}$$

which is non-negative for $\alpha \in [0,1]$.

We now show the second statement of the lemma. The second derivative of $f_k(\epsilon,\alpha)$ is

$$\frac{d^2 f_k(\epsilon,\alpha)}{d\alpha^2} = (1-2\epsilon)\frac{k(k-1)}{2^k}\binom{k-1}{\frac{k-1}{2}}\alpha(1-\alpha^2)^{\frac{k-3}{2}}$$

which is non-negative for all $\alpha, \epsilon \in [0, 1/2]$. Thus $f_k(\epsilon,\alpha)$ is convex in $\alpha \in [0,1/2]$. Since $f_k(\epsilon, 0) = 0$ and $f_k(\epsilon,1) = \epsilon$, the convexity of $f_k(\epsilon,\alpha)$ will imply the lemma if we can show $df_k(\epsilon,\alpha)/d\alpha < 0$ at $\alpha = 0$. For $\epsilon < \beta_k$,

$$\frac{df_k(\epsilon,\alpha)}{d\alpha} \leq \frac{1}{2} - \frac{(1-\alpha^2)^{\frac{k-1}{2}}}{2}$$

with equality if and only if $\alpha = 1$, and thus at $\alpha = 0$, $df_k(\epsilon,\alpha)/d\alpha < 0$. □

5.2 Negative Result

Suppose we simply wish to “remember” an input bit for L computation steps. That is, to design a noisy circuit of depth L with one input x whose output is x with high probability. This would seem to be a prerequisite for computing complex functions. If the computation components are ϵ-noisy, k-input gates, the obvious method is to take the majority of k independent copies of the best circuit for remembering the input bit for $L-1$ steps. This construction results in a depth L formula of majority gates which has error probability $m^L_{\epsilon,k}(0)$ where $m^L_{\epsilon,k}$ is the L-fold composition of $m_{\epsilon,k}$. By lemma 5.1.1, if $\epsilon \geq \beta_k$, this technique will not work for arbitrarily large L. In fact, for k odd,

$$\text{if } \epsilon \geq \beta_k \text{ then } \lim_{L\to\infty} m^L_{\epsilon,k}(0) = 1/2.$$

This is the intuitive reason why β_k (which is derived from the behavior of noisy, k-input majority gates) is the noise level threshold for computation using any noisy, k-input gate.

To make this intuition precise, we show that if $\epsilon \geq \beta_k$ then, for any fixed $\delta < 1/2$, there are Boolean functions which cannot be computed by formula with error δ. In particular, theorem 5.2.1 shows that all functions which depend on n variables cannot be computed with error δ, for arbitrarily large n.[2]

Theorem 5.2.1 *For k odd, if $\epsilon \geq \beta_k$ then any formula using ϵ-noisy, k-input gates for computing a Boolean function which depends on at least $k^{L-1}+1$ variables errs with probability $\geq m^L_{\epsilon,k}(0)$ on some input.*

[2] Recall that a function depends on an argument x if there exists a setting of the other arguments such that the function restricted to that setting is not a constant.

Note that this implies reliable computation is impossible if $\epsilon \geq \beta_k$ (since $\lim_{L\to\infty} m^L_{\epsilon,k}(0) = 1/2$) and hence $\mathcal{F}_k \leq \beta_k$.

Proof: Let f be a Boolean function which depends on at least $k^{L-1}+1$ variables. Let F be a formula for f composed of ϵ-noisy, k-input gates. Since f depends on $k^{L-1}+1$ variables, there exists some variable x which is an input to only gates at depth $\geq L$ in F.[3] Thus any path from the input x to the output of the formula must pass through at least L gates. Fix the inputs other than x so that either $f = x$ or $f = 1 - x$; without loss of generality say $f = x$. Let F_x be the formula F with inputs other than x fixed as above.

Consider the two conditional probabilities $\mathbf{P}[F_x = 1|x = 0]$ and $\mathbf{P}[F_x = 0|x = 1]$. The maximum of these two quantities is a lower bound on the error probability of F.

Following Hajek and Weller, one may view these conditional probabilities geometrically as the point $(\mathbf{P}[F_x = 1|x = 0], \mathbf{P}[F_x = 0|x = 1])$ in the unit square. In general, if Y is a Boolean random variable, let

$$\lambda^Y = (\lambda^Y_0, \lambda^Y_1) = (\mathbf{P}[Y = 1|x = 0], \mathbf{P}[Y = 0|x = 1]).$$

For example, the ϵ-noisy, k-input majority gate with all inputs equal to x, produces an output Y described by the point $\lambda^Y = (m_{\epsilon,k}(0), m_{\epsilon,k}(0)) = (\epsilon, \epsilon)$. In this case, the probability that Y differs from x is ϵ.

The gate whose output is F_x (the top gate in the formula) does not receive x directly as input. The value of x must pass through at least $L-1$ noisy gates to reach this top gate. Each gate adds noise to the value of x, but the computation performed by the gate may compensate for this noise.

We show that if $\epsilon \geq \beta_k$ then each gate cannot compensate for the added noise. In fact, the space of points λ^Y, describing possible distributions at the gate's output, contracts as we pass x through more and more noisy gates. In particular, let $S(a)$ be the convex hull of the points $\{(0,1),(1,0),(a,a),(1-a,1-a)\}$. We show (lemma 5.3.1) that if the inputs to an ϵ-noisy, k-input gate are described by points in $S(a)$, then the output must lie in $S(m_{\epsilon,k}(a))$. See figure 5.3.

Using this lemma, we show by induction on L that the point describing the output of F_x lies within $S(m^L_{\epsilon,k}(0))$. This establishes the theorem since the error of any random variable whose point lies in $S(a)$ is at least a.

For $L = 1$, the formula consists of at least one gate. The points describing inputs to the top gate of the formula F_x lie within $S(0)$ (trivially) and thus, by lemma 5.3.1, the point describing the output lies within $S(m_{\epsilon,k}(0))$.

For $L > 1$, the formula consists of a top gate with at most k inputs. Each of these inputs is either constant with respect to x or the output of a formula in which x is an input to gates at depth $\geq L-1$. In the first case, the point describing the input lies within $S(a)$ for all a. In the

[3] The depth of a gate is the number gates on the path from its input to the output of the formula.

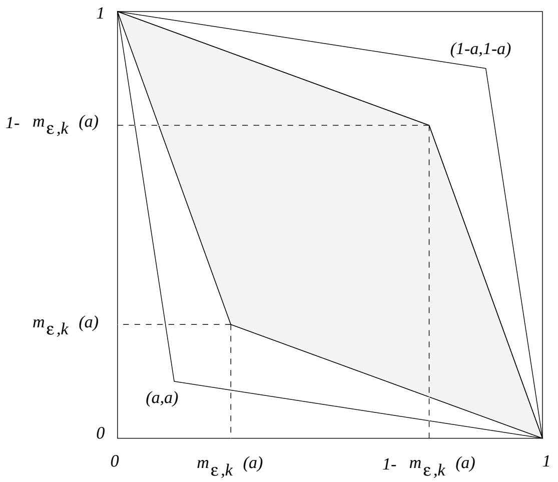


Figure 5.3: Contraction of $S(a)$ to $S(m_{\epsilon,k}(a))$ caused by one noisy gate.

second, the point describing the input lies in $S(m_{\epsilon,k}^{L-1}(0))$ by induction. Thus, by lemma 5.3.1, the point describing the output lies within $S(m_{\epsilon,k}^{L}(0))$. □

5.3 Contraction of $\mathbf{S(a)}$

Lemma 5.3.1 *If $\epsilon \geq \beta_k$ and $\lambda^{Y_1}, \lambda^{Y_2}, ..., \lambda^{Y_k} \in S(a)$ with $0 \leq a \leq 1/2$ then for all ϵ-noisy, k-input gates g with inputs $Y_1, Y_2, ..., Y_k$ and output Y, $\lambda^Y \in S(m_{\epsilon,k}(a))$.*

Proof: We will show in lemmas 5.4.1 and 5.4.2 that we may assume that $\lambda^{Y_i} = (a, a)$ for all i and that g is an ϵ-noisy, k-input threshold gate. A k-input threshold gate outputs 1 if and only if the number of inputs equal to 1 is at least t. The threshold t is an integer between 0 and k inclusive.

We show that the output Y of the gate g has $\lambda^Y \in S(m_{\epsilon,k}(a))$. By symmetry, we need only consider those threshold gates g with threshold $t \geq \lceil k/2 \rceil$. We will show that λ^Y lies within the convex hull of the vertices $\{(0,1), (1/2, 1/2), (m_{\epsilon,k}(a), m_{\epsilon,k}(a))\}$. Since $t \geq \lceil k/2 \rceil$, $\lambda_0^Y \leq \lambda_1^Y$. Also, $a \leq 1 - a$ implies $\lambda_0^Y + \lambda_1^Y \leq 1$. Thus we need only show that,

$$\lambda_0^Y + m_{\epsilon,k}(a)(\lambda_1^Y - \lambda_0^Y) \geq m_{\epsilon,k}(a) \tag{5.4}$$

when $\epsilon \geq \beta_k$.

Let V be the pre-noise output of gate g. That is, $\lambda_b^Y = \epsilon + (1 - 2\epsilon)\lambda_b^V$ for $b \in \{0, 1\}$. Then (5.4) becomes,

$$\lambda_0^V + m_{\epsilon,k}(a)(\lambda_1^V - \lambda_0^V) \geq \phi_k$$

where $\lambda_0^V = \phi_k(k-t,a)$, $\lambda_1^V = \phi_k(t-1,a)$, and

$$\phi_k(l,a) = \sum_{i=0}^{l} \binom{k}{i} (1-a)^i a^{k-i}.$$

Since $\epsilon \geq \beta_k$ implies $m_{\epsilon,k}(a) > a$, it suffices to show $\lambda_0^V + a(\lambda_1^V - \lambda_0^V) \geq \phi_k$ or,

$$a(\lambda_1^V - \phi_k) \geq (1-a)(\phi_k - \lambda_0^V)$$

Substituting the values of ϕ_k, λ_0^V, and λ_1^V yields,[4]

$$a(\phi_k(t-1,a) - \phi_k(\lfloor k/2 \rfloor, a)) \geq (1-a)(\phi_k(\lfloor k/2 \rfloor, a) - \phi_k(k-t,a))$$

The inequality holds term by term since $i \geq \lceil k/2 \rceil$ implies $a\binom{k}{i}(1-a)^i a^{k-i} \geq (1-a)\binom{k}{k-i}(1-a)^{k-i}a^i$. □

5.4 Reduction lemmas

The above proof relies on two lemmas which are rather straight-forward extensions of similar lemmas for $k = 3$ given by Hajek and Weller in [11].

An ϵ-noisy, k-input gate g takes as input $Y_1, \ldots, Y_k$, described by points $\lambda^{Y_1}, \ldots, \lambda^{Y_k}$, and outputs Y described by λ^Y. Thus the gate g defines a mapping $g : [0,1]^{2k} \rightarrow [0,1]^2$. Lemma 5.3.1 states that if $\epsilon \geq \beta_k$ then the union over all g of $g(S(a)^k)$ is contained in $S(m_{\epsilon,k}(a))$. The purpose of the following two lemmas is to show that it suffices to prove that the union over all threshold gates g of $g((a,a)^k)$ is contained in $S(m_{\epsilon,k}(a))$. (Note: $(a,a)^k$ is the point $(a,a),(a,a),...,(a,a)$ in $[0,1]^{2k}$.) The method is to show that the set of image points has the same convex hull in both cases. Thus, since $S(m_{\epsilon,k}(a))$ is convex, showing containment of either set implies containment of the other.

Lemma 5.4.1 *If C is the convex hull of the union over all g of $g(S(a)^k)$ and C_a is the convex hull of the union over all g of $g((a,a)^k)$ then*

$$C = C_a$$

Proof: The mapping from $S(a)^k$ to $[0,1]^2$ defined by g is affine, $[0,1]^2 \rightarrow [0,1]^2$, in each λ^{Y_i} when the others are fixed. Thus the image of $S(a)^k$ is contained in the convex hull of the image of the set of vertices of $S(a)^k$. Each vertex is of the form $(\lambda^{Y_1}, \lambda^{Y_2}, ..., \lambda^{Y_k})$ with $\lambda^{Y_i} \in \{(1,0),(0,1),(a,a),(1-a,1-a)\}$. If $\lambda^{Y_i} \in \{(1,0),(0,1)\}$ then the same value of λ^Y can be obtained with $\lambda^{Y_i} = (a,a)$ by modifying the gate g to ignore the value of Y_i. Similarly, if $\lambda^{Y_i} = (1-a,1-a)$ then the same value of λ^Y can be obtained with $\lambda^{Y_i} = (a,a)$ by modifying the gate g to negate input Y_i. The lemma follows. □

[4] If $t = \lceil k/2 \rceil$ both sides of the inequality are zero.

Lemma 5.4.2 *If C_a is the convex hull of the union over all g of $g((a,a)^k)$ and $C_{a,t}$ is the convex hull of the union over threshold gates g of $g((a,a)^k)$ then*

$$C_a = C_{a,t}$$

Proof: Note that $\lambda^{Y_i} = (a,a)$ for all i. To establish the lemma, it suffices to show that for any constants r and s, $r\lambda_0^Y + s\lambda_1^Y$ is minimized when g is some threshold function.

Again let V be the pre-noise output of gate g, so $\lambda_b^Y = \epsilon + (1-2\epsilon)\lambda_b^V$ for $b \in \{0,1\}$. Thus to minimize $r\lambda_0^Y + s\lambda_1^Y$, we minimize $r\lambda_0^V + s\lambda_1^V$.

$$\lambda_0^V = \sum_{(Y_1,Y_2,...,Y_k)\in S_1} \mathbf{P}[(Y_1,Y_2,...,Y_k)|x=0]$$

$$\lambda_1^V = \sum_{(Y_1,Y_2,...,Y_k)\in S_0} \mathbf{P}[(Y_1,Y_2,...,Y_k)|x=1]$$

where S_b is the set of k-bit vectors representing inputs for which $V = b$. A gate g which minimizes $r\lambda_0^V + s\lambda_1^V$ has $(Y_1,Y_2,...,Y_k) \in S_1$ if and only if $r\mathbf{P}[(Y_1,Y_2,...,Y_k)|x=0] < s\mathbf{P}[(Y_1,Y_2,...,Y_k)|x=1]$. From the fact that $\lambda^{Y_i} = (a,a)$,

$$\mathbf{P}[(Y_1,Y_2,...,Y_k)|x=0] = a^t(1-a)^{k-t}$$

$$\mathbf{P}[(Y_1,Y_2,...,Y_k)|x=1] = a^{k-t}(1-a)^t$$

where t is the number of ones in the vector $(Y_1,Y_2,...,Y_k)$. Thus the relation $r\mathbf{P}[(Y_1,Y_2,...,Y_k)$-$|x=0] < s\mathbf{P}[(Y_1,Y_2,...,Y_k)|x=1]$ holds monotonically in t and the lemma follows. □

5.5 Positive Result

The preceding section shows that the ability of an ϵ-noisy, k-input majority gate to decrease error probability is necessary for reliable computation using k-input gates. In this section, we show that this is also a sufficient condition.

For $\epsilon < \beta_k$, we show that there exists $\delta < 1/2$ such that given any Boolean function, we can construct a formula using ϵ-noisy, k-input gates which $(1-\delta)$-reliably computes the function. One obvious idea is to use von Neumann's technique of taking the noisy majority of k independent copies of a computation in order to decrease the error probability. This process can be repeated to decrease the error probability still further, but there is a limit. It will decrease error if and only if the original error is in the interval $(\nu_{\epsilon,k}, 1/2)$ where

$$\nu_{\epsilon,k} = \lim_{L\to\infty} m_{\epsilon,k}^L(a)$$

(By lemma 5.1.1, the limit exists and is the same for any $a \in [0,1/2)$.)

Once the error probability is back to a reasonable level, more computation can be done. Such a scheme works as long as computation can be performed at the "reasonable" error level achieved by the majority gates. In other words, computation at this level of error must result in an output which is correct on all inputs with probability strictly greater than $1/2$.

Hajek and Weller found a noisy, 3-input gate which computes reliably given very noisy inputs. Strangely, it requires the error of all of its inputs to be close to ν, not just close to or less than ν. In fact, the probability of an incorrect output bit can increase from below 1/2 to above 1/2 by *decreasing* the error of some of the inputs. Thus, if we know only that the noise at each gate is at most ϵ, an adversary could decrease the noise of some gates to below ϵ and ruin the reliability of the output. The construction takes advantage of the precise ϵ noise at the gates to obtain a reliable formula.

Hajek and Weller's noisy, 3-input computation gate is used to simulate the computation of a noiseless 2-input NAND gate. It is called an XNAND gate. A noiseless XNAND gate outputs 1 for inputs (0,0,0), (1,0,0), (0,0,1), and (0,1,1); and outputs 0 otherwise.

Let x and y be the inputs to a NAND gate. Let X be a noisy version of x; and Y_1 and Y_2 independent noisy versions of y. The output of XNAND on input (X, Y_1, Y_2) is intended to be a reliable version of NAND on input (x, y). The following lemma makes this connection precise.

Lemma 5.5.1 (Lemma 3.1 from [11]) *For $\epsilon, \nu \in [0, 1/2)$ there is a $\delta < 1/2$ and an open interval I with $\nu \in I \subset [0, 1/2]$ so that the following is true. If $\mathbf{P}[X \neq x]$, $\mathbf{P}[Y_1 \neq y]$, $\mathbf{P}[Y_2 \neq y] \in I$, and if Z is the output of an ϵ-noisy* XNAND *gate with input (X, Y_1, Y_2), then $\mathbf{P}[Z \neq \text{NAND}\,(x, y)] < \delta$.*

Proof: If $\mathbf{P}[X \neq x] = \mathbf{P}[Y_1 \neq y] = \mathbf{P}[Y_2 \neq y] = \nu$ then $\mathbf{P}[Z \neq \text{NAND}\,(x, y)]$ equals $(1-\nu)(2\epsilon - 1) + 1 - \epsilon$ if $(x, y) = (0, 0)$ or $(1, 1)$ and equals $(2\nu^2 - 2\nu + 1)(2\epsilon - 1) + 1 - \epsilon$ if $(x, y) = (1, 0)$ or $(0, 1)$. In either case, if $\epsilon, \nu \in [0, 1/2)$ then $\mathbf{P}[Z \neq \text{NAND}\,(x, y)] < 1/2$. Since $\mathbf{P}[Z \neq \text{NAND}\,(x, y)]$ is a continuous function of $(\mathbf{P}[X \neq x], \mathbf{P}[Y_1 \neq y], \mathbf{P}[Y_2 \neq y])$, the proof is complete. □

We use the XNAND gate in conjunction with k-input majority gates (k odd) to show that reliable computation by precisely ϵ-noisy, k-input gates is possible if $\epsilon < \beta_k$.

Theorem 5.5.1 *For k odd and $0 \leq \epsilon < \beta_k$, there exists $\delta < 1/2$ such that any Boolean function can be $(1-\delta)$-reliably computed by a formula using ϵ-noisy, k-input gates.*

Note that this implies $\mathcal{F}_k \geq \beta_k$.

Proof: The proof is a simple extension of Proposition 3 from Hajek and Weller [11]. For k odd ($k \geq 3$), an ϵ-noisy XNAND gate can be implemented by an ϵ-noisy, k-input gate which ignores all but three of its inputs. Use δ and I with $\nu = \nu_{\epsilon,k}$ from the proof of lemma 5.5.1 and choose L large enough so that $[m^L_{\epsilon,k}(0), m^L_{\epsilon,k}(\delta)] \subset I$.

Start with a formula composed of 2-input NAND gates which computes the function. The idea is to replace the noiseless formula with a formula composed of ϵ-noisy, k-input majority gates and XNAND gates. The replacement is performed inductively. If the formula is trivial, i.e. a single input or constant, then we are done. Otherwise, suppose the top NAND gate has two inputs x and y. By induction, replace the formulas computing x and y with three noisy formulas: one which computes a noisy version U of x, and two which compute independent noisy versions, V_1 and V_2, of y. The induction insures that the error probabilities of these noisy versions lie within $[0, \delta]$.

By replicating their formulas, make k^L independent copies of each of U, V_1,and V_2. Use L levels of ϵ-noisy, k-input gates to combine the copies of U into one noisy version X of x whose error probability lies within I. Do the same with the copies of V_1 and V_2 to obtain Y_1 and Y_2 with error in I. By lemma 5.5.1, the output of an XNAND gate with these inputs will be a $(1 - \delta)$-reliable version of the original output. □

Chapter 6

Future Work

6.1 Information Theory

One of the most obvious open problems is to extend the upper bound on the fraction of information that can cross a noisy channel to more general channels. Suppose the random variable X represents a randomly chosen message, and Y a corrupted version of this message. If Z is the output of a noisy channel on input Y then Z carries no more information about X than Y did. What is the maximum fraction of information carried by Y that makes it through? In other words, we desire an upper bound on $I(X;Z)/I(X;Y)$ over all possible joint distributions on X and Y. The bound then will be solely in terms of the channel between Y and Z.

We obtain a tight upper bound on $I(X;Z)/I(X;Y)$ only in the case when X and Y are Boolean random variables and Z is the output of a binary channel on input Y. If Z is the output of a 2-input, m-output channel, we show an upper bound which generalizes the binary channel bound but is not tight for all channels.

In the most general situation, X takes one of r possible values, Y takes one of n possible values, and Z is the output of an n-input, m-output channel on input Y. What is the maximum over joint distributions on X and Y of $I(X;Z)/I(X;Y)$?

Often this ratio is 1. For example, if X is a Boolean random variable and the n inputs to the channel can be partitioned so that the outputs of two of the partitions do not overlap, then the channel will perfectly preserve one bit of information. A joint distribution on X and Y that achieves this places Y in one of the two partitions if $X = 0$ and in the other if $X = 1$. Since the outputs of these two partitions do not overlap, $I(X;Z) = 1$.

We conjecture that when X is Boolean, the distributions on X and Y which maximize $I(X;Z)/I(X;Y)$ place all of their probability on only two of the n values of Y. That is, the behavior of the general n-input, m-output channel is governed by the “best” 2-input, m-output channel.

6.2 Depth and Size Bounds

The techniques used in this thesis to obtain lower bounds on the depth and size of reliable circuits treat circuit components as black boxes. The functions these components compute are immaterial to the analysis. The components are simply characterized by the number of their inputs and their noise. In addition, the function computed by the circuit as a whole is only specified by the number of its inputs and, in the case of size bounds, its sensitivity.

Such gross complexity measures are suitable for our analyses, since our arguments are based on understanding how noise affects measures of correlation. But they place limitations on the precision of the bounds we can obtain. One potential way to improve these bounds is to use more detailed complexity measures, particularly on the function computed by the circuit.

6.3 Threshold Bounds

The tight threshold result of chapter 5 holds only when the number k of inputs to a noisy gate is odd, and the computation is performed by a formula. In this case, we can argue that for noise above a certain threshold, the best way to preserve a single input value through L levels of noisy computation is to feed k^L copies of the input into a formula of noisy, k-input majority gates. This is the fundamental idea behind the proof.

For $k = 2$, there are two possible “majority” gates: the “and” and “or” gates. Given inputs with symmetric error, one introduces bias towards 1 and the other bias towards 0. Repeatedly using the same one skews the error probability more and more. In order to dampen the bias, Valiant [26] proposed alternating “and” and “or” gates level by level. This scheme decreases error if the gate noise is below a certain threshold. However, it is not clear that this is the best way to decrease error for noise values at and above that threshold.

The second deficiency in the exact threshold results is that they hold only for formulas. We do not know the exact threshold for reliable circuit computation. We know that since a formula is a circuit, the exact threshold for formulas provides a lower bound on the threshold for circuits. In addition, we establish in chapter 3 an upper bound on this threshold which has the same asymptotic dependence on k as the lower bound. The gap is extremely small. Almost surely, the true threshold for circuits and formulas is the same. It would be very interesting if this were not true.

6.4 Markov Chains

Markov chains which converge rapidly to their stationary distribution have provided efficient methods to obtain approximate solutions to many problems whose exact solution is presumably difficult. In each case, proving a bound on the rate of convergence is the crucial step in analyzing the efficiency of the algorithm. See Dyer, Frieze, and Kannan [4] for a specific example; and Vazirani [27] and Sinclair [24] for a survey of results.

Convergence is typically argued in terms of the L_1 or L_2 distance between the current distribution on the chain and the final stationary distribution. Each step of the Markov chain is shown to reduce this distance by a multiplicative constant.

Another way to view the convergence of a Markov chain is as a process of information loss. Initially, the system is in a known state, but over time information about the initial state decays, in exactly the same manner as information carried by a signal decays as the signal crosses noisy channels. If, after some time, the current state of the Markov chain contains little information about the initial state then, intuitively, the current distribution on states should appear "close" to the stationary distribution. Sampling according to the current distribution is then approximately equivalent to sampling according to the stationary distribution. Since both n state Markov chains and noisy channels with n inputs and n outputs can be viewed as stochastic matrices, one is tempted to prove rapid information loss by analyzing the factor of information decay of n-input, n-output channels.

The problem with this approach is that for many n-input, n-output channels (including those that represent typical Markov chains of interest), there are input distributions which preserve information perfectly in crossing the channel. This observation was made in discussing the general n-input, m-output channel bound above. It appears that in order to show rapid information loss in this way, one must take advantage of the fact that not all input distributions are possible.

Another possibility is to approach information loss in a more indirect manner; similar to the way in which convergence rates are bounded by showing bounds on conductance. Roughly, conductance is the probability of escaping from the most secluded part of the chain in one step.[1] Large conductance implies rapid convergence to the stationary distribution: since a random walk is never trapped in some small part of the chain, it can reach the stationary distribution quickly. One idea is to derive a direct relation between information loss and conductance. A more exciting possibility is that a different structural property of the Markov chain (other than conductance) governs information loss. If such a property is easy to verify, it may prove useful in obtaining convergence results for Markov chains in which conductance bounds are unknown.

6.5 Quadratic Dynamical Systems

A dynamical system consists of a fixed function f which maps points from an n-dimensional space S to S. The system evolves by applying the function f to the current "configuration" of the system (a point in S) to obtain a new configuration in S (for a more complete discussion see [21]). Markov chains are one of the simplest examples of a dynamical system. For

[1] Conductance is the minimum over all bipartitions of the chain of the probability of escaping the smaller partition in one time step (assuming initially the uniform distribution on the smaller partition).

Markov chains the space S is the set of all probability distributions on n states and the function f is represented by the transition matrix of the Markov chain (a linear function).

The study of dynamical systems focuses on properties of the *trajectory* of a point p in the space S. The trajectory of p is the sequence $p, f(p), f(f(p)), \ldots$. One of the most fundamental questions is whether the trajectory approaches a unique limiting value independent of p and if so at what rate. In the case of Markov chains, the conditions under which the trajectory converges are known and bounds on the rate at which this convergence occurs can be obtained from conductance bounds or eigenvalue methods.

The situation changes dramatically for nonlinear f. Even in the case when f is a symmetric quadratic function which is known to converge, bounds on the rate of convergence are still unknown.[2] Traditional techniques using conductance or eigenvalues have so far failed to provide bounds on the convergence rate. In this case, as opposed to Markov chains, the only method known to prove convergence argues that the entropy of the system strictly increases unless the system is in the stationary configuration. At the heart of the argument, a step of the quadratic system is broken down into smaller steps, one of which is a linear map that strictly increases entropy. It may be possible to show that information irrevocably decays at this point by viewing the linear map as a noisy channel. The hope is that a bound on this decay will translate into a bound on the rate of convergence.

[2] An example of such a system is Maxwell's (space homogeneous) kinetic gas model. See [21].

Appendix A

Information Theory

This appendix is meant as a quick review of some of the properties of entropy and information. The results presented here should provide the reader with some intuition for the quantities which play the major roles in information theory.

A.1 Definitions

The *entropy* of a random variable X with distribution $\boldsymbol{p}_X$ is

$$H(X) \equiv -\sum_x \boldsymbol{p}_X(x) \log \boldsymbol{p}_X(x).$$

The *conditional entropy* of a random variable X given a random variable Y is

$$H(X|Y) \equiv -\sum_{x,y} \boldsymbol{p}_{X,Y}(x,y) \log \boldsymbol{p}_{X|Y=y}(x).$$

For distributions $\boldsymbol{p}$ and $\boldsymbol{q}$, the *Kullback-Leibler divergence* (or relative entropy) from $\boldsymbol{q}$ to $\boldsymbol{p}$ is

$$D(\boldsymbol{p}||\boldsymbol{q}) \equiv \sum_x \boldsymbol{p}(x) \log \frac{\boldsymbol{p}(x)}{\boldsymbol{q}(x)}$$

The *mutual information* between two random variables X and Y is

$$\begin{aligned} I(X;Y) &\equiv \sum_{x,y} \boldsymbol{p}_{X,Y}(x,y) \log \frac{\boldsymbol{p}_{X,Y}(x,y)}{\boldsymbol{p}_X(x)\boldsymbol{p}_Y(y)} \\ &= \sum_x \boldsymbol{p}_X(x) D(\boldsymbol{p}_{Y|X=x}||\boldsymbol{p}_Y) \\ &= D(\boldsymbol{p}_{X,Y}||\boldsymbol{p}_X\boldsymbol{p}_Y) \end{aligned}$$

$$
\begin{aligned}
&= H(X) - H(X|Y) \\
&= H(Y) - H(Y|X)
\end{aligned}
$$

The *conditional mutual information* between X and Y given Z is

$$
\begin{aligned}
I(X;Y|Z) &\equiv \sum_{x,y,z} \boldsymbol{p}_{X,Y,Z}(x,y,z) \log \frac{\boldsymbol{p}_{X,Y|Z=z}(x,y)}{\boldsymbol{p}_{X|Z=z}(x)\boldsymbol{p}_{Y|Z=z}(y)} \\
&= H(X|Z) - H(X|Y,Z) \\
&= H(Y|Z) - H(Y|X,Z)
\end{aligned}
$$

We can also condition on a particular value of Z,

$$
I(X;Y|Z=z) \equiv \sum_{x,y} \boldsymbol{p}_{X,Y|Z=z}(x,y) \log \frac{\boldsymbol{p}_{X,Y|Z=z}(x,y)}{\boldsymbol{p}_{X|Z=z}(x)\boldsymbol{p}_{Y|Z=z}(y)}
$$

With this definition,

$$
I(X;Y|Z) = \sum_{z} \boldsymbol{p}_Z(z) I(X;Y|Z=z)
$$

A.2 Properties of Entropy and Information

Perhaps the best way to read this section is to start at the end with Fano's inequality and the data processing inequality. Understanding these inequalities is crucial to understanding most of the thesis. Also, their proofs tie together all of the previous results in the section. The first part of the section is provided as reference for understanding these inequalities and some of the results in the thesis.

Chain Rule for Entropy and Information

Theorem A.2.1 (Chain rule for entropy)

$$
H(Y_1,\ldots,Y_n|Q) = \sum_{i=1}^{n} H(Y_i|Y_1,\ldots,Y_{i-1},Q)
$$

Proof: Let $X = Y_1$ and $Y = Y_2,\ldots,Y_n$.

$$
\begin{aligned}
H(X,Y|Q) &= -\sum_{x,y,q} \boldsymbol{p}_{X,Y,Q}(x,y,q) \log \boldsymbol{p}_{X,Y|Q=q}(x,y) \\
&= -\sum_{x,y,q} \boldsymbol{p}_{X,Q}(x,q)\boldsymbol{p}_{Y|X=x,Q=q}(y)(\log \boldsymbol{p}_{X|Q=q}(x) + \log \boldsymbol{p}_{Y|X=x,Q=q}(y))
\end{aligned}
$$

$$\begin{aligned} &= -\sum_{x,q} \boldsymbol{p}_{X,Q}(x,q) \log \boldsymbol{p}_{X|Q=q}(x) \\ &\quad - \sum_{x,y,q} \boldsymbol{p}_{X,Q}(x,q)\boldsymbol{p}_{Y|X=x,Q=q}(y) \log \boldsymbol{p}_{Y|X=x,Q=q}(y) \\ &= H(X|Q) + H(Y|X,Q) \end{aligned}$$

The general result follows by repeating the argument on $H(Y|X,Q)$. □

Theorem A.2.2 (Chain rule for information)

$$I(X;Y_1,\ldots,Y_n|Q) = \sum_{i=1}^{n} I(X;Y_i|Y_1,\ldots,Y_{i-1},Q)$$

Proof:

$$\begin{aligned} I(X;Y_1,\ldots,Y_n|Q) &= H(Y_1,\ldots,Y_n|Q) - H(Y_1,\ldots,Y_n|X,Q) \\ &= \sum_{i=1}^{n} H(Y_i|Y_1,\ldots,Y_{i-1},Q) - \sum_{i=1}^{n} H(Y_i|Y_1,\ldots,Y_{i-1},X,Q) \\ &= \sum_{i=1}^{n} I(X;Y_i|Y_1,\ldots,Y_{i-1},Q) \end{aligned}$$

□

Non-negativity of Information

A function $f(x)$ is *convex* on the interval (a,b) if for all $x_1, x_2 \in (a,b)$ and $p \in [0,1]$,

$$f(px_1 + (1-p)x_2) \leq pf(x_1) + (1-p)f(x_2)$$

The function is *strictly convex* if, in addition, the preceding inequality is strict for $p \in (0,1)$. A function $f(x)$ is *concave* if $-f(x)$ is convex.

The concavity of the log function implies all of the inequalities in this section.

Theorem A.2.3 (Divergence is non-negative) *For probability distributions $\boldsymbol{p}$ and $\boldsymbol{q}$,*

$$D(\boldsymbol{p}||\boldsymbol{q}) \geq 0$$

with equality if and only if $\boldsymbol{p}(x) = \boldsymbol{q}(x)$ for all x.

Proof: The proof follows from the concavity of the log function.

$$\begin{aligned}
-D(\boldsymbol{p}||\boldsymbol{q}) &= \sum_x \boldsymbol{p}(x) \log \frac{\boldsymbol{q}(x)}{\boldsymbol{p}(x)} \\
&\leq \log \left(\sum_x \boldsymbol{p}(x) \frac{\boldsymbol{q}(x)}{\boldsymbol{p}(x)} \right) \\
&= 0
\end{aligned}$$

□

It follows from this theorem and the fact that information can be written as a divergence, that

Corollary A.2.1 (Information is non-negative)

$$I(X;Y) \geq 0$$

with equality if and only if X and Y are independent.

From this corollary and the fact that $I(X;Y) = H(X) - H(X|Y)$, we see that

Corollary A.2.2 (Conditioning reduces entropy)

$$H(X) \geq H(X|Y)$$

Another implication of the non-negativity of divergence is that the entropy of a random variable taking one of m possible values is at most $\log m$.

Corollary A.2.3 *If X is a random variable taking one of m possible values then $H(X) \leq \log(m)$ with equality if and only if X has distribution $\frac{\bar{1}}{m}$ (the uniform distribution on m values).*

Proof:

$$\log(m) - H(X) = D(\boldsymbol{p}_X || \frac{\bar{1}}{m}) \geq 0$$

□

Data Processing Inequality

The data processing inequality states that functions of a random variable Y carry no more information about a random variable X than Y does. The functions can be deterministic or randomized; the only restriction being that they depend on X only through Y. If Z is a function of Y then the restriction that the function depend on X only through Y is precisely that Z is independent of X given Y.

Theorem A.2.4 (Data processing inequality) *Let X, Y, and Z be random variables such that Z is independent of X given Y. Then*

$$I(X;Y) \geq I(X;Z)$$

Also $I(X;Y) \geq I(X;Y|Z)$.

Proof: We use the chain rule for information to write,

$$I(X;YZ) = I(X;Y) + I(X;Z|Y) = I(X;Z) + I(X;Y|Z)$$

Since X and Z are independent given Y, $I(X;Z|Y) = 0$. Thus, $I(X;Y) = I(X;Z) + I(X;Y|Z)$ and, since information is non-negative, $I(X;Y) \geq I(X;Z)$. The same reasoning implies $I(X;Y) \geq I(X;Y|Z)$. □

Theorem A.2.5 (Sub-additivity of Information) *If $Y_1, \ldots, Y_n$ are independent random variables given random variable X then*

$$I(X;Y_1, \ldots, Y_n) \leq \sum_{i=1}^{n} I(X;Y_i)$$

Proof: The chain rule of information implies,

$$I(X;Y_1, \ldots, Y_n) = \sum_{i=1}^{n} I(X;Y_i|Y_1, \ldots, Y_{i-1})$$

From the preceding proof (with $X = Y_i$, $Y = X$, and $Z = X_1, \ldots, X_{i-1}$) the ith term in the sum is bounded by $I(X;Y_i)$. □

Fano's Inequality

Suppose two random variables X and Y are very likely to have the same value. Intuitively, we would expect the mutual information between X and Y to be large. Fano's inequality makes this intuition precise.

Theorem A.2.6 (Fano's inequality) *Let X be a random variable taking m possible values. Let E and Y be random variables such that*

$$E = \begin{cases} 0 & \text{if } Y = X \\ 1 & \text{if } Y \neq X \end{cases}$$

Then

$$I(X;Y) \geq H(X) - H(E) - \mathbf{P}[E = 1]\log(m-1)$$

Proof: We use the chain rule for entropy to write,

$$H(E, X|Y) = H(E|Y) + H(X|Y, E) = H(X|Y) + H(E|Y, X)$$

Now the mutual information between X and Y can be written as,

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= H(X) - H(E|Y) - H(X|Y, E) + H(E|Y, X) \end{aligned}$$

Since E is determined by the values of X and Y, $H(E|Y, X) = 0$. Also, since conditioning reduces entropy, $H(E|Y) \leq H(E)$. Thus,

$$\begin{aligned} I(X;Y) &\leq H(X) - H(E) - H(X|Y, E) \\ &= H(X) - H(E) - \mathbf{P}[E = 0]H(X|Y, E = 0) - \mathbf{P}[E = 1]H(X|Y, E = 1) \end{aligned}$$

If $E = 0$ then $X = Y$ and so $H(X|Y, E = 0) = 0$. If $E = 1$ then $X \neq Y$ and so, given Y, X can take one of only $m - 1$ values. Thus $H(X|Y, E = 1) \leq \log(m - 1)$ which completes the proof. □

For our applications, X will be a Boolean random variable, and, in addition, we will only know that $\mathbf{P}[E = 1] \leq \delta$ (or $\mathbf{P}[E = 0] \leq \delta$) for some $\delta < 1/2$. In this case, $I(X;Y) \geq H(X) - H(E) \geq 1 + \delta \log \delta + (1 - \delta) \log(1 - \delta)$.

For background on information theory, the texts of Gallager [10] and Cover and Thomas [1] as well as Shannon's original paper [23] are recommended.

Appendix B

L_C Distance

In this appendix, we discuss the use of L_c distance in place of mutual information for proving lower bounds on noisy formula depth. This particular application provides the framework for presenting several results on the use of L_c distance as a measure of correlation. Some of the results are also used in chapter 4 to prove lower bounds on the size of noisy circuits.

Let X and Y be Boolean random variables. One measure of the correlation between X and Y is the mutual information $I(X;Y)$. This is the measure used by Pippenger in proving the first lower bounds on noisy formula depth [16]. Mutual information is an attractive measure for this purpose because it possesses three properties:

1. If X and Y are equal with high probability then $I(X;Y)$ is large.
2. If Y is subjected to random noise (unrelated to X) then $I(X;Y)$ decreases by at least a multiplicative factor related to the noise.
3. If Y_1 and Y_2 carry information about X and are independent given X then $I(g(Y_1,Y_2);X) \leq I(Y_1;X) + I(Y_2;X)$ for any Boolean function g.

These are the only properties of information needed to prove Pippenger's noisy formula lower bound.

Another potential measure of correlation between X and Y is the L_c distance between the conditional distributions on Y given $X = 0$ and $X = 1$. One can think of the random variable X as the outcome of a coin flip and Y as a random variable related to this outcome. Thus if the coin comes up heads, Y has one distribution, while tails causes Y to have a different distribution. For X and Y to be highly correlated, these two distributions should be far apart according to their L_c distance. Recall that the L_c distance between two vectors $\boldsymbol{p}$ and $\boldsymbol{q}$ of dimension m is

$$||\boldsymbol{p} - \boldsymbol{q}||_c = \left(\sum_{i=1}^{m} |\boldsymbol{p}(i) - \boldsymbol{q}(i)|^c \right)^{1/c} .$$

It is trivial to check that the L_c distance obeys the first two properties of mutual information stated above. In fact, the multiplicative factor in property two for L_c distance is identical to the bound obtained by Pippenger for mutual information. For noise $\epsilon = (1-\xi)/2$, the factor is ξ. Thus, if the L_c distance obeys property 3, then it can replace the mutual information in Pippenger's proof, to yield a lower bound on the depth of noisy circuits with the same multiplicative increase and the same threshold bound obtained by Pippenger. However, the improved depth lower bound presented in chapter 3, which depends upon a ξ^2 drop in information, appears to be beyond the reach of an argument using L_c distance.

The remainder of the appendix is devoted to proving that the L_c distance measure obeys the third property which we call the *sub-additivity* property.

The following theorem shows that L_c distance, for $c \geq 1$, has the sub-additivity property. The proof is similar to the proof of the sub-additivity of mutual information (see theorem A.2.5). The first step is to show that the distance between the distributions $\boldsymbol{p}_{Y_1,\ldots,Y_k|X=0}$ and $\boldsymbol{p}_{Y_1,\ldots,Y_k|X=1}$ is at most the sum of the distances between $\boldsymbol{p}_{Y_i|X=0}$ and $\boldsymbol{p}_{Y_i|X=1}$.

Lemma B.0.1 *Let X be a Boolean random variable. If $Y_1,\ldots,Y_k$ are (not necessarily Boolean) random variables which are mutually independent given X then*

$$||\boldsymbol{p}_{Y_1,\ldots,Y_k|X=0} - \boldsymbol{p}_{Y_1,\ldots,Y_k|X=1}||_c \leq \sum_{i=1}^{k} ||\boldsymbol{p}_{Y_i|X=0} - \boldsymbol{p}_{Y_i|X=1}||_c$$

for $c \geq 1$.

Proof: The lemma follows from the case $k=2$ by a simple inductive argument. We prove the case $k=2$ using the triangle inequality for L_c distance (line B.1).[1]

In particular, let $p_i + \epsilon_i = \mathbf{P}[Y_1 = i|X=0]$, $p_i - \epsilon_i = \mathbf{P}[Y_1 = i|X=1]$, $q_i + \delta_i = \mathbf{P}[Y_2 = i|X=0]$, and $q_i - \delta_i = \mathbf{P}[Y_2 = i|X=1]$. Observe that the p_i (and q_i) are non-negative and sum to 1. With this notation,

$$\begin{aligned}
||\boldsymbol{p}_{Y_1,Y_2|X=0} - \boldsymbol{p}_{Y_1,Y_2|X=1}||_c &= \left(\sum_{i,j} |(p_i+\epsilon_i)(q_j+\delta_j) - (p_i-\epsilon_i)(q_j-\delta_j)|^c\right)^{1/c} \\
&= \left(\sum_{i,j} |2p_i\delta_j + 2q_j\epsilon_i|^c\right)^{1/c} \\
&\leq \left(\sum_{i,j} (|2p_i\delta_j| + |2q_j\epsilon_i|)^c\right)^{1/c}
\end{aligned}$$

[1] This is a version of Minkowski's inequality. See theorem 25 in [13].

$$
\begin{aligned}
&\leq \left(\sum_{i,j} |2p_i\delta_j|^c\right)^{1/c} + \left(\sum_{i,j} |2q_j\epsilon_i|^c\right)^{1/c} \qquad \text{(B.1)}\\
&\leq \left(\sum_{j} |2\delta_j|^c\right)^{1/c} + \left(\sum_{i} |2\epsilon_i|^c\right)^{1/c}\\
&= ||\boldsymbol{p}_{Y_1|X=0} - \boldsymbol{p}_{Y_1|X=1}||_c + ||\boldsymbol{p}_{Y_2|X=0} - \boldsymbol{p}_{Y_2|X=1}||_c
\end{aligned}
$$

□

To follow the method of proving sub-additivity of information, we would have to show the equivalent of the data processing inequality:

$$||\boldsymbol{p}_{g(Y_1,\ldots,Y_k)|X=0} - \boldsymbol{p}_{g(Y_1,\ldots,Y_k)|X=1}||_c \leq ||\boldsymbol{p}_{Y_1,\ldots,Y_k|X=0} - \boldsymbol{p}_{Y_1,\ldots,Y_k|X=1}||_c$$

for any function g. This statement is true for $c = 1$, but not for $c > 1$. Nevertheless, it is sufficient to know the data processing inequality for $c = 1$ to establish sub-additivity for any L_c distance if g is a Boolean function. We first prove the data processing inequality for $c = 1$ and then prove sub-additivity.

Lemma B.0.2 (Data Processing Inequality for L_1 distance) *Let X be a Boolean random variable and let Y and Z be random variables such that Z is independent of X given Y. Then*

$$||\boldsymbol{p}_{Z|X=0} - \boldsymbol{p}_{Z|X=1}||_1 \leq ||\boldsymbol{p}_{Y|X=0} - \boldsymbol{p}_{Y|X=1}||_1$$

Proof: The independence of Z and X given Y implies that $\boldsymbol{p}_{Z|Y} = \boldsymbol{p}_{Z|XY}$. Thus $\boldsymbol{p}_{Z|X} = \sum_y \boldsymbol{p}_{Z|Y=y}\boldsymbol{p}_{Y|X}(y)$. This allows us to write,

$$
\begin{aligned}
||\boldsymbol{p}_{Z|X=0} - \boldsymbol{p}_{Z|X=1}||_1 &= \sum_z |\boldsymbol{p}_{Z|X=0}(z) - \boldsymbol{p}_{Z|X=1}(z)|\\
&= \sum_z \left|\sum_y \boldsymbol{p}_{Z|Y=y}(z)\boldsymbol{p}_{Y|X=0}(y) - \sum_y \boldsymbol{p}_{Z|Y=y}(z)\boldsymbol{p}_{Y|X=1}(y)\right|\\
&\leq \sum_z \sum_y \boldsymbol{p}_{Z|Y=y}(z)|\boldsymbol{p}_{Y|X=0}(y) - \boldsymbol{p}_{Y|X=1}(y)|\\
&= \sum_y |\boldsymbol{p}_{Y|X=0}(y) - \boldsymbol{p}_{Y|X=1}(y)|
\end{aligned}
$$

□

Theorem B.0.7 (Sub-additivity of L_c distance) *Let X be a Boolean random variable. Let g be a Boolean function of k random variables $Y_1, \ldots, Y_k$ which are independent given X. Let $Z = g(Y_1, \ldots, Y_k)$. For $c \geq 1$,*

$$||\boldsymbol{p}_{Z|X=0} - \boldsymbol{p}_{Z|X=1}||_c \leq \sum_{i=1}^{k} ||\boldsymbol{p}_{Y_i|X=0} - \boldsymbol{p}_{Y_i|X=1}||_c$$

Proof: For $\boldsymbol{p}$ and $\boldsymbol{q}$ probability distributions over $\{0,1\}$, $||\boldsymbol{p}-\boldsymbol{q}||_c = 2^{1/c-1}||\boldsymbol{p}-\boldsymbol{q}||_1$. Thus it suffices to prove the theorem for $c = 1$.

From lemma B.0.2,

$$||\boldsymbol{p}_{Z|X=0} - \boldsymbol{p}_{Z|X=1}||_1 \leq ||\boldsymbol{p}_{Y_1,\ldots,Y_k|X=0} - \boldsymbol{p}_{Y_1,\ldots,Y_k|X=1}||_1.$$

The theorem follows using lemma B.0.1. □

Appendix C

Bounding Noisy Circuit Size via Information

In this section, we explain how the signal decay theorem is used to obtain a slightly weaker lower bound on noisy circuit size than the one presented in chapter 4. The purpose of presenting this proof is to show an application of the signal decay theorem for general 2-input, m-output channels. The bounds obtained, in this case, are not as strong as those obtained using L_1 distance. This is in contrast to the improved bounds on circuit depth obtained when using mutual information in place of L_1 distance.

Theorem C.0.8 *For $\epsilon \in (0, 1/2]$ and $\delta \in [0, 1/2)$, if a Boolean function f is $(1-\delta)$-reliably computed by a circuit with ϵ-noisy, k-input gates then the number of gates in the circuit is at least*

$$\frac{s \log s + 2s \log(2\Delta)}{k \log t}$$

where s is the sensitivity of the function f, $\Delta = 1 + \delta \log \delta + (1-\delta)\log(1-\delta)$, $t = \frac{\omega^3+(1-\omega)^3}{\omega(1-\omega)}$, and $\omega = \frac{1-\sqrt[k]{1-2\epsilon}}{2}$.

Proof: Without loss of generality, we assume that the function f has maximum sensitivity on the input $\boldsymbol{e}_0$; that $f(\boldsymbol{e}_0) = 0$; and that on input $\boldsymbol{e}_0$, f is sensitive to $x_1, ..., x_s$ (i.e. $f(\boldsymbol{e}_i) = 1$ for $i = 1, ..., s$). Let m_i be the number of input wires from input x_i.

Imagine flipping a fair coin to choose one of two possible tests to perform on a $(1-\delta)$-reliable circuit for f, and let X denote the outcome of the coin flip. One test, when $X = 0$, is to give the circuit the input $\boldsymbol{e}_0$. The other test, when $X = 1$, is to give the circuit the input $\boldsymbol{e}_i$ with probability α_i for $i = 1, \ldots, s$. Since the circuit is $(1-\delta)$-reliable, if $X = 0$ the output Z of circuit is 0 with probability $\geq 1-\delta$. Similarly, if $X = 1$ then $Z = 1$ with probability $\geq 1-\delta$. By Fano's inequality, $I(X;Z) \geq \Delta$.

Let Y be the vector of input values on the noisy input wires from the inputs $x_1, \ldots, x_s$. The circuit does not receive the value X. It only sees Y. Thus Z is independent of X given Y, and the data processing inequality implies that $I(X;Y) \geq I(X;Z)$.

We next upper bound $I(X;Y)$ in order to obtain the theorem. The key observation is that the relation between X and Y is captured by a 2-input, M-output noisy channel where $M = 2^{\Sigma m_i}$. The random variable Y takes one of M possible values depending on the wire noise and the value of X. Let $\boldsymbol{p}_{Y|0}$ and $\boldsymbol{p}_{Y|1}$ denote the distributions on Y given $X = 0$ and $X = 1$ respectively. These distributions must differ significantly so that $I(X;Y) \geq \Delta$. We show using the signal decay theorem for 2-input, M-output channels (theorem 2.5.1) that this information is upper bounded by a function of the wire noise and the number of wires from the inputs.

Since $I(X;X) = 1$, theorem 2.5.1 implies that

$$I(X;Y) = \frac{I(X;Y)}{I(X;X)} \leq 1 - \left(\sum_y \sqrt{\boldsymbol{p}_{Y|0}(y) \boldsymbol{p}_{Y|1}(y)} \right)^2 .$$

Theorem 2.5.1 applies to every 2-input, M-output channel, regardless of the conditional distributions on the input to the channel (conditioned on the value of X). For this application, the input to the channel is the actual value X, and thus the conditional distributions on the input are trivial. One would expect that in this case a better bound holds, though at the current time, we have not examined this possibility.

We now show that the bound obtained using theorem 2.5.1 is itself upper bounded by a function of the L_1 distance between the distributions $\boldsymbol{p}_{Y|0}$ and $\boldsymbol{p}_{Y_1}$. The theorem will then follow by the reasoning presented in theorem 4.2.1.

Let $p_y + \epsilon_y = \boldsymbol{p}_{Y|0}(y)$ and $p_y - \epsilon_y = \boldsymbol{p}_{Y|1}(y)$. Observe that the p_y are non-negative and sum to 1.

$$\begin{aligned} 1 - \left(\sum_y \sqrt{\boldsymbol{p}_{Y|0}(y) \boldsymbol{p}_{Y|1}(y)} \right)^2 &= 1 - \left(\sum_y \sqrt{(p_y + e_y)(p_y - e_y)} \right)^2 \\ &\leq 1 - \left(\sum_y p_y - |e_y| \right)^2 \\ &= 2 \sum_y |e_y| - \left(\sum_y |e_y| \right)^2 \end{aligned}$$

The first term in the bound is the L_1 distance between $\boldsymbol{p}_{Y|0}$ and $\boldsymbol{p}_{Y|1}$ and by the argument in theorem 4.2.1 is at most

$$\left(\sum_{i=1}^{s} \frac{1}{t^{m_i}} \right)^{-1/2} .$$

The second, negative, term is at most $-(1-2\delta)^2$. Taken together, the lower and upper bound on information imply,

$$(\Delta + (1-2\delta)^2)^2 \leq \left(\sum_{i=1}^{s} \frac{1}{t^{m_i}} \right)^{-1}$$

which implies the theorem. □

Bibliography

[1] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, 1991.

[2] R. L. Dobrushin and S. I. Ortyukov. Lower bound for the redundancy of self-correcting arrangements of unreliable functional elements. *Problems of Information Transmission*, 13:59–65, 1977.

[3] R. L. Dobrushin and S. I. Ortyukov. Upper bound for the redundancy of self-correcting arrangements of unreliable functional elements. *Problems of Information Transmission*, 13:203–218, 1977.

[4] M. Dyer, A. Frieze, and R. Kannan. A randomized polynomial time algorithm for approximating the volume of convex bodies. In *Proceedings of the 21st Annual Symposium on Theory of Computing*, pages 375–381, 1989.

[5] P. Elias. Computation in the presence of noise. *IBM Journal of Research and Development*, 2(4):346–353, October 1958.

[6] W. Evans and L. J. Schulman. Signal propagation, with application to a lower bound on the depth of noisy formulas. In *Proceedings of the 34th Annual Symposium on Foundations of Computer Science*, pages 594–603, 1993.

[7] T. Feder. Reliable computation by networks in the presence of noise. *IEEE Transactions on Information Theory*, 35(3):569–571, May 1989.

[8] P. Gács and A. Gál. Lower bounds for the complexity of reliable boolean circuits with noisy gates. *IEEE Transactions on Information Theory*, 40(2):579–583, March 1994.

[9] A. Gál. Lower bounds for the complexity of reliable boolean circuits with noisy gates. In *Proceedings of the 32nd Annual Symposium on Foundations of Computer Science*, pages 594–601, 1991.

[10] R. G. Gallager. *Information Theory and Reliable Communication*. Wiley, 1968.

[11] B. Hajek and T. Weller. On the maximum tolerable noise for reliable computation by formulas. *IEEE Transactions on Information Theory*, 37(2):388–391, March 1991.

[12] R. W. Hamming. Error detecting and error correcting codes. *Bell Syst. Tech. J.*, 29:147–160, April 1950. Also in Key Papers in the Development of Coding Theory, E. R. Berlekamp (Ed), IEEE Press, N.Y., pages 9-12, 1974.

[13] G. Hardy, J. E. Littlewood, and G. Pólya. *Inequalities*. Cambridge University Press, second edition, 1952.

[14] W. S. McCulloch and W. Pitts. A logical calculus of the ideas immanent in nervous activity. *Bull. of Math. Biophysics*, 5:115–133, 1943.

[15] N. Pippenger. On networks of noisy gates. In *Proceedings of the 26th Annual Symposium on Foundations of Computer Science*, pages 30–36, 1985.

[16] N. Pippenger. Reliable computation by formulas in the presence of noise. *IEEE Transactions on Information Theory*, 34(2):194–197, March 1988.

[17] N. Pippenger. Analysis of error correction by majority voting. *Advances in Computing Research*, 5:171–198, 1989.

[18] N. Pippenger. Invariance of complexity measures for networks with unreliable gates. *J. ACM*, 36:531–539, 1989.

[19] N. Pippenger. Developments in "the synthesis of reliable organisms from unreliable components". In *Proceedings of Symposia in Pure Mathematics*, volume 50, pages 311–324, 1990.

[20] N. Pippenger, G. D. Stamoulis, and J. N. Tsitsiklis. On a lower bound for the redundancy of reliable networks with noisy gates. *IEEE Transactions on Information Theory*, 37(3):639–643, May 1991.

[21] Y. Rabinovich, A. Sinclair, and A. Wigderson. Quadratic dynamical systems. In *Proceedings of the 33nd Annual Symposium on Foundations of Computer Science*, pages 304–313, 1992.

[22] R. Reischuk and B. Schmeltz. Reliable computation with noisy circuits and decision trees — a general $n \log n$ lower bound. In *Proceedings of the 32nd Annual Symposium on Foundations of Computer Science*, pages 602–611, 1991.

[23] C. E. Shannon. A mathematical theory of communication. *Bell System Tech. J.*, 27:379–423; 623–656, 1948.

[24] A. Sinclair. *Algorithms for Random Generation and Counting : A Markov Chain Approach*. Progress in Theoretical Computer Science. Birkhäuser, 1993.

[25] A. M. Turing. On the computable numbers, with an application to the Entscheidungsproblem. *Proc. London Math. Soc.*, pages 230–265, 1936.

[26] L. G. Valiant. Short monotone formulae for the majority function. *Journal of Algorithms*, 5:363–366, 1984.

[27] U. Vazirani. Rapidly mixing Markov chains. *Proceedings of Symposia in Applied Mathematics*, 44:99–121, 1991.

[28] J. von Neumann. Probabilistic logics and the synthesis of reliable organisms from unreliable components. In C. E. Shannon and J. McCarthy, editors, *Automata Studies*, pages 43–98. Princeton University Press, 1956.