Signal Propagation, with Application to a Lower Bound on the Depth of Noisy Formulas

William Evans* Leonard J. Schulman†

Department of Computer Science
University of California at Berkeley
Berkeley CA 94720

Abstract

We study the decay of an information signal propagating through a series of noisy channels. We obtain exact bounds on such decay, and as a result provide a new lower bound on the depth of formulas with noisy components. This improves upon previous work of Pippenger and significantly decreases the gap between his lower bound and the classical upper bound of von Neumann. We also discuss connections between our work and the study of mixing rates of Markov chains.

1 Introduction

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 [6]. In this case the computation is a signal propagating through time.

Von Neumann recognized that, rather than being technological and passing, this signal decay was an essential difficulty for large-scale computations. Consequently he was interested in whether, and at what cost, a computer with noisy components might simulate one with ideal, noiseless components [10].

In this paper 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 lemma" for information. Let X be a random variable denoting the message chosen at the source. Let X be input to a communication channel, and let the random variable 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 $I(X;Y)$ (definitions below) 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 lemma states that no matter what the properties of the second channel, $I(X;Z) \leq I(X;Y)$.

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

If the second channel is noisy then one may expect that this inequality be 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 $Y \rightarrow Z$ 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 channels) as a key step in his method for showing a lower bound on the depth, and an upper bound on the

*Research supported in part by NSF grant CCR 92-01092
†Research supported by an NSF postdoctoral fellowship.

maximum tolerable component noise, of noisy formulas [7].

In this paper we improve Pippenger's inequality, and obtain the exact upper bound on the maximum achievable "signal strength ratio" $I(X;Z)/I(X;Y)$, for every binary channel.

As a corollary we improve both Pippenger's lower bounds on depth, and upper bounds on tolerable component noise, for noisy formulas.

1.1 Formula Depth

Among the fundamental concerns in computation are the depth and size of circuits required to compute Boolean functions. 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 considered circuits composed of computational gates each with a bounded number of inputs. In the noisy circuit each gate failed (produced a 0 instead of a 1 or vice versa) independently with probability ϵ.

Von Neumann provided the following positive, but qualified, response to this question: Every circuit with noiseless gates can be simulated by a circuit with noisy gates, whose depth is at most a constant times the depth of the original circuit, provided that the probability of error in each component of the circuit is no more than some $\epsilon < 1/2$. (ϵ depends on the number of inputs to a gate.)

This answer has two especially interesting features. The first is the limit ϵ on component failure, above which the construction fails. The second is that the construction requires a slow-down (i.e. increase in the 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 with out-degree 1) [7]. Shortly afterward Feder extended Pippenger's bound to general noisy circuits [3].

Using more precise information theoretic bounds developed in this paper, we improve Pippenger's result to show:

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

- *If $\xi^2 > 1/k$ then*

$$c > \frac{\log n\Delta}{\log(k\xi^2)}$$

- *If $\xi^2 \leq 1/k$ then $n \leq 1/\Delta$*

This result is the best known except in the case of $k = 3$ where, by different methods, Hajek and Weller have shown that to achieve $\delta < 1/2$ for arbitrary n, ξ must be greater than $2/3$ [5]. This matches the threshold for ξ in von Neumann's construction.

The application of our information theoretic analysis to the lower bound on formula depth follows the outline of Pippenger's argument which, very briefly, has the following structure. First he observes that in order for a function depending essentially on all inputs to be computed with high reliability, the mutual information (in the Shannon sense, defined below) between each input variable and the output must be high. Next he shows that when some intermediate result in the computation is affected by random noise, the mutual information between it and the input strictly decreases. Computation gates can compensate for this loss, up to a point, by combining the information of their predecessors; but the necessity of doing this forces the formula to be large.

Thus the argument depends essentially on two properties of mutual information. First, a subadditivity property. Let X be a random variable representing the value of one input to the formula, with the remaining variables fixed. Subadditivity means that the mutual information between X and the output Y of a gate, before being affected by noise, is no more than the sum of the mutual information between X and the inputs to the gate. Pippenger establishes this claim using the data processing lemma for mutual information, and the fact that in a formula the inputs to the gate are conditionally independent given X.

After the gate performs its computation, its output Y is affected by random noise. Let Z be the resulting random variable. (Z equals Y with probability $(1+\xi)/2$, and $\overline{Y}$ with probability $(1-\xi)/2$.) The second property of mutual information that we need is that the ratio $I(X;Z)/I(X;Y)$ is bounded by a function of the noise parameter ξ. Pippenger shows that

Figure 1: Bounds on $I(X;Z)/I(X;Y)$.

the input signal fed into the path, and the output signal which emerges from the path after being affected by a noisy channel at each level.

With this formulation in mind, we view the input value $X = 0$ or $X = 1$ as the outcome of a random experiment. The random variable Y, representing the value of the signal partway through the path, 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 two conditional distributions are. After passing through a noisy channel, the distributions become less distinguishable and so the mutual information decreases.

Thus our investigation may be viewed in another way. The two possible values $0, 1$ of a random variable along the path may be regarded as the states of a two-state Markov chain, and propagation of the signal through each noisy channel may be viewed as a time step of the Markov chain. In this light, we are interested in showing a rapid mixing rate for the chain. In particular, we are interested in showing that two distributions on $\{0, 1\}$, corresponding to the conditional distributions given $X = 0$ or $X = 1$, quickly become statistically indistinguishable.

Such bounds are usually demonstrated with respect to the L_1, L_2 or L_∞ norm, but it is not obvious that these measures satisfy the subadditivity property which is required in order to decompose the formula into a sum over paths. Nevertheless, by considering the norm *after* processing at the gate, we prove in theorem 4 that indeed any L_c norm, c finite, is subadditive. (For mutual information, subadditivity is

evident already before processing.) Using this theorem, we can show Pippenger's lower bound via these norms, without reference to information theory (see section 6). However, this method does not appear to be strong enough to show the lower bound of theorem 1, which we argue through mutual information.

Mixing rates of large Markov chains have been studied extensively, in terms of combinatorial properties related to the connectivity of the chains. (For background see the survey papers of Vazirani [9] and of Dyer and Frieze [2].) By contrast we focus on the detailed properties of small, connected chains.

2 Definitions

We use p_X to denote a probability distribution on random variable X. Similarly, $p_{Y|X=x}$ or $p_{Y|x}$ denotes a probability distribution on random variable Y conditioned on $X = x$.

The *entropy* of a distribution p is

$$H(p) = -\sum_x p(x) \log p(x).$$

For distributions p and q, the *Kullback Liebler divergence* (or relative entropy) from q to p is

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

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

$$I(X;Y) = \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p_X(x)p_Y(y)}$$

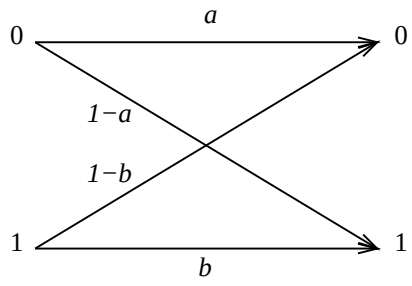
A *binary channel* takes a binary value input and produces an output bit according to a probability distribution which depends on the input value.

If the input to the channel is a random variable X with distribution p_X then the channel outputs a random variable Y whose distribution is $p_Y = p_X \cdot A$. In particular, $p_Y(0) = p_X(0)a + p_X(1)(1-b)$ and $p_Y(1) = p_X(0)(1-a) + p_X(1)b$.

For background on information theory the texts of Gallager [4] and Cover and Thomas [1] as well as Shannon's original paper [8] are recommended.

3 Reduction to Weak Signal Case

Our first step relies upon a geometric interpretation of mutual information. Let $p_{Y|0}$ and $p_{Y|1}$ be the distributions on Y conditional on each input possibility $X = 0, 1$; and let p_Y be the average of these distributions (with the weights $p_X(0), p_X(1)$). From the definitions we have:



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

Figure 2: A general binary channel and its corresponding row-stochastic matrix.

$$\begin{aligned} I(X;Y) &= \sum_x p_X(x) \sum_y p_{Y|x}(y) \log \frac{p_{Y|x}(y)}{p_Y(y)} \\ &= H(p_Y) - \sum_x p_X(x) H(p_{Y|x}) = h \end{aligned}$$

Here h is the altitude marked in figure 3. Thus mutual information can be interpreted as a discrete second derivative of the entropy function H.

Now suppose we pass the random variable Y through a channel A and obtain the output Z. For each $x = 0, 1$, the distribution $p_{Z|x}$ equals $p_{Y|x} \cdot A$. Just as for Y, the mutual information $I(X;Z)$ is the discrete second derivative among the points $H(p_{Z|0})$, $H(p_Z)$ and $H(p_{Z|1})$, where p_Z is the average of $p_{Z|0}$ and $p_{Z|1}$ weighted by $p_X(0)$ and $p_X(1)$. Thus, $I(X;Z)$ is the altitude h' in figure 3. 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 $p_{Y|0}, p_{Y|1}$ and all weights 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 1 *For any strictly concave functions f, g on the interval $[0,1]$, and any $p \in [0,1]$, the ratio*

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

Figure 3: Visualization of $I(X;Y)$ and $I(X;Z)$

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

The lemma holds for more general functions f and g but for brevity we restrict ourselves to the above statement.

Proof: Let x^* and y^* be a closest pair of points which achieve the maximum ratio r. We obtain a contradiction by finding a closer pair x,y which achieve at least ratio r. Say $x^* < y^*$ (otherwise reflect the interval $[0,1]$ about $1/2$).

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 can 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 $p_{Y|X=0}$ and $p_{Y|X=1}$ which are almost indistinguishable:

Corollary 1 *The ratio* $I(X;Z)/I(X;Y)$ *is maximized in the limit* $|p_{Y|X=0} - p_{Y|X=1}| \to 0$.

(Recall that $p_{Y|X=0}$ and $p_{Y|X=1}$ correspond to points on the unit interval. The distance function is induced from the interval.)

Proof: Fix any weights $p_X(0)$ and $p_X(1)$. Then $I(X;Y)$ and $I(X;Z)$ are the discrete second derivatives of strictly concave functions, namely the restrictions of the entropy function to various subintervals of $[0,1]$. □

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 $p_{Y|X=0}$ and $p_{Y|X=1}$ 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.

4 Signal Decay

In the previous section, we showed that the ratio $r = I(X;Z)/I(X;Y)$ is maximized for a pair of infinitesimally close distributions. This greatly simplifies the task of identifying the maximizing distributions since instead of having to consider two independent parameters (specifying the distributions), we can range over just one parameter (specifying one of the distributions), and express r with a series expansion in terms of the distance between the two distributions.

Another simplification that results from restricting to the case of infinitesimally separated distributions is that, if p and $p+\epsilon$ represent a close pair of distributions on Y, then

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

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

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 reparameterization and then solve for the maximum.

Let p be a probability distribution, $p = (p(0), p(1))$. Define $\sigma(p) = (\sqrt{p(0)}, \sqrt{p(1)})$. Geometrically σ 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 $(0,1)$.

Write the L_2 distance of two vectors $u = (u(0), u(1))$ and $v = (v(0), v(1))$ in $\Re^2$ as: $\|u-v\|_2 = \left(\sum_{i=0,1}(u(i)-v(i))^2\right)^{1/2}$.

Let $\epsilon = (\epsilon(0), \epsilon(1))$ be such that $\epsilon(0) + \epsilon(1) = 0$. Thus both p and $p+\epsilon$ are probability distributions. If $\epsilon(0), \epsilon(1) << p(0), p(1)$ then $D(p+\epsilon\|p)$ is approximated by a power series expansion in ϵ. However, the coefficients of this expansion vary depending on p. The map σ has the property that the first term of the power series expansion no longer has a dependence on the probability distribution about which the expansion is being taken. In fact, after reparameterization by σ, the first term in the divergence is simply proportional to the square of the L_2 distance between the two vectors on the circle corresponding to the two distributions:

Lemma 2 $D(p+\epsilon\|p) = \frac{1}{4}\|\sigma(p+\epsilon) - \sigma(p)\|_2^2 + O(\epsilon^3)$

Proof: By series expansion. □

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

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

Theorem 2 *Let X and Y be boolean random variables. Let the channel A be*

$$A = \begin{bmatrix} a & 1-a \\ 1-b & b \end{bmatrix} = \begin{bmatrix} \sin^2\alpha & \cos^2\alpha \\ \sin^2\beta & \cos^2\beta \end{bmatrix}$$

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

$$I(X;Z)/I(X;Y) \le \sin^2(\alpha-\beta).$$

Observe that α and β are the angles (at the origin) to the points specifying the most extreme possible distributions on Z.

Proof: As discussed, it suffices to show that

$$\frac{D((p+\epsilon)\cdot A\|p\cdot A)}{D(p+\epsilon\|p)} \le \sin^2(\alpha-\beta) \qquad (1)$$

for any distribution $p = (p(0), p(1))$ on Y. The resulting distribution on Z is $p_Z(0) = p(0)\sin^2\alpha + (1-$

$p(0))\sin^2\beta, p_Z(1) = 1 - p(0)\sin^2\alpha - (1-p(0))\sin^2\beta$. Substituting A into the ratio (1), and expanding in terms of ϵ, we find that

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

By differentiation one can determine that this expression is maximized for the distribution p specified by

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

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

5 Depth of Noisy Formulas

Let f be a Boolean function which depends essentially on n arguments. Pippenger's argument shows that any formula 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 the noise level. This implies a lower bound on the factor by which the depth of a formula 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 essentially on k inputs, and no gate that depends essentially on more than k inputs. The function f which is the d-fold composition of g depends essentially on $n = k^d$ inputs and can be computed by a depth d formula in the perfect gate model. Pippenger's result implies a ratio of R between the depths of formulas for f in the noisy-gate and perfect-gate models.

Theorem 3 (Pippenger) *Let f be a function which depends essentially on n inputs. Let F be a formula of depth c using gates with at most k inputs, where each gate fails independently with probability $(1-\xi)/2 > 0$. Suppose F computes the function f with probability $\geq 1-\delta$, where $\delta < 1/2$. Let $\Delta = 1 - H(\delta)$.*

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

The idea of the proof is to show that for each input to F the following two conditions hold. On one hand, since F correctly computes f with high probability, the information between the output and the input must be large. On the other hand, since each gate fails with probability $(1-\xi)/2$, the information between the output and the input along any one path between them of length l is exponentially small in l. Since each gate in F has at most k inputs, this implies that the depth of F must be large in order to have many paths between each input and the output.

Notice that every gate increases the distance (path length) of its inputs to the output. However, it also increases the number of paths from inputs to 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 sustainable computation and we cannot compute functions of an arbitrary number of inputs.

Using theorem 2, we improve on Pippenger's bounds for the threshold and for the factor by which the depth must increase.

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

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

Proof: The proof method follows Pippenger, but the proof is included in its entirety for completeness (see the appendix). □

Our result improves on Pippenger's in two ways. First, we increase the threshold below which computation in the noisy gate model is infeasible. For $k = 3$ this threshold is known exactly. Von Neumann shows that a noisy formula which is correct with probability $1 - \delta > 1/2$ is possible if $\xi > 2/3$. Hajek and Weller show that such computation for arbitrary n is impossible if $\xi \leq 2/3$. Their result applies only to $k = 3$, therefore the best lower bound on the threshold for $k > 3$ is Pippenger's bound of $\xi > 1/k$. We improve this bound to $\xi > 1/\sqrt{k}$.

Second, we increase the factor by which the depth of the noisy formula must increase. To compute a function which depends essentially on n inputs, Pippenger shows that a noisy formula 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 $1/\log(k\xi^2)$.

6 Subadditivity under any L_c Norm

Both Pippenger's result and our result rely on the subadditivity of mutual information. That is, the mutual information between the output Y of a gate and a

random variable X, is at most the sum over the inputs Y_i to the gate, of the mutual information between Y_i and X (assuming that the Y_i are conditionally independent given X). The corresponding statement for L_c norms, c finite, is that the L_c distance between the conditional distributions $p_{Y|X=0}$ and $p_{Y|X=1}$ is at most the sum of the L_c distances between the conditional distributions $p_{Y_i|X=0}$ and $p_{Y_i|X=1}$. For our purpose we wish subadditivity to hold for any gate, regardless of the Boolean function computed by the gate.

Recall that the L_c distance of two vectors p, q each of length 2, is

$$\|p-q\|_c = (|p(0)-q(0)|^c + |p(1)-q(1)|^c)^{1/c}.$$

The following theorem proves that all L_c, for finite c, have the subadditivity property. The proof is fundamentally different from the proof of the subadditivity of mutual information. In the case of mutual information, one can show that the random variable which is the cross product of the inputs to the gate, has at most the sum of the information at those inputs. Thus subadditivity is shown *before* the processing at the gate. Then by the data processing lemma, any computation performed by the gate will not increase the information.

In the case of L_c norms, the argument does not include a data processing lemma. Rather, given the set of conditional distributions at the inputs, we identify the gate whose computation boosts the norm of the conditional distributions at the output the most. We then prove subadditivity of the norm *after* the processing at the gate.

Theorem 4 *Let g be a Boolean function of k inputs, each a random variable Y_i. Let $p_i = p_{Y_i|X=0}$ and $q_i = p_{Y_i|X=1}$ be the conditional distributions on Y_i given the value of X, and assume that the p_i are mutually independent; as are the q_i. Let $Z = g(Y_1, \cdots, Y_k)$. For any finite c,*

$$\|p_{Z|X=0} - p_{Z|X=1}\|_c \leq \sum_{i=1}^{k} \|p_i - q_i\|_c$$

Proof:

For p and q probability distributions over $\{0,1\}$, $\|p-q\|_c = \frac{2^{1/c}}{2}\|p-q\|_1$. Thus we need only prove the theorem for $c=1$.

Since the conditional distributions p_i (and q_i) are independent, the conditional distribution on the product $Y_1 \times \cdots \times Y_k$ given X is the product of the individual distributions p_i (or q_i). For example, the probability of $Y_1 = y_1, \cdots, Y_k = y_k$ given $X = 0$ is $\prod p_i(y_i)$. Let $p(y) = \prod p_i(y_i)$ and $q(y) = \prod q_i(y_i)$ where $y = y_1 \cdots y_k$ (p and q are probability distributions over $\{0,1\}^k$).

It follows that the maximum value of $\|p_{Z|X=0} - p_{Z|X=1}\|_1$ is achieved for the function (gate) g which assigns 0 to all vectors $y \in \{0,1\}^k$ with $p(y) > q(y)$, and assigns 1 to all vectors y with $p(y) \leq q(y)$.

We prove the theorem by induction on k. The case $k=1$ is trivial. Suppose the theorem holds for k inputs. We add a new input Y_{k+1} with conditional distributions p_{k+1} given $X=0$ and q_{k+1} given $X=1$. The right side of the inequality increases by $\|p_{k+1} - q_{k+1}\|_1$. Thus to prove the inequality we need only show that

$$\begin{aligned}
&\sum_{y\in\{0,1\}^k} (|p_{k+1}(0)p(y) - q_{k+1}(0)q(y)| \\
&\quad + |p_{k+1}(1)p(y) - q_{k+1}(1)q(y)|) \\
&\leq \sum_{y\in\{0,1\}^k} |p(y)-q(y)| + \|p_{k+1} - q_{k+1}\|_1
\end{aligned}$$

This is true if, for all y,

$$\begin{aligned}
&|p_{k+1}(0)p(y) - q_{k+1}(0)q(y)| \qquad (2)\\
&\quad + |p_{k+1}(1)p(y) - q_{k+1}(1)q(y)| \\
&\leq |p(y)-q(y)| + \min\{p(y), q(y)\}\|p_{k+1} - q_{k+1}\|_1
\end{aligned}$$

since $\sum_y \min\{p(y), q(y)\} \leq 1$. We now show (2) is true by case analysis. To simplify notation, let $a = p(y)$, $b = q(y)$, $r = p_{k+1}(0)$, and $s = q_{k+1}(0)$. Thus we wish to show that

$$\begin{aligned}
&|ra - sb| + |(1-r)a - (1-s)b| \qquad (3)\\
&\leq |a-b| + 2\min\{a,b\}|r-s|
\end{aligned}$$

First, we may assume without loss of generality that $a > b$. Observe that $ra - sb$ and $(1-r)a - (1-s)b$ cannot both be negative because that would imply $a < b$. We are left with three cases.

Case 1: $ra - sb$ and $(1-r)a - (1-s)b$ are both positive. In this case, the sum of their absolute values is $a-b$ which cancels with the $a-b$ on the right side of (3) leaving $0 \leq 2|r-s|\min\{a,b\}$.

Case 2: $ra - sb$ is negative and $(1-r)a - (1-s)b$ is positive. In this case,

$$\begin{aligned}
&|ra - sb| + |(1-r)a - (1-s)b| - |a-b| \\
&= sb - ra + (1-r)a - (1-s)b - a + b \\
&= -2ra + 2sb \leq 2b(s-r)
\end{aligned}$$

$$\leq 2|r-s|\min\{a,b\}$$

Case 3: Follows from case 2 by symmetry with $1-r$ replacing r, and $1-s$ replacing s. □

The L_c distance between any pair of distributions, upon being passed through a noisy channel, decreases by a factor of ξ. Hence the L_c norm 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, provided by Pippenger. Thus that result can be argued with the tools generally employed in the study of mixing of Markov chains. In addition this approach is technically attractive as we need not invoke a probability distribution on each input X. However the results of the present paper, which depend upon a ξ^2 drop in the signal strength, appear to be beyond the reach of the argument through L_c norms.

Acknowledgments

Thanks to N. Pippenger for helpful consultations.

References

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

[2] M. Dyer and A. Frieze. Computing the volume of convex bodies: A case where randomness provably helps. *Proceedings of Symposia in Applied Mathematics*, 44:123–169, 1991.

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

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

[5] 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.

[6] 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.

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

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

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

[10] 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.

A Proof of Theorem 1

Let $x_1 \cdots x_n$ be the inputs to the function f. Since f depends essentially 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 $\bar{x_i}$. Let F_i be the formula F 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 $F_i(X)$ be the random variable which is the output of F_i when $x_i = X$. Note that since F_i contains noisy gates, the random variable $F_i(X)$ is not determined entirely by X. However, by our assumption that F is correct with high probability, either $\Pr(F_i(X) = X) \geq 1-\delta$ or $\Pr(F_i(X) = \bar{X}) \geq 1-\delta$. By Fano's Lemma (see [1]),

$$I(F_i(X); X) \geq \Delta \tag{4}$$

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

To upper bound this information, it is convenient to think of a formula G with random input X as a k-ary tree in which leaves correspond to the input X or the constants 0 or 1 and internal nodes correspond to noisy gates. We claim that

$$I(G(X); X) \leq \sum_{P \in G} \xi^{2|P|} \tag{5}$$

where P is the set of paths in G from leaves with input X to the root and $|P|$ is the number of nodes (gates) along this path.

The proof is by induction on the number of gates in G. If G is a constant leaf, both sides of the inequality vanish. If G is a X leaf then $I(G(X); X) = H(p_X) = 1$ and $|P| = 0$ so $\sum_P \xi^{2|P|} = 1$. Otherwise, let $G_1(X) \cdots G_k(X)$ be the inputs to the gate at the root of G. Let g be the function computed by this gate, in the absence of noise.

$$\begin{aligned} & I(g(G_1(X), \cdots, G_k(X)); X) \\ & \leq I(G_1(X), \cdots, G_k(X); X) \\ & \leq \sum_{i=1}^{k} I(G_i(X); X) \end{aligned}$$

The first inequality follows from the data processing lemma. The second inequality uses the fact that, since the gates fail independently and G is a formula, $G_1(X) \cdots G_k(X)$ are conditionally independent given X.

Since the gate is noisy, its output $G(X)$ is $g(G_1(X), \cdots, G_k(X))$ complimented with probability $(1-\xi)/2$. This corresponds to passing the result of g through the binary symmetric channel

$$\begin{bmatrix} \frac{1+\xi}{2} & \frac{1-\xi}{2} \\ \frac{1-\xi}{2} & \frac{1+\xi}{2} \end{bmatrix}$$

By theorem 2,

$$I(G(X); X) \leq \xi^2 I(g(G_1(X), \cdots, G_k(X)); X)$$
$$\leq \xi^2 \sum_{i=1}^{k} I(G_i(X); X)$$

Applying the inductive hypothesis to each term of the sum, we obtain the required bound.

Combining the bounds (4) and (5) and summing over all F_i gives

$$n\Delta \leq \sum_{P \in F} \xi^{2|P|} \tag{6}$$

The theorem follows easily. First suppose $\xi^2 > 1/k$. Then

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

where c is the depth of F. This says that when $\xi^2 > 1/k$, the expression $\sum_{P \in F} \xi^{2|P|}$ is maximized for F equal to the complete k-ary tree of depth c. If F is not complete, adding k children to a leaf at depth $l < c$ increases the sum by $k\xi^{2(l+1)} - \xi^{2l}$. Since $\xi^2 > 1/k$ this is strictly positive. Note that even if we allow gates with less than k inputs, the bound still holds. Combining this with (6), we obtain

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

which implies the first result of the theorem.

For the second result, suppose $\xi^2 \leq 1/k$. We claim that there exists $1 \leq i \leq n$ such that

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

The claim follows by an averaging argument and the fact that $\sum_{P \in F} 1/k^{|P|} \leq 1$ (which can be proven by induction).

Combining (4) and (5) with the above claim, we obtain

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

which implies the second result of the theorem. □