Signal Propagation and Noisy Circuits

William S. Evans, *Member, IEEE*, and Leonard J. Schulman

***Abstract*— The information carried by a signal decays when the signal is corrupted by random noise. This occurs when a message is transmitted over a noisy channel, 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 rate at which information decreases as a signal 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 is required to 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. We provide a lower bound on the multiplicative increase in circuit depth necessary for reliable computation, and an upper bound on the maximum level of noise at which reliable computation is possible.**

A preliminary version of this work appeared in the first author's thesis [1].

***Index Terms*—Data processing inequality, mutual information, noisy circuit complexity.**

I. 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 [2].

Von Neumann recognized that, rather than being technological and passing, this signal decay was an essential difficulty for large-scale computations, that by their nature rely on the propagation of long chains of events [3]. Von Neumann's goal was to subject noisy computation to the same thermodynamical treatment that communication had received in the contemporary work of Shannon [4]. Surprisingly, it took over 35 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 [5].

In this paper, we investigate the propagation of information signals in noisy media. We study a basic question that is relevant to any such propagation, whether in communication or in computation. To set the framework we recall the well-known "data processing inequality" 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 inequality states that no matter what the properties of the second channel, $I(X;Z) \leq I(X;Y)$

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

If the second channel is noisy then one may expect that this inequality will be strict, and further, that the signal decay will 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 $I(X;Z)/I(X;Y)$.

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 that 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 symmetric binary 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 [5].

Manuscript received October 1, 1997; revised May 17, 1999. The work of W. S. Evans was supported in part by NSF under Grant CCR 92-01092. The work of L. J. Schulman was supported by an NSF Post-Doctoral Fellowship. The material in this paper was presented in part at the 34th Symposium on Foundations of Computer Science, Palo Alto, CA, 1993 and at the IEEE International Symposium on Information Theory, Whistler, BC, Canada, 1995.

W. Evans is with the Department of Computer Science, University of Arizona, Tucson, AZ USA (e-mail: will@cs.arizona.edu).

L. Schulman is with the College of Computing, Georgia Institute of Technology, Atlanta, GA USA (e-mail: schulman@cc.gatech.edu).

Communicated by R. L. Cruz, Associate Edotpr for Communication Networks.

Publisher Item Identifier S 0018-9448(99)08521-1.

In this paper we improve Pippenger's inequality, and obtain the exact upper bound on the maximum achievable "information propagation factor" $I(X; Z)/I(X; Y)$, for any binary channel. This may be considered a *quantified* data processing inequality. The inequality is also shown to hold under certain conditioning events, and in this form, we employ it to obtain lower bounds on the complexity of reliable circuits with noisy components.

A. Circuit Depth

We apply our bound on the information propagation factor to obtain lower bounds on the depth of *noisy circuits*. Von Neumann introduced this model of computation in an attempt to capture the limitations of physical circuits. In his definition, a noisy circuit is composed of gates that fail (produce a 0 instead of a 1 or *vice versa*) independently with probability ϵ. This is the definition we adopt. It is, perhaps, unreasonable to assume that a physical circuit can rely on its gates to fail with exact probability ϵ. Alternative noisy circuit models that weaken this assumption have been proposed [6]. Our goal, however, is to show lower bounds, for which the strong von Neumann model is an appropriate choice since the lower bounds automatically apply to all weaker models.

To study the limitations of physical circuits, von Neumann asked whether noisy circuits can compute the same functions as circuits with noiseless gates; and if so, at what cost in depth (latency)? 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 ϵ, the probability of error in each component of the circuit, is less than some ϵ_0. (Von Neumann's construction using three-input majority gates required $\epsilon < 0.0073$, but, as he argued, $\epsilon_0 = 1/6$ is the true limit of his method.) The simulation is, of course, not perfect. The guarantee is only that the noisy circuit is δ-*reliable*; that it produces the correct answer on every input with probability at least $1 - \delta$ for a fixed $\delta < 1/2$.

This answer has two especially interesting features. The first is the existence of a limit ϵ_0 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 for noisy formulas (circuits whose gates have out-degree 1) [5]. Shortly afterward, Feder extended Pippenger's bound to general noisy circuits [7].

In this paper, we improve both Pippenger's and Feder's results. The key component in the improved result is our precise bound on the information propagation factor. We discuss this bound in Section III. We then discuss the lower bound on circuit depth in Section IV. First we introduce some notation.

II. Notation

We use $\underline{\Pr}\{X\}$ to denote $(\Pr\{X = 0\}, \Pr\{X = 1\})$, the probability distribution on the random variable X. The entropy of a distribution $\underline{\Pr}\{X\}$ is denoted $H(\underline{\Pr}\{X\})$ or $H(X)$, and in the special case of a binary-valued random variable with distribution $(q, 1-q)$ we abbreviate by $H(q) = H(q, 1-q)$. A *binary channel* is characterized by a row-stochastic matrix $A = \left(\begin{smallmatrix} 1-a & a \\ b & 1-b \end{smallmatrix}\right)$ (Fig. 1). Let Y denote the input random variable, and Z the output random variable of a binary channel. Conditional on input 0, the output distribution is $\underline{\Pr}\{Z|Y = 0\} = (1, 0) \cdot A$; conditional on input 1 it is $\underline{\Pr}\{Z|Y = 1\} = (0, 1) \cdot A$; and given input distribution $\underline{\Pr}\{Y\}$, it is the weighted combination $\underline{\Pr}\{Z\} = \underline{\Pr}\{Y\} \cdot A$.

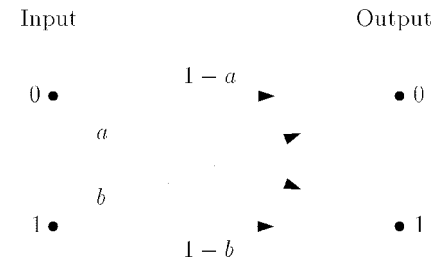

Fig. 1. Binary channel.

III. Quantified Data Processing Inequality

Our first step relies upon a geometric interpretation of mutual information. The mutual information between X and Y is

$$\begin{aligned} I(X;Y) &= \sum_x \Pr\{X=x\} \sum_y \Pr\{Y=y|X=x\} \\ &\quad \cdot \log \frac{\Pr\{Y=y|X=x\}}{\Pr\{Y=y\}} \\ &= H(Y) - \sum_x \Pr\{X=x\} H(Y|X=x). \end{aligned}$$

The distribution $\underline{\Pr}\{Y\}$ is the weighted average of the distributions $\underline{\Pr}\{Y|X = x\}$. Consider a distribution as an element of the hyperplane of points whose coordinates sum to 1. The entropy function H defines a surface above this hyperplane. The mutual information between X and Y is the difference between the height of this surface at $\underline{\Pr}\{Y\}$, and the averaged height

$$\sum_x \Pr\{X=x\} H(Y|X=x).$$

In the case of binary-valued distributions, the mutual information $I(X; Y)$ is simply the height α of the entropy surface above the line passing through $H(\Pr\{Y = 0|X = 0\})$ and $H(\Pr\{Y = 0|X = 1\})$, at the point $\Pr\{Y = 0\}$ (Fig. 2).

Define the *discrete second derivative* of a function f on $\mathbb{R}$ to be

$$f_2(x, y, p) = pf(x) + (1-p)f(y) - f(px + (1-p)y)$$

for x, y in the domain of f and $p \in [0, 1]$. Observe that for binary-valued X

$$\begin{aligned} &I(X;Y) \\ &\quad = -H_2(\Pr\{Y=0|X=0\} \Pr\{Y=0|X=1\}, \Pr\{X=0\}). \end{aligned}$$

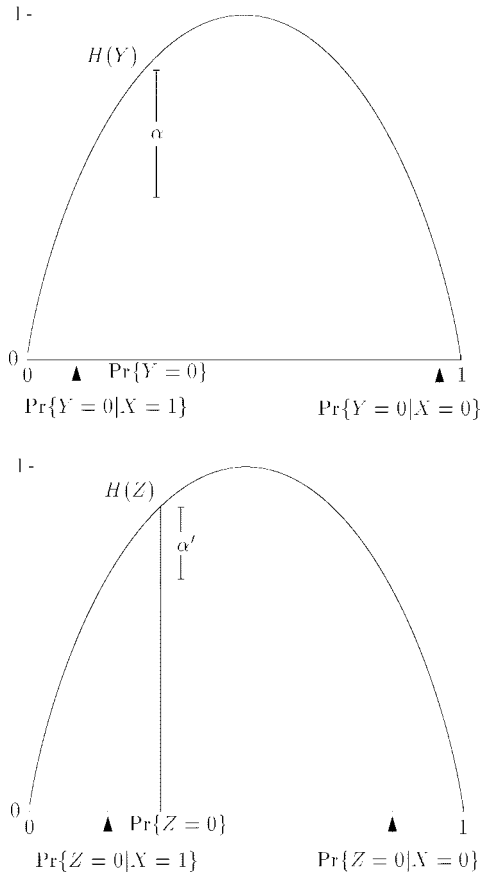

Fig. 2. $I(X;\, Y)$ and $I(X;\, Z)$.

(The definition obviously extends beyond binary-valued X but this will not be needed in the paper.) Thus mutual information is (after reversing sign) a discrete second derivative of the entropy function.

If we input the random variable Y to a channel A, we obtain an output variable Z with conditional distributions

$$\underline{\Pr}\{Z|X=0\} = \underline{\Pr}\{Y|X=0\} \cdot A$$

and

$$\underline{\Pr}\{Z|X=1\} = \underline{\Pr}\{Y|X=1\} \cdot A$$

and overall distribution

$$\begin{aligned}\underline{\Pr}\{Z\} = {} & \Pr\{X=0\}\underline{\Pr}\{Z|X=0\} \\ & + \Pr\{X=1\}\underline{\Pr}\{Z|X=1\}.\end{aligned}$$

Just as for $I(X;\, Y)$, the mutual information $I(X;\, Z)$ is given by a discrete second derivative

$$\begin{aligned}&I(X;\, Z) \\ &\quad = -H_2(\Pr\{Z=0|X=0\}, \Pr\{Z=0|X=1\}, \Pr\{X=0\})\end{aligned}$$

(α', Fig. 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 $\underline{\Pr}\{Y|X=0\}$, $\underline{\Pr}\{Y|X=1\}$, and all weights $\underline{\Pr}\{X\}$ of the ratio α'/α.

We will find the maximum ratio α'/α 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.

Lemma 1: If the functions $f, g : [0, 1] \to \mathbb{R}$ have negative second derivatives on $(0, 1)$ then

$$\sup_{\substack{x \neq y \in [0,1] \\ p \in (0,1)}} \frac{f_2(x, y, p)}{g_2(x, y, p)} = \sup_{t \in (0,1)} \frac{f''(t)}{g''(t)}.$$

Equality is attained in the limit $|x - y| \to 0$ where x and y approach a value of t achieving $\sup_{t \in (0,1)}(f''(t)/g''(t))$.

Proof: Let

$$c = \sup_{t \in (0,1)} (f''(t)/g''(t)).$$

If c is finite, observe that $(cg - f)'' = cg'' - f'' \leq 0$ on $(0, 1)$. This implies $cg - f$ is concave on $[0, 1]$, thus $(cg - f)_2(x, y, p) \leq 0$ (for $x \neq y \in [0, 1]$, $p \in (0, 1)$) and, consequently, $f_2(x, y, p)/g_2(x, y, p) \leq c$. If c is infinite this inequality is trivial. Equality of the suprema is observed by taking a series of points t for which $f''(t)/g''(t)$ approaches the limit c (finite or infinite). For each point $t \in (0, 1)$

$$\lim_{h \to 0} \frac{f_2(t, t+h, p)}{g_2(t, t+h, p)} = \frac{f''(t)}{g''(t)}. \qquad \square$$

We employ the lemma with $f(t) = H((t, 1-t) \cdot A)$ and $g(t) = H(t, 1-t)$. We have

$$I(X;\, Z) = -H_2(z_0, z_1, p) = -f_2(y_0, y_1, p)$$

and

$$I(X;\, Y) = -H_2(y_0, y_1, p) = -g_2(y_0, y_1, p)$$

where $p = \Pr\{X = 0\}$, $y_i = \Pr\{Y = 0|X = i\}$, and $z_i = \Pr\{Z = 0|X = i\}$. Since H is strictly concave, the lemma implies that the information propagation factor $I(X;\, Z)/I(X;\, Y)$ is maximized for pairs of distributions $\underline{\Pr}\{Y|X=0\}$ and $\underline{\Pr}\{Y|X=1\}$ that are almost indistinguishable

$$|\Pr\{Y=0|X=0\} - \Pr\{Y=0|X=1\}| \to 0.$$

In fact, 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 maximum ratio is achieved only in the limit of very close distributions. Thus a (nontrivial) noisy channel performs at its peak efficiency only when it is carrying a very weak signal.

For example, suppose we transmit one bit of information over a long cable and each meter of the cable introduces some random noise that is symmetric in the sense that it affects 0's and 1's with the same probability. We will later see that in this symmetric case, the information propagation factor is maximized when each of the distributions $\underline{\Pr}\{Y|X=0\}$ and $\underline{\Pr}\{Y|X=1\}$ are asymptotically close to the uniform distribution (0's and 1's equally likely). This is also the distribution each signal approaches as it travels along this cable. Lemma 1 implies that the greatest rate of information loss (the smallest information propagation factor) occurs in the first part of the cable. For a cable, this can also be observed

just by examining powers of the matrix describing a short stretch of cable, but the lemma carries the conclusion also to more complicated cases in which information is recombined, as in a circuit. One may also conclude that, in certain cases, if several signals carry information about an event, it may be best to propagate each signal separately rather than combine the information into a single, clearer signal. This is because the information carried by each separate, weak signal can propagate at close to the maximum propagation factor, while the information carried by a strong signal decays more rapidly. (The particulars of the case must be considered, however, since only certain weak signals approach the minimum loss.)

Theorem 1: Let X and Y be binary random variables. Let the channel A be

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

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

$$\frac{I(X;\, Z)}{I(X;\, Y)} \leq \sin^2 \theta$$

where θ is the angle between the vectors $(\sqrt{1-a},\, \sqrt{a})$ and $(\sqrt{b},\, \sqrt{1-b})$.

Proof: Let f and g be as above. Let

$$A(t) = (1-a)t + b(1-t).$$

The ratio

$$f''(t)/g''(t) = (1-a-b)^2 \frac{t(1-t)}{A(t)(1-A(t))}$$

is maximized at

$$t = \sqrt{b(1-b)}/(\sqrt{b(1-b)} + \sqrt{a(1-a)}).$$

The value of the ratio for this value of t is

$$1 - (\sqrt{b(1-a)} + \sqrt{a(1-b)})^2.$$

Now Lemma 1 implies

$$\begin{aligned} \frac{I(X;\, Z)}{I(X;\, Y)} &= \frac{f_2(y_0,\, y_1,\, p)}{g_2(y_0,\, y_1,\, p)} \\ &\leq 1 - \left(\sqrt{b(1-a)} + \sqrt{a(1-b)}\right)^2 \\ &= 1 - \cos^2 \theta = \sin^2 \theta. \end{aligned}$$

□

Note that for symmetric channels the maximum occurs at $t = 1/2$, implying that the conditional distributions on Y given X for which the information propagation factor is maximized, are close to the uniform distribution, which is also the stationary distribution. For asymmetric channels, the maximum occurs away from the stationary distribution.

Theorem 1 extends in a useful way under certain conditioning events: if Q is a random variable such that Z is independent of $(Q,\, X)$ given Y, then the theorem holds under conditioning by Q.

Corollary 1: Let X and Y be binary random variables. Let the channel A be

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

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

$$\frac{I(X;\, Z|Q)}{I(X;\, Y|Q)} \leq \sin^2 \theta$$

where θ is the angle between the vectors $(\sqrt{1-a},\, \sqrt{a})$ and $(\sqrt{b},\, \sqrt{1-b})$.

Proof: Since Z is independent of $(Q,\, X)$ given Y, $\underline{\Pr}\{Z|QXY\} = \underline{\Pr}\{Z|Y\}$ and thus

$$\underline{\Pr}\{Z|Q=q,\, X=x\} = \underline{\Pr}\{Y|Q=q,\, X=x\} \cdot A$$

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 1. It follows from the theorem that

$$\frac{I(X;\, Z|Q=q)}{I(X;\, Y|Q=q)} \leq \sin^2 \theta.$$

The corollary follows since

$$\begin{aligned} \frac{I(X;\, Z|Q)}{I(X;\, Y|Q)} &= \frac{\sum_q \Pr\{Q=q\} I(X;\, Z|Q=q)}{\sum_q \Pr\{Q=q\} I(X;\, Y|Q=q)} \\ &\leq \max_q \frac{I(X;\, Z|Q=q)}{I(X;\, Y|Q=q)} \leq \sin^2 \theta. \end{aligned}$$

□

IV. Noisy Circuit Depth

Our lower bound on circuit depth follows the general outline of Pippenger's lower bound on formula depth [5]. The complications introduced by adopting a circuit rather than a formula model require a careful application of the conditioned version of the quantified data processing Theorem 1. Using this theorem also results in a better lower bound than that obtained by either Pippenger or Feder [7]. We begin with a sketch of Pippenger's argument.

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 that is highly correlated with X. By Fano's lemma, if X is a random variable then the mutual information carried by the output about X must be high.

On the other hand, one shows that 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 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. Pippenger established this for formulas, by showing that the total information sent is bounded by the sum of the information sent over each path from X to the output.

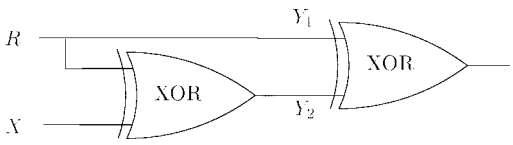

Fig. 3. If R is the output of a very long chain of noisy gates then R is essentially a random bit. Thus $I(Y_1; X) + I(Y_2; X)$ is close to zero, while $I(Y1, Y2; X)$ is close to 1.

This supports the view of information as a kind of fluid that 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.

Such a statement requires two inequalities to hold. One of the inequalities is the data processing inequality which states that $I(Y; X) \leq I(Y_1, \cdots, Y_k; X)$ where $Y_1, \cdots, Y_k$ are the inputs to a gate with prenoise output Y. This of course holds for circuits as well. The second inequality is $I(Y_1, \cdots, Y_k; X) \leq \sum_i I(Y_i; X)$. This holds for formulas since the Y_i are mutually independent given X; but it may not be true for circuits. Fig. 3 shows a circuit in which $I(Y_1, \cdots, Y_k; X)$ is 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 for circuits.

Rather than going through the intermediate step of decomposing the circuit into disjoint paths, we directly upper-bound the information between the values carried by any set of wires and the input X. For circuits composed of gates that err with probability $(1-\xi)/2$ ("$(1-\xi)/2$-noisy gates"), the bound we obtain is 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 of the circuit is bounded by the sum of $\xi^{2|P|}$ over all paths P from that input to the output. Since we consider a set of paths, rather than an individual path, the argument for the ξ^2 drop in information at every noisy gate is more complicated; this is addressed using Corollary 1.

Lemma 2: Let G be a circuit composed of $(1-\xi)/2$-noisy gates. Suppose each input to G is X (a binary 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 of the circuit, inputs to the circuit (X and constants 0 and 1), and the output terminal; and whose edges are wires. Direct a wire (edge) from vertex h to g if the output of h is the input of g. Number the input vertices 0 and number the gate vertices distinctly from 1 to the number of gates in G so that each wire starts from its smaller numbered endpoint. Such 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 that 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 that contain wires numbered $\leq t$. Consider a set W containing wires numbered $\leq t+1$. Let Z be the binary random value carried by the wires numbered $t+1$ in W. (Several wires may be numbered $t+1$ if gate $t+1$ has several outputs.) Since each gate has a distinct number and noise occurs at the gate, Z is well defined. Let $W_1, \cdots, W_m$ be the wires in W numbered $\leq t$. Now $I(W; X) = I(Z, W_1, \cdots, W_m; X)$. Expanding we have

$$\begin{aligned} &I(Z, W_1, \cdots, W_m; X) \\ &\qquad = I(Z; X|W_1, \cdots, W_m) + I(W_1, \cdots, W_m; X). \end{aligned}$$

Let Y be the prenoise 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 $W_1, \cdots, W_m$, since they are the output of gates numbered $\leq t$, are independent of Z given Y. Thus Corollary 1 implies

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

since the square of the sine of the angle between $(\sqrt{1-a}, \sqrt{a})$ and $(\sqrt{b}, \sqrt{1-b})$ for $a = b = (1-\xi)/2$ is ξ^2.

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

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

Therefore,

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

Since $Y_1, \cdots, Y_k$ 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

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

$$\leq \sum_{\substack{P \text{ from } X \text{ to} \\ \{Y_1,\cdots,Y_k\}}} \xi^{2(|P|+1)} + \sum_{\substack{P \text{ from } X \text{ to} \\ \{W_1,\cdots,W_m\}}} \xi^{2|P|}$$

$$= \sum_{P \text{ from } X \text{ to } Z} \xi^{2|P|} + \sum_{\substack{P \text{ from } X \text{ to} \\ \{W_1,\cdots,W_m\}}} \xi^{2|P|}$$

$$= \sum_{\substack{P \text{ from } X \text{ to} \\ \{Z,W_1,\cdots,W_m\}}} \xi^{2|P|} \leq \sum_{P \text{ from } X \text{ to } W} \xi^{2|P|}. \qquad \square$$

A. Noisy Circuit Depth Lower Bound

We say that a function depends on an argument if for some assignment to the remaining arguments the function remains undetermined. We show that any circuit C that reliably computes a binary function that depends on n arguments using $(1-\xi)/2$-noisy k-input gates must have depth at least $R \log_k n$, for an $R > 1$ depending on k and ξ.

On the other hand, there is a function that depends on $n = k^d$ arguments that can be computed by a circuit using noiseless k-input gates of depth d: take the d-fold composition of a gate that depends on its k inputs. Since the function depends on k^d arguments and the circuit uses only k-input gates, its minimum noiseless circuit depth is d. Our result implies that any reliable noisy circuit for such a function has depth at least Rd. Thus there are functions whose shallowest noisy circuits are deeper by a factor of R than their shallowest noiseless circuit.

Theorem 2: Let f be a function that depends on n inputs. Let C be a circuit of depth d using gates with at most k inputs, where each gate fails independently with probability $(1-\xi)/2$. Suppose C δ-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 $d \geq \log(n\Delta)/\log(k\xi^2)$.
- If $\xi^2 \leq 1/k$ then $n \leq 1/\Delta$.

Proof: Let $x_1, \cdots, x_n$ 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 restricted to this setting for the $n-1$ inputs other than x_i. Let X be a uniformly distributed binary random variable. Let $C_i(X)$ be the random variable that is the output of C_i when $x_i = X$. By Fano's inequality [8, Theorem 2.11.1]

$$I(C_i(X); X) \geq \Delta. \tag{1}$$

We apply Lemma 2 with $G = C_i$ and $W = C_i(X)$ to obtain the upper bound

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

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

Combining the bounds (1) and (2) and summing over all C_i gives

$$n\Delta \leq \sum_{P \in C} \xi^{2|P|}. \tag{3}$$

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

Lemma 3: For all circuits C of depth d that are composed of k-input gates, if $\xi^2 > 1/k$ then

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

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 d, since this tree has $\sum_P \xi^{2|P|} = k^d \xi^{2d}$.

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 $l < d$ 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 $\xi^2 > 1/k$. $\square$

Combining the result of Lemma 3 with (3), we obtain

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

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, since the degrees of the gates are bounded, the paths of the circuit must also become increasingly long. If the gates are too noisy, the additional paths will not compensate for the loss in signal quality. In a large enough circuit, the output will have little dependence on most of the inputs. There is a threshold on the noise level, above which we cannot reliably compute functions of an arbitrary number of inputs.

In order to bound 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 the output. The claim follows by an averaging argument and the fact that $\sum_{P \in C} 1/k^{|P|} \leq 1$ (the Kraft inequality).

Combining (1) and (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. $\square$

This theorem improves on the results of Pippenger and Feder in two ways. First, we increase the lower bound on the threshold for ξ. Second, we increase the factor by which the depth of the reliable circuit must increase. To compute a function that depends on n inputs, Feder shows that a reliable circuit must have depth greater than $\log_k n$ by at least a factor $1/(1+\log_k \xi)$ (the same factor provided by Pippenger for formulas). Our result is that this factor must be at least $1/(1+\log_k \xi^2)$. See Fig. 4.

Our lower bound on the depth of reliable circuits should be compared with the depth of reliable circuits constructed by von Neumann's method. Von Neumann devotes a correction level composed of three-input majority gates to increase re-

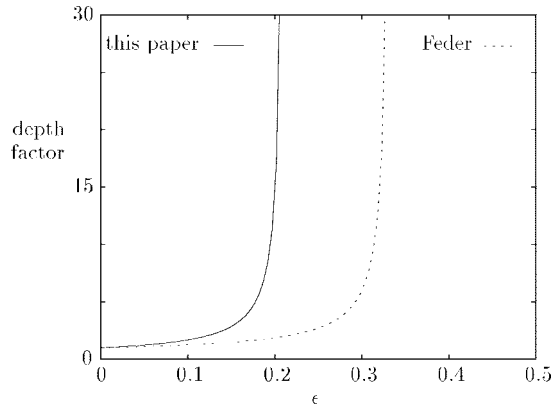

Fig. 4. Lower bounds on factor of increase in circuit depth using three-input, ϵ-noisy gates.

liability after several computation levels. Pippenger analyzes this method when computation is performed by three-input parity gates and determines that depth increases by a factor asymptotic to $1+2/\log_3(1/\epsilon)$ as $\epsilon \to 0$ [9]. Our result implies the factor must be at least asymptotic to $1 + 4\epsilon/\ln 3$.

Von Neumann's method using three-input majority gates works only for $\epsilon < 1/6$. Our bound on the noise threshold shows that for k-input gates, reliable computation by circuits is possible only for $\epsilon < (1 - 1/\sqrt{k})/2$. For formulas, when $k = 3$, Hajek and Weller obtained the stronger result that reliable computation is impossible for $\epsilon \geq 1/6$ [10]. For even $k > 2$, Theorem 2 provides the best known threshold bound for both circuits and formulas, but for odd $k \geq 3$, we have obtained tight bounds on the noise threshold for formulas by extending the method of Hajek and Weller [1].

Our depth bounds can be easily extended to the case of asymmetric noise, in which a gate fails with different probabilities if its prenoise output is 0 or 1. If Y is the prenoise output of the gate then the noisy output Z of the gate is the output of an arbitrary binary channel A on input Y. We use Theorem 1 to bound the fraction of information preserved in crossing this more general channel.

Theorem 3: Let f be a function that depends on n inputs. Let C be a circuit of depth d using gates with at most k inputs. The proper outcome of each gate is subjected to the channel $A = \left(\begin{smallmatrix} 1-a & a \\ b & 1-b \end{smallmatrix}\right)$. Suppose C δ-reliably computes the function f where $\delta < 1/2$. Let

$$\Delta = 1 + \delta \log \delta + (1 - \delta) \log(1 - \delta).$$

Let

$$\varphi = 1 - (\sqrt{b(1-a)} + \sqrt{a(1-b)})^2.$$

- If $\varphi > 1/k$ then $d \geq \log(n\Delta)/\log(k\varphi)$.
- If $\varphi \leq 1/k$ then $n \leq 1/\Delta$.

Proof: The proof is identical to the proof of Theorem 2 with the bound φ (from Theorem 1) replacing ξ^2. □

ACKNOWLEDGMENT

The authors wish to thank N. Pippenger for helpful consultations.

REFERENCES

[1] W. Evans, "Information theory and noisy computation," Ph.D. dissertation, Univ. Calif. Berkeley, 1994.

[2] R. W. Hamming, "Error detecting and error correcting codes," *Bell Syst. Tech. J.*, vol. 29, pp. 147–160, Apr. 1950; also in *Key Papers in the Development of Coding Theory*, E. R. Berlekamp, Ed. New York: IEEE Press, 1974, pp. 9–12.

[3] J. von Neumann, "Probabilistic logics and the synthesis of reliable organisms from unreliable components," in *Automata Studies*, C. E. Shannon and J. McCarthy, Eds. Princeton, NJ: Princeton Univ. Press, 1956, pp. 43–98.

[4] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423; 623–656, 1948.

[5] N. Pippenger, "Reliable computation by formulas in the presence of noise," *IEEE Trans. Inform. Theory*, vol. 34, pp. 194–197, Mar. 1988.

[6] ———, "Invariance of complexity measures for networks with unreliable gates," *J. Assoc. Comput. Mach.*, vol. 36, pp. 531–539, 1989.

[7] T. Feder, "Reliable computation by networks in the presence of noise," *IEEE Trans. Inform. Theory*, vol. 35, pp. 569–571, May 1989.

[8] T. M. Cover and J. A. Thomas, *Elements of Information Theory.* New York: Wiley, 1991.

[9] N. Pippenger, "Analysis of error correction by majority voting," *Adv. Comput. Res.*, vol. 5, pp. 171–198, 1989.

[10] B. Hajek and T. Weller, "On the maximum tolerable noise for reliable computation by formulas," *IEEE Trans. Inform. Theory*, vol. 37, pp. 388–391, Mar. 1991.