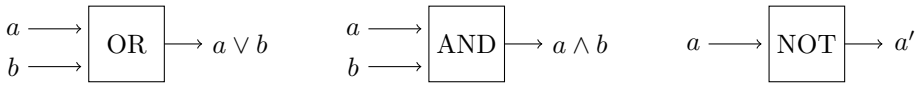
Information Theoretic Methods in Statistics and Computer Science
Instructor: Yury Polyanskiy
Lecture Date: 27.11.2019

Scribe: Yunus Inan

1 Strong Data Processing Inequality and Applications

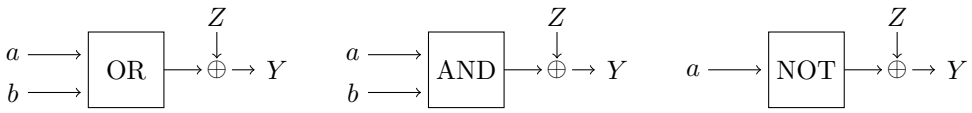
1.1 Motivation: Computing a boolean function with noisy gates

A boolean function with n inputs is defined as $f : \{0,1\}^n \to \{0,1\}$. Note that a boolean function can be described as a network of primitive logic gates of the three following kinds:



Side Note: In 1938, Shannon has found that any boolean function f can be represented with primitive logic gates. [Sha38]

Now suppose there are additive noise components on the output of each primitive gate. In this case, we have a network of the following noisy gates.



Here, $Z \sim \text{Bern}(\delta)$ and assumed to be independent of the inputs. In other words, with probability δ, the output of a gate will be flipped no matter what input is given to that gate. Hence, we sometimes refer to these gates as δ-noisy gates.

An interesting question is: Can we compute any boolean function f with δ-noisy gates? Note that any circuit that consists of noisy gates simulates a random boolean function. Therefore, we want to approximate f with high probability with a noisy circuit C. In a mathematically convenient way, we want

$$\mathbb{P}\big(C(X_1,\ldots,X_n) \neq f(X_1,\ldots,X_n)\big) \leq \frac{1}{2} - \epsilon_0 \tag{1.1}$$

where $C(X_1,\ldots,X_n)$ is the output of the noisy circuit and $\epsilon_0 > 0$ is a constant independent of the inputs $X_1,\ldots,X_n$. Von Neumann has proven this is indeed possible for sufficiently small δ values.

Theorem 1 (Von Neumann, 1957)**.** *There exists $\delta^* > 0$ such that for all $\delta < \delta^*$ it is possible to compute every boolean function f via δ-noisy 3-majority gates.*

Von Neumann's original estimate $\delta^* > 0.087$ was subsequently improved by Pippenger. The main (still open) question of this area is to find the largest δ^* for which the above theorem holds.

Intuitively, the condition in (1.1) implies the output should be correlated with the inputs. Otherwise, a uniformly random guess of the output would yield an error probability of $\frac{1}{2}$. This requires the mutual information between the inputs and the output to be greater than zero. We now give a theorem of Evans and Schulman that gives an upper bound to the mutual information between any of the inputs and the output. We will prove the theorem in Section 1.3.

Theorem 2 ([ES99])**.** *Suppose an n-input noisy boolean circuit composed of gates with at most K inputs and with noise components having at most δ probability of error. Then, the mutual information between any input X_i and output Y is upper bounded as*

$$I(X_i;Y) \leq \big(K(1-2\delta)^2\big)^{d_i} \log 2$$

where d_i is the minimum length between X_i and Y (i.e, the minimum number of gates required to be passed through until reaching Y).

Theorem 2 implies that noisy computation is only possible for $\delta < \frac{1}{2} - \frac{1}{2\sqrt{K}}$. This is the best known threshold. An illustration is given below:

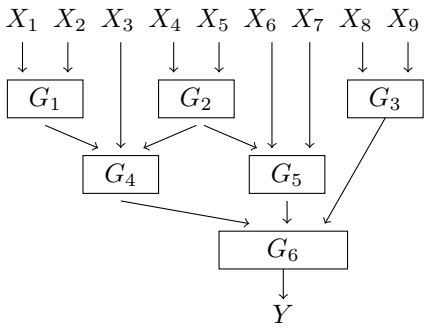
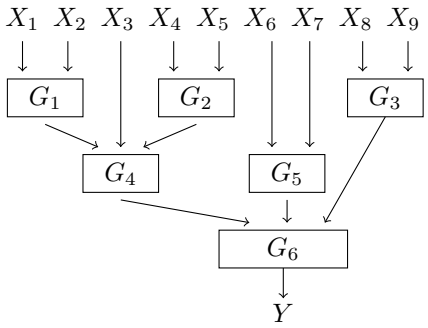


Figure 1.1: An example of a 9-input Boolean Circuit

The above 9-input circuit has gates with at most 3 inputs. The 3-input gates are G_4, G_5 and G_6. The minimum distance between X_3 and Y is $d_3 = 2$, and the minimum distance between X_5 and Y is $d_5 = 3$. If G_i's are δ-noisy gates, we can invoke Theorem 2 between any input and the output.

Not surprisingly, Theorem 2 also tells there are some circuits that are not computable with δ-noisy gates. For instance, take $f(X_1, \ldots, X_n) = \mathrm{XOR}(X_1, \ldots, X_n)$. Then for at least one input X_i, we have $d_i \geq \frac{\log n}{\log K}$. This shows that $I(X_i; Y) \to 0$ as $n \to \infty$, hence X_i and Y will be almost independent for large n. Note that $\mathrm{XOR}(X_1, \ldots, X_n) = \mathrm{XOR}\big(\mathrm{XOR}(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n), X_i\big)$. Therefore, it is impossible to compute an n-input XOR with δ-noisy gates for large n.

Computation with formulas: Note that the graph structure given in Figure 1.1 contains some undirected loops. A *formula* is a type of boolean circuits that does not contain any undirected loops unlike the case in Figure 1.1. In other words, for a formula the underlying graph structure forms a tree. Removing one of the outputs of G_2 of Figure 1.1, we obtain a formula as given below.



In Theorem 1 of [EP98], it is shown that we can compute reliably any boolean function f that is represented with a formula with at most K-input gates with K odd and every gate are at most δ-noisy and $\delta < \delta_f^*$, and no such computation is possible for $\delta > \delta_f^*$, where

$$\delta_f^* = \frac{1}{2} - \frac{2^{K-1}}{K\binom{K-1}{\frac{K-1}{2}}}$$

where the approximation holds for large K. This threshold is better than the upper-bound on the threshold given by Theorem 2 for general boolean circuits. However, for large K we have

$$\delta_f^* \approx \frac{1}{2} - \frac{\sqrt{\pi/2}}{2\sqrt{K}}, K \gg 1$$

showing that the estimate of Evans-Schulman $\delta^* \leq \frac{1}{2} - \frac{1}{2\sqrt{K}}$ is order-tight for large K. This demonstrates the tightness of Theorem 2.

1.2 Strong Data Processing Inequality

Definition 1. (Contraction coefficient for $P_{Y|X}$) For a fixed conditional distribution (or kernel) $P_{Y|X}$, define

$$\eta_f = \eta_f(P_{Y|X}) = \sup_{\substack{P_X, Q_X \\ D_f(P_X||Q_X)>0}} \frac{D_f(P_Y||Q_Y)}{D_f(P_X||Q_X)}.$$

From DPI, we know $\eta_f D_f(P_X||Q_X) \geq D_f(P_Y||Q_Y)$. This is called Strong data processing inequality (SDPI). The reason it is called such is that contrary to the ordinary DPI, which only shows that the f-divergence decreases, SDPI helps to quantify the multiplicative decrease between the two f-divergences.

Some remarks:

1. η_f is very hard to compute in general.
2. Suppose $P_{Y|X}$ is a kernel for a time-homogeneous Markov chain with stationary distribution π (i.e., $P_{Y|X} = P_{X_{t+1}|X_t}$). Then for any initial distribution q, SDPI gives the following bound:

$$D_f(qP^n||\pi) \leq \eta_f^n D_f(q||\pi)$$

   These type of exponential decreases are frequently encountered in the Markov chains literature.

We have stated that η_f is very hard to compute for most f. However, the following theorem states that for $f = \frac{1}{2}|1-t|$, i.e. for total variation, η_f can be characterized in a simple fashion. We simply denote this coefficient as η_{TV}.

Theorem 3 ([Dob56])**.** $\eta_{\mathrm{TV}} = \sup_{x \neq x'} \mathrm{TV}(P_{Y|X=x}, P_{Y|X=x'})$.

Before proving theorem 3, we give the following definition and lemma.

Definition 2 (Coupling)**.** Suppose there are two random variables X and X' with distributions P_X and $P_{X'}$. Fix a joint distribution $P_{XX'}$ on (X, X') with marginals P_X and $P_{X'}$ respectively. (X, X') is called a coupling of X and X'.

Lemma 1 (Characterizations of Total Variation)**.** *For any distributions P_X and Q_X on a discrete alphabet $\mathcal{X}$, the total variation $\mathrm{TV}(P_X, Q_X)$ can be characterized in two ways:*

(i) $\mathrm{TV}(P_X, Q_X) = \sup_{E \subseteq \mathcal{X}} P_X(E) - Q_X(E)$

(ii)

$$\mathrm{TV}(P_X, Q_X) = \inf_{\substack{P_{X_0 X_0'} : P_{X_0} = P_X \\ P_{X_0'} = Q_X}} \mathbb{P}(X_0 \neq X_0')$$

Note that (ii) of Lemma 1 implies that the total variation is the infimal value of $\mathbb{P}(X_0 \neq X_0')$ over all possible couplings (X_0, X_0') with marginals P_X and Q_X.

Proof of Theorem 3. We consider the following two cases

- $\eta_{\mathrm{TV}} \geq \sup_{x_0 \neq x_0'} \mathrm{TV}(P_{Y|X=x_0}, P_{Y|X=x_0'})$:

  This case is obvious. Take $P_X = \delta_{x_0}$ and $Q_X = \delta_{x_0'}$.[1] Then from the definition of η_{TV}, we have $\eta_{\mathrm{TV}} \geq \mathrm{TV}(P_{Y|X=x_0}, P_{Y|X=x_0'})$ for any x_0 and x_0', $x_0 \neq x_0'$.

[1] δ_{x_0} is the probability distribution with $\mathbb{P}(X = x_0) = 1$

- $\eta_{\mathrm{TV}} \le \sup_{x_0 \neq x_0'} \mathrm{TV}(P_{Y|X=x_0}, P_{Y|X=x_0'})$:

  Define $\tilde{\eta} \triangleq \sup_{x_0 \neq x_0'} \mathrm{TV}(P_{Y|X=x_0}, P_{Y|X=x_0'})$. We consider the discrete alphabet case. Fix any P_X, Q_X and $P_Y = P_X \circ P_{Y|X}$, $Q_Y = Q_X \circ P_{Y|X}$. Observe that for any $E \subseteq \mathcal{Y}$

  $$P_{Y|X=x_0}(E) - P_{Y|X=x_0'}(E) \le \tilde{\eta} \mathbb{1}\{x_0 \neq x_0'\}. \tag{1.2}$$

  Now suppose there are random variables X_0 and X_0' having some marginals P_X and Q_X respectively. Consider any coupling (X_0,X_0'). Then averaging (1.2) and taking the supremum, we obtain

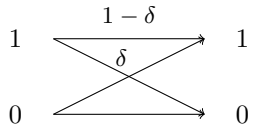
  $$\sup_{E \subseteq \mathcal{Y}} P_Y(E) - Q_Y(E) = \mathrm{TV}(P_Y, Q_Y) \le \tilde{\eta} \mathbb{P}(X_0 \neq X_0')$$

  where the equality follows from Lemma 1 (i). To complete the proof, we use Lemma 1 (ii): Total variation is the infimal value of $\mathbb{P}(X_0 \neq X_0')$ for any possible couplings. Therefore for the coupling (X_0, X_0') attaining the infimal value, we have

  $$\mathrm{TV}(P_Y, Q_Y) \le \tilde{\eta} \mathrm{TV}(P_X, Q_X).$$

□

Example 1 (η_{TV} of a Binary Symmetric Channel)**.** Consider the Binary Symmetric Channel with crossover probability δ (BSC(δ)).



Then η_{TV} of the BSC(δ) is given by

$$\begin{aligned}\eta_{\mathrm{TV}}(\mathrm{BSC}(\delta)) =& \mathrm{TV}(\mathrm{Bern}(\delta), \mathrm{Bern}(1-\delta)) \\ =& \frac{1}{2}\big(|\delta - (1-\delta)| + |1 - \delta - \delta|\big) = |1 - 2\delta|.\end{aligned}$$

We sometimes want to relate η_f with the f-mutual informations instead of f-divergences. This relation is given in the following theorem.

Theorem 4.

$$\eta_f(P_{Y|X}) = \sup_{P_{UX}:\ U \to X \to Y} \frac{I_f(U;Y)}{I_f(U;X)}.$$

Recall that for any Markov chain $U \to X \to Y$, DPI states that $I_f(U;Y) \le I_f(U;X)$ and Theorem 4 gives the stronger bound $I_f(U;Y) \le \eta_f I_f(U;X)$.

Sketch of Proof for Theorem 4. Similar to Theorem 3, we consider the two directions:

- $\eta_f \ge \sup_{\substack{P_{UX}: \\ U \to X \to Y}} \frac{I_f(U;Y)}{I_f(U;X)}$:

  For any u_0, we have $D_f(P_{Y|U=u_0}||P_Y) \le \eta_f D_f(P_{X|U=u_0}||P_X)$. Averaging the above expression over any P_U, we obtain

  $$I_f(U;Y) \le \eta_f I_f(U;X)$$

- $\eta_f \le \sup_{\substack{P_{UX}: \\ U \to X \to Y}} \frac{I_f(U;Y)}{I_f(U;X)}$:

  Fix $\tilde{P}_X$, $\tilde{Q}_X$ and let $U \sim \mathrm{Bern}(\lambda)$ for some $\lambda \in [0,1]$. Define the conditional distibution $P_{X|U}$ as $P_{X|U=1} = \tilde{P}_X$, $P_{X|U=0} = \tilde{Q}_X$. Take $\lambda \to 0$, then

  $$I_f(U;X) = \lambda D_f(\tilde{P}_X||\tilde{Q}_X) + o(\lambda)$$

and avoiding some technical subtleties we obtain

$$I_f(U;Y) = \lambda D_f(\tilde{P}_Y||\tilde{Q}_Y) + o(\lambda)$$

The ratio $\frac{I_f(U;Y)}{I_f(U;X)}$ will then converge to $\frac{D_f(\tilde{P}_Y||\tilde{Q}_Y)}{D_f(\tilde{P}_X||\tilde{Q}_X)} \geq \eta_f$ (some technicalities avoided here as well).

□

Some Important Theorems:

1- For any f, $\eta_f \leq \eta_{\mathrm{TV}}$.

2- $\eta_{\mathrm{KL}} = \eta_{\chi^2}$.

Sketch of Proof for 2. Observe the following two conditions:

- $\eta_{\mathrm{KL}} \geq \eta_{\chi^2}$ by locality. Recall that every f-divergence behaves locally as χ^2.
- Using the identity $D(P||Q) = \int_0^\infty \chi^2(P||Q_t)dt$ where $Q_t = \frac{tP+Q}{1+t}$, we have

$$D(P_Y||Q_Y) = \int_0^\infty \chi^2(P_Y||Q_{Y\,t})dt \leq \eta_{\chi^2} \int_0^\infty \chi^2(P_X||Q_{X\,t})dt = \eta_{\chi^2} D(P_X||Q_X).$$

□

3- $\eta_{\chi^2} = \sup_{P_X,f,g} \rho(f(X), g(Y))$, where $\rho(X,Y) \triangleq \frac{E[XY]}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}$ is the correlation coefficient between X and Y.

4- For binary-input channels, denote $P_0 = P_{Y|X=0}$ and $P_1 = P_{Y|X=1}$. Then we have

$$\eta_{\mathrm{KL}} = \sup_{0<\beta<1} \mathrm{LC}_\beta(P_0||P_1)$$

where [2]

$$\mathrm{LC}_\beta(P||Q) = D_f(P||Q),\ f(x) = \bar{\beta}\beta\frac{(1-x)^2}{\bar{\beta}x+\beta}$$

is the Le Cam divergence of order β.

4'- In particular,

$$\frac{1}{2}H^2(P_0,P_1) \leq \eta_{\mathrm{KL}} \leq H^2(P_0,P_1)$$

where $H^2(P,Q) = D_f(P||Q),\ f(x) = (1-\sqrt{x})^2$ is the Hellinger distance.

5- Suppose a binary-input channel with transition probabilities $P_0 = P_{Y|X=0}$, $P_1 = P_{Y|X=1}$. If there exists a bijection $\Phi : \mathcal{Y} \to \mathcal{Y}$ satisfying $P_0 \circ \Phi^{-1} = P_1$ and $P_1 \circ \Phi^{-1} = P_0$, we call the channel as a binary-input symmetric channel (BMS).

For any BMS, we have $\eta_{\mathrm{KL}} = I_{\chi^2}(X;Y)$ for $X \sim \mathrm{Bern}(0.5)$.

Example 2. (η_{KL} of a BSC(δ)) Suppose we have the same BSC as in Example 1. Then,

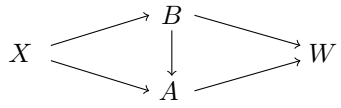
$$\eta_{\mathrm{KL}} = \eta_{\chi^2} = \sup_{\alpha\neq\beta} \frac{\chi^2(\mathrm{Bern}(\alpha\bar{\delta}+\bar{\alpha}\delta)||\mathrm{Bern}(\beta\bar{\delta}+\bar{\beta}\delta))}{\chi^2(\mathrm{Bern}(\alpha)||\mathrm{Bern}(\beta))} = (1-2\delta)^2.$$

[2] $\bar{\beta} \triangleq 1-\beta$

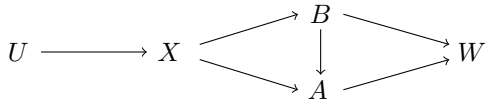
1.3 Information Percolation for Directed Acyclic Graphs

In this section, we are concerned about the amount of information percolation in a directed acyclic graph (DAG) $G = (V, E)$. In the following context the vertex set V refers to a set of vertices v, each associated with a random variable X_v and the edge set E refers to a set of directed edges whose configuration allows us to factorize the joint distribution over V. Throughout the section, we consider Shannon mutual information, i.e., $f = x \log x$. Let us give a detailed example below.

Example 3. Suppose we have a graph $G = (V, E)$ as below and define $\eta \triangleq \eta(P_{W|AB})$:



Now, prepend another random variable $U \sim \text{Bern}(\lambda)$ at the beginning, the new graph $G' = (V', E')$ is shown below: We want to verify the relation



$$I(U; BW) \leq \bar{\eta} I(U; B) + \eta I(U; AB). \tag{1.3}$$

Recall that from chain rule we have $I(U; BW) = I(U; B) + I(U; W|B) \geq I(U; B)$. Hence, if (1.3) is correct, then $\eta \to 0$ implies $I(U; BW) \approx I(U; B)$ and symmetrically $I(U; AW) \approx I(U; A)$. Therefore for small δ, observing W, A or W, B does not give advantage over observing solely A or B, respectively.

Observe that G' forms a Markov chain $U \to X_0 \to (A, B) \to W$, which allows us to factorize the joint distribution over E' as

$$P_{UXABW} = P_U P_{X|U} P_{AB|X} P_{W|AB}.$$

Now consider the joint distribution conditioned on $B = b$, i.e., $P_{UXAW|B}$. The conditional joint distribution yields the conditional Markov chain $U \to X \to A \to W|B = b$. Note that given B and A, X is independent of W and this results in the factorization

$$P_{X|AB} P_{W|AB} = P_{XW|AB},$$

from which follows the mentioned conditional Markov chain. Using the conditional Markov chain, SDPI gives us for any b,

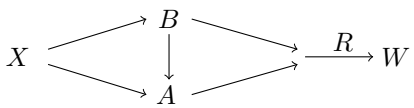
$$I(U; W|B = b) \leq \eta I(U; A|B = b).$$

Averaging over b and adding $I(U; B)$ to both sides we obtain

$$\begin{aligned} I(U; WB) &\leq \eta I(U; A|B) + I(U; B) \\ &= \eta I(U; AB) + \bar{\eta} I(U; B). \end{aligned}$$

Now, we provide another example which has in some sense an analogous setup to Example 3.

Example 4 (Percolation)**.** Take the graph $G = (V, E)$ in example 3 with a small modification.



Now, suppose X,A,B,W are some cities and the edge set E represents the roads between these cities. Let R be a random variable denoting the state of the road connecting to W with $\mathbb{P}(R \text{ is open}) = \eta$ and $\mathbb{P}(R \text{ is closed}) = \bar{\eta}$. For any $Y \in V$, let the event $\{X \to Y\}$ indicate that one can drive from X to Y. Then

$$\mathbb{P}(X \to B \text{ or } W) = \eta \mathbb{P}(X \to A \text{ or } B) + \bar{\eta}\mathbb{P}(X \to B). \tag{1.4}$$

Observe the resemblance between (1.3) and (1.4). For any X and A, we will refer to the probability $\mathbb{P}(X \to A)$ as $\text{perc}(X \to A)$, i.e., percolation probability from X to A.

We will now give a theorem that relates η_{KL} to percolation probability on a DAG under the following settting: Consider a DAG $G = (V, E)$.

- All edges are open
- Every vertex is open with probability $p(v) = \eta_{\text{KL}}\big(P_{X_v|X_{\text{Pa}(v)}}\big)$ where $\text{Pa}(v)$ denotes the set of parents of v.

Note that $P_{X_v|X_{\text{Pa}(v)}}$ describe the stochastic recipee for producing X_v based on its parent variables. We assume that in addition to a DAG we also have been given all these constituent channels (or at least bounds on their η_{KL} coefficients).

Theorem 5 ([PW17])**.** *Let $G = (V, E)$ be a DAG and let X_0 be a node with in-degree equal to zero (i.e. a source node). Note that for any $S \subset V$ we can inductively stitch together constituent channels $P_{X_v|X_{Pa(v)}}$ and obtain $P_{X_S|X_0}$. Then we have*

$$\eta_{\text{KL}}(P_{X_S|X_0}) \leq \textit{perc}(X_0 \to X_S).$$

Sketch of Proof. The graph in Example 3 satisfies this relation. The proof follows from an induction on the size of G. It can be shown that for every DAG G, the base case reduces to the graph in Example 3. □

Note that the above setting includes the case for δ-noisy gates as well. Suppose the output of the gate is open with probability $p(v) = \eta_{\text{KL}}\big(P_{X_v|X_{\text{Pa}(v)}}\big) = \eta_{\text{KL}}(\text{BSC}(\delta))$ as the conditional distribution $P_{X_v|X_{\text{Pa}(v)}}$ is no different than those of a $\text{BSC}(\delta)$. Therefore, we have $\eta_{\text{KL}}(\delta\text{-noisy gate}) \leq p(v) = (1-2\delta)^2$.

We are now in the position to prove Theorem 2.

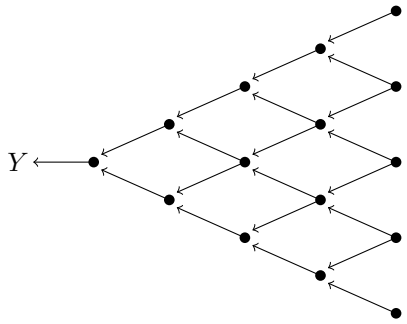
Proof of Theorem 2. First observe the noisy boolean circuit is a form of DAG. From SPDI, for each i, we have $I(X_i; Y) \leq \eta_{\text{KL}}(P_{Y|X_i})H(X_i)$. From Theorem 5, we know $\eta_{\text{KL}}(P_{Y|X_i}) \leq \text{perc}(X_i \to Y)$. We now want to upper bound $\text{perc}(X_i \to Y)$. Recall that the minimum distance between X_i and Y is d_i. Any vertex having minimum distance d_i from Y has the percolation probability $p(v)^{d_i} \leq (1-2\delta)^{2d_i}$ if the error probability is smaller than δ for all gates. For a noisy-circuit that consists of gates with at most K-inputs, the number of possible paths of length d_i from Y is at most K^{d_i}. Since the circuit is DAG, we know that any possible path connecting X_i to Y must contain some of the K^{d_i} possible paths. Therefore, the probability of percolation $\text{perc}(X_i \to Y)$ is upper bounded by $K^{d_i}(1-2\delta)^{2d_i}$. We obtain the final result by upper bounding $H(X_i) \leq \log 2$ as

$$I(X_i; Y) \leq \eta_{\text{KL}}(P_{Y|X_i})H(X_i) \leq K^{d_i}(1-2\delta)^{2d_i} \log 2$$

□

We conclude the section with an example illustrating that Theorem 5 may give stronger bounds when compared to Theorem 2.

Example 5. Suppose we have the topological restriction on the placement of gates (namely that the inputs to each gets should be from nearest neighbors to the left), resulting in the following circuit of 2-input δ-noisy gates.



Note that each gate may be a simple passthrough (i.e. serve as router) or a constant output. Theorem 2 states that if $(1-2\delta)^2 < \frac{1}{2}$, then noisy computation within arbitrary topology is not possible. Theorem 5 improves this to $(1-2\delta)^2 < p_c$, where p_c is the oriented site-percolation threshold for the particular graph we have. Namely, if each vertex is open with probability $p < p_c$ then with probability 1 the connected component emanating from any given node (and extending to the right) is finite. For the example above the site percolation threshold is estimated as $p \approx 0.705$ (so called Stavskaya automata).

References

[Dob56] R. Dobrushin. "Central Limit Theorem for Nonstationary Markov Chains, I". In: *Theory Probab. Appl.* 1.1 (1956), pp. 65–80. DOI: 10.1137/1101006.

[EP98] W. Evans and N. Pippenger. "On the maximum tolerable noise for reliable computation by formulas". In: *IEEE Transactions on Information Theory* 44.3 (1998), pp. 1299–1305. ISSN: 1557-9654. DOI: 10.1109/18.669417.

[ES99] W. S. Evans and L. J. Schulman. "Signal propagation and noisy circuits". In: *IEEE Transactions on Information Theory* 45.7 (1999), pp. 2367–2373. ISSN: 1557-9654. DOI: 10.1109/18.796377.

[PW17] Yury Polyanskiy and Yihong Wu. "Strong Data-Processing Inequalities for Channels and Bayesian Networks". In: *Convexity and Concentration*. Ed. by Eric Carlen, Mokshay Madiman, and Elisabeth M. Werner. New York, NY: Springer New York, 2017, pp. 211–249. ISBN: 978-1-4939-7005-6.

[Sha38] C. E. Shannon. "A symbolic analysis of relay and switching circuits". In: *Electrical Engineering* 57.12 (1938), pp. 713–723. ISSN: 2376-7804. DOI: 10.1109/EE.1938.6431064.