

3.18 POINTWISE UNIVERSALITY OF THE NORMAL FORM

Yaser S. Abu-Mostafa

California Institute of Technology
Pasadena, CA 91125

1. Motivation.

The problems posed here arise in the context of combinational complexity of Boolean functions whose truth tables cannot be concisely specified [2]. This class of functions arises in the study of computation and decision-making based on natural data, such as the case of pattern recognition in uncontrolled environments. The main feature of these functions is the lack of a structure that would allow an efficient systematic implementation. This leaves us with a large number of essentially unrelated cases to account for, which puts a lower bound on the complexity of these functions. However, an exhaustive solution is not necessary either, since the essential dimensionality of the data is typically far less than the actual dimensionality.

As an example, consider the problem of recognizing a tree in a visual scene. The input data is a matrix of binary pixels representing the scene, and the Boolean function decides the presence or absence of a tree. It is clear that a visual scene is not a totally random binary matrix; there are many correlations that reduce the entropy. On the other hand, the presence or absence of a tree cannot be formalized in a simple way; the visual object "tree," apart from being a fuzzy notion [12], is an assembly of a large number of loosely related observations. To define a tree is to capture these observations in a model, but the partial randomness due to the way natural objects are made precludes a concise model.

The formalization of these ideas involves defining and relating several quantitative measures on Boolean functions. These measures are the cost C of implementing a function, the entropy H of the data, the randomness

R of the function, and the complexity K which measures the relative complexity of the function as far as simple decomposition is concerned. The measures are based on combinational complexity [11] which is the actual cost of decision-making, Shannon's entropy [10] which measures the essential dimensionality of data, Kolmogorov-Chaitin complexity [4,7] which measures the randomness of strings, and compositional complexity [1] which is defined in terms of the standard pattern recognition system that makes a global decision based on local features. These notions are made precise in the next section.

2. Definitions.

Let N be a positive integer, and consider the set F_N of all Boolean functions f from $\{0,1\}^N$ to $\{0,1\}$. The cardinality of F_N is given by $|F_N| = 2^{2^N}$. The independent Boolean variables will be called s_1, \dots, s_N . All logarithms and exponentials are to the base 2. The four measures, C , H , R , and K , assign to Boolean functions in F_N values ranging from 0 to N bits (approximately), with most of the functions assigned values close to N .

Let n be a non-negative integer. An n -input *universal gate* is a switching device with n input lines and 1 output line that can simulate any Boolean function of n variables, for example, a PROM with n address lines and 1 data line. The *cost* of this gate is defined as 2^n "cells." A combinational circuit Γ is a loop-free interconnection of universal gates where the variables s_1, \dots, s_N are supplied. The cost of Γ is the sum of the costs of its gates (wires are free, unlimited fan-out). Γ simulates f if f is the output of one of the gates in Γ .

Definition. The (*normalized*) *cost* C is a real-valued function defined on F_N by

$$C(f) = \log \min\{ \text{cost of } \Gamma : \Gamma \text{ simulates } f \} \quad \text{bits} .$$

$C(f)$ differs by at most a constant from the cost based on any other complete basis of switching devices such as 2-input NAND gates. It is clear

that $C(f) \leq N$ bits, since an N -input PROM with cost 2^N cells can simulate any function in F_N .

Definition. Let $h(f) \leq 2^{N-1}$ be the number of 1's, or the number of 0's, in the Karnaugh map of f . The (*deterministic*) entropy H is a real-valued function defined on F_N by

$$H(f) = \log \left[1 + h(f) \right] \quad \text{bits.}$$

Clearly, $H(f) \leq N$ bits. The entropy of the constant functions is $\log(1 + 0) = 0$ bits, of the N -input AND function is $\log(1 + 1) = 1$ bit, and of the N -input XOR function is $\log(2^{N-1} + 1) \approx N$ bits. This entropy measure is related to Shannon's entropy (of the ensemble $\{0,1\}^N$ under some probability distribution) by considering only the typical blocks in the Karnaugh map of f .

Let $\tau(f)$ be a listing of the truth table of f , that is, $\tau(f) = \tau_0, \tau_1, \dots, \tau_{2^N-1}$ where τ_k is the value of f when the inputs are the N -bit binary representation of the number k . Let U be a universal Turing machine with input alphabet $\{0,1\}$, and let \mathbf{p} denote the binary program supplied to the tape of U . If, given \mathbf{p} , U halts and leaves the binary string \mathbf{w} on the tape, we say that $\mathbf{w} = U(\mathbf{p})$. $|\mathbf{p}|$ denotes the length of \mathbf{P} .

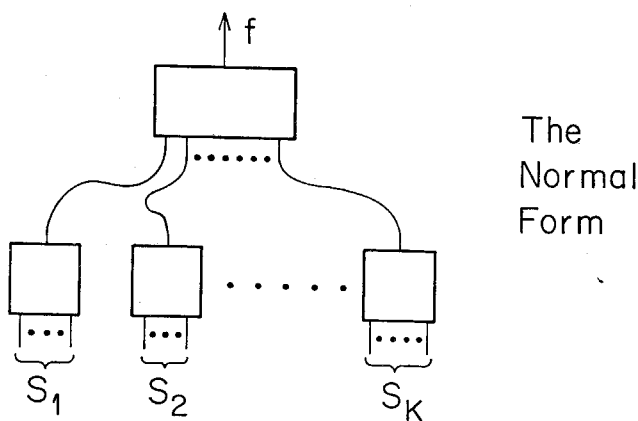
Definition. The *randomness* R is a real-valued function defined on F_N by

$$R(f) = \log \min \{ |\mathbf{p}| : U(\mathbf{p}) = \tau(f) \} \quad \text{bits.}$$

A legal program \mathbf{p} for U consists of an encoding of a Turing machine followed by an input string, hence $|\mathbf{p}|$ is positive and the logarithm is valid. Also, since any string $\tau(f)$ can be generated by a program whose length is a constant (the code of a trivial Turing machine) + the length of the string (namely, 2^N), $R(f)$ is at most $\approx N$ bits. In contrast with the other measures, $R(f)$ is an uncomputable function.

A normal form is a simple decomposition of the Boolean function $f(s_1, \dots, s_N)$ into $f = g(h_1, \dots, h_K)$, where the h_k 's are Boolean functions depending only on variables within subsets S_1, \dots, S_K of $\{s_1, \dots, s_N\}$. A normal form is characterized by the (not-necessarily-

distinct) subsets S_1, \dots, S_K and said to admit a function f if f can be decomposed as above with the h_k 's depending on the variables within the S_k 's, respectively. The number of functions in F_N admitted by a normal form is denoted by $N(S_1 \cdots S_K)$. For example, if $K = N$ and $S_K = \{s_k\}$, then $N(S_1 \cdots S_N) = 2^{2^N}$. In general, $N(S_1 \cdots S_K)$ expresses the power of the normal form $S_1 \cdots S_K$.



Definition. The (*normal-form*) complexity K is a real-valued function defined on F_N by

$$K(f) = \log \log \min \{N(S_1 \cdots S_K) : S_1 \cdots S_K \text{ admits } f\} \quad \text{bits.}$$

Since any normal form admits the two constant functions, taking the logarithm twice is valid. Also, since $|F_N| = 2^{2^N}$, $K(f) \leq N$ bits. Having a large value of $K(f)$ means that f cannot be expressed as a function of few arguments each of which depends on few variables. A circuit simulation of the normal form $S_1 \cdots S_K$ consists of K primary universal gates with $|S_1|, \dots, |S_K|$ inputs, followed by a secondary universal gate with K inputs (see figure). The cost of this circuit is directly related to

$\log N(S_1 \cdots S_K)$ [2], since a universal gate of n inputs costs 2^n cells and simulates 2^{2^n} functions. Therefore, $K(f)$ can be thought of as the (normalized) cost of normal-form simulation of f .

3. Known Relations.

In this section, we state the known pairwise relations between the four measures C , H , R , and K . We shall say that " $A(f) \leq B(f) + o(N)$ for all f " means: Given $\epsilon > 0$ there is a positive integer N_o such that $N \geq N_o$ and $f \in F_N$ implies that $A(f) \leq B(f) + \epsilon N$. We shall also say that " $A(f) \leq B(f) + o(N)$ for almost all f " means: Given $\epsilon > 0$ there is a positive integer N_o such that $N \geq 0$ and $0 < \alpha \leq 1$ implies that the ratio between $|\{f \in F_N : A(f) > B(f) + \epsilon N \text{ and } (\alpha - \epsilon)N \leq A(f) < (\alpha + \epsilon)N\}|$ and $|\{f \in F_N : (\alpha - \epsilon)N \leq A(f) \leq (\alpha + \epsilon)N\}|$ is less than ϵ . The following relations are proved [2,3] by simulation, enumeration, and construction.

$$R1: C(f) \leq H(f) + o(N) \text{ for all } f.$$

$$R2: C(f) \leq R(f) + o(N) \text{ for almost all } f.$$

$$R3: C(f) \leq K(f) + o(N) \text{ for all } f.$$

$$R4: H(f) \leq C(f) + o(N) \text{ for almost all, but not all, } f.$$

$$R5: H(f) \leq R(f) + o(N) \text{ for almost all, but not all, } f.$$

$$R6: H(f) \leq K(f) + o(N) \text{ for almost all, but not all, } f.$$

$$R7: R(f) \leq C(f) + o(N) \text{ for all } f.$$

$$R8: R(f) \leq H(f) + o(N) \text{ for all } f.$$

$$R9: R(f) \leq K(f) + o(N) \text{ for all } f.$$

$$R10: K(f) \leq C(f) + o(N) \text{ for almost all } f.$$

$$R11: K(f) \leq H(f) + o(N) \text{ for almost all } f.$$

$$R12: K(f) \leq R(f) + o(N) \text{ for almost all } f.$$

4. Problems.

Relations R1-R12 of the previous section raise a number of questions about how strongly C , H , R , and K are related. The following questions address stronger versions of relations R2, R10, R11, and R12:

Q1: Is $C(f) \leq R(f) + o(N)$ for all f ?

Q2: Is $K(f) \leq C(f) + o(N)$ for all f ?

Q3: Is $K(f) \leq H(f) + o(N)$ for all f ?

Q4: Is $K(f) \leq R(f) + o(N)$ for all f ?

The answers to these questions, combined with relations R1-R12, determine the exact asymptotic relations between C , H , R , and K . For example, is $|C(f) - K(f)| = o(N)$ for all f ? In other words, is the difference between the minimum cost of an unrestricted simulation and the minimum cost of a normal-form simulation of *any* function f asymptotically negligible w.r.t. N ? Relations R3 and R10 give an affirmative answer to the question in an "almost always" sense. An affirmative answer in an "always" sense would mean that the normal form is a *point-wise universal* (asymptotically optimal for *every* function) structure for simulation of Boolean functions. If the answer is affirmative, more specific questions about the size of the error term $o(N)$ can be addressed. For example, it is easy to see that $|C(f) - K(f)| = \Omega(\sqrt{N})$ for some simple functions such as the N -input XOR. Is $|C(f) - K_M(f)| = O(N^{1/M})$ for all f , where $K_M(f)$ is based on an M -stage normal form instead of a two-stage normal form?

The answers to Q1-Q4 also yield the answers to other questions of interest. Is $|C(f) - R(f)| = o(N)$ for all f ? An affirmative answer to Q3 bounds the cost of normal-form simulation of a function by the essential dimensionality (entropy) of the function. This would mean that the standard pattern recognition system is asymptotically optimal for the typical pattern recognition problem. Other questions related to the size of the error term $o(N)$ (which is $O(\log N)$ for some, and $O(\sqrt{N})$ for other, of the relations R1-R12) are also of theoretical and practical interest.

REFERENCES

- [1] H. Abelson et al., "Compositional Complexity of Boolean Functions," *Discrete Appl. Math.*, 4, pp. 1-10 (1982).
- [2] Y. Abu-Mostafa, *Complexity of Information Extraction*, Ph.D. Thesis, Caltech, 1983.
- [3] Y. Abu-Mostafa, "Complexity of Random Problems," invited paper, *IEEE International Symposium on Information Theory*, 1985.
- [4] G. Chaitin, "A Theory of Program Size Formally Identical to Information Theory," *J. Assoc. Comput. Mach.*, 22, pp. 329-340 (1975).
- [5] R. Duda and P. Hart, *Pattern Classification and Scene Analysis*, Wiley-Interscience, New York, 1973.
- [6] Z. Kohavi, *Switching and Finite Automata Theory*, 2nd ed., McGraw-Hill, New York, 1978.
- [7] A. Kolmogorov, "Three Approaches for Defining the Concept of Information Quantity," *Inf. Transmission*, 1, pp. 3-11 (1965).
- [8] P. Martin-Lof, "The Definition of Random Sequences," *Inf. Control*, 9, pp. 602-619 (1966).
- [9] R. McEliece, *The Theory of Information and Coding*, Addison-Wesley, Reading, MA, 1977.
- [10] C. Shannon, "A Mathematical Theory of Communication," *Bell Syst. Tech. J.*, 28, pp. 59-98 (1949).
- [11] C. Shannon, "The Synthesis of Two-Terminal Switching Circuits", *Bell Syst. Tech. J.*, 28, pp. 59-98 (1949).
- [12] L. Zadeh, "Fuzzy Sets," *Inf. Control*, 8, pp. 338-353 (1965).