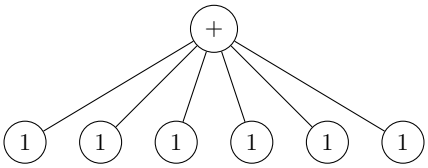
SHANNON ENTROPY IN INTEGER FACTORIZATION

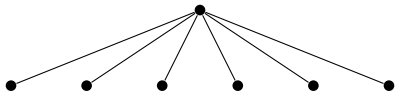
ADAM MARKS

1. Introduction

Consider the following graph of the computational expression $1+1+1+1+1+1$.



If we restrict ourselves to the consideration of such graphs, rooted trees in which a leaf vertex represents the number 1 and all other vertices represents an n-ary sum, we can simplify the diagram to the following.



We take the integers $\mathbb{Z}$ to be built from the natural numbers $\mathbb{N}$, themselves constructed from the Peano axioms, with successor function s giving the definition of 1 as $s(0)$. The multiplication of natural numbers reduces to iterated addition, and the natural numbers themselves reduce to compositions of the successor function. With this viewpoint, any positive integer n can be written as graph in the form above, and any such graph corresponds to a positive integer.

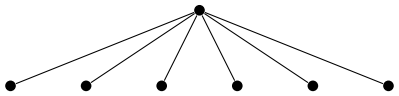
We call these graphs *addition trees*, and by measuring their symmetry we obtain a measure of the Shannon entropy in various factorizations of integers. From here, we consider implications for the computational complexity of constructing integer factorizations.

University of California, Irvine
Date: Draft v1, September 10, 2020.

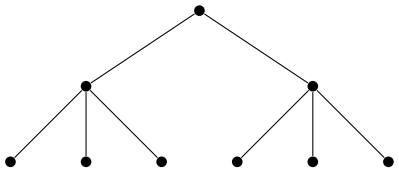
2. Factorization of six

As a motivating example, let T_0, T_1, T_2 be representations of the number 6 as follows.

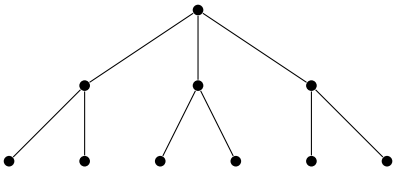
$T_0 = 1 + 1 + 1 + 1 + 1 + 1$



$T_1 = 2 \cdot (1 + 1 + 1)$



$T_2 = 3 \cdot (1 + 1)$



Intuitively, we may expect that T_1 and T_2 contain more information than T_0. Indeed, we can formalize this notion by using the automorphism group of each graph as a quantification of symmetry, and applying the measurement of Shannon entropy to a randomly selected vertex.

Consider $\mathrm{Aut}(T_1)$, the automorphism group of T_1. $\mathrm{Aut}(T_1) \cong S_2 \times S_3 \times S_3$, where S_n is the symmetric group on n vertices. Labelling the vertices $V = \{v_0, \ldots, v_8\}$, left to right, top to bottom, the orbits of the action of $\mathrm{Aut}(T_1)$ on the vertices are $V_0 = \{v_0\}, V_1 = \{v_1, v_2\}, V_2 = \{v_3, v_4, v_5, v_6, v_7, v_8\}$. If a vertex v is chosen at random, and X is a random variable representing the orbit of the chosen vertex, the entropy of X is given by Shannon's formula as

$$H(X) = -\sum_{i=0}^{2} p_i \log_2 p_i,$$

where p_i is the probability that $v \in V_i$.

$$p_i = \frac{|V_i|}{|V|}.$$

Denoting by $H(T)$ the entropy $H(X)$ when X is the discrete random variable constructed in this way for graph T, we thus have

$$H(T_1) = -\left(\frac{1}{9}\log_2\frac{1}{9} + \frac{2}{9}\log_2\frac{2}{9} + \frac{6}{9}\log_2\frac{6}{9}\right) \approx 1.224 \text{ bits.}$$

For T_2 we have, $\mathrm{Aut}(T_2) \cong S_3 \times S_2 \times S_2 \times S_2$, and

$$H(T_2) = -\left(\frac{1}{10}\log_2\frac{1}{10} + \frac{3}{10}\log_2\frac{3}{10} + \frac{6}{10}\log_2\frac{6}{10}\right) \approx 1.295 \text{ bits.}$$

And for T_0 we have $\mathrm{Aut}(T_0) \cong S_6$, and

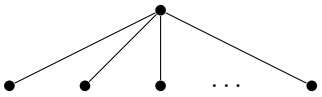
$$H(T_0) = -\left(\frac{1}{7}\log_2\frac{1}{7} + \frac{6}{7}\log_2\frac{6}{7}\right) \approx 0.592 \text{ bits.}$$

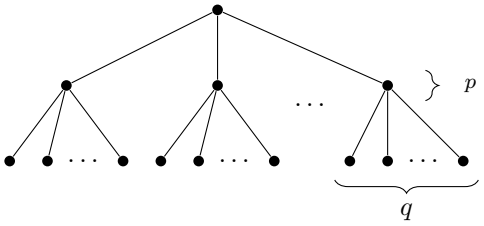
Noting that T_1 and T_2 represent the number 6 in factored forms (factoring out 2 or 3 respectively), we have the situation that the unfactored form T_0 contains less information than the factored forms, as desired. Also note that factoring out the larger integer produces more information than factoring out the smaller integer. Indeed,

$$H(T_0) < H(T_1) < H(T_2).$$

3. Factorization of pq

Let p and q be primes. Generalizing the foregoing approach, let $m = pq$, let T_m be the unfactored addition tree, and let T_p the addition tree with p factored out. We have:

T_m  $\mathrm{Aut}(T_m) \cong S_m$

T_p  $\mathrm{Aut}(T_p) \cong S_p \times S_q^p$

So

$$H(T_m) = -\left(\frac{1}{m+1}\log_2\frac{1}{m+1} + \frac{m}{m+1}\log_2\frac{m}{m+1}\right),$$

and

$$H(T_p) = -\left(\frac{1}{m+p+1}\log_2\frac{1}{m+p+1} + \frac{p}{m+p+1}\log_2\frac{p}{m+p+1} + \frac{m}{m+p+1}\log_2\frac{m}{m+p+1}\right).$$

4. TBD / Future work

(a) Analyze the asymptotic behavior of $H(T_p)$.
(b) Relate addition trees to Turing machines. Can a conclusion be drawn about the time complexity of factoring? In general, if a piece of data D contains N bits of entropy, is there a lower bound on the number of steps required to materialize D in a Turing machine?
(c) Can we relate these questions to the problem spaces P and NP?

References

[1] Sheldon Ross, *A First Course in Probability,* Macmillan Publishing Company, New York, 1988.