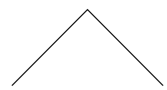
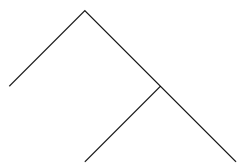
information theory → information in stochastic processes ★★★

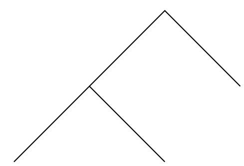
Random Binary Tree

Consider the following process which generates a random binary tree: we begin with the base tree:



Then, we pick one of the leaves (a node with no further branches) of the tree at random, and add a pair of branches/leaves to that node. Thus, with equal probability, the first step of our random tree generation will choose one of the following 2 trees:





To generate a larger tree, we again pick one of the leaves in the tree at random, and attach 2 branches to it. In general, we can denote the state space of random trees with n leaves with Ω_n, and denote with T_n the random variable corresponding to the tree drawn from the random tree ensemble described above.

Our goal in this problem will be to compute the entropy rate of this stochastic process. It will turn out that it will be easier to think of an equivalent way of generating T_n, the tree with n leaves. Consider a process where you drawn uniformly a number k from $\{1, \ldots, n-1\}$, and assign the left half of the tree to have k leaves, and the right half to have $n-k$ leaves. Then, going to the left half of the tree, one can imagine assigning that node's left branch (if $k > 1$) to point to m nodes with probability uniform on $\{1, \ldots, k-1\}$, etc. The process will eventually truncate because once we draw a 1, then we have found a leaf, and each step reduces the number of leaves in the branch by at least 1.

(a) Prove that the two random tree ensembles described above are in fact, equivalent. This does not actually take a lot of effort: justify why if you can prove that

$$\mathrm{P}_{\Omega_N}(\text{left side of tree has } k \text{ leaves}) = \frac{1}{N-1} \qquad (1 \le k \le N-1)$$

you are done. Then, prove this statement.

(b) Give a basic sense of what this random tree ensemble looks like, by depicting the state space Ω_4 for 4 leaved trees, as well as giving the probability distribution on Ω_4.

(c) Now, let us look at the entropy

$$H_n \equiv \mathrm{H}(T_n).$$

Show, using the alternative generation of random trees described above, that

$$H_n = \log(n-1) + \frac{2}{n-1}\sum_{k=1}^{n-1} H_k.$$

(d) Use the result of the previous part to find a recursive relation between H_n and H_{n-1}.

(e) Numerically determine the finite quantity

$$\alpha \equiv \lim_{n \to \infty} \frac{H_n}{n}$$

up to 3 significant digits.