

## Homework Set #2 Solutions

### 1. Entropy and pairwise independence.

Let  $X, Y, Z$  be three binary Bernoulli( $\frac{1}{2}$ ) random variables that are pairwise independent; that is,  $I(X; Y) = I(X; Z) = I(Y; Z) = 0$ .

- (a) Under this constraint, what is the minimum value for  $H(X, Y, Z)$ ?
- (b) Give an example achieving this minimum.
- (c) Now suppose that  $X, Y, Z$  are three random variables each uniformly distributed over the alphabet  $\{1, 2, \dots, m\}$ . Again, they are pairwise independent. What is the minimum value for  $H(X, Y, Z)$ ?

### Solution: Entropy and pairwise independence.

- (a) Due to pairwise independence and the marginal distributions,  $H(X, Y) = H(X) + H(Y) = 2$ . Thus,

$$\begin{aligned} H(X, Y, Z) &= H(X, Y) + H(Z|X, Y) \\ &= H(X) + H(Y) + H(Z|X, Y) \\ &\geq H(X) + H(Y) \\ &= 2 \end{aligned}$$

with equality if and only if  $Z$  is a deterministic function of  $X$  and  $Y$ . This minimum is achieved by the example in part (b).

- (b) Let  $Z = X \oplus Y$ , where  $\oplus$  denotes XOR. It is easy to check that all the marginal distributions are satisfied, as well as pairwise independence.
- (c) Here is one possible solution. Without loss of generality, relabel the alphabet to be  $\{0, 1, 2, \dots, m-1\}$  instead of  $\{1, 2, \dots, m\}$ . Let  $Z = X + Y \pmod{m}$ . Then

$$H(X, Y, Z) = H(X) + H(Y) = \log m + \log m = 2 \log m.$$

One can then justify that under this construction, all the marginal distributions are satisfied, and we also have pairwise independence. For example, to argue pairwise

independence between  $Z$  and  $X$ ,

$$\begin{aligned} I(X; Z) &= H(Z) - H(Z|X) \\ &= \log m - H(X + Y|X) \\ &= \log m - H(Y|X) \\ &= \log m - H(Y) \\ &= \log m - \log m \\ &= 0. \end{aligned}$$

Thus,  $X$  and  $Z$  are independent.

## 2. The value of a question.

Let  $X \sim p(x)$ ,  $x = 1, 2, \dots, m$ .

We are given a set  $S \subseteq \{1, 2, \dots, m\}$ . We ask whether  $X \in S$  and receive the answer

$$Y = \begin{cases} 1, & \text{if } X \in S \\ 0, & \text{if } X \notin S. \end{cases}$$

Suppose  $\Pr\{X \in S\} = \alpha$ .

Find the decrease in uncertainty  $H(X) - H(X|Y)$ .

Apparently any set  $S$  with a given probability  $\alpha$  is as good as any other.

**Solution: The value of a question.**

$$\begin{aligned} H(X) - H(X|Y) &= I(X; Y) \\ &= H(Y) - H(Y|X) \\ &= H(\alpha) - H(Y|X) \\ &= H(\alpha) \end{aligned}$$

since  $H(Y|X) = 0$ .

## 3. Random questions.

One wishes to identify a random object  $X \sim p(x)$ . A question  $Q \sim r(q)$  is asked at random according to  $r(q)$ . This results in a deterministic answer  $A = A(x, q) \in \{a_1, a_2, \dots\}$ . Suppose the object  $X$  and the question  $Q$  are independent. Then  $I(X; Q, A)$  is the uncertainty in  $X$  removed by the question-answer  $(Q, A)$ .

(a) Show  $I(X; Q, A) = H(A|Q)$ . Interpret.

(b) Now suppose that two i.i.d. questions  $Q_1, Q_2 \sim r(q)$  are asked, eliciting answers  $A_1$  and  $A_2$ . Show that two questions are less valuable than twice the value of a single question in the sense that  $I(X; Q_1, A_1, Q_2, A_2) \leq 2I(X; Q_1, A_1)$ .

**Solution: Random questions.**

We assume that  $X, Q_1, Q_2$  are jointly independent. We also assume  $A(X, Q)$  is a deterministic function; i.e., the answer is a function of the question  $Q$  and the object  $X$ .

(a)

$$\begin{aligned} I(X; Q, A) &\stackrel{(a)}{=} I(Q; X) + I(A; X|Q) \\ &\stackrel{(b)}{=} I(A; X|Q) \\ &\stackrel{(c)}{=} H(A|Q) - H(A|X, Q) \\ &\stackrel{(d)}{=} H(A|Q) \end{aligned}$$

where

- (a) Chain rule.
- (b)  $X$  and  $Q$  are independent.
- (c) Definition of mutual information.
- (d) Answer is a deterministic function of  $X$  and  $Q$ .

The interpretation is as follows. The uncertainty removed in  $X$  given  $(Q, A)$  is the same as the uncertainty in the answer given the question.

- (b) Using the result from part (a) and the fact that questions are independent, we can easily obtain the desired relationship.

$$\begin{aligned} I(X; Q_1, A_1, Q_2, A_2) &\stackrel{(a)}{=} I(X; Q_1) + I(X; A_1|Q_1) + I(X; Q_2|A_1, Q_1) \\ &\quad + I(X; A_2|A_1, Q_1, Q_2) \\ &\stackrel{(b)}{=} I(X; A_1|Q_1) + H(Q_2|A_1, Q_1) - H(Q_2|X, A_1, Q_1) \\ &\quad + I(X; A_2|A_1, Q_1, Q_2) \\ &\stackrel{(c)}{=} I(X; A_1|Q_1) + I(X; A_2|A_1, Q_1, Q_2) \\ &= I(X; A_1|Q_1) + H(A_2|A_1, Q_1, Q_2) - H(A_2|X, A_1, Q_1, Q_2) \\ &\stackrel{(d)}{=} I(X; A_1|Q_1) + H(A_2|A_1, Q_1, Q_2) \\ &\stackrel{(e)}{\leq} I(X; A_1|Q_1) + H(A_2|Q_2) \\ &\stackrel{(f)}{\leq} H(A_1|Q_1) - H(A_1|X, Q_1) + H(A_2|Q_2) \\ &\stackrel{(g)}{\leq} H(A_1|Q_1) + H(A_2|Q_2) \\ &\stackrel{(h)}{\leq} 2H(A_1|Q_1) \\ &\stackrel{(i)}{=} 2I(X; A_1, Q_1) \end{aligned}$$

- (a) Chain rule.
- (b)  $X$  and  $Q_1$  are independent.
- (c)  $Q_2$  is independent of  $X$ ,  $Q_1$ , and  $A_1$ . (d)  $A_2$  is completely determined given  $Q_2$  and  $X$ .
- (e) Conditioning decreases entropy.
- (f) Expand the entropy.
- (g)  $A_1$  is completely determined given  $Q_1$  and  $X$ .
- (h) Identically distributed.
- (i) Invoke the result of part (a).

In the above solution, the mutual information chain rule was applied directly. Here is an alternative solution which some might prefer. Starting over, we want to prove that

$$I(X; Q_1, A_1, Q_2, A_2) \leq 2I(X; Q_1, A_1).$$

To make the inequality look simpler, we can break the left-hand side as

$$I(X; Q_1, A_1, Q_2, A_2) = I(X; Q_1, A_1) + I(X; Q_2, A_2 | Q_1, A_1).$$

Subtracting  $I(X; Q_1, A_1)$  from both sides shows that it suffices to argue that

$$I(X; Q_2, A_2 | Q_1, A_1) \leq I(X; Q_1, A_1).$$

This is argued by the following chain of inequalities:

$$\begin{aligned} I(X; Q_2, A_2 | Q_1, A_1) &\stackrel{(a)}{=} H(A_2 | Q_2, Q_1, A_1) \\ &\stackrel{(b)}{\leq} H(A_2 | Q_2) \\ &\stackrel{(c)}{=} H(A_1 | Q_1) \\ &\stackrel{(d)}{=} I(X; Q_1, A_1) \\ I(X; Q_1, A_1, Q_2, A_2) &\stackrel{(e)}{\leq} 2I(X; Q_1, A_1). \end{aligned}$$

where

- (a) Follows from the result of Part (a), conditioning on  $Q_1$  and  $A_1$ .
- (b) Conditioning decreases entropy.
- (c) Identically distributed.
- (d) Follows from the result of Part (a).
- (e) Add  $I(X; Q_1, A_1)$  to both sides of the inequality.

#### 4. Bottleneck.

Suppose a (non-stationary) Markov chain starts in one of  $n$  states, necks down to  $k < n$  states, and then fans back to  $m > k$  states. Thus  $X_1 \rightarrow X_2 \rightarrow X_3$ ,  $X_1 \in \{1, 2, \dots, n\}$ ,  $X_2 \in \{1, 2, \dots, k\}$ ,  $X_3 \in \{1, 2, \dots, m\}$ , and  $p(x_1, x_2, x_3) = p(x_1)p(x_2|x_1)p(x_3|x_2)$ .

- (a) Show that the dependence of  $X_1$  and  $X_3$  is limited by the bottleneck by proving that  $I(X_1; X_3) \leq \log k$ .
- (b) Evaluate  $I(X_1; X_3)$  for  $k = 1$ , and conclude that no dependence can survive such a bottleneck.

**Solution: Bottleneck.**

- (a) From the data processing inequality, and the fact that entropy is maximum for a uniform distribution, we get

$$\begin{aligned} I(X_1; X_3) &\leq I(X_1; X_2) \\ &= H(X_2) - H(X_2 | X_1) \\ &\leq H(X_2) \\ &\leq \log k. \end{aligned}$$

Thus, the dependence between  $X_1$  and  $X_3$  is limited by the size of the bottleneck. That is  $I(X_1; X_3) \leq \log k$ .

- (b) For  $k = 1$ ,  $0 \leq I(X_1; X_3) \leq \log 1 = 0$  so that  $I(X_1, X_3) = 0$ . Thus, for  $k = 1$ ,  $X_1$  and  $X_3$  are independent.

**5. Conditional mutual information.**

Consider a sequence of  $n$  binary random variables  $X_1, X_2, \dots, X_n$ . Each  $n$ -sequence with an even number of 1's has probability  $2^{-(n-1)}$  and each  $n$ -sequence with an odd number of 1's has probability 0. Find the mutual informations

$$I(X_1; X_2), \quad I(X_2; X_3 | X_1), \dots, I(X_{n-1}; X_n | X_1, \dots, X_{n-2}).$$

**Solution: Conditional mutual information.**

Consider a sequence of  $n$  binary random variables  $X_1, X_2, \dots, X_n$ . Each sequence of length  $n$  with an even number of 1's is equally likely and has probability  $2^{-(n-1)}$ .

Any  $n - 1$  or fewer of these are independent. Thus, for  $k \leq n - 1$ ,

$$I(X_{k-1}; X_k | X_1, X_2, \dots, X_{k-2}) = 0.$$

However, given  $X_1, X_2, \dots, X_{n-2}$ , once we know either  $X_{n-1}$  or  $X_n$ , we know the other.

$$\begin{aligned} I(X_{n-1}; X_n | X_1, X_2, \dots, X_{n-2}) &= H(X_n | X_1, X_2, \dots, X_{n-2}) - H(X_n | X_1, X_2, \dots, X_{n-1}) \\ &= 1 - 0 = 1 \text{ bit.} \end{aligned}$$

## 6. Fano's inequality.

Let  $\Pr(X = i) = p_i, i = 1, 2, \dots, m$  and let  $p_1 \geq p_2 \geq p_3 \geq \dots \geq p_m$ . The minimal probability of error predictor of  $X$  is  $\hat{X} = 1$ , with resulting probability of error  $P_e = 1 - p_1$ . Maximize  $H(\mathbf{p})$  subject to the constraint  $1 - p_1 = P_e$  to find a bound on  $P_e$  in terms of  $H$ . This is Fano's inequality in the absence of conditioning.

### Solution: Fano's Inequality.

When there is no information, the minimal probability of error predictor is  $\hat{X} = 1$ , the most probable value of  $X$ . The probability of error in this case is  $P_e = 1 - p_1$ . Hence if we fix  $P_e$ , we fix  $p_1$ . We maximize the entropy of  $X$  for a given  $P_e$  to obtain an upper bound on the entropy for a given  $P_e$ . The entropy,

$$\begin{aligned} H(\mathbf{p}) &= -p_1 \log p_1 - \sum_{i=2}^m p_i \log p_i \\ &= -p_1 \log p_1 - \sum_{i=2}^m P_e \frac{p_i}{P_e} \log \frac{p_i}{P_e} - P_e \log P_e \\ &= H(P_e) + P_e H\left(\frac{p_2}{P_e}, \frac{p_3}{P_e}, \dots, \frac{p_m}{P_e}\right) \\ &\leq H(P_e) + P_e \log(m-1), \end{aligned}$$

since the maximum of  $H\left(\frac{p_2}{P_e}, \frac{p_3}{P_e}, \dots, \frac{p_m}{P_e}\right)$  is attained by an uniform distribution. Hence any  $X$  that can be predicted with a probability of error  $P_e$  must satisfy

$$H(X) \leq H(P_e) + P_e \log(m-1),$$

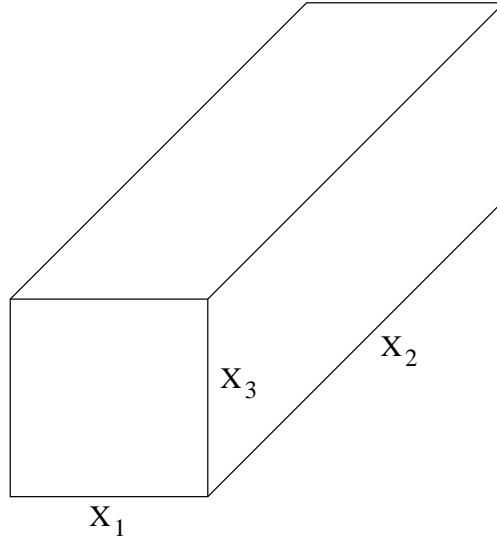
which is the unconditional form of Fano's inequality. We can weaken this inequality to obtain an explicit lower bound for  $P_e$ ,

$$P_e \geq \frac{H(X) - 1}{\log(m-1)}.$$

## 7. Random box size.

An  $n$ -dimensional rectangular box with sides  $X_1, X_2, X_3, \dots, X_n$  is to be constructed. The volume is  $V_n = \prod_{i=1}^n X_i$ . The edge-length  $l$  of an  $n$ -cube with the same volume as the random box is  $l = V_n^{1/n}$ . Let  $X_1, X_2, \dots$  be i.i.d. uniform random variables over the interval  $[0, a]$ .

Find  $\lim_{n \rightarrow \infty} V_n^{1/n}$ , and compare to  $(EV_n)^{1/n}$ . Clearly the expected edge length does not capture the idea of the volume of the box.



**Solution: Random box size.**

The volume  $V_n = \prod_{i=1}^n X_i$  is a random variable. Since the  $X_i$  are i.i.d., uniformly distributed on  $[0, a]$ , we have:

$$\log_e V_n^{\frac{1}{n}} = \frac{1}{n} \log_e V_n = \frac{1}{n} \sum \log_e X_i \rightarrow E(\log_e(X))$$

by the Strong Law of Large Numbers, since  $X_i$  and  $\log_e(X_i)$  are i.i.d. and  $E(\log_e(X)) < \infty$ . Now

$$E(\log_e(X_i)) = \frac{1}{a} \int_0^a \log_e(x) dx = \log_e(a) - 1$$

Hence, since  $e^x$  is a continuous function,

$$\lim_{n \rightarrow \infty} V_n^{\frac{1}{n}} = e^{\lim_{n \rightarrow \infty} \frac{1}{n} \log_e V_n} = \frac{a}{e} < \frac{a}{2}.$$

Thus the “effective” edge length of this solid is  $\frac{a}{e}$ . Note that since the  $X_i$ ’s are independent,  $E(V_n) = \prod E(X_i) = (\frac{a}{2})^n$ . Also  $\frac{a}{2}$  is the arithmetic mean of the random variable, and  $\frac{a}{e}$  is the geometric mean.

**8. An AEP-like limit and the AEP.**

(a) Let  $X_1, X_2, \dots$  be i.i.d. drawn according to probability mass function  $p(x)$ . Find

$$\lim_{n \rightarrow \infty} [p(X_1, X_2, \dots, X_n)]^{\frac{1}{n}}.$$

(b) Let  $X_1, X_2, \dots$  be drawn *i.i.d.* according to the following distribution:

$$X_i = \begin{cases} 2, & \frac{1}{2} \\ 3, & \frac{1}{3} \\ 4, & \frac{1}{6} \end{cases}$$

Find the limiting behavior of the product

$$(X_1 X_2 \cdots X_n)^{1/n}.$$

(c) Evaluate the limit of  $p(X_1, X_2, \dots, X_n)^{\frac{1}{n}}$  for the distribution in part b.

**Solution: An AEP-like limit and the AEP.**

(a)  $X_1, X_2, \dots$ , i.i.d.  $\sim p(x)$ . Hence  $\log(X_i)$  are also i.i.d. and

$$\begin{aligned} \lim(p(X_1, X_2, \dots, X_n))^{\frac{1}{n}} &= \lim 2^{\log(p(X_1, X_2, \dots, X_n))^{\frac{1}{n}}} \\ &= 2^{\lim \frac{1}{n} \sum \log p(X_i)} \\ &= 2^{E(\log(p(X)))} \\ &= 2^{-H(X)} \end{aligned}$$

by the strong law of large numbers.

(b) Let

$$p_n = (X_1 X_2 \cdots X_n)^{\frac{1}{n}}.$$

Then

$$\log p_n = \frac{1}{n} \sum_{i=1}^n \log X_i \rightarrow E \log X,$$

with probability 1, by the strong law of large numbers. Thus  $p_n \rightarrow 2^{E \log X}$  with prob. 1. We can easily calculate  $E \log X = \frac{1}{2} \log 2 + \frac{1}{3} \log 3 + \frac{1}{6} \log 4 = \frac{5}{6} + \frac{\log 3}{3}$ , and therefore  $p_n \rightarrow 2^{\frac{5}{6} + \frac{\log 3}{3}} = 2.5698$ .

(c) Let

$$p_n = p(X_1, X_2, \dots, X_n)^{\frac{1}{n}}.$$

Then by independence we have

$$\log p_n = \frac{1}{n} \sum_{i=1}^n \log p(X_i),$$

and by the strong law of large numbers we get

$$\frac{1}{n} \sum_{i=1}^n \log p(X_i) \rightarrow E \log p(X),$$

with probability 1. Thus  $p_n \rightarrow 2^{-H(X)}$  with prob. 1. We can easily calculate  $H(X) = -\frac{1}{2} \log \frac{1}{2} - \frac{1}{3} \log \frac{1}{3} - \frac{1}{6} \log \frac{1}{6} = 1.4591$ , and therefore  $p_n \rightarrow 2^{-H(X)} = .3637$ .

9. **AEP.**

Let  $(X_i, Y_i)$  be *i.i.d.*  $\sim p(x, y)$ . We form the log likelihood ratio of the hypothesis that  $X$  and  $Y$  are independent vs. the hypothesis that  $X$  and  $Y$  are dependent. What is the limit of

$$\frac{1}{n} \log \frac{p(X^n)p(Y^n)}{p(X^n, Y^n)}?$$

**Solution: AEP.**

Firstly, we should argue that the  $X_i$ 's are jointly independent, and the  $Y_i$ 's are jointly independent. We are told that the *pairs*  $(X_i, Y_i)$  are independent. Hence,

$$\Pr(X_1 = x_1, Y_1 = y_1, \dots, X_n = x_n, Y_n = y_n) = \prod_{i=1}^n \Pr(X_i = x_i, Y_i = y_i)$$

Summing over both sides:

$$\begin{aligned} \Pr(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) &= \sum_{(y_1, \dots, y_n) \in \mathcal{Y}^n} \Pr(X_1 = x_1, Y_1 = y_1, \dots, X_n = x_n, Y_n = y_n) \\ &= \sum_{(y_1, \dots, y_n) \in \mathcal{Y}^n} \prod_{i=1}^n \Pr(X_i = x_i, Y_i = y_i) \\ &= \prod_{i=1}^n \sum_{y_i \in \mathcal{Y}} \Pr(X_i = x_i, Y_i = y_i) \\ &= \prod_{i=1}^n \Pr(X_i = x_i) \end{aligned}$$

and similarly for the  $Y_i$ 's. Hence, we can write  $\Pr(X^n) = \prod_{i=1}^n \Pr(X_i)$ .

We then have

$$\frac{1}{n} \log \frac{p(X^n)p(Y^n)}{p(X^n, Y^n)} = \frac{1}{n} \sum_{i=1}^n \log \frac{p(X_i)p(Y_i)}{p(X_i, Y_i)}.$$

By the strong law of large numbers,

$$\frac{1}{n} \sum_{i=1}^n \log \frac{p(X_i)p(Y_i)}{p(X_i, Y_i)} \rightarrow E \left( \log \frac{p(X)p(Y)}{p(X, Y)} \right),$$

with probability 1. Thus we have

$$\frac{1}{n} \log \frac{p(X^n)p(Y^n)}{p(X^n, Y^n)} \rightarrow -I(X; Y).$$

10. **Entropy of a disjoint mixture.**

Let  $X_1$  and  $X_2$  be discrete random variables drawn according to probability mass functions  $p_1(\cdot)$  and  $p_2(\cdot)$  over the respective alphabets  $\mathcal{X}_1 = \{1, 2, \dots, m\}$  and  $\mathcal{X}_2 = \{m + 1, \dots, n\}$ . Notice that these sets do not intersect. Let

$$X = \begin{cases} X_1, & \text{with probability } \alpha, \\ X_2, & \text{with probability } 1 - \alpha. \end{cases}$$

- (a) Find  $H(X)$  in terms of  $H(X_1)$  and  $H(X_2)$  and  $\alpha$ .
- (b) Maximize over  $\alpha$  to show that  $2^{H(X)} \leq 2^{H(X_1)} + 2^{H(X_2)}$  and interpret using the notion that  $2^{H(X)}$  is the effective alphabet size.
- (c) Let  $X_1$  and  $X_2$  be uniformly distributed over their alphabets. What is the maximizing  $\alpha$  and the associated  $H(X)$ ?

**Solution: Entropy of a disjoint mixture.**

- (a) We can do this problem by writing down the definition of entropy and expanding the various terms. Instead, we will use the algebra of entropies for a simpler proof. Since  $X_1$  and  $X_2$  have disjoint support sets, we can write

$$X = \begin{cases} X_1 & \text{with probability } \alpha \\ X_2 & \text{with probability } 1 - \alpha \end{cases}$$

Define a function of  $X$ ,

$$\theta = f(X) = \begin{cases} 1 & \text{when } X = X_1 \\ 2 & \text{when } X = X_2 \end{cases}$$

Then as in problem 1, we have

$$\begin{aligned} H(X) &= H(X, f(X)) = H(\theta) + H(X|\theta) \\ &= H(\theta) + p(\theta = 1)H(X|\theta = 1) + p(\theta = 2)H(X|\theta = 2) \\ &= H(\alpha) + \alpha H(X_1) + (1 - \alpha)H(X_2) \end{aligned}$$

where  $H(\alpha) = -\alpha \log \alpha - (1 - \alpha) \log(1 - \alpha)$ .

An alternative solution is to expand the entropy directly:

$$\begin{aligned} H(X) &= - \sum_x p(x) \log p(x) \\ &= - \left( \sum_x \alpha p_1(x) \log(\alpha p_1(x)) + \sum_x (1 - \alpha) p_2(x) \log((1 - \alpha) p_2(x)) \right) \\ &= - \left( \alpha \sum_x p_1(x) (\log \alpha + \log p_1(x)) + (1 - \alpha) \sum_x p_2(x) (\log(1 - \alpha) + \log p_2(x)) \right) \\ &= \alpha H(X_1) + (1 - \alpha) H(X_2) + H(\alpha). \end{aligned}$$

- (b) Letting,  $F(\alpha) = H(\alpha) + \alpha H(X_1) + (1 - \alpha)H(X_2)$ , we see that  $F$  is a concave function of  $\alpha$  so that we can maximize it by setting its derivative to 0. Thus solving

$$\begin{aligned} F'(\alpha) &= -\log \alpha + \log(1 - \alpha) + H(X_1) - H(X_2) \\ &= 0, \end{aligned}$$

for  $\alpha$  yields,

$$\alpha^* = \frac{2^{H(X_1)}}{2^{H(X_1)} + 2^{H(X_2)}}.$$

Evaluating  $F$  at this maximizing  $\alpha$  yields,

$$F(\alpha^*) = \log(2^{H(X_1)} + 2^{H(X_2)}).$$

Therefore,

$$\begin{aligned} H(X) = H(\alpha) + \alpha H(X_1) + (1 - \alpha)H(X_2) &= F(\alpha) \\ &\leq F(\alpha^*) \\ &= \log(2^{H(X_1)} + 2^{H(X_2)}), \end{aligned}$$

so that  $2^{H(X)} \leq 2^{H(X_1)} + 2^{H(X_2)}$ , with equality when  $\alpha = \alpha^*$ .

Interpretation of the inequality: The effective alphabet size of the mixture of two distributions is less than the sum of the effective alphabet sizes of the constituent distributions.

- (c) Since  $X_1$  and  $X_2$  are uniform over their alphabets, their entropy is given by the log of the alphabet size. Thus,  $H(X_1) = \log |\mathcal{X}_1| = \log m$ , and  $H(X_2) = \log |\mathcal{X}_2| = \log(n - m)$ . Substituting these values to  $\alpha^*$  from part b gives:

$$\begin{aligned} \alpha^* &= \frac{2^{H(X_1)}}{2^{H(X_1)} + 2^{H(X_2)}} \\ &= \frac{2^{\log m}}{2^{\log m} + 2^{\log(n-m)}} \\ &= \frac{m}{m + (n - m)} \\ &= \frac{m}{n} \end{aligned}$$

The associated  $H(X)$  is:

$$\begin{aligned} H(X) = \log(2^{H(X_1)} + 2^{H(X_2)}) &= \log(m + (n - m)) \\ &= \log(n) \end{aligned}$$

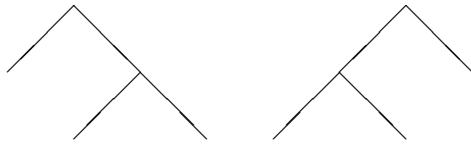
Which makes sense because in this case the appropriate choice of  $\alpha$  results in  $X$  uniform over its alphabet with entropy equal to the log of the alphabet size.

11. **Entropy of a random tree.**

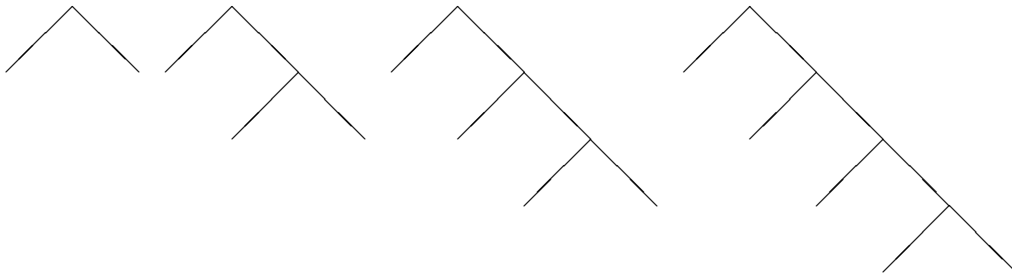
Consider the following method of generating a random tree with  $n$  nodes. First expand the root node:



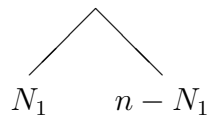
Then expand one of the two terminal nodes at random:



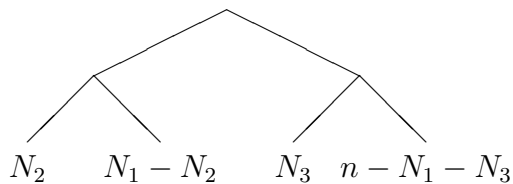
At time  $k$ , choose one of the  $k - 1$  terminal nodes according to a uniform distribution and expand it. Continue until  $n$  terminal nodes have been generated. Thus a sequence leading to a five node tree might look like this:



Surprisingly, the following method of generating random trees yields the same probability distribution on trees with  $n$  terminal nodes. First choose an integer  $N_1$  uniformly distributed on  $\{1, 2, \dots, n - 1\}$ . We then have the picture.



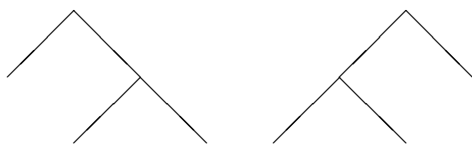
Then choose an integer  $N_2$  uniformly distributed over  $\{1, 2, \dots, N_1 - 1\}$ , and independently choose another integer  $N_3$  uniformly over  $\{1, 2, \dots, (n - N_1) - 1\}$ . The picture is now:



Continue the process until no further subdivision can be made. (The equivalence of these two tree generation schemes follows, for example, from Polya's urn model.)

Now let  $T_n$  denote a random  $n$ -node tree generated as described. The probability distribution on such trees seems difficult to describe, but we can find the entropy of this distribution in recursive form.

First some examples. For  $n = 2$ , we have only one tree. Thus  $H(T_2) = 0$ . For  $n = 3$ , we have two equally probable trees:



Thus  $H(T_3) = \log 2$ . For  $n = 4$ , we have five possible trees, with probabilities  $1/3, 1/6, 1/6, 1/6, 1/6$ .

Now for the recurrence relation. Let  $N_1(T_n)$  denote the number of terminal nodes of  $T_n$  in the right half of the tree. Justify each of the steps in the following:

$$H(T_n) \stackrel{(a)}{=} H(N_1, T_n) \tag{1}$$

$$\stackrel{(b)}{=} H(N_1) + H(T_n|N_1) \tag{2}$$

$$\stackrel{(c)}{=} \log(n-1) + H(T_n|N_1) \tag{3}$$

$$\stackrel{(d)}{=} \log(n-1) + \frac{1}{n-1} \sum_{k=1}^{n-1} [H(T_k) + H(T_{n-k})] \tag{4}$$

$$\stackrel{(e)}{=} \log(n-1) + \frac{2}{n-1} \sum_{k=1}^{n-1} H(T_k). \tag{5}$$

$$\tag{6}$$

(f) Use this to show that

$$(n-1)H_n = nH_{n-1} + (n-1)\log(n-1) - (n-2)\log(n-2), \tag{7}$$

or

$$\frac{H_n}{n} = \frac{H_{n-1}}{n-1} + c_n, \tag{8}$$

for appropriately defined  $c_n$ . Since  $\sum c_n = c < \infty$ , you have proved that  $\frac{1}{n}H(T_n)$  converges to a constant. Thus the expected number of bits necessary to describe the random tree  $T_n$  grows linearly with  $n$ .

**Solution: Entropy of a random tree.**

- (a)  $H(T_n, N_1) = H(T_n) + H(N_1|T_n) = H(T_n) + 0$  by the chain rule for entropies and since  $N_1$  is a function of  $T_n$ .
- (b)  $H(T_n, N_1) = H(N_1) + H(T_n|N_1)$  by the chain rule for entropies.
- (c)  $H(N_1) = \log(n-1)$  since  $N_1$  is uniform on  $\{1, 2, \dots, n-1\}$ .
- (d)

$$H(T_n|N_1) = \sum_{k=1}^{n-1} P(N_1 = k)H(T_n|N_1 = k) \quad (9)$$

$$= \frac{1}{n-1} \sum_{k=1}^{n-1} H(T_n|N_1 = k) \quad (10)$$

by the definition of conditional entropy. Since conditional on  $N_1$ , the left subtree and the right subtree are chosen independently,  $H(T_n|N_1 = k) = H(T_k, T_{n-k}|N_1 = k) = H(T_k) + H(T_{n-k})$ , so

$$H(T_n|N_1) = \frac{1}{n-1} \sum_{k=1}^{n-1} (H(T_k) + H(T_{n-k})). \quad (11)$$

- (e) By a simple change of variables,

$$\sum_{k=1}^{n-1} H(T_{n-k}) = \sum_{k=1}^{n-1} H(T_k). \quad (12)$$

- (f) Hence if we let  $H_n = H(T_n)$ ,

$$(n-1)H_n = (n-1)\log(n-1) + 2 \sum_{k=1}^{n-1} H_k \quad (13)$$

$$(n-2)H_{n-1} = (n-2)\log(n-2) + 2 \sum_{k=1}^{n-2} H_k \quad (14)$$

$$(15)$$

Subtracting the second equation from the first, we get

$$(n-1)H_n - (n-2)H_{n-1} = (n-1)\log(n-1) - (n-2)\log(n-2) + 2H_{n-1} \quad (16)$$

or

$$\frac{H_n}{n} = \frac{H_{n-1}}{n-1} + \frac{\log(n-1)}{n} - \frac{(n-2)\log(n-2)}{n(n-1)} \quad (17)$$

$$= \frac{H_{n-1}}{n-1} + C_n \quad (18)$$

where

$$C_n = \frac{\log(n-1)}{n} - \frac{(n-2)\log(n-2)}{n(n-1)} \quad (19)$$

$$= \frac{\log(n-1)}{n} - \frac{\log(n-2)}{(n-1)} + \frac{2\log(n-2)}{n(n-1)} \quad (20)$$

Substituting the equation for  $H_{n-1}$  in the equation for  $H_n$  and proceeding recursively, we obtain a telescoping sum

$$\frac{H_n}{n} = \sum_{j=3}^n C_j + \frac{H_2}{2} \quad (21)$$

$$= \sum_{j=3}^n \frac{2\log(j-2)}{j(j-1)} + \frac{1}{n}\log(n-1). \quad (22)$$

Since the last term,  $\lim_{n \rightarrow \infty} \frac{1}{n}\log(n-1) = 0$

$$\lim_{n \rightarrow \infty} \frac{H_n}{n} = \sum_{j=3}^{\infty} \frac{2\log(j-2)}{j(j-1)} \quad (23)$$

$$\leq \sum_{j=3}^{\infty} \frac{2\log(j-1)}{(j-1)^2} \quad (24)$$

$$= \sum_{j=2}^{\infty} \frac{2}{j^2} \log j \quad (25)$$

For sufficiently large  $j$ ,  $\log j \leq \sqrt{j}$  and hence the sum in (25) is dominated by the sum  $\sum_j j^{-\frac{3}{2}}$  which converges. Hence the above sum converges to some constant, and the number of bits required to describe a random  $n$ -node tree grows linearly with  $n$ . In fact, computer evaluation shows that the limit is:

$$\lim_{n \rightarrow \infty} \frac{H_n}{n} = \sum_{j=3}^{\infty} \frac{2}{j(j-1)} \log(j-2) = 1.736 \text{ bits per branch.} \quad (26)$$

### Additional Remarks:

Note that the Polya tree generation algorithm described in this problem does not generate trees uniformly at random, because there are more ways to create balanced trees than there are to create imbalanced ones. It is interesting to compare the expected number of bits the Polya algorithm requires to describe a random tree  $T_n$  with that of an “ideal” process that can generate rooted ordered binary trees uniformly at random. To determine  $H(U_n)$ , the entropy for  $n$ -leaf trees generated uniformly at random,

recall that the number of distinct rooted ordered trees with  $n + 1$  leaves is the Catalan number<sup>1</sup>  $C_n$ :

$$C_n := \frac{1}{n+1} \binom{2n}{n}.$$

Using Stirling's approximation, when  $n$  is large,

$$C_n \doteq \frac{1}{n+1} \frac{2^{2n}}{\sqrt{\pi n}}.$$

Thus,

$$\begin{aligned} \lim_{n \rightarrow \infty} \frac{H(U_n)}{n} &= \lim_{n \rightarrow \infty} \frac{\log_2 \frac{1}{n+1} \binom{2n}{n}}{n} \\ &\doteq \lim_{n \rightarrow \infty} \frac{\log_2 \frac{1}{n+1} \frac{2^{2n}}{\sqrt{\pi n}}}{n} \\ &= \lim_{n \rightarrow \infty} \frac{2n - \log_2((n+1)\sqrt{\pi n})}{n} \\ &= 2 \text{ bits per branch.} \end{aligned}$$

Since  $1.736 < 2$ , the Polya tree generation scheme has an entropy rate that falls below that of an ideal tree generator.<sup>2</sup>

Lastly, if you live in the heartland of Wasilla and don't have access to a computer, you can still bound  $\lim_{n \rightarrow \infty} \frac{H(T_n)}{n}$  with only a pencil:

$$\begin{aligned} \lim_{n \rightarrow \infty} \frac{H(T_n)}{n} &= \sum_{j=3}^{\infty} \frac{2}{j(j-1)} \log_2(j-2) \\ &= 2 \sum_{j=3}^{\infty} \frac{1}{j(j-1)} \log_2(j-2) \\ &= 2 \sum_{j=2}^{\infty} \frac{1}{j(j+1)} \log_2(j-1) \\ &= 2 \sum_{j=2}^{\infty} \left( \left( -\frac{1}{j+1} \right) - \left( -\frac{1}{j} \right) \right) \log_2(j-1). \end{aligned}$$

---

<sup>1</sup>Catalan numbers are the answer to a surprising number of seemingly different combinatorial questions, including the number of ways to write  $n$  pairs of parentheses on a paper such that the left and right parentheses are correctly matched, the number of ways to tile a staircase shape of height  $n$  with  $n$  rectangles, the number of ways a convex polygon with  $n + 2$  sides can be cut into triangles by connecting vertices with straight lines, the number of ways ...

<sup>2</sup>Question: The asymptotic effective alphabet size for the ideal tree generator is four. What is the connection between this fact, and the fact that the asymptotic ratio of Catalan numbers is also four?

Applying summation by parts,

$$\begin{aligned}
\lim_{n \rightarrow \infty} \frac{H(T_n)}{n} &= 2 \lim_{j \rightarrow \infty} \left[ \log_2(j-1) \cdot \left(-\frac{1}{j}\right) \right] - 2 \log_2(2-1) \cdot \left(-\frac{1}{2}\right) \\
&\quad + 2 \sum_{j=2}^{\infty} \frac{1}{j+1} \log_2 \frac{j}{j-1} \\
&= 2 \sum_{j=2}^{\infty} \frac{1}{j+1} \log_2 \frac{j}{j-1} \\
&= -2 \sum_{j=2}^{\infty} \frac{1}{j+1} \log_2 \left(1 - \frac{1}{j}\right).
\end{aligned}$$

By the Cauchy-Schwarz inequality,

$$\begin{aligned}
\lim_{n \rightarrow \infty} \frac{H(T_n)}{n} &\leq 2 \sqrt{\sum_{j=2}^{\infty} \left(\frac{1}{j+1}\right)^2} \sqrt{\sum_{j=2}^{\infty} \left(\log_2 \left(1 - \frac{1}{j}\right)\right)^2} \\
&= 2 \sqrt{\sum_{j=3}^{\infty} \frac{1}{j^2}} \sqrt{\sum_{j=2}^{\infty} \left(\log_2 \left(1 - \frac{1}{j}\right)\right)^2} \\
&= \left(2 \sqrt{\frac{\pi^2}{6} - 1 - \frac{1}{4}}\right) \sqrt{\sum_{j=2}^{\infty} \left(\log_2 \left(1 - \frac{1}{j}\right)\right)^2} \\
&= \left(2 \sqrt{\frac{\pi^2}{6} - \frac{5}{4}}\right) \sqrt{\sum_{j=2}^{\infty} \left(\log_2 \left(1 - \frac{1}{j}\right)\right)^2} \\
&= \left(2(\log_2 e) \sqrt{\frac{\pi^2}{6} - \frac{5}{4}}\right) \sqrt{\sum_{j=2}^{\infty} \left(\log \left(1 - \frac{1}{j}\right)\right)^2}.
\end{aligned}$$

We now apply a staircase upper bound to the sum inside the radical (see Figure 1):

$$\begin{aligned}
\sum_{j=2}^{\infty} \left(\log \left(1 - \frac{1}{j}\right)\right)^2 &\leq (\log(1 - 1/2))^2 + \int_2^{\infty} \left(\log \left(1 - \frac{1}{x}\right)\right)^2 dx \\
&= (\log 2)^2 + \int_2^{\infty} \left(\log \left(1 - \frac{1}{x}\right)\right)^2 dx \\
&= (\log 2)^2 + \int_{\frac{1}{2}}^1 \frac{\log^2(u)}{(u-1)^2} du
\end{aligned}$$

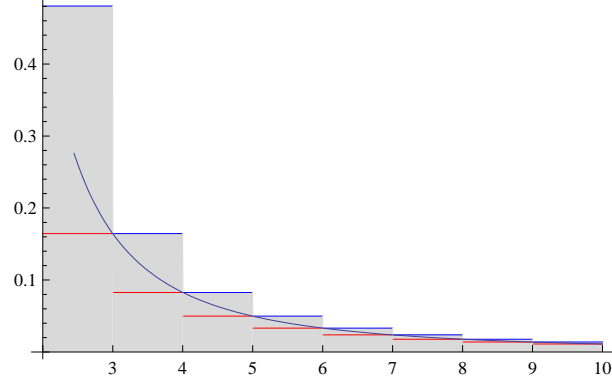


Figure 1: Staircase bounding.

Using residue calculus, the second integral evaluates to  $\frac{\pi^2}{6} - 2(\log 2)^2$ . Thus,

$$\begin{aligned} \lim_{n \rightarrow \infty} \frac{H(T_n)}{n} &\leq \left( 2(\log_2 e) \sqrt{\frac{\pi^2}{6} - \frac{5}{4}} \right) \sqrt{(\log 2)^2 + \frac{\pi^2}{6} - 2(\log 2)^2} \\ &\leq 2(\log_2 e) \left( \sqrt{\frac{\pi^2}{6} - \frac{5}{4}} \right) \left( \sqrt{\frac{\pi^2}{6} - (\log 2)^2} \right) = 1.95674 < 2 = \lim_{n \rightarrow \infty} \frac{H(U_n)}{n}. \end{aligned}$$