ASSIGNMENT 4

Exercise 1 (Binary erasure channel)**.** A binary erasure channel with erasure probability β, denoted BEC(β) has output alphabet $\{0, 1, e\}$ and transitions given by $P(0|0) = P(1|1) = 1-\beta$, and $P(e|0) = P(e|1) = \beta$ where e is the erasure symbol. Show that the capacity of BEC(β) is $1 - \beta$.

Exercise 2 (Z-channel)**.** The Z-channel has binary input and output alphabets and transition probabilities $p(y|x)$ given by:

$$p(0|0) = 1 - p(1|0) = 1,$$
$$p(0|1) = p(1|1) = 1/2.$$

Find the capacity of the Z-channel and the maximizing input probability distribution.

Exercise 3 (Binary multiplier channel)**.** Consider the channel $Y = X \cdot Z$, where X and Z are independent binary random variables and $Z \sim Ber(\alpha)$. [i.e., $P(Z = 1) = \alpha$].

a. Find the capacity of this channel and the maximizing distribution on X.

b. Now suppose that the receiver can observe Z as well as Y. What is the capacity?

Exercise 4 (Unused symbols)**.** Show that the capacity of the channel with transition probability matrix

$$Q = \begin{pmatrix} \frac{2}{3} & \frac{1}{3} & 0 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ 0 & \frac{1}{3} & \frac{2}{3} \end{pmatrix}$$

is achieved by a distribution that places zero probability on one of input symbols. What is the capacity of this channel?

Exercise 5 (Choice of channels)**.** Find the capacity C of the union of two channels $(\mathcal{X}_1, p_1(y_1|x_1), \mathcal{Y}_1)$ and $(\mathcal{X}_2, p_2(y_2|x_2), \mathcal{Y}_2)$, where at each time, one can send a symbol over channel 1 or channel 2 but not both. Assume that the output alphabets are distinct and do not intersect. Show that $2^C = 2^{C_1} + 2^{C_2}$. Thus, 2^C is the effective alphabet size of a channel with capacity C.

Exercise 6. (Collision entropy) For i.i.d. random variables X and Y on $\mathcal{X}$ with distribution P, what is $\mathbb{P}[X = Y]$ in terms of $H_2(P)$ where

$$H_2(P) = -\log \sum_{x \in \mathcal{X}} P(x)^2.$$

$H_2(P)$ is called the Rényi entropy of order 2 or the "collision entropy" of the source.

Exercise 7. (Shotgun DNA sequencing)[1] DNA sequencing is the basic workhorse of modern day biology and medicine. Shotgun sequencing is the dominant technique used: many randomly located

[1] A. Motahari, G. Bresler, and D. Tse, "Information theory of DNA shotgun sequencing." IEEE Transactions on Information Theory 59.10 (2013): 6273-6289.

short fragments called reads are extracted from the DNA sequence, and these reads are assembled to reconstruct the original sequence. A basic question is: given a sequencing technology and the statistics of the DNA sequence, what is the minimum number of reads required for reliable reconstruction?

The DNA sequence $s = s_1 s_2 \cdots s_G$ is modeled as an i.i.d. random process of length G with each symbol taking values according to a probability distribution $p = (p_1, p_2, p_3, p_4)$ on the nucleotide alphabet $\{A, C, G, T\}$. A read is a substring of length L from the DNA sequence. The objective of DNA sequencing is to reconstruct the whole sequence s based on N reads from the sequence. The starting location of each read is uniformly distributed on the DNA sequence and are independent from one read to another. We seek to understand the fundamental limits on the two quantities N and L.

a. *Covering:* Argue that for the perfect reconstruction of s, for a fixed L, the collection of reads should *cover* the entire sequence and hence a necessary condition is that $N \geq G/L$.

b. *An improvement via the coupon collector problem:* The well-known "coupon collector problem" is the following. Suppose we repeatedly and independently sample a random variable that is uniformly distributed over $\{1, 2, \ldots, n\}$. How many samples do we need to ensure the sampling of all n numbers? The answer to this question is roughly $n \log n$ (`https://en.wikipedia.org/wiki/Coupon_collector%27s_problem`).

   Now, consider a modified DNA read technique where in each read, you get to observe L independent locations (instead of contiguous locations). Can you use the coupon collector result to get an estimate on the necessary number of reads N for this modified problem? What does it say about the required number of reads for the original problem?

c. Suppose we have two DNA sequences, the first sequence generated by a uniform distribution on $\{A, C, T, G\}$ and the second by a distribution $(0.5, 0.4, 0.05, 0.05)$. The DNA sequences and the corresponding reads are as follows:

   1. Sequence: $ACTGCATAGT$, Reads: TGC, CAT, ACT, TAG, AGT.
   2. Sequence: $ACACATACGC$, Reads: ACA, CAC, TAC, ACG, CGC

   Impossible to reconstruct?

   i. Which among the two sequences can you reconstruct (uniquely) from the reads? Why?

   ii. Calculate the Rényi entropy of order 2 for both the distributions. Observe that larger the value of H_2, smaller the probability of "collisions" or repeats.

Fig. 4. Two pairs of interleaved repeats of length $L-1$ create ambiguity: from the reads, it is impossible to know whether the sequences $\mathbf{x}$ and $\mathbf{y}$ are as shown, or swapped.

d. We observe that even if we have access to all length-L reads of the sequence, *repeats* make reconstruction impossible (see figure). Denoting by S_i^L the length-L subsequence starting at position i, and R_L the number of length-L repeats, we have

$$\mathbb{E}[R_L] = \sum_{1 \leq i < j \leq G} \mathbb{P}[S_i^L = S_j^L].$$

Justify the following:

$$\mathbb{E}[R_L] > \left(\frac{G^2}{2} - GL\right) e^{-LH_2(P)}.$$

Hint– For a given sequence generated by (p_1, p_2, p_3, p_4), what is the probability that two specific physically disjoint length-ℓ subsequences are identical? In the sum, drop the terms in which S_i^L and S_j^L overlap.

e. *Phase transition:* For $G \gg L$, the above bound may be approximated as

$$\mathbb{E}[R_L] \approx \frac{G^2}{2} e^{-LH_2(P)}.$$

Let $G, L \to \infty$ with $L/\ln G = \overline{L}$, a constant. Conclude that the expected number of repeats approaches zero if

$$\overline{L} > 2/H_2(P)$$

and approaches infinity if

$$\overline{L} < 2/H_2(P).$$

Interpret this result as a prescription for *how large L should be* in order for reconstruction to be successful. Observe that N does not play any role here.

f. Assuming that $\overline{L} > 2/H_2(P)$ and N equals the estimate obtained in part b, conclude that the number of reads (of length L) per nucleotide, given by N/G is roughly $H_2(P)$.