PROBLEM 1. (7 points)

Suppose $X_1, X_2, \ldots$ are i.i.d. random variables with $\Pr(X_1 = 0) = \Pr(X_1 = 1) = 1/2$. Let $Y_n = \sum_{i=1}^{n} X_i$.

Hint: No explicit computation is necessary. For each part you only need to show that the quantity for $n+1$ is at least as large as the quantity for n. "Conditioning reduces entropy" is your friend.

(a) (2 points) Show that $H(Y_n)$ is nondecreasing in n.

Solution: This follows immediately, as

$$\begin{aligned} H(Y_{n+1}) &\geq H(Y_{n+1} \mid X_{n+1}) && \text{[conditioning reduces entropy]} \\ &= H(Y_n \mid X_{n+1}) && [Y_{n+1} = Y_n + X_{n+1}] \\ &= H(Y_n). && [Y_n \text{ is independent of } X_{n+1}] \end{aligned}$$

(b) (2 points) Show that $H(X^n|Y_n)$ is nondecreasing in n.

Solution: This follows from

$$\begin{aligned} H(X^{n+1} \mid Y_{n+1}) &\geq H(X^n \mid Y_{n+1}) && \text{[chain rule, then } H(X_{n+1} \mid Y_{n+1}) \geq 0] \\ &\geq H(X^n \mid Y_{n+1}, Y_n) && \text{[conditioning reduces entropy]} \\ &= H(X^n \mid X_{n+1}, Y_n) && [(Y_{n+1}, Y_n) \mapsto (X_{n+1}, Y_n) \text{ is invertible}] \\ &= H(X^n \mid Y_n). && [X^n \text{ is independent of } X_{n+1}] \end{aligned}$$

(c) (3 points) Show that $H(X_n|Y_n)$ is nondecreasing in n.

Solution: Since $Y_{n+1} = Y_n + X_{n+1}$, with X_{n+1} independent of Y_n, we have $X_n \multimap Y_n \multimap Y_{n+1}$, i.e., X_n, Y_n, Y_{n+1} form a Markov chain. Hence, by the data processing inequality, we have $I(X_n; Y_n) \geq I(X_n; Y_{n+1})$. This gives us the desired result, as

$$\begin{aligned} H(X_{n+1} \mid Y_{n+1}) &= H(X_n \mid Y_{n+1}) && [X_{n+1}, X_n \text{ are identical w.r.t. } Y_{n+1}] \\ &= H(X_n) - I(X_n; Y_{n+1}) \\ &\geq H(X_n) - I(X_n; Y_n) \\ &= H(X_n \mid Y_n). \end{aligned}$$

Remark: This makes formal the intuitive idea that as more independent random variables are added, the influence of each on the sum reduces, and conversely, the uncertainty in each knowing the sum increases.

PROBLEM 2. (9 points)

Suppose $X_1, X_2, \ldots$ is a binary (i.e., $X_n \in \{0, 1\}$) stationary process with entropy rate H. Define the following quantities:

$$\begin{aligned} p &= \Pr(X_1 = 1), \\ \alpha &= \Pr(X_2 = 1 \mid X_1 = 1), \\ \beta &= \Pr(X_2 = 0 \mid X_1 = 0). \end{aligned}$$

(Note that this does not necessarily imply that X is a Markov process.)

(a) (2 points) Show that $p = p\alpha + (1 - p)(1 - \beta)$.

Solution: This follows directly, as

$$\begin{aligned} p &= \Pr(X_1 = 1) = \Pr(X_2 = 1) && [X_n \text{ is stationary}] \\ &= \sum_{i=0,1} \Pr(X_1 = i) \Pr(X_2 = 1 \mid X_1 = i). && [\text{total probability} = p\alpha + (1 - p)(1 - \beta)] \end{aligned}$$

(b) (2 points) Show that $H \leq ph_2(\alpha) + (1 - p)h_2(\beta)$, where h_2 is the binary entropy function given by $h_2(x) = -x \log(x) - (1 - x) \log(1 - x)$.

Solution: Observe that the right-hand side is exactly $H(X_2 \mid X_1)$. The entropy rate for a stationary process is given by

$$\begin{aligned} H &= \lim_{n\to\infty} H(X_n \mid X_{n-1}, \ldots, X_1) \\ &\leq \lim_{n\to\infty} H(X_n \mid X_{n-1}) && [\text{conditioning reduces entropy}] \\ &= H(X_2 \mid X_1), && [X \text{ is stationary}] \end{aligned}$$

and we are done.

(c) (2 points) Show that among all such stationary processes the Markov process has the largest entropy rate. [Recall that for a Markov process, $\Pr(X_n = x_n \mid X^{n-1} = x^{n-1}) = \Pr(X_n = x_n \mid X_{n-1} = x_{n-1})$.]

Solution: In the solution to part (b), observe that the only inequality is in the "conditioning reduces entropy" step. Equality holds here if and only if X_n is conditionally independent of $X_{n-2}, X_{n-3}, \ldots, X_1$ given X_{n-1}, or equivalently, if X is a Markov process. In this case, we have $H = ph_2(\alpha) + (1 - p)h_2(\beta)$, while for all other such stationary processes, we have that $H \leq ph_2(\alpha) + (1 - p)h_2(\beta)$, which completes the proof.

(d) (3 points) Suppose we have a stationary binary process for which every '1' is immediately followed by a '0'. Show that the entropy rate of this process is at most $\max_{a\in[0,1]} \dfrac{h_2(a)}{1 + a}$.

Solution: Observe that this process is a special case of the above with $\alpha = 0$, but β is some number in $[0, 1]$. Then, from part (a), we have $p = (1-p)(1-\beta)$, and from part (b), we have $H \leq (1-p)h_2(\beta)$. Let $b = 1-\beta$, then we have $h_2(b) = h_2(1-\beta) = h_2(\beta)$, since h_2 is symmetric about $\frac{1}{2}$, and

$$1 + b = 1 + \frac{p}{1 - p} = \frac{1}{1 - p} \implies 1 - p = \frac{1}{1 + b}.$$

Substituting these in $H \leq (1-p)h_2(\beta)$, we have $H \leq \frac{h_2(b)}{1+b}$. This holds for some b such that $\beta = 1-b \in [0,1]$, i.e., some $b \in [0,1]$, hence we have $H \leq \frac{h_2(b)}{1+b} \leq \max_{a\in[0,1]} \frac{h_2(a)}{1+a}$.

Remark: Among all stationary processes with a fixed value of the marginal distribution, the i.i.d. process has the largest entropy rate, since the uncertainty is maximized if each element in the sequence is independent. Parts (a)-(c) naturally extend this to all stationary processes with a fixed value of the marginals and one-step transition probabilities — the uncertainty is maximized if there is no more dependence than is necessitated by the fixed transition probabilities. The process in part (d) represents a run-length limited setup, where no two '1's may appear together. Such a constraint often occurs in practice: `https://en.wikipedia.org/wiki/Run-length_limited#Need_for_RLL_coding`.

PROBLEM 3. (9 points)

Suppose $U_1, U_2, \ldots$ are i.i.d. random variables on the alphabet $\mathcal{U}$ with distribution p_U, and define $H := H(U_1)$. Suppose $S_1, S_2, \ldots$ are sets with $S_n \subseteq \mathcal{U}^n$, and define $p_n := \Pr(U^n \in S_n)$. Pick $\epsilon > 0$ and let $T_n = T(n, p_U, \epsilon)$ be the typical sets as defined in class. Let $A_n = T_n \cap S_n$.

(a) (3 points) Show that, for large enough n, $\Pr(U^n \in A_n) \geq p_n - \epsilon$.

Solution: Since $p_n = \Pr(U^n \in S_n)$, the above statement is equivalent to

$$\Pr(U^n \in S_n) - \Pr(U^n \in A_n) \leq \epsilon.$$

The left-hand side can be written as

$$\begin{aligned} \Pr(U^n \in S_n) - \Pr(U^n \in A_n) &= \Pr(U^n \in S_n \setminus A_n) && [A_n \subseteq S_n] \\ &= \Pr(U^n \in S_n \cap T_n^c) && [A_n = T_n \cap S_n] \\ &\leq \Pr(U^n \in T_n^c). \end{aligned}$$

Since the U_i are i.i.d. with distribution p_U, we know that $\lim_{n\to\infty} \Pr(U^n \in T_n) = 1$. Let n_0 be such that $\Pr(U^n \in T_n) \geq 1 - \epsilon$ for all $n \geq n_0$. Then we have that for large enough n, $\Pr(U^n \in T_n^c) \leq \epsilon$, which completes the proof.

(b) (2 points) Show that, for large enough n, $|A_n| \geq (p_n - \epsilon)2^{n(1-\epsilon)H}$.
Hint: Any $u^n \in A_n$ also belongs to T_n.

Solution: Observing that any $u^n \in A_n$ also belongs to T_n, we have that $\Pr(U^n = u^n) \leq 2^{-n(1-\epsilon)H}$ for all $u^n \in A_n$. Combining this with the result from part (a), we have, for large enough n,

$$\begin{aligned} p_n - \epsilon \leq \Pr(U^n \in A_n) &= \sum_{u^n \in A_n} \Pr(U^n = u^n) \\ &\leq \sum_{u^n \in A_n} 2^{-n(1-\epsilon)H} && [u^n \in A_n \subseteq T_n] \\ &= |A_n| 2^{-n(1-\epsilon)H}, \end{aligned}$$

and multiplying both sides by $2^{n(1-\epsilon)H}$, we are done.

(c) (2 points) Show that if $p := \lim_{n\to\infty} p_n > 0$, then $\lim_{n\to\infty} \frac{1}{n} \log |S_n| \geq H$.
Hint: Pick $\epsilon < p$, use (b), and note that S_n includes A_n.

Solution: As given in the hint, pick a positive $\epsilon < p$ (this is possible since $p > 0$). Let n_1 be such that $p_n > \epsilon$ for all $n \geq n_1$ (again, this is possible since the limit is positive) — this is just to ensure that all subsequent logarithms are well-defined. Since $A_n \subseteq S_n$, we have, for large enough n (i.e., $n \geq \max\{n_0, n_1\}$),

$$\log |S_n| \geq \log |A_n| \geq \log(p_n - \epsilon) + n(1-\epsilon)H,$$

and hence, $\lim_{n\to\infty} \frac{1}{n} \log |S_n| \geq (1-\epsilon)H$. Since ϵ can be made arbitrarily small, we are done.

(d) (2 points) Fix $\rho \geq 0$ and let $k_n = \lfloor n\rho \rfloor$. Consider assigning k_n-bit representations to n-letter words u^n via a function $f_n : \mathcal{U}^n \to \{0,1\}^{k_n}$ and attempting to recover the

n-letter word u^n from the representation via a function $g_n : \{0,1\}^{k_n} \to \mathcal{U}^n$. Suppose $\lim_{n\to\infty} \Pr(U^n = g_n(f_n(U^n)) > 0$. Show that $\rho \geq H$.
Hint: Let $S_n := \{u^n : u^n = g_n(f_n(u^n))\}$.

Solution: As given in the hint, let $S_n := \{u^n : u^n = g_n(f_n(u^n))\}$ and $p_n := \Pr(U^n \in S_n)$. Then, we have $p = \lim_{n\to\infty} p_n = \lim_{n\to\infty} \Pr(U^n = g_n(f_n(U^n)) > 0$. By part (c), we have $\lim_{n\to\infty} \frac{1}{n} \log |S_n| \geq H$, i.e.,

$$\begin{aligned} H &\leq \lim_{n\to\infty} \frac{1}{n} \log |S_n| \leq \lim_{n\to\infty} \frac{1}{n} \log 2^{k_n} \\ &= \lim_{n\to\infty} \frac{1}{n} \lfloor n\rho \rfloor = \rho, \qquad [n\rho - 1 < \lfloor n\rho \rfloor \leq n\rho, \text{ sandwich theorem}] \end{aligned}$$

and we are done. We used the fact that $|S_n| \leq 2^{k_n}$, which is true because the range of f_n is $\{0,1\}^{k_n}$, hence $g_n(f_n(u^n))$ can take at most 2^{k_n} values.

Remark: We saw in the lectures that typical sets have probability nearly 1 and size nearly 2^{nH}. In this problem, we not only show that this is as small as such a large-probability set can be, but also that any sequence of sets with nonzero(!) probability in the limit, must have size growing as 2^{nH}, which is a surprising result. Part (d) is an application of this fact — the source coding theorem says that by compressing to a rate more than H, we can recover U^n with probability nearly 1, but here we see that to have even a small, nonzero probability of correctly recovering U^n, we must have a rate at least H. Equivalently, if the rate is even slightly below H, the probability of recovering U^n correctly goes to 0.

PROBLEM 4. (12 points)

Consider two binary codes, c_1 and c_2 for the *nonnegative* integers $\{0, 1, 2, \dots\}$. The code c_1 is defined as $c_1(n) = 1^n0$; e.g, $c_1(3) = 1110$. The code c_2 is given as follows: $c_2(0) =$ null, $c_2(1) = 0, c_2(2) = 1, c_2(3) = 00, c_2(4) = 01, c_2(5) = 10, c_2(6) = 11, c_2(7) = 000$, and so on. Observe that $\text{length}(c_1(n)) = 1 + n$, and $\text{length}(c_2(n)) = \lfloor \log(1 + n) \rfloor$.

(a) (2 points) Is c_1 injective? Is it prefix-free? Is c_2 injective? Is it prefix-free?

Solution: Clearly, c_1 is both injective and prefix-free, but c_2 is only injective, not prefix-free.

(b) (2 points) With $n_1 = \text{length}(c_2(n))$, and $n_2 = \text{length}(c_2(n_1))$, consider the code formed by a concatenation $c(n) = c_1(n_2)c_2(n_1)c_2(n)$. Explain why c is prefix-free.

Solution: Suppose there exist nonnegative integers n, m such that $c(n)$ is a prefix of $c(m)$, with n_1, n_2, m_1, m_2 defined analogously. Since $c(n)$ and $c(m)$ start with $c_1(n_2)$ and $c_1(m_2)$ respectively, and c_1 is a prefix-free code, the only way for $c(n)$ to be a prefix of $c(m)$ is if $c_1(n_2) = c_1(m_2)$, i.e., $n_2 = m_2$, or equivalently, $\text{length}(c_2(n_1)) = \text{length}(c_2(m_1))$. Now, since the initial segments match, we require that the remainder of $c(n)$ is a prefix of $c(m)$, i.e., $c_2(n_1)c_2(n)$ is a prefix of $c_2(m_1)c_2(m)$. However, the lengths of $c_2(n_1)$ and $c_2(m_1)$ are equal, and among codewords of the same length, no codeword of c_2 is a prefix of the other. Hence, we must have $n = m$, and c is prefix-free.

(c) (3 points) Show that there is a prefix-free code c_3 for *positive* integers $\{1, 2, \dots\}$ with $\text{length}(c_3(n)) \le \log(n) + 2\log(1 + \log(n)) + 1$.

Solution: We use the same code as in part (c), except that we encode $n - 1$ instead of n, since we now need a code for positive integers only. Formally, we define $c_3(n) = c_1(n_2)c_2(n_1)c_2(n - 1)$, with $n_1 = \text{length}(c_2(n - 1)) = \lfloor \log(n) \rfloor$, and $n_2 = \text{length}(c_2(n_1)) = \lfloor \log(1 + n_1) \rfloor$, for $n \in \{1, 2, \dots\}$, which we now know to be prefix-free. All that is left to show is the bound on the length of $c_3(n)$, which comes from

$$\begin{aligned}
\text{length}(c_3(n)) &= \text{length}(c_1(n_2)) + \text{length}(c_2(n_1)) + \text{length}(c_2(n-1)) \\
&= 1 + n_2 + \lfloor \log(1 + n_1) \rfloor + \lfloor \log(n) \rfloor \\
&= 1 + 2\lfloor \log\left(1 + \lfloor \log(n) \rfloor\right) \rfloor + \lfloor \log(n) \rfloor \\
&\le 1 + 2\log(1 + \log(n)) + \log(n),
\end{aligned}$$

and we are done.

Suppose that $\dots, U_{-2}, U_{-1}, U_0, U_1, U_2, \dots$ are i.i.d. from an alphabet $\mathcal{U}$, and we observe U_i at time i. Let $N_i = \inf\{j > 0 : U_{i-j} = U_i\}$, i.e., the symbol U_i we observed at time i, was most recently observed at time $i - N_i$.

(d) (2 points) Show that $\Pr(N_i > j \mid U_i = u) = (1 - p_U(u))^j$, $j = 0, 1, \dots$. Conclude that $\mathbb{E}[N_i \mid U_i = u] = 1/p_U(u)$. [Fact: If $X \in \{0, 1, 2, \dots\}$, then $\mathbb{E}[X] = \sum_{i=0}^{\infty} \Pr(X > i)$.]

Solution: For $j = 0$, this is trivial, since both sides are 1. For $j \ge 1$, the event $\{N_i > j\}$ is identical to the event that none of $U_{i-1}, U_{i-2}, \dots, U_{i-j}$ is equal to U_i,

hence

$$\begin{aligned}\Pr(N_i > j \mid U_i = u) &= \Pr(U_{i-j} \neq u, \ldots, U_{i-1} \neq u \mid U_i = u) \\ &= \Pr(U_{i-j} \neq u) \cdots \Pr(U_{i-1} \neq u) \qquad [U_i \text{ are i.i.d.}] \\ &= (1 - p_U(u))^j.\end{aligned}$$

We can use this to compute the expectation conditioned on $\{U_i = u\}$ as

$$\mathbb{E}[N_i \mid U_i = u] = \sum_{j=0}^{\infty} \Pr(N_i > j \mid U_i = u) = \sum_{j=0}^{\infty} (1 - p_U(u))^j = \frac{1}{p_U(u)}.$$

Suppose that we have already described the "past" $(\ldots, U_{-2}, U_{-1}, U_0)$ in binary, we now describe $U_1, U_2, \ldots$, by giving a binary descriptions of $N_1, N_2, \ldots$, by the code c_3 above.

(e) (3 points) With $H = H(U_i) = H(U_1)$, show that $\mathbb{E}[\text{length}(c_3(N_i))] \leq H + 2\log(1 + H) + 1$.
Hint: First condition on $\{U_i = u\}$.

Solution: We first condition on $\{U_i = u\}$, and use part (c) to get

$$\begin{aligned}\mathbb{E}[\text{length}(c_3(N_i)) \mid U_i = u] &\leq \mathbb{E}[\log N_i + 2\log(1 + \log N_i) + 1 \mid U_i = u] \\ &\leq \log \mathbb{E}[N_i \mid U_i = u] + 2\log(1 + \log \mathbb{E}[N_i \mid U_i = u]) + 1 \\ &= \log \frac{1}{p_U(u)} + 2\log\left(1 + \log \frac{1}{p_U(u)}\right) + 1,\end{aligned}$$

where the second inequality follows from the concavity of log. We can now compute the unconditional expectation as

$$\begin{aligned}\mathbb{E}[\text{length}(c_3(N_i))] &= \sum_{u \in \mathcal{U}} p_U(u)\, \mathbb{E}[\text{length}(c_3(N_i)) \mid U_i = u] \\ &\leq \sum_{u \in \mathcal{U}} p_U(u) \left[\log \frac{1}{p_U(u)} + 2\log\left(1 + \log \frac{1}{p_U(u)}\right) + 1\right] \\ &\leq H + 2\log(1 + H) + 1,\end{aligned}$$

where the last step follows, once again, by the concavity of log, and we are done.

Remark: This problem describes a universal compression scheme. It has the feature that the only codes needed are for positive integers, irrespective of the alphabet $\mathcal{U}$, but seems to induce an overhead of $2\log(1+H)+1$ bits. By taking $V_1 = (U_1, \ldots, U_n), V_2 = (U_{n+1}, \ldots, U_{2n})$ and so on, and applying part (e) to these V_i, as n goes to infinity, the overhead can be made as small as needed, hence this scheme can also compress any i.i.d. source to its entropy. Further, the result on the expectation conditioned on $\{U_i = u\}$ in part (d) holds even for stationary ergodic processes, so this method compresses any stationary ergodic process to its entropy rate.