Problem 1.

(a) From the multinomial formula, for any non-negative $x_1, \ldots, x_K$ with $x_1 + \cdots + x_K = 1$ we have

$$1 = (x_1 + \cdots + x_K)^n = \sum_{n_1, \ldots, n_K: n_1 + \cdots + n_K = n} \binom{n}{n_1, \ldots, n_K} x_1^{n_1} \cdots x_K^{n_K} \geq \binom{n}{n_1, \ldots, n_K} x_1^{n_1} \cdots x_K^{n_K}.$$ 

(b) From (a), for any non-negative $x_1, \ldots, x_n$ that sum to 1,

$$\log |S_{n_1, \ldots, n_K}| = \log \binom{n}{n_1, \ldots, n_K} \leq n \sum_{i=1}^K \frac{n_i}{n} \log \frac{1}{x_i}.$$ 

Choose now $x_i = n_i/n$ to obtain the desired result.

(c) Consider the following code: given $(u_1, \ldots, u_n)$, compute $n_1, \ldots, n_K$. Describe each of the $n_i \in \{0, 1, \ldots n\}, i = 1, \ldots, K - 1$, using $\lceil \log(1 + n) \rceil$ bits in the usual binary encoding of integers (no need to describe $n_K$ since the $n_i$’s sum to $n$). At this moment the decoder will know that the sequence $u_1, \ldots, u_n$ belongs to $S_{n_1, \ldots, n_K}$, and thus with further $\lceil \log |S_{n_1, \ldots, n_K}| \rceil$ bits we can describe which element of $S_{n_1, \ldots, n_K}$ we were given.

An alternative solution consists of verifying that the given codeword lengths satisfy the Kraft’s inequality: let $\ell_0 := (K - 1)\lceil \log(n + 1) \rceil$ and $\ell_1(u_1, \ldots, u_n) := \lceil \log |S_{n_1, \ldots, n_K}| \rceil$ (with $n_1, \ldots, n_K$ as before) so that the codeword lengths are

$$\ell(u_1, \ldots, u_n) = \ell_0 + \ell_1(u_1, \ldots, u_n).$$ 

Then,

$$\sum_{u_1, \ldots, u_n} 2^{-\ell(u_1, \ldots, u_n)} = \sum_{n_1, \ldots, n_K: n_1 + \cdots + n_K = n} \sum_{u_1, \ldots, u_n \in S_{n_1, \ldots, n_K}} 2^{-\ell_0 - \ell_1(u_1, \ldots, u_n)} \leq \sum_{n_1, \ldots, n_K: n_1 + \cdots + n_K = n} 2^{-\ell_0} \sum_{u_1, \ldots, u_n \in S_{n_1, \ldots, n_K}} 1/|S_{n_1, \ldots, n_K}| = 2^{-\ell_0} \sum_{n_1, \ldots, n_K: n_1 + \cdots + n_K = n} 1.$$ 

The last sum contains at most $(n + 1)^{K-1}$ terms: for each of $n_1, \ldots, n_{K-1}$ there are at most $(n + 1)$ choices and once $(n_1, \ldots, n_{K-1})$ is chosen there is but a single choice for $n_K$. As $2^{\ell_0} \geq (n + 1)^{K-1}$ we see that the Kraft’s inequality is satisfied and a prefix-free code with the specified lengths exists.

(d) We have

$$0 \leq E[D((X_1, \ldots, X_K)\| (\mu_1, \ldots, \mu_K))] = \sum_i E[X_i \log(X_i/\mu_i)] = -E[h(X_1, \ldots, X_n)] + \sum_i E[X_i] \log(1/\mu_i) = -E[h(X_1, \ldots, X_n)] + h(\mu_1, \ldots, \mu_n).$$
(e) Let $N_i$ be the number of occurrences of the symbol $i$ in the sequence $U_1, \ldots, U_n$. By (c) and (b)

$$\text{length} \left( C_n(U_1, \ldots, U_n) \right) \leq (K - 1) \lceil \log(1 + n) \rceil + \lceil nh(N_1/n, \ldots, N_K/n) \rceil$$

$$\leq K + (K - 1) \log(1 + n) + nh(N_1/n, \ldots, N_K/n)$$

Note that $E[N_i] = np_i$ where $p_i = \Pr(U = i)$, and thus by (d) we have

$$\frac{1}{n} E[\text{length}(C_n(U_1, \ldots, U_n))] \leq \frac{K + (K - 1) \log(1 + n)}{n} + h(p_1, \ldots, p_K).$$

Noting that $H(U) = h(p_1, \ldots, p_K)$, we demonstrate what was asked.

Observe that in constructing the code $C_n$ we did not use any knowledge of the statistics of $U$, but for i.i.d. sources, we see that for large $n$ the code performs as well a code that is designed with the knowledge of the statistics. The ‘universality penalty’ we pay is $O((K \log n)/n)$. 

Problem 2.

(a) Since \( \{X_i : i \in \mathbb{Z}\} \) is stationary, \((U_1, \ldots, U_n) = (f(X_1), \ldots, f(X_n))\) has the same statistics as \((f(X_{k+1}), \ldots, f(X_{k+n})) = (U_{k+1}, \ldots, U_{k+n})\). Thus the process \(\{U_i : i \in \mathbb{Z}\}\) is stationary. Consequently, the sequence \(a_i\) is non-increasing, and \(\lim a_i\) exists and is equal to the entropy rate of the process \(\{U_i : i \in \mathbb{Z}\}\).

(b) Since \(\{X_i : i \in \mathbb{Z}\}\) is Markov, conditional on \(X_1\) the sequence \((X_2, \ldots, X_{i+1})\) is independent of \(X_0\). Since \((U_2, \ldots, U_{i+1})\) is a function of \((X_2, \ldots, X_{i+1})\) we thus see that conditional on \(X_1\), the sequence \((U_2, \ldots, U_{i+1})\) is also independent of \(X_0\). Consequently, \(I(X_0; U_2, \ldots, U_{i+1}|X_1) = 0\).

(c) By stationarity \(b_i = H(U_{i+1}|U_i, \ldots, U_2, X_1)\). Thus,

\[
b_i - H(U_{i+1}|U_i, \ldots, U_2, X_1, X_0) = I(X_0; U_{i+1}|U_i, \ldots, U_2, X_1).
\]

But from (b) and the chain rule we have

\[
0 = I(X_0; U_2, \ldots, U_{i+1}|X_1) = \sum_{j=2}^{i+1} I(X_0; U_j|U_2, \ldots, U_{j-1}, X_1)
\]

and conclude that each term in the sum above, in particular \(I(X_0; U_{i+1}|U_i, \ldots, U_2, X_1)\), equals zero. We thus find that \(b_i = H(U_{i+1}|U_i, \ldots, U_2, X_1, X_0)\) as claimed.

(d) From (c) and the fact that \(U_1\) is a function of \(X_1\)

\[
b_i = H(U_{i+1}|U_i, \ldots, U_2, X_1, X_0) = H(U_{i+1}|U_i, \ldots, U_1, X_1, X_0) \leq H(U_{i+1}|U_i, \ldots, U_1, X_0) = b_{i+1}.
\]

(e) Observe that \(d_i = I(X_0; U_i|U_1, \ldots, U_{i-1})\). So \(d_i \geq 0\), and by the chain rule \(\sum_{i=1}^n d_i = I(X_0; U_1, \ldots, U_n)\).

(f) Since \(a_i \geq a_{i+1}\) (see comments in (a)) and \(b_i \leq b_{i+1}\) (by (d)), \(d_{i+1} = a_{i+1} - b_{i+1} \leq a_i - b_i = d_i\).

(g) From (f) and (e)

\[
n d_n \leq d_1 + \cdots + d_n = I(X_0; U_1, \ldots, U_n) \leq H(X_0) \leq \log |\mathcal{X}|.
\]

Thus \(\lim_{n \to \infty} d_n = 0\). Consequently, \(\lim_{n \to \infty} b_n = \lim_{n \to \infty} a_n\).

A process \(\{U_i : i \in \mathbb{Z}\}\) as in this problem is called a ‘hidden Markov process.’ Observe that for a stationary process the sequence \(a_n\) converges to the entropy rate \(H\) from above, but in general there is no way how large one should take \(n\) to get a good estimate of \(H\). We now see that for hidden Markov processes we have another sequence \(b_n\) that converges to \(H\) from below, and taking \(n = \log |\mathcal{X}|/\epsilon\) guarantees that \(b_n \leq H \leq a_n\) with \(a_n - b_n \leq \epsilon\).
Problem 3.

(a) Note that when $W \neq w_0$, we have $W' = W$, and when $W = w_0$ we have $W' = w_0u$ for some $u \in U$. Thus

$$\text{length}(W') - \text{length}(W) = \mathbf{1}(W = w_0).$$

Thus $E[\text{length}(W')] - E[\text{length}(W)]$ equals $\Pr(W = w_0) = p_0$.

(b) We have

$$H(W') - H(W) = \sum_{u \in U} p(w_0u) \log \frac{1}{p(w_0u)} - p_0 \log \frac{1}{p_0}$$

The first sum equals

$$\sum_{u} p_0 p(u) \log \frac{1}{p_0 p(u)} = p_0 \left[ \log \frac{1}{p_0} + H(U) \right],$$

consequently $H(W') - H(W) = p_0 H(U)$.

(c) The only dictionary with $k = 1$ interior node is $D = U$. For this dictionary $\text{length}(W) = 1$ and $H(W) = H(U)$ so $S_1$ is true.

(d) Any dictionary $D'$ with $k + 1$ interior nodes is obtained from a dictionary $D$ with $k$ interior nodes by the construction described in the problem. Consequently, from (b), hypothesis $S_k$, and (a)

$$H(W') = H(W) + p_0 H(U) = E[\text{length}(W')] H(U) + p_0 H(U) = E[\text{length}(W')] H(U)$$

proving $S_{k+1}$. The statement that $S_k$ is true for all $k$ then follows by induction.

In class we had proved this relationship between $H(W)$, $H(U)$ and $E[\text{length}(W)]$ by a more complicated proof than the one described in this problem.