Problem Set 2 Solution

Solutions courtesy of Joy A. Thomas, with editing by Frank R. Kschischang.

2.14 *Entropy of a sum.*

(a) Let $\mathcal{X} = \{x_1, \ldots, x_r\}$ and let $\mathcal{Y} = \{y_1, \ldots, y_s\}$. Note that for any fixed $x \in \mathcal{X}$, $H(Y + x \mid X = x) = H(Y \mid X = x) = -\sum_{y \in \mathcal{Y}} p(y \mid x) \log p(y \mid x)$. Now, since $Z = X + Y$ we have

$$\begin{aligned} H(Z \mid X) &= \sum_{x \in \mathcal{X}} p(x) H(Z \mid X = x) \\ &= -\sum_{x \in \mathcal{X}} p(x) H(Y + x \mid X = x) \\ &= -\sum_{x \in \mathcal{X}} p(x) H(Y \mid X = x) \\ &= H(Y \mid X). \end{aligned}$$

If X and Y are independent, then $H(Y \mid X) = H(Y)$ and, in this case,

$$H(Z) \geq H(Z \mid X) = H(Y \mid X) = H(Y).$$

Reversing the roles of X and Y we can similarly show that $H(Z) \geq H(X)$.

(b) Let

$$X = -Y = \begin{cases} 0 & \text{with probability } \frac{1}{2}, \\ 1 & \text{with probability } \frac{1}{2}. \end{cases}$$

Then $H(X) = H(Y) = 1$, yet $Z = 0$ with probability 1 and hence $H(Z) = 0$.

(c) Since Z is a function of X and Y, we have $H(Z) \leq H(X, Y)$ and, in turn, $H(X, Y) \leq H(X) + H(Y)$. To have equality in both inequalities, we need, first, that X and Y can be recovered uniquely from Z, and, second, that X and Y are independent. For example, if X is uniformly distributed over $\{0, 1\}$ and Y is uniformly distributed over $\{0, 2\}$ (and independent of X), then Z is uniformly distributed over $\{0, 1, 2, 3\}$, and we have $H(Z) = H(X) + H(Y)$.

2.16 *Bottleneck.*

(a) The data-processing inequality provides that $I(X_1; X_3) \leq I(X_1; X_2)$. However, $I(X_1; X_2) = H(X_2) - H(X_2 \mid X_1) \leq H(X_2) \leq \log k$, where the first inequality follows since $H(X_2 \mid X_1) \geq 0$, and the second inequality follows from the fact that the uniform distribution on an alphabet of size k has the maximum possible entropy of $\log k$. It follows that $I(X_1; X_3) \leq \log k$.

(b) If $k = 1$, $\log k = 0$, and we have $I(X_1; X_3) \leq 0$. However, since $I(X_1; X_3) \geq 0$, we conclude that $I(X_1; X_3) = 0$, i.e., X_1 and X_3 are independent when $k = 1$.

2.17 *Pure randomness and bent coins.*

$$\begin{aligned} nH(p) &\overset{(a)}{=} H(X_1, \ldots, X_n) \\ &\overset{(b)}{\geq} H(Z_1, Z_2, \ldots, Z_K, K) \\ &\overset{(c)}{=} H(K) + H(Z_1, \ldots, Z_K \mid K) \\ &\overset{(d)}{=} H(K) + E(K) \\ &\overset{(e)}{\geq} E(K). \end{aligned}$$

(a) Since the X_i's are independent, we have $H(X_1, \ldots, X_n) = H(X_1) + \cdots + H(X_n)$, and since they are identically distributed with entropy $H(p)$, we have $H(X_1, \ldots, X_n) = nH(p)$.

(b) Random variables K and $Z_1, \ldots, Z_K$ are functions of $X_1, \ldots, X_n$ and since the entropy of a function of a random variable is not greater than the entropy of the random variable, the result follows.

(c) This is the chain rule for entropy.

(d) We have $H(Z_1, \ldots, Z_K \mid K) = \sum_k H(Z_1, \ldots, Z_K \mid K = k)p(k) = \sum_k kp(k) = E(K)$, where the second-last inequality follows since the Z_i's are independent and each of unit entropy.

(e) Follows from fact that $H(K) \geq 0$.

Since p is unknown, the only way to generate pure random bits is to use the fact that all sequences with the same number of ones are equally likely. for example, the sequences 0001, 0010, 0100, and 1000 are equally likely and can be used to generate 2 pure random bits. An example of a mapping to generate random bits is

$$\begin{array}{cccc} 0000 \to \epsilon & & & \\ 0001 \to 00 & 0010 \to 01 & 0100 \to 10 & 1000 \to 11 \\ 0011 \to 00 & 0110 \to 01 & 1100 \to 10 & 1001 \to 11 \\ 1010 \to 0 & 0101 \to 1 & & \\ 1110 \to 11 & 1101 \to 10 & 1011 \to 01 & 0111 \to 00 \\ 1111 \to \epsilon & & & \end{array}$$

The resulting expected number of bits is

$$\begin{aligned} E(K) &= 4pq^3 \cdot 2 + 4p^2q^2 \cdot 2 + 2p^2q^2 \cdot 1 + 4p^3q \cdot 2 \\ &= 8pq^3 + 10p^2q^2 + 8p^3q, \end{aligned}$$

where $q = 1 - p$. For example, for $p \approx \frac{1}{2}$, the expected number of pure random bits is close to 1.625. This is substantially less then the 4 pure random bits that could be generated if p were known to be exactly $\frac{1}{2}$.

We will now analyze the efficiency of this scheme of generating random bits for long sequences of bent coin flips. Let n be the number of bent coin flips. The algorithm

that we will use is the obvious extension of the above method of generating pure bits using the fact that all sequences with the same number of ones are equally likely.

Consider all sequences with k ones. There are $\binom{n}{k}$ such sequences, which are all equally likely. If $\binom{n}{k}$ were a power of 2, then we could generate $\log_2 \binom{n}{k}$ pure random bits from such a set. However, in the general case, $\binom{n}{k}$ is not a power of 2 and the best we can to is the divide the set of $\binom{n}{k}$ elements into subsets of sizes which are powers of 2. The largest set would have a size $2^{\lfloor \log \binom{n}{k} \rfloor}$ and could be used to generate $\lfloor \log \binom{n}{k} \rfloor$ random bits. We could divide the remaining elements into the largest set which is a power of 2, etc. The worst case would occur when $\binom{n}{k} = 2^{l+1} - 1$, in which case the subsets would be of sizes $2^l, 2^{l-1}, 2^{l-2}, \ldots, 1$.

Instead of analyzing the scheme exactly, we will just find a lower bound on number of random bits generated from a set of size $\binom{n}{k}$. Let $l = \lfloor \log \binom{n}{k} \rfloor$. Then at least half of the elements belong to a set of size 2^l and would generate l random bits, at least $\frac{1}{4}$th belong to a set of size 2^{l-1} and generate $l-1$ random bits, etc. On average, the number of bits generated is

$$E[K \mid k \text{ 1's in sequence}] \geq \frac{1}{2}l + \frac{1}{4}(l-1) + \cdots + \frac{1}{2^l}1 \triangleq S.$$

Note that $2S = l + \frac{l-1}{2} + \frac{l-2}{4} + \cdots + \frac{1}{2^{l-1}}$, thus

$$\begin{aligned} S &= 2S - S \\ &= l - \left(\frac{1}{2} + \frac{1}{4} + \cdots + \frac{1}{2^l} \right) \\ &\geq l - 1, \end{aligned}$$

since the infinite series sums to 1. Thus the fact that $\binom{n}{k}$ is not a power of 2 will cost at most 1 bit on the average in the number of random bits that are produced.

Hence, the expected number of pure random bits produced by this algorithm is

$$\begin{aligned} E(K) &\geq \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} \lfloor \log \binom{n}{k} - 1 \rfloor \\ &\geq \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} \left(\log \binom{n}{k} - 2 \right) \\ &= \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} \log \binom{n}{k} - 2 \\ &\geq \sum_{n(p-\epsilon) \leq k \leq n(p+\epsilon)} \binom{n}{k} p^k q^{n-k} \log \binom{n}{k} - 2. \end{aligned}$$

Now for sufficiently large n, the probability that the number of 1's in the sequence is close to np is near 1 (by the weak law of large numbers). For such sequences, $\frac{k}{n}$ is close to p and hence there exists a δ such that

$$\binom{n}{k} \geq 2^{n(H(\frac{k}{n}) - \delta)} \geq 2^{n(H(p) - 2\delta)}$$

using Stirling's approximation for the binomial coefficients and the continuity of the entropy function. If we assume that n is large enough so that the probability that $n(p-\epsilon) \leq k \leq n(p+\epsilon)$ is greater than $1-\epsilon$, then we see that

$$E(K) \geq (1-\epsilon)n(H(p)-2\delta)-2,$$

which is very good since $nH(p)$ is an upper bound on the number of pure random bits that can be produced from the bent coin sequence.

2.19 *Infinite entropy.* Let us first show that A is finite. We will assume logarithms are to base e. For $x > 1$, let $f(x) = \frac{1}{x \log^2 x} = -\frac{\mathrm{d}}{\mathrm{d}x}\frac{1}{\log(x)}$. Then $f(x)$ is positive, decreasing monotonically with x, and we have $\sum_{n=3}^{\infty} f(n) \leq \int_{x=2}^{\infty} f(x)\mathrm{d}x = \frac{1}{\log(3)}$, and so $A \leq \frac{1}{2\log^2 2} + \frac{1}{\log 3} < \infty$. Now

$$\begin{aligned}
H(X) &= -\sum_{n=2}^{\infty} p(n)\log p(n) \\
&= \sum_{n=2}^{\infty} p(n)\log(An\log^2 n) \\
&= \sum_{n=2}^{\infty} p(n)(\log(A)+\log(n)+\log(\log^2 n)) \\
&= \log(A) + \sum_{n=2}^{\infty} p(n)\log(n) + \sum_{n=2}^{\infty} p(n)\log(\log^2 n).
\end{aligned}$$

The first term is finite. Regardless of the base of the logarithm, the last sum has only finitely many negative terms, and so cannot diverge to $-\infty$. (For example, if the base of the logarithm is 2, the last sum has only nonnegative terms.) The middle sum diverges since

$$\begin{aligned}
\sum_{n=2}^{\infty} \frac{\log n}{An\log^2 n} &= \frac{1}{A}\sum_{n=2}^{\infty}\frac{1}{n\log n} \\
&\geq \frac{1}{A}\int_{x=2}^{\infty}\frac{1}{x\log x}\mathrm{d}x \\
&\geq \frac{1}{A}\left(\lim_{a\to\infty}(\log(\log a)) - \log(\log(2))\right) \\
&= \infty.
\end{aligned}$$

Thus $H(X) = +\infty$.

The divergence of the middle sum can also be seen by an appeal to Cauchy's conden-

sation test. If $a_2, a_3, a_4, \ldots$ is a positive non-increasing sequence, then

$$\begin{aligned}\sum_{n=2}^{\infty} a_n &= \underbrace{a_2} + \underbrace{a_3 + a_4} + \underbrace{a_5 + \cdots + a_8} + \cdots \\ &\geq a_2 + 2 \cdot a_4 + 4 \cdot a_8 + \cdots \\ &= \sum_{n=1}^{\infty} 2^{n-1} a_{2^n}.\end{aligned}$$

Taking logarithms to the base two, we see that

$$\sum_{n=2}^{\infty} \frac{1}{n \log n} \geq \sum_{n=1}^{\infty} 2^{n-1} \frac{1}{2^n \cdot n} = \frac{1}{2} \sum_{n=1}^{\infty} \frac{1}{n} = \infty.$$

2.21 *Markov's inequality for probabilities.* Define the random variable $Y = \log\left(\frac{1}{p(X)}\right)$ as a function of X. Since $p(x) \leq 1$, Y takes non-negative values and thus Markov's inequality applies to Y. In particular, for any $a > 0$, $P[Y \geq a] \leq \frac{E(Y)}{a}$. Note that $E(Y) = \sum_{x \in \mathcal{X}} p(x) \log(1/p(x)) = H(X)$. Taking $a = \log(1/d)$, we get

$$\begin{aligned}P(Y \geq \log(1/d)) &= P(\log(1/p(X)) \geq \log(1/d)) \\ &= P(1/p(X) \geq 1/d) \\ &= P(p(X) \leq d) \leq \frac{H(X)}{\log(1/d)},\end{aligned}$$

from which the result follows.

2.27 *Grouping rule for entropy.* By definition,

$$H(\mathbf{p}) = \sum_{i=1}^{m} p_i \log_2 \left(\frac{1}{p_i}\right),$$

and

$$H(\mathbf{q}) = \sum_{i=1}^{m-2} p_i \log_2 \left(\frac{1}{p_i}\right) + (p_{m-1} + p_m) \log_2 \left(\frac{1}{p_{m-1} + p_m}\right).$$

Therefore,

$$\begin{aligned}H(\mathbf{p}) - H(\mathbf{q}) &= -p_{m-1} \log_2 p_{m-1} - p_m \log_2 p_m + p_{m-1} \log_2(p_{m-1} + p_m) + p_m \log_2(p_{m-1} + p_m) \\ &= -p_{m-1} \log_2 \left(\frac{p_{m-1}}{p_{m-1} + p_m}\right) - p_m \log_2 \left(\frac{p_m}{p_{m-1} + p_m}\right) \\ &= -(p_{m-1} + p_m) \left[\frac{p_{m-1}}{p_{m-1} + p_m} \log_2 \left(\frac{p_{m-1}}{p_{m-1} + p_m}\right) + \frac{p_m}{p_{m-1} + p_m} \log_2 \left(\frac{p_m}{p_{m-1} + p_m}\right)\right] \\ &= (p_{m-1} + p_m) H\left(\frac{p_{m-1}}{p_{m-1} + p_m}, \frac{p_m}{p_{m-1} + p_m}\right),\end{aligned}$$

from which the result follows.

Another approach to this same problem is as follows. Let $X \in \{1, \ldots, m\}$ be a random variable distributed according to $\mathbf{p}(x)$. Define random variables $Y = f(X) \in \{1, \ldots, m-1\}$ and $Z = g(X) \in \{0, 1\}$ according to

$$f(x) = \begin{cases} m-1 & \text{if } x = m, \\ x & \text{otherwise;} \end{cases} \qquad g(x) = \begin{cases} 1 & \text{if } x = m, \\ 0 & \text{otherwise.} \end{cases}$$

Then Y is distributed according to $\mathbf{q}(x)$. Since the map taking X to $(f(X), g(X))$ is invertible, we have $H(X) = H(Y, Z) = H(Y) + H(Z \mid Y)$. The latter term is computed as

$$\begin{aligned} H(Z \mid Y) &= \sum_{y \in \{1,\ldots,m-1\}} H(Z \mid Y = y)\mathbf{q}(y) \\ &= 0 + H(Z \mid Y = m-1)q_{m-1} \\ &= (p_{m-1} + p_m) H\left(\frac{p_{m-1}}{p_{m-1} + p_m}, \frac{p_m}{p_{m-1} + p_m}\right). \end{aligned}$$

Here we have used the fact that there is zero uncertainty in Z given $Y = y$, unless $Y = m - 1$.

3.4 *AEP.*

(a) The definition of A^n is exactly the definition of the typical set, thus by the AEP we have that $P(X^n \in A^n) \to 1$ as $n \to \infty$.

(b) First, by the weak law of large numbers, we have $P(X^n \in B^n) \to 1$, and from part (a) we have $P(X^n \in A^n) \to 1$. Thus, for any $\epsilon_1 > 0$, there exists an integer n_1 such that $P(X^n \in B^n) > 1 - \epsilon_1$ for all $n \geq n_1$. Similarly, for any $\epsilon_2 > 0$, there exists an integer n_2 such that $P(X^n \in A^n) > 1 - \epsilon_2$ for all $n \geq n_2$. Take any $n \geq \max(n_1, n_2)$. Then

$$\begin{aligned} P(X^n \in A^n \cap B^n) &= P(X^n \in A^n) + P(X^n \in B^n) - P(X^n \in A^n \cup B^n) \\ &> (1 - \epsilon_2) + (1 - \epsilon_1) - P(X^n \in A^n \cup B^n) \\ &\geq 1 - \epsilon_1 - \epsilon_2. \end{aligned}$$

Since ϵ_1 and ϵ_2 can be chosen arbitrarily, we see that, indeed, $P(X^n \in A^n \cap B^n) \to 1$ as $n \to \infty$.

(c)

$$\begin{aligned} 1 &\geq P(X^n \in A^n \cap B^n) \\ &= \sum_{x \in A^n \cap B^n} p(x) \\ &\geq \sum_{x \in A^n \cap B^n} 2^{-n(H+\epsilon)} \\ &= |A^n \cap B^n| 2^{-n(H+\epsilon)}. \end{aligned}$$

Therefore,

$$|A^n \cap B^n| \leq 2^{n(H+\epsilon)}.$$

(d) By part (c), $P(X^n \in A^n \cap B^n) \to 1$ for n sufficiently large, thus $P(X^n \in A^n \cap B^n) \geq \frac{1}{2}$ for n sufficiently large. For any such n, we then have

$$\begin{aligned}\frac{1}{2} &\leq \sum_{x \in A^n \cap B^n} p(x) \\ &\leq \sum_{x \in A^n \cap B^n} 2^{-n(H-\epsilon)} \\ &= |A^n \cap B^n| 2^{-n(H-\epsilon)}.\end{aligned}$$

Rearranging, we have $|A^n \cap B^n| \geq \left(\frac{1}{2}\right) 2^{n(H-\epsilon)}$.

3.5 *Sets defined by probabilities.*

(a)

$$\begin{aligned}1 &\geq P(X^n \in C_n(t)) \\ &= \sum_{x \in C_n(t)} p(x) \\ &\geq \sum_{x \in C_n(t)} 2^{-nt} \\ &= |C_n(t)| 2^{-nt}.\end{aligned}$$

Therefore, $|C_n(t)| \leq 2^{nt}$.

(b) First, if $t \geq H + \epsilon$ for $\epsilon > 0$, then $A_\epsilon^{(n)} \subseteq C_n(t)$, and thus

$$t \geq H + \epsilon$$

is a sufficient condition for $P(X^n \in C_n(t)) \to 1$.

Now, by our results on high probability sets, if $P(X^n \in C_n(t)) \to 1$, then

$$|C_n(t)| > 2^{n(H-\epsilon)}.$$

From part (a), we also have $|C_n(t)| \leq 2^{nt}$, therefore, we have that

$$t > H - \epsilon$$

is a necessary condition for $P(X^n \in C_n(t)) \to 1$.

The above analysis leaves open whether $t = H$ is sufficient; indeed, this will depend on the distribution of X. For example, if $P(X = 0) = P(X = 1) = \frac{1}{2}$, then $t = H$ is sufficient, since all sequences have probability 2^{-n}. However, if $P(X = 0) = p$ and $P(X = 1) = 1 - p$, with $0 < p < 1/2$, then $t = H$ is not sufficient. Roughly speaking, this follows from the fact that the shape of the distribution of $-\frac{1}{n} \log p(X_1, X_2, \ldots, X_n)$ converges to that of a Gaussian with mean $\mu = H(X)$ and variance σ, where $\sigma \to 0$ as $n \to \infty$. Thus, when $t = H$, the total probability of the sequences for which $p(x^n) \geq 2^{-nt}$ is essentially $\frac{1}{2}$, and thus $P(X^n \in C_n(t))$ does not converge to 1.

3.7 *AEP and source coding.*

(a) There are

$$\binom{100}{0} + \binom{100}{1} + \binom{100}{2} + \binom{100}{3} = 1 + 100 + 4950 + 161700 = 166751$$

binary sequences containing three or fewer ones. Since $2^{17} < 166751 < 2^{18}$, the fixed codeword length will need to be 18 bits. Thus we would achieve a rate of 0.18 bit/symbol. The source entropy is 0.0454 bit/symbol, so this coding scheme is operating at a rate that is approximately 400% of entropy (and so is not very efficient).

(b) The probability of observing a sequence containing three or fewer ones is, for $p = 0.005$,

$$P(\text{correct encoding}) = \sum_{i=0}^{3} \binom{100}{i} p^i (1-p)^{100-i} = 0.998327,$$

thus

$$P(\text{encoding failure}) = 1 - P(\text{correct encoding}) = 1.67327 \times 10^{-3}.$$

(c) Let W be the number of ones observed in the source sequence of $n = 100$ bits. Then $E(W) = 0.5$, and $\text{VAR}(W) = np(1-p) = 0.4975$. Chebysev's inequality gives

$$P(|W - E(W)| \geq b) \leq \frac{\text{VAR}(W)}{b^2}$$

Setting $b = 3.5$, we find

$$P(|W - 0.5| \geq 3.5) \leq \frac{0.4975}{3.5^2} = 0.0406122$$

This bound is very loose.

A better bound would be a Chernoff bound. Let X be a random variable. Apply, for any real number $s \neq 0$, Markov's inequality to the random variable e^{sX}. We then find, for any t,

$$P(e^{sX} \geq e^{st}) \leq \frac{E(e^{sX})}{e^{st}}.$$

For $s > 0$, this gives the Chernoff bound

$$P(X \geq t) \leq e^{-st} E(e^{sX}).$$

For the binomial random variable W of (b), we have, for $n = 100$ and $p = 0.005$,

$$E(e^{sW}) = ((1-p) + pe^s)^n,$$

thus, for any $s > 0$,

$$P(W \geq t) \leq e^{-st}((1-p) + pe^s)^n.$$

The right hand side is minimized by choosing $e^s = t(1-p)/(p(n-t))$, giving

$$P(W \geq t) \leq \left(\frac{np}{t}\right)^t \left(\frac{n-np}{n-t}\right)^{n-t}.$$

Substituting $n = 100$, $p = 0.005$ and $t = 4$, we get

$$P(W \geq 4) \leq 7.59657 \times 10^{-3},$$

which is tighter than the Chebyshev bound.

3.13 *Calculation of typical set.*

(a) $H(X) = H(0.6, 0.4) \approx 0.970951$.

(b) Let $p = P(X = 1)$. A particular sequence $(x_1, \ldots, x_n)$ containing exactly w ones occurs with probability $p^w(1-p)^{n-w}$. To qualify for membership in the typical set $A_\epsilon^{(n)}$, such a sequence must have probability in the interval $[2^{-n(H(X)+\epsilon)}, 2^{-n(H(X)-\epsilon)}]$. Noting that $2^{-nH(X)} = p^{np}(1-p)^{n(1-p)}$, we find that, for a sequence x containing exactly w ones,

$$\begin{aligned} x \in A_\epsilon^{(n)} &\text{ iff } p^w(1-p)^{n-w} \in [p^{np}(1-p)^{n(1-p)}2^{-n\epsilon}, p^{np}(1-p)^{n(1-p)}2^{n\epsilon}] \\ &\text{iff } p^{w-np}(1-p)^{n-w-n(1-p)} \in [2^{-n\epsilon}, 2^{n\epsilon}] \\ &\text{iff } \left(\frac{p}{1-p}\right)^{w-np} \in [2^{-n\epsilon}, 2^{n\epsilon}] \\ &\text{iff } (w-np)\log_2(p/(1-p)) \in [-n\epsilon, n\epsilon] \\ &\text{iff } w \in [np - n\epsilon/\eta, np + n\epsilon/\eta], \end{aligned}$$

where $\eta = |\log_2(p/(1-p))|$. For $n = 25$, $\epsilon = 0.1$, and $\eta = |\log_2(0.6/0.4)| = 0.585$, we have $np - n\epsilon/\eta = 10.726$ and $np + n\epsilon/\eta = 19.274$, thus the typical set contains those sequences with w ones, where $11 \leq w \leq 19$. The probability of the typical set is approximately 0.936246 and its cardinality is exactly 26 366 510.

(c) There are 20 457 889 elements in the smallest set that has probability 0.9: these are all the 16 777 216 sequences with $13 \leq w \leq 25$ ones and any 3 680 673 of the sequences with 12 ones.

(d)

$$|A_\epsilon^{(n)} \cap B_\delta^{(n)}| = 3680673 + \sum_{w=13}^{19} \binom{25}{w} = 20\,389\,483$$

Thus 99.6656% of the highest probability sequences also fall into the typical set.