

ECE 1502 — Information Theory
Problem Set 3 solutions
 October 17, 2007

2.14 *Entropy of a sum.*

(a) $Z = X + Y$. Hence $p(Z = z|X = x) = p(Y = z - x|X = x)$.

$$\begin{aligned}
 H(Z|X) &= \sum_x p(x) H(Z|X = x) \\
 &= - \sum_x p(x) \sum_z p(Z = z|X = x) \log p(Z = z|X = x) \\
 &= \sum_x p(x) \sum_y p(Y = z - x|X = x) \log p(Y = z - x|X = x) \\
 &= \sum_x p(x) H(Y|X = x) \\
 &= H(Y|X).
 \end{aligned}$$

If X and Y are independent, then $H(Y|X) = H(Y)$. Since $I(X; Z) \geq 0$, we have $H(Z) \geq H(Z|X) = H(Y|X) = H(Y)$. Similarly we can show that $H(Z) \geq H(X)$.

(b) Consider the following joint distribution for X and Y Let

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

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

(c) We have

$$H(Z) \leq H(X, Y) \leq H(X) + H(Y)$$

because Z is a function of (X, Y) and $H(X, Y) = H(X) + H(Y|X) \leq H(X) + H(Y)$. We have equality iff (X, Y) is a function of Z and $H(Y) = H(Y|X)$, i.e., X and Y are independent.

Bottleneck.

2.16 (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 $k = 1$, $I(X_1; X_3) \leq \log 1 = 0$ and since $I(X_1, X_3) \geq 0$, $I(X_1, X_3) = 0$. Thus, for $k = 1$, X_1 and X_3 are independent.

2.17 Pure randomness and bent coins.

$$\begin{aligned}
 nH(p) &\stackrel{(a)}{=} H(X_1, \dots, X_n) \\
 &\stackrel{(b)}{\geq} H(Z_1, Z_2, \dots, Z_K) \\
 &\stackrel{(c)}{=} H(Z_1, Z_2, \dots, Z_K, K) \\
 &\stackrel{(d)}{=} H(K) + H(Z_1, \dots, Z_K|K) \\
 &\stackrel{(e)}{=} H(K) + E(K) \\
 &\stackrel{(f)}{\geq} EK .
 \end{aligned}$$

- (a) Since X_1, X_2, \dots, X_n are i.i.d. with probability of $X_i = 1$ being p , the entropy $H(X_1, X_2, \dots, X_n)$ is $nH(p)$.
- (b) Z_1, \dots, Z_K is a function of X_1, X_2, \dots, X_n , and since the entropy of a function of a random variable is less than the entropy of the random variable, $H(Z_1, \dots, Z_K) \leq H(X_1, X_2, \dots, X_n)$.
- (c) K is a function of Z_1, Z_2, \dots, Z_K , so its conditional entropy given Z_1, Z_2, \dots, Z_K is 0. Hence $H(Z_1, Z_2, \dots, Z_K, K) = H(Z_1, \dots, Z_K) + H(K|Z_1, Z_2, \dots, Z_K) = H(Z_1, Z_2, \dots, Z_K)$.
- (d) Follows from the chain rule for entropy.
- (e) By assumption, Z_1, Z_2, \dots, Z_K are pure random bits (given K), with entropy 1 bit per symbol. Hence

$$H(Z_1, Z_2, \dots, Z_K|K) = \sum_k p(K = k)H(Z_1, Z_2, \dots, Z_k|K = k) \quad (1)$$

$$= \sum_k p(k)k \quad (2)$$

$$= EK. \quad (3)$$

- (f) Follows from the non-negativity of discrete entropy.
- (g) Since we do not know p , 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

$$\begin{aligned}
 0000 &\rightarrow \Lambda \\
 0001 &\rightarrow 00 & 0010 &\rightarrow 01 & 0100 &\rightarrow 10 & 1000 &\rightarrow 11 \\
 0011 &\rightarrow 00 & 0110 &\rightarrow 01 & 1100 &\rightarrow 10 & 1001 &\rightarrow 11 \\
 1010 &\rightarrow 0 & 0101 &\rightarrow 1 \\
 1110 &\rightarrow 11 & 1101 &\rightarrow 10 & 1011 &\rightarrow 01 & 0111 &\rightarrow 00 \\
 1111 &\rightarrow \Lambda
 \end{aligned} \quad (4)$$

The resulting expected number of bits is

$$EK = 4pq^3 \times 2 + 4p^2q^2 \times 2 + 2p^2q^2 \times 1 + 4p^3q \times 2 \quad (5)$$

$$= 8pq^3 + 10p^2q^2 + 8p^3q. \quad (6)$$

For example, for $p \approx \frac{1}{2}$, the expected number of pure random bits is close to 1.625. This is substantially less than the 4 pure random bits that could be generated if p were 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 $\log \binom{n}{k}$ 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 do is to 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 $2^l, 2^{l-1}, 2^{l-2}, \dots, 1$.

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 the average, the number of bits generated is

$$E[K|k \text{ 1's in sequence}] \geq \frac{1}{2}l + \frac{1}{4}(l-1) + \dots + \frac{1}{2^l}1 \quad (7)$$

$$= l - \frac{1}{4} \left(1 + \frac{1}{2} + \frac{2}{4} + \frac{3}{8} + \dots + \frac{l-1}{2^{l-2}} \right) \quad (8)$$

$$\geq l - 1, \quad (9)$$

since the infinite series sums to 1.

Hence 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

$$EK \geq \sum_{k=0}^n \binom{n}{k} p^k q^{n-k} \lfloor \log \binom{n}{k} - 1 \rfloor \quad (10)$$

$$\geq \sum_{k=0}^n \binom{n}{k} p^k q^{n-k} \left(\log \binom{n}{k} - 2 \right) \quad (11)$$

$$= \sum_{k=0}^n \binom{n}{k} p^k q^{n-k} \log \binom{n}{k} - 2 \quad (12)$$

$$\geq \sum_{n(p-\epsilon) \leq k \leq n(p+\epsilon)} \binom{n}{k} p^k q^{n-k} \log \binom{n}{k} - 2. \quad (13)$$

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

$$\binom{n}{k} \geq 2^{n(H(\frac{k}{n})-\delta)} \geq 2^{n(H(p)-2\delta)} \quad (14)$$

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 $EK \geq (1-\epsilon)n(H(p)-2\delta)-2$, 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.* By definition, $p_n = \Pr(X = n) = 1/An \log^2 n$ for $n \geq 2$. Therefore

$$\begin{aligned} H(X) &= - \sum_{n=2}^{\infty} p(n) \log p(n) \\ &= - \sum_{n=2}^{\infty} \left(1/An \log^2 n\right) \log \left(1/An \log^2 n\right) \\ &= \sum_{n=2}^{\infty} \frac{\log(An \log^2 n)}{An \log^2 n} \\ &= \sum_{n=2}^{\infty} \frac{\log A + \log n + 2 \log \log n}{An \log^2 n} \\ &= \log A + \sum_{n=2}^{\infty} \frac{1}{An \log n} + \sum_{n=2}^{\infty} \frac{2 \log \log n}{An \log^2 n}. \end{aligned}$$

The first term is finite. For base 2 logarithms, all the elements in the sum in the last term are nonnegative. (For any other base, the terms of the last sum eventually all become positive.) So all we have to do is bound the middle sum, which we do by comparing with an integral.

$$\sum_{n=2}^{\infty} \frac{1}{An \log n} > \int_2^{\infty} \frac{1}{Ax \log x} dx = K \ln \ln x \Big|_2^{\infty} = +\infty.$$

We conclude that $H(X) = +\infty$.

2.21 *Markov inequality applied to entropy.*

$$P(p(X) < d) \log \frac{1}{d} = \sum_{x:p(x)<d} p(x) \log \frac{1}{d} \quad (15)$$

$$\leq \sum_{x:p(x)<d} p(x) \log \frac{1}{p(x)} \quad (16)$$

$$\leq \sum_x p(x) \log \frac{1}{p(x)} \quad (17)$$

$$= H(X) \quad (18)$$

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,

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

Rearranging, we have

$$H(\mathbf{p}) = H(\mathbf{q}) + (p_{m-1} + p_m) H_2 \left(\frac{p_{m-1}}{p_{m-1} + p_m} \right).$$

3.1 *Markov's inequality and Chebyshev's inequality.*

(a) If X has distribution $F(x)$,

$$\begin{aligned} EX &= \int_0^\infty x dF \\ &= \int_0^\delta x dF + \int_\delta^\infty x dF \\ &\geq \int_\delta^\infty x dF \\ &\geq \int_\delta^\infty \delta dF \\ &= \delta \Pr\{X \geq \delta\}. \end{aligned}$$

Rearranging sides and dividing by δ we get,

$$\Pr\{X \geq \delta\} \leq \frac{EX}{\delta}. \tag{19}$$

One student gave a proof based on conditional expectations. It goes like

$$\begin{aligned} EX &= E(X|X \geq \delta) \Pr\{X \geq \delta\} + E(X|X < \delta) \Pr\{X < \delta\} \\ &\geq E(X|X \geq \delta) \Pr\{X \geq \delta\} \\ &\geq \delta \Pr\{X \geq \delta\}, \end{aligned}$$

which leads to (19) as well.

Given δ , the distribution achieving

$$\Pr\{X \geq \delta\} = \frac{EX}{\delta},$$

is

$$X = \begin{cases} \delta & \text{with probability } \frac{\mu}{\delta} \\ 0 & \text{with probability } 1 - \frac{\mu}{\delta}, \end{cases}$$

where $\mu \leq \delta$. In particular, the only nonnegative random variable X that satisfies Markov's inequality, with equality, for all $t > 0$, has pdf $f_X(x) = \delta(x)$.

(b) Letting $X = (Y - \mu)^2$ in Markov's inequality,

$$\begin{aligned} \Pr\{(Y - \mu)^2 > \epsilon^2\} &\leq \Pr\{(Y - \mu)^2 \geq \epsilon^2\} \\ &\leq \frac{E(Y - \mu)^2}{\epsilon^2} \\ &= \frac{\sigma^2}{\epsilon^2}, \end{aligned}$$

and noticing that $\Pr\{(Y - \mu)^2 > \epsilon^2\} = \Pr\{|Y - \mu| > \epsilon\}$, we get,

$$\Pr\{|Y - \mu| > \epsilon\} \leq \frac{\sigma^2}{\epsilon^2}.$$

(c) Letting Y in Chebyshev's inequality from part (b) equal \bar{Z}_n , and noticing that $E\bar{Z}_n = \mu$ and $\text{Var}(\bar{Z}_n) = \frac{\sigma^2}{n}$ (ie. \bar{Z}_n is the sum of n iid r.v.'s, $\frac{Z_i}{n}$, each with variance $\frac{\sigma^2}{n^2}$), we have,

$$\Pr\{|\bar{Z}_n - \mu| > \epsilon\} \leq \frac{\sigma^2}{n\epsilon^2}.$$

3.4 AEP.

(a) The definition of A^n is exactly the definition of the typical set, thus by Theorem 3.1.2 we have $\Pr\{X^n \in A^n\} \rightarrow 1$.

(b) First, by the weak law of large numbers, we have $\Pr\{X^n \in B^n\} \rightarrow 1$, and from part (a) we have $\Pr\{X^n \in A^n\} \rightarrow 1$. Thus, $\Pr\{X^n \in B^n\} > 1 - \epsilon_1$, and $\Pr\{X^n \in A^n\} > 1 - \epsilon$, for n sufficiently large.

Now,

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

Therefore, $\Pr\{X^n \in A^n \cap B^n\} \rightarrow 1$.

(c)

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

Therefore,

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

- (d) By part (c), $\Pr\{X^n \in A^n \cap B^n\} \rightarrow 1$ for n sufficiently large. Thus, clearly $\Pr\{X^n \in A^n \cap B^n\} \geq \frac{1}{2}$ for n sufficiently large. Then,

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

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)

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

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 $\Pr\{X^n \in C_n(t)\} \rightarrow 1$.

Now, by Theorem 3.3.1 (see also question 3.11), if $\Pr\{X^n \in C_n(t)\} \rightarrow 1$, 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 such that $\Pr\{X^n \in C_n(t)\} \rightarrow 1$.

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

3.13 *Calculation of typical set.*

(a) $H(X) = H_2(0.6) \approx 0.97095$

(b) For $n = 25$ and $\epsilon = 0.1$, the typical sequences are those for which

$$8.7155275 \times 10^{-9} \leq p(x_1, x_2, \dots, x_{25}) \leq 2.7889688 \times 10^{-7}.$$

The typical set are those sequences with $11 \leq k \leq 19$ ones.

The probability of the typical set is approximately 0.936246227, and the cardinality of the typical set is 26366510.

(c) There are 20457889 elements in the smallest set that has probability 0.9; all those sequences with $13 \leq k \leq 25$ ones and any 3680673 of the sequences with 12 ones.

(d)

$$|A_\epsilon^n \cap B_\delta^n| = 3680673 + \sum_{k=13}^{19} \binom{25}{k} = 20389483$$