Problem Set 1 Solution

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

2.1 *Coin flips.*

(a) The number X of tosses till the first head appears has the geometric distribution with parameter $p = 1/2$, where $P(X = n) = pq^{n-1}$, $n \in \{1, 2, \ldots\}$. Hence the entropy of X is

$$\begin{aligned} H(X) &= -\sum_{n=1}^{\infty} pq^{n-1} \log(pq^{n-1}) \\ &= -\left[\sum_{n=0}^{\infty} pq^n \log p + \sum_{n=0}^{\infty} npq^n \log q\right] \\ &= \frac{-p \log p}{1-q} - \frac{pq \log q}{p^2} \\ &= \frac{-p \log p - q \log q}{p} \\ &= H(p)/p \text{ bits.} \end{aligned}$$

If $p = 1/2$, then $H(X) = 2$ bits.

(b) Intuitively, it seems clear that the best questions are those that have equally likely chances of receiving a yes or a no answer. Consequently, one possible guess is that the most "efficient" series of questions is: Is $X = 1$? If not, is $X = 2$? If not, is $X = 3$? ...with a resulting expected number of questions equal to $\sum_{n=1}^{\infty} n(1/2^n) = 2$. This should reinforce the intuition that $H(X)$ is a measure of the uncertainty of X. Indeed in this case, the entropy is exactly the same as the average number of questions needed to define X, and in general $E(\#\text{ of questions}) \geq H(X)$. This problem has an interpretation as a source coding problem. Let $0 =$no, $1 =$yes, $X =$Source, and $Y =$Encoded Source. Then the set of questions in the above procedure can be written as a collection of (X, Y) pairs: $(1, 1)$, $(2, 01)$, $(3, 001)$, etc. In fact, this intuitively derived code is the optimal (Huffman) code minimizing the expected number of questions.

2.3 *Minimum entropy.* We wish to find *all* probability vectors $\mathbf{p} = (p_1, p_2, \ldots, p_n)$ which minimize

$$H(\mathbf{p}) = -\sum_i p_i \log p_i.$$

Now $-p_i \log p_i \geq 0$, with equality iff $p_i = 0$ or 1. Hence the only possible probability vectors which minimize $H(\mathbf{p})$ are those with $p_i = 1$ for some i and $p_j = 0, j \neq i$. There are n such vectors, i.e., $(1, 0, \ldots, 0)$, $(0, 1, 0, \ldots, 0)$, $\ldots$, $(0, \ldots, 0, 1)$, and the minimum value of $H(\mathbf{p})$ is 0.

2.4 *Entropy of functions of a random variable.*

(a) $H(X, g(X)) = H(X) + H(g(X) \mid X)$ by the chain rule for entropies.

(b) $H(g(X) \mid X) = 0$ since for any particular value of X, g(X) is fixed, and hence $H(g(X) \mid X) = \sum_x p(x) H(g(X) \mid X = x) = \sum_x 0 = 0$.

(c) $H(X, g(X)) = H(g(X)) + H(X \mid g(X))$ again by the chain rule.

(d) $H(X \mid g(X)) \geq 0$, with equality iff X is a function of $g(X)$, i.e., $g(.)$ is one-to-one. Hence $H(X, g(X)) \geq H(g(X))$.

Combining parts (b) and (d), we obtain $H(X) \geq H(g(X))$.

2.5 *Zero conditional entropy.* Assume that there exists an x, say x_0 and two different values of y, say y_1 and y_2 such that $p(x_0, y_1) > 0$ and $p(x_0, y_2) > 0$. Then $p(x_0) \geq p(x_0, y_1) + p(x_0, y_2) > 0$, and $p(y_1 \mid x_0)$ and $p(y_2 \mid x_0)$ are not equal to 0 or 1. Thus

$$\begin{aligned} H(Y \mid X) &= -\sum_x p(x) \sum_y p(y \mid x) \log p(y \mid x) \\ &\geq p(x_0)(-p(y_1 \mid x_0) \log p(y_1 \mid x_0) - p(y_2 \mid x_0) \log p(y_2 \mid x_0)) \\ &> 0, \end{aligned}$$

since $-t \log t \geq 0$ for $0 \leq t \leq 1$, and is strictly positive for t not equal to 0 or 1. Therefore the conditional entropy $H(Y \mid X)$ is 0 if and only if Y is a function of X.

2.8 *Drawing with and without replacement.* Intuitively, it is clear that if the balls are drawn with replacement, the number of possible choices for the i-th ball is larger, and therefore the conditional entropy is larger. But computing the conditional distributions is slightly involved. It is easier to compute the unconditional entropy.

- With replacement. In this case the conditional distribution of each draw is the same for every draw. Thus

$$X_i = \begin{cases} \text{red} & \text{with prob. } \frac{r}{r+w+b} \\ \text{white} & \text{with prob. } \frac{w}{r+w+b} \\ \text{black} & \text{with prob. } \frac{b}{r+w+b} \end{cases}$$

  and therefore

$$\begin{aligned} H(X_i \mid X_{i-1}, \ldots, X_1) &= H(X_i) \\ &= \log(r + w + b) - \frac{r}{r+w+b} \log r - \frac{w}{r+w+b} \log w - \frac{b}{r+w+b} \log b. \end{aligned}$$

- Without replacement. The unconditional probability of the i-th ball being red is still $r/(r+w+b)$, of being blue is $b/(r+w+b)$, etc. Thus the unconditional entropy $H(X_i)$ is still the same as with replacement. The conditional entropy $H(X_i \mid X_{i-1}, \ldots, X_1)$, however, is less than the unconditional entropy, and therefore the entropy of drawing without replacement is lower.

2.10 *Entropy of a disjoint mixture.*

(a) We can do this problem by writing down the definition of entropy and expanding the various terms. Instead, we will use the algebra of entropies for a simpler proof. Since X_1 and X_2 have disjoint support sets, we can write

$$X = \begin{cases} X_1 & \text{with probability } \alpha, \\ X_2 & \text{with probability } 1-\alpha. \end{cases}$$

Define a function of X,

$$\theta = f(X) = \begin{cases} 1 & \text{when } X = X_1, \\ 2 & \text{when } X = X_2. \end{cases}$$

Then we have

$$\begin{aligned} H(X) &= H(X, f(X)) = H(\theta) + H(X \mid \theta) \\ &= H(\theta) + p(\theta = 1)H(X \mid \theta = 1) + p(\theta = 2)H(X \mid \theta = 2) \\ &= H(\alpha) + \alpha H(X_1) + (1-\alpha)H(X_2) \end{aligned}$$

where $H(\alpha) = -\alpha \log \alpha - (1-\alpha)\log(1-\alpha)$.

(b) To maximize $H(X)$, we take the the derivative with respect to α and find the α that makes the derivative zero. We have

$$\frac{\mathrm{d}}{\mathrm{d}\alpha} H(X) = \log_2\left(\frac{1-\alpha}{\alpha}\right) + H(X_1) - H(X_2)$$

and the derivative is zero when

$$\alpha = \frac{2^{H(X_1)}}{2^{H(X_1)} + 2^{H(X_2)}}$$

Let $A_1 = 2^{H(X_1)}$ and let $A_2 = 2^{H(X_2)}$. Since $\alpha = A_1/(A_1 + A_2)$ maximizes $H(X)$, we have

$$\begin{aligned} H(X) &\le -\alpha \log_2\left(\frac{A_1}{A_1 + A_2}\right) - (1-\alpha)\log_2\left(\frac{A_2}{A_1 + A_2}\right) + \alpha \log_2 A_1 + (1-\alpha)\log_2 A_2 \\ &= \alpha \log_2(A_1 + A_2) + (1-\alpha)\log_2(A_1 + A_2) \\ &= \log_2(A_1 + A_2) \end{aligned}$$

Thus

$$2^{H(X)} \le 2^{H(X_1)} + 2^{H(X_2)}.$$

If we interpret 2^H as an effective alphabet size, we see that the effective alphabet size of a union of disjoint alphabets at most the sum of their effective alphabet sizes.

2.11 *Measure of correlation.* X_1 and X_2 are identically distributed and

$$\rho = 1 - \frac{H(X_2 \mid X_1)}{H(X_1)}$$

(a)

$$\begin{aligned}\rho &= \frac{H(X_1) - H(X_2 \mid X_1)}{H(X_1)} \\ &= \frac{H(X_2) - H(X_2 \mid X_1)}{H(X_1)} \text{ (since } H(X_1) = H(X_2)) \\ &= \frac{I(X_1; X_2)}{H(X_1)}.\end{aligned}$$

(b) Since $0 \leq H(X_2 \mid X_1) \leq H(X_2) = H(X_1)$, we have

$$0 \leq \frac{H(X_2 \mid X_1)}{H(X_1)} \leq 1$$

$$0 \leq \rho \leq 1.$$

(c) $\rho = 0$ iff $I(X_1; X_2) = 0$ iff X_1 and X_2 are independent.

(d) $\rho = 1$ iff $H(X_2 \mid X_1) = 0$ iff X_2 is a function of X_1. By symmetry, X_1 is a function of X_2, i.e., X_1 and X_2 have a one-to-one relationship.

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

(a) We give three proofs of Markov's Inequality.

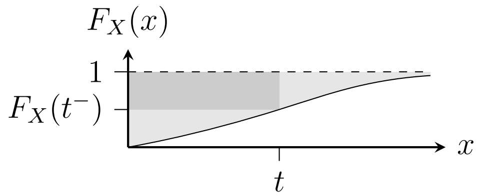
First, suppose X has probability density function $f_X(x)$. Then

$$\begin{aligned}E(X) &= \int_0^\infty x f_X(x) \mathrm{d}x \\ &= \int_0^t x f_X(x) \mathrm{d}x + \int_t^\infty x f_X(x) \mathrm{d}x \\ &\geq \int_t^\infty x f_X(x) \mathrm{d}x \\ &\geq \int_t^\infty t f_X(x) \mathrm{d}x \\ &= tP(X \geq t).\end{aligned}$$

Second, suppose X has cumulative distribution function $F_X(x) = P(X \leq x)$. Then, for any $t > 0$,

$$\begin{aligned}E(X) &= \int_0^\infty (1 - F_X(x)) \mathrm{d}x \\ &\geq t(1 - F_X(t^-)) \\ &= tP(X \geq t)\end{aligned}$$

This is illustrated in the figure below. The mean value $E(X)$ corresponds to the shaded area bounded between $F_X(x)$ and 1, while the quantity $t(1-F_X(t^-))$ corresponds to darker shaded rectangular area indicated. (Here $F_X(t^-) = \lim_{\epsilon \to 0^+} F_X(t-\epsilon)$).



To prove that $E(X) = \int_0^\infty (1 - F_X(x))\mathrm{d}x$, let $u(x) = x$ and let $v(x) = 1 - F_X(x)$. Suppose that $F_X(x)$ is differentiable with derivate $f_X(x)$. Then from $\int v\mathrm{d}u = uv - \int u\mathrm{d}v$ it follows that

$$\begin{aligned}\int_0^\infty (1 - F_X(x))\mathrm{d}x &= x(1 - F_X(x))|_0^\infty + \int_0^\infty x f_X(x)\mathrm{d}x \\ &= 0 + E(X).\end{aligned}$$

One may also reason as follows. Let $u(x)$ be the unit step function, given as

$$u(x) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases}$$

Then any non-negative real number x can be written as

$$x = \int_0^x \mathrm{d}s = \int_0^\infty (1 - u(s - x))\mathrm{d}s.$$

Substitute the random variable X for x and take the expected value of both sides. We then have

$$E(X) = E\left(\int_0^\infty (1 - u(s - X))\mathrm{d}s\right) = \int_0^\infty E(1 - u(s - X))\mathrm{d}s = \int_0^\infty (1 - P(X \leq s))\mathrm{d}s.$$

To change the order of expectation and integration in the second equality, we appealed to the Fubini-Tonelli theorem.

The third proof of Markov's Inequality is as follows. For any real number t and for any real number x, we have $x = xu(x - t) + x(1 - u(x - t))$. Substitute the random variable X for x and take the expectation of both sides. For $t > 0$ we get

$$\begin{aligned}E[X] &= E[Xu(X - t)] + E[X(1 - u(X - t))] \\ &\geq E[tu(X - t)] + 0 \\ &= tP[X \geq t]\end{aligned}$$

To achieve equality in Markov's inequality, we fix $t > 0$ and define a random variable X taking values in $\{0, t\}$. Let $p = P[X = t]$, so that $E[X] = pt$. Then $P[X \geq t] = p = E[X]/t$, i.e., Markov's inequality holds with equality.

(b) Chebyshev's Inequality follows immediately by applying Markov's Inequality to the random variable $X = (Y - \mu)^2$. Since X is non-negative, Markov's Inequality applies, and we have

$$P(X \geq \epsilon^2) = P((Y-\mu)^2 \geq \epsilon^2) = P(|Y-\mu| \geq \epsilon) \leq \frac{E(X)}{\epsilon^2} = \frac{\sigma^2}{\epsilon^2}.$$

(c) If $X_1, \ldots, X_n$ are random variables then the variance of their sum is given by

$$\mathrm{VAR}(X_1 + \cdots + X_n) = \sum_{i=1}^{n} \mathrm{VAR}(X_i) + \sum_{\substack{i,j \\ i \neq j}} \mathrm{COV}(X_i, X_j),$$

where $\mathrm{COV}(X_i, X_j) = E((X_i - E(X_i))(X_j - E(X_j)))$ denotes the covariance of X_i and X_j. When the X_i's are pairwise uncorrelated (for example, when they are pairwise independent), then $\mathrm{COV}(X_i, X_j) = 0$ for $i \neq j$, and we get

$$\mathrm{VAR}(X_1 + \cdots + X_n) = \sum_{i=1}^{n} \mathrm{VAR}(X_i).$$

For the problem at hand, since $Z_1, \ldots, Z_n$ are i.i.d. with mean μ and variance σ^2, we have

$$\mathrm{VAR}(Z_1 + \cdots + Z_n) = n\sigma^2.$$

Since $\mathrm{VAR}(\alpha X) = \alpha^2 \mathrm{VAR}(X)$ for any real number α, we get

$$\mathrm{VAR}(\bar{Z}_n) = \mathrm{VAR}\left(\frac{1}{n}(Z_1 + \cdots + Z_n)\right) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

It is easy to see that $E(\bar{Z}_n) = \mu$, thus Chebyshev's Inequality gives

$$P[|\bar{Z}_n - \mu| \geq \epsilon] \leq \frac{\sigma^2}{n\epsilon^2}.$$