Solutions 1

Exercise 1.1. See Examples 1.2 and 1.11 in the course notes.

Exercise 1.2. Observe that the Hamming distance of two vectors is the minimum number of bit flips required to transform one into the other. Using this, the first three conditions are trivial to verify. As for the triange inequality

$$d(x,z) \leq d(x,y) + d(y,z), \tag{1}$$

consider each position i of the vectors x, y and z. If $x_i = z_i$, the corresponding position contributes 0 to the left-hand-side of equation (1). In this case, either $y_i = x_i = z_i$, thus contributing 0 to the right-hand-side as well, or $y_i \neq x_i, z_i$, thus contributing 2 to the right-hand-side. If $x_i \neq z_i$, so that the corresponding i contributes 1 to the left-hand-side of equation (1), then y_i must be different from at least one of x_i and z_i, thus contributing at least 1 to the right-hand-side. Summing over all values of i we readily obtain the triangle inequality.

Exercise 1.3. This is very similar to the case of $\mathrm{BSC}(\varepsilon)$ considered in the courses notes. For a received vector $y \in \Sigma^n$ and any codeword z, we have

$$p(y|z) = \prod_{i=1}^{n} p(y_i|z_i).$$

From the definition of our channel, $p(y_i|z_i) = \varepsilon/(q-1)$ for $y_i \neq z_i$ (this is the case for $d(y,z)$ coordinates) and $p(y_i|z_i) = 1-\varepsilon$ for $y_i = z_i$ (this is the case for $n-d(y,z)$ coordinates). Therefore

$$p(y|z) = \left(\frac{\varepsilon}{q-1}\right)^{d(y,z)} (1-\varepsilon)^{n-d(y,z)} = (1-\varepsilon)^n \left(\frac{\varepsilon/(q-1)}{1-\varepsilon}\right)^{d(y,z)}.$$

Since $\varepsilon \leq (q-1)/q$, the ratio $\frac{\varepsilon/(q-1)}{1-\varepsilon} \leq 1$, so that the codeword z that maximizes $p(y|z)$ is the one that minimizes $d(y,z)$.

Exercise 1.4.

1. Let $A(n) := H\left(\frac{1}{n}, \ldots, \frac{1}{n}\right)$. We first show that

$$A(s^m) = mA(s). \tag{2}$$

To see this, note that $A(s^m) = H\left(\frac{1}{s^m}, \ldots, \frac{1}{s^m}\right)$ corresponds to a choice between s^m equally likely events. We can group each s of these events together using Axiom 3. For example, grouping the first s events gives us

$$A(s^m) = H\left(\frac{1}{s^{m-1}}, \frac{1}{s^m}, \ldots, \frac{1}{s^m}\right) + \frac{1}{s^{m-1}} A(s).$$

Similarly grouping all the other events s by s, we obtain

$$A(s^m) = A(s^{m-1}) + A(s).$$

We can now repeat this procedure recursively to obtain

$$\begin{aligned} A(s^m) &= A(s^{m-1}) + A(s) \\ &= A(s^{m-2}) + 2A(s) \\ &= \cdots \\ &= mA(s). \end{aligned}$$

Now for s and t integers, and for n arbitrarily large, we can always find m such that

$$s^m \leq t^n < s^{m+1}. \tag{3}$$

On one hand, this gives us

$$\frac{m}{n} \leq \frac{\log t}{\log s} \leq \frac{m}{n} + \frac{1}{n}. \tag{4}$$

On the other hand, from Axiom 2, we know that A is a monotonic increasing function of its argument, so that equation (3) gives us

$$A(s^m) \leq A(t^n) < A(s^{m+1}).$$

From equation (2), this is equivalent to saying that

$$mA(s) \leq nA(t) < (m+1)A(s),$$

which gives us

$$\frac{m}{n} \leq \frac{A(t)}{A(s)} < \frac{m}{n} + \frac{1}{n}. \tag{5}$$

As we let n grow to infinity, equations (4) and (5) gives us that

$$\lim_{n\to\infty} \frac{A(t)}{A(s)} = \frac{\log t}{\log s},$$

so that $A(t)$ must be of the form

$$A(t) = K \log t$$

for a constant K, where K must be positive to satisfy Axiom 2.

2. Suppose the p_i are commensurable probabilities, so that $p_i = \frac{n_i}{\sum n_i}$. Consider choosing an event from $\sum n_i$ equiprobable events. From the expression we derived above for $A(n)$, we know that the entropy of this choice is $K \log \sum n_i$. But using Axiom 3, we can also view this choice in the following equivalent manner: we can break down a choice from $\sum n_j$ equiprobable events into a choice from n events with probabilities $p_1, \ldots, p_n$, then if the ith event is chosen, we have a second choice between n_i equiprobable events. The entropy of this event is

$$H(p_1, \ldots, p_n) + \sum p_i K \log n_i.$$

We thus obtain

$$\begin{aligned} H(p_1,\ldots,p_n) &= K\left(\log\sum n_i - \sum p_i \log n_i\right) \\ &= K\left(\sum p_i \log\sum n_i - \sum p_i \log n_i\right) \\ &= -K\sum p_i \log\frac{n_i}{\sum n_i} \\ &= -K\sum p_i \log p_i. \end{aligned}$$

3. Now suppose the p_i are incommensurable. Since the rationals are dense in the reals, we can approximate the p_i with rational numbers. We can thus find rationals $\tilde{p}_1,\ldots,\tilde{p}_{n-1}$ such that $|p_i - \tilde{p}_i| < \varepsilon$ for any $\varepsilon > 0$. Define $\tilde{p}_n$ as $1 - \sum_{i=1}^{n-1}\tilde{p}_i$. This ensures that $(\tilde{p}_1,\ldots,\tilde{p}_n)$ is indeed a probability distribution, and $|p_n - \tilde{p}_n| < (n-1)\epsilon$ can be made as small as we want. By continuity of H (Axiom 1), $H(p_1,\ldots,p_n)$ tends to $H(\tilde{p}_1,\ldots,\tilde{p}_n) = -K\sum\tilde{p}_i\log\tilde{p}_i$. But by continuity of the function $f(x_1,\ldots,x_n) = -K\sum x_i\log x_i$ (defined over real probability vectors $(x_1,\ldots,x_n)$), $-K\sum\tilde{p}_i\log\tilde{p}_i$ tends to $-K\sum p_i\log p_i$. Thus the expression holds in general.

Exercise 1.5.

1. •

$$\begin{aligned} I(X;Y) &= \sum_{x,y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \\ &= \sum_{x,y} p(x,y)\log\frac{p(x|y)}{p(x)} \\ &= -\sum_{x,y} p(x,y)\log p(x) + \sum_{x,y} p(x,y)\log p(x|y) \\ &= -\sum_{x} p(x)\log p(x) + \sum_{x,y} p(x,y)\log p(x|y) \\ &= H(X) - H(X|Y). \end{aligned}$$

We prove similarly that

$$I(X;Y) = H(Y) - H(Y|X).$$

•

$$\begin{aligned} I(X;Y) &= \sum_{x,y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \\ &= -\sum_{x,y} p(x,y)\log p(x) - \sum_{x,y} p(x,y)\log p(y) + \sum_{x,y} p(x,y)\log p(x|y) \\ &= -\sum_{x} p(x)\log p(x) - \sum_{y} p(y)\log p(y) + \sum_{x,y} p(x,y)\log p(x|y) \\ &= H(X) + H(Y) - H(X,Y). \end{aligned}$$

We can clearly see that $I(X;Y)$ is symmetric in its arguments.

- $$\begin{aligned} I(X;X) &= \sum_x p(x,x) \log \frac{p(x,x)}{p(x)p(x)} \\ &= \sum_x p(x) \log \frac{1}{p(x)} \\ &= H(X). \end{aligned}$$

  We could also obtain this formula by noting that $I(X;X) = H(X) - H(X|X) = H(X)$.

2. Using the chain rule for two variables, we have

$$\begin{aligned} H(X_1, X_2) &= H(X_1) + H(X_2|X_1) \\ H(X_1, X_2, X_3) &= H(X_1) + H(X_2, X_3|X_1) \\ &= H(X_1) + H(X_2|X_1) + H(X_3|X_2, X_1) \\ &\vdots \\ H(X_1, \ldots, X_n) &= H(X_1) + H(X_2|X_1) + \cdots + H(X_n|X_{n-1}, \ldots, X_1) \\ &= \sum_{i=1}^{n} H(X_i|X_{i-1}, \ldots, X_1). \end{aligned}$$

3. To prove the chain rule for relative entropy, note that

$$\begin{aligned} D(p(x,y)||q(x,y)) &= \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{q(x,y)} \\ &= \sum_x \sum_y p(x,y) \log \frac{p(x)p(y|x)}{q(x)q(y|x)} \\ &= \sum_x \sum_y p(x,y) \log \frac{p(x)}{q(x)} + \sum_x \sum_y p(x,y) \log \frac{p(y|x)}{q(y|x)} \\ &= D(p(x)||q(x)) + D(p(y|x)||q(y|x)). \end{aligned}$$

Exercise 1.6.

1. Let χ be the support set of the random variable X and let $A = \{x : p(x) > 0\}$ be the

support set of the probability distribution $p(x)$. We have

$$\begin{aligned} -D(p||q) &= -\sum_{x\in A} p(x)\log\frac{p(x)}{q(x)} \\ &= \sum_{x\in A} p(x)\log\frac{q(x)}{p(x)} \\ &\leq \log\sum_{x\in A} p(x)\frac{q(x)}{p(x)} \\ &= \log\sum_{x\in A} q(x) \\ &\leq \log\sum_{x\in\chi} q(x) \\ &= \log 1 = 0, \end{aligned}$$

with equality if and only if $q(x)/p(x) = 1$ everywhere, since $\log t$ is a strictly concave function of t. Therefore

$$D(p||q) \geq 0 \tag{6}$$

with equality if and only if $p(x) = q(x)$ for all x.

For any pair X, Y of random variables, $I(X;Y) = D(p(x,y)||p(x)p(y))$. Equation (6) gives us

$$I(X;Y) \geq 0, \tag{7}$$

with equality if and only if $p(x,y) = p(x)p(y)$ for all values x, y, that is, if and only if X and Y are independent.

2. Let X take values over χ with some probability distribution p, and let u be the uniform distribution over χ, so that $u(x) = \frac{1}{|\chi|}$ for all x. Consider the quantity

$$D(p||u) = \sum_{\chi} p(x)\log\frac{p(x)}{u(x)} = \sum_{\chi} p(x)\log p(x) - \sum_{\chi} p(x)\log u(x) = \log|\chi| - H(X).$$

From equation (6), we have that $H(X) \leq \log\chi|$, with equality if p and u are the same distribution.

3. From equation (7), we have

$$I(X;Y) = H(X) - H(X|Y) \geq 0,$$

so that

$$H(X|Y) \leq H(X),$$

with equality if and only if $I(X;Y) = 0$, i.e., if and only if X and Y are independent. Thus conditioning reduces entropy.
In the previous exercise, we saw the chain rule for the entropy of n variables:

$$H(X_1,\ldots,X_n) = \sum_{i=1}^{n} H(X_i|X_{i-1},\ldots,X_1).$$

Each conditional entropy term $H(X_i|X_{i-1},\ldots,X_1)$ is such that

$$H(X_i|X_{i-1},\ldots,X_1) \leq H(X_i),$$

with equality if and only if X_i is independent from the $(i-1)$-tuple $X_1,\ldots,X_{i-1}$. We finally get

$$H(X_1,\ldots,X_n) \leq \sum_{i=1}^{n} H(X_i),$$

with equality if and only if the X_i are independent.