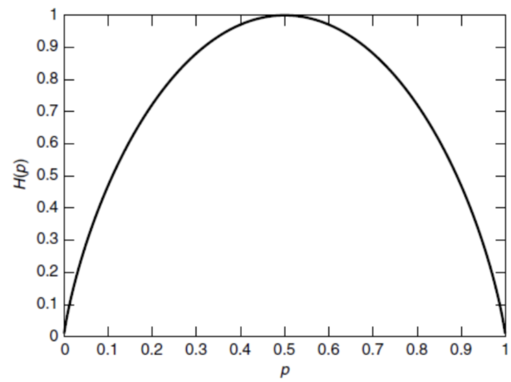
Solution 5 - Basics of Information Theory

1.



$$H(X) = -p \log p - (1-p) \log(1-p) \stackrel{\text{def}}{=} H(p)$$

In particular, $H(X) = 1$ bit when $p = \frac{1}{2}$. The graph of the function $H(p)$ is shown in the figure. The figure illustrates some of the basic properties of entropy: It is a concave function of the distribution and equals 0 when $p = 0$ or 1. This makes sense, because when $p = 0$ or 1, the variable is not random and there is no uncertainty. Similarly, the uncertainty is maximum when $p = \frac{1}{2}$, which also corresponds to the maximum value of the entropy.

2.

$$\begin{aligned}
H(X,Y) &\triangleq -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x,y) \\
&= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x) - \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(y|x) \\
&= -\sum_{x\in\mathcal{X}} p(x)\log p(x) - \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(y|x) \quad \text{(by marginalizing out } y) \\
&= H(X) + H(Y|X)
\end{aligned}$$

3. The marginal distribution of X is $(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8})$ and the marginal distribution of Y is $(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$, and hence

$$H(X) = -\sum_{i=1}^{4}[p(X=i)\log p(X=i)] = \frac{7}{4} \text{ bits}$$

$$H(Y) = -\sum_{i=1}^{4}[p(Y=i)\log p(Y=i)] = 2 \text{ bits}$$

$$
\begin{aligned}
H(X|Y) &\triangleq \sum_{y\in\mathcal{Y}} p(y)H(X|Y=y) \\
&= \sum_{i=1}^{4} p(Y=i)H(X|Y=i) \\
&= \frac{1}{4}H(\frac{1}{2},\frac{1}{4},\frac{1}{8},\frac{1}{8}) + \frac{1}{4}H(\frac{1}{4},\frac{1}{2},\frac{1}{8},\frac{1}{8}) + \frac{1}{4}H(\frac{1}{4},\frac{1}{4},\frac{1}{4},\frac{1}{4}) + \frac{1}{4}H(1,0,0,0) \\
&= \frac{1}{4}\times\frac{7}{4} + \frac{1}{4}\times\frac{7}{4} + \frac{1}{4}\times 2 + \frac{1}{4}\times 0 \\
&= \frac{11}{8} \text{ bits}
\end{aligned}
$$

Similarly, $H(Y|X) = \frac{13}{8}$ bits and $H(X,Y) = \frac{27}{8}$ bits.

4. Labeling the points from left-to-right as A, B, and C,

   A Leaving A out, the training will fit a line exactly from B to C: $f(x) = 2x - 1$. $f(-2) = -5$ compared with the training point A $y = 1$ yields an error estimate of $(-5-1)^2 = 36$.

   B Leaving B out, the training will fit a line exactly from A to C: $f(x) = 0.5x + 2$. $f(1) = 2.5$ compared with the training point B $y = 1$ yields an error estimate of $(2.5-1)^2 = 2.25$.

   C Leaving C out, the training will fit a line exactly from A to B: $f(x) = 0x + 1$. $f(2) = 1$ compared with the training point B $y = 3$ yields an error estimate of $(3-1)^2 = 4$.

   The average is $(36 + 2.25 + 4)/3 \approx 14.08$.