Problem 4.

As shown in the notes, the entropy for a node y having N_y instances is defined as:

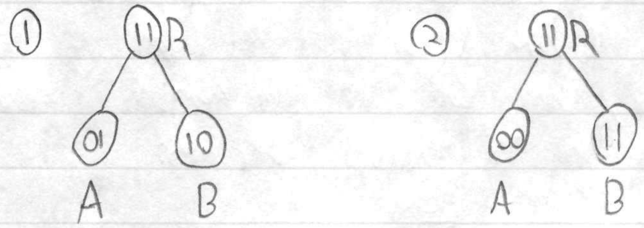
$$H(y) = -\sum_{i=1}^{K} P_{yk} \log_2 P_{yk} \quad , \quad K = \# \text{ classes}$$

The information gain, or reduction in entropy, is then defined as:

$$IG(y,V) = H(y) - \sum_{i=1}^{|V|} \frac{N_i}{N_y} H(i)$$, where feature V has $|V|$ distinct values, resulting in a $|V|$-way split

a) To show that the decrease in entropy by a split on a binary feature $|V|=2$, can never be greater than 1 bit, we will use a simple example: Let's say we have only 2 datapoints, and the labels for these datapoints are [0,1]. The label is binary (0,1).

Our "feature" is as follows: if the element is present in a given node, its label is 1, and 0 otherwise. It is clear that only two cases exist for root node R with child nodes A and B.



① $H(R) = -\left(\left(\frac{1}{2}\log_2\frac{1}{2}\right) + \left(\frac{1}{2}\log_2\frac{1}{2}\right)\right) = -\left(\left(-\frac{1}{2}\right) + \left(-\frac{1}{2}\right)\right) = 1$

$H(A) = -\left(\left(1\log_2 1\right)\right) + \left(1\log_2 1\right)\right) = -(0+0) = 0$

$H(B) = -\left(\left(0\log_2 0\right)\right) + \left(0\log_2 0\right)\right) = 0$

$IG = H(R) - P(A)H(A) - P(B)H(B) = 1 - 0 - 0 = 1$

② $H(R) = 1$ (same as before)

$$H(A) = -\left(\left(\tfrac{1}{2}\log_2 \tfrac{1}{2}\right) + \left(\tfrac{1}{2}\log_2 \tfrac{1}{2}\right)\right) = -\left(\left(-\tfrac{1}{2}\right) + \left(-\tfrac{1}{2}\right)\right) = 1$$

$H(B) = 1$ (same as $H(A)$)

$$IG = H(R) - P(A)H(A) - P(B)H(B) = 1 - \tfrac{1}{2}(1) - \tfrac{1}{2}(1) = 0$$

As we can see, these are the two most extreme cases. Case 1, in which the labelling goes from perfectly random to 100% correct, and case 2, in which there is no change in classification accomplished by the split.

Thus, the information gain is bounded by $0 \le IG \le 1$ bit

b) Now, let us generalize to the case of arbitrary branching $B > 1$.

We have already established that a perfect classification in all child nodes will yield a weighted entropy of 0. We have also established that the greatest reduction in entropy will occur when we go from a perfectly random guess of the label to a perfect classification. To do this, we will set the number of label classes $K = B$. Therefore, our equation for entropy becomes:

$$H(y) = -\sum_{i=1}^{B} P(B)\log_2 P(B) = -\sum_{i=1}^{B} \frac{1}{B}\log_2 \frac{1}{B}$$

Since we are randomly guessing the label, $P(B) = \frac{1}{B}$

$$= -B \cdot \frac{1}{B}\log_2 \frac{1}{B} = -\log_2 \frac{1}{B} = \log_2 \left(\frac{1}{B}\right)^{-1} = \log_2 B$$

$IG = \log_2 B - 0 = \log_2 B$ ... Finally: $0 \le IG \le \log_2 B$ bits