All students should do the first two problems. CS 2110 students, and CS 1511 students shooting for a grade of A, should additionally do the third and fourth problems. In any case writeup your solutions up using LaTeX and turn in the pdf in gradescope.

The first two homework problems riff off the wikipedia article on conditional entropy, `https://en.wikipedia.org/wiki/Conditional_entropy`, and the wikipedia article on Shannon's noisy channel theorem: `https://en.wikipedia.org/wiki/Noisy-channel_coding_theorem`. You should read the article on conditional entropy, stopping when it starts on conditional differential entropy. You should read the article on noisy channel theorem up through the statement of the theorem, you don't need to read the proof.

1. (2 points) Assume x is a bit that a sender wants to send to a receiver over a noisy channel, and let y be the bit received by the receiver. Because the channel is noisy, y may not equal x.Assume the noisy channel has the following properties $\mathsf{P}(y = 0 \mid x = 0) = .9$, $\mathsf{P}(y = 1 \mid x = 0) = .1$, $\mathsf{P}(y = 0 \mid x = 1) = .2$, and $\mathsf{P}(y = 1 \mid x = 1) = .8$. Let Y be the probability distribution for the received bit.

   (a) Let X be the probability distribution where x is 0 with probability 1/3 and x is 1 with probability 2/3.

      i. What is H(X) for this example?

      ii. What is the probability distribution Y? That is, what is the probability that y=0 and what is the probability that y=1?

      iii. What is H(Y) for this example?

      iv. What is $H(X \mid Y)$ for this example?

      v. What is $H(Y \mid X)$ for this example?

      vi. What is $I(X; Y)$ for this example?

   (a) Let X be the probability distribution where x is 0 with probability p and x is 1 with probability $1 - p$. Assume the noisy channel has the following properties $\mathsf{P}(y = 0 \mid x = 0) = q$, $\mathsf{P}(y = 1 \mid x = 0) = 1 - q$, $\mathsf{P}(y = 0 \mid x = 1) = 1 - q$, and $\mathsf{P}(y = 1 \mid x = 1) = q$. So the channel flips the sent bit with probability $1 - q$.

      Compute the following as a function of p and q.

      i. What is H(X) ?

      ii. What is the probability distribution Y? That is, what is the probability that y=0 and what is the probability that y=1?

      iii. What is H(Y) ?

      iv. What is $H(X \mid Y)$?

      v. What is $H(Y \mid X)$?

      vi. What is $I(X; Y)$ for this example?

   (b) According to Shannon's noisy channel theorem, what is capacity of this channel as a function of q?

      Hint: You want to take the maximum value of $I(X; Y)$ over all possible choices of p.

2. (4 points) The following properties are stated without proof in the wikipedia article on conditional entropy, `https://en.wikipedia.org/wiki/Conditional_entropy`. You should prove them from the definitions and prior properties.

   (a) $H(Y \mid X) \leq H(Y)$

   (b) $H(X,Y) \leq H(X \mid Y) + H(Y \mid X) + I(X;Y)$

   (c) $H(X,Y) \leq H(X) + H(Y) - I(X;Y)$

   (d) $I(X;Y) \leq H(X)$

   Hint: If you are unsure how to get started, look at the proof of the chain rule on the wikipedia page for inspiration. The proofs of these properties should be similar, a few lines long, using the entropic definitions and basic facts about probability.

3. (6 points) Let X be a random variable where the cardinality of its support is at most k. So in other words, X can only take on at most k distinct values.

   - Give an example of a distribution for X for which it is the case that $H(X) = \lg k$, where here lg is the logarithm base 2.
   - Prove that it must be the case that $H(X) \leq \lg k$.

4. (8 points) This homework deals with some aspects raised on the wikipedia page for Shannon-Fano coding `https://en.wikipedia.org/wiki/Shannon-Fano_coding`. See also the wikipedia page on Shannon's source coding theorem `https://en.wikipedia.org/wiki/Shannon's_source_coding_theorem`.

   (a) The article discusses two versions of Shannon's algorithm, and notes that they may produce different encodings for the same input. Show however that the expected word length will always be the same for both methods. Assume that the probabilities are distinct.

   (b) Prove or disprove that both methods (Shannon's algorithm and Fano's code: binary splitting) for implementing Shannon's encoding produce codewords of the same length for each letter/symbol in the source alphabet. Assume that the probabilities are distinct.

   (c) When all probabilities are of the form 2^{-i} for some integer $i \geq 1$, prove or disprove that Shannon's algorithm produces codes whose expected length in bits is equal the entropy.

   (d) When all probabilities are of the form 2^{-i} for some integer $i \geq 1$, prove or disprove that Fano's code: binary splitting produces codes whose expected length in bits is equal the entropy.

   (e) When all probabilities are of the form 2^{-i} for some integer $i \geq 1$, prove or disprove that Huffman's encoding produces codes whose expected length in bits is equal the entropy.

(f) The wikipedia page notes that in Shannon's encoding the expected length of a code word is at most one more than the entropy. This is proved by noting that length of the code word for a symbol with probability p_i is at most $1 - \lg p_i$. The wikipedia page notes that in Fano's binary splitting encoding the expected length of the code word is also at most one more than the entropy. The goal here is to determined whether this can be proved using the same method as was used to prove this for Shannons encoding. That is, prove or disprove that Fano's binary splitting encoding also has the property that the length of the code word for a symbol with probability p_i is at most $1 - \lg p_i$.