Assignment 6

1. Suppose we have two binary random variables $x_1, x_2 \in \{0, 1\}$. Their probability distribution is given by $P(x_1 = 0, x_2 = 0) = 0$, $P(x_1 = 1, x_2 = 0) = 1/4$, $P(x_1 = 0, x_2 = 1) = 1/2$, and $P(x_1 = 1, x_2 = 1) = 1/4$.

   a) What is the entropy of this probability distribution?

   b) Find the maximum-entropy distribution over x_1, x_2 consistent with first-order constraints (that is, in which $E[x_1]$ and $E[x_2]$ are the same as for the true distribution). What is the entropy of this distribution?

2. Download the file `google-10000-english.txt` linked on the course website. This is a list of the 10,000 most common English words, according to Google. The words are all lower case, so, including spaces, there are 27 possible characters. We will examine the probability distribution $P(c)$ over these characters, where $c = \{$‘ ’,‘a’,‘b’, …$\}$.

   a) What is the maximum entropy of a probability distribution over 27 characters?

   b) What is the actual entropy of the distribution over these characters using the 10,000 word dataset?

   c) What is the conditional entropy $P(c_i|c_{i-1} = C)$, where c_i is the ith character and c_{i-1} is the character preceding it, for $C = \{$‘ ’,‘a’,‘b’, …$\}$? For what characters is this conditional entropy minimal or maximal?

   If using python, you may find the following code useful:

```
x = np.array(list(open("google-10000-english.txt").read()))
x = x[:-1]
chars,inds,counts = np.unique(x,return_inverse=True,return_counts=True)
```

   This will yield 1) `chars`, containing the 27 unique characters, 2) `inds`, a vector whose ith element is an integer from 0 to 26 corresponding to the ith character in the file, and 3) `counts`, the frequencies of each character.