INFORMATION THEORY
University of Amsterdam, 2018
Team Name: Transfer

Homework set 4

Contents

Question 1

(a)

- $h(0.4) \approx 0.971$. So the typical set is bounded by $0.971 - 0.1 < ... < 0.971 + 0.1$. Using the table we can see that all sequences with $11 \leq k \leq 19$ are part of the typical set.
- The probability of the typical set is equal to sum of all the probabilities of the sequences inside the typical set, which is 0.936246.
- We can count the elements of the typical set by taking the sum of the elements per sequence inside the typical set, which gives us: $26, 366, 510$.

(b)

"The smallest set is the smallest set of sequences of length n which has total probability greater that $1 - \delta$."

We need to use high probability sequences, which are sequences with many ones, since $p(1) = 0.6$. Therefore we start at the bottom for $k = 25$ and go upwards until the probability of the consecutive sequences add up to 0.9.

We do so by selecting the sequences for $13 \leq k \leq 25$, which gives us a total probability of 0.846232. Thus we still miss $0.9 - 0.846232 = 0.053768$ probability, which has to be taken from $k = 12$. However, the probability of $k = 12$ is 0.075967, so we only need $0.053678/0.075967 = 0.7077810$ times the number of the sequences of $k = 12$, which have a total number of $5200300 * 0.7077810 \approx 3,680,673$.

This give us $16,777,216 + 3,680,673 = 20,457,889$ elements in the smallest set.

(c)

The intersection of the typical set and the smallest set consist of all the sequences with $13 \leq k \leq 19$, and a fraction of the sequences with $k = 12$. Thus there are $16,708,810 + 3,680,673 = 20,389,483$ elements in the intersection. This intersection has a probability of $0.816780 + 0.053768 = 0.870548$.

Question 2

We give the three given properties numbers for reference:
1. $I(A;B) = 0$.
2. $I(A;C|B) = I(A;B|C)$.
3. $H(A|BC) = 0$.

Proof. We have

$$\begin{aligned} H(C) &= H(C|AB) + I(B;C) + I(A;C|B) \\ &\geq I(A;C|B) \\ &= 0 + 0 + I(A;C|B) \\ &= H(A|BC) + I(A;B) + I(A;C|B) \\ &= H(A) \end{aligned}$$

where we used an entropy diagram in the first step, the non-negativity of mutual information and conditional entropy in the second step, properties 1 and 3 in the fourth step and again an entropy diagram in the last step. All in all, we have shown $H(C) \geq H(A)$, even without using property 2. □

Since we have shown the weak version, we need to provide an example showing that the inequality need not be strict. Therefore, let $\mathcal{X} = \{x\}$ be a one-point-space, and $A = B = C = \text{const}_x$ all be random variables that are constant x. From Practice Problem set 2, Exercise 1 (a), we know that the entropies are all zero: $H(A) = H(B) = H(C) = 0$. So clearly, the inequality is not strict here. Furthermore, all three properties 1, 2 and 3 trivially hold, since all the conditional entropies and mutual informations are bounded from above by some unconditional entropy, which is in our case always 0.

Question 3

We have

$$T_n \approx c^n$$
$$(T_n)^{\frac{1}{n}} \approx c$$
$$\frac{1}{n}\log T_n \approx \log c$$

We focus now on the left hand side

$$\frac{1}{n}\log T_n = \frac{1}{n}\log(C_1 C_2 .. C_n) = \frac{1}{n}\sum_{i=1}^{n}\log C_i$$

We can consider the random variables $X_i := \log C_i$. We have $\mu_i = E[X_i] = E[\log C_i]$
It is easy to see now that we can apply the weak law of the large numbers

$$\frac{1}{n}\sum_{i=1}^{n}\log C_i = \frac{1}{n}\sum_{i=1}^{n} X_i \to^p \mu_i$$

The value of μ_i is

$$\mu_i = E[X_i] = E[\log C_i] = \frac{3}{4}\log\frac{2}{3} + \frac{1}{4}\log\frac{3}{5} \tag{1}$$
$$= \log\left((\frac{2}{3})^{\frac{3}{4}}(\frac{3}{5})^{\frac{1}{4}}\right) \tag{2}$$
$$= \log\left((\frac{8}{27})^{\frac{1}{4}}(\frac{3}{5})^{\frac{1}{4}}\right) \tag{3}$$
$$= \log\left((\frac{8}{45})^{\frac{1}{4}}\right) \tag{4}$$

Which means, thanks to the first set of approximations, that $c \approx (\frac{8}{45})^{\frac{1}{4}}$

Question 4

(a)

Let $s = 26$ be the size of the alphabet. Then, since they are both uniformly chosen at random, $H(M) = \log s^m = m \log s$ and $H(K) = \log s^k = k \log m$. Since given the ciphertext and key we can find the message, $H(M \mid CK) = 0$. Similarly, $H(C \mid MK) = 0$. Since if $k > m$, the ciphertext is actually calculated with an initial part of the key of length m and we are not concerned with finding the key, we can w.l.o.g. assume that $k \leq m$. Therefore, we also have that $H(K \mid MC) = 0$.

Now,

$$I(M;C) = H(M) - H(M \mid C) = m \log s - H(M \mid C),$$

where

$$H(M \mid C) = H(M \mid CK) + I(M;K \mid C) = 0 + I(M;K \mid C).$$

For now, let $a = I(M;K \mid C) \geq 0$. Since the message and key are chosen independently, $I(M;K) = 0$. Since

$$I(M;K) = I(M;K \mid C) + R(M;K;C) = a + R(M;K;C) = 0,$$

we have that $R(M;K;C) = -a$.

Since the ciphertext is as random as the message, $H(C) = H(M) = m \log s$. This allows us to calculate a:

$$\begin{aligned} H(C) = m \log s &= H(C \mid MK) + I(M;C \mid K) + I(K;C \mid M) + R(M;C;K) \\ &= 0 + m \log s + k \log s - a, \end{aligned}$$

so $a = k \log s$. Finally,

$$I(M;C) = H(M) - H(M \mid C) = m \log s - I(M;K \mid C) = m \log s - k \log s = (m-k) \log s.$$

Since $s = 26$, $\log s > 0$, so $I(M;C) = 0$ if and only if $m - k = 0$, i.e. when $k = m$. See the entropy diagram in Figure 1 for visual support with our calculations.

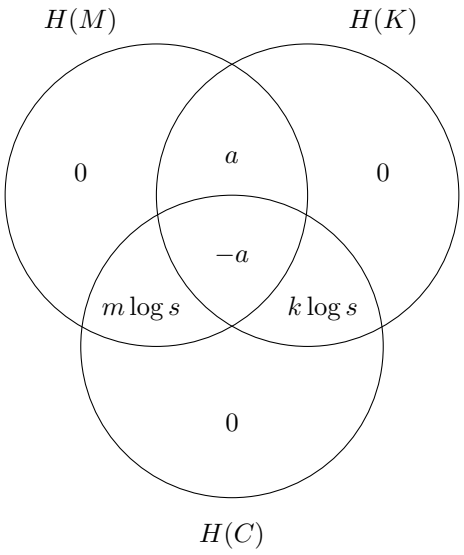

Figure 1: Entropy diagram for assignment 4(a)

(b)

Let k be the key-length and m be the message length. For simplicity, we assume that m is a multiple of k, i.e. $m = n \cdot k$ for some $n \in \mathbb{N}$. Since m is very large compared to k, this assumption does not change the probabilities too much (for example, we could just cut of the end of the message to achieve this).

We furthermore assume that, in the formula we will prove, P_K is actually the distribution of the key *prior* to creating the key, whereas P_C is the distribution of the cipher *after* creating the key, the message and the resulting cipher. That is, $P_C(x)$ is the probability that, when going uniformly to a random position of the cipher (each having probability $\frac{1}{m}$), the resulting letter is x. We can write the position of a letter in C as $i \cdot k + j$ for some $i \in \{0, \dots, n-1\}$ and a $j \in \{1, \dots, k\}$. Let C_{ik+j} be the resulting letter on the position. This letter can by construction be written as $C_{ik+j} = M_{ik+j} + K_j$. We get

$$\begin{aligned}
P_C(x) &= \sum_{i=0}^{n-1} \sum_{j=1}^{k} \frac{1}{m} P(C_{ik+j} = x) \\
&= \frac{1}{m} \sum_{i=0}^{n-1} \sum_{j=1}^{k} P(M_{ik+j} + K_j = x) \\
&= \frac{1}{m} \sum_{j=1}^{k} \sum_{i=0}^{n-1} P(M_{ik+j} = x - K_j) \\
&\approx \frac{1}{m} \sum_{j=1}^{k} \sum_{i=0}^{n-1} P_M(x - K_j) \\
&= \frac{n}{m} \sum_{j=1}^{k} P_M(x - K_j) \\
&= \frac{1}{k} \sum_{j=1}^{k} P_M(x - K_j).
\end{aligned}$$

Some steps require explanation: In the third to last step, we replaced the "empirical distribution" by the prior distribution of M. This does obviously change the statement (before that, $P(M_{ik+j} = x - K_j)$ was either 1 or 0, depending on whether the fact is true or not), but due to the large message-size, it is at least approximately correct. In the second to last step, we used that the arguments in the inner sum do not depend on i anymore, and in the last step we used $m = k \cdot n$.

Now we are in the position to compute the collision probability for P_C:

$$\begin{aligned}
Coll(P_C) &= \sum_{x\in\mathcal{X}} (P_C(x))^2 \\
&\approx \sum_{x\in\mathcal{X}} \left(\frac{1}{k}\sum_{j=1}^{k} P_M(x-K_j)\right)^2 \\
&= \frac{1}{k^2}\sum_{x\in\mathcal{X}}\sum_{j,j'=1}^{k} P_M(x-K_j)P_M(x-K_{j'}) \\
&= \frac{1}{k^2}\left(\sum_{x\in\mathcal{X}}\sum_{j=1}^{k} P_M(x-K_j)^2 + \sum_{x\in\mathcal{X}}\sum_{j\neq j'=1}^{k} P(M=x-K_j)\cdot P(M=x-K_{j'})\right) \\
&\approx \frac{1}{k^2}\left(\sum_{j=1}^{k}\sum_{x\in\mathcal{X}} P_M(x-K_j)^2 + \sum_{x\in\mathcal{X}}\sum_{j\neq j'=1}^{k} P_K(x-M)P_K(x-M)\right) \\
&= \frac{1}{k^2}\left(k\cdot Coll(P_M) + \sum_{x\in\mathcal{X}}\sum_{j\neq j'=1}^{k}\frac{1}{|\mathcal{X}|^2}\right) \\
&= \frac{1}{k}\cdot Coll(P_M) + \frac{1}{k^2}\cdot|\mathcal{X}|\cdot k(k-1)\cdot\frac{1}{|\mathcal{X}|^2} \\
&= \frac{1}{k}Coll(P_M) + \frac{k-1}{k}\frac{1}{|\mathcal{X}|} \\
&= \frac{1}{k}Coll(P_M) + \frac{k-1}{k}Coll(P_K).
\end{aligned}$$

Some steps require again some explanation: In the fifth step, we switched the roles of M and K_j and went from the empirical distribution of the key to the prior distribution. That made that step again only approximately correct. In the fourth to last step, we used that the distribution of K is uniform. In the third to last step we used that there are precisely $k\cdot(k-1)$ summands in the inner sum of the right-hand-side, and all these summands are equal. And in the last step, we used that $Coll(P_K) = \frac{1}{|\mathcal{X}|}$ since P_K is uniform. That finishes the proof.

(c)

Since the key is uniformly chosen at random from the alphabet, it is clear that for any letter x of the alphabet, $P_K(x) = 1/26$. Therefore,

$$Coll(P_K) = 26 \cdot (1/26)^2 = 26/26^2 = 1/26.$$

By counting the occurrences of each letter in the ciphertext, and dividing that count by the total amount of letters, we obtain a probability distribution for the ciphertext. By plugging this into the definition of collision probability given in the assignment, we can find $Coll(P_C)$. Since we calculated the probability distribution programmatically, we will also calculate k programmatically. So, instead of showing an approximation here, we will directly plug this calculation into the calculation of k to find an answer that is as correct as possible.

We can approximate k by rewriting the given approximation of $Coll(P_C)$ and plugging in the three collision probabilities we know by now:

$$\begin{aligned}
Coll(P_C) &\approx \frac{1}{k}Coll(P_M) + \frac{k-1}{k}Coll(P_K) \\
&\approx \frac{1}{k}Coll(P_M) + Coll(P_K) - \frac{1}{k}Coll(P_K) \\
Coll(P_C) - Coll(P_K) &\approx \frac{1}{k}(Coll(P_M) - Coll(P_K)) \\
k(Coll(P_C) - Coll(P_K) &\approx Coll(P_M) - Coll(P_K) \\
k &\approx \frac{Coll(P_M) - Coll(P_K)}{Coll(P_C) - Coll(P_K)} \approx 3.0148 \approx 3.
\end{aligned}$$

(d)

So our key is of length 3 and we know a few words that *might* occur in the encrypted message. Since we don't know which word occurs in the text, let alone *where* in the text it appears, we will have to use brute force to find the original message.

Because of the length of our key, every third letter of a word needs to be shifted according to the same key character. First, we loop through each $word$. Then we loop through each possible position j in the ciphertext, so $0 \leq j < |ciph| - |word|$. For each j, we store the currently obtained shifts for each key character, initially unset. Then, we loop through each pair of characters of $word$ that are 3 characters apart. More precisely, we loop through each possible position i in $word$ for which we can check a pair of letters that should be shifted by the same key character. So, $0 \leq i < |word| - k$ and the left character is at position i and the right character at position $i + k$. Then, we calculate the shifts for both characters compared to their relative position ($j+i$ and $j+i+k$) in the ciphertext. If they are equal and the current shift of the relevant key character (at position $i + j \mod k$) is still unset, we set it to the found shift. If this shift is already set, but equal to the current shift, we continue. Otherwise, we break out of the loop and discard the current position in the ciphertext.

If all possible positions i for the current j and $word$ are consistent (so the loop isn't broken), the currently obtained shifts are printed, converting it to a possible key word, as well as printing the text deciphered according to this key. However, at this point we discovered that only two of the given words could possibly occur in the text: `QUEEN` and `GRACE`. Notice that these words are thus short, that there is no sixth character to compare the third character with. Therefore, we still needed to use brute force to find the shift of the third key character.

Finally, we found that the keyword must be `BAT`, giving us the deciphered message:

```
OLOVELIESTOFQUEENSYOURSISTHEHONOUROFCUTTINGTHERIBBONWHICHPROVESTHATIHA
VELOSTMYBETBYJUPITERIYIELDWITHAGOODGRACEBEFORESOMUCHGRACE.
```