Assignment 4 - Information Theory and Inference

Consider the "noisy typewriter" channel. In this channel, symbols from

$$\mathcal{A}_Z = \{\texttt{A},\texttt{B},\texttt{C},\ldots,\texttt{Y},\texttt{Z},\texttt{-}\}$$

with $|\mathcal{A}_Z| = 27$ (we are considering the English alphabet plus the space – symbol) are mapped onto the same alphabet ($\mathcal{A}_Y = \mathcal{A}_Z$), with the following probabilistic rule: each letter is mapped with equal probability onto the same letter, the preceding letter or the following letter. In other words, we have

$$P(Y=\texttt{A}|Z=\texttt{B}) = 1/3, \quad P(Y=\texttt{B}|Z=\texttt{B}) = 1/3, \quad P(Y=\texttt{C}|Z=\texttt{B}) = 1/3$$

$$P(Y=\texttt{B}|Z=\texttt{C}) = 1/3, \quad P(Y=\texttt{C}|Z=\texttt{C}) = 1/3, \quad P(Y=\texttt{D}|Z=\texttt{C}) = 1/3$$

Note that the preceding symbol of A is – and the following symbol of – is A. We saw that one can reliably send one of 9 characters through this channel by selecting non-confusable codewords, e.g. {B,E,H,K,N,Q,T,W,Z}. What if you want to reliably send *all of the 27 characters* ($\mathcal{A}_X = \mathcal{A}_Z$) by using the channel?

1. One could use a repetition code R_3, that uses codewords of length 3:

$$E(\texttt{A}) = \texttt{AAA} \qquad E(\texttt{B}) = \texttt{BBB} \qquad E(\texttt{C}) = \texttt{CCC}\ldots$$

and decoding provided by the MAP estimator. What is the information transmission rate of this code? What is the error probability (Bayes risk)?

2. There is an efficient code using codewords of length 3 to code sequences of length 2 (*pairs* of characters). Try to find it. What is the information transmission rate?

1. The information transmission rate is defined as $R = \dfrac{\log_2 |S|}{L(E)}$

In this case $S = A_x \Rightarrow |S| = 27$ and the length of the codeword is $L(E) = 3$

$$\Rightarrow R = \frac{\log_2(27)}{3} = \frac{\log(3^3)}{3} = \frac{3\log_2(3)}{3}$$

$$\boxed{R = \log_2(3)}$$

The decoding with the MAP estimator consists in finding z that maximizes $P(Y|Z)$

Each letter is mapped with the same probabilistic rule, so we can compute the error for a fixed letter : y_1

If y_1 is mapped on A_z, $z \in \{y_0, y_1, y_2\}$, $P(y_i) = \frac{1}{3}$

So, for one coded letter y_1 with the repetion code $R_3 \to y_1 y_1 y_1$

can receive one of 27 possible combination of $\{y_0, y_1, y_2\}$ that makes a triple.

(each element of $y_1 y_1 y_1$ can become y_0, y_1 or y_2 with equal probability $\frac{1}{3}$.

So the probability to receive one of the 27 possible triples is $(1/3)^3$

Now let's imagine that we want to send "B", it is encoded as "BBB".

The triple received $(y_0, y_1, y_2) = Y$ has $\underbrace{27 \text{ different possibilities}}_{\text{combinations of } \{A,B,C\}}$ as each y_i can be one of the element of $\{A, B, C\} = S(B)$.

To decode and find the original letter, the decoder consider each of the 27 letters of the alphabet.

x' can possibly be the original letter if (and only if) all the letters from Y are on $\{x'-1, x', x'+1\} = S(x')$

→ If only one x' satisfies this condition, the decoder choose it and the error is 0 if it's right, 1 if not

→ If there are multiple possible x', the decoder selects one uniformly among them.

We can analyse the 27 possible cases (when "B" is sent):

If $x' = A$ or $x' = C$ we would miss either C or A in $S(x')$

1) a) $y_i \in \{A, B, C\} \to$ ABC, ACB, BAC, BCA, CAB, CBA : $\boxed{\text{6 combinations.}}$

Only $x' = B$ satisfies the condition for which all the y_i are in $S(x') = \{x'-1, x', x'+1\}$ ie. in $\{A, B, C\}$

$\Rightarrow \boxed{\text{error} = 0}$

b) $y_i \in \{A, C\} \to$ AAC, ACA, CAA, CCA, CAC, ACC : $\boxed{\text{6 combinations}}$

Only $x' = B$ satisfies the condition for which all the y_i are in $\{x'-1, x', x'+1\}$ ie. in $\{A, B, C\}$

$\Rightarrow \boxed{\text{error} = 0}$

2) a) $y_i \in \{A, B\} \to$ AAB, ABA, BAA, BBA, BAB, ABB $\boxed{\text{: 6 combinations}}$

Both (and only) $x' = A$ and $x' = B$ satisfy the condition, ie all y_i are in $\{-, A, B\}$ or in $\{A, B, C\}$

$\to$ 2 possible x', only one is the right one (B) $\Rightarrow \boxed{\text{error} = \frac{1}{2}}$

b) $y_i \in \{B, C\} \to$ BBC, BCB, CBB, CCB, CBC, BCC: $\boxed{\text{6 combinations}}$

Both (and only) $x' = B$ and $x' = C$ satisfy the condition, ie all y_i are in $\{A, B, C\}$ or in $\{B, C, D\}$

$\to$ 2 possible x', only one is the right one (B) $\Rightarrow \boxed{\text{error} = \frac{1}{2}}$

3) a) $y_i \in \{A\} \to$ AAA: $\boxed{\text{1 combination}}$ $x' = -$ or $x' = A$ or $x' = B$ are the only 3 possible x' that satisfy y_i in $\{Z, -, A\}$ or in $\{-, A, B\}$ or in $\{A, B, C\}$.

Only one x' is the right one out of 3 $\Rightarrow \boxed{\text{error} = \frac{2}{3}}$

b) $y_i \in \{B\} \to$ BBB: $\boxed{\text{1 combination}}$ $x' = A$ or $x' = B$ or $x' = C$ (following the same reasoning as previously) $\Rightarrow \boxed{\text{error} = 2/3}$

c) $y_i \in \{C\} \to$ CCC : $\boxed{\text{1 combination}}$ $x' = B$ or $x' = C$ or $x' = D$ (following the same reasoning as previously) $\Rightarrow \boxed{\text{error} = 2/3}$

because each triple has a proba 1/27 to be received

$$\Rightarrow P_{error} = \frac{1}{27}\left[6.0 + 6.0 + 6\cdot\frac{1}{2} + 6\cdot\frac{1}{2} + 1\cdot\frac{2}{3} + 1\cdot\frac{2}{3} + 1\cdot\frac{2}{3}\right]$$

$$\boxed{P_{error} = \frac{8}{27}}$$

2) If we consider the same alphabet $A_{\mathbb{Z}}$, used to create sequences of length 2, then the number of possible sequences that can be sent is $27^2 = \boxed{729}$

We want to use a code length of 3

$$\Rightarrow R = \frac{\log_2(27^2)}{3} = \frac{2\log_2(3^3)}{3} = \frac{2\times 3\times \log(3)}{3}$$

$$\Rightarrow \boxed{R = 2\log_2(3)}$$ the information transmission rate.

In order to find an efficient code, we need to find one with $\boxed{\text{error} = 0}$, i.e the decoder find only one possible x' and doesn't have to pick one among several x'.

The goal is to have 729 coded words, such as the $S(x') = \{x'-1, x', x'+1\}$ is unique for each coded word and are all totally disjoints from each other.

For that we can use $\boxed{\text{the sub-alphabet } S_9 = \{A, D, G, J, M, P, S, V, Y\}}$ containing 9 letters, spaced by 3 from the original alphabet.

It respects the criteria just enounced previously; for example:

$$\left.\begin{array}{l} x' = A \Rightarrow S(A) = \{-, A, B\} \\ x' = D \Rightarrow S(D) = \{C, D, E\} \\ x' = G \Rightarrow S(G) = \{F, G, H\} \end{array}\right\} \longrightarrow \text{no intersection.}$$

⋮

We are restricted to use code words of length 3. Thanks to that, we can create the necessary amount of codewords with $S_9 \Rightarrow |S_9| = 9$

$\Rightarrow \boxed{9^3 = 729 \text{ coded words}}$ that are combinations of 3 elements of S_9

We have then 729 original words (x_1, x_2), the pairs we want to encode, and 729 possible coded words, that are triple (y_1, y_2, y_3). $y_i \in S_9$ and $x_i \in A_z$

A bijection can then be exploited to associate for each of the 729 (x_1, x_2) an unique tuple (y_1, y_2, y_3) from the 729 possible tuples.
The association could be arbitrarily defined "manually", by creating a dictionary, or even by finding a bijective function $(x_1, x_2) \mapsto (y_1, y_2, y_3)$.

Thanks to that bijection, we know now that when decoding $Y = (y_1, y_2, y_3)$, the decoder will find only $\boxed{\text{one possible } x' \text{ that can generate this } Y}$, therefore $\boxed{\text{the error will be } 0}$