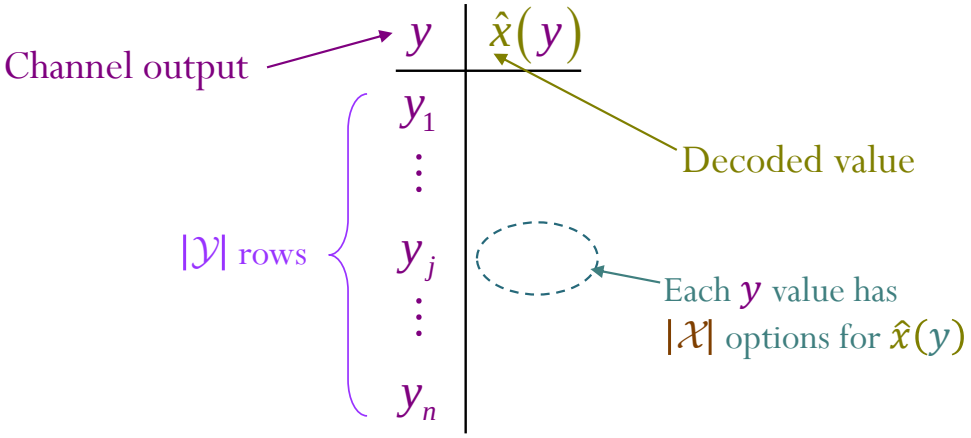
ECS452 2017/2 Part I.5 Dr.Prapun

3.3 Optimal Decoding for DMC

From the previous section, we now know how to compute the error probability for any given decoder. Here, we will attempt to find the "best" decoder. Of course, by "best", we mean "having minimum value of error probability". It is interesting to first consider the question of how many reasonable decoders we can use.

3.30. How many "reasonable" decoders are there?: Recall from 3.22 that a decoder $\hat{x}(\cdot)$ is a function that map each observed valued of the channel output y to the guessed value of the channel input. Therefore we can think of a decoder as a table:



We have already seen this table representation in Example 3.27 and Example 3.28. Such table has $|\mathcal{Y}|$ rows. For each value of y, we need to specify what is the value of $\hat{x}(y)$. To have a chance of correct guessing, any "reasonable" decoder would select the value of $\hat{x}(y)$ from $\mathcal{X}$. Therefore, there are $|\mathcal{X}|^{|\mathcal{Y}|}$ reasonable decoders.

Example 3.31. The naive decoder in Example 3.26 is not a reasonable decoder. The channel input X is either 0 or 1. So, it does not make sense have a guess value of $\hat{x}(2) = 2$ or $\hat{x}(3) = 3$.

Example 3.32. "Reasonable" Decoder for BSC: For BSC in Example 3.2, any decoder has to answer two important questions:

(a) What should be the guess value of X when $Y = 0$ is observed? $\hat{x} = 0$ or $\hat{x} = 1$

(b) What should be the guess value of X when $Y = 1$ is observed? $\hat{x} = 0$ or $\hat{x} = 1$

Essentially, any reasonable decoder for the BSC needs to complete this table:

y	$\hat{x}(y)$
0	
1	

So, only four reasonable decoders for BSC:

y	$\hat{x}(y)$
0	0
1	1

Naive $\hat{x}(y) = y$

y	$\hat{x}(y)$
0	1
1	0

$\hat{x}(y) = \bar{y} = 1 - y$

y	$\hat{x}(y)$
0	0
1	0

$\hat{x}(y) = 0$

y	$\hat{x}(y)$
0	1
1	1

$\hat{x}(y) = 1$

Example 3.33. For the DMC defined in Example 3.26, how many reasonable decoders are there?

$|\mathcal{X}|^{|\mathcal{Y}|} = 2^3 = 8$

We calculate the error probability of three decoders in Example 3.26, Example 3.27, and Example 3.28. There are still many reasonable possibilities to evaluate. Using `MATLAB`, we can find the error probability for all possible reasonable decoders:

y	$\hat{x}_{D1}(y)$	$\hat{x}_{D2}(y)$	$\hat{x}_{D3}(y)$	$\hat{x}_{D4}(y)$	$\hat{x}_{D5}(y)$	$\hat{x}_{D6}(y)$	$\hat{x}_{D7}(y)$	$\hat{x}_{D8}(y)$
1	0	0	0	0	1	1	1	1
2	0	0	1	1	0	0	1	1
3	0	1	0	1	0	1	0	1
$P(\mathcal{E})$	0.80	0.62	0.52	0.34	0.66	0.48	0.38	0.20

$\hat{x}_{optimal}(y)$ smallest (optimal)

Ex. 3.27 Ex. 3.28

3.34. For general DMC, it would be tedious to list all possible decoders. It is even more time-consuming to try to calculate the error probability for all of them. Therefore, in this section, we will derive a visual construction and a formula of the "optimal" decoder.

3.35. From the recipe 3.29 for finding $P(\mathcal{C})$ and $P(\mathcal{E})$, we see that $P(\mathcal{C})$ is the sum of our circled numbers. So, to maximize $P(\mathcal{C})$, we want to circle the largest number. For row y in the decoding table, whatever the value we select for $\hat{x}(y)$ will determine which number will be circled in the column corresponding to y in matrix $\mathbf{P}$. To maximize $P(\mathcal{C})$, we want to circle the largest number in the column. This means $\hat{x}(y)$ should be the same as the x value that maximizes the probability value in the corresponding column.

Example 3.36. For the DMC and the input probablities defined in Example 3.26, the joint pmf matrix $\mathbf{P}$ was found to be

$\hat{x}$	1	1	1
$x \setminus y$	1	2	3
0	0.1	0.04	0.06
1	0.24	0.32	0.24

Therefore, the optimal decoder is

$P(\mathcal{C}) = 0.24 + 0.32 + 0.24$
$= 0.8$

$P(\mathcal{E}) = 1 - 0.8 = 0.2$

y	$\hat{x}(y)$
1	1
2	1
3	1

3.37. Deriving the optimal decoder: Mathematically, we first note that to minimize $P(\mathcal{E})$, we need to maximize $P(\mathcal{C})$. Here, we apply the total probability theorem by using the events $[Y = y]$ to partition the sample

space:

$$P(\mathcal{C}) = \sum_y P(\mathcal{C} \,|[Y = y]) \, P[Y = y].$$

Event $\mathcal{C}$ is the event $[\hat{X} = X]$. Therefore,

$$P(\mathcal{C}) = \sum_y P\left[\hat{X} = X \,|Y = y\right] P[Y = y].$$

Now, recall that our decoder is a function of Y; that is $\hat{X} = \hat{x}(Y)$. So,

$$\begin{aligned} P(\mathcal{C}) &= \sum_y P[\hat{x}(Y) = X \,|Y = y] \, P[Y = y] \\ &= \sum_y P[X = \hat{x}(y) \,|Y = y] \, P[Y = y] \end{aligned}$$

In this form, we see[13] that for each $Y = y$, we should maximize $P[X = \hat{x}(y) \,|Y = y]$. Therefore, for each y, the decoder $\hat{x}(y)$ should output the value of x which maximizes[14] $P[X = x|Y = y]$:

$$\hat{x}_{\text{optimal}}(y) = \arg\max_x P[X = x \,|Y = y].$$

In other words, the *optimal* decoder is the decoder that maximizes the "a posteriori probability" $P[X = x \,|Y = y]$.

Definition 3.38. The optimal decoder derived in 3.37 is called the **maximum a posteriori probability (MAP) decoder**:

$$\hat{x}_{\text{MAP}}(y) = \hat{x}_{\text{optimal}}(y) = \arg\max_x P[X = x \,|Y = y]. \tag{6}$$

3.39. After the fact, it is quite intuitive that this should be the best decoder. Recall that the decoder don't have a direct access to the X value.

[13] We also see that any decoder that produces random results (on the support of X) can not be better than our optimal decoder. Outputting the value of x which does not maximize the a posteriori probability reduces the contribution in the sum that gives $P(\mathcal{C})$.

[14] For those who are not familiar with the "arg max" (arguments of the maximum) function,

$$\arg\max_x f(x) = \text{the } x \text{ value that maximizes } f(x).$$

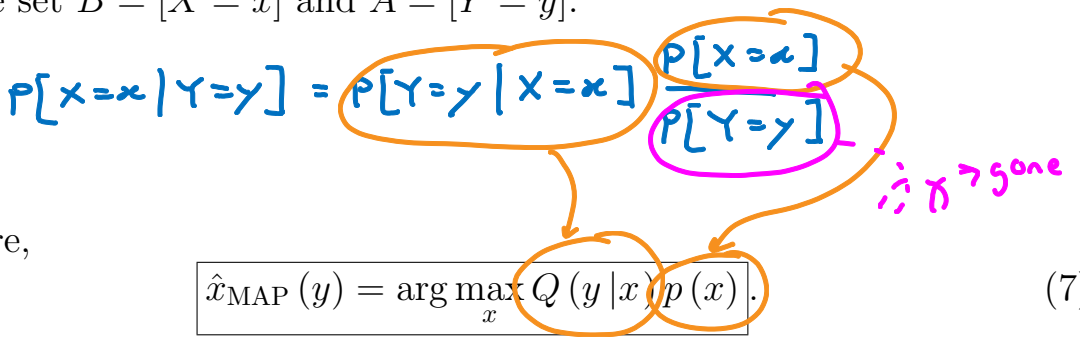
The corresponding maximum value of $f(x)$ is $\max_x f(x)$. In other words, in contrast to global maximum, referring to the largest outputs of a function, arg max refers to the inputs, or arguments, at which the function outputs are as large as possible. For example, for $f(x) = 5 - x^2$, we have $\arg\max_x f(x) = 0$ and $\max_x f(x) = 5$.

- Without knowing the value of Y, to minimize the error probability, it should guess the most likely value of X which is the value of x that maximize $P[X = x]$.
- Knowing $Y = y$, the decoder can update its probability about x from $P[X = x]$ to $P[X = x|Y = y]$. Therefore, the decoder should guess the value of the most likely x value conditioned on the fact that $Y = y$.

3.40. We should manipulate Formula (3.38) for the MAP decoder a bit further because, in practice, we usually only know $p(x)$ and $Q(y|x)$. To connect these terms to $P[X = x|Y = y]$ required in (3.38), fist, recall "Form 1" of the Bayes' theorem:

$$P(B|A) = P(A|B)\frac{P(B)}{P(A)}.$$

Here, we set $B = [X = x]$ and $A = [Y = y]$.



Therefore,

$$\boxed{\hat{x}_{\text{MAP}}(y) = \arg\max_x Q(y|x)p(x)}. \tag{7}$$

Note that the term $P[Y = y]$ does not depend on x and it is positive; therefore, it does not change the the answer of arg max and hence can be ignored.

3.41. A recipe for finding the MAP decoder (optimal decoder) and its corresponding error probability:

(a) Find the **P** matrix by scaling elements in each row of the **Q** matrix by their corresponding prior probability $p(x)$.

(b) Select (by circling) the maximum value in each column (for each value of y) in the **P** matrix.

- If there are multiple max values in a column, select one. This won't affect the optimality of your answer.

(i) The corresponding x value is the value of $\hat{x}$ for that y.

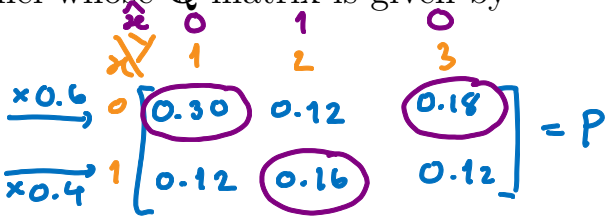
(ii) The sum of the selected values from the $\mathbf{P}$ matrix is $P(\mathcal{C})$.

(c) $P(\mathcal{E}) = 1 - P(\mathcal{C})$.

Example 3.42. We have applied recipe 3.41 back when we try to find the optimal decoder in Example 3.36.

Example 3.43. Find the MAP decoder and its corresponding error probability for the DMC channel whose $\mathbf{Q}$ matrix is given by

$$\mathbf{Q} = \begin{array}{c|ccc} x \setminus y & 1 & 2 & 3 \\ 0 & 0.5 & 0.2 & 0.3 \\ 1 & 0.3 & 0.4 & 0.3 \end{array}$$



y	$\hat{x}(y)$
1	0
2	1
3	0

and $\underline{\mathbf{p}} = [0.6, 0.4]$. Note that the DMC is the same as in Example 3.26 but the input probabilities are different.

$$P(\mathcal{C}) = 0.30 + 0.16 + 0.18 = 0.64$$

$$P(\mathcal{E}) = 1 - 0.64 = 0.36$$

Definition 3.44. In many scenarios, the MAP decoder is too complicated or the prior probabilities are unknown. In such cases, we may consider using a *suboptimal* decoder that ignores the prior probability term in (7). This decoder is called the **maximum likelihood (ML)** decoder:

$$\boxed{\hat{x}_{\mathrm{ML}}(y) = \arg\max_{x} Q(y|x)}. \tag{8}$$

3.45. ML decoder is the same as the MAP decoder when X is a uniform random variable. In other words, when the prior probabilities $p(x)$ are uniform, the ML decoder is optimal.

3.46. A recipe for finding the ML decoder and its corresponding error probability:

(a) Select (by circling) the maximum value in each column (for each value of y) in the $\mathbf{Q}$ matrix.

- If there are multiple max values in a column, select one. Different choices will lead to different $P(\mathcal{E})$. However, if the information about $\underline{\mathbf{p}}$ is not available at the decoder, it can not determine which choice is better anyway.

- The corresponding x value is the value of $\hat{x}$ for that y.

(b) Find the $\mathbf{P}$ matrix by scaling elements in each row of the $\mathbf{Q}$ matrix by their corresponding prior probability $p(x)$.

(c) In the $\mathbf{P}$ matrix, select the elements corresponding to the selected positions in the $\mathbf{Q}$ matrix.

(d) The sum of the selected values from the $\mathbf{P}$ matrix is $P(\mathcal{C})$.

(e) $P(\mathcal{E}) = 1 - P(\mathcal{C})$.

Example 3.47. Find the ML decoder and its corresponding error probability for the DMC channel in Example 3.26 whose $\mathbf{Q}$ matrix is

$$\mathbf{Q} = \begin{array}{c|ccc} x \setminus y & 1 & 2 & 3 \\ 0 & 0.5 & 0.2 & 0.3 \\ 1 & 0.3 & 0.4 & 0.3 \end{array}$$

$\xrightarrow{\times 0.2}$ $\xrightarrow{\times 0.8}$ $\mathbf{P} = \begin{bmatrix} 0.10 & 0.04 & 0.06 \\ 0.24 & 0.32 & 0.24 \end{bmatrix}$

y	$\hat{x}_{\mathrm{ML}}(y)$
1	0
2	1
3	0

and $\underline{\mathbf{p}} = [0.2, 0.8]$.

$P(\mathcal{C}) = 0.10 + 0.32 + 0.06 = 0.48$

$P(\mathcal{E}) = 1 - 0.48 = 0.52$

Example 3.48. Find the ML decoder and the corresponding error probability for a communication over BSC with $p = 0.1$ and $p_0 = 0.8$.

Note that

- the prior probabilities p_0 (and p_1) is not used when finding $\hat{x}_{\mathrm{ML}}$.
- the ML decoder and the MAP decoder are the same in this example
  - ML decoder can be optimal even when the prior probabilities are not uniform.

3.49. In general, for BSC, it's straightforward to show that

(a) when $p < 0.5$, we have $\hat{x}_{\mathrm{ML}}(y) = y$ with corresponding $P(\mathcal{E}) = p$.

(b) when $p > 0.5$, we have $\hat{x}_{\mathrm{ML}}(y) = 1-y$ with corresponding $P(\mathcal{E}) = 1-p$.

(c) when $p = 0.5$, all four reasonable decoders have the same $P(\mathcal{E}) = 1/2$.

- In fact, when $p = 0.5$, the channel completely destroys any connection between X and Y. In particular, in this scenario, $X \perp\!\!\!\perp Y$. So, the value of the observed y is useless.