Final Exam - Moed A

Total time for the exam: 3 hours!

Important: For **True / False** questions, copy the statement to your notebook and write clearly true or false. You should prove the statement if true, and provide counterexample otherwise.

1) **Multipath Gaussian channel. (24 Points)**
Consider a Gaussian noise channel of power constraint P, where the signal takes two different paths and the received noisy signals, Y_1 and Y_2, are feed into a finite impulse response (FIR) filter which coherently combines the input signals, namely, $Y = a \cdot Y_1 + b \cdot Y_2$, where $a, b \in \mathbb{R}$. The system model is shown in the figure below.

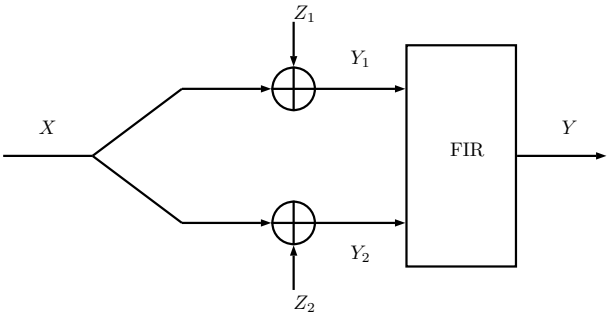


Fig. 1: Channel model

We assume that Z_1 and Z_2 are jointly Gaussian, with zero means, and covariance matrix

$$K = \begin{bmatrix} N & N\rho \\ N\rho & N \end{bmatrix}.$$

a) *Given* $a, b \in \mathbb{R}$, find the capacity $C(a, b, \rho)$ of the channel described above.
Solution:
We have $Y_1 = X + Z_1$ and $Y_2 = X + Z_2$. Also, $Y = (a + b) \cdot X + aZ_1 + bZ_2$. Using the results derived in the class, we know that the capacity of this Gaussian channel is given by

$$\begin{aligned} C(a, b, \rho) &= \frac{1}{2} \log \left(1 + \frac{\text{var}\left[(a + b) \cdot X\right]}{\text{var}\left[aZ_1 + bZ_2\right]} \right) \\ &= \frac{1}{2} \log \left(1 + \frac{(a + b)^2}{a^2 + b^2 + 2ab\rho} \frac{P}{N} \right). \end{aligned}$$

b) Evaluate your result in the previous item for $\rho = 0, -1$, and 1? Explain your results when $a = b$.
Solution:
We have:

$$\begin{aligned} C(a, b, 0) &= \frac{1}{2} \log \left(1 + \frac{(a + b)^2}{a^2 + b^2} \frac{P}{N} \right), \\ C(a, b, 1) &= \frac{1}{2} \log \left(1 + \frac{P}{N} \right), \\ C(a, b, -1) &= \frac{1}{2} \log \left(1 + \frac{(a + b)^2}{(a - b)^2} \frac{P}{N} \right). \end{aligned}$$

Also, when $a = b$ we get

$$\begin{aligned} C(a, a, 0) &= \frac{1}{2} \log \left(1 + \frac{2P}{N} \right), \\ C(a, a, 1) &= \frac{1}{2} \log \left(1 + \frac{P}{N} \right), \\ C(a, a, -1) &= \infty. \end{aligned}$$

The above result make sense. Indeed when $\rho = 0$, we get $Y = 2aX + a(Z_1 + Z_2)$, and we see that the SNR is $2P/N$ independently of a. On the other hand when $\rho = 1$, this essentially means that $Z_1 = Z_2$, and thus $Y = 2aX + 2aZ_1$, which implies that the SNR is P/N independently of a (this is true also if $a \neq b$). When $\rho = -1$, and $a = b$ we see that the capacity is infinite, which makes sense as in this case $Z_1 = -Z_2$, and thus summing up Y_1 and Y_2 cancels the

noise, so unbounded amount of information can be transmitted.

c) What is the best filter in the sense of maximizing the capacity, i.e., solve $\max_{a,b} C(a,b,\rho)$, where the maximization is over all $a,b \in \mathbb{R}$. Explain your result. (You may use the inequality $\frac{(a+b)^2}{a^2+b^2+2ab\rho} \leq \frac{2}{1+\rho}$, for any $a,b \in \mathbb{R}$ and $\rho \in [-1,]$.)
Solution:
We want to solve the following optimization problem:

$$\max_{a,b\in\mathbb{R}} \frac{1}{2}\log\left(1+\frac{(a+b)^2}{a^2+b^2+2ab\rho}\frac{P}{N}\right).$$

Since $\log(\cdot)$ is monotonically increasing we can focus on the term inside the logarithm. Accordingly, taking partial derivatives with respect to a and b, we get that the optimal solution appears on $a=b$. In this case we get

$$\max_{a,b\in\mathbb{R}} \frac{1}{2}\log\left(1+\frac{(a+b)^2}{a^2+b^2+2ab\rho}\frac{P}{N}\right) = \frac{1}{2}\log\left(1+\frac{2}{1+\rho}\frac{P}{N}\right).$$

To see that this is indeed the maximum, we can use the following inequality:

$$\frac{(a+b)^2}{a^2+b^2+2ab\rho} \leq \frac{2}{1+\rho},$$

for any $\rho \in [-1,1]$, and $a,b \in \mathbb{R}$. The above result implies that choosing $a=b$, and an arbitrary $a \neq 0$ is optimal, independently of ρ. Indeed, since all we want to do is to maximize the SNR, it is known that this can be achieved by the matched filter, which in turn coincides with the above result. Given $a=b$, it makes sense that $a \neq 0$ can be chosen arbitrarily as it does not affect the SNR (both the signal and the noise are multiplied by the same coefficient a).

2) **True or False on entropy identities (24 Points)**:

a) **True/False**: For any discrete random variables, X_1, X_2, and X_3,

$$H(X_1,X_2,X_3) \leq \frac{1}{2}\left[H(X_1,X_2)+H(X_2,X_3)+H(X_3,X_1)\right].$$

Solution: **True.**
Using chain rule, $H(X1,X2,X3)$ can be expanded in the following two way

$$\begin{aligned}
2H(X_1,X_2,X_3) =& H(X_1,X_2)+H(X_3|X_1,X_2)+H(X_2,X_3)+H(X_1|X_2,X_3)\\
\leq& H(X_1,X_2)+H(X_2,X_3)+H(X_3|X_1,X_2)+H(X_1)\\
\leq& H(X_1,X_2)+H(X_2,X_3)+H(X_3|X_1)+H(X_1)\\
=& H(X_1,X_2)+H(X_2,X_3)+H(X_3,X_1).
\end{aligned}$$

b) **True/False**: For any discrete random variables, X_1, X_2, and X_3,

$$H(X_1,X_2,X_3) \geq \frac{1}{2}\left[H(X_1,X_2|X_3)+H(X_2,X_3|X_1)+H(X_3,X_1|X_2)\right].$$

Solution: **True.**

$$\begin{aligned}
H(X_1,X_2|X_3)+H(X_2,X_3|X_1)+H(X_3,X_1|X_2) &= H(X_1,X_2|X_3)+H(X_3)+H(X_2,X_3|X_1)+H(X_1)\\
&\quad + H(X_3,X_1|X_2)+H(X_2)-[H(X_1)+H(X_2)+H(X_3)]\\
&= 3H(X_1,X_2,X_3)-[H(X_1)+H(X_2)+H(X_3)]\\
&\leq 3H(X_1,X_2,X_3)-H(X_1,X_2,X_3)\\
&= 2H(X_1,X_2,X_3).
\end{aligned}$$

c) **True/False**: For two probability distributions, p_{XY} and q_{XY}, that are defined on $\mathcal{X}\times\mathcal{Y}$, the following holds:

$$D(p_{XY}||q_{XY}) \geq D(p_X||q_X).$$

Solution: **True.**
Consider the definition of the conditional divergence,

$$D\big(P_{X|Z}\big|\big|Q_{X|Z}\big|P_Z\big) = \sum_{(x,z)\in\mathcal{X}\times\mathcal{Z}} P_{X,Z}(x,z)\log\left(\frac{P_{X|Z}(x|z)}{Q_{X|Z}(x|z)}\right).$$

We recall that

$$D\big(P_{X,Y}\big|\big|Q_{X,Y}\big) = D\big(P_X\big|\big|Q_X\big)+D\big(P_{Y|X}\big|\big|Q_{Y|X}\big|P_X\big),$$

where

$$D\big(P_{Y|X}\big|\big|Q_{Y|X}\big|P_X\big) = \sum_{x\in\mathcal{X}} P_X(x)D\big(P_{Y|X=x}\big|\big|Q_{Y|X=x}\big),$$

which is non-negative. We conclude that

$$D(p_{XY}||q_{XY}) \geq D(p_X||q_X).$$

d) Given are two channels with identical inputs and outputs alphabets ($|\mathcal{X}_i| = |\mathcal{Y}_i|$ for $i = 1, 2$). Their capacities are denoted by C_1 and C_2, respectively, and the capacity of their cascaded version is C_{12} or C_{21} depending on the channel ordering.

i) **True/False**: If $|\mathcal{X}_i| = |\mathcal{Y}_i| = 2$ for all i, then $C_{12} = C_{21} = 0$ if and only if $C_1 = 0$ or $C_2 = 0$.
Solution: **True.**
One direction of the proof is trivial. The capacity of a channel is zero if and only if $X \perp\!\!\!\perp Y$ for any input distribution. For a binary channel with transition probabilities, say α and β, this is translated to the fact that the capacity is zero if and only if $\alpha + \beta = 1$. Now, the transition probabilities of the channels are denoted by $p(y = 0|x = 1), p(y = 1|x = 0) = (\alpha, \beta)$ and (δ,γ), and the cascaded channel parameters are $(\bar{\alpha}\gamma + \alpha\bar{\delta}, \beta\bar{\gamma} + \bar{\beta}\delta)$. It is not difficult to check that the capacity of the cascaded channel is 0 if and only if $\alpha + \beta = 1$ or $\delta + \gamma = 1$.

ii) **True/False**: If $|\mathcal{X}_i| = |\mathcal{Y}_i| = 3$ for all i, then $C_{12} = 0$ if and only if $C_1 = 0$ or $C_2 = 0$.
Solution: **False.**
Consider the following channels:
First channel: $p(y = 0|x = 0) = p(y = 1|x = 1) = p(y = 1|x = 2) = 1$. Second channel: $p(y = 0|x = 0) = p(y = 0|x = 1) = p(y = 2|x = 2) = 1$. The capacity of each channel is 1 bit but the capacity of their cascaded version (in the above ordering) is 0.

3) **Shannon code (24 Points)**: Consider the following method for generating a code for a random variable X which takes on m values $\{1, 2, \ldots, m\}$ with probabilities $p_1, p_2, \ldots, p_m$. Assume that the probabilities are ordered so that $p_1 \geq p_2 \geq \cdots \geq p_m$. Define

$$F_i = \begin{cases} 0 & i = 1 \\ \sum_{k=1}^{i-1} p_k & i = 2, 3, \ldots, m+1 \end{cases}, \tag{1}$$

namely, the sum of the probabilities of all symbols less than i. Then the codeword for i is the number $F_i \in [0, 1]$ rounded off to l_i bits, where $l_i = \lceil \log \frac{1}{p_i} \rceil$.

a) **True/False**

i) The code constructed by this process is prefix-free.
Solution: **True.**
By the choice of l_i, we have

$$2^{-l_i} \leq p_i < 2^{-(l_i - 1)}. \tag{2}$$

Thus F_j, $j > i$ differs from F_i by at least 2^{-l_i}, and will therefore differ from F_i is at least one place in the first l_i bits of the binary expansion of F_i. Thus the codeword for F_j, $j > i$, which has length $l_j \geq l_i$, differs from the codeword for F_i at least once in the first l_i places. Thus no codeword is a prefix of any other codeword.

ii) The average length L satisfies $H(X) \leq L < H(X) + 1$.
Solution: **True.**
Since $l_i = \lceil \log \frac{1}{p_i} \rceil$, we have

$$\log \frac{1}{p_i} \leq l_i < \log \frac{1}{p_i} + 1 \tag{3}$$

which implies that

$$H(X) \leq L = \sum p_i l_i < H(X) + 1. \tag{4}$$

b) Construct the code for the probability distribution $(0.5, 0.25, 0.125, 0.125)$.
Solution:
We build the following table

Symbol	Probability	F_i in decimal	F_i in binary	l_i	Codeword
1	0.5	0.0	0.0	1	0
2	0.25	0.5	0.10	2	10
3	0.125	0.75	0.110	3	110
4	0.125	0.875	0.111	3	111

The Shannon code in this case achieves the entropy bound (1.75 bits) and is optimal.

c) **True/False** For dyadic distribution, i.e. $\forall i : p_i = 2^{-l_i}$ for some positive integer $l_i's$, the average length for this code matches $H(X)$.
Solution: **True.**
Without loss of generality, assume

$$p_1 \geq p_2 \geq \cdots \geq p_m$$

$$l_1 \leq l_2 \leq \cdots \leq l_m$$

and by construction of the Shannon codes, $l_i = \lceil \log \frac{1}{p_i} \rceil = \log \frac{1}{p_i}$ and thus the expected code length is

$$E[l(X)] = \sum_{i=1}^{m} p_i \cdot l_i \tag{5}$$

$$= \sum_{i=1}^{m} p_i \cdot \log \frac{1}{p_i} \tag{6}$$

$$= H(X). \tag{7}$$

4) **Mixture of exponential distributions and EM (28 Points)**: A mixture of exponential distributions is defined by a vector of parameters $\lambda = [\lambda_1, \ldots, \lambda_k]$ and a latent discrete random variable Z which denotes the number of the exponential distributions hence $Z \in \{1, 2, ..., k\}$. The probability density function (pdf) of the mixture is:

$$P(x; \lambda) = \sum_{j=1}^{K} P(Z = j) f(x|z = j; \lambda_j) \tag{8}$$

where $f(x|z = j; \lambda_j) = \lambda_j \exp(-\lambda_j \cdot x)$, for $x \geq 0$, is the exponential pdf with parameter λ_j.

a) Assume $K = 1$, namely, $f(x; \lambda) = \lambda \exp(-\lambda x)$, for $x \geq 0$. Given a set of observations $\{x_i\}_{i=1}^{n}$ drawn i.i.d from $f(x; \lambda)$, find the maximum likelihood estimator (MLE) of λ.
Solution:
The joint p.d.f of the observations is $\prod_{i=1}^{m} f(x_i; \lambda)$. Furthermore, since log is a monotonic increasing function on $(0, \infty)$, we can take $\hat{\lambda}$ that maximizes $\sum_{i=1}^{m} \log f(x; \lambda)$. Therefore:

$$\frac{\partial}{\partial \lambda} \sum_i \log \lambda \exp^{-\lambda x_i} = \sum_i (\frac{1}{\lambda} - x_i) \tag{9}$$

$$= \frac{m}{\lambda} - \sum_{i=1}^{m} x_i \tag{10}$$

$$= 0 \tag{11}$$

Therefore,

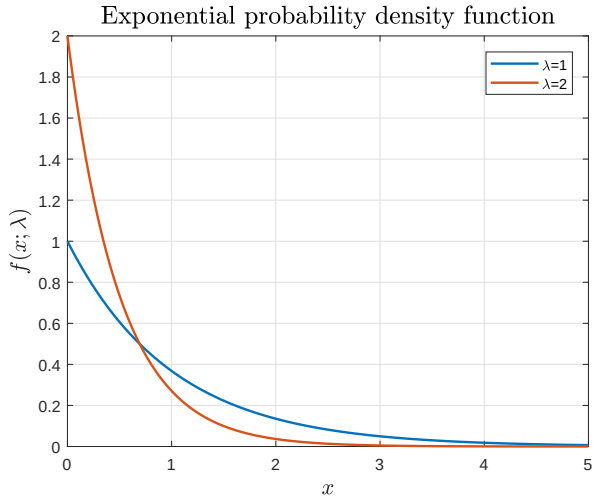
$$\hat{\lambda}^{(MLE)} = \frac{m}{\sum_{i=1}^{m} x_i}$$

and the second derivative is

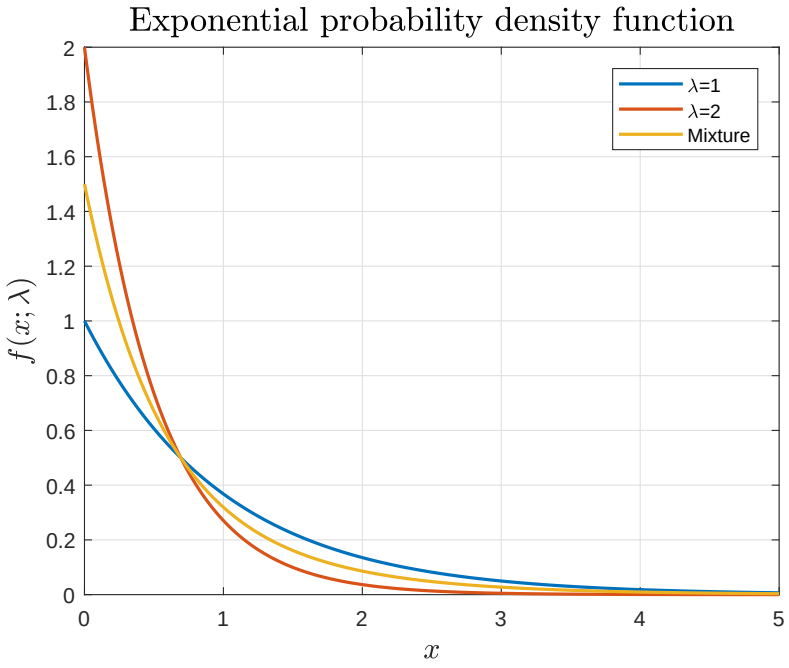
$$\frac{\partial^2}{\partial \lambda^2} \sum_i \log \lambda \exp^{-\lambda x_i} = -\frac{m}{\lambda^2}$$

which implies that this is a maxima.

b) The exponential distribution is used to model a continuous time interval between two Poisson events. The average rate of the Poisson events is denoted by λ, and the exponential distribution determines how much time will pass until the arrival of a new event.



The figure above has the pdf $f(x|z = j; \lambda_j)$ for $\lambda_j = 1$ and for $\lambda_j = 2$. Draw the mixture of these two distributions, where $P_Z(1) = P_Z(2) = 0.5$.
Solution:
The p.d.f is the average between the two given density functions.



c) Recall from class that the EM algorithm aims to maximize the log-likelihood function: The E-step at iteration (t):

$$w(i,j) \triangleq Q_i^{(t)}(j) = P_{Z|X;\theta}\left(Z = j | X = x_i; \theta^{(t-1)}\right) \tag{12}$$

where $\theta^{(t-1)}$ are the parameters of the exponential mixture model at iteration $t-1$. The M-step at iteration (t):

$$\theta^{(t)} = \arg\max_{\theta} \sum_{i=1}^{n} \mathbb{E}_{Q_{Z_i}^{(t)}}[\log(P_{X,Z}(x_i, z_i; \theta))] \tag{13}$$

$$= \arg\max_{\theta} \sum_{i=1}^{n} \sum_{j=1}^{K} Q_i^{(t)}(j) \log P(x_i, z = j; \theta) \tag{14}$$

You are given training samples $\{x_i\}_{i=1}^n$ from a mixture of exponential distributions. Write explicitly the estimation formulas of the EM algorithm for exponential mixtures, using the training samples.

Solution:

We identify θ as the parameters of the distribution, λ and $\{\phi\}_j = 1^K$. Denote

$$\phi(j) = \hat{P}_Z(j)$$

For the expectation step, use the given parameters ϕ and λ to calculate

$$Q_i(j) = \frac{\phi(j)\lambda_j \exp^{-\lambda_j x_i}}{\sum_{l=1}^{K} \phi(l)\lambda_l \exp^{-\lambda_l x_i}}. \tag{15}$$

Then, in the maximization step, we consider $Q_i(j)$ as constant and maximize over ϕ_j and λ. Consider the following derivation

$$\hat{\theta} \overset{(a)}{=} \arg\max_{\phi_j, \lambda_j} \sum_{i=1}^{m} \sum_{j=1}^{K} Q_i(j)(\log \phi_j \lambda_j \exp^{-\lambda_j x_i}) \tag{16}$$

$$= \arg\max_{\phi_j, \lambda_j} \sum_{i=1}^{m} \sum_{j=1}^{K} Q_i(j)(\log \phi_j + \log \lambda_j - \lambda_j x_i) \tag{17}$$

where (a) follows from (14). To maximize w.r.t. $\phi(j)$, we need to maximize $\sum_{i=1}^{m} \sum_{j=1}^{K} Q_i(j) \log \phi_j$ under the constraint that $\sum_{j=1}^{K} \phi_j = 1$. We construct the Lagrangian

$$\mathcal{L}(\phi) = \sum_{i=1}^{m} \sum_{j=1}^{K} Q_i(j) \log \phi_j + \alpha(\sum_{j=1}^{K} \phi_j - 1) \tag{18}$$

Taking derivating and finding ϕ_j results in

$$\phi_j = -\frac{\sum_{i=1}^{m} Q_i(j)}{\alpha} \tag{19}$$

Using the constraint on the sum, we can easily find that

$$-\alpha\phi_j = \sum_{i=1}^{m} Q_i(j) \tag{20}$$

$$\Rightarrow -\alpha = \sum_{i=1}^{m}\sum_{j=1}^{K} Q_i(j) \tag{21}$$

$$= \sum_{i=1}^{m} 1 = m \tag{22}$$

Therefore,

$$\phi(j) = \frac{1}{m}\sum_{i} Q_i(j)$$

For the parameter λ, note that

$$\frac{\partial}{\partial\lambda_j}\sum_{i}\sum_{j=1}^{K} Q_i(j)(\log\lambda_j - \lambda_j x_i) = \sum_{i} Q_i(j)(\frac{1}{\lambda_j} - x_i) \tag{23}$$

therefore, by taking maximum over λ we have

$$\lambda_j = \frac{\sum_i Q_i(j)}{\sum_i Q_i(j)x_i} \tag{24}$$

for $j = 1, \dots K$.

d) Assume that you have 10 agents in a company that provide services via telephone. Whenever an agent completes a service call, he fills a service form in the company's database and the duration of the call is recorded. After each call, the agent immediately answers a new call. The duration of a service call may depend on the costumer's need, the agent and, the day of the week. Hence, for each agent, these calls are modeled by a mixture of exponential distributions.

You are given a train data $\mathcal{S} = \{(x_i, y_i)\}_{i=1}^{n}$ where x_i is the duration of the i-th call and $y_i \in \{1, \dots, 10\}$ is the agent number. You are also given a test data $\mathcal{T} = \{x_j\}_{j=1}^{L}$.

Write a pseudo code that fits a mixture of K exponential distributions for each agent in the company, using T iterations of expectation maximization. Then, it assigns an agent number for each one of the test samples.

Solution:

Let K denote the number of exponential distributions in a mixture, T the number of iterations in the EM algorithm, X a set of samples and Y a set of labels (corresponding to the samples). Then the algorithm can be summarized in the following pseudo-code (next page):

Algorithm 1 Exponential mixtures for multiclass

function FITMIXTURE(X,K,T)
  $\hat{\lambda}, \phi \leftarrow$ initial values
  for $t = 1, \ldots, T$ **do**
    for $j = 1, \ldots, K$ **do**
      $w(j,i) = \frac{\phi(j)\lambda_j \exp^{-\lambda_j x_i}}{\sum_{l=1}^{K} \phi(l)\lambda_l \exp^{-\lambda_l x_i}}$ for each $i \in \{1, \ldots, m\}$
    for $j = 1, \ldots, K$ **do**
      $\phi(j) \leftarrow \frac{1}{m} \sum_i w(j,i)$
      $\lambda_j \leftarrow \frac{\sum_{i=1} w(j,i)}{\sum_i w(j,i) x_i}$
  return λ, ϕ
function TRAIN(X,Y,K,T)
  for $l = 1, \ldots, 10$ **do**
    $x^{(l)} \leftarrow$ all x_i for i s.t. $y_i = l$
    $(\lambda^{(l)}, \phi^{(l)}) =$FitMixture $(x^{(l)}, K, T)$
  $\theta \leftarrow \{(\lambda^{(l)}, \phi^{(l)})\}_{l=1}^{10}$ **return** θ
function TEST(X, θ)
  $\{(\lambda^{(l)}, \phi^{(l)})\}_{l=1}^{10} \leftarrow \theta$
  for x in X **do**
    for $l = 1, \ldots, 10$ **do**
      $P(x|l) = \sum_{j=1}^{K} \phi^{(l)}(j) \lambda_j^{(l)} \exp^{-\lambda_j^{(l)} x}$
    $\hat{y}(x) = \arg\max_l P(x|l)$
procedure RUNALGORITHM($X_{train}, Y_{train}, X_{test}, K, T$)
  $\theta \leftarrow$ Train($X_{train}, Y_{train}, K, T$)
  $\hat{Y}_{test} \leftarrow$ Test(X_{test}, θ)

Remark: The decision criteria here is according to maximum likelihood.

Good Luck!