EE566 ADAPTATION AND LEARNING **Instructor: Ali H. Sayed**
Homework Assignment #3 **Due: April 28, 2025**

---

1) (**Chapter 28**) Consider an unknown $M-$dimensional vector $\boldsymbol{x} \sim f_{\boldsymbol{x}}(x)$ that we wish to infer from a collection of N independent and identically distributed observations $\{\boldsymbol{y}_n,\ n = 1, 2 \ldots, N\}$. The conditional distribution of each observation $\boldsymbol{y}_n$ given $\boldsymbol{x}$ is uniform across all observations and denoted by $\boldsymbol{y}_n|\boldsymbol{x} \sim f_{\boldsymbol{y}|\boldsymbol{x}}(y_n|x)$. Show that the MAP estimator for $\boldsymbol{x}$ given the N observations $\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N\}$ can be found by solving

$$\widehat{x}_{\text{MAP}} \ = \ \underset{x\in\mathbb{R}^M}{\text{argmin}} \left\{ -\frac{1}{N}\sum_{n=1}^{N} \ln f_{\boldsymbol{y}|\boldsymbol{x}}(y_n|x) - \frac{1}{N}\ln f_{\boldsymbol{x}}(x) \right\}$$

Solution: Using the iid assumption we have that

$$f_{\{\boldsymbol{y}_n\}_{n=1}^N|\boldsymbol{x}}\Big(\{y_n\}_{n=1}^N|x\Big) = \prod_{n=1}^{N} f_{\boldsymbol{y}_n|\boldsymbol{x}}(y_n|x) = \prod_{n=1}^{N} f_{\boldsymbol{y}|\boldsymbol{x}}(y_n|x)$$

Using Bayes' rule we have

$$\begin{aligned}
\widehat{x}_{\text{MAP}} &= \underset{x\in\mathbb{R}^M}{\text{argmax}}\ f_{\boldsymbol{x}|\{\boldsymbol{y}_n\}_{n=1}^N}\Big(x|\{y_n\}_{n=1}^N\Big) \\
&= \underset{x\in\mathbb{R}^M}{\text{argmax}} \left\{ \frac{f_{\{\boldsymbol{y}_n\}_{n=1}^N|\boldsymbol{x}}\Big(\{y_n\}_{n=1}^N|x\Big) \times f_{\boldsymbol{x}}(x)}{f_{\{\boldsymbol{y}_n\}_{n=1}^N}\Big(\{y_n\}_{n=1}^N\Big)} \right\} \\
&= \underset{x\in\mathbb{R}^M}{\text{argmax}} \left\{ f_{\{\boldsymbol{y}_n\}_{n=1}^N|\boldsymbol{x}}\Big(\{y_n\}_{n=1}^N|x\Big) \times f_{\boldsymbol{x}}(x) \right\} \\
&= \underset{x\in\mathbb{R}^M}{\text{argmax}} \left\{ \left(\prod_{n=1}^{N} f_{\boldsymbol{y}|\boldsymbol{x}}(y_n|x) \right) \times f_{\boldsymbol{x}}(x) \right\}
\end{aligned}$$

Since the logarithm function is monotonically increasing, we can also write

$$\begin{aligned}
\widehat{x}_{\text{MAP}} &= \underset{x\in\mathbb{R}^M}{\text{argmax}}\ \ln \left[\left(\prod_{n=1}^{N} f_{\boldsymbol{y}|\boldsymbol{x}}(y_n|x) \right) \times f_{\boldsymbol{x}}(x) \right] \\
&= \underset{x\in\mathbb{R}^M}{\text{argmax}} \left\{ \sum_{n=1}^{N} \ln f_{\boldsymbol{y}|\boldsymbol{x}}(y_n|x) + \ln f_{\boldsymbol{x}}(x) \right\} \\
&= \underset{x\in\mathbb{R}^M}{\text{argmin}} \left\{ -\sum_{n=1}^{N} \ln f_{\boldsymbol{y}|\boldsymbol{x}}(y_n|x) - \ln f_{\boldsymbol{x}}(x) \right\}
\end{aligned}$$

Normalizing by the sample size we conclude that

$$\widehat{x}_{\text{MAP}} \ = \ \underset{x\in\mathbb{R}^M}{\text{argmin}} \left\{ -\frac{1}{N}\sum_{n=1}^{N} \ln f_{\boldsymbol{y}|\boldsymbol{x}}(y_n|x) - \frac{1}{N}\ln f_{\boldsymbol{x}}(x) \right\}$$

□

2) (**Chapter 29**) We examined the correlation coefficient ρ_{xy} between two scalar random variables $\{\boldsymbol{x}, \boldsymbol{y}\}$ in Prob. 3.13 in the text. The correlation between these random variables given a third scalar variable $\boldsymbol{z}$ is defined as follows. Let $\widetilde{\boldsymbol{x}} = \boldsymbol{x} - \widehat{\boldsymbol{x}}$ denote the residual that remains from estimating $\boldsymbol{x}$ from $\boldsymbol{z}$ using a linear regression model, say, $\widehat{\boldsymbol{x}} = \alpha\boldsymbol{z} + \beta$. Likewise, let $\widetilde{\boldsymbol{y}} = \boldsymbol{y} - \widehat{\boldsymbol{y}}$ denote the residual

that remains from estimating $\boldsymbol{y}$ from $\boldsymbol{z}$ using a linear regression model as well, say, $\widehat{\boldsymbol{y}} = a\boldsymbol{z} + b$. Then, the (conditional) correlation is defined by

$$\rho_{xy|z} \triangleq \rho_{\tilde{x},\tilde{y}}$$

That is, it is equal to the standard correlation between the residual variables after the effect of $\boldsymbol{z}$ has been removed using linear regression. Show that if $\boldsymbol{x}$ and $\boldsymbol{y}$ are conditionally independent of each other given $\boldsymbol{z}$, then $\rho_{xy|z} = 0$. Is the converse true?

Solution: We know that

$$\begin{aligned}\widehat{\boldsymbol{x}} - \bar{x} &= \frac{\sigma_{xz}}{\sigma_z^2}(z - \bar{z}) \\ \widehat{\boldsymbol{y}} - \bar{y} &= \frac{\sigma_{yz}}{\sigma_z^2}(z - \bar{z})\end{aligned}$$

so that the scalars (α, β) and (a, b) are given by

$$\begin{aligned}\alpha &= \frac{\sigma_{xz}}{\sigma_z^2}, \quad \beta = \bar{x} - \frac{\sigma_{xz}}{\sigma_z^2}\bar{z} \\ a &= \frac{\sigma_{yz}}{\sigma_z^2}, \quad b = \bar{y} - \frac{\sigma_{yz}}{\sigma_z^2}\bar{z}\end{aligned}$$

Now given $\boldsymbol{z}$, the regression error $\widetilde{\boldsymbol{x}}$ is solely dependent on $\boldsymbol{x}$ while $\widetilde{\boldsymbol{y}}$ is solely dependent on $\boldsymbol{y}$ since

$$\begin{aligned}\widetilde{\boldsymbol{x}} &= \boldsymbol{x} - \alpha\boldsymbol{z} - \beta \\ \widetilde{\boldsymbol{y}} &= \boldsymbol{y} - a\boldsymbol{z} - b\end{aligned}$$

It follows that $\widetilde{\boldsymbol{x}}$ and $\widetilde{\boldsymbol{y}}$ are also conditionally independent of each other given $\boldsymbol{z}$, from which we conclude that $\rho_{\tilde{x},\tilde{y}} = 0$ and, therefore, $\rho_{xy|z} = 0$.

The converse is not true. If $\rho_{xy|z} = 0$ then $\rho_{\tilde{x},\tilde{y}} = 0$. However, we know that this latter condition does not imply that $\widetilde{\boldsymbol{x}}$ and $\widetilde{\boldsymbol{y}}$ are independent of each other. Now given by $\boldsymbol{x}$ and $\boldsymbol{y}$ can be expressed in terms of these errors as

$$\begin{aligned}\boldsymbol{x} &= \widetilde{\boldsymbol{x}} + \alpha\boldsymbol{z} + \beta \\ \boldsymbol{y} &= \widetilde{\boldsymbol{y}} + a\boldsymbol{z} + b\end{aligned}$$

we conclude that $\boldsymbol{x}$ and $\boldsymbol{y}$ are not necessarily independent of each other given $\boldsymbol{z}$.

□

3) (**Chapter 31**) Consider N i.i.d. scalar random variables $\{\boldsymbol{y}_n\}$ distributed according to $\boldsymbol{y} \sim f_{\boldsymbol{y}}(y) = \frac{1}{2\sigma}e^{-|y|/\sigma}$. Show that the ML estimate for σ is given by

$$\widehat{\sigma} = \frac{1}{N}\sum_{n=1}^{N}|y_n|$$

Verify that this estimator is unbiased and asymptotically efficient.

Solution: The log-likelihood function is given by

$$\begin{aligned}
\ell(\{y_n\};\sigma) &= \ln\left[\prod_{n=1}^{N}\frac{1}{2\sigma}e^{-|y_n|/\sigma}\right] \\
&= \ln\left[\left(\frac{1}{2\sigma}\right)^N \exp\left\{-\sum_{n=1}^{N}\frac{|y_n|}{\sigma}\right\}\right] \\
&= -N\ln(2\sigma) - \sum_{n=1}^{N}\frac{|y_n|}{\sigma}
\end{aligned}$$

Differentiating relative to σ leads to

$$-N\frac{1}{\widehat{\sigma}} + \sum_{n=1}^{N}\frac{|y_n|}{\widehat{\sigma}^2} = 0$$

leading to

$$\widehat{\sigma} = \frac{1}{N}\sum_{n=1}^{N}|y_n|$$

Next note that

$$\begin{aligned}
\mathbb{E}\,|\boldsymbol{y}| &= \int_{-\infty}^{\infty}|y|f_{\boldsymbol{y}}(y)dy \\
&= \frac{1}{2\sigma}\int_{-\infty}^{\infty}|y|e^{-|y|/\sigma}dy \\
&= \frac{1}{2\sigma}\left[\int_{-\infty}^{0}-ye^{y/\sigma}dy + \int_{0}^{\infty}ye^{-y/\sigma}dy\right] \\
&= \frac{1}{\sigma}\int_{0}^{\infty}ye^{-y/\sigma}dy \\
&= \int_{0}^{\infty}\frac{y}{\sigma}e^{-y/\sigma}dy \\
&= \sigma\int_{0}^{\infty}y'e^{-y'}dy', \quad y' = y/\sigma \\
&= \sigma\times\Gamma(2) \\
&= \sigma
\end{aligned}$$

where we used the value of the gamma function at 2, which is equal to 1. It follows from the observation that $\Gamma(n+1) = n!$ for integer values of n. Therefore, the proposed estimator is unbiased since

$$\mathbb{E}\,\widehat{\boldsymbol{\sigma}} = \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\,|\boldsymbol{y}_n| = \sigma$$

Similarly, note that

$$\begin{aligned}
\mathbb{E}\,|\boldsymbol{y}|^2 &= \int_{-\infty}^{\infty} |y|^2 f_{\boldsymbol{y}}(y)dy \\
&= \frac{1}{2\sigma}\int_{-\infty}^{\infty} |y|^2 e^{-|y|/\sigma}dy \\
&= \frac{1}{2\sigma}\left[\int_{-\infty}^{0} y^2 e^{y/\sigma}dy + \int_{0}^{\infty} y^2 e^{-y/\sigma}dy\right] \\
&= \frac{1}{\sigma}\int_{0}^{\infty} y^2 e^{-y/\sigma}dy \\
&= \sigma\int_{0}^{\infty} \frac{y}{\sigma^2} e^{-y/\sigma}dy \\
&= \sigma^2\int_{0}^{\infty} z^2 e^{-z}dz, \quad z = y/\sigma \\
&= \sigma^2 \times \Gamma(3) \\
&= \sigma^2 \times 2! \\
&= 2\sigma^2
\end{aligned}$$

Consequently, the mean-square-error of $\widehat{\boldsymbol{\sigma}}$ is given by

$$\begin{aligned}
\text{MSE}(\widehat{\boldsymbol{\sigma}}) &= \mathbb{E}\,(\widehat{\boldsymbol{\sigma}} - \sigma)^2 \\
&= \text{var}(\widehat{\boldsymbol{\sigma}}) \\
&= \frac{1}{N^2}\sum_{n=1}^{n} \text{var}(|\boldsymbol{y}_n|) \\
&= \frac{1}{N^2}\sum_{n=1}^{n}\Big(\mathbb{E}\,|\boldsymbol{y}_n|^2 - [\mathbb{E}\,|\boldsymbol{y}_n|]^2\Big) \\
&= \frac{1}{N^2}\sum_{n=1}^{n}\Big(2\sigma^2 - \sigma^2\Big) \\
&= \frac{\sigma^2}{N}
\end{aligned}$$

Observe that the variance tends to 0 as $N \to \infty$ and, therefore, the ML estimator is asymptotically efficient.

□

4) (**Chapter 50**) Consider two $N \times M$ matrices D and H. Introduce the SVD representation

$$D^{\mathsf{T}}H = U\Sigma V^{\mathsf{T}}$$

where U and V are $M \times M$ orthogonal, and Σ is diagonal. Show that the $M \times M$ orthogonal matrix $\Theta^\star$ that solves

$$\Theta^\star = \underset{\Theta\Theta^{\mathsf{T}}=I_M}{\text{argmin}} \|D - H\Theta\|_{\text{F}}$$

is given by $\Theta^\star = VU^{\mathsf{T}}$.

Solution: Using

$$\|A\|_{\text{F}}^2 = \text{Tr}(A^{\mathsf{T}}A)$$

we write

$$\begin{aligned}
\|D-H\Theta\|_{\mathrm{F}}^2 &= \mathrm{Tr}\Big((D-H\Theta)^{\mathsf{T}}(D-H\Theta)\Big) \\
&= \mathrm{Tr}(D^{\mathsf{T}}D)+\mathrm{Tr}(\Theta^{\mathsf{T}}H^{\mathsf{T}}H\Theta)-2\mathrm{Tr}(D^{\mathsf{T}}H\Theta) \\
&= \mathrm{Tr}(D^{\mathsf{T}}D)+\mathrm{Tr}(\Theta\Theta^{\mathsf{T}}H^{\mathsf{T}}H)-2\mathrm{Tr}(D^{\mathsf{T}}H\Theta) \\
&= \mathrm{Tr}(D^{\mathsf{T}}D)+\mathrm{Tr}(H^{\mathsf{T}}H)-2\mathrm{Tr}(D^{\mathsf{T}}H\Theta) \quad \text{(since } \Theta \text{ is orthogonal)} \\
&= \|D\|_{\mathrm{F}}^2+\|H\|_{\mathrm{F}}^2-2\mathrm{Tr}(D^{\mathsf{T}}H\Theta)
\end{aligned}$$

where we used the fact that $\mathrm{Tr}(AB)=\mathrm{Tr}(BA)$. Ignoring the terms that do not depend on Θ, the optimization problem reduces to solving

$$\Theta^{\star}=\underset{\Theta\Theta^{\mathsf{T}}=I_M}{\mathrm{argmax}}\ \mathrm{Tr}(D^{\mathsf{T}}H\Theta)$$

The matrix $D^{\mathsf{T}}H$ has dimensions $M\times M$. We introduce its SVD:

$$D^{\mathsf{T}}H=U\Sigma V^{\mathsf{T}}$$

where U and V are $M\times M$ orthogonal, and Σ is diagonal with entries $\sigma_1\geq\sigma_2\geq\ldots\geq\sigma_M$. Thus,

$$\begin{aligned}
\mathrm{Tr}(D^{\mathsf{T}}H\Theta) &= \mathrm{Tr}(U\Sigma V^{\mathsf{T}}\Theta) \\
&= \mathrm{Tr}(\Sigma V^{\mathsf{T}}\Theta U) \\
&\overset{\Delta}{=} \mathrm{Tr}(\Sigma X) \\
&= \sum_{m=1}^{M}\sigma_m X_{mm}
\end{aligned}$$

where we introduced the matrix $X=V^{\mathsf{T}}\Theta U$ and denoted its diagonal entries by $\{X_{mm}\}$. It is obvious that X is an orthogonal matrix as well since

$$XX^{\mathsf{T}}V^{\mathsf{T}}\Theta UU^{\mathsf{T}}\Theta^{\mathsf{T}}V=I_M$$

and, moreover, X and Θ define each other uniquely since

$$\Theta=VXU^{\mathsf{T}}$$

Now recall from Prob. **??** that $X_{mm}\leq 1$. Therefore, it holds that

$$\sum_{m=1}^{M}\sigma_m X_{mm}\leq\sum_{m=1}^{M}\sigma_m$$

Equality holds if, and only if, $X_{mm}=1$ for all m. Since X is unitary and its rows have unit norm, this is only possible when X is the identity matrix. Therefore, the optimal choices for X is $X=I_M$ from which we conclude that

$$\Theta^{\star}=VU^{\mathsf{T}}$$

□

5) (**Chapter 51**) Consider a collection of scalar measurements $\{x(n)\}_{n=0}^{N-1}$ sampled independently from the Gaussian distribution $\mathcal{N}(\theta^o,1)$ with unknown mean θ^o. We wish to estimate θ^o by seeking the solution to a regularized least-squares problem with a regularization term that penalizes proximity to a prior estimate denoted by θ_1, namely,

$$\theta^{\star}\ \overset{\Delta}{=}\ \arg\min_{\theta\in\mathbb{R}}\left\{\frac{1}{N}\sum_{n=0}^{N-1}(x(n)-\theta)^2+\rho(\theta-\theta_1)^2\right\},\quad \rho>0$$

(a) Let $\bar{\theta} = \frac{1}{N}\sum_{n=0}^{N-1} x(n)$. Verify that

$$\theta^\star = \frac{1}{1+\rho}\bar{\theta} \; + \; \frac{\rho}{1+\rho}\theta_1$$

(b) Show that

$$\mathbb{E}\,(\boldsymbol{\theta}^\star - \theta^o)^2 = \Big(\frac{1}{1+\rho}\Big)^2 \frac{1}{N} + \Big(\frac{\rho}{1+\rho}\Big)^2 \mathbb{E}\,(\boldsymbol{\theta}_1 - \theta^o)^2$$

(c) Verify that the optimal choice for the regularization parameter is

$$\rho = \frac{1/N}{\mathbb{E}\,(\boldsymbol{\theta}_1 - \theta^o)^2}$$

for which the mean-square error expression in part (b) simplifies to

$$\mathbb{E}\,(\boldsymbol{\theta}^\star - \theta^o)^2 \approx \min\left\{\frac{1}{N},\; \mathbb{E}\,(\boldsymbol{\theta}_1 - \theta^o)^2\right\}$$

(d) Conclude that a large ρ is needed when the prior estimate has a small mean-square-error, while a small ρ is needed otherwise. In other words, conclude that the size of the regularization parameter depends on the quality of the initial estimate, θ_1.

Solution:

(a) Differentiating relative to θ and setting the gradient vector to 0 gives

$$-\frac{2}{N}(x(n) - \theta^\star) + 2\rho(\theta^\star - \theta_1) = 0$$

Solving for $\theta^\star$ leads to the desired expression.

(b) The MSE is given by

$$\mathbb{E}\,(\boldsymbol{\theta}^\star - \theta^o)^2 \;\; = \;\; \Big(\frac{1}{1+\rho}\Big)^2 \mathbb{E}\,(\boldsymbol{\theta}^\star - \theta^o)^2 \; + \; \Big(\frac{\rho}{1+\rho}\Big)^2 \mathbb{E}\,(\boldsymbol{\theta}_1 - \theta^o)^2$$

where

$$\begin{aligned}
\mathbb{E}\,(\boldsymbol{\theta}^\star - \theta^o)^2 &= \mathbb{E}\Big|\frac{1}{N}\sum_{n=0}^{N-1}\boldsymbol{x}(n) - \theta^o\Big|^2 \\
&= \frac{1}{N^2}\mathbb{E}\Big|\sum_{n=0}^{N-1}(\boldsymbol{x}(n) - \theta^o)\Big|^2 \\
&= \frac{1}{N^2} \times N \\
&= 1/N
\end{aligned}$$

and the desired result follows.

(c) Differentiating the MSE expression over ρ gives

$$\frac{-2}{(1+\rho)^3}\frac{1}{N} + \frac{2\rho(1+\rho) - 2\rho^2}{(1+\rho)^3}\mathbb{E}\,(\boldsymbol{\theta}_1 - \theta^o)^2 = 0$$

from which the desired result follows:

$$\rho = \frac{1/N}{\mathbb{E}\,(\boldsymbol{\theta}_1 - \theta^o)^2}$$

Substituting into the MSE expression gives

$$\mathbb{E}\,(\boldsymbol{\theta}^{\star}-\theta^{o})^2 \;=\; \frac{\frac{1}{N}\times\mathbb{E}\,(\boldsymbol{\theta}_1-\theta^{o})^2}{\frac{1}{N}+\mathbb{E}\,(\boldsymbol{\theta}_1-\theta^{o})^2}$$

When $1/N$ is small we can approximate the above expression by

$$\mathbb{E}\,(\boldsymbol{\theta}^{\star}-\theta^{o})^2 \;=\; \frac{\mathbb{E}\,(\boldsymbol{\theta}_1-\theta^{o})^2}{1+\mathbb{E}\,(\boldsymbol{\theta}_1-\theta^{o})^2/(1/N)} \approx 1/N$$

On the other hand, when $1/N$ is large we can write

$$\mathbb{E}\,(\boldsymbol{\theta}^{\star}-\theta^{o})^2 \;=\; \frac{1/N}{(1/N)/\mathbb{E}\,(\boldsymbol{\theta}_1-\theta^{o})^2+1} \approx \mathbb{E}\,(\boldsymbol{\theta}_1-\theta^{o})^2$$

In other words, we can approximate the MSE expression by

$$\mathbb{E}\,(\boldsymbol{\theta}^{\star}-\theta^{o})^2 \approx \min\left\{\frac{1}{N},\; \mathbb{E}\,(\boldsymbol{\theta}_1-\theta^{o})^2\right\}$$

(d) The value of the optimal ρ in part (c) has $\mathbb{E}\,(\boldsymbol{\theta}_1-\theta^{o})^2$ in the denominator, and the result follows. This means that the choice of the regularization parameter depends on the quality of the prior estimate, θ_1.

□

6) (**Chapter 52**) Consider feature vectors $h \in \mathbb{R}^M$ and a collection of K classifiers (or experts) denoted by $\{E_1(h), E_2(h), \ldots, E_K(h)\}$. Each feature vector $h \in \mathbb{R}^M$ can belong to one of R classes denoted by $r = 1, 2, \ldots, R$. Introduce an $R \times K$ matrix $\mathcal{E}$, which summarizes the opinion of the experts about the class of h. Each row of index r corresponds to one of the labels, and each column of index c corresponds to one of the classifiers or experts. The entry $\mathcal{E}_{rc}$ indicates the level of confidence that expert E_c has about feature h belonging to class r. For illustration purposes, we exhibit a matrix $\mathcal{E}$ corresponding to $R = 4$ labels and $K = 5$ experts:

$$\mathcal{E} = \left[\begin{array}{c|ccccc} \text{labels} & E_1 & E_2 & E_3 & E_4 & E_5 \\ \hline r=1 & \mathcal{E}_{11} & \mathcal{E}_{12} & \mathcal{E}_{13} & \mathcal{E}_{14} & \mathcal{E}_{15} \\ r=2 & \mathcal{E}_{21} & \mathcal{E}_{22} & \mathcal{E}_{23} & \mathcal{E}_{24} & \mathcal{E}_{25} \\ r=3 & \mathcal{E}_{31} & \mathcal{E}_{32} & \mathcal{E}_{33} & \mathcal{E}_{34} & \mathcal{E}_{35} \\ r=4 & \mathcal{E}_{41} & \mathcal{E}_{42} & \mathcal{E}_{43} & \mathcal{E}_{44} & \mathcal{E}_{45} \end{array}\right]$$

(a) Explain how the $K-$NN classifier with R labels is a special case of this construction. What would the classifiers in this case be and how would the entries of $\mathcal{E}$ be chosen?

(b) Explain how the weighted $K-$NN classifier with R labels from Example 52.2 is a special case of this construction. What would the classifiers in this case be and how would the entries of $\mathcal{E}$ be chosen?

(c) Verify that in both cases the ultimate label for h is chosen in terms of the $\infty-$norm of matrix $\mathcal{E}$ as follows:

$$r^{\star} = \underset{1\le r\le R}{\text{argmax}}\; \|\mathcal{E}\|_{\infty}$$

Solution:

(a) For the $K-$NN classifier, each of the K neighbors defines one of the experts. Moreover, the value of $\mathcal{E}_{rc}$ will be 1 if the cth neighbor belongs to class r and 0 otherwise.

(b) For the weighted $K-$NN classifier with R labels, each of the K neighbors defines one of the experts. Moreover, the value of $\mathcal{E}_{rc}$ will be w_c defined by (52.22c), while all other values on the cth column of $\mathcal{E}$ will be 0.

(c) For the $K-$NN classifier, if we add the entries on each row r we find the number of votes that label r receives from the K neighbors. The row with the highest number of votes determines $r^\star$. Likewise, for the weighted $K-$NN classifier, if we add the entries on each row r we find the overall weight that label r receives from the K neighbors. The row with the highest aggregate weight determines $r^\star$. Both scenarios amount to determining $r^\star$ by determining the row of $\mathcal{E}$ with the highest sum of its entries. This corresponds to the $\infty-$norm of $\mathcal{E}$.

□