**Solution Set #2**

*Warning: Homeworks will not be graded if submitted after the deadline. For all problems, show detailed reasoning.*

**1. (Bias-variance tradeoff)** *Consider a problem of estimating $\theta$ from a set of samples $\{x^{(1)}, \ldots, x^{(m)}\}$ generated i.i.d. according to Bernoulii($\theta$), where $0 \leq \theta \leq 1$. As shown in (5.28) in page 125 of the textbook, the sample-mean estimator $\hat{\theta}_m = \frac{1}{m}\sum_{i=1}^{m} x^{(i)}$ given in (5.22) is unbiased, i.e.,* $\text{bias}(\hat{\theta}_m) = \mathbb{E}[\hat{\theta}_m] - \theta = 0$. *The variance of the sample-mean estimator is $\frac{1}{m}\theta(1-\theta)$ as shown in (5.52) in page 129.*

(a) *Let's consider a trivial estimator $\tilde{\theta}_m = 1$, i.e., it always says 1 regardless of the samples it observes. Calculate the bias and variance of the trivial estimator.*

**Solution**

The bias of the trivial estimator, $\tilde{\theta}_m = 1$, is

$$\text{bias}(\tilde{\theta}_m) = \mathbb{E}[\tilde{\theta}_m] - \theta = 1 - \theta$$

Therefore, this is an unbiased estimator only for $\theta = 1$. Its variance is

$$\text{var}(\tilde{\theta}_m) = \mathbb{E}[(\tilde{\theta}_m - 1)^2] = 0.$$

(b) *Assume $m = 1$. Can you think of an estimator that minimizes the worst-case MSE over all $\theta$? Namely, find an estimator $\bar{\theta}_1$ based on $x^{(1)}$ that minimizes the following.*

$$\max_{0 \leq \theta \leq 1} \mathbb{E}[(\bar{\theta}_1 - \theta)^2]$$

*What is the minimum worse-case MSE that the estimator achieves?*

**Solution**

The estimator $\bar{\theta}_1$ is a function of $x^{(1)}$. Thus we can write $\bar{\theta}_1$ as $\bar{\theta}_1(x^{(1)})$. Because of $x^{(1)} \sim \text{Bernoulli}(\theta)$, there are two values for the estimator, $\bar{\theta}_1(0)$ and $\bar{\theta}_1(1)$. Let $\theta_0$ and $\theta_1$ denote $\bar{\theta}_1(0)$ and $\bar{\theta}_1(1)$, respectively. Then the MSE of the estimator is

$$\begin{aligned}
\mathbb{E}[(\bar{\theta}_1 - \theta)^2] &= \sum_{x \in \mathcal{X}} (\bar{\theta}_1(x) - \theta)^2 p(x) \\
&= (\theta_0 - \theta)^2(1-\theta) + (\theta_1 - \theta)^2\theta \\
&= (1 + 2\theta_0 - 2\theta_1)\theta^2 + (\theta_1^2 - \theta_0^2 - 2\theta_0)\theta + \theta_0^2 \\
&= g(\theta, \theta_0, \theta_1)
\end{aligned}$$

Now, we want to find $\theta_0^*$ and $\theta_1^*$ that minimize $\max_{0 \leq \theta \leq 1} g(\theta, \theta_0, \theta_1)$, i.e.

$$(\theta_0^*, \theta_1^*) = \underset{(\theta_0, \theta_1)}{\text{argmin}} \max_{0 \leq \theta \leq 1} g(\theta, \theta_0, \theta_1)$$

First, we can see that

$$\begin{aligned}
\min_{0\le\theta_0,\theta_1\le 1}\max_{0\le\theta\le 1} g(\theta,\theta_0,\theta_1) &\overset{(a)}{\ge} \min_{0\le\theta_0,\theta_1\le 1}\max_{\theta\in\{0,\frac{1}{2},1\}} g(\theta,\theta_0,\theta_1)\\
&= \min_{0\le\theta_0,\theta_1\le 1}\max\left\{\theta_0^2, \frac{(\theta_0-\frac{1}{2})^2+(\theta_1-\frac{1}{2})^2}{2}, (1-\theta_1)^2\right\}\\
&\overset{(b)}{=} \min_{0\le u,v\le 1}\max\left\{u^2, \frac{(u-\frac{1}{2})^2+(v-\frac{1}{2})^2}{2}, v^2\right\}\\
&\overset{(c)}{=} \min_{0\le u\le v\le 1}\max\left\{\frac{(u-\frac{1}{2})^2+(v-\frac{1}{2})^2}{2}, v^2\right\}\\
&\overset{(d)}{\ge} \frac{1}{16},
\end{aligned}$$

where (a) follows since a function maximized over a larger set must be bigger than or equal to that maximized over a smaller set, and in this case, $\{0,\frac{1}{2},1\}\subseteq[0,1]$, (b) follows by defining $u=\theta_0$ and $v=1-\theta_1$, (c) follows since we only need to consider $v\ge u$ due to symmetry (i.e., the case $u\ge v$ can be analyzed similarly and will give the same answer), and finally (d) follows since we can split the minimization in the right hand side of (c) into two domains for $v$, namely $v\ge\frac{1}{4}$ and $v<\frac{1}{4}$. For the former, $\min_{0\le u\le v\le 1,\, v\ge\frac{1}{4}}\max\left\{\frac{(u-\frac{1}{2})^2+(v-\frac{1}{2})^2}{2}, v^2\right\}\ge\min_{0\le u\le v\le 1,\, v\ge\frac{1}{4}} v^2=\frac{1}{16}$. For the latter, it can be easily seen that $\min_{0\le u\le v<\frac{1}{4}}\max\left\{\frac{(u-\frac{1}{2})^2+(v-\frac{1}{2})^2}{2}, v^2\right\}>\frac{1}{16}$. Therefore, (d) holds.

On the other hand, we get

$$\begin{aligned}
\min_{0\le\theta_0,\theta_1\le 1}\max_{0\le\theta\le 1} g(\theta,\theta_0,\theta_1) &\overset{(e)}{\le} \max_{0\le\theta\le 1} g(\theta,\frac{1}{4},\frac{3}{4})\\
&\overset{(f)}{=} \frac{1}{16},
\end{aligned}$$

where (e) follows since fixing $\theta_0=\frac{1}{4}$ and $\theta_1=\frac{3}{4}$ can only increase the value of the left hand side of (e) and (f) follows since $g(\theta,\frac{1}{4},\frac{3}{4})=\frac{1}{16}$ for all $\theta$.

Therefore, by combining steps (a) $\sim$ (f), we conclude that

$$\min_{0\le\theta_0,\theta_1\le 1}\max_{0\le\theta\le 1} g(\theta,\theta_0,\theta_1)=\frac{1}{16}$$

is the minimum possible worst-case MSE and $\theta_0^*=\frac{1}{4},\theta_1^*=\frac{3}{4}$ achieve it.

(c) *Assume $m=1$ and plot the mean squared error (MSE)[1], the squared bias, and the variance for all three estimators, i.e., the sample-mean estimator, the trivial estimator in (a), and the*

[1]Note that the MSE is equal to $\text{Bias}(\hat{\theta}_m)^2+\text{Var}(\hat{\theta}_m)$ as shown in (5.54) because

$$\begin{aligned}
\text{MSE} &= \mathbb{E}[(\hat{\theta}_m-\theta)^2]\\
&= \mathbb{E}[(\hat{\theta}_m-\mathbb{E}(\hat{\theta}_m)+\mathbb{E}(\hat{\theta}_m)-\theta)^2]\\
&= \text{Var}(\hat{\theta}_m)+2\mathbb{E}[\hat{\theta}_m-\mathbb{E}(\hat{\theta}_m)]\mathbb{E}[\mathbb{E}(\hat{\theta}_m)-\theta]+(\mathbb{E}(\hat{\theta}_m)-\theta)^2\\
&= \text{Var}(\hat{\theta}_m)+2\{\mathbb{E}(\hat{\theta}_m)-\mathbb{E}(\hat{\theta}_m)\}\cdot\{\mathbb{E}(\hat{\theta}_m)-\theta\}+\text{Bias}(\hat{\theta}_m)^2\\
&= \text{Var}(\hat{\theta}_m)+\text{Bias}(\hat{\theta}_m)^2.
\end{aligned}$$

*estimator you found in (b) for all values of* $0 \leq \theta \leq 1$. *For which values of* $\theta$, *do you observe a bias-variance tradeoff similar to Fig. 5.6, i.e., the trivial estimator is in the underfitting zone, the sample-mean estimator is in the overfitting zone, and the optimal estimator you found in (b) is in between the two?*

**Solution** i) trivial estimator

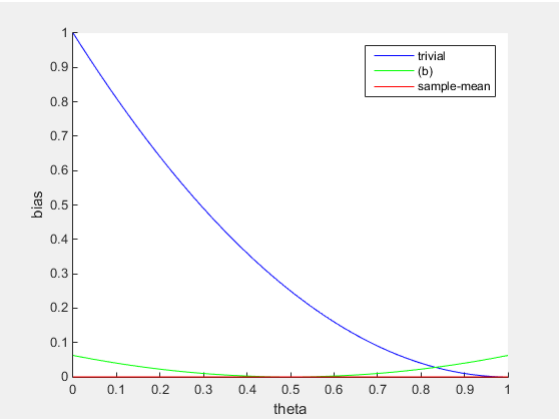
$$\text{bias}(\tilde{\theta_1})^2 = (1-\theta)^2$$
$$\text{var}(\tilde{\theta_1}) = 0$$
$$\text{MSE}(\tilde{\theta_1}) = (1-\theta)^2$$

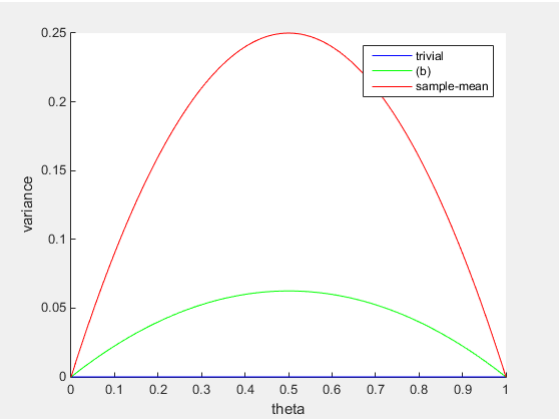
ii) sample-mean estimator

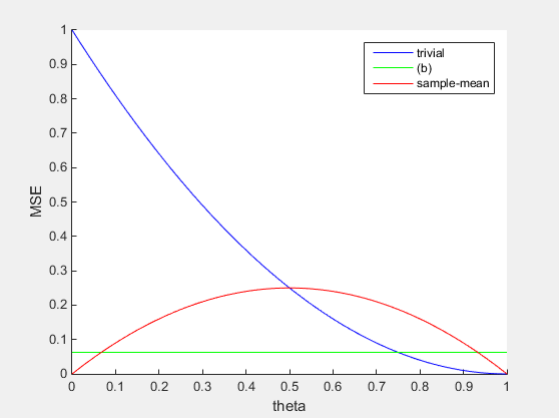
$$\text{bias}(\hat{\theta_1})^2 = 0$$
$$\text{var}(\hat{\theta_1}) = \theta(1-\theta)$$
$$\text{MSE}(\hat{\theta_1}) = \theta(1-\theta)$$

iii) estimator in (b)

$$\text{bias}(\bar{\theta_1})^2 = \left(\tfrac{1-\theta}{4} + \tfrac{3}{4}\theta - \theta\right)^2 = \left(\tfrac{1}{4} - \tfrac{\theta}{2}\right)^2$$
$$\text{var}(\bar{\theta_1}) = \text{MSE}(\bar{\theta_1}) - \text{bias}(\bar{\theta_1})^2 = \tfrac{1}{16} - \left(\tfrac{1}{4} - \tfrac{\theta}{2}\right)^2 = \tfrac{\theta}{4} - \tfrac{\theta^2}{4}$$
$$\text{MSE}(\bar{\theta_1}) = \tfrac{1}{16}$$




To find the range of $\theta$ for which a bias-variance tradeoff is observed such that the trivial estimator is in the underfitting regime, the sample-mean estimator is in the overfitting regime and the estimator in (b) is in between the two, we calculate the following conditions

that compare biases, variances, and MSE's of the three estimators such that the bias keeps decreasing, the variance keeps increasing, and the MSE decreases and then increases as we go from the trivial estimator to the estimator in (b) and to the sample-mean estimator.

i) bias

$$(1-\theta)^2 > (\frac{1}{4} - \frac{\theta}{2})^2 > 0$$

ii) variance

$$0 < \frac{\theta}{4} - \frac{\theta^2}{4} < \theta(1-\theta)$$

iii) MSE

$$(1-\theta)^2 > \frac{1}{16}$$
$$\frac{1}{16} < \theta(1-\theta)$$

Combining above conditions, then we get

$$\frac{2-\sqrt{3}}{4} < \theta < \frac{3}{4}$$

Note that the trivial estimator is in the underfitting regime because it is not even using any observation and the sample-mean estimator is in the overfitting regime because it relies too much on the observed data, i.e., concludes $\theta$ has extreme values of either 0 or 1, even though only one sample was observed. The estimator in (b) achieves a good balance between the two since it is based on the observed data but does not rely too much on it and minimizes the worst-case MSE.

(d) *Assume $m = 1$. Can you think of an estimator that minimizes the average MSE over all $\theta$? Namely, find an estimator $\bar{\theta}_1$ based on $x^{(1)}$ that minimizes the following.*

$$\int_0^1 \mathbb{E}[(\bar{\theta}_1 - \theta)^2] d\theta$$

*How is this different from the estimator in (b)?*

**Solution** Using

$$\mathbb{E}[(\bar{\theta}_1 - \theta)^2] \quad = \quad (1 + 2\theta_0 - 2\theta_1)\theta^2 + (\theta_1^2 - \theta_0^2 - 2\theta_0)\theta + \theta_0^2$$

we obtained in the solution for (a), we get

$$\begin{aligned}\int_0^1 \mathbb{E}[(\bar{\theta}_1 - \theta)^2] d\theta &= \int_0^1 (1 + 2\theta_0 - 2\theta_1)\theta^2 + (\theta_1^2 - \theta_0^2 - 2\theta_0)\theta + \theta_0^2 d\theta \\ &= \frac{1}{2}(\theta_0 - \frac{1}{3})^2 + \frac{1}{2}(\theta_1 - \frac{2}{3})^2 + \frac{1}{18}\end{aligned}$$

Therefore, $\theta_0^*$ and $\theta_1^*$ that minimize $\int_0^1 \mathbb{E}[(\bar{\theta}_1 - \theta)^2] d\theta$ are $\frac{1}{3}$ and $\frac{2}{3}$, respectively. Thus,

$$\bar{\theta}_1 = \begin{cases} \frac{1}{3} & \text{if} \quad x^{(1)} = 0 \\ \frac{2}{3} & \text{if} \quad x^{(1)} = 1 \end{cases}$$

$$\text{bias}(\bar{\theta}_1)^2 = \left(\mathbb{E}(\bar{\theta}_1) - \theta\right)^2 = \left(\tfrac{1-\theta}{3} + \tfrac{2}{3}\theta - \theta\right)^2 = \left(\tfrac{1}{3} - \tfrac{2}{3}\theta\right)^2$$
$$\text{var}(\bar{\theta}_1) = \left(\tfrac{1}{3} - \mathbb{E}(\bar{\theta}_1)\right)^2 (1-\theta) + \left(\tfrac{2}{3} - \mathbb{E}(\bar{\theta}_1)\right)^2 \theta = \tfrac{1}{9}\theta(1-\theta)$$
$$\text{MSE}(\bar{\theta}_1) = \text{bias}(\bar{\theta}_1)^2 + \text{var}(\bar{\theta}_1) = \tfrac{1}{9}\theta(1-\theta) + \left(\tfrac{1}{3} - \tfrac{2}{3}\theta\right)^2 = \tfrac{1}{9}(3\theta^2 - 3\theta + 1)$$

The estimator in (b) has a constant MSE, i.e., $\frac{1}{16}$ and minimizes the worst case MSE. The MSE of the estimator in (d) is between $\frac{1}{36}$ and $\frac{1}{9}$ and its average is $\int_0^1 \frac{1}{9}(3\theta^2 - 3\theta + 1)d\theta = \frac{1}{18}$, therefore achieving a lower minimum MSE than the one in (b) but its worst case MSE is bigger than that achieved by the estimator in (b).

Note that the worst-case MSE is uniquely defined but the average MSE is not since there are many different ways of taking average, e.g., arithmetic average, geometric average, harmonic average, weighted average, etc. In this problem, the average was taken assuming $\theta$ is uniform between 0 and 1, which is a natural choice, but is not the only choice.

**2. (Linear regression for Problem 1)** *Let's consider Problem 1 as a linear regression problem, i.e., we want to build a model to output* $\hat{y} = \mathbf{w}^T\mathbf{x} = \sum_{i=1}^n w_i x_i$ *so that it gives a good estimate of* $\theta$*, where* $\{x_1, \ldots, x_n\}$ *are generated i.i.d. Bernoulli(*$\theta$*). Assume* $0 < \theta < 1$.

(a) *Let's generate a training set consisting of* $m$ *examples with a fixed* $\theta$*. Then, the training set is given by* $(\mathbf{X}, \mathbf{y})$*, where* $\mathbf{X}$ *is an* $m \times n$ *matrix whose elements are all i.i.d. Bernoulli(*$\theta$*) random variables and* $\mathbf{y}$ *is an* $m \times 1$ *vector whose entries are all equal to* $\theta$*. Find the optimum* $\mathbf{w}^*$ *that minimizes the mean square error* $\text{MSE} = \frac{1}{m}\|\hat{\mathbf{y}} - \mathbf{y}\|^2$*. Will it be equivalent to the sample-mean estimator in Problem 1 as* $m, n \to \infty$*?*[2] *You don't need to perform rigorous analysis for convergence of random variables. For example, you may use the approximations* $\frac{1}{m}\sum_{i=1}^m X_{i,j}^2 \sim \theta$*,* $\forall j$ *and* $\frac{1}{m}\sum_{i=1}^m X_{i,j}X_{i,k} \sim \theta^2$*,* $\forall j \neq k$*, which become accurate as* $m \to \infty$*.*

**Solution** As derived in page 19 of lecture note #4, $\mathbf{w}^* = (X^TX)^{-1}X^Ty$. By direct calculation, we can approximate $\mathbf{w}^*$:

$$\begin{aligned}
\mathbf{w}^* &= (X^TX)^{-1}X^Ty \\
&= \begin{pmatrix} \sum_{i=1}^m X_{i,1}^2 & \sum_{i=1}^m X_{i,1}X_{i,2} & \cdots & \sum_{i=1}^m X_{i,1}X_{i,n} \\ \sum_{i=1}^m X_{i,2}X_{i,1} & \sum_{i=1}^m X_{i,2}^2 & \cdots & \sum_{i=1}^m X_{i,2}X_{i,n} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^m X_{i,n}X_{i,1} & \cdots & \cdots & \sum_{i=1}^m X_{i,n}^2 \end{pmatrix}^{-1} X^Ty \\
&\approx \begin{pmatrix} m\theta & m\theta^2 & \cdots & m\theta^2 \\ m\theta^2 & m\theta & \cdots & m\theta^2 \\ \vdots & \vdots & \ddots & \vdots \\ m\theta^2 & \cdots & \cdots & m\theta \end{pmatrix}^{-1} X^Ty = \frac{1}{m\theta}\begin{pmatrix} 1 & \theta & \cdots & \theta \\ \theta & 1 & \cdots & \theta \\ \vdots & \vdots & \ddots & \vdots \\ \theta & \cdots & \cdots & 1 \end{pmatrix}^{-1} X^Ty
\end{aligned}$$

By applying the matrix inversion lemma, we get the following for an invertible matrix $A$ and column vectors $\mathbf{u}$ and $\mathbf{v}$:

$$(A + \mathbf{u}\mathbf{v}^T)^{-1} = A^{-1} - \frac{A^{-1}\mathbf{u}\mathbf{v}^TA^{-1}}{1 + \mathbf{v}^TA^{-1}\mathbf{u}}.$$

[2] Note that since the correct answer $\theta$ is already given in the training set for this problem, an optimum estimator is the one that simply outputs $\theta$. However, the goal of this problem is to see whether a linear regression can learn to solve the problem.

Since $\begin{pmatrix} 1 & \theta & \cdots & \theta \\ \theta & 1 & \cdots & \theta \\ \vdots & \vdots & \ddots & \vdots \\ \theta & \cdots & \cdots & 1 \end{pmatrix} = (1-\theta)I + \Theta$ where $I$ is the identity matrix and $\Theta$ is a $n \times n$ matrix whose elements are all $\theta$, its inverse matrix is:

$$\begin{pmatrix} 1 & \theta & \cdots & \theta \\ \theta & 1 & \cdots & \theta \\ \vdots & \vdots & \ddots & \vdots \\ \theta & \cdots & \cdots & 1 \end{pmatrix}^{-1} = \frac{1}{1-\theta} I - \frac{1}{(1-\theta)^2 + (1-\theta)n\theta}\Theta.$$

Finally, using the above result, $\mathbf{w}^*$ is:

$$\begin{aligned}
\mathbf{w}^* &\approx \frac{1}{m\theta}\left(\frac{1}{1-\theta} I - \frac{1}{(1-\theta)^2 + (1-\theta)n\theta}\Theta\right)\mathbf{X}^T\mathbf{y} \\
&= \frac{1}{m\theta}\left(\frac{1}{1-\theta} I - \frac{1}{(1-\theta)^2 + (1-\theta)n\theta}\Theta\right)\begin{pmatrix} \theta\sum_{i=1}^m X_{i,1} \\ \theta\sum_{i=1}^m X_{i,2} \\ \vdots \\ \theta\sum_{i=1}^m X_{i,n} \end{pmatrix} \\
&\approx \frac{1}{m\theta}\left(\frac{1}{1-\theta} I - \frac{1}{(1-\theta)^2 + (1-\theta)n\theta}\Theta\right)\begin{pmatrix} m\theta^2 \\ m\theta^2 \\ \vdots \\ m\theta^2 \end{pmatrix} \\
&= \frac{\theta}{1-\theta}\begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} - \frac{\theta}{(1-\theta)^2 + (1-\theta)n\theta}\begin{pmatrix} n\theta \\ n\theta \\ \vdots \\ n\theta \end{pmatrix} \\
&= \frac{\theta}{1-\theta+n\theta}\begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \\
&\approx \frac{1}{n}\begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}
\end{aligned}$$

where the first two approximations becomes accurate for sufficiently large $m$ and the third approximation becomes accurate for sufficiently large $n$. Therefore, we can conclude that $\mathbf{w}^*$ gives the sample-mean estimator for sufficiently large $m$ and $n$.

(b) *Now, let's assume $n = 1$ and generate $m$ examples each time choosing $\theta$ uniformly between 0 and 1.*[3] *Furthermore, let's introduce a bias term in the model, i.e., the model is now $\hat{y} = wx + b$. Find the optimum $w$ and $b$ that minimizes the MSE $\frac{1}{m}\|\hat{\mathbf{y}} - \mathbf{y}\|^2$ as $m \to \infty$. Note that you can still use (5.12) to solve this problem by assuming $\hat{y} = w_1 x + w_2 x_2$, where $x_2$ is fixed to 1. Then, $\mathbf{X}$ is an $m \times 2$ vector, where its $i$-th entry in the first column is Bernoulli($\theta_i$) and its second column is all 1 and $\mathbf{y}$ is an $m \times 1$ vector whose $i$-th element*

[3] In Problem 1, the number of samples was $m$, which corresponds to $n$ in Problem 2. $m$ in Problem 2 is the number of examples, where each example $n$ dimensional.

*is $\theta_i$, where $\theta_i$'s are i.i.d. Uniform($[0,1]$). How does this compare with the estimator you found in Problem 1 (d)?*

**Solution** By Problem 2 (a), we know that $\mathbf{w}^* = (X^T X)^{-1} X^T y$. Therefore,

$$\begin{aligned}
\begin{pmatrix} w \\ b \end{pmatrix} &= \left( \begin{pmatrix} x_1 & x_2 & \cdots & x_m \\ 1 & 1 & \cdots & 1 \end{pmatrix} \begin{pmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_m & 1 \end{pmatrix} \right)^{-1} \begin{pmatrix} x_1 & x_2 & \cdots & x_m \\ 1 & 1 & \cdots & 1 \end{pmatrix} \begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_m \end{pmatrix} \\
&= \frac{1}{m\sum_{i=1}^m x_i^2 - (\sum_{i=1}^m x_i)^2} \begin{pmatrix} m\sum_{i=1}^m x_i\theta_i - (\sum_{i=1}^m x_i)(\sum_{i=1}^m \theta_i) \\ (\sum_{i=1}^m x_i^2)(\sum_{i=1}^m \theta_i) - (\sum_{i=1}^m x_i)(\sum_{i=1}^m x_i\theta_i) \end{pmatrix}
\end{aligned}$$

For sufficiently large $m$,

$$\begin{aligned}
&\frac{1}{m}\sum_{i=1}^m x_i \approx \mathrm{E}[X] = \int_0^1 \sum_{x=0}^1 x p(X = x|\theta) d\theta = \int_0^1 \theta d\theta = \frac{1}{2} \\
&\frac{1}{m}\sum_{i=1}^m x_i^2 = \frac{1}{m}\sum_{i=1}^m x_i \approx \frac{1}{2} \\
&\frac{1}{m}\sum_{i=1}^m \theta_i \approx \mathrm{E}[\theta] = \int_0^1 \theta d\theta = \frac{1}{2} \\
&\frac{1}{m}\sum_{i=1}^m x_i\theta_i \approx \mathrm{E}[X\theta] = \int_0^1 \sum_{x=0}^1 x\theta p(X = x|\theta) d\theta = \int_0^1 \theta^2 d\theta = \frac{1}{3}.
\end{aligned}$$

Therefore, the approximation of $\begin{pmatrix} w \\ b \end{pmatrix}$ for sufficiently large $m$ is,

$$\begin{pmatrix} w \\ b \end{pmatrix} = \begin{pmatrix} 1/3 \\ 1/3 \end{pmatrix}.$$

Also, this solution yields the following estimator,

$$\hat{y} = \begin{cases} \frac{1}{3} & \text{if} \quad x = 0 \\ \frac{2}{3} & \text{if} \quad x = 1 \end{cases},$$

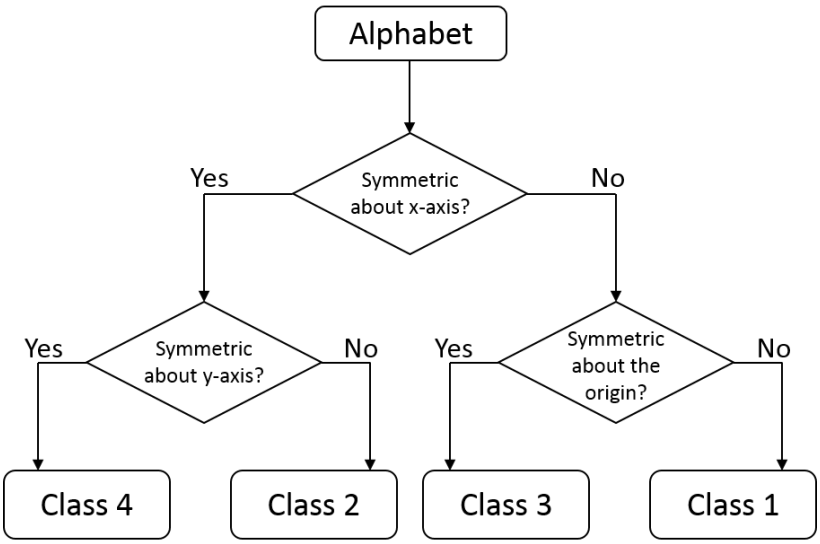
which is the same as the estimator in Problem 1(d).

**3. (Unsupervised learning)** *Design a machine learning algorithm that can automatically classify {A,B,C,D,E,H,I,K,M,N,O,S,T,U,V,W,X,Y,Z}, into the following four categories.*

- A,M,T,U,V,W,Y
- B,C,D,E,K
- N,S,Z
- H,I,O,X

**Solution**

Let Class $i$ denote the $i$-th class from the top. First, let's consider a hand-crafted algorithm shown below.




However, this is not a machine learning algorithm but just a hand-crafted algorithm by a human, where the "learning" part was done by a human brain. (Think about what was going on inside your brain the first moment when you realized the meaning of patterns given in this problem.) By definition, a machine learning algorithm learns from data without being explicitly programmed. In unsupervised learning, correct labels are not even given. Therefore, in unsupervised machine learning, a machine needs to automatically find hidden patterns by observing many examples. Since it does not know a priori which patterns to look for, it can try to look for as many patterns as possible as long as its model capacity allows. If its model is simple, it may be able to find certain edges with certain angles. If its model is complex enough, then it may be able to find horizontal symmetry as in 'A', vertical symmetry as in 'B', point symmetry as in 'O', whether the input has only straight line segments as in 'F', whether the input has curved segments as in 'P', whether the input has only curved segments as in 'S', etc. A deep feedforward neural network with enough complexity will be able to learn such patterns from data in an unsupervised manner (assuming enough data is given). Although we have not learned how such unsupervised learning can be done in deep neural networks, there are ways to do it, e.g., deep belief networks and autoencoders. There is no guarantee that such a neural network will only respond to the four categories given in this problem. In fact, it is more likely that such a network will respond to other categories as well such as whether the input contains straight line segments, etc. But, the point here is that a subset of outputs of such a network, if properly designed and trained, is likely to correspond to the four categories given in this example.

In case of a human brain, if the unconscious part of the brain can detect such patterns and forward the information to the conscious part of the brain, then it can make the job of the conscious part of the brain easier, i.e., the job of finding such patterns "interesting" and taking appropriate actions, e.g., drawing the flow chart given in this solution.

**4. (Deep learning for XOR)** *In page 26 of Lecture notes #7, we saw that the following point becomes a global minimum if $a = 1$, but it is locally convex as $a$ changes in the neighborhood of $-1$.*[4] *Is the following point with $a = -1$ a local minimum or a saddle point?*

$$\mathbf{W} = \begin{pmatrix} a & -a \\ -a & 1 \end{pmatrix}, \quad \mathbf{c} = \begin{pmatrix} 0 \\ a-1 \end{pmatrix}, \quad \mathbf{w} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad b = 0$$

[4]There was a typo in page 26, i.e., $\mathbf{c}$ should be $(0, a-1)^T$ not $(0,0)^T$.

**Solution** Generally, we consider

$$\mathbf{W} = \begin{pmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{pmatrix}, \quad \mathbf{c} = \begin{pmatrix} c_1 \\ c_2 \end{pmatrix}, \quad \mathbf{w} = \begin{pmatrix} w_1 \\ w_2 \end{pmatrix}, \quad b = b.$$

Then, the loss function is given in the follownig matrix form:

$$J(\theta) = \frac{1}{4} \| \big( g(\mathbf{X}\mathbf{W} + \mathbf{c}')\mathbf{w} + b' \big) - \mathbf{Y} \|_2^2$$

where

$$\mathbf{X} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 1 \end{pmatrix}, \quad \mathbf{Y} = \begin{pmatrix} 0 \\ 1 \\ 1 \\ 0 \end{pmatrix}, \quad \mathbf{c}' = \begin{pmatrix} c^T \\ c^T \\ c^T \\ c^T \end{pmatrix}, \quad \mathbf{b}' = \begin{pmatrix} b \\ b \\ b \\ b \end{pmatrix}.$$

First, let's compute $g(\mathbf{X}\mathbf{W} + \mathbf{c}')$.

$$g(\mathbf{X}\mathbf{W} + \mathbf{c}') = g \begin{pmatrix} c_1 & c_2 \\ w_{21} + c_1 & w_{22} + c_2 \\ w_{11} + c_1 & w_{12} + c_2 \\ w_{11} + w_{21} + c_1 & w_{12} + w_{22} + c_2 \end{pmatrix}$$

Note that in the neighborhood of $\mathbf{W} = \begin{pmatrix} -1 & 1 \\ 1 & 1 \end{pmatrix}$ and $\mathbf{c} = \begin{pmatrix} 0 \\ -2 \end{pmatrix}$, some elements of $g(\mathbf{X}\mathbf{W} + \mathbf{c}')$ are strictly smaller or larger than 0. For example, when $c_2$ is around $-2$, $g(c_2) = 0$. Therefore, in the neighborhood of the point, we get

$$g(\mathbf{X}\mathbf{W} + \mathbf{c}') = \begin{pmatrix} g(c_1) & 0 \\ w_{21} + c_1 & 0 \\ 0 & 0 \\ g(w_{11} + w_{21} + c_1) & g(w_{12} + w_{22} + c_2) \end{pmatrix}$$

Then, we get

$$\begin{aligned} J(\theta) &= \frac{1}{4} \| \big( g(\mathbf{X}\mathbf{W} + \mathbf{c}')\mathbf{w} + b' \big) - \mathbf{Y} \|_2^2 \\ &= \frac{1}{4} \left\| \begin{pmatrix} w_1 g(c_1) + b \\ (w_{21} + c_1)w_1 + b - 1 \\ b - 1 \\ w_1 g(w_{11} + w_{21} + c_1) + w_2 g(w_{12} + w_{22} + c_2) + b \end{pmatrix} \right\|_2^2, \end{aligned}$$

from which we get

$$\begin{aligned} \frac{\partial J(\theta)}{\partial b} &= \frac{1}{4} \big( 2(w_1 g(c_1) + b) + 2(w_1(w_{21} + c_1) + b - 1) \\ &\quad + 2(b - 1) + 2(w_1 g(w_{11} + w_{21} + c_1) + w_2 g(w_{12} + w_{22} + c_2) + b) \big) \end{aligned}$$

around $\mathbf{W}^* = \begin{pmatrix} -1 & 1 \\ 1 & 1 \end{pmatrix}$, $\mathbf{c}^* = \begin{pmatrix} 0 \\ -2 \end{pmatrix}$, $\mathbf{w}^* = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ and $b^* = 0$. Therefore,

$$\frac{\partial J(\theta)}{\partial b}\Big|_{\substack{\mathbf{W}=\mathbf{W}^* \\ \mathbf{c}=\mathbf{c}^* \\ \mathbf{w}=\mathbf{w}^* \\ \mathbf{b}=\mathbf{b}^*}} = \frac{1}{4}(2(1 \times 0 + 0) + 2(1 + 0 - 1) + 2(0 - 1) + 2 \times 0) = -\frac{1}{2}.$$

This point is not a critical point. So, it is neither a local minimum nor a saddle point.