Mid Term Exam (Group MT-09)
Machine Learning - 1 [CS/DS 864]

Darshan Bhat - MT2015038
Keerthan Pai K - MT2015053

October 19, 2016

Note: We have discussed with Group MT-03(Freeze Francis and Mohammed Haroon), but the representaions are different.

Q.1

The given description of $H_i's$ follows the structure of a feed-forward neural network with one hidden layer and one node in the output layer.

For given n input points, each H_i hypothesis class can produce $G_{H_i}(n)$ labellings. In the extreme case, the labellings produced are all distinct at the output of hidden layer. Now if H_1 can produce $G_{H_1}(n)$ different labellings of n input points, and H_2 can produce $G_{H_2}(n)$ different labellings independent of H_1 class ,then together we have $G_{H_1}(n) \times G_{H_2}(n)$ different possible labellings of n points at maximum.

With k hypothesis classes at the hidden layer, total number of distinct labellings is upper bounded by $\prod_{i=1}^{k} G_{H_i}(n)$. Now these labellings are fed to the output layer with H_0 hypothesis class. Each input can now be labelled with one of the $G_{H_0}(n)$ labellings. This implies a total of $G_{H_0}(n) \times \prod_{i=0}^{k} G_{H_i}(n)$ labellings are possible at the output layer.

That is,

$$G_H(n) \leq \prod_{i=0}^{k} G_{H_i}(n)$$

We need to prove

$$d_H \leq 2D \log_2 D$$

given that,

$$D > e \log_2 D$$

Proof:

$$\begin{aligned}
2^{d_H} &\leq \prod_{i=0}^{k} \left(\frac{ed_H}{d_{H_i}}\right)^{d_{H_i}} \\
&\leq \prod_{i=0}^{k} \left(\frac{ed_H}{2}\right)^{d_{H_i}} \\
&\leq \left(\frac{ed_H}{2}\right)^{\sum_{i=0}^{k} d_{H_i}} \\
2^{d_H} &\leq \left(\frac{ed_H}{2}\right)^{D} \\
d_H &\leq D\log_2\left(\frac{ed_H}{2}\right)
\end{aligned}$$

This holds for all valid d_H values. We will prove our claim by contradiction.

Let $d_H > 2D\log_2 D$
i.e. $d_H = 2D\log_2 D + 1$, and we know that $D > e\log_2 D$

$$\begin{aligned}
\therefore 2D\log_2 D + 1 &\leq D\log_2\left(\frac{e2D\log_2 D + e}{2}\right) \\
&\leq D\log_2\left(\frac{2eD\log_2 D + 2eD\log_2 D}{2}\right) \\
&\leq D\log_2\left(\frac{4eD\log_2 D}{2}\right) \\
&\leq D\log_2\left(2eD\log_2 D\right) \\
\implies 2D\log_2 D &< D\log_2\left(2eD\log_2 D\right) \\
2\log_2 D &< \log_2 2 + \log_2 e + \log_2 D + \log_2(\log_2 D) \\
log_2 D &< \log_2 2 + \log_2 e + \log_2(\log_2 D) \\
\implies log_2 D &\leq \log_2(e\log_2 D) \\
\implies D &\leq e\log_2 D
\end{aligned}$$

This contradicts with the given $D > e\log_2 D$.

Hence,

$$d_H \leq 2D\log_2 D$$

Q.2

2.a) We can prove that H matrix is idempotent.

$$H = X\left(X^TX\right)^{-1}X^T$$
$$H^2 = X\left(X^TX\right)^{-1}X^TX\left(X^TX\right)^{-1}X^T$$
$$H^2 = X\left(X^TX\right)^{-1}(X^TX)\left(X^TX\right)^{-1}X^T$$
$$H^2 = X\left(X^TX\right)^{-1}X^T$$
$$H^2 = H$$

Now consider,

$$Hx = \lambda x$$
$$H^2x = \lambda x$$
$$HHx = \lambda x$$
$$H\lambda x = \lambda x$$
$$\lambda Hx = \lambda x$$
$$\lambda^2 x = \lambda x$$
$$(\lambda^2 - \lambda)x = 0$$

Solving for λ gives $\lambda = 0$ and $\lambda = 1$.

2.b)
The proof is by induction on the size n of the matrix A. The result is trivial for $n = 1$. Now let $n > 1$ and assume the result is true for any matrix of size $n-1$.

Let λ be the eigenvalue of A, that is, $det(\lambda I - A) = 0$, then $\lambda I - A$ must be non-invertible. This means that there exist a non-zero real vector u such that $Au = \lambda u$. We can always normalize u so that $u^Tu = 1$. Thus, $\lambda = u^TAu$ is real. That is, the eigenvalues of a symmetric matrix are always real.

Now consider the eigenvalue λ_1 and an associated eigenvector u_1. Using the Gram-Schmidt orthogonalization procedure, we can compute a $n\times(n-1)$ matrix V_1 such that $[u_1, V_1]$ is orthogonal. By induction, we can write the $(n-1)\times(n-1)$ symmetric matrix $V_1^TAV_1$ as $Q_1\Lambda_1Q_1^T$, where Q_1 is a $(n-1)\times(n-1)$ matrix of eigenvectors, and $\Lambda_1 = \mathbf{diag}(\lambda_2, \ldots, \lambda_n)$ are the $n-1$ eigenvalues of $V_1^TAV_1$. Finally, we define the $n\times(n-1)$ matrix $U_1 := V_1Q_1$. By construction the matrix $U := [u_1, U_1]$ is orthogonal.

We have

$$U^T A U = \begin{pmatrix} u_1^T \\ U_1^T \end{pmatrix} A \begin{pmatrix} u_1 & U_1 \end{pmatrix} = \begin{pmatrix} u_1^T A u_1 & u_1^T A U_1 \\ U_1^T A u_1 & U_1^T A U_1 \end{pmatrix} = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \Lambda_1 \end{pmatrix},$$

where we have exploited the fact that $U_1^T A u_1 = \lambda_1 U_1^T u_1 = 0$, and $U_1^T A U_1 = \lambda_1$.

We have exhibited an orthogonal $n \times n$ matrix U such that $U^T A U$ is diagonal. This proves the theorem.

2.c)
The trace of a matrix is the sum of its diagonal entries. This has the property that $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$ for any two matrices of same size. $X^T X$ is a $(p+1) \times (p+1)$ matrix.

Now consider,

$$\begin{aligned} \mathrm{Tr}(H) &= \mathrm{Tr}(X(X^T X)^{-1} X^T) \\ &= \mathrm{Tr}((X^T X)^{-1} X^T X) \\ &= \mathrm{Tr}(I_{p+1}) \\ &= p + 1 \end{aligned}$$

$$\mathrm{Tr}(H) = \lambda_1 + \lambda_2 + ... + \lambda_{p+1}$$

Since $\lambda \in$ 0,1, we can conclude that number of eigen values which are 1's is $p + 1$.

Q.3

To show that X^tX is positive definite.
Consider,

$$v^tX^tXv = (Xv)^t(Xv) \tag{1}$$

Let $Xv = z$ with the dimension $n \times 1$.

$$(1) \Rightarrow z^tz = ||z||^2 > 0$$

So, $X^tX > 0$ and thus it is positive semi definite.
We can model linear regression as as a system of linear equations. The vector equation is equivalent to a matrix equation of the form

$$X\mathbf{w} = y$$

where X is an $n \times p$ matrix, w is a column vector with p entries, and y is a column vector with n entries.

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}, \quad \mathbf{w} = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_p \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

Solution for such a system exist only if $\mathbf{y}$ is in the column space of $\mathbf{X}$. It may happend that there exist no solution. In such case we can find the best $\mathbf{w}$ such that $\mathbf{Xw}$ is closest to $\mathbf{b}$ by using least squares approximation.
We know that $\mathbf{Xw}$ is in the column space of $\mathbf{X}$ and $\mathbf{y}$ is not in the plane of $\mathbf{Xw}$.This can be achieved only when $\mathbf{Xw}$ is the projection of $\mathbf{y}$

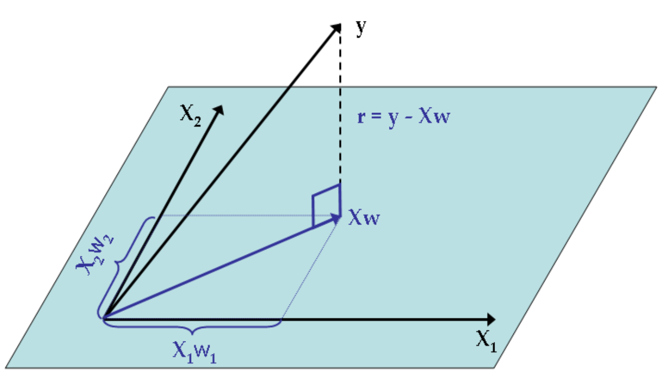

Figure 1:

$$\therefore \mathbf{Xw} \perp (y - \mathbf{Xw})$$

$$
\begin{aligned}
\therefore (\mathbf{X}\mathbf{w})^t \ \ (y - \mathbf{X}\mathbf{w}) &= 0 \\
\mathbf{w}^t\mathbf{X}^t \ \ (y - \mathbf{X}\mathbf{w}) &= 0 \\
\mathbf{w}^t\mathbf{X}^t y - \mathbf{w}^t\mathbf{X}^t\mathbf{X}\mathbf{w} &= 0 \\
\mathbf{w}^t\mathbf{X}^t y &= \mathbf{w}^t\mathbf{X}^t\mathbf{X}\mathbf{w} \\
\mathbf{X}^t y &= \mathbf{X}^t\mathbf{X}\mathbf{w} \\
\mathbf{w} &= (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t y
\end{aligned}
$$

Alternatively, the closed form solution can be derived using the positive semidefinite property of the matrix $\boldsymbol{X}^T X$. Consider the loss function,

$$
\begin{aligned}
S(\boldsymbol{w}) &= \left\|\mathbf{y} - \mathbf{X}\boldsymbol{w}\right\|^2 \\
&= (\mathbf{y} - \mathbf{X}\boldsymbol{w})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{w}) \\
&= \mathbf{y}^{\mathrm{T}}\mathbf{y} - \boldsymbol{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y} - \mathbf{y}^{\mathrm{T}}\mathbf{X}\boldsymbol{w} + \boldsymbol{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{w}
\end{aligned}
$$

When $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is positive definite, the quantity

$$S(\boldsymbol{w}) = \mathbf{y}^{\mathrm{T}}\mathbf{y} - 2\boldsymbol{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y} + \boldsymbol{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{w}$$

can be written as

$$\langle \boldsymbol{w}, \boldsymbol{w} \rangle - 2\langle \boldsymbol{w}, (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} \rangle + \langle (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}, (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} \rangle + C,$$

where C depends only on $\mathbf{y}$ and $\mathbf{X}$, and $\langle \cdot, \cdot \rangle$ is the inner product defined by

$$\langle x, y \rangle = x^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})y.$$

It follows that $S(\boldsymbol{w})$ is equal to

$$\langle \boldsymbol{w} - (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}, \boldsymbol{w} - (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} \rangle + C$$

and therefore minimized exactly when

$$\boldsymbol{w} - (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} = 0.$$

Hence the the best value of w is,

$$\boldsymbol{w} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

Q.4

4.a) y is of the form $y = g(x) + \epsilon_x$.

Let $\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$

$$So,\ Y = g + \epsilon \tag{1}$$

where $\quad \mathbf{g} = \begin{bmatrix} g(x_1) \\ g(x_2) \\ \vdots \\ g(x_n) \end{bmatrix}, \quad \epsilon = \begin{bmatrix} \epsilon_{x_1} \\ \epsilon_{x_2} \\ \vdots \\ \epsilon_{x_n} \end{bmatrix}$

$$\begin{aligned} E &= (Y^T - w^T X^T)(Y - Xw) \\ &= Y^T Y - Y^T X w - w^T X^T Y + w^T X^T X w \end{aligned}$$

To get optimal weights $\hat{w}$, $\frac{\delta E}{\delta w} = 0$

$$i.e. - X^T Y - X^T Y + 2X^T X \hat{w} = 0$$

$$X^T X \hat{w} = X^T Y$$

$$\therefore \hat{w} = (X^T X)^{-1} X^T Y$$

From equation (1),

$$\hat{w} = (X^T X)^{-1} X^T (g + \epsilon)$$

Estimated output $\hat{y} = X\hat{w}$

$$\hat{y} = X(X^T X)^{-1} X^T g + X(X^T X)^{-1} X^T \epsilon$$

From the given data, $g = Xw^*$

So,

$$\hat{Y} = X(X^T X)^{-1} X^T (Xw^*) + X(X^T X)^{-1} X^T \epsilon$$

$$\hat{Y} = Xw * + \hat{H}\epsilon \tag{2}$$

From the given data,

$$y = x^T w^* + \epsilon$$

From (2) $\hat{y}$ at a given point x will be

$$\hat{y} = x^T w^* + x^T (X^T X)^{-1} X^T \epsilon$$

$$y - \hat{y} = \epsilon - x^T (X^T X)^{-1} X^T \epsilon$$

4.b) Let

$$\mathbf{X} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad \mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}$$

We have,

$$\begin{aligned} x^T A x &= \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \\ &= \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n \\ \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n \end{bmatrix} \\ &= \begin{matrix} a_{11}x_1^2 + a_{12}x_1x_2 + \cdots + a_{1n}x_1x_n + a_{21}x_2x_1 + a_{22}x_2^2 + \cdots + a_{2n}x_2x_n \\ + \cdots\cdots + a_{n1}x_nx_1 + \cdots + a_{nn}x_n^2 \end{matrix} \end{aligned} \tag{1}$$

Also,

$$xx^T = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix} = \begin{bmatrix} x_1^2 & x_1x_2 & \cdots & x_1x_n \\ x_2x_1 & x_2^2 & \cdots & x_2x_n \\ \vdots & \vdots & \ddots & \vdots \\ x_nx_1 & x_nx_2 & \cdots & x_n^2 \end{bmatrix}$$

$$xx^T A = \begin{bmatrix} x_1^2 & x_1x_2 & \cdots & x_1x_n \\ x_2x_1 & x_2^2 & \cdots & x_2x_n \\ \vdots & \vdots & \ddots & \vdots \\ x_nx_1 & x_nx_2 & \cdots & x_n^2 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}$$

$$trace(xx^T A) = \begin{matrix} a_{11}x_1^2 + a_{21}x_1x_2 + \cdots + a_{n1}x_1x_n + a_{12}x_2x_1 + a_{22}x_2^2 + \cdots + a_{n2}x_2x_n \\ + \cdots\cdots + a_{1n}x_nx_1 + \cdots + a_{nn}x_n^2 \end{matrix} \tag{2}$$

Since A is symmetric, $a_{ij} = a_{ji}$
So, (1) = (2) that is,

$$x^T A x = tr(xx^T A)$$

Since Expectation E(.) and trace tr(.) both are linear operators,

$$\begin{aligned} tr[E[M]] &= \sum_i E[M_{ii}] \\ &= E[\sum_i M_{ii}] \\ &= E[tr[M]] \end{aligned}$$

4.c) x is a d-dimensional vector in R^d. We need to estimate the covariance matrix $E_D[xx^T]$ where expectation is over all possible datasets which is simply the expectattion of xx^T in the domain of x. The matrix xx^T will be of the order $d \times d$.

$$xx^T = \begin{bmatrix} x_1^2 & x_1x_2 & \cdots & x_1x_d \\ x_2x_1 & x_2^2 & \cdots & x_2x_d \\ \vdots & \vdots & \ddots & \vdots \\ x_dx_1 & x_px_2 & \cdots & x_d^2 \end{bmatrix}$$

$$E[xx^T] = \begin{bmatrix} E[x_1^2] & E[x_1x_2] & \cdots & E[x_1x_d] \\ E[x_2x_1] & E[x_2^2] & \cdots & E[x_2x_d] \\ \vdots & \vdots & \ddots & \vdots \\ E[x_dx_1] & E[x_px_2] & \cdots & E[x_d^2] \end{bmatrix}$$

This matrix can be estimated by X^TX, where X is the $n \times d$ matrix. Each row of X is one sampled d-dimensional vector x_i. So we have n x_i samples arranged row wise in X. So,

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nd} \end{bmatrix}$$

$$X^T = \begin{bmatrix} x_{11} & x_{21} & \cdots & x_{n1} \\ x_{12} & x_{22} & \cdots & x_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1d} & x_{2d} & \cdots & x_{nd} \end{bmatrix}$$

$$X^TX_{d\times d} = \begin{bmatrix} (x_{11}^2 + x_{21}^2 + x_{31}^2 + \cdots) & (x_{11}x_{12} + x_{21}x_{22} + \cdots) & \cdots \\ (x_{12}x_{11} + x_{22}^2 + x_{32}x_{31} + \cdots)(& x_{12}^2 + x_{22}^2 + x_{32}^2 + \cdots) & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}$$

The i, j^{th} entry in X^TX will be

$$[X^TX]_{ij} = \sum_{k=1}^{n} x_{ki}x_{kj}$$

We see here that the i, j^{th} entry in X^TX is filled with i, j^{th} entry in $[x_ix_j^T]$ matrix $fori = 1...n$ i.e. we sample independent points from R^d space with some probability distribution (iid points). we calculate the underlying xx^T matrix for each of these points. Then we add every i, j^{th} from n xx^T matrices and copy to $[X^TX]_{ij}$ entry. Since the points are iid, clearly we are taking an average of $[xx^T]_{ij}$ with these samples.

$$\begin{array}{cc} & j \\ i & \begin{bmatrix} \cdot \\ a_{ij}^1 \end{bmatrix} \end{array} \quad \begin{array}{cc} & j \\ i & \begin{bmatrix} \cdot \\ a_{ij}^2 \end{bmatrix} \end{array} \quad \cdots \begin{array}{cc} & j \\ i & \begin{bmatrix} \cdot \\ a_{ij}^n \end{bmatrix} \end{array} \qquad \begin{bmatrix} X^T X \end{bmatrix} \qquad = \sum_{k=1}^{n} a_{ij}^k$$

from the law of large numbers $[X^T X]$ should converge to $nE_D[xx^T]$. From this we get the bound,

$$nE_D[xx^T] = (1 + O(\sqrt[2]{n}))X^T X$$

$$\implies nE_D[xx^T].(X^T X)^{-1} = (1 + O(\sqrt[2]{n}))I$$

4.d) The true risk of linear regression with square error loss function is given by,

$$E_{D,\epsilon}\left[\|y - \hat{y}\|^2\right]$$

Consider

$$\begin{aligned}
E_D\left[\|y - \hat{y}\|^2\right] &= E_D\left[(y - \hat{y})^T (y - \hat{y})\right] \\
&= E_D\left[(e - x^T(X^T X)^{-1} X^T \epsilon)^T \, (e - x^T(X^T x)^{-1} X^T \epsilon)\right] \\
&= E_D\left[(e^T - \epsilon^T X(X^T X)^{-1} x) \, (e - x^T(X^T X)^{-1} X^T \epsilon)\right] \\
&= E_D\left[e^T e - e^T x^T (X^T X)^{-1} X^T \epsilon - \epsilon^T X (X^T X)^{-1} x e + \epsilon^T X (X^T X)^{-1} x x^T (X^T X)^{-1} X^T \epsilon\right] \\
&= E_D\left[tr\begin{bmatrix} e^T e - e^T x^T (X^T X)^{-1} X^T \epsilon - \epsilon^T X (X^T X)^{-1} x e \\ + \, \epsilon^T X (X^T X)^{-1} x x^T (X^T X)^{-1} X^T \epsilon \end{bmatrix}\right] \qquad (1)
\end{aligned}$$

Consider,

$$\begin{aligned}
tr\left[E_D\left[e^T x^T (X^T X)^{-1} X^T \epsilon\right]\right] &= tr\left[E_D\left[e^T x^T\right] (X^T X)^{-1} X^T \epsilon\right] \\
&= tr\left[E_D\left[e^T\right] E_D\left[x^T\right] (X^T X)^{-1} X^T \epsilon\right] \\
&= tr\left[0\right] \qquad\qquad (\because E_D\left[e^T\right] = 0) \\
&= 0
\end{aligned}$$

Similarly,

$$\begin{aligned}
tr\left[E_D\left[\epsilon^T X (X^T X)^{-1} x e\right]\right] &= tr\left[\epsilon^T X (X^T X)^{-1} E_D\left[x\right] E_D\left[e\right]\right] \\
&= 0
\end{aligned}$$

Therefore,

$$\begin{aligned}
(1) &= E_D\left[tr\left[e^T e + \epsilon^T X (X^T X)^{-1} x x^T (X^T X)^{-1} X^T \epsilon\right]\right] \\
&= \sigma^2 + E_D\left[tr\left[\underbrace{\epsilon^T X (X^T X)^{-1}} x x^T (X^T X)^{-1} X^T \epsilon\right]\right] \\
&= \sigma^2 + E_D\left[tr\left[x x^T (X^T X)^{-1} X^T \epsilon \epsilon^T X (X^T X)^{-1}\right]\right] \\
&= \sigma^2 + tr\left[E_D\left[x x^T\right] (X^T X)^{-1} X^T \epsilon \epsilon^T X (X^T X)^{-1}\right]
\end{aligned}$$

Now,

$$
\begin{aligned}
E_\epsilon\left[E_D\left[\|y-\hat{y}\|^2\right]\right] &= \sigma^2 + E_\epsilon\left[tr\left[E_D\left[xx^T\right](X^TX)^{-1}X^T\epsilon\epsilon^TX(X^TX)^{-1}\right]\right] \\
&= \sigma^2 + tr\left[E_D\left[xx^T\right](X^TX)^{-1}X^TE_\epsilon\left[\epsilon\epsilon^T\right]X(X^TX)^{-1}\right] \\
&= \sigma^2 + tr\left[E_D\left[xx^T\right](X^TX)^{-1}X^T(\sigma^2 I)X(X^TX)^{-1}\right] \\
&= \sigma^2 + \frac{\sigma^2}{n}tr\left[nE_D\left[xx^T\right]\underline{(X^TX)^{-1}X^TX}(X^TX)^{-1}\right] \\
&= \sigma^2 + \frac{\sigma^2}{n}tr\left[nE_D\left[xx^T\right](X^TX)^{-1}\right] \\
&= \sigma^2 + \frac{\sigma^2}{n}\left[tr\left(1+o\left(\sqrt{n}\right)\right)I\right] \\
&= \sigma^2 + \frac{\sigma^2}{n}\left(d+1+o\left(\sqrt{n}\right)(d+1)\right) \\
&= \sigma^2\left(1+\frac{d+1}{n}+o\left(\frac{d+1}{\sqrt{n}}\right)\right)
\end{aligned}
$$

Q.5

Bounding $R(\hat{h})$ under $R_v(\hat{h})$ and intern $R(\hat{h_l})$:

We are selecting $\hat{h}$ from the set $H = \{\hat{h_1}, \hat{h_2}, \ldots, \hat{h_l}, \ldots, \hat{h_k}\}$ for which $R_v(\hat{h_i})$ is the minimum. This is same as ERM principle for which the VC theorem bound holds.
So,

$$P\left(|R(\hat{h}) - R_v(\hat{h})| \geq \epsilon\right) \leq 4G_H(2n)e^{-2\epsilon^2(1-\alpha)n}$$

$G_H(2n)$ is bounded by $|H| = k$.

$$P\left(|R(\hat{h}) - R_v(\hat{h})| \geq \epsilon\right) \leq 4ke^{-2\epsilon^2(1-\alpha)n}$$

Let $4ke^{-2\epsilon^2(1-\alpha)n} = \frac{\delta}{2}$.
With probability of atleast $1 - \frac{\delta}{2}$,

$$R(\hat{h}) \leq R_v(\hat{h}) + \epsilon$$

Since $R_v(\hat{h_l}) \leq R_v(\hat{h})$,

$$R(\hat{h}) \leq R_v(\hat{h_l}) + \epsilon \tag{1}$$

Bounding $R_v(\hat{h_l})$ in terms of $R(\hat{h_l})$:

Since ERM principle holds for this case also,

$$P\left(|R_v(\hat{h_l}) - R(\hat{h_l})| \geq \epsilon\right) \leq 4ke^{-2\epsilon^2(1-\alpha)n}$$

So with the probability of atleast $1 - \frac{\delta}{2}$,

$$R_v(\hat{h_l}) \leq R(\hat{h_l}) + \epsilon \tag{2}$$

Bounding $R(\hat{h_l})$ in terms of $R_T(\hat{h_l})$ in turn by $R_T(\hat{h^*})$:

We selected $\hat{h_l}$ from H_l class by ERM principle with training set risk as empherical risk. Here $G_{H_l}(2n)$ is bounded by c_l,

$$\Rightarrow P\left(|R(\hat{h_l}) - R_T(\hat{h_l})| \geq \epsilon'\right) \leq 4c_l e^{-2(\epsilon')^2\alpha n}$$

Let $4c_l e^{-2(\epsilon')^2(\alpha)n}$ be $\frac{\delta}{2}$.
Then with probability of atleast $\frac{\delta}{2}$,

$$R(\hat{h_l}) \leq R_T(\hat{h_l}) + \epsilon'$$

Since $R_T(\hat{h_l}) \leq R_T(h^*)$,

$$R(\hat{h_l}) \leq R_T(h^*) + \epsilon' \tag{3}$$

<u>*Bounding $R_T(h^*)$ using $R(h*)$:*</u>

Again with V.C, theorem,

$$P\left(|R(h^*) - R_T(h^*)| \geq \epsilon'\right) \leq 4c_l e^{-2(\epsilon')^2 \alpha n}$$

Then with probability of atleast $\frac{\delta}{2}$,

$$R_T(h^*) \leq R(h^*) + \epsilon' \tag{4}$$

From (1),(2),(3) and (4)

$$R(\hat{h}) \leq R(h^*) + 2\epsilon + 2\epsilon' \tag{5}$$

<u>*To calculate ϵ and ϵ' :*</u>

$$\frac{\delta}{2} = 4ke^{-2\epsilon^2(1-\alpha)n}$$

$$\epsilon = \sqrt{\frac{1}{2(1-\alpha)n}.ln\frac{8k}{\delta}} \tag{6}$$

$$\frac{\delta}{2} = 4c_l e^{-2(\epsilon')^2(\alpha)n}$$

$$\epsilon` = \sqrt{\frac{1}{2(\alpha)n}.ln\frac{8c_l}{\delta}} \tag{7}$$

Substituting (6) and (7) in (5),

$$R(\hat{h}) \leq R(\hat{h^*}) + 2\sqrt{\frac{1}{2(1-\alpha)n}.ln\frac{8k}{\delta}} + 2\sqrt{\frac{1}{2(\alpha)n}.ln\frac{8c_l}{\delta}} \tag{8}$$

Both validation set equations $\{(1),(2)\}$ and training set equations $\{(3),(4)\}$ together hold with probability

$$(1-\frac{\delta}{2}).(1-\frac{\delta}{2}) = (1-\frac{\delta}{2})^2 = 1 + \frac{\delta^2}{4} - 2\frac{\delta}{2} \geq (1-\delta)$$

With probability at least $(1-\delta)$ equation (8) holds.