COL341: Fundamentals of Machine Learning Spring 2023

Homework 1

Lecturer: Prof. Chetan Arora Saket Kandoi 2021MT60265

Question 1

(a) Given,

$$H = X(X^TX)^{-1}X^T$$

To show:

$$H = H^T \tag{1}$$

Let us begin by calculating H^T,

$$\begin{aligned}
H^T &= (X(X^TX)^{-1}X^T)^T \\
&= (X^T)^T((X^TX)^{-1})^TX^T && (\because (A_1A_2\ldots A_n)^T = A_n^T\ldots A_2^TA_1^T) \\
&= X((X^TX)^{-1})^TX^T && (\because (A^T)^T = A) \\
&= X((X^TX)^T)^{-1}X^T && (\because I = (A^{-1}A)^T = A^T(A^{-1})^T \implies (A^T)^{-1} = (A^{-1})^T) \\
&= X(X^T(X^T)^T)^{-1}X^T \\
&= X(X^TX)^{-1}X^T \\
&= H
\end{aligned}$$

Hence,H is symmetric.

(b) To show:

$$H^K = H \text{ for any positive integer } K. \tag{2}$$

We will prove the above statement using mathematical induction.
Base case:
For $K = 1$, $H = H$ is trivial. For $K = 2$,

$$\begin{aligned}
H^2 &= H \cdot H \\
&= (X(X^TX)^{-1}X^T) \cdot (X(X^TX)^{-1}X^T) \\
&= X(X^TX)^{-1}X^TX(X^TX)^{-1}X^T \\
&= X(X^TX)^{-1}(X^TX)(X^TX)^{-1}X^T \\
&= X((X^TX)^{-1}(X^TX))(X^TX)^{-1}X^T \\
&= X(I)(X^TX)^{-1}X^T && (\because A^{-1}A = AA^{-1} = I) \\
&= X(X^TX)^{-1}X^T && (\because XI = IX = X) \\
&= H
\end{aligned}$$

Induction Hypothesis: $H^K = H$ for $K = n$ $(n \geq 2)$

Inductive Step:

$$\begin{aligned}
H^{n+1} &= H^n \cdot H \\
&= H \cdot H && \text{(By Induction Hypothesis)} \\
&= H^2 \\
&= H
\end{aligned}$$

Hence, by the principle of mathematical induction, $H^K = H$ is true for all $K \in \mathbb{N}$.

(c) Given I is a $N \times N$ Identity matrix,
To show:

$$(I - H)^K = (I - H) \text{ for any positive integer } K. \tag{3}$$

We will prove the above statement using mathematical induction.
Base case:
For $K = 1$, $(I - H) = (I - H)$ is trivial. For $K = 2$,

$$\begin{aligned}
(I - H)^2 &= (I - H) \cdot (I - H) \\
&= I \cdot I - I \cdot H - H \cdot I + H \cdot H \\
&= I^2 - 2H + H^2 \\
&= I - 2H + H^2 \\
&= I - 2H + H && \text{(By Equation (2))} \\
&= I - H
\end{aligned}$$

Induction Hypothesis: $(I - H)^K = I - H$ for $K = n$ $(n \geq 2)$

Inductive Step:

$$\begin{aligned}
(I - H)^{n+1} &= (I - H)^n \cdot (I - H) \\
&= (I - H) \cdot (I - H) && \text{(By Induction Hypothesis)} \\
&= (I - H)^2 \\
&= I - H
\end{aligned}$$

Hence, by the principle of mathematical induction, $(I - H)^K = (I - H)$ is true for all $K \in \mathbb{N}$.

(d) To show:

$$trace(H) = d + 1 \tag{4}$$

Substituting H,

$$\begin{aligned}
trace(H) &= trace(X(X^TX)^{-1}X^T) \\
&= trace((X(X^TX)^{-1})X^T) \\
&= trace(X^T(X(X^TX)^{-1})) && (\because trace(AB) = trace(BA)) \\
&= trace((X^TX)(X^TX)^{-1}) \\
&= trace(I) \text{ where } I \text{ is a } (d+1)\times(d+1) \text{ Identity matrix} \\
&= d+1
\end{aligned}$$

Question 2

Given,

$$y = \mathbf{w}^{*T}\mathbf{x} + \epsilon$$

For the data $D = \{(\mathbf{x_1}, y_1), \ldots, (\mathbf{x_N}, y_N)\}$,

$$\begin{aligned}
y_i &= \mathbf{w}^{*T}\mathbf{x_i} + \epsilon_i \text{ for all } 1 \le i \le N \\
\mathbf{y} &= X\mathbf{w}^* + \epsilon && (5)
\end{aligned}$$

(a) We know

$$\begin{aligned}
\hat{\mathbf{y}} &= H\mathbf{y} \\
&= H\cdot(X\mathbf{w}^* + \epsilon) \\
&= H\cdot X\mathbf{w}^* + H\epsilon \\
&= (X(X^TX)^{-1}X^T)\cdot X\mathbf{w}^* + H\epsilon \\
&= X((X^TX)^{-1}(X^TX))\mathbf{w}^* + H\epsilon \\
&= X(I)\mathbf{w}^* + H\epsilon \\
&= X\mathbf{w}^* + H\epsilon && (6)
\end{aligned}$$

(b) By equations (5) and (6), we know

$$\begin{aligned}
\mathbf{y} &= X\mathbf{w}^* + \epsilon \\
\hat{\mathbf{y}} &= X\mathbf{w}^* + H\epsilon \\
\hat{\mathbf{y}} - \mathbf{y} &= X\mathbf{w}^* + H\epsilon - (X\mathbf{w}^* + \epsilon) \\
&= H\epsilon - \epsilon \\
&= (H - I)\epsilon \text{ where } I \text{ is a } N\times N \text{ Identity matrix} && (7)
\end{aligned}$$

Hence, required matrix is $(H - I)$.

(c)

$$\begin{aligned}
E_{in}(\mathbf{w_{lin}}) &= \frac{1}{N}\sum_{n=1}^{N}(\hat{y}_i - y_i)^2 \\
&= \frac{1}{N}||\hat{\mathbf{y}} - \mathbf{y}||^2 \\
&= \frac{1}{N}||(H-I)\epsilon||^2 \\
&= \frac{1}{N}(((H-I)\epsilon)^T(H-I)\epsilon) \\
&= \frac{1}{N}(\epsilon^T(H-I)^T(H-I)\epsilon) \\
&= \frac{1}{N}(\epsilon^T(H^T-I^T)(H-I)\epsilon) \\
&= \frac{1}{N}(\epsilon^T(H-I)(H-I)\epsilon) && \text{(by Equation (1))} \\
&= \frac{1}{N}(\epsilon^T(H-I)^2\epsilon) \\
&= \frac{1}{N}(\epsilon^T(I-H)^2\epsilon) \\
&= \frac{1}{N}(\epsilon^T(I-H)\epsilon) && \text{(by Equation (3))} \qquad (8)
\end{aligned}$$

(d)

$$\begin{aligned}
\mathbb{E}_D\left[E_{in}(\mathbf{w_{lin}})\right] &= \mathbb{E}_D\left[\frac{1}{N}(\epsilon^T(I-H)\epsilon)\right] \\
&= \frac{1}{N}\mathbb{E}_D\left[\epsilon^T(I-H)\epsilon\right] \\
&= \frac{1}{N}\mathbb{E}_D\left[\epsilon^T\epsilon - \epsilon^T H\epsilon\right] \\
&= \frac{1}{N}\mathbb{E}_D\left[\sum_{i=1}^{N}\epsilon_i^2 - \sum_{i=1}^{N}\sum_{j=1}^{N}\epsilon_i h_{ij}\epsilon_j\right] \\
&= \frac{1}{N}\left(\mathbb{E}_D\left[\sum_{i=1}^{N}\epsilon_i^2\right] - \mathbb{E}_D\left[\sum_{i=1}^{N}\sum_{j=1}^{N}\epsilon_i h_{ij}\epsilon_j\right]\right) \\
&= \frac{1}{N}\left(\sum_{i=1}^{N}\mathbb{E}_D\left[\epsilon_i^2\right] - \sum_{i=1}^{N}\sum_{j=1}^{N}\mathbb{E}_D\left[\epsilon_i h_{ij}\epsilon_j\right]\right) \qquad (9)
\end{aligned}$$

Since ϵ_i's are independent, expectation of each ϵ_i^2 depends only corresponding x_i and y_i. Let us first simplify the first summation,

$$\begin{aligned}\sum_{i=1}^{N} \mathbb{E}_D\left[\epsilon_i^2\right] &= \sum_{i=1}^{N} \mathbb{E}_{(x_i,y_i)}\left[\epsilon_i^2\right] \\ &= \sum_{i=1}^{N} \sigma_i^2 - \mu_i^2 \\ &= \sum_{i=1}^{N} \sigma^2 - 0^2 \\ &= N\sigma^2 \end{aligned} \tag{10}$$

Now the second summation,

$$\sum_{i=1}^{N}\sum_{j=1}^{N} \mathbb{E}_D\left[\epsilon_i h_{ij} \epsilon_j\right] = \sum_{i=1}^{N}\sum_{j=1, j\neq i}^{N} \mathbb{E}_D\left[\epsilon_i h_{ij} \epsilon_j\right] + \sum_{i=1}^{N} \mathbb{E}_D\left[\epsilon_i^2 h_{ii}\right]$$

Let us approach these terms of both these summations separately,

$$\begin{aligned}\mathbb{E}_D\left[\epsilon_i^2 h_{ii}\right] &= \mathbb{E}_{(x_1,y_1)}\left[\epsilon_i^2 h_{ii}\right] \\ \mathbb{E}_D\left[\epsilon_i h_{ij} \epsilon_j\right] &= \mathbb{E}_{\{(x_i,y_i),(x_j,y_j)\}}\left[\epsilon_i h_{ij} \epsilon_j\right]\end{aligned}$$

This is the lowest form to which the expression can be reduced with the given data. However, if we assume H is not randomly distributed (H is completely constructed over terms of x_i, which is not a random variable), we can further simply,

$$\begin{aligned}\mathbb{E}_D\left[\epsilon_i^2 h_{ii}\right] &= h_{ii}\mathbb{E}_{(x_i,y_i)}\left[\epsilon_i^2\right] \\ &= h_{ii}\sigma^2 \\ \mathbb{E}_D\left[\epsilon_i h_{ij} \epsilon_j\right] &= h_{ij}\mathbb{E}_{\{(x_i,y_i),(x_j,y_j)\}}\left[\epsilon_i \epsilon_j\right] \\ &= h_{ij}\mathbb{E}_{(x_i,y_i)}\left[\epsilon_i\right]\mathbb{E}_{(x_j,y_j)}\left[\epsilon_j\right] \\ &= h_{ij} \cdot 0 \cdot 0 \\ &= 0\end{aligned}$$

Hence,

$$\begin{aligned}\sum_{i=1}^{N}\sum_{j=1}^{N} \mathbb{E}_D\left[\epsilon_i h_{ij} \epsilon_j\right] &= \sum_{i=1}^{N} h_{ii}\sigma^2 \\ &= \sigma^2 \sum_{i=1}^{N} h_{ii} \\ &= \sigma^2 \cdot trace(H)\end{aligned} \tag{11}$$

Substituting in equation (9),

$$
\begin{aligned}
\mathbb{E}_D\left[E_{in}(\mathbf{w_{lin}})\right] &= \frac{1}{N}(N\sigma^2 - \sigma^2 trace(H)) \\
&= \sigma^2\left(1 - \frac{trace(H)}{N}\right) \\
&= \sigma^2\left(1 - \frac{d+1}{N}\right) \text{ by Equation (4)}
\end{aligned} \tag{12}
$$

(e)

$$
\begin{aligned}
E_{test}(\mathbf{w_{lin}}) &= \frac{1}{N}\sum_{n=1}^{N}(\hat{y}_i - y'_i)^2 \\
&= \frac{1}{N}||\hat{\mathbf{y}} - \mathbf{y}'||^2 \\
&= \frac{1}{N}||X\mathbf{w}^* + H\epsilon - (X\mathbf{w}^* + \epsilon')||^2 \\
&= \frac{1}{N}||H\epsilon - \epsilon'||^2 \\
&= \frac{1}{N}((H\epsilon - \epsilon')^T(H\epsilon - \epsilon')) \\
&= \frac{1}{N}((\epsilon^T H^T - \epsilon'^T)(H\epsilon - \epsilon')) \\
&= \frac{1}{N}((\epsilon^T H - \epsilon'^T)(H\epsilon - \epsilon')) \text{ By equation (1)} \\
&= \frac{1}{N}(\epsilon^T H^2\epsilon - \epsilon^T H\epsilon' - \epsilon'^T H\epsilon + \epsilon'^T\epsilon') \\
\mathbb{E}_{D,\epsilon'}\left[E_{out}(\mathbf{w_{lin}})\right] &= \mathbb{E}_{D,\epsilon'}\left[\frac{1}{N}(\epsilon^T H^2\epsilon - \epsilon^T H\epsilon' - \epsilon'^T H\epsilon + \epsilon'^T\epsilon')\right] \\
&= \frac{1}{N}\mathbb{E}_{D,\epsilon'}\left[\epsilon^T H^2\epsilon - \epsilon^T H\epsilon' - \epsilon'^T H\epsilon + \epsilon'^T\epsilon'\right] \\
&= \frac{1}{N}\mathbb{E}_{D,\epsilon'}\left[\epsilon^T H^2\epsilon + \epsilon'^T\epsilon'\right] \\
&= \frac{1}{N}\mathbb{E}_{D,\epsilon'}\left[\sum_{i=1}^{N}\sum_{j=1}^{N}\epsilon_i h_{ij}\epsilon_j + \sum_{i=1}^{N}\epsilon_i'^2\right] \\
&= \frac{1}{N}\left(\sum_{i=1}^{N}\sum_{j=1}^{N}\mathbb{E}_D\left[\epsilon_i h_{ij}\epsilon_j\right] + \sum_{i=1}^{N}\mathbb{E}_D\left[\epsilon_i'^2\right]\right)
\end{aligned}
$$

Substituting from equations (10) and (11),

$$
\mathbb{E}_{D,\epsilon'}\left[E_{out}(\mathbf{w_{lin}})\right] = \frac{1}{N}(\sigma^2 trace(H) + N(\sigma'^2 - \mu'^2))
$$

If we assume both ϵ and ϵ' have the same mean and variance,

$$\begin{aligned}\mathbb{E}_{D,\epsilon'}\left[E_{out}(\mathbf{w_{lin}})\right] &= \frac{1}{N}(\sigma^2 trace(H) + N\sigma^2)\\ &= \sigma^2\left(\frac{d+1}{N}+1\right)\end{aligned}$$

Question 3

Given,
$\sum = \mathbb{E}_{\mathbf{x}}[\mathbf{x}\mathbf{x}^{\mathbf{T}}]$ is non-singular, i.e,

$$||\mathbb{E}_{\mathbf{x}}[\mathbf{x}\mathbf{x}^{\mathbf{T}}]|| \neq 0 \tag{13}$$

(a) For the purposes of legibility, ϵ_{test} will be used to denote the noise realization for the test point and ϵ will be used to denote the vector of noise realizations on the data.

$$\begin{aligned}y - g(\mathbf{x}) &= (\mathbf{w}^{*\prime})^{\mathbf{T}}\mathbf{x} + \epsilon_{test} - (\mathbf{w_{lin}}^{\mathbf{T}}\mathbf{x})\\ \mathbf{w_{lin}} &= X^{\dagger}\mathbf{y}\\ &= X^{\dagger}(X\mathbf{w}^* + \epsilon) \text{ By equation (5)}\\ &= ((X^TX)^{-1}X^T)\cdot(X\mathbf{w}^* + \epsilon)\\ \mathbf{w_{lin}}^T &= (X\mathbf{w}^* + \epsilon)^T \cdot ((X^TX)^{-1}X^T)^T\\ &= (\mathbf{w}^{*T}X^T + \epsilon^T)\cdot((X^T)^T((X^TX)^{-1})^T)\\ &= (\mathbf{w}^{*T}X^T + \epsilon^T)\cdot(X((X^TX)^T)^{-1})\\ &= (\mathbf{w}^{*T}X^T + \epsilon^T)\cdot(X(X^T(X^T)^T)^{-1})\\ &= (\mathbf{w}^{*T}X^T + \epsilon^T)\cdot(X(X^TX)^{-1})\\ &= \mathbf{w}^{*T}X^TX(X^TX)^{-1} + \epsilon^TX(X^TX)^{-1}\\ &= \mathbf{w}^{*T} + \epsilon^TX(X^TX)^{-1}\\ \mathbf{w_{lin}}^T\mathbf{x} &= \mathbf{w}^{*T}\mathbf{x} + \epsilon^TX(X^TX)^{-1}\mathbf{x}\\ &= \mathbf{w}^{*T}\mathbf{x} + (\mathbf{x}^T(\epsilon^TX(X^TX)^{-1})^T)^T\\ &= \mathbf{w}^{*T}\mathbf{x} + (\mathbf{x}^T((X^TX)^{-1})^TX^T\epsilon)^T\\ &= \mathbf{w}^{*T}\mathbf{x} + (\mathbf{x}^T((X^TX)^T)^{-1}X^T\epsilon)^T\\ &= \mathbf{w}^{*T}\mathbf{x} + (\mathbf{x}^T(X^TX)^{-1}X^T\epsilon)^T\end{aligned}$$

Since all terms in the given expression are scalars,

$$\mathbf{w_{lin}}^T\mathbf{x} = \mathbf{w}^{*T}\mathbf{x} + \mathbf{x}^T(X^TX)^{-1}X^T\epsilon \tag{14}$$

Hence,

$$y - g(\mathbf{x}) = (\mathbf{w}^{*\prime})^{\mathbf{T}}\mathbf{x} + \epsilon_{test} - (\mathbf{w}^{*T}\mathbf{x} + \mathbf{x}^T(X^TX)^{-1}X^T\epsilon)$$

If the test data set also comes from the same genuine linear relationship,

$$y - g(\mathbf{x}) = \epsilon_{test} - \mathbf{x}^T(X^TX)^{-1}X^T\epsilon \tag{15}$$

(b)

$$\begin{aligned}
E_{out} &= \mathbb{E}_{\mathbf{x},\epsilon_{test}}\left[(y - g(\mathbf{x}))^2\right] \\
&= \mathbb{E}_{\mathbf{x},\epsilon_{test}}\left[(\epsilon_{test} - \mathbf{x}^T(X^TX)^{-1}X^T\epsilon)^2\right] \text{ By equation (15)} \\
&= \mathbb{E}_{\mathbf{x},\epsilon_{test}}\left[\epsilon_{test}^2 - 2\epsilon_{test}\mathbf{x}^T(X^TX)^{-1}X^T\epsilon + (\mathbf{x}^T(X^TX)^{-1}X^T\epsilon)^2\right] \\
&= \mathbb{E}_{\epsilon_{test}}\left[\epsilon_{test}^2\right] - 2\mathbb{E}_{\epsilon_{test}}\left[\epsilon_{test}\right]\mathbb{E}_D\left[\mathbf{x}^T(X^TX)^{-1}X^T\epsilon\right] + \mathbb{E}_D\left[(\mathbf{x}^T(X^TX)^{-1}X^T\epsilon)(\mathbf{x}^T(X^TX)^{-1}X^T\epsilon)^T\right]
\end{aligned}$$

Since $\mathbf{x}^T(X^TX)^{-1}X^T\epsilon$ is a scalar, it is equal to its transpose. Also $a = trace(a)$ for any scalar a.

$$\begin{aligned}
E_{out} &= \mathbb{E}_{\epsilon_{test}}\left[\epsilon_{test}^2\right] - 2\cdot 0\cdot\mathbb{E}_D\left[\mathbf{x}^T(X^TX)^{-1}X^T\epsilon\right] + \mathbb{E}_D\left[trace((\mathbf{x}^T(X^TX)^{-1}X^T\epsilon)(\mathbf{x}^T(X^TX)^{-1}X^T\epsilon)^T)\right] \\
&= \sigma^2 + trace\left(\mathbb{E}_D\left[\mathbf{x}^T(X^TX)^{-1}X^T\epsilon\epsilon^TX((X^TX)^{-1})^T\mathbf{x}\right]\right) \\
&= \sigma^2 + trace\left(\mathbb{E}_D\left[\mathbf{x}\mathbf{x}^T(X^TX)^{-1}X^T\epsilon\epsilon^TX((X^TX)^{-1})^T\right]\right) \ (trace(AB) = trace(BA)) \\
&= \sigma^2 + trace\left(\mathbb{E}_{\mathbf{x}}\left[\mathbf{x}\mathbf{x}^T\right](X^TX)^{-1}X^T\epsilon\epsilon^TX((X^TX)^T)^{-1}\right) \\
&= \sigma^2 + trace\left(\Sigma(X^TX)^{-1}X^T\epsilon\epsilon^TX(X^TX)^{-1}\right)
\end{aligned} \tag{16}$$

(c) ϵ is a $N \times 1$ matrix, and its transpose is a $1 \times N$ matrix. Hence, the matrix product results in a $N \times N$ matrix.

$$\begin{aligned}
\epsilon\epsilon^T &= \begin{vmatrix} \epsilon_1 \\ \epsilon_2 \\ \dots \\ \epsilon_N \end{vmatrix} \cdot \begin{vmatrix} \epsilon_1 & \epsilon_2 & \dots & \epsilon_N \end{vmatrix} \\
&= \begin{vmatrix} \epsilon_1\epsilon_1 & \epsilon_1\epsilon_2 & \dots & \epsilon_1\epsilon_N \\ \epsilon_2\epsilon_1 & \epsilon_2\epsilon_2 & \dots & \epsilon_2\epsilon_N \\ \dots & \dots & \dots & \dots \\ \epsilon_N\epsilon_1 & \epsilon_N\epsilon_2 & \dots & \epsilon_N\epsilon_N \end{vmatrix}
\end{aligned}$$

Expectation over ϵ will apply to each term in the matrix, with

$$\mathbb{E}_\epsilon[\epsilon_i\epsilon_j] = \begin{cases} 0 & : i \neq j \\ \sigma^2 & : i = j \end{cases}$$

$$\begin{aligned}
\mathbb{E}_\epsilon[\epsilon\epsilon^T] &= \begin{vmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \sigma^2 \end{vmatrix} \\
\mathbb{E}_\epsilon[\epsilon\epsilon^T] &= \sigma^2 I
\end{aligned} \tag{17}$$

(d) Taking expectation over ϵ on equation (16),

$$\begin{aligned}
E_{out} &= E_\epsilon \left[\sigma^2 + trace\left(\Sigma(X^TX)^{-1}X^T\epsilon\epsilon^TX(X^TX)^{-1}\right)\right] \\
&= \sigma^2 + E_\epsilon \left[trace\left(\Sigma(X^TX)^{-1}X^T\epsilon\epsilon^TX(X^TX)^{-1}\right)\right] \\
&= \sigma^2 + trace\left(E_\epsilon\left[\Sigma(X^TX)^{-1}X^T\right]E_\epsilon\left[\epsilon\epsilon^T\right]E_\epsilon\left[X(X^TX)^{-1}\right]\right) \\
&= \sigma^2 + trace\left(\Sigma(X^TX)^{-1}X^TE_\epsilon\left[\epsilon\epsilon^T\right]X(X^TX)^{-1}\right) \\
&= \sigma^2 + trace\left(\Sigma(X^TX)^{-1}X^T\sigma^2IX(X^TX)^{-1}\right) \text{ By equation (17)} \\
&= \sigma^2 + \sigma^2 trace\left(\Sigma(X^TX)^{-1}X^TX(X^TX)^{-1}\right) \\
&= \sigma^2\left(1 + trace\left(\Sigma(X^TX)^{-1}\right)\right) \\
&= \sigma^2\left(1 + trace\left(\frac{1}{N}\cdot N\cdot\Sigma(X^TX)^{-1}\right)\right) \\
&= \sigma^2\left(1 + \frac{1}{N}trace\left(N\cdot\Sigma(X^TX)^{-1}\right)\right) \\
&= \sigma^2\left(1 + \frac{1}{N}trace\left(\Sigma(\frac{1}{N}X^TX)^{-1}\right)\right) \qquad (18)
\end{aligned}$$

If $\frac{1}{N}X^TX \approx \Sigma$, we have

$$\begin{aligned}
E_{out} &= \sigma^2\left(1 + \frac{1}{N}trace\left(I\right)\right) \text{ where } I \text{ is a } d+1\times d+1 \text{ Identity matrix} \\
&= \sigma^2\left(1 + \frac{d+1}{N}\right)
\end{aligned}$$

(e) By the definition of convergence in probability, for sufficiently small δ, $(\frac{1}{N}X^TX)^{-1}$ lies within δ of Σ^{-1} with required high probability with respect to a given norm,i.e.,

$$\left\|\left(\frac{1}{N}X^TX\right)^{-1} - \Sigma^{-1}\right\| \leq \delta \qquad (19)$$

$$\begin{aligned}
E_{out} &= \sigma^2\left(1 + \frac{1}{N}trace\left(I\right) + \frac{c}{N}\right) \text{ where } c \text{ is a constant} \\
&= \sigma^2\left(1 + \frac{d+1}{N} + O\left(\frac{1}{N}\right)\right) \qquad (20)
\end{aligned}$$