MSE of ML estimator

DISCLAIMER: THIS DOCUMENT HAS NOT BEEN THOROUGHLY CHECKED FOR ACCURACY. PLEASE PROCEED WITH CAUTION. YOU CAN USE THIS AS A ROUGH OUTLINE UNTIL THIS DISCLAIMER IS REMOVED.

Given a data-point $\mathbf{x}_i$, we have:

$$y_i = \mathbf{w}^T \mathbf{x}_i + \epsilon_i$$

Here, $\epsilon_i \sim \mathcal{N}\left(0, \sigma^2\right)$. Also, we assume that ϵ_i and ϵ_j to be independent, hence cov$(\epsilon_i, \epsilon_j) = 0$. Now, we "treat" $\mathbf{x}_i$ as fixed and y_i as a random variable that is governed by the following conditional distribution:

$$y_i \mid \mathbf{x}_i \ \sim \mathcal{N}\left(\mathbf{w}^T \mathbf{x}_i, \sigma^2\right)$$

Here, $\mathbf{w}$ is also fixed. But the difference between $\mathbf{w}$ and $\mathbf{x}_i$ is that $\mathbf{x}_i$ is known and $\mathbf{w}$ is unknown. We can add all the y_is into a random vector $\mathbf{y} = \begin{bmatrix} y_1 & \cdots & y_n \end{bmatrix}^T$. The conditional distribution of this random vector given the data-matrix is:

$$\mathbf{y} \mid \mathbf{X} \sim \mathcal{N}\left(\mathbf{X}^T \mathbf{w}, \sigma^2 \mathbf{I}\right)$$

We wish to estimate $\mathbf{w}$. The ML (maximum likelihood) estimator of $\mathbf{w}$ is $\widehat{\mathbf{w}}$ and given by:

$$\widehat{\mathbf{w}} = \left(\mathbf{X}\mathbf{X}^T\right)^{-1} \mathbf{X}\mathbf{y}$$

Here, we are deriving the result for the special case of $\mathbf{X}\mathbf{X}^T$ being invertible. This happens when the rows of $\mathbf{X}$ are linearly independent, that is, when there is no linear dependence among the features. The estimator $\widehat{\mathbf{w}}$ is also a random vector since it is a function of the random vector $\mathbf{y}$. The estimator will turn into an estimate when we replace the random vector $\mathbf{y}$ with its realization. Let us first compute some useful quantities. Using the linearity of expectation:

$$\begin{aligned} E[\widehat{\mathbf{w}}] &= \left[\left(\mathbf{X}\mathbf{X}^T\right)^{-1} \mathbf{X}\right] E[\mathbf{y}] \\ &= \left(\mathbf{X}\mathbf{X}^T\right)^{-1} \mathbf{X}\mathbf{X}^T \mathbf{w} \\ &= \mathbf{w} \end{aligned}$$

Since $E[\widehat{\mathbf{w}}] = \mathbf{w}$, we have an unbiased estimator. Since the bias is zero, the MSE actually captures the variance in the estimator:

$$
\begin{aligned}
E\left[||\widehat{\mathbf{w}}-\mathbf{w}||^2\right] &= E\left[||\widehat{\mathbf{w}}-E[\widehat{\mathbf{w}}]||^2\right] \\
&= \texttt{trace}\Big(\texttt{cov}(\widehat{\mathbf{w}})\Big)
\end{aligned}
$$

The trace of the covariance matrix is the sum of the variances of the d components of the random vector $\widehat{\mathbf{w}}$. Let $\mathbf{A} = \left(\mathbf{X}\mathbf{X}^T\right)^{-1}\mathbf{X}$, then $\widehat{\mathbf{w}} = \mathbf{A}\mathbf{y}$. We note the following facts:

- $\mathbf{A}\mathbf{X}^T = \mathbf{I}$
- $\mathbf{X}\mathbf{A}^T = \mathbf{I}$
- $\widehat{\mathbf{w}}\widehat{\mathbf{w}}^T = \mathbf{A}\mathbf{y}\mathbf{y}^T\mathbf{A}^T$

Now, let us compute the covariance matrix $\texttt{cov}(\widehat{\mathbf{w}})$:

$$
\texttt{cov}(\widehat{\mathbf{w}}) = E\left[\widehat{\mathbf{w}}\widehat{\mathbf{w}}^T\right] - E[\widehat{\mathbf{w}}]E[\widehat{\mathbf{w}}]^T
$$

We will again use the linearity of expectation at several places. We will also use the following fact:

$$
\begin{aligned}
\texttt{cov}(\mathbf{y}) &= E\left[\mathbf{y}\mathbf{y}^T\right] - E[\mathbf{y}]E[\mathbf{y}]^T \\
\sigma^2\mathbf{I} &= E\left[\mathbf{y}\mathbf{y}^T\right] - \mathbf{X}^T\mathbf{w}\mathbf{w}^T\mathbf{X}
\end{aligned}
$$

Now, we continue to expand the RHS of the covariance matrix for $\widehat{\mathbf{w}}$:

$$
\begin{aligned}
E\left[\widehat{\mathbf{w}}\widehat{\mathbf{w}}^T\right] - E[\widehat{\mathbf{w}}]E[\widehat{\mathbf{w}}]^T &= E\left[\mathbf{A}\mathbf{y}\mathbf{y}^T\mathbf{A}^T\right] - \mathbf{w}\mathbf{w}^T \\
&= \mathbf{A}E\left[\mathbf{y}\mathbf{y}^T\right]\mathbf{A}^T - \mathbf{w}\mathbf{w}^T \\
&= \mathbf{A}\left[\texttt{cov}(\mathbf{y}) + E[\mathbf{y}]E[\mathbf{y}]^T\right]\mathbf{A}^T - \mathbf{w}\mathbf{w}^T \\
&= \mathbf{A}\left[\sigma^2\mathbf{I} + \mathbf{X}^T\mathbf{w}\mathbf{w}^T\mathbf{X}\right]\mathbf{A}^T - \mathbf{w}\mathbf{w}^T \\
&= \sigma^2 \cdot \mathbf{A}\mathbf{A}^T + \mathbf{A}\mathbf{X}^T\left(\mathbf{w}\mathbf{w}^T\right)\mathbf{X}\mathbf{A}^T - \mathbf{w}\mathbf{w}^T \\
&= \sigma^2 \cdot \mathbf{A}\mathbf{A}^T + \mathbf{w}\mathbf{w}^T - \mathbf{w}\mathbf{w}^T \\
&= \sigma^2 \cdot \mathbf{A}\mathbf{A}^T \\
&= \sigma^2\left[\left(\mathbf{X}\mathbf{X}^T\right)^{-1}\mathbf{X}\right]\left[\mathbf{X}^T\left(\mathbf{X}\mathbf{X}^T\right)^{-1}\right] \\
&= \sigma^2\left(\mathbf{X}\mathbf{X}^T\right)^{-1}
\end{aligned}
$$

We can now compute the MSE as:

$$
\begin{aligned}
E\left[||\widehat{\mathbf{w}}-\mathbf{w}||^2\right] &= E\left[||\widehat{\mathbf{w}}-E[\widehat{\mathbf{w}}]||^2\right] \\
&= \texttt{trace}\Big(\texttt{cov}(\widehat{\mathbf{w}})\Big) \\
&= \sigma^2\texttt{trace}\left[\left(\mathbf{X}\mathbf{X}^T\right)^{-1}\right]
\end{aligned}
$$