Question 1: Select ALL correct choices.

Question 2: A linear ML model can be written as:

$$f(\mathbf{x}, \mathbf{w}) = \sum_{i=0}^{n} w_i x_i = \mathbf{w}^T \mathbf{x}$$

The loss function can be written as:

$$J(\mathbf{w}) = \frac{1}{m} \sum_{i=1}^{n} \left[f(\mathbf{x}^{(i)}, \mathbf{w}) - y^{(i)} \right]^2$$

[2.1] Show analytically that the optimal weight vector that minimizes the cost function $J(\mathbf{w})$ is:

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

Solution.

Since our model is linear, we can write the cost function as:

$$J(\mathbf{w}) = \frac{1}{m} \sum_{i=1}^{n} \left[\mathbf{w}^T \mathbf{x}^{(i)} - y^{(i)} \right]^2 \tag{1}$$

The product $\mathbf{w}^T \mathbf{x}^{(i)} = \mathbf{X}\mathbf{w}, \forall i \in \{1, 2, \cdots, n\}$ since the LHS implies the matrix mulplitcation of the RHS, as $\mathbf{X}$ is the matrix of all input entries $\mathbf{x}^{(i)}$. So, Eq. 1 can be written as:

$$J(\mathbf{w}) = \frac{1}{m} \left[\mathbf{X}\mathbf{w} - \mathbf{y} \right]^2 \tag{2}$$

The inside of the brackets is just a vector, and the square of a vector is the norm of a vector, so we can reduce Eq. 2:

$$\begin{aligned} J(\mathbf{w}) &= \frac{1}{m} \|\mathbf{X}\mathbf{w} - \mathbf{y}\| &(3)\\ &= \frac{1}{m} (\mathbf{X}\mathbf{w} - \mathbf{y})^T (\mathbf{X}\mathbf{w} - \mathbf{y}) &(4)\\ &= \frac{1}{m} (\mathbf{w}^T \mathbf{X}^T \mathbf{X}\mathbf{w} - \mathbf{w}^T \mathbf{X}^T \mathbf{y} - \mathbf{y}^T \mathbf{X}\mathbf{w} + \mathbf{y}^T \mathbf{y}) &(5) \end{aligned}$$

Now to optimize the cost with respect to weights, we can take the gradient of $J(\mathbf{w})$ w.r.t. the weights $\mathbf{w}^{(i)}$.

$$\begin{aligned} \nabla_{\mathbf{w}} J(\mathbf{w}) &= \frac{1}{m} \nabla_{\mathbf{w}} (\mathbf{w}^T \mathbf{X}^T \mathbf{X}\mathbf{w} - \mathbf{w}^T \mathbf{X}^T \mathbf{y} - \mathbf{y}^T \mathbf{X}\mathbf{w} + \mathbf{y}^T \mathbf{y}) &(6)\\ &= \frac{1}{m} \nabla_{\mathbf{w}} (\mathbf{w}^T \mathbf{X}^T \mathbf{X}\mathbf{w} - \mathbf{w}^T \mathbf{X}^T \mathbf{y} - \mathbf{y}^T \mathbf{X}\mathbf{w}) &(7)\\ &= \frac{1}{m} [\nabla_{\mathbf{w}} (\mathbf{w}^T \mathbf{X}^T \mathbf{X}\mathbf{w}) - \nabla_{\mathbf{w}} (\mathbf{w}^T \mathbf{X}^T \mathbf{y}) - \nabla_{\mathbf{w}} (\mathbf{y}^T \mathbf{X}\mathbf{w}))] &(8) \end{aligned}$$

The gradients $\nabla_{\mathbf{w}}$ are simply derivatives of each matrix function w.r.t. $\mathbf{w}$, which can be computed using equations (69) and (81) from *The Matrix Cookbook* [2].

$$= \frac{1}{m}\left[\frac{\partial}{\partial \mathbf{w}}(\mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w}) - \frac{\partial}{\partial \mathbf{w}}(\mathbf{w}^T\mathbf{X}^T\mathbf{y}) - \frac{\partial}{\partial \mathbf{w}}(\mathbf{y}^T\mathbf{X}\mathbf{w}))\right] \tag{9}$$

$$= \frac{1}{m}((\mathbf{X}^T\mathbf{X} + (\mathbf{X}^T\mathbf{X})^T)\mathbf{w} - \mathbf{X}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}) \tag{10}$$

$$= \frac{1}{m}((\mathbf{X}^T\mathbf{X} + \mathbf{X}^T(\mathbf{X}^T)^T)\mathbf{w} - 2\mathbf{X}^T\mathbf{y}) \tag{11}$$

$$= \frac{1}{m}((\mathbf{X}^T\mathbf{X} + \mathbf{X}^T\mathbf{X})\mathbf{w} - 2\mathbf{X}^T\mathbf{y}) \tag{12}$$

$$= \frac{1}{m}(2\mathbf{X}^T\mathbf{X}\mathbf{w} - 2\mathbf{X}^T\mathbf{y}) \tag{13}$$

$$= \frac{2}{m}(\mathbf{X}^T\mathbf{X}\mathbf{w} - \mathbf{X}^T\mathbf{y}) \tag{14}$$

We find the minimum when $\nabla_{\mathbf{w}}J(\mathbf{w}) = 0$,

$$\frac{2}{m}(\mathbf{X}^T\mathbf{X}\mathbf{w} - \mathbf{X}^T\mathbf{y}) = 0 \tag{15}$$

$$\mathbf{X}^T\mathbf{X}\mathbf{w} - \mathbf{X}^T\mathbf{y} = 0 \tag{16}$$

$$\mathbf{X}^T\mathbf{X}\mathbf{w} = \mathbf{X}^T\mathbf{y} \tag{17}$$

$$(\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{X}\mathbf{w}) = (\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{y}) \tag{18}$$

$$\mathbf{I}\mathbf{w} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{19}$$

$$\boxed{\mathbf{w} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{w}^*} \tag{20}$$

Without advanced methods, multipliying two $(n \times n)$ matricies is of $O(n^3)$ complexity [1]. For very large data sets, computing $\mathbf{w}$ would be enormously computationally expensive.

References

[1] Andy He and Evan Williams. Computational complexity of matrix multiplication. `https://www.cs.cornell.edu/courses/cs6810/2023fa/Matrix.pdf`, Fall 2023. Accessed: 2024-09-10.

[2] Kaare Brandt Petersen and Michael Syskind Pederson. The matrix cookbook. Distributed by University of Waterloo, November 2012.