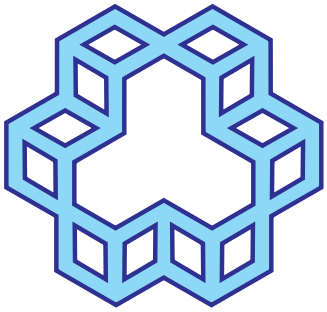

K. N. Toosi University of Technology

Faculty of Physics
Educational Group of
Atomic-Molecular and Astronomy

# Answer to Question 2
# (Minimizer in Linear Regression)

by
Ali Bagheri

Teacher
Dr. Mohammad Hossein Zhoolideh

Academic Year 1401-1402
(First Semester)

## Question

Show that the minimizer for least-squares linear regression with $L_2$ regularization is $\mathbf{w} = \left(\lambda\mathbf{I} + \mathbf{\Phi}^T\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^T\mathbf{t}$.

**Note:** To solve this question, you can use chapter 3 of the Pattern Recognition and Machine Learning (Bishop) book.

### Answer

In Section 1.1 of the Pattern Recognition and Machine Learning (Bishop) book, the idea of adding an regularization term to an error function in order to control the overfitting is introduced, so that the total error function to be minimized takes the form

$$J(\mathbf{w}) = E_D(\mathbf{w}) + \lambda E_W(\mathbf{w})$$

where $\lambda$ is the regularization coefficient that controls the relative importance of the data-dependent error ($E_D(\mathbf{w})$) and the regularization term ($E_W(\mathbf{w})$). One of the simplest forms of regularizer is given by the sum-of-squares of the weight vector elements

$$E_W(\mathbf{w}) = \frac{1}{2}\mathbf{w}^T\mathbf{w}$$

If we also consider the sum-of-squares error function given by

$$E_D(\mathbf{w}) = \frac{1}{2}\left(\mathbf{t} - \mathbf{w}\mathbf{\Phi}\right)^2,$$

then the total error function becomes

$$\begin{aligned} J(\mathbf{w}) &= E_D(\mathbf{w}) + \lambda E_W(\mathbf{w}) \\ &= \frac{1}{2}\left(\mathbf{t} - \mathbf{w}\mathbf{\Phi}\right)^2 + \frac{\lambda}{2}\mathbf{w}^T\mathbf{w} \end{aligned}$$

Now we take the derivative of this equation in terms of $\mathbf{w}$ and set it equal to zero

$$\begin{aligned} \frac{\partial}{\partial\mathbf{w}}J(\mathbf{w}) &= -\left(\mathbf{t} - \mathbf{w}\mathbf{\Phi}\right)\mathbf{\Phi} + \lambda\mathbf{w} \\ &= 0 \end{aligned}$$

As a result, we have

$$\begin{aligned} \mathbf{w} &= (\lambda\mathbf{I} + \mathbf{\Phi}^2)^{-1}\mathbf{\Phi}^T\mathbf{t} \\ &= \left(\lambda\mathbf{I} + \mathbf{\Phi}^T\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^T\mathbf{t} \end{aligned}$$

# A Proof of some important equations

$$A = \begin{pmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \dots & a_{nn} \end{pmatrix}_{n\times n}$$

$$B = \begin{pmatrix} b_{11} & \dots & b_{1n} \\ \vdots & \ddots & \vdots \\ b_{n1} & \dots & b_{nn} \end{pmatrix}_{n\times n}$$

$$AB = \begin{pmatrix} a_{11}b_{11} + \dots + a_{1n}b_{n1} & \dots & a_{11}b_{1n} + \dots + a_{1n}b_{nn} \\ \vdots & \ddots & \vdots \\ a_{n1}b_{11} + \dots + a_{nn}b_{n1} & \dots & a_{n1}b_{1n} + \dots + a_{nn}b_{nn} \end{pmatrix}_{n\times n}$$

$$= \begin{pmatrix} \sum_{j=1}^{n} a_{1j}b_{j1} & \dots & \sum_{j=1}^{n} a_{1j}b_{jn} \\ \vdots & \ddots & \vdots \\ \sum_{j=1}^{n} a_{nj}b_{j1} & \dots & \sum_{j=1}^{n} a_{nj}b_{jn} \end{pmatrix}_{n\times n}$$

$$tr(AB) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}b_{ji} \equiv f(AB)$$

$$\nabla_A f(AB) = \begin{pmatrix} \frac{\partial f(AB)}{\partial a_{11}} & \dots & \frac{\partial f(AB)}{\partial a_{1n}} \\ \vdots & \ddots & \vdots \\ \frac{\partial f(AB)}{\partial a_{n1}} & \dots & \frac{\partial f(AB)}{\partial a_{nn}} \end{pmatrix}$$

$$= \begin{pmatrix} b_{11} & \dots & b_{n1} \\ \vdots & \ddots & \vdots \\ b_{1n} & \dots & b_{nn} \end{pmatrix} = B^T$$

$$\rightarrow \nabla_A f(AB) = B^T \tag{1}$$

$$A = \begin{pmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \dots & a_{nn} \end{pmatrix}_{n \times n}$$

$$A^T = \begin{pmatrix} a_{11} & \dots & a_{n1} \\ \vdots & \ddots & \vdots \\ a_{1n} & \dots & a_{nn} \end{pmatrix}_{n \times n}$$

$$\nabla_{A^T} f(A) = \begin{pmatrix} \frac{\partial f(A)}{\partial a_{11}} & \dots & \frac{\partial f(A)}{\partial a_{n1}} \\ \vdots & \ddots & \vdots \\ \frac{\partial f(A)}{\partial a_{1n}} & \dots & \frac{\partial f(A)}{\partial a_{nn}} \end{pmatrix}_{n \times n}$$

$$\nabla_A f(A) = \begin{pmatrix} \frac{\partial f(A)}{\partial a_{11}} & \dots & \frac{\partial f(A)}{\partial a_{1n}} \\ \vdots & \ddots & \vdots \\ \frac{\partial f(A)}{\partial a_{n1}} & \dots & \frac{\partial f(A)}{\partial a_{nn}} \end{pmatrix}_{n \times n}$$

$$\rightarrow \nabla_{A^T} f(A) = \left(\nabla_A f(A)\right)^T \tag{2}$$

In the same way, we can prove the following equation:

$$\nabla_A tr\left(ABA^TC\right) = CAB + C^TAB^T \tag{3}$$

Combining Equations (2) and (3), we find that

$$\begin{aligned} \nabla_{A^T} tr\left(ABA^TC\right) &= \left(\nabla_A tr\left(ABA^TC\right)\right)^T \\ &= \left(CAB + C^TAB^T\right)^T \\ &= B^TA^TC^T + BA^TC \end{aligned} \tag{4}$$

$$
\begin{aligned}
\nabla_\theta J(\theta) &= \nabla_\theta \frac{1}{2}(X_{(n\times m)}\theta_{(m\times 1)} - \vec{y}_{(n\times 1)})^T (X_{(n\times m)}\theta_{(m\times 1)} - \vec{y}_{(n\times 1)}) \\
&= \frac{1}{2}\nabla_\theta \Big( \theta^T_{(1\times m)} X^T_{(m\times n)} X_{(n\times m)} \theta_{(m\times 1)} \\
&\qquad - \theta^T_{(1\times m)} X^T_{(m\times n)} \vec{y}_{(n\times 1)} \\
&\qquad - \vec{y}^T_{(1\times n)} X_{(n\times m)} \theta_{(m\times 1)} \\
&\qquad + \vec{y}^T_{(1\times n)} \vec{y}_{(n\times 1)} \Big)
\end{aligned} \tag{5}
$$

The above equation is in the form of the following equation:

$$
\frac{1}{2}\nabla_\theta \Big( A_{(1\times 1)} - B_{(1\times 1)} - B^T_{(1\times 1)} + C_{(1\times 1)} \Big)
$$

In fact, the result of a $1 \times 1$ matrix is a polynomial expression, and since each polynomial expression is equal to its transpose and also the trace of each polynomial expression is equal to that expression itself, the equation can be written as follows:

$$
\begin{aligned}
\nabla_\theta J(\theta) &= \frac{1}{2}\nabla_\theta tr\left(\theta^T X^T X\theta - 2\vec{y}^T X\theta + \vec{y}^T\vec{y}\right) \\
&= \frac{1}{2}\nabla_\theta \left( tr\left(\theta^T X^T X\theta\right) - 2tr\left(\vec{y}^T X\theta\right) + tr\left(\vec{y}^T\vec{y}\right)\right) \\
&= \frac{1}{2}\left( \nabla_\theta tr\left(\theta^T X^T X\theta\right) - 2\nabla_\theta tr\left(\vec{y}^T X\theta\right) + \nabla_\theta tr\left(\vec{y}^T\vec{y}\right)\right) \\
&= \frac{1}{2}\left( \nabla_\theta tr\left(\theta^T X^T X\theta\right) - 2\nabla_\theta tr\left(\vec{y}^T X\theta\right)\right)
\end{aligned}
$$

From equations (1) and (4), we know:

$$
\begin{aligned}
\nabla_{A^T} tr\left(ABA^TC\right) &= B^T A^T C^T + BA^T C \\
\xrightarrow{A^T=\theta,\, B=B^T=X^TX,\, C=I} \nabla_\theta tr\left(\theta^T X^T X\theta\right) &= X^T X\theta I + X^T X\theta I \\
&= X^T X\theta + X^T X\theta
\end{aligned}
$$

$$
\begin{aligned}
\nabla_A tr\left(AB\right) &= \nabla_A tr\left(BA\right) = B^T \\
\xrightarrow{A=\theta,\, B=\vec{y}^T X} \nabla_\theta tr\left(\vec{y}^T X\theta\right) &= X^T\vec{y}
\end{aligned}
$$

As a result, we have:

$$
\begin{aligned}
\nabla_\theta J(\theta) &= \frac{1}{2}\left( X^T X\theta + X^T X\theta - 2X^T\vec{y}\right) \\
&= X^T X\theta - X^T\vec{y}
\end{aligned} \tag{6}
$$

By comparing equation (5) with equation (6), we find that in the matrix derivative, when we want to take a derivative with respect to $\theta$, $\theta^T$ is the same as $\theta$ itself.