1 Summary

Let's start from linear regression with L2 regularization [1].

$$\boldsymbol{y} = \boldsymbol{w}^T \phi(\boldsymbol{x})$$

Note that $\phi(\boldsymbol{x}) = X$ where X is a $m \times D$ matrix where each training sample has m features.

To choose the optimal $\boldsymbol{w}$, we minimize the sum of squares:

$$\boldsymbol{w}^* = \arg\min_{\boldsymbol{w}} ||\boldsymbol{y} - \boldsymbol{w}^T \phi(\boldsymbol{x})||^2$$

To avoid overfitting, we add the regularization parameter:

$$\boldsymbol{w}^* = \arg\min_{\boldsymbol{w}} ||\boldsymbol{y} - \boldsymbol{w}^T \phi(\boldsymbol{x})||^2 + \lambda ||\boldsymbol{w}||^2$$

Lets say that $L(\boldsymbol{w}) = ||\boldsymbol{y} - \boldsymbol{w}^T \phi(\boldsymbol{x})||^2 + \lambda ||\boldsymbol{w}||^2$

What is this derivative? We will compute the following gradient:

$$\nabla_{\boldsymbol{w}} L(\boldsymbol{w}) = \frac{\partial L(\boldsymbol{w})}{\partial \boldsymbol{w}} = (\frac{\partial L(\boldsymbol{w})}{\partial w_1}, ..., \frac{\partial L(\boldsymbol{w})}{\partial w_i})$$

Let's go look into one element of this gradient $\nabla_{\boldsymbol{w}}$:

$$\frac{\partial L(\boldsymbol{w})}{\partial w_i} = \frac{\partial(||\boldsymbol{y} - \boldsymbol{w}^T \phi(\boldsymbol{x})||^2 + \lambda ||\boldsymbol{w}||^2)}{\partial w_i}$$

Keep in mind that

$$\begin{aligned}
||\boldsymbol{y} - X\boldsymbol{w}||^2 &= \boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{y}^T X\boldsymbol{w} - (X\boldsymbol{w})^T\boldsymbol{y} + (X\boldsymbol{w})^T(X\boldsymbol{w}) && (1)\\
&= \boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{y}^T X\boldsymbol{w} + (X\boldsymbol{w})^T(X\boldsymbol{w}) && (2)\\
&= \sum_i y_i^2 - 2\boldsymbol{y}^T X\boldsymbol{w} + \sum_j (\boldsymbol{x}_j\boldsymbol{w})^2 && (3)\\
&= \sum_i y_i^2 - 2\sum_i a_i w_i + \sum_j (\sum_i x_{ji} w_i)^2 && (4)\\
&= \sum_i y_i^2 - 2\sum_i a_i w_i + \sum_j (x_{j1}w_1 + ... + x_{ji}w_i)^2 && (5)
\end{aligned}$$

where $\boldsymbol{a} = \boldsymbol{y}^T X$, $a_i = \boldsymbol{y}^T X_i$

Since $\forall k$ s.t. $k \neq i$, $\frac{\partial w_k}{\partial w_i} = 0$,

$$\frac{\partial L(\boldsymbol{w})}{\partial w_i} = -2a_i - \frac{\partial \sum_j ((x_{ji}w_i)(x_{j1}w_1 + ... + x_{ji-1}w_{i-1}) + (x_{j1}w_1 + ... + x_{ji-1}w_{i-1})(x_{ji}w_i) + (x_{ji}w_i)^2)}{\partial w_i} \quad (6)$$

$$= -2a_i - \sum_j ((x_{ji})(x_{j1}w_1 + ... + x_{ji-1}w_{i-1}) + (x_{j1}w_1 + ... + x_{ji-1}w_{i-1})(x_{ji}) + (2x_{ji}^2 w_i)) \quad (7)$$

$$= -2a_i - \sum_j 2((x_{ji})(x_{j1}w_1 + ... + x_{ji-1}w_{i-1}) + (2x_{ji}^2 w_i)) \quad (8)$$

$$= -2(a_i - \sum_j ((x_{ji})(x_{j1}w_1 + ... + x_{ji-1}w_{i-1}) + (x_{ji}^2 w_i))) \quad (9)$$

$$= -2(a_i - \sum_j ((x_{ji})(x_{j1}w_1 + ... + x_{ji-1}w_{i-1} + x_{ji}w_i))) \quad (10)$$

$$= -2(a_i - \sum_j ((x_{ji})(x_{j1}, ..., x_{ji-1}, x_{ji})\boldsymbol{w}) \quad (11)$$

$$= -2(\boldsymbol{y}^T X_i - X_j^T X_i \boldsymbol{w}) \quad (12)$$

In the matrix representation:

$$\nabla_{\boldsymbol{w}} L(\boldsymbol{w}) = -2(\boldsymbol{y}^T X - X^T X w)$$

Also in general:

$$\nabla_{\boldsymbol{w}} w^T X^T X w = 2X^T X w$$

Note that $X^T X$ is symmetric and it is part of the common matrix derivative pattern[1]

References

[1] Lecture 11: Regularization. `http://grandmaster.colorado.edu/~cketelsen/files/csci5622/videos/lesson11/lesson11.pdf`. [Online; accessed 19-Nov-2016].

[1] `https://en.wikipedia.org/wiki/Matrix_calculus#Scalar-by-vector_identities`