Deriving the normal equation for linear regression page 109 section 5.1.4

The normal equation is once again mentioned on page 172 chapter 6.1 in the XOR example, but I will show the one mentioned on page 109 to give context in chapter 6. The author doesn't show the full derivation on page 109 chapter 5.1.4, instead just applies the formula for $\frac{\partial x^T A x}{\partial x}$ after using the foil rule to distribute the terms, which is fine and makes the derivation very short. But it's nice to know where things come from, and deriving it by yourself is also instructive. With that said, there are apparently different ways to go about it. I will present them to show different techniques and perspectives. At the end I will show a very simplified approach that I learnt from the MIT course Matrix Calculus 18.S096.

Instead of rewriting all the equations and setting up the stage, I will just refer to the original equations on page 109 chapter 5.1.4. Note that $X^{(\text{train})}$ is a matrix of the training dataset (features are columns and data points are rows), $y^{(\text{train})}$ is a vector (the labels), and w is the weight vector for the linear regression. On page 109 they have on the second line in the derivation of the normal equation

$$\nabla_w \left[w^T X^{(\text{train})T} X^{(\text{train})} w - 2 w^T X^{(\text{train})T} y^{(\text{train})} + y^{(\text{train})T} y^{(\text{train})} \right] = 0$$

To go from the first line to the second line above, just distribute the terms. Note that the cross term $-2 w^T X^{(\text{train})T} y^{(\text{train})}$ is easily obtained by observing that $-\left(X^{(\text{train})} w\right)^T y^{(\text{train})}$ and $-y^{(\text{train})T} X^{(\text{train})} w$ can both be converted to the same by taking the transpose. This is because $X^{(\text{train})}$ is a matrix, $y^{(\text{train})}$ is a vector, and w is a vector, so both expressions result in a scalar, and taking the transpose of a scalar doesn't change the original value. We therefore have $-\left(y^{(\text{train})T} X^{(\text{train})} w\right)^T = -w^T X^{(\text{train})T} y^{(\text{train})}$, and adding them together gives $-2 w^T X^{(\text{train})T} y^{(\text{train})}$.

The only non-trivial part of the equation that needs some exploring is the first term $w^T X^{(\text{train})T} X^{(\text{train})} w = w^T A w$, where $A = X^{(\text{train})T} X^{(\text{train})}$. If we let $x = w$, then taking the derivative with respect to x leads us to proving this well-known relation:

$$\frac{\partial (x^T A x)}{\partial (x)} = (A + A^T) x$$

I will show three ways, of which two are basically the same. The first is the tedious element wise way, which requires some mental bookkeeping, and the only reason I'm showing it is to demonstrate how tedious it is compared to the other methods.

Scalar sum approach - working element wise

We write the quadratic form as (easy to see, Ax is just the dot product, and then we do another dot product between x^T and Ax)

$$x^T A x = \sum_j^n x_j \sum_i^n x_i A_{ji}$$

and take the derivative w.r.t x_k. Using the product rule we get

$$\begin{aligned} \frac{\partial x^T A x}{\partial x_k} &= \sum_j^n \frac{\partial x_j}{\partial x_k} \sum_i^n x_i A_{ji} + \sum_j^n x_j \sum_i^n \frac{\partial x_i}{\partial x_k} A_{ji} \\ &= \sum_i^n x_i A_{ki} + \sum_j^n x_j A_{jk} \end{aligned}$$

The first term $\sum_j^n \frac{\partial x_j}{\partial x_k} \sum_i^n x_i A_{ji}$ will be zero everywhere except when $j = k$ in the outer sum so that $x_j = x_k$ and $A_{ji} = A_{ki}$. This means only when the outer sum $j = k$, will we have a non-zero term, and multiplying this instance with the inner sum will get us the result above, that's why the inner sum remains the same because it's not affected, only the outer sum. The second term $\sum_j^n x_j \sum_i^n \frac{\partial x_i}{\partial x_k} A_{ji}$ will be zero everywhere except when $i = k$ in the inner sum. The outer sum remains the same because it's unaffected.

Now putting this into a column vector we get

$$\begin{bmatrix} \sum_i^n x_i A_{1i} + \sum_j^n x_j A_{j1} \\ \sum_i^n x_i A_{2i} + \sum_j^n x_j A_{j2} \\ \vdots \\ \sum_i^n x_i A_{ni} + \sum_j^n x_j A_{jn} \\ \end{bmatrix} = Ax + \left(x^T A\right)^T = (A + A^T)x$$

The $\sum_j^n x_j A_{j1}$ can be a little unclear, but try to draw this. It is the dot product between the row vector x^T and the 1st column vector in A denoted as $A_{:1}$, where : means all the rows. Putting this together we obviously get $x^T A$, but because we need to make them compatible to sum then we transpose them $\left(x^T A\right)^T = A^T x$ so that $Ax + A^T x = (A + A^T)x$. If A is symmetric meaning $A^T = A$ then we get $A + A^T = A + A = 2A$ and therefore $(A + A^T)x = 2Ax$ in the end.

Differential approach

Differential approach is nicer, as we don't have to tediously work with the matrix or vectors element wise. No need to keep track of the content, just treat the matrices and vectors as objects or functions as a whole, not as tables of elements. As you might have seen in the previous approach, you really need to be aware of the contents of the matrix or vectors, which can lead to some headache if the structure is even more complicated.

Recall from the definition of derivative that we have

$$f'(x) = \lim_{\delta x \to 0} \frac{f(x + \delta x) - f(x)}{\delta x}$$

where δx is a finite change in x. Moving δx to the left side gives

$$f'(x)\delta x \approx f(x + \delta x) - f(x) - \text{higher order terms}$$

where the higher order terms is added to approximate the infinitely close gap between the RHS and LHS (because of the limit), which is just the change in function f. But we can always ignore the higher order terms since they decay faster than the linear term δx when it goes to zero, which practically means for small numbers the higher order terms have no effect. If we now make use of the idea of differentials, which encodes the limit going to zero implicitly as an arbitrarily small quantity, we can write the differential df exact like so

$$df = f(x + dx) - f(x) = f'(x)dx$$

which is interpreted as the change in output function f denoted as df by the change in input x denoted as dx multiplied by the derivative $f'(x)$ that tells us the rate of change in x. Notice that we have an equality in this relation instead of approximation. This relation can be extended to vectors and matrices, where we just let x be a vector or matrix. We can now use the differential formula df to compute the derivative of $x^T Ax$. If we let

$$f(x) = x^T Ax$$

then we get the following derivation:

$$\begin{aligned}
df &= f(x+dx) - f(x) \\
&= \text{substitute } \ x^T Ax \ \text{ into the expression} \\
&= (x+dx)^T A(x+dx) - x^T Ax \\
&= \text{use foil rule to distribute the terms} \\
&= (x^T Ax) + (x^T Adx) + dx^T Ax + dx^T Adx - x^T Ax \\
&= \text{use } \ \left(dx^T Ax\right)^T = x^T A^T dx \\
&= \text{and note that } \ dx^T Adx \ \text{ is higher order term (quadratic in dx) so ignore} \\
&= x^T Adx + x^T A^T dx \\
&= x^T (A + A^T) dx \\
&= f'(x)dx
\end{aligned}$$

Notice that the differential formula says $df = f'(x)dx$, and since we have that exact form at the last line we can pattern match and extract the derivative which is $f'(x) = x^T(A + A^T) = (A + A^T)x$. We can easily transpose between them because the result is a vector, so it doesn't really change the content of it. The reason to drop the higher order term $dx^T Adx$ is because when $dx \to 0$, it will decrease faster to zero than dx, meaning it will not have any effect on the result, therefore we can skip the higher order terms in dx.

Differential approach with product rule

Omitting the proof, but one can show that the differential product rule is

$$f(x) = g(x)h(x) \implies df = dgh + gdh$$

which is the exact same form as regular product rule for derivatives, except here we are working with differentials. To prove this just apply the differential approach to the product gh. Using this rule for differential and recalling that the differential formula is

$$df = f(x+dx) - f(x) = f'(x)dx$$

we can proceed to use the product rule to derive the relation given that we have the function

$$f(x) = x^T Ax$$

we then get

$$df = dx^T(Ax) + x^T d(Ax) = dx^T(Ax) + x^T A\, dx = \{\text{tranpose one of them}\} = x^T A^T dx + x^T A\, dx = x^T(A^T + A)dx$$

Extracting the derivative out of that we get

$$f'(x) = \frac{\partial f}{\partial x} = x^T(A + A^T) = (A + A^T)x$$

And that's it. We can see here that we didn't need to use any fancy formulas, just our knowledge of basic derivative and product rule sufficed. We also didn't need to directly work with the elements of the matrices or vectors, just treat them as objects without looking into them saves us a ton of tedious bookkeeping.

Connecting the result to the normal equation

Since we said $x = w$, $A = X^{(\text{train})\,T} X^{(\text{train})}$ and $\left(X^{(\text{train})\,T} X^{(\text{train})}\right)^T = X^{(\text{train})\,T} X^{(\text{train})}$ is symmetric we therefore get

$$(A + A^T)x = 2Ax = 2X^{(\text{train})\,T} X^{(\text{train})} w$$

Moreover, the second term in

$$\nabla w\big(w^T X^{(\text{train})\,T} X^{(\text{train})} w - 2w^T X^{(\text{train})\,T} y^{(\text{train})} + y^{(\text{train})\,T} y^{(\text{train})}\big) = 0$$

gives after taking the derivative w.r.t w

$$\nabla w\big(-2w^T X^{(\text{train})\,T} y^{(\text{train})}\big) = -2X^{(\text{train})\,T} y^{(\text{train})}$$

We can once again prove this with the differential product rule approach without having to directly look at the elements of the vectors/matrices. Let

$$f(w) = -2w^T X^{(\text{train})\,T} y^{(\text{train})}$$

then we get the following using product rule w.r.t w

$$\begin{aligned}
df &= -2dw^T\big(X^{(\text{train})\,T} y^{(\text{train})}\big) - 2w^T d\big(X^{(\text{train})\,T} y^{(\text{train})}\big) \\
&= \{\text{transpose the first term and second term is zero because independent of w}\} \\
&= -2\big(X^{(\text{train})\,T} y^{(\text{train})}\big) dw \\
&= f'(w) dw
\end{aligned}$$

and we can extract the derivative directly, which is $-2X^{(\text{train})\,T} y^{(\text{train})}$ as earlier stated. Finally the third term in the equation

$$\nabla w\big(w^T X^{(\text{train})\,T} X^{(\text{train})} w - 2w^T X^{(\text{train})\,T} y^{(\text{train})} + y^{(\text{train})\,T} y^{(\text{train})}\big) = 0$$

gives

$$\nabla w\big(y^{(\text{train})\,T} y^{(\text{train})}\big) = 0$$

because it's independent of w so it will become zero when taking the derivative w.r.t w. Putting the results together and solving for w we get

$$\begin{aligned}
&2X^{(\text{train})\,T} X^{(\text{train})} w - 2X^{(\text{train})\,T} y^{(\text{train})} = 0 \\
&\implies X^{(\text{train})\,T} X^{(\text{train})} w - X^{(\text{train})\,T} y^{(\text{train})} = 0 \\
&\implies X^{(\text{train})\,T} X^{(\text{train})} w = X^{(\text{train})\,T} y^{(\text{train})} \\
\implies w &= \left(X^{(\text{train})\,T} X^{(\text{train})}\right)^{-1} \left(X^{(\text{train})\,T} y^{(\text{train})}\right)
\end{aligned}$$

which is the normal equation for linear regression presented in the book.