Deriving the Gradient and Hessian of Linear and Quadratic Functions in Matrix Notation

Mark Schmidt

February 6, 2019

1 Gradient of Linear Function

Consider a linear function of the form
\[ f(w) = a^T w, \]
where \( a \) and \( w \) are length-\( d \) vectors. We can derive the gradient in matrix notation as follows:

1. Convert to summation notation:
\[ f(w) = \sum_{j=1}^{d} a_j w_j, \]
where \( a_j \) is element \( j \) of \( a \) and \( w_j \) is element \( j \) of \( w \).

2. Take the partial derivative with respect to a generic element \( k \):
\[ \frac{\partial}{\partial w_k} \left[ \sum_{j=1}^{d} a_j w_j \right] = a_k. \]

3. Assemble the partial derivatives into a vector:
\[ \nabla f(w) = \begin{bmatrix} \frac{\partial}{\partial w_1} \\ \frac{\partial}{\partial w_2} \\ \vdots \\ \frac{\partial}{\partial w_d} \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_d \end{bmatrix}. \]

4. Convert to matrix notation:
\[ \nabla f(w) = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_d \end{bmatrix} = a. \]

So our final result is that
\[ \nabla f(w) = a. \]

This generalizes the scalar case where \( \frac{d}{dw}[\alpha w] = \alpha \). We can also consider general linear functions of the form
\[ f(w) = a^T w + \beta, \]
for a scalar \( \beta \). But in this case we still have \( \nabla f(w) = a \) since the y-intercept \( \beta \) does not depend on \( w \).
2 Gradient of Quadratic Function

Consider a quadratic function of the form
\[ f(w) = w^T A w, \]
where \( w \) is a length-\( d \) vector and \( A \) is a \( d \) by \( d \) matrix. We can derive the gradeint in matrix notation as follows

1. Convert to summation notation:
\[ f(w) = w^T \begin{bmatrix} \sum_{j=1}^{n} a_{1j} w_j \\ \sum_{j=1}^{n} a_{2j} w_j \\ \vdots \\ \sum_{j=1}^{n} a_{dj} w_j \end{bmatrix} Aw = \sum_{i=1}^{d} \sum_{j=1}^{d} w_i a_{ij} w_j, \]
where \( a_{ij} \) is the element in row \( i \) and column \( j \) of \( A \). To help with computing the partial derivatives, it helps to re-write it in the form
\[ f(w) = \sum_{i=1}^{d} \sum_{j=1}^{d} w_i a_{ij} w_j = \sum_{i=1}^{d} (a_{ii} w_i^2 + \sum_{j \neq i} w_i a_{ij} w_j). \]

2. Take the partial derivative with respect to a generic element \( k \):
\[ \frac{\partial}{\partial w_k} \left[ \sum_{i=1}^{d} (a_{ii} w_i^2 + \sum_{j \neq i} w_i a_{ij} w_j) \right] = 2a_{kk} w_k + \sum_{j \neq k} w_j a_{jk} + \sum_{j \neq k} a_{kj} w_j. \]
The first term comes from the \( a_{kk} \) term that is quadratic in \( w_k \), while the two sums come from the terms that are linear in \( w_k \). We can move one \( a_{kk} w_k \) into each of the sums to simplify this to
\[ \frac{\partial}{\partial w_k} \left[ \sum_{i=1}^{d} (a_{ii} w_i^2 + \sum_{j \neq i} w_i a_{ij} w_j) \right] = \sum_{j=1}^{d} w_j a_{jk} + \sum_{j=1}^{d} a_{kj} w_j. \]

3. Assemble the partial derivatives into a vector:
\[ \nabla f(w) = \begin{bmatrix} \frac{\partial}{\partial w_1} \\ \frac{\partial}{\partial w_2} \\ \vdots \\ \frac{\partial}{\partial w_d} \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^{d} w_j a_{1j} + \sum_{j=1}^{d} a_{1j} w_j \\ \sum_{j=1}^{d} w_j a_{2j} + \sum_{j=1}^{d} a_{2j} w_j \\ \vdots \\ \sum_{j=1}^{d} w_j a_{dj} + \sum_{j=1}^{d} a_{dj} w_j \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^{d} w_j a_{1j} \\ \sum_{j=1}^{d} w_j a_{2j} \\ \vdots \\ \sum_{j=1}^{d} w_j a_{dj} \end{bmatrix} + \begin{bmatrix} \sum_{j=1}^{d} a_{1j} w_j \\ \sum_{j=1}^{d} a_{2j} w_j \\ \vdots \\ \sum_{j=1}^{d} a_{dj} w_j \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^{d} w_j a_{1j} \\ \sum_{j=1}^{d} w_j a_{2j} \\ \vdots \\ \sum_{j=1}^{d} w_j a_{dj} \end{bmatrix} + \begin{bmatrix} \sum_{j=1}^{d} a_{1j} w_j \\ \sum_{j=1}^{d} a_{2j} w_j \\ \vdots \\ \sum_{j=1}^{d} a_{dj} w_j \end{bmatrix} = \nabla w = A^T w + Aw = (A^T + A)w. \]

4. Convert to matrix notation:
\[ \nabla f(w) = \begin{bmatrix} \sum_{j=1}^{d} w_j a_{1j} \\ \sum_{j=1}^{d} w_j a_{2j} \\ \vdots \\ \sum_{j=1}^{d} w_j a_{dj} \end{bmatrix} + \begin{bmatrix} \sum_{j=1}^{d} a_{1j} w_j \\ \sum_{j=1}^{d} a_{2j} w_j \\ \vdots \\ \sum_{j=1}^{d} a_{dj} w_j \end{bmatrix} = A^T w + Aw = (A^T + A)w. \]
So our final result is that
\[ \nabla f(w) = (A^T + A)w. \]
Note that if \( A \) is symmetric (\( A^T = A \)) then we have \((A^T + A) = (A + A) = 2A\) so we have
\[ \nabla f(w) = 2Aw. \]
This generalizes the scalar case where \( \frac{d}{dw} [\alpha w^2] = 2\alpha w \). We can also consider general quadratic functions of the form
\[ f(w) = \frac{1}{2} w^T A w + b^T w + \gamma. \]
Using the above results we have
\[ \nabla f(w) = \frac{1}{2} (A^T + A)w + b, \]
and if \( A \) is symmetric then
\[ \nabla f(w) = Aw + b. \]

3 Hessian of Linear Function

For a linear function of the form,
\[ f(w) = a^T w, \]
we show above the partial derivatives are given by
\[ \frac{\partial f}{\partial w_k} = a_k. \]
Since these first partial derivatives don’t depend on any \( w_k \), the second partial derivatives are thus given by
\[ \frac{\partial^2 f}{\partial w_k \partial w_k'} = 0, \]
which means that the Hessian matrix is the zero matrix,
\[
\begin{bmatrix}
\frac{\partial}{\partial w_1 \partial w_1} f(w) & \frac{\partial}{\partial w_1 \partial w_2} f(w) & \cdots & \frac{\partial}{\partial w_1 \partial w_d} f(w) \\
\frac{\partial}{\partial w_2 \partial w_1} f(w) & \frac{\partial}{\partial w_2 \partial w_2} f(w) & \cdots & \frac{\partial}{\partial w_2 \partial w_d} f(w) \\
\vdots & \vdots & \ddots & \vdots \\
\frac{\partial}{\partial w_d \partial w_1} f(w) & \frac{\partial}{\partial w_d \partial w_2} f(w) & \cdots & \frac{\partial}{\partial w_d \partial w_d} f(w)
\end{bmatrix} = 
\begin{bmatrix}
0 & 0 & \cdots & 0 \\
0 & 0 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & 0
\end{bmatrix},
\]
and using 0 to denote the zero matrix we have
\[ \nabla^2 f(w) = 0. \]

4 Hessian of Quadratic Function

For a quadratic function of the form,
\[ f(w) = w^T A w, \]
we show above the partial derivatives are given by linear functions,
\[
\frac{\partial f}{\partial w_k} = \sum_{j=1}^{d} w_j a_{jk} + \sum_{j=1}^{d} a_{kj} w_j.
\]
The second partial derivatives are thus constant functions of the form
\[ \frac{\partial^2 f}{\partial w_k \partial w_{k'}} = a_{k'k} + a_{kk'}, \]
which means that the Hessian matrix has a simple form
\[
\nabla^2 f(w) = \begin{bmatrix}
\frac{\partial}{\partial w_1} f(w) & \frac{\partial}{\partial w_2} f(w) & \cdots & \frac{\partial}{\partial w_d} f(w) \\
\vdots & \vdots & \ddots & \vdots \\
\frac{\partial}{\partial w_d} f(w) & \frac{\partial}{\partial w_1} f(w) & \cdots & \frac{\partial}{\partial w_d} f(w)
\end{bmatrix} = \begin{bmatrix}
a_{11} + a_{11} & a_{21} + a_{12} & \cdots & a_{d1} + a_{1d} \\
a_{12} + a_{21} & a_{22} + a_{22} & \cdots & a_{d2} + a_{2d} \\
\vdots & \vdots & \ddots & \vdots \\
a_{1d} + a_{d1} & a_{2d} + a_{d2} & \cdots & a_{dd} + a_{dd}
\end{bmatrix}.
\]
This gives a result of
\[ \nabla^2 f(w) = A + A^T, \]
and if \( A \) is symmetric this simplifies to
\[ \nabla^2 f(w) = 2A. \]

5 Example of Least Squares

The least squares objective function has the form
\[ f(w) = \frac{1}{2} \| Xw - y \|^2, \]
which can be written as
\[ f(w) = \frac{1}{2} w^T X^T X w - w^T X^T y + \frac{1}{2} y^T y. \]
By using that \( X^T X \) and symmetric and using the results above we have that
\[ \nabla f(w) = X^T X w - X^T y, \]
and that
\[ \nabla^2 f(w) = X^T X. \]