Deriving the Gradient of Linear and Quadratic Functions in Matrix Notation

Mark Schmidt

October 21, 2016

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 gradeint 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 results 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 β. But in this case we still have $\nabla f(w) = a$ since the y-intercept β 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**:

$$f(w) = w^T \underbrace{\begin{bmatrix} \sum_{j=1}^n a_{1j} w_j \\ \sum_{j=1}^n a_{2j} w_j \\ \vdots \\ \sum_{j=1}^n a_{dj} w_j \end{bmatrix}}_{Aw} = \sum_{i=1}^d \sum_{j=1}^d w_i a_{ij} w_j.$$

   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:

$$\frac{\partial}{\partial w_k} \left[\sum_{i=1}^d (a_{ii} w_i^2 + \sum_{j \neq i} w_i a_{ij} w_j). \right] = 2 a_{kk} w_k + \sum_{j \neq k} w_j a_{jk} + \sum_{j \neq k} a_{kj} w_j.$$

   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**:

$$\nabla f(w) = \begin{bmatrix} \frac{\partial}{\partial w_1} \\ \frac{\partial}{\partial w_2} \\ \vdots \\ \frac{\partial}{\partial w_d} \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^d w_j a_{j1} + \sum_{j=1}^d a_{1j} w_j \\ \sum_{j=1}^d w_j a_{j2} + \sum_{j=1}^d a_{2j} w_j \\ \vdots \\ \sum_{j=1}^d w_j a_{jd} + \sum_{j=1}^d a_{dj} w_j \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^d w_j a_{j1} \\ \sum_{j=1}^d w_j a_{j2} \\ \vdots \\ \sum_{j=1}^d w_j a_{jd} \end{bmatrix} + \begin{bmatrix} \sum_{j=1}^d a_{1j} w_j \\ \sum_{j=1}^d a_{2j} w_j \\ \vdots \\ \sum_{j=1}^d a_{dj} w_j \end{bmatrix}$$

4. **Convert to matrix notation**:

$$\nabla f(w) = \begin{bmatrix} \sum_{j=1}^d w_j a_{j1} \\ \sum_{j=1}^d w_j a_{j2} \\ \vdots \\ \sum_{j=1}^d w_j a_{jd} \end{bmatrix} + \begin{bmatrix} \sum_{j=1}^d a_{1j} w_j \\ \sum_{j=1}^d a_{2j} w_j \\ \vdots \\ \sum_{j=1}^d a_{dj} w_j \end{bmatrix} = A^T w + A w = (A^T + A) w.$$

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

$$\nabla f(w) = Aw + b.$$