Differentiating Gram Schmidt

Wouter Edeling

Centrum Wiskunde & Informatica, Scientific Computing group, 1098 XG Amsterdam, the Netherlands.

October 21, 2021

Gram-Schmidt orthogonalization is a well-known technique where a non-orthogonal (yet independent) set of d column vectors $\mathbf{q}_i \in \mathbb{R}^D$ is made orthogonal via

$$\mathbf{w}_i = \mathbf{q}_i - \sum_{j=1}^{i-1} \left(\frac{\mathbf{w}_j^T \mathbf{q}_i}{\mathbf{w}_j^T \mathbf{w}_j} \right) \mathbf{w}_j, \quad i = 1, \cdots, d. \tag{1}$$

That is, we start with $\mathbf{w}_1 := \mathbf{q}_1$, and for all subsequent vectors $\mathbf{q}_i$ we subtract the projections of $\mathbf{q}_i$ onto each vector $\mathbf{w}_j$ which has previously been orthogonalized. This leaves us with a orthogonal basis $[\mathbf{w}_1(\mathbf{q}_1) \;\; \mathbf{w}_2(\mathbf{q}_1, \mathbf{q}_2) \;\; \cdots \;\; \mathbf{w}_d(\mathbf{q}_1, \mathbf{q}_2 \cdots, \mathbf{q}_d)]$. Finally, to obtain an orthonormal basis, each column vector can be divided by its length:

$$\left\{ \frac{\mathbf{w}_1(\mathbf{q}_1)}{\|\mathbf{w}_1(\mathbf{q}_1)\|_2}, \;\; \frac{\mathbf{w}_2(\mathbf{q}_1, \mathbf{q}_2)}{\|\mathbf{w}_2(\mathbf{q}_1, \mathbf{q}_2)\|_2}, \;\; \cdots \;\; , \frac{\mathbf{w}_d(\mathbf{q}_1, \mathbf{q}_2 \cdots, \mathbf{q}_d)}{\|\mathbf{w}_d(\mathbf{q}_1, \mathbf{q}_2 \cdots, \mathbf{q}_d)\|_2} \right\} \tag{2}$$

We will derive an expression for the derivatives of (1) with respect to the $\mathbf{q}_i$ vectors, and likewise for the normalized vectors (2).

1 Derivatives of unnormalized Gram-Schmidt vectors

We will first compute the derivative of $\mathbf{w}_i$, before it is normalized by its length $\|\mathbf{w}_i\|_2$:

$$\frac{\partial \mathbf{w}_i}{\partial \mathbf{q}_k} = \frac{\partial \mathbf{q}_i}{\partial \mathbf{q}_k} - \sum_{j=1}^{i-1} \frac{\partial}{\partial \mathbf{q}_k} \left[\left(\frac{\mathbf{w}_j^T \mathbf{q}_i}{\mathbf{w}_j^T \mathbf{w}_j} \right) \mathbf{w}_j \right], \quad i = 1, \cdots, d, \quad k = 1, \cdots, d. \tag{3}$$

A brute-force computation of this derivative using a computer algebra system will show that this quickly becomes a complicated expression with a very large number of terms. However, we can find a simple expression for this derivative by computing it in an iterative fashion:

$i = 1$:

$$\mathbf{w}_1 = \mathbf{q}_1 \Rightarrow \frac{\partial \mathbf{w}_1}{\partial \mathbf{q}_1} = I_D =: D_{11}, \quad \text{and} \quad \frac{\partial \mathbf{w}_1}{\partial \mathbf{q}_i} = 0 \quad \text{for } i > 1. \tag{4}$$

$i = 2$:

$$\mathbf{w}_2 = \mathbf{q}_2 - \left(\frac{\mathbf{w}_1^T \mathbf{q}_2}{\mathbf{w}_1^T \mathbf{w}_1} \right) \mathbf{w}_1 \tag{5}$$

First we compute the 'shear' derivative $\partial \mathbf{w}_2/\partial \mathbf{q}_1$ (defined as $\partial \mathbf{w}_i/\partial \mathbf{q}_k$ where $i \neq k$):

$$\begin{aligned}
\frac{\partial \mathbf{w}_2}{\partial \mathbf{q}_1} &= -\frac{\partial}{\partial \mathbf{q}_1}\left[\left(\frac{\mathbf{w}_1^T\mathbf{q}_2}{\mathbf{w}_1^T\mathbf{w}_1}\right)\mathbf{w}_1\right] \\
&= -\frac{\partial}{\partial \mathbf{w}_1}\left[\left(\frac{\mathbf{w}_1^T\mathbf{q}_2}{\mathbf{w}_1^T\mathbf{w}_1}\right)\mathbf{w}_1\right]\frac{\partial \mathbf{w}_1}{\partial \mathbf{q}_1} \\
&= \underbrace{-\left[\frac{1}{\mathbf{w}_1^T\mathbf{w}_1}\,\mathbf{w}_1\mathbf{q}_2^T - \frac{2\mathbf{w}_1^T\mathbf{q}_2}{(\mathbf{w}_1^T\mathbf{w}_1)^2}\,\mathbf{w}_1\mathbf{w}_1^T + \frac{\mathbf{w}_1^T\mathbf{q}_2}{\mathbf{w}_1^T\mathbf{w}_1}\,I_D\right]}_{=:D_{21}}\frac{\partial \mathbf{w}_1}{\partial \mathbf{q}1} \\
&= D_{21}\frac{\partial \mathbf{w}_1}{\partial \mathbf{q}_1} = D_{21}D_{11}. \qquad (6)
\end{aligned}$$

Here, the second equality is just the chain rule, and in the third we used the following matrix calculus identity:

$$D_{ij} := -\frac{\partial}{\partial \mathbf{w}_j}\left[\left(\frac{\mathbf{w}_j^T\mathbf{q}_i}{\mathbf{w}_j^T\mathbf{w}_j}\right)\mathbf{w}_j\right] = -\left[\frac{1}{\mathbf{w}_j^T\mathbf{w}_j}\,\mathbf{w}_j\mathbf{q}_i^T - \frac{2\mathbf{w}_j^T\mathbf{q}_i}{(\mathbf{w}_j^T\mathbf{w}_j)^2}\,\mathbf{w}_j\mathbf{w}_j^T + \frac{\mathbf{w}_j^T\mathbf{q}_i}{\mathbf{w}_j^T\mathbf{w}_j}\,I_D\right], \qquad (7)$$

which is derived in Appendix A. Hence, we can compute D_{21}, and the matrix $D_{11} := \partial \mathbf{w}_1/\partial \mathbf{q}_1 = I_D$ was already computed in the previous iteration. The matrix-matrix product of D_{21} and D_{11} yields $\partial \mathbf{w}_2/\partial \mathbf{q}_1$.

We now compute the 'normal' derivative $\partial \mathbf{w}_2/\partial \mathbf{q}_2$:

$$\begin{aligned}
\frac{\partial \mathbf{w}_2}{\partial \mathbf{q}_2} &= I_D - \frac{\partial}{\partial \mathbf{q}_2}\left[\left(\frac{\mathbf{w}_1^T\mathbf{q}_2}{\mathbf{w}_1^T\mathbf{w}_1}\right)\mathbf{w}_1\right] \\
&= I_D - \frac{\mathbf{w}_1\mathbf{w}_1^T}{\mathbf{w}_1^T\mathbf{w}_1} \\
&= D_{11} - \frac{\mathbf{w}_1\mathbf{w}_1^T}{\mathbf{w}_1^T\mathbf{w}_1} =: D_{22} \qquad (8)
\end{aligned}$$

In the second equality we made use of the identity:

$$\frac{\partial}{\partial \mathbf{q}_i}\left[\left(\frac{\mathbf{w}_j^T\mathbf{q}_i}{\mathbf{w}_j^T\mathbf{w}_j}\right)\mathbf{w}_j\right] = \frac{\mathbf{w}_j\mathbf{w}_j^T}{\mathbf{w}_j^T\mathbf{w}_j}, \qquad (9)$$

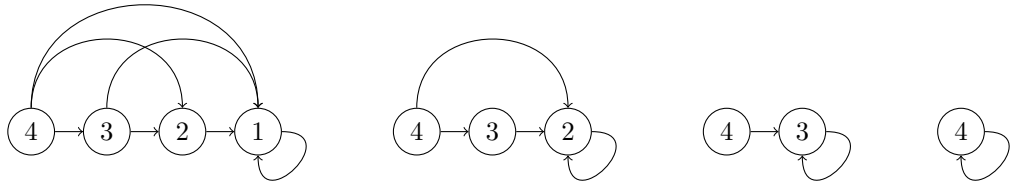
also derived in Appendix A. This holds if $\mathbf{w}_j$ does not depend upon $\mathbf{q}_i$, which is true in (8) since $\mathbf{w}_1 = \mathbf{w}_1(\mathbf{q}_1)$. Also, since $\mathbf{w}_2 = \mathbf{w}_2(\mathbf{q}_1, \mathbf{q}_2)$, we have $\partial \mathbf{w}_2/\partial \mathbf{q}_k = 0$ for $k > 2$. Further note that in (8), we can write $\partial \mathbf{w}_2/\partial \mathbf{q}_2$ as the difference between $D_{11} := \partial \mathbf{w}_1/\partial \mathbf{q}_1$ and $\mathbf{w}_1\mathbf{w}_1^T/\mathbf{w}_1^T\mathbf{w}_1$, and that we have defined this expression for $\partial \mathbf{w}_2/\partial \mathbf{q}_2$ as D_{22}.

$i = 3$:

$$\mathbf{w}_3 = \mathbf{q}_3 - \left(\frac{\mathbf{w}_1^T\mathbf{q}_3}{\mathbf{w}_1^T\mathbf{w}_1}\right)\mathbf{w}_1 - \left(\frac{\mathbf{w}_2^T\mathbf{q}_3}{\mathbf{w}_2^T\mathbf{w}_2}\right)\mathbf{w}_2 \qquad (10)$$

We again compute the 'shear' derivatives first:

$$\begin{aligned}
\frac{\partial \mathbf{w}_3}{\partial \mathbf{q}_1} &= -\frac{\partial}{\partial \mathbf{q}_1}\left[\left(\frac{\mathbf{w}_1^T\mathbf{q}_3}{\mathbf{w}_1^T\mathbf{w}_1}\right)\mathbf{w}_1\right] - \frac{\partial}{\partial \mathbf{q}_1}\left[\left(\frac{\mathbf{w}_2^T\mathbf{q}_3}{\mathbf{w}_2^T\mathbf{w}_2}\right)\mathbf{w}_2\right] \\
&= \underbrace{-\frac{\partial}{\partial \mathbf{w}_1}\left[\left(\frac{\mathbf{w}_1^T\mathbf{q}_3}{\mathbf{w}_1^T\mathbf{w}_1}\right)\mathbf{w}_1\right]}_{=:D_{31}}\frac{\partial \mathbf{w}_1}{\partial \mathbf{q}_1}\underbrace{-\frac{\partial}{\partial \mathbf{w}_2}\left[\left(\frac{\mathbf{w}_2^T\mathbf{q}_3}{\mathbf{w}_2^T\mathbf{w}_2}\right)\mathbf{w}_2\right]}_{=:D_{32}}\frac{\partial \mathbf{w}_2}{\partial \mathbf{q}_1} \\
&= D_{31}\frac{\partial \mathbf{w}_1}{\partial \mathbf{q}_1} + D_{32}\frac{\partial \mathbf{w}_2}{\partial \mathbf{q}_1} = D_{31}D_{11} + D_{32}D_{21}D_{11}. \qquad (11)
\end{aligned}$$



(a) All pairwise paths from 4 to 1, always ending with $(1,1)$. Corresponds to indices of $\partial \mathbf{w}_4/\partial \mathbf{q}_1$.

(b) All pairwise paths from 4 to 2, always ending with $(2,2)$. Corresponds to indices of $\partial \mathbf{w}_4/\partial \mathbf{q}_2$.

(c) All pairwise paths from 4 to 3, always ending with $(3,3)$. Corresponds to indices of $\partial \mathbf{w}_4/\partial \mathbf{q}_3$.

(d) The only pairwise paths from 4 to 4. Corresponds to index of $\partial \mathbf{w}_4/\partial \mathbf{q}_4$.

Figure 1: The directed graphs that generate the indices of the D_{ij} matrices of (14).

And

$$\begin{aligned}\frac{\partial \mathbf{w}_3}{\partial \mathbf{q}_2} &= -\frac{\partial}{\partial \mathbf{q}_2}\left[\left(\frac{\mathbf{w}_1^T\mathbf{q}_3}{\mathbf{w}_1^T\mathbf{w}_1}\right)\mathbf{w}_1\right] - \frac{\partial}{\partial \mathbf{q}_2}\left[\left(\frac{\mathbf{w}_2^T\mathbf{q}_3}{\mathbf{w}_2^T\mathbf{w}_2}\right)\mathbf{w}_2\right] \\ &= -\frac{\partial}{\partial \mathbf{w}_1}\left[\left(\frac{\mathbf{w}_1^T\mathbf{q}_3}{\mathbf{w}_1^T\mathbf{w}_1}\right)\mathbf{w}_1\right]\underbrace{\frac{\partial \mathbf{w}_1}{\partial \mathbf{q}_2}}_{0} \underbrace{- \frac{\partial}{\partial \mathbf{w}_2}\left[\left(\frac{\mathbf{w}_2^T\mathbf{q}_3}{\mathbf{w}_2^T\mathbf{w}_2}\right)\mathbf{w}_2\right]}_{:=D_{32}}\frac{\partial \mathbf{w}_2}{\partial \mathbf{q}_2} \\ &= D_{32}\frac{\partial \mathbf{w}_2}{\partial \mathbf{q}_2} = D_{32}D_{22}. \end{aligned} \tag{12}$$

The 'normal' derivative is given by:

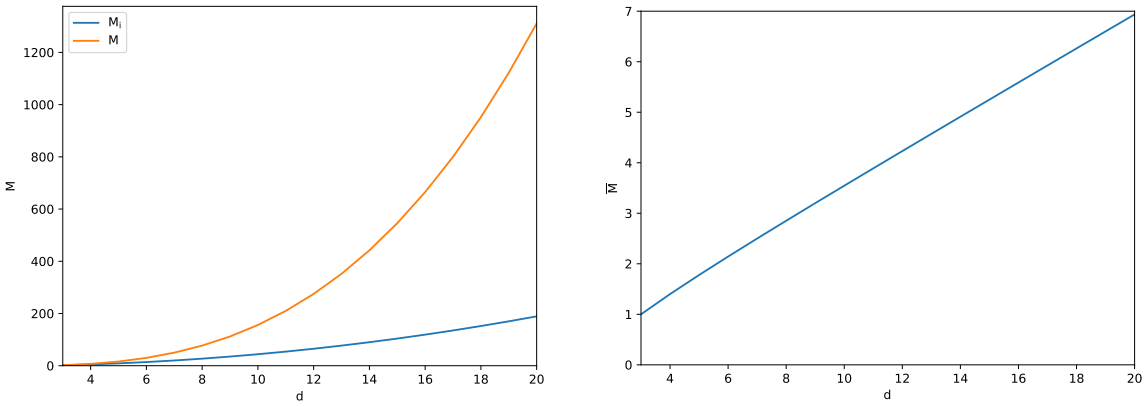
$$\begin{aligned}\frac{\partial \mathbf{w}_3}{\partial \mathbf{q}_3} &= I_D - \frac{\partial}{\partial \mathbf{q}_3}\left[\left(\frac{\mathbf{w}_1^T\mathbf{q}_3}{\mathbf{w}_1^T\mathbf{w}_1}\right)\right] - \frac{\partial}{\partial \mathbf{q}_3}\left[\left(\frac{\mathbf{w}_2^T\mathbf{q}_3}{\mathbf{w}_2^T\mathbf{w}_2}\right)\right] \\ &= \underbrace{I_D - \frac{\mathbf{w}_1\mathbf{w}_1^T}{\mathbf{w}_1^T\mathbf{w}_1}}_{D_{22}} - \frac{\mathbf{w}_2\mathbf{w}_2^T}{\mathbf{w}_2^T\mathbf{w}_2} = D_{22} - \frac{\mathbf{w}_2\mathbf{w}_2^T}{\mathbf{w}_2^T\mathbf{w}_2} =: D_{33} \end{aligned} \tag{13}$$

Finally, consider the case for $i = 4$ in shorthand notation only:

$$\begin{aligned}\frac{\partial \mathbf{w}_4}{\partial \mathbf{q}_1} &= D_{41}D_{11} + D_{42}D_{21}D_{11} + D_{43}D_{31}D_{11} + D_{43}D_{32}D_{21}D_{11} \\ \frac{\partial \mathbf{w}_4}{\partial \mathbf{q}_2} &= D_{42}D_{22} + D_{43}D_{32}D_{22} \\ \frac{\partial \mathbf{w}_4}{\partial \mathbf{q}_3} &= D_{43}D_{33} \\ \frac{\partial \mathbf{w}_4}{\partial \mathbf{q}_3} &= D_{33} - \frac{\mathbf{w}_3\mathbf{w}_3^T}{\mathbf{w}_3^T\mathbf{w}_3} =: D_{44} \end{aligned} \tag{14}$$

The structure is now apparent. The gradients $\partial \mathbf{w}_i/\partial \mathbf{q}_k$, when completely expanded as in (14), are determined by a series of matrix-matrix multiplications, the indices of which come from all pairwise paths of a directed graph from $i \to k$, ending with $k \to k$. To see this, consider the graphs of Figure 1, and compare this to the indices of the D_{ij} matrices appearing in the expressions of (14).

The graph gives some insight into the structure of each derivative, as it allows us to directly expand all terms that make up each gradient. However, we will not directly use it in practice to compute the gradients. Instead, we will start with the gradient of $\mathbf{w}_1$, then compute the gradients of $\mathbf{w}_2$ etc. This is because at any given $\mathbf{w}_i$, we can reuse the results from the previous iteration, thereby avoiding repeated



(a) The total number of matrix multiplication M_i at a given i, and the total cumulative cost M.

(b) The total number of matrix-matrix multiplications, divided by the number of gradients that are computed.

Figure 2: The number of matrix-matrix multiplications as a function of d.

matrix multiplications. This is clear when we write the 'shear' gradients as

$$\boxed{\frac{\partial \mathbf{w}_i}{\partial \mathbf{q}_k} = \sum_{j=1}^{i-1} D_{ij} \frac{\partial \mathbf{w}_j}{\partial \mathbf{q}_k}, \quad i \neq k, \;\; i > k} \tag{15}$$

At any given $i > 1$, all $\partial \mathbf{w}_j / \partial \mathbf{q}_k$ terms above have either already been computed at the previous iterations, or are zero when $k > j$. If we count the minimum number of matrix-matrix multiplications that are required to compute all shear gradients at a given i, we find that when $k = 1$, we get $i-1$ matrix-matrix multiplications. As an example, consider $\partial \mathbf{w}_4 / \partial \mathbf{q}_1 = D_{41} \partial \mathbf{w}_1 / \partial \mathbf{q}_1 + D_{42} \partial \mathbf{w}_2 / \partial \mathbf{q}_1 + D_{43} \partial \mathbf{w}_3 / \partial \mathbf{q}_1$, which requires $i-1 = 3$ matrix multiplications. Technically however, D_{41} is multiplied by $\partial \mathbf{w}_1 / \partial \mathbf{q}_1 = I_D$, so the first product we never have to compute. When $k = 2$, we still have $i - 1$ terms in (15). However, the first term will include $\partial \mathbf{w}_1 / \partial \mathbf{q}_2 = 0$. Likewise, when $k = 3$ the first two terms will be zero. Thus the total number of required matrix multiplications M_i is $M_i = (i-1) + (i-2) + \cdots + 2 + 1 - 1$, where we subtracted 1 at the end because we do not count multiplication by I_D. Finally, the total number of matrix multiplication M to compute all non-zero gradients of $\mathbf{w}_i$, for all $i = 1, \cdots, d$, is therefore given by

$$M = \sum_{i=3}^{d} M_i = \sum_{i=3}^{d} \sum_{j=2}^{i-1} j. \tag{16}$$

We start counting at $i = 3$ because $i = 1$ and $i = 2$ require no matrix-matrix multiplications, see (4) and (6). In Figure 2(a) we plot M versus d. which shows that for $d = 20$ we would already need to perform more than 1200 matrix-matrix multiplications. That said, it should be noted that this is the cost of computing *all* non-zero gradients $\partial \mathbf{w}_i / \partial \mathbf{q}_k$ which require matrix-matrix multiplication. In the case of $d = 20$, the total number of such gradients is given by $2 + 3 + \cdots + 19 = 189$. Hence, it is not particularly surprising that a large number of multiplications are required. Figure 2(b) shows the total number of multiplications divided by the number of gradients which are computed, which suggests that the average number of matrix-matrix multiplications scales linearly with d. Still, for high d the number of matrix-matrix multiplications M can be significant.

The preceding pertained to the shear gradients. From (6), (13) and (14), we can see that the normal gradients are computed as:

$$\boxed{D_{ii} := \frac{\partial \mathbf{w}_i}{\partial \mathbf{q}_i} = D_{i-1,i-1} - \frac{\mathbf{w}_{i-1} \mathbf{w}_{i-1}^T}{\mathbf{w}_{i-1}^T \mathbf{w}_{i-1}}, \quad i > 1}, \tag{17}$$

with $D_{11} := I_D$. Note that no matrix-matrix multiplication is involved in computing these derivatives.

Finally, let us note that we verified the correctness of (15) and (17), by comparing our numerical result with the results from a computer algebra system, which used symbolic math to directly compute the derivatives from the definition of the $\mathbf{w}_i$ (Eq. (1)).

2 Derivatives of normalized Gram-Schmidt vectors

Equations (15) and (17) give simple iterative expressions for $\partial \mathbf{w}_i / \partial \mathbf{q}_k$. However, Gram-Schmidt vectors are often normalized, and so we need to compute $\partial(\mathbf{w}_i / \|\mathbf{w}_i\|_2)/\partial \mathbf{q}_k$. As shown in Appendix A, we can just premultiply $\partial \mathbf{w}_i / \partial \mathbf{q}_k$ with a matrix which only depends upon $\mathbf{w}_i$, to obtain the gradient of the corresponding normed vector:

$$\boxed{\frac{\partial}{\partial \mathbf{q}_k}\left(\frac{\mathbf{w}_i}{\|\mathbf{w}_i\|}\right) = \left[\frac{I_D}{\|\mathbf{w}_i\|_2} - \frac{\mathbf{w}_i \mathbf{w}_i^T}{\|\mathbf{w}_i\|_2^3}\right] \frac{\partial \mathbf{w}_i}{\partial \mathbf{q}_k}}. \tag{18}$$

3 Verification and validation

All derivative expressions were validated with the use of a symbolic math library. We also used this library to verify our Python implementation of the derivatives. These results can be recreated by running a Jupyter notebook available at [1].

Acknowledgements

Funding: European Union Horizon 2020 research and innovation programme under grant agreement #800925 (VECMA project).

References

[1] W.N. Edeling. Gram_Schmidt_Derivatives Github repo. https://github.com/wedeling/Gram_Schmidt_Derivatives.

[2] K.B. Petersen and M.S. Pedersen. The matrix cookbook, version 20121115. *Technical Univ. Denmark, Kongens Lyngby, Denmark, Tech. Rep*, 3274, 2012.

A Matrix calculus identities

We will derive a number of useful matrix calculus identities here. Let $\mathbf{w}$ and $\mathbf{q}$ be two vectors in $\mathbb{R}^{D\times 1}$. Then

$$\frac{\partial}{\partial \mathbf{w}}\left(\frac{\mathbf{w}^T\mathbf{q}}{\mathbf{w}^T\mathbf{w}}\right) = \frac{1}{\mathbf{w}^T\mathbf{w}}\,\mathbf{q}^T - \frac{2\,\mathbf{w}^T\mathbf{q}}{(\mathbf{w}^T\mathbf{w})^2}\,\mathbf{w}^T \in \mathbb{R}^{1\times D} \tag{19}$$

Proof: This follows directly from the quotient rule:

$$\frac{\partial}{\partial \mathbf{w}}\left(\frac{\mathbf{w}^T\mathbf{q}}{\mathbf{w}^T\mathbf{w}}\right) = \frac{\frac{\partial}{\partial \mathbf{w}}[\mathbf{w}^T\mathbf{q}]\mathbf{w}^T\mathbf{w} - \mathbf{w}^T\mathbf{q}\frac{\partial}{\partial \mathbf{w}}[\mathbf{w}^T\mathbf{w}]}{(\mathbf{w}^T\mathbf{w})^2} = \frac{\mathbf{q}^T\,(\mathbf{w}^T\mathbf{w}) - (\mathbf{w}^T\mathbf{q})\,2\mathbf{w}^T}{(\mathbf{w}^T\mathbf{w})^2}. \tag{20}$$

which yields (19). Going one step further, we have to following identity:

$$\boxed{\frac{\partial}{\partial \mathbf{w}}\left[\left(\frac{\mathbf{w}^T\mathbf{q}}{\mathbf{w}^T\mathbf{w}}\right)\mathbf{w}\right] = \frac{1}{\mathbf{w}^T\mathbf{w}}\,\mathbf{w}\mathbf{q}^T - \frac{2\mathbf{w}^T\mathbf{q}}{(\mathbf{w}^T\mathbf{w})^2}\,\mathbf{w}\mathbf{w}^T + \frac{\mathbf{w}^T\mathbf{q}}{\mathbf{w}^T\mathbf{w}}\,I_D} \in \mathbb{R}^{D\times D} \tag{21}$$

Proof: Apply the product rule:

$$\begin{aligned}\frac{\partial}{\partial \mathbf{w}}\left[\left(\frac{\mathbf{w}^T\mathbf{q}}{\mathbf{w}^T\mathbf{w}}\right)\mathbf{w}\right] &= \mathbf{w}\frac{\partial}{\partial \mathbf{w}}\left(\frac{\mathbf{w}^T\mathbf{q}}{\mathbf{w}^T\mathbf{w}}\right) + \left(\frac{\mathbf{w}^T\mathbf{q}}{\mathbf{w}^T\mathbf{w}}\right) I_D \\ &= \mathbf{w}\left[\frac{1}{\mathbf{w}^T\mathbf{w}}\,\mathbf{q}^T - \frac{2\,\mathbf{w}^T\mathbf{q}}{(\mathbf{w}^T\mathbf{w})^2}\,\mathbf{w}^T\right] + \left(\frac{\mathbf{w}^T\mathbf{q}}{\mathbf{w}^T\mathbf{w}}\right) I_D \\ &= \frac{1}{\mathbf{w}^T\mathbf{w}}\,\mathbf{w}\mathbf{q}^T - \frac{2\mathbf{w}^T\mathbf{q}}{(\mathbf{w}^T\mathbf{w})^2}\,\mathbf{w}\mathbf{w}^T + \frac{\mathbf{w}^T\mathbf{q}}{\mathbf{w}^T\mathbf{w}}\,I_D \end{aligned} \tag{22}$$

where in the second equality we inserted (19). The final identity we need is given by

$$\boxed{\frac{\partial}{\partial \mathbf{q}}\left[\left(\frac{\mathbf{w}^T\mathbf{q}}{\mathbf{w}^T\mathbf{w}}\right)\mathbf{w}\right] = \frac{1}{\mathbf{w}^T\mathbf{w}}\,\mathbf{w}\mathbf{w}^T} \in \mathbb{R}^{D\times D} \tag{23}$$

Proof: Apply the product rule:

$$\begin{aligned}\frac{\partial}{\partial \mathbf{q}}\left[\left(\frac{\mathbf{w}^T\mathbf{q}}{\mathbf{w}^T\mathbf{w}}\right)\mathbf{w}\right] &= \mathbf{w}\frac{\partial}{\partial \mathbf{q}}\left(\frac{\mathbf{w}^T\mathbf{q}}{\mathbf{w}^T\mathbf{w}}\right) + \left(\frac{\mathbf{w}^T\mathbf{q}}{\mathbf{w}^T\mathbf{w}}\right)\underbrace{\frac{\partial \mathbf{w}}{\partial \mathbf{q}}}_{0} \\ &= \frac{1}{\mathbf{w}^T\mathbf{w}}\,\mathbf{w}\frac{\partial}{\partial \mathbf{q}}\left(\mathbf{w}^T\mathbf{q}\right) = \frac{1}{\mathbf{w}^T\mathbf{w}}\,\mathbf{w}\mathbf{w}^T \end{aligned} \tag{24}$$

Note that we assume that $\mathbf{w}$ does not depend upon $\mathbf{q}$, and that we can therefore take $1/\mathbf{w}^T\mathbf{w}$ out of the differentiation operator. This is perhaps confusing, as we do not make this assumption elsewhere. However, in the context of deriving the derivatives of the Gram-Schmidt vectors, this assumption always holds when we apply (23), see e.g.(8).

Finally, a standard formula (see e.g. [2]), for the gradient of a normed vector is:

$$\frac{\partial}{\partial \mathbf{w}}\left(\frac{\mathbf{w}}{\|\mathbf{w}\|_2}\right) = \frac{I_D}{\|\mathbf{w}\|_2} - \frac{\mathbf{w}\mathbf{w}^T}{\|\mathbf{w}\|_2^3}. \tag{25}$$

In our case, a useful related identity is:

$$\boxed{\frac{\partial}{\partial \mathbf{q}}\left(\frac{\mathbf{w}}{\|\mathbf{w}\|_2}\right) = \left[\frac{I_D}{\|\mathbf{w}\|_2} - \frac{\mathbf{w}\mathbf{w}^T}{\|\mathbf{w}\|_2^3}\right]\frac{\partial \mathbf{w}}{\partial \mathbf{q}}}, \tag{26}$$

which follows directly from the chain rule.