Proof of Theorem 1 in OP-TSOD

Theorem 1. *Given* $\boldsymbol{A} \in \mathbb{R}^{n_1 \times n_2}$, $\boldsymbol{G} \in \mathbb{R}^{n_1 \times d}$, *let* $\boldsymbol{M} = \boldsymbol{A}^\top \boldsymbol{G} = \boldsymbol{Q}_T \boldsymbol{R}_T$ *be the QR decomposition of* $\boldsymbol{A}^\top \boldsymbol{G}$, *where* $\boldsymbol{Q}_T \in \mathbb{R}^{n_2 \times d}$ *is an unitary matrix and is calculated by implementing Gram-Schmidt process on matrix* $\boldsymbol{M} = \boldsymbol{A}^\top \boldsymbol{G}$ *, and* $\boldsymbol{R}_T \in \mathbb{R}^{d \times d}$ *is an upper triangle matrix,* $\boldsymbol{u}_i \in \mathbb{R}^{n_2}$ *and* $\boldsymbol{w}_i \in \mathbb{R}^{n_2}$ *denotes the* i*-th column vector of* $\boldsymbol{M}$ *and* $\boldsymbol{Q}_T$ *respectively. Then the gradient of* $f(\boldsymbol{G}; \boldsymbol{A})$ *is given by*

$$\nabla f(\boldsymbol{G}; \boldsymbol{A}) = \boldsymbol{A}\boldsymbol{D}\boldsymbol{W}, \tag{1}$$

where

$$\begin{aligned}
\boldsymbol{D} &= 2(\boldsymbol{H}\boldsymbol{A} + \boldsymbol{A}^\top \boldsymbol{H}^\top)\boldsymbol{Q}_T, \boldsymbol{D} \in \mathbb{R}^{n_2 \times d};\\
\boldsymbol{H} &= \boldsymbol{Q}_T \boldsymbol{Q}_T^\top \boldsymbol{A}^\top - \boldsymbol{A}^\top, \boldsymbol{H} \in \mathbb{R}^{n_2 \times n_1};\\
\boldsymbol{W} &= \{\frac{\partial \boldsymbol{w}_i}{\partial \boldsymbol{u}_j}\}_{1 \le i \le d; 1 \le j \le d}, \boldsymbol{W} \in \mathbb{R}^{d \times d};\\
\frac{\partial \boldsymbol{w}_i}{\partial \boldsymbol{u}_j} &= \begin{cases} 0, & i < j,\\ \frac{1}{||\boldsymbol{u}_i - \sum_{k=1}^{i-1} (\boldsymbol{u}_i, \boldsymbol{w}_k)\boldsymbol{w}_k||}, & i = j,\\ -\frac{\sum_{k=j+1}^{i} (\boldsymbol{u}_i, \boldsymbol{w}_{k-1}) \frac{\partial \boldsymbol{w}_{k-1}}{\partial \boldsymbol{u}_j}}{||\boldsymbol{u}_i - \sum_{k=1}^{i} (\boldsymbol{u}_i, \boldsymbol{w}_k)\boldsymbol{w}_k||}, & i > j. \end{cases}
\end{aligned}$$

Eqn. (1) can be further simplified as,

$$\nabla f(\boldsymbol{G}; \boldsymbol{A}) = [(\boldsymbol{A}\boldsymbol{Q}_T)(\boldsymbol{A}\boldsymbol{Q}_T)^\top(\boldsymbol{A}\boldsymbol{Q}_T) - (\boldsymbol{A}\boldsymbol{A}^\top)(\boldsymbol{A}\boldsymbol{Q}_T)]\boldsymbol{W}. \tag{2}$$

Proof. The gradient $\frac{\partial f(\boldsymbol{G};\boldsymbol{A})}{\partial \boldsymbol{G}}$ of the objective function $f(\boldsymbol{G}; \boldsymbol{A})$ w.r.t. $\boldsymbol{G}$ can be calculated by using the chain rule,

$$\frac{\partial f}{\partial \boldsymbol{G}} = \frac{\partial f}{\partial \boldsymbol{Q}_T} \frac{\partial \boldsymbol{Q}_T}{\partial \boldsymbol{M}} \frac{\partial \boldsymbol{M}}{\partial \boldsymbol{G}}, \tag{3}$$

where $\boldsymbol{M} = \boldsymbol{A}^\top \boldsymbol{G}$. Thus the next step is to calculate $\frac{\partial f}{\partial \boldsymbol{Q}_T}$, $\frac{\partial \boldsymbol{Q}_T}{\partial \boldsymbol{M}}$ and $\frac{\partial \boldsymbol{M}}{\partial \boldsymbol{G}}$ respectively. For easy understanding of the derivation process, we denote $\boldsymbol{H} = \boldsymbol{Q}_T \boldsymbol{Q}_T^\top \boldsymbol{A}^\top - \boldsymbol{A}^\top$, and $\boldsymbol{H} \in \mathbb{R}^{n_2 \times n_1}$.

Step 1: Calculate $\frac{\partial f}{\boldsymbol{Q}_T}$: Note that $\frac{\partial f}{\partial \boldsymbol{Q}_T}$ will produce a matrix. Following the trace trick in [4], we can derive the derivative as follow,

$$
\begin{aligned}
df &= tr[df] \\
&= tr[(\frac{\partial f}{\partial \boldsymbol{H}})^\top d\boldsymbol{H}] \\
&= tr[2\boldsymbol{H}^\top d[\boldsymbol{Q}_T\boldsymbol{Q}_T^\top \boldsymbol{A}^\top - \boldsymbol{A}^\top]] \\
&= tr[2\boldsymbol{H}^\top d[\boldsymbol{Q}_T\boldsymbol{Q}_T^\top \boldsymbol{A}^\top]] \\
&= tr[2\boldsymbol{H}^\top[(d\boldsymbol{Q}_T)\boldsymbol{Q}_T^\top A^\top + \boldsymbol{Q}_T(d\boldsymbol{Q}_T)\boldsymbol{A}^\top]] \\
&= tr[2\boldsymbol{A}^\top \boldsymbol{H}^\top[(d\boldsymbol{Q}_T)\boldsymbol{Q}_T^\top + \boldsymbol{Q}_T(d\boldsymbol{Q}_T^\top)]] \\
&= tr[2[\boldsymbol{A}^\top \boldsymbol{H}^\top + \boldsymbol{H}\boldsymbol{A}](d\boldsymbol{Q}_T)\boldsymbol{Q}_T] \\
&= tr[2\boldsymbol{Q}_T^\top[\boldsymbol{A}^\top \boldsymbol{H}^\top + \boldsymbol{H}\boldsymbol{A}](d\boldsymbol{Q}_T)]
\end{aligned}
$$

Then $\frac{\partial f}{\partial \boldsymbol{Q}_T} = 2(\boldsymbol{H}\boldsymbol{A} + \boldsymbol{A}^\top \boldsymbol{H}^\top)\boldsymbol{Q}_T$. We denote $\boldsymbol{D} = 2(\boldsymbol{H}\boldsymbol{A} + \boldsymbol{A}^\top \boldsymbol{H}^\top)\boldsymbol{Q}_T$, then $\frac{\partial f}{\partial \boldsymbol{Q}_T} = \boldsymbol{D}$.

Step 2: Calculate $\frac{\partial \boldsymbol{M}}{\partial \boldsymbol{G}}$: Note that $\frac{\partial \boldsymbol{M}}{\partial \boldsymbol{G}}$ is a matrix-matrix derivative and thus will produce a tensor, which can be calculated by using the vectorization approach [4, 5], *i.e.* $\frac{\partial \boldsymbol{M}}{\partial \boldsymbol{G}} = \frac{\partial vec(\boldsymbol{M})}{\partial vec(\boldsymbol{G})} = \boldsymbol{I}_d \bigotimes \boldsymbol{A} \in \mathbb{R}^{n_1 d \times n_2 d}$ following [2, 4, 5], where $vec(\boldsymbol{M})$ reshapes the matrix $\boldsymbol{M} \in \mathbb{R}^{n_2 \times d}$ to a $n_2 d \times 1$ vector, $\bigotimes$ is the Kronecker product [5], $I_d \in \mathbb{R}^{d \times d}$ is an identity matrix and $\boldsymbol{A} \in \mathbb{R}^{n_1 \times n_2}$.

Step 3: Calculate $\frac{\partial \boldsymbol{Q}_T}{\partial \boldsymbol{M}}$: Denote $\boldsymbol{M} = \left[\boldsymbol{u}_1\ \boldsymbol{u}_2\ ...\ \boldsymbol{u}_d\right]; \boldsymbol{Q}_T = \left[\boldsymbol{w}_1\ \boldsymbol{w}_2\ ...\ \boldsymbol{w}_d\right]$. Similar to the calculation of $\frac{\partial \boldsymbol{M}}{\partial \boldsymbol{G}}$, $\frac{\partial \boldsymbol{Q}_T}{\partial \boldsymbol{M}}$ is also a matrix-matrix derivative. We hence use the same vectorization approach. Since $\boldsymbol{Q}_T$ is obtained through QR decomposition of $\boldsymbol{M}$ by the Schmidt orthogonalization process [3], every column vector in $\boldsymbol{Q}_T$ can be simplified as a combination of column vector in $\boldsymbol{M}$ and given by the following process:

$$
\begin{aligned}
\boldsymbol{w}_1 &= \frac{\boldsymbol{u}_1}{||\boldsymbol{u}_1||}; \\
\boldsymbol{w}_2 &= \frac{\boldsymbol{u}_2 - (\boldsymbol{u}_2, \boldsymbol{w}_1)\boldsymbol{w}_1}{||\boldsymbol{u}_2 - (\boldsymbol{u}_2, \boldsymbol{w}_1)\boldsymbol{w}_1||}; \\
&...
\end{aligned}
$$

Then, we can calculate the vector-vector derivative of $\boldsymbol{Q}_T$ w.r.t. $\boldsymbol{G}$ as follows,

$$
\begin{aligned}
&\frac{\partial \boldsymbol{w}_1}{\partial \boldsymbol{u}_1} = \frac{1}{||\boldsymbol{u_1}||}; \frac{\partial \boldsymbol{w}_1}{\partial \boldsymbol{u}_i} = 0, i = \{2, 3, ..., d\};\\
&\frac{\partial \boldsymbol{w}_2}{\partial \boldsymbol{u}_1} = \frac{-(\boldsymbol{u}_2, \boldsymbol{w}_1)\frac{\partial \boldsymbol{w}_1}{\partial \boldsymbol{u}_1}}{||\boldsymbol{u}_2 - (\boldsymbol{u}_2, \boldsymbol{w}_1)\boldsymbol{w}_1||}\\
&\frac{\partial \boldsymbol{w}_2}{\partial \boldsymbol{u}_2} = \frac{1}{||\boldsymbol{u}_2 - (\boldsymbol{u}_2, \boldsymbol{w}_1)\boldsymbol{w}_1||}; \frac{\partial \boldsymbol{w}_2}{\partial \boldsymbol{u}_i} = 0, i = \{3, 4, ..., d\};\\
&\frac{\partial \boldsymbol{w}_3}{\partial \boldsymbol{u}_1} = \frac{-(\boldsymbol{u}_3, \boldsymbol{w}_2)\frac{\partial \boldsymbol{w}_2}{\partial \boldsymbol{u}_1} - (\boldsymbol{u}_3, \boldsymbol{w}_1)\frac{\partial \boldsymbol{w}_2}{\partial \boldsymbol{u}_1}}{||\boldsymbol{u}_3 - (\boldsymbol{u}_2, \boldsymbol{w}_2)\boldsymbol{w}_2 - (\boldsymbol{u}_3, \boldsymbol{w}_1)\boldsymbol{w}_1||};\\
&\frac{\partial \boldsymbol{w}_3}{\partial \boldsymbol{u}_2} = \frac{-(\boldsymbol{u}_3, \boldsymbol{w}_2)\frac{\partial \boldsymbol{w}_2}{\partial \boldsymbol{u}_2}}{||\boldsymbol{u}_3 - (\boldsymbol{u}_2, \boldsymbol{w}_2)\boldsymbol{w}_2 - (\boldsymbol{u}_3, \boldsymbol{w}_1)\boldsymbol{w}_1||};\\
&\frac{\partial \boldsymbol{w}_3}{\partial \boldsymbol{u}_3} = \frac{1}{||\boldsymbol{u}_3 - (\boldsymbol{u}_2, \boldsymbol{w}_2)\boldsymbol{w}_2 - (\boldsymbol{u}_3, \boldsymbol{w}_1)\boldsymbol{w}_1||};\\
&\frac{\partial \boldsymbol{w}_3}{\partial \boldsymbol{u}_i} = 0, i = \{4, 5, ..., d\}.\\
&\ldots\ldots
\end{aligned}
$$

Combining the above vector-vector derivatives will produce $\frac{\partial \boldsymbol{Q}_T}{\partial \boldsymbol{M}} = \frac{\partial vec(\boldsymbol{Q}_T)}{\partial vec(\boldsymbol{G})} = \boldsymbol{I}_{n_2} \bigotimes \boldsymbol{W} \in \mathbb{R}^{n_2 d \times n_2 d}$ following [1], where $\boldsymbol{I}_{n_2}$ is an identity matrix, and

$$
\begin{aligned}
\boldsymbol{W} =& \{\frac{\partial \boldsymbol{w}_i}{\partial \boldsymbol{u}_j}\}_{1 \le i \le d; 1 \le j \le d}, \boldsymbol{W} \in \mathbb{R}^{d \times d};\\
\frac{\partial \boldsymbol{w}_i}{\partial \boldsymbol{u}_j} =& \begin{cases} 0, & i < j,\\ \frac{1}{||\boldsymbol{u}_i - \sum_{k=1}^{i-1}(\boldsymbol{u}_i, \boldsymbol{w}_k)\boldsymbol{w}_k||}, & i = j,\\ -\frac{\sum_{k=j+1}^{i}(\boldsymbol{u}_i, \boldsymbol{w}_{k-1})\frac{\partial \boldsymbol{w}_{k-1}}{\partial \boldsymbol{u}_j}}{||\boldsymbol{u}_i - \sum_{k=1}^{i}(\boldsymbol{u}_i, \boldsymbol{w}_k)\boldsymbol{w}_k||}, & i > j. \end{cases}
\end{aligned}
$$

Finally, $\frac{\partial f(\boldsymbol{G};\boldsymbol{A})}{\partial vec(\boldsymbol{G})} = (1 \bigotimes \boldsymbol{D})(\boldsymbol{I}_{n_2} \bigotimes \boldsymbol{W})(\boldsymbol{I}_d \bigotimes \boldsymbol{A}) \in \mathbb{R}^{1 \times n_1 d}$, where $vec(\boldsymbol{G})$ is the vector form of matrix $\boldsymbol{G}$. Finally, unstacking the vector form back to the matrix form will produce $\nabla f(\boldsymbol{G}; \boldsymbol{A}) = \boldsymbol{ADW} \in R^{n_1 \times d}$ following [2,4,5].

References

1. Abadir, K.M., Magnus, J.R.: Matrix algebra, vol. 1. Cambridge University Press (2005)
2. Bodewig, E.: Matrix calculus. Elsevier (2014)
3. Gander, W.: Algorithms for the qr decomposition. Res. Rep **80**(02), 1251–1268 (1980)
4. Graham, A.: Kronecker products and matrix calculus with applications. Courier Dover Publications (2018)
5. Whitcomb, L.L.: Notes on kronecker products. Available: spray. me. jhu. edu/llw/courses/me530647/kron_1. pdf (2013)