Supplementary Document to “Eliminating the Invariance on the Loss Landscape of Linear Autoencoders”, Proof of the Theorems

Reza Oftadeh, Jiayi Shen, Zhangyang Wang, Dylan Shell*

1 Preliminaries

Before we present the proof for the main theorems, the following three lemmas introduce some notations and basic relations that are required for the proofs. Note that the theorems' numbering are different than the numbering in the main article.

Lemma 1. *The constant matrices $\boldsymbol{T}_p \in \mathbb{R}^{p\times p}$ and $\boldsymbol{S}_p \in \mathbb{R}^{p\times p}$ are defined as*

$$(\boldsymbol{T}_p)_{ij} = (p-i+1)\,\delta_{ij},\ \ i.e.\ \boldsymbol{T}_p = \operatorname{diag}(p, p-1, \cdots, 1)\,, \tag{1}$$

$$(\boldsymbol{S}_p)_{ij} = p - \max(i,j) + 1,\ \ i.e.\ \boldsymbol{S}_p = \begin{bmatrix} p & p-1 & \cdots & 2 & 1 \\ p-1 & p-1 & \cdots & 2 & 1 \\ \vdots & \vdots & \ddots & 2 & 1 \\ 2 & 2 & 2 & 2 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix},\ \ e.g.\ S_4 = \begin{bmatrix} 4 & 3 & 2 & 1 \\ 3 & 3 & 2 & 1 \\ 2 & 2 & 2 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}. \tag{2}$$

Clearly, the diagonal matrix $\boldsymbol{T}_p$ is positive definite. Another matrix that will appear in the formulation is $\hat{\boldsymbol{S}}_p := \boldsymbol{T}_p^{-1}\boldsymbol{S}_p\boldsymbol{T}_p^{-1}$

$$\left(\hat{\boldsymbol{S}}_p\right)_{ij} = \left(\boldsymbol{T}_p^{-1}\boldsymbol{S}_p\boldsymbol{T}_p^{-1}\right)_{ij} = \frac{1}{p-\min(i,j)+1}\ \ i.e.\ \boldsymbol{T}_p^{-1}\boldsymbol{S}_p\boldsymbol{T}_p^{-1} = \begin{bmatrix} \frac{1}{p} & \frac{1}{p} & \cdots & \frac{1}{p} & \frac{1}{p} \\ \frac{1}{p} & \frac{1}{p-1} & \cdots & \frac{1}{p-1} & \frac{1}{p-1} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \frac{1}{p} & \frac{1}{p-1} & \cdots & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{p} & \frac{1}{p-1} & \cdots & \frac{1}{2} & 1 \end{bmatrix},$$

$$e.g.\ \hat{\boldsymbol{S}}_4 = \begin{bmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{4} & \frac{1}{3} & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{4} & \frac{1}{3} & \frac{1}{2} & 1 \end{bmatrix}.$$

The following properties of Hadamard product and matrices $\boldsymbol{T}_p$ and $\boldsymbol{S}_p$ are used throughout:

*All authors are with Department of Computer Science and Engineering, Texas A&M University, Texas. Correspondence to: Reza Oftadeh <reza.oftadeh@tamu.edu>.

1. *For any arbitrary matrix* $\boldsymbol{A} \in \mathbb{R}^{n\times p}$,

$$\sum_{i=1}^{p} \boldsymbol{I}_{i;p} = \boldsymbol{T}_p, \text{ and} \tag{3}$$

$$\sum_{i=1}^{p} \boldsymbol{I}_{i;p} \boldsymbol{A}' \boldsymbol{A} \boldsymbol{I}_{i;p} = \boldsymbol{S}_p \circ (\boldsymbol{A}'\boldsymbol{A}), \tag{4}$$

   where, $\circ$ *is the Hadamard (element-wise) product.*

2. *For any matrices* $\boldsymbol{M}_1, \boldsymbol{M}_2 \in \mathbb{R}^{p\times p}$ *and diagonal matrices* $\mathscr{D}, \mathscr{E} \in \mathbb{R}^{p\times p}$,

$$\mathscr{D}(\boldsymbol{M}_1 \circ \boldsymbol{M}_2)\mathscr{E} = (\mathscr{D}\boldsymbol{M}_1\mathscr{E}) \circ \boldsymbol{M}_2 = \boldsymbol{M}_1 \circ (\mathscr{D}\boldsymbol{M}_2\mathscr{E}).$$

   Moreover, if $\boldsymbol{\Pi}_1, \boldsymbol{\Pi}_2 \in \mathbb{R}^{p\times p}$ *are permutation matrices then*

$$\boldsymbol{\Pi}_1(\boldsymbol{M}_1 \circ \boldsymbol{M}_2)\boldsymbol{\Pi}_2 = (\boldsymbol{\Pi}_1\boldsymbol{M}_1\boldsymbol{\Pi}_2) \circ (\boldsymbol{\Pi}_1\boldsymbol{M}_2\boldsymbol{\Pi}_2).$$

3. $\boldsymbol{S}_p$ *is invertible and its inverse is a symmetric tridiagonal matrix*

$$(\boldsymbol{S}_p^{-1})_{ij} = \begin{cases} 1 & i=j=1 \\ 2 & i=j\neq 1 \\ -1 & |i-j|=1 \\ 0 & \text{otherwise} \end{cases}, \ \text{i.e. } \boldsymbol{S}_p^{-1} = \begin{bmatrix} 1 & -1 & \cdots & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & -1 & 2 \end{bmatrix}.$$

4. $\boldsymbol{S}_p$ *is positive definite.*

5. *For any matrix* $\boldsymbol{A} \in \mathbb{R}^{n\times p}$, $\boldsymbol{S}_p \circ (\boldsymbol{A}'\boldsymbol{A})$ *is positive semidefinite. If (not necessarily full rank)* $\boldsymbol{A}$ *has no zero column then* $\boldsymbol{S}_p \circ (\boldsymbol{A}'\boldsymbol{A})$ *is positive definite.*

6. *For any diagonal matrix* $\mathscr{D} \in \mathbb{R}^{p\times p}$

$$\boldsymbol{S}_p \circ \mathscr{D} = \boldsymbol{T}_p\mathscr{D}, \text{ and} \tag{5}$$

$$\hat{\boldsymbol{S}}_p \circ \mathscr{D} = \boldsymbol{T}_p^{-1}\mathscr{D}. \tag{6}$$

7. *Let* $\mathscr{D}, \mathscr{E} \in \mathbb{R}^{p\times p}$ *be positive semidefinite matrices, where* $\mathscr{E}$ *has no zero diagonal element, and* $\mathscr{D}$ *is of rank* $r \leq p$. *Also, let for any* $r \leq p$, $\mathbb{J}_r = \{i_1, \cdots, i_r\} (1 \leq i_1 < \cdots < i_r < n)$ *be any* ordered $r-$*index set. Then* $\mathscr{D}$ *and* $\mathscr{E}$ *satisfy*

$$\mathscr{E}\left(\hat{\boldsymbol{S}}_p \circ \mathscr{D}\right) = \left(\hat{\boldsymbol{S}}_p \circ \mathscr{E}\right)\mathscr{D},$$

   if and only if, the following two conditions are satisfied:

   (a) *The matrix* $\mathscr{D}$ *is diagonal with* $p-r$ *zero diagonal elements and* r *positive diagonal elements indexed by the set* $\mathbb{J}_r$. *That is for any* $i \in \mathbb{J}_r : (\mathscr{D})_{ii} > 0$ *and the rest of elements of* $\mathscr{D}$ *are zero.*

   (b) *For any* $i, j \in \mathbb{J}_r$ *and* $i \neq j$ *we have* $(\mathscr{E})_{i,j} = 0$.

   Clearly, if $\mathscr{D}$ *is positive definite then* $\mathbb{J}_r = \mathbb{N}_p$ *and hence, both* $\mathscr{D}$ *and* $\mathscr{E}$ *are diagonal.*

Proof. . The proof of the properties are as follows.

1. eq. (3) is trivial. For eq. (4) note that $\boldsymbol{A}\boldsymbol{I}_{i;p}$ selects the first i columns of $\boldsymbol{A}$ (zeros out the rest), and similarly, $\boldsymbol{I}_{i;p}\boldsymbol{A}'$ selects the first i rows of $\boldsymbol{A}$ (zeros out the rest). Therefore, $\boldsymbol{I}_{i;p}\boldsymbol{A}'\boldsymbol{A}\boldsymbol{I}_{i;p}$ is a $p \times p$ matrix that its Leading Principal Submatrix of order i (LPS$_i$) [1] is the same as the LPS$_i$ of $\boldsymbol{A}'\boldsymbol{A}$ (and the rest of the elements are zero). Hence, $\sum_{i=1}^{p} \boldsymbol{I}_{i;p}\boldsymbol{A}'\boldsymbol{A}\boldsymbol{I}_{i;p}$ (counting backwards) adds LPS$_p$ of $\boldsymbol{A}'\boldsymbol{A}$ (i.e. $\boldsymbol{A}'\boldsymbol{A}$ itself) with LPS$_{p-1}$ that doubles LPS$_{p-1}$ part of the result and then adds LPS$_{p-2}$ that triples the LPS$_{p-2}$ part of result, the process continues until by the last addition LPS$_1$is added to the result for the p^{th}times. This is exactly the same as evaluating $\boldsymbol{S}_p \circ (\boldsymbol{A}'\boldsymbol{A})$.
2. This is a standard result (Horn & Johnson, 2012), and no proof is needed.
3. Directly compute $\boldsymbol{S}_p\boldsymbol{S}_p^{-1}$:

$$
\begin{aligned}
\left(\boldsymbol{S}_p\boldsymbol{S}_p^{-1}\right)_{ij} &= \sum_{k=1}^{p} (\boldsymbol{S}_p)_{ik}(\boldsymbol{S}_p^{-1})_{kj} \xrightarrow{\forall |k-j|>1 : (\boldsymbol{S}_p^{-1})_{kj}=0} \\
&= \begin{cases} (\boldsymbol{S}_p)_{i,j-1}(\boldsymbol{S}_p^{-1})_{j-1,j} + (\boldsymbol{S}_p)_{i,j}(\boldsymbol{S}_p^{-1})_{j,j} + (\boldsymbol{S}_p)_{i,j+1}(\boldsymbol{S}_p^{-1})_{j+1,j} & 2 \le j \,\& \\ & j \le p-1 \\ (\boldsymbol{S}_p)_{i,p-1}(\boldsymbol{S}_p^{-1})_{p-1,p} + (\boldsymbol{S}_p)_{i,p}(\boldsymbol{S}_p^{-1})_{p,p} & j = p \\ (\boldsymbol{S}_p)_{i,1}(\boldsymbol{S}_p^{-1})_{1,1} + (\boldsymbol{S}_p)_{i,2}(\boldsymbol{S}_p^{-1})_{2,1} & j = 1 \end{cases} \\
&= \begin{cases} -(\boldsymbol{S}_p)_{i,j-1} + 2(\boldsymbol{S}_p)_{i,j} - (\boldsymbol{S}_p)_{i,j+1} & 2 \le j \le p-1 \\ -(\boldsymbol{S}_p)_{i,p-1} + 2(\boldsymbol{S}_p)_{i,p} & j = p \\ (\boldsymbol{S}_p)_{i,1} - (\boldsymbol{S}_p)_{i,2} & j = 1 \end{cases} \\
&= \begin{cases} \max(i, j-1) - 2\max(i, j) + \max(i, j+1) & 2 \le j \le p-1 \\ -(p - \max(i, p-1) + 1) + 2(p - \max(i, p) + 1) & j = p \\ -\max(i, 1) + \max(i, 2) & j = 1 \end{cases} \\
&= \begin{cases} \max(i, j-1) - 2\max(i, j) + \max(i, j+1) & 2 \le j \le p-1 \\ 1 - p + \max(i, p-1) & j = p \\ \max(i, 2) - \max(i, 1) & j = 1 \end{cases} \\
&= \begin{cases} \begin{cases} 1 & i = j \\ 0 & i \ne j \end{cases} & 1 < j < p \\ \begin{cases} 1 & i = p \\ 0 & i \ne p \end{cases} & j = p \\ \begin{cases} 1 & i = 1 \\ 0 & i \ge 2 \end{cases} & j = 1 \end{cases} = (\boldsymbol{I}_p)_{ij}.
\end{aligned}
$$

4. Firstly, note that $\boldsymbol{S}_p^{-1}$ is symmetric and nonsingular so all the eigenvalues are real and nonzero. It is also a diagonally dominant matrix (Horn & Johnson (2012), Def 6.1.9) since

$$\forall i \in \{1, \cdots, p\} : C_i \coloneqq |(\boldsymbol{S}_p^{-1})_{ii}| \ge \sum_{j=1, j\ne i} |(\boldsymbol{S}_p^{-1})_{ij}| \eqqcolon R_i,$$

[1]For a $p \times p$ matrix, the leading principal submatrix of order i is an $i \times i$ matrix derived by removing the last $p - i$ rows and columns of the original matrix (Horn & Johnson (2012), P17)

where the inequality is strict for the first and the last row and it is equal for the rows in the middle. Moreover, by Gersgorin circle theorem (Horn & Johnson (2012), Thm 6.1.1) for every eigenvalue l_i of $\boldsymbol{S}_p^{-1}$ there exists i such that $l_i \in [C_i - R_i, C_i + R_i]$. Since $\forall i : C_i \geq R_i$ we have all the eigenvalues are non-negative. They are also nonzero, hence, $\boldsymbol{S}_p^{-1}$ is positive definite, which implies $\boldsymbol{S}_p$ is also positive definite.

5. For any matrix $\boldsymbol{A} \in \mathbb{R}^{n\times p}$, $\boldsymbol{A}'\boldsymbol{A}$ is positive semidefinite. Also, $\boldsymbol{S}_p$ is positive definite so by Schur product theorem (Horn & Johnson (2012), Thm 7.5.3(a)), $\boldsymbol{S}_p \circ (\boldsymbol{A}'\boldsymbol{A})$ is positive semidefinite. Moreover, if all diagonal elements of $\boldsymbol{A}'\boldsymbol{A}$ are positive (i.e. $\boldsymbol{A}$ has no zero column) by the extension of Schur product theorem (Horn & Johnson (2012), Thm 7.5.3(b)) it is positive definite. This can also be easily deduced using the Oppenheim inequality (Horn & Johnson (2012), Thm 7.8.16); that is for positive semidefinite matrices $\boldsymbol{S}_p$ and $\boldsymbol{A}'\boldsymbol{A}$: $\det(\boldsymbol{S}_p)\prod_i(\boldsymbol{A}'\boldsymbol{A})_{ii} \leq \det(\boldsymbol{S}_p \circ (\boldsymbol{A}'\boldsymbol{A}))$. Since, $\boldsymbol{S}_p$ is positive definite, $\det(\boldsymbol{S}_p) > 0$ (in fact it is 1 for any p) and if $\boldsymbol{A}'\boldsymbol{A}$ has no zero diagonal then $\det(\boldsymbol{S}_p \circ (\boldsymbol{A}'\boldsymbol{A})) > 0$ and therefore, $\boldsymbol{S}_p \circ (\boldsymbol{A}'\boldsymbol{A})$ is positive definite.

6. Clearly, the matrix $\boldsymbol{T}_p$ is achieved by setting the off-diagonal elements of $\boldsymbol{S}_p$ to zero. Hence, for any diagonal matrix $\mathscr{D} \in \mathbb{R}^{p\times p}$: $\boldsymbol{S}_p \circ \mathscr{D} = \boldsymbol{T}_p \circ \mathscr{D}$. For the diagonal matrices Hadamard product and matrix product are interchangeable so the latter may also be written as $\boldsymbol{T}_p\mathscr{D}$. The same argument applies for the second identity.

7. This property can easily be proved by induction on p and careful bookkeeping of indices.

□

Lemma 2 (Simultaneous diagonalization by congruence)**.** *Let $\boldsymbol{M}_1, \boldsymbol{M}_2 \in \mathbb{R}^{p\times p}$, where $\boldsymbol{M}_1$ is positive definite and $\boldsymbol{M}_2$ is positive semidefinite. Also, let $\mathscr{D}, \mathscr{E} \in \mathbb{R}^{r\times r}$ be positive definite diagonal matrices with $r \leq p$. Further, assume there is a $\boldsymbol{C} \in \mathbb{R}^{r\times p}$ of rank $r \leq p$ such that*

$$\boldsymbol{C}\boldsymbol{M}_1\boldsymbol{C}' = \mathscr{D} \text{ and}$$
$$\boldsymbol{C}\boldsymbol{M}_2\boldsymbol{C}' = \mathscr{D}\mathscr{E}.$$

Then there exists a nonsingular $\bar{\boldsymbol{C}} \in \mathbb{R}^{p\times p}$ that its first r rows are the matrix $\boldsymbol{C}$ and

$$\bar{\boldsymbol{C}}\boldsymbol{M}_1\bar{\boldsymbol{C}}' = \bar{\mathscr{D}} \text{ and}$$
$$\bar{\boldsymbol{C}}\boldsymbol{M}_2\bar{\boldsymbol{C}}' = \bar{\mathscr{D}}\bar{\mathscr{E}},$$

where, $\bar{\mathscr{D}} = \bar{\mathscr{D}} \oplus \boldsymbol{I}_{r-p}$ is a $p \times p$ diagonal matrix and $\bar{\mathscr{E}} = \mathscr{E} \oplus \underline{\mathscr{E}}$ is another $p \times p$ diagonal matrix, in which $\underline{\mathscr{E}} \in \mathbb{R}^{p-r\times p-r}$ is a nonnegative diagonal matrix. Clearly, the rank of $\boldsymbol{M}_2$ is r plus the number of nonzero diagonal elements of $\underline{\mathscr{E}}$.

Proof. The proof is rather straightforward since this lemma is the direct consequence of Theorem 7.6.4 in Horn & Johnson (2012). The theorem basically states that if $\boldsymbol{M}_1, \boldsymbol{M}_2 \in \mathbb{R}^{p\times p}$ is symmetric and $\boldsymbol{M}_1$ is positive definite then there exists an invertible $S \in \mathbb{R}^{p\times p}$ such that $S\boldsymbol{M}_1S' = \boldsymbol{I}_p$ and $S\boldsymbol{M}_2S'$ is a diagonal matrix with the same inertia as $\boldsymbol{M}_2$. Here, we have $\boldsymbol{M}_2$ that is positive semidefinite and $\boldsymbol{C} \in \mathbb{R}^{r\times p}$ of rank $r \leq p$ such that

$$\left(\mathscr{D}^{\frac{-1}{2}}\boldsymbol{C}\right)\boldsymbol{M}_1\left(\mathscr{D}^{\frac{-1}{2}}\boldsymbol{C}\right)' = \boldsymbol{I}_r \text{ and}$$

$$\left(\mathscr{D}^{\frac{-1}{2}}\boldsymbol{C}\right)\boldsymbol{M}_2\left(\mathscr{D}^{\frac{-1}{2}}\boldsymbol{C}\right)' = \mathscr{E}.$$

Therefore, since S is of full rank p and $\mathscr{D}^{\frac{-1}{2}}\boldsymbol{C}$ is of rank $r \leq p$, there exists $p-r$ rows in S that are linearly independent of rows of $\mathscr{D}^{\frac{-1}{2}}\boldsymbol{C}$. Establish $\bar{\boldsymbol{C}} \in \mathbb{R}^{p\times p}$ by adding those $p-r$ rows to $\boldsymbol{C}$. Then $\bar{\boldsymbol{C}}$ has p linearly independent rows so it is nonsingular, and fulfills the lemma's proposition that is

$$\begin{aligned}\bar{\boldsymbol{C}}\boldsymbol{M}_1\bar{\boldsymbol{C}}' &= \bar{\mathscr{D}} \text{ and}\\ \bar{\boldsymbol{C}}\boldsymbol{M}_2\bar{\boldsymbol{C}}' &= \bar{\mathscr{D}}\bar{\mathscr{E}},\end{aligned}$$

where, $\bar{\mathscr{D}} = \bar{\mathscr{D}} \oplus \boldsymbol{I}_{r-p}$ is a $p \times p$ diagonal matrix and $\bar{\mathscr{E}} = \mathscr{E} \oplus \underline{\mathscr{E}}$ is another $p \times p$ diagonal matrix, in which $\underline{\mathscr{E}} \in \mathbb{R}^{p-r\times p-r}$ is a nonnegative diagonal matrix. □

Lemma 3. *Let $\boldsymbol{A}$ and $\boldsymbol{B}$ define a critical point of L. Further, let $\boldsymbol{V} \in \mathbb{R}^{n\times p}$ and $\boldsymbol{W} \in \mathbb{R}^{p\times n}$ are such that* $\|\boldsymbol{V}\|_F, \|\boldsymbol{W}\|_F = O(\varepsilon)$ for some $\varepsilon > 0$. *Then*

$$\begin{aligned}L(\boldsymbol{A}+\boldsymbol{V},\boldsymbol{B}+\boldsymbol{W}) - L(\boldsymbol{A},\boldsymbol{B}) =& \langle \boldsymbol{V}\boldsymbol{T}_p\boldsymbol{B}\boldsymbol{\Sigma}_{xx}\boldsymbol{B}', \boldsymbol{V}\rangle_F\\ &-2\langle \boldsymbol{\Sigma}_{yx}\boldsymbol{W}'\boldsymbol{T}_p - \boldsymbol{A}\left(\boldsymbol{S}_p \circ (\boldsymbol{B}\boldsymbol{\Sigma}_{xx}\boldsymbol{W}' + \boldsymbol{W}\boldsymbol{\Sigma}_{xx}\boldsymbol{B}')\right), \boldsymbol{V}\rangle_F\\ &+\langle (\boldsymbol{S}_p \circ (\boldsymbol{A}'\boldsymbol{A}))\boldsymbol{W}\boldsymbol{\Sigma}_{xx}, \boldsymbol{W}\rangle_F + O(\varepsilon^3).\end{aligned} \tag{7}$$

Further, for $\boldsymbol{W} = \bar{\boldsymbol{W}} := (\boldsymbol{S}_p \circ (\boldsymbol{A}'\boldsymbol{A}))^{-1}\boldsymbol{T}_p\boldsymbol{V}'\boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}$, the above equation becomes

$$\begin{aligned}L(\boldsymbol{A}+\boldsymbol{V},\boldsymbol{B}+\bar{\boldsymbol{W}}) - L(\boldsymbol{A},\boldsymbol{B}) =& \operatorname{Tr}(\boldsymbol{V}'\boldsymbol{V}\boldsymbol{T}_p\boldsymbol{B}\boldsymbol{\Sigma}_{xx}\boldsymbol{B}') - \operatorname{Tr}\left(\boldsymbol{V}'\boldsymbol{\Sigma}\boldsymbol{V}\boldsymbol{T}_p(\boldsymbol{S}_p \circ (\boldsymbol{A}'\boldsymbol{A}))^{-1}\boldsymbol{T}_p\right)\\ &+2\operatorname{Tr}\Big(\boldsymbol{V}'\boldsymbol{A}\Big(\boldsymbol{S}_p \circ \Big(\boldsymbol{B}\boldsymbol{\Sigma}_{xy}\boldsymbol{V}\boldsymbol{T}_p(\boldsymbol{S}_p \circ (\boldsymbol{A}'\boldsymbol{A}))^{-1}\\ &+\ (\boldsymbol{S}_p \circ (\boldsymbol{A}'\boldsymbol{A}))^{-1}\boldsymbol{T}_p\boldsymbol{V}'\boldsymbol{\Sigma}_{yx}\boldsymbol{B}'\Big)\Big)\Big) + O(\varepsilon^3).\end{aligned} \tag{8}$$

Finally, in case the critical $\boldsymbol{A}$ is of full rank p and so, $(\boldsymbol{A},\boldsymbol{B}) = (\boldsymbol{U}_{\mathbb{I}_p}\boldsymbol{\Pi}\boldsymbol{D}, \hat{\boldsymbol{B}}(\boldsymbol{U}_{\mathbb{I}_p}\boldsymbol{\Pi}\boldsymbol{D}))$, for the encoder direction $\boldsymbol{V}$ with $\|\boldsymbol{V}\|_F = O(\varepsilon)$ and $\boldsymbol{W} = \bar{\boldsymbol{W}}$ we have,

$$\begin{aligned}L(\boldsymbol{A}+\boldsymbol{V},\boldsymbol{B}+\boldsymbol{W}) - L(\boldsymbol{A},\boldsymbol{B}) =& \operatorname{Tr}\left(\boldsymbol{V}'\boldsymbol{V}\boldsymbol{\Pi}'\boldsymbol{\Lambda}_{\mathbb{I}_p}\boldsymbol{\Pi}\boldsymbol{T}_p\boldsymbol{D}^{-2}\right) - \operatorname{Tr}\left(\boldsymbol{V}'\boldsymbol{\Sigma}\boldsymbol{V}\boldsymbol{T}_p\boldsymbol{D}^{-2}\right)\\ &+2\operatorname{Tr}\left(\boldsymbol{V}'\boldsymbol{U}_{\mathbb{I}_p}\boldsymbol{\Pi}\boldsymbol{D}\left(\boldsymbol{S}_p \circ \left(\boldsymbol{D}^{-1}\boldsymbol{\Pi}'\boldsymbol{U}_{\mathbb{I}_p}'\boldsymbol{\Sigma}\boldsymbol{V}\boldsymbol{D}^{-2}\right)\right)\right)\\ &+2\operatorname{Tr}\left(\boldsymbol{V}'\boldsymbol{U}_{\mathbb{I}_p}\boldsymbol{\Pi}\boldsymbol{D}\left(\boldsymbol{S}_p \circ \left(\boldsymbol{D}^{-2}\boldsymbol{V}'\boldsymbol{\Sigma}\boldsymbol{U}_{\mathbb{I}_p}\boldsymbol{\Pi}\boldsymbol{D}^{-1}\right)\right)\right)\\ &+O(\varepsilon^3).\end{aligned} \tag{9}$$

Proof. As described in section 3.1, the second order Taylor expansion for the loss $L(\boldsymbol{A},\boldsymbol{B})$ is then given by eq. (53), i.e.

$$\begin{aligned}L(\boldsymbol{A}+\boldsymbol{V},\boldsymbol{B}+\boldsymbol{W}) - L(\boldsymbol{A},\boldsymbol{B}) =& d_{\boldsymbol{A}}L(\boldsymbol{A},\boldsymbol{B})\boldsymbol{V} + d_{\boldsymbol{B}}L(\boldsymbol{A},\boldsymbol{B})\boldsymbol{W} + \frac{1}{2}d^2_{\boldsymbol{A}}L(\boldsymbol{A},\boldsymbol{B})\boldsymbol{V}^2\\ &+d_{\boldsymbol{A}\boldsymbol{B}}L(\boldsymbol{A},\boldsymbol{B})\boldsymbol{V}\boldsymbol{W} + \frac{1}{2}d^2_{\boldsymbol{B}}L(\boldsymbol{A},\boldsymbol{B})\boldsymbol{W}^2 + R_{\boldsymbol{V},\boldsymbol{W}}(\boldsymbol{A},\boldsymbol{B}).\end{aligned}$$

If $\|\boldsymbol{V}\|_F, \|\boldsymbol{W}\|_F = O(\varepsilon)$ then $\|R(\boldsymbol{V},\boldsymbol{W})\| = O(\varepsilon^3)$. Moreover, when $\boldsymbol{A}$ and $\boldsymbol{B}$ define a critical point of L we have $d_{\boldsymbol{A}}L(\boldsymbol{A},\boldsymbol{B})\boldsymbol{V} = d_{\boldsymbol{B}}L(\boldsymbol{A},\boldsymbol{B})\boldsymbol{W} = 0$. By setting the derivatives $d^2_{\boldsymbol{A}}L(\boldsymbol{A},\boldsymbol{B})\boldsymbol{V}^2$, $d_{\boldsymbol{AB}}L(\boldsymbol{A},\boldsymbol{B})\boldsymbol{VW}$, $d^2_{\boldsymbol{B}}L(\boldsymbol{A},\boldsymbol{B})\boldsymbol{W}^2$ that are given by eq. (59), eq. (58), and eq. (56) respectively, the above equation simplifies to

$$\begin{aligned}
L(\boldsymbol{A}+\boldsymbol{V},\boldsymbol{B}+\boldsymbol{W}) - L(\boldsymbol{A},\boldsymbol{B}) =& \langle \boldsymbol{V}\left(\boldsymbol{S}_p \circ (\boldsymbol{B\Sigma}_{xx}\boldsymbol{B}')\right), \boldsymbol{V}\rangle_F \\
&-2\langle \boldsymbol{\Sigma}_{yx}\boldsymbol{W}'\boldsymbol{T}_p - \boldsymbol{A}\left(\boldsymbol{S}_p \circ (\boldsymbol{B\Sigma}_{xx}\boldsymbol{W}' + \boldsymbol{W\Sigma}_{xx}\boldsymbol{B}')\right), \boldsymbol{V}\rangle_F \\
&+\langle (\boldsymbol{S}_p \circ (\boldsymbol{A}'\boldsymbol{A}))\, \boldsymbol{W\Sigma}_{xx}, \boldsymbol{W}\rangle_F + O(\varepsilon^3).
\end{aligned}$$

Now, based on the first item in Corollary 1, $\boldsymbol{B\Sigma}_{xx}\boldsymbol{B}'$ is a $p \times p$ diagonal matrix, so based on eq. (5): $\boldsymbol{S}_p \circ (\boldsymbol{B\Sigma}_{xx}\boldsymbol{B}') = \boldsymbol{T}_p\boldsymbol{B\Sigma}_{xx}\boldsymbol{B}'$. The substitution then yields eq. (7). Finally, in the above equation replace $\boldsymbol{W}$ with $\bar{\boldsymbol{W}} = (\boldsymbol{S}_p \circ (\boldsymbol{A}'\boldsymbol{A}))^{-1}\boldsymbol{T}_p\boldsymbol{V}'\boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}$. We have

$$\begin{aligned}
&L(\boldsymbol{A}+\boldsymbol{V},\boldsymbol{B}+\bar{\boldsymbol{W}}) - L(\boldsymbol{A},\boldsymbol{B}) = \\
&= \langle \boldsymbol{VT}_p\boldsymbol{B\Sigma}_{xx}\boldsymbol{B}', \boldsymbol{V}\rangle_F - 2\langle \boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}\boldsymbol{VT}_p\left(\boldsymbol{S}_p \circ (\boldsymbol{A}'\boldsymbol{A})\right)^{-1}\boldsymbol{T}_p, \boldsymbol{V}\rangle_F \\
&+2\langle \boldsymbol{A}\Big(\boldsymbol{S}_p \circ \Big(\boldsymbol{B\Sigma}_{xx}\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}\boldsymbol{VT}_p(\boldsymbol{S}_p\circ(\boldsymbol{A}'\boldsymbol{A}))^{-1} + (\boldsymbol{S}_p\circ(\boldsymbol{A}'\boldsymbol{A}))^{-1}\boldsymbol{T}_p\boldsymbol{V}'\boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xx}\boldsymbol{B}'\Big)\Big), \boldsymbol{V}\rangle_F \\
&+\langle (\boldsymbol{S}_p \circ (\boldsymbol{A}'\boldsymbol{A}))\left(\boldsymbol{S}_p \circ (\boldsymbol{A}'\boldsymbol{A})\right)^{-1}\boldsymbol{T}_p\boldsymbol{V}'\boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xx}, \left(\boldsymbol{S}_p \circ (\boldsymbol{A}'\boldsymbol{A})\right)^{-1}\boldsymbol{T}_p\boldsymbol{V}'\boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}\rangle_F + O(\varepsilon^3) \\
&= \operatorname{Tr}\left(\boldsymbol{V}'\boldsymbol{VT}_p\boldsymbol{B\Sigma}_{xx}\boldsymbol{B}'\right) - \operatorname{Tr}\left(\boldsymbol{V}'\boldsymbol{\Sigma VT}_p\left(\boldsymbol{S}_p \circ (\boldsymbol{A}'\boldsymbol{A})\right)^{-1}\boldsymbol{T}_p\right) \\
&+ 2\operatorname{Tr}\left(\boldsymbol{V}'\boldsymbol{A}\left(\boldsymbol{S}_p \circ \left(\boldsymbol{B\Sigma}_{xy}\boldsymbol{VT}_p\left(\boldsymbol{S}_p \circ (\boldsymbol{A}'\boldsymbol{A})\right)^{-1} + \left(\boldsymbol{S}_p \circ (\boldsymbol{A}'\boldsymbol{A})\right)^{-1}\boldsymbol{T}_p\boldsymbol{V}'\boldsymbol{\Sigma}_{yx}\boldsymbol{B}'\right)\right)\right) + O(\varepsilon^3),
\end{aligned}$$

which is eq. (8). For the final equation, we have

$$\begin{aligned}
\boldsymbol{T}_p\boldsymbol{B\Sigma}_{xx}\boldsymbol{B}' =& \boldsymbol{T}_p\boldsymbol{D}^{-1}\boldsymbol{\Pi}'\boldsymbol{U}'_{\mathbb{I}_p}\underbrace{\boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xx}\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}}\boldsymbol{U}_{\mathbb{I}_p}\boldsymbol{\Pi D}^{-1} \\
=& \boldsymbol{T}_p\boldsymbol{D}^{-1}\boldsymbol{\Pi}'\underbrace{\boldsymbol{U}'_{\mathbb{I}_p}\boldsymbol{\Sigma U}_{\mathbb{I}_p}}\boldsymbol{\Pi D}^{-1} = \boldsymbol{T}_p\boldsymbol{D}^{-1}\underbrace{\boldsymbol{\Pi}'\boldsymbol{\Lambda}_{\mathbb{I}_p}\boldsymbol{\Pi}}\boldsymbol{D}^{-1} \\
=& \boldsymbol{\Pi}'\boldsymbol{\Lambda}_{\mathbb{I}_p}\boldsymbol{\Pi T}_p\boldsymbol{D}^{-2}, \text{ and} \qquad (10) \\
\boldsymbol{T}_p\left(\boldsymbol{S}_p \circ (\boldsymbol{A}'\boldsymbol{A})\right)^{-1}\boldsymbol{T}_p =& \boldsymbol{T}_p\left(\boldsymbol{S}_p \circ \left(\boldsymbol{D}\underbrace{\boldsymbol{\Pi}'\boldsymbol{U}'_{\mathbb{I}_p}\boldsymbol{U}_{\mathbb{I}_p}\boldsymbol{\Pi}}\boldsymbol{D}\right)\right)^{-1}\boldsymbol{T}_p \\
=& \boldsymbol{T}_p\left(\boldsymbol{S}_p \circ \boldsymbol{D}^2\right)^{-1}\boldsymbol{T}_p = \boldsymbol{T}_p\boldsymbol{T}_p^{-1}\boldsymbol{D}^{-2}\boldsymbol{T}_p = \boldsymbol{T}_p\boldsymbol{D}^{-2}. \qquad (11)
\end{aligned}$$

Replace the above in eq. (8) and simplify:

$$\begin{aligned}
L(\boldsymbol{A}+\boldsymbol{V},\boldsymbol{B}+\boldsymbol{W}) - L(\boldsymbol{A},\boldsymbol{B}) =& \operatorname{Tr}\left(\boldsymbol{V}'\boldsymbol{VT}_p\boldsymbol{B\Sigma}_{xx}\boldsymbol{B}'\right) - \operatorname{Tr}\left(\boldsymbol{V}'\boldsymbol{\Sigma VT}_p\left(\boldsymbol{S}_p \circ (\boldsymbol{A}'\boldsymbol{A})\right)^{-1}\boldsymbol{T}_p\right) \\
&+2\operatorname{Tr}\Big(\boldsymbol{V}'\boldsymbol{A}\Big(\boldsymbol{S}_p \circ \Big(\boldsymbol{B\Sigma}_{xy}\boldsymbol{VT}_p\left(\boldsymbol{S}_p \circ (\boldsymbol{A}'\boldsymbol{A})\right)^{-1} \\
&+ \left(\boldsymbol{S}_p \circ (\boldsymbol{A}'\boldsymbol{A})\right)^{-1}\boldsymbol{T}_p\boldsymbol{V}'\boldsymbol{\Sigma}_{yx}\boldsymbol{B}'\Big)\Big)\Big) + O(\varepsilon^3) \xrightarrow[\text{eq. (11)}]{\text{eq. (10)}} \\
L(\boldsymbol{A}+\boldsymbol{V},\boldsymbol{B}+\boldsymbol{W}) - L(\boldsymbol{A},\boldsymbol{B}) =& \operatorname{Tr}\left(\boldsymbol{V}'\boldsymbol{V\Pi}'\boldsymbol{\Lambda}_{\mathbb{I}_p}\boldsymbol{\Pi T}_p\boldsymbol{D}^{-2}\right) - \operatorname{Tr}\left(\boldsymbol{V}'\boldsymbol{\Sigma VT}_p\boldsymbol{D}^{-2}\right) \\
&+2\operatorname{Tr}\left(\boldsymbol{V}'\boldsymbol{A}\left(\boldsymbol{S}_p \circ \left(\boldsymbol{B\Sigma}_{xy}\boldsymbol{VD}^{-2} + \boldsymbol{D}^{-2}\boldsymbol{V}'\boldsymbol{\Sigma}_{yx}\boldsymbol{B}'\right)\right)\right)
\end{aligned}$$

$$
\begin{aligned}
&+O(\varepsilon^3) \xRightarrow[\boldsymbol{B}=\hat{\boldsymbol{B}}(\boldsymbol{U}_{\mathbb{I}_p}\boldsymbol{\Pi}\boldsymbol{D})]{\boldsymbol{A}=\boldsymbol{U}_{\mathbb{I}_p}\boldsymbol{\Pi}\boldsymbol{D}} \\
L(\boldsymbol{A}+\boldsymbol{V},\boldsymbol{B}+\boldsymbol{W}) - L(\boldsymbol{A},\boldsymbol{B}) &= \operatorname{Tr}\left(\boldsymbol{V}'\boldsymbol{V}\boldsymbol{\Pi}'\Lambda_{\mathbb{I}_p}\boldsymbol{\Pi}\boldsymbol{T}_p\boldsymbol{D}^{-2}\right) - \operatorname{Tr}\left(\boldsymbol{V}'\boldsymbol{\Sigma}\boldsymbol{V}\boldsymbol{T}_p\boldsymbol{D}^{-2}\right) \\
&+2\operatorname{Tr}\left(\boldsymbol{V}'\boldsymbol{U}_{\mathbb{I}_p}\boldsymbol{\Pi}\boldsymbol{D}\left(\boldsymbol{S}_p \circ \left(\boldsymbol{D}^{-1}\boldsymbol{\Pi}'\boldsymbol{U}_{\mathbb{I}_p}'\boldsymbol{\Sigma}\boldsymbol{V}\boldsymbol{D}^{-2}\right)\right)\right) \\
&+2\operatorname{Tr}\left(\boldsymbol{V}'\boldsymbol{U}_{\mathbb{I}_p}\boldsymbol{\Pi}\boldsymbol{D}\left(\boldsymbol{S}_p \circ \left(\boldsymbol{D}^{-2}\boldsymbol{V}'\boldsymbol{\Sigma}\boldsymbol{U}_{\mathbb{I}_p}\boldsymbol{\Pi}\boldsymbol{D}^{-1}\right)\right)\right) \\
&+O(\varepsilon^3),
\end{aligned}
$$

which finalizes the proof. □

2 Proof of Main Results

Proposition 1. *For any fixed matrix $\boldsymbol{A} \in \mathbb{R}^{n\times p}$ the function $L(\boldsymbol{A},\boldsymbol{B})$ is convex in the coefficients of $\boldsymbol{B}$ and attains its minimum for any $\boldsymbol{B}$ satisfying the equation*

$$(\boldsymbol{S}_p \circ (\boldsymbol{A}'\boldsymbol{A}))\boldsymbol{B}\boldsymbol{\Sigma}_{xx} = \boldsymbol{T}_p\boldsymbol{A}'\boldsymbol{\Sigma}_{yx}, \tag{12}$$

where $\boldsymbol{T}_p$ and $\boldsymbol{S}_p$ are constant matrices defined by Eqs 1 and 2. Further, if $\boldsymbol{A}$ has no zero column, then $L(\boldsymbol{A},\boldsymbol{B})$ is strictly convex in $\boldsymbol{B}$ and has a unique minimum when the critical $\boldsymbol{B}$ is

$$\boldsymbol{B} = \hat{\boldsymbol{B}}(\boldsymbol{A}) = (\boldsymbol{S}_p \circ (\boldsymbol{A}'\boldsymbol{A}))^{-1}\boldsymbol{T}_p\boldsymbol{A}'\boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}, \tag{13}$$

and in the autoencoder case it becomes

$$\boldsymbol{B} = \hat{\boldsymbol{B}}(\boldsymbol{A}) = (\boldsymbol{S}_p \circ (\boldsymbol{A}'\boldsymbol{A}))^{-1}\boldsymbol{T}_p\boldsymbol{A}'. \tag{13'}$$

Proof. For this proof we use the first and second order derivatives for $L(\boldsymbol{A},\boldsymbol{B})$ wrt $\boldsymbol{B}$ derived in Lemma 5. From eq. (56), we have that for a given $\boldsymbol{A}$ the second derivative wrt to $\boldsymbol{B}$ of the cost $L(\boldsymbol{A},\boldsymbol{B})$ at $\boldsymbol{B}$, and in the direction $\boldsymbol{W}$ is the quadratic form

$$d^2_{\boldsymbol{B}^2}L(\boldsymbol{A},\boldsymbol{B})\boldsymbol{W}^2 = 2\operatorname{Tr}\left(\boldsymbol{W}'\left(\boldsymbol{S}_p \circ \boldsymbol{A}'\boldsymbol{A}\right)\boldsymbol{W}\boldsymbol{\Sigma}_{xx}\right).$$

The matrix $\boldsymbol{\Sigma}_{xx}$ is positive-definite and by Lemma 1, $\boldsymbol{S}_p \circ \boldsymbol{A}'\boldsymbol{A}$ is positive-semidefinite. Hence, $d^2_{\boldsymbol{B}^2}L(\boldsymbol{A},\boldsymbol{B})\boldsymbol{W}^2$ is clearly non-negative for all $\boldsymbol{W} \in \mathbb{R}^{p\times n}$. Therefore, $L(\boldsymbol{A},\boldsymbol{B})$ is convex in coefficients of $\boldsymbol{B}$ for a fixed matrix $\boldsymbol{A}$. Also the critical points of $L(\boldsymbol{A},\boldsymbol{B})$ for a fixed $\boldsymbol{A}$ is a matrix $\boldsymbol{B}$ that satisfies $\forall \boldsymbol{W} \in \mathbb{R}^{p\times n} : d_{\boldsymbol{B}}L(\boldsymbol{A},\boldsymbol{B})\boldsymbol{W} = 0$ and hence, from eq. (54) we have

$$-2\langle \boldsymbol{T}_p\boldsymbol{A}'\boldsymbol{\Sigma}_{yx} - \left(\boldsymbol{S}_p \circ (\boldsymbol{A}'\boldsymbol{A})\right)\boldsymbol{B}\boldsymbol{\Sigma}_{xx}, \boldsymbol{W}\rangle_F = 0.$$

Setting $\boldsymbol{W} = \boldsymbol{T}_p\boldsymbol{A}'\boldsymbol{\Sigma}_{yx} - \left(\boldsymbol{S}_p \circ (\boldsymbol{A}'\boldsymbol{A})\right)\boldsymbol{B}\boldsymbol{\Sigma}_{xx}$ we have

$$\boldsymbol{T}_p\boldsymbol{A}'\boldsymbol{\Sigma}_{yx} - \left(\boldsymbol{S}_p \circ (\boldsymbol{A}'\boldsymbol{A})\right)\boldsymbol{B}\boldsymbol{\Sigma}_{xx} = 0.$$

For a fixed $\boldsymbol{A}$, the cost $L(\boldsymbol{A},\boldsymbol{B})$ is convex in $\boldsymbol{B}$, so any matrix $\boldsymbol{B}$ that satisfies the above equation corresponds to a minimum of $L(\boldsymbol{A},\boldsymbol{B})$. Further, if $\boldsymbol{A}$ has no zero column then by Lemma 1, $\boldsymbol{S}_p \circ \boldsymbol{A}'\boldsymbol{A}$ is positive definite. Hence, $\forall \boldsymbol{W} \in \mathbb{R}^{p\times n} : d^2_{\boldsymbol{B}^2}L(\boldsymbol{A},\boldsymbol{B})\boldsymbol{W}^2 = 2\operatorname{Tr}\left(\boldsymbol{W}'\left(\boldsymbol{S}_p \circ \boldsymbol{A}'\boldsymbol{A}\right)\boldsymbol{W}\boldsymbol{\Sigma}_{xx}\right)$ is positive. Therefore, the cost $L(\boldsymbol{A},\boldsymbol{B})$ becomes strictly convex and the unique global minimum is achieved at $\boldsymbol{B} = \hat{\boldsymbol{B}}(\boldsymbol{A})$ as defined in eq. (13). □

Proposition 2. *For any fixed matrix $\boldsymbol{B} \in \mathbb{R}^{p\times n}$ the function $L(\boldsymbol{A},\boldsymbol{B})$ is a convex function in $\boldsymbol{A}$. Moreover, for a fixed $\boldsymbol{B}$, the matrix $\boldsymbol{A}$ that satisfies*

$$\boldsymbol{A}\left(\boldsymbol{S}_p \circ (\boldsymbol{B}\boldsymbol{\Sigma}_{xx}\boldsymbol{B}')\right) = \boldsymbol{\Sigma}_{yx}\boldsymbol{B}'\boldsymbol{T}_p \tag{14}$$

is a critical point of $L(\boldsymbol{A},\boldsymbol{B})$.

Proof. For this proof we use the first and second order derivatives for $L(\boldsymbol{A},\boldsymbol{B})$ wrt $\boldsymbol{A}$ derived in Lemma 6. For a fixed $\boldsymbol{B}$, based on eq. (59) the second derivative wrt to $\boldsymbol{A}$ of $L(\boldsymbol{A},\boldsymbol{B})$ at $\boldsymbol{A}$, and in the direction $\boldsymbol{V}$ is the quadratic form

$$d^2_{\boldsymbol{A}^2} L(\boldsymbol{A},\boldsymbol{B})\boldsymbol{V}^2 = 2\langle \boldsymbol{V}\left(\boldsymbol{S}_p \circ (\boldsymbol{B}\boldsymbol{\Sigma}_{xx}\boldsymbol{B}')\right), \boldsymbol{V}\rangle_F = 2\,\mathrm{Tr}\left(\boldsymbol{V}\left(\boldsymbol{S}_p \circ (\boldsymbol{B}\boldsymbol{\Sigma}_{xx}\boldsymbol{B}')\right)\boldsymbol{V}'\right).$$

The matrix $\boldsymbol{\Sigma}_{xx}$ is positive-definite and by Lemma 1, $\boldsymbol{S}_p \circ (\boldsymbol{B}\boldsymbol{\Sigma}_{xx}\boldsymbol{B}')$ is positive-semidefinite. Hence, $d^2_{\boldsymbol{A}^2} L_1(\boldsymbol{A},\boldsymbol{B})\boldsymbol{V}^2$is non-negative for all $\boldsymbol{V} \in \mathbb{R}^{n\times p}$. Therefore, $L(\boldsymbol{A},\boldsymbol{B})$ is convex in coefficients of $\boldsymbol{A}$ for a fixed matrix $\boldsymbol{B}$. Based on eq. (57) the critical point of $L(\boldsymbol{A},\boldsymbol{B})$ for a fixed $\boldsymbol{B}$ is a matrix $\boldsymbol{A}$ that satisfies for all $\boldsymbol{V} \in \mathbb{R}^{n\times p}$

$$\begin{aligned} d_{\boldsymbol{A}} L(\boldsymbol{A},\boldsymbol{B})\boldsymbol{V} = \langle -2\left(\boldsymbol{\Sigma}_{yx}\boldsymbol{B}'\boldsymbol{T}_p - \boldsymbol{A}\left(\boldsymbol{S}_p \circ (\boldsymbol{B}\boldsymbol{\Sigma}_{xx}\boldsymbol{B}')\right)\right), \boldsymbol{V}\rangle_F = 0 \implies \\ \boldsymbol{\Sigma}_{yx}\boldsymbol{B}'\boldsymbol{T}_p = \boldsymbol{A}\left(\boldsymbol{S}_p \circ (\boldsymbol{B}\boldsymbol{\Sigma}_{xx}\boldsymbol{B}')\right), \end{aligned}$$

which is eq. (14). □

Theorem 1. *Let $\boldsymbol{A} \in \mathbb{R}^{n\times p}$ and $\boldsymbol{B} \in \mathbb{R}^{p\times n}$ be such that $\boldsymbol{A}$ is of rank $r \le p$. Under the given assumptions, the matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ define a critical point of $L(\boldsymbol{A},\boldsymbol{B})$ if and only if for any given r-index set $\mathbb{I}_r$, and a nonsingular diagonal matrix $\boldsymbol{D} \in \mathbb{R}^{r\times r}$, $\boldsymbol{A}$ and $\boldsymbol{B}$ are of the form*

$$\boldsymbol{A} = \boldsymbol{U}_{\mathbb{I}_r}\boldsymbol{C}\boldsymbol{D}, \tag{15}$$

$$\boldsymbol{B} = \boldsymbol{D}^{-1}\boldsymbol{\Pi}_{\boldsymbol{C}}\boldsymbol{U}'_{\mathbb{I}_r}\boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}, \tag{16}$$

where, $\boldsymbol{C} \in \mathbb{R}^{r\times p}$ is of full rank r with nonzero and normalized columns such that $\boldsymbol{\Pi}_{\boldsymbol{C}} := \left(\boldsymbol{S}_p \circ (\boldsymbol{C}'\boldsymbol{C})\right)^{-1}\boldsymbol{T}_p\boldsymbol{C}'$ is a rectangular permutation matrix of rank r and $\boldsymbol{C}\boldsymbol{\Pi}_{\boldsymbol{C}} = \boldsymbol{I}_r$. For all $1 \le r \le p$, such $\boldsymbol{C}$ always exists. In particular, if matrix $\boldsymbol{A}$ is of full rank p, i.e. $r = p$, the two given conditions on $\boldsymbol{\Pi}_{\boldsymbol{C}}$ are satisfied iff the invertible matrix $\boldsymbol{C}$ is any squared $p\times p$ permutation matrix $\boldsymbol{\Pi}$. In this case $(\boldsymbol{A},\boldsymbol{B})$ define a critical point of $L(\boldsymbol{A},\boldsymbol{B})$ iff they are of the form

$$\boldsymbol{A} = \boldsymbol{U}_{\mathbb{I}_p}\boldsymbol{\Pi}\boldsymbol{D}, \tag{17}$$

$$\boldsymbol{B} = \boldsymbol{D}^{-1}\boldsymbol{\Pi}'\boldsymbol{U}'_{\mathbb{I}_p}\boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}. \tag{18}$$

Proof. Before we start, a reminder on notation and some useful identities that are used throughout the proof. The matrix $\boldsymbol{\Sigma} := \boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}$ has an eigenvalue decomposition $\boldsymbol{\Sigma} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}'$, where the i^{th} column of $\boldsymbol{U}$, denoted as $\boldsymbol{u}_i$, is an eigenvector of $\boldsymbol{\Sigma}$ corresponding to the i^{th} largest eigenvalue of $\boldsymbol{\Sigma}$, denoted as λ_i. Also, $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \cdots, \lambda_n)$ is the diagonal vector of ordered eigenvalues of $\boldsymbol{\Sigma}$, with $\lambda_1 > \lambda_2 > \cdots > \lambda_n > 0$. We use the following notation to organize a subset of eigenvectors of $\boldsymbol{\Sigma}$ into a rectangular matrix. Let for any $r \le p$, $\mathbb{I}_r = \{i_1, \cdots, i_r\}(1 \le i_1 < \cdots < i_r < n)$ be any *ordered* $r-$index set. Define $\boldsymbol{U}_{\mathbb{I}_r} \in \mathbb{R}^{n\times p}$ as $\boldsymbol{U}_{\mathbb{I}_r} = [\boldsymbol{u}_{i_1}, \cdots, \boldsymbol{u}_{i_r}]$. That is the columns of $\boldsymbol{U}_{\mathbb{I}_r}$ are the ordered orthonormal eigenvectors of $\boldsymbol{\Sigma}$ associated with eigenvalues $\lambda_{i_1} < \cdots < \lambda_{i_r}$. The following identities are then easy to verify:

$$\boldsymbol{U}'_{\mathbb{I}_r}\boldsymbol{U}_{\mathbb{I}_r} = \boldsymbol{I}_r,$$

$$\boldsymbol{\Sigma U}_{\mathbb{I}_r} = \boldsymbol{U}_{\mathbb{I}_r}\boldsymbol{\Lambda}_{\mathbb{I}_r}, \tag{19}$$

$$\boldsymbol{U}'_{\mathbb{I}_r}\boldsymbol{\Sigma U}_{\mathbb{I}_r} = \boldsymbol{\Lambda}_{\mathbb{I}_r}. \tag{20}$$

The sufficient condition:

Let $\boldsymbol{A} \in \mathbb{R}^{n\times p}$of rank $r \le p$ and no zero column be given by eq. (15), $\boldsymbol{B} \in \mathbb{R}^{p\times n}$ given by eq. (16), and the accompanying conditions are met. Notice that $\boldsymbol{U}'_{\mathbb{I}_r}\boldsymbol{U}_{\mathbb{I}_r} = \boldsymbol{I}_r$ implies that $\boldsymbol{DC'CD} = \boldsymbol{DC'U}'_{\mathbb{I}_r}\boldsymbol{U}_{\mathbb{I}_r}\boldsymbol{CD} = \boldsymbol{A'A}$, so

$$\begin{aligned}
\boldsymbol{B} &= \boldsymbol{D}^{-1}\boldsymbol{\Pi_C U}'_{\mathbb{I}_r}\boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1} \xRightarrow[\boldsymbol{D}^{-1}\boldsymbol{D}=\boldsymbol{I}_p]{\boldsymbol{\Pi_C}:=\left(\boldsymbol{S}_p\circ(\boldsymbol{C'C})\right)^{-1}\boldsymbol{T}_p\boldsymbol{C}'} \\
\boldsymbol{B} &= \boldsymbol{D}^{-1}\left(\boldsymbol{S}_p\circ(\boldsymbol{C'C})\right)^{-1}\boldsymbol{D}^{-1}\boldsymbol{DT}_p\boldsymbol{C'U}'_{\mathbb{I}_r}\boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1} \xRightarrow[\boldsymbol{DT}_p=\boldsymbol{T}_p\boldsymbol{D}]{\text{Lemma 1-2}} \\
\boldsymbol{B} &= \left(\boldsymbol{S}_p\circ\underbrace{(\boldsymbol{DC'CD})}\right)^{-1}\boldsymbol{T}_p\underbrace{\boldsymbol{DC'U}'_{\mathbb{I}_r}}\boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1} \xRightarrow[\boldsymbol{DC'CD}=\boldsymbol{A'A}]{\boldsymbol{A}'=\boldsymbol{D'C'U}'_{\mathbb{I}_r}} \\
\boldsymbol{B} &= \left(\boldsymbol{S}_p\circ(\boldsymbol{A'A})\right)^{-1}\boldsymbol{T}_p\boldsymbol{A}'\boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1} = \hat{\boldsymbol{B}}(\boldsymbol{A}),
\end{aligned}$$

which is eq. (13). Therefore, based on Proposition 1, for the given $\boldsymbol{A}$, the matrix $\boldsymbol{B}$ defines a critical point of $L(\boldsymbol{A},\boldsymbol{B})$. For the gradient wrt to $\boldsymbol{A}$, first note that with $\boldsymbol{B}$ given by eq. (16) we have

$$\begin{aligned}
\boldsymbol{B\Sigma}_{xx}\boldsymbol{B}' &= \boldsymbol{D}^{-1}\boldsymbol{\Pi_C U}'_{\mathbb{I}_r}\boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xx}\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}\boldsymbol{U}_{\mathbb{I}_r}\boldsymbol{\Pi}'_{\boldsymbol{C}}\boldsymbol{D}^{-1} \\
&= \boldsymbol{D}^{-1}\boldsymbol{\Pi_C}\underbrace{\boldsymbol{U}'_{\mathbb{I}_r}\boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}\boldsymbol{U}_{\mathbb{I}_r}}\boldsymbol{\Pi}'_{\boldsymbol{C}}\boldsymbol{D}^{-1} \xRightarrow{\text{eq. (20)}} \\
\boldsymbol{B\Sigma}_{xx}\boldsymbol{B}' &= \boldsymbol{D}^{-1}\boldsymbol{\Pi_C}\boldsymbol{\Lambda}_{\mathbb{I}_r}\boldsymbol{\Pi}'_{\boldsymbol{C}}\boldsymbol{D}^{-1}.
\end{aligned} \tag{21}$$

The matrix $\boldsymbol{\Pi_C}$ is a rectangular permutation matrix so $\boldsymbol{\Pi_C\Lambda}_{\mathbb{I}_r}\boldsymbol{\Pi}'_{\boldsymbol{C}}$ is diagonal so as $\boldsymbol{D}^{-1}\boldsymbol{\Pi_C\Lambda}_{\mathbb{I}_r}\boldsymbol{\Pi}'_{\boldsymbol{C}}\boldsymbol{D}^{-1}$. Therefore, $\boldsymbol{B\Sigma}_{xx}\boldsymbol{B}'$ is diagonal and by eq. (5) in Lemma 1-6 we have

$$\begin{aligned}
\boldsymbol{S}_p\circ(\boldsymbol{B\Sigma}_{xx}\boldsymbol{B}') &= \boldsymbol{T}_p\boldsymbol{B\Sigma}_{xx}\boldsymbol{B}' = \boldsymbol{B\Sigma}_{xx}\boldsymbol{B}'\boldsymbol{T}_p \\
&= \boldsymbol{D}^{-1}\boldsymbol{\Pi_C\Lambda}_{\mathbb{I}_r}\boldsymbol{\Pi}'_{\boldsymbol{C}}\boldsymbol{D}^{-1}\boldsymbol{T}_p \xRightarrow{\boldsymbol{A}\times} \\
\boldsymbol{A}\left(\boldsymbol{S}_p\circ(\boldsymbol{B\Sigma}_{xx}\boldsymbol{B}')\right) &= \boldsymbol{AD}^{-1}\boldsymbol{\Pi_C\Lambda}_{\mathbb{I}_r}\boldsymbol{\Pi}'_{\boldsymbol{C}}\boldsymbol{D}^{-1}\boldsymbol{T}_p \xRightarrow{\boldsymbol{A}=\boldsymbol{U}_{\mathbb{I}_r}\boldsymbol{CD}} \\
\boldsymbol{A}\left(\boldsymbol{S}_p\circ(\boldsymbol{B\Sigma}_{xx}\boldsymbol{B}')\right) &= \boldsymbol{U}_{\mathbb{I}_r}\boldsymbol{CDD}^{-1}\boldsymbol{\Pi_C\Lambda}_{\mathbb{I}_r}\boldsymbol{\Pi}'_{\boldsymbol{C}}\boldsymbol{D}^{-1}\boldsymbol{T}_p \xRightarrow{\boldsymbol{A}=\boldsymbol{U}_{\mathbb{I}_r}\boldsymbol{CD}} \\
&= \boldsymbol{U}_{\mathbb{I}_r}\underbrace{\boldsymbol{C\Pi_C}}\boldsymbol{\Lambda}_{\mathbb{I}_r}\boldsymbol{\Pi}'_{\boldsymbol{C}}\boldsymbol{D}^{-1}\boldsymbol{T}_p \xRightarrow{\boldsymbol{C\Pi_C}=\boldsymbol{I}_r} \\
\boldsymbol{A}\left(\boldsymbol{S}_p\circ(\boldsymbol{B\Sigma}_{xx}\boldsymbol{B}')\right) &= \underbrace{\boldsymbol{U}_{\mathbb{I}_r}\boldsymbol{\Lambda}_{\mathbb{I}_r}}\boldsymbol{\Pi}'_{\boldsymbol{C}}\boldsymbol{D}^{-1}\boldsymbol{T}_p \xRightarrow{\text{eq. (19)}} \\
&= \boldsymbol{\Sigma U}_{\mathbb{I}_r}\boldsymbol{\Pi}'_{\boldsymbol{C}}\boldsymbol{D}^{-1}\boldsymbol{T}_p \\
&= \boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}\boldsymbol{U}_{\mathbb{I}_r}\boldsymbol{\Pi}'_{\boldsymbol{C}}\boldsymbol{D}^{-1}\boldsymbol{T}_p \\
&= \boldsymbol{\Sigma}_{yx}\underbrace{\left(\boldsymbol{D}^{-1}\boldsymbol{\Pi_C U}'_{\mathbb{I}_r}\boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}\right)'}\boldsymbol{T}_p \\
&= \boldsymbol{\Sigma}_{yx}\boldsymbol{B}'\boldsymbol{T}_p,
\end{aligned}$$

which is eq. (14). Therefore, based on Proposition Proposition 2, for the given $\boldsymbol{B}$, the matrix $\boldsymbol{A}$ define a critical point of $L(\boldsymbol{A},\boldsymbol{B})$. Hence, $\boldsymbol{A}$ and $\boldsymbol{B}$ together define a critical point of $L(\boldsymbol{A},\boldsymbol{B})$.

The necessary condition:

Based on Proposition 1 and Proposition 2, for $\boldsymbol{A}$ (with no zero column) and $\boldsymbol{B}$, to define a critical point of $L(\boldsymbol{A},\boldsymbol{B})$, $\boldsymbol{B}$ has to be $\hat{\boldsymbol{B}}(\boldsymbol{A})$ given by eq. (13), and $\boldsymbol{A}$ has to satisfy eq. (14). That is

$$\begin{aligned}
\boldsymbol{A}\left(\boldsymbol{S}_p \circ \left(\hat{\boldsymbol{B}}\boldsymbol{\Sigma}_{xx}\hat{\boldsymbol{B}}'\right)\right) =& \boldsymbol{\Sigma}_{yx}\hat{\boldsymbol{B}}'\boldsymbol{T}_p \xrightarrow{\hat{\boldsymbol{B}}(\boldsymbol{A}) \text{ on RHS}} \\
\boldsymbol{A}\left(\boldsymbol{S}_p \circ \left(\hat{\boldsymbol{B}}\boldsymbol{\Sigma}_{xx}\hat{\boldsymbol{B}}'\right)\right) =& \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{yx}\boldsymbol{A}\boldsymbol{T}_p(\boldsymbol{S}_p\circ(\boldsymbol{A}'\boldsymbol{A}))^{-1}\boldsymbol{T}_p \xrightarrow[\boldsymbol{\Sigma}=\boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{yx}]{\times \boldsymbol{A}'} \\
\boldsymbol{A}\left(\boldsymbol{S}_p \circ \left(\hat{\boldsymbol{B}}\boldsymbol{\Sigma}_{xx}\hat{\boldsymbol{B}}'\right)\right)\boldsymbol{A}' =& \boldsymbol{\Sigma}\boldsymbol{A}\boldsymbol{T}_p(\boldsymbol{S}_p\circ(\boldsymbol{A}'\boldsymbol{A}))^{-1}\boldsymbol{T}_p\boldsymbol{A}' \xrightarrow[\times \boldsymbol{U},\boldsymbol{U}'\times]{\boldsymbol{\Sigma}=\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}''} \\
\boldsymbol{U}'\boldsymbol{A}\left(\boldsymbol{S}_p \circ \left(\hat{\boldsymbol{B}}\boldsymbol{\Sigma}_{xx}\hat{\boldsymbol{B}}'\right)\right)\boldsymbol{A}'\boldsymbol{U} =& \boldsymbol{U}'\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}'\boldsymbol{A}\boldsymbol{T}_p(\boldsymbol{S}_p\circ(\boldsymbol{A}'\boldsymbol{A}))^{-1}\boldsymbol{T}_p\boldsymbol{A}'\boldsymbol{U} \xrightarrow{\boldsymbol{U}'\boldsymbol{U}=\boldsymbol{I}_n} \\
\boldsymbol{U}'\boldsymbol{A}\left(\boldsymbol{S}_p \circ \left(\hat{\boldsymbol{B}}\boldsymbol{\Sigma}_{xx}\hat{\boldsymbol{B}}'\right)\right)\boldsymbol{A}'\boldsymbol{U} =& \boldsymbol{\Lambda}\boldsymbol{\Delta},
\end{aligned} \tag{22}$$

where, $\boldsymbol{\Delta} := \boldsymbol{U}'\boldsymbol{A}\boldsymbol{T}_p(\boldsymbol{S}_p\circ(\boldsymbol{A}'\boldsymbol{A}))^{-1}\boldsymbol{T}_p\boldsymbol{A}'\boldsymbol{U}$ is symmetric and positive semidefinite. The LHS of the above equation is symmetric so the RHS is symmetric too, so $\boldsymbol{\Lambda}\boldsymbol{\Delta} = (\boldsymbol{\Lambda}\boldsymbol{\Delta})' = \boldsymbol{\Delta}'\boldsymbol{\Lambda}' = \boldsymbol{\Delta}\boldsymbol{\Lambda}$. Therefore, $\boldsymbol{\Delta}$ commutes with the diagonal matrix of eigenvalues $\boldsymbol{\Lambda}$. Since, eigenvalues are assumed to be distinct, $\boldsymbol{\Delta}$ has to be diagonal as well. By Lemma 1 $\boldsymbol{T}_p(\boldsymbol{S}_p\circ(\boldsymbol{A}'\boldsymbol{A}))^{-1}\boldsymbol{T}_p$ is positive definite and $\boldsymbol{U}$ is an orthogonal matrix. Therefore, $r = \operatorname{rank}(\boldsymbol{A}) = \operatorname{rank}(\boldsymbol{\Delta}) = \operatorname{rank}(\boldsymbol{U}'\boldsymbol{\Delta}\boldsymbol{U})$, which implies that the diagonal matrix $\boldsymbol{\Delta}$, has r nonzero and *positive* diagonal entries. There exists an $r-$index set $\mathbb{I}_r$ corresponding to the nonzero diagonal elements of $\boldsymbol{\Delta}$. Forming a diagonal matrix $\boldsymbol{\Delta}_{\mathbb{I}_r} \in \mathbb{R}^{r\times r}$ by filling its diagonal entries (in order) by the nonzero diagonal elements of $\boldsymbol{\Delta}$ we have

$$\begin{aligned}
\boldsymbol{U}\boldsymbol{\Delta}\boldsymbol{U}' =& \boldsymbol{U}_{\mathbb{I}_r}\boldsymbol{\Delta}_{\mathbb{I}_r}\boldsymbol{U}'_{\mathbb{I}_r} \xrightarrow{\text{Def of } \boldsymbol{\Delta}} \\
\boldsymbol{U}\boldsymbol{U}'\boldsymbol{A}\boldsymbol{T}_p(\boldsymbol{S}_p\circ(\boldsymbol{A}'\boldsymbol{A}))^{-1}\boldsymbol{T}_p\boldsymbol{A}'\boldsymbol{U}\boldsymbol{U}' =& \boldsymbol{U}_{\mathbb{I}_r}\boldsymbol{\Delta}_{\mathbb{I}_r}\boldsymbol{U}'_{\mathbb{I}_r} \xrightarrow{\boldsymbol{U}\boldsymbol{U}'=\boldsymbol{I}_n} \\
\boldsymbol{A}\boldsymbol{T}_p(\boldsymbol{S}_p\circ(\boldsymbol{A}'\boldsymbol{A}))^{-1}\boldsymbol{T}_p\boldsymbol{A}' =& \boldsymbol{U}_{\mathbb{I}_r}\boldsymbol{\Delta}_{\mathbb{I}_r}\boldsymbol{U}'_{\mathbb{I}_r},
\end{aligned} \tag{23}$$

which indicates that the matrix $\boldsymbol{A}$ has the same column space as $\boldsymbol{U}_{\mathbb{I}_r}$. Therefore, there exists a full rank matrix $\tilde{\boldsymbol{C}} \in \mathbb{R}^{r\times p}$ such that $\boldsymbol{A} = \boldsymbol{U}_{\mathbb{I}_r}\tilde{\boldsymbol{C}}$. Since $\boldsymbol{A}$ has no zero column, $\tilde{\boldsymbol{C}}$ has no zero column. Further, by normalizing the columns of $\tilde{\boldsymbol{C}}$ we can write $\boldsymbol{A} = \boldsymbol{U}_{\mathbb{I}_r}\boldsymbol{C}\boldsymbol{D}$, where $\boldsymbol{D} \in \mathbb{R}^{p\times p}$ is diagonal that contains the norms of columns of $\tilde{\boldsymbol{C}}$. Therefore, $\boldsymbol{A}$ is exactly in the form given by eq. (15). The matrix $\boldsymbol{C}$ has to satisfy eq. (23) that is

$$\begin{aligned}
\boldsymbol{A}\boldsymbol{T}_p(\boldsymbol{S}_p\circ(\boldsymbol{A}'\boldsymbol{A}))^{-1}\boldsymbol{T}_p\boldsymbol{A}' =& \boldsymbol{U}_{\mathbb{I}_r}\boldsymbol{\Delta}_{\mathbb{I}_r}\boldsymbol{U}'_{\mathbb{I}_r} \xrightarrow{\boldsymbol{A}=\boldsymbol{U}_{\mathbb{I}_r}\boldsymbol{C}} \\
\boldsymbol{U}_{\mathbb{I}_r}\boldsymbol{C}\boldsymbol{D}\boldsymbol{T}_p(\boldsymbol{S}_p\circ(\boldsymbol{A}'\boldsymbol{A}))^{-1}\boldsymbol{T}_p\boldsymbol{D}\boldsymbol{C}'\boldsymbol{U}'_{\mathbb{I}_r} =& \boldsymbol{U}_{\mathbb{I}_r}\boldsymbol{\Delta}_{\mathbb{I}_r}\boldsymbol{U}'_{\mathbb{I}_r} \xrightarrow[\boldsymbol{A}'\boldsymbol{A}=\boldsymbol{D}\boldsymbol{C}'\boldsymbol{C}\boldsymbol{D}]{\times\boldsymbol{U}_{\mathbb{I}_r},\boldsymbol{U}_{\mathbb{I}_r}\times} \\
\boldsymbol{C}\boldsymbol{D}\boldsymbol{T}_p(\boldsymbol{S}_p\circ(\boldsymbol{D}\boldsymbol{C}'\boldsymbol{C}\boldsymbol{D}))^{-1}\boldsymbol{T}_p\boldsymbol{C}'\boldsymbol{D} =& \boldsymbol{\Delta}_{\mathbb{I}_r} \xrightarrow{\text{Lemma 1-2}} \\
\boldsymbol{C}\boldsymbol{T}_p\boldsymbol{D}\boldsymbol{D}^{-1}(\boldsymbol{S}_p\circ(\boldsymbol{C}'\boldsymbol{C}))^{-1}\boldsymbol{D}^{-1}\boldsymbol{D}\boldsymbol{T}_p\boldsymbol{C}' =& \boldsymbol{\Delta}_{\mathbb{I}_r} \implies \\
\boldsymbol{C}\boldsymbol{T}_p(\boldsymbol{S}_p\circ(\boldsymbol{C}'\boldsymbol{C}))^{-1}\boldsymbol{T}_p\boldsymbol{C}' =& \boldsymbol{\Delta}_{\mathbb{I}_r}.
\end{aligned} \tag{24}$$

Now that the structure of $\boldsymbol{A}$ has been identified, evaluate $\hat{\boldsymbol{B}}(\boldsymbol{A})$ of eq. (13) by setting $\boldsymbol{A} = \boldsymbol{U}_{\mathbb{I}_r}\boldsymbol{C}\boldsymbol{D}$, that is

$$\boldsymbol{B} = \hat{\boldsymbol{B}}(\boldsymbol{A}) = (\boldsymbol{S}_p\circ(\boldsymbol{A}'\boldsymbol{A}))^{-1}\boldsymbol{T}_p\boldsymbol{A}'\boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}$$

$$
\begin{aligned}
&=(\boldsymbol{S}_p \circ (\boldsymbol{D}\boldsymbol{C}'\boldsymbol{C}\boldsymbol{D}))^{-1}\boldsymbol{T}_p\boldsymbol{D}\boldsymbol{C}'\boldsymbol{U}'_{\mathbb{I}_r}\boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1} \xRightarrow{\text{Lemma 1-2}} \\
\boldsymbol{B} &=\boldsymbol{D}^{-1}(\boldsymbol{S}_p \circ (\boldsymbol{C}'\boldsymbol{C}))^{-1}\boldsymbol{T}_p\boldsymbol{C}'\boldsymbol{U}'_{\mathbb{I}_r}\boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1},
\end{aligned}
$$

which by defining $\boldsymbol{\Pi_C} \coloneqq (\boldsymbol{S}_p \circ (\boldsymbol{C}'\boldsymbol{C}))^{-1}\boldsymbol{T}_p\boldsymbol{C}'$ gives eq. (19) for $\boldsymbol{B}$ as claimed. While $\boldsymbol{C}$ has to satisfy eq. (24), $\boldsymbol{A}$ and $\boldsymbol{B}$ in the given form have to satisfy eq. (22) that provides another condition for $\boldsymbol{C}$ as follows. First, note that

$$
\begin{aligned}
\boldsymbol{S}_p \circ \left(\hat{\boldsymbol{B}}\boldsymbol{\Sigma}_{xx}\hat{\boldsymbol{B}}'\right) &= \boldsymbol{S}_p \circ \left(\boldsymbol{D}^{-1}(\boldsymbol{S}_p \circ (\boldsymbol{C}'\boldsymbol{C}))^{-1}\boldsymbol{T}_p\boldsymbol{C}'\boldsymbol{U}'_{\mathbb{I}_r}\boldsymbol{\Sigma}\boldsymbol{U}_{\mathbb{I}_r}\boldsymbol{C}\boldsymbol{T}_p(\boldsymbol{S}_p \circ (\boldsymbol{C}'\boldsymbol{C}))^{-1}\boldsymbol{D}^{-1}\right) \\
&= \boldsymbol{S}_p \circ \left(\boldsymbol{D}^{-1}(\boldsymbol{S}_p \circ (\boldsymbol{C}'\boldsymbol{C}))^{-1}\boldsymbol{T}_p\boldsymbol{C}'\boldsymbol{\Lambda}_{\mathbb{I}_r}\boldsymbol{C}\boldsymbol{T}_p(\boldsymbol{S}_p \circ (\boldsymbol{C}'\boldsymbol{C}))^{-1}\boldsymbol{D}^{-1}\right) \xRightarrow{\text{Lemma 1-2}} \\
&= \boldsymbol{D}^{-1}\left(\boldsymbol{S}_p \circ \left((\boldsymbol{S}_p \circ (\boldsymbol{C}'\boldsymbol{C}))^{-1}\boldsymbol{T}_p\boldsymbol{C}'\boldsymbol{\Lambda}_{\mathbb{I}_r}\boldsymbol{C}\boldsymbol{T}_p(\boldsymbol{S}_p \circ (\boldsymbol{C}'\boldsymbol{C}))^{-1}\right)\right)\boldsymbol{D}^{-1}
\end{aligned}
$$

Now, replace $\boldsymbol{A}$ and $\boldsymbol{B}$ in eq. (22) by their respective identities that we just derived. Performing the same process for eq. (22) we have

$$
\begin{aligned}
\boldsymbol{U}'\boldsymbol{A}\left(\boldsymbol{S}_p \circ \left(\hat{\boldsymbol{B}}\boldsymbol{\Sigma}_{xx}\hat{\boldsymbol{B}}'\right)\right)\boldsymbol{A}'\boldsymbol{U} &= \boldsymbol{\Lambda}\boldsymbol{\Delta} \xRightarrow[\times \boldsymbol{U}',\, \boldsymbol{U}\times]{\boldsymbol{A}=\boldsymbol{U}_{\mathbb{I}_r}\boldsymbol{C}\boldsymbol{D}} \\
\boldsymbol{U}_{\mathbb{I}_r}\boldsymbol{C}\left(\boldsymbol{S}_p \circ \left((\boldsymbol{S}_p \circ (\boldsymbol{C}'\boldsymbol{C}))^{-1}\boldsymbol{T}_p\boldsymbol{C}'\boldsymbol{\Lambda}_{\mathbb{I}_r}\boldsymbol{C}\boldsymbol{T}_p(\boldsymbol{S}_p \circ (\boldsymbol{C}'\boldsymbol{C}))^{-1}\right)\right)\boldsymbol{C}'\boldsymbol{U}'_{\mathbb{I}_r} &= \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{\Delta}\boldsymbol{U}' \xRightarrow[\boldsymbol{U}'_{\mathbb{I}_r}\times]{\times \boldsymbol{U}_{\mathbb{I}_r}} \\
\boldsymbol{C}\left(\boldsymbol{S}_p \circ \left((\boldsymbol{S}_p \circ (\boldsymbol{C}'\boldsymbol{C}))^{-1}\boldsymbol{T}_p\boldsymbol{C}'\boldsymbol{\Lambda}_{\mathbb{I}_r}\boldsymbol{C}\boldsymbol{T}_p(\boldsymbol{S}_p \circ (\boldsymbol{C}'\boldsymbol{C}))^{-1}\right)\right)\boldsymbol{C}' &= \boldsymbol{U}'_{\mathbb{I}_r}\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{\Delta}\boldsymbol{U}'\boldsymbol{U}_{\mathbb{I}_r} \implies \\
\boldsymbol{C}\left(\boldsymbol{S}_p \circ \left((\boldsymbol{S}_p \circ (\boldsymbol{C}'\boldsymbol{C}))^{-1}\boldsymbol{T}_p\boldsymbol{C}'\boldsymbol{\Lambda}_{\mathbb{I}_r}\boldsymbol{C}\boldsymbol{T}_p(\boldsymbol{S}_p \circ (\boldsymbol{C}'\boldsymbol{C}))^{-1}\right)\right)\boldsymbol{C}' &= \boldsymbol{\Lambda}_{\mathbb{I}_r}\boldsymbol{\Delta}_{\mathbb{I}_r}. \qquad (25)
\end{aligned}
$$

Now we have to find $\boldsymbol{C}$ such that it satisfies eq. (24) and eq. (25). To make the process easier to follow, lets have them in one place. The matrix $\boldsymbol{C} \in \mathbb{R}^{r\times p}$ have to satisfy

$$
\begin{aligned}
\boldsymbol{C}\boldsymbol{T}_p\left(\boldsymbol{S}_p \circ (\boldsymbol{C}'\boldsymbol{C})\right)^{-1}\boldsymbol{T}_p\boldsymbol{C}' &=\boldsymbol{\Delta}_{\mathbb{I}_r} \text{ and} \qquad (26) \\
\boldsymbol{C}\left(\boldsymbol{S}_p \circ \left((\boldsymbol{S}_p \circ (\boldsymbol{C}'\boldsymbol{C}))^{-1}\boldsymbol{T}_p\boldsymbol{C}'\boldsymbol{\Lambda}_{\mathbb{I}_r}\boldsymbol{C}\boldsymbol{T}_p(\boldsymbol{S}_p \circ (\boldsymbol{C}'\boldsymbol{C}))^{-1}\right)\right)\boldsymbol{C}' &=\boldsymbol{\Lambda}_{\mathbb{I}_r}\boldsymbol{\Delta}_{\mathbb{I}_r}. \qquad (27)
\end{aligned}
$$

Since $\boldsymbol{C}$ is a rectangular matrix, solving above equations for $\boldsymbol{C}$ in this form seems intractable. We use a trick to temporarily extend $\boldsymbol{C}$ into an invertible square matrix as follows.

- Temporarily, let $\boldsymbol{M}_1 = \boldsymbol{T}_p\left(\boldsymbol{S}_p \circ (\boldsymbol{C}'\boldsymbol{C})\right)^{-1}\boldsymbol{T}_p$, and $\boldsymbol{M}_2 = \boldsymbol{S}_p \circ \left((\boldsymbol{S}_p \circ (\boldsymbol{C}'\boldsymbol{C}))^{-1}\boldsymbol{T}_p\boldsymbol{C}'\boldsymbol{\Lambda}_{\mathbb{I}_r}\boldsymbol{C}\boldsymbol{T}_p(\boldsymbol{S}_p \circ (\boldsymbol{C}'\boldsymbol{C}))^{-1}\right)$. Then $\boldsymbol{M}_1$ is positive definite and $\boldsymbol{M}_2$ is positive semidefinite, so they are simultaneously diagonalizable by congruence that is based on Lemma 2 and eq. (26) and eq. (27), there exists a nonsingular $\bar{\boldsymbol{C}} \in \mathbb{R}^{p\times p}$ such that $\boldsymbol{C}$ consists of the first r rows of $\bar{\boldsymbol{C}}$ and

$$
\begin{aligned}
\bar{\boldsymbol{C}}\boldsymbol{T}_p\left(\boldsymbol{S}_p \circ (\boldsymbol{C}'\boldsymbol{C})\right)^{-1}\boldsymbol{T}_p\bar{\boldsymbol{C}}' &=\bar{\boldsymbol{\Delta}}_{\mathbb{I}_r}, \qquad (28) \\
\bar{\boldsymbol{C}}\left(\boldsymbol{S}_p \circ \left((\boldsymbol{S}_p \circ (\boldsymbol{C}'\boldsymbol{C}))^{-1}\boldsymbol{T}_p\boldsymbol{C}'\boldsymbol{\Lambda}_{\mathbb{I}_r}\boldsymbol{C}\boldsymbol{T}_p\left(\boldsymbol{S}_p \circ (\boldsymbol{C}'\boldsymbol{C})\right)^{-1}\right)\right)\bar{\boldsymbol{C}}' &=\bar{\boldsymbol{\Lambda}}_{\mathbb{I}_r}\bar{\boldsymbol{\Delta}}_{\mathbb{I}_r}, \qquad (29)
\end{aligned}
$$

  where, $\bar{\boldsymbol{\Delta}}_{\mathbb{I}_r} = \boldsymbol{\Delta}_{\mathbb{I}_r} \oplus I_{r-p}$ is a $p\times p$ diagonal matrix and $\bar{\boldsymbol{\Lambda}}_{\mathbb{I}_r} = \boldsymbol{\Lambda}_{\mathbb{I}_r} \oplus \underline{\boldsymbol{\Lambda}}$ is another $p\times p$ diagonal matrix, in which $\underline{\boldsymbol{\Lambda}} \in \mathbb{R}^{r-p\times r-p}$ is a nonnegative diagonal matrix.

- Substitute $\bar{\boldsymbol{\Delta}}_{\mathbb{I}_r}$ from eq. (28) in eq. (29), then left multiply by $\bar{\boldsymbol{C}}'^{-1}$, and right multiply by $\bar{\boldsymbol{C}}'I_{r;p}$:

$$
\bar{\boldsymbol{C}}\left(\boldsymbol{S}_p \circ \left((\boldsymbol{S}_p \circ (\boldsymbol{C}'\boldsymbol{C}))^{-1}\boldsymbol{T}_p\boldsymbol{C}'\boldsymbol{\Lambda}_{\mathbb{I}_r}\boldsymbol{C}\boldsymbol{T}_p\left(\boldsymbol{S}_p \circ (\boldsymbol{C}'\boldsymbol{C})\right)^{-1}\right)\right)\bar{\boldsymbol{C}}' =
$$

$$\bar{\boldsymbol{\Lambda}}_{\mathbb{I}_r}\bar{\boldsymbol{C}}\boldsymbol{T}_p\left(\boldsymbol{S}_p\circ(\boldsymbol{C}'\boldsymbol{C})\right)^{-1}\boldsymbol{T}_p\bar{\boldsymbol{C}}' \xrightarrow[\times\bar{\boldsymbol{C}}'^{-1}]{\bar{\boldsymbol{C}}'I_{r;p}\times}$$

$$\bar{\boldsymbol{C}}'I_{r;p}\bar{\boldsymbol{C}}\left(\boldsymbol{S}_p\circ\left(\left(\boldsymbol{S}_p\circ(\boldsymbol{C}'\boldsymbol{C})\right)^{-1}\boldsymbol{T}_p\boldsymbol{C}'\boldsymbol{\Lambda}_{\mathbb{I}_r}\boldsymbol{C}\boldsymbol{T}_p\left(\boldsymbol{S}_p\circ(\boldsymbol{C}'\boldsymbol{C})\right)^{-1}\right)\right)=$$
$$\bar{\boldsymbol{C}}'I_{r;p}\bar{\boldsymbol{\Lambda}}_{\mathbb{I}_r}\bar{\boldsymbol{C}}\boldsymbol{T}_p\left(\boldsymbol{S}_p\circ(\boldsymbol{C}'\boldsymbol{C})\right)^{-1}\boldsymbol{T}_p.$$

- Now we can revert back everything to $\boldsymbol{C}$ again. Since $\boldsymbol{C}$ consists of the first r rows of $\bar{\boldsymbol{C}}$ we have $\bar{\boldsymbol{C}}'I_{r;p}\bar{\boldsymbol{C}}=\boldsymbol{C}'\boldsymbol{C}$, and $\bar{\boldsymbol{C}}'I_{r;p}\bar{\boldsymbol{\Lambda}}_{\mathbb{I}_r}\bar{\boldsymbol{C}}=\boldsymbol{C}'\boldsymbol{\Lambda}_{\mathbb{I}_r}\boldsymbol{C}$, which turns the above equation into

$$\boldsymbol{C}'\boldsymbol{C}\left(\boldsymbol{S}_p\circ\left(\boldsymbol{I}_p\left(\boldsymbol{S}_p\circ(\boldsymbol{C}'\boldsymbol{C})\right)^{-1}\boldsymbol{T}_p\boldsymbol{C}'\boldsymbol{\Lambda}_{\mathbb{I}_r}\boldsymbol{C}\boldsymbol{T}_p\left(\boldsymbol{S}_p\circ(\boldsymbol{C}'\boldsymbol{C})\right)^{-1}\boldsymbol{I}_p\right)\right)=$$
$$\boldsymbol{I}_p\boldsymbol{C}'\boldsymbol{\Lambda}_{\mathbb{I}_r}\boldsymbol{C}\boldsymbol{T}_p\left(\boldsymbol{S}_p\circ(\boldsymbol{C}'\boldsymbol{C})\right)^{-1}\boldsymbol{T}_p.$$

- In the above equation, replace $\boldsymbol{I}_p$ by $\boldsymbol{T}_p^{-1}\boldsymbol{T}_p$ in LHS and by $\boldsymbol{T}_p^{-1}\left(\boldsymbol{S}_p\circ(\boldsymbol{C}'\boldsymbol{C})\right)\boldsymbol{T}_p^{-1}\boldsymbol{T}_p\left(\boldsymbol{S}_p\circ(\boldsymbol{C}'\boldsymbol{C})\right)^{-1}\boldsymbol{T}_p$ in the RHS. Use $\boldsymbol{\Pi_C}\coloneqq\left(\boldsymbol{S}_p\circ(\boldsymbol{C}'\boldsymbol{C})\right)^{-1}\boldsymbol{T}_p\boldsymbol{C}'$ to shrink it into :

$$\boldsymbol{C}'\boldsymbol{C}\left(\boldsymbol{S}_p\circ\left(\boldsymbol{T}_p^{-1}\boldsymbol{T}_p\boldsymbol{\Pi_C}\boldsymbol{\Lambda}_{\mathbb{I}_r}\boldsymbol{\Pi}_{\boldsymbol{C}}'\boldsymbol{T}_p\boldsymbol{T}_p^{-1}\right)\right)=\boldsymbol{T}_p^{-1}\left(\boldsymbol{S}_p\circ(\boldsymbol{C}'\boldsymbol{C})\right)\boldsymbol{T}_p^{-1}\boldsymbol{T}_p\boldsymbol{\Pi_C}\boldsymbol{\Lambda}_{\mathbb{I}_r}\boldsymbol{\Pi}_{\boldsymbol{C}}'\boldsymbol{T}_p.$$

- By the second property of Lemma 1 we can collect diagonal matrices $\boldsymbol{T}_p^{-1}$'s around $\boldsymbol{S}_p$ to arrive at

$$(\boldsymbol{C}'\boldsymbol{C})\left(\hat{\boldsymbol{S}}_p\circ\left(\boldsymbol{T}_p\boldsymbol{\Pi_C}\boldsymbol{\Lambda}_{\mathbb{I}_r}\boldsymbol{\Pi}_{\boldsymbol{C}}'\boldsymbol{T}_p\right)\right)=\left(\hat{\boldsymbol{S}}_p\circ(\boldsymbol{C}'\boldsymbol{C})\right)\left(\boldsymbol{T}_p\boldsymbol{\Pi_C}\boldsymbol{\Lambda}_{\mathbb{I}_r}\boldsymbol{\Pi}_{\boldsymbol{C}}'\boldsymbol{T}_p\right),$$

  where, $\hat{\boldsymbol{S}}_p\coloneqq\boldsymbol{T}_p^{-1}\boldsymbol{S}_p\boldsymbol{T}_p^{-1}$.

- Define $p\times p$ matrices $\mathscr{E}_r\coloneqq\boldsymbol{C}'\boldsymbol{C}$ and $\mathscr{D}_r\coloneqq\boldsymbol{T}_p\boldsymbol{\Pi_C}\boldsymbol{\Lambda}_{\mathbb{I}_r}\boldsymbol{\Pi}_{\boldsymbol{C}}'\boldsymbol{T}_p$. Substitute in the above to arrive at:

$$\mathscr{E}_r\left(\hat{\boldsymbol{S}}_p\circ\mathscr{D}_r\right)=\left(\hat{\boldsymbol{S}}_p\circ\mathscr{E}_r\right)\mathscr{D}_r.$$

  Both $\mathscr{D}_r$ and $\mathscr{E}_r$ in the above identity are positive semidefinite. Moreover, since by assumption $\boldsymbol{C}$ has no zero columns, $\mathscr{E}_r$ has no zero diagonal element. Then the 7^{th} property of Lemma 1 implies the following two conclusions:

  1. The matrix $\mathscr{D}_r$ is diagonal. The rank of $\mathscr{D}_r$ is r so it has exactly r positive diagonal elements and the rest is zero. This argument is true for $\boldsymbol{T}_p^{-1}\mathscr{D}_r\boldsymbol{T}_p^{-1}=\boldsymbol{\Pi_C}\boldsymbol{\Lambda}_{\mathbb{I}_r}\boldsymbol{\Pi}_{\boldsymbol{C}}'$. Since $\boldsymbol{\Lambda}_{\mathbb{I}_r}$ is a diagonal positive definite matrix, the $p\times r$ matrix $\boldsymbol{\Pi_C}\coloneqq\left(\boldsymbol{S}_p\circ(\boldsymbol{C}'\boldsymbol{C})\right)^{-1}\boldsymbol{T}_p\boldsymbol{C}'$ of rank r should have $p-r$ zero rows. Let $\mathbb{J}_r$ be an $r-$index set corresponding to nonzero diagonal elements of $\boldsymbol{\Pi_C}\boldsymbol{\Lambda}_{\mathbb{I}_r}\boldsymbol{\Pi}_{\boldsymbol{C}}'$. Then the matrix $\boldsymbol{\Pi_C}[\mathbb{J}_r,\mathbb{N}_r]$ ($r\times r$ submatrix of $\boldsymbol{\Pi_C}$ consist of its $\mathbb{J}_r$ rows) is nonsingular.
  2. For every $i,j\in\mathbb{J}_r$ and $i\neq j$, $(\mathscr{E}_r)_{i,j}=0$. Since $\mathscr{E}_r\coloneqq\boldsymbol{C}'\boldsymbol{C}$ and so $(\mathscr{E}_r)_{i,j}$ is the inner product of i^{th} and j^{th} columns of $\boldsymbol{C}$, we conclude that the columns of $\boldsymbol{C}[\mathbb{N}_r,\mathbb{J}_r]$ ($r\times r$ submatrix of $\boldsymbol{C}$ consist of its $\mathbb{J}_r$ columns) are orthogonal or in other words $\boldsymbol{C}[\mathbb{N}_r,\mathbb{J}_r]'\boldsymbol{C}[\mathbb{N}_r,\mathbb{J}_r]$ is diagonal. The columns of $\boldsymbol{C}$ are normalized. Therefore, $\boldsymbol{C}[\mathbb{N}_r,\mathbb{J}_r]'\boldsymbol{C}[\mathbb{N}_r,\mathbb{J}_r]=\boldsymbol{I}_r$ and hence, $\boldsymbol{C}[\mathbb{N}_r,\mathbb{J}_r]$ is an orthogonal matrix.

- We use the two conclusions to solve the original eq. (26) and eq. (27). First use $\boldsymbol{\Pi_C}\coloneqq\left(\boldsymbol{S}_p\circ(\boldsymbol{C}'\boldsymbol{C})\right)^{-1}\boldsymbol{T}_p\boldsymbol{C}'$ to shrink them into :

$$\boldsymbol{C}\boldsymbol{T}_p\boldsymbol{\Pi_C}=\boldsymbol{\Delta}_{\mathbb{I}_r},\tag{30}$$

$$\boldsymbol{C}\left(\boldsymbol{S}_p \circ (\boldsymbol{\Pi_C}\boldsymbol{\Lambda}_{\mathbb{I}_r}\boldsymbol{\Pi}'_{\boldsymbol{C}})\right)\boldsymbol{C}' = \boldsymbol{\Lambda}_{\mathbb{I}_r}\boldsymbol{\Delta}_{\mathbb{I}_r}. \tag{31}$$

Next, by the first conclusion, the matrix $\boldsymbol{T}_p^{-1}\mathscr{D}_r\boldsymbol{T}_p^{-1} = \boldsymbol{\Pi_C}\boldsymbol{\Lambda}_{\mathbb{I}_r}\boldsymbol{\Pi}'_{\boldsymbol{C}}$ is diagonal and so eq. (31) becomes

$$\begin{aligned} \underbrace{\boldsymbol{C}\boldsymbol{T}_p\boldsymbol{\Pi_C}}\boldsymbol{\Lambda}_{\mathbb{I}_r}\boldsymbol{\Pi}'_{\boldsymbol{C}}\boldsymbol{C}' &= \boldsymbol{\Lambda}_{\mathbb{I}_r}\boldsymbol{\Delta}_{\mathbb{I}_r} \xRightarrow{\text{eq. (30)}} \\ \boldsymbol{\Delta}_{\mathbb{I}_r}\boldsymbol{\Lambda}_{\mathbb{I}_r}\boldsymbol{\Pi}'_{\boldsymbol{C}}\boldsymbol{C}' &= \boldsymbol{\Lambda}_{\mathbb{I}_r}\boldsymbol{\Delta}_{\mathbb{I}_r} \implies \\ \boldsymbol{\Pi}'_{\boldsymbol{C}}\boldsymbol{C}' = \boldsymbol{C}\boldsymbol{\Pi_C} &= \boldsymbol{I}_r, \end{aligned} \tag{32}$$

which is one of the two claimed conditions. What is left is to show that $\boldsymbol{\Pi_C}$ is a rectangular permutation matrix. From the first conclusion we also have $\boldsymbol{\Pi_C}$ has exactly r nonzero columns indexed by $\mathbb{J}_r$ so

$$\boldsymbol{C}[\mathbb{N}_r, \mathbb{J}_r]\boldsymbol{\Pi_C}[\mathbb{J}_r, \mathbb{N}_r] = \boldsymbol{I}_r.$$

By the second conclusion $\boldsymbol{C}[\mathbb{N}_r, \mathbb{J}_r]$ is an orthogonal matrix therefore, $\boldsymbol{\Pi_C}[\mathbb{J}_r, \mathbb{N}_r]$ is the orthogonal matrix $\boldsymbol{C}[\mathbb{N}_r, \mathbb{J}_r]'$. Moreover, we had $\boldsymbol{T}_p^{-1}\mathscr{D}_r\boldsymbol{T}_p^{-1} = \boldsymbol{\Pi_C}\boldsymbol{\Lambda}_{\mathbb{I}_r}\boldsymbol{\Pi}'_{\boldsymbol{C}}$ is a $p\times p$ diagonal matrix with exactly r nonzero diagonal elements. Hence, $\boldsymbol{\Pi_C}[\mathbb{N}_r, \mathbb{J}_r]\boldsymbol{\Lambda}_{\mathbb{I}_r}\boldsymbol{\Pi}'_{\boldsymbol{C}}[\mathbb{N}_r, \mathbb{J}_r]$ is an $r\times r$ positive definite diagonal matrix with $\boldsymbol{\Lambda}_{\mathbb{I}_r}$ having distinct diagonal elements, and $\boldsymbol{\Pi_C}[\mathbb{N}_r, \mathbb{J}_r]$ being orthogonal. Therefore, $\boldsymbol{\Pi_C}[\mathbb{J}_r, \mathbb{N}_r]$ (as well as $\boldsymbol{C}[\mathbb{N}_r, \mathbb{J}_r]$) should be a square permutation matrix. Putting back the zero columns, we conclude that $\boldsymbol{C}$ should be such that $\boldsymbol{\Pi_C} := (\boldsymbol{S}_p \circ (\boldsymbol{C}'\boldsymbol{C}))^{-1}\boldsymbol{T}_p\boldsymbol{C}'$ is a rectangular permutation matrix and $\boldsymbol{C}\boldsymbol{\Pi_C} = \boldsymbol{I}_r$. Note that it is possible to further analyze these conditions and determine the exact structure of $\boldsymbol{C}$. However, this is not needed in general for the critical point analysis of the next theorem except for the case where $r = p$ and $\boldsymbol{C}$ is a square invertible matrix. In this case, square matrix $\boldsymbol{\Pi_C}$ is of full rank p, $\mathbb{J}_r = \mathbb{N}_p$ and therefore, $\boldsymbol{C}[\mathbb{N}_r, \mathbb{J}_r] = \boldsymbol{C}[\mathbb{N}_p, \mathbb{N}_p] = \boldsymbol{C}$. Hence, $\boldsymbol{C}$ is any square permutation matrix $\boldsymbol{\Pi}$, $\boldsymbol{C}'\boldsymbol{C} = \boldsymbol{\Pi}'\boldsymbol{\Pi} = \boldsymbol{I}_p$ and $\boldsymbol{\Pi_C} := (\boldsymbol{S}_p \circ (\boldsymbol{C}'\boldsymbol{C}))^{-1}\boldsymbol{T}_p\boldsymbol{C}' = \boldsymbol{T}_p^{-1}\boldsymbol{T}_p\boldsymbol{\Pi}' = \boldsymbol{\Pi}'$, which verifies eq. (17) and eq. (18) for $\boldsymbol{A}$ and $\boldsymbol{B}$ when $\boldsymbol{A}$ is of full rank p.

□

Corollary 1. *Let $(\boldsymbol{A}, \boldsymbol{B})$ be a critical point of $L(\boldsymbol{A}, \boldsymbol{B})$ under the given assumptions and rank$\boldsymbol{A} = r \leq p$. Then the following hold:*

1. *The matrix $\boldsymbol{B}\boldsymbol{\Sigma}_{xx}\boldsymbol{B}'$ is a $p \times p$ diagonal matrix of rank r.*
2. *For all $1 \leq r \leq p$, for any critical pair $(\boldsymbol{A}, \boldsymbol{B})$, the global map $\boldsymbol{G} := \boldsymbol{A}\boldsymbol{B}$ becomes*

$$\boldsymbol{G} = \boldsymbol{U}_{\mathbb{I}_r}\boldsymbol{U}'_{\mathbb{I}_r}\boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}. \tag{33}$$

   For the autoencoder case ($\boldsymbol{Y} = \boldsymbol{X}$) the global map is simply $\boldsymbol{G} = \boldsymbol{U}_{\mathbb{I}_r}\boldsymbol{U}'_{\mathbb{I}_r}$.

3. *$(\boldsymbol{A}, \boldsymbol{B})$ is also the critical point of the classical loss $\tilde{L}(\boldsymbol{A}, \boldsymbol{B}) = \sum_{i=1}^{p} \|\boldsymbol{Y} - \boldsymbol{A}\boldsymbol{B}\boldsymbol{X}\|_F^2$.*

Proof. 1. We already show in the proof Theorem 1 that for critical $(\boldsymbol{A}, \boldsymbol{B})$ the matrix $\boldsymbol{B}\boldsymbol{\Sigma}_{xx}\boldsymbol{B}'$ is given by eq. (21) that is

$$\boldsymbol{B}\boldsymbol{\Sigma}_{xx}\boldsymbol{B}' = \boldsymbol{D}^{-1}\boldsymbol{\Pi_C}\boldsymbol{\Lambda}_{\mathbb{I}_r}\boldsymbol{\Pi}'_{\boldsymbol{C}}\boldsymbol{D}^{-1}.$$

The matrix $\boldsymbol{\Pi_C}$ is a $p \times r$ rectangular permutation matrix so $\boldsymbol{\Pi_C}\boldsymbol{\Lambda}_{\mathbb{I}_r}\boldsymbol{\Pi}'_{\boldsymbol{C}}$ is diagonal as well as $\boldsymbol{D}^{-1}\boldsymbol{\Pi_C}\boldsymbol{\Lambda}_{\mathbb{I}_r}\boldsymbol{\Pi}'_{\boldsymbol{C}}\boldsymbol{D}^{-1}$. Therefore, $\boldsymbol{B}\boldsymbol{\Sigma}_{xx}\boldsymbol{B}'$ is diagonal. The diagonal matrix $\boldsymbol{\Lambda}_{\mathbb{I}_r}$ is of rank r therefore, $\boldsymbol{B}\boldsymbol{\Sigma}_{xx}\boldsymbol{B}'$ is of rank r.

2. Again by Theorem 1 critical $(\boldsymbol{A}, \boldsymbol{B})$ is of the form given by eq. (15) and eq. (16) with the proceeding conditions on the invariance $\boldsymbol{C}$. Therefore, the global map is

$$\begin{aligned} \boldsymbol{G} = \boldsymbol{A}\boldsymbol{B} &= \boldsymbol{U}_{\mathbb{I}_r}\boldsymbol{C}\boldsymbol{D}\boldsymbol{D}^{-1}\boldsymbol{\Pi}_{\boldsymbol{C}}\boldsymbol{U}'_{\mathbb{I}_r}\boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1} \\ &= \boldsymbol{U}_{\mathbb{I}_r}\boldsymbol{C}\boldsymbol{\Pi}_{\boldsymbol{C}}\boldsymbol{U}'_{\mathbb{I}_r}\boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1} \xRightarrow{\boldsymbol{C}\boldsymbol{\Pi}_{\boldsymbol{C}}=\boldsymbol{I}_r} \\ \boldsymbol{G} &= \boldsymbol{U}_{\mathbb{I}_r}\boldsymbol{U}'_{\mathbb{I}_r}\boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}. \end{aligned}$$

3. Based on Baldi & Hornik (1989) $(\boldsymbol{A}, \boldsymbol{B})$ define a critical point of $\tilde{L}(\boldsymbol{A}, \boldsymbol{B}) = \sum_{i=1}^{p} \|\boldsymbol{Y} - \boldsymbol{A}\boldsymbol{B}\boldsymbol{X}\|_F^2$ iff they satisfy

$$\begin{aligned} \boldsymbol{A}'\boldsymbol{A}\boldsymbol{B}\boldsymbol{\Sigma}_{xx} &= \boldsymbol{A}'\boldsymbol{\Sigma}_{yx} \text{ and} && (34) \\ \boldsymbol{A}\boldsymbol{B}\boldsymbol{\Sigma}_{xx}\boldsymbol{B}' &= \boldsymbol{\Sigma}_{yx}\boldsymbol{B}'. && (35) \end{aligned}$$

Again by assumption $(\boldsymbol{A}, \boldsymbol{B})$ define a critical point of $L(\boldsymbol{A}, \boldsymbol{B})$ so by Theorem 1 they are of the form given by eq. (15) and eq. (16) with the proceeding conditions on the invariance $\boldsymbol{C}$. Hence,

$$\begin{aligned} \boldsymbol{A}'\boldsymbol{A}\boldsymbol{B}\boldsymbol{\Sigma}_{xx} &= \boldsymbol{D}\boldsymbol{C}'\underbrace{\boldsymbol{U}'_{\mathbb{I}_r}\boldsymbol{U}_{\mathbb{I}_r}}\boldsymbol{C}\underbrace{\boldsymbol{D}\boldsymbol{D}^{-1}}\boldsymbol{\Pi}_{\boldsymbol{C}}\boldsymbol{U}'_{\mathbb{I}_r}\boldsymbol{\Sigma}_{yx}\underbrace{\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xx}} \\ &= \boldsymbol{D}\boldsymbol{C}'\underbrace{\boldsymbol{C}\boldsymbol{\Pi}_{\boldsymbol{C}}}\boldsymbol{U}'_{\mathbb{I}_r}\boldsymbol{\Sigma}_{yx} \xRightarrow{\boldsymbol{C}\boldsymbol{\Pi}_{\boldsymbol{C}}=\boldsymbol{I}_r} \\ \boldsymbol{A}'\boldsymbol{A}\boldsymbol{B}\boldsymbol{\Sigma}_{xx} &= \boldsymbol{D}\boldsymbol{C}'\boldsymbol{U}'_{\mathbb{I}_r}\boldsymbol{\Sigma}_{yx} = \boldsymbol{A}'\boldsymbol{\Sigma}_{yx}. \end{aligned}$$

Hence, eq. (34) is satisfied. For the second equation we use the first property of this corollary that is $\boldsymbol{B}\boldsymbol{\Sigma}_{xx}\boldsymbol{B}'$ is diagonal and satisfy eq. (14) of Proposition 2 that is

$$\begin{aligned} \boldsymbol{A}\left(\boldsymbol{S}_p \circ (\boldsymbol{B}\boldsymbol{\Sigma}_{xx}\boldsymbol{B}')\right) &= \boldsymbol{\Sigma}_{yx}\boldsymbol{B}'\boldsymbol{T}_p \xRightarrow{\boldsymbol{B}\boldsymbol{\Sigma}_{xx}\boldsymbol{B}' \text{ is diagonal}} \\ \boldsymbol{A}\boldsymbol{T}_p\boldsymbol{B}\boldsymbol{\Sigma}_{xx}\boldsymbol{B}' &= \boldsymbol{\Sigma}_{yx}\boldsymbol{B}'\boldsymbol{T}_p \xRightarrow{\boldsymbol{B}\boldsymbol{\Sigma}_{xx}\boldsymbol{B}' \text{ is diagonal}} \\ \boldsymbol{A}\boldsymbol{B}\boldsymbol{\Sigma}_{xx}\boldsymbol{B}'\boldsymbol{T}_p &= \boldsymbol{\Sigma}_{yx}\boldsymbol{B}'\boldsymbol{T}_p \implies \\ \boldsymbol{A}\boldsymbol{B}\boldsymbol{\Sigma}_{xx}\boldsymbol{B}' &= \boldsymbol{\Sigma}_{yx}\boldsymbol{B}'. \end{aligned}$$

Hence, the second condition, eq. (35) is also satisfied. Therefore, any critical point of $L(\boldsymbol{A}, \boldsymbol{B})$ is a critical point of $\tilde{L}(\boldsymbol{A}, \boldsymbol{B})$.

□

Lemma 4. *The loss function $L(\boldsymbol{A}, \boldsymbol{B})$ can be written as*

$$\begin{aligned} L(\boldsymbol{A}, \boldsymbol{B}) = p\,\mathrm{Tr}(\boldsymbol{\Sigma}_{yy}) &- 2\,\mathrm{Tr}\left(\boldsymbol{A}\boldsymbol{T}_p\boldsymbol{B}\boldsymbol{\Sigma}_{xy}\right) \\ &+ \mathrm{Tr}\left(\boldsymbol{B}'\left(\boldsymbol{S}_p \circ (\boldsymbol{A}'\boldsymbol{A})\right)\boldsymbol{B}\boldsymbol{\Sigma}_{xx}\right). \end{aligned} \tag{36}$$

Proof. We have

$$\begin{aligned} L(\boldsymbol{A}, \boldsymbol{B}) &= \sum_{i=1}^{p} \|\boldsymbol{Y} - \boldsymbol{A}I_{i;p}\boldsymbol{B}\boldsymbol{X}\|_F^2 = \sum_{i=1}^{p} \langle \boldsymbol{Y} - \boldsymbol{A}I_{i;p}\boldsymbol{B}\boldsymbol{X}, \boldsymbol{Y} - \boldsymbol{A}I_{i;p}\boldsymbol{B}\boldsymbol{X}\rangle_F \\ &= \sum_{i=1}^{p} \left(\langle \boldsymbol{Y}, \boldsymbol{Y}\rangle_F + \langle \boldsymbol{Y}, -\boldsymbol{A}I_{i;p}\boldsymbol{B}\boldsymbol{X}\rangle_F + \langle -\boldsymbol{A}I_{i;p}\boldsymbol{B}\boldsymbol{X}, \boldsymbol{Y}\rangle_F\right. \end{aligned}$$

$$
\begin{aligned}
&+\langle -\boldsymbol{A}I_{i;p}\boldsymbol{B}\boldsymbol{X}, -\boldsymbol{A}I_{i;p}\boldsymbol{B}\boldsymbol{X}\rangle_F) \\
&= p\langle \boldsymbol{Y}, \boldsymbol{Y}\rangle_F - 2\langle \boldsymbol{Y}, \boldsymbol{A}\left(\sum_{i=1}^{p} I_{i;p}\right)\boldsymbol{B}\boldsymbol{X}\rangle_F + \sum_{i=1}^{p}\langle \boldsymbol{A}I_{i;p}\boldsymbol{B}\boldsymbol{X}, \boldsymbol{A}I_{i;p}\boldsymbol{B}\boldsymbol{X}\rangle_F \xRightarrow{\text{eq. (3)}} \\
&= p\,\mathrm{Tr}(\boldsymbol{Y}\boldsymbol{Y}') - 2\,\mathrm{Tr}\left(\boldsymbol{A}\boldsymbol{T}_p\boldsymbol{B}\boldsymbol{X}\boldsymbol{Y}'\right) + \sum_{i=1}^{p}\mathrm{Tr}\left(\boldsymbol{X}'\boldsymbol{B}'I_{i;p}\boldsymbol{A}'\boldsymbol{A}I_{i;p}\boldsymbol{B}\boldsymbol{X}\right) \\
&= p\,\mathrm{Tr}(\boldsymbol{\Sigma}_{yy}) - 2\,\mathrm{Tr}\left(\boldsymbol{A}\boldsymbol{T}_p\boldsymbol{B}\boldsymbol{\Sigma}_{xy}\right) + \mathrm{Tr}\left(\boldsymbol{X}\boldsymbol{X}'\boldsymbol{B}'\sum_{i=1}^{p}\left(I_{i;p}\boldsymbol{A}'\boldsymbol{A}I_{i;p}\right)\boldsymbol{B}\right) \xRightarrow{\text{eq. (4)}} \\
&= p\,\mathrm{Tr}(\boldsymbol{\Sigma}_{yy}) - 2\,\mathrm{Tr}\left(\boldsymbol{A}\boldsymbol{T}_p\boldsymbol{B}\boldsymbol{\Sigma}_{xy}\right) + \mathrm{Tr}\left(\boldsymbol{B}'\left(\boldsymbol{S}_p \circ (\boldsymbol{A}'\boldsymbol{A})\right)\boldsymbol{B}\boldsymbol{\Sigma}_{xx}\right),
\end{aligned}
$$

which is eq. (36). □

Theorem 2. *Let $\boldsymbol{A}^* \in \mathbb{R}^{n\times p}$ and $\boldsymbol{B}^* \in \mathbb{R}^{p\times n}$ such that $\boldsymbol{A}^*$ is of rank $r \le p$. Under the given assumptions, $(\boldsymbol{A}^*, \boldsymbol{B}^*)$ define a local minima of the proposed loss function iff they are of the form*

$$\boldsymbol{A}^* = \boldsymbol{U}_{1:p}\boldsymbol{D}_p, \tag{37}$$

$$\boldsymbol{B}^* = \boldsymbol{D}_p^{-1}\boldsymbol{U}'_{1:p}\boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}, \tag{38}$$

where the i^{th} column of $\boldsymbol{U}_{1:p}$ is a unit eigenvector of $\boldsymbol{\Sigma} := \boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}$ corresponding the i^{th} largest eigenvalue and $\boldsymbol{D}_p$ is a diagonal matrix with nonzero diagonal elements. In other words, $\boldsymbol{A}^*$ contains ordered unnormalized eigenvectors of $\boldsymbol{\Sigma}$ corresponding to the p largest eigenvalues. Moreover, all the local minima are global minima with the value of the loss function at those global minima being

$$L(\boldsymbol{A}^*, \boldsymbol{B}^*) = p\ \mathrm{Tr}(\boldsymbol{\Sigma}_{yy}) - \sum_{i=1}^{p}(p-i+1)\,\lambda_i, \tag{39}$$

where λ_i is the i^{th} largest eigenvalue of $\boldsymbol{\Sigma}$.

Proof. The full rank matrices $\boldsymbol{A}^*$ and $\boldsymbol{B}^*$ given by eq. (37) and eq. (38) are clearly of the form given by Theorem 1 with $\mathbb{I}_p = \mathbb{N}_p := \{1, 2, \cdots, p\}$, and $\boldsymbol{\Pi}_p = \boldsymbol{I}_p$. Hence, they define a critical point of $L(\boldsymbol{A}, \boldsymbol{B})$. We want to show that these are the only local minima, that is any other critical $(\boldsymbol{A}, \boldsymbol{B})$ is a saddle points. The proof is similar to the second partial derivative test. However, in this case the Hessian is a forth order tensor. Therefore, the second order Taylor approximation of the loss, derived in Lemma 3, is used directly. To prove the necessary condition, we show that at any other critical point $(\boldsymbol{A}, \boldsymbol{B})$, where the first order derivatives are zero, there exists infinitesimal direction along which the second derivative of loss is negative. Next, for the sufficient condition we show that the any critical point of the form $(\boldsymbol{A}^*, \boldsymbol{B}^*)$ is a local and global minima.

The necessary condition:

Recall that $\boldsymbol{U}_{\mathbb{I}_p}$ is the matrix of eigenvectors indexed by the $p-$index set $\mathbb{I}_p$ and $\boldsymbol{\Pi}$ is a $p \times p$ permutation matrix. Since all the index sets $\mathbb{I}_r$, $r \le p$ are assumed to be ordered, the only way to have $\boldsymbol{U}_{\mathbb{N}_p} = \boldsymbol{U}_{\mathbb{I}_p}\boldsymbol{\Pi}$ is by having $\mathbb{I}_p = \mathbb{N}_p$ and $\boldsymbol{\Pi} = \boldsymbol{I}_p$. Let $\boldsymbol{A}$ (with no zero column) and $\boldsymbol{B}$ define an arbitrary critical point of $L(\boldsymbol{A}, \boldsymbol{B})$. Then Based on the previous theorem, either $\boldsymbol{A} = \boldsymbol{U}_{\mathbb{I}_r}\boldsymbol{C}$ with $r < p$ or $\boldsymbol{A} = \boldsymbol{U}_{\mathbb{I}_p}\boldsymbol{\Pi}\boldsymbol{D}$ while in both cases $\boldsymbol{B} = \hat{\boldsymbol{B}}(\boldsymbol{A})$ given by eq. (13). If $(\boldsymbol{A}, \boldsymbol{B})$ is not of the form of $(\boldsymbol{A}^*, \boldsymbol{B}^*)$ then there are three possibilities either 1) $\boldsymbol{A} = \boldsymbol{U}_{\mathbb{I}_r}\boldsymbol{C}\boldsymbol{D}$ with $r < p$, or 2) $\boldsymbol{A} = \boldsymbol{U}_{\mathbb{I}_p}\boldsymbol{\Pi}\boldsymbol{D}$ with $\mathbb{I}_p \ne \mathbb{N}_p$ or 2) $\boldsymbol{A} = \boldsymbol{U}_{\mathbb{N}_p}\boldsymbol{\Pi}\boldsymbol{D}$ but $\boldsymbol{\Pi} \ne \boldsymbol{I}_p$. The first two cases corresponds to not having the "right" and/or "enough" eigenvectors, and the third corresponds to not having the "right" ordering. We introduce the following notation and investigate each case separately.

Let $\varepsilon > 0$ and $\boldsymbol{U}_{i;j} \in \mathbb{R}^{n\times p}$ be a matrix of all zeros except the i^{th} column, which contains $\boldsymbol{u}_j$; the eigenvector of $\boldsymbol{\Sigma}$ corresponding to the j^{th} largest eigenvalue. Therefore,

$$\boldsymbol{U}'_{i;j}\boldsymbol{\Sigma}\boldsymbol{U}_{i;j} = \boldsymbol{U}'_{i;j}\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}'\boldsymbol{U}_{i;j} = \lambda_j \boldsymbol{E}_i, \tag{40}$$

where, $\boldsymbol{E}_i \in \mathbb{R}^{p\times p}$ is matrix of zeros except the i^{th} diagonal element that contains 1. In what follows, for each case we define a encoder direction $\boldsymbol{V} \in \mathbb{R}^{n\times p}$ with $\|\boldsymbol{V}\|_F = O(\varepsilon)$, and set the decoder direction $\boldsymbol{W} \in \mathbb{R}^{p\times n}$ as $\boldsymbol{W} = \bar{\boldsymbol{W}} := (\boldsymbol{S}_p \circ (\boldsymbol{A}'\boldsymbol{A}))^{-1}\boldsymbol{T}_p\boldsymbol{V}'\boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}$. Then we use eq. (8) and eq. (9) of Lemma 3, to show that the given direction $(\boldsymbol{V},\boldsymbol{W})$ infinitesimally reduces the loss and hence, in every case the corresponding critical $(\boldsymbol{A},\boldsymbol{B})$ is a saddle point.

1. For the case $\boldsymbol{A} = \boldsymbol{U}_{\mathbb{I}_r}\boldsymbol{C}\boldsymbol{D}$, with $r < p$, note that based on the first item in Corollary 1, $\boldsymbol{B}\boldsymbol{\Sigma}_{xx}\boldsymbol{B}'$ is a $p \times p$ diagonal matrix of rank r so it has $p - r$ zero diagonal elements. Pick an $i \in \mathbb{N}_p$ such that $(\boldsymbol{B}\boldsymbol{\Sigma}_{xx}\boldsymbol{B}')_{ii}$ is zero and a $j \in \mathbb{N}_p \setminus \mathbb{I}_r$. Set $\boldsymbol{V} = \varepsilon\boldsymbol{U}_{i;j}\boldsymbol{D}$ and $\boldsymbol{W} = \bar{\boldsymbol{W}}$. Clearly,

$$\begin{aligned}
\boldsymbol{V}'\boldsymbol{A} =& \varepsilon\boldsymbol{D}\boldsymbol{U}'_{i;j}\boldsymbol{U}_{\mathbb{I}_r}\boldsymbol{C}\boldsymbol{D} = 0, && (41)\\
\boldsymbol{V}'\boldsymbol{V}\boldsymbol{T}_p\boldsymbol{B}\boldsymbol{\Sigma}_{xx}\boldsymbol{B}' =& \varepsilon^2\boldsymbol{D}\underbrace{\boldsymbol{U}'_{i;j}\boldsymbol{U}_{i;j}}\boldsymbol{D}\boldsymbol{T}_p\boldsymbol{B}\boldsymbol{\Sigma}_{xx}\boldsymbol{B}',\\
=& \varepsilon^2\boldsymbol{D}\boldsymbol{E}_i\boldsymbol{D}\boldsymbol{T}_p\boldsymbol{B}\boldsymbol{\Sigma}_{xx}\boldsymbol{B}' = \varepsilon^2\boldsymbol{D}^2\boldsymbol{T}_p\boldsymbol{E}_i(\boldsymbol{B}\boldsymbol{\Sigma}_{xx}\boldsymbol{B}') = 0 \text{ and} && (42)\\
\boldsymbol{V}'\boldsymbol{\Sigma}\boldsymbol{V} =& \varepsilon^2\boldsymbol{D}\boldsymbol{U}'_{i;j}\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}'\boldsymbol{U}_{i;j}\boldsymbol{D} = \varepsilon^2\lambda_j\boldsymbol{D}^2\boldsymbol{E}_i. && (43)
\end{aligned}$$

   Notice, $\|\boldsymbol{V}\|_F, \|\boldsymbol{W}\|_F = O(\varepsilon)$, so based on eq. (8) of Lemma 3, we have

$$\begin{aligned}
& L(\boldsymbol{A}+\boldsymbol{V},\boldsymbol{B}+\boldsymbol{W}) - L(\boldsymbol{A},\boldsymbol{B}) =\\
& \operatorname{Tr}(\boldsymbol{V}'\boldsymbol{V}\boldsymbol{T}_p\boldsymbol{B}\boldsymbol{\Sigma}_{xx}\boldsymbol{B}') - \operatorname{Tr}\left(\boldsymbol{V}'\boldsymbol{\Sigma}\boldsymbol{V}\boldsymbol{T}_p(\boldsymbol{S}_p\circ(\boldsymbol{A}'\boldsymbol{A}))^{-1}\boldsymbol{T}_p\right)\\
& + 2\operatorname{Tr}\left(\boldsymbol{V}'\boldsymbol{A}\left(\boldsymbol{S}_p\circ\left(\boldsymbol{B}\boldsymbol{\Sigma}_{xy}\boldsymbol{V}\boldsymbol{T}_p(\boldsymbol{S}_p\circ(\boldsymbol{A}'\boldsymbol{A}))^{-1} + (\boldsymbol{S}_p\circ(\boldsymbol{A}'\boldsymbol{A}))^{-1}\boldsymbol{T}_p\boldsymbol{V}'\boldsymbol{\Sigma}_{yx}\boldsymbol{B}'\right)\right)\right)\\
& + O(\varepsilon^3) \xRightarrow[eq.\ (42)]{eq.\ (41)}\\
& L(\boldsymbol{A}+\boldsymbol{V},\boldsymbol{B}+\boldsymbol{W}) - L(\boldsymbol{A},\boldsymbol{B}) =\\
& -\operatorname{Tr}\left(\boldsymbol{V}'\boldsymbol{\Sigma}\boldsymbol{V}\boldsymbol{T}_p(\boldsymbol{S}_p\circ(\boldsymbol{A}'\boldsymbol{A}))^{-1}\boldsymbol{T}_p\right) + O(\varepsilon^3) \xRightarrow[\boldsymbol{A}'\boldsymbol{A}=\boldsymbol{D}\boldsymbol{C}'\boldsymbol{C}\boldsymbol{D}]{eq.\ (43)}\\
& L(\boldsymbol{A}+\boldsymbol{V},\boldsymbol{B}+\boldsymbol{W}) - L(\boldsymbol{A},\boldsymbol{B}) =\\
& -\varepsilon^2\lambda_j\operatorname{Tr}\left(\boldsymbol{D}^2\boldsymbol{E}_i\boldsymbol{D}^{-1}\left(\left(\underbrace{\boldsymbol{T}_p^{-1}\boldsymbol{S}_p\boldsymbol{T}_p^{-1}}\right)\circ(\boldsymbol{C}'\boldsymbol{C})\right)^{-1}\boldsymbol{D}^{-1}\right) + O(\varepsilon^3) =\\
& -\varepsilon^2\lambda_j\left(\left(\hat{\boldsymbol{S}}_p\circ(\boldsymbol{C}'\boldsymbol{C})\right)^{-1}\right)_{ii} + O(\varepsilon^3).
\end{aligned}$$

   Therefore, since $\left(\hat{\boldsymbol{S}}_p\circ(\boldsymbol{C}'\boldsymbol{C})\right)^{-1}$ is a positive definite matrix, as $\varepsilon \to 0$, we have $L(\boldsymbol{A}+\boldsymbol{V},\boldsymbol{B}+\boldsymbol{W}) \leq L(\boldsymbol{A},\boldsymbol{B})$. Hence, any $(\boldsymbol{A},\boldsymbol{B}) = (\boldsymbol{U}_{\mathbb{I}_r}\boldsymbol{C}\boldsymbol{D}, \hat{\boldsymbol{B}}(\boldsymbol{U}_{\mathbb{I}_r}\boldsymbol{C}\boldsymbol{D}))$ with $r < p$ is a saddle point.

2. Next, consider the case where $\boldsymbol{A} = \boldsymbol{U}_{\mathbb{I}_p}\boldsymbol{\Pi}\boldsymbol{D}$ with $\mathbb{I}_p \neq \mathbb{N}_p$. Then there exists at least one $j \in \mathbb{I}_p \setminus \mathbb{N}_p$ and $i \in \mathbb{N}_p \setminus \mathbb{I}_p$ such that $i < j$ (so $\lambda_i > \lambda_j$). Let σ be the permutation corresponding to the permutation matrix $\boldsymbol{\Pi}$. Also, let $\varepsilon > 0$ and $\boldsymbol{U}_{\sigma(j);i} \in \mathbb{R}^{n\times p}$ be a matrix of all zeros except the $\sigma(j)^{\text{th}}$ column, which

contains $\boldsymbol{u}_i$; the eigenvector of $\boldsymbol{\Sigma}$ corresponding to the i^{th} largest eigenvalue. Set $\boldsymbol{V} = \varepsilon \boldsymbol{U}_{\sigma(j);i}\boldsymbol{D}$ and $\boldsymbol{W} = \bar{\boldsymbol{W}}$. Then, since $i \notin \mathbb{I}_p$ we have

$$\boldsymbol{V}'\boldsymbol{U}_{\mathbb{I}_p} = \varepsilon \boldsymbol{D}\boldsymbol{U}'_{\sigma(j);i}\boldsymbol{U}_{\mathbb{I}_p} = 0, \tag{44}$$

$$\boldsymbol{V}'\boldsymbol{V} = \varepsilon^2 \boldsymbol{D}\boldsymbol{U}'_{\sigma(j);i}\boldsymbol{U}_{\sigma(j);i}\boldsymbol{D} = \varepsilon^2 \boldsymbol{D}^2 \boldsymbol{E}_{\sigma(j)}, \text{ and} \tag{45}$$

$$\boldsymbol{V}'\boldsymbol{\Sigma}\boldsymbol{V} = \varepsilon^2 \boldsymbol{D}\boldsymbol{U}'_{\sigma(j);i}\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}'\boldsymbol{U}_{\sigma(j);i}\boldsymbol{D} = \varepsilon^2 \lambda_i \boldsymbol{D}^2 \boldsymbol{E}_{\sigma(j)}. \tag{46}$$

Since $\|\boldsymbol{V}\|_F, \|\boldsymbol{W}\|_F = O(\varepsilon)$, based on eq. (9) of Lemma 3, we have

$$\begin{aligned}
L(\boldsymbol{A}+\boldsymbol{V}, \boldsymbol{B}+\boldsymbol{W}) - L(\boldsymbol{A}, \boldsymbol{B}) =& \operatorname{Tr}\left(\boldsymbol{V}'\boldsymbol{V}\boldsymbol{\Pi}'\boldsymbol{\Lambda}_{\mathbb{I}_p}\boldsymbol{\Pi}\boldsymbol{T}_p\boldsymbol{D}^{-2}\right) - \operatorname{Tr}\left(\boldsymbol{V}'\boldsymbol{\Sigma}\boldsymbol{V}\boldsymbol{T}_p\boldsymbol{D}^{-2}\right) \\
&+2\operatorname{Tr}\left(\boldsymbol{V}'\boldsymbol{U}_{\mathbb{I}_p}\boldsymbol{\Pi}\boldsymbol{D}\left(\boldsymbol{S}_p \circ \left(\boldsymbol{D}^{-1}\boldsymbol{\Pi}'\boldsymbol{U}'_{\mathbb{I}_p}\boldsymbol{\Sigma}\boldsymbol{V}\boldsymbol{D}^{-2}\right)\right)\right) \\
&+2\operatorname{Tr}\left(\boldsymbol{V}'\boldsymbol{U}_{\mathbb{I}_p}\boldsymbol{\Pi}\boldsymbol{D}\left(\boldsymbol{S}_p \circ \left(\boldsymbol{D}^{-2}\boldsymbol{V}'\boldsymbol{\Sigma}\boldsymbol{U}_{\mathbb{I}_p}\boldsymbol{\Pi}\boldsymbol{D}^{-1}\right)\right)\right) \\
&+O(\varepsilon^3) \xrightarrow[eq.\ (45), eq.\ (46)]{eq.\ (44)} \\
L(\boldsymbol{A}+\boldsymbol{V}, \boldsymbol{B}+\boldsymbol{W}) - L(\boldsymbol{A}, \boldsymbol{B}) =& \operatorname{Tr}\left(\varepsilon^2\boldsymbol{D}^2\boldsymbol{E}_{\sigma(j)}\boldsymbol{\Pi}'\boldsymbol{\Lambda}_{\mathbb{I}_p}\boldsymbol{\Pi}\boldsymbol{T}_p\boldsymbol{D}^{-2}\right) \\
&- \operatorname{Tr}\left(\varepsilon^2\lambda_i\boldsymbol{D}^2\boldsymbol{E}_{\sigma(j)}\boldsymbol{T}_p\boldsymbol{D}^{-2}\right) + O(\varepsilon^3) \\
=& \varepsilon^2 \operatorname{Tr}\Big(\underbrace{\boldsymbol{E}_{\sigma(j)}\boldsymbol{\Pi}'\boldsymbol{\Lambda}_{\mathbb{I}_p}\boldsymbol{\Pi}}\boldsymbol{T}_p\Big) - \varepsilon^2\lambda_i \operatorname{Tr}\big(\boldsymbol{E}_{\sigma(j)}\boldsymbol{T}_p\big) + O(\varepsilon^3) \\
=& \varepsilon^2\lambda_j \operatorname{Tr}\left(\boldsymbol{E}_{\sigma(j)}\boldsymbol{T}_p\right) - \varepsilon^2\lambda_i \operatorname{Tr}\left(\boldsymbol{E}_{\sigma(j)}\boldsymbol{T}_p\right) + O(\varepsilon^3) \\
=& -\varepsilon^2(p - \sigma(j) + 1)(\lambda_i - \lambda_j) + O(\varepsilon^3).
\end{aligned}$$

Note that in the above, the diagonal matrix $\boldsymbol{\Pi}'\boldsymbol{\Lambda}_{\mathbb{I}_p}\boldsymbol{\Pi}$ has the same diagonal elements as $\boldsymbol{\Lambda}_{\mathbb{I}_p}$ but they are permuted by σ. So $\boldsymbol{E}_{\sigma(j)}\boldsymbol{\Pi}'\boldsymbol{\Lambda}_{\mathbb{I}_p}\boldsymbol{\Pi}$ selects $\sigma(j)^{\text{th}}$ diagonal element of $\boldsymbol{\Pi}'\boldsymbol{\Lambda}_{\mathbb{I}_p}\boldsymbol{\Pi}$ that is the j^{th}diagonal element of $\boldsymbol{\Lambda}_{\mathbb{I}_p}$, which is nothing but λ_j. Now, since $i < j$ so $\lambda_i > \lambda_j$ and $\sigma(j) \leq p$, as $\varepsilon \to 0$, we have $L(\boldsymbol{A}+\boldsymbol{V}, \boldsymbol{B}+\boldsymbol{W}) \leq L(\boldsymbol{A}, \boldsymbol{B})$. Hence, any $(\boldsymbol{A}, \boldsymbol{B}) = (\boldsymbol{U}_{\mathbb{I}_p}\boldsymbol{\Pi}\boldsymbol{D}, \hat{\boldsymbol{B}}(\boldsymbol{U}_{\mathbb{I}_p}\boldsymbol{\Pi}\boldsymbol{D}))$ is a saddle point.

3. Finally consider the case where $\boldsymbol{A} = \boldsymbol{U}_{\mathbb{N}_p}\boldsymbol{\Pi}\boldsymbol{D}$ with $\boldsymbol{\Pi} \neq \boldsymbol{I}_p$. Since $\boldsymbol{\Pi} \neq \boldsymbol{I}_p$, the permutation σ of the set $\mathbb{N}_p$, corresponding to the permutation matrix $\boldsymbol{\Pi}$, has at least a cycle $(i_1 i_2 \cdots i_k)$, where $1 < i_1 < i_2 \cdots < i_k < p$ and $2 \leq k \leq p$. Hence, $\boldsymbol{\Pi}$ can be decomposed as $\boldsymbol{\Pi} = \boldsymbol{\Pi}_{(i_1 i_2 \cdots i_k)}\hat{\boldsymbol{\Pi}}$, where $\hat{\boldsymbol{\Pi}}$ is the permutation matrix corresponding to other cycles of σ. The cycle $(i_1 i_2 \cdots i_k)$ can be decomposed into transpositions as $(i_1 i_2 \cdots i_k) = (i_k i_{k-1}) \cdots (i_k i_1)$, which in matrix form is $\boldsymbol{\Pi}_{(i_1 i_2 \cdots i_k)} = \boldsymbol{\Pi}_{(i_k i_1)}\boldsymbol{\Pi}_{(i_k i_2)} \cdots \boldsymbol{\Pi}_{(i_k i_{k-1})}$. Therefore, $\boldsymbol{\Pi}$ can be decomposed as $\boldsymbol{\Pi} = \boldsymbol{\Pi}_{(i_k i_1)}\tilde{\boldsymbol{\Pi}}$, where $\tilde{\boldsymbol{\Pi}} = \boldsymbol{\Pi}_{(i_k i_2)} \cdots \boldsymbol{\Pi}_{(i_k i_{k-1})}\hat{\boldsymbol{\Pi}}$. Note that $\boldsymbol{\Pi}_{(i_k i_1)}$, the permutation matrix corresponding to transposition $(i_k i_1)$ is a symmetric involutory matrix, i.e. $\boldsymbol{\Pi}^2_{(i_k i_1)} = \boldsymbol{I}_p$. Set $\boldsymbol{V} = \varepsilon(\boldsymbol{U}_{i_1;i_1} - \boldsymbol{U}_{i_k;i_k})\tilde{\boldsymbol{\Pi}}\boldsymbol{D}$ and $\boldsymbol{W} = \bar{\boldsymbol{W}}$. Again we replace $\boldsymbol{V}$ and $\boldsymbol{W}$ in eq. (9) of Lemma 3. There are some tedious steps to simplify the equation, which is given in section 2.1. The final result is as follows. With the given $\boldsymbol{V}$ and $\boldsymbol{W}$, the third and forth terms of the RHS of eq. (9) are canceled and the first two terms are simplified to

$$\operatorname{Tr}\left(\boldsymbol{V}'\boldsymbol{V}\boldsymbol{\Pi}'\boldsymbol{\Lambda}_{\mathbb{N}_p}\boldsymbol{\Pi}\boldsymbol{T}_p\boldsymbol{D}^{-2}\right) = \varepsilon^2\lambda_{i_k}(p - i_1 + 1) + \varepsilon^2\lambda_{i_1}(p - i_m + 1), \text{ and} \tag{47}$$

$$\operatorname{Tr}\left(\boldsymbol{V}'\boldsymbol{\Sigma}\boldsymbol{V}\boldsymbol{T}_p\boldsymbol{D}^{-2}\right) = \varepsilon^2\lambda_{i_1}(p - i_1 + 1) + \varepsilon^2\lambda_{i_k}(p - i_m + 1), \tag{48}$$

in which, $m = \max\{k-1, 2\}$. This means that If the selected cycle is just a transposition $(i_1 i_2)$ then $i_m = i_2$. But if for the selected cycle $(i_1 i_2 \cdots i_k)$, k is greater than 2 then $i_m = i_{k-1}$. Using above equations, eq. (9) yields

$$\begin{aligned}L(\boldsymbol{A}+\boldsymbol{V},\boldsymbol{B}+\boldsymbol{W})-L(\boldsymbol{A},\boldsymbol{B}) =&\operatorname{Tr}\left(\boldsymbol{V}'\boldsymbol{V}\boldsymbol{\Pi}'\boldsymbol{\Lambda}_{\mathbb{I}_p}\boldsymbol{\Pi}\boldsymbol{T}_p\boldsymbol{D}^{-2}\right)-\operatorname{Tr}\left(\boldsymbol{V}'\boldsymbol{\Sigma}\boldsymbol{V}\boldsymbol{T}_p\boldsymbol{D}^{-2}\right)+O(\varepsilon^3)\\
=&\varepsilon^2\lambda_{i_k}(p-i_1+1)+\varepsilon^2\lambda_{i_1}(p-i_m+1)\\
&-\varepsilon^2\lambda_{i_1}(p-i_1+1)-\varepsilon^2\lambda_{i_k}(p-i_m+1)+O(\varepsilon^3)\\
=&-\varepsilon^2 i_1\lambda_{i_k}-\varepsilon^2 i_m\lambda_{i_1}+\varepsilon^2 i_1\lambda_{i_1}+\varepsilon^2 i_m\lambda_{i_k}\\
=&-\varepsilon^2\left((\lambda_{i_1}-\lambda_{i_k})(i_m-i_1)\right)+O(\varepsilon^3).\end{aligned}\tag{49}$$

By the above definition of i_m, we have $i_m - i_1 > 0$ and since $i_1 < i_k$, $\lambda_{i_1} - \lambda_{i_k} > 0$. Hence, the first term in the above equation is negative and as $\varepsilon \to 0$, we have $L(\boldsymbol{A}+\boldsymbol{V},\boldsymbol{B}+\boldsymbol{W}) - L(\boldsymbol{A},\boldsymbol{B}) < 0$. Therefore, any any $(\boldsymbol{A},\boldsymbol{B}) = (\boldsymbol{U}_{\mathbb{I}_p}\boldsymbol{\Pi}\boldsymbol{D}, \hat{\boldsymbol{B}}(\boldsymbol{U}_{\mathbb{I}_p}\boldsymbol{\Pi}\boldsymbol{D}))$ with $\boldsymbol{\Pi} \neq \boldsymbol{I}_p$ is a saddle point.

The Sufficient condition:

From Lemma 4 we know that the loss $L(\boldsymbol{A},\boldsymbol{B})$ can be written in the form of eq. (36). Use this equation to evaluate loss at $(\boldsymbol{A}^*,\boldsymbol{B}^*) = \left(\boldsymbol{U}_{\mathbb{N}_p}\boldsymbol{D}_p, \boldsymbol{D}_p^{-1}\boldsymbol{U}'_{\mathbb{N}_p}\boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}\right)$ as follows

$$\begin{aligned}
&L(\boldsymbol{A}^*,\boldsymbol{B}^*) = p\operatorname{Tr}(\boldsymbol{\Sigma}_{yy}) - 2\operatorname{Tr}\left(\boldsymbol{A}^*\boldsymbol{T}_p\boldsymbol{B}^*\boldsymbol{\Sigma}_{xy}\right) + \operatorname{Tr}\left(\boldsymbol{B}^{*'}\left(\boldsymbol{S}_p\circ\left(\boldsymbol{A}^{*'}\boldsymbol{A}^*\right)\right)\boldsymbol{B}^*\boldsymbol{\Sigma}_{xx}\right) \implies\\
&L(\boldsymbol{A}^*,\boldsymbol{B}^*) = p\operatorname{Tr}(\boldsymbol{\Sigma}_{yy}) - 2\operatorname{Tr}\left(\boldsymbol{U}_{\mathbb{N}_p}\boldsymbol{D}_p\boldsymbol{T}_p\boldsymbol{D}_p^{-1}\boldsymbol{U}'_{\mathbb{N}_p}\underbrace{\boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}}\right)\\
&\qquad + \operatorname{Tr}\left(\left(\boldsymbol{S}_p\circ\left(\boldsymbol{D}_p\underbrace{\boldsymbol{U}'_{\mathbb{N}_p}\boldsymbol{U}_{\mathbb{N}_p}}\boldsymbol{D}_p\right)\right)\boldsymbol{D}_p^{-1}\boldsymbol{U}'_{\mathbb{N}_p}\underbrace{\boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xx}\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}}\boldsymbol{U}_{\mathbb{N}_p}\boldsymbol{D}_p^{-1}\right) \implies\\
&L(\boldsymbol{A}^*,\boldsymbol{B}^*) = p\operatorname{Tr}(\boldsymbol{\Sigma}_{yy}) - 2\operatorname{Tr}\left(\boldsymbol{T}_p\underbrace{\boldsymbol{D}_p\boldsymbol{D}_p^{-1}}\underbrace{\boldsymbol{U}'_{\mathbb{N}_p}\boldsymbol{\Sigma}\boldsymbol{U}_{\mathbb{N}_p}}\right)\\
&\qquad + \operatorname{Tr}\left(\left(\underbrace{\boldsymbol{S}_p\circ(\boldsymbol{I}_p)}\right)\underbrace{\boldsymbol{D}_p\boldsymbol{D}_p^{-1}}\underbrace{\boldsymbol{U}'_{\mathbb{N}_p}\boldsymbol{\Sigma}\boldsymbol{U}_{\mathbb{N}_p}}\underbrace{\boldsymbol{D}_p^{-1}\boldsymbol{D}_p}\right) \implies\\
&L(\boldsymbol{A}^*,\boldsymbol{B}^*) = p\operatorname{Tr}(\boldsymbol{\Sigma}_{yy}) - 2\operatorname{Tr}\left(\boldsymbol{T}_p\boldsymbol{\Lambda}_{\mathbb{N}_p}\right) + \operatorname{Tr}\left(\boldsymbol{T}_p\boldsymbol{\Lambda}_{\mathbb{N}_p}\right) \implies\\
&L(\boldsymbol{A}^*,\boldsymbol{B}^*) = p\operatorname{Tr}(\boldsymbol{\Sigma}_{yy}) - \operatorname{Tr}\left(\boldsymbol{T}_p\boldsymbol{\Lambda}_{\mathbb{N}_p}\right) = p\ \operatorname{Tr}(\boldsymbol{\Sigma}_{yy}) - \sum_{i=1}^{p}(p-i+1)\,\lambda_i,
\end{aligned}$$

which is eq. (39), as claimed. Notice that the above value is independent of the diagonal matrix $\boldsymbol{D}_p$. From the necessary condition we know that any critical point not in the form of $(\boldsymbol{A}^*,\boldsymbol{B}^*)$ is a saddle point. Hence, due to the convexity of the loss at least one $(\boldsymbol{A}^*,\boldsymbol{B}^*)$ is a global minimum but since the value of the loss at $(\boldsymbol{A}^*,\boldsymbol{B}^*)$ is independent of $\boldsymbol{D}_p$ all these critical points yield the same value for the loss. Therefore, any critical point in the form of $(\boldsymbol{A}^*,\boldsymbol{B}^*)$ is a local and global minima. □

2.1 Supplementary details of the proof of Theorem 2

To verify eq. (47), eq. (48), and eq. (49) in the proof of Theorem 2, we want to replace $\boldsymbol{V}$ and $\boldsymbol{W}$ in eq. (9) of Lemma 3 with $\boldsymbol{V} = \varepsilon(\boldsymbol{U}_{i_1;i_1} - \boldsymbol{U}_{i_k;i_k})\tilde{\boldsymbol{\Pi}}\boldsymbol{D}$ and $\boldsymbol{W} = \bar{\boldsymbol{W}}$ and simplify. eq. (9) is

$$\begin{aligned}L(\boldsymbol{A}+\boldsymbol{V},\boldsymbol{B}+\boldsymbol{W})-L(\boldsymbol{A},\boldsymbol{B}) =& \operatorname{Tr}\left(\boldsymbol{V}'\boldsymbol{V}\boldsymbol{\Pi}'\boldsymbol{\Lambda}_{\mathbb{I}_p}\boldsymbol{\Pi}\boldsymbol{T}_p\boldsymbol{D}^{-2}\right)-\operatorname{Tr}\left(\boldsymbol{V}'\boldsymbol{\Sigma}\boldsymbol{V}\boldsymbol{T}_p\boldsymbol{D}^{-2}\right)\\
&+2\operatorname{Tr}\left(\boldsymbol{V}'\boldsymbol{U}_{\mathbb{I}_p}\boldsymbol{\Pi}\boldsymbol{D}\left(\boldsymbol{S}_p\circ\left(\boldsymbol{D}^{-1}\boldsymbol{\Pi}'\boldsymbol{U}'_{\mathbb{I}_p}\boldsymbol{\Sigma}\boldsymbol{V}\boldsymbol{D}^{-2}\right)\right)\right)\end{aligned}$$

$$+2\,\mathrm{Tr}\left(\boldsymbol{V}'\boldsymbol{U}_{\mathbb{I}_p}\boldsymbol{\Pi}\boldsymbol{D}\left(\boldsymbol{S}_p\circ\left(\boldsymbol{D}^{-2}\boldsymbol{V}'\boldsymbol{\Sigma}\boldsymbol{U}_{\mathbb{I}_p}\boldsymbol{\Pi}\boldsymbol{D}^{-1}\right)\right)\right)$$
$$+O(\varepsilon^3).$$

We investigate each term on the RHS separately. but before note that

$$\boldsymbol{E}_i\tilde{\boldsymbol{\Pi}}\boldsymbol{T}_p\tilde{\boldsymbol{\Pi}}' = \left(\tilde{\boldsymbol{\Pi}}\boldsymbol{T}_p\tilde{\boldsymbol{\Pi}}'\right)_{i,i}\boldsymbol{E}_i = (\boldsymbol{T}_p)_{\tilde{\sigma}^{-1}(i),\tilde{\sigma}^{-1}(i)}\boldsymbol{E}_i = (p-\tilde{\sigma}^{-1}(i)+1)\boldsymbol{E}_i, \tag{50}$$

where, $\tilde{\sigma}$ and its function inverse $\tilde{\sigma}^{-1}$ are permutations corresponding to $\tilde{\boldsymbol{\Pi}}$ and $\tilde{\boldsymbol{\Pi}}'$ respectively. $\tilde{\boldsymbol{\Pi}}\boldsymbol{T}_p\tilde{\boldsymbol{\Pi}}'$ is a diagonal matrix where diagonal elements of $\boldsymbol{T}_p$ are ordered based on $\tilde{\sigma}^{-1}$. Moreover, recall that we decomposed the permutation matrix $\boldsymbol{\Pi}$ in $\boldsymbol{A}$ with a cycle $(i_1i_2\cdots i_k)$ as $\boldsymbol{\Pi}=\boldsymbol{\Pi}_{(i_1i_k)}\underbrace{\boldsymbol{\Pi}_{(i_ki_2)}\cdots\boldsymbol{\Pi}_{(i_ki_{k-1})}\hat{\boldsymbol{\Pi}}}=\boldsymbol{\Pi}_{(i_1i_k)}\tilde{\boldsymbol{\Pi}}$, where $i_1,i_2,\cdots i_k$ are fixed points of $\hat{\boldsymbol{\Pi}}$. Therefore, with $\tilde{\sigma}$ being the permutation corresponding to $\tilde{\boldsymbol{\Pi}}$ we have

$$\tilde{\sigma}(i_1)=i_1 \implies \tilde{\sigma}^{-1}(i_1)=i_1, \text{ and} \tag{51}$$
$$\tilde{\sigma}(i_{k-1})=i_m \implies \tilde{\sigma}^{-1}(i_k)=i_m, \tag{52}$$

where, $m=\max\{k-1,2\}$. This means that If the selected cycle is just a transposition (i_1i_2) then $i_m=i_2$. But if for the selected cycle $(i_1i_2\cdots i_k)$, k is greater than 2 then $i_m=i_{k-1}$.

For the first term we have

$$\begin{aligned}
\boldsymbol{V}'\boldsymbol{V} &= \varepsilon^2\boldsymbol{D}\tilde{\boldsymbol{\Pi}}'(\boldsymbol{U}'_{i_1;i_1}-\boldsymbol{U}'_{i_k;i_k})(\boldsymbol{U}_{i_1;i_1}-\boldsymbol{U}_{i_k;i_k})\tilde{\boldsymbol{\Pi}}\boldsymbol{D} \xrightarrow{\boldsymbol{U}'_{i_1;i_1}\boldsymbol{U}_{i_k;i_k}=0}\\
\boldsymbol{V}'\boldsymbol{V} &= \varepsilon^2\boldsymbol{D}\tilde{\boldsymbol{\Pi}}'(\boldsymbol{U}'_{i_1;i_1}\boldsymbol{U}_{i_1;i_1}+\boldsymbol{U}'_{i_k;i_k}\boldsymbol{U}_{i_k;i_k})\tilde{\boldsymbol{\Pi}}\boldsymbol{D} \xrightarrow[\boldsymbol{U}'_{i_k;i_k}\boldsymbol{U}_{i_k;i_k}=\boldsymbol{E}_{i_k}]{\boldsymbol{U}'_{i_1;i_1}\boldsymbol{U}_{i_1;i_1}=\boldsymbol{E}_{i_1}}\\
\boldsymbol{V}'\boldsymbol{V} &= \varepsilon^2\boldsymbol{D}\tilde{\boldsymbol{\Pi}}'(\boldsymbol{E}_{i_1}+\boldsymbol{E}_{i_k})\tilde{\boldsymbol{\Pi}}\boldsymbol{D} \xrightarrow{\tilde{\boldsymbol{\Pi}}'(\boldsymbol{E}_{i_1}+\boldsymbol{E}_{i_k})\tilde{\boldsymbol{\Pi}}\text{ is diagonal}}\\
\boldsymbol{V}'\boldsymbol{V} &= \varepsilon^2\tilde{\boldsymbol{\Pi}}'(\boldsymbol{E}_{i_1}+\boldsymbol{E}_{i_k})\tilde{\boldsymbol{\Pi}}\boldsymbol{D}^2 \implies\\
\mathrm{Tr}\left(\boldsymbol{V}'\boldsymbol{V}\boldsymbol{\Pi}'\boldsymbol{\Lambda}_{\mathbb{N}_p}\boldsymbol{\Pi}\boldsymbol{T}_p\boldsymbol{D}^{-2}\right) &= \mathrm{Tr}\left(\overbrace{\boldsymbol{V}'\boldsymbol{V}}\boldsymbol{D}^{-2}\tilde{\boldsymbol{\Pi}}'\boldsymbol{\Pi}_{(i_1i_k)}\boldsymbol{\Lambda}_{\mathbb{N}_p}\boldsymbol{\Pi}_{(i_1i_k)}\tilde{\boldsymbol{\Pi}}\boldsymbol{T}_p\right)\\
&= \mathrm{Tr}\left(\varepsilon^2\tilde{\boldsymbol{\Pi}}'(\boldsymbol{E}_{i_1}+\boldsymbol{E}_{i_k})\underbrace{\tilde{\boldsymbol{\Pi}}\boldsymbol{D}^2\boldsymbol{D}^{-2}\tilde{\boldsymbol{\Pi}}'}_{\boldsymbol{I}_p}\boldsymbol{\Pi}_{(i_1i_k)}\boldsymbol{\Lambda}_{\mathbb{N}_p}\boldsymbol{\Pi}_{(i_1i_k)}\tilde{\boldsymbol{\Pi}}\boldsymbol{T}_p\right)\\
&= \varepsilon^2\,\mathrm{Tr}\left((\boldsymbol{E}_{i_1}+\boldsymbol{E}_{i_k})\boldsymbol{\Pi}_{(i_1i_k)}\boldsymbol{\Lambda}_{\mathbb{N}_p}\boldsymbol{\Pi}_{(i_1i_k)}\tilde{\boldsymbol{\Pi}}\boldsymbol{T}_p\tilde{\boldsymbol{\Pi}}'\right)\\
&= \varepsilon^2\,\mathrm{Tr}\left(\lambda_{i_k}\boldsymbol{E}_{i_1}\tilde{\boldsymbol{\Pi}}\boldsymbol{T}_p\tilde{\boldsymbol{\Pi}}'+\lambda_{i_1}\boldsymbol{E}_{i_k}\tilde{\boldsymbol{\Pi}}\boldsymbol{T}_p\tilde{\boldsymbol{\Pi}}'\right) \xrightarrow{\text{eq. (50)}}\\
\mathrm{Tr}\left(\boldsymbol{V}'\boldsymbol{V}\boldsymbol{\Pi}'\boldsymbol{\Lambda}_{\mathbb{N}_p}\boldsymbol{\Pi}\boldsymbol{T}_p\boldsymbol{D}^{-2}\right) &= \varepsilon^2\lambda_{i_k}(p-\tilde{\sigma}^{-1}(i_1)+1)\boldsymbol{E}_{i_1}+\varepsilon^2\lambda_{i_1}(p-\tilde{\sigma}^{-1}(i_k)+1)\boldsymbol{E}_{i_k} \xrightarrow[\text{eq. (52)}]{\text{eq. (51)}}\\
\mathrm{Tr}\left(\boldsymbol{V}'\boldsymbol{V}\boldsymbol{\Pi}'\boldsymbol{\Lambda}_{\mathbb{N}_p}\boldsymbol{\Pi}\boldsymbol{T}_p\boldsymbol{D}^{-2}\right) &= \varepsilon^2\lambda_{i_k}(p-i_1+1)\boldsymbol{E}_{i_1}+\varepsilon^2\lambda_{i_1}(p-i_m+1)\boldsymbol{E}_{i_k},
\end{aligned}$$

which is eq. (47) as claimed.

For the second term we have

$$\boldsymbol{V}'\boldsymbol{\Sigma}\boldsymbol{V} = \varepsilon^2\boldsymbol{D}\tilde{\boldsymbol{\Pi}}'(\boldsymbol{U}'_{i_1;i_1}-\boldsymbol{U}'_{i_k;i_k})\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}'(\boldsymbol{U}_{i_1;i_1}-\boldsymbol{U}_{i_k;i_k})\tilde{\boldsymbol{\Pi}}\boldsymbol{D}$$

$$
\begin{aligned}
&= \varepsilon^2 \boldsymbol{D}\tilde{\boldsymbol{\Pi}}'(\underbrace{\boldsymbol{U}'_{i_1;i_1}\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}'\boldsymbol{U}_{i_1;i_1}}_{\lambda_{i_1}\boldsymbol{E}_{i_1}} - \underbrace{\boldsymbol{U}'_{i_1;i_1}\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}'\boldsymbol{U}_{i_k;i_k}}_{0} \\
&- \underbrace{\boldsymbol{U}'_{i_k;i_k}\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}'\boldsymbol{U}_{i_1;i_1}}_{0} + \underbrace{\boldsymbol{U}'_{i_k;i_k}\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}'\boldsymbol{U}_{i_k;i_k}}_{\lambda_{i_k}\boldsymbol{E}_{i_k}})\tilde{\boldsymbol{\Pi}}\boldsymbol{D} \\
&= \varepsilon^2 \tilde{\boldsymbol{\Pi}}'(\lambda_{i_1}\boldsymbol{E}_{i_1} + \lambda_{i_k}\boldsymbol{E}_{i_k})\tilde{\boldsymbol{\Pi}}\boldsymbol{D}^2 \implies \\
\operatorname{Tr}\left(\boldsymbol{V}'\boldsymbol{\Sigma}\boldsymbol{V}\boldsymbol{T}_p\boldsymbol{D}^{-2}\right) &= \operatorname{Tr}\left(\varepsilon^2 \tilde{\boldsymbol{\Pi}}'(\lambda_{i_1}\boldsymbol{E}_{i_1} + \lambda_{i_k}\boldsymbol{E}_{i_k})\tilde{\boldsymbol{\Pi}}\boldsymbol{D}^2\boldsymbol{T}_p\boldsymbol{D}^{-2}\right) \\
&= \varepsilon^2 \operatorname{Tr}\left(\lambda_{i_1}\boldsymbol{E}_{i_1}\tilde{\boldsymbol{\Pi}}\boldsymbol{T}_p\tilde{\boldsymbol{\Pi}}' + \lambda_{i_k}\boldsymbol{E}_{i_k}\tilde{\boldsymbol{\Pi}}\boldsymbol{T}_p\tilde{\boldsymbol{\Pi}}'\right) \xRightarrow{\text{eq. (50)}} \\
\operatorname{Tr}\left(\boldsymbol{V}'\boldsymbol{\Sigma}\boldsymbol{V}\boldsymbol{T}_p\boldsymbol{D}^{-2}\right) &= \varepsilon^2\lambda_{i_1}(p - \tilde{\sigma}^{-1}(i_1) + 1) + \varepsilon^2\lambda_{i_k}(p - \tilde{\sigma}^{-1}(i_k) + 1) \xRightarrow[\text{eq. (52)}]{\text{eq. (51)}} \\
\operatorname{Tr}\left(\boldsymbol{V}'\boldsymbol{\Sigma}\boldsymbol{V}\boldsymbol{T}_p\boldsymbol{D}^{-2}\right) &= \varepsilon^2\lambda_{i_1}(p - i_1 + 1) + \varepsilon^2\lambda_{i_k}(p - i_m + 1),
\end{aligned}
$$

which is eq. (48) as claimed.

Finally, we have to show that the third and the forth terms of the eq. (9) are canceled. First, observe that

$$
\begin{aligned}
\operatorname{Tr}\left(\boldsymbol{V}'\boldsymbol{U}_{\mathbb{N}_p}\boldsymbol{\Pi}\boldsymbol{D}\left(\boldsymbol{S}_p \circ \left(\boldsymbol{D}^{-1}\boldsymbol{\Pi}'\boldsymbol{U}'_{\mathbb{N}_p}\boldsymbol{\Sigma}\boldsymbol{V}\boldsymbol{D}^{-2}\right)\right)\right) &= \\
\operatorname{Tr}\left(\varepsilon\boldsymbol{D}\tilde{\boldsymbol{\Pi}}'(\boldsymbol{U}'_{i_1;i_1} - \boldsymbol{U}'_{i_k;i_k})\boldsymbol{U}_{\mathbb{N}_p}\boldsymbol{\Pi}\left(\boldsymbol{S}_p \circ \left(\boldsymbol{\Pi}'\boldsymbol{U}'_{\mathbb{N}_p}\boldsymbol{\Sigma}\boldsymbol{V}\boldsymbol{D}^{-2}\right)\right)\right) &= \\
\varepsilon\operatorname{Tr}\left(\tilde{\boldsymbol{\Pi}}'(\boldsymbol{E}_{i_1} - \boldsymbol{E}_{i_k})\boldsymbol{\Pi}\left(\boldsymbol{S}_p \circ \left(\boldsymbol{\Pi}'\boldsymbol{U}'_{\mathbb{N}_p}\boldsymbol{\Sigma}\boldsymbol{V}\boldsymbol{D}^{-2}\right)\right)\boldsymbol{D}\right) &= \\
\varepsilon^2\operatorname{Tr}\left(\tilde{\boldsymbol{\Pi}}'(\boldsymbol{E}_{i_1} - \boldsymbol{E}_{i_k})\boldsymbol{\Pi}\left(\boldsymbol{S}_p \circ \left(\boldsymbol{\Pi}'(\lambda_{i_1}\boldsymbol{E}_{i_1} - \lambda_{i_k}\boldsymbol{E}_{i_k})\tilde{\boldsymbol{\Pi}}\right)\right)\right) &= \\
\varepsilon^2\operatorname{Tr}\left((\boldsymbol{E}_{i_1} - \boldsymbol{E}_{i_k})\left(\left(\boldsymbol{\Pi}\boldsymbol{S}_p\tilde{\boldsymbol{\Pi}}'\right) \circ \left(\boldsymbol{\Pi}\boldsymbol{\Pi}'(\lambda_{i_1}\boldsymbol{E}_{i_1} - \lambda_{i_k}\boldsymbol{E}_{i_k})\tilde{\boldsymbol{\Pi}}\tilde{\boldsymbol{\Pi}}'\right)\right)\right) &= \\
\varepsilon^2\operatorname{Tr}\left(\left(\boldsymbol{\Pi}\boldsymbol{S}_p\tilde{\boldsymbol{\Pi}}'\right) \circ \left((\boldsymbol{E}_{i_1} - \boldsymbol{E}_{i_k})\left(\lambda_{i_1}\boldsymbol{E}_{i_1} - \lambda_{i_k}\boldsymbol{E}_{i_k}\right)\right)\right) &= \\
\varepsilon^2\operatorname{Tr}\left(\left(\boldsymbol{\Pi}\boldsymbol{S}_p\tilde{\boldsymbol{\Pi}}'\right) \circ \left(\lambda_{i_1}\boldsymbol{E}_{i_1} + \lambda_{i_k}\boldsymbol{E}_{i_k}\right)\right)&, \text{ and}
\end{aligned}
$$

$$
\begin{aligned}
\operatorname{Tr}\left(\boldsymbol{V}'\boldsymbol{U}_{\mathbb{N}_p}\boldsymbol{\Pi}\boldsymbol{D}\left(\boldsymbol{S}_p \circ \left(\boldsymbol{D}^{-2}\boldsymbol{V}'\boldsymbol{\Sigma}\boldsymbol{U}_{\mathbb{N}_p}\boldsymbol{\Pi}\boldsymbol{D}^{-1}\right)\right)\right) &= \\
\operatorname{Tr}\left(\varepsilon\boldsymbol{D}\tilde{\boldsymbol{\Pi}}'(\boldsymbol{U}'_{i_1;i_1} - \boldsymbol{U}'_{i_k;i_k})\boldsymbol{U}_{\mathbb{N}_p}\boldsymbol{\Pi}\left(\boldsymbol{S}_p \circ \left(\boldsymbol{D}^{-1}\boldsymbol{V}'\boldsymbol{\Sigma}\boldsymbol{U}_{\mathbb{N}_p}\boldsymbol{\Pi}\boldsymbol{D}^{-1}\right)\right)\right) &= \\
\varepsilon\operatorname{Tr}\left(\tilde{\boldsymbol{\Pi}}'(\boldsymbol{E}_{i_1} - \boldsymbol{E}_{i_k})\boldsymbol{\Pi}\left(\boldsymbol{S}_p \circ \left(\boldsymbol{D}^{-1}\boldsymbol{V}'\boldsymbol{\Sigma}\boldsymbol{U}_{\mathbb{N}_p}\boldsymbol{\Pi}\right)\right)\right) &= \\
\varepsilon^2\operatorname{Tr}\left((\boldsymbol{E}_{i_1} - \boldsymbol{E}_{i_k})\boldsymbol{\Pi}\left(\boldsymbol{S}_p \circ \left(\tilde{\boldsymbol{\Pi}}'(\lambda_{i_1}\boldsymbol{E}_{i_1} - \lambda_{i_k}\boldsymbol{E}_{i_k})\boldsymbol{\Pi}\right)\right)\tilde{\boldsymbol{\Pi}}'\right) &= \\
\varepsilon^2\operatorname{Tr}\left((\boldsymbol{E}_{i_1} - \boldsymbol{E}_{i_k})\left(\left(\boldsymbol{\Pi}\boldsymbol{S}_p\tilde{\boldsymbol{\Pi}}'\right) \circ \left(\boldsymbol{\Pi}\tilde{\boldsymbol{\Pi}}'(\lambda_{i_1}\boldsymbol{E}_{i_1} - \lambda_{i_k}\boldsymbol{E}_{i_k})\boldsymbol{\Pi}\tilde{\boldsymbol{\Pi}}'\right)\right)\right) &= \\
\varepsilon^2\operatorname{Tr}\left((\boldsymbol{E}_{i_1} - \boldsymbol{E}_{i_k})\left(\left(\boldsymbol{\Pi}\boldsymbol{S}_p\tilde{\boldsymbol{\Pi}}'\right) \circ \left(\boldsymbol{\Pi}_{(i_1 i_k)}(\lambda_{i_1}\boldsymbol{E}_{i_1} - \lambda_{i_k}\boldsymbol{E}_{i_k})\boldsymbol{\Pi}_{(i_1 i_k)}\right)\right)\right) &= \\
\varepsilon^2\operatorname{Tr}\left((\boldsymbol{E}_{i_1} - \boldsymbol{E}_{i_k})\left(\left(\boldsymbol{\Pi}\boldsymbol{S}_p\tilde{\boldsymbol{\Pi}}'\right) \circ \left((\lambda_{i_1}\boldsymbol{E}_{i_k} - \lambda_{i_k}\boldsymbol{E}_{i_1})\right)\right)\right) &= \\
\varepsilon^2\operatorname{Tr}\left(\left(\boldsymbol{\Pi}\boldsymbol{S}_p\tilde{\boldsymbol{\Pi}}'\right) \circ \left((\boldsymbol{E}_{i_1} - \boldsymbol{E}_{i_k})(\lambda_{i_1}\boldsymbol{E}_{i_k} - \lambda_{i_k}\boldsymbol{E}_{i_1})\right)\right) &=
\end{aligned}
$$

$$-\varepsilon^2 \operatorname{Tr}\left(\left(\mathbf{\Pi} \boldsymbol{S}_p \tilde{\mathbf{\Pi}}'\right) \circ (\lambda_{i_1} \boldsymbol{E}_{i_k} + \lambda_{i_k} \boldsymbol{E}_{i_1})\right) =$$
$$-\varepsilon^2 \operatorname{Tr}\left(\left(\mathbf{\Pi} \boldsymbol{S}_p \tilde{\mathbf{\Pi}}'\right) \circ (\lambda_{i_1} \boldsymbol{E}_{i_k} + \lambda_{i_k} \boldsymbol{E}_{i_1})\right).$$

Now, note that in both cases the matrices that are multiplied elementwise with $\mathbf{\Pi} \boldsymbol{S}_p \tilde{\mathbf{\Pi}}'$ are diagonal and hence, we only need to look at diagonal elements of $\mathbf{\Pi} \boldsymbol{S}_p \tilde{\mathbf{\Pi}}'$. Moreover,

$$\mathbf{\Pi} \boldsymbol{S}_p \tilde{\mathbf{\Pi}}' = \mathbf{\Pi}_{(i_1 i_k)} \mathbf{\Pi}_{(i_k i_2)} \cdots \mathbf{\Pi}_{(i_k i_{k-1})} \hat{\mathbf{\Pi}} \boldsymbol{S}_p \hat{\mathbf{\Pi}}' \mathbf{\Pi}_{(i_k i_{k-1})} \cdots \mathbf{\Pi}_{(i_k i_2)},$$

where, $i_1 \cdots i_k$ are fixed points of permutation corresponding to $\hat{\mathbf{\Pi}}$ so $\hat{\mathbf{\Pi}} \boldsymbol{S}_p \hat{\mathbf{\Pi}}'$ has the same values at diagonal positions i_1 and i_k as the original matrix $\boldsymbol{S}_p$. The only permutation that is only on the left side is $\mathbf{\Pi}_{(i_1 i_k)}$ which exchanges the i_1 and i_k rows of $\boldsymbol{S}_p$. Since $\boldsymbol{S}_p$ is such that the elements at each row before the diagonal element are the same and $i_k > i_1$, we have the i_1 and i_k diagonal elements of $\mathbf{\Pi} \boldsymbol{S}_p \tilde{\mathbf{\Pi}}'$ have the same value. Let that value be denoted as s. Then the sum of the above two equations yields $m(\lambda_{i_1} + \lambda_{i_k}) - m(\lambda_{i_1} + \lambda_{i_k}) = 0$, as claimed.

3 Derivatives of the Loss function

3.1 First and Second Order Fréchet Derivative

In order to derive and analyze the critical points of the cost function which is a real-valued function of matrices we use the first and second order Fréchet derivatives as described in chapter 4 of Zeidler (1995). For a function $f : \mathbb{R}^{n\times m} \to \mathbb{R}$ the first order Fréchet derivative at the point $\boldsymbol{A} \in \mathbb{R}^{n\times m}$ is a linear functional $df(\boldsymbol{A}) : \mathbb{R}^{n\times m} \to \mathbb{R}$ such that

$$\lim_{\boldsymbol{V}\to 0} \frac{|f(\boldsymbol{A}+\boldsymbol{V}) - f(\boldsymbol{A}) - df(\boldsymbol{A})\boldsymbol{V}|}{\|\boldsymbol{V}\|_F} = 0,$$

where we used the shorthand $df(\boldsymbol{A})\boldsymbol{V} \equiv (df(\boldsymbol{A}))(\boldsymbol{V})$. Similarly, the 2nd derivative is a bilinear functional $d^2 f(\boldsymbol{A}) : \mathbb{R}^{n\times m} \times \mathbb{R}^{n\times m} \to \mathbb{R}$ such that

$$\lim_{\boldsymbol{V}\to 0} \frac{|df(\boldsymbol{A}+\boldsymbol{V})K - df(\boldsymbol{A})K - d^2 f(\boldsymbol{A})\boldsymbol{V}K|}{\|\boldsymbol{V}\|_F} = 0,$$

for all $\|K\|_F \leq 1$, where again $d^2 f(\boldsymbol{A})\boldsymbol{V}K \equiv (d^2 f(\boldsymbol{A}))(\boldsymbol{V}, K)$. The generalized Taylor formula then becomes:

$$f(\boldsymbol{A}+\boldsymbol{V}) = f(\boldsymbol{A}) + df(\boldsymbol{A})\boldsymbol{V} + \frac{1}{2} d^2 f(\boldsymbol{A})\boldsymbol{V}^2 + o(\|\boldsymbol{V}\|^2),$$

Moreover, we derive functions $\nabla f : \mathbb{R}^{n\times m} \to \mathbb{R}^{n\times m}$ and $\boldsymbol{H}(\boldsymbol{A}) : \mathbb{R}^{n\times m} \to \mathbb{R}^{n\times m}$ such that $df(\boldsymbol{A})\boldsymbol{V} = \langle \nabla f(\boldsymbol{A}), \boldsymbol{V}\rangle_F$ and $d^2 f(\boldsymbol{A})\boldsymbol{V}^2 = \langle \boldsymbol{H}(\boldsymbol{A})\boldsymbol{V}, \boldsymbol{V}\rangle_F$, where again $H(\boldsymbol{A})\boldsymbol{V} \equiv H(\boldsymbol{A})(\boldsymbol{V})$. Then clearly, $\boldsymbol{A} \in \mathbb{R}^{n\times m}$is a critical point of f iff $\nabla f(\boldsymbol{A}) = 0$ and for such $\boldsymbol{A}$s the sign of the bilinear form $\langle \boldsymbol{H}(\boldsymbol{A})\boldsymbol{V}, \boldsymbol{V}\rangle$over *directions* $\boldsymbol{V}$ determines the type of the critical point.

Extending the generalized Taylor theorem of Zeidler (1995), the second order Taylor expansion for the loss $L(\boldsymbol{A}, \boldsymbol{B})$ is then given by

$$L(\boldsymbol{A}+\boldsymbol{V}, \boldsymbol{B}+\boldsymbol{W}) - L(\boldsymbol{A}, \boldsymbol{B}) = d_{\boldsymbol{A}} L(\boldsymbol{A}, \boldsymbol{B})\boldsymbol{V} + d_{\boldsymbol{B}} L(\boldsymbol{A}, \boldsymbol{B})\boldsymbol{W} + \frac{1}{2} d^2_{\boldsymbol{A}} L(\boldsymbol{A}, \boldsymbol{B})\boldsymbol{V}^2$$

$$+d_{\boldsymbol{AB}}L(\boldsymbol{A},\boldsymbol{B})\boldsymbol{V}\boldsymbol{W}+\frac{1}{2}d^2_{\boldsymbol{B}}L(\boldsymbol{A},\boldsymbol{B})\boldsymbol{W}^2+R_{\boldsymbol{V},\boldsymbol{W}}(\boldsymbol{A},\boldsymbol{B}), \tag{53}$$

where, if $\|\boldsymbol{V}\|_F, \|\boldsymbol{W}\|_F = O(\varepsilon)$ then $\|R(\boldsymbol{V},\boldsymbol{W})\| = O(\varepsilon^3)$. Clearly, as at critical points where $d_{\boldsymbol{A}}L(\boldsymbol{A},\boldsymbol{B})\boldsymbol{V}+d_{\boldsymbol{B}}L(\boldsymbol{A},\boldsymbol{B})\boldsymbol{W}=0$, as $\varepsilon\to 0$ we have $R_{\boldsymbol{V},\boldsymbol{W}}(\boldsymbol{A},\boldsymbol{B})\to 0$ and the sign of the sum of the second order partial Fréchet derivatives determines the type of the critical point very much similar to second partial derivative test for two variable functions. However, here for local minima we have to show the sign is positive in all directions and for saddle points have to show the sign is positive in some directions and negative at least in on direction. Finally, note that the smoothness of the loss entails that Fréchet derivative and directional derivative (Gateaux) both exist and (foregoing some subtleties in definition) are the same.

3.2 First and Second Order Derivative of the Loss wrt to $\boldsymbol{B}$

Lemma 5. *The first and second (partial Fréchet) derivative of the loss $L(\boldsymbol{A},\boldsymbol{B})$ wrt to $\boldsymbol{B}$ is derived as follows.*

$$d_{\boldsymbol{B}}L(\boldsymbol{A},\boldsymbol{B})\boldsymbol{W}= -2\operatorname{Tr}\left(\boldsymbol{W}'\left(\boldsymbol{T}_p\boldsymbol{A}'\boldsymbol{\Sigma}_{yx}-\left(\boldsymbol{S}_p\circ(\boldsymbol{A}'\boldsymbol{A})\right)\boldsymbol{B}\boldsymbol{\Sigma}_{xx}\right)\right) \tag{54}$$

$$= -2\langle \boldsymbol{T}_p\boldsymbol{A}'\boldsymbol{\Sigma}_{yx}-\left(\boldsymbol{S}_p\circ(\boldsymbol{A}'\boldsymbol{A})\right)\boldsymbol{B}\boldsymbol{\Sigma}_{xx},\boldsymbol{W}\rangle_F. \tag{55}$$

$$d^2_{\boldsymbol{B}^2}L(\boldsymbol{A},\boldsymbol{B})\boldsymbol{W}^2 = 2\langle\left(\boldsymbol{S}_p\circ(\boldsymbol{A}'\boldsymbol{A})\right)\boldsymbol{W}\boldsymbol{\Sigma}_{xx},\boldsymbol{W}\rangle_F = 2\operatorname{Tr}\left(\boldsymbol{W}'\left(\boldsymbol{S}_p\circ(\boldsymbol{A}'\boldsymbol{A})\right)\boldsymbol{W}\boldsymbol{\Sigma}_{xx}\right). \tag{56}$$

Proof. Directly compute

$$\begin{aligned}
L(\boldsymbol{A},\boldsymbol{B}+\boldsymbol{W}) &= \sum_{i=1}^{p}\|\boldsymbol{Y}-\boldsymbol{A}\boldsymbol{I}_{i;p}(\boldsymbol{B}+\boldsymbol{W})\boldsymbol{X}\|_F^2\\
&= \sum_{i=1}^{p}\langle \boldsymbol{Y}-\boldsymbol{A}\boldsymbol{I}_{i;p}(\boldsymbol{B}+\boldsymbol{W})\boldsymbol{X},\boldsymbol{Y}-\boldsymbol{A}\boldsymbol{I}_{i;p}(\boldsymbol{B}+\boldsymbol{W})\boldsymbol{X}\rangle_F\\
&= \sum_{i=1}^{p}\langle \boldsymbol{Y}-\boldsymbol{A}\boldsymbol{I}_{i;p}\boldsymbol{B}\boldsymbol{X},\boldsymbol{Y}-\boldsymbol{A}\boldsymbol{I}_{i;p}\boldsymbol{B}\boldsymbol{X}\rangle_F+\sum_{i=1}^{p}\langle \boldsymbol{Y}-\boldsymbol{A}\boldsymbol{I}_{i;p}\boldsymbol{B}\boldsymbol{X},-\boldsymbol{A}\boldsymbol{I}_{i;p}\boldsymbol{W}\boldsymbol{X}\rangle_F\\
&+\sum_{i=1}^{p}\langle -\boldsymbol{A}\boldsymbol{I}_{i;p}\boldsymbol{W}\boldsymbol{X},\boldsymbol{Y}-\boldsymbol{A}\boldsymbol{I}_{i;p}\boldsymbol{B}\boldsymbol{X}\rangle_F+\sum_{i=1}^{p}\langle -\boldsymbol{A}\boldsymbol{I}_{i;p}\boldsymbol{W}\boldsymbol{X},-\boldsymbol{A}\boldsymbol{I}_{i;p}\boldsymbol{W}\boldsymbol{X}\rangle_F\\
&= L(\boldsymbol{A},\boldsymbol{B})-\sum_{i=1}^{p}2\langle \boldsymbol{Y}-\boldsymbol{A}\boldsymbol{I}_{i;p}\boldsymbol{B}\boldsymbol{X},\boldsymbol{A}\boldsymbol{I}_{i;p}\boldsymbol{W}\boldsymbol{X}\rangle+O(\|\boldsymbol{W}\|_F^2) \implies
\end{aligned}$$

$$\begin{aligned}
L(\boldsymbol{A},\boldsymbol{B}+\boldsymbol{W})-L(\boldsymbol{A},\boldsymbol{B}) &= -2\sum_{i=1}^{p}\langle \boldsymbol{Y}-\boldsymbol{A}\boldsymbol{I}_{i;p}\boldsymbol{B}\boldsymbol{X},\boldsymbol{A}\boldsymbol{I}_{i;p}\boldsymbol{W}\boldsymbol{X}\rangle_F+O(\|\boldsymbol{W}\|_F^2)\overset{\boldsymbol{W}\to 0}{\Longrightarrow}\\
d_{\boldsymbol{B}}L(\boldsymbol{A},\boldsymbol{B})\boldsymbol{W} &= -2\sum_{i=1}^{p}\operatorname{Tr}(\boldsymbol{X}'\boldsymbol{W}'\boldsymbol{I}_{i;p}\boldsymbol{A}'(\boldsymbol{Y}-\boldsymbol{A}\boldsymbol{I}_{i;p}\boldsymbol{B}\boldsymbol{X}))\\
&= -2\operatorname{Tr}\left(\boldsymbol{W}'\left(\left(\sum_{i=1}^{p}\boldsymbol{I}_{i;p}\right)\boldsymbol{A}'\boldsymbol{Y}\boldsymbol{X}'-\left(\sum_{i=1}^{p}\boldsymbol{I}_{i;p}\boldsymbol{A}'\boldsymbol{A}\boldsymbol{I}_{i;p}\right)\boldsymbol{B}\boldsymbol{X}\boldsymbol{X}'\right)\right)\\
&= -2\operatorname{Tr}\left(\boldsymbol{W}'\left(\boldsymbol{T}_p\boldsymbol{A}'\boldsymbol{Y}\boldsymbol{X}'-\left(\boldsymbol{S}_p\circ(\boldsymbol{A}'\boldsymbol{A})\right)\boldsymbol{B}\boldsymbol{X}\boldsymbol{X}'\right)\right),
\end{aligned}$$

which can be written as the given form. For the second derivative wrt $\boldsymbol{B}$ we have

$$
\begin{aligned}
d_{\boldsymbol{B}} L(\boldsymbol{A},\boldsymbol{B})\boldsymbol{W} &= -2\langle \boldsymbol{T}_p\boldsymbol{A}'\boldsymbol{\Sigma}_{yx} - (\boldsymbol{S}_p \circ (\boldsymbol{A}'\boldsymbol{A}))\,\boldsymbol{B}\boldsymbol{\Sigma}_{xx}, \boldsymbol{W}\rangle_F \implies \\
d_{\boldsymbol{B}} L(\boldsymbol{A},\boldsymbol{B}+\bar{\boldsymbol{W}})\boldsymbol{W} &= -2\langle \boldsymbol{T}_p\boldsymbol{A}'\boldsymbol{\Sigma}_{yx} - (\boldsymbol{S}_p \circ (\boldsymbol{A}'\boldsymbol{A}))\,(\boldsymbol{B}+\bar{\boldsymbol{W}})\boldsymbol{\Sigma}_{xx}, \boldsymbol{W}\rangle_F \\
&= -2\langle \boldsymbol{T}_p\boldsymbol{A}'\boldsymbol{\Sigma}_{yx} - (\boldsymbol{S}_p \circ (\boldsymbol{A}'\boldsymbol{A}))\,\boldsymbol{B}\boldsymbol{\Sigma}_{xx}, \boldsymbol{W}\rangle_F \\
&\quad + 2\langle (\boldsymbol{S}_p \circ (\boldsymbol{A}'\boldsymbol{A}))\,\bar{\boldsymbol{W}}\boldsymbol{\Sigma}_{xx}, \boldsymbol{W}\rangle_F \implies \\
d_{\boldsymbol{B}} L(\boldsymbol{A},\boldsymbol{B}+\bar{\boldsymbol{W}})\boldsymbol{W} - d_{\boldsymbol{B}} L(\boldsymbol{A},\boldsymbol{B})\boldsymbol{W} &= 2\langle (\boldsymbol{S}_p \circ (\boldsymbol{A}'\boldsymbol{A}))\,\bar{\boldsymbol{W}}\boldsymbol{\Sigma}_{xx}, \boldsymbol{W}\rangle_F,
\end{aligned}
$$

which by having $\bar{\boldsymbol{W}} \to 0$ results in the second order partial derivative. □

3.3 First and Second Order Derivative of the Loss wrt to $\boldsymbol{A}$

Lemma 6. *The first and second (partial Fréchet) derivative of the loss $L(\boldsymbol{A},\boldsymbol{B})$ wrt to $\boldsymbol{A}$ is derived as follows.*

$$
d_{\boldsymbol{A}} L(\boldsymbol{A},\boldsymbol{B})\boldsymbol{V} = -2\langle \boldsymbol{\Sigma}_{yx}\boldsymbol{B}'\boldsymbol{T}_p - \boldsymbol{A}\left(\boldsymbol{S}_p \circ (\boldsymbol{B}\boldsymbol{\Sigma}_{xx}\boldsymbol{B}')\right), \boldsymbol{V}\rangle_F, \tag{57}
$$

$$
d^2_{\boldsymbol{A}\boldsymbol{B}} L(\boldsymbol{A},\boldsymbol{B})\boldsymbol{V}\boldsymbol{W} = -2\langle \boldsymbol{\Sigma}_{yx}\boldsymbol{W}'\boldsymbol{T}_p - \boldsymbol{A}\left(\boldsymbol{S}_p \circ (\boldsymbol{B}\boldsymbol{\Sigma}_{xx}\boldsymbol{W}')\right) - \boldsymbol{A}\left(\boldsymbol{S}_p \circ (\boldsymbol{W}\boldsymbol{\Sigma}_{xx}\boldsymbol{B}')\right), \boldsymbol{V}\rangle_F, \tag{58}
$$

$$
d^2_{\boldsymbol{A}^2} L(\boldsymbol{A},\boldsymbol{B})\boldsymbol{V}^2 = 2\langle \boldsymbol{V}\left(\boldsymbol{S}_p \circ (\boldsymbol{B}\boldsymbol{\Sigma}_{xx}\boldsymbol{B}')\right), \boldsymbol{V}\rangle_F. \tag{59}
$$

Proof. Directly compute

$$
\begin{aligned}
L(\boldsymbol{A}+\boldsymbol{V},\boldsymbol{B}) &= \sum_{i=1}^{p}\langle \boldsymbol{Y} - (\boldsymbol{A}+\boldsymbol{V})\boldsymbol{I}_{i;p}\boldsymbol{B}\boldsymbol{X}, \boldsymbol{Y} - (\boldsymbol{A}+\boldsymbol{V})\boldsymbol{I}_{i;p}\boldsymbol{B}\boldsymbol{X}\rangle_F \\
&= \sum_{i=1}^{p}\langle \boldsymbol{Y} - \boldsymbol{A}\boldsymbol{I}_{i;p}\boldsymbol{B}\boldsymbol{X}, \boldsymbol{Y} - \boldsymbol{A}\boldsymbol{I}_{i;p}\boldsymbol{B}\boldsymbol{X}\rangle_F - \sum_{i=1}^{p}\langle \boldsymbol{Y} - \boldsymbol{A}\boldsymbol{I}_{i;p}\boldsymbol{B}\boldsymbol{X}, \boldsymbol{V}\boldsymbol{I}_{i;p}\boldsymbol{B}\boldsymbol{X}\rangle_F \\
&\quad + \sum_{i=1}^{p}\langle -\boldsymbol{V}\boldsymbol{I}_{i;p}\boldsymbol{B}\boldsymbol{X}, \boldsymbol{Y} - \boldsymbol{A}\boldsymbol{I}_{i;p}\boldsymbol{B}\boldsymbol{X}\rangle_F + \sum_{i=1}^{p}\langle -\boldsymbol{V}\boldsymbol{I}_{i;p}\boldsymbol{B}\boldsymbol{X}, -\boldsymbol{V}\boldsymbol{I}_{i;p}\boldsymbol{B}\boldsymbol{X}\rangle_F \\
&= L(\boldsymbol{A},\boldsymbol{B}) - \sum_{i=1}^{p} 2\langle \boldsymbol{Y} - \boldsymbol{A}\boldsymbol{I}_{i;p}\boldsymbol{B}\boldsymbol{X}, \boldsymbol{V}\boldsymbol{I}_{i;p}\boldsymbol{B}\boldsymbol{X}\rangle_F + \sum_{i=1}^{p}\langle \boldsymbol{V}\boldsymbol{I}_{i;p}\boldsymbol{B}\boldsymbol{X}, \boldsymbol{V}\boldsymbol{I}_{i;p}\boldsymbol{B}\boldsymbol{X}\rangle_F
\end{aligned}
$$

$$
\begin{aligned}
L(\boldsymbol{A}+\boldsymbol{V},\boldsymbol{B}) - L(\boldsymbol{A},\boldsymbol{B}) &= -\sum_{i=1}^{p} 2\langle \boldsymbol{Y} - \boldsymbol{A}\boldsymbol{I}_{i;p}\boldsymbol{B}\boldsymbol{X}, \boldsymbol{V}\boldsymbol{I}_{i;p}\boldsymbol{B}\boldsymbol{X}\rangle_F + O(\|\boldsymbol{V}\|_F^2) \overset{\boldsymbol{V}\to 0}{\Longrightarrow} \\
d_{\boldsymbol{A}} L(\boldsymbol{A},\boldsymbol{B})\boldsymbol{V} &= -\sum_{i=1}^{p} 2\langle \boldsymbol{Y} - \boldsymbol{A}\boldsymbol{I}_{i;p}\boldsymbol{B}\boldsymbol{X}, \boldsymbol{V}\boldsymbol{I}_{i;p}\boldsymbol{B}\boldsymbol{X}\rangle_F \\
&= -2\operatorname{Tr}(\boldsymbol{V}'(\boldsymbol{\Sigma}_{yx}\boldsymbol{B}'\sum_{i=1}^{p}\boldsymbol{I}_{i;p} - \boldsymbol{A}\sum_{i=1}^{p}\boldsymbol{I}_{i;p}\boldsymbol{B}\boldsymbol{\Sigma}_{xx}\boldsymbol{B}'\boldsymbol{I}_{i;p})) \implies \\
d_{\boldsymbol{A}} L(\boldsymbol{A},\boldsymbol{B})\boldsymbol{V} &= -2\langle \boldsymbol{\Sigma}_{yx}\boldsymbol{B}'\boldsymbol{T}_p - \boldsymbol{A}\left(\boldsymbol{S}_p \circ (\boldsymbol{B}\boldsymbol{\Sigma}_{xx}\boldsymbol{B}')\right), \boldsymbol{V}\rangle_F \implies \\
d_{\boldsymbol{A}} L(\boldsymbol{A}+\bar{\boldsymbol{V}},\boldsymbol{B})\boldsymbol{V} &= -2\langle \boldsymbol{\Sigma}_{yx}\boldsymbol{B}'\boldsymbol{T}_p - (\boldsymbol{A}+\bar{\boldsymbol{V}})\left(\boldsymbol{S}_p \circ (\boldsymbol{B}\boldsymbol{\Sigma}_{xx}\boldsymbol{B}')\right), \boldsymbol{V}\rangle_F \\
d_{\boldsymbol{A}} L(\boldsymbol{A}+\bar{\boldsymbol{V}},\boldsymbol{B})\boldsymbol{V} - d_{\boldsymbol{A}} L(\boldsymbol{A},\boldsymbol{B})\boldsymbol{V} &= 2\langle \bar{\boldsymbol{V}}\left(\boldsymbol{S}_p \circ (\boldsymbol{B}\boldsymbol{\Sigma}_{xx}\boldsymbol{B}')\right), \boldsymbol{V}\rangle_F \overset{\bar{\boldsymbol{V}}\to 0}{\Longrightarrow}
\end{aligned}
$$

$$
\begin{aligned}
d^2_{\boldsymbol{A}^2}L(\boldsymbol{A},\boldsymbol{B})(\boldsymbol{V},\bar{\boldsymbol{V}}) &= 2\langle \bar{\boldsymbol{V}}\left(\boldsymbol{S}_p \circ \left(\boldsymbol{B}\boldsymbol{\Sigma}_{xx}\boldsymbol{B}'\right)\right), \boldsymbol{V}\rangle_F \implies \\
d^2_{\boldsymbol{A}^2}L(\boldsymbol{A},\boldsymbol{B})\boldsymbol{V}^2 &= 2\langle \boldsymbol{V}\left(\boldsymbol{S}_p \circ \left(\boldsymbol{B}\boldsymbol{\Sigma}_{xx}\boldsymbol{B}'\right)\right), \boldsymbol{V}\rangle_F
\end{aligned}
$$

$$
\begin{aligned}
d_{\boldsymbol{A}}L(\boldsymbol{A},\boldsymbol{B}+\boldsymbol{W})\boldsymbol{V} =& -2\langle \boldsymbol{\Sigma}_{yx}(\boldsymbol{B}+\boldsymbol{W})'\boldsymbol{T}_p, \boldsymbol{V}\rangle_F \\
& -2\langle -\boldsymbol{A}\left(\boldsymbol{S}_p \circ \left((\boldsymbol{B}+\boldsymbol{W})\boldsymbol{\Sigma}_{xx}(\boldsymbol{B}+\boldsymbol{W})'\right)\right), \boldsymbol{V}\rangle_F \\
& -2\langle \boldsymbol{\Sigma}_{yx}\boldsymbol{B}'\boldsymbol{T}_p - \boldsymbol{A}\left(\boldsymbol{S}_p \circ \left(\boldsymbol{B}\boldsymbol{\Sigma}_{xx}\boldsymbol{B}'\right)\right), \boldsymbol{V}\rangle_F \\
=& d_{\boldsymbol{A}}L(\boldsymbol{A},\boldsymbol{B})\boldsymbol{V} - 2\langle \boldsymbol{\Sigma}_{yx}\boldsymbol{W}'\boldsymbol{T}_p, \boldsymbol{V}\rangle_F \\
& -2\langle -\boldsymbol{A}\left(\boldsymbol{S}_p \circ \left(\boldsymbol{B}\boldsymbol{\Sigma}_{xx}\boldsymbol{W}'\right)\right) - \boldsymbol{A}\left(\boldsymbol{S}_p \circ \left(\boldsymbol{W}\boldsymbol{\Sigma}_{xx}\boldsymbol{B}'\right)\right), \boldsymbol{V}\rangle_F + O(\|\boldsymbol{W}\|_F^2) \implies
\end{aligned}
$$

$$
\begin{aligned}
d_{\boldsymbol{A}}L(\boldsymbol{A},\boldsymbol{B}+\boldsymbol{W})\boldsymbol{V} - d_{\boldsymbol{A}}L(\boldsymbol{A},\boldsymbol{B})\boldsymbol{V} =& -2\langle \boldsymbol{\Sigma}_{yx}\boldsymbol{W}'\boldsymbol{T}_p, \boldsymbol{V}\rangle_F \\
& -2\langle -\boldsymbol{A}\left(\boldsymbol{S}_p \circ \left(\boldsymbol{B}\boldsymbol{\Sigma}_{xx}\boldsymbol{W}'\right)\right) - \boldsymbol{A}\left(\boldsymbol{S}_p \circ \left(\boldsymbol{W}\boldsymbol{\Sigma}_{xx}\boldsymbol{B}'\right)\right), \boldsymbol{V}\rangle_F \\
& + O(\|\boldsymbol{W}\|_F^2) \overset{\boldsymbol{W}\to 0}{\implies}
\end{aligned}
$$

$$
d^2_{\boldsymbol{AB}}L(\boldsymbol{A},\boldsymbol{B})\boldsymbol{V}\boldsymbol{W} = -2\langle \boldsymbol{\Sigma}_{yx}\boldsymbol{W}'\boldsymbol{T}_p - \boldsymbol{A}\left(\boldsymbol{S}_p \circ \left(\boldsymbol{B}\boldsymbol{\Sigma}_{xx}\boldsymbol{W}'\right)\right) - \boldsymbol{A}\left(\boldsymbol{S}_p \circ \left(\boldsymbol{W}\boldsymbol{\Sigma}_{xx}\boldsymbol{B}'\right)\right), \boldsymbol{V}\rangle_F.
$$

□

References

Baldi, P. and Hornik, K. Neural networks and principal component analysis: Learning from examples without local minima. *Neural networks*, 2(1):53–58, 1989.

Horn, R. and Johnson, C. *Matrix Analysis*. Cambridge University Press, 2012. ISBN 9781139788885.

Zeidler, E. *Applied Functional Analysis: Main Principles and Their Applications*. Applied Mathematical Sciences. Springer New York, 1995. ISBN 9780387944227.