1 Theory

1.1 General matrix statements

1. Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ be an orthogonal matrix, i.e., all columns and rows sum to 1 and they are pairwise orthogonal. Writing the matrix as

$$\mathbf{A} = \begin{bmatrix}\mathbf{v}_1 & \dots & \mathbf{v}_n\end{bmatrix} = \begin{bmatrix}\mathbf{u}_1 & \dots & \mathbf{u}_n\end{bmatrix}^\mathsf{T},$$

we can write these conditions as

$$\mathbf{v}_i^\mathsf{T}\mathbf{v}_j = \mathbf{u}_i^\mathsf{T}\mathbf{u}_j = \begin{cases}1, & i = j,\\ 0, & i \neq j.\end{cases}$$

A matrix inverse $\mathbf{A}^{-1}$ is a matrix for which $\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$. To show that $\mathbf{A}^\mathsf{T}$ is a matrix inverse, we need to show these two properties.

First,

$$\mathbf{A}\mathbf{A}^\mathsf{T} = \begin{bmatrix}\mathbf{u}_1^\mathsf{T}\\ \vdots \\ \mathbf{u}_n^\mathsf{T}\end{bmatrix}\begin{bmatrix}\mathbf{u}_1 & \dots & \mathbf{u}_n\end{bmatrix} = \begin{bmatrix}\mathbf{u}_1^\mathsf{T}\mathbf{u}_1 & \mathbf{u}_1^\mathsf{T}\mathbf{u}_2 & \dots & \mathbf{u}_1^\mathsf{T}\mathbf{u}_n\\ \mathbf{u}_2^\mathsf{T}\mathbf{u}_1 & \mathbf{u}_2^\mathsf{T}\mathbf{u}_2 & \dots & \mathbf{u}_2^\mathsf{T}\mathbf{u}_n\\ \vdots & \vdots & \ddots & \vdots\\ \mathbf{u}_n^\mathsf{T}\mathbf{u}_1 & \mathbf{u}_n^\mathsf{T}\mathbf{u}_2 & \dots & \mathbf{u}_n^\mathsf{T}\mathbf{u}_n\end{bmatrix} = \mathbf{I}. \tag{1}$$

For the second equality, we write

$$\mathbf{A}^\mathsf{T}\mathbf{A} = \begin{bmatrix}\mathbf{v}_1^\mathsf{T}\\ \vdots \\ \mathbf{v}_n^\mathsf{T}\end{bmatrix}\begin{bmatrix}\mathbf{v}_1 & \dots & \mathbf{v}_n\end{bmatrix},$$

from which we can deduce $\mathbf{A}^\mathsf{T}\mathbf{A} = \mathbf{I}$ because the equation is the same as (1), only the vectors $\mathbf{u}_i$ are replaced with $\mathbf{v}_i$, but their dot products are defined the same.

We have shown that $\mathbf{A}\mathbf{A}^\mathsf{T} = \mathbf{A}^\mathsf{T}\mathbf{A} = \mathbf{I}$ for orthogonal matrices and thus

$$\mathbf{A}^{-1} = \mathbf{A}^\mathsf{T}.$$

■

2. To show that $(\mathbf{AB})^\mathsf{T} = \mathbf{B}^\mathsf{T}\mathbf{A}^\mathsf{T}$, we show that all elements of both matrix products are the same. Elements of the matrix product on the left-hand side are

$$(\mathbf{AB})^\mathsf{T}_{ij} = (\mathbf{AB})_{ji} = \mathbf{A}_j^\mathsf{T}\mathbf{B}^{(i)},$$

where we use the bottom index to refer to the matrix row and the upper one to refer to the column. Similarly, we compute

$$(\mathbf{B}^\mathsf{T}\mathbf{A}^\mathsf{T})_{ij} = (\mathbf{B}^\mathsf{T})_i^\mathsf{T}(\mathbf{A}^\mathsf{T})^{(j)} = \left(\mathbf{B}^{(i)}\right)^\mathsf{T}\mathbf{A}_j = \mathbf{A}_j^\mathsf{T}\mathbf{B}^{(i)},$$

where we just reversed the elements of the dot product in the last equality. This finishes the proof.

■

3. From the matrix product $\mathbf{ABC}$, we can obtain sizes of all matrices: $\mathbf{A} \in \mathbb{R}^{n\times l}$, $\mathbf{B} \in \mathbb{R}^{l\times m}$ and $\mathbf{C} \in \mathbb{R}^{m\times n}$. From the trace definition, we write

$$\operatorname{tr}(\mathbf{ABC}) = \sum_{i=1}^{n} (\mathbf{ABC})_{ii} = \sum_{i=1}^{n}\sum_{j=1}^{m} (\mathbf{AB})_{ij} c_{ji} = \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{l} a_{ik} b_{kj} c_{ji}. \tag{2}$$

By rearranging summations and matrix elements, we can obtain

$$\operatorname{tr}(\mathbf{ABC}) = \ldots = \sum_{k=1}^{l}\sum_{i=1}^{n}\sum_{j=1}^{m} b_{kj} c_{ji} a_{ik} = \operatorname{tr}(\mathbf{BCA}).$$

To obtain the last equality, we used the logic from (2) in reverse. Similarly, we obtain

$$\operatorname{tr}(\mathbf{ABC}) = \ldots = \sum_{j=1}^{m}\sum_{k=1}^{l}\sum_{i=1}^{n} c_{ji} a_{ik} b_{kj} = \operatorname{tr}(\mathbf{CAB}),$$

proving

$$\operatorname{tr}(\mathbf{ABC}) = \operatorname{tr}(\mathbf{BCA}) = \operatorname{tr}(\mathbf{CAB}).$$

■

1.2 Eigendecomposition of symmetric squared real matrices

Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ be a symmetric matrix and λ_i its eigenvalues with corresponding eigenvectors $\mathbf{v}_i$. Because the matrix is symmetric,

$$\lambda_i \mathbf{v}_i^\top \mathbf{v}_j = (\lambda_i \mathbf{v}_i)^\top \mathbf{v}_j = (\mathbf{A}\mathbf{v}_i)^\top \mathbf{v}_j = \mathbf{v}_i^\top \mathbf{A}^\top \mathbf{v}_j = \mathbf{v}_i^\top \mathbf{A} \mathbf{v}_j = \lambda_j \mathbf{v}_i^\top \mathbf{v}_j,$$

and $(\lambda_i - \lambda_j)\mathbf{v}_i^\top \mathbf{v}_j = 0$. From this, we see that the eigenspaces with distinct eigenvalues have to be orthogonal. For their basis, we can find orthonormal vectors, which will also be mutually orthogonal.

Combining the eigenvector equations $\mathbf{A}\mathbf{v}_i = \lambda_i \mathbf{v}_i$ for all vectors, we get

$$\mathbf{A}\begin{bmatrix}\mathbf{v}_1 & \ldots & \mathbf{v}_n\end{bmatrix} = \begin{bmatrix}\lambda_1\mathbf{v}_1 & \ldots & \lambda_n\mathbf{v}_n\end{bmatrix} = \begin{bmatrix}\mathbf{v}_1 & \ldots & \mathbf{v}_n\end{bmatrix}\begin{bmatrix}\lambda_1 & & \\ & \ddots & \\ & & \lambda_n\end{bmatrix}.$$

By defining $\mathbf{V} = \begin{bmatrix}\mathbf{v}_1 & \ldots & \mathbf{v}_n\end{bmatrix}$ and $\mathbf{\Lambda} = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$, we can write

$$\mathbf{AV} = \mathbf{V\Lambda}.$$

Because $\mathbf{V}$ is orthogonal, we can multiply this equation from the right with its transpose and get

$$\mathbf{A} = \mathbf{V\Lambda V}^\top,$$

which is the eigendecomposition of $\mathbf{A}$.

■

1.3 Gram matrix

For the matrix $\mathbf{X} \in \mathbb{R}^{n \times m}$, the Gram matrix is defined as $\mathbf{X}^\mathsf{T}\mathbf{X} \in \mathbb{R}^{m \times m}$.

1. The Gram matrix is symmetric:

$$\left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^\mathsf{T} = \mathbf{X}^\mathsf{T}\left(\mathbf{X}^\mathsf{T}\right)^\mathsf{T} = \mathbf{X}^\mathsf{T}\mathbf{X}.$$

$\blacksquare$

2. Let $(\mathbf{X}^\mathsf{T}\mathbf{X})\mathbf{v} = \lambda\mathbf{v}$. Now, we compute

$$\|\mathbf{X}\mathbf{v}\|^2 = (\mathbf{X}\mathbf{v})^\mathsf{T}(\mathbf{X}\mathbf{v}) = \mathbf{v}^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{v} = \mathbf{v}^\mathsf{T}(\mathbf{X}^\mathsf{T}\mathbf{X})\mathbf{v} = \mathbf{v}^\mathsf{T}\lambda\mathbf{v} = \lambda\|\mathbf{v}\|^2.$$

Factoring out the eigenvalue, we obtain the inequality:

$$\lambda = \frac{\|\mathbf{X}\mathbf{v}\|^2}{\|\mathbf{v}\|^2} \geq 0.$$

It follows from the fact that the norm is non-negative and the norm of the eigenvectors is positive. Because all eigenvalues of the Gram matrix are non-negative, it is positive semi-definite.

$\blacksquare$

3. Again, let $(\mathbf{X}^\mathsf{T}\mathbf{X})\mathbf{v} = \lambda\mathbf{v}$. We can check that $\mathbf{u} = \mathbf{X}\mathbf{v}$ is an eigenvector of $\mathbf{X}\mathbf{X}^\mathsf{T}$,

$$\mathbf{X}\mathbf{X}^\mathsf{T}(\mathbf{X}\mathbf{v}) = \mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})\mathbf{v} = \mathbf{X}\lambda\mathbf{v} = \lambda(\mathbf{X}\mathbf{v}),$$

with the same eigenvalue λ. This holds for non-zero eigenvalues because $\mathbf{u}$ has to be non-zero and matrices $\mathbf{X}^\mathsf{T}\mathbf{X}$ and $\mathbf{X}$ have the same nullspace:

$$\begin{aligned} \mathbf{X}\mathbf{v} = \mathbf{0} &\implies (\mathbf{X}^\mathsf{T}\mathbf{X})\mathbf{v} = \mathbf{X}^\mathsf{T}\mathbf{0} = \mathbf{0}, \\ (\mathbf{X}^\mathsf{T}\mathbf{X})\mathbf{v} = \mathbf{0} &\implies \mathbf{v}^\mathsf{T}(\mathbf{X}^\mathsf{T}\mathbf{X})\mathbf{v} = \|\mathbf{X}\mathbf{v}\|^2 = 0. \end{aligned}$$

$\blacksquare$

1.4 Filling a gap in our Courant-Fisher theorem proof

Let $\mathcal{S}_1, \mathcal{S}_2 \subseteq \mathbb{R}^n$ be two subspaces. We define the basis of their intersection as

$$\mathcal{B}_{\mathcal{S}_1 \cap \mathcal{S}_2} = \{\mathbf{w}_1, \ldots, \mathbf{w}_k\},$$

where $k = \dim(\mathcal{S}_1 \cap \mathcal{S}_2)$. Since the intersection only spans the vectors in common, we need to expand its basis to obtain the basis of the two subspaces:

$$\begin{aligned} \mathcal{B}_{\mathcal{S}_1} &= \{\mathbf{w}_1, \ldots, \mathbf{w}_k, \mathbf{v}_{k+1}, \ldots, \mathbf{v}_l\}, \\ \mathcal{B}_{\mathcal{S}_2} &= \{\mathbf{w}_1, \ldots, \mathbf{w}_k, \mathbf{u}_{k+1}, \ldots, \mathbf{u}_m\}. \end{aligned}$$

We also used $l = \dim \mathcal{S}_1$ and $m = \dim \mathcal{S}_2$.

From the inclusion-exclusion principle, we have

$$|\mathcal{B}_{\mathcal{S}_1 \cup \mathcal{S}_2}| = |\mathcal{B}_{\mathcal{S}_1}| + |\mathcal{B}_{\mathcal{S}_2}| - |\mathcal{B}_{\mathcal{S}_1 \cap \mathcal{S}_2}|,$$

and thus

$$\dim(\mathcal{S}_1 \cup \mathcal{S}_2) = \dim \mathcal{S}_1 + \dim \mathcal{S}_2 - \dim(\mathcal{S}_1 \cap \mathcal{S}_2).$$

The union cannot span more than $\mathbb{R}^n$, so we have

$$\dim \mathcal{S}_1 + \dim \mathcal{S}_2 - \dim(\mathcal{S}_1 \cap \mathcal{S}_2) \leq n,$$

and

$$\dim \mathcal{S}_1 + \dim \mathcal{S}_2 - n \leq \dim(\mathcal{S}_1 \cap \mathcal{S}_2).$$

Finally, if $\dim \mathcal{S}_1 + \dim \mathcal{S}_2 > n$, we get

$$\dim(\mathcal{S}_1 \cap \mathcal{S}_2) \geq 1.$$

$\blacksquare$

1.5 Vector and matrix derivatives

1. To obtain the derivative of the dot product, we first compute the derivative with respect to a single vector element:

$$\frac{\partial}{\partial x_i} \mathbf{a}^\mathsf{T} \mathbf{x} = \frac{\partial}{\partial x_i}(a_1 x_1 + \ldots + a_n x_n) = a_i.$$

From the definition of the vector derivative, we get

$$\frac{\partial}{\partial \mathbf{x}} \mathbf{a}^\mathsf{T} \mathbf{x} = \begin{bmatrix} \frac{\partial}{\partial x_1} \mathbf{a}^\mathsf{T} \mathbf{x} & \ldots & \frac{\partial}{\partial x_n} \mathbf{a}^\mathsf{T} \mathbf{x} \end{bmatrix} = \begin{bmatrix} a_1 & \ldots & a_n \end{bmatrix} = \mathbf{a}.$$

$\blacksquare$

2. We approach similarly in the next case:

$$\mathbf{a}^\mathsf{T} \mathbf{X} \mathbf{b} = \begin{bmatrix} a_1 & \ldots & a_m \end{bmatrix} \begin{bmatrix} x_{11} & \ldots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{m1} & \ldots & x_{mn} \end{bmatrix} \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix} = \sum_{i=1}^{m} \sum_{j=1}^{n} a_i b_j x_{ij}.$$

From the matrix derivative definition, we get

$$\frac{\partial}{\partial \mathbf{X}} \mathbf{a}^\mathsf{T} \mathbf{X} \mathbf{b} = \left[\frac{\partial}{\partial x_{ij}} \sum_{i=1}^{m} \sum_{j=1}^{n} a_i b_j x_{ij} \right]_{ij} = [a_i b_j]_{ij} = \mathbf{a} \mathbf{b}^\mathsf{T}.$$

$\blacksquare$

3. For the matrix $\mathbf{X} \in \mathbb{R}^{n \times n}$, the trace of its square is the sum of dot products of corresponding rows and columns:

$$\mathrm{tr}(\mathbf{X}^2) = \sum_{i=1}^{n} \sum_{j=1}^{m} x_{ij} x_{ji}.$$

Each matrix element appears twice in the summation, meaning:

$$\frac{\partial}{\partial x_{ij}} \operatorname{tr}\left(\mathbf{X}^2\right) = \frac{\partial}{\partial x_{ij}}(x_{ij}x_{ji} + x_{ji}x_{ij}) = \frac{\partial}{\partial x_{ij}} 2x_{ij}x_{ji} = 2x_{ji}.$$

It follows that

$$\frac{\partial}{\partial \mathbf{X}} \operatorname{tr}\left(\mathbf{X}^2\right) = 2\mathbf{X}^\mathsf{T}.$$

■

4. From the product $\mathbf{AXBX}^\mathsf{T}\mathbf{C}$, we can obtain the matrix sizes: $\mathbf{A} \in \mathbb{R}^{n\times m}$, $\mathbf{B} \in \mathbb{R}^{l\times l}$, $\mathbf{C} \in \mathbb{R}^{m\times n}$ and $\mathbf{X} \in \mathbb{R}^{m\times l}$. First, we write out the expression we are differentiating:

$$\operatorname{tr}\left(\mathbf{AXBX}^\mathsf{T}\mathbf{C}\right) = \sum_{i=1}^{n} (\mathbf{AXBX}^\mathsf{T}\mathbf{C})_{ii} = \sum_{i=1}^{n}\sum_{j=1}^{l} (\mathbf{AXB})_{ij}(\mathbf{X}^\mathsf{T}\mathbf{C})_{ji}.$$

Using

$$(\mathbf{AXB})_{ij} = \sum_{k=1}^{l} (\mathbf{AX})_{ik} b_{kj} = \sum_{k=1}^{l}\sum_{p=1}^{m} a_{ip}x_{pk}b_{kj},$$

and

$$(\mathbf{X}^\mathsf{T}\mathbf{C})_{ij} = \sum_{r=1}^{m} x_{ri}c_{rj},$$

we get

$$\operatorname{tr}\left(\mathbf{AXBX}^\mathsf{T}\mathbf{C}\right) = \sum_{i=1}^{n}\sum_{j=1}^{l}\sum_{k=1}^{l}\sum_{p=1}^{m}\sum_{r=1}^{m} a_{ip}x_{pk}b_{kj}x_{rj}c_{ri}.$$

When we differentiate this expression with respect to a single matrix element $x_{i_0j_0}$, we can see that the only non-zero expressions we get are for $(p,k) = (i_0, j_0)$ and $(r,j) = (i_0, j_0)$:

$$\begin{aligned}\frac{\partial}{\partial x_{i_0j_0}} \operatorname{tr}\left(\mathbf{AXBX}^\mathsf{T}\mathbf{C}\right) &= \frac{\partial}{\partial x_{i_0j_0}} \sum_{i,j,k,p,r} a_{ip}x_{pk}b_{kj}x_{rj}c_{ri} = \sum_{i,j,k,p,r} a_{ip}b_{kj}c_{ri}\frac{\partial}{\partial x_{i_0j_0}} x_{pk}x_{rj} \\ &= \sum_{i,j,r} a_{ii_0}b_{j_0j}c_{ri}x_{rj} + \sum_{i,k,p} a_{ip}b_{kj_0}c_{i_0i}x_{pk}.\end{aligned}$$

By reordering summations and matrix elements, we can obtain the matrix expression:

$$\begin{aligned}\frac{\partial}{\partial x_{i_0j_0}} \operatorname{tr}\left(\mathbf{AXBX}^\mathsf{T}\mathbf{C}\right) &= \sum_{r=1}^{m}\sum_{i=1}^{n} a_{ii_0}c_{ri}\sum_{j=1}^{l} x_{rj}b_{j_0j} + \sum_{p=1}^{m}\sum_{i=1}^{n} c_{i_0i}a_{ip}\sum_{k=1}^{l} x_{pk}b_{kj_0} \\ &= \sum_{r=1}^{m} (\mathbf{A}^\mathsf{T}\mathbf{C}^\mathsf{T})_{i_0r}(\mathbf{XB}^\mathsf{T})_{rj_0} + \sum_{p=1}^{m} (\mathbf{CA})_{i_0p}(\mathbf{XB})_{pj_0} \\ &= (\mathbf{A}^\mathsf{T}\mathbf{C}^\mathsf{T}\mathbf{XB}^\mathsf{T})_{i_0j_0} + (\mathbf{CAXB})_{i_0j_0},\end{aligned}$$

meaning that

$$\frac{\partial}{\partial \mathbf{X}} \operatorname{tr}\left(\mathbf{AXBX}^\mathsf{T}\mathbf{C}\right) = \mathbf{A}^\mathsf{T}\mathbf{C}^\mathsf{T}\mathbf{XB}^\mathsf{T} + \mathbf{CAXB}.$$

■