Useful Knowledge
In
Matrix Computation

Runze Mao

June 8, 2017

1 Formalism

Name	Meaning
$A_{(j)}$	The j-th row vector of matrix A
A_j	The j-th column vector of matrix A
A_{ij}	The element at i-th row, j-th column of A
$\|A\|_{\mathcal{F}}$	Frobenius norm of matrix A
$tr(A)$	The trace of matrix A
$A \geq 0$	A is a non-negative matrix, which means $A_{ij} \geq 0$
I	Identity matrix

Table 1: Several notations in this article.

2 Useful Knowledge

2.1

Theorem 1 $A \in R^{m\times n},\ B \in R^{n\times p}$, then:

$$\partial \frac{\|AB\|_{\mathcal{F}}^2}{\partial B} = 2A^T AB$$

Proof Assume

$$y = \|AB\|_{\mathcal{F}}^2 = \sum_{i=1}^{m}\sum_{j=1}^{p}(A_{(i)} \cdot B_j)^2$$

Then

$$\frac{\partial y}{\partial B_{ij}} = \partial \frac{\sum_{k=1}^{m}(A_{(k)} \cdot B_j)^2}{\partial B_{ij}} \tag{2.1.1}$$

For a specific k, we have

$$\begin{aligned}\frac{\partial(A_{(k)}\cdot B_j)^2}{\partial B_{ij}} &= \partial\frac{(\sum_{l=1}^{n}A_{kl}B_{lj})^2}{\partial B_{ij}}\\ &= \partial\frac{(A_{ki}B_{ij}+\sum_{l\neq i}A_{kl}B_{lj})^2}{\partial B_{ij}}\\ &= \partial\frac{A_{ki}^2B_{ij}^2+2A_{ki}B_{ij}\cdot\sum_{l\neq i}A_{kl}B_{lj}+constant}{\partial B_{ij}}\\ &= 2A_{ki}^2B_{ij}+2A_{ki}\cdot\sum_{l\neq i}A_{kl}B_{lj}\\ &= 2A_{ki}\cdot\sum_{l=1}^{n}A_{kl}B_{lj}\\ &= 2A_{ki}A_{(k)}\cdot B_j\end{aligned}$$

Therefore

$$\begin{aligned}eq(2.1.1) &= 2(A_{1i}A_{(1)}\cdot B_j+A_{2i}A_{(2)}\cdot B_j+\ldots+A_{mi}A_{(m)}\cdot B_j)\\ &= 2\,[A_{1i},\ldots,A_{mi}]\cdot\begin{bmatrix}A_{(1)}\\ \vdots\\ A_{(m)}\end{bmatrix}\cdot B_j\\ &= 2A_i^TAB_j\end{aligned}$$

Thus, we finally get

$$\frac{\partial y}{\partial B}=2\begin{bmatrix}A_1^TAB_1 & \ldots & A_1^TAB_p\\ \vdots & \ddots & \vdots\\ A_n^TAB_1 & \ldots & A_n^TAB_p\end{bmatrix}=2\begin{bmatrix}A_1^T\\ \vdots\\ A_n^T\end{bmatrix}\cdot[AB_1,\ldots,AB_p]=2A^TAB$$

From this conclusion, we can further infer that

$$\partial\frac{\|AB\|_{\mathcal{F}}^2}{\partial A}=\partial\frac{\|(B^TA^T)^T\|_{\mathcal{F}}^2}{\partial A}=\partial\frac{\|B^TA^T\|_{\mathcal{F}}^2}{\partial A}=(\partial\frac{\|B^TA^T\|_{\mathcal{F}}^2}{\partial A^T})^T=2ABB^T$$

2.2

Theorem 2 $A\in R^{m\times n},\ B\in R^{n\times m}$, then:

$$\partial\frac{tr(AB)}{\partial B}=A^T$$

Proof First, it should be pointed out that

$$tr(AB)=\sum_{i=1}^{m}(AB)_{ii}=\sum_{i=1}^{m}A_{(i)}\cdot B_i=\sum_{i=1}^{m}\sum_{j=1}^{n}A_{ij}B_{ji}$$

Therefore

$$\partial\frac{tr(AB)}{\partial B_{xy}} = \partial\frac{\sum_{i=1}^{m}\sum_{j=1}^{n} A_{ij}B_{ji}}{\partial B_{xy}} = A_{yx}$$

Thus, we get that

$$\partial\frac{tr(AB)}{\partial B} = A^T$$

Notice that

$$\because tr(BA) = \sum_{i=1}^{n}\sum_{j=1}^{m} B_{ij}A_{ji} = tr(AB) \tag{2.2.1}$$

$$\therefore \partial\frac{tr(AB)}{\partial A} = \partial\frac{tr(BA)}{\partial A} = B^T$$

2.3

Theorem 3 When a non-negative matrix $A \in R^{m\times n}$ is involved in Lagrange multiplier, a constraint must be added that $A \geq 0$. Thus, $-\sum_{i,j}\lambda_{ij}A_{ij}$ should be added to the original objective function. This can be simplified formally to $-tr(LA^T)$ where $L \in R^{m\times n}$, because of the fact that $-tr(LA^T) = -\sum_{i,j} L_{ij}A_{ij}$

2.4

Theorem 4 Given an invertible matrix A, we have $(A^{-1})^T = (A^T)^{-1}$.

Proof

$$\because (A^{-1})^T A^T = (AA^{-1})^T = I$$

$$\therefore (A^{-1})^T = (A^T)^{-1}$$

2.5

For square matrices X and Y, we say $X = Y^{\frac{1}{2}}$ if

$$X^2 = XX = Y$$

And $Y^{-\frac{1}{2}}$ is simply $(Y^{\frac{1}{2}})^{-1} = (Y^{-1})^{\frac{1}{2}}$.*(cannot prove yet)*

Theorem 5 Given $A \in R^{m\times n}$, let $\tilde{A} = A(A^TA)^{-\frac{1}{2}}$, then we have

$$\tilde{A}^T\tilde{A} = I$$

Proof Let $B = (A^TA)^{-\frac{1}{2}}$, which implies that

$$BB = (A^TA)^{-1}$$

Thus,

$$B^TB^T = (BB)^T = ((A^TA)^{-1})^T = ((A^TA)^T)^{-1} = (A^TA)^{-1}$$

Therefore, $B^T = (A^TA)^{-\frac{1}{2}} = B$, which means B is symmetric. Then, we can draw that

$$\tilde{A}^T\tilde{A} = B^TA^TAB = BA^TAB = B^{-1}(BBA^TA)B = I$$

2.6

Theorem 6 Given two invertible matrices A and B,

$$(AB)^{-1} = B^{-1}A^{-1}$$

2.7

Theorem 7 If A is a symmetric matrix, then A^{-1} is also symmetric.

Proof

$$A^{-1} = (A^T)^{-1} = (A^{-1})^T$$

2.8

Theorem 8 Assume $A \in R^{m\times n}$, then $\|A^TA\|_{\mathcal{F}}^2 = \|AA^T\|_{\mathcal{F}}^2$.

Proof Let $X = A^TA, Y = AA^T$.

$$\because \|A^TA\|_{\mathcal{F}}^2 = \sum_{i,j} X_{ij}^2 = tr(XX^T)$$
$$\therefore \|A^TA\|_{\mathcal{F}}^2 = tr(A^TAA^TA)$$

According to $eq(2.2.1)$, $tr(A^TAA^TA) = tr(AA^TAA^T)$.

$$\therefore \|A^TA\|_{\mathcal{F}}^2 = tr(YY^T) = \sum_{i,j} Y_{ij}^2 = \|AA^T\|_{\mathcal{F}}^2$$

2.9

Theorem 9 Given a symmetric matrix $A \in R^{n\times n}$, the eigenvectors of A corresponding to different eigenvalues are orthogonal.

Proof Assume v_1, v_2 are two eigenvectors of A, and their corresponding eigenvalues are λ_1, λ_2, respectively. Then it is clear that

$$Av_1 = \lambda_1 v_1$$
$$Av_2 = \lambda_2 v_2$$

Transpose both equations, and left-multiply v_2, v_1, respectively, by them:

$$v_1^TA^Tv_2 = \lambda_1 v_1^T v_2 \tag{2.9.1}$$
$$v_2^TA^Tv_1 = \lambda_2 v_2^T v_1 \tag{2.9.2}$$

Notice that $v_1^Tv_2 = v_2^Tv_1$, and that $v_2^TA^Tv_1 = v_1^TAv_2$. Since A is symmetric, it can be further infered that $v_2^TA^Tv_1 = v_1^TA^Tv_2$. Minus $eq(2.9.2)$ by $eq(2.9.1)$, we get:

$$v_1^TA^Tv_2 - v_2^TA^Tv_1 = \lambda_1 v_1^Tv_2 - \lambda_2 v_2^T v_1$$
$$0 = (\lambda_1 - \lambda_2)v_1^Tv_2$$

Therefore, if $\lambda_1 \neq \lambda_2$, then $v_1^Tv_2 = 0$.

2.10

Theorem 10 Given a symmetric non-negative matrix $W \in R^{n\times n}$ with distinct eigenvalues, the solution to $\underset{H^TH=I,H\geq 0}{argmin} \quad tr(H^TWH)$, where $H \in R^{n\times k}$, is given by the first k eigenvectors of W.

Proof Assume that $v_1, v_2, \ldots, v_n$ are the eigenvectors of W, ordered descendingly by their corresponding eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$. Each v_i is length normalized, so it can be infered that $v_i^T v_j = \begin{cases} 0, & i \neq j \\ 1, & i = j \end{cases}$. Assume that $H = [h_1, h_2, \ldots, h_k]$ is a solution to the objective function, it can be proved that substituting v_j for $h_j (1 \leq j \leq k)$ gives a better solution.

To prove this, assume $h_i = \sum_{j=1}^{n} a_{ij} v_j$. Since $h_i^T h_i = 1$, so $\sum_{j=1}^{n} a_{ij}^2 = 1$. We have

$$\begin{aligned}
h_1^T W h_1 &= \big(\sum_{j=1}^{n} a_{1j} v_j\big)^T W \big(\sum_{j=1}^{n} a_{1j} v_j\big) \\
&= \big(\sum_{j=1}^{n} a_{1j} v_j\big)^T \big(\sum_{j=1}^{n} a_{1j} \lambda_j v_j\big) \\
&= \sum_{j=1}^{n} a_{1j}^2 \lambda_j \\
&\leq \sum_{j=1}^{n} a_{1j}^2 \lambda_1 = \lambda_1 = v_1^T W v_1
\end{aligned}$$

So $h_1^T W h_1 \leq v_1^T W v_1$, which means substituting v_1 for h_1 leads to a better solution. However, this would destroy the orthogonality among the vectors in H because v_1 may not be orthogonal to $h_2, h_3, etc.$ To retain orthogonality, the remaining vectors should be picked from the subspace orthogonal to v_1.

Assume that $h_1, \ldots, h_{s-1}$ has been replaced by $v_1, \ldots, v_{s-1}$, respectively. The s-th vector in H should be orthogonal to all of the previous ones, which means that $\forall i \in [1, s-1], h_s^T v_i = 0, \therefore (\sum_{j=1}^{n} a_{sj} v_j)^T v_i = a_{si} = 0$. Therefore, $h_s = \sum_{j=1}^{n} a_{sj} v_j$ can be reduced to $h_s = \sum_{j=s}^{n} a_{sj} v_j$. Thus,

$$\begin{aligned}
h_s^T W h_s &= \big(\sum_{j=s}^{n} a_{sj} v_j\big)^T W \big(\sum_{j=s}^{n} a_{sj} v_j\big) \\
&= \sum_{j=s}^{n} a_{sj}^2 \lambda_j \\
&\leq \lambda_s = v_s^T W v_s
\end{aligned}$$

Therefore, we can substitute v_s for h_s and achieve a better solution. After k substitution, H consists of the first k eigenvectors of W, and the optimal solution is achieved.

2.11

It's my personal habit to adopt **denominator layout** rather than **numerator layout**.

With denominator-layout notation, the derivatives of matrices are as follows (if x or y is a vector, then $x \in R^{n\times 1}, y \in R^{m\times 1}$) :

vector by scalar

$$\frac{\partial \mathbf{y}}{\partial x} = \left[\frac{\partial y_1}{\partial x}, \ldots, \frac{\partial y_m}{\partial x}\right]$$

scalar by vector

$$\frac{\partial y}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y}{\partial x_1} \\ \vdots \\ \frac{\partial y}{\partial x_n} \end{bmatrix}$$

vector by vector

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_m}{\partial x_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_1}{\partial x_n} & \cdots & \frac{\partial y_m}{\partial x_n} \end{bmatrix}$$

which is the transpose of the Jacobian matrix.

scalar by matrix

$$A \in R^{p\times n}, \quad \frac{\partial y}{\partial A} = \begin{bmatrix} \frac{\partial y}{\partial a_{11}} & \cdots & \frac{\partial y}{\partial a_{1n}} \\ \frac{\partial y}{\partial a_{21}} & \cdots & \frac{\partial y}{\partial a_{2n}} \\ \vdots & \ddots & \vdots \\ \frac{\partial y}{\partial a_{p1}} & \cdots & \frac{\partial y}{\partial a_{pn}} \end{bmatrix}$$

From Wikipedia (index: **Matrix Calculus**): *When taking derivatives with an aggregate (vector or matrix) denominator in order to find a maximum or minimum of the aggregate, it should be kept in mind that using numerator layout will produce results that are transposed with respect to the aggregate.*

The result produced by numerator layout is just the transpose of that by denominator layout.

2.12

Fix differentiable functions $f : R^m \to R^p$ and $g : R^n \to R^m$ and a point x in R^n. Let D denote the total derivative, and $f \circ g$ denote the composite of f and g. Then according to the chain rule,

$$D_x(f \circ g) = D_{g(x)} f \circ D_x g$$

Because the derivatives are all linear transformation, they can be rewritten as matrices. The matrix corresponding to a derivative is the Jacobian matrix, and the composite of two derivatives corresponds to the product of their Jacobian matrices.

$$J_{f\circ g}(x) = J_f(g(x))J_g(x)$$

If the denominator layout is adopted, the derivative should be the transpose of the Jacobian matrix. Thus,

$$D_x(f \circ g) = J_g(x)^T J_f(g(x))^T$$

(*All from Wikipeida, with index* ***Chain Rule***.)