Matrix Notes: derivative

Yzy

2020 年 8 月 22 日

目录

1 基本符号

∇ 可看作一个运算符号 (是一个梯度算子)，它作用到一个多元函数上，就得到一个向量，这个向量的每个分量，是这个函数关于每个自变量的偏导数，比如：
$\nabla\varphi(\theta)=\left(\frac{\partial\varphi(\theta)}{\partial\theta_1},\ldots,\frac{\partial\varphi(\theta)}{\partial\theta_d}\right)\ \theta=(\theta_1,\ldots,\theta_d)$

2 矩阵乘法的四种理解

内积：一个行向量乘以一个列向量，结果是一个数；

外积：一个列向量乘以一个行向量，结果是一个矩阵；(外积是一种特殊的克罗内克积 kronecker product)

假设 $\boldsymbol{a}^T$ 为行向量，$\boldsymbol{b}$ 是列向量，即 $\boldsymbol{a} = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}, \boldsymbol{b} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$

则内积为：$\boldsymbol{a}^T \cdot \boldsymbol{b} = a_1b_1 + a_2b_2$，外积为：$\boldsymbol{b} \otimes \boldsymbol{a}^T = \begin{bmatrix} b_1a_1 & b_1a_2 \\ b_2a_1 & b_2a_2 \end{bmatrix}$

矩阵是由向量组成，从不同的角度对矩阵进行抽象，可以将矩阵乘法转换为向量乘法。

给定一个矩阵 $\boldsymbol{A}_{2\times2}$(规定行数和列数以便于理解)：

可以将其看成**由两个行向量组成的列向量** $\begin{bmatrix} \boldsymbol{a}_1^T \\ \boldsymbol{a}_2^T \end{bmatrix}$；

也可以看成**由两个列向量组成的行向量** $\begin{bmatrix} \boldsymbol{a}_1 & \boldsymbol{a}_2 \end{bmatrix}$。

2.1 角度一：行乘列 (外积)

$\boldsymbol{A}$ 是由行向量组成的列向量，$\boldsymbol{B}$ 是由列向量组成的行向量。

$$\boldsymbol{AB} = \begin{bmatrix} \boldsymbol{a}_1^T \\ \boldsymbol{a}_2^T \end{bmatrix} \begin{bmatrix} \boldsymbol{b}_1 & \boldsymbol{b}_2 \end{bmatrix}$$

此时矩阵乘积变为了两个新的向量的外积形式，按照外积定义则有：

$$\boldsymbol{AB} = \begin{bmatrix} \boldsymbol{a}_1^T\boldsymbol{b}_1 & \boldsymbol{a}_1^T\boldsymbol{b}_2 \\ \boldsymbol{a}_2^T\boldsymbol{b}_1 & \boldsymbol{a}_2^T\boldsymbol{b}_2 \end{bmatrix}$$

注意到这里每一个 $\boldsymbol{a}_i^T\boldsymbol{b}_j$ 都是一个内积，即一个标量，作为 $\boldsymbol{AB}$ 矩阵中第 i 行第 j 列的元素。因此，矩阵乘积可以看成是两个向量的外积，并且外积矩阵中的每一个元素是一个内积, 这是最直接的理解方式。

2.2 角度二：列空间

$\boldsymbol{A}$ 和 $\boldsymbol{B}$ 都是由列向量组成的行向量

$$\boldsymbol{AB} = \begin{bmatrix} \boldsymbol{a}_1 & \boldsymbol{a}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{b}_1 & \boldsymbol{b}_2 \end{bmatrix}$$

令 $\boldsymbol{C}=\boldsymbol{AB}$，考虑 $\boldsymbol{C}$ 的每一个列向量：

$$\boldsymbol{c}_1 = A\boldsymbol{b}_1 = \begin{bmatrix} \boldsymbol{a}_1 & \boldsymbol{a}_2 \end{bmatrix} \boldsymbol{b}_1 = \boldsymbol{a}_1 b_{11} + \boldsymbol{a}_2 b_{12}$$

同理：

$$\boldsymbol{c}_2 = \boldsymbol{a}_1 b_{21} + \boldsymbol{a}_2 b_{22}$$

因此，矩阵 $\boldsymbol{C}$ 的每一个列向量，是 $\boldsymbol{A}$ 的列向量的一个线性组合，该线性组合中的系数是 $\boldsymbol{b}_i$ 的各个元素。从这个角度说 $\boldsymbol{C}$ 的每一列都存在于 $\boldsymbol{A}$ 的列向量空间内。

2.3 角度三：行空间

$\boldsymbol{A}$ 是由行向量组成的列向量，$\boldsymbol{B}$ 也是由行向量组成的列向量

$$\boldsymbol{AB} = \begin{bmatrix} \boldsymbol{a}_1^T \\ \boldsymbol{a}_2^T \end{bmatrix} \begin{bmatrix} \boldsymbol{b}_1^T \\ \boldsymbol{b}_2^T \end{bmatrix}$$

令 $\boldsymbol{C}=\boldsymbol{AB}$，考虑 $\boldsymbol{C}$ 的每一个行向量：

$$\boldsymbol{c}_1^T = \boldsymbol{a}_1^T \boldsymbol{B} = \boldsymbol{a}_1^T \begin{pmatrix} \boldsymbol{b}_1^T \\ \boldsymbol{b}_2^T \end{pmatrix} = a_{11}\boldsymbol{b}_1^T + a_{12}\boldsymbol{b}_2^T$$

同理：

$$\boldsymbol{c}_2^T = a_{21}\boldsymbol{b}_1^T + a_{22}\boldsymbol{b}_2^T$$

因此，矩阵 $\boldsymbol{C}$ 的每一个行向量，是 $\boldsymbol{B}$ 的行向量的一个线性组合，该线性组合中的系数是 $\boldsymbol{a}_i^T$ 的各个元素。从这个角度说 $\boldsymbol{C}$ 的每一个行向量都存在于 $\boldsymbol{B}$ 的行向量空间内。

2.4 角度四：内积

$\boldsymbol{A}$ 是由列向量组成的行向量，$\boldsymbol{B}$ 是由行向量组成的列向量

$$\boldsymbol{AB} = \begin{bmatrix} \boldsymbol{a}_1 & \boldsymbol{a}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{b}_1^T \\ \boldsymbol{b}_2^T \end{bmatrix}$$

按照内积定义有：

$$\boldsymbol{AB} = \boldsymbol{a}_1\boldsymbol{b}_1^T + \boldsymbol{a}_2\boldsymbol{b}_2^T$$

注意到 $\boldsymbol{a}_i\boldsymbol{b}_j^T$ 是一个外积形式，结果为一个矩阵。因此 $\boldsymbol{C}$ 是由各个外积矩阵相加得到的。

2.5 总结

$\boldsymbol{C}=\boldsymbol{AB}$ 是一个矩阵

- 既是列向量组成的行向量，每个列向量是 $\boldsymbol{A}$ 的列空间的线性组合
- 又是行向量组成的列向量，每个行向量是 $\boldsymbol{B}$ 的行空间的线性组合
- 既是一个**内积**，内积的每个成分是一个外积
- 又是一个**外积**，外积矩阵的每一个元素是一个内积

3 矩阵向量求导

更多基本公式和例子详见:https://www.cnblogs.com/pinard/p/10791506.html 或 https://zhuanlan.zhihu.com/p/24709748 或 https://github.com/LynnHo/Matrix-Calculus

3.1 求导布局

1. 分子布局 (numerator layout): 求导结果的维度以分子为主
2. 分母布局 (denominator layout): 求导结果的维度以分母为主
3. 混合布局：即如果是向量或者矩阵对标量求导，则使用分子布局为准；如果是标量对向量或者矩阵求导，则以分母布局为准。对于向量对向量求导 (有些分歧), 以分子布局的雅克比矩阵为主。

定义：

- x—> 标量 $\mathbf{x}$—>n 维向量 $\mathbf{X}$—>$m \times n$ 矩阵
- y—> 标量 $\mathbf{y}$—>m 维向量 $\mathbf{Y}$—>$p \times q$ 矩阵

两者相差一个**转置**，例子：标量 **y** 对矩阵 **X** 求导，那么如果按分母布局，则求导结果的维度和矩阵 **X** 的维度 m×n 是一致的。如果是分子布局，则求导结果的维度为 n×m。

所以，**标量**对**向量或矩阵**求导，**向量或矩阵**对**标量**求导这四种情况，对应的分子布局和分母布局的排列方式已经确定了。

向量对向量的求导：(只讨论列向量，行向量求导只相差一个转置)
m 维列向量 **y** 对 n 维列向量 **x** 求导，一共有 mn 个标量对标量求导。

分子布局，结果矩阵的第一维度以分子为准，即为 $m \times n$ 矩阵 (**雅克比矩阵**)：

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}}(or\frac{\partial \mathbf{y}}{\partial \mathbf{x}^{\mathbf{T}}}) = \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \cdots & \frac{\partial y_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y_m}{\partial x_1} & \frac{\partial y_m}{\partial x_2} & \cdots & \frac{\partial y_m}{\partial x_n} \end{pmatrix}$$

分母布局，结果矩阵的第一维度以分母为准，即为 $n \times m$ 矩阵 (**梯度矩阵**)：

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}}(or\frac{\partial \mathbf{y^T}}{\partial \mathbf{x}}) = \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_2}{\partial x_1} & \cdots & \frac{\partial y_m}{\partial x_1} \\ \frac{\partial y_1}{\partial x_2} & \frac{\partial y_2}{\partial x_2} & \cdots & \frac{\partial y_m}{\partial x_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y_1}{\partial x_n} & \frac{\partial y_2}{\partial x_n} & \cdots & \frac{\partial y_m}{\partial x_n} \end{pmatrix}$$

上述两种布局简单思路就是，以谁做布局就用谁的维度作为结果矩阵的第一个维度，即看成列，另一个向量则看成行。

3.2 矩阵向量求导—定义法

3.2.1 标量对向量求导

标量对向量求导，严格来说是**实值函数**对向量的求导。即定义实值函数 $f: R^n \to R$，也即 $f(\boldsymbol{x}) = y$(这里 f,y 均代表标量)，自变量 $\boldsymbol{x}$ 是 n 维向量，y 是标量。

所谓标量对向量的求导，其实就是**标量对向量里的每个分量分别求导，最后把求导的结果排列在一起，按一个向量表示而已**。所以**将实值函数对向量的每一个分量来求导，最后找到规律，得到求导的结果向量**。

例子： $y = \mathbf{a}^T\mathbf{x}$，求解 $\frac{\partial \mathbf{a}^T\mathbf{x}}{\partial \mathbf{x}}$

根据定义，我们先对 $\boldsymbol{x}$ 的第 i 个分量进行求导，这是一个标量对标量的求导，如：$\frac{\partial \mathbf{a}^T\mathbf{x}}{\partial x_i} = \frac{\partial \sum_{j=1}^n a_j x_j}{\partial x_i} = \frac{\partial a_i x_i}{\partial x_i} = a_i$

所以对向量的第 i 个分量的求导结果就等于向量 a 的第 i 个分量。由于是分母布局，最后所有求导结果的分量组成的是一个 n 维向量。即为向量 a：$\frac{\partial \mathbf{a}^T\mathbf{x}}{\partial \mathbf{x}} = \mathbf{a}$

同理：$\frac{\partial \mathbf{x}^T\mathbf{a}}{\partial \mathbf{x}} = \mathbf{a}$

例子： $y = \mathbf{x}^T\mathbf{A}\mathbf{x}$，求解 $\frac{\partial \mathbf{x}^T\mathbf{A}\mathbf{x}}{\partial \mathbf{x}}$

对 $\boldsymbol{x}$ 的第 k 个分量求导：

$\frac{\partial \mathbf{x}^T\mathbf{A}\mathbf{x}}{\partial x_k} = \frac{\partial \sum_{i=1}^n \sum_{j=1}^n x_i A_{ij} x_j}{\partial x_k} = \sum_{i=1}^n A_{ik}x_i + \sum_{j=1}^n A_{kj}x_j = \mathbf{A}^T\mathbf{x} + \mathbf{A}\mathbf{x}$

3.2.2 标量对向量求导的基本法则

- **常量**对向量的求导结果为 0
- 线性法则：如果 f, g 都是实值函数，c_1, c_2 为常数，则：$\frac{\partial (c_1 f(\mathbf{x}) + c_2 g(\mathbf{x}))}{\partial \mathbf{x}} = c_1 \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} + c_2 \frac{\partial g(\mathbf{x})}{\partial \mathbf{x}}$
- 乘法法则：如果 f, g 都是实值函数，则：$\frac{\partial f(\mathbf{x}) g(\mathbf{x})}{\partial \mathbf{x}} = f(\mathbf{x})\frac{\partial g(\mathbf{x})}{\partial \mathbf{x}} + \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} g(\mathbf{x})$

- 除法法则:如果 f, g 都是实值函数,且 $g(x)\ 0$,则: $\frac{\partial f(\mathbf{x})/g(\mathbf{x})}{\partial \mathbf{x}} = \frac{1}{g^2(\mathbf{x})}\left(g(\mathbf{x})\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} - f(\mathbf{x})\frac{\partial g(\mathbf{x})}{\partial \mathbf{x}}\right)$

3.2.3 标量对矩阵求导

例子: $y = \mathbf{a}^T\mathbf{X}\mathbf{b}$, 求解 $\frac{\partial \mathbf{a}^T\mathbf{X}\mathbf{b}}{\partial \mathbf{X}}$, 其中, $\boldsymbol{a}$ 是 m 维向量,$\boldsymbol{b}$ 是 n 维向量,$\boldsymbol{X}$ 是 m×n 的矩阵
对矩阵 $\boldsymbol{X}$ 的任意一个位置的 $\boldsymbol{X}_{ij}$ 求导:
$\dfrac{\partial \mathbf{a}^T\mathbf{X}\mathbf{b}}{\partial X_{ij}} = \dfrac{\partial \sum_{p=1}^{m}\sum_{q=1}^{n} a_p X_{pq} b_q}{\partial X_{ij}} = \dfrac{\partial a_i X_{ij} b_j}{\partial X_{ij}} = a_i b_j$(从行空间视角做矩阵乘法) 即求导结果在 (i.j) 位置的求导结果是 a 向量第 i 个分量和 b 第 j 个分量的乘积,将所有的位置的求导结果排列成一个 m×n 的矩阵,即为 ab^T,这样最后的求导结果为: $\dfrac{\partial \mathbf{a}^T\mathbf{X}\mathbf{b}}{\partial \mathbf{X}} = ab^T$ **标量对矩阵求导也有和第二节对向量求导类似的基本法则**

3.2.4 向量对向量求导

例子: $\mathbf{y} = \mathbf{A}\mathbf{x}$, 其中 $\mathbf{A}$ 为 n×m 的矩阵, $\mathbf{x}, \mathbf{y}$ 分别为 m,n 维向量。求解 $\frac{\partial \mathbf{A}\mathbf{x}}{\partial \mathbf{x}}$:
$(n \times 1)$ 对 $(m \times 1)$ 求导,按照分子布局,结果应是一个 $n \times m$ 矩阵
先求矩阵的第 i 行和向量的内积对向量的第 j 分量求导,用定义法求解过程如下: $\dfrac{\partial \mathbf{A_i}\mathbf{x}}{\partial \mathbf{x_j}} = \dfrac{\partial A_{ij} x_j}{\partial \mathbf{x_j}} = A_{ij}$
所以矩阵 A 的第 i 行和向量的内积对向量的第 j 分量求导的结果是矩阵 A 的 (i,j) 位置的值。由于是分子布局,所以排列出的结果是 A, 而不是 A^T

3.3 矩阵微分

标量的导数和微分: $df = f'(x)dx$(单变量), 若是多变量, 则: $df = \sum_{i=1}^{n}\dfrac{\partial f}{\partial x_i}dx_i = \left(\dfrac{\partial f}{\partial \mathbf{x}}\right)^T d\mathbf{x}$(多元向量值函数的微分) 可以看出标量对向量的求导 $\dfrac{\partial f}{\partial \mathbf{x}}$(列向量) 与标量的向量微分 df **有一个转置的关系**!

推广到**变元是矩阵的情况**:$df = \sum_{i=1}^{m}\sum_{j=1}^{n}\dfrac{\partial f}{\partial X_{ij}}dX_{ij} = \operatorname{tr}\left(\left(\dfrac{\partial f}{\partial \mathbf{X}}\right)^T d\mathbf{X}\right)$,其中迹函数等于主对角线的和: $\operatorname{tr}\left(A^T B\right) = \sum_{i,j} A_{ij} B_{ij}$($A, B$ **必须同型,且此处是逐元素点乘**。)

此公式可以把 A 和 B 均看做列向量来理解,矩阵同理!▯积是两个矩

阵相同的对应位置上元素乘积之和。因为两个矩阵相乘，A 中第 A_{ij} 个元素乘以 B 中第 B_{ij} 个元素的积，全部在形成的矩阵对角线上。(可以举例子理解)

矩阵微分和它的导数也有一个转置的关系，只是在外面套了一个迹函数。由于标量的迹函数就是它本身，那么矩阵微分和向量微分可以统一表示，即：$df=\operatorname{tr}\left(\left(\frac{\partial f}{\partial \mathbf{X}}\right)^T d\mathbf{X}\right) \quad df=\operatorname{tr}\left(\left(\frac{\partial f}{\partial \mathbf{x}}\right)^T d\mathbf{x}\right)$

3.3.1 矩阵向量求导—微分法 (标量对向量或矩阵的求导)

求解方法 (因变量须为标量)：若标量函数 f 是矩阵 $\boldsymbol{X}$ 经加减乘法、逆、行列式、逐元素函数等运算构成，则使用相应的运算法则对 f 求微分，再使用迹函数技巧给 df 套上迹并将其它项交换至 $d\boldsymbol{X}$ 左侧, 那么对于迹函数里面在 $d\boldsymbol{X}$ 左边的部分，只需要加一个转置便可以得到导数。

迹函数常用技巧：

1. 标量的迹等于自己：$\operatorname{tr}(x)=x$
2. 转置不变：$\operatorname{tr}\left(A^T\right)=\operatorname{tr}(A)$
3. 交换律：$\operatorname{tr}(AB)=\operatorname{tr}(BA)$，需要满足 A, B^T 同维度，实质上就是 A, B 相容，并且由于前面加了 $tr()$，所以最后结果是一个方阵。其证明过程保存在 chrome 矩阵分析书签中, 其中需要注意比较 $tr(AB)$ 和 $tr(BA)$ 的代数表达式时，交换两个求和符号实质上就是从 AB->BA 的矩阵相乘 (行乘列)，仔细比较下标即可发现。
4. 加减法：$\operatorname{tr}(X+Y)=\operatorname{tr}(X)+\operatorname{tr}(Y), \operatorname{tr}(X-Y)=\operatorname{tr}(X)-\operatorname{tr}(Y)$
5. 矩阵乘法和迹交换：$\operatorname{tr}\left((A \odot B)^T C\right) = \operatorname{tr}\left(A^T(B \odot C)\right)$，需要满足 A, B, C 同维度。(中间是哈达玛积，此式利用了前面提到的迹函数等于主对角线的和一式)

例子：$y=\mathbf{a}^T \mathbf{X} \mathbf{b}$，求解 $\frac{\partial y}{\partial \mathbf{X}}$

1. 对 f 求微分：$dy=d\mathbf{a}^T \mathbf{X} \mathbf{b}+\mathbf{a}^T d\mathbf{X} \mathbf{b}+\mathbf{a}^T \mathbf{X} d\mathbf{b}=\mathbf{a}^T d\mathbf{X} \mathbf{b}$(因变量部分不是自变量的函数，因此导数为 0，直接省略了)
2. 两边套上迹函数：$dy=\operatorname{tr}(dy)=\operatorname{tr}\left(\mathbf{a}^T d\mathbf{X} \mathbf{b}\right)=\operatorname{tr}\left(\mathbf{b}\mathbf{a}^T d\mathbf{X}\right)$

3. 根据矩阵导数和微分的定义，迹函数里面在 $d\boldsymbol{X}$ 左边的部分 $\boldsymbol{b}\boldsymbol{a}^T$, 加上一个转置即为要求的导数，即：$\frac{\partial f}{\partial \mathbf{X}}=\left(\mathbf{b}\mathbf{a}^T\right)^T=\boldsymbol{a}\boldsymbol{b}^T$

例子： $y=\mathbf{a}^T\exp(\mathbf{X}\mathbf{b})$，求解 $\frac{\partial y}{\partial \mathbf{X}}$

1. $dy=\operatorname{tr}(dy)=\operatorname{tr}\left(\mathbf{a}^T \operatorname{d}\exp(\mathbf{X}\mathbf{b})\right)$(因变量部分不是自变量的函数，因此导数为 0。$d\boldsymbol{a}^T=0$ 省略了)

2. $=\operatorname{tr}\left(\mathbf{a}^T(\exp(\mathbf{X}\mathbf{b})\odot d(\mathbf{X}\mathbf{b}))\right)$(这里使用了逐元素求导 $d\sigma(X)=\sigma'(X)\odot dX$，$exp'(x)=exp(x)$)

3. $=\operatorname{tr}\left((\mathbf{a}\odot\exp(\mathbf{X}\mathbf{b}))^T d\mathbf{X}\mathbf{b}\right)$(这里 $d(\boldsymbol{X}b)$ 展开后 $=d\boldsymbol{X}\cdot\boldsymbol{b}+\boldsymbol{X}\cdot d\boldsymbol{b}$，后一项为 0 所以省略)

4. $=\ \operatorname{tr}\left((\mathbf{a}\odot\exp(\mathbf{X}\mathbf{b}))^T d\mathbf{X}\cdot\mathbf{b}\right)$(这里验证前后两部分的维度呈转置关系，可以利用迹的交换律把 $\boldsymbol{b}$ 放到最前面)

5. $=\operatorname{tr}\left(\mathbf{b}(\mathbf{a}\odot\exp(\mathbf{X}\mathbf{b}))^T d\mathbf{X}\right)$

所以结果为：$\dfrac{\partial y}{\partial \mathbf{X}}=(\mathbf{a}\odot\exp(\mathbf{X}\mathbf{b}))b^T$

3.3.2 向量矩阵的迹函数对向量矩阵求导（也是标量对向量矩阵求导）

$\dfrac{\partial\operatorname{tr}(AB)}{\partial A}=B^T(A,B^T$ 同型)，按照矩阵微分定义 (张贤达矩阵分析与应用 3.2.1: $\mathrm{d}(\operatorname{tr}\boldsymbol{U})=\operatorname{tr}(\mathrm{d}\boldsymbol{U})$)：

$d\operatorname{tr}(AB)=\operatorname{tr}(d(AB))=\operatorname{tr}(dA\cdot B+A\cdot dB)=\operatorname{tr}(dA\cdot B)=\operatorname{tr}(B\cdot dA)$；同理，$\dfrac{\partial\operatorname{tr}(AB)}{\partial B}=A^T$。

求解 $\dfrac{\partial tr\left(W^TAW\right)}{\partial W}$(张贤达一书 3.2.1: $d\left(\boldsymbol{X}^{\mathrm{T}}\right)=(d\boldsymbol{X})^{\mathrm{T}}$)：

$d\left(\operatorname{tr}\left(W^TAW\right)\right)=\operatorname{tr}\left(dW^TAW+W^TAdW\right)=\operatorname{tr}\left(dW^TAW\right)+\operatorname{tr}\left(W^TAdW\right)=$
$\operatorname{tr}\left((dW)^TAW\right)+\operatorname{tr}\left(W^TAdW\right)=\operatorname{tr}\left(W^TA^TdW\right)+\operatorname{tr}\left(W^TAdW\right)=\operatorname{tr}\left(W^T\left(A+A^T\right)dW\right)$

所以结果为：$\dfrac{\partial\operatorname{tr}\left(W^TAW\right)}{\partial W}=\left(A+A^T\right)W$

求解 $\dfrac{\partial\operatorname{tr}\left(B^TX^TCXB\right)}{\partial X}$：

$d\left(\operatorname{tr}\left(B^TX^TCXB\right)\right)=\operatorname{tr}\left(B^TdX^TCXB\right)+\operatorname{tr}\left(B^TX^TCdXB\right)=\operatorname{tr}\left((dX)^TCXBB^T\right)+$
$\operatorname{tr}\left(BB^TX^TCdX\right)=\operatorname{tr}\left(BB^TX^TC^TdX\right)+\operatorname{tr}\left(BB^TX^TCdX\right)=\operatorname{tr}\left(\left(BB^TX^TC^T+BB^TX^TC\right)dX\right)$

因此：$\dfrac{\partial\operatorname{tr}\left(B^TX^TCXB\right)}{\partial X}=\left(C+C^T\right)XBB^T$

3.4 矩阵向量求导—链式法则

3.4.1 向量对向量求导 (分子布局)

假设多个向量存在依赖关系，如：$\mathbf{x} \to \mathbf{y} \to \mathbf{z}$，则有以下链式求导法则：$\dfrac{\partial \mathbf{z}}{\partial \mathbf{x}} = \dfrac{\partial \mathbf{z}}{\partial \mathbf{y}} \dfrac{\partial \mathbf{y}}{\partial \mathbf{x}}$；但是要求所有有依赖关系的变量都是向量，不能是矩阵或标量！

从矩阵维度相容的角度来理解上述链式法则：假设 $\mathbf{x}, \mathbf{y}, \mathbf{z}$ 分别是 m,n,p 维向量，则求导结果 $\frac{\partial \mathbf{z}}{\partial \mathbf{x}}$ 是一个 p×m 的雅克比矩阵，而右边 $\frac{\partial \mathbf{z}}{\partial \mathbf{y}}$ 是一个 p×n 的雅克比矩阵，$\frac{\partial \mathbf{y}}{\partial \mathbf{x}}$ 是一个 n×m 的矩阵，两个雅克比矩阵的乘积维度刚好是 p×m，和左边相容。

3.4.2 标量对多个向量求导

机器学习算法中，最终要优化的一般是一个标量损失函数，因此最后求导的目标是标量，如：$\mathbf{x} \to \mathbf{y} \to z$，按照上一小节易发现维度不相容。

假设 $\boldsymbol{x}, \boldsymbol{y}$ 分别是 m,n 维向量, 那么 $\dfrac{\partial z}{\partial \boldsymbol{x}}$ 的求导结果是一个 m×1 的向量，$\dfrac{\partial z}{\partial \boldsymbol{y}}$ 是一个 n×1 的向量，$\dfrac{\partial \mathbf{y}}{\partial \mathbf{x}}$ 是一个 n×m 的雅克比矩阵, 右边的向量和矩阵是没法直接乘的。

1. 假如把标量求导的部分都做一个转置,则维度就可以相容了:$\left(\dfrac{\partial z}{\partial \mathbf{x}}\right)^T = \left(\dfrac{\partial z}{\partial \mathbf{y}}\right)^T \dfrac{\partial \mathbf{y}}{\partial \mathbf{x}}$
2. 然后两边再转置得到**标量对多个向量求导的链式法则:**$\dfrac{\partial z}{\partial \mathbf{x}} = \left(\dfrac{\partial \mathbf{y}}{\partial \mathbf{x}}\right)^T \dfrac{\partial z}{\partial \mathbf{y}}$
3. 如果是**标量对更多的向量求导**, 比如 $\mathbf{y}_1 \to \mathbf{y}_2 \to \ldots \to \mathbf{y}_n \to z$，则其链式求导表达式：$\dfrac{\partial z}{\partial \mathbf{y}_1} = \left(\dfrac{\partial \mathbf{y_n}}{\partial \mathbf{y}_{\mathbf{n}-1}} \dfrac{\partial \mathbf{y_{n-1}}}{\partial \mathbf{y_{n-2}}} \cdots \dfrac{\partial \mathbf{y_2}}{\partial \mathbf{y_1}}\right)^T \dfrac{\partial z}{\partial \mathbf{y_n}}$

另一种求复合函数导数的思路：

假设已求得 $\dfrac{\partial f}{\partial Y}$，而 Y 是 X 的函数，如何求 $\dfrac{\partial f}{\partial X}$？可以直接从微分入手建立复合法则：先写出 $df = \operatorname{tr}\left(\frac{\partial f^T}{\partial Y} dY\right)$，再将 dY 用 dX 表示出来代入，并使用迹技巧将其他项交换至 dX 左侧，即可得到 $\dfrac{\partial f}{\partial X}$。

举个例子，$Y=AXB$, 此时

$$df=\operatorname{tr}\left(\frac{\partial f^T}{\partial Y}dY\right)=\operatorname{tr}\left(\frac{\partial f^T}{\partial Y}AdXB\right)=\operatorname{tr}\left(B\frac{\partial f^T}{\partial Y}AdX\right)=\operatorname{tr}\left(\left(A^T\frac{\partial f}{\partial Y}B^T\right)^T dX\right),$$

可得 $\frac{\partial f}{\partial X}=A^T\frac{\partial f}{\partial Y}B^T$！

最小二乘法例子：

(链式)： $l=(X\boldsymbol{\theta}-\boldsymbol{y})^T(X\boldsymbol{\theta}-\boldsymbol{y})$, 优化的损失函数 l 是一个标量, 模型参数 $\boldsymbol{\theta}$ 是一个向量。假设向量 $\boldsymbol{z}=X\boldsymbol{\theta}-\boldsymbol{y}$($\boldsymbol{z}$ 是向量), 则 $l=\boldsymbol{z}^T\boldsymbol{z}$, $\boldsymbol{\theta}\to\boldsymbol{z}\to l$ 存在链式求导关系，因此：$\frac{\partial l}{\partial\boldsymbol{\theta}}=\left(\frac{\partial\boldsymbol{z}}{\partial\boldsymbol{\theta}}\right)^T\frac{\partial l}{\partial\mathbf{z}}=X^T(2\boldsymbol{z})=2X^T(X\boldsymbol{\theta}-y)$。

其中用到的求导公式：$\frac{\partial(X\boldsymbol{\theta}-y)}{\partial\boldsymbol{\theta}}=X$(分子是向量，向量对向量求导，定义法，使用分子布局)，$\frac{\partial\boldsymbol{z}^T\boldsymbol{z}}{\partial\boldsymbol{z}}=2\boldsymbol{z}$(分子是标量, 标量对向量求导，使用分母布局)。

(微分法)： 因为 l 是一个标量，所以可以用前面的微分法来进行求导！

1. $d(l)=\operatorname{tr}(d(l))=\operatorname{tr}\left[d(\theta^TX^TX\theta)-d(\theta^TX^Ty)-d(y^TX\theta)\right]$
   $=\operatorname{tr}\left[d(\theta^T)X^TX\theta+\theta^TX^TXd\theta-d(\theta^T)X^Ty-y^TXd\theta\right]$
   $=\operatorname{tr}[d(\theta^T)X^TX\theta]+\operatorname{tr}(\theta^TX^TXd\theta)-\operatorname{tr}[d(\theta^T)X^Ty]-\operatorname{tr}(y^TXd\theta)$
   $=\operatorname{tr}(\theta^TX^TXd\theta)+\operatorname{tr}(\theta^TX^TXd\theta)-\operatorname{tr}(y^TXd\theta)-\operatorname{tr}(y^TXd\theta)$
   $=2\operatorname{tr}(\theta^TX^TXd\theta)-2\operatorname{tr}(y^TXd\theta)$
   $=\operatorname{tr}[2(\theta^TX^T-y^T)Xd\theta]$
   $=\operatorname{tr}\left\{[2X^T(X\theta-y)]^Td\theta\right\}$(这里的 $\boldsymbol{\theta}$ 和 $\boldsymbol{y}$ 均为列向量)
2. 所以由微分法可得：$\frac{\partial l}{\partial\boldsymbol{\theta}}=2X^T(X\boldsymbol{\theta}-\boldsymbol{y})$
3. **第 1 步繁琐，应该为：** $d(l)=(\boldsymbol{X}d\boldsymbol{\theta})^T(\boldsymbol{X}\boldsymbol{\theta}-\boldsymbol{y})+(\boldsymbol{X}\boldsymbol{\theta}-\boldsymbol{y})^T(\boldsymbol{X}d\boldsymbol{\theta})=2(\boldsymbol{X}\boldsymbol{\theta}-\boldsymbol{y})^T\boldsymbol{X}d\boldsymbol{\theta}$，$\frac{\partial l}{\partial\boldsymbol{w}}=2\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{\theta}-\boldsymbol{y})$
4. $\frac{\partial l}{\partial\boldsymbol{\theta}}=\mathbf{0}$ 即 $\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\theta}=\boldsymbol{X}^T\boldsymbol{y}$, 故 $\boldsymbol{\theta}$ 的最小二乘估计为 $\boldsymbol{\theta}=\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}$

3.5 标量对多个矩阵求导

假设有这样的依赖关系：$\mathbf{X}\to\mathbf{Y}\to z$，回顾多元复合函数的求导法则(只要涉及到求导自变量的中间变量都要进行求偏导，且最后结果累加)，有：

$\frac{\partial z}{\partial x_{ij}}=\sum_{k,l}\frac{\partial z}{\partial Y_{kl}}\frac{\partial Y_{kl}}{\partial X_{ij}}=\operatorname{tr}\left(\left(\frac{\partial z}{\partial Y}\right)^T\frac{\partial Y}{\partial X_{ij}}\right)$

这里没有给出基于矩阵整体的链式求导法则，主要原因是矩阵对矩阵的求导是比较复杂的定义，目前也未涉及。因此**只能给出对矩阵中一个标量的链式求导方法**。这个方法并不实用，因为我们并不想每次都基于定义法来求导最后再去排列求导结果。

这里的 $\sum_{k,l}$ 是什么意思？**应该是指标量 z 对整个矩阵 Y 逐元素的进行求偏导，再通过链式法则让后面的对应矩阵 Y 元素对自变量 X_{ij} 求导，然后把所有结果累加起来！**。

例 1： A, B, X, Y 都是矩阵，z 是标量，其中 $z = f(Y), Y = AX + B$，求 $\dfrac{\partial z}{\partial X}$

1. 这里使用**定义法**,先用上面的标量链式求导公式 $\dfrac{\partial z}{\partial x_{ij}} = \sum_{k,l} \dfrac{\partial z}{\partial Y_{kl}} \dfrac{\partial Y_{kl}}{\partial X_{ij}}$
2. 后半部分的求导：$\dfrac{\partial Y_{kl}}{\partial X_{ij}} = \dfrac{\partial \sum_s (A_{ks} X_{sl})}{\partial X_{ij}} = \dfrac{\partial A_{ki} X_{il}}{\partial X_{ij}} = A_{ki} \delta_{lj}$，其中 $\delta_{lj} = \begin{cases} 1 & l = j \\ 0 & l \neq j \end{cases}$ 这里 $\sum_s (A_{ks} X_{sl})$ 表示行乘列运算，代表结果矩阵 Y 中的元素 Y_{kl}。
3. 故最终转化为：$\dfrac{\partial z}{\partial x_{ij}} = \sum_{k,l} \dfrac{\partial z}{\partial Y_{kl}} A_{ki} \delta_{lj} = \sum_k \dfrac{\partial z}{\partial Y_{kj}} A_{ki}$
4. 然后将标量对矩阵中单个元素求偏导的结果进行排列，首先 $\sum_k \dfrac{\partial z}{\partial Y_{kj}}$ 是求偏导结果矩阵 $\dfrac{\partial z}{\partial Y}$ 的第 j 列，然后 $\sum_k A_{ki}$ 转置过来就是后半部分结果矩阵 A^T 的第 i 行, 所以排列成矩阵即为：$\dfrac{\partial z}{\partial X} = A^T \dfrac{\partial z}{\partial Y}$
5. 总结：
   $z = f(Y), Y = AX + B \rightarrow \dfrac{\partial z}{\partial X} = A^T \dfrac{\partial z}{\partial Y}$;
   当 $\boldsymbol{x}$ 是一个向量的时候也成立：
   $z = f(\mathbf{y}), \mathbf{y} = A\mathbf{x} + \mathbf{b} \rightarrow \dfrac{\partial z}{\partial \mathbf{x}} = A^T \dfrac{\partial z}{\partial \mathbf{y}}$
6. 如果要求导的自变量在左边，线性变换在右边，也有类似稍有不同的结论如下：
   $z = f(Y), Y = XA + B \rightarrow \dfrac{\partial z}{\partial X} = \dfrac{\partial z}{\partial Y} A^T$
   $z = f(\mathbf{y}), \mathbf{y} = X\mathbf{a} + \mathbf{b} \rightarrow \dfrac{\partial z}{\partial \mathbf{X}} = \dfrac{\partial z}{\partial \mathbf{y}} a^T$

(对上述第 6 步额外两条公式求解)

例 2： $z=f(Y), Y=XA+B$，求 $\dfrac{\partial z}{\partial X}$

1. **定义法：** $\dfrac{\partial z}{\partial x_{ij}}=\sum_{k,l}\dfrac{\partial z}{\partial Y_{kl}}\dfrac{\partial Y_{kl}}{\partial X_{ij}}$

2. 后半部分的求导：$\dfrac{\partial Y_{kl}}{\partial X_{ij}}=\dfrac{\partial\sum_s\left(X_{ks}A_{sl}\right)}{\partial X_{ij}}=\dfrac{\partial X_{kj}A_{jl}}{\partial X_{ij}}=\delta_{kj}A_{jl}$，其中 $\delta_{kj}=\begin{cases}1 & k=i\\0 & k\neq i\end{cases}$ 这里 $\sum_s(X_{ks}A_{sl})$ 表示行乘列运算，代表结果矩阵 Y 中的元素 Y_{kl}

3. 最终转化为：$\dfrac{\partial z}{\partial x_{ij}}=\sum_{k,l}\dfrac{\partial z}{\partial Y_{kl}}\delta_{kj}A_{jl}=\sum_l\dfrac{\partial z}{\partial Y_{il}}A_{jl}$

4. 将标量对矩阵中单个元素求偏导的结果进行排列，首先看 $\dfrac{\partial z}{\partial x_{ij}}$ 在最终结果矩阵中的下标为 ij，说明是两个矩阵相乘 (第 i 行乘第 j 列的结果)，所以 $\sum_l\dfrac{\partial z}{\partial Y_{il}}$ 是求偏导结果矩阵 $\dfrac{\partial z}{\partial Y}$ 的第 i 行，然后 $\sum_l A_{jl}$ 转置过来就是后半部分结果矩阵 A^T 的第 j 列, 所以排列成矩阵即为: $\dfrac{\partial z}{\partial X}=\dfrac{\partial z}{\partial Y}A^T$

例 3： $z=f(\mathbf{y}), \mathbf{y}_{m\times 1}=X_{m\times n}\mathbf{a}_{n\times 1}+\mathbf{b}_{m\times 1}$ 求 $\dfrac{\partial z}{\partial \mathbf{X}}$(a 是向量，X 是矩阵)

1. 还是定义法：$\dfrac{\partial z}{\partial X_{ij}}=\sum_k\dfrac{\partial z}{\partial \boldsymbol{y}_k}\dfrac{\partial \boldsymbol{y}_k}{\partial X_{ij}}$

2. 其中后半部分求导：$\dfrac{\partial \boldsymbol{y}_k}{\partial X_{ij}}=\dfrac{\partial\sum_l\left(X_{kl}a_l\right)}{\partial X_{ij}}=\dfrac{\partial X_{kj}a_j}{\partial X_{ij}}=\delta_{kj}a_j$，其中 $\delta_{kj}=\begin{cases}1 & k=i\\0 & k\neq i\end{cases}$ 这里 $\sum_l(X_{kl}a_l)$ 表示行乘列运算，代表结果向量 y 中的元素 $\boldsymbol{y}_k$

3. 最终转化为：$\dfrac{\partial z}{\partial X_{ij}}=\sum_k\dfrac{\partial z}{\partial \boldsymbol{y}_k}\delta_{kj}a_j=\dfrac{\partial z}{\partial \boldsymbol{y}_i}a_j$

4. 将标量对矩阵中单个元素求偏导的结果进行排列，首先看 $\dfrac{\partial z}{\partial X_{ij}}$ 在最终结果矩阵中的下标为 ij，这里说明是两个向量相乘 (第 i 列乘第 j 行的结果)，所以结果是矩阵！所以 $\dfrac{\partial z}{\partial \boldsymbol{y}_i}$ 是求偏导结果矩阵 ($m\times 1$ 向

量)$\dfrac{\partial z}{\partial \boldsymbol{y}}$ 的第 i 行，然后 a_j 是后半部分结果矩阵 $a^T(n \times 1$ 向量) 的第 j 列, 所以排列成矩阵即为：$\dfrac{\partial z}{\partial X} = \dfrac{\partial z}{\partial \boldsymbol{y}} a^T$

4 矩阵对矩阵求导

矩阵对矩阵求导的定义，计算过程使用了向量化技巧并涉及到 Kronecker 积!

TODO: 看完矩阵应用与分析 (1.10 和第 3 章)，矩阵求导术 (下) 以及 liujianping 博客矩阵求导 (5)! 并做笔记

dot product (scalar product) 点积，也叫数量积或标量积，是内积的一种特殊形式。

inner product 矩阵 A,B 的内积是逐元素相乘再加起来 $sum(sum(A.*B))$，等价于 $tr(A^TB)$

$\operatorname{tr}\left(A^TB\right)=\sum_{i,j}A_{ij}B_{ij}(A,B$ **必须同型，且此处是逐元素点乘**。)，所以两个矩阵的内积与两个相邻的内积是类似的，都是逐元素点乘然后求和!

4.1 定义

4.2 微分法

4.3 实例