矩阵求导

I. 标量对矩阵求导

1.1 基础知识

标量 f 对于矩阵 X 的导数，定义如 **公式1**，即 f 对 X 逐元素求导排成与 X 尺寸相同的矩阵：

$$\frac{\partial f}{\partial X} = \left[\frac{\partial f}{\partial X_{ij}} \right] \tag{1}$$

在一元微积分中，标量对标量的导数和微分有如下联系：

$$df = f'(x)dx \tag{2}$$

1.2 梯度和微分

而在多元微积分中的梯度（标量对向量的导数）也和微分有联系：

$$df = \sum_{i=1}^{n} \frac{\partial f}{\partial x_{ij}} dx_i = \frac{\partial f}{\partial \vec{x}}^T d\vec{x} \tag{3}$$

其中，第一个等式来自全微分公式，第二个等式表达了梯度和微分的关系：

- 全微分 df 是梯度向量 $\frac{\partial f}{\partial x}(n \times 1)$ 与微分向量 $dx(n \times 1)$ 的内积
- 内积是两个同样 size 的矩阵的相同位置元素的乘积之和

1.3 标量对矩阵的导数和微分

于是，我们可以把标量对矩阵的导数和微分建立联系：

$$df = \sum_{i=1}^{m} \sum_{i=1}^{n} \frac{\partial f}{\partial x_{ij}} dx_{ij} = tr(\frac{\partial f}{\partial X}^T dX) \tag{4}$$

和梯度类似的，第一个等式来自全微分公式，第二个等式表明了矩阵导数和微分的联系

- 全微分 df 是导数 $\frac{\partial f}{\partial x}(m \times n)$ 与微分矩阵 $dX(m \times n)$ 的内积

矩阵的迹 trace 是方阵的对角线元素之和。对于同样 size 的矩阵 A 和 B，$tr(A^T B)$ 是矩阵A和B的内积，即

$$tr(A^T B) = \sum_{i,j} A_{ij} B_{ij} \tag{5}$$

1.4 矩阵微分运算法则

定义常用的矩阵微分运算法则：

1. 矩阵加减：$d(X \pm Y) = dX \pm dY$
2. 矩阵乘法：$d(XY) = (dX) \cdot Y + X \cdot dY$
3. 矩阵转置：$d(X^T) = (dX)^T$
4. 矩阵的迹：$d(tr(X)) = tr(dX)$
5. 逆矩阵：$dX^{-1} = -X^{-1} dX X^{-1}$，证明：$dI = d(XX^{-1}) = (dX)X^{-1} + xdX^{-1} = \vec{O}$
6. 行列式：$d\,|X| = tr(X^{\#} dX)$，其中 $X^{\#}$ 是 X 的伴随矩阵
    - 如果 X 可逆，则这个等式可以写成 $d\,|X| = |X|\, tr(X^{-1} dX)$
    - 本等式可用laplace展开证明，详见张贤达《矩阵分析与应用》第279页
7. 逐元素乘法：$d(X \odot Y) = dX \odot Y + X \odot dY$，其中 $\odot$ 表示同样 size 的矩阵 X 和 Y 的逐元素相乘

8. 逐元素函数：$d\sigma(X)=\sigma'(X)\odot dX$，其中 $\sigma(X)=[\sigma(X_{ij})]$ 是逐元素标量函数运算，$\sigma'(X)=[\sigma'(X_{ij})]$ 是逐元素求导数。例如：
   - $X=\begin{bmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{bmatrix}$， $dsin(X)=\begin{bmatrix} cosX_{11}dX_{11} & cosX_{12}dX_{12} \\ cosX_{21}dX_{21} & cosX_{22}dX_{22} \end{bmatrix}=cos(X)\odot dX$

1.5 迹运算

根据矩阵导数和微分的联系 $df=tr(\frac{\partial f}{\partial X}^T dX)$，在求出左侧的微分 df 之后，可以写成右侧的形式并且得到导数，需要一些迹运算技巧：

1. 标量套上迹：$a=tr(a)$
2. 转置：$tr(A^T)=tr(A)$
3. 线性：$tr(A\pm B)=tr(A)\pm tr(B)$
4. 矩阵乘法交换：$tr(AB)=tr(BA)$，其中 A 和 B^T 的 size 相同，显然两侧都等于 $\sum_{i,j}A_{ij}B_{ij}$
5. 矩阵乘法/逐元素乘法 的交换：$tr(A^T(B\odot C))=tr((A\odot B)^TC)$，其中A、B、C的 size 相同，两侧都等于 $\sum_{i,j}A_{ij}B_{ij}C_{ij}$
6. 同样的，$tr(A\cdot(B\odot C))=tr((A^T\odot B)^TC)$

1.6 导数的运算

标量函数 f 是由矩阵 X 经过加减乘法、逆、行列式、逐元素函数等运算构成的。

1. 我们可以使用 1.4 矩阵微分运算法则 中的运算法则求出 f 的微分 df （df 是一个包含 dX 的多项式）。
2. 之后给 df 套上迹，并且把其他项（也就是下面式子中的 A）交换到 dX 的左侧，对照导数和微分的联系 $df=tr(\frac{\partial f}{\partial X}^T dX)$，就可以函数 f 关于矩阵 X 的导数：

$$df=tr(df)=tr(A\cdot dX)=tr(\frac{\partial f}{\partial X}^T dX) \tag{6}$$

$$\frac{\partial f}{\partial X}=A^T \tag{7}$$

为了避免以后看不懂，说明一下，df 可以先被放到 $tr()$ 里面，之后被整理成 $A\cdot dX$ 的形式，又已经证明 $df=tr(\frac{\partial f}{\partial X}^T dX)$，那么显然有 $\frac{\partial f}{\partial X}=A^T$。

特别的，如果矩阵退化成向量，那么根据梯度和微分的联系 $df=\frac{\partial f}{\partial \vec{x}}^T d\vec{x}$，同样能得到导数。

1.7 复合函数求导

假设已知 $\frac{\partial f}{\partial Y}$，而Y是X的函数，如何求 $\frac{\partial f}{\partial X}$？

- 不可以使用标量求导的链式法则 $\frac{\partial f}{\partial X}=\frac{\partial f}{\partial Y}\frac{\partial Y}{\partial X}$，因为矩阵对矩阵的导数 $\frac{\partial Y}{\partial X}$ 还没有定义。
- 链式法则的源头仍然是微分，所以直接从微分入手建立复合法则
- **首先写出 $df=tr(\frac{\partial f}{\partial Y}^T dY)$，之后将 dY 用 dX 表示出来带入，使用迹运算技巧将其他项交换到**

dX **的左侧，就可以得到** $\frac{\partial f}{\partial X}$

例如，对于 $Y = AXB$，已知 $\frac{\partial f}{\partial Y}$，求 $\frac{\partial f}{\partial X}$：

1. $dY = (dA)XB + A(dXB) = (dA)XB + A(dX)B + AX(dB) = A(dX)B$, 因为A和B是常量，所以 $dA = 0, dB = 0$
2. $df = tr(\frac{\partial f}{\partial Y}^T dY) = tr(\frac{\partial f}{\partial Y}^T A(dX)B) = tr(B\frac{\partial f}{\partial Y}^T A(dX)) = tr((A^T \frac{\partial f}{\partial Y} B^T)^T dX)$
3. $\frac{\partial f}{\partial X} = A^T \frac{\partial f}{\partial Y} B^T$

在这个过程中，使用了矩阵乘法交换的迹运算技巧，交换了 $\frac{\partial f}{\partial Y}^T A(dX)$ 和 B

1.8 示例

查看III. 参考资料中标量对矩阵求导，本笔记来自那篇文章。

原文中除了有上述推理，还包括了标量对矩阵求导的例子。

II. 矩阵对矩阵求导

2.1 基本特征

矩阵对矩阵的导数，应该有如下这些特征：

1. 矩阵 $F(p \times q)$ 和矩阵 $X(m \times n)$ 的导数应该包含 **mnpq 个偏导数** $\frac{\partial F_{kl}}{\partial X_{ij}}$，这样才能不损失信息
2. 导数和微分有简明的联系，因为在导数的计算和应用中需要这个联系
3. 导数有简明的从整体出发的算法

2.2 矩阵对矩阵导数和微分

先定义向量 $\vec{f}(p \times 1)$ 对向量 $\vec{x}(m \times 1)$ 的导数：

$$\frac{\partial \vec{f}}{\partial \vec{x}} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_2}{\partial x_1} & \cdots & \frac{\partial f_p}{\partial x_1} \\ \frac{\partial f_1}{\partial x_2} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_p}{\partial x_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_1}{\partial x_m} & \frac{\partial f_2}{\partial x_m} & \cdots & \frac{\partial f_p}{\partial x_m} \end{bmatrix} \tag{8}$$

$$d\vec{f} = \frac{\partial \vec{f}}{\partial \vec{x}}^T d\vec{x} \tag{9}$$

再定义**矩阵的按列优先的向量化**：

$$vec(X) = [X_{11}, \cdots, X_{m1}, X_{12}, \cdots, X_{m2}, \cdots, X_{1n}, \cdots, X_{mn}]^T \qquad (mn \times 1) \tag{10}$$

在此之上，定义矩阵 $F_{p\times q}$ 对于矩阵 $X_{m\times n}$ 的导数为：

$$\frac{\partial F}{\partial X} = \frac{\partial vec(F)}{\partial vec(X)} \qquad (mn \times pq) \tag{11}$$

导数和微分的联系如下:

$$vec(dF) = \frac{\partial F}{\partial X}^T vec(dX) \tag{12}$$

也就是先把矩阵 F 和 X 向量化，使用向量间导数来计算出矩阵间导数。需要注意的是向量化后的矩阵失去了关于 size 的信息，但是因为矩阵函数中，输入矩阵和函数结果的 size 都是唯一且确定的，所以失去的 size 信息不会造成影响。

关于克罗内克积，见2.4 运算法则。

2.3 补充说明

2.3.1 两种方法的转换

按照2.2 矩阵间导数定义，标量 f 对矩阵 $X_{m\times n}$ 的导数 $\frac{\partial f}{\partial X}$ 是 $mn \times 1$ 向量，和1.3 标量对矩阵的导数和微分中相关的定义不兼容，但是二者可以相互转换。

使用记号 $\triangledown_X f$ 来表示1.3 标量对矩阵的导数和微分中定义的 $m \times n$ 矩阵，于是有：

$$\frac{\partial f}{\partial X} = vec(\triangledown_X f) \tag{13}$$

虽然本部分的方法也可以用于标量对矩阵求导的情况，但是直接使用1.3 标量对矩阵的导数和微分中的方法更加方便。

2.3.2 黑塞矩阵 (Hessian Matrix)

标量 f 对于矩阵 X $(m \times n)$ 的二阶导数，又称为黑塞矩阵，定义如下:

$$\triangledown_X^2 f = \frac{\partial^2 f}{\partial X^2} = \frac{\partial \triangledown_X f}{\partial X} \qquad (mn \times mn) \tag{14}$$

黑塞矩阵是一个对称矩阵，对于向量 $\frac{\partial f}{\partial X}$ 或者矩阵 $\triangledown_X f$ 求导都可以得到黑塞矩阵，但是从矩阵 $\triangledown_X f$ 出发更方便。

2.3.3 定义的利弊

$$\frac{\partial F}{\partial X} = \frac{\partial F}{\partial vec(X)} = \frac{\partial F}{\partial vec(X)} = \frac{\partial vec(F)}{\partial vec(X)} \tag{15}$$

使用上述公式，求导时矩阵被向量化，弊端在于一定程度上破坏了矩阵的结构，会导致结果的形式变得复杂，但是好处在于多元微积分中关于梯度、黑塞矩阵的结论可以沿用过来，只需要将矩阵向量化。

例如在优化问题中，牛顿法的更新 ΔX，满足 $vec(\Delta X) = -(\nabla_X^2 f)^{-1} vec(\triangle_X f)$

2.3.4 其他定义

矩阵对矩阵的导数还有其他的定义，比如:

1. 对于 $F_{p\times q}$ 中的每个元素，用 $X(m\times n)$ 做标量对矩阵求导，结果看作：F 中原来某个元素的位置，替换成一个和 X 同样 size 的矩阵:

$$\frac{\partial F}{\partial X} = \left[\frac{\partial F_{kl}}{\partial X}\right] \qquad (mp\times nq) \tag{16}$$

2. 对于 $X_{m\times n}$ 中的每个元素，用这个元素对于 $F_{p\times q}$ 中的每一个元素做标量对标量求导，那么用这个元素对于 F 求导的结果就是一个和 F 同样 size 的矩阵。结果看作： X 中原来某个元素的位置，替换成一个和 F 同样 size 的矩阵:

$$\frac{\partial F}{\partial X} = \left[\frac{\partial F}{\partial X_{ij}}\right] \qquad (mp\times nq) \tag{17}$$

这两种定义，能够兼容1.3 标量对矩阵的导数和微分中的定义，但是微分和导数的联系不够简明（dF 等于 $\frac{\partial F}{\partial X}$ 中逐个 $m\times n$ 子块分别和 dX 做内积），不便于计算和运用。这些都是糟糕的定义，好的定义必须能够配合微分运算。

2.4 运算法则

我们需要利用导数和微分的联系 $vec(dF) = \frac{\partial F}{\partial X}^T vec(dX)$ 来计算导数，计算矩阵微分的方法同1.4 矩阵微分运算法则，从微分得到导数 需要一些向量化的技巧:

1. 线性：$vec(A+B) = vec(A) + vec(B)$
2. 矩阵乘法：$vec(AXB) = (B^T\otimes A)\cdot vec(X)$，其中 $\otimes$ 表示克罗内克积
   - $A_{m\times n}$ **与** $B_{p\times q}$ **的Kronecker积是：** $A\otimes B = [A_{ij}B] \quad (mp\times nq)$
   - 本等式的证明参见张贤达《矩阵分析与应用》P107～108
3. 转置：$vec(A^T) = K_{mn} vec(A)$，
   - 设A是 $m\times n$ 矩阵，则 K_{mn} 是交换矩阵 (commutation matrix)
   - K_{mn} 将按列优先的向量化变成按行优先的向量化，例如：

$$K_{22} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \qquad vec(A^T) = \begin{bmatrix} A_{11} \\ A_{12} \\ A_{21} \\ A_{22} \end{bmatrix}, \qquad vec(A) = \begin{bmatrix} A_{11} \\ A_{21} \\ A_{21} \\ A_{22} \end{bmatrix}$$

4. 逐元素乘法：$vec(A\odot X) = diag(A)\cdot vec(X)$，其中 $diag(A)$ $(mn\times mn)$ 是用 A 中的元素以按列优先的形式排成的对角阵

2.5 导数的运算

若矩阵函数 F 是矩阵 X 经加减乘法、逆、行列式、逐元素函数等运算构成的：

1. 使用相应的运算法则对 F 求它的微分 dF
2. 对 dF 做向量化得到 $vec(dF)$
3. 其它项交换至 $vec(dX)$ 左侧，对照导数与微分的联系公式 $vec(dF) = \frac{\partial F}{\partial X}^T vec(dX)$ ，即能得到矩阵 F 关于矩阵 X 的导数。

$$vec(dF) = A \cdot vec(X) = \frac{\partial F}{\partial X}^T vec(dX) \tag{18}$$

$$\frac{\partial f}{\partial X} = A^T \tag{19}$$

特别地，若矩阵退化为向量，对照导数与微分的联系 $d\vec{f} = \frac{\partial \vec{f}}{\partial \vec{x}}^T d\vec{x}$，同样能得到导数。

2.6 复合函数求导

假设已知 $\frac{\partial F}{\partial Y}$，而Y是X的函数，如何求 $\frac{\partial F}{\partial X}$？可以从导数和微分的联系入手，进而推出链式法则：

$$vec(dF) = \frac{\partial F}{\partial Y}^T dY = \frac{\partial F}{\partial Y}^T \frac{\partial Y}{\partial X}^T dX = \frac{\partial F}{\partial X}^T vec(dX) \tag{20}$$

$$\frac{\partial F}{\partial X} = (\frac{\partial F}{\partial Y}^T \frac{\partial Y}{\partial X}^T)^T = \frac{\partial Y}{\partial X}\frac{\partial F}{\partial Y} \tag{21}$$

注意链式法则中 $\frac{\partial Y}{\partial X}$ 和 $\frac{\partial F}{\partial Y}$ 的顺序不能对调，这两者是矩阵。

2.7 恒等变形

和标量对矩阵的导数相比，矩阵对矩阵的导数形式更加复杂，从不同角度出发常会得到形式不同的结果。有一些Kronecker积和交换矩阵相关的恒等式，可用来做等价变形：

1. $(A \otimes B)^T = A^T \otimes B^T$
2. $vec(\mathbf{ab}^T) = \mathbf{b} \otimes \mathbf{a}$
3. $(A \otimes B)(C \otimes D) = (AC) \otimes (BD)$
    - 可以对 $F = D^T B^T XAC$ 求导来证明
    - 直接求导得到 $\frac{\partial F}{\partial X} = (AC) \otimes (BD)$
    - 引入 $Y = B^T XA$，有 $\frac{\partial F}{\partial Y} = C \otimes D$，$\frac{\partial Y}{\partial X} = A \otimes B$，由链式法则可以得到 $\frac{\partial F}{\partial X} = (A \otimes B)(C \otimes D)$
4. $K_{mn} = (K_{nm})^T$，$K_{mn}K_{nm} = I$
5. $K_{pm}(A \otimes B)K_{nq} = B \otimes A$，其中 $A\ (m \times n)$， $B\ (p \times q)$。可以对 AXB^T 不同向量化证明：
    - 一方面，$vec(AXB^T) = (B \otimes A) \cdot vec(X)$
    - 另一方面，
      $vec(AXB^T) = K_{pm}vec(BX^T AT^T) = K_{pm}(A \otimes B)vec(X^T) = K_{pm}(A \otimes B)K_{nq} \cdot vec(X)$

2.8 示例

查看III. 参考资料中矩阵对矩阵求导，本笔记来自那篇文章。

原文中除了有上述推理，还包括了矩阵对矩阵求导的例子。

2.9 一个例子

这个例子包含了本文中两种求导方式的差异，以及相互之间的转换，还有一些我自己的理解。说明一些前提条件：

- 标量 L 对矩阵 $X_{m\times n}$ 求导的场合，使用本文中提到的两种方法，结果是不一样的
- 使用1.3 标量对矩阵的导数和微分中定义的方法的场合，结果计作 $\nabla_X f$ 或者 $\frac{\partial L^S}{\partial X} \quad (m\times n)$
- 使用2.2 矩阵对矩阵导数和微分中所定义的方法的场合，结果计作 $\frac{\partial L}{\partial X} \quad (mn\times 1)$
- 根据[2.3.1 两种方法的转换][# 2.3.1 两种方法的转换]，$\frac{\partial L}{\partial X} = vec(\nabla_X f) = vec((\frac{\partial L^S}{\partial X}))$
- 在已知 $X_{m\times n}$ 的 size 的场合，定义 $reshape()$ 函数，把被 vec 向量化的矩阵变回原来的 size，显然 $reshape = vec^{-1}$
- 本问题来自斋藤康毅（日）的《深度学习入门：基于Python的理论与实现》的P146
- 使用 $\frac{\partial L^S}{\partial X}$ 这个写法完全是为了和书上尽量保持类似，书上没有出现矩阵对矩阵求导，所以它使用了 $\frac{\partial L}{\partial X}$ 这样的写法来表示标量对矩阵求导，而不是使用 $\nabla_X f$ 这个符号。为了避免和书上的符号差别太大，就做了上述妥协。

例子：L是某个神经网络的误差，是一个标量，我们现在有一张计算图告诉我们 $Y_{1\times 3} = X_{1\times 2}W_{2\times 3} + B_{1\times 3}$，并且已知 $\frac{\partial L^S}{\partial Y}$ ，求证：

$$(\frac{\partial L^S}{\partial X})_{1\times 2} = (\frac{\partial L^S}{\partial Y})_{1\times 3} \cdot (W^T)_{3\times 2} \tag{2.9.1}$$

$$(\frac{\partial L^S}{\partial W})_{2\times 3} = (X^T)_{2\times 1} \cdot (\frac{\partial L^S}{\partial Y})_{1\times 3} \tag{2.9.2}$$

证明 2.9.1:

$$\begin{gathered}
vec(dY) = vec(dX\cdot W) = vec(I_{1\times 1}\cdot dX_{1\times 2}\cdot W_{2\times 3}) = (W^T\otimes I_{1\times 1})\cdot vec(dX) = W^T\cdot vec(dX) \\
\frac{\partial Y}{\partial X} = (W^T)^T = W \\
\frac{\partial L}{\partial X} = \frac{\partial Y}{\partial X}\frac{\partial L}{\partial Y} \\
\frac{\partial L^S}{\partial X} = reshape(\frac{\partial L}{\partial X}) = (\frac{\partial L}{\partial X})^T = (\frac{\partial Y}{\partial X}\frac{\partial L}{\partial Y})^T = (\frac{\partial L}{\partial Y})^T(\frac{\partial Y}{\partial X})^T = reshape(\frac{\partial L}{\partial Y})\cdot W^T = \frac{\partial L^S}{\partial Y}\cdot W^T \\
PS\text{：这里}reshape\text{函数和}transpose\text{等价只是因为“}\frac{\partial L}{\partial X}\text{和}\frac{\partial L}{\partial Y}\text{这两个式子的结果是行向量”这一巧合罢了}
\end{gathered}$$

证明2.9.2：

III. 参考资料

- 标量对矩阵求导：https://zhuanlan.zhihu.com/p/24709748
- 矩阵对矩阵求导：https://zhuanlan.zhihu.com/p/24863977
- 全微分相关知识：https://www.youtube.com/watch?v=h_iMoNz00Mo
- 克罗内克积：https://zh.wikipedia.org/wiki/克罗内克积
- Latex数学公式使用:
    - 常用符号：http://mohu.org/info/symbols/symbols.htm
    - 在线示例：https://www.codecogs.com/latex/eqneditor.php