闲话矩阵求导

矩阵求导，想必许多领域能见到。统计学，经济学，优化，机器学习等等，在对目标问题建立数学模型之后，问题往往被抽象为关于矩阵的优化问题。于是免不了需要对矩阵进行求导等操作。

简单的向量和矩阵求导，大多数熟悉这些计算的人，应该都能直接写下，然而复杂的矩阵函数求导则没那么简单，著名的matrix cookbook为广大的研究者们提供了一本大字典，里面有着各种简单到复杂矩阵和向量的求导法则，但是如果你的好奇心和我一样重，那么你肯定不会满足于查字典这种方法，特别是在推导公式一气呵成满纸乱飞的时候，查字典岂不是大煞风景？

事实上，所有求导的法则都可以从最基本的求导规则推导出来。不知你有没发现，不同的文献中，同样的式子求导的结果有时候会不一样，仔细观察会发现刚好相差一个转置，于是我们得先说说求导的两个派别（布局）。

1 布局(Layout)

不知道为什么会是这个名字，总之矩阵求导有两种布局，分子布局(numerator layout)和分母布局(denominator layout)。
为了阐明这两种布局的区别，我们先来看最简单的求导规则。
首先是向量yy对标量xx求导，我们假定所有的向量都是列向量，

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}$$

在分子布局下，
在分子布局下，

$$\frac{\partial y}{\partial x}=\begin{bmatrix}\frac{\partial y_1}{\partial x}\\ \frac{\partial y_2}{\partial x}\\ \vdots\\ \frac{\partial y_m}{\partial x}\end{bmatrix}$$

而在分母布局下，

$$\frac{\partial y}{\partial x}=\begin{bmatrix}\frac{\partial y_1}{\partial x} & \frac{\partial y_2}{\partial x} & \cdots & \frac{\partial y_m}{\partial x}\end{bmatrix}$$

2 基本的求导规则（定义）

这一部分，我们将看到一些基本的求导规则，这些与其说是规则，倒不如说是定义。因此这一部分是需要好好理解并且记忆（如果你看一遍还记不住的话）的。
标量y对向量x求导:

$$\frac{\partial y}{\partial x}=\begin{bmatrix}\frac{\partial y}{\partial x_1}\\ \frac{\partial y}{\partial x_2}\\ \vdots\\ \frac{\partial y}{\partial x_m}\end{bmatrix}$$

注意到，标量对向量求导和向量对标量求导刚好反过来。
向量对向量求导，

$$x=\begin{bmatrix}x_1\\ x_2\\ \vdots\\ x_n\end{bmatrix}$$

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}$$

$$\frac{\partial y}{\partial x} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_2}{\partial x_1} & \cdots & \frac{\partial y_m}{\partial x_1} \\ \frac{\partial y_1}{\partial x_2} & \frac{\partial y_2}{\partial x_2} & \cdots & \frac{\partial y_m}{\partial x_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y_1}{\partial x_n} & \frac{\partial y_2}{\partial x_n} & \cdots & \frac{\partial y_m}{\partial x_n} \end{bmatrix}$$

标量对矩阵求导，

$$\frac{\partial y}{\partial X} = \begin{bmatrix} \frac{\partial y}{\partial x_{11}} & \frac{\partial y}{\partial x_{12}} & \cdots & \frac{\partial y}{\partial x_{1q}} \\ \frac{\partial y}{\partial x_{21}} & \frac{\partial y}{\partial x_{22}} & \cdots & \frac{\partial y}{\partial x_{2q}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y}{\partial x_{p1}} & \frac{\partial y}{\partial x_{p2}} & \cdots & \frac{\partial y}{\partial x_{pq}} \end{bmatrix}$$

矩阵对标量求导，

$$\frac{\partial y}{\partial x} = \begin{bmatrix} \frac{\partial y_{11}}{\partial x} & \frac{\partial y_{21}}{\partial x} & \cdots & \frac{\partial y_{m1}}{\partial x} \\ \frac{\partial y_{12}}{\partial x} & \frac{\partial y_{22}}{\partial x} & \cdots & \frac{\partial y_{m2}}{\partial x} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y_{1n}}{\partial x} & \frac{\partial y_{2n}}{\partial x} & \cdots & \frac{\partial y_{mn}}{\partial x} \end{bmatrix}$$

事实上，直观上看，凡是对标量求导，结果的形式都要转置，而标量对向量和矩阵求导则位置保持不动。这样总结方便我们记忆。

总的来说，涉及矩阵和向量的求导不外乎五大类别，
- 向量对标量
- 标量对向量
- 向量对向量
- 矩阵对标量
- 标量对矩阵

这些定义我在上面都已经一一列出。接下来是时候去看一些更加复杂的东西了。

3 维度分析

接下来我们来看一些常见的求导，
首先是$\frac{\partial Ax}{\partial x}$,注意到 $(Ax)_i = a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{in}x_n$,于是利用向量对向量求导法则，我们有

$$\frac{\partial Ax)}{\partial x} = \begin{bmatrix} \frac{\partial (Ax)_1}{\partial x_1} & \frac{\partial (Ax)_2}{\partial x_1} & \cdots & \frac{\partial (Ax)_m}{\partial x_1} \\ \frac{\partial (Ax)_1}{\partial x_2} & \frac{\partial (Ax)_2}{\partial x_2} & \cdots & \frac{\partial (Ax)_m}{\partial x_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial (Ax)_1}{\partial x_n} & \frac{\partial (Ax)_2}{\partial x_n} & \cdots & \frac{\partial (Ax)_m}{\partial x_n} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{21} & \cdots & a_{m1} \\ a_{12} & a_{22} & \cdots & a_{m2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1n} & a_{2n} & \cdots & a_{mn} \end{bmatrix} = A^T$$

理论上对于任意的表达式，我们都可以通过定义出发，利用上面这种形式推导得到。
但是对于一些复杂的求导，这个时候恐怕逐项展开分析就不是很靠谱了。
我们先来看求导分类的前三类，对于这三类问题，我们来看一个非常强大的方法，通过分析维度来得到结果。
考虑$\frac{\partial Au}{\partial x}$,$A$与$x$无关，所以$A$肯定可以先提出求导式，至于去哪了暂时不清楚。
假如$A \in R^{m \times n}, u \in R^{n \times 1}, x \in R^{p \times 1}$
我们知道最后结果肯定和$\frac{\partial u}{\partial x}$有关，注意到$\frac{\partial u}{\partial x} \in R^{p \times n}$,于是$A$只能转置以后添在后面，因此

$$\frac{\partial Au}{\partial x} = \frac{\partial u}{\partial x} A^T$$

同样对于$\frac{\partial au}{\partial x}$,a和x相关的标量，假定$u \in R^{m \times 1}, x \in R^{n \times 1}$根据乘积法则（非精确版本），前一个部分肯定是$a$ $\frac{\partial u}{\partial x}$,后一部分为 $\frac{\partial a}{\partial x} \in R^{n \times 1}$ 和u的某种形式的积,分析维度发现只能是$\frac{\partial a}{\partial x} u^T$

于是

$$\frac{\partial au}{\partial x} = a\frac{\partial u}{\partial x} + \frac{\partial a}{\partial x}u^T$$

我们发现，虽然乘积法则的精准形式无法应用于矩阵求导中，然而这种非精确的乘积法则可以准确的告诉我们哪些项一定会出现在结果中，然后通过分析维度，我们就可以写出结果。
再看$\frac{\partial x^T Ax}{\partial x}$,其中$A$和$x$无关，
为了分析这个问题，我们考虑一个更一般的问题，

$$\frac{\partial x^T Ay}{\partial x}, x \in R^{m\times 1}, y \in R^{n\times 1}$$

我们利用非精确的乘积法则，可以将这个分成两部分

$$\frac{\partial (x^T A)y}{\partial x}$$

于是结果和两部分相关，一个是

$$\frac{\partial y}{\partial x} \in R^{m\times n}$$

,另一个是

$$\frac{\partial x^T A}{\partial x} = A \in R^{m\times n}$$

,同样通过分析维度，我们可以得到

$$\frac{\partial (x^T A)y}{\partial x} = \frac{\partial y}{\partial x}A^T x + Ay$$

因此

$$\frac{\partial x^T Ax}{\partial x} = (A^T + A)x$$

最后看一个式子

$$\frac{\partial a^T x x^T b}{\partial x}, a, b, x \in R^{m \times 1}$$

$$\frac{\partial a^T x x^T b}{\partial x} = \frac{\partial (a^T x)(x^T b)}{\partial x}$$

注意到

$$\frac{\partial (a^T x)}{\partial x} = a, \frac{\partial (x^T b)}{\partial x} = b$$

所以(注意到$x^T b \in R$),

$$\frac{\partial a^T x x^T b}{\partial x} = \frac{\partial (a^T x)(x^T b)}{\partial x} = a x^T b + b a^T x = (a b^T + b a^T) x$$

4 标量对矩阵求导（微分形式）

接下来看五种类型中剩下的两类，在实际的问题中，主要是矩阵的迹对矩阵的求导问题。 正如我们在前面看到的，在矩阵的求导中，不存在精确的乘积法则，我们只是通过非精确的乘积法则分析出单项式中含有的项，再通过维度分析得到结果。 但是，有一种情形下，乘积法则是精确成立的，我们现在就来看这一种情形——迹的微分。因为在微分形式下，

- 乘积法则成立
- 迹和微分可交换

好了，现在你应该已经忘记分子布局了吧，不过不要紧，所有之前的结果转置一下，就得到了分子布局下的结果。

接下来请注意，当我们谈论微分的时候，只有在分子布局下才是有意义的。

（Warning：微分只有分子布局，没有分母布局）

首先我们指出

$$dY = tr(A dX)$$

等价于

$$\frac{\partial Y}{\partial X} = A$$

注意这是分子布局下的，对应分母布局下应该为

$$\frac{\partial Y}{\partial X} = A^T$$

为了方便记忆，防止混淆，我们干脆将

$$dY = tr(AdX)$$

和

$$\frac{\partial Y}{\partial X} = A^T$$

直接等同起来。
于是所有的迹形式对矩阵的求导都先转化为微分形式，比如

$$dtr(AX) = tr(d(AX)) = tr(AdX)$$

其实很简单，我们再看几个例子来加深理解：
先回忆一些非常有用的迹的性质：
- 矩阵的迹和转置的迹相同（转置性质）
- 矩阵乘积的迹和矩阵乘积轮换对称后的迹相同（循环排列）
考虑

$$\begin{aligned}
dtr(X^TAX) &= tr(d(X^TAX)) \\
&= tr(X^TAdX + d(X^TA)X) \\
&= tr(X^TAdX + d(X^TA)X) \\
&= tr(X^TAdX + d(A^TX)^TX) \\
&= tr(X^TAdX) + tr(d(A^TX)^TX) \\
&= tr(X^TAdX) + tr(d(A^TX)^TX) \\
&= tr(X^TAdX) + tr(X^Td(A^TX)) \\
&= tr(X^TAdX) + tr(X^TA^TdX) \\
&= tr(X^TAdX + X^TA^TdX) \\
&= tr((X^TA + X^TA^T)dX)
\end{aligned}$$

所以

$$\frac{\partial tr(X^T AX)}{\partial X} = (X^T A + X^T A^T)^T = (A + A^T)X$$

这是一份简短的矩阵求导介绍，它的目的是告诉你如何更好的快速推导这些公式，避免查阅手册的麻烦。当然如果你觉得你完全是一个工程师，查阅手册感觉很方便，那么继续按照你的方式生活吧。如果你觉得很有用，那么请继续：**Have fun with math!**