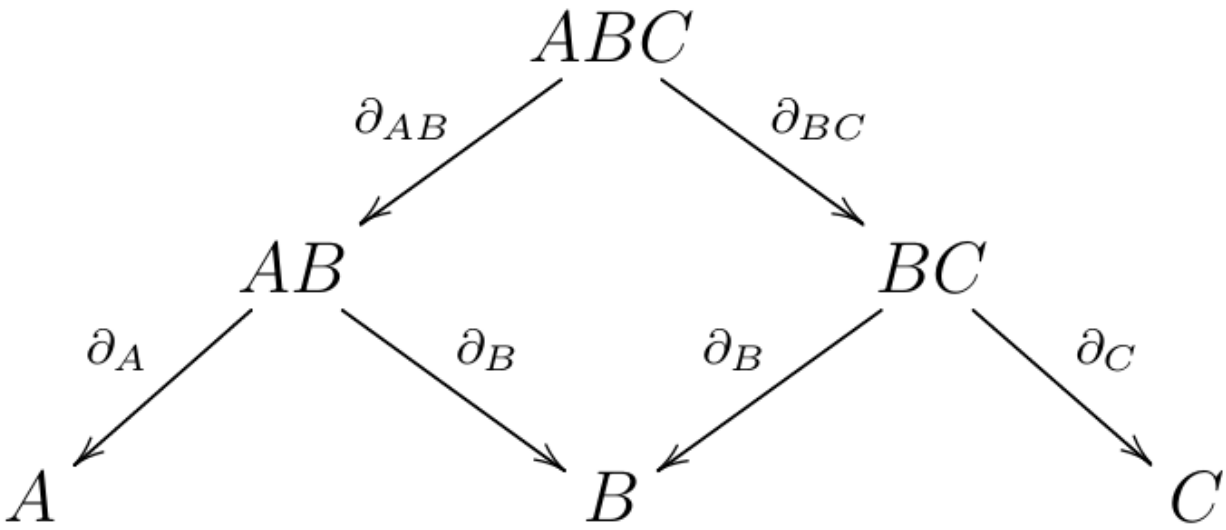


绝对不会出 bug 的矩阵求导——定义，推导，动机；非交换链式法则

街头王子
什么都懂一点，除了数学

关注他

多元微积分已经把多元函数的求导法则发展得淋漓尽致了，若在函数 $y=f(x)$ 中， x,y 都是向量，那么导数 $\frac{dy}{dx}$ 就是一个矩阵，叫做**Jacobi矩阵**. 似乎没有什么需要发展的了，直到近几年大火的深度学习中出现了大量的矩阵求导，有时我们会面对着 x,y 都是矩阵的情况，而且映射 f 大多是矩阵的和积，我暂时称之为“线性的”，这时如果我们沿用老旧的理论，就会发现导数中大量的项是重复的，如果转而使用更紧凑的形式来代替，就会大大提升计算效率，避免不必要的计算资源浪费. 让我们举例说明. 为了书写方便，我会使用爱因斯坦求和约定，所有的求和号都会省略掉，不熟悉的读者可以看我的这篇文章

街头王子：爱因斯坦求和约定 & More
23 赞同 · 5 评论　文章

设 $A=(a_{ij})\in\mathbb{R}^{m\times n}$ ， $B=(b_{ij})\in\mathbb{R}^{n\times k}$ ， $Y=AB=(y_{ij})\in\mathbb{R}^{m\times k}$.

易知 $y_{ik}=a_{ij}b_{jk}$ ，让我们对标量 a_{ij} 求导：

可见，当 $i \neq j$ 时导数一定为0，因此所有含有信息的导数是 $\frac{dy_{ik}}{da_{ij}}$ （注意这里 i 不是哑指标，不需要进行求和），而它的值又和指标 i 无关，因此真正有效的信息仅仅是 b_{jk} ，即 $n \times k$ 个实数值. 而如果用老旧的理论，把 Y, A 看作向量，那么 $\frac{dY}{dA}$ 是一个 $mk \times mn$ 矩阵，这显然是存储空间的浪费. 在矩阵求导理论中，我们定义 $(\frac{dY}{dA})_{kj} = \frac{dy_{ik}}{da_{ij}} = b_{jk}$ ，即 $\frac{dY}{dA} = B^T$ ，具体为什么这样是合理的，我们放到后面说.

现在让我们对 b_{ij} 求导，看看结果会不会不一样：

$$\frac{dy_{ik}}{db_{jl}} = \begin{cases} 0 & k \neq l \\ a_{ij} & k = l \end{cases}$$

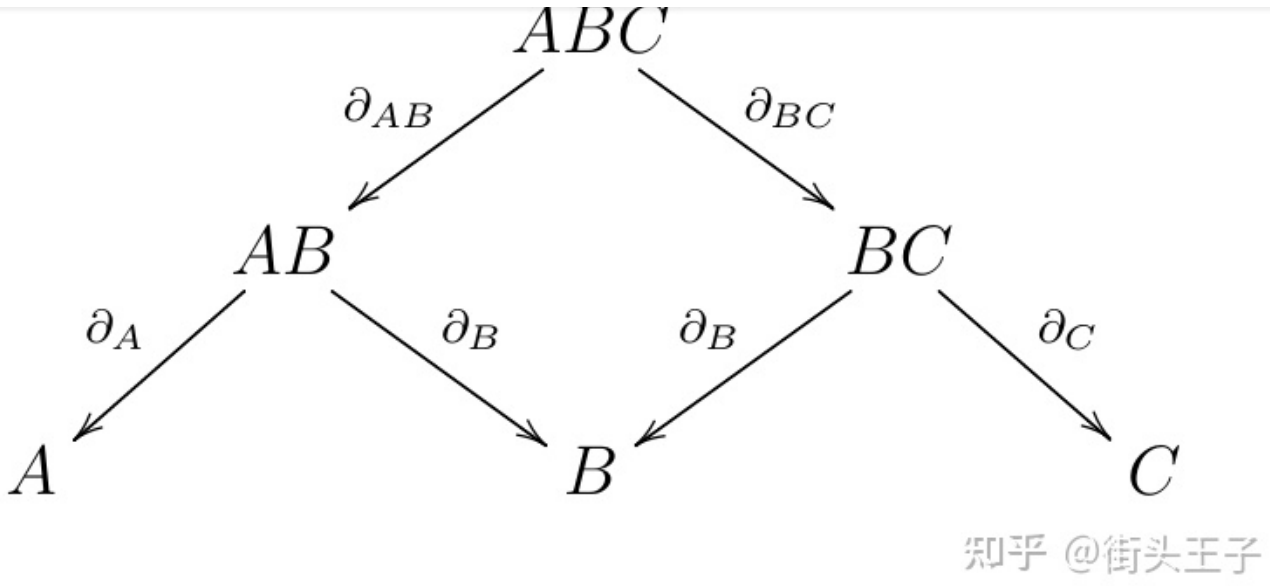
因此仅有 $\frac{dy_{ik}}{db_{jk}}$ 是有效的信息，类似上面，我们定义 $\frac{dY}{dB} = A^T$. 这个转置为什么会出现，现在还没有很好的解释，我们后面会说到.

我们需要的是定义一种只适用于矩阵乘法（加法是自然的）的求导法则，使得矩阵对矩阵的导数仍然是一个矩阵，而不是拥有四个下标的庞然大物. 设 $\mathcal{M}$ 是所有双下标数组（矩阵），则我们要定义这样一个映射

$$d : \mathcal{M} \times \mathcal{M} \to \mathcal{M}$$

并且 $\frac{dY}{dX} := d(Y, X)$.

但这无疑是一个大胆的尝试，必须做到几点，才能保证理论是自洽的.



链式法则的求导顺序；若要ABC对B求导，有两条路径可以选择

首先是链式法则的结合性，假设我们要计算 $\frac{d(ABC)}{dB}$ ，根据链式法则我们可以按 $\frac{d(ABC)}{d(AB)}$ ---> $\frac{d(AB)}{dB}$ 的顺序计算，也可以按 $\frac{d(ABC)}{d(BC)}$ ---> $\frac{d(BC)}{dB}$ 的顺序来. 我们希望按照这两条路径进行求导的结果一样，不然的话，如果把 $A(BC)$ 和 $(AB)C$ 进行求导得到的结果就会不一样，这样矩阵的结合律也丧失了，这是我们不希望看到的.

第二，当输出是标量，也就是1x1矩阵时，我们希望对某个矩阵 B 的导数的形状和它本身相同. 考虑向量的特殊情形，就是梯度和向量本身的维度相等，且行列一致（都是行向量或者都是列向量）. 这个要求的动机来源于机器学习，他们会用复杂的线性代数搭建一个输出为标量的损失函数 $\mathcal{L}$ ，依赖于一些参数 W_i ，通常是矩阵，并希望最小化 $\mathcal{L}$. 用梯度下降法优化参数的公式为

$$W_i \leftarrow W_i - \alpha \frac{d\mathcal{L}}{dW_i}$$

其中 α 是步长. 所以我们当然希望 $\frac{d\mathcal{L}}{dW_i}$ 和 W_i 的维度保持一致了，否则就无法做减法了.

总结一下，我们要定义映射 d 及其链式法则，使得下列条件满足：

1. 按 $\frac{d(ABC)}{d(AB)}$ ---> $\frac{d(AB)}{dB}$ 和按 $\frac{d(ABC)}{d(BC)}$ ---> $\frac{d(BC)}{dB}$ 的顺序求导的结果一致
2. 设 a, c 是（列）向量，则 $\frac{d(a^T Bc)}{dB}$ 和 B 的维度一致.

定义2. （链式法则）

设 A,B,C 是矩阵，则

$$\frac{d(ABC)}{dB}=\frac{d(ABC)}{d(BC)}\frac{d(BC)}{dB}=\frac{d(AB)}{dB}\frac{d(ABC)}{dAB}$$

我们需要验证定义2和定义1的兼容性，将定义1代入上式，得

$\frac{d(ABC)}{d(BC)}=A^T,\frac{d(ABC)}{d(AB)}=C^T,\frac{d(BC)}{d(B)}=C^T,\frac{d(AB)}{d(B)}=A^T$ ，故

$$\frac{d(ABC)}{d(BC)}\frac{d(BC)}{dB}=A^TC^T=\frac{d(AB)}{dB}\frac{d(ABC)}{dAB}$$

当我们把定义1的式子代入两种表达式中时，它们求得的结果都是 A^TC^T ，从而两种定义1的存在不会影响定义2的等式的成立，故它们是兼容的.

定理3.（导数的良定义）

设 $A_1,\dots,A_n$ 是一列矩阵，则 $\frac{d(A_1\cdots A_n)}{dA_j}$ 被唯一地定义， $\forall 1\le j\le n$.

证明. 两个矩阵的情形由定义1直接给出. 假设命题对n-1个矩阵成立，则

$$\frac{d(A_1\cdots A_n)}{dA_j}=\frac{d([A_1\cdots A_{j-1}]A_j[A_{j+1}\cdots A_n])}{dA_j}$$

$$=\frac{d(A_1\cdots A_n)}{d(A_j\cdots A_n)}\frac{d(A_j\cdots A_n)}{dA_j}=\frac{d(A_1\cdots A_j)}{dA_j}\frac{d(A_1\cdots A_n)}{d(A_1\cdots A_j)}$$

而在最后一行中，若将分母看成单个矩阵，则分子可成为 $\le n-1$ 个矩阵之积，从而被唯一定义，故任意n个矩阵之积对其中某个矩阵的导数，如上所述，被唯一定义，证毕.

定理4.（导数的维度匹配）

设 $A_1,\dots,A_n$ 是一列矩阵， b,c 是列向量，则 $\frac{d(b^TA_1\cdots A_nc)}{dA_j}$ 的形状与 A_j 相同，

$\forall 1\le j\le n$.

dA

与 A 的形状相同，证毕.

综上，我们定义了当函数的输入和输出都是矩阵，且函数关系简化为数个矩阵的乘积时的矩阵求导法则和非交换的链式法则，在要求保持矩阵乘法的结合律，并且当输出是标量是导数的维度和原矩阵匹配的情况下，这种定义方式几乎是唯一的. 由于不需要将输入和输出空间中的矩阵展开成向量，且所得的导数仍然是一个矩阵，存储空间的要求被大大缩小了，适用于实际场景中对机器学习模型快速求导的要求，事实上这是目前正在广泛使用的求导方式. 然而需注意，只有当输入和输出之间仅由矩阵的乘法和加法决定的时候，这种快速计算的方式才能生效. 当输入和输出之间呈一般光滑函数的关系时，信息量无法压缩成一个小矩阵. 压缩了空间，就必然带来适用范围的减小，但在一种特定的使用场景，总能发挥独特的作用.

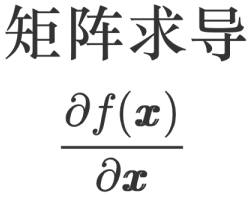
编辑于 02-23

机器学习　矩阵计算　多元微积分

推荐阅读

矩阵求导公式

小雨姑娘　发表于小雨姑娘的...



对称矩阵的求导，以多元正态分布的极大似然估计为例（矩阵...

Iterator

机器学习

“矩阵求
区域。虽
所讲的多
些公式以
是必要的
而且易出

Towser

1 条评论　切换为时间排序

🤣🤣🤣

赞