向量、矩阵与张量的基本求导方法

姚兴虎

2019 年 5 月 15 日

本文并非是作为工具书来查询各种各样的向量矩阵与张量之间的求导公式，而是试图从最简单的层面来对这些求导公式的推导过程进行介绍。

1 将导数表达式简化为最简单的形式

涉及向量与矩阵之间的求导的运算难点往往在于你要一次性的完成多项任务。这些任务包括并行的对对应元素进行求导，考虑表达式中的求和符号以及连乘符号，考虑链式法则。将一个看起来很棘手的求导问题分解为最简单的形式有助于我们仔细理解计算的整个过程。

1.1 考虑单个分量的导数并对求和符号展开

当拿到一个复杂的矩阵求导问题时，首先考虑其每个标量元素对其他标量元素的导数值往往能够简化问题。当将一个复杂的问题转化到最简单的标量间的求导问题时，我们能够对矩阵间的求和、微分等数学运算进行简化。

例如我们考虑这样的一个问题: 假设我们有一个维数为 $C\times D$ 的矩阵 W, 一个长度为 D 的列向量 $\mathbf{x}=[x_1,x_2,\cdots,x_D]^T$, 一个长度为 C 的列向量 $\mathbf{y}=[y_1,y_2,\cdots,y_C]^T$，满足如下的关系，

$$\mathbf{y}=W\mathbf{x}.$$

我们要求向量 $\mathbf{y}$ 对向量 $\mathbf{x}$ 的导数。该求导运算的结果是由 $\mathbf{x}$ 的每个分量对 $\mathbf{y}$ 的每个分量组合而成，因此求导后会得到一个规模为 $C\times D$ 的矩阵。我们考虑导数的一个分量 $\frac{\partial y_3}{\partial x_7}$ 的计算，其为一个标量对另一个标量的求导运算。我们只需要找到 y_3 与 x_7 的关系，然后按照最普通的求导法则进行计算即可。事实上，我们注意到:

$$y_3=\sum_{j=1}^{D}W_{3,j}x_j$$

通过这一转化，我么便将原始的矩阵表达式转换为普通的标量表达式。

1.2 去掉求和符号

尽管从上面的表达式中直接计算导数并不困难，但是当求和符号 $\sum$ 和连乘符号 $\prod$ 里面的内容较为复杂时，将其展开将会有助于我们的计算。于是我们可以将连乘符号展开以确保我们的每步计算都精确无误，从而可以得到:

$$y_3=W_{3,1}x_1+W_{3,2}x_2+\cdots+W_{3,7}x_7+\cdots+W_{3,D}x_D.$$

这一公式就是最简单的线性等式，于是我们有:

$$\begin{aligned}\frac{\partial y_3}{\partial x_7} &= \frac{\partial}{\partial x_7}[W_{3,1}x_1 + W_{3,2}x_2 + \cdots + W_{3,7}x_7 + \cdots + W_{3,D}x_D] \\ &= W_{3,7}.\end{aligned}$$

通过关注 $\mathbf{x}$ 和 $\mathbf{y}$ 的单个分量之间的关系，我们能将求导的计算过程化简为最简单的导数计算问题，当我们在计算复杂的导数时，通过这一方法可以帮助我们理清思路。

1.3 补全计算结果

我们最初的目的是为了计算向量 $\mathbf{y}$ 对向量 $\mathbf{x}$ 的导数，通过上面的分析我们注意到，其计算结果是一个维数为 $C\times D$ 的矩阵，该矩阵具有如下的形式:

$$\begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \frac{\partial y_1}{\partial x_3} & \cdots & \frac{\partial y_1}{\partial x_D} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \frac{\partial y_2}{\partial x_3} & \cdots & \frac{\partial y_2}{\partial x_D} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y_C}{\partial x_1} & \frac{\partial y_C}{\partial x_2} & \frac{\partial y_C}{\partial x_3} & \cdots & \frac{\partial y_C}{\partial x_D} \end{bmatrix}$$

在我们的这个例子中，这一矩阵被称为 *Jacobian matrix*. 我们对每个分量执行上面一小节的对应过程可以得到如下的关系:

$$\frac{\partial y_i}{\partial x_j} = W_{i,j}.$$

这意味着，导数计算结果所对应的矩阵为:

$$\frac{d\mathbf{y}}{d\mathbf{x}} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \frac{\partial y_1}{\partial x_3} & \cdots & \frac{\partial y_1}{\partial x_D} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \frac{\partial y_2}{\partial x_3} & \cdots & \frac{\partial y_2}{\partial x_D} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y_C}{\partial x_1} & \frac{\partial y_C}{\partial x_2} & \frac{\partial y_C}{\partial x_3} & \cdots & \frac{\partial y_C}{\partial x_D} \end{bmatrix} = \begin{bmatrix} W_{1,1} & W_{1,2} & W_{1,3} & \cdots & W_{1,D} \\ W_{2,1} & W_{2,2} & W_{2,3} & \cdots & W_{2,D} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ W_{C,1} & W_{C,2} & W_{C,3} & \cdots & W_{C,D} \end{bmatrix}.$$

也就是说，经过这些详细的计算过程，我们可以得到对于 $\mathbf{y} = W\mathbf{x}$，我们有:

$$\frac{d\mathbf{y}}{d\mathbf{x}} = W.$$

1.4 行向量间的求导

行向量对行向量的求导过程与上面的分析过程没有本质的区别，我们考虑，

$$\mathbf{y} = \mathbf{x}W.$$

其中，$\mathbf{y} = [y_1, y_2, \cdots, y_C]$ 是长度为 C 的行向量，$\mathbf{x} = [x_1, x_2, \cdots, x_D]$ 是长度为 D 的行向量，W 的维数为 $D\times C$. 同样，我们有，

$$y_3 = \sum_{j=1}^{D} x_j W_{j,3}.$$

于是，

$$\frac{\partial y_3}{\partial x_7} = W_{7,3}.$$

从而，

$$\frac{d\mathbf{y}}{d\mathbf{x}} = W.$$

1.5 处理更高的维度

我们同样考虑，

$$\mathbf{y} = \mathbf{x}W.$$

其中，$\mathbf{y} = [y_1, y_2, \cdots, y_C]$ 是长度为 C 的行向量，$\mathbf{x} = [x_1, x_2, \cdots, x_D]$ 是长度为 D 的行向量，W 的维数为 $D{\times}C$. 在这一小节，我么考虑行向量 y 对矩阵 W 的求导表达式:

$$\frac{d\mathbf{y}}{dW}.$$

在这种情形下，$\mathbf{y}$ 沿着一个坐标轴变化，W 沿着两个坐标轴变化，因此很自然的会想到其导数计算结果应该是一个三维张量，事实上，经过之后的推导我们可以发现这一求导结果完全可以用二维的矩阵进行表示。

我们还是首先从向量 $\mathbf{y}$ 的一个分量对矩阵进行求导开始进行推导，事实上，我们有:

$$y_j = x_1 W_{1,j} + x_2 W_{2,j} + \cdots + x_D W_{D,j}.$$

于是我们可以看出:

$$\frac{\partial y_j}{\partial W_{i.j}} = x_i,$$

而 y_j 对矩阵中其他元素的偏导数为 0. 于是我们可以将求导结果定义为新的二维矩阵:

$$\frac{d\mathbf{y}}{dW} = \begin{bmatrix} \frac{\partial y_1}{\partial W_{1,1}} & \frac{\partial y_1}{\partial W_{2,1}} & \frac{\partial y_1}{\partial W_{3,1}} & \cdots & \frac{\partial y_1}{\partial W_{D,1}} \\ \frac{\partial y_2}{\partial W_{1,2}} & \frac{\partial y_2}{\partial W_{2,2}} & \frac{\partial y_2}{\partial W_{3,2}} & \cdots & \frac{\partial y_2}{\partial W_{D,2}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y_C}{\partial W_{1,C}} & \frac{\partial y_C}{\partial W_{2,C}} & \frac{\partial y_C}{\partial W_{3,C}} & \cdots & \frac{\partial y_C}{\partial W_{D,C}} \end{bmatrix} = \begin{bmatrix} x_1 & x_2 & x_3 & \cdots & x_D \\ x_1 & x_2 & x_3 & \cdots & x_D \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_1 & x_2 & x_3 & \cdots & x_D \end{bmatrix}.$$

可以看出行向量对矩阵的求导结果同样可以用一个矩阵来表示，这种表示方法在实现神经网络时非常有用。

1.6 矩阵对矩阵求导 (多个数据点)

我们考虑将多个数据点进行叠加从而得到如下的关系式:

$$Y = XW.$$

其中 X 的维度是 $N{\times}D$,W 的维度是 $D{\times}C$,Y 的维度是 $N{\times}C$, 并且矩阵 Y 的每行是由矩阵 X 的对应行的元素经过线性变换得到的。同样考虑其中的分量我们可以得到:

$$Y_{i,j} = \sum_{k=1}^{D} X_{i,k} W_{k,j}.$$

从这一表达式中我们可以看出对于偏微分

$$\frac{\partial Y_{a,b}}{\partial X_{c,d}},$$

当 $a \neq c$ 时值均为 0. 这也就是说，既然 Y 的当前行是由 X 的对应行算出来的，那么当求导时行数不一致时其导数值均为 0. 于是我们可以得到

$$\frac{\partial Y_{i,j}}{\partial X_{i,k}} = W_{k,j}$$

从而我们有:

$$\frac{\partial Y_{i,:}}{\partial X_{i,:}} = W.$$

这一结果是上一小节结果的自然推广。

1.7 将链式法则考虑进去

在这里我们考虑两个列向量 $\mathbf{x}, \mathbf{y}$ 之间的线性变换，其关系如下所示：

$$\mathbf{y} = VW\mathbf{x},$$

我们想要计算向量 $\mathbf{y}$ 对向量 $\mathbf{x}$ 的导数值。我们可以简单的观察到矩阵 V 与矩阵 W 的乘积依然是一个矩阵，因此借用上面的结论我们立马能得到求导结果:

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = VW = U.$$

在这里为了对链式法则进行应用我们采取一种更为复杂的方式进行求导运算, 我们先定义中间变量 $\mathbf{m} = W\mathbf{x}$, 其中 $\mathbf{m}$ 具有 M 个分量，于是我们有 $\mathbf{y} = V\mathbf{m}$, 根据链式法则我们可以得到:

$$\frac{d\mathbf{y}}{d\mathbf{x}} = \frac{d\mathbf{y}}{d\mathbf{m}} \frac{d\mathbf{m}}{d\mathbf{x}}$$

我们考虑单个分量之间的标量求导，可以写为

$$\frac{dy_i}{dx_j} = \frac{dy_i}{d\mathbf{m}} \frac{d\mathbf{m}}{dx_j} = \sum_{k=1}^{M} \frac{dy_i}{dm_k} \frac{dm_k}{dx_j} = \sum_{k=1}^{M} V_{i,k} W k, j$$

将上述分量合并为矩阵便可得到同样的求导结果 VW。当涉及到利用链式法则进行求导运算时，我们可以采取类似的方法首相将各个中间变量找出来，然后考虑单个分量之间的求导计算，最后将涉及到的相关中间结果进行相加即可。