矩阵求导术（上）

长躯鬼侠
数学爱好者

矩阵求导的技术，在统计学、控制论、机器学习等领域有广泛的应用。鉴于我看过的一些资料或言之不详、或繁乱无绪，本文来做个科普，分作两篇，上篇讲标量对矩阵的求导术，下篇讲矩阵对矩阵的求导术。本文使用小写字母x表示标量，粗体小写字母 $\boldsymbol{x}$ 表示（列）向量，大写字母X表示矩阵。

首先来琢磨一下定义，标量f对矩阵X的导数，定义为 $\frac{\partial f}{\partial X}=\left[\frac{\partial f}{\partial X_{ij}}\right]$，即f对X逐元素求导排成与X尺寸相同的矩阵。然而，这个定义在计算中并不好用，实用上的原因是对函数较复杂的情形难以逐元素求导；哲理上的原因是逐元素求导破坏了**整体性**。试想，为何要将f看做矩阵X而不是各元素 X_{ij} 的函数呢？答案是用矩阵运算更整洁。所以在求导时不宜拆开矩阵，而是要找一个从整体出发的算法。

为此，我们来回顾，一元微积分中的导数（标量对标量的导数）与微分有联系：$df=f'(x)dx$；多元微积分中的梯度（标量对向量的导数）也与微分有联系：$df=\sum_{i=1}^{n}\frac{\partial f}{\partial x_i}dx_i=\frac{\partial f}{\partial \boldsymbol{x}}^T d\boldsymbol{x}$，这里第一个等号是全微分公式，第二个等号表达了梯度与微分的联系：全微分 df 是梯度向量 $\frac{\partial f}{\partial \boldsymbol{x}}$ (n×1)与微分向量 $d\boldsymbol{x}$ (n×1)的内积；受此启发，我们将矩阵导数与微分建立联系：$df=\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{\partial f}{\partial X_{ij}}dX_{ij}=\text{tr}\left(\frac{\partial f}{\partial X}^T dX\right)$。其中tr代表迹(trace)是方阵对角线元素之和，满足性质：对尺寸相同的矩阵A,B，$\text{tr}(A^TB)=\sum_{i,j}A_{ij}B_{ij}$，即 $\text{tr}(A^TB)$ 是矩阵A,B的**内积**。与梯度相似，这里第一个等号是全微分公式，第二个等号表达了矩阵导数与微分的联系：全微分 df 是导数 $\frac{\partial f}{\partial X}$ (m×n)与微分矩阵 dX (m×n)的内积。

然后来建立运算法则。回想遇到较复杂的一元函数如 $f=\log(2+\sin x)e^{\sqrt{x}}$，我们是如何求导的呢？通常不是从定义开始求极限，而是先建立了初等函数求导和四则运算、复合等法则，再来运用这些法则。故而，我们来创立常用的矩阵微分的运算法则：

1. 加减法：$d(X\pm Y)=dX\pm dY$；

$d(X^T) = (dX)^T$；迹：$d\text{tr}(X) = \text{tr}(dX)$。
2. 逆：$dX^{-1} = -X^{-1}dXX^{-1}$。此式可在 $XX^{-1} = I$ 两侧求微分来证明。
3. 行列式：$d|X| = \text{tr}(X^{\#}dX)$，其中 $X^{\#}$ 表示X的伴随矩阵，在X可逆时又可以写作 $d|X| = |X|\text{tr}(X^{-1}dX)$。此式可用Laplace展开来证明，详见张贤达《矩阵分析与应用》第279页。
4. 逐元素乘法：$d(X \odot Y) = dX \odot Y + X \odot dY$，$\odot$ 表示尺寸相同的矩阵X,Y逐元素相乘。
5. 逐元素函数：$d\sigma(X) = \sigma'(X) \odot dX$，$\sigma(X) = [\sigma(X_{ij})]$ 是逐元素标量函数运算，$\sigma'(X) = [\sigma'(X_{ij})]$ 是逐元素求导数。例如 $X = \begin{bmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{bmatrix}, d\sin(X) = \begin{bmatrix} \cos X_{11}dX_{11} & \cos X_{12}dX_{12} \\ \cos X_{21}dX_{21} & \cos X_{22}dX_{22} \end{bmatrix} = \cos(X) \odot dX$。

我们试图利用矩阵导数与微分的联系 $df = \text{tr}\left(\frac{\partial f}{\partial X}^T dX\right)$，在求出左侧的微分 df 后，该如何写成右侧的形式并得到导数呢？这需要一些迹技巧(trace trick)：

1. 标量套上迹：$a = \text{tr}(a)$
2. 转置：$\text{tr}(A^T) = \text{tr}(A)$。
3. 线性：$\text{tr}(A \pm B) = \text{tr}(A) \pm \text{tr}(B)$。
4. 矩阵乘法交换：$\text{tr}(AB) = \text{tr}(BA)$，其中 A 与 B^T 尺寸相同。两侧都等于 $\sum_{i,j} A_{ij}B_{ji}$。
5. 矩阵乘法/逐元素乘法交换：$\text{tr}(A^T(B \odot C)) = \text{tr}((A \odot B)^T C)$，其中 A, B, C 尺寸相同。两侧都等于 $\sum_{i,j} A_{ij}B_{ij}C_{ij}$。

观察一下可以断言，**若标量函数f是矩阵X经加减乘法、逆、行列式、逐元素函数等运算构成，则使用相应的运算法则对f求微分，再使用迹技巧给df套上迹并将其它项交换至dX左侧，对照导数与微分的联系 $df = \text{tr}\left(\frac{\partial f}{\partial X}^T dX\right)$，即能得到导数。**

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

在建立法则的最后，来谈一谈复合：假设已求得 $\frac{\partial f}{\partial Y}$，而Y是X的函数，如何求 $\frac{\partial f}{\partial X}$ 呢？在微积分中有标量求导的链式法则 $\frac{\partial f}{\partial x} = \frac{\partial f}{\partial y}\frac{\partial y}{\partial x}$，但这里我们**不能随意沿用标量的链式法则**，因为矩阵对矩阵的导数 $\frac{\partial Y}{\partial X}$ 截至目前仍是未定义的。于是我们继续追本溯源，链式法则是从何而来？源头仍然是微分。我们直接从微分入手建立复合法则：先写出 $df = \text{tr}\left(\frac{\partial f}{\partial Y}^T dY\right)$，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 $\frac{\partial f}{}$

最常见的情形是 $Y = AXB$，此时 $df = \text{tr}\left(\frac{\partial f}{\partial Y}^T dY\right) = \text{tr}\left(\frac{\partial f}{\partial Y}^T AdXB\right) = \text{tr}\left(B\frac{\partial f}{\partial Y}^T AdX\right) = \text{tr}\left((A^T\frac{\partial f}{\partial Y}B^T)^T dX\right)$，可得到 $\frac{\partial f}{\partial X} = A^T\frac{\partial f}{\partial Y}B^T$。注意这里 $dY = (dA)XB + AdXB + AXdB = AdXB$，由于 A, B 是常量，$dA = 0, dB = 0$，以及我们使用矩阵乘法交换的迹技巧交换了 $\frac{\partial f}{\partial Y}^T AdX$ 与 B。

接下来演示一些算例。特别提醒要依据已经建立的运算法则来计算，不能随意套用微积分中标量导数的结论，比如认为AX对X的导数为A，这是没有根据、意义不明的。

例1：$f = \boldsymbol{a}^T X\boldsymbol{b}$，求 $\frac{\partial f}{\partial X}$。其中 $\boldsymbol{a}$ 是 $m\times 1$ 列向量，X 是 $m\times n$ 矩阵，$\boldsymbol{b}$ 是 $n\times 1$ 列向量，f 是标量。

解：先使用矩阵乘法法则求微分，$df = d\boldsymbol{a}^T X\boldsymbol{b} + \boldsymbol{a}^T dX\boldsymbol{b} + \boldsymbol{a}^T Xd\boldsymbol{b} = \boldsymbol{a}^T dX\boldsymbol{b}$，注意这里的 $\boldsymbol{a}, \boldsymbol{b}$ 是常量，$d\boldsymbol{a} = \boldsymbol{0}, d\boldsymbol{b} = \boldsymbol{0}$。由于df是标量，它的迹等于自身，$df = \text{tr}(df)$，套上迹并做矩阵乘法交换：$df = \text{tr}(\boldsymbol{a}^T dX\boldsymbol{b}) = \text{tr}(\boldsymbol{b}\boldsymbol{a}^T dX) = \text{tr}((\boldsymbol{a}\boldsymbol{b}^T)^T dX)$，注意这里我们根据 $\text{tr}(AB) = \text{tr}(BA)$ 交换了 $\boldsymbol{a}^T dX$ 与 $\boldsymbol{b}$。对照导数与微分的联系 $df = \text{tr}\left(\frac{\partial f}{\partial X}^T dX\right)$，得到 $\frac{\partial f}{\partial X} = \boldsymbol{a}\boldsymbol{b}^T$。

注意：这里不能用 $\frac{\partial f}{\partial X} = \boldsymbol{a}^T\frac{\partial X}{\partial X}\boldsymbol{b} = ?$，导数与矩阵乘法的交换是不合法则的运算（而微分是合法的）。有些资料在计算矩阵导数时，会略过求微分这一步，这是逻辑上解释不通的。

例2：$f = \boldsymbol{a}^T\exp(X\boldsymbol{b})$，求 $\frac{\partial f}{\partial X}$。其中 $\boldsymbol{a}$ 是 $m\times 1$ 列向量，X 是 $m\times n$ 矩阵，$\boldsymbol{b}$ 是 $n\times 1$ 列向量，exp表示逐元素求指数，f 是标量。

解：先使用矩阵乘法、逐元素函数法则求微分：$df = \boldsymbol{a}^T(\exp(X\boldsymbol{b})\odot(dX\boldsymbol{b}))$，再套上迹并做交换：$df = \text{tr}(\boldsymbol{a}^T(\exp(X\boldsymbol{b})\odot(dX\boldsymbol{b}))) = \text{tr}((\boldsymbol{a}\odot\exp(X\boldsymbol{b}))^T dX\boldsymbol{b})$
$= \text{tr}(\boldsymbol{b}(\boldsymbol{a}\odot\exp(X\boldsymbol{b}))^T dX) = \text{tr}(((\boldsymbol{a}\odot\exp(X\boldsymbol{b}))\boldsymbol{b}^T)^T dX)$，注意这里我们先根据 $\text{tr}(A^T(B\odot C)) = \text{tr}((A\odot B)^T C)$ 交换了 $\boldsymbol{a}$、$\exp(X\boldsymbol{b})$ 与 $dX\boldsymbol{b}$，再根据 $\text{tr}(AB) = \text{tr}(BA)$ 交换了 $(\boldsymbol{a}\odot\exp(X\boldsymbol{b}))^T dX$ 与 $\boldsymbol{b}$。对照导数与微分的联系 $df = \text{tr}\left(\frac{\partial f}{\partial X}^T dX\right)$，得到 $\frac{\partial f}{\partial X} = (\boldsymbol{a}\odot\exp(X\boldsymbol{b}))\boldsymbol{b}^T$。

例3：$f = \text{tr}(Y^T MY), Y = \sigma(WX)$，求 $\frac{\partial f}{\partial X}$。其中 W 是 $l\times m$ 矩阵，X 是 $m\times n$ 矩阵，Y 是 $l\times n$ 矩阵，M 是 $l\times l$ 对称矩阵，σ 是逐元素函数，f 是标量。

解：先求 $\frac{\partial f}{\partial Y}$，求微分，使用矩阵乘法、转置法则：$df = \text{tr}((dY)^T MY) + \text{tr}(Y^T MdY) = \text{tr}(Y^T M^T dY) + \text{tr}(Y^T MdY) = \text{tr}(Y^T(M + M^T)dY)$，对照导数与微分的联系，得到 $\frac{\partial f}{\partial Y} = (M + M^T)Y = 2MY$，注意这里M是对称矩阵。为求 $\frac{\partial f}{\partial X}$，写出 $df = \text{tr}\left(\frac{\partial f}{\partial Y}^T dY\right)$，再将dY用dX表示出来代入，并使用矩阵乘法/逐元素乘法交换：$df = \text{tr}\left(\frac{\partial f}{\partial Y}^T (\sigma'(WX) \odot (WdX))\right) = \text{tr}\left(\left(\frac{\partial f}{\partial Y} \odot \sigma'(WX)\right)^T WdX\right)$，对照导数与微分的联系，得到 $\frac{\partial f}{\partial X} = W^T \left(\frac{\partial f}{\partial Y} \odot \sigma'(WX)\right) = W^T((2M\sigma(WX)) \odot \sigma'(WX))$。

例4【线性回归】：$l = \|X\boldsymbol{w} - \boldsymbol{y}\|^2$， 求 $\boldsymbol{w}$ 的最小二乘估计，即求 $\frac{\partial l}{\partial \boldsymbol{w}}$ 的零点。其中 $\boldsymbol{y}$ 是 $m \times 1$ 列向量，X 是 $m \times n$ 矩阵，$\boldsymbol{w}$ 是 $n \times 1$ 列向量，l 是标量。

解：这是标量对向量的导数，不过可以把向量看做矩阵的特例。先将向量模平方改写成向量与自身的内积：$l = (X\boldsymbol{w} - \boldsymbol{y})^T(X\boldsymbol{w} - \boldsymbol{y})$，求微分，使用矩阵乘法、转置等法则：$dl = (Xd\boldsymbol{w})^T(X\boldsymbol{w} - \boldsymbol{y}) + (X\boldsymbol{w} - \boldsymbol{y})^T(Xd\boldsymbol{w}) = 2(X\boldsymbol{w} - \boldsymbol{y})^T Xd\boldsymbol{w}$，注意这里 $Xd\boldsymbol{w}$ 和 $X\boldsymbol{w} - \boldsymbol{y}$ 是向量，两个向量的内积满足 $\boldsymbol{u}^T\boldsymbol{v} = \boldsymbol{v}^T\boldsymbol{u}$。对照导数与微分的联系 $dl = \frac{\partial l}{\partial \boldsymbol{w}}^T d\boldsymbol{w}$，得到 $\frac{\partial l}{\partial \boldsymbol{w}} = 2X^T(X\boldsymbol{w} - \boldsymbol{y})$。$\frac{\partial l}{\partial \boldsymbol{w}} = \boldsymbol{0}$ 即 $X^T X\boldsymbol{w} = X^T\boldsymbol{y}$，得到 $\boldsymbol{w}$ 的最小二乘估计为 $\boldsymbol{w} = (X^T X)^{-1}X^T\boldsymbol{y}$。

例5【方差的最大似然估计】：样本 $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$，求方差 Σ 的最大似然估计。写成数学式是：$l = \log|\Sigma| + \frac{1}{N}\sum_{i=1}^N (\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T\Sigma^{-1}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})$，求 $\frac{\partial l}{\partial \Sigma}$ 的零点。其中 $\boldsymbol{x}_i$ 是 $m \times 1$ 列向量，$\bar{\boldsymbol{x}} = \frac{1}{N}\sum_{i=1}^N \boldsymbol{x}_i$ 是样本均值，Σ 是 $m \times m$ 对称正定矩阵，l 是标量，log表示自然对数。

解：首先求微分，使用矩阵乘法、行列式、逆等运算法则，第一项是 $d\log|\Sigma| = |\Sigma|^{-1}d|\Sigma| = \text{tr}(\Sigma^{-1}d\Sigma)$，第二项是 $\frac{1}{N}\sum_{i=1}^N (\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T d\Sigma^{-1}(\boldsymbol{x}_i - \bar{\boldsymbol{x}}) = -\frac{1}{N}\sum_{i=1}^N (\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T\Sigma^{-1}d\Sigma\Sigma^{-1}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})$。再给第二项套上迹做交换：$\text{tr}\left(\frac{1}{N}\sum_{i=1}^N (\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T\Sigma^{-1}d\Sigma\Sigma^{-1}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})\right) = \frac{1}{N}\sum_{i=1}^N \text{tr}((\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T\Sigma^{-1}d\Sigma\Sigma^{-1}(\boldsymbol{x}_i - \bar{\boldsymbol{x}}))$ $= \frac{1}{N}\sum_{i=1}^N \text{tr}\left(\Sigma^{-1}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T\Sigma^{-1}d\Sigma\right) = \text{tr}(\Sigma^{-1}S\Sigma^{-1}d\Sigma)$，其中先交换迹与求和，然后将 $\Sigma^{-1}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})$ 交换到左边，最后再交换迹与求和，并定义 $S = \frac{1}{N}\sum_{i=1}^N (\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T$ 为样本方

差矩阵。得到 $dl = \text{tr}\left(\left(\Sigma^{-1} - \Sigma^{-1}S\Sigma^{-1}\right)d\Sigma\right)$ 。对照导数与微分的联系，有 $\frac{\partial l}{\partial \Sigma} = (\Sigma^{-1} - \Sigma^{-1}S\Sigma^{-1})^T$ ，其零点即 Σ 的最大似然估计为 $\Sigma = S$ 。

例6【多元logistic回归】： $l = -\boldsymbol{y}^T \log \text{softmax}(W\boldsymbol{x})$ ，求 $\frac{\partial l}{\partial W}$ 。其中 $\boldsymbol{y}$ 是除一个元素为1外其它元素为0的 $m \times 1$ 列向量， W 是 $m \times n$ 矩阵， $\boldsymbol{x}$ 是 $n \times 1$ 列向量， l 是标量；log表示自然对数， $\text{softmax}(\boldsymbol{a}) = \frac{\exp(\boldsymbol{a})}{\boldsymbol{1}^T \exp(\boldsymbol{a})}$ ，其中 $\exp(\boldsymbol{a})$ 表示逐元素求指数， $\boldsymbol{1}$ 代表全1向量。

解1：首先将softmax函数代入并写成 $l = -\boldsymbol{y}^T\left(\log(\exp(W\boldsymbol{x})) - \boldsymbol{1}\log(\boldsymbol{1}^T\exp(W\boldsymbol{x}))\right) = -\boldsymbol{y}^T W\boldsymbol{x} + \log(\boldsymbol{1}^T\exp(W\boldsymbol{x}))$ ，这里要注意逐元素log满足等式 $\log(\boldsymbol{u}/c) = \log(\boldsymbol{u}) - \boldsymbol{1}\log(c)$ ，以及 $\boldsymbol{y}$ 满足 $\boldsymbol{y}^T\boldsymbol{1} = 1$ 。求微分，使用矩阵乘法、逐元素函数等法则： $dl = -\boldsymbol{y}^T dW\boldsymbol{x} + \frac{\boldsymbol{1}^T\left(\exp(W\boldsymbol{x}) \odot (dW\boldsymbol{x})\right)}{\boldsymbol{1}^T\exp(W\boldsymbol{x})}$ 。再套上迹并做交换，注意可化简 $\boldsymbol{1}^T\left(\exp(W\boldsymbol{x}) \odot (dW\boldsymbol{x})\right) = \exp(W\boldsymbol{x})^T dW\boldsymbol{x}$ ，这是根据等式 $\boldsymbol{1}^T(\boldsymbol{u} \odot \boldsymbol{v}) = \boldsymbol{u}^T\boldsymbol{v}$ ，故

$$dl = \text{tr}\left(-\boldsymbol{y}^T dW\boldsymbol{x} + \frac{\exp(W\boldsymbol{x})^T dW\boldsymbol{x}}{\boldsymbol{1}^T\exp(W\boldsymbol{x})}\right) = \text{tr}(-\boldsymbol{y}^T dW\boldsymbol{x} + \text{softmax}(W\boldsymbol{x})^T dW\boldsymbol{x}) = \text{tr}(\boldsymbol{x}(\text{softmax}(W\boldsymbol{x}) - \boldsymbol{y})^T dW)$$

。对照导数与微分的联系，得到 $\frac{\partial l}{\partial W} = (\text{softmax}(W\boldsymbol{x}) - \boldsymbol{y})\boldsymbol{x}^T$ 。

解2：定义 $\boldsymbol{a} = W\boldsymbol{x}$ ，则 $l = -\boldsymbol{y}^T\log\text{softmax}(\boldsymbol{a})$ ，先同上求出 $\frac{\partial l}{\partial \boldsymbol{a}} = \text{softmax}(\boldsymbol{a}) - \boldsymbol{y}$ ，再利用复合法则： $dl = \text{tr}\left(\frac{\partial l}{\partial \boldsymbol{a}}^T d\boldsymbol{a}\right) = \text{tr}\left(\frac{\partial l}{\partial \boldsymbol{a}}^T dW\boldsymbol{x}\right) = \text{tr}\left(\boldsymbol{x}\frac{\partial l}{\partial \boldsymbol{a}}^T dW\right)$ ，得到 $\frac{\partial l}{\partial W} = \frac{\partial l}{\partial \boldsymbol{a}}\boldsymbol{x}^T$ 。

最后一例留给经典的神经网络。神经网络的求导术是学术史上的重要成果，还有个专门的名字叫做BP算法，我相信如今很多人在初次推导BP算法时也会颇费一番脑筋，事实上使用矩阵求导术来推导并不复杂。为简化起见，我们推导二层神经网络的BP算法。

例7【二层神经网络】： $l = -\boldsymbol{y}^T\log\text{softmax}(W_2\sigma(W_1\boldsymbol{x}))$ ，求 $\frac{\partial l}{\partial W_1}$ 和 $\frac{\partial l}{\partial W_2}$ 。其中 $\boldsymbol{y}$ 是除一个元素为1外其它元素为0的的 $m \times 1$ 列向量， W_2 是 $m \times p$ 矩阵， W_1 是 $p \times n$ 矩阵， $\boldsymbol{x}$ 是 $n \times 1$ 列向量， l 是标量；log表示自然对数， $\text{softmax}(\boldsymbol{a}) = \frac{\exp(\boldsymbol{a})}{\boldsymbol{1}^T\exp(\boldsymbol{a})}$ 同上， σ 是逐元素sigmoid函数 $\sigma(a) = \frac{1}{1+\exp(-a)}$ 。

解：定义 $\boldsymbol{a}_1 = W_1\boldsymbol{x}$ ， $\boldsymbol{h}_1 = \sigma(\boldsymbol{a}_1)$ ， $\boldsymbol{a}_2 = W_2\boldsymbol{h}_1$ ，则 $l = -\boldsymbol{y}^T\log\text{softmax}(\boldsymbol{a}_2)$ 。在前例中已求出 $\frac{\partial l}{\partial \boldsymbol{a}_2} = \text{softmax}(\boldsymbol{a}_2) - \boldsymbol{y}$ 。使用复合法则，

$dl = \text{tr}\left(\frac{\partial l}{\partial \boldsymbol{a}_2}^T d\boldsymbol{a}_2\right) = \text{tr}\left(\frac{\partial l}{\partial \boldsymbol{a}_2}^T dW_2\boldsymbol{h}_1\right) + \text{tr}\left(\frac{\partial l}{\partial \boldsymbol{a}_2}^T W_2 d\boldsymbol{h}_1\right)$ ，使用矩阵乘法交换的迹技巧从

第一项得到 $\frac{\partial l}{\partial W_2}=\frac{\partial l}{\partial \boldsymbol{a}_2}\boldsymbol{h}_1^T$，从第二项得到 $\frac{\partial l}{\partial \boldsymbol{h}_1}=W_2^T\frac{\partial l}{\partial \boldsymbol{a}_2}$。接下来对第二项继续使用复合来求 $\frac{\partial l}{\partial \boldsymbol{a}_1}$，并利用矩阵乘法和逐元素乘法交换的迹技巧：

$dl_2=\text{tr}\left(\frac{\partial l}{\partial \boldsymbol{h}_1}^T d\boldsymbol{h}_1\right)=\text{tr}\left(\frac{\partial l}{\partial \boldsymbol{h}_1}^T(\sigma'(\boldsymbol{a}_1)\odot d\boldsymbol{a}_1)\right)=\text{tr}\left(\left(\frac{\partial l}{\partial \boldsymbol{h}_1}\odot\sigma'(\boldsymbol{a}_1)\right)^T d\boldsymbol{a}_1\right)$，得到 $\frac{\partial l}{\partial \boldsymbol{a}_1}=\frac{\partial l}{\partial \boldsymbol{h}_1}\odot\sigma'(\boldsymbol{a}_1)$。为求 $\frac{\partial l}{\partial W_1}$，再用一次复合法则：

$dl_2=\text{tr}\left(\frac{\partial l}{\partial \boldsymbol{a}_1}^T d\boldsymbol{a}_1\right)=\text{tr}\left(\frac{\partial l}{\partial \boldsymbol{a}_1}^T dW_1\boldsymbol{x}\right)=\text{tr}\left(\boldsymbol{x}\frac{\partial l}{\partial \boldsymbol{a}_1}^T dW_1\right)$，得到 $\frac{\partial l}{\partial W_1}=\frac{\partial l}{\partial \boldsymbol{a}_1}\boldsymbol{x}^T$。

推广：样本 $(\boldsymbol{x}_1,y_1),\dots,(\boldsymbol{x}_N,y_N)$，$l=-\sum_{i=1}^N \boldsymbol{y}_i^T\log\text{softmax}(W_2\sigma(W_1\boldsymbol{x}_i+\boldsymbol{b}_1)+\boldsymbol{b}_2)$，其中 $\boldsymbol{b}_1$ 是 $p\times 1$ 列向量，$\boldsymbol{b}_2$ 是 $m\times 1$ 列向量，其余定义同上。

解1：定义 $\boldsymbol{a}_{1,i}=W_1\boldsymbol{x}_i+\boldsymbol{b}_1$，$\boldsymbol{h}_{1,i}=\sigma(\boldsymbol{a}_{1,i})$，$\boldsymbol{a}_{2,i}=W_2\boldsymbol{h}_{1,i}+\boldsymbol{b}_2$，则 $l=-\sum_{i=1}^N \boldsymbol{y}_i^T\log\text{softmax}(\boldsymbol{a}_{2,i})$。先同上可求出 $\frac{\partial l}{\partial \boldsymbol{a}_{2,i}}=\text{softmax}(\boldsymbol{a}_{2,i})-\boldsymbol{y}_i$。使用复合法则，

$dl=\text{tr}\left(\sum_{i=1}^N\frac{\partial l}{\partial \boldsymbol{a}_{2,i}}^T d\boldsymbol{a}_{2,i}\right)=\text{tr}\left(\sum_{i=1}^N\frac{\partial l}{\partial \boldsymbol{a}_{2,i}}^T dW_2\boldsymbol{h}_{1,i}\right)+\underbrace{\text{tr}\left(\sum_{i=1}^N\frac{\partial l}{\partial \boldsymbol{a}_{2,i}}^T W_2 d\boldsymbol{h}_{1,i}\right)}_{dl_2}+\text{tr}\left(\sum_{i=1}^N\frac{\partial l}{\partial \boldsymbol{a}_{2,i}}^T d\boldsymbol{b}_2\right)$

，从第一项得到得到 $\frac{\partial l}{\partial W_2}=\sum_{i=1}^N\frac{\partial l}{\partial \boldsymbol{a}_{2,i}}\boldsymbol{h}_{1,i}^T$，从第二项得到 $\frac{\partial l}{\partial \boldsymbol{h}_{1,i}}=W_2^T\frac{\partial l}{\partial \boldsymbol{a}_{2,i}}$，从第三项得到到 $\frac{\partial l}{\partial \boldsymbol{b}_2}=\sum_{i=1}^N\frac{\partial l}{\partial \boldsymbol{a}_{2,i}}$。接下来对第二项继续使用复合法则，得到 $\frac{\partial l}{\partial \boldsymbol{a}_{1,i}}=\frac{\partial l}{\partial \boldsymbol{h}_{1,i}}\odot\sigma'(\boldsymbol{a}_{1,i})$。为求 $\frac{\partial l}{\partial W_1},\frac{\partial l}{\partial \boldsymbol{b}_1}$，再用一次复合法则：

$dl_2=\text{tr}\left(\sum_{i=1}^N\frac{\partial l}{\partial \boldsymbol{a}_{1,i}}^T d\boldsymbol{a}_{1,i}\right)=\text{tr}\left(\sum_{i=1}^N\frac{\partial l}{\partial \boldsymbol{a}_{1,i}}^T dW_1\boldsymbol{x}_i\right)+\text{tr}\left(\sum_{i=1}^N\frac{\partial l}{\partial \boldsymbol{a}_{1,i}}^T d\boldsymbol{b}_1\right)$，得到 $\frac{\partial l}{\partial W_1}=\sum_{i=1}^N\frac{\partial l}{\partial \boldsymbol{a}_{1,i}}\boldsymbol{x}_i^T$，$\frac{\partial l}{\partial \boldsymbol{b}_1}=\sum_{i=1}^N\frac{\partial l}{\partial \boldsymbol{a}_{1,i}}$。

解2：可以用矩阵来表示N个样本，以简化形式。定义 $X=[\boldsymbol{x}_1,\cdots,\boldsymbol{x}_N]$，$A_1=[\boldsymbol{a}_{1,1},\cdots,\boldsymbol{a}_{1,N}]=W_1X+\boldsymbol{b}_1\boldsymbol{1}^T$，$H_1=[\boldsymbol{h}_{1,1},\cdots,\boldsymbol{h}_{1,N}]=\sigma(A_1)$，$A_2=[\boldsymbol{a}_{2,1},\cdots,\boldsymbol{a}_{2,N}]=W_2H_1+\boldsymbol{b}_2\boldsymbol{1}^T$，注意这里使用全1向量来扩展维度。先同上求出 $\frac{\partial l}{\partial A_2}=[\text{softmax}(\boldsymbol{a}_{2,1})-\boldsymbol{y}_1,\cdots,\text{softmax}(\boldsymbol{a}_{2,N})-\boldsymbol{y}_N]$。使用复合法则，

$dl=\text{tr}\left(\frac{\partial l}{\partial A_2}^T dA_2\right)=\text{tr}\left(\frac{\partial l}{\partial A_2}^T dW_2H_1\right)+\underbrace{\text{tr}\left(\frac{\partial l}{\partial A_2}^T W_2dH_1\right)}_{dl_2}+\text{tr}\left(\frac{\partial l}{\partial A_2}^T d\boldsymbol{b}_2\boldsymbol{1}^T\right)$，从第一项得到 $\frac{\partial l}{\partial W_2}=\frac{\partial l}{\partial A_2}H_1^T$，从第二项得到 $\frac{\partial l}{\partial H_1}=W_2^T\frac{\partial l}{\partial A_2}$，从第三项得到到 $\frac{\partial l}{\partial \boldsymbol{b}_2}=\frac{\partial l}{\partial A_2}\boldsymbol{1}$。接下来对第二项继续使用复合法则，得到

则：$dl_2 = \text{tr}\left(\frac{\partial l}{\partial A_1}^T dA_1\right) = \text{tr}\left(\frac{\partial l}{\partial A_1}^T dW_1 X\right) + \text{tr}\left(\frac{\partial l}{\partial A_1}^T db_1 \mathbf{1}^T\right)$，得到 $\frac{\partial l}{\partial W_1} = \frac{\partial l}{\partial A_1} X^T$，$\frac{\partial l}{\partial b_1} = \frac{\partial l}{\partial A_1}\mathbf{1}$。

下篇见zhuanlan.zhihu.com/p/24...。

编辑于 03-06

机器学习 矩阵分析 优化

刚开始搞CV时候，发现典型的CVPR paper总有这么几个矩阵求导公式，与那些只知道刷baseline的文章相比，顿时逼格高大上啊！唉，作者你要是早几年写这个专栏一定能解救一批苦逼的PhD啊！

3

花花 2017-02-05

写得真好！

赞

郑建东 2017-02-12

感谢大神指导，有点茅塞顿开的感觉。

赞

小成 2017-02-14

小白求教：
请问在例3【logistic回归】的解答中第一行：
首先将softmax函数代入并写成 l = ... ，这个公式中右侧第二项为什么不用和第一项一样乘上y的转置？

1

长躯鬼侠 (作者) 回复 小成 2017-02-18

这里我跳了一步，y是除一个元素为1外其它元素为0的向量，log()是个标量，化简一下就是这样。

赞

小成 回复 长躯鬼侠 (作者) 2017-02-19

明白啦，感谢大神回复!

赞

查看全部 6 条回复

郭洋洋 2017-02-15

十分好用 NBNBNBNBNBNBNB

赞

Martins3 2017-02-18

很强，很酷

赞

shiryaev 2017-02-27

cal.cs.illinois.edu/~jo...

这篇更基础点，希望能给更多的小伙伴帮忙

4

相思作坊半世離殇 回复 shiryaev 2017-03-20

感谢分享资料。看了这篇pdf和楼主写的文章，明白了许多

1

bb just 回复 相思作坊半世離殇 2019-06-04

链接打不开了，请问你还有这个质量么。

赞

查看全部 7 条回复

黄角兽 2017-03-08

功德无量🙏

赞

排骨郎 2017-03-16

写的很好

赞

我不是那样的烟火 2017-03-27

虽然我也是理解一点的，可是，你这篇文章简直不要太赞好吧

赞

Winston 2017-05-12

很实用，感谢！

赞

赵小呆 2017-05-24

之前看cs229,上面写要拆开运算，结果每次写cs229的习题都要运算好久，还容易出错。楼主伟大！但是这里有一个问题，矩阵运算和常量运算有时运算法则一致，有时不一致，楼主有推荐资料解决这个问题吗？

赞

长躯鬼侠 (作者) 回复 赵小呆 2017-05-25

我不确定你具体想问哪种情况，或许你可以举个例子来讨论。简略地讲，求微分的运算法则，对矩阵和标量基本一致；从微分获得导数，对矩阵需要特别的技术，比如迹技巧。

赞

亦然

想问一下复数矩阵求导且对非方阵求导应做哪些改变和处理?

赞

穆文 2017-07-30

文章讲的很赞，非常感谢分享。说两点我的理解，跟作者探讨下

1. 我的理解就是，其实矩阵不过是另一套『数学语言』，把原来标量世界的各种运算法则，转换到矩阵世界，用矩阵、矩阵的乘积、矩阵秩、矩阵迹、行列式等，来刻画新的运算法则，不知这个理解对不对?

2. 对原文『不能随意套用微积分中标量导数的结论，比如认为AX对X的导数为A，这是没有根据、意义不明的』这句话持保留态度，这里X是矩阵的时候，没有明确定义，但如果X是向量，结论是A没错吧？因为AX刻画了原始空间X到新的空间AX的映射，即向量对向量求导，是有明确定义

赞

长躯鬼侠 (作者) 回复 穆文 2017-07-31

我在这里讲的是标量对矩阵的导数。而向量对向量的导数，是我的下篇里矩阵对矩阵导数的一种特殊情况，按下篇中的定义有Ax对x的导数是A^T。

赞

陌烛 2017-08-26

你好，请问下，上篇中，例三中的log都是当作ln使用的吗？写作log用作ln?

赞

长躯鬼侠 (作者) 回复 陌烛 2017-08-28

是的，log当做ln用了。

1

陌烛 回复 长躯鬼侠 (作者) 2017-09-07

知道了，蟹蟹

赞

X'plore 2017-09-14

df = \sum_{i} \frac{\partial f}{\partial x_i}dx_i = \frac{\partial f}{\partial \boldsymbol{x}}^T d\boldsymbol{x} 请问右面这一步如何理解，为什么等于转置的偏微分?

赞

长躯鬼侠 (作者) 回复 X'plore 2017-09-14

根据定义

赞

武宽Quin Wu 2017-09-1

博主您好，关于您提到的一个标量f对矩阵X的求导，我有一点不解的是，您这里没有提到wiki中所谓的 分子布局 跟分母布局，如果是分子布局的话。\frac{\partial f}{\partial X} = \frac{\partial f}{\partial X_{ji}}吧，而不应该是X_{ij}。在分母布局下，\frac{\partial f}{\partial X}就是您所描述的 \frac{\partial f}{\partial X_{ij}}。是这样嘛?

1

长躯鬼侠 (作者) 回复 武宽Quin Wu 2017-09-16

我使用的定义第二个。

1

徐敬 2017-09-21

厉害了

赞

侯佐 2017-10-03

很厉害，解决了很长时间的疑问

赞

上一页 1 2 3 4 ... 11 下一页