矩阵求导极简教程

王礼萍

2019 年 10 月 21 日

目录

第一章　致谢

支撑笔者在繁忙的学业中抽空完成这个教程的动力永远来自我最亲爱的 Congcong, 因为, 首先是为你而写, 你永远是我取悦的第一个读者!

To Congcong, with my full love!

第二章　记号

- $\mathbb{R}$: 实数集
- $\mathbb{R}^n$: n 维实向量构成的线性空间
- $\mathbb{R}^{n\times q}$: $n \times q$ 矩阵构成的线性空间
- $\phi : \mathbb{R} \to \mathbb{R}$ 的函数
- $f : \mathbb{R}^n \to \mathbb{R}$ 的函数
- $F : \mathbb{R}^{n\times q} \to \mathbb{R}^{m\times p}$ 的函数
- $B(c; r)$ 其中 $c \in \mathbb{R}^n$: 以 c 为中心, r 为半径的开球

第三章　绪论

首先界定下本教程标题中” 矩阵导数” 的含意. 第一, 这里的矩阵, 不仅指二维的矩阵, 也包括向量 (可以将长度为 n 的列向量看成 $1 \times n$ 的矩阵). 第二, 关于矩阵导数, 一方面, 函数本身可以是一个矩阵 (或者向量), 另一方面, 函数的自变量可以是一个矩阵 (或者向量).

笔者刚开始接触机器学习与深度学习时, 在相关教材和论文中多次见到矩阵函数或者对矩阵变量求导的操作, 但是大多采用的是 ad-hoc 的方法 (推过 BP 神经网络的读者应该会有比较深的体会). 而一元微积分中有一套成熟的求导技巧, 一旦熟练掌握则非常简单, 难道多元求导就没有一套成熟的体系吗?

笔者查阅了很多相关书籍, 最终找到了一本专门讲解矩阵求导的书籍 [1], 不过这本书好几百页, 从一元微积分开始讲起, 介绍了非常丰富的矩阵论中的结论, 对于一般读者是一个较大的负担. 因此, 笔者致力于提供一个 minimum volume 的矩阵求导的入门教程, 以满足日常学习和科研的需求.

第四章　矩阵论知识补充

为了后面的推导, 需要在本科线性代数的基础上补充与矩阵相关的若干概念和性质.

4.1　基础结论

定理 1. 对于 n-阶方阵 A, B 有

$$\mathrm{tr}(A^T B) = \sum_{i,j} a_{ij} b_{ij}$$

特殊地, 对于 $B = A$, 有

$$\mathrm{tr}(A^T A) = \sum_{i,j} a_{ij}^2 = \|A\|_F^2$$

4.2　拉直操作与 Kronecker 乘积

定义 1. 对于 $m \times n$ 的矩阵 A, 定义拉直操作 $\mathrm{vec}\, A = (a_1^T, a_2^T, \ldots, a_n^T)^T$, 其中 a_i 代表矩阵 A 的第 i 列.

定义 2. 对于矩阵 A, B*(对于维度没有任何要求)*, 定义 *Kronecker* 乘积

$$A \otimes B = \Big(a_{ij} B\Big)$$

第五章　矩阵导数定义

学习过一元微积分/高等数学的读者应该会记得, 连续不一定可导, 但可导一定连续, 而且引入导数之前, 我们事实上先引入的是微分的概念. 为了引入矩阵导数, 同样地, 我们需要先引入函数连续性和微分.

对于一般的函数, 自变量, 函数值都可以是标量, 向量和矩阵, 所以一共有 9 种情况, 为了简化讨论, 我们将考虑如下两种情况, 其他六种情况其实是这三种情况的特例.

- $f : S \to \mathbb{R}^m$, 其中 $S \subset \mathbb{R}^n$
- $F : S \to \mathbb{R}^{m\times p}$, 其中 $S \subset \mathbb{R}^{n\times q}$

本章后面将针对这两种情况, 按照函数的连续性, 微分以及导数的顺序, 进行介绍.

5.1　连续性

定义 3. *对于函数 $\phi : S \to \mathbb{R}$, 其中 $S \subset \mathbb{R}^n$, 对于 $c \in S$. 如果 $\forall\epsilon > 0$, $\exists\delta > 0$ 使得对于所有的 $c+u \in S$, 其中 $\|u\| < \delta$, 都有 $|\phi(c+u) - \phi(c)| < \epsilon$, 则称 ϕ 在 c 处连续. 进一步地, 如果 ϕ 在 S 中每一点都连续, 则称 ϕ 在 S 上连续.*

定义 4. *对于函数 $f : S \to \mathbb{R}^m$, 其中 $S \subset \mathbb{R}^n$, 记*

$$f(x) = (f_1(x), \ldots, f_m(x))'$$

所以函数 f 其实定义了 m 个实数值函数 $f_1, \ldots, f_m$. 如果 $f_1, \ldots, f_m$ 都在 $c \in S$ 上连续, 则称 f 在 c 处连续. 进一步地, 如果 f 在 S 中每一点都连续, 则称 f 在 S 上连续.

5.2 微分与导数

定义 5. 对于函数 $\phi : S \to \mathbb{R}$, 其中 $S \subset \mathbb{R}^n$, 设 c 是 S 的一个内点, $B(c,r) \subset S$, 如果存在 $a(c)$ 使得对于任意的 $c+u \in B(c;r)$ 有

$$\phi(c+u) = \phi(c) + a(c)u + r_c(u)$$

而且

$$\lim_{u\to 0} \frac{r_c(u)}{\|u\|} = 0$$

则称 ϕ 在 c 处可微, 记 $d\phi(c;u) = a(c)u$, 称为函数 ϕ 在 c 处的微分, 记 $D\phi = a(c)$, 称为函数 ϕ 在 c 处的导数.

对于实数值函数 ϕ, 实际中经常会用到它的梯度 (gradient), 即

$$\nabla\phi(c) = (\frac{\partial\phi}{\partial c_1}, \ldots, \frac{\partial\phi}{\partial c_n})^T$$

可以证明 $\nabla\phi(c) = D\phi(c)^T = a(c)$.

定义 6. 对于函数 $f : S \to \mathbb{R}^m$, 其中 $S \subset \mathbb{R}^n$, 设 c 是 S 的一个内点, $B(c,r) \subset S$, 如果存在 $A(c)$ 使得对于任意的 $c+u \in B(c;r)$ 有

$$f(c+u) = f(c) + A(c)u + r_c(u)$$

而且

$$\lim_{u\to 0} \frac{r_c(u)}{\|u\|} = 0$$

则称 f 在 c 处可微, 记 $df(c;u) = A(c)u$, 称为函数 f 在 c 处的微分, 记 $Df = a(c)$, 称为函数 f 在 c 处的导数.

我们还会经常看到所谓的 Jacobian 矩阵, 即

$$\left(\frac{\partial f_i}{\partial c_j}\right)$$

可以证明: 如果函数 f 在 c 处可微, 则 Df 就是 Jacobian 矩阵.

对于更为一般的矩阵函数 $F : S \to \mathbb{R}^{m\times p}$, 其中 $S \subset \mathbb{R}^{n\times q}$, 为了定义微分, 我们需要将自变量和函数值同时进行拉直操作, 从而通过**??**完成定义.

定义 7. 对于 $F : S \to \mathbb{R}^{m\times p}$, 其中 $S \subset \mathbb{R}^{n\times q}$, 设 C 是 S 的一个内点, $B(C,r) \subset S$, 如果存在 $A(C)$ 使得对于任意的 $C+U \in B(C;r)$ 有

$$\operatorname{vec} F(C+U) = \operatorname{vec} F(C) + A(C)\operatorname{vec} U + R_C(U)$$

而且

$$\lim_{U\to 0}\frac{R_C(U)}{\|U\|}=0$$

则称 F 在 C 处可微. 设 $dF(C;U)\in\mathbb{R}^{m\times p}$, 则根据 $\operatorname{vec}dF(C;U)=A(C)\operatorname{vec}U$ 可以唯一确定出 $dF(C;U)$, 称为函数 F 在 C 处的微分, 记 $DF=A(C)$, 称为函数 F 在 C 处的导数.

我们考虑一个特殊的 $F:S\to\mathbb{R}$, 其中 $S\subset\mathbb{R}^{n\times q}$, 它将矩阵映射到一个函数. 根据上面的定义, F 的导数是一个长度为 nq 的行向量, 但是我们通常要求的是一个梯度矩阵

$$\nabla F(C)=\left(\frac{\partial F}{\partial C_{ij}}\right)$$

可以证明:

$$\operatorname{vec}\nabla F(C)=DF(c)^T$$

事实上, 对于一般的 F, 也可以推导出 $\frac{\partial F_{ij}}{\partial C_{kl}}$ 与 DF 的关系, 但是比较繁琐, 而且用得不多, 故略去.

观察微分的定义, 我们可以发现几点

- 微分可以是标量, 向量或者矩阵, 但是其形状始终和函数值形状相同
- 微分与选取的 u 有关, 本身并不要求 $u\to 0$

5.3 微分的性质

引入微分的目的主要在于微分具有一系列非常统一和优美的性质, 这些性质有助于我们以一种非常统一的方式来求解函数导数 (六中会有详细介绍), 现在我们就来介绍这些性质, 这些性质根据微分的定义很容易证明, 读者可以参考 [1] 的第八章.

5.3.1 实值函数

对于可微实值函数 u,v 和常数 α 有:

- $d\alpha=0$
- $d(\alpha u)=\alpha du$
- $d(u+v)=du+dv$

- $d(uv) = (du)v + udv$
- $d(\frac{u}{v}) = \frac{(du)v - udv}{v^2}$

用一元可微实值函数 f 复合有

$$df(u) = f'(u)du$$

因而有

- $de^u = e^u du$
- $d\log u = \frac{1}{u}du$

等结论.

5.3.2 矩阵值函数

对于向量/矩阵值函数 U, V, 常矩阵 A, 常数 α, 有如下结论 (运算能够进行的条件下)

对于基本运算有:

- $dA = 0$
- $d(\alpha U) = \alpha dU$
- $d(U + V) = dV + dV$
- $d(UV) = (dU)V + UdV$
- $d(U \otimes V) = dU \otimes V + U \otimes dV$
- $d(U \odot V) = dU \otimes V + U \odot dV$

对于常见矩阵操作有:

- $dU^T = (dU)^T$
- $d\operatorname{vec} U = \operatorname{vec} dU$
- $d\operatorname{tr} U = \operatorname{tr} dU$

用一元可微实值函数 f 复合有

第六章　具体实践

在实际使用中, 一般需要求解的是梯度或者 Jacobian 矩阵, 这时就有两种路线

- 直接计算诸如 $\frac{\partial f_i}{\partial c_j}$ 这类逐元素导数
- 利用5, 6, 7中定义的微分求出相应的导数, 进而得到梯度或者 Jacobian 矩阵

第一种方法, 过程比较简单, 但是函数形式各异, 尤其是当函数涉及到复杂的矩阵计算时, 很难算出 $\frac{\partial f_i}{\partial c_j}$, 即使计算成功, 最后也往往难以凑出一个紧凑的矩阵表达. 另外, 对于比较难以处理复合函数.

下面针对不同函数类型, 通过具体例子来阐述第二种方法, 虽然过程稍显繁琐, 但是构成了一个成熟的框架 (套路), 因而比较方便.

本章接下来将根据不同函数类型, 通过具体例子来阐述如何求导.

6.1　向量映到实数

对于函数 $\phi : S \to \mathbb{R}$, 其中 $S \subset \mathbb{R}^n$, 我们通常希望得到

$$\nabla\phi(x) = (\frac{\partial\phi(x)}{\partial x_1}, \ldots, \frac{\partial\phi(x)}{\partial x_n})^T$$

例 1. $\phi(x) = a^T x$, 其中 $a, x \in \mathbb{R}^n$, 且 a 为常向量, 求 $\nabla\phi(x)$.

解 *1.*

$$a^T x = \sum_i a_i x_i,$$

所以 $\nabla\phi(x) = a$.

例 2. $\phi(x) = x^T Ax$, 其中 $x \in \mathbb{R}^n$,A 是 $\mathbb{R}^{n\times n}$ 中的一个常矩阵, 求 $\nabla\phi(X)$.

解 *1.*

$$x^T Ax = \sum_i \sum_j x_i a_{ij} x_j$$

所以 $\frac{\partial\phi(x)}{\partial x_i} = \sum_j a_{ij}x_j + \sum_j x_j a_{ji}$, 因而 $\nabla\phi(x) = (A + A^T)x$.

解 *2.* 先求微分,

$$\begin{aligned} d(x^T Ax) &= (dx)^T Ax + x^T Adx \\ &= x^T A^T dx + x^T Adx \\ &= x^T(A + A^T)dx \end{aligned}$$

所以, $\phi(x) = (x^T(A + A^T))^T = (A + A^T)x$. 特殊地, 如果 $A = A^T$, 则 $\nabla\phi(x) = 2Ax$.

例 3. 求 $\phi(x) = \|x\|_2^2$ 的梯度.

6.2 矩阵映到实数

对于函数 $\phi : S \to \mathbb{R}$, 其中 $S \subset \mathbb{R}^{n\times q}$, 我们通常希望得到

$$\nabla\phi(X) = \begin{pmatrix} \frac{\partial\phi}{\partial x_{11}} \cdots \frac{\partial\phi}{\partial x_{1q}} \\ \cdots \frac{\partial\phi}{\partial x_{ij}} \cdots \\ \frac{\partial\phi}{\partial x_{n1}} \cdots \frac{\partial\phi}{\partial x_{nq}} \end{pmatrix}_{n\times q}$$

例 4. $\phi(X) = a^T Xa$, 其中 $X \in \mathbb{R}^{n\times n}$, 且 a 为常向量, 求 $\nabla\phi(X)$.

解 1.

$$a^T Xa =$$

例 5. $\phi(X) = a^T X^{-1}a$, 其中 X 是 $\in \mathbb{R}^{n\times n}$ 中可逆矩阵, 且 a 为常向量, 求 $\nabla\phi(X)$.

例 6. 求 $\phi(X) = \det X$ 的梯度, 其中 $\det X$ 表示可逆矩阵 X 的行列式.

解 *1.* 对 X 按第 1 行展开得

$$\det X = \sum_j c_{1j} X_{1j}$$

所以 $\frac{\partial \phi(X)}{\partial X_{1j}} = c_{1j}$, 同理有 $\frac{\partial \phi(X)}{\partial X_{ij}} = c_{ij}$, 因而有

$$\nabla \phi(X) = (X^{\#})^T$$

如果 X 可逆, 则 $X^{\#} = (\det X) X^{-1}$, 所以 $\nabla \phi(X) = (\det X)(X^{-1})^T$

例 7. 求 $\log\det X$ 的梯度, 其中 $\det X$ 表示可逆矩阵 X 的行列式.

解. 由前例知 $\nabla \log\det X = X^{-T}$

例 8. 求 $\operatorname{tr} X$ 的梯度

解. $\nabla \operatorname{tr} X = I$

例 9. 求 $\operatorname{tr} AA^T = \operatorname{tr} A^T A = \|A\|_F^2$ 的梯度.

6.3 向量映到向量

对于函数 $f : S \to \mathbb{R}^m$, 其中 $S \subset R^n$
我们通常要求的是 Jacobian 矩阵

$$\frac{\partial y}{\partial x} = \begin{pmatrix} \frac{\partial y_1}{\partial x_1} \cdots \frac{\partial y_1}{\partial x_n} \\ \cdots \frac{\partial y_i}{\partial x_j} \cdots \\ \frac{\partial y_m}{\partial x_1} \cdots \frac{\partial y_m}{\partial x_n} \end{pmatrix}_{m \times n}$$

例 10. 机器学习中经常遇到线性层 $y = Wx + w_0$, 求 $\frac{\partial y}{\partial x}$

解. *sdf*

6.4 实数映到矩阵

对于 $F : S \to \mathbb{R}^{m \times p}$, 其中 $S \subset \mathbb{R}$ 我们希望求解

$$\frac{\partial F}{\partial x} = (\frac{\partial F_{ij}}{\partial x})_{m \times p}.$$

这种情况一般利用逐元素求导即可, 比较简单, 但是如果如果结合矩阵微分和矩阵的性质, 往往能得到更多有用的结论.

例 11. 设 $F(x) : \mathbb{R} \to \mathbb{R}^{n\times n}$ 是一个矩阵函数, 且 $F(x)$ 作为矩阵可逆, 求 $\frac{\partial F^{-1}}{\partial x}$.

解 *1.* 由

$$\begin{aligned}\frac{\partial I}{\partial x} &= \frac{\partial (FF^{-1})}{\partial x} \\ &= (\frac{\partial F}{\partial x})F^{-1} + F\frac{\partial F^{-1}}{\partial x} \\ &= 0\end{aligned}$$

得 $\frac{\partial F^{-1}}{\partial x} = -F^{-1}\frac{\partial F}{\partial x}F^{-1}$.

第七章　更多的例子

Bibliography

[1] Jan R. Magnus and Heinz Neudecker. *Matrix Differential Calculus with Applications in Statistics and Econometrics*. Second. John Wiley, 1999. ISBN: 0471986321 9780471986324 047198633X 9780471986331.