Toy Matrix gradient

Renhe W.

Contents

1 矩阵梯度

1.1 实值函数相对于实向量的梯度

对于 $n\times 1$ 向量 $\boldsymbol{X}$ 的梯度算子记作 $\nabla_{\boldsymbol{X}}$，定义为:

$$\nabla_{\boldsymbol{X}}=[\frac{\partial}{\partial x_1},\frac{\partial}{\partial x_2},\cdots,\frac{\partial}{\partial x_n}]^T=\frac{\partial}{\partial \boldsymbol{X}}. \tag{1.1}$$

因此，$n\times 1$ 实向量 $\boldsymbol{X}$ 为变元的实际标量函数 $f(\boldsymbol{X})$ 相对于 $\boldsymbol{X}$ 的梯度是一个 $n\times 1$ 的一列向量，定义为：

$$\nabla_{\boldsymbol{X}}f(\boldsymbol{X})=\begin{bmatrix}\frac{\partial f(\boldsymbol{X})}{\partial x_1}, & \frac{\partial f(\boldsymbol{X})}{\partial x_2}, & \cdots, & \frac{\partial f(\boldsymbol{X})}{\partial x_n}\end{bmatrix}^T=\begin{bmatrix}\frac{\partial f(\boldsymbol{X})}{\partial x_1}\\ \frac{\partial f(\boldsymbol{X})}{\partial x_2}\\ \vdots\\ \frac{\partial f(\boldsymbol{X})}{\partial x_n}\end{bmatrix}=\frac{\partial f(\boldsymbol{X})}{\partial \boldsymbol{X}}. \tag{1.2}$$

梯度方向的负方向为变元 $\boldsymbol{X}$ 的梯度流 (gradient flow)，记作：

$$\dot{\boldsymbol{X}}=-\nabla_{\boldsymbol{X}}f(\boldsymbol{X}). \tag{1.3}$$

可以看出：

1. 一个以向量为变元函数的梯度为一个向量；
2. 梯度的每个分量代表变量函数在该分量方向上的变化率.

梯度向量最重要的性质之一是，它指出了当变元增大时函数 f 的最大增大率。相反，梯度的负值（负梯度）指出了当变元增大时函数 f 的最大减小率。根据这样一种性质，即可设计出求一函数极小值的迭代算法。

类似地，实值函数 $f(x)$ 相对于 $1\times n$ 一行向量 x^T 的梯度为 $1{\times}n$ 一行行向量，定义为:

$$\nabla_{\boldsymbol{X}^T}f(\boldsymbol{X})=\begin{bmatrix}\frac{\partial f(\boldsymbol{X})}{\partial x_1}, & \frac{\partial f(\boldsymbol{X})}{\partial x_2}, & \cdots, & \frac{\partial f(\boldsymbol{X})}{\partial x_n}\end{bmatrix}=\frac{\partial f(\boldsymbol{X})}{\partial \boldsymbol{X}^T}. \tag{1.4}$$

m 维一行向量函数 $f(\boldsymbol{X})=[f_1(\boldsymbol{X}),\cdots,f_m(\boldsymbol{X})]$ 对 n 维一列向量 $\boldsymbol{X}$ 的梯度为 $n\times m$ 矩阵[1]:

$$\begin{bmatrix}\frac{\partial f_1(\boldsymbol{X})}{\partial x_1} & \frac{\partial f_2(\boldsymbol{X})}{\partial x_1} & \cdots & \frac{\partial f_n(\boldsymbol{X})}{\partial x_1}\\ \frac{\partial f_1(\boldsymbol{X})}{\partial x_2} & \frac{\partial f_2(\boldsymbol{X})}{\partial x_2} & \cdots & \frac{\partial f_n(\boldsymbol{X})}{\partial x_2}\\ \vdots & \vdots & \cdots & \vdots\\ \frac{\partial f_1(\boldsymbol{X})}{\partial x_n} & \frac{\partial f_2(\boldsymbol{X})}{\partial x_n} & \cdots & \frac{\partial f_n(\boldsymbol{X})}{\partial x_n}\end{bmatrix}. \tag{1.5}$$

下面举几个例子：

[1] $f(x)$ 为列，X 为行.

Example 1.1. 若 $f(x)=[x_1,\cdots,x_n]$，则 $\frac{\partial \boldsymbol{X}^T}{\partial \boldsymbol{X}}=\boldsymbol{I}$：

由 *(1.5)*，有：

$$\begin{bmatrix} \frac{\partial x_1}{\partial x_1} & \frac{\partial x_2}{\partial x_1} & \cdots & \frac{\partial x_n}{\partial x_1} \\ \frac{\partial x_1}{\partial x_2} & \frac{\partial x_2}{\partial x_2} & \cdots & \frac{\partial x_n}{\partial x_2} \\ \vdots & \vdots & \cdots & \vdots \\ \frac{\partial x_1}{\partial x_n} & \frac{\partial x_2}{\partial x_n} & \cdots & \frac{\partial x_n}{\partial x_n} \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} = \boldsymbol{I}.$$

1.2 实值函数的梯度矩阵

在最优化问题中，需要最优化的对象可能是某个加权矩阵。因此，有必要分析实值函数相对于矩阵变元的梯度。实值函数 $f(A)$ 相对于 $m\times n$ 是矩阵 A 的梯度为一 $m\times n$ 矩阵，简称梯度矩阵，定义为：

$$\frac{\partial f(\boldsymbol{A})}{\partial \boldsymbol{A}} = \begin{bmatrix} \frac{\partial f(\boldsymbol{A})}{\partial A_{11}} & \frac{\partial f(\boldsymbol{A})}{\partial A_{12}} & \cdots & \frac{\partial f(\boldsymbol{A})}{\partial A_{1n}} \\ \frac{\partial f(\boldsymbol{A})}{\partial A_{21}} & \frac{\partial f(\boldsymbol{A})}{\partial A_{22}} & \cdots & \frac{\partial f(\boldsymbol{A})}{\partial A_{2n}} \\ \vdots & \vdots & \cdots & \vdots \\ \frac{\partial f(\boldsymbol{A})}{\partial A_{m1}} & \frac{\partial f(\boldsymbol{A})}{\partial A_{m2}} & \cdots & \frac{\partial f(\boldsymbol{A})}{\partial A_{mn}} \end{bmatrix}. \tag{1.6}$$

式中 A_{ij} 为矩阵 A 的元素.

实值函数相对于矩阵变元的梯度具有以下性质：

Example 1.2. 若 $f(\boldsymbol{A})=c$ 是常数，其中 $\boldsymbol{A}$ 为 $m\times n$ 矩阵，则梯度 $\frac{\partial c}{\partial A}=\boldsymbol{O}_{m\times n}$.

Example 1.3. 若 $\boldsymbol{A}\in R^{m\times n}$，$x\in R^{m\times 1}$，$y\in R^{n\times 1}$，则 $\frac{\partial x^T\boldsymbol{A}y}{\partial \boldsymbol{A}}=xy^T$.

ANS:

$$\boldsymbol{A} = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & A_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ A_{m1} & A_{m2} & \cdots & A_{mn} \end{bmatrix}, x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix}, y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}.$$

$\frac{\partial x^T\boldsymbol{A}y}{\partial \boldsymbol{A}}=\frac{\partial \sum_{i=1}^m\sum_{j=1}^n x_iA_{ij}y_j}{\partial \boldsymbol{A}}$，由 *(1.6)*，$\frac{\partial \sum_{i=1}^m\sum_{j=1}^n x_iA_{ij}y_j}{\partial \boldsymbol{A}} = \begin{bmatrix} x_1y_1 & x_1y_2 & \cdots & x_1y_n \\ x_2y_1 & x_2y_2 & \cdots & x_2y_n \\ \vdots & \vdots & \cdots & \vdots \\ x_my_1 & x_my_2 & \cdots & x_my_n \end{bmatrix} = xy^T.$