Some Important Properties for Matrix Calculus

Dawen Liang
Carnegie Mellon University
dawenl@andrew.cmu.edu

1 Introduction

Matrix calculation plays an essential role in many machine learning algorithms, among which matrix calculus is the most commonly used tool. In this note, based on the properties from the differential calculus, we show that they are all adaptable to the matrix calculus[1]. And in the end, an example on least-square linear regression is presented.

2 Notation

A matrix is represented as a bold upper letter, e.g. $\mathbf{X}$, where $\mathbf{X}_{m,n}$ indicates the numbers of rows and columns are m and n, respectively. A vector is represented as a bold lower letter, e.g. $\mathbf{x}$, where it is a $n \times 1$ column vector in this note. An important concept for a $n \times n$ matrix $\mathbf{A}_{n,n}$ is the *trace* $\mathrm{Tr}(\mathbf{A})$, which is defined as the sum of the diagonal:

$$\mathrm{Tr}(\mathbf{A}) = \sum_{i=1}^{n} \mathbf{A}_{ii} \tag{1}$$

where $\mathbf{A}_{ii}$ index the element at the ith row and ith column.

3 Properties

The derivative of a matrix is usually referred as the *gradient*, denoted as ∇. Consider a function $f : \mathbb{R}^{m \times n} \to \mathbb{R}^{p \times q}$, the gradient for $f(\mathbf{A})$ w.r.t. $\mathbf{A}_{m,n}$ is:

$$\nabla_{\mathbf{A}} f(\mathbf{A}) = \frac{\partial f(\mathbf{A})}{\partial \mathbf{A}} = \begin{pmatrix} \frac{\partial f}{\partial \mathbf{A}_{11}} & \frac{\partial f}{\partial \mathbf{A}_{12}} & \cdots & \frac{\partial f}{\partial \mathbf{A}_{1n}} \\ \frac{\partial f}{\partial \mathbf{A}_{21}} & \frac{\partial f}{\partial \mathbf{A}_{22}} & \cdots & \frac{\partial f}{\partial \mathbf{A}_{2n}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f}{\partial \mathbf{A}_{m1}} & \frac{\partial f}{\partial \mathbf{A}_{m2}} & \cdots & \frac{\partial f}{\partial \mathbf{A}_{mn}} \end{pmatrix}$$

This definition is very similar to the differential derivative, thus a few simple properties hold (the matrix $\mathbf{A}$ below is square matrix and has the same dimension with the vectors):

$$\nabla_{\mathbf{x}} \mathbf{b}^{\mathrm{T}} \mathbf{A} \mathbf{x} = \mathbf{b}^{\mathrm{T}} \mathbf{A} \tag{2}$$

[1] Some of the detailed derivations which are omitted in this note can be found at http://www.cs.berkeley.edu/~jduchi/projects/matrix_prop.pdf

$$\nabla_{\mathbf{A}}\mathbf{XAY} = \mathbf{Y}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}} \tag{3}$$

$$\nabla_{\mathbf{x}}\mathbf{x}^{\mathrm{T}}\mathbf{Ax} = \mathbf{Ax} + \mathbf{A}^{\mathrm{T}}\mathbf{x} \tag{4}$$

$$\nabla_{\mathbf{A}^{\mathrm{T}}} f(\mathbf{A}) = (\nabla_{\mathbf{A}} f(\mathbf{A}))^{\mathrm{T}} \tag{5}$$

where superscript T denotes the transpose of a matrix or a vector.

Now let us turn to the properties for the derivative of the trace. First of all, a few useful properties for trace:

$$\mathrm{Tr}(\mathbf{A}) = \mathrm{Tr}(\mathbf{A}^{\mathrm{T}}) \tag{6}$$

$$\mathrm{Tr}(\mathbf{ABC}) = \mathrm{Tr}(\mathbf{BCA}) = \mathrm{Tr}(\mathbf{CAB}) \tag{7}$$

$$\mathrm{Tr}(\mathbf{A} + \mathbf{B}) = \mathrm{Tr}(\mathbf{A}) + \mathrm{Tr}(\mathbf{B}) \tag{8}$$

which are all easily derived. Note that the second one be extended to more general case with arbitrary number of matrices.

Thus, for the derivatives,

$$\nabla_{\mathbf{A}}\mathrm{Tr}(\mathbf{AB}) = \mathbf{B}^{\mathrm{T}} \tag{9}$$

Proof:

Just extend $\mathrm{Tr}(\mathbf{AB})$ according to the trace definition (Eq. 1).

$$\nabla_{\mathbf{A}}\mathrm{Tr}(\mathbf{ABA}^{\mathrm{T}}\mathbf{C}) = \mathbf{CAB} + \mathbf{C}^{\mathrm{T}}\mathbf{AB}^{\mathrm{T}} \tag{10}$$

Proof:

$$\begin{aligned}
&\nabla_{\mathbf{A}}\mathrm{Tr}(\mathbf{ABA}^{\mathrm{T}}\mathbf{C}) \\
=&\nabla_{\mathbf{A}}\mathrm{Tr}(\underbrace{(\mathbf{AB})}_{u(\mathbf{A})}\underbrace{(\mathbf{A}^{\mathrm{T}}\mathbf{C})}_{v(\mathbf{A}^{\mathrm{T}})}) \\
=&\nabla_{\mathbf{A}:u(\mathbf{A})}\mathrm{Tr}(u(\mathbf{A})v(\mathbf{A}^{\mathrm{T}})) + \nabla_{\mathbf{A}:v(\mathbf{A}^{\mathrm{T}})}\mathrm{Tr}(u(\mathbf{A})v(\mathbf{A}^{\mathrm{T}})) \\
=&(v(\mathbf{A}^{\mathrm{T}}))^{\mathrm{T}}\nabla_{\mathbf{A}}u(\mathbf{A}) + (\nabla_{\mathbf{A}^{\mathrm{T}}:v(\mathbf{A}^{\mathrm{T}})}\mathrm{Tr}(u(\mathbf{A})v(\mathbf{A}^{\mathrm{T}}))^{\mathrm{T}} \\
=&\mathbf{C}^{\mathrm{T}}\mathbf{AB}^{\mathrm{T}} + ((u(\mathbf{A}))^{\mathrm{T}}\nabla_{\mathbf{A}^{\mathrm{T}}}v(\mathbf{A}^{\mathrm{T}}))^{\mathrm{T}} \\
=&\mathbf{C}^{\mathrm{T}}\mathbf{AB}^{\mathrm{T}} + (\mathbf{B}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{C}^{\mathrm{T}})^{\mathrm{T}} \\
=&\mathbf{CAB} + \mathbf{C}^{\mathrm{T}}\mathbf{AB}^{\mathrm{T}}
\end{aligned}$$

Here we make use of the property of the derivative of product: $(u(x)v(x))' = u'(x)v(x) + u(x)v'(x)$. The notation $\nabla_{\mathbf{A}:u(\mathbf{A})}$ means to calculate the derivative w.r.t. $\mathbf{A}$ only on $u(\mathbf{A})$. Same applies to $\nabla_{\mathbf{A}^{\mathrm{T}}:v(\mathbf{A}^{\mathrm{T}})}$. Here chain rule is used. Note that the conversion from $\nabla_{\mathbf{A}:v(\mathbf{A}^{\mathrm{T}})}$ to $\nabla_{\mathbf{A}^{\mathrm{T}}:v(\mathbf{A}^{\mathrm{T}})}$ is based on Eq. 5.

4 An Example on Least-square Linear Regression

Now we will derive the solution for least-square linear regression in matrix form, using the properties shown above. We know that the least-square linear regression has a closed-form solution (often referred as *normal equation*).

Assume we have N data points $\{\mathbf{x}^{(i)}, y^{(i)}\}^{1:N}$, and the linear regression function $h_\theta(\mathbf{x})$ is parametrized by θ. We can rearrange the data to matrix form:

$$\mathbf{X} = \begin{bmatrix} (\mathbf{x}^{(1)})^{\mathrm{T}} \\ (\mathbf{x}^{(2)})^{\mathrm{T}} \\ \vdots \\ (\mathbf{x}^{(N)})^{\mathrm{T}} \end{bmatrix} \qquad \mathbf{y} = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(N)} \end{bmatrix}$$

Thus the error can be represented as:

$$\mathbf{X}\theta - \mathbf{y} = \begin{bmatrix} h_\theta(\mathbf{x}^{(1)}) - \mathbf{y}^{(1)} \\ h_\theta(\mathbf{x}^{(2)}) - \mathbf{y}^{(2)} \\ \vdots \\ h_\theta(\mathbf{x}^{(N)}) - \mathbf{y}^{(N)} \end{bmatrix}$$

The squared error $E(\theta)$, according to the numerical definition:

$$E(\theta) = \frac{1}{2}\sum_{i=1}^{N}(h_\theta(\mathbf{x}^{(i)}) - \mathbf{y}^{(i)})^2$$

which is equivalent to the matrix form:

$$E(\theta) = \frac{1}{2}(\mathbf{X}\theta - \mathbf{y})^{\mathrm{T}}(\mathbf{X}\theta - \mathbf{y})$$

Take the derivative:

$$\begin{aligned} \nabla_\theta E(\theta) &= \nabla\frac{1}{2}\underbrace{(\mathbf{X}\theta - \mathbf{y})^{\mathrm{T}}(\mathbf{X}\theta - \mathbf{y})}_{1\times 1 \text{ matrix, thus } \mathrm{Tr}(\cdot)=(\cdot)} \\ &= \frac{1}{2}\nabla\mathrm{Tr}(\theta^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\theta - \mathbf{y}^{\mathrm{T}}\mathbf{X}\theta - \theta^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y} + \mathbf{y}^{\mathrm{T}}\mathbf{y}) \\ &= \frac{1}{2}\nabla\mathrm{Tr}(\theta^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\theta) - \nabla\mathrm{Tr}(\mathbf{y}^{\mathrm{T}}\mathbf{X}\theta) - \nabla\mathrm{Tr}(\theta^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y}) \\ &= \frac{1}{2}\nabla\mathrm{Tr}(\theta\,\mathbf{I}\,\theta^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}) - (\mathbf{y}^{\mathrm{T}}\mathbf{X})^{\mathrm{T}} - \mathbf{X}^{\mathrm{T}}\mathbf{y} \end{aligned}$$

The first term can be computed using Eq. 10, where $\mathbf{A} = \theta$, $\mathbf{B} = \mathbf{I}$, and $\mathbf{C} = \mathbf{X}^{\mathrm{T}}\mathbf{X}$ (Note that in this case, $\mathbf{C} = \mathbf{C}^{\mathrm{T}}$). Plug back to the derivation:

$$\begin{aligned} \nabla_\theta E(\theta) &= \frac{1}{2}(\mathbf{X}^{\mathrm{T}}\mathbf{X}\theta + \mathbf{X}^{\mathrm{T}}\mathbf{X}\theta - 2\mathbf{X}^{\mathrm{T}}\mathbf{y}) \\ &= \frac{1}{2}(2\mathbf{X}^{\mathrm{T}}\mathbf{X}\theta - 2\mathbf{X}^{\mathrm{T}}\mathbf{y}) \\ &\xRightarrow{\text{Set to 0}} \mathbf{X}^{\mathrm{T}}\mathbf{X}\theta = \mathbf{X}^{\mathrm{T}}\mathbf{y} \\ \theta_{\mathrm{LS}} &= (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} \end{aligned}$$

The normal equation is obtained in matrix form.