The Art of Matrix Derivative

Starfly
starfly3119@gmail.com

Beihang University — February 11, 2020

Introduction

Matrix derivative may be confusing if you are new into this field. Recently, I am having Andrew Ng's Neural Networks and Deep Learning. I hope this article may help you. In the first part, I will figure out the formulas in the course step by step. In the second part, I will tell you how to build a systematic method of matrix derivative and this part is translated from an answer in Zhihu.

1 Backward Propagation

1.1 Overview

In Week4 course, for the l-th layer, we have:

Algorithm 1: `l-th layer Back Propagation`

Input: $da^{[l]}$
Result: $da^{[l-1]}, dW^{[l]}, db^{[l]}$

$dz^{[l]} = da^{[l]} \odot g^{[l]'}(z^{[l]})$;
$dW^{[l]} = dz^{[l]} \cdot a^{[l-1]^T}$;
$db^{[l]} = dz^{[l]}$;
$da^{[l-1]} = W^{[l]^T} \cdot dz^{[l]}$;
return $da^{[l-1]}, dW^{[l]}, db^{[l]}$;

To avoid the collision of the notations, use the following ones:

Table 1: Notations

Andrew Ng	This article	Illustration
Loss	L	The loss function
$da^{[l]}$	$\frac{\partial L}{\partial a^{[l]}}$	The partial derivative of L with respect to $a^{[l]}$
$dW^{[l]}$	$\frac{\partial L}{\partial W^{[l]}}$	The partial derivative of L with respect to $W^{[l]}$
$db^{[l]}$	$\frac{\partial L}{\partial b^{[l]}}$	The partial derivative of L with respect to $b^{[l]}$
$dz^{[l]}$	$\frac{\partial L}{\partial z^{[l]}}$	The partial derivative of L with respect to $z^{[l]}$

Therefore, the above Algorithm can be denoted as:

Algorithm 2: `l-th layer Back Propagation`

Input: $\frac{\partial L}{\partial a^{[l]}}$
Result: $\frac{\partial L}{\partial a^{[l-1]}}, \frac{\partial L}{\partial W^{[l]}}, \frac{\partial L}{\partial b^{[l]}}$

$\frac{\partial L}{\partial z^{[l]}} = \frac{\partial L}{\partial a^{[l]}} \odot g^{[l]'}(z^{[l]})$;
$\frac{\partial L}{\partial W^{[l]}} = \frac{\partial L}{\partial z^{[l]}} \cdot a^{[l-1]^T}$;
$\frac{\partial L}{\partial b^{[l]}} = \frac{\partial L}{\partial z^{[l]}}$;
$\frac{\partial L}{\partial a^{[l-1]}} = W^{[l]^T} \cdot \frac{\partial L}{\partial z^{[l]}}$;
return $\frac{\partial L}{\partial a^{[l-1]}}, \frac{\partial L}{\partial W^{[l]}}, \frac{\partial L}{\partial b^{[l]}}$;

Now let's figure out the above formulas step by step.

1.2 Example

Remark: Don't worry if you are confused about some process in the following step. You can quickly glance at this example and read the **Part II** first. And then it may be much eaiser for you to understand the whole process.

Take a two layer neural netork as an example:

For an training example (x, y), according to the forward propagation, we have:

$$z^{[1]} = W^{[1]}x + b^{[1]} \tag{1}$$
$$a^{[1]} = g^{[1]}(z^{[1]}) \tag{2}$$
$$z^{[2]} = W^{[2]}a^{[1]} + b^{[2]} \tag{3}$$
$$a^{[2]} = g^{[2]}(z^{[2]}) \tag{4}$$

The loss funtion with respect to (x, y) is:

$$L = Loss(a^{[2]}, y) \tag{5}$$

Differentiate for L:

$$dL = Loss'(a^{[2]}, y)da^{[2]} \tag{6}$$

Therefore:

$$\frac{\partial L}{\partial a^{[2]}} = Loss'(a^{[2]}, y) \tag{7}$$

Apply:

$$\begin{aligned} dL &= \frac{\partial L}{\partial a^{[2]}} da^{[2]} \\ &= \frac{\partial L}{\partial a^{[2]}} d(g^{[2]}(z^{[2]})) \\ &= \frac{\partial L}{\partial a^{[2]}} g^{[2]'}(z^{[2]})dz^{[2]} \end{aligned} \tag{8}$$

Therefore:

$$\frac{\partial L}{\partial z^{[2]}} = \frac{\partial L}{\partial a^{[2]}} g^{[2]'}(z^{[2]}) \tag{9}$$

Apply:

$$\begin{aligned}
dL &= \frac{\partial L}{\partial z^{[2]}} dz^{[2]} \\
&= \frac{\partial L}{\partial z^{[2]}} d(W^{[2]} a^{[1]} + b^{[2]}) \\
&= tr(\frac{\partial L}{\partial z^{[2]}} dW^{[2]} a^{[1]}) + tr(\frac{\partial L}{\partial z^{[2]}} W^{[2]} da^{[1]}) + tr(\frac{\partial L}{\partial z^{[2]}} db^{[2]}) \\
&= tr(a^{[1]} \frac{\partial L}{\partial z^{[2]}} dW^{[2]}) + tr(\frac{\partial L}{\partial z^{[2]}} W^{[2]} da^{[1]}) + tr(\frac{\partial L}{\partial z^{[2]}} db^{[2]})
\end{aligned} \tag{10}$$

Here we get:

$$\frac{\partial L}{\partial W^{[2]}} = \frac{\partial L}{\partial z^{[2]}} a^{[1]^T} \tag{11}$$

$$\frac{\partial L}{\partial a^{[1]}} = W^{[2]^T} \frac{\partial L}{\partial z^{[2]}} \tag{12}$$

$$\frac{\partial L}{\partial b^{[2]}} = \frac{\partial L}{\partial z^{[2]}} \tag{13}$$

Use L_2 to denote the second component of the (10) equation.
Apply:

$$\begin{aligned}
dL_2 &= tr(\frac{\partial L^T}{\partial a_1} da_1) \\
&= tr(\frac{\partial L^T}{\partial a_1} d(g^{[1]}(z^{[1]}))) \\
&= tr(\frac{\partial L^T}{\partial a_1} g^{[1]'}(z^{[1]}) \odot dz^{[1]}) \\
&= tr((\frac{\partial L}{\partial a_1} \odot g^{[1]'}(z^{[1]}))^T dz^{[1]})
\end{aligned} \tag{14}$$

Therefore:

$$\frac{\partial L}{\partial z^{[1]}} = \frac{\partial L_2}{\partial z^{[1]}} = \frac{\partial L}{\partial a_1} \odot g^{[1]'}(z^{[1]}) \tag{15}$$

Again, apply:

$$\begin{aligned}
dL_2 &= tr(\frac{\partial L^T}{\partial z^{[1]}} dz^{[1]}) \\
&= tr(\frac{\partial L^T}{\partial z^{[1]}} d(W^{[1]} a^{[0]} + b^{[1]})) \\
&= tr(\frac{\partial L^T}{\partial z^{[1]}} dW^{[1]} a^{[0]}) + tr(\frac{\partial L^T}{\partial z^{[1]}} W^{[1]} da^{[0]}) + tr(\frac{\partial L^T}{\partial z^{[1]}} db^{[1]}) \\
&= tr(a^{[0]} \frac{\partial L^T}{\partial z^{[1]}} dW^{[1]}) + tr(\frac{\partial L^T}{\partial z^{[1]}} W^{[1]} da^{[0]}) + tr(\frac{\partial L^T}{\partial z^{[1]}} db^{[1]})
\end{aligned} \tag{16}$$

Here we get:

$$\frac{\partial L}{\partial W^{[1]}} = \frac{\partial L_2}{\partial W^{[1]}} = \frac{\partial L}{\partial z^{[1]}} a^{[0]^T} \tag{17}$$

$$\frac{\partial L}{\partial b^{[1]}} = \frac{\partial L_2}{\partial b^{[1]}} = \frac{\partial L}{\partial z^{[1]}} \tag{18}$$

where $a^{[0]} = x$

2 Matrix derivative

2.1 Core relationship

2.1.1 Single variable calculus

In single variable calculus, we have:

$$df = f'(x)dx \tag{19}$$

where df is the **differential**, $f'(x)$ is the **derivative** of f with respect to x.

2.1.2 Multi-variable calculus

In multi-variable calculus, we have:

$$df = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} dx_i = \frac{\partial f}{\partial \mathbf{x}}^T d\mathbf{x} \tag{20}$$

where f is a **scalar-valued** function, df is the **total differential**, $\frac{\partial f}{\partial x_i}$ is the **partial derivative** of f with respect to x_i, **x** is an n-by-1 vector, $\frac{\partial f}{\partial \mathbf{x}}$ is the **gradient** of f.

The above formula tells us the **total differential** is the **inner product** of **gradient vector** $\frac{\partial f}{\partial x}$ and **dx**.

> Question 1
>
> What if X is an m-by-n matrix?

We have:

$$df = \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{\partial f}{\partial X_{ij}} dX_{ij} \tag{21}$$

where f is a **scalar-valued** function, X is an m-by-n matrix. Inspired by the (20) equation, we can give the following formula:

$$df = \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{\partial f}{\partial X_{ij}} dX_{ij} = tr(\frac{\partial f}{\partial X}^T dX) \tag{22}$$

Here we get our core relationship:

$$df = tr(\frac{\partial f}{\partial X}^T dX) \tag{23}$$

Similarily, we can conclude that the **total differential** is the **inner product** of $\frac{\partial f}{\partial X}$ and dX.

2.1.3 Example

$f(X)$ is a **scalar-valued** function. $X = \begin{bmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{bmatrix}$

$$df = \frac{\partial f}{\partial X_{11}} dX_{11} + \frac{\partial f}{\partial X_{12}} dX_{12} + \frac{\partial f}{\partial X_{21}} dX_{21} + \frac{\partial f}{\partial X_{22}} dX_{22} \tag{24}$$

$$\frac{\partial f}{\partial X} = \begin{bmatrix} \frac{\partial f}{\partial X_{11}} & \frac{\partial f}{\partial X_{12}} \\ \frac{\partial f}{\partial X_{21}} & \frac{\partial f}{\partial X_{22}} \end{bmatrix} \tag{25}$$

$$dX = \begin{bmatrix} dX_{11} & dX_{12} \\ dX_{21} & dX_{22} \end{bmatrix} \tag{26}$$

$$\frac{\partial f}{\partial X}^T dX = \begin{bmatrix} \frac{\partial f}{\partial X_{11}} dX_{11} + \frac{\partial f}{\partial X_{21}} dX_{21} & \frac{\partial f}{\partial X_{11}} dX_{12} + \frac{\partial f}{\partial X_{21}} dX_{22} \\ \frac{\partial f}{\partial X_{12}} dX_{11} + \frac{\partial f}{\partial X_{22}} dX_{21} & \frac{\partial f}{\partial X_{12}} dX_{12} + \frac{\partial f}{\partial X_{22}} dX_{22} \end{bmatrix} \tag{27}$$

$$tr(\frac{\partial f}{\partial X}^T dX) = \frac{\partial f}{\partial X_{11}} dX_{11} + \frac{\partial f}{\partial X_{12}} dX_{12} + \frac{\partial f}{\partial X_{21}} dX_{21} + \frac{\partial f}{\partial X_{22}} dX_{22} \tag{28}$$

From this example, we find that the equation(23) is right.

2.2 The rule of matrix differential

X and Y are two matrix with the same size:

(1) $d(X \pm Y) = dX \pm dY$;

(2) $d(XY) = (dX)Y + XdY$;

(3) $d(X^T) = (dX)^T$;

(4) $dtr(X) = tr(dX)$;

(5) If X in invertible, $dX^{-1} = -X^{-1}dXX^{-1}$;

(6) $d|X| = tr(X^* dX)$, where X^* is the adjoint matrix of X;

(7) $d(X \odot Y) = dX \odot Y + X \odot dY$, $\odot$ is the element-wise product;

(8) $d\sigma(X) = \sigma'(X) \odot dX$, $\sigma(X)$ is an element-wise function.

2.3 Trace tricks

(1) If a is a scalar, $a = tr(a)$;

(2) $tr(A^T) = tr(A)$, where A is an n-by-n matrix;

(3) $tr(A \pm B) = tr(A) \pm tr(B)$, where both A and B are n-by-n matrix;

(4) $tr(AB) = tr(BA)$, where A and B^T have the same size;

(5) $tr(A^T(B \odot C)) = tr((A \odot B)^T C)$, where A,B and C have the same size.

2.4 General method

In this section, I will clarify the general method of figuring out the derivative of a **scalar-valued** function with respect to a matrix X.

(a) Apply **the rule of matrix differential** to get the total differential of a **scalar-valued** function;

(b) Apply **trace tricks**;

(c) Apply the **core relationship** equation(23).

Remark: If the matrix X degenerate to be a vector, apply equation(20) in the (c) step instead.

2.5 Examples

In this section, I will give various examples to illustrate the usage of the general method.

2.5.1 Composition function

If f is a **scalar-valued** function with respect to a matrix Y, and $Y = AXB$, A, B are constant matrics, what is the derivative of f with respect to X?

Question 2

We have chain rule in computing the derivative of composition function, what about matrix derivative?

Actually, we haven't given the definition of **the derivative of a matrix with respect to a matrix**, so we cannot apply the chain rule directly. Let's see how to figure it out.

Apply **core relationship**:

$$\begin{aligned} df &= tr(\frac{\partial f^T}{\partial Y}dY) \\ &= tr(\frac{\partial f^T}{\partial Y}d(AXB)) \\ &= tr(\frac{\partial f^T}{\partial Y}AdXB) \end{aligned} \tag{29}$$

Apply **trace tricks(4)**:

$$df = tr(B\frac{\partial f^T}{\partial Y}AdX) \tag{30}$$

Therefore:

$$\frac{\partial f}{\partial X} = A^T\frac{\partial f}{\partial Y}B^T \tag{31}$$

Remark: The idea of the above example is used several times in the **PartI** process.

2.5.2 Warm up 1

If the **scalar-valued** function $f = a^T Xb$, where a is an m-by-1 column vector, X is an m-by-n matrix, b is an n-by-1 column vector, try to figure out $\frac{\partial f}{\partial X}$.

$$\begin{aligned} df &= d(a^T Xb) \\ &= a^T dXb \\ &= tr(a^T dXb) \\ &= tr(ba^T dX) \end{aligned} \tag{32}$$

Therefore:

$$\frac{\partial f}{\partial X} = ab^T \tag{33}$$

2.5.3 Warm up 2

If the **scalar-valued** function $f = a^T exp(Xb)$, where a is an m-by-1 column vector, X is an m-by-n matrix, b is an n-by-1 column vector, $exp(\cdot)$ is an element-wise function, try to figure out $\frac{\partial f}{\partial X}$.

$$\begin{aligned} df &= a^T(exp(Xb) \odot (dXb) \\ &= tr(a^T(exp(Xb) \odot (dXb)) \\ &= tr((a \odot exp(Xb))^T dXb) \\ &= tr(b(a \odot exp(Xb))^T dX) \end{aligned} \tag{34}$$

Therefore:

$$\frac{\partial f}{\partial X} = (a \odot exp(Xb))b^T \tag{35}$$

2.5.4 Warm up 3

If the **scalar-valued** function $f = tr(Y^T MY), Y = \sigma(WX)$, where W is l-by-m matrix, X is m-by-n matrix, Y is l-by-n matrix, M is l-by-l symmetric matrix, try to figure out $\frac{\partial f}{\partial X}$.

$$\begin{aligned} df &= tr((dY)^T MY) + tr(Y^T MdY) \\ &= tr(Y^T M^T dY) + tr(Y^T MdY) \\ &= tr(2Y^T MdY) \end{aligned} \tag{36}$$

Therefore:

$$\frac{\partial f}{\partial Y} = 2MY \tag{37}$$

Apply **core relationship**:

$$\begin{aligned} df &= tr(\frac{\partial f^T}{\partial Y} dY) \\ &= tr(\frac{\partial f^T}{\partial Y}(\sigma'(WX) \odot (WdX))) \\ &= tr((\frac{\partial f}{\partial Y} \odot \sigma'(WX))^T (WdX)) \end{aligned} \tag{38}$$

Therefore:

$$\begin{aligned} \frac{\partial f}{\partial X} &= W^T(\frac{\partial f}{\partial Y} \odot \sigma'(WX)) \\ &= W^T((2M\sigma(WX)) \odot \sigma'(WX)) \end{aligned} \tag{39}$$

2.5.5 Linear regression

In Andrew Ng's Machine Learning course, he introduces two methods in linear regression. One of them is **Normal equation**. Let see how to figure out the **Normal equation** method.

If a **scalar-valued** function $l = \|Xw - y\|^2$, where y is an m-by-1 column vector, X is an m-by-n matrix, w is an n-by-1 column vector. Try to figure the **Least square estimation** of w.

$$l = \|Xw - y\|^2 = (Xw - y)^T (Xw - y) \tag{40}$$

$$\begin{aligned} dl &= (Xdw)^T (Xw - y) + (Xw - y)^T (Xdw) \\ &= 2(Xw - y)^T X dw \end{aligned} \tag{41}$$

Therefore:

$$\frac{\partial l}{\partial w} = 2X^T(Xw - y) \tag{42}$$

Set $\frac{\partial l}{\partial w} = 0$, we get:

$$XX^T w = X^T y \tag{43}$$

Therefore, the **Least square estimation** of w is:

$$\hat{w} = (X^T X)^{-1} X^T y \tag{44}$$