CS2109S Matrix calculus Cheatsheet

Notation	Meaning
$x, y, \epsilon \in \mathbb{R}$	scalar
$oldsymbol{x},oldsymbol{y},oldsymbol{\epsilon}\in\mathbb{R}^d$	vector
$oldsymbol{X},oldsymbol{Y},oldsymbol{W}\in\mathbb{R}^{d imes n}$	matrix
$x_i(X_{ij})$	entries of vector(matrix)
c,k	#classes
d,m	$\# { m features}$
n	#samples
in, out	#input/output features
b, w_0	bias
l,ϵ	loss, error

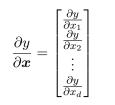
Scalar calculus

The entries of vector and matrix follow these rules as well. $f, g : \mathbb{R} \to \mathbb{R}$ are functions of $x. a \in \mathbb{R}$.

Expression	Derivative w.r.t. x
a	0
$a \cdot f \\ x^n$	$a \cdot \frac{df}{dx}$ nx^{n-1}
x^n	nx^{n-1}
f + g	$\frac{df}{df} + \frac{dg}{dg}$
f-g	$\frac{dy}{dx} - \frac{dy}{dx}$
$f\cdot g$	$f \cdot \frac{dg}{dx} + \frac{df}{dx} \cdot g$
f(g(x))	$\frac{df(u)}{du}\frac{du}{dx}$, let $u = g(x)$

Matrix calculus

Using denominator layout $(y \in \mathbb{R}, x, \frac{\partial y}{\partial x} \in \mathbb{R}^d)$



Hessian formulation

For vector-valued functions $\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x}) : \mathbb{R}^d \to \mathbb{R}^c$

$$\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}} = \begin{bmatrix} \frac{\partial y_1}{\partial \boldsymbol{x}} & \frac{\partial y_2}{\partial \boldsymbol{x}} & \dots & \frac{\partial y_c}{\partial \boldsymbol{x}} \end{bmatrix}$$

Common vector derivatives

 $\boldsymbol{b} \in \mathbb{R}^d$ and $\boldsymbol{A} \in \mathbb{R}^{d \times c}$ are not functions on $\boldsymbol{x} \in \mathbb{R}^d$

f(x)	$rac{\partial oldsymbol{f}}{\partial oldsymbol{x}}$
$A^T x$	A
$oldsymbol{b}^Toldsymbol{x}$	\boldsymbol{b}
$oldsymbol{x}^Toldsymbol{b}$	b
$oldsymbol{x}^Toldsymbol{x}$	$2\boldsymbol{x}$
$\boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x}$	2Ax

Chain rule in matrix form

Since matrix multiplication does not commute, the order of the derivatives matters in the chain rule.

For instance, given $y = b^T (X^T a)$. Let $h = X^T a$ and k = Xb. Then,

$$\frac{\partial y}{\partial \boldsymbol{a}} = \frac{\partial \boldsymbol{h}}{\partial \boldsymbol{a}} \frac{\partial y}{\partial \boldsymbol{h}} = \boldsymbol{X}\boldsymbol{b}, \quad \frac{\partial y}{\partial \boldsymbol{X}} = \frac{\partial y}{\partial \boldsymbol{k}} (\frac{\partial \boldsymbol{k}}{\partial \boldsymbol{X}})^T = \boldsymbol{a}\boldsymbol{b}^T.$$

The order the derivatives might vary, but it can be determined by a shape consistency check.

Matrix chain rule, for back propagation

Given $\boldsymbol{Y} = \boldsymbol{W}^T \boldsymbol{X}, \, \epsilon = l(\boldsymbol{Y})$ as a loss function.

• Derivative to update weight:

$$\frac{\partial \epsilon}{\partial \boldsymbol{W}} = \boldsymbol{X} \cdot (\frac{\partial \epsilon}{\partial \boldsymbol{Y}})^T$$

• Derivative to be carried to the previous layer:

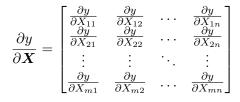
$$\frac{\partial \epsilon}{\partial \boldsymbol{X}} = \boldsymbol{W} \cdot \frac{\partial \epsilon}{\partial \boldsymbol{Y}}$$

Note: Use the shape of matrices to determine the order.

Matrix derivatives

Scalar-by-Matrix: the shape of a scalar-by-matrix derivative is the same as that of the matrix.

For instance:



The shape of the resulting derivative is the same as the shape of X.

Notation for models

- Linear regression Hypothesis function: $h(\boldsymbol{w}^T \boldsymbol{x})$, where $\boldsymbol{x}, \boldsymbol{w} \in \mathbb{R}^d$.
- Logistic regression Apply the same shape as above.

SVM Apply the same shape as above.

Neural Network One layer: $\boldsymbol{Y} = \boldsymbol{W}^T \boldsymbol{X}$, where $\boldsymbol{X} \in \mathbb{R}^{in \times n}, \boldsymbol{W} \in \mathbb{R}^{in \times out}, \boldsymbol{Y} \in \mathbb{R}^{out \times n}$.