COS 302 Precept 6

Spring 2020

Princeton University

Outline

Overview of Matrix Gradients

Review of Differentiation Basics

Examples of Various Gradients

Outline

Different Flavors of Gradients

Type	Scalar	Vector	Matrix
Scalar	$\frac{\partial y}{\partial x}$	$\frac{\partial \mathbf{y}}{\partial x}$	$\frac{\partial \mathbf{Y}}{\partial x}$
Vector	$\frac{\partial y}{\partial \mathbf{x}}$	$\frac{\partial \mathbf{y}}{\partial \mathbf{x}}$	
Matrix	$\frac{\partial y}{\partial \mathbf{X}}$		

[1] source: https://www.comp.nus.edu.sg/ cs5240/lecture/matrix-differentiation.pdf

Challenges of Vector/Matrix Calculus

- Good news: most of the rules you know and love from single variable calculus generalize well (not in all cases but in some).
- Bad news: confusing notation (many more variables lead to very tedious book-keeping) and more identities to memorize.

Numerator vs Denominator Layout

- Two main conventions used in vector/matrix calculus: numerator and denominator layouts.
- Numerator layout makes the dimension of the derivative be the numerator dimension by denominator dimension.
- For example, if y is a scalar and $\boldsymbol{x} \in \mathbb{R}^N$ then

$$\frac{\partial y}{\partial \boldsymbol{x}} \in \mathbb{R}^{1 \times N}$$

How to Become a Differentiation Master

1. Identify what is the flavor of your derivative.
2. What is the dimension of the derivative?
3. Identify any differentiation rules you will need for this case.
4. Can any part of the derivative be reduced to a particular identity?
5. Identify the partial derivatives.

Outline

Overview of Matrix Gradients

Review of Differentiation Basics

Examples of Various Gradients

Definitions

Names	Notation & Expression
Difference Quotient	$\frac{\delta y}{\delta x} = \frac{f(x+\delta x)-f(x)}{\delta x}$
Derivative	$\frac{df}{dx} = \lim_{h\to 0} \frac{f(x+h)-f(x)}{h}$
Partial Derivative	$\frac{\partial f}{\partial dx_i} = \lim_{h\to 0} \frac{f(x_1,\ldots,x_i+h,\ldots,x_n)-f(x)}{h}$
Gradient	$\nabla_{\boldsymbol{x}} f = \frac{df}{d\boldsymbol{x}} = \left[\frac{\partial f}{\partial dx_1} \ldots \frac{\partial f}{\partial dx_n}\right]$
Jacobian	$\nabla_{\boldsymbol{x}} \boldsymbol{f} = \frac{d\boldsymbol{f}(\boldsymbol{x})}{d\boldsymbol{x}} = \left[\frac{\partial \boldsymbol{f}(\boldsymbol{x})}{\partial dx_1} \ldots \frac{\partial \boldsymbol{f}(\boldsymbol{x})}{\partial dx_n}\right]$

Differentiation Rules (Scalar-Scalar)

- Sum Rule

$$(f(x)+g(x))' = f'(x)+g'(x)$$

- Product Rule

$$(f(x)g(x))' = f'(x)g(x)+f(x)g'(x)$$

- Quotient Rule

$$\left(\frac{f(x)}{g(x)}\right)' = \frac{f'(x)g(x)-f(x)g'(x)}{(g(x))^2}$$

- Chain Rule

$$(g(f(x)))' = (g\circ f)'(x) = g'(f(x))f'(x)$$

Differentiation Rules (Scalar-Vector)

- Sum Rule

$$\frac{\partial}{\partial \boldsymbol{x}}(f(\boldsymbol{x}) + g(\boldsymbol{x})) = \frac{\partial f}{\partial \boldsymbol{x}} + \frac{\partial g}{\partial \boldsymbol{x}}$$

- Product Rule

$$\frac{\partial}{\partial \boldsymbol{x}}(f(\boldsymbol{x})g(\boldsymbol{x})) = \frac{\partial f}{\partial \boldsymbol{x}}g(\boldsymbol{x}) + f(\boldsymbol{x})\frac{\partial g}{\partial \boldsymbol{x}}$$

- Chain Rule

$$\frac{\partial}{\partial \boldsymbol{x}}(g(f(\boldsymbol{x}))) = \frac{\partial}{\partial \boldsymbol{x}}(g \circ f)(\boldsymbol{x}) = \frac{\partial g}{\partial f}\frac{\partial f}{\partial \boldsymbol{x}}$$

Outline

Gradient of Matrix Multiplication

Consider the matrix $\boldsymbol{A} \in \mathbb{R}^{M \times N}$ and the vector $\boldsymbol{x} \in \mathbb{R}^N$. Define the vector function $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x}$ where $\boldsymbol{f} : \mathbb{R}^N \to \mathbb{R}^M$, what is $\frac{d\boldsymbol{f}}{d\boldsymbol{x}}$?

Gradient of Matrix Multiplication cont.

1. Dimension of gradient: $\frac{d\boldsymbol{f}}{d\boldsymbol{x}} \in \mathbb{R}^{M \times N}$
2. One of these $M \times N$ partial derivatives will look like:

$$f_i(\boldsymbol{x}) = \sum_{j=1}^{N} A_{ij} x_j \implies \frac{\partial f_i}{\partial x_j} = A_{ij}$$

3. Collecting all of these partial derivatives:

$$\frac{d\boldsymbol{f}}{d\boldsymbol{x}} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_N} \\ \vdots & & \vdots \\ \frac{\partial f_M}{\partial x_1} & \cdots & \frac{\partial f_M}{\partial x_N} \end{bmatrix} = \begin{bmatrix} A_{11} & \cdots & A_{1N} \\ \vdots & & \vdots \\ A_{M1} & \cdots & A_{MN} \end{bmatrix} = \boldsymbol{A} \in \mathbb{R}^{M \times N}$$

Chain Rule Example

Consider the scalar function $h : \mathbb{R} \to \mathbb{R}$ where $h(t) = f(g(t))$ with $f : \mathbb{R}^2 \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}^2$ such that

$$f(\boldsymbol{x}) = \exp(x_1 x_2^2) \quad ,$$

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = g(t) = \begin{bmatrix} t \cos t \\ t \sin t \end{bmatrix} .$$

What is $\frac{dh}{dt}$?

Chain Rule Example cont.

Even though the gradient is a scalar we need to compute two vector gradients (gradients) because of the chain rule:

$$\frac{dh}{dt} = \frac{\partial f}{\partial \boldsymbol{x}}\frac{\partial \boldsymbol{x}}{\partial t} = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} \end{bmatrix} \begin{bmatrix} \frac{\partial x_1}{\partial t} \\ \frac{\partial x_2}{\partial t} \end{bmatrix}$$

$$= \begin{bmatrix} \exp(x_1 x_2^2)x_2^2 & 2\exp(x_1 x_2^2)x_1 x_2 \end{bmatrix} \begin{bmatrix} \cos t - t\sin t \\ \sin t + t\cos t \end{bmatrix}$$

$$= \exp(x_1 x_2^2)(x_2^2(\cos t - x_2) + 2x_1 x_2(\sin t + x_1)),$$

where $x_1 = t\cos t$ and $x_2 = t\sin t$

Least Squares by Chain Rule

Consider the matrix $\boldsymbol{X} \in \mathbb{R}^{N \times D}$ where $N > D$ and the two vectors $\boldsymbol{y} \in \mathbb{R}^N$ and $\boldsymbol{\beta} \in \mathbb{R}^D$. Before we saw in class that the over-determined system of linear equations:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}$$

does not always have a solution. Instead of solving the problem directly we can try to find an approximate solution $\hat{\boldsymbol{\beta}}$.

Least Squares by Chain Rule cont.

If we picked a random $\boldsymbol{\beta}$, and multiplied it by $\boldsymbol{X}$, we probably wouldn't get a vector that was very close to $\boldsymbol{y}$. Specifically, the error vector

$$\boldsymbol{e}(\boldsymbol{\beta}) = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}$$

would probably not be close to the zero vector. A good choice of $\boldsymbol{\beta}$ is one that minimizes the Euclidean distance between $\boldsymbol{y}$ and $\boldsymbol{X}\boldsymbol{\beta}$. Specifically, one that minimizes the function $L(\boldsymbol{e}) = \|\boldsymbol{e}\|^2$.

Least Squares by Chain Rule cont.

To find the best $\boldsymbol{\beta}$, let's take the gradient of L with respect to $\boldsymbol{\beta}$ and set it equal to zero.

1. $\frac{\partial L}{\partial \boldsymbol{\beta}} \in \mathbb{R}^{1 \times D}$
2. $\frac{\partial L}{\partial \boldsymbol{\beta}} = \frac{\partial L}{\partial \boldsymbol{e}} \frac{\partial \boldsymbol{e}}{\partial \boldsymbol{\beta}}$
3. $\frac{\partial L}{\partial e_i} = \frac{\partial}{\partial e_i} \sum_{i=1}^{N} e_i^2 = 2e_i$. Since, $\frac{\partial L}{\partial \boldsymbol{e}} \in \mathbb{R}^{1 \times N}$ we have $\frac{\partial L}{\partial \boldsymbol{e}} = 2\boldsymbol{e}^T$
4. $\frac{\partial \boldsymbol{e}}{\partial \boldsymbol{\beta}} = -\boldsymbol{X} \in \mathbb{R}^{N \times D}$
5. $\frac{\partial L}{\partial \boldsymbol{\beta}} = -2\boldsymbol{e}^T \boldsymbol{X} = -2(\boldsymbol{y}^T - \boldsymbol{\beta}^T \boldsymbol{X}^T)\boldsymbol{X} = 0 \implies \boldsymbol{\beta} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y}$