2.1. Differential Calculus

Lecture based on "Dive into Deep Learning" **http://D2L.AI** (Zhang et al., 2020)

Prof. Dr. Christoph Lippert

Digital Health & Machine Learning

In deep learning, our goal is to fit a model that performs well on data that we have never seen before.

We do so by **train** models by minimizing a **loss function** on training data.

Thus we can decompose the task of fitting models into two key concerns:

i. **optimization**: fitting a model to observed training data
ii. **generalization**: the mathematical principles and practitioners' wisdom that guide as to how to produce models whose validity extends beyond the exact set of data points used to train them.

We begin by addressing the calculation of derivatives, a crucial step in nearly all deep learning optimization algorithms.

In deep learning, we typically choose loss functions that are differentiable with respect to our model's parameters.

Put simply, this means that for each parameter, we can determine how rapidly the loss would increase or decrease, were we to **increase** or **decrease** that parameter by an infinitesimally small amount.

Let $f : \mathbb{R} \to \mathbb{R}$.

The **derivative** of f is defined as

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h},$$

if this limit exists.

- If $f'(a)$ exists, f is said to be **differentiable** at a.
- If f is differentiable at every number of an interval, then this function is differentiable on this interval.
- The derivative $f'(x)$ is the **instantaneous** rate of change of $f(x)$ with respect to x.
- It is based on the variation h in x, which approaches 0.

Given $y = f(x)$, where x and y are the independent variable and the dependent variable of the function f, respectively.

The following expressions are equivalent:

$$f'(x) = y' = \frac{dy}{dx} = \frac{df}{dx} = \frac{d}{dx}f(x) = Df(x) = D_x f(x),$$

where symbols $\frac{d}{dx}$ and D are **differentiation operators**.

We can use the following rules to differentiate common functions:

(i) $DC = 0$ (C is a constant),

(ii) $Dx^n = nx^{n-1}$ (the **power rule**, $n \in \mathbb{R}$),

(iii) $De^x = e^x$,

(iv) $D\ln(x) = 1/x$.

Let functions f and g both be differentiable and let C be a constant, then the following holds:

i. the **constant multiple rule**

$$\frac{d}{dx}[Cf(x)] = C\frac{d}{dx}f(x)$$

ii. the **sum rule**

$$\frac{d}{dx}[f(x) + g(x)] = \frac{d}{dx}f(x) + \frac{d}{dx}g(x)$$

iii. the **product rule**

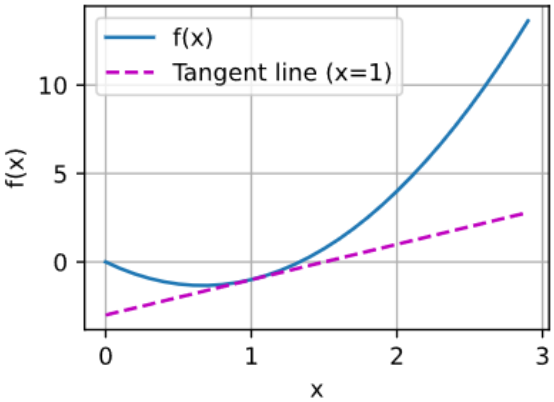
$$\frac{d}{dx}[f(x)g(x)] = f(x)\frac{d}{dx}[g(x)] + g(x)\frac{d}{dx}[f(x)]$$

iv. the **quotient rule**

$$\frac{d}{dx}\left[\frac{f(x)}{g(x)}\right] = \frac{g(x)\frac{d}{dx}[f(x)] - f(x)\frac{d}{dx}[g(x)]}{[g(x)]^2}$$

Example

$u = f(x) = 3x^2 - 4x.$



Let $y = f(x_1, x_2, \ldots, x_n)$ be a **multivariate** function.

The **partial derivative** of y with respect to x_i is

$$\frac{\partial y}{\partial x_i} = \lim_{h \to 0} \frac{f(x_1, \ldots, x_{i-1}, x_i + h, x_{i+1}, \ldots, x_n) - f(x_1, \ldots, x_i, \ldots, x_n)}{h}.$$

- For notation of partial derivatives, the following are equivalent:

$$\frac{\partial y}{\partial x_i} = \frac{\partial f}{\partial x_i} = f_{x_i} = f_i = D_i f = D_{x_i} f.$$

Let $f : \mathbb{R}^n \to \mathbb{R}$.

The input vector is $\mathbf{x} = [x_1, x_2, \ldots, x_n]^\top$.

The **gradient** vector of $f(\mathbf{x})$ with respect to $\mathbf{x}$ is

$$\nabla_{\mathbf{x}} f(\mathbf{x}) = \left[\frac{\partial f(\mathbf{x})}{\partial x_1}, \frac{\partial f(\mathbf{x})}{\partial x_2}, \ldots, \frac{\partial f(\mathbf{x})}{\partial x_n} \right]^\top,$$

where $\nabla_{\mathbf{x}} f(\mathbf{x})$ is often replaced by $\nabla f(\mathbf{x})$ when there is no ambiguity.

Let $\mathbf{x}$ be an n-dimensional vector, the following rules are often used when differentiating multivariate functions:

(i) $\nabla_{\mathbf{x}} \mathbf{A}\mathbf{x} = \mathbf{A}^\top$ for all $\mathbf{A} \in \mathbb{R}^{m \times n}$

(ii) $\nabla_{\mathbf{x}} \mathbf{x}^\top \mathbf{A} = \mathbf{A}$ for all $\mathbf{A} \in \mathbb{R}^{n \times m}$

(iii) $\nabla_{\mathbf{x}} \mathbf{x}^\top \mathbf{A}\mathbf{x} = (\mathbf{A} + \mathbf{A}^\top)\mathbf{x}$ for all $\mathbf{A} \in \mathbb{R}^{n \times n}$

(iv) $\nabla_{\mathbf{x}} \|\mathbf{x}\|^2 = \nabla_{\mathbf{x}} \mathbf{x}^\top \mathbf{x} = 2\mathbf{x}$.

Similarly, for any matrix $\mathbf{X}$, we have $\nabla_{\mathbf{X}} \|\mathbf{X}\|_F^2 = 2\mathbf{X}$.

Layout convention

$$\frac{\partial y}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y}{\partial x_1} \\ \frac{\partial y}{\partial x_2} \\ \vdots \\ \frac{\partial y}{\partial x_n} \end{bmatrix}$$

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_2}{\partial x_1} & \cdots & \frac{\partial y_m}{\partial x_1} \\ \frac{\partial y_1}{\partial x_2} & \frac{\partial y_2}{\partial x_2} & \cdots & \frac{\partial y_m}{\partial x_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y_1}{\partial x_n} & \frac{\partial y_2}{\partial x_n} & \cdots & \frac{\partial y_m}{\partial x_n} \end{bmatrix}$$

$$\frac{\partial \mathbf{y}}{\partial x} = \begin{bmatrix} \frac{\partial y_1}{\partial x} & \frac{\partial y_2}{\partial x} & \cdots & \frac{\partial y_m}{\partial x} \end{bmatrix}$$

$$\frac{\partial y}{\partial \mathbf{X}} = \begin{bmatrix} \frac{\partial y}{\partial x_{11}} & \frac{\partial y}{\partial x_{12}} & \cdots & \frac{\partial y}{\partial x_{1q}} \\ \frac{\partial y}{\partial x_{21}} & \frac{\partial y}{\partial x_{22}} & \cdots & \frac{\partial y}{\partial x_{2q}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y}{\partial x_{p1}} & \frac{\partial y}{\partial x_{p2}} & \cdots & \frac{\partial y}{\partial x_{pq}} \end{bmatrix}$$

See also ‘Denominator layout’ under https://en.wikipedia.org/wiki/Matrix_calculus

Let $y = f(u)$ and $u = g(x)$ be differentiable.
The **chain rule** states that

$$\frac{dy}{dx} = \frac{dy}{du}\frac{du}{dx}.$$

Suppose that the differentiable function y has variables $u_1, u_2, \ldots, u_m$, where each differentiable function u_i has variables $x_1, x_2, \ldots, x_n$.
Then the chain rule gives

$$\frac{dy}{dx_i} = \frac{dy}{du_1}\frac{du_1}{dx_i} + \frac{dy}{du_2}\frac{du_2}{dx_i} + \cdots + \frac{dy}{du_m}\frac{du_m}{dx_i}$$

for any $i = 1, 2, \ldots, n$.

Summary

- Differential calculus can be applied to optimization problems in deep learning.
- A **derivative** can be interpreted as the instantaneous rate of change of a function with respect to its variable.

  It is also the slope of the tangent line to the curve of the function.
- A **gradient** is a vector whose components are the partial derivatives of a multivariate function with respect to all its variables.
- The **chain rule** enables us to differentiate **composite** functions.