University of Stuttgart
Institute for Natural Language Processing

Thang Vu

Math for Machine Learning

Outline

Difference Quotient

1

Motivation

- In machine learning, we want to *build* a model which maps from inputs x to desired outputs y,
  E.g., predict how many ice cream scoops y are being sold depending on the temperature x.
  Model $f(x)$: e.g., a line: $f(x) = a \cdot x + b \approx y$
- Training such a model boils down to finding a *good* set of parameters (here: $a, b \in \mathbb{R}$)
- Good parameters are parameters that, e.g., minimize the distance between our model prediction $f(x)$ and real observed outcomes y

Motivation

- We need to choose a measure of distance between our model predictions $f(x_i)$ and observed n real outcomes (x_i, y_i) (called *loss function*):
  E.g., $\mathcal{L} = \frac{1}{N} \sum_1^N [f(x) - y]^2$
- Then, we can minimize this distance: By taking derivatives w.r.t. to the model parameters a, b

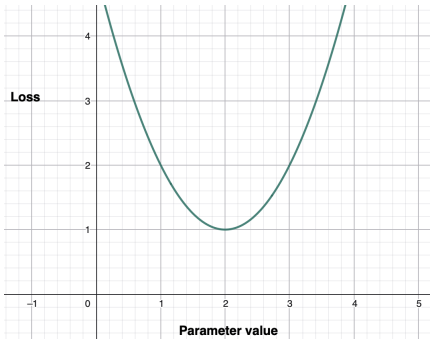

Figure: Loss for different model parameters

Motivation

- Since our model can become arbitrarily complex (e.g., neural network), there may not be an analytical solution
- But:
  The derivative (*gradient*) of f points into the direction of the steepest ascent of f.
  $\Rightarrow$ follow the negative gradient (*Gradient Descent*).

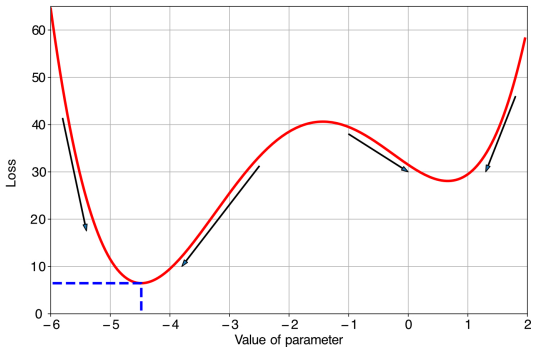

Figure: Deisenroth, Faisal, and Ong (2020)

Live Voting

Motivation

Figure: The importance of derivatives

Difference Quotient

- Given $f(x) = y$ with $x, y \in \mathbb{R}$
- Average slope of $f(x)$ between x_0 and $x_0 + \delta x$: line through $f(x_0)$ and $f(x_0 + \delta x)$.
- This line (secant) is a simple linear function
- Its slope is given by the *difference quotient*:

$$\frac{\delta y}{\delta x} = \frac{f(x_0 + \delta x) - f(x_0)}{\delta x}$$

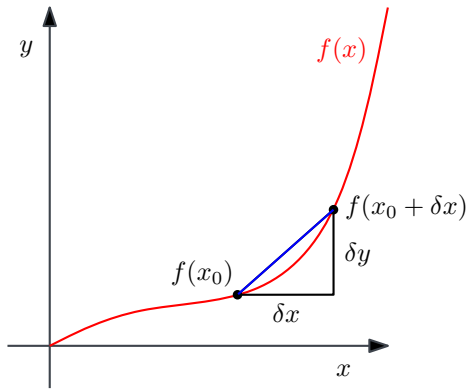

Figure: Deisenroth, Faisal, and Ong (2020)

Derivative

- If δx is infinitely small: the average slope of f between x_0 and $x_0 + \delta x$ is the tangent of f at x_0
- We call this tangent the *derivative* of f at x_0
- It points into the direction (w.r.t. x) of the steepest ascent of f

Derivative

More formally, for $h > 0$ the *derivative* of f is defined as

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

Derivative: Example

$$f(x) = x^2$$

$$\frac{df}{dx} = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h} = ?$$

Derivative: Example

$$f(x) = x^2$$

$$\begin{aligned}
\frac{df}{dx} &= \lim_{h \to 0} \frac{f(x+h) - f(x)}{h} \\
&= \lim_{h \to 0} \frac{(x+h)^2 - x^2}{h} = \lim_{h \to 0} \frac{x^2 + 2xh + h^2 - x^2}{h} \\
&= \lim_{h \to 0} \frac{2xh + h^2}{h} = \lim_{h \to 0} 2x + h = 2x
\end{aligned}$$

Live Voting

Differentiation Rules

- $\frac{df(x)}{dx} = \frac{df}{dx} = \frac{d}{dx}f(x) = \frac{d}{dx}f = f'(x) = f(x)' = f'$
- $\frac{dg}{dx} = 0$ if g does not depend on x, e.g. a constant 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: $(g(f(x)))' = g'(f(x))f'(x)$

Common Derivatives

For $x \in \mathbb{R}, n \in \mathbb{N}$

- $\frac{d}{dx} x^n = n x^{n-1}$
- $\frac{d}{dx} \frac{1}{x^n} = \frac{d}{dx} x^{-n} = -n x^{-n-1} = -\frac{n}{x^{n+1}}$
- $\frac{d}{dx} e^x = e^x$
- $\frac{d}{dx} \ln x = \frac{1}{x}$
- $\frac{d}{dx} \sin x = \cos x$
- $\frac{d}{dx} \cos x = -\sin x$
- $\frac{d}{dx} \tan x = \frac{1}{\cos^2 x}$

Derivative – Example: Sum Rule

$$\frac{d}{dx} 2x = ?$$

Derivative – Example: Sum Rule

$$\frac{d}{dx}2x = \frac{d}{dx}(x+x) = \left(\frac{d}{dx}x\right) + \left(\frac{d}{dx}x\right) = 1 + 1 = 2$$

Derivative – Example: Product Rule

$$\frac{d}{dx}x^2 = ?$$

Derivative – Example: Product Rule

$$\begin{aligned}\frac{d}{dx}x^2 &= \frac{d}{dx}(x \cdot x) = \left(\frac{d}{dx}x\right) \cdot x + x \cdot \left(\frac{d}{dx}x\right) \\ &= 1 \cdot x + x \cdot 1 = x + x = 2x\end{aligned}$$

Derivative – Example: Quotient Rule

$$\frac{d}{dx}\frac{x^2}{2x} = ?$$

Derivative – Example: Quotient Rule

$$\begin{aligned}\frac{d}{dx}\frac{x^2}{2x} &= \frac{\left(\frac{d}{dx}x^2\right)\cdot 2x - x^2\cdot\left(\frac{d}{dx}2x\right)}{(2x)^2} \\ &= \frac{2x\cdot 2x - x^2\cdot 2}{4x^2} = \frac{4x^2-2x^2}{4x^2} \\ &= \frac{2x^2}{4x^2} = \frac{1}{2}\end{aligned}$$

Derivative – Example: Chain Rule

$$\frac{d}{dx}(2 + x^2)^2 = ?$$

Derivative – Example: Chain Rule

$$\begin{aligned}\frac{d}{dx}(2+x^2)^2 &= \frac{d}{dx}g(f(x)) \text{ where } g(f(x)) = f(x)^2, f(x) = 2+x^2 \\ &= \frac{d}{df(x)}g(f(x)) \cdot \frac{d}{dx}f(x) = 2f(x) \cdot 2x \\ &= 2(2+x^2) \cdot 2x = 8x + 4x^3\end{aligned}$$

Live Voting

Partial Derivative

2

Partial derivative

For $f : \mathbb{R}^n \to \mathbb{R}$ with n arguments $x_1, .., x_n$, its *partial* derivative is the standard derivative w.r.t. only one of its arguments:

Partial Derivative

$$\frac{\partial}{\partial x_i} f(x_1, .., x_n) = \lim_{h \to 0} \frac{f(x_1, .., x_i + h, .., x_n) - f(x)}{h}$$

We collect them in a row vector and call it the *gradient* of f:

Gradient

$$\nabla_x f = \text{grad} f = \frac{df}{dx} = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} & \frac{\partial f(x)}{\partial x_2} & \cdots & \frac{\partial f(x)}{\partial x_n} \end{bmatrix} \in \mathbb{R}^{1 \times n}$$

Partial Derivative: Example

$$f(x, g) = 2x + 3g$$

$$\frac{\partial}{\partial x} f(x, g) = ?$$

Partial Derivative: Example

$$f(x, g) = 2x + 3g$$

$$\frac{\partial}{\partial x} f(x, g) = 2$$

Partial Derivative: Example

$$f(x, g) = 2x + 3g$$

$$\frac{\partial}{\partial x} f(x, g) = 2$$

$$\frac{\partial}{\partial g} f(x, g) = ?$$

Partial Derivative: Example

$$f(x, g) = 2x + 3g$$

$$\frac{\partial}{\partial x} f(x, g) = 2$$

$$\frac{\partial}{\partial g} f(x, g) = 3$$

Live Voting

Gradient: Example

$$f(x_1, x_2) = x_1^2 x_2 + x_1 x_2^3$$

$$\frac{df}{dx} = ?$$

Gradient: Example

$$f(x_1, x_2) = x_1^2 x_2 + x_1 x_2^3$$

$$\frac{df}{dx} = \begin{bmatrix} \frac{\partial f(x_1,x_2)}{\partial x_1} & \frac{\partial f(x_1,x_2)}{\partial x_2} \end{bmatrix}$$

Gradient: Example

$$f(x_1, x_2) = x_1^2 x_2 + x_1 x_2^3$$

$$\frac{df}{dx} = \begin{bmatrix} \frac{\partial f(x_1,x_2)}{\partial x_1} & \frac{\partial f(x_1,x_2)}{\partial x_2} \end{bmatrix}$$

$$\frac{\partial f(x_1, x_2)}{\partial x_1} = ?$$

Gradient: Example

$$f(x_1, x_2) = x_1^2 x_2 + x_1 x_2^3$$

$$\frac{df}{dx} = \begin{bmatrix} \frac{\partial f(x_1,x_2)}{\partial x_1} & \frac{\partial f(x_1,x_2)}{\partial x_2} \end{bmatrix}$$

$$\frac{\partial f(x_1, x_2)}{\partial x_1} = 2x_1 x_2 + x_2^3$$

Gradient: Example

$$f(x_1, x_2) = x_1^2 x_2 + x_1 x_2^3$$

$$\frac{df}{dx} = \begin{bmatrix} \frac{\partial f(x_1,x_2)}{\partial x_1} & \frac{\partial f(x_1,x_2)}{\partial x_2} \end{bmatrix}$$

$$\frac{\partial f(x_1, x_2)}{\partial x_1} = 2x_1 x_2 + x_2^3$$

$$\frac{\partial f(x_1, x_2)}{\partial x_2} = ?$$

Gradient: Example

$$f(x_1, x_2) = x_1^2 x_2 + x_1 x_2^3$$

$$\frac{df}{dx} = \begin{bmatrix} \frac{\partial f(x_1,x_2)}{\partial x_1} & \frac{\partial f(x_1,x_2)}{\partial x_2} \end{bmatrix}$$

$$\frac{\partial f(x_1, x_2)}{\partial x_1} = 2x_1 x_2 + x_2^3$$

$$\frac{\partial f(x_1, x_2)}{\partial x_2} = x_1^2 + 3x_1 x_2^2$$

Gradient: Example

$$f(x_1, x_2) = x_1^2 x_2 + x_1 x_2^3$$

$$\frac{df}{dx} = \begin{bmatrix} \frac{\partial f(x_1,x_2)}{\partial x_1} & \frac{\partial f(x_1,x_2)}{\partial x_2} \end{bmatrix} = \begin{bmatrix} 2x_1x_2 + x_2^3 & x_1^2 + 3x_1x_2^2 \end{bmatrix}$$

$$\frac{\partial f(x_1, x_2)}{\partial x_1} = 2x_1x_2 + x_2^3$$

$$\frac{\partial f(x_1, x_2)}{\partial x_2} = x_1^2 + 3x_1x_2^2$$

Live Voting

Gradient of Vector-Valued Functions

3

Vector-Valued Functions

We can generalize the concept of gradients to vector-valued functions, that is, $f : \mathbb{R}^n \to \mathbb{R}^m$, $f(x) = \begin{bmatrix} f_1(x) \\ \vdots \\ f_m(x) \end{bmatrix} \in \mathbb{R}^m$.

Viewing this as a vector of functions $[f_1, \ldots, f_m]^\top$ allows us to apply the same rules for differentiation we just learned:

$$\frac{\partial f}{\partial x_i} = \begin{bmatrix} \frac{\partial f_1}{\partial x_i} \\ \vdots \\ \frac{\partial f_m}{\partial x_i} \end{bmatrix} = \begin{bmatrix} \lim_{h \to 0} \frac{f_1(x_1, \ldots, x_i + h, \ldots, x_n) - f_1(x)}{h} \\ \vdots \\ \lim_{h \to 0} \frac{f_m(x_1, \ldots, x_i + h, \ldots, x_n) - f_m(x)}{h} \end{bmatrix}$$

Jacobian

We know that the gradient of f with respect to a vector expands as the row vector of partial derivatives.
Applying this to our vector-valued function yields the Jacobian J of f:

Jacobian

$$
\begin{aligned}
J = \nabla_x f &= \frac{d}{dx} f(x) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \cdots \frac{\partial f}{\partial x_n} \end{bmatrix} \\
&= \begin{bmatrix}
\frac{\partial}{\partial x_1} f_1(x) & \frac{\partial}{\partial x_2} f_1(x) & \cdots & \frac{\partial}{\partial x_n} f_1(x) \\
\frac{\partial}{\partial x_1} f_2(x) & \frac{\partial}{\partial x_2} f_2(x) & \cdots & \frac{\partial}{\partial x_n} f_2(x) \\
\vdots & \vdots & \ddots & \vdots \\
\frac{\partial}{\partial x_1} f_m(x) & \frac{\partial}{\partial x_2} f_m(x) & \cdots & \frac{\partial}{\partial x_n} f_m(x)
\end{bmatrix}
\end{aligned}
$$

Jacobian: Example

Let $f(x) = \begin{bmatrix} x_1^2 + 5 \\ 2x_1 + x_2 \end{bmatrix}$

Then $J_f = \begin{bmatrix} \frac{\partial}{\partial x_1} f_1(x) & \frac{\partial}{\partial x_2} f_1(x) \\ \frac{\partial}{\partial x_1} f_2(x) & \frac{\partial}{\partial x_2} f_2(x) \end{bmatrix} = \begin{bmatrix} 2x & 0 \\ 2 & 1 \end{bmatrix}$

Live Voting

References I

Deisenroth, Marc Peter, A. Aldo Faisal, and Cheng Soon Ong (2020). *Mathematics for Machine Learning*. Cambridge University Press. DOI: 10.1017/9781108679930.