Vector Calculus

Mark van der Wilk

Department of Computing
Imperial College London

@markvanderwilk
m.vdwilk@imperial.ac.uk

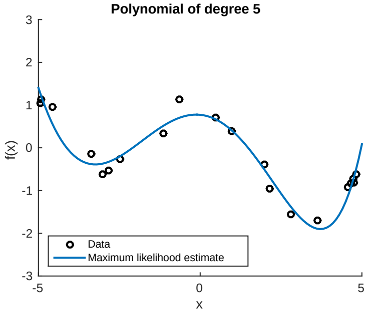
October 12, 2021

Reading Material

Lecture notes, Chapter 5

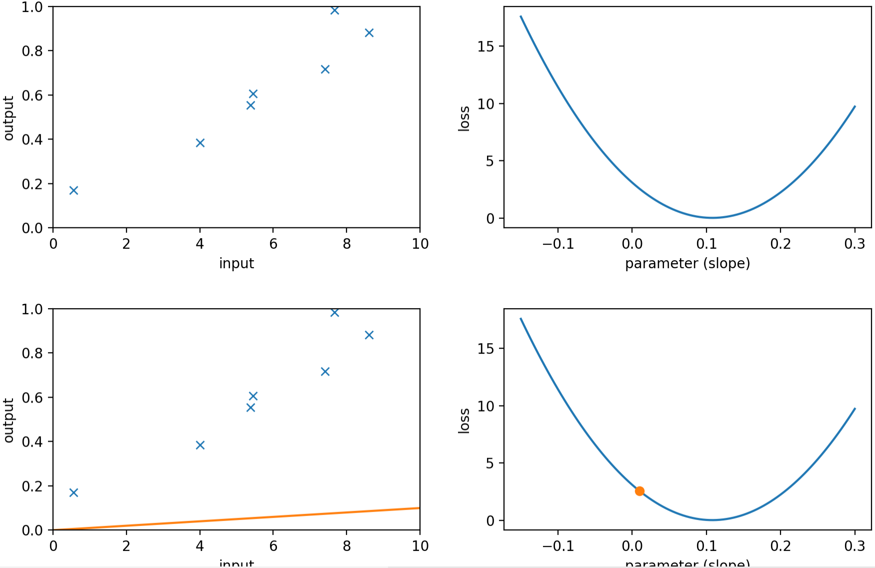
https://mml-book.com

Curve Fitting (Regression) in Machine Learning (2)

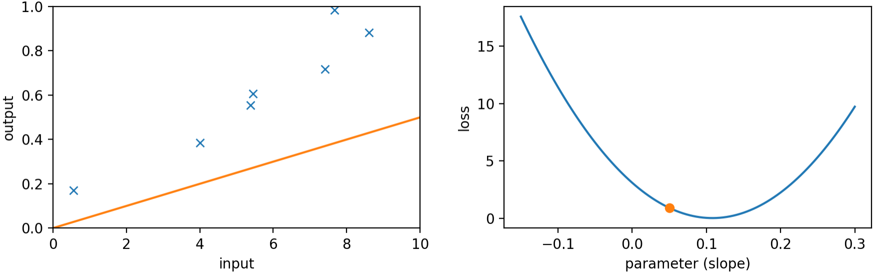


- Training the model means finding parameters $\boldsymbol{\theta}^*$, such that $f(\boldsymbol{x}_i, \boldsymbol{\theta}^*) \approx y_i$
- Define a loss function, e.g., $\sum_{i=1}^{N}(y_i - f(\boldsymbol{x}_i, \boldsymbol{\theta}))^2$, which we want to optimize
- Adjust $\boldsymbol{\theta}$ until loss is as small as we can get it: **Minimisation** / **optimisation**.

Example: Minimising the loss



Example: Minimising the loss



Two questions for now:

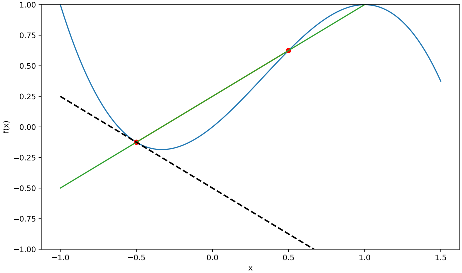
- How should we change a to make the loss smaller?
- How do we know when we can't get better?

Scalar Differentiation $f : \mathbb{R} \to \mathbb{R}$

- Derivative defined as the limit of the difference quotient

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

▶▶ Slope of the secant line through $f(x)$ and $f(x+h)$



Some Examples

$$
\begin{aligned}
f(x) &= x^n & f'(x) &= nx^{n-1} \\
f(x) &= \sin(x) & f'(x) &= \cos(x) \\
f(x) &= \tanh(x) & f'(x) &= 1 - \tanh^2(x) \\
f(x) &= \exp(x) & f'(x) &= \exp(x) \\
f(x) &= \log(x) & f'(x) &= \frac{1}{x}
\end{aligned}
$$

Differentiation Rules

- Sum Rule

$$\big(f(x) + g(x)\big)' = f'(x) + g'(x) = \frac{df(x)}{dx} + \frac{dg(x)}{dx}$$

- Product Rule

$$\big(f(x)g(x)\big)' = f'(x)g(x) + f(x)g'(x) = \frac{df(x)}{dx}g(x) + f(x)\frac{dg(x)}{dx}$$

- Chain Rule

$$(g \circ f)'(x) = \big(g(f(x))\big)' = g'(f(x))f'(x) = \frac{dg(f(x))}{df}\frac{df(x)}{dx}$$

- Quotient Rule

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

Example: Scalar Chain Rule

$$(g \circ f)'(x) = \big(g(f(x))\big)' = g'(f(x))f'(x) = \frac{dg}{df}\frac{df}{dx}$$

Beginner

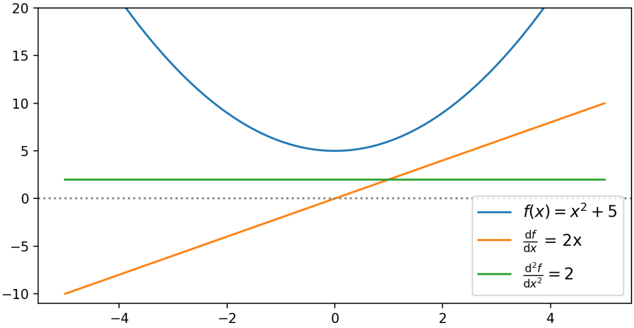
$$\begin{aligned} g(z) &= 6z + 3 \\ z = f(x) &= -2x + 5 \\ (g \circ f)'(x) &= \underbrace{(6)}_{dg/df}\underbrace{(-2)}_{df/dx} \\ &= -12 \end{aligned}$$

Advanced

$$\begin{aligned} g(z) &= \tanh(z) \\ z = f(x) &= x^n \\ (g \circ f)'(x) &= \underbrace{(1 - \tanh^2(x^n))}_{dg/df}\underbrace{nx^{n-1}}_{df/dx} \end{aligned}$$

Work it out with your neighbors

Finding minima

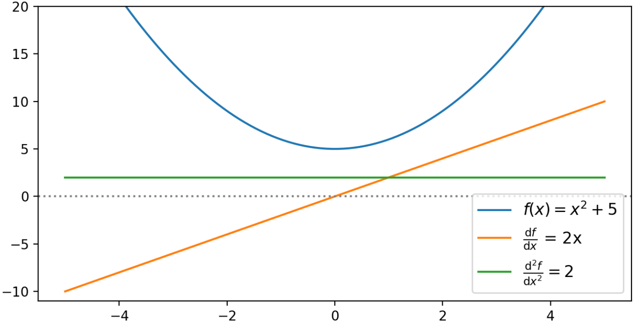


Q1: How should we change the input to reduce the output?

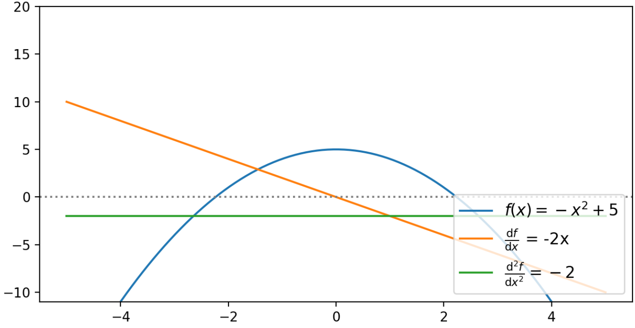
- Find the derivative function and compute it at a point to find the point's gradient.
- Increase for negative gradients. Decrease for positive gradients.

This is the idea behind **gradient descent**.

Finding minima I



Finding minima II



- At a minimum, there is no change we can make that lowers our function value $\implies$ gradient must be zero.
- Zero gradient is not enough!

Finding minima III

- For minimum, $f(x)$ must go from decreasing to increasing
  $\implies$ gradient of gradient positive

Local and global minima

Board.

Example: Linear regression

For the example from earlier, find optimal a:

$$f(x) = a \cdot x \qquad L(a) = \sum_{n=1}^{N} (f(x_n) - y_n)^2 \tag{2}$$

$$\frac{\mathrm{d}L}{\mathrm{d}a} = \sum_{n=1}^{N} 2(ax_n - y_n)x_n = \sum_{n=1}^{N} 2ax_n^2 - 2x_ny_n = 0 \tag{3}$$

$$2a \sum_n x_n^2 = \sum_n 2x_ny_n \tag{4}$$

$$a = \frac{\sum_n x_ny_n}{\sum_n x_n^2} \tag{5}$$

$$\frac{\mathrm{d}^2L}{\mathrm{d}a^2} = \sum_{n=1}^{N} 2x_n^2 \geqslant 0 \tag{6}$$

Summary

You have seen:

- That derivatives are useful for finding minima of functions
- How to differentiate simple functions
- An example of solving for the minimum point
- How to identify minima

Linear regression: multiple parameters

What happens when our function has multiple parameters?

$$f(x) = \theta_3 x^3 + \theta_2 x^2 + \theta_1 x + \theta_0 \tag{7}$$

Think of a **vector** as parameterising our function:

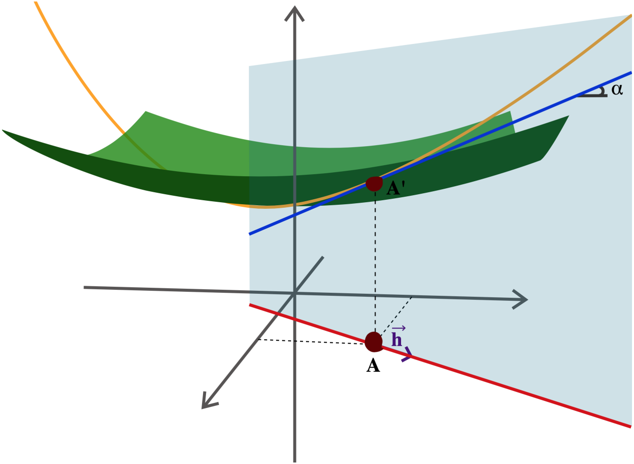
$$f(x) = \boldsymbol{\theta}^\mathsf{T} \boldsymbol{\phi}(x) \qquad\qquad \boldsymbol{\phi}(x) = \begin{bmatrix} x^3 & x^2 & x & 1 \end{bmatrix}^\mathsf{T} \tag{8}$$

We want to:

- Understand how a function (e.g. loss) changes when we change $\boldsymbol{\theta}$.
- Characterise what an optimum is for a function of a vector.

Both can be analysed by
turning the multi-D problem into many 1D problems.

Directional derivative



How does the function change if we move in a particular *direction*?

Directional derivative

Define **directional derivative** $\nabla_{\mathbf{v}} L(\boldsymbol{\theta})$ as how much the function changes if we move in direction $\mathbf{v}$:

$$\nabla_{\mathbf{v}} L(\boldsymbol{\theta}) = \lim_{h \to 0} \frac{L(\boldsymbol{\theta} + h\mathbf{v}) - L(\boldsymbol{\theta})}{h}$$

$$\nabla_{\mathbf{v}} L(\boldsymbol{\theta}) = \lim_{h \to 0} \frac{L(\theta_1 + hv_1, \theta_2 + hv_2) - L(\theta_1, \theta_2)}{h}$$

$$= \lim_{h \to 0} \frac{L(\theta_1 + hv_1, \theta_2 + hv_2) - L(\theta_1, \theta_2 + hv_2)}{h} + \frac{L(\theta_1, \theta_2 + hv_2) - L(\theta_1, \theta_2)}{h}$$

$$= \lim_{h \to 0} \frac{L(\theta_1 + h', \theta_2 + h'\frac{v_2}{v_1}) - L(\theta_1, \theta_2 + h'\frac{v_2}{v_1})}{h'/v_1} + \frac{L(\theta_1, \theta_2 + h'') - L(\theta_1, \theta_2)}{h''/v_2}$$

$$= \frac{\partial L}{\partial \theta_1} v_1 + \frac{\partial L}{\partial \theta_2} v_2$$

- Can find gradient in **any** direction with the **partial derivatives**

Multivariate Differentiation $f : \mathbb{R}^N \to \mathbb{R}$

$$y = f(\boldsymbol{x}), \quad \boldsymbol{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix} \in \mathbb{R}^N$$

- Partial derivative (change one coordinate at a time):

$$\frac{\partial f}{\partial x_i} = \lim_{h \to 0} \frac{f(x_1, \ldots, x_{i-1}, x_i + h, x_{i+1}, \ldots, x_N) - f(\boldsymbol{x})}{h}$$

- Jacobian vector (gradient) collects all partial derivatives:

$$\frac{\mathrm{d}f}{\mathrm{d}\boldsymbol{x}} = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \cdots & \frac{\partial f}{\partial x_N} \end{bmatrix} \in \mathbb{R}^{1 \times N}$$

Note: By convention, we define this to be a **row vector**.

Multivariate Differentiation $f : \mathbb{R}^N \to \mathbb{R}$

Derivative w.r.t. vector

Since we can find the directional derivative *in any direction* with the Jacobian, we **define** this vector to be the derivative of a function w.r.t. a vector.

Steepest descent direction

Directional derivative:

$$\nabla_{\boldsymbol{v}} f(\boldsymbol{\theta}) = \frac{\mathrm{d}f}{\mathrm{d}\boldsymbol{\theta}}\boldsymbol{v} \tag{9}$$

(inner product, row vector times column vector)

What is the direction where the function changes the most?

$$\frac{\mathrm{d}f}{\mathrm{d}\boldsymbol{\theta}}\boldsymbol{v} = |\frac{\mathrm{d}f}{\mathrm{d}\boldsymbol{\theta}}||\boldsymbol{v}| \cos\beta \tag{10}$$

- Choose unit vector $\boldsymbol{v}$
- Angle between vectors β should be zero $\implies \cos\beta = 1$.

Steepest descent points in direction of Jacobian/gradient vector.

Example: Multivariate Differentiation

Beginner

$$f : \mathbb{R}^2 \to \mathbb{R}$$
$$f(x_1, x_2) = x_1^2 x_2 + x_1 x_2^3 \in \mathbb{R}$$

Advanced

$$f : \mathbb{R}^2 \to \mathbb{R}$$
$$f(x_1, x_2) = (x_1 + 2x_2^3)^2 \in \mathbb{R}$$

Partial derivatives? Gradient?

Work it out with your neighbors

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

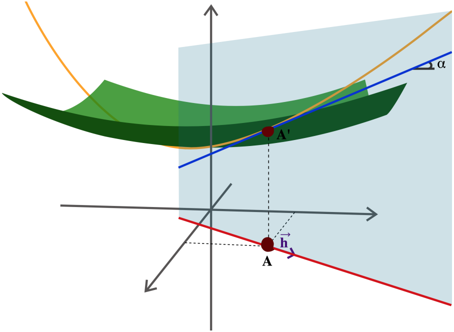
$$\frac{\partial f(x_1, x_2)}{\partial x_1} = 2(x_1 + 2x_2^3) \overbrace{(1)}^{\frac{\partial}{\partial x_1}(x_1 + 2x_2^3)}$$
$$\frac{\partial f(x_1, x_2)}{\partial x_2} = 2(x_1 + 2x_2^3) \underbrace{(6x_2^2)}_{\frac{\partial}{\partial x_2}(x_1 + 2x_2^3)}$$

Gradient $\frac{\mathrm{d}f}{\mathrm{d}\boldsymbol{x}} = \left[\frac{\partial f(x_1,x_2)}{\partial x_1} \quad \frac{\partial f(x_1,x_2)}{\partial x_2}\right] \in \mathbb{R}^{1\times 2}$

$$\frac{\mathrm{d}f}{\mathrm{d}\boldsymbol{x}} = \left[2x_1 x_2 + x_2^3 \quad x_1^2 + 3x_1 x_2^2\right]$$
$$\frac{\mathrm{d}f}{\mathrm{d}\boldsymbol{x}} = \left[2(x_1 + 2x_2^3) \quad 12(x_1 + 2x_2^3)x_2^2\right]$$

Optima, minima, maxima

What is an optimum for a function of a vector?



- Directional derivative should be zero *in all directions* $\implies \frac{\mathrm{d}f}{\mathrm{d}x} = \mathbf{0}$.
- For minimum: second directional derivative should be positive *in all directions*.

Summary

Motivation: Want to optimise functions of several variables

- Directional derivative
- Partial derivatives $\implies$ gradient vector
- Steepest descent direction
- At an optimum $\frac{\mathrm{d}f}{\mathrm{d}\boldsymbol{x}} = \mathbf{0}$

Next time: Derivatives of vectors and chain rules.