Foundations of Machine Learning in Python

Moritz Wolter

July 12, 2022

High Performance Computing and Analytics Lab, Uni Bonn

Overview

Derivatives and Gradients

Optimization

Optimization

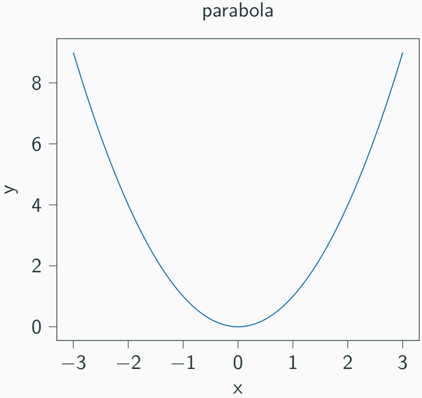
Traditionally, optimization means minimizing using a cost function $f(x)$. Given the cost, we must find the cheapest point x^* on the function, or in other words,

$$x^* = \min_{x \in \mathbb{R}} f(x) \tag{1}$$

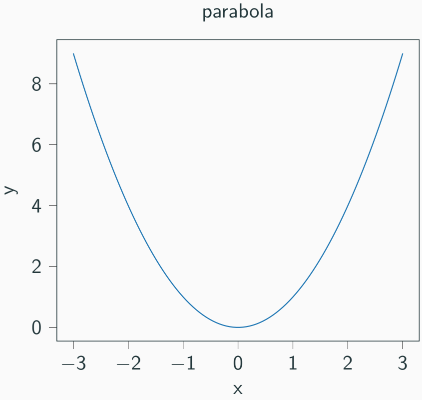
Functions

Functions are mathematical mappings. Consider for example the quadratic funtion, $f(x) : \mathbb{R} \to \mathbb{R}$:

$$f(x) = x^2 \tag{2}$$



Where is the minimum?



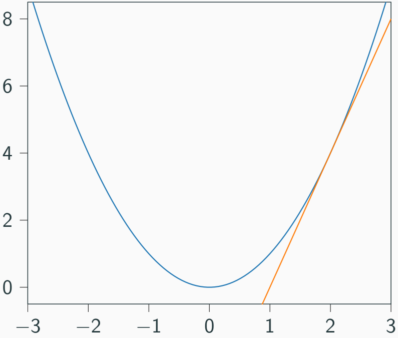
In this case we immediately see it's at zero. Finding it algorithmically requires derivate information.

Derivatives and Gradients

The derivative

$$\frac{df(x)}{dx} = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h} \tag{3}$$

parabola with derivative at two



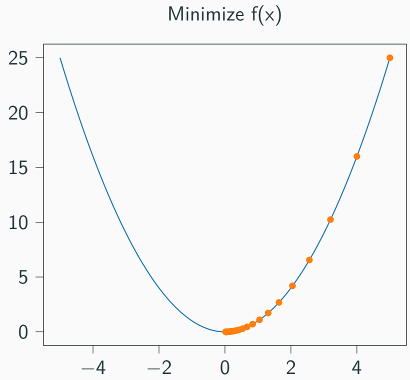
Derivation of the parabola derivative

$$
\begin{aligned}
\lim_{h\to 0}\frac{(x+h)^2-x^2}{h} &\Leftrightarrow \lim_{h\to 0}\frac{x^2+2xh+h^2-x^2}{h} &\quad (4)\\
&\Leftrightarrow \lim_{h\to 0}\frac{2xh+h^2}{h} &\quad (5)\\
&\Leftrightarrow \lim_{h\to 0}\frac{h(2x+h)}{h} &\quad (6)\\
&\Leftrightarrow \lim_{h\to 0} 2x+h &\quad (7)\\
&\Leftrightarrow 2x &\quad (8)
\end{aligned}
$$

Steepest descent on the parabola

$$x_n = x_{n-1} - \alpha \cdot \frac{df}{dx} \tag{9}$$

Working with $x_0 = 5$ and $\alpha = 0.1$ for 25 steps leads to:



Multidimensional problems

$$f(x, y) = (a - x)^2 + b(y - x^2)^2 \tag{10}$$

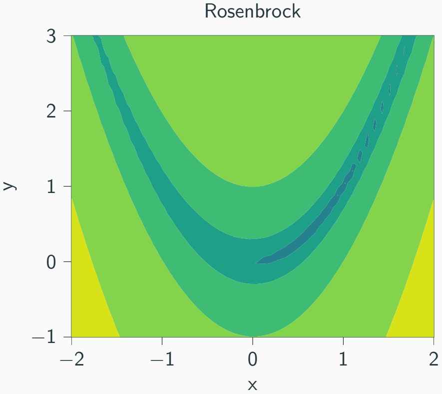


Figure: Rosenbrock function with a=1 and b=100 .

The gradient

$$\nabla f = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{pmatrix} \tag{11}$$

Rosenbrock gradient

$$\nabla f(x, y) = \begin{pmatrix} -2a + 2x - 4byx + 4bx^3 \\ 2by - 2bx^2 \end{pmatrix} \tag{12}$$

Gradient descent

$x_0 = [0.1, 3.]$, $\alpha = 0.01$

$$x_n = x_{n-1} - \alpha \cdot \nabla f(x) \tag{13}$$

Gradient descent on the Rosenbrock function

Rosenbrock Optimization

Gradient descent with momentum

Rosenbrock Optimization

Optimization

Second order optimization

TODO