UNIVERSITY OF LEEDS

Class: Machine Learning

Elements of Local Optimisation

Instructor: Matteo Leonetti

Learning outcomes

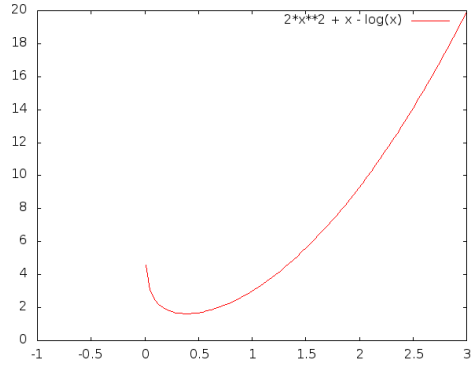
UNIVERSITY OF LEEDS

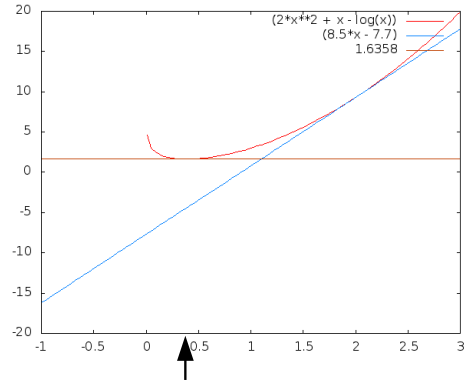
- Describe the difference between zero, first, and second-order optimisation methods.
- Apply gradient descent to a given objective function.
- Choose an appropriate step size for gradient descent.

Goal

UNIVERSITY OF LEEDS

Find the minimum point of a given function:





The minimum is at 0.39

Optimisation is the field that studies how to find the minimum (or maximum) of a function.

Stationary points (minimum, maximum, and saddle points) have an important property: the derivative of the function in those points is always 0.

Local Methods

UNIVERSITY OF LEEDS

Local methods move from the current point along a given direction by a certain step, until a local minimum is found. Different methods determine the direction and the step differently.

There are in general infinitely many directions, some will improve the current point (go down) others will make the current solution worse (of higher value, by going up).

The best possible direction is the direction of steepest descent, which is the anti-gradient (the gradient is the direction of steepest ascent).

Gradient descent

UNIVERSITY OF LEEDS

First order: gradient descent

$$x_{t+1} = x_t - \eta \nabla f(x_t)$$

step parameter

Second order: Newton's method

$$f(x_n + \Delta x) \approx f(x_n) + f'(x_n)\Delta x + \frac{1}{2} f''(x_n)\Delta x^2$$

Taylor's expansion

$$\frac{\partial}{\partial \Delta x} f(x_n + \Delta x) = f'(x_n) + f''(x_n)\Delta x = 0$$

Optimal step

$$\Delta x = \frac{-f'(x_n)}{f''(x_n)}$$

Many dimensions: $x_{t+1} = x_t - H^{-1}|_{x_n} \nabla f$

Question

UNIVERSITY OF LEEDS

The current point is <1,0>, compute the next point following gradient descent on the function $f(x,y) = x^3 + 2y^2 - y$ with step size 0.1.

Question

UNIVERSITY OF LEEDS

We want to compute: $x_{t+1} = \langle 1,0 \rangle - 0.1 \nabla f(x_t)$

$\nabla f = \langle 3x^2, 4y - 1 \rangle$ Evaluated in <1,0> is <3,-1>

$$x_{t+1} = \langle 1,0 \rangle - 0.1 \cdot \langle 3, -1 \rangle = \langle 0.7, 0.1 \rangle$$

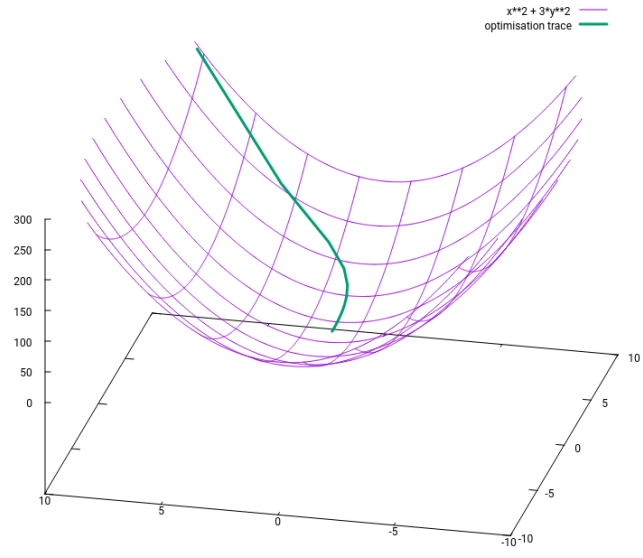
$f(1,0) = 1$

Our solution has improved!

$f(0.7, 0.1) = 0.263$

In 3D

UNIVERSITY OF LEEDS



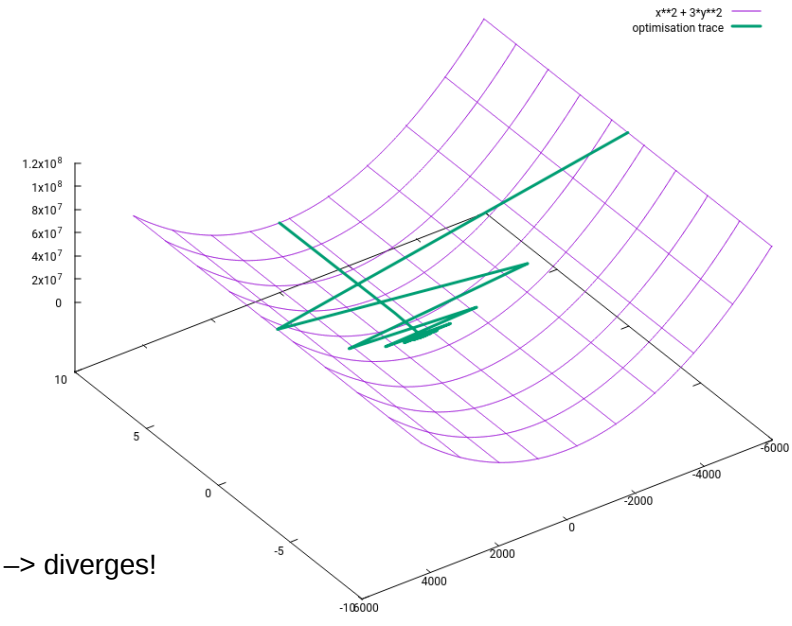
Step size: 0.1

The step size determines how much the algorithm will move the point along the gradient. In first-order methods it is usually chosen as a small constant <= 0.1.

A smaller step makes the algorithm slower, but a step that's too large will make it bounce between solutions indefinitely, or even overshoot (example in next slide).

In 3D

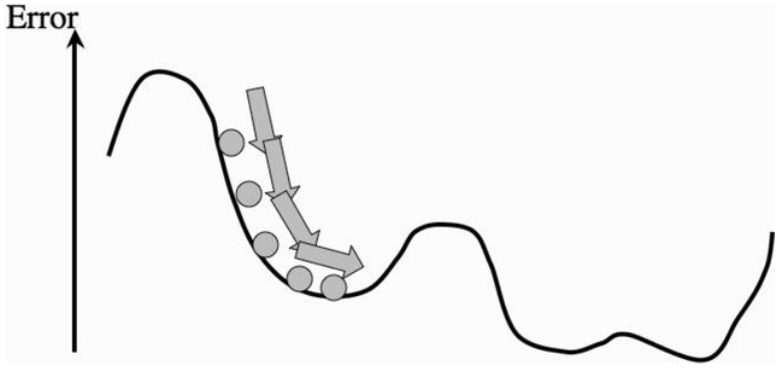
UNIVERSITY OF LEEDS



Step size: 0.14 –> diverges!

Local Minima

UNIVERSITY OF LEEDS



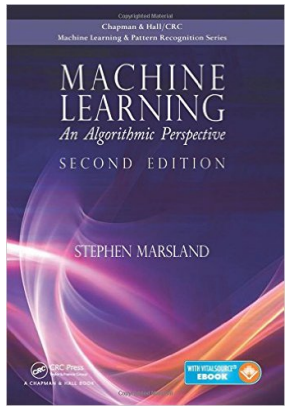
Local methods stop when the gradient is close to zero, which means that they are close to a stationary point.

There is no guarantee that such a point is the *global* minimum. Local methods will, in general, converge to a *local* minimum of the objective function.

A local minimum is a point such that all the points around it have a higher value of the objective function.

Conclusion

UNIVERSITY OF LEEDS



Sections 9.0, 9.1