EEE319 Optimisation
Lecture 9 Nonlinear Programming (3)

Prof. Xinheng Wang

xinheng.wang@xjtlu.edu.cn

Office: EE512

Nonlinear programming

- Last week
  - Nonlinear programming
    - Multiple variables
      - Newton Method

- Today
  - Nonlinear programming
    - Gradient descent

Gradient descent

- Recall from lecture 2
- Gradient descent minimises a function by iteratively moving a little bit in the direction of negative gradient, where negative gradient is a vector pointing at the greatest decrease of a function. It is represented mathematically as:

$$x_{i+1} = x_i - \alpha \nabla f(x_i)$$

where α is called learning rate, which is normally a constant, and ∇f is the derivative, specifically it is the gradient of function f.

- Question: How gradient descent equation is derived?

Gradient descent

- The underlying technique is still Taylor series:
- Recall Taylor series: A Taylor series is a representation of a function as an infinite sum of terms that are calculated from the values of the function's derivatives at a single point. It is expressed as:

$$f(x) = f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \frac{f'''(a)}{3!}(x-a)^3 + \cdots$$

where *a* is a real or complex number.

- This can be written in a sigma notation

$$f(x) = \sum_{i=0}^{\infty} \frac{f^{(i)}(a)}{i!}(x-a)^i$$

where $f^{(i)}(a)$ denotes the *i*th derivative of evaluated at the point *a*, where *i*! denotes the factorial of *n*.

Gradient descent

- Because variable a is confusing with the step length α used in gradient descent, let us rewrite the Taylor series as:

$$f(\theta) = f(\theta_0) + \frac{f'\theta_0}{1!}(\theta - \theta_0) + \frac{f''(\theta_0)}{2!}(\theta - \theta_0)^2 + \frac{f'''(\theta_0)}{3!}(\theta - \theta_0)^3 + \cdots$$

- If only the first two terms are considered, Taylor series could be approximately as:

$$f(\theta) \approx f(\theta_0) + (\theta - \theta_0)f'(\theta_0)$$

- $f'(\theta_0)$ could be replaced with $\nabla f(\theta_0)$, which is the first-order derivative, then the equation will become

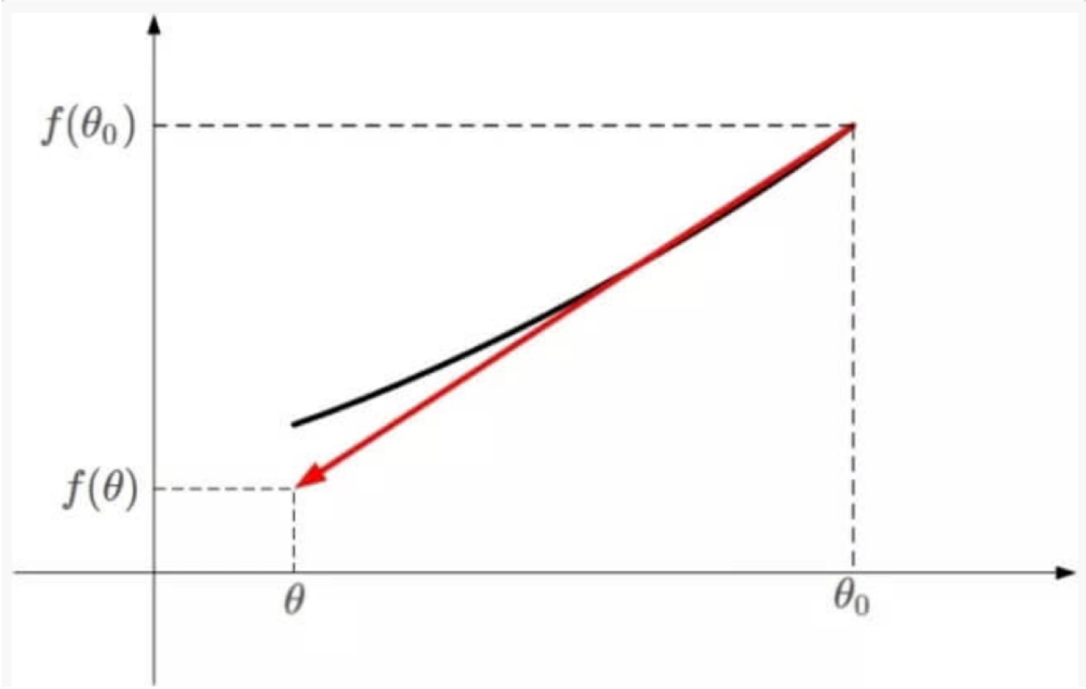
$$f(\theta) \approx f(\theta_0) + (\theta - \theta_0)\nabla f(\theta_0)$$

Gradient descent

- Let us illustrate the derivative in the figure on the right. In this figure, black line is the curve, red line is the derivative at point θ_0, $f(\theta)$ and $f(\theta_0)$ are the values at points θ and θ_0, respectively and they are very close. In another word, $\theta - \theta_0$ is a small value. Therefore,

$$\nabla f\theta_0 = \frac{f(\theta) - f(\theta_0)}{\theta - \theta_0}$$

$$f(\theta) = f(\theta_0) + (\theta - \theta_0)\nabla f(\theta_0)$$



Same equation can be determined.

Gradient descent

- Since $\theta - \theta_0$ is a small value, we use αv to represent it, where α is the learning rate/step length as talked earlier and v is the unit vector of $\theta - \theta_0$.

$$\theta - \theta_0 = \alpha v$$

- A new equation is derived as

$$f(\theta) = f(\theta_0) + \alpha v \nabla f(\theta_0)$$

- In order to minimise a function, we expect the value from the current one is always smaller than previous one. That is

$$f(\theta) < f(\theta_0)$$

- In another word,

$$f(\theta) - f(\theta_0) < 0$$

- Subsequently

$$\alpha v \nabla f(\theta_0) < 0$$

Gradient descent

- Let us have a look at

$$\alpha v \nabla f(\theta_0) < 0$$

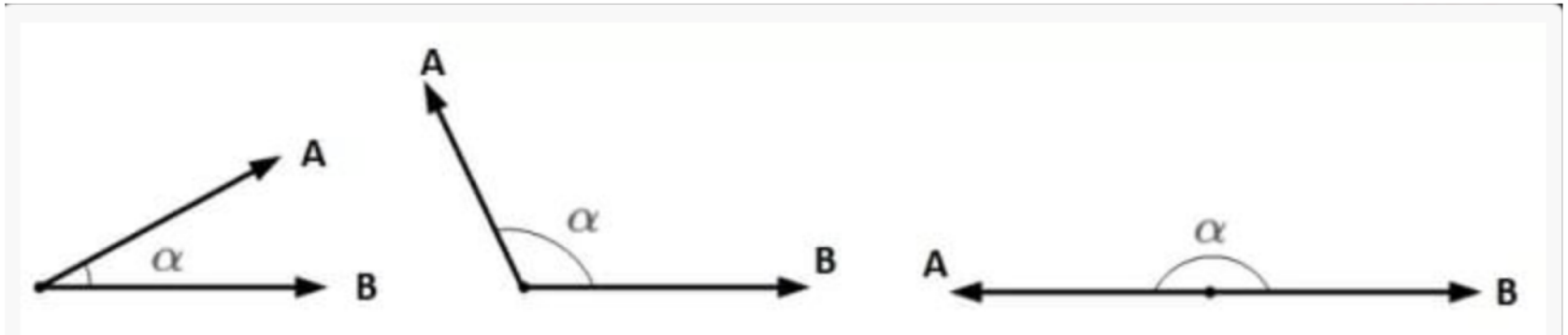
- Since α is a scaler, this can be ignored, the above question becomes

$$v \nabla f(\theta_0) < 0$$

where both of v and $\nabla f(\theta_0)$ are vectors.

Gradient descent

- Let us have a quick at the multiplication of two vectors, as illustrated in the following figure



$$\mathbf{A} \cdot \mathbf{B} = \|\mathbf{A}\|\|\mathbf{B}\| cos\alpha$$

- Only when **A** and **B** are on opposite positions, where $cos\alpha = -1$, it makes $\mathbf{A} \cdot \mathbf{B}$ a minimum negative value.

Gradient descent

- In order to make $v\nabla f\theta_0 < 0$ hold, we just enable $\boldsymbol{v}$ the opposite of $\nabla f\theta_0$. Therefore

$$v = -\frac{\nabla f(\theta_0)}{\|\nabla f\theta_0\|}$$

- The reason to divide by the magnitude of $\nabla f\theta_0$ is only because $\boldsymbol{v}$ is a unit vector.
- Recall $\theta - \theta_0 = \alpha v$
- Therefore

$$\theta = \theta_0 - \alpha\frac{\nabla f(\theta_0)}{\|\nabla f(\theta_0)\|}$$

- Since $\|\nabla f\theta_0\|$ is a scalar, we merge it into α. A new equation is obtained as:

$$\theta = \theta_0 - \alpha\nabla f(\theta_0)$$

This is the equation of gradient descent.

Gradient descent

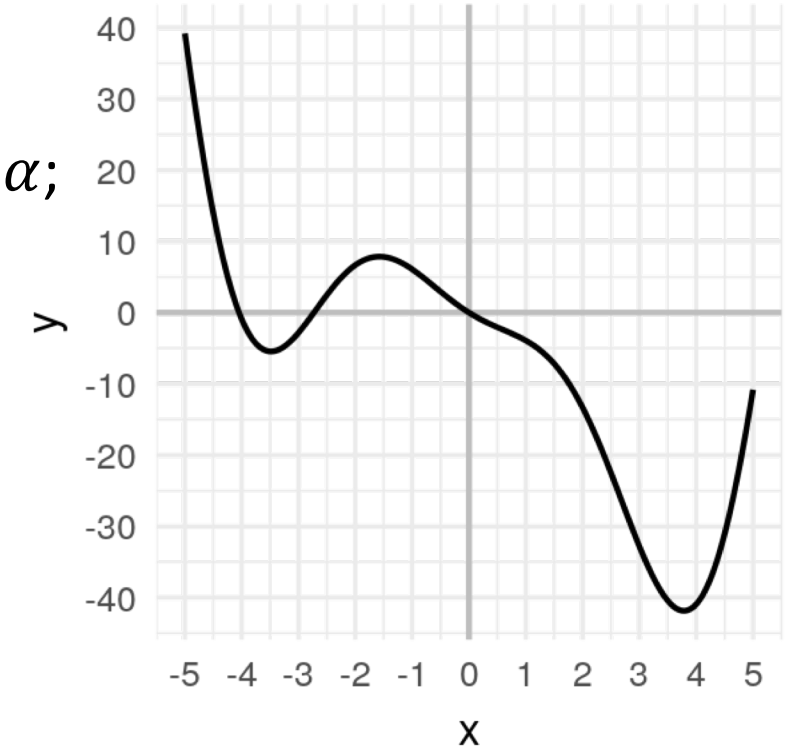
- A few remarks
  (1) Application of gradient descent is similar to Newton Method;
  (2) How to select the value of learning rate α is a concern.

- Let us illustrate this by an example [1]
- Find the minimum value of function $f(x) = 2x^2 \cos(x) - 5x$ at interval [-5, 5]

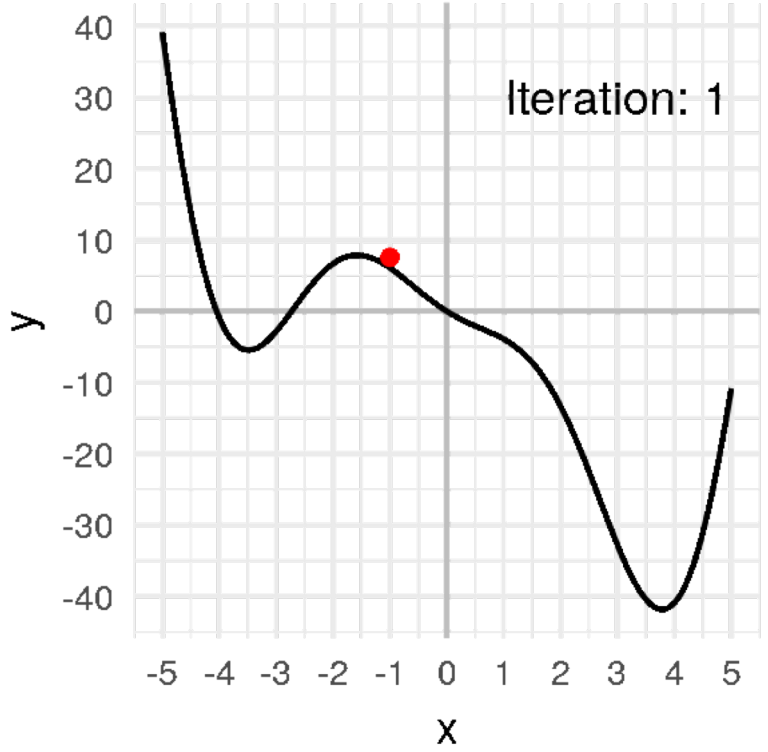
1. https://www.charlesbordet.com/en/gradient-descent/#

Gradient descent

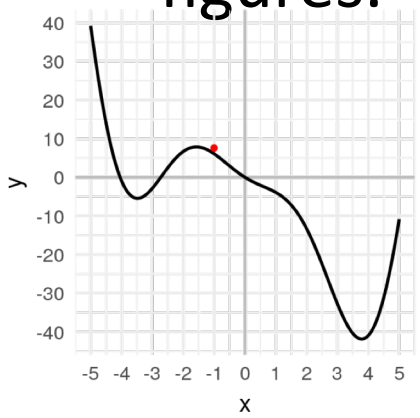
- A few steps
  - Starting with an initial guess of x_0;
  - Compute the derivative f' at x_0;
  - Compute $x_1 = x_0 - \alpha f'$ with a selected value of α;
  - Iterate this process until it converges.

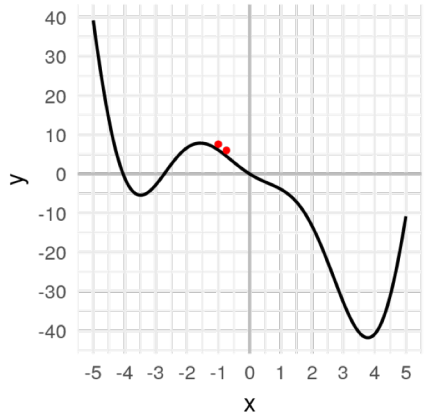


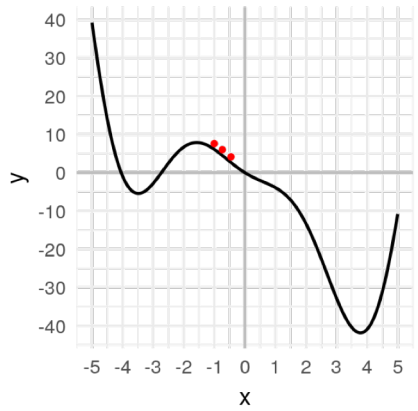
Gradient descent



- Initial value $x_0 = -1$;
- $f(x_0)$=6.08
- $f'(x_0)$=-5.23913
- Select $\alpha = 0.05$
- Compute $x_1 = x_0 - \alpha f'$ and repeat the iterations, illustrated in the figures.







Gradient descent

- How about a different value of α, i.e., $\alpha = 0.001$?

Gradient descent

- How about a different value of α, i.e., $\alpha = 0.2$?

Gradient descent

- The way to select a step length
  - Fixed step size
    - As demonstrated in the example where a fixed value is chosen
  - Backtracking line search
  - Exact line search

Gradient descent

- The way to select a step length
  - Backtracking line search
    - Fix a parameter $0 < \beta < 1$
    - At each iteration, an initial step size α is selected, then following inequality is evaluated

$$f(x_i - \alpha \nabla f(x_i)) \leq f(x_i) - \frac{\alpha}{2} \|\nabla f(x_i)\|^2$$

    - If this inequality is verified, then the current step size is kept. If not, the step size is updated as $\beta\alpha$.

Gradient descent

- The way to select a step length
  - Exact line search
    - This is an optimal way to find the step length, where α is chosen to minimise the function $x - \alpha \nabla f(x)$, though it is rarely used.

THANK YOU

VISIT US

WWW.XJTLU.EDU.CN

FOLLOW US

@XJTLU

Xi'an Jiaotong-Liverpool University
西交利物浦大學