Pattern Recognition

Optimization methods

Krystian Mikolajczyk

Blackboard

Optimization methods

Find x that is a minimum of $f(x)$, $\quad argmin_x f$

$$\frac{df(x)}{dx} = 0$$

- Gradient descent
  - Follow the negative gradient towards local minimum
- Newton-Raphson
  - Linearize the function in a point, get the roots of the linearized function, repeat for the root

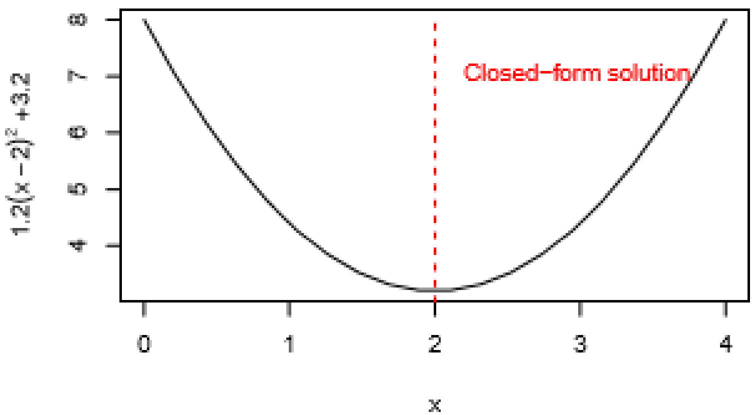
Optimization methods

Find x that is a minimum of $f(x)$, $\text{argmin}_x f$

$$f'(x)=0$$

- Take the derivative, set equal to zero and solve for x

$$f(x) = 1.2(x-2)^2 + 3.2$$
$$\frac{df(x)}{dx} = 1.2(2)(x-2) = 2.4(x-2)$$
$$\frac{df(x)}{dx} = 0 = 2.4(x-2)$$
$$x = 2$$

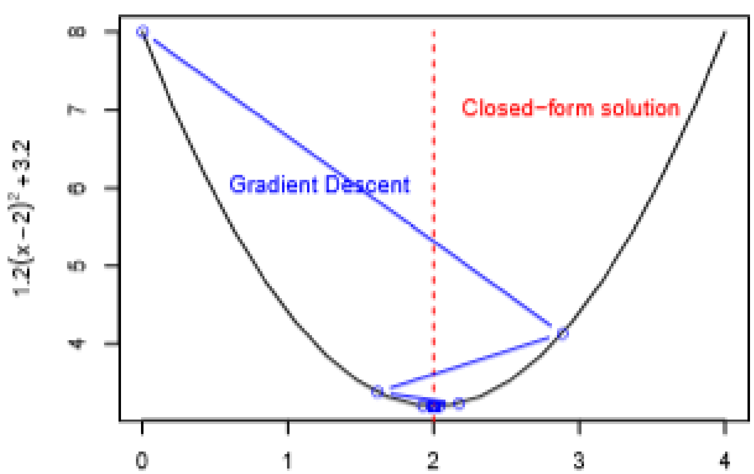


If derivative cannot be directly solved?

Gradient descent

- If derivative cannot be directly solved?
- Start with some *x*, use derivative at that value to find the direction to update *x*
- A constant controlling how far we move, eg. $\gamma = 0.6$

$$f(x) = 1.2(x-2)^2 + 3.2$$
$$\frac{df(x)}{dx} = 2.4(x-2)$$
$$x(0) = 0 \text{ (for example)}$$
$$x(n) = x(n-1) - \gamma 2.4(x-2)$$

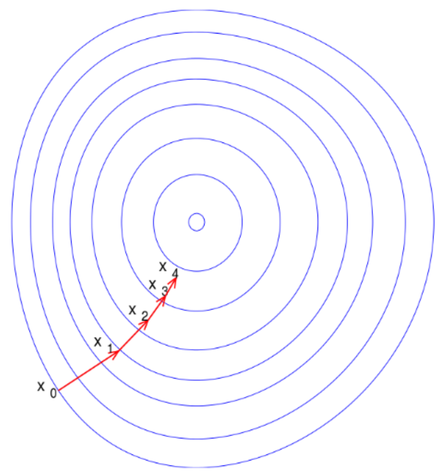


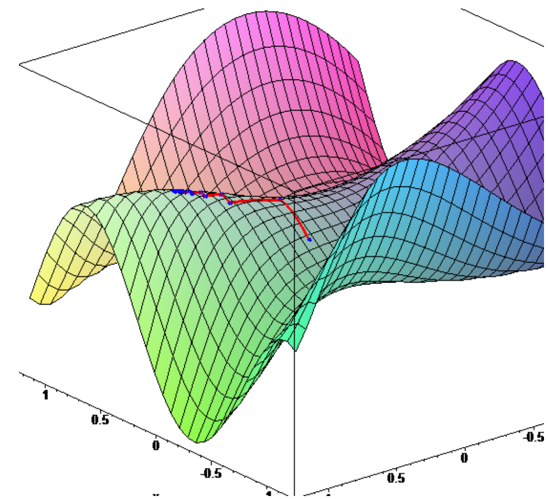
Gradient descent

- Non-convex multi-dimensional optimisation problem
  - Typical in practice

$$\mathbf{x}_{n+1} = \mathbf{x}_n - \gamma_n \nabla F(\mathbf{x}_n)$$

$$F(\mathbf{x}_0) \geq F(\mathbf{x}_1) \geq F(\mathbf{x}_2) \geq \cdots,$$





γ must be set somehow but should be small enough

Newton-Raphson

- In general - iterative method for finding the roots of a differentiable function *f*
  - solutions to the equation *f(x)=0*
- In optimization – for finding local minimum of a twice differentiable function *f*
  - solutions to the equation $argmin_x f(x)$
  - solutions to $f'(x)=0$, stationary point of *f*
  - Taylor expansion $f_T(x) = f_T(x_n + \Delta x) \approx f(x_n) + f'(x_n)\Delta x + \frac{1}{2} f''(x_n)\Delta x^2$

$$0 = \frac{\mathrm{d}}{\mathrm{d}\Delta \mathrm{x}} \left(f(x_n) + f'(x_n)\Delta x + \frac{1}{2} f''(x_n)\Delta x^2 \right) = f'(x_n) + f''(x_n)\Delta x \qquad \Delta x = -\frac{f'(x_n)}{f''(x_n)}$$

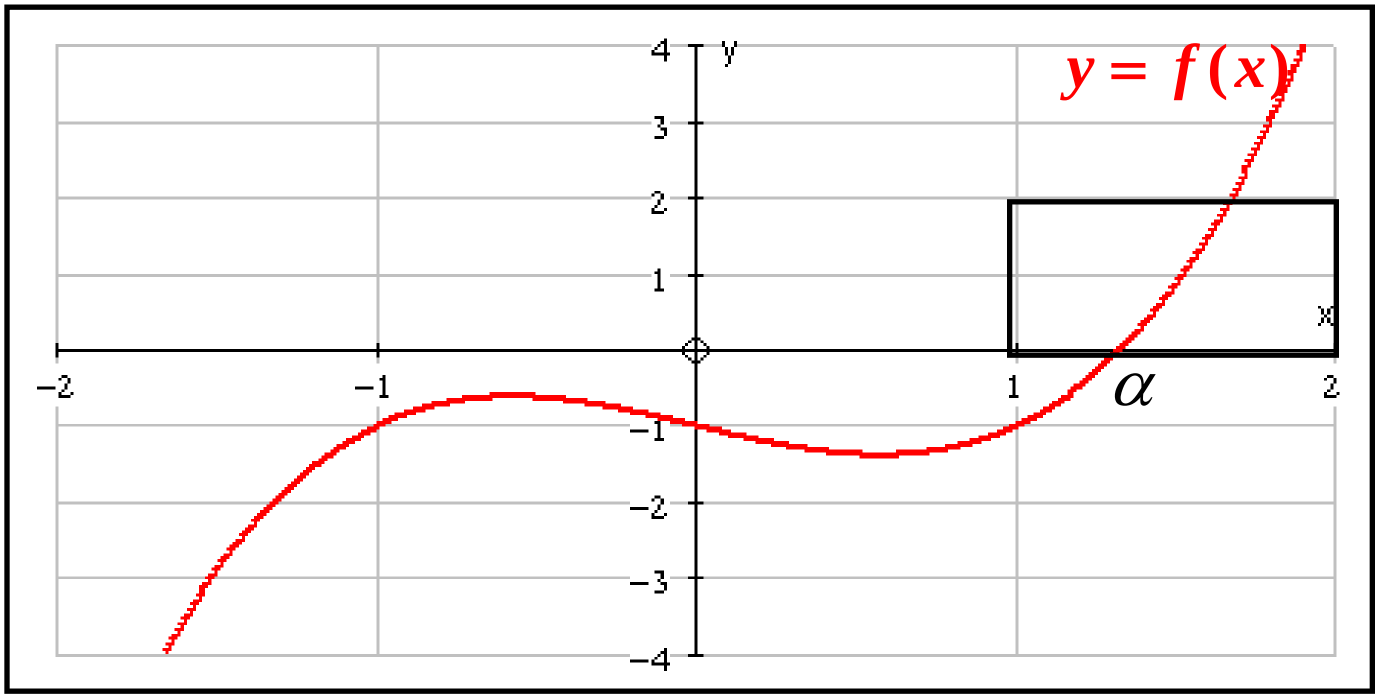
$$x_{n+1} = x_n + \Delta x = x_n - \frac{f'(x_n)}{f''(x_n)}$$

Multi-dimensional $\mathbf{x}_{n+1} = \mathbf{x}_n - \gamma[\mathbf{H} f(\mathbf{x}_n)]^{-1} \nabla f(\mathbf{x}_n)$

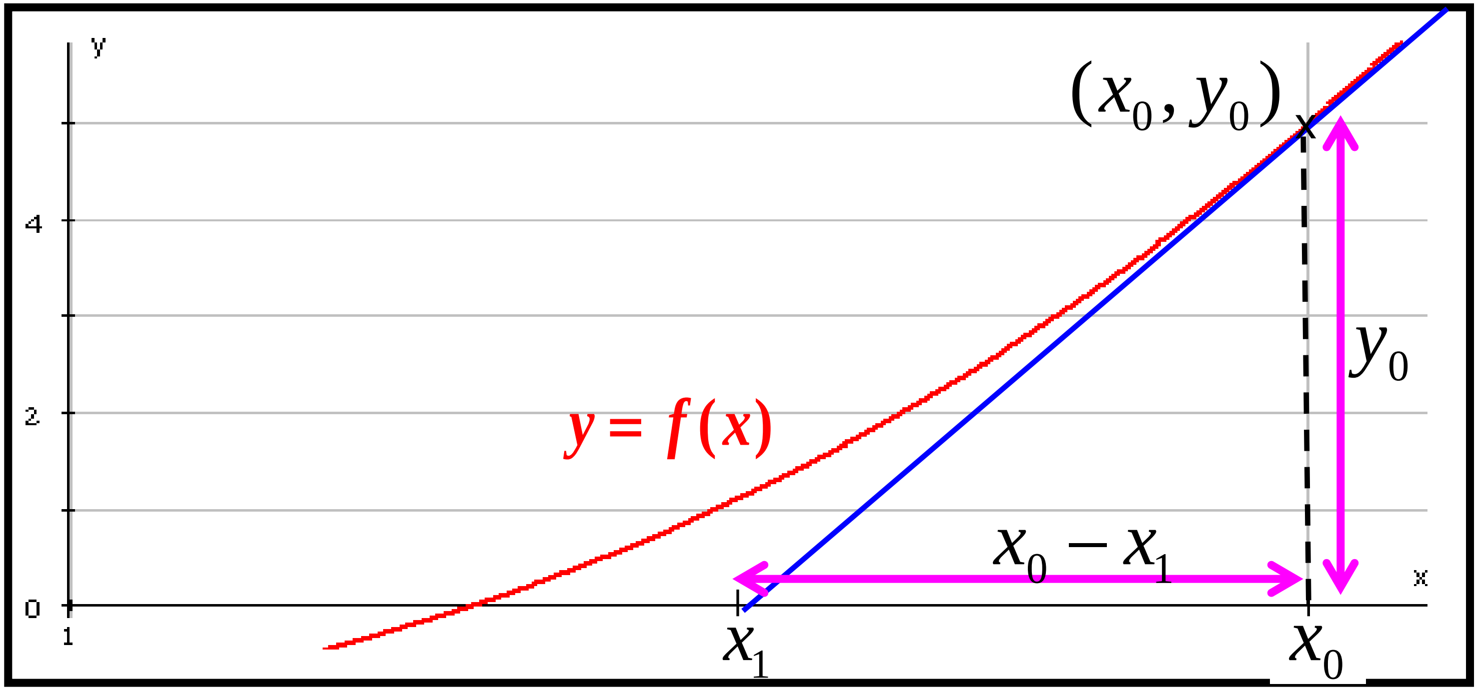
Newton-Raphson

Find an approximate solution to the equation

$$f(x) = x^3 - x - 1 = 0$$



Newton-Raphson



To carry out the iteration we need to find the points where the tangents meet the x-axis.

The grad. of the tangent $= \dfrac{\text{the change in } y}{\text{the change in } x}$

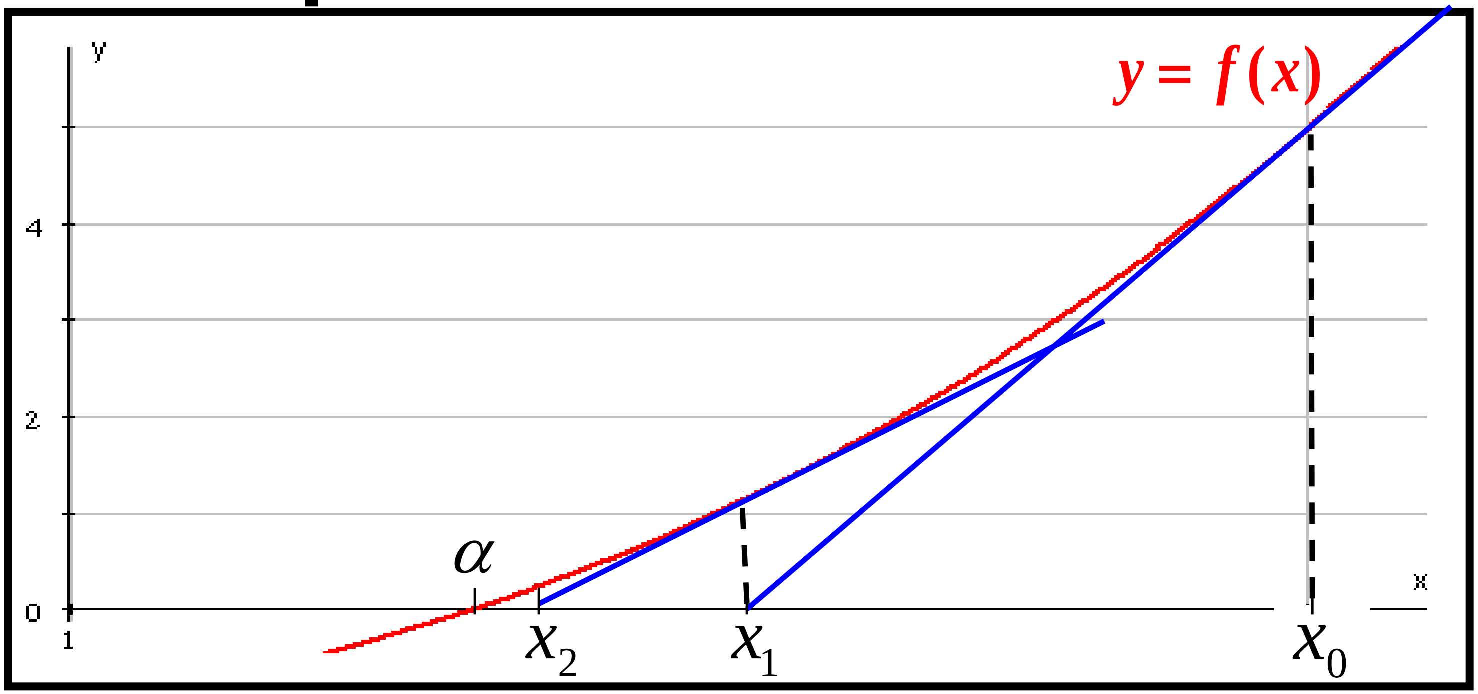
$$y = f(x) \quad \frac{dy}{dx} \text{ at } x_0 = \frac{y_0}{x_0 - x_1} = \frac{f(x_0)}{x_0 - x_1} = f'(x_0) \qquad \Rightarrow \quad x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$$

$$\boxed{x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}}$$

We must know how to differentiate $f(x)$

Convergence is often very fast

Newton-Raphson



Suppose our first estimate is given at $x_0 = 2$

We draw the tangent to the curve (linearize) at x_0

Find point x_1 where the tangent crosses the X-axis

Repeating . . .

Each point x_1 , . . . x_2 is closer to α

Newton-Raphson

- To find a minimum of *f(x)* we use 1st and 2nd order derivative

$$f(x) = 1.2(x-2)^2 + 3.2$$

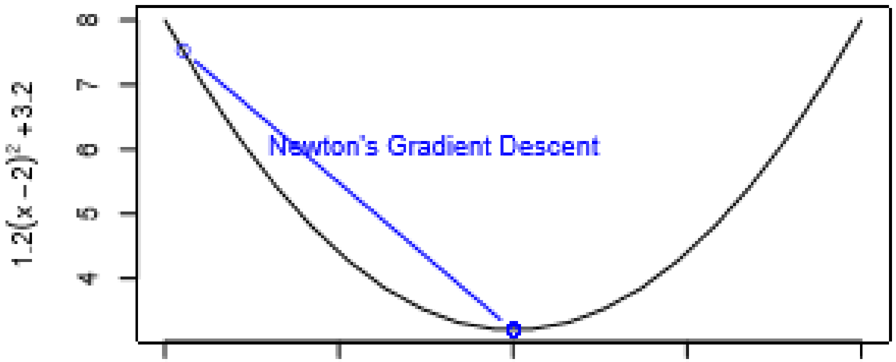
$$\frac{df(x)}{dx} = f' = 2.4(x-2)$$

$$\frac{d^2 f(x)}{dx^2} = f'' = 2.4$$

$$x(n) = x(n-1) - \frac{f'}{f''}$$

$$x(n) = x(n-1) - \frac{2.4(x-2)}{2.4}$$

$$x(n) = x(n-1) - (x-2)$$



Newton vs gradient descent

- Gradient descent and Newton Raphson
  - Newton's method uses curvature information (second order derivative) to take a faster route
  - Newton converges faster than gradient descent
- If we cannot directly solve for x or obtain 2nd order derivative?
  - cannot do Newton
    - In multidimensional problems eg. N, the second derivative is a Hessian matrix NxN, may be too big
    - Second derivative may lead far away
  - can still do gradient descent
- gradient descent guarantees convergence to a local minimum
  - convergence can be slow near the minimum point
  - increasingly 'zigzags' as the gradients point nearly orthogonally to the shortest direction to a minimum point