Numerical Optimization

Raja Damanik, M.Sc.

What we will learn

- Optimization & Application
- Newton Method
- Steepest Descent
- Constrained optimization problem setting

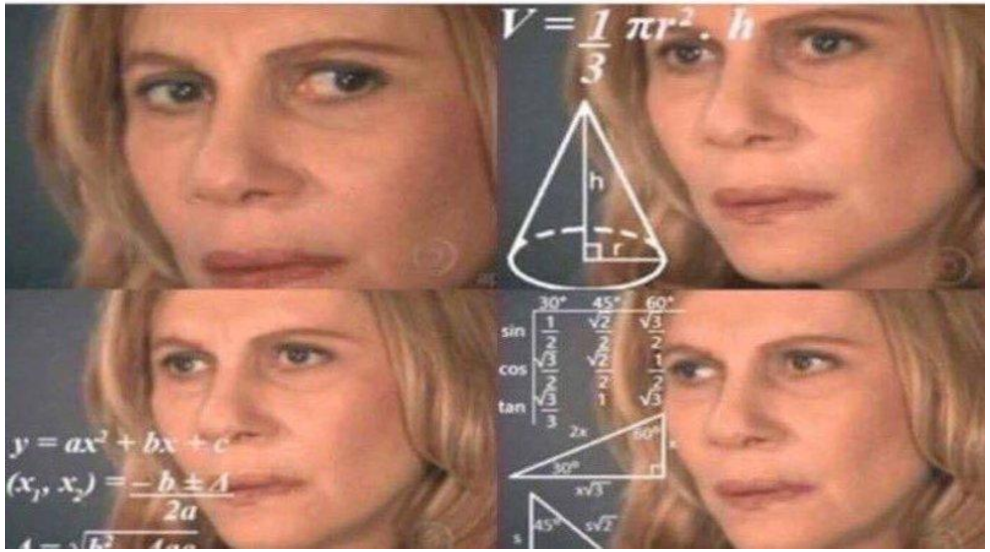
What is optimization

- Given a function f:
  - Find the maximum / minimum possible value of f.
  - Find the input that realizes this maximum / minimum value.

- f is called the **objective function**
- The input of f can be thought of factors that affect the value of the objective function

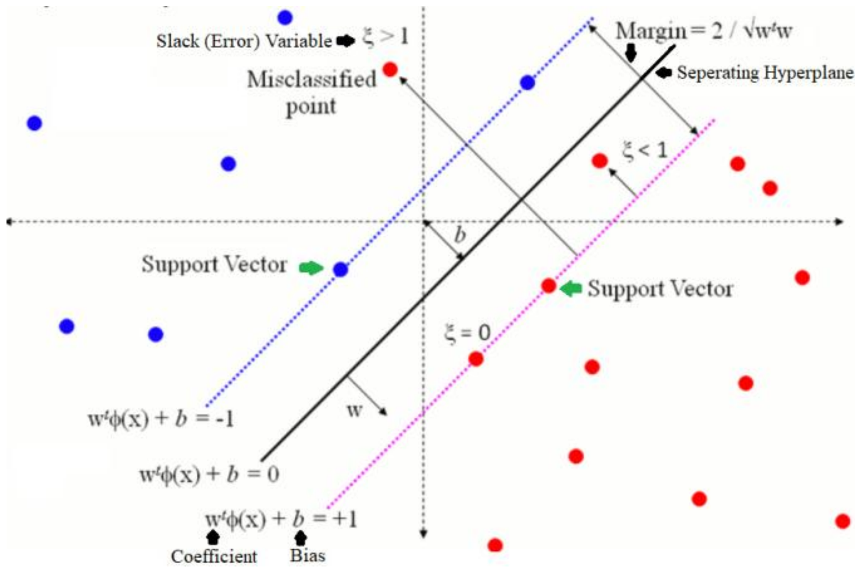
What is optimization

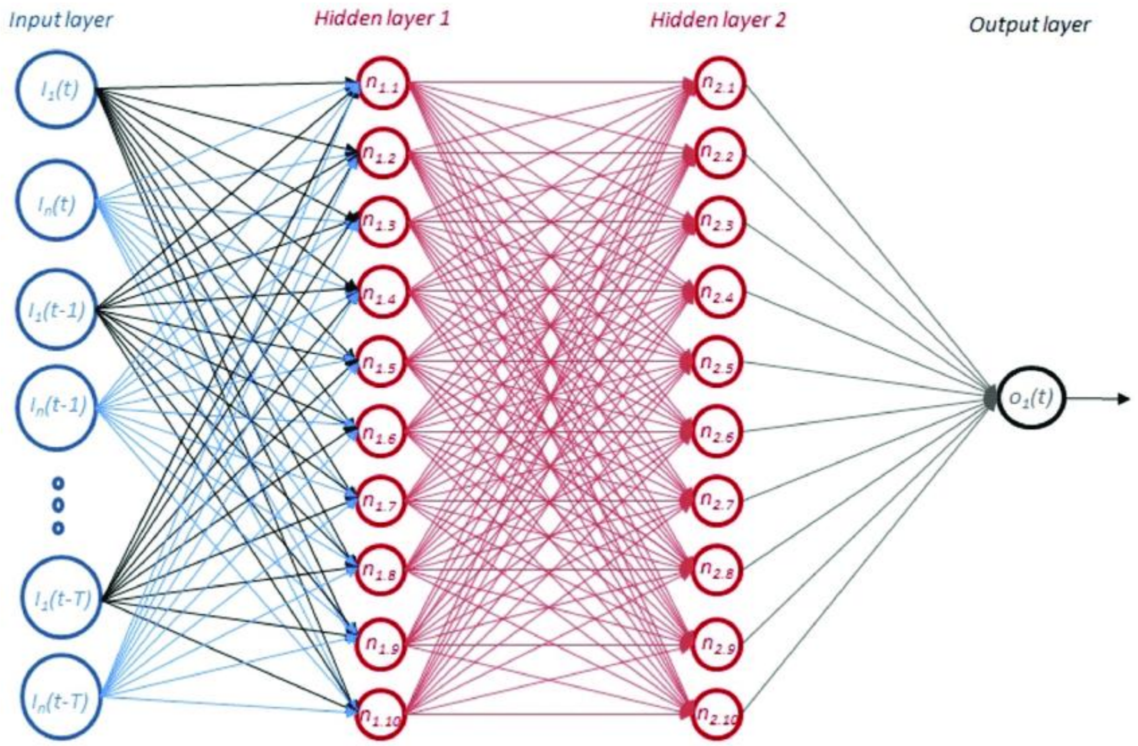
- In real life, one needs to carefully:
  - Identify what to optimize, i.e. what is the objective
  - What are factors that will affect the objective
  - How does each factor affect the objective

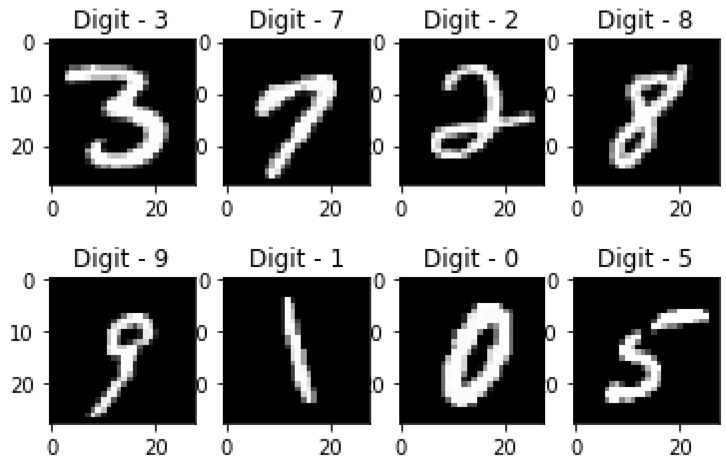


Application

- Backbone of machine learning

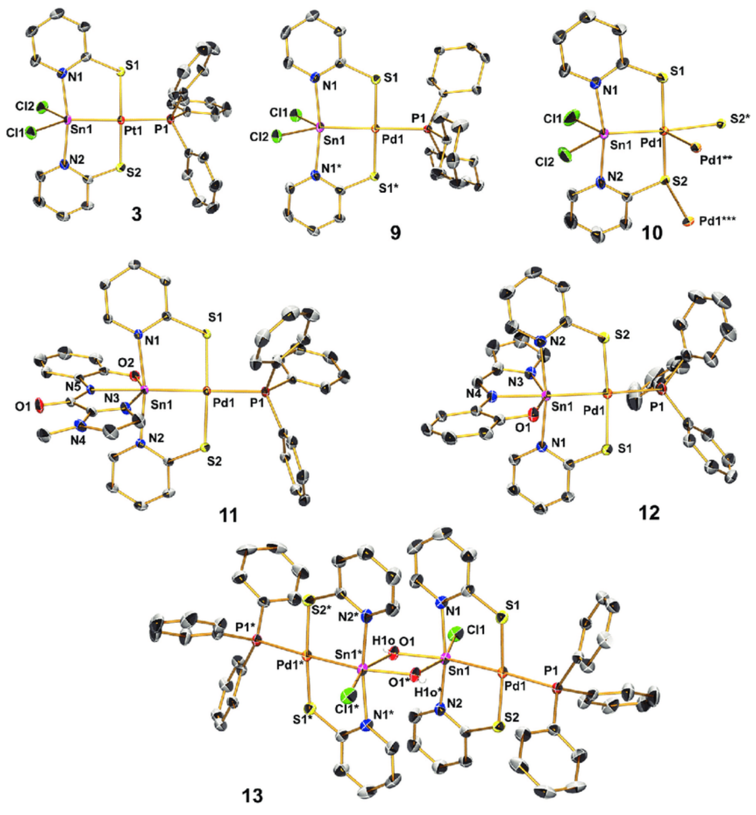


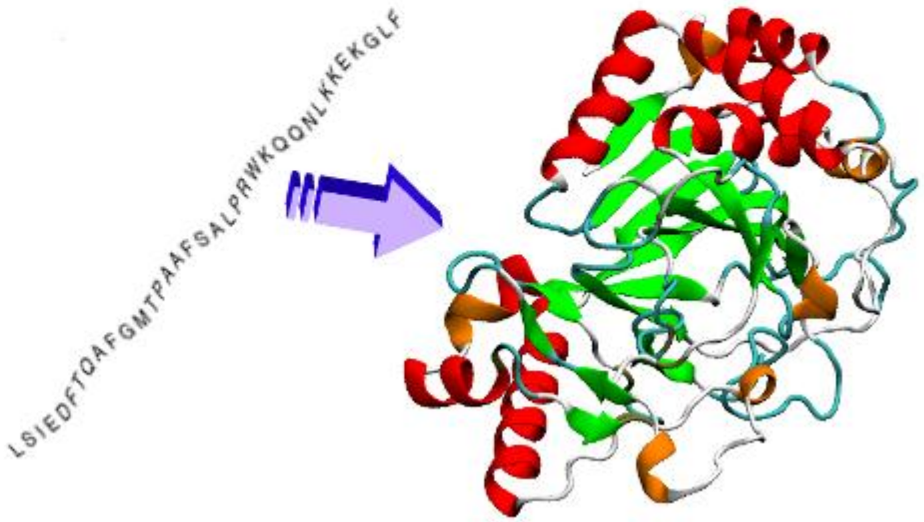




Application

- Understanding protein folding -> very important research!



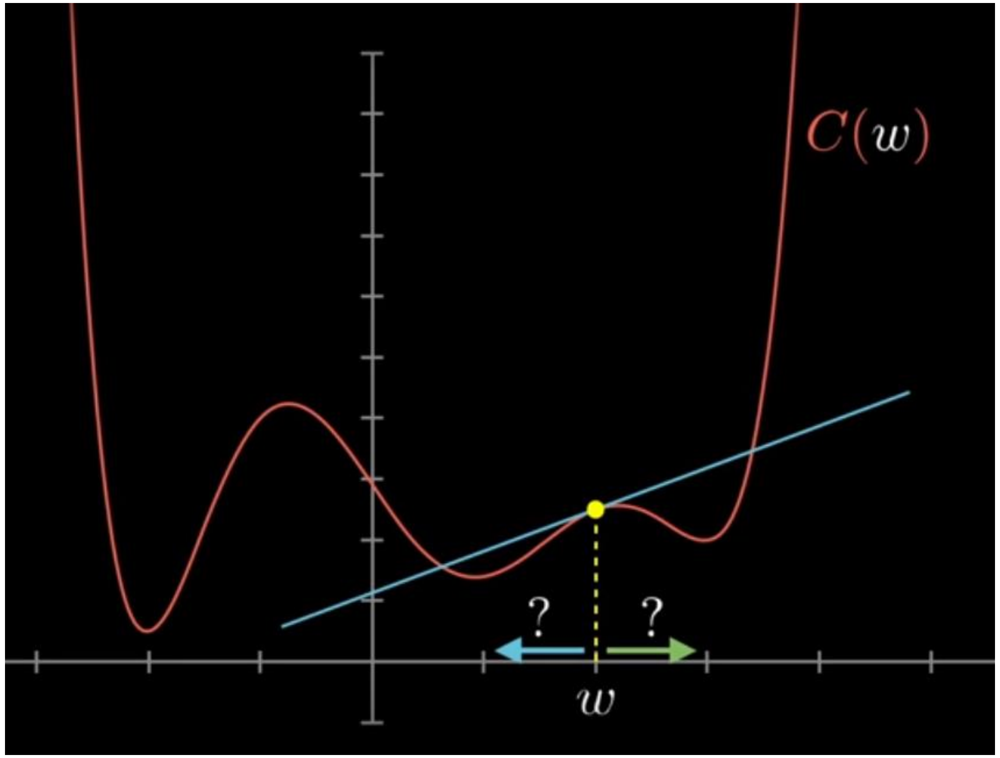


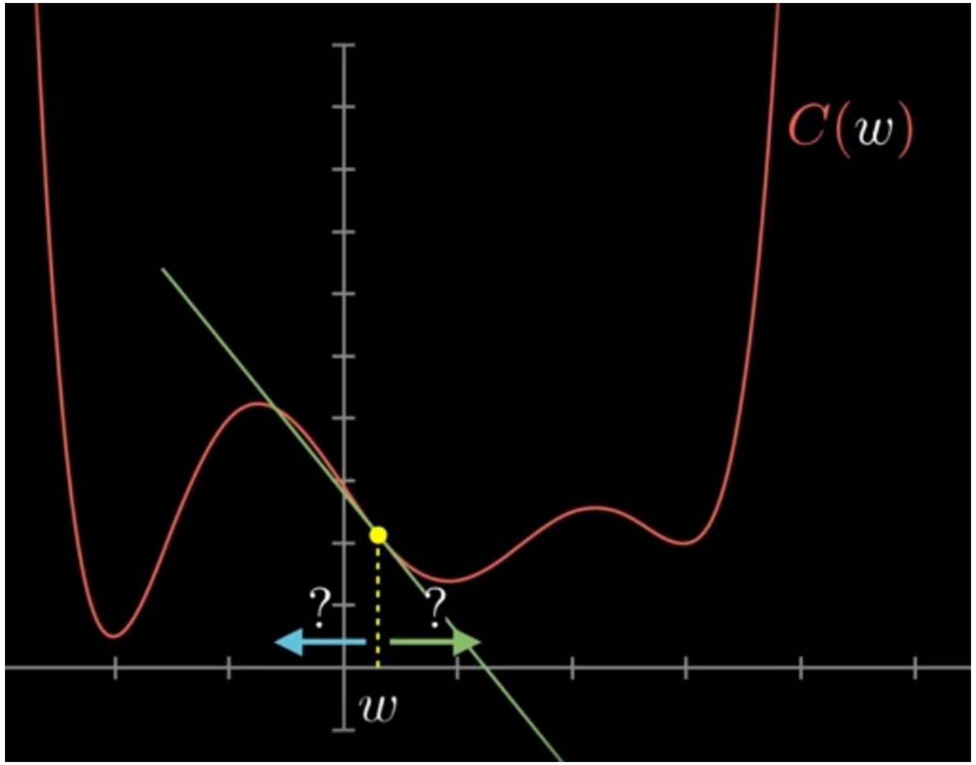
Application

- Motion planning in robotics
- Investment

Gradient

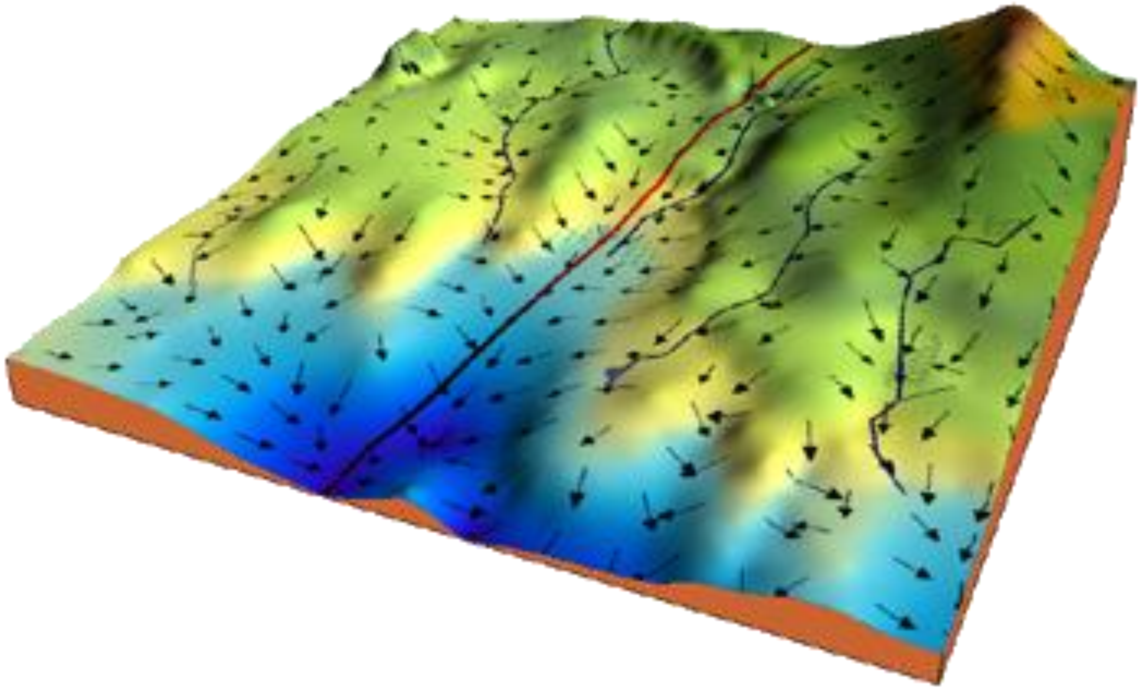
- Gradient in graph of $y = f(x)$.





Gradient

- Gradient for a surface $z = f(x, y)$.



Gradient

- Given

$$f: \mathbb{R}^n \to \mathbb{R}.$$

- Define ∇f (read: *grad f*) by

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

Gradient

- At point (x_0, y_0), the vector

$$\nabla f(x_0, y_0)$$

is the direction where f has steepest-ascend.

- At point (x_0, y_0), the vector

$$-\nabla f(x_0, y_0)$$

is the direction where f has steepest-descend.

Gradient

- If (x_0, y_0) are **local maxima / minima** then it must be the case that $\nabla f(x_0, y_0) = 0$.
- But… if $\nabla f(x_0, y_0) = 0$, it is not necessary that (x_0, y_0) are local minima / maxima. It can be a saddle point.
- Text in blue suggests: If we want to find local maxima / minima, try to solve $\nabla f = 0$. The local maxima / minima must be somewhere there.
- Text in red suggests: There must be another test to identify which is what: local minima / local maxima / saddle-point.

Hessian

- Recall that in our setting,

$$f : \mathbb{R}^n \to \mathbb{R}$$

and

$$\nabla f : \mathbb{R}^n \to \mathbb{R}^n$$

- Now, the derivative of ∇f, as we know is just $J_{\nabla f}$.

Hessian

- We give $J_{\nabla f}$, a new name:

Hessian of f.

$$H_f\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_n} \end{pmatrix}$$

Analogy

$\boldsymbol{f}: \mathbb{R} \to \mathbb{R}$	$\boldsymbol{f}: \mathbb{R}^{\boldsymbol{n}} \to \mathbb{R}$
$f'(x)$	$\nabla f(\mathbf{x})$
$f''(x)$	$H_f(\mathbf{x})$
$f'''(x)$	☺

Properties of Hessian matrix

- Hessian matrix H_f must be **symmetric** due to symmetricity of partial derivatives

$$\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i}$$

Second-derivative test

Suppose $\nabla f(x^*) = 0$. Then:

- If $H_f(x^*)$ is positive definite, $f(x^*)$ is local minima.
- If $H_f(x^*)$ is negative definite, $f(x^*)$ is local maxima.
- Otherwise, $f(x^*)$ is a saddle-point.

Unconstrained optimization

- Given a function

$$f: \mathbb{R}^n \to \mathbb{R}$$

- Find global optimum of f, that is:

$$x^* \text{ such that } f(x^*) \leq f(x) \text{ for any } x.$$

Assumption

- Function f is continuous & smooth (differentiable as many times as we like)
- We are interested in local minima

Strategy

- Numerical methods can only find local minima at most.

Newton Method

- Goal: Find the local minima of $f(x)$.
- Idea:

Solve the non-linear system

$$\nabla f(\mathbf{x}) = \mathbf{0}$$

Newton method

1. Set initial guess x_0
2. Compute ∇f and H_f
3. At each iteration:
    1. Compute matrix $H = H_f(x_k)$ and vector $b = \nabla f(x_k)$
    2. Compute vector s from linear-system $Hs = b$
    3. Update $x_{k+1} = x_k - s$

Stopping criterion: $||\nabla f(x_k)|| < \mathrm{TOL}$

Properties of Newton Method

- Once the initial guess is close to the solution, it converges very fast (quadratic convergences). ☺

- But, there is a trade-off because in each iteration:
  - Computing ∇f and H_f requires $O(n^2)$ scalar-function calls
  - Computing s requires $O(n^3)$ scalar-function calls.

Strategy

- Instead of computing the Hessian H_f, we use an approximation to H_f that we update at each iteration.

- Three approximation-methods to H_f:
  - SR-1
  - DFP
  - BFGS

Quasi-Newton Method

1. Set initial guess x_0 and $B_0 = I$
2. At each iteration:
   a. Compute $p_k = -B_k \nabla f(x_k)$
   b. Define a scalar function in α: $g(\alpha) = f(x_k + \alpha p_k)$.
   c. Using line-search, find step-size α^* that minimizes the scalar-function.
   d. Update $x_{k+1} = x_k + \alpha^* p_k$
   e. Update B_{k+1} using a suitable update formula (SR-1, DFP, BFGS)
3. Jika sudah dicapai konvergensi, berhenti

Updating B_k

Let $\delta^k = x_{k+1} - x_k$ and $\gamma^k = \nabla f(x_{k+1}) - \nabla f(x_k)$

- SR-1:

$$B^{k+1} = B^k + \frac{(\delta^k - B^k\gamma^k)(\delta^k - B^k\gamma^k)^T}{(\delta^k - B^k\gamma^k)^T\gamma^k}$$

- DFP:

$$B^{k+1} = B^k + \frac{\delta^k(\delta^k)^T}{(\delta^k)^T\gamma^k} - \frac{B^k\gamma^k(\gamma^k)^T B^k}{(\gamma^k)^T B^k\gamma^k}$$

- BFGS:

$$B^{k+1} = B^k + \left(1 + \frac{(\gamma^k)^T B^k\gamma^k}{(\delta^k)^T\gamma^k}\right)\frac{\delta^k(\delta^k)^T}{(\delta^k)^T\gamma^k} - \left(\frac{\delta^k(\gamma^k)^T B^k + B^k\gamma^k(\delta^k)^T}{(\delta^k)^T\gamma^k}\right)$$

Steepest Descent

- Goal: Find the local minima of $f(x)$.
- Idea:

  Move from x_k to x_{k+1} towards the direction of steepest descent

Steepest Descent

1. Set initial guess x_0
2. Compute ∇f
3. At k-th iteration:
   - Compute $p_k = -\nabla f(x_k)$
   - Define a scalar function in α,

$$g(\alpha) = f(x_k + \alpha p_k)$$

   - Using line-search, find scalar α^* that minimizes $g(\alpha)$
   - Update $x_{k+1} = x_k + \alpha^* p_k$

Stopping criterion: $||\nabla f(x_k)|| < \mathrm{TOL}$

Line-Search: Finding step-size α

- One can use Exact Line Search to find α: solve non-linear scalar equation $g'(\alpha) = 0$ using:
  - Bisecation method
  - Newton method (1-dimension version)
  - Secant method (1-dimension version)

- One can also use simpler & more relaxed method, e.g. Armijo Backtracking Line-Search.

Backtracking Line-Search

1. Set initial guess α_0 (e.g. $\alpha_0 = 1$) and $\ell = 0$
2. Until $f(x_k + \alpha_i p_k) < f(x_k)$:
    1. Set $\alpha_{i+1} = \frac{1}{2}\alpha_i$ (or multiply by some constant in the interval $(0,1)$
    2. Increment ℓ by one
3. Set $\alpha^* = \alpha_\ell$.

Armijo-Backtracking Line-Search

1. Set initial guess α_0 (e.g. $\alpha_0 = 1$) and $\ell = 0$
2. Until $f(x_k + \alpha_i p_k) \leq f(x_k) + 0.1\alpha_i \nabla f(x_k)^\top p_k$:
    1. Set $\alpha_{i+1} = \frac{1}{2}\alpha_i$ (or multiply by some constant in the interval $(0,1)$
    2. Increment ℓ by one
3. Set $\alpha^* = \alpha_\ell$.

Note: In case of steepest-descent method, $p_k := -\nabla f(x_k)$

CONSTRAINED OPTIMIZATION

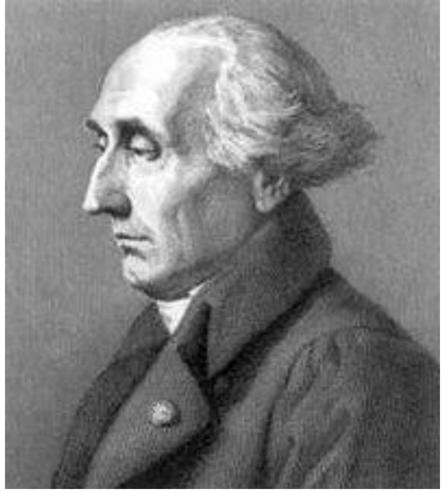
- Let $f: \mathbb{R}^n \to \mathbb{R}$ and $g: \mathbb{R}^n \to \mathbb{R}^m$.
- Find the maximum and minimum of $f(\boldsymbol{x})$ but with requirements that $\boldsymbol{x}$ satisfies $g(\boldsymbol{x}) = \boldsymbol{0}$

- **Example:**

If $x^2 + y^2 = 1$, what is the maximum and minimum of $x + 2y$?

If $x^2 + y^2 = 1$ and $y^2 - z^2 = 3$, what is the maximum and minimum of xyz?

Lagrange’s Multiplier



- Define Lagrange function

$$L(\boldsymbol{x}, \boldsymbol{\lambda}) = f(\boldsymbol{x}) - \lambda^{\mathsf{T}} g(\boldsymbol{x})$$

where $\boldsymbol{x} \in \mathbb{R}^n, \boldsymbol{\lambda} \in \mathbb{R}^m$. **$\lambda$ is called the Lagrange’s multiplier.**

- Find $\boldsymbol{x}^*$ **and $\boldsymbol{\lambda}$** which satisfy the non-linear system

$$\frac{d}{dx_i} L(\boldsymbol{x}, \boldsymbol{\lambda}) = 0$$

$$\frac{d}{d\lambda_i} L(\boldsymbol{x}, \boldsymbol{\lambda}) = 0$$

- The system will find among the points in the constraint where f is **maximum and minimum**. To decide which one is maximum & minimum

Example 1

- If x and y satisfies $x^2 + y^2 = 1$, find the maximum and minimum value of

$$x + 2y.$$

Example 2

- **If $x^2 + y^2 = 1$ and $y^2 - z^2 = 3$, what is the maximum and minimum of xyz?**

Idea of Lagrange Multipliers

- Draw on the input space (xy −plane) the curve of $g(x, y) = 0$.
- Draw countour plot $f(x, y) = c$ for various c on the input space.
- If $f(x, y)$ is maximum / minimum, the curve of $g(x, y) = 0$ and $f(x, y) = c$ must be tangent.
- Hence, their gradient (tangent line) must be parallel, namely
$$\nabla f(x^*, y^*) = \lambda \nabla g(x, y).$$