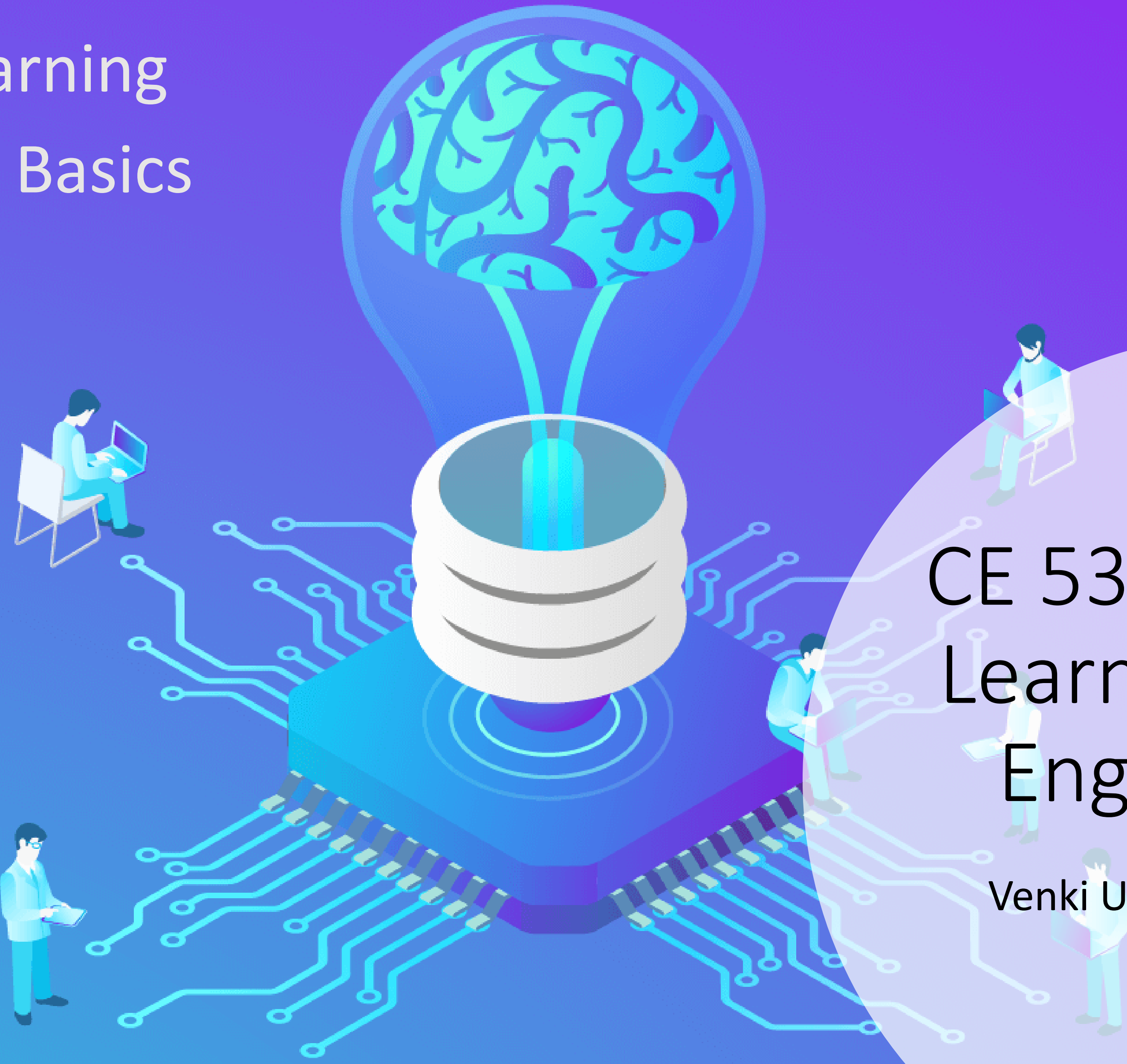
Machine Learning Optimization Basics

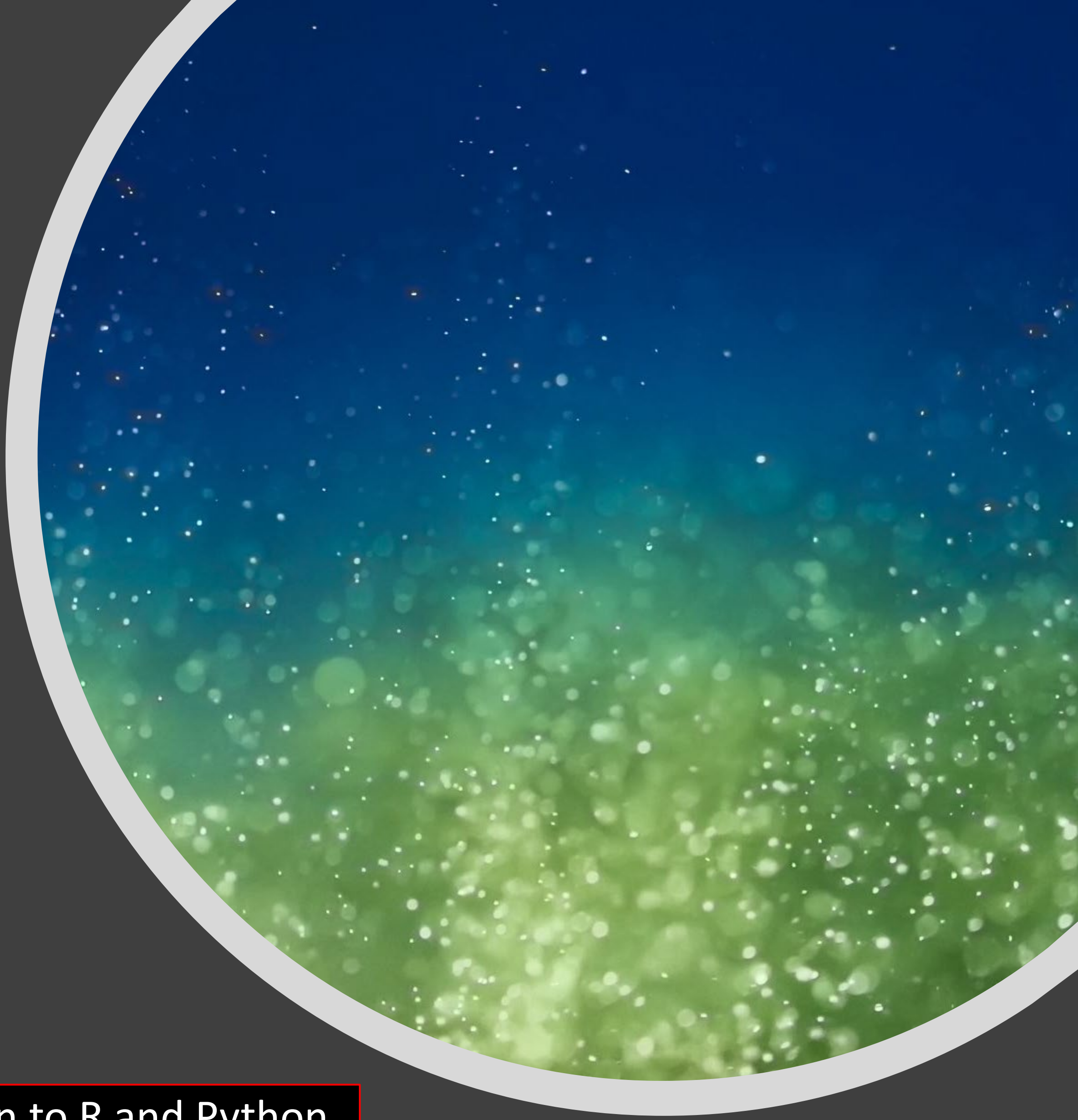


CE 5331 Machine Learning for Civil Engineers

Venki Uddameri, Ph.D. , P.E.

Recap

- Why study machine learning
- How machine learning is useful in civil engineering
- Data Types
- Exploratory Data Analysis
  - Visualization
  - Summary Statistics
- Basic Probability
  - Common Continuous and Discrete Distributions



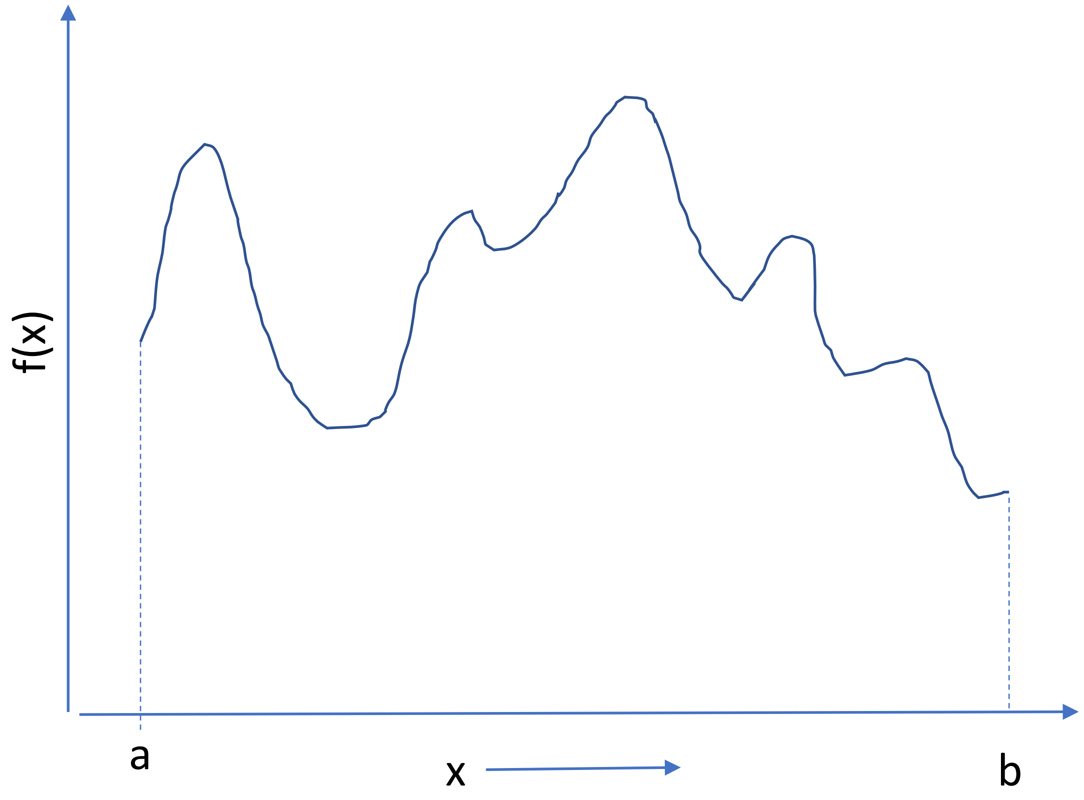
Introduction to R and Python
Anaconda and R Studio

Goals

- Present an Overview of Optimization
- Discuss Gradient Descent Approach
- Optimization in Python and R

Functions

- Functions map a relationship between input(s) and output(s)
  - Functions can be of many types
  - Mathematical, logical, rule-based
- The mapping of inputs → outputs can be linear or nonlinear
- The function can be one-dimensional or multi-dimensional



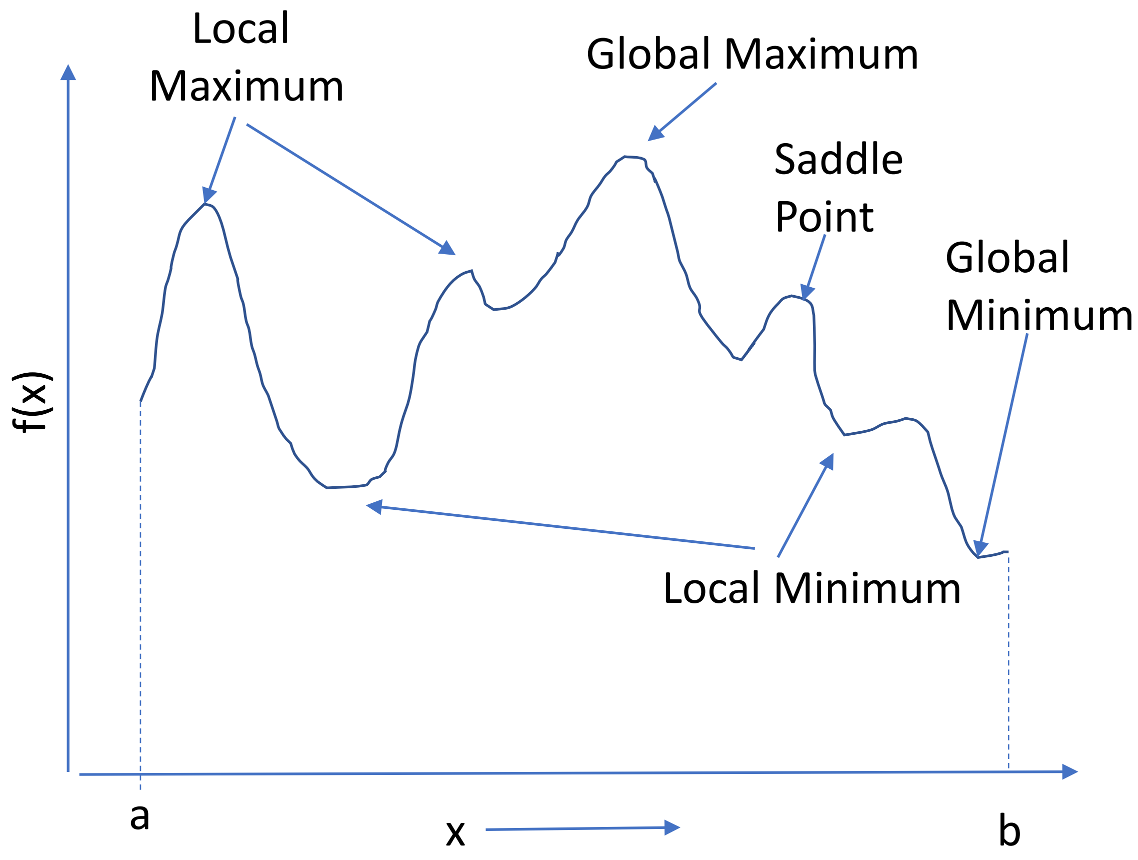
The function f(x) is defined over an interval $a \le x \le b$

A function can defined over an open or a closed interval

[a,b] is the closed interval – includes both a and b
(a,b) is the open-interval – does not include a and b

Optimization

- Optimization means finding either a ‘maximum’ or a ‘minimum’ of a function
- This optimum value is typically defined over a range of interest



Local Maximum: Value is higher than other values in the vicinity
Local Minimum: Value is lower than other values in the vicinity

Global Maximum: Highest value in the range (a,b)
Global Minimum: Lowest value in the range (a,b)

A Saddle Point Occurs when the value of the function is higher on one side and lower on the other (the slope is zero at that point)

Convex Function

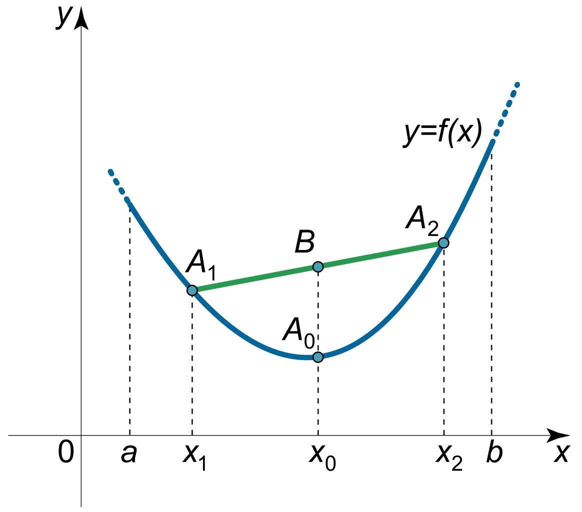
- A function is said to be strictly convex when a line connecting any two points of the function lies strictly above the function
- A function is convex if the second derivative is greater than 0

A convex function is also referred to as convex upwards function

A global minimum value exists if the function is convex

For a convex function the local minimum is also its global minimum

Convex Function



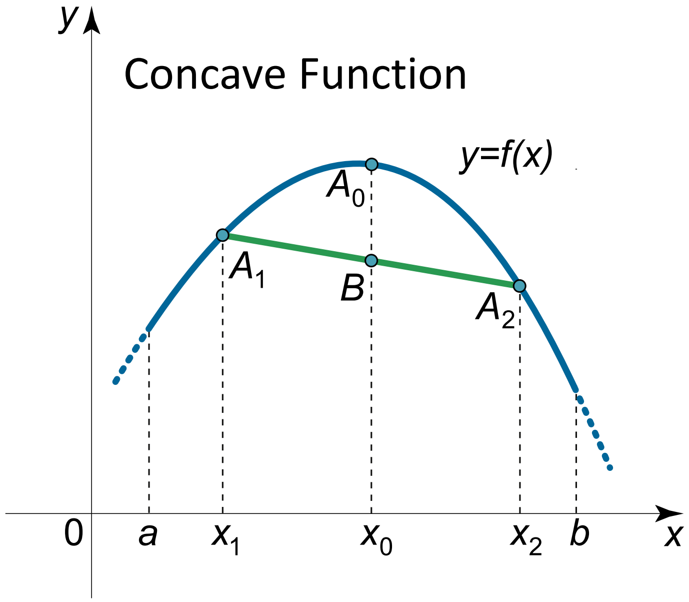
$$\frac{d^2 f(x)}{dx^2} > 0$$ Strictly Convex

$$\frac{d^2 f(x)}{dx^2} \geq 0$$ Convex

Concave Function

- For a concave function a straight line joining any two points lies below the function

- The second derivative of a concave function is greater than zero



A concave function is also called a convex downwards function

$$\frac{d^2 f(x)}{dx^2} < 0$$ Strictly Concave

A global maximum value exists if the function is convex
For a convex function the local maximum is also its global maximum

$$\frac{d^2 f(x)}{dx^2} \leq 0$$ Concave

Global Optimum Values

- A point at which a function will have a maximum or minimum is called a stationary point.
- At the stationary point the first derivative of the function is equal to zero
  - Necessary Condition
- If the second-derivative is > 0 then it is a global minimum
  - Convex function
- If the second-derivative is < 0 then it is a global maximum
  - Concave function

$$\left.\frac{df(x)}{dx}\right|_{x=x_o} = 0$$

Necessary Condition

$$\left.\frac{d^2 f(x)}{dx^2}\right|_{x=x_o} < 0$$

Concave function (global maximum)

$$\left.\frac{d^2 f(x)}{dx^2}\right|_{x=x_o} > 0$$

Convex function (global minimum)

We need to investigate further if the second-derivative is equal to zero

Locally Optimal Values

- If the second-derivative is equal to zero at the stationary point then:
- Calculate the first non-zero higher order derivative that is non-negative
  - Lets us call this the nth order derivative
  - The n can either be odd or even

$$\left.\frac{d^2 f(x)}{dx^2}\right|_{x=x_o} = 0$$

$$\left.\frac{d^n f(x)}{dx^n}\right|_{x=x_o} \neq 0$$

'n' is even

'n' is odd then saddle point

$$\left.\frac{d^n f(x)}{dx^n}\right|_{x=x_o} > 0$$

Local Minimum

$$\left.\frac{d^n f(x)}{dx^n}\right|_{x=x_o} < 0$$

Local Maximum

Functions of Multiple Variables

- Let f(X) be a function of several variables X = [x_1, x_2, x_3,.., x_n]
- A Hessian Matrix or simply Hessian or H-Matrix is a fundamental construct to study optimization over multiple variables
- A Hessian Matrix is constructed by taking the second-derivatives of the function
  - Diagonal elements have second-order derivatives with respect to one variable
  - Off-diagonal terms have cross-derivative terms

$$H = \begin{bmatrix} \frac{\partial^2 F(X)}{\partial x_1^2} & \frac{\partial^2 F(X)}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 F(X)}{\partial x_1 \partial x_n} \\ \frac{\partial^2 F(X)}{\partial x_2 \partial x_1} & \frac{\partial^2 F(X)}{\partial x_2^2} & \cdots & \frac{\partial^2 F(X)}{\partial x_2 \partial x_n} \\ \vdots & \vdots & & \vdots \\ \frac{\partial^2 F(X)}{\partial x_n \partial x_1} & \frac{\partial^2 F(X)}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 F(X)}{\partial x_n^2} \end{bmatrix}$$

The Hessian Matrix is often computed numerically at a given point

Convexity of a function with multiple Variables

- The eigen values of the Hessian Matrix tell us whether the function is convex or concave
  - If all the eigen values are positive then the function is strictly convex
  - If all the eigen values are negative then the function is strictly concave

Eigen values are obtained by solving the following characteristic equation

$$|\lambda I - H[f(X)]| = 0$$

Eigen Vectors

Identity Matrix

Hessian Matrix of f(X)

Convexity or Concavity cannot be ascertained if the eigen values are neither all positive or negative

Unconstrained Optimization

- For $X = X_o$ to be a stationary point the first derivative of f(X) wrt all variables must be equal to zero
- The Eigen values of the Hessian Matrix provide the sufficient conditions for local or global optima

$$\underbrace{\frac{\partial f(X)}{\partial x_1} = \frac{\partial f(X)}{\partial x_2} = \dots = \frac{\partial f(X)}{\partial x_n} = 0}_{\text{Necessary Condition}}$$

Solve a system of nonlinear equations to obtain unknown values of X that satisfy the above criteria

Compute the Hessian Matrix at the solution and obtain its Eigen Values to check for optimality conditions

Lagrange Multipliers

- Lagrange Multipliers technique is used to convert constrained optimization problems to unconstrained optimization problems
- The constraints are first expressed as "equality constraints" and folded into the objective functions

$$Min : Z = f(x_1, x_2, x_3)$$

Obj. function

$$subject\ to :$$

$$g(x_1, x_2, x_3) = b_1$$
$$h(x_1, x_2, x_3) = b_2$$

Constraints

Fold the Constraints into the obj. function to form the Lagrangian fn. (L)

$$\boxed{L = f(x_1, x_2, x_3) + \lambda_1 [b_1 - g(x_1, x_2, x_3)] + \lambda_2 [b_2 - h(x_1, x_2, x_3)]}$$

$$\frac{\partial L}{\partial x_1} = \frac{\partial L}{\partial x_2} = \frac{\partial L}{\partial x_3} = \frac{\partial L}{\partial \lambda_1} = \frac{\partial L}{\partial \lambda_2} = 0$$

Find derivatives and solve the system of nonlinear equations

$$\lambda_i = \frac{\Delta f(x)}{\Delta b_i}$$

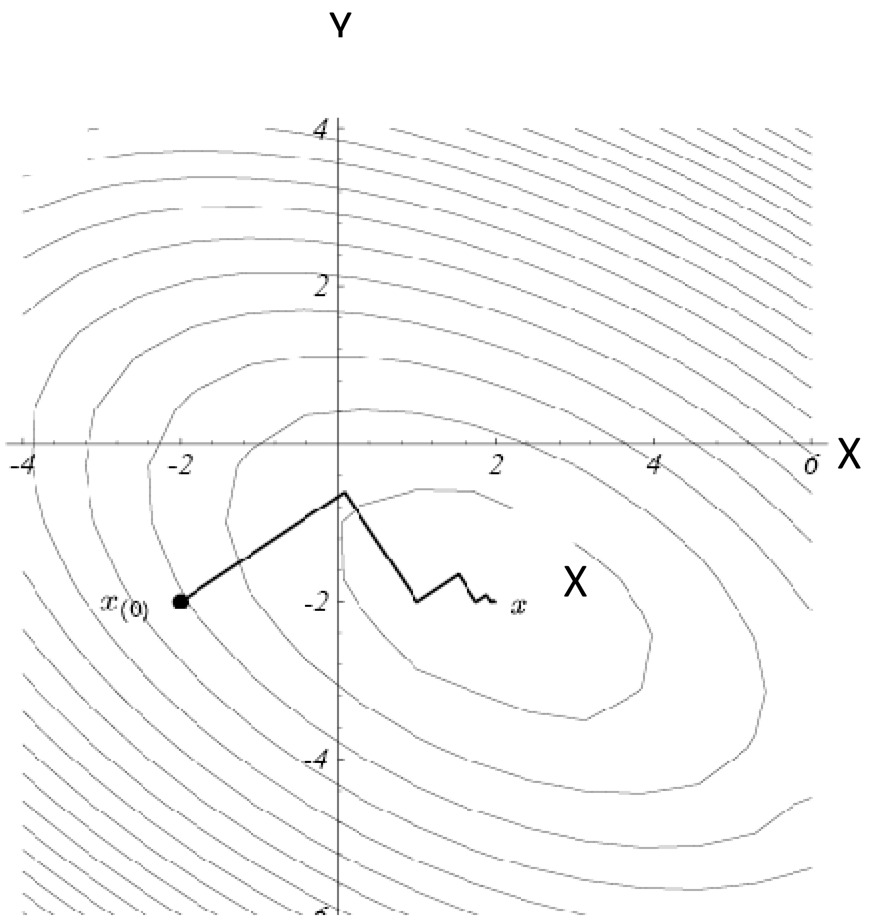
Lagrange Multipliers denote how the objective function changes with a change in the value of the constraint

From Lagrange Multiplier to Gradient Descent

- Lagrange multipliers generally work when the derivatives can be analytically solved
- For some functions this might not be possible
- For others while an analytical solution may be possible, the system of equations have to be solved numerically
- Gradient Descent is an algorithm used to perform nonlinear optimization
  - I will introduce a basic version here
  - There are several useful and important modifications that I will discuss later

Gradient Descent - Introduction

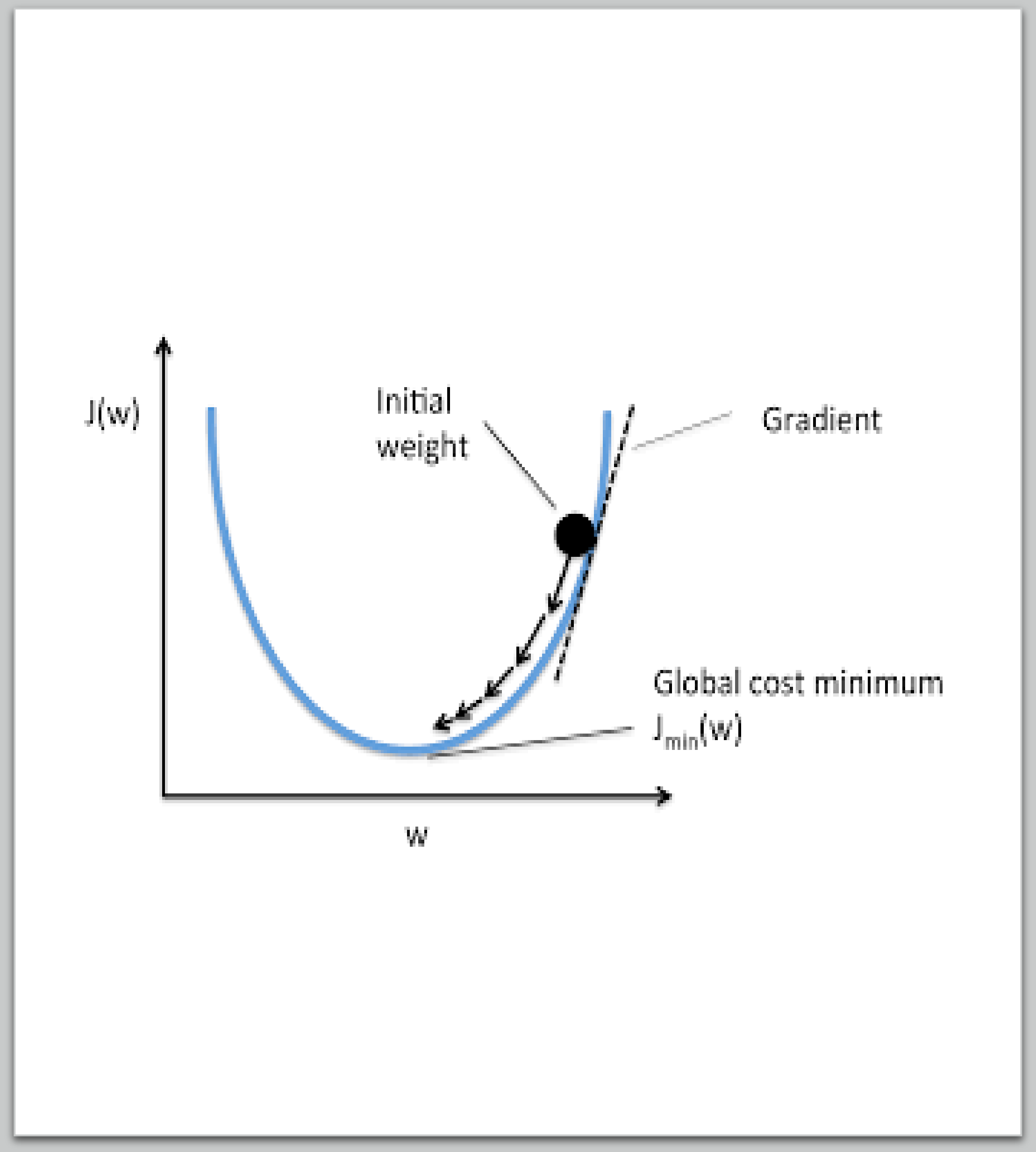
- Consider an objective function that we want to minimize
  - There are two decision variables x and y
  - Remember any constrained problem can be recast as an unconstrained problem
  - In LM method we take partial derivatives wrt x and y and set them to zero
    - What is we cannot do so or not able to solve the problem?
- We need to search the x-y space to find where the value is optimal



How should be proceed with the search?

Gradient Descent

- We want to search in a way that we can get to the optimum location in the least amount of steps
  - For maximization problems – Fastest Way to the top
  - For minimization problem – Slide down to the valley floor the fastest



Gradient Descent

- Gradient descent makes use of the concept of total derivative
    - Total Derivative is the best linear approximation of the function at any specified point

$$dz = \frac{\partial z}{\partial x}dx + \frac{\partial z}{\partial y}dy$$

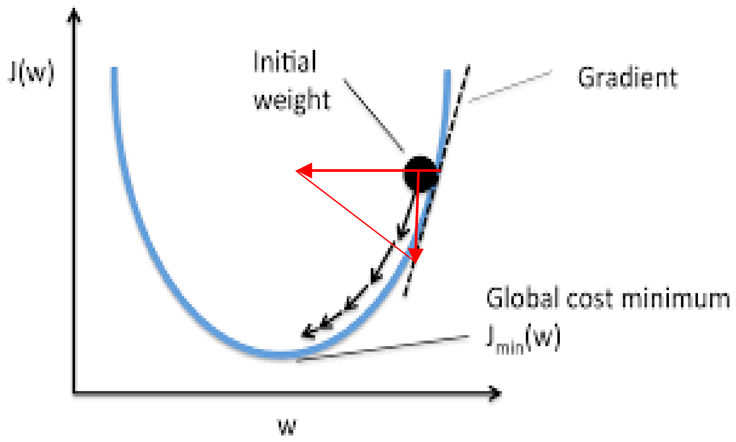
Now at any point c (x_c, y_c)

$$\Delta z = \left.\frac{\partial z}{\partial x}\right|_c \Delta x + \left.\frac{\partial z}{\partial y}\right|_c \Delta y$$

What is the maximum distance that we move?
How much should Δx and Δy be?

$$\boxed{\begin{aligned} &Max:\ \Delta z = Max:\ \left.\frac{\partial z}{\partial x}\right|_c \Delta x + \left.\frac{\partial z}{\partial y}\right|_c \Delta y \\ &Subject\ to:\ (\Delta x)^2 + (\Delta y)^2 = s^2 \end{aligned}}$$

Solve Using Lagrange Multipliers



Gradient Descent

- Taking derivatives wrt x and y of L and solve

$$L = \frac{\partial z}{\partial x}\bigg|_c \Delta x + \frac{\partial z}{\partial y}\bigg|_c \Delta y + \lambda\left[s^2 + \Delta x^2 + \Delta y^2\right]$$

$$\frac{\partial L}{\partial \Delta x} = \frac{\partial z}{\partial x} - 2\lambda\Delta x$$

$$\frac{\partial L}{\partial \Delta y} = \frac{\partial z}{\partial y} - 2\lambda\Delta y$$

$$\frac{\partial L}{\partial \Delta\lambda} = s^2 - (\Delta x)^2 - (\Delta y)^2$$

Equate to zero and solve

$$\boxed{\lambda = \frac{1}{2\Delta x}\frac{\partial z}{\partial x} \qquad \lambda = \frac{1}{2\Delta y}\frac{\partial z}{\partial y}}$$

$$\Delta x = \Delta y\left(\frac{\frac{\partial z}{\partial x}}{\frac{\partial z}{\partial y}}\right)\Bigg|_c$$

$$s^2 = (\Delta x)^2 + (\Delta y)^2$$

$$s^2 = \left(\Delta y\left(\frac{\frac{\partial z}{\partial x}}{\frac{\partial z}{\partial y}}\right)\Bigg|_c\right)^2 + (\Delta y)^2$$

$$(\Delta y)^2 = \frac{s^2}{\left(\left(\frac{\frac{\partial z}{\partial x}}{\frac{\partial z}{\partial y}}\right)\Big|_c\right)^2 + 1}$$

$$(\Delta y) = \frac{s\frac{\partial z}{\partial y}}{\left[\left(\frac{\partial z}{\partial x}\right)^2 + \left(\frac{\partial z}{\partial y}\right)^2\right]^{0.5}}$$

$$(\Delta x) = \frac{s\frac{\partial z}{\partial x}}{\left[\left(\frac{\partial z}{\partial x}\right)^2 + \left(\frac{\partial z}{\partial y}\right)^2\right]^{0.5}}$$

Where s is the step-size

$$x_{new} = X_c + \Delta X$$

Accuracy depends upon step size and How partials are computed

You should know

- What is optimization
- Why is it useful for machine learning
- What are convex and concave functions
- Optimization – Analytical solutions
- What is Lagrange Multiplier Method
- What is Gradient Descent