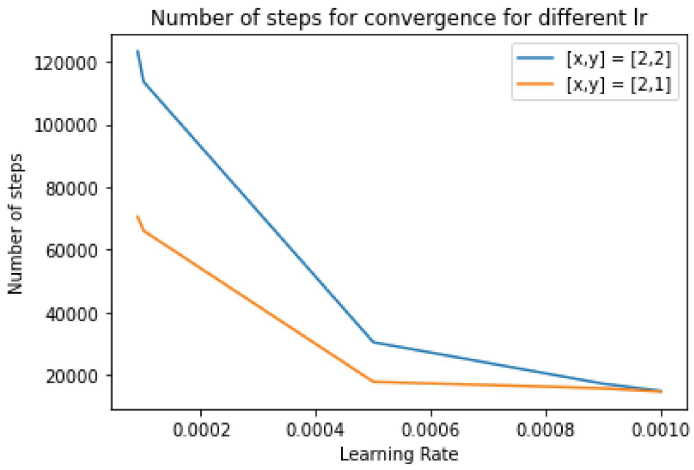
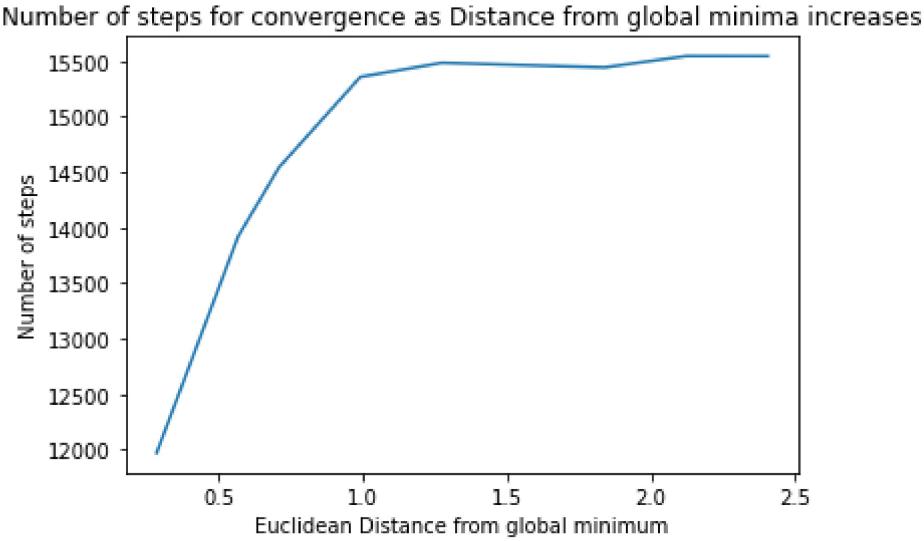
Problem Sheet 3:

Provide a graph which demonstrates the effect of learning rate on convergence of steepest descent algorithm (learning-rate(x) vs num iterations for convergence(y)). What do you think will be the optimal learning rate for this function?



- As we see in the above graph, as the learning rate increases from 2e-4 to 1e-3, the number of steps required to convergence is reduced.
- Increasing learning rate further leads to oscillation and it fails to converge.
- Also, if the initial guess is closer to global minimum, the number of steps is lesser.
- Optimal learning rate is 1e-3

Provide a graph which demonstrates the convergence rate of steepest descent for various initial guesses. (The distance of initial guess from global minimum(x) ||globM in – initial Guess||2 vs number of iterations(y).



- As distance of initial guess from global minimum increases, no of steps to converge increases. Learning rate of 1e-3 with tolerance 1e-6 was used for initial guess = [1.2,1.2], [1.4,1.4], [1.5,1.5], [1.7,1.7], [1.9,1.9], [2.3,2.3], [2.5,2.5], [2.7,2.7], [2.9,2.9].

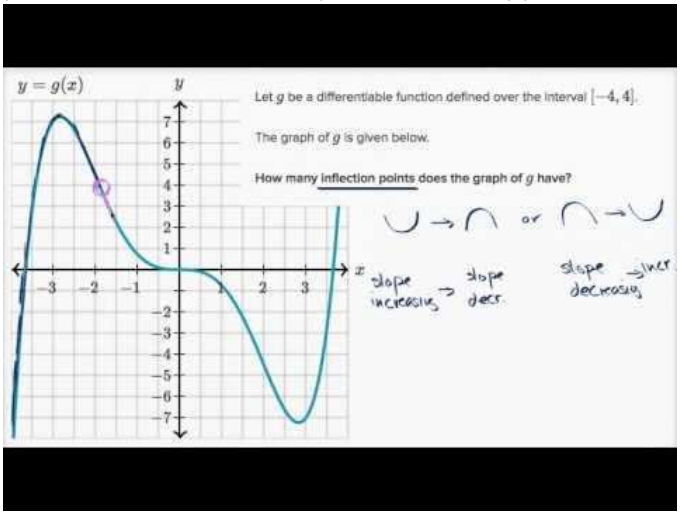
State true or false. The steepest descent method (i.e., gradient descent) requires multiple random starts to ensure the global minimum has reached if the optimization function is convex in nature. (2 Point)

- False: If optimization function is convex, gradient descent will converge with any random initial guess. Convex functions are smooth and gradient direction will always point in direction towards global minima irrespective of starting point.

You are given some function f(X): R n → R, X ∈ Ω ⊂ Rn. • If the Hessian matrix G = ∇2 f is positive definite, does f have a minimum? Justify your answer. • If G is negative definite, does f have a minimum value? Justify

- If Hessian matrix is positive definite, then function f(x) is a convex function. If a function f(x) is convex then f will have a minimal.
- Similarly, if hessian matrix is negative definite, then f(x) is a concave function and f will have a maximum.

Consider a function which has a stationary point but also has zero Hessian at the stationary point. Can you use the gradient descent method to find the stationary point for such a function?

- Yes, by doing Gradient descent, our function value will decrease until we hit a point where gradient is zero. Since hessian is zero, the stationary point is the point of inflection. So, convergence or identification of stationary points depends on the initial guess. If initial guess is a point above the inflection point, stationary points will be identified.
- 

Give a linear model which best approximates the following data (least square fit / linear regression). Use the analytical formula for linear regression

A simple linear regression model can be given as:

$y = wo + w1*x$

$Y = [1\ x]\ [wo\ w1]^T$

$Y = XW^T$

$Loss = (Y - XW^T)^T * (Y - XW^T)$

$Loss = Y^T Y - 2WX^T Y + WX^T XW^T$

$DLoss/dW = 0 - 2X^T Y + 2WX^T X => 0$

$W = (X^T X)^{-1} X^T Y$

Here $X = [1\ x] = [[1\ 1], [1\ 2], [1\ 3]]$

$W = [[1.33\ 0.33\ -0.67], [-0.5\ 0.0\ 0.5]] * [1.5\ 1.5\ 3]$

$W = [0.48, 0.75]$

$w0 = 0.48$, $w1 = 0.75$

Q.4 Usually quadratic function is given by

A.c) $x^T Ax - bx + c$

a)

$\nabla f(x) = Ax - b$ $\quad \nabla^2 f(x) = A \cdot = H \cdot$

$= Ax^{[i]} - b$

$= Ax^* + AMS - b$ $\quad \because x^{[i]} = x^* + MS$

$= b + AMS - b$ $\quad \because Ax^* = b \rightarrow$ gradient $= 0$

$= MAS$

$= MHS$ $\quad \rightarrow H = A^{-1}$

$\boxed{g^{[i]} = M\lambda S}$.

Q.4. (b) $f(x_0+\alpha d_k) = f(x_0) + \alpha d_k \nabla f(x_0)^T + \frac{1}{2} \alpha d_k^T \nabla^2 f(x_0)^2 \alpha d_k$

$d_k = -g$, $\nabla f(x_0) = g$, $\nabla^2 f(x_0) = H$

$= f(x_0) + \alpha d_1$.

$$f(x_0+\alpha d_k) = f(x_0) - \alpha g^T g + \frac{1}{2} \alpha^2 g^T H g$$

$\neq (f(x_0)) H$

Function converges if $f(x_0+\alpha d_k) = f(x_0)$.

diff. w.r.t α.

i.e $0 = -\alpha g^T g + \frac{1}{2} \alpha g^T H g$

optimal α :- $\boxed{\alpha = \frac{g^T g}{g^T H g}}$.

As $x_{k+1} = x_k + \alpha_k d_k$

Similarly $e_k \rightarrow$ error $(x_k - x^*)$

$e_{k+1} = e_k + \alpha_k d_k$

$= e_k - \frac{g^T g}{g^T H g} \cdot g$

$= e_k - \frac{g^T g \cdot (\lambda e)}{g^T H e \lambda}$

$= e_k - \frac{g^T g}{g^T g} \frac{\lambda e}{\lambda}$

$= e - e$

$\boxed{e_{k+1} = 0}$.

$g = Ax - b$

$g = Ax - Ax^*$

$g = Ae$

$g = \lambda e$

Since error = 0, convergence is reached. For one dimensional x, convergence reached in 1 step.