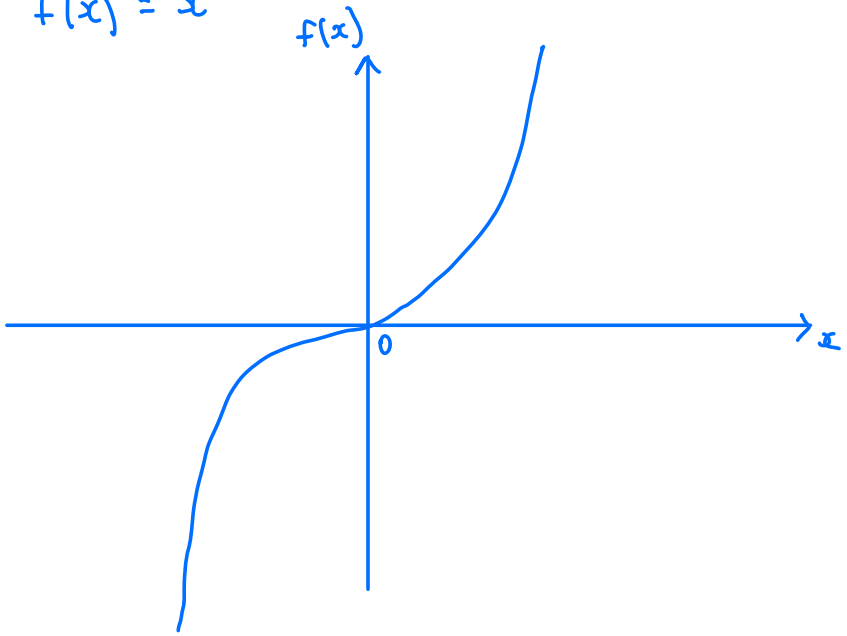
1. a.

$f(x) = x^3$



$\nabla f(x) = 3x^2$

At $x_0 = 0$ $\nabla f(x_0) = 0$, but as can be seen from the graph, it is neither a local minimum nor a local maximum of f.

It is a "SADDLE POINT".

IN 2 VARIABLES

$y = x_1^2 - x_2^2$

$\nabla f(x_1) = 2x_1$

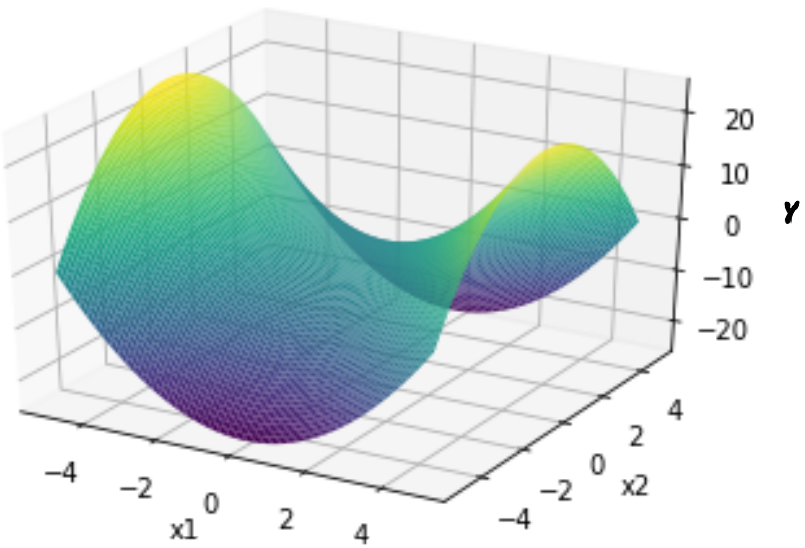
$\nabla f(x_2) = -2x_2$

At $x_0 = (0,0)$

$\nabla f(x_1)_0 = 0$

$\nabla f(x_2)_0 = 0$

We can observe the following plot to notice that it is neither a local minimum nor a maximum. It increases along one dimension and decreases along another.



The code to generate the above plot :

```
def f(x1, x2):
    return (x1 ** 2) - (x2 ** 2)

x1 = np.arange(-5, 5, 0.1)
x2 = np.arange(-5, 5, 0.1)

X1, X2 = np.meshgrid(x1, x2)
Y = f(X1, X2)

fig = plt.figure()
ax = plt.axes(projection='3d')
ax.plot_surface(X1, X2, Y, rstride=1, cstride=1,
                cmap='viridis', edgecolor='none')
ax.set_xlabel('x1')
ax.set_ylabel('x2')
ax.set_zlabel('y');
```

b. By Smoothness Upper Bound:

$$f(w_{k+1}) \leq f(w_k) - \frac{1}{2\beta} \|\nabla f(w_k)\|_2^2$$

So for $k=0$

$$f(w_1) \leq f(w_0) - \frac{1}{2\beta} \|\nabla f(w_0)\|_2^2 \quad -(1)$$

$$f(w_2) \leq f(w_1) - \frac{1}{2\beta} \|\nabla f(w_1)\|_2^2 \quad -(2)$$

Substitute (1) in (2)

$$f(w_2) \leq f(w_0) - \frac{1}{2\beta} \left(\|\nabla f(w_0)\|_2^2 + \|\nabla f(w_1)\|_2^2 \right)$$

If we continue for t iterations

$$f(w_{t+1}) \leq f(w_0) - \frac{1}{2\beta} \left(\|\nabla f(w_0)\|_2^2 + \|\nabla f(w_1)\|_2^2 + \cdots + \|\nabla f(w_t)\|_2^2 \right)$$

Let f^* be the global optimum

So $f(w_i) \geq f^*$

$$f(w_{t+1}) \geq f^*$$

On the RHS $\sum_{i=0}^{t} \|\nabla f(w_i)\|_2^2$ can be simplified by considering the smallest of the values.

$$f^* \leq f(w_0) - \frac{1}{2\beta} \times (t+1) \times \min_{i=\{0,\dots t\}} \left(\|\nabla f(w_i)\|_2^2\right)$$

$$\min_{i \in \{0,\dots t\}} \left(\|\nabla f(w_i)\|_2^2\right) \leq \frac{2\beta\left[f(w_0) - f^*\right]}{t+1}$$

So this proves that if we run gradient descent for t iterations, we can find a point w, such that

$$\|\nabla f(w)\|^2 \leq \frac{2\beta\left[f(w_0) - f^*\right]}{t+1}$$

i.e. $\|\nabla f(w)\|^2 \leq O(1/t)$

or

$\|\nabla f(w)\|$ is less than a constant, ε

when $\sqrt{\frac{2\beta\left[f(w_0) - f^*\right]}{t+1}} \leq \varepsilon$

As numerator is a constant, we can always increase the number of iterations to get the LHS $\leq \varepsilon$.

So let us now express number of iterations in terms of $f(w_0)$, β, ε and f^*.

We know $\|\nabla f(w)\| \leq \varepsilon$

only when t is such that

$$\frac{2\beta [f(w_0) - f^*]}{t+1} \leq \varepsilon^2$$

$$t+1 \geq \frac{2\beta}{\varepsilon^2} [f(w_0) - f^*]$$

$$t \geq \frac{2\beta}{\varepsilon^2} [f(w_0) - f^*] - 1$$

So, $\|\nabla f(w)\| \leq \varepsilon$

for all

$$t \geq \frac{2\beta}{\varepsilon^2} [f(w_0) - f^*]$$

2. f is α- convex

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\alpha}{2} \|y - x\|_2^2 \quad \forall x, y \in \mathbb{R}^n$$

So for $y = x^*$, $x = x_k$

$$f(x^*) \geq f(x_k) + \langle \nabla f(x_k), x^* - x_k \rangle + \frac{\alpha}{2} \|x^* - x_k\|_2^2$$

$$f(x_k) \leq f(x^*) - \langle \nabla f(x_k), x^* - x_k \rangle - \frac{\alpha}{2} \|x^* - x_k\|_2^2 \quad -(1)$$

f is β - SMOOTH

$$f(x_{k+1}) \leq f(x_k) - \frac{1}{2\beta} \|\nabla f(x_k)\|^2 \quad -(2)$$

(1) in (2)

$$f(x_{k+1}) \leq f(x^*) - \langle \nabla f(x_k), x^* - x_k \rangle - \frac{\alpha}{2} \|x^* - x_k\|_2^2$$

$$- \frac{1}{2\beta} \|\nabla f(x_k)\|_2^2 \quad -(3)$$

From Gradient Descent we know,

$$x_{k+1} = x_k - t \nabla f(x_k)$$

$$\nabla f(x_k) = \frac{1}{t} (x_k - x_{k+1})$$

$$t = 1/\beta$$

$$\nabla f(x_k) = \beta (x_k - x_{k+1}) \quad - (4)$$

(4) in (3)

$$f(x_{k+1}) \leq f(x^*) + \beta \langle x_k - x_{k+1}, x_k - x^* \rangle$$

$$- \frac{\alpha}{2} \|x^* - x_k\|_2^2 - \frac{1}{2\beta} \|\beta (x_k - x_{k+1})\|_2^2 \quad - (5)$$

we know

$$\|u - v\|^2 = \|u\|^2 + \|v\|^2 - 2\langle u, v \rangle$$

$$\langle u, v \rangle = \frac{1}{2} \left(\|u\|^2 + \|v\|^2 - \|u - v\|^2 \right)$$

$$\langle (x_k - x_{k+1}), (x_k - x^*) \rangle = \frac{1}{2} \left[\|x_k - x_{k+1}\|_2^2 + \|x_k - x^*\|_2^2 - \|x^* - x_{k+1}\|_2^2 \right]$$

$$- (6)$$

(6) in (5)

$$f(x_{k+1}) - f(x^*) \leq \frac{\beta}{2} \left[\|x_k - x_{k+1}\|_2^2 + \|x_k - x^*\|_2^2 - \|x^* - x_{k+1}\|_2^2 \right]$$

$$- \frac{\alpha}{2} \|x^* - x_k\|_2^2 - \frac{\beta}{2} \|(x_k - x_{k+1})\|_2^2$$

$$f(x_{k+1}) - f(x^*) \leq \frac{(\beta - \alpha)}{2} \|x_k - x^*\|_2^2 - \frac{\beta}{2} \|x^* - x_{k+1}\|_2^2 \quad - (7)$$

f is α- convex

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\alpha}{2} \|y - x\|_2^2 \quad \forall x, y \in \mathbb{R}^n$$

If $y = x_{k+1}$ $x = x^*$

$$f(x_{k+1}) \geq f(x^*) + \langle \nabla f(x^*), x_{k+1} - x^* \rangle + \frac{\alpha}{2} \|x_{k+1} - x^*\|_2^2$$

We know $\nabla f(x^*) = 0$ [optimal point]

$$f(x_{k+1}) - f(x^*) \geq \frac{\alpha}{2} \|x_{k+1} - x^*\|_2^2 \quad - (8)$$

From (7) and (8)

$$\frac{\alpha}{2} \|x_{k+1} - x^*\|_2^2 \leq \frac{(\beta - \alpha)}{2} \|x_k - x^*\|_2^2 - \frac{\beta}{2} \|x^* - x_{k+1}\|_2^2$$

$$\|x_{k+1} - x^*\|_2^2 \leq \frac{(\beta - \alpha)}{\alpha + \beta} \|x_k - x^*\|_2^2$$

$$(\alpha + \beta) \|x_{k+1} - x^*\|_2^2 \leq (\beta - \alpha) \|x_k - x^*\|_2^2 \quad - (9)$$

As $\alpha > 0$ $(\alpha + \beta) \|x_{k+1} - x^*\|_2^2 \geq \beta \|x_{k+1} - x^*\|_2^2 \quad - (10)$

Combine (9) and (10)

$$\beta \|x_{k+1} - x^*\|_2^2 \leq (\beta - \alpha) \|x_k - x^*\|_2^2$$

$$\|x_{k+1} - x^*\|_2^2 \leq \left(1 - \frac{\alpha}{\beta}\right) \|x_k - x^*\|_2^2$$

For $k = 0$

$$\|x_1 - x^*\|_2^2 \leq \left(1 - \frac{\alpha}{\beta}\right)\|x_0 - x^*\|_2^2 \quad - (11)$$

$k = 1$

$$\|x_2 - x^*\|_2^2 \leq \left(1 - \frac{\alpha}{\beta}\right)\|x_1 - x^*\|_2^2 \quad - (12)$$

Apply (11) in (12)

$$\|x_2 - x^*\|_2^2 \leq \left(1 - \frac{\alpha}{\beta}\right)^2 \|x_0 - x^*\|_2^2$$

Similarly

$k = 2$ $$\|x_3 - x^*\|_2^2 \leq \left(1 - \frac{\alpha}{\beta}\right)\|x_2 - x^*\|_2^2$$

$$\leq \left(1 - \frac{\alpha}{\beta}\right)^3 \|x_0 - x^*\|_2^2$$

So for k

$$\|x_k - x^*\|_2^2 \leq \left(1 - \frac{\alpha}{\beta}\right)^k \|x_0 - x^*\|_2^2 \; //$$

3. Assume x_0 is a local minimum

i.e. there exists $\delta > 0$ such that for all x

with $\|x - x_0\| < \delta$, $f(x_0) \leq f(x)$. - (1)

Let y be any other point in $\mathbb{R}^d$.

Let us take the point very close to x_0

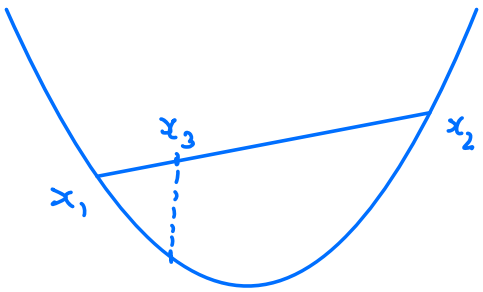
as $x_k = (1-\gamma)x_0 + \gamma y$, where γ is a very small positive number. ($\gamma \to 0$)

From (1)

$$f(x_0) \leq f(x_k)$$

$$f(x_0) \leq f((1-\gamma)x_0 + \gamma y) \quad - (2)$$

By Convexity, we know that



$$f\left((1-\gamma)x_1 + \gamma x_2\right) \leq (1-\gamma) f(x_1) + \gamma f(x_2) \quad - (3)$$

Apply ③ in ②

$$f(x_0) \le f((1-\gamma)x_0 + \gamma y) \le (1-\gamma) f(x_0) + \gamma f(y)$$

$$f(x_0) \le (1-\gamma) f(x_0) + \gamma f(y)$$

$$\gamma f(x_0) \le \gamma f(y)$$ γ is a small positive number.

$$f(x_0) \le f(y)$$

So, the value at x_0 is less than the value at any y, $y \in \mathbb{R}^d$.

So x_0 is a GLOBAL MINIMUM //

4. a. $L(w) = \frac{1}{n} \sum_{i=1}^{n} l(y_i, \sigma(\langle w, x_i \rangle))$

$$= \frac{1}{n} \sum_{i=1}^{n} \left(-y_i \log\left(\frac{1}{1+e^{-\langle w, x_i \rangle}} \right) + (1-y_i) \log\left(1 - \frac{1}{1+e^{-\langle w, x_i \rangle}} \right) \right)$$

Let $z_i = \langle w, x_i \rangle$

$$a_i = \frac{1}{1+e^{-z_i}}$$

$$L_i(w) = -y_i \log a_i + (1-y_i) \log(1-a_i)$$

$$\frac{\partial L_i}{\partial w} = \frac{\partial L_i}{\partial a_i} \cdot \frac{\partial a_i}{\partial z_i} \cdot \frac{\partial z_i}{\partial w}$$

$$\frac{\partial L_i}{\partial a_i} = \frac{-y_i}{a_i} + \frac{(1-y_i)}{1-a_i}$$

$$\frac{\partial a_i}{\partial z_i} = \frac{-1(-e^{-z_i})}{(1+e^{-z_i})^2} = \frac{1+e^{-z_i}-1}{(1+e^{-z_i})^2} = a_i - a_i^2$$

$$= a_i(1-a_i)$$

$$\frac{\partial z_i}{\partial w} = x_i$$

$$\frac{\partial L_i}{\partial w} = \left[-y_i (1 - a_i) + (1 - y_i) a_i \right] x_i$$

$$= \left(-y_i + y_i a_i + a_i - a_i y_i \right) x_i$$

$$= (a_i - y_i) x_i$$

$$\frac{\partial L}{\partial w} = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial L_i}{\partial w} = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{1}{1 + e^{-\langle w, x_i \rangle}} - y_i \right) x_i$$

$$\boxed{\frac{\partial L}{\partial w} = \frac{1}{n} \cdot x^T (\sigma(xw) - y)}$$

b) i) Gradient Descent, NAG:

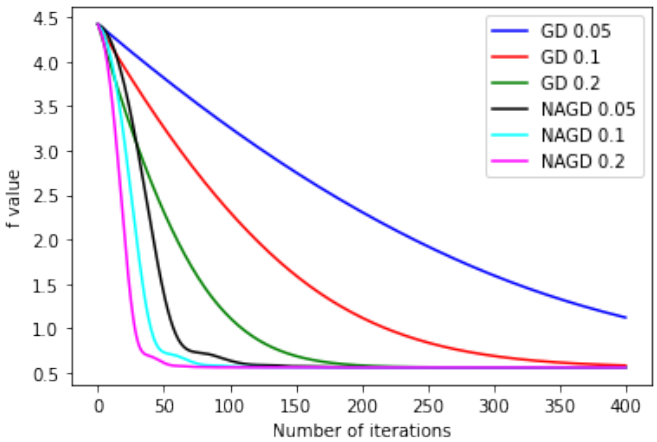
```
def gradient_descent(xinit,steps,gradient):
    """Run gradient descent.
    Return an array with the rows as the iterates.
    """
    xs = [xinit]
    x = xinit
    for step in steps:
        x = x - step*gradient(x)
        xs.append(x)
    return np.array(xs)

def nagd(winit,gradient,eta=0.1,nsteps=100):
    """Run Nesterov's accelrated graident descent.
    Return an array with the rows as the iterates.
    """
    ws = [winit]
    u = v = w = winit
    for i in range(nsteps):
        etai = (i+1)*eta/2
        alphai = 2/(i+3)
        w = v - eta*gradient(v)
        u = u - etai*gradient(v)
        v = alphai*u + (1-alphai)*w
        ws.append(w)
    return np.array(ws)
```

Cost and Gradient:

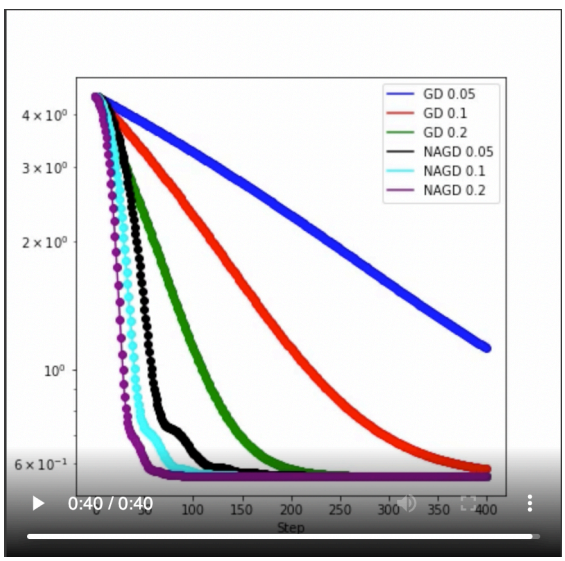
```
def sigmoid(x):
    return 1/(1 + np.exp(-x))

def logistic_cost(X, Y, w):
    n, d = X.shape
    a = sigmoid(X.dot(w))
    cost = - (1/n)* (Y.T @ (np.log(a)) + (1 - Y).T @ (np.log(1 - a)))
    return np.squeeze(cost)

def logistic_grad(X, Y, w):
    n, d = X.shape
    return (1/n) * X.T.dot(sigmoid(X.dot(w)) - Y)
```

ii. Value of f against the number of iterations.

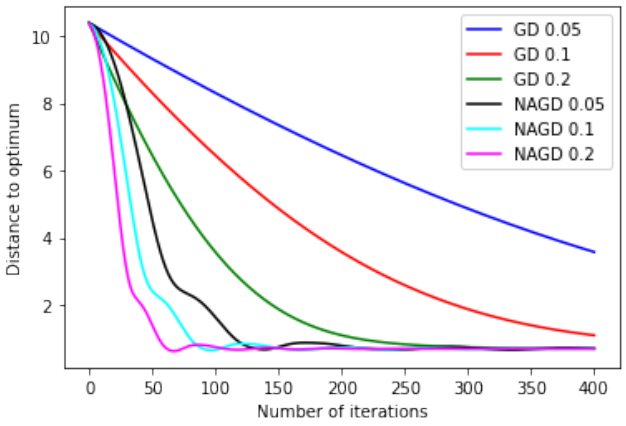
3 Step sizes are used 0.05, 0.1 and 0.2.



Same plot from the animation generated:



iii. The distance of the current iterate w_k to w_0 plotted:



Same plot from the animation generated: