Q1

(a) From the visualization, what do you think is the minimal value of this function and where does it occur?

> **Solution:** Since $(x-1)^2$ and $(y-3)^2$ are both always nonnegative, the minimum function value of $f(x, y)$ is attained when both are equal to zero. This occurs at $(1, 3)$, which is also where the gradient field shows the smallest vectors, in magnitude.

(b) Calculate the gradient $\nabla f = \begin{bmatrix} \frac{\partial f}{\partial x} & \frac{\partial f}{\partial y} \end{bmatrix}^T$.

> **Solution:**
>
> $$\begin{bmatrix} \frac{\partial f}{\partial x} & \frac{\partial f}{\partial y} \end{bmatrix}^T = \begin{bmatrix} 2(x-1) & 2(y-3) \end{bmatrix}^T$$

(c) When $\nabla f = \vec{0}$, what are the values of x and y?

> **Solution:**
>
> $$\nabla f = \vec{0} \implies 2(x-1) = 2(y-3) = 0 \implies x = 1,\ y = 3$$

Q2

Solution:

$$\theta^{t+1} \leftarrow \theta^t - \alpha \frac{\partial L}{\partial \theta}\bigg|_{\theta=\theta^t}$$

$$\frac{\partial L}{\partial \theta} = \frac{1}{n}\sum_{i=1}^{n} 2\theta x_i^2$$

Q3

> **Solution:** Yes, walking in a straight line between any two points on the graph will keep us at or above the graph.

Q4

(a) Suppose we have just one observation in our training data, $(x_1 = 1, x_2 = 2, y = 4)$. Assume that we set the learning rate α to 1. An incomplete version of the gradient descent update equation for θ is shown below. $\theta_0^{(t)}$ and $\theta_1^{(t)}$ denote the guesses for θ_0 and θ_1 at timestep t, respectively.

$$\begin{bmatrix} \theta_0^{(t+1)} \\ \theta_1^{(t+1)} \end{bmatrix} = \begin{bmatrix} \theta_0^{(t)} \\ \theta_1^{(t)} \end{bmatrix} - \begin{bmatrix} A \\ B \end{bmatrix}$$

Express both A and B in terms of $\theta_0^{(t)}$, $\theta_1^{(t)}$, and any necessary constants.

Solution: Note, our empirical risk here is $R(\theta) = 4 - \theta_0 \cdot 0.5 - \theta_0 \cdot \theta_1 - \sin(\theta_1) \cdot 2$. Taking partial derivatives with respect to θ_0 and θ_1 yield A and B respectively:

$$A = -0.5 - \theta_1^{(t)}$$

$$B = -\theta_0^{(t)} - 2\cos(\theta_1^{(t)})$$

(b) Assume we initialize both $\theta_0^{(0)}$ and $\theta_1^{(0)}$ to 0. Determine $\theta_0^{(1)}$ and $\theta_1^{(1)}$ (i.e. the guesses for θ_0 and θ_1 after one iteration of gradient descent).

Solution: From the above, we have

$$\theta_0^{(1)} = \theta_0^{(0)} - (-0.5 - \theta_1^{(0)}) = \theta_0^{(0)} + \theta_1^{(0)} + 0.5 = 0 + 0 + 0.5 = 0.5$$

$$\theta_1^{(1)} = \theta_1^{(0)} - (-\theta_0^{(0)} - 2\cos(\theta_1^{(0)})) = \theta_0^{(0)} + \theta_1^{(0)} + 2\cos(\theta_1^{(0)}) = 0 + 0 + 2\cos(0) = 2$$

(c)

Solution: $\theta_0^{(t)}$ approaches infinity. To see this, we can look at the update equations for $\theta_0^{(t)}$ and $\theta_1^{(t)}$:

$$\theta_0^{(t+1)} = \theta_0^{(t)} + \theta_1^{(t)} + 0.5$$

$$\theta_1^{(t+1)} = \theta_0^{(t)} + \theta_1^{(t)} + 2\cos(\theta_1^{(t)})$$

Loosely speaking, both $\theta_0^{(t)}$ and $\theta_1^{(t)}$ double on each iteration, and so they both increase towards positive infinity as we increase our number of iterations t.