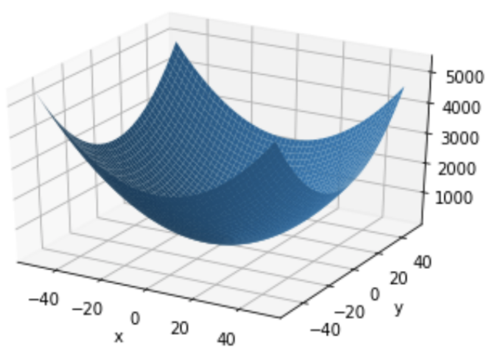
MSSE Data C200, Spring 2021 | March 10th, 2021

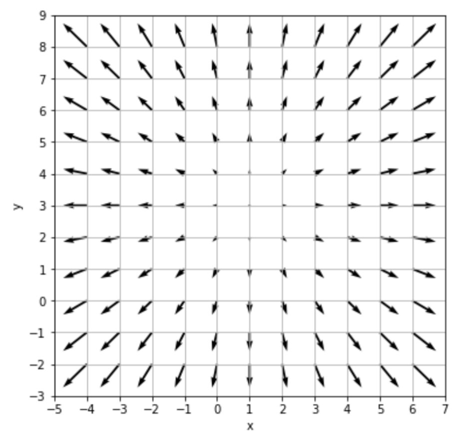
Discussion #10

Name:

Visualizing Gradients

1. On the left is a 3D plot of $f(x, y) = (x-1)^2 + (y-3)^2$. On the right is a plot of its **gradient field**. Note that the arrows show the relative magnitudes of the gradient vector.





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

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

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

Gradient Descent Algorithm

2. Given the following loss function and $\vec{x} = [x_i]_{i=1}^n$, $\vec{y} = [y_i]_{i=1}^n$, and θ^t, explicitly write out the update equation for θ^{t+1} in terms of x_i, y_i, θ^t, and α, where α is the constant learning rate.

$$L(\theta, \vec{x}, \vec{y}) = \frac{1}{n}\sum_{i=1}^{n}\left(\theta^2 x_i^2 - \log(y_i)\right)$$

Convexity

3. Convexity allows optimization problems to be solved more efficiently and for global optimums to be realized. Mainly, it gives us a nice way to minimize loss (i.e. gradient descent). There are three ways to informally define convexity.

   a. Walking in a straight line between points on the function keeps you at or above the function. This works for any function.

   b. The tangent line at any point lies at or below the function, globally. To use this definition, the function must be differentiable.

   c. The second derivative is non-negative everywhere (in other words, the function is ”concave up” everywhere). To use this definition, the function must be twice differentiable.

   Is the function described in Question 1 convex? Make an argument visually.

GPA Descent

4. Consider the following non-linear model with two parameters:

$$f_\theta(x) = \theta_0 \cdot 0.5 + \theta_0 \cdot \theta_1 \cdot x_1 + \sin(\theta_1) \cdot x_2$$

For some nonsensical reason, we decide to use the residuals of our model as the loss function. That is, the loss for a single observation is

$$L(\theta) = y_i - f_\theta(x_i)$$

We want to use gradient descent to determine the optimal model parameters, $\hat{\theta}_0$ and $\hat{\theta}_1$.

(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.

(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).

(c) What happens to $\theta_0^{(t)}$ as $t \to \infty$ (i.e. as we run more and more iterations of gradient descent)?