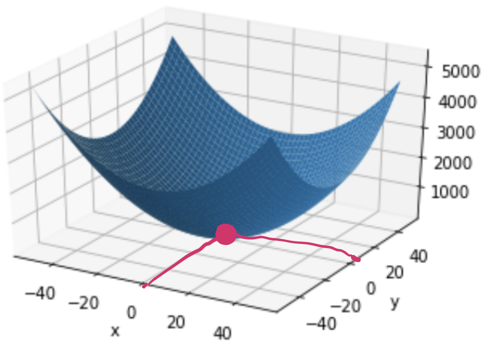
DS 100/200: Principles and Techniques of Data Science **Date: March 6, 2020**

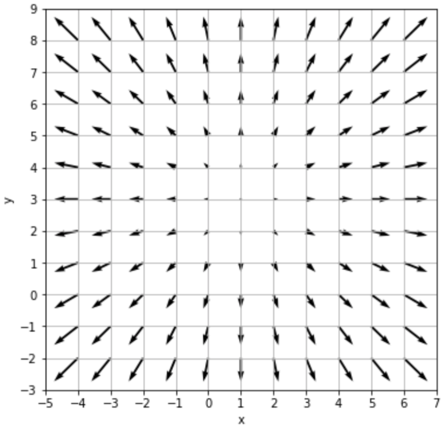
Discussion #7

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?

$f(x,y) = (x-1)^2 + (y-3)^2$

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

$\frac{\partial f}{\partial x} = 2(x-1) + 0 = 2(x-1)$

$\frac{\partial f}{\partial y} = 2(y-3)$

$\boxed{\nabla f = \begin{bmatrix} 2(x-1) \\ 2(y-3) \end{bmatrix}}$

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

$\boxed{\begin{matrix} x = 1 \\ y = 3 \end{matrix}}$

Gradient Descent Algorithm

2. Given the following loss function and $\mathbf{x} = (x_i)_{i=1}^n$, $\mathbf{y} = (y_i)_{i=1}^n$, β^t, explicitly write out the update equation for β^{t+1} in terms of x_i, y_i, β^t, and α, where α is the constant step size.

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

$$\nabla_\beta L(\beta, x, y) = \frac{\partial L}{\partial \beta} L(\beta, x, y) = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial \beta}\left[(\beta^2 x_i^2 - \log(y_i))\right]$$

$$= \frac{1}{n}\sum_{i=1}^{n} 2\beta x_i^2 + \beta^2(0)$$

$$= \frac{1}{n}\sum_{i=1}^{n} 2\beta x_i^2$$

$$\beta^{(t+1)} = \beta^{(t)} - \alpha\left[\frac{1}{n}\sum_{i=1}^{n} 2\beta^{(t)} x_i^2\right]$$

3. (a) The learning rate α can *potentially* affect which of the following? Select all that apply. Assume nothing about the function being minimized other than that its gradient exists. You may assume the learning rate is positive.

   ☒ A. The speed at which we converge to a minimum.

   ☒ B. Whether gradient descent converges.

   □ C. The direction in which the step is taken. ← gradient controls this

   ☒ D. Whether gradient descent converges to a local minimum or a global minimum.

   (see picture at the end of the solutions)

   (b) Suppose we run gradient descent with a fixed learning rate of $\alpha = 0.1$ to minimize the 2D function $f(x, y) = 5 + x^2 + y^2 + 5xy$.

   The gradient of this function is

$$\nabla_{x,y} f(x, y) = \begin{bmatrix} 2x + 5y \\ 2y + 5x \end{bmatrix}$$

   If our starting guess is $x^{(0)} = 1, y^{(0)} = 2$, what will be our next guess $x^{(1)}, y^{(1)}$?

$$\begin{bmatrix} x^{(1)} \\ y^{(1)} \end{bmatrix} = \begin{bmatrix} x^{(0)} \\ y^{(0)} \end{bmatrix} - \alpha \nabla f(x, y) \Big|_{x = x^{(0)},\ y = y^{(0)}}$$

$x^{(1)} =$ $\boxed{-0.2}$ $y^{(1)} =$ $\boxed{1.1}$

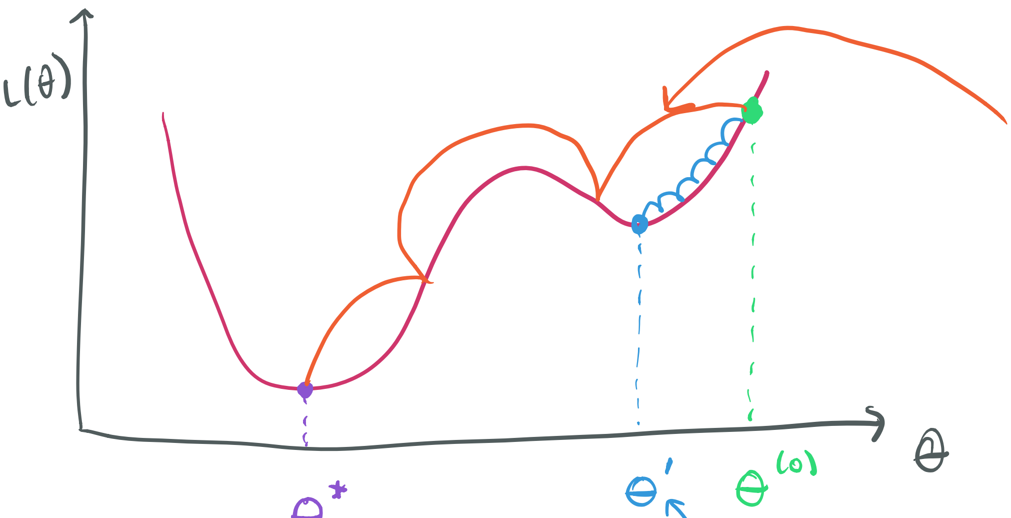
$$\begin{bmatrix} x^{(0)} \\ y^{(0)} \end{bmatrix} - \alpha \nabla f(x, y) = \begin{bmatrix} 1 \\ 2 \end{bmatrix} - (0.1)\begin{bmatrix} 2(1) + 5(2) \\ 2(2) + 5(1) \end{bmatrix} = \begin{bmatrix} -0.2 \\ 1.1 \end{bmatrix}$$

(c) Suppose we are performing gradient descent to minimize the empirical risk of a linear regression model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2$ on a dataset with 100 observations. Let $\mathcal{D}$ be the number of components in the gradient, e.g. $\mathcal{D} = 2$ for the equation in part b. What is $\mathcal{D}$ for the gradient used to optimize this linear regression model?

○ A. 2 ○ B. 3 ● C. 4 ○ D. 8 ○ E. 100 ○ F. 200 ○ G. 300
○ H. 400 ○ I. 800

There are 4 parameters, so the gradient will have 4 components.

How α affects converging to local or global minimum



converge to θ^* when α is big since your big steps will overshoot the local minimum θ'

want to converge to this

starting here

converge to this when α is small (small steps)