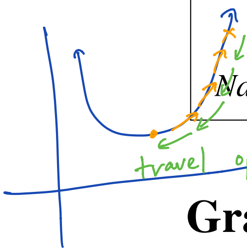
DS 100: Principles and Techniques of Data Science	**Date: September 26, 2018**

Discussion #5

Name: Armaan (the star student)



travel opposite to the steepest ascent

$(x-1)^2 + (y-3)^2 = K$ is equation of circle with radius $\sqrt{K}$

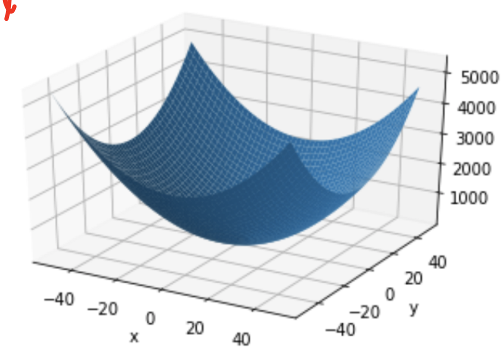
Gradients

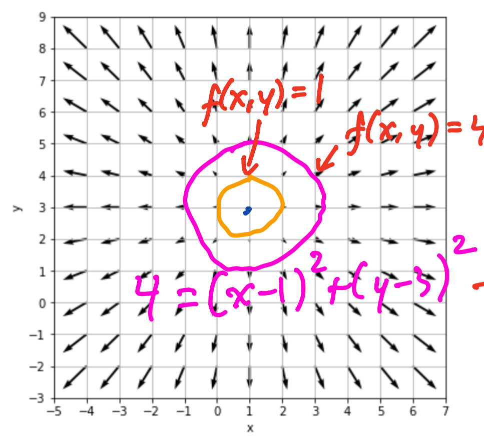
1. On the left is a 3D plot of $f(x, y) = (x-1)^2 + (y-3)^2$. On the right is a plot of its gradient field. Note that the arrows show the relative magnitudes of the gradient vector.

single var

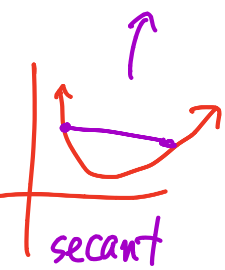
$f(x) = x^3 + 4x$

$f'(x) = 3x^2 + 4$





convex!



secant is always above $f(x)$

yes! (a) Is this function convex? Make a visual argument—it doesn't have to be formal.

(b) Superimpose a contour plot of this function for $f(x, y) = 0, 1, 2, 3, 4, 5$ onto the gradient field.

(c) What do you notice about the relationship between the level curves and the gradient vectors?

magnitudes and angles increase

(d) In areas where the contour lines are close together, the function values are

○ Slowly changing ⊗ Quickly changing

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

(1, 3)

(f) Calculate the gradient $\nabla f = \left[\frac{\partial f}{\partial x} \quad \frac{\partial f}{\partial y}\right]^T$.

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

$\nabla f = 0$

$\rightarrow 2(x-1) = 0$

$2(y-3) = 0$

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

(h) If you started at a random point on the surface generated by this function, which direction would you want to go relative to the gradient field to reach the minimum of the function?

$\nabla f = \left[2(x-1) \quad 2(y-3)\right]^T$

row vec, so I transposed

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

$\frac{\partial f}{\partial x} = 2(x-1) \quad \frac{\partial f}{\partial y} = 2(y-3)$

2. In this question, we will explore some basic properties of the gradient.

   Note: In this class, we use the following conventions:

   - x represents a scalar
   - X represents a random variable
   - $\mathbf{x}$ represents a vector
   - $\mathbf{X}$ represents a matrix or a random vector (context will tell)

   (a) Determine the derivative of $f(x) = a_0 + a_1 x$ and gradient of $g(x_1, x_2) = a_0 + a_1 x_1 + a_2 x_2$.

   $$\frac{d}{dx} f(x) = a_1 \qquad \nabla g(x) = \begin{bmatrix} \frac{\partial g}{\partial x_1} \\ \frac{\partial g}{\partial x_2} \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}$$

   (b) Suppose $\mathbf{x} = \begin{bmatrix} x_1 & x_2 & \dots & x_n \end{bmatrix}^T$, and $h(\mathbf{x}) = \mathbf{a}^T\mathbf{x}$, where $\mathbf{a}, \mathbf{x} \in \mathbb{R}^n$. Determine ∇h.

   $$a = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \qquad x = \begin{bmatrix} 3 \\ 4 \end{bmatrix}$$

   $$a^T x = \begin{array}{c} \begin{matrix} a_1 & a_2 \end{matrix} \\ \begin{bmatrix} 1 & 2 \end{bmatrix} \end{array} \begin{bmatrix} 3 \\ 4 \end{bmatrix} = \underbrace{1 \cdot 3}_{a_1 x_1} + \underbrace{2 \cdot 4}_{a_2 x_2} = 11$$

   $$h(x) = a_1 x_1 + a_2 x_2 + \dots + a_n x_n$$

   $$\nabla h(x) = \begin{bmatrix} a_1 & a_2 & \dots & \dots & a_n \end{bmatrix}^T = \boxed{a}$$

   (c) Determine the gradient of $f(\mathbf{x}) = \mathbf{x}^T\mathbf{x}$. *(Hint: f is a scalar-valued function. How can you write $\mathbf{x}^T\mathbf{x}$ as a sum of scalars?)*

   $$x^T x = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = x_1^2 + x_2^2 + \dots + x_n^2$$

   $$\frac{\partial x^T x}{x_1} = 2x_1, \quad \frac{\partial x^T x}{x_2} = 2x_2$$

   $$\nabla x^T x = \begin{bmatrix} 2x_1 \\ 2x_2 \\ \vdots \\ 2x_n \end{bmatrix} = \boxed{2x}$$

   (d) Suppose $\mathbf{A} \in \mathbb{R}^{n \times n}$. It is a fact that $\nabla \mathbf{x}^T\mathbf{A}\mathbf{x} = (\mathbf{A} + \mathbf{A}^T)\mathbf{x}$. Show that this formula holds even when $\mathbf{A}, \mathbf{x}$ are scalars. (Why?)

   $$\underbrace{\begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix}}_{x^T} \underbrace{\begin{bmatrix} | & | & | & | \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}}_{x}$$

   $$= \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} \ \end{bmatrix}$$

   = scalar!

   scalar

   $$x^T A x \longrightarrow A x^2$$

   $$\frac{d\,Ax^2}{dx} = 2Ax$$

   $$(A + A^T)x = (A + A)x = 2Ax$$

Loss Minimization

3. Consider the following loss function:

$$L(\theta, x) = \begin{cases} 4(\theta - x) & \theta \geq x \\ x - \theta & \theta < x \end{cases}$$

Given a sample of $x_1, ..., x_n$, find the optimal θ that minimizes the the average loss.

- Take partial derivative w.r.t. θ, set to 0, optimize

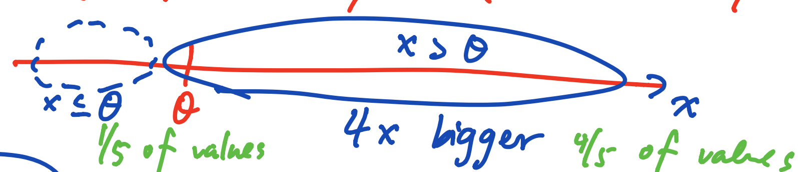
$$\frac{\partial L}{\partial \theta} = \begin{cases} 4 & \theta \geq x \\ -1 & \theta < x \end{cases}$$

$$L = L(\theta, x_1) + L(\theta, x_2) + \dots + L(\theta, x_n)$$

$$\frac{\partial L}{\partial \theta} = \frac{\partial L}{\partial \theta}(x_1) + \frac{\partial L}{\partial \theta}(x_2) + \dots + \frac{\partial L}{\partial \theta}(x_n) = 0$$

each of these is equal to 4 or -1

$$4(\# \text{ of } x \leq \theta) = 1(\# \text{ of } x > \theta)$$



$\therefore \theta_{optimal} = 20\%$ ile

$$L(\theta, x) = (\theta - x_1)^2 + (\theta - x_2)^2 + \dots + (\theta - x_n)^2$$

$$= \sum_{i=1}^{n} (\theta - x_i)^2$$

$$\frac{\partial L}{\partial \theta} = \cancel{2}(\theta - x_1) + \cancel{2}(\theta - x_2) + \dots + \cancel{2}(\theta - x_n) = 0$$

$$(\theta - x_1) + (\theta - x_2) + \dots + (\theta - x_n) = 0$$

$$n\theta = x_1 + x_2 + \dots + x_n$$

$$\boxed{\theta = \frac{x_1 + x_2 + \dots + x_n}{n}}$$

Gradient Descent Algorithm

4. Given the following loss function and $\mathbf{x} = (x_i)_{i=1}^n$, $\mathbf{y} = (y_i)_{i=1}^n$, 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 step size.

move in opposite direction of gradient

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

In general:

$$\theta^{t+1} = \theta^t - \alpha \frac{\partial L}{\partial \theta}(\theta^t)$$

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

$$\rightarrow \boxed{\theta^{t+1} = \theta^t - \alpha \cdot \frac{1}{n} \sum_{i=1}^{n} 2\theta_t x_i^2}$$

iterative

θ^t: estimate at step t

5. (a) In your own words, describe how to use the update equation in the gradient descent algorithm.

   (b) Say that x and y are your model parameters and f as defined in question 1 is your loss function. Describe in your own words what happens "visually" as the gradient descent algorithm runs.