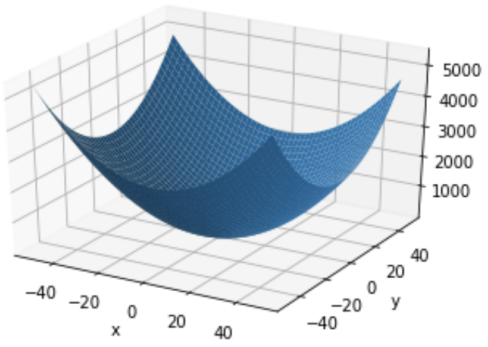
DS 100: Principles and Techniques of Data Science **Date: September 26, 2018**

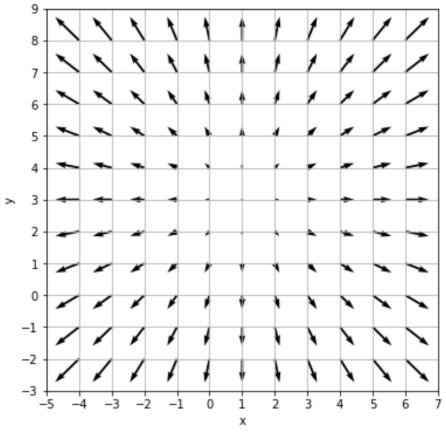
Discussion #5 Solutions

Name:

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.



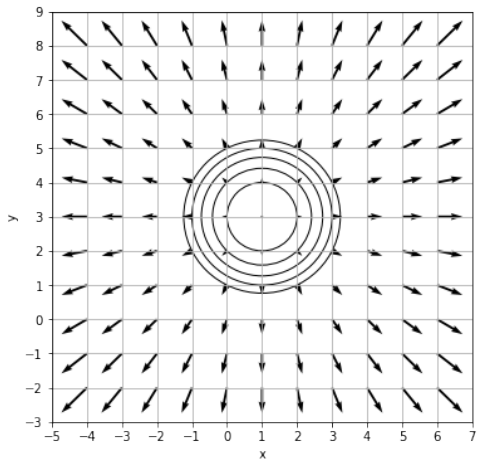


(a) Is this function convex? Make a visual argument—it doesn't have to be formal.

> **Solution:** Yes. From looking at the surface, connecting any two points remains above the surface. Additionally, the second derivative is positive in x and y.

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

> **Solution:** The contour plots are concentric circles centered at $(1, 3)$ with radii of $0, \sqrt{1}, \sqrt{2}, \sqrt{3}, \sqrt{4}, \sqrt{5}$.



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

Solution: The gradient vectors increase in magnitude as the level curves increase. Additionally, the gradient vectors always lie perpendicular to the level curves, as they represent the direction in which the function curves away from each level set.

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

◯ Slowly changing √ Quickly changing

Solution: Quickly changing. The change in function value between contour lines is constant, but with contour lines closer together, the constant difference occurs over a shorter area. This corresponds with a larger gradient, meaning the function values are quickly changing.

(e) 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 nonnegative, the minimum function value of $f(x,y)$ is attained when both are equal to zero. This occurs at $(1,3)$ where the gradient field shows the smallest (in magnitude) vectors.

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

Solution:

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

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

Solution:

$$\nabla f = \mathbf{0} \implies 2(x-1) = 2(y-3) = 0 \implies x = 1,\ y = 3.$$

If the gradient is equal to zero, then the function must be at a local minima. The only minima in this case is the global minima, meaning it must be at $(1, 3)$, due to part (e).

(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?

Solution: You would want to go opposite of the direction of the gradient field. This also agrees with the notion of gradient descent, opposing the direction of steepest ascent is the direction of steepest descent.

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 $g(x_1, x_2) = a_0 + a_1 x_1 + a_2 x_2$.

   Solution:

   $$\frac{df}{dx} = a_1$$

   $$\nabla g = \begin{bmatrix} \frac{\partial g}{\partial x_1} & \frac{\partial g}{\partial x_2} \end{bmatrix}^T = \begin{bmatrix} a_1 & a_2 \end{bmatrix}^T$$

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

   Solution: Note that $h(\mathbf{x}) = \mathbf{a}^T\mathbf{x}$ is just a concise way of writing

   $$h(\mathbf{x}) = \sum_{i=1}^{n} a_i x_i = a_1 x_1 + a_2 x_2 + ... + a_n x_n$$

   So as in (a), we have

   $$\nabla h = \begin{bmatrix} \frac{\partial h}{\partial x_1} & \frac{\partial h}{\partial x_2} & ... & \frac{\partial h}{\partial x_n} \end{bmatrix}^T = \begin{bmatrix} a_1 & a_2 & ... & a_n \end{bmatrix}^T = \mathbf{a}$$

   (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?)*

   Solution: $f(\mathbf{x})$ can also be expanded as $\sum_{i=1}^{n} = x_i^2 = x_1^2 + x_2^2 + ... + x_n^2$

   $$\nabla f = \begin{bmatrix} 2x_1 & 2x_2 & ... & 2x_n \end{bmatrix}^T = 2\mathbf{x}$$

   (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?)

   Solution: For the LHS, $\nabla \mathbf{x}^T\mathbf{A}\mathbf{x} = \frac{d}{dx} xAx = \frac{d}{dx} Ax^2$ since the transpose of a scalar is itself and we gain commutativity of A and x.
   For the RHS, $(\mathbf{A} + \mathbf{A}^T)\mathbf{x} = (A + A)x = 2Ax$ for the same reasons as before.

This is the statement we are used to in univariate calculus: $\frac{d}{dx}Ax^2 = 2Ax$.

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.

Solution:

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

$$\sum_{i=1}^{n} \frac{\partial L(\theta, x_i)}{\partial \theta} = \sum_{\theta < x_i} -1 + \sum_{\theta \geq x_i} 4 = 0 \implies \#\{x_i > \theta\} = 4\#\{x_i \leq \theta\}$$

We know that

$$\#\{x_i > \theta\} + \#\{x_i \leq \theta\} = n$$

Which implies that

$$n - \#\{x_i \leq \theta\} = 4\#\{x_i \leq \theta\} \implies 0.20n = \#\{x_i \leq \theta\}$$

Thus, θ is the 20th percentile of $x_1, ..., x_n$

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.

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

Solution:

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

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

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.