M462-562-Homework 3: written part

Due: February 28 (Tuesday).

Problems:

1. Convex functions are essential in data analysis because they have desirable mathematical properties that allow us to efficiently minimize them using optimization methods such as gradient descent. One of the key properties of a convex function is that any local minimum of the function is also a global minimum. This is a highly desirable property for optimization, as it ensures that the optimization algorithm will always converge to the optimal solution.

   A function $f : \mathbb{R}^n \to \mathbb{R}$ is **convex** if for any vectors $x_1, x_2 \in \mathbb{R}^n$ and scalar $t \in (0, 1)$, we have

   $$tf(x_1) + (1-t)f(x_2) \geq f(tx_1 + (1-t)x_2).$$

   In words, a function is convex when its curve lies below any chord joining two of its points. (See this this picture).

   Your goal is to show that the function

   $$f(\theta) = \|y - X\theta\|^2,$$

   in a least squares problem, is a convex function. This is, you are going to show that for any vectors θ_1 and θ_2, and for any scalar t with $0 \leq t \leq 1$, the inequality

   $$tf(\theta_1) + (1-t)f(\theta_2) \geq f(t\theta_1 + (1-t)\theta_2).$$

   holds

   Step 1. Show that

   $$f(\theta) = \|y\|^2 - 2y^T X\theta + \theta^T X^T X\theta.$$

   Step 2. Show that, for any two vectors θ_1, θ_2 and scalar t, we have

   $$f(t\theta_1+(1-t)\theta_2)-(tf(\theta_1) + (1-t)f(\theta_2)) = -t(1-t)\|X(\theta_1-\theta_2)\|^2.$$

   Step 3. Conclude that the function $f(\theta)$ is convex.

2. Consider the function

   $$f\left(\begin{bmatrix} x \\ y \end{bmatrix}\right) = x^2 + by^2 \qquad \text{with } b < 1,$$

and the gradient descent iteration

$$\begin{bmatrix} x_k \\ y_k \end{bmatrix} = \begin{bmatrix} x_{k-1} \\ y_{k-1} \end{bmatrix} - s\nabla f\left(\begin{bmatrix} x_{k-1} \\ y_{k-1} \end{bmatrix}\right), \qquad \text{for } k = 1, 2, \ldots,$$

where $s > 0$ is the learning rate.

Part 1. Starting at $\begin{bmatrix} x_0 \\ y_0 \end{bmatrix} = \begin{bmatrix} b \\ 1 \end{bmatrix}$, find a formula for $\begin{bmatrix} x_k \\ y_k \end{bmatrix}$.

Part 2. For what values of the learning rate s does gradient descent converge to the minimum of f.

Part 3. For what values of the learning rate s does gradient descent approaches the minimum in a zig-zag path.