6. Optimization and gradients

Gradient ascent/descent methods are typical tools for maximizing/minimizing functions. Consider the function $L(x, \theta)$, where $x = [x_1, x_2]^T$ and $\theta = [\theta_1, \theta_2]^T$. We want to select θ such that we maximize/minimize the value of L.

6. (a)

1/1 point (graded)

The gradient $\nabla_\theta L(x, \theta)$ is a vector with two components:

$$\frac{\partial}{\partial \theta_j} L(x, \theta), j = 1, 2.$$

Let $L(x, \theta) = \log(1 + \exp(-\theta \cdot x))$. Evaluate the gradient. Which of the following is its j^{th} component?

- $\dfrac{\exp(-\theta \cdot x)}{1 + \exp(-\theta \cdot x)}$
- $\dfrac{-x_j \exp(-\theta \cdot x)}{1 + \exp(-\theta \cdot x)}$ ✔
- $\dfrac{-x_j}{1 + \exp(-\theta \cdot x)}$

Note on notation: In this course, we will sometimes abuse notation and use x_j to mean the **vector** whose j^{th} component is x_j (roughly, "x_j for the whole range of j").

STANDARD NOTATION

Solution:

The derivative of $\log(x) = \frac{1}{x}$ and the derivative of $e^{cx} = ce^{cx}$. Applying these rules with the chain rule gives the correct answer.

Submit You have used 1 of 1 attempt

Answers are displayed within the problem

6. (b)

0/1 point (graded)

The direction of the derivative of a function gives us the direction of the change in the function with changes in its variables. Under stochastic gradient ascent/descent methods, we make an educated guess about the next values of the variables to try. This corresponds to intelligently choosing values for θ in $L(x, \theta)$. Given $\theta' = \theta + \epsilon \cdot \nabla_\theta L(x, \theta)$, where ϵ is a small positive real number, is the value of $L(x, \theta')$ greater or smaller than the value of $L(x, \theta)$?

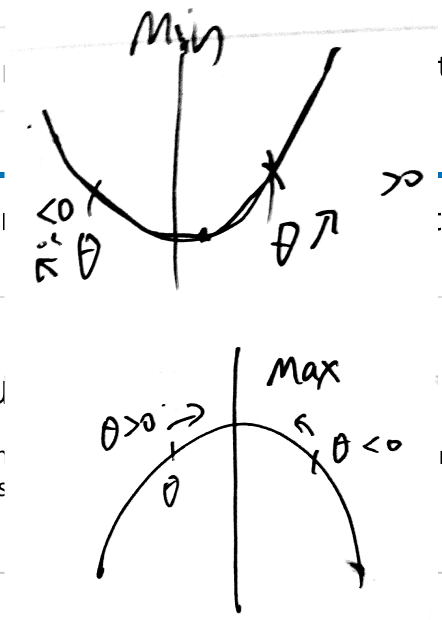
- greater ✔
- smaller ✖

STANDARD NOTATION

Solution:

Consider the one-dimensional case. If the gradient is positive, we obtain θ' by moving from θ in the positive direction. This increases $L(x, \theta)$. If the gradient is negative, we move in the negative direction, again increasing $L(x, \theta)$. This analysis extends to higher dimensions. Note that if we used the function above to continue updating θ, we would (in theory) maximize $L(x, \theta)$. Alternatively if our update rule was $\theta' = \theta - \epsilon \cdot \nabla_\theta L(x, \theta)$, we would minimize the function. There are more complications in higher dimensions, but this is the basic idea behind stochastic gradient descent, which forms the backbone of modern machine learning.

Sub tempt



A he problem

Discu

Show Discussion

Topic: Ur roject 0 (1 week):Homework 0 / 6. Optimization and gradients