Explanatory Notes for 6.390

Shaunticlair Ruiz (Current TA)

Fall 2022

Gradient Descent in n-D

This idea can be built up in **any number** of dimensions: each variable θ_k creates a **different** line we can use to **approximate**.

And, we can combine them into a **flat** hyperplane: so, we can **add up** all of the different **derivatives**.

Key Equation 1

In **n-D**, you can optimize your function J using this rule:

$$\theta_{new} = \theta_{old} - \eta \underbrace{\begin{bmatrix} \partial J/\partial \theta_1 \\ \partial J/\partial \theta_2 \\ \vdots \\ \partial J/\partial \theta_d \end{bmatrix}}_{\text{Using } \theta_{old}}$$

This is our **generalized gradient descent** rule.

The Gradient

We call this **gradient** descent because that right term **is** the gradient!

Definition 2

The gradient can be written as

$$\nabla_\theta J = \begin{bmatrix} \partial J/\partial \theta_1 \\ \partial J/\partial \theta_2 \\ \vdots \\ \partial J/\partial \theta_d \end{bmatrix} = \frac{dJ}{d\theta}$$

So, our rule can be rewritten (for the last time) as:

Key Equation 3

The **gradient descent** rule can be generally written as:

$$\theta_{new} = \theta_{old} - \eta \nabla_\theta J(\theta_{old})$$

θ_{old} is the input to $\nabla_\theta J$, not multiplication!

$$\theta_{new} = \theta_{old} - \eta J'(\theta_{old})$$

In fact, the gradient is the best direction we could choose!

Check one of the other "topics" documents for a proof, or just check the full explanatory notes!

Concept 4

The **gradient** ∇J is the **direction of greatest increase** for J.

That means means the opposite direction $-\nabla J$ is the **direction of greatest decrease** in J.

This is the single **most important concept** in this entire chapter!