Explanatory Notes for 6.390

Shaunticlair Ruiz (Current TA)

Fall 2022

Our procedure

So, we have our parameter **update**, $\Delta\theta$. We'll start at $t = 0$.

Before, we represented the i^{th} **data point** with $x^{(i)}$. We'll reuse this **notation**.

Notation 1

Here, we're changing θ over **time**: each step happens at $t = \{1, 2, 3, \ldots\}$ so we need **notation** for that.

We'll **reuse** the notation from $x^{(i)}$, for the i^{th} data point.

In this case, we'll do $\theta^{(t)}$: the value of θ after t **steps** are taken.

Earlier, we **introduced** θ_{old} and θ_{new}: these are $\theta^{(t-1)}$ and $\theta^{(t)}$.

Example: After **10 steps** of 1-D gradient descent, we have gone from $\theta^{(0)}$ to $\theta^{(10)}$.

So, we move the **first** time using $J'(\theta^{(0)})$.

Once we've moved in parameter space **one** time, though, our **derivative** has changed: we're in a different part of the **surface**.

So, we'll take a **second** step with a **new** derivative, $J'(\theta^{(1)})$.

We want to do this **repeatedly**. We'll take our equation

$$\theta_{new} = \theta_{old} + \Delta\theta \tag{1}$$

And combine it with our **chosen** step size.

Key Equation 2

In **1-D**, **Gradient Descent** is implemented as follows:

At each time step t, we **improve** our hypothesis θ using the following rule:

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

Using $\theta^{(t)}$ notation:

$$\theta^{(t)} = \theta^{(t-1)} - \eta J'(\theta^{(t-1)})$$

We repeat until we reach whatever our chosen **termination condition** is.

We can also write it as:

$$\theta_{new} = \theta_{old} - \eta \left(\frac{dJ}{d\theta_{old}} \right)$$

We've got our gradient descent **update** rule in 1-D!

Termination Conditions

When do we **stop**? We can't let it run forever.

We have some options:

- Stop after a **fixed** T steps.
  - This has the advantage of being **simple**, but how do you know what the **correct** number of steps is?
- Stop when θ **isn't changing** much: $\left|\Delta\theta\right| < \epsilon$, for example.
  - If our θ isn't changing much, our algorithm isn't **improving** our hypothesis much. So, it makes sense to stop: we've stabilized.
- Stop when the **derivative is small**: $\left|J'(\theta)\right| < \epsilon$.
  - Mathematically **equivalent** to our last choice. But a different **perspective**: if the slope is small, our surface is relatively **flat**, and we're near a **minimum** (probably).
  - "The derivative is **small**" is weaker, but in the same spirit as "the derivative is **zero**", $J'(\theta) = 0$, from last chapter.