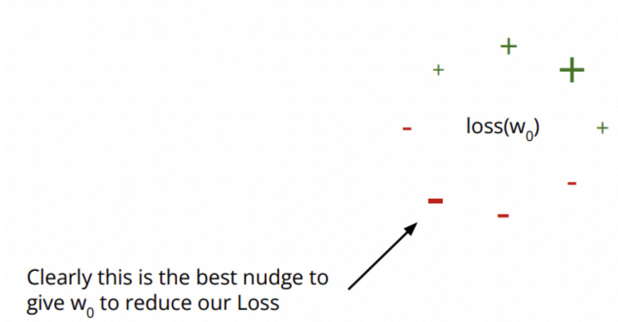
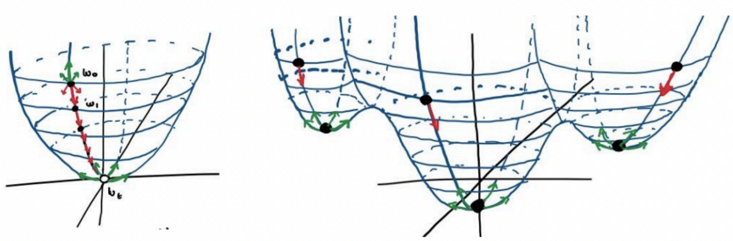
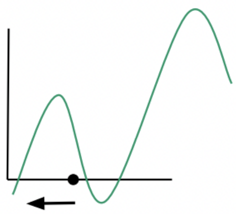
CAS CS 506
Lec 20

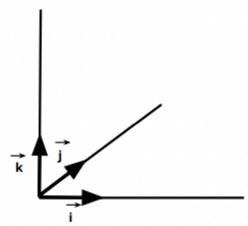
Gradient Descent

1. Gradient Descent (Intuition)
    a. Optimization method when there is no closed form solution to finding the extrema of a function
    b. What to do if we don't have an optimization method?
    c. Example: Logistic Regression
    d. Goal: find a sequence of w_i's and b's that converge toward a minimum
    e. Consider a random weight w_0. What happens to Loss(w_0) as you nudge w_0 slightly?
    f. 

    g. As such we can define the following sequence:
        i. w_1 = best nudge to w_0
        ii. w_2 = best nudge to w_1
        iii. …
        iv. Until we reach w_t that looks like
        D
        v. At this point we can stop updating w
    h. 
2. Gradients
    a. Intuitively the best nudge should be in the direction of the largest rate of change (steepness) of the function
    b. Rate of change → think derivatives

c. 

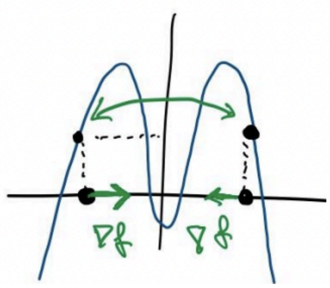
d. Intuitively, the rate of change of a multi-dimensional function should be a combination of the rate change in each dimension. For a 3-dimensional function, the rate of change would be

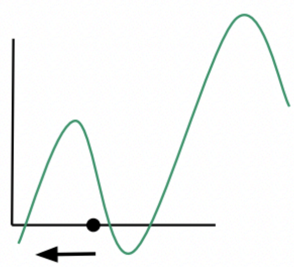


$$\nabla f(x, y, z) = \frac{\partial f}{\partial x}\vec{i} + \frac{\partial f}{\partial y}\vec{j} + \frac{\partial f}{\partial z}\vec{k}$$

e. However, the gradient expresses the instantaneous rate of change. At p, change in f_p is the steepest but the highest value of f will depend on how many units we step in that direction. If we step too many units away, the instantaneous change in f is no longer representative of what values f will take

f. Example



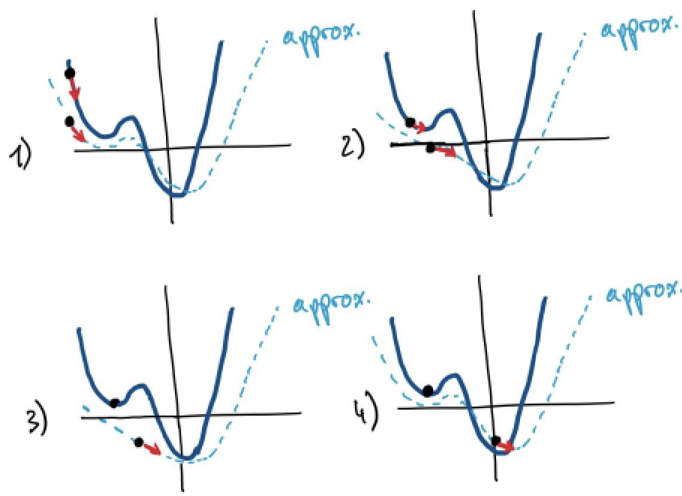


g. Given a smooth function f for which there exists no closed form solution for finding its maximum, we can find a local maximum through the following steps:
   i. Define a step size a (tuning parameter)
   ii. Initialize p to be random
   iii. p_{new} = a * change in f_p + p
   iv. p - p_{new}
   v. Repeat 3 & 4 until p ~ p_{new}

h. To find a local minimum, just use -change in f_p

i. Notes about a:
   i. If a is too large, GD may overshoot the maximum, take a long time to or never be able to converge
   ii. If a is too small, GD may take too long to converge

j. Stochastic Gradient Descent
   i. Recall the cost is computed for the entire dataset. This has some limitations:
      1. Expensive to run
      2. Result we get depends only on the initial starting point
   ii. Goal: Approximate the gradient of the cost using a sample of the data (batch)



3. Note
   a. The magnitude of change in f_p depends on p. As p gets closer to the min/max, the size of change in f_p decreases
   b. This also means that points p that contain more information have larger gradients. So the order with which this process is exposed to examples matters