gradient descent

can optimize many different types of lines and other shapes. Y axis is the loss function the x axis is the intercept. The optimal value is found by taking big steps when fat away from the minimum and smaller steps as the minimum value is approached. This method is effective when finding the minimum value where the derivate is not 0. The step size is changed based on how close the curve is to zero.

New intercept = old intercept - step size

stochastic gradient descent

takes one sample for each step and uses that sample to calculate the derivative. This reduces the numbers calculated by a factor of the number of steps so an example with 3 data points runs 3 times faster than it would using gradient descent. It is even more efficient with clusters because it calculates just one point from each cluster instead of every point.

Learning rate – the fraction used to adjust our weights and biases