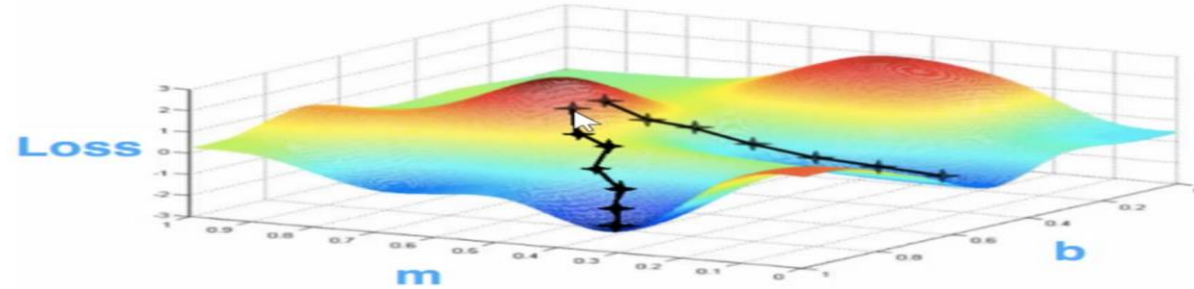


batch_size

Gradient Descent

$f(x)$ = nonlinear function of x



- ✓ Stochastic = random
- ✓ True Gradient Descent: Draw a single sample and corresponding target at each iteration. Very slow
- ✓ Batch Gradient Descent: Draw all samples and corresponding targets in each iteration. Accurate but expensive
- ✓ Mini-Batch Stochastic Gradient Descent: Randomly draw a batch of training samples and corresponding targets. This is an effective middle path
- ✓ batch_size value indicates the number of samples and corresponding targets to be drawn at random in each training iteration (also called epoch)
- ✓ After each iteration of "batch_size" number of samples backpropagation starts
- ✓ Given backpropagation is computationally expensive, increasing batch_size value would imply faster training speed due of lower number of backpropagations
- ✓ However, larger batch_size value increases the probability of getting stuck in a local minima
- ✓ Standard values of batch_size: 32, 64, 128 i.e. 2^5 , 2^6 , 2^7
- ✓ Many vectorized operation implementations work faster when inputs are sized in powers of 2