HW to Chapter 11 “Learning Rates Decay and Hyperparameters”

Non-programming Assignment

1. What is learning rate decay and why is it needed?

Learning rate decay is a technique used to reduce the learning rate gradually as the training progresses. This is important because while a large learning rate might speed up the initial training phase, it can also cause overshooting, leading to oscillations around the minimum of the loss function. By reducing the learning rate over time, we ensure more precise convergence, especially in later stages of training when the network is approaching the optimal parameters.

1. What are saddle and plateau problems?

A **saddle point** is a region in the loss surface where gradients are very small but not necessarily at the minimum. This can make it hard for gradient descent to make progress.

A plateau refers to regions of the loss surface that are flat, meaning the gradients are very small or zero, causing the optimization process to stall or move very slowly.

1. Why should we avoid grid approach in hyperparameter choice?

The grid approach systematically searches a predefined set of hyperparameters. However, this method can be inefficient because the number of combinations grows exponentially as more hyperparameters are added. Instead, a random search approach is often preferred because it explores the parameter space more broadly and can find good combinations faster.

1. What is mini batch and how is it used?

A mini-batch is a subset of the training data used to compute the gradient and update the weights during training. This allows for more frequent updates compared to full-batch gradient descent, without the noise introduced by stochastic gradient descent. Mini-batch gradient descent strikes a balance between efficiency and performance.