CSCI699: Theory of Machine Learning Fall 2023

Problem Set 3 (Empirical portion)

Due: November 29 by 11:59 pm

Discussion is allowed and encouraged but everyone should write solutions on their own. Please also mention any collaborators you had substantial discussions with. You are also allowed to consult general resources on the internet (such as one of the books, or other lecture notes online), but you should not search for any the solutions themselves online.

If you use the LaTeX template, then please only keep your answers and remove the questions before submitting. Homeworks should be written in Latex and submitted via Gradescope. When you submit on Gradescope, make sure to mark the page which contains each answer.

Problem 4: Training and Generalization in the Overparameterized Regime

(14 pts) In class, we saw some theoretical perspectives for understanding neural networks. We also saw that there are still many mysteries about how neural networks work. Investigating empirical phenomenon has become a powerful tool both for uncovering such mysteries and guiding the search for plausible hypothesis about how neural networks work. We will undertake some basic empirical investigations in this question.

The goal of the problem is to investigate some perhaps surprising phenomenon related to optimization and generalization in the overparameterized regime and the effect of label noise. You can use this Colab notebook for the experiments.

Each question describes a setting, and you are generally asked about your observations, relate them to some concepts you have seen in class, and tinker with some parameters to observe their effect.

For parts (a)-(c), you will train a MLP with 1 hidden layer with 390 units on the MNIST dataset, which contains images of handwritten digits (0-9). We consider a binary classification task to predict whether the digit is < 5 or not. For part (d), we consider some synthetic data for a binary classification task, where the class conditioned distributions are Gaussian, and train a linear model.

(a) (1 pt) The first experiment involves training a MLP on the MNIST data without label noise. This will be a baseline for the other cases. State your observations on the accuracy and generalization of the model.

(b) In this case, you will train the model on samples with random labels, i.e., the labels are flipped with probability 0.5. (The labels for the test set are unchanged).

  i. (2 pts) What do you observe about the training accuracy in this case? What does this imply about the observations from (a)?

  ii. (1 pt) Try changing the number of hidden units and discuss your observations.

(c) In this case, you will train the model on samples with 20% label noise, and test on samples with no label noise.

i. (1 pt) Before running the experiment, discuss some possible things you might expect a predictor to do (in terms of train and test accuracy).

ii. (2 pts) What do you observe during different stages of training?

iii. (2 pts) Try reducing the batch size and using momentum, and discuss your observations.

iv. (2 pts) Try changing the number of hidden units and discuss your observations.

(d) In this case, we will essentially repeat (c) with a linear model on a synthetic dataset. The data is generated as follows:

$$y \sim \text{Unif}(\{-1, 1\}), \quad \mathbf{x} \sim \mathcal{N}(y\boldsymbol{\mu}, \sigma^2 I_d),$$

where $\boldsymbol{\mu} \in \mathbb{R}^d$ with $\mu_i \sim 0.1\text{Unif}([0, 1])$, for every $i \in [d]$. Next, the labels are flipped with probability 0.2.

i. (1 pts) Discuss your observations on the train and test accuracy.

ii. (2 pts) Try increasing the number of training samples n and reducing the dimension d, and discuss your observations.