

# Homework #3

CSE 446/546: Machine Learning  
Prof. Matt Golub & Hunter Schafer  
Due: Feb 26, 2024 11:59pm  
Points A: 82; B: 5

Please review all homework guidance posted on the website before submitting to Gradescope. Reminders:

- All code must be written in Python and all written work must be typeset (e.g.  $\text{\LaTeX}$ ).
- Make sure to read the “What to Submit” section following each question and include all items.
- Please provide succinct answers and supporting reasoning for each question. Similarly, when discussing experimental results, concisely create tables and/or figures when appropriate to organize the experimental results. All explanations, tables, and figures for any particular part of a question must be grouped together.
- For every problem involving generating plots, please include the plots as part of your PDF submission.
- When submitting to Gradescope, please link each question from the homework in Gradescope to the location of its answer in your homework PDF. Failure to do so may result in deductions of up to 10% of the value of each question not properly linked. For instructions, see [https://www.gradescope.com/get\\_started#student-submission](https://www.gradescope.com/get_started#student-submission).

**Important:** By turning in this assignment (and all that follow), you acknowledge that you have read and understood the collaboration policy with humans and AI assistants alike: <https://courses.cs.washington.edu/courses/cse446/24wi/assignments/>. Any questions about the policy should be raised at least 24 hours before the assignment is due. There are no warnings or second chances. If we suspect you have violated the collaboration policy, we will report it to the college of engineering who will complete an investigation. Not adhering to these reminders may result in point deductions.

## Conceptual Questions

A1. The answers to these questions should be answerable without referring to external materials. Briefly justify your answers with a few words.

- a. [2 points] True or False: Training deep neural networks requires minimizing a convex loss function, and therefore gradient descent will provide the best result.
- b. [2 points] True or False: It is a good practice to initialize all weights to zero when training a deep neural network.
- c. [2 points] True or False: We use non-linear activation functions in a neural network's hidden layers so that the network learns non-linear decision boundaries.
- d. [2 points] True or False: Given a neural network, the time complexity of the backward pass step in the backpropagation algorithm can be prohibitively larger compared to the relatively low time complexity of the forward pass step.
- e. [2 points] True or False: Neural Networks are the most extensible model and therefore the best choice for any circumstance.

### What to Submit:

- **Parts a-e:** 1-2 sentence explanation containing your answer.

#### Solution:

**a**

False. Deep neural network usually does not have convex loss function, and thus, the gradient descent may not reach the global minimum.

**b**

FALSE. A good initialization can improve the algorithm a lot.

**c**

True. We want to learn the non-linear patterns in the data, and if we only use linear activation function, we will end up with a combination of linear functions.

**d**

False. The time complexity of the forward and backward pass step are the same.

**e**

False. For example, for some convex optimization problem, gradient descent can do well.

## Kernels

A2. [5 points] Suppose that our inputs  $x$  are one-dimensional and that our feature map is infinite-dimensional:  $\phi(x)$  is a vector whose  $i$ th component is:

$$\frac{1}{\sqrt{i!}} e^{-x^2/2} x^i ,$$

for all nonnegative integers  $i$ . (Thus,  $\phi$  is an infinite-dimensional vector.) Show that  $K(x, x') = e^{-\frac{(x-x')^2}{2}}$  is a kernel function for this feature map, i.e.,

$$\phi(x) \cdot \phi(x') = e^{-\frac{(x-x')^2}{2}} .$$

Hint: Use the Taylor expansion of  $z \mapsto e^z$ . (This is the one dimensional version of the Gaussian (RBF) kernel).

## What to Submit:

- Proof.

## Solution:

$$\phi(x)\phi(x') = \sum_{i=0}^{\infty} \frac{1}{\sqrt{i!}} e^{-x^2/2} x^i \cdot \frac{1}{\sqrt{i!}} e^{-x'^2/2} x'^i = \frac{1}{i!} e^{-(x^2+x'^2)/2} (xx')^i$$

Given that  $e^z = \sum_{i=0}^{\infty} z^i / i!$ , we have

$$\phi(x)\phi(x') = \sum_{i=0}^{\infty} \frac{1}{i!} e^{-(x^2+x'^2)/2} (xx')^i = \sum_{i=0}^{\infty} e^{-(x^2+x'^2)/2} e^{xx'} = \sum_{i=0}^{\infty} e^{-(x^2+x'^2-2xx')/2} = \sum_{i=0}^{\infty} e^{-(x-x')^2/2}$$

Thus,  $K(x, x') = e^{-\frac{(x-x')^2}{2}}$  is a kernel function for this feature map.

A3. This problem will get you familiar with kernel ridge regression using the polynomial and RBF kernels. First, let's generate some data. Let  $n = 30$  and  $f_*(x) = 6 \sin(\pi x) \cos(4\pi x^2)$ . For  $i = 1, \dots, n$  let each  $x_i$  be drawn uniformly at random from  $[0, 1]$ , and let  $y_i = f_*(x_i) + \epsilon_i$  where  $\epsilon_i \sim \mathcal{N}(0, 1)$ . For any function  $f$ , the true error and the train error are respectively defined as:

$$\mathcal{E}_{\text{true}}(f) = \mathbb{E}_{X,Y} [(f(X) - Y)^2], \quad \hat{\mathcal{E}}_{\text{train}}(f) = \frac{1}{n} \sum_{i=1}^n (f(x_i) - y_i)^2.$$

Now, our goal is, using kernel ridge regression, to construct a predictor:

$$\hat{\alpha} = \arg \min_{\alpha} \|K\alpha - y\|_2^2 + \lambda \alpha^\top K \alpha, \quad \hat{f}(x) = \sum_{i=1}^n \hat{\alpha}_i k(x_i, x)$$

where  $K \in \mathbb{R}^{n \times n}$  is the kernel matrix such that  $K_{i,j} = k(x_i, x_j)$ , and  $\lambda \geq 0$  is the regularization constant.

- a. **[10 points]** Using leave-one-out cross validation, find a good  $\lambda$  and hyperparameter settings for the following kernels:

- $k_{\text{poly}}(x, z) = (1 + x^\top z)^d$  where  $d \in \mathbb{N}$  is a hyperparameter,
- $k_{\text{rbf}}(x, z) = \exp(-\gamma \|x - z\|_2^2)$  where  $\gamma > 0$  is a hyperparameter<sup>1</sup>.

We strongly recommend implementing either [grid search](#) or [random search](#). **Do not use sklearn**, but actually implement of these algorithms. Reasonable values to look through in this problem are:  $\lambda \in 10^{[-5, -1]}$  and  $d \in [5, 25]$ . You do **not** need to search over  $\gamma$  (you can use the heuristic given in the footnote), but if you would like to, a reasonable place to start would be to sample from a narrow gaussian distribution centered at the value described in the footnote.

Report the values of  $d$ ,  $\lambda$ , and  $\gamma$  for both kernels.

- b. **[10 points]** Let  $\hat{f}_{\text{poly}}(x)$  and  $\hat{f}_{\text{rbf}}(x)$  be the functions learned using the hyperparameters you found in part a. For a single plot per function  $\hat{f} \in \{\hat{f}_{\text{poly}}(x), \hat{f}_{\text{rbf}}(x)\}$ , plot the original data  $\{(x_i, y_i)\}_{i=1}^n$ , the true  $f(x)$ , and  $\hat{f}(x)$  (i.e., define a fine grid on  $[0, 1]$  to plot the functions).

## What to Submit:

- **Part a:** Report the values of  $d$ ,  $\gamma$  and the value of  $\lambda$  for both kernels as described.
- **Part b:** Two plots. One plot for each function.

<sup>1</sup>Given a dataset  $x_1, \dots, x_n \in \mathbb{R}^d$ , a heuristic for choosing a range of  $\gamma$  in the right ballpark is the inverse of the median of all  $\binom{n}{2}$  squared distances  $\|x_i - x_j\|_2^2$ .

- **Code** on Gradescope through coding submission.

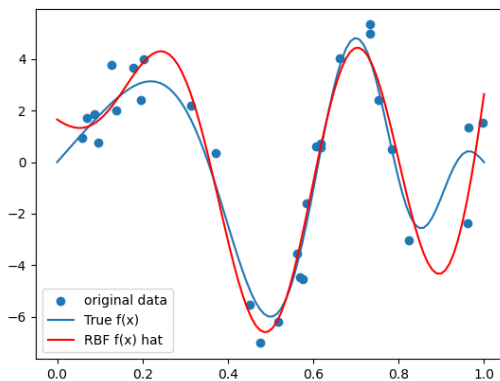
### Solution:

**a**

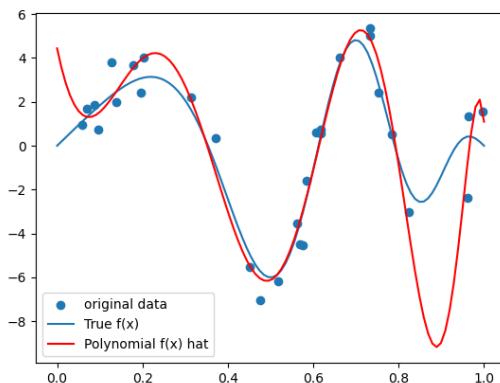
For the polynomial kernel, the optimal  $\lambda$  is  $4.037 * 10^{-5}$ , and the optimal  $d$  is 16.  
For the RBF kernel, the optimal  $\lambda$  is 0.002, and the optimal  $\gamma$  is 10.542.

**b**

This is the figure for RBF kernel.



This is the figure for polynomial kernel.



## Perceptron

B1. One of the oldest algorithms used in machine learning (from the early 60's) is an online algorithm for learning a linear threshold function called the Perceptron Algorithm. It works as follows:

1. Start with the all-zeroes weight vector  $\mathbf{w}_1 = 0$ , and initialize  $t$  to 1. Also let's automatically scale all examples  $\mathbf{x}$  to have (Euclidean) norm 1, since this doesn't affect which side of the plane they are on.
2. Given example  $\mathbf{x}$ , predict positive iff  $\mathbf{w}_t \cdot \mathbf{x} > 0$ .
3. On a mistake, update as follows:
  - Mistake on positive:  $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t + \mathbf{x}$ .

- Mistake on negative:  $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - \mathbf{x}$ .

4.  $t \leftarrow t + 1$ .

If we make a mistake on a positive  $\mathbf{x}$  we get  $\mathbf{w}_{t+1} \cdot \mathbf{x} = (\mathbf{w}_t + \mathbf{x}) \cdot \mathbf{x} = \mathbf{w}_t \cdot \mathbf{x} + 1$ , and similarly if we make a mistake on a negative  $\mathbf{x}$  we have  $\mathbf{w}_{t+1} \cdot \mathbf{x} = (\mathbf{w}_t - \mathbf{x}) \cdot \mathbf{x} = \mathbf{w}_t \cdot \mathbf{x} - 1$ . So, in both cases we move closer (by 1) to the value we wanted. Here is a [link](#) if you are interested in more details.

Now consider the linear decision boundary for classification (labels in  $\{-1, 1\}$ ) of the form  $\mathbf{w} \cdot \mathbf{x} = 0$  (i.e., no offset). Now consider the following loss function evaluated at a data point  $(\mathbf{x}, y)$  which is a variant on the hinge loss.

$$\ell((\mathbf{x}, y), \mathbf{w}) = \max\{0, -y(\mathbf{w} \cdot \mathbf{x})\}.$$

- a. [2 points] Given a dataset of  $(\mathbf{x}_i, y_i)$  pairs, write down a single step of subgradient descent with a step size of  $\eta$  if we are trying to minimize

$$\frac{1}{n} \sum_{i=1}^n \ell((\mathbf{x}_i, y_i), \mathbf{w})$$

for  $\ell(\cdot)$  defined as above. That is, given a current iterate  $\tilde{\mathbf{w}}$  what is an expression for the next iterate?

- b. [2 points] Use what you derived to argue that the Perceptron can be viewed as implementing SGD applied to the loss function just described (for what value of  $\eta$ )?
- c. [1 point] Suppose your data was drawn i.i.d. and that there exists a  $\mathbf{w}^*$  that separates the two classes perfectly. Provide an explanation for why hinge loss is generally preferred over the loss given above. Murphy, Chapter 17.3 (Support vector machines) will be helpful here.

## What to Submit:

- **Part a:** Expression for a single step of subgradient descent
- **Part b:** A 1-2 sentence explanation.
- **Part c:** A 1-2 sentence explanation.

## Solution:

a

$$\tilde{\mathbf{w}}_{t+1} = \tilde{\mathbf{w}}_t - \eta \nabla \frac{1}{n} \sum_{i=1}^n \ell((x_i, y_i), w)$$

$$\nabla \frac{1}{n} \sum_{i=1}^n \ell((x_i, y_i), w) = \begin{cases} 0, & \text{if } \ell((x_i, y_i), w) \leq 0 \\ -y \cdot x_i, & \text{if } \ell((x_i, y_i), w) > 0 \end{cases}$$

b

When we have the  $n = 1$  and the  $\eta = 1$ , then, our derivation above becomes:

If  $-y(w \cdot x_i) \leq 0$ ,

$$\tilde{\mathbf{w}}_{t+1} = \tilde{\mathbf{w}}_t$$

If  $-y(w \cdot x_i) > 0$  and  $y = 1$ ;

$$\tilde{\mathbf{w}}_{t+1} = \tilde{\mathbf{w}}_t - x_i$$

If  $-y(w \cdot x_i) > 0$  and  $y = -1$ ;

$$\tilde{\mathbf{w}}_{t+1} = \tilde{\mathbf{w}}_t + x_i$$

**c**

Hinge loss is the  $\max(0, 1 - y(w \cdot x))$ .

When we have  $-y(w \cdot x_i) \leq 0$ , we stop updating  $w$  based on the above loss function since they correctly classify the data point. While with Hinge loss, when  $-1 \leq -y(w \cdot x_i) \leq 0$ , we still update the  $w$ , so that we can make our classification enough margin.

## Introduction to PyTorch

### Resources

For questions A.4 and A.5, you will use PyTorch. In [Section materials \(Week 6\)](#) there is a notebook that you might find useful. Additionally make use of [PyTorch Documentation](#), when needed.

A4. PyTorch is a great tool for developing, deploying and researching neural networks and other gradient-based algorithms. In this problem we will explore how this package is built, and re-implement some of its core components. Firstly start by reading `README.md` file provided in `intro_pytorch` subfolder. A lot of problem statements will overlap between here, readme's and comments in functions.

- a. **[10 points]** You will start by implementing components of our own PyTorch modules. You can find these in folders: `layers`, `losses` and `optimizers`. Almost each file there should contain at least one problem function, including exact directions for what to achieve in this problem. Lastly, you should implement functions in `train.py` file.
- b. **[5 points]** Next we will use the above module to perform hyper-parameter search<sup>2</sup>. Here we will also treat loss function as a hyper-parameter. However, because cross-entropy and MSE require different shapes we are going to use two different files: `crossentropy_search.py` and `mean_squared_error_search.py`. For each you will need to build and train (in provided order) 5 models:
  - Linear neural network (Single layer, no activation function)
  - NN with one hidden layer (2 units) and sigmoid activation function after the hidden layer
  - NN with one hidden layer (2 units) and ReLU activation function after the hidden layer
  - NN with two hidden layer (each with 2 units) and Sigmoid, ReLU activation functions after first and second hidden layers, respectively
  - NN with two hidden layer (each with 2 units) and ReLU, Sigmoid activation functions after first and second hidden layers, respectively

For each loss function, submit a plot of losses from training and validation sets. All models should be on the same plot (10 lines per plot), with two plots total (1 for MSE, 1 for cross-entropy).

- c. **[5 points]** For each loss function, report the best performing architecture (best performing is defined here as achieving the lowest validation loss at any point during the training), and plot its guesses on test set. You should use function `plot_model_guesses` from `train.py` file. Lastly, report accuracy of that model on a test set.

---

<sup>2</sup>In this problem, the hyper-parameters required to search are (1) model architectures and (2) loss functions. Classic hyper-parameters like batch size and learning rates are not required to be searched over as long as the loss curve converges.

## The Softmax function

One of the activation functions we ask you to implement is softmax. For a prediction  $\hat{y} \in \mathbb{R}^k$  corresponding to single datapoint (in a problem with  $k$  classes):

$$\text{softmax}(\hat{y}_i) = \frac{\exp(\hat{y}_i)}{\sum_j \exp(\hat{y}_j)}$$

### What to Submit:

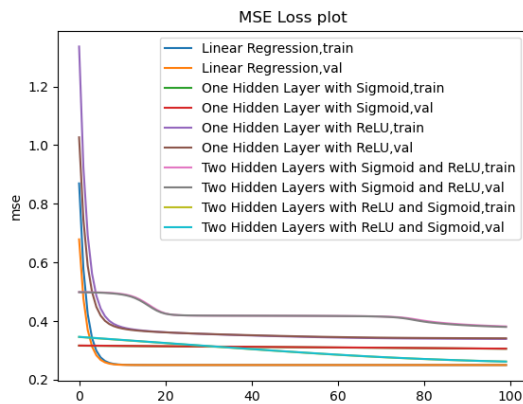
- **Part b:** 2 plots (one per loss function), with 10 lines each, showing both training and validation loss of each model. Make sure plots are titled, and have proper legends.
- **Part c:** Names of best performing models (i.e. descriptions of their architectures), and their accuracy on test set.
- **Part c:** 2 scatter plots (one per loss function), with predictions of best performing models on test set.
- **Code** on Gradescope through coding submission

### Solution:

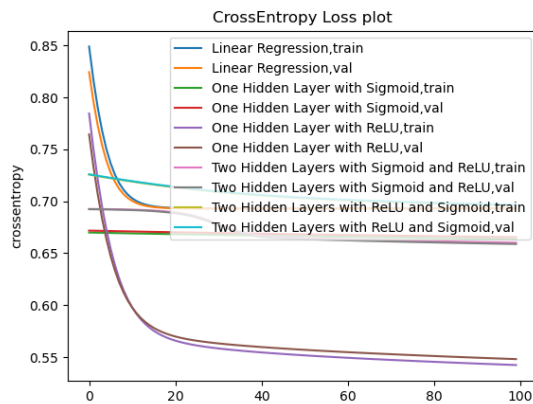
a: Please see the code uploaded.

b:

loss plot with mse loss is

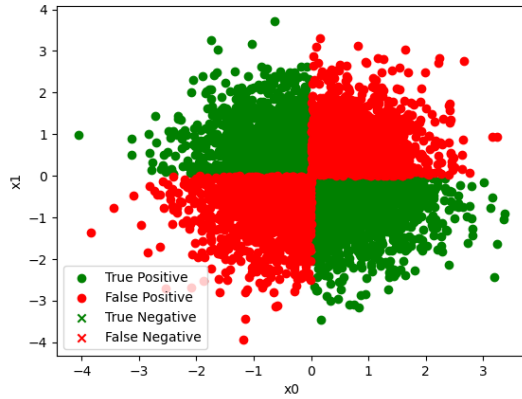


loss plot with crossentropy loss is

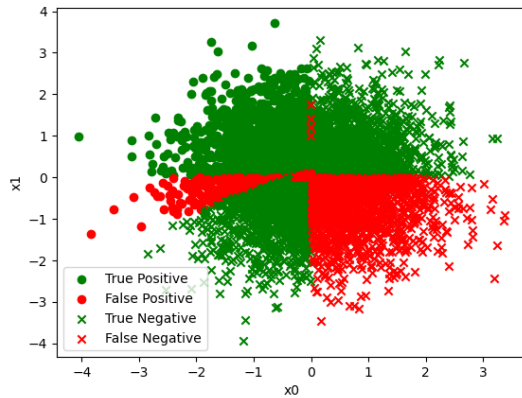


C:

With the MSE, the best model is Linear Regression and the accuracy is 0.5032, the plot on test data is:



With crossentropy, the best model is One Hidden Layer with ReLU and the accuracy is 0.6856, the plot on test data is:



## Neural Networks for MNIST

A5. In Homework 1, we used ridge regression to train a classifier for the MNIST dataset. In Homework 2, we used logistic regression to distinguish between the digits 2 and 7. Now, in this problem, we will use PyTorch to build a simple neural network classifier for MNIST to further improve our accuracy.

We will implement two different architectures: a shallow but wide network, and a narrow but deeper network. For both architectures, we use  $d$  to refer to the number of input features (in MNIST,  $d = 28^2 = 784$ ),  $h_i$  to refer to the dimension of the  $i$ -th hidden layer and  $k$  for the number of target classes (in MNIST,  $k = 10$ ). For the non-linear activation, use ReLU. Recall from lecture that

$$\text{ReLU}(x) = \begin{cases} x, & x \geq 0 \\ 0, & x < 0. \end{cases}$$

### Weight Initialization

Consider a weight matrix  $W \in \mathbb{R}^{n \times m}$  and  $b \in \mathbb{R}^n$ . Note that here  $m$  refers to the input dimension and  $n$  to the output dimension of the transformation  $x \mapsto Wx + b$ . Define  $\alpha = \frac{1}{\sqrt{m}}$ . Initialize all your weight matrices and



biases according to  $\text{Unif}(-\alpha, \alpha)$ .

## Training

For this assignment, use the Adam optimizer from `torch.optim`. Adam is a more advanced form of gradient descent that combines momentum and learning rate scaling. It often converges faster than regular gradient descent in practice. You can use either Gradient Descent or any form of Stochastic Gradient Descent. Note that you are still using Adam, but might pass either the full data, a single datapoint or a batch of data to it. Use cross entropy for the loss function and ReLU for the non-linearity.

## Implementing the Neural Networks

- a. **[10 points]** Let  $W_0 \in \mathbb{R}^{h \times d}$ ,  $b_0 \in \mathbb{R}^h$ ,  $W_1 \in \mathbb{R}^{k \times h}$ ,  $b_1 \in \mathbb{R}^k$  and  $\sigma(z): \mathbb{R} \rightarrow \mathbb{R}$  some non-linear activation function applied element-wise. Given some  $x \in \mathbb{R}^d$ , the forward pass of the wide, shallow network can be formulated as:

$$\mathcal{F}_1(x) := W_1 \sigma(W_0 x + b_0) + b_1$$

Use  $h = 64$  for the number of hidden units and choose an appropriate learning rate. Train the network until it reaches 99% accuracy on the training data and provide a training plot (loss vs. epoch). Finally evaluate the model on the test data and report both the accuracy and the loss.

- b. **[10 points]** Let  $W_0 \in \mathbb{R}^{h_0 \times d}$ ,  $b_0 \in \mathbb{R}^{h_0}$ ,  $W_1 \in \mathbb{R}^{h_1 \times h_0}$ ,  $b_1 \in \mathbb{R}^{h_1}$ ,  $W_2 \in \mathbb{R}^{k \times h_1}$ ,  $b_2 \in \mathbb{R}^k$  and  $\sigma(z): \mathbb{R} \rightarrow \mathbb{R}$  some non-linear activation function. Given some  $x \in \mathbb{R}^d$ , the forward pass of the network can be formulated as:

$$\mathcal{F}_2(x) := W_2 \sigma(W_1 \sigma(W_0 x + b_0) + b_1) + b_2$$

Use  $h_0 = h_1 = 32$  and perform the same steps as in part a.

- c. **[5 points]** Compute the total number of parameters of each network and report them. Then compare the number of parameters as well as the test accuracies the networks achieved. Is one of the approaches (wide, shallow vs. narrow, deeper) better than the other? Give an intuition for why or why not.

**Using PyTorch:** For your solution, you may not use any functionality from the `torch.nn` module except for `torch.nn.functional.relu` and `torch.nn.functional.cross_entropy`. You must implement the networks  $\mathcal{F}_1$  and  $\mathcal{F}_2$  from scratch. For starter code and a tutorial on PyTorch refer to the sections 6 and 7 material.

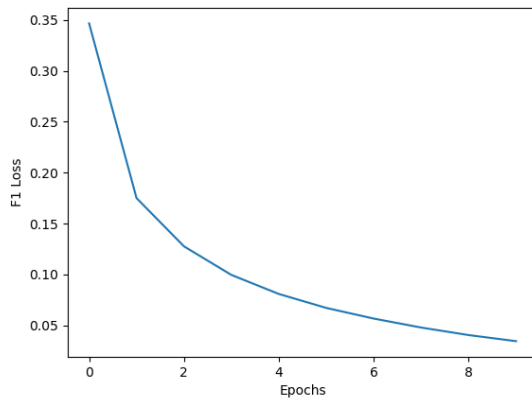
## What to Submit:

- **Parts a-b:** Provide a plot of the training loss versus epoch. In addition, evaluate the model trained on the test data and report the accuracy and loss.
- **Part c:** Report the number of parameters for the network trained in part (a) and for the network trained in part (b). Provide a comparison of the two networks as described in part (c) in 1-2 sentences.
- **Code** on Gradescope through coding submission.

### Solution:

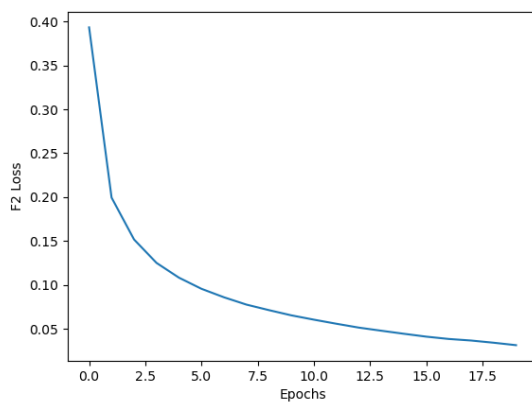
**a**

The training loss plot of F1 model is



**b**

The training loss plot of F2 model is



**c**

The number of parameters in F1 and F2 are:

num of parameter in f1: 50890 and num of parameter in f2: 26506.

The test loss in F1 and F2 are:

loss f1 test: 0.11024830490350723 and loss f2 test 0.2087639719247818.

The test accuracy in F1 and F2 are:

accuracy f1 test: 0.968 and accuracy f2 test 0.9556.

Thus, F1 is better than F2 in terms of accuracy, but F1 has more parameter.

## Administrative

A6.

- a. [2 points] About how many hours did you spend on this homework? There is no right or wrong answer :)

50 + hours...