ECE 364: Programming Methods for Machine Learning, Spring 2026

Midterm 1 – March 10, 2026

- **You will have 75 minutes (1.25 hours) to solve all the problems. Most have multiple parts.** Don't spend too much time on questions you don't understand and focus on answering as much as you can!
- ***BUDGET YOUR TIME WISELY***. I highly recommend working on the questions you know first and the questions you need to think about second.
- *No* resources are allowed for use during the exam except a cheatsheet and scratch paper on the back of the exam. **Do not tear out the cheatsheet or the scratch paper!** It messes with the auto-scanner.
- You should write your answers *completely* in the space given for the question. We will not grade parts of any answer written outside of the designated space.
- Please *use a dark-colored pen* unless you are *absolutely* sure your pencil writing is forceful enough to be legible when scanned. We reserve the right to take off points if we have difficulty reading the uploaded document.
- **Don't cheat.** C'mon, be cool, be honest.
- **Good luck!**

Name: _____

NetID: _____

Date: _____

1. **True/False** (12 points)

For each of the following problems, circle TRUE if the statement is *always* true, circle FALSE otherwise. There is no partial credit for these questions.

a. Linear regression models assume that the relationship between the independent variable(s) x and the dependent variable y can be represented as a linear combination of the input features.

Solution:

(TRUE) FALSE

b. Logistic regression is primarily used to predict continuous numerical values, such as predicting the exact price of a house.

Solution:

TRUE (FALSE)

c. In PyTorch, a Tensor's `.requires_grad` attribute must be set to `True` to track operations on it for automatic gradient computation.

Solution:

(TRUE) FALSE

d. The Rectified Linear Unit (ReLU) activation function is defined by the mathematical expression $f(x) = \max(0, x)$.

Solution:

(TRUE) FALSE

e. In PyTorch, executing `loss.backward()` computes the gradients and immediately updates the weights of the model to minimize the loss.

Solution:

TRUE (FALSE)

f. It is **not** advisable to pair the mean squared error loss function with the Sigmoid activation function because the gradients at large input values tend to be small.

Solution:

(TRUE) FALSE

g. The Swish activation function, defined as $f(x) = x \cdot \text{sigmoid}(\beta x)$, is strictly increasing for all positive values of β.

Solution:

TRUE (FALSE)

h. In ReLU, if a neuron's weights are updated such that the function always outputs zero for all inputs in the training set, its gradient effectively becomes zero permanently.

Solution:

(TRUE) FALSE

2. **Lost in the Layers: Navigating Data Slices** (14 points)

For each of the code segments, answer the following questions based on *the final state* of the variables.

(a)
```
import torch
data = torch.zeros((3, 2))
bias = torch.tensor([[10.], [20.], [30.]])
result = data + bias
result[result > 15] -= 5
```

i. What will be the shape of `result`?

Solution:
3×2 (3 rows by 2 columns)

ii. What does `result` contain?

Solution:
$$\begin{bmatrix} 10 & 10 \\ 15 & 15 \\ 25 & 25 \end{bmatrix}$$

(b)
```
import torch
a = torch.arange(42).view(3, -1) # arange(N) gives int vector [0...N-1]
b = a[:, ::5]
b.add_(1)
```

i. What does `a[:, 0]` return?

Solution:
$$\begin{bmatrix} 1 & 15 & 29 \end{bmatrix}$$

ii. What is the shape of `b`?

Solution:
3×3 (3 rows by 3 columns)

iii. What does `b` contain?

Solution:
$$\begin{bmatrix} 1 & 6 & 11 \\ 15 & 20 & 25 \\ 29 & 34 & 39 \end{bmatrix}$$

3. **Feeling out of place** (6 points)

Suppose we want to find the value of x where $\ln(x^2) = 3$. We write the following gradient descent code:

```
x_gd = torch.tensor(6.0, requires_grad=True)
target = 3
alpha = 0.001
epochs = 20000
for i in range(epochs):
    f = torch.log(torch.pow(x_gd, 2))
    loss = (target - f)**2
    loss.backward()
    with torch.no_grad():
        x_gd -= alpha * x_gd.grad
```

...but there's an error! After much debugging, you narrow the issue down to `x_gd`. For some reason `x_gd` keeps becoming infinite! What do you need to change/add/delete, so that the gradient descent code can work appropriately?

Solution:

The issue is that you are not zeroing out the gradient. You need to add the line `x_gd.grad = None`, which zeros out the gradient. If you want, you can play around with the code below:

```

import torch
import numpy as np
import matplotlib.pyplot as plt

x_vals = []
f_vals = []
x_gd = torch.tensor(6.0, requires_grad=True)
target=3
alpha = 0.001
epochs = 20000
for i in range(epochs):
    f = torch.log(torch.pow(x_gd,2))
    loss = (target-f)**2
    loss.backward()
    with torch.no_grad():
        x_gd -= alpha*x_gd.grad
        x_gd.grad = None  ### Need to add this
        x_vals.append(x_gd.data.item())
        f_vals.append(f.data.item())

fig, ax1 = plt.subplots()
color = 'tab:red'
ax1.set_xlabel('iterations')
ax1.set_ylabel('x values', color=color)
ax1.plot(range(epochs), x_vals, color=color)
ax1.tick_params(axis='y', labelcolor=color)

ax2 = ax1.twinx()  # instantiate a second Axes that shares the same x-axis
color = 'tab:blue'
ax2.set_ylabel('f(x)', color=color)  # we already handled the x-label with ax1
ax2.plot(range(epochs), f_vals, color=color, linestyle='--')
ax2.tick_params(axis='y', labelcolor=color)
```

```
34
35 fig.tight_layout()  # otherwise the right y-label is slightly clipped
36 plt.show()
37
38 print('x = ' + str(x_gd.data.item()) + ' when f(x) = ' + str(target) )
39
```

4. **Jacobians, Hessians, and Why My Brain Hurts (Matrix Calculus)** (6 points)
Given:

$$f(x) = a^T x$$

where $x \in \mathbb{R}^n$, $a \in \mathbb{R}^n$

Find the solution to the following partial derivative:

$$\frac{\partial f}{\partial x} =$$

You must explain why your answer is correct for full credit.

Solution:

This question is very straight forward:

$$a^T x = \sum_{i=1}^{n} a_i x_i$$

Compute partial derivative with respect of x_j:

$$\frac{\partial f}{\partial x_j} = a_j$$

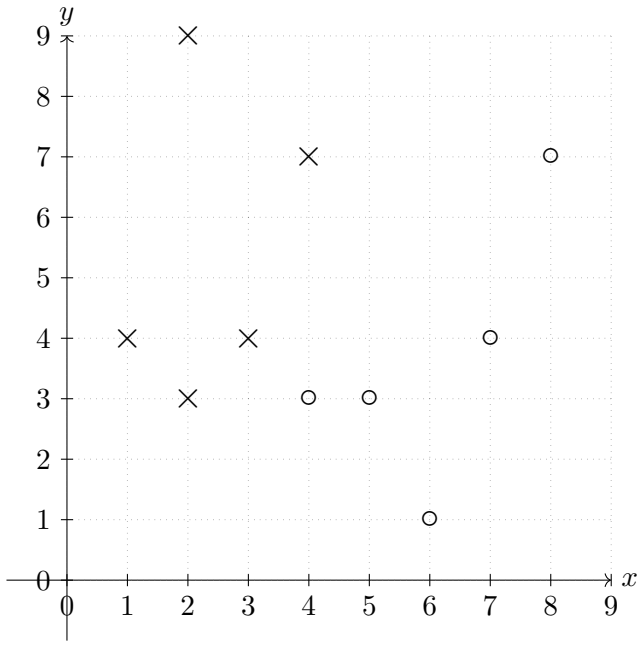
Put element wise partial derivative into a vector:

$$\frac{\partial f}{\partial x} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} = a$$

5. **The Art of the SVM Margin** (12 points)

Support vectors are the critical data points in an SVM that lie closest to the decision boundary. They determine both the optimal separating boundary and the margin width between classes. Removing these points from the training set would shift the boundary, potentially altering the classification results.

Consider a SVM that separates the following 2D points into two classes (X and O):



With points:

- **Class X:** $\{(1,4),(2,3),(2,9),(3,4),(4,7)\}$
- **Class O:** $\{(4,3),(5,3),(6,1),(7,4),(8,7)\}$

(a) Drawing the Boundary: Sketch the decision boundary for an SVM on the provided graph. Choose the decision boundary that maximizes the margin and minimizes the loss. Additionally, sketch the margin boundaries for each class (these are the lines that pass through the closest data points from each class, positioned half a margin away from the decision boundary). Label each line clearly. You should draw a total of three lines. Include the approximate equation for each line.

Solution:
Class $\times$ boundary: a line with the equation $y = x + 1$,
Class $\circ$ boundary: a line with the equation $y = x - 1$, and
Decision boundary: a line with the equation $y = x$.

(b) If we remove $(8, 7)$ belonging to class $\circ$ from the dataset, what happens to the decision boundary? Briefly explain your reasoning.

Solution:
The decision boundary does not shift. After removing $(8, 7)$, the following candidates are available for support vectors:

1. $(2, 3)$, $(3, 4)$, and $(4, 3)$.
2. $(2, 3)$, $(4, 3)$, and $(7, 4)$.

The first option will result in more margin. Hence, the decision boundary remains the same.

(c) If we remove $(2, 3)$ and $(3, 4)$ belonging to class $\times$ from the dataset, what happens to the decision boundary? Briefly explain your reasoning.

Solution:
After removing the mentioned points, the following candidates are available for support vectors:

1. $(1, 4)$, $(4, 7)$, $(4, 3)$, and $(8, 7)$.

With this option, the margin boundary for class $\times$ will shift towards the $\times$'s. Hence, the decision boundary will shift towards the $\times$'s.

6. **Where to Draw the Line? A Classifier's Dilemma** (10 points)

Let g be a logical function, defined on the feature space $\{+1, -1\}^3$, which maps:

- $g(+1, +1, +1) = +1$
- $g(-1, +1, +1) = +1$
- $g(-1, -1, -1) = +1$
- $g(+1, -1, +1) = -1$.

Given a linear classifier $h(x) = \text{sign}(\boldsymbol{w} \cdot \boldsymbol{x} + b)$, where $\boldsymbol{w} \in \mathbb{R}^{1\times 3}$, $\boldsymbol{x} \in \mathbb{R}^{3\times 1}$, and $b \in \mathbb{R}^{1\times 1}$, give a valid $(\boldsymbol{w}, b)$ pair that matches the ground truth g. Let $\text{sign}(z) = +1$ for $z \geq 0$ and -1 otherwise. Give your solution and show that it is valid.

Solution:

Let $\boldsymbol{w} = \begin{bmatrix} -1 & +1 & 0 \end{bmatrix}$ and $b = 1$. Then the ($\boldsymbol{w}$, b) pair would match the ground truth g for a linear classifier $h(x) = \text{sign}(\boldsymbol{w} \cdot \boldsymbol{x} + b)$. We can check by doing the following calculations, seeing that this mapping leads to:

$$h(+1,+1,+1) = \text{sign}(\begin{bmatrix} -1 & +1 & 0 \end{bmatrix} \cdot \begin{bmatrix} +1 \\ +1 \\ +1 \end{bmatrix} + 1) = \text{sign}(+1) = +1 = g(+1,+1,+1)$$

$$h(-1,+1,+1) = \text{sign}(\begin{bmatrix} -1 & +1 & 0 \end{bmatrix} \cdot \begin{bmatrix} -1 \\ +1 \\ +1 \end{bmatrix} + 1) = \text{sign}(3) = +1 = g(-1,+1,+1)$$

$$h(-1,-1,-1) = \text{sign}(\begin{bmatrix} -1 & +1 & 0 \end{bmatrix} \cdot \begin{bmatrix} -1 \\ -1 \\ -1 \end{bmatrix} + 1) = \text{sign}(+1) = +1 = g(-1,-1,-1)$$

$$h(+1,-1,+1) = \text{sign}(\begin{bmatrix} -1 & +1 & 0 \end{bmatrix} \cdot \begin{bmatrix} +1 \\ -1 \\ +1 \end{bmatrix} + 1) = \text{sign}(-1) = -1 = g(+1,-1,+1)$$

and matches the mapping of g where g is a logical function defined on the feature space $\{+1,-1\}^3$.

7. **Stumbling Down the Gradient: My Life Story** (12 points)

Consider the following function

$$f(x,y) = \frac{1}{2e}\left(x^2+y^2\right) - \left(\frac{1}{2}+\frac{1}{e}\right)(x+y)$$

(a) Determine the gradient $\nabla f(x,y)$.

Solution:

$$\frac{df}{dx} = \frac{x}{e} - \left(\frac{1}{2}+\frac{1}{e}\right)$$
$$\frac{df}{dy} = \frac{y}{e} - \left(\frac{1}{2}+\frac{1}{e}\right)$$

Hence,

$$\nabla f(x,y) = \begin{bmatrix} \frac{x}{e} - \left(\frac{1}{2}+\frac{1}{e}\right) \\ \frac{y}{e} - \left(\frac{1}{2}+\frac{1}{e}\right) \end{bmatrix}$$

(b) Let the starting point for gradient descent at $k=0$ be $\left(x^{(0)}, y^{(0)}\right) = (1,1)$ and the step size be $\alpha = 2(e-1)$. Here, e is Euler's number. Apply gradient descent to obtain the values of x and y at iterations $k=1$ and $k=2$.

Solution:

Gradient descent update can be written as

$$\left(x^{(k)}, y^{(k)}\right) = \left(x^{(k-1)}, y^{(k-1)}\right) - \alpha \nabla f\left(x,y\right)|_{\left(x^{(k-1)}, y^{(k-1)}\right)}.$$

At $k=1$, we have

$$\nabla f(x,y)|_{(1,1)} = \begin{bmatrix} \frac{1}{e} - \left(\frac{1}{2}+\frac{1}{e}\right) \\ \frac{1}{e} - \left(\frac{1}{2}+\frac{1}{e}\right) \end{bmatrix} = \begin{bmatrix} -\frac{1}{2} \\ -\frac{1}{2} \end{bmatrix}$$
$$\left(x^{(1)}, y^{(1)}\right) = \begin{bmatrix} 1 \\ 1 \end{bmatrix} - 2(e-1)\begin{bmatrix} -\frac{1}{2} \\ -\frac{1}{2} \end{bmatrix} = \begin{bmatrix} 1+(e-1) \\ 1+(e-1) \end{bmatrix} = \begin{bmatrix} e \\ e \end{bmatrix}.$$

At $k=2$, we have

$$\nabla f(x,y)|_{(e,e)} = \begin{bmatrix} \frac{e}{e} - \left(\frac{1}{2}+\frac{1}{e}\right) \\ \frac{e}{e} - \left(\frac{1}{2}+\frac{1}{e}\right) \end{bmatrix} = \begin{bmatrix} \frac{1}{2} - \frac{1}{e} \\ \frac{1}{2} - \frac{1}{e} \end{bmatrix}$$
$$\left(x^{(2)}, y^{(2)}\right) = \begin{bmatrix} e \\ e \end{bmatrix} - 2(e-1)\begin{bmatrix} \frac{1}{2} - \frac{1}{e} \\ \frac{1}{2} - \frac{1}{e} \end{bmatrix} = \begin{bmatrix} e - \frac{(e-1)(e-2)}{e} \\ e - \frac{(e-1)(e-2)}{e} \end{bmatrix} = \begin{bmatrix} 3 - \frac{2}{e} \\ 3 - \frac{2}{e} \end{bmatrix}.$$

8. **torch.nn models, Dataloaders and Optimizers, oh my!** (15 points)

A real estate agency is modeling home prices in a historic district using three factors: (1) area, (2) number of bedrooms, and (3) age of home in years. They have collected data on 50 homes. Some examples from the dataset are as follows:

Area (sqft)	# Bedrooms	Age (years)	Price ($)
1,500	2	5	250,000
1,800	3	6	310,000
2,200	3	4	360,000
2,500	4	2	420,000
3,000	5	5	510,000

The prices are modeled as

$$\text{Price} = w_0 + w_1 \times \text{Area} + w_2 \times \text{No. of Bedrooms} + w_3 \times \text{Property Age} \tag{1}$$

Based on this setup, answer the questions below.

i. Complete the following dataset class. The i^{th} element (accessed as `dataset[i]`) must return a tuple `(x, y)`. `x` is a 1-D tensor with three elements: (1) area, (2) bedrooms, and (3) age. `y` is a 0-D (or scalar) containing the corresponding price.

Solution:

```
import torch
from torch.utils.data import Dataset

class HomePriceDataset(Dataset):
    def __init__(self):
        super().__init__()
        # i-th elements of area_list, bedrooms_list, age_list, and
        # price_list contain the area, the number of bedrooms, the
        # age, and the price of the i-th home.
        self.area_list = [1500, 1800, 2200, 2500, 3000]
        self.bedrooms_list = [2, 3, 3, 4, 5]
        self.age_list = [5, 6, 4, 2, 5]
        self.price_list = [250000, 310000, 360000, 420000, 510000]

    def __len__(self):
        # Implement this
        return len(self.price_list)

    def __getitem__(self, index: int):
        # Implement this (x is the input, and y is the output)
        x = torch.tensor([
            self.area_list[index],
            self.bedrooms_list[index],
            self.age_list[index],
        ])
        y = torch.tensor(self.price_list[index])
        return x, y
```

ii. Complete the following class to implement the model shown in Eq. 1. The `forward` method accepts a batched input tensor of size $B \times 3$ (where B is the batch size). It must return a tensor of predicted prices with size B.

Solution:

The model can be implemented as

```python
import torch
import torch.nn as nn

class HomePriceModel(nn.Module):
    def __init__(self, num_features: int):
        super().__init__()
        # Implement the model
        self.w = nn.Parameter(torch.ones(num_features))
        self.b = nn.Parameter(torch.zeros(1))

    def forward(self, x: torch.Tensor):
        """
        x is a tensor of shape (B, 3).
        """
        # Implement the forward method
        return x @ self.w + self.b
```

Alternatively, we can also use `nn.Linear` to implement the model, which is very concise.

```python
import torch
import torch.nn as nn

class HomePriceModel(nn.Module):
    def __init__(self, num_features: int):
        super().__init__()
        # Implement the model
        self.model = nn.Linear(num_features, 1, bias=True)

    def forward(self, x: torch.Tensor):
        """
        x is a tensor of shape (B, 3).
        """
        # Implement the forward method
        return self.model(x).squeeze(-1)
```

iii. Write the code to train `HomePriceModel` on `HomePriceDataset` for 10 epochs. Use the SGD optimizer with a learning rate of 0.01, mean-squared error loss, and a batch size of 2. No validation step is required.

Solution:

```
import torch.nn as nn
from torch.optim import SGD

dataset = HomePriceDataset()
dataloader = DataLoader(dataset, batch_size=2)
model = HomePriceModel(3)
loss_fn = nn.MSELoss()
optim = SGD(model.parameters(), lr=0.01)

for epoch in range(10):
    for batch in dataloader:
        optim.zero_grad()
        output = model(batch[0])
        loss = loss_fn(output, batch[1])
        loss.backward()
        optim.step()
```

9. **The Computational Graph That Ate My Sanity** (18 points)

Recall the softmax function from the lectures. A slight variation is the log-softmax function which applies to the output of softmax. It can be defined as

$$\text{LogSoftmax}\left(x_i\right) = \log \frac{e^{x_i}}{\sum_j e^{x_j}} = x_i - \log \sum_j e^{x_j}$$

Consider a 3-dimensional tensor $g(x, y, z)$.LogSoftmax(x) can be written as

$$g(x, y, z) = x - \log\left(e^x + e^y + e^z\right)$$

The below computation graph depicts $g(x, y, z)$

(a) For the graph in Figure 1, we have $w_1 = x, w_2 = y$, and $w_3 = z$. Express each intermediate node $(w_4, w_5, w_6, w_7, w_8)$ value in terms of inputs (x, y, z) to construct $g(x, y, z)$. Note that $g(x, y, z) = w_9$ and $w_9 = w_1 - w_8$.

Solution:

$$\begin{aligned}
w_4 &= \exp(x) \\
w_5 &= \exp(y) \\
w_6 &= \exp(z) \\
w_7 &= (\exp(x) + \exp(y) + \exp(z)) \\
w_8 &= \log(\exp(x) + \exp(y) + \exp(z)
\end{aligned}$$

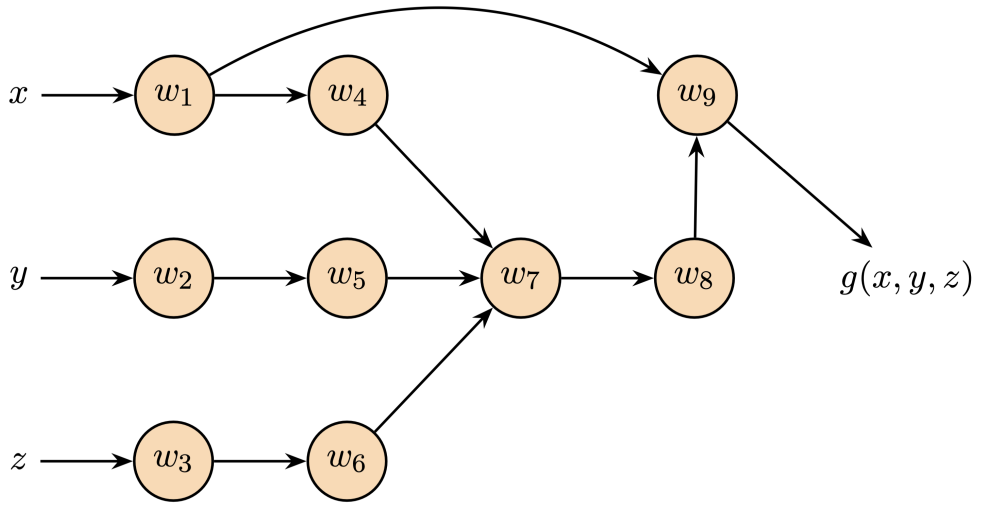


Figure 1: Computation graph for question 9

(b) Determine the partial derivatives of each successor node with respect to its predecessors, e.g., $\frac{\partial w_4}{\partial w_1}$, $\frac{\partial w_5}{\partial w_2}$, $\frac{\partial w_6}{\partial w_3}$, etc. The answers should be in term of x, y, z.

Solution:

$$\frac{\partial w_4}{\partial w_1} = \exp(x)$$
$$\frac{\partial w_5}{\partial w_2} = \exp(y)$$
$$\frac{\partial w_6}{\partial w_3} = \exp(z)$$
$$\frac{\partial w_7}{\partial w_4} = 1$$
$$\frac{\partial w_7}{\partial w_5} = 1$$
$$\frac{\partial w_7}{\partial w_6} = 1$$
$$\frac{\partial w_8}{\partial w_7} = \frac{1}{\exp(x) + \exp(y) + \exp(z)}$$
$$\frac{\partial w_9}{\partial w_1} = \frac{\exp(y) + \exp(z)}{\exp(x) + \exp(y) + \exp(z)}$$
$$\frac{\partial w_9}{\partial w_8} = -1$$

(c) Determine the adjoints at each node $\bar{w}_i = \frac{\partial f}{\partial w_i}$, the answers should be in term of x, y, z.

Solution:

$$
\begin{aligned}
\bar{w_9} &= 1 \\
\bar{w_8} &= -1 \\
\bar{w_7} &= \frac{-1}{\exp(x) + \exp(y) + \exp(z)} \\
\bar{w_6} &= \frac{-1}{\exp(x) + \exp(y) + \exp(z)} \\
\bar{w_5} &= \frac{-1}{\exp(x) + \exp(y) + \exp(z)} \\
\bar{w_4} &= \frac{-1}{\exp(x) + \exp(y) + \exp(z)} \\
\bar{w_3} &= -\frac{\exp(z)}{\exp(x) + \exp(y) + \exp(z)} \\
\bar{w_2} &= -\frac{\exp(y)}{\exp(x) + \exp(y) + \exp(z)} \\
\bar{w_1} &= 1 - \frac{\exp(x)}{\exp(x) + \exp(y) + \exp(z)}
\end{aligned}
$$

Solution:

We verified the solutions via PyTorch. Feel free to check!

```
import torch
import numpy as np

x = torch.tensor([1.5], requires_grad=True) # make sure gradients are
    computed when backpropagation is called
y = torch.tensor([np.pi/3], requires_grad=True)

w1 = x
w2 = y
w3 = z
w4 = torch.exp(w1)
w5 = torch.exp(w2)
w6 = torch.exp(w3)
w7 = w4 + w5 + w6
w8 = torch.log(w7)
w9 = w1 - w8
g = w9

# manual gradients
with torch.no_grad():
    # adjoints
    w9bar = 1
    w8bar = -1
    w7bar = -1/w7
    w6bar = -1/(torch.exp(x)+torch.exp(y)+torch.exp(z))
    w5bar = -1/(torch.exp(x)+torch.exp(y)+torch.exp(z))
    w4bar = -1/(torch.exp(x)+torch.exp(y)+torch.exp(z))
```

```
    w3bar = -torch.exp(z)/(torch.exp(x)+torch.exp(y)+torch.exp(z))
    w2bar = -torch.exp(y)/(torch.exp(x)+torch.exp(y)+torch.exp(z))
    w1bar = 1-torch.exp(x)/(torch.exp(x)+torch.exp(y)+torch.exp(z))

# automatic gradients via backpropagation
w1.retain_grad(), w2.retain_grad(), w3.retain_grad(), w4.retain_grad(), w5
    .retain_grad(), w6.retain_grad(), w7.retain_grad(), w8.retain_grad(),
    w9.retain_grad() # making sure PyTorch populates all gradients
g.backward() # initiate backpropagation from f as the seed node

print("Making sure the overall equation is correct: ")
g_manual = x - torch.log(torch.exp(x)+torch.exp(y)+torch.exp(z))
print('g: Manual = {}, PyTorch = {}'.format(g_manual, f))

print('Comparing our calculations to PyTorch Autograd:')
print('w1: Manual = {}, PyTorch = {}'.format(w1bar, w1.grad))
print('w2: Manual = {}, PyTorch = {}'.format(w2bar, w2.grad))
print('w3: Manual = {}, PyTorch = {}'.format(w3bar, w3.grad))
print('w4: Manual = {}, PyTorch = {}'.format(w4bar, w4.grad))
print('w5: Manual = {}, PyTorch = {}'.format(w5bar, w5.grad))
print('w6: Manual = {}, PyTorch = {}'.format(w6bar, w6.grad))
print('w7: Manual = {}, PyTorch = {}'.format(w7bar, w7.grad))
print('w8: Manual = {}, PyTorch = {}'.format(w8bar, w8.grad))
print('w9: Manual = {}, PyTorch = {}'.format(w9bar, w9.grad))

```

This page is for additional scratch work!

This page is for additional scratch work!

PyTorch Cheatsheet - Part 1

Useful activation function and torch.nn.functional

- Linear function: $y = WX + b$ where W and X are vectors of size N (number of dimensions to the input).

```
torch.nn.Linear(in_features, out_features, bias=True, device=None, dtype=None)
```

- Sigmoid function: $\frac{1}{1+e^{-z}}$ where z is the logit(s).

```
torch.nn.functional.sigmoid(input)
```

- Softmax function: $p(Y = t|x) = \frac{\exp(w_t^T x)}{\sum_{y \in \{0,\dots,C-1\}} \exp(w_y^T x)}$

```
torch.nn.functional.softmax(input, dim=None, _stacklevel=3, dtype=None)[source]
```

Loss Functions

- Mean squared error: $\ell(x, t; w) = (y - t)^2$

```
torch.nn.MSELoss(size_average=None, reduce=None, reduction='mean')
```

- Minimum log-likelihood: $\ell(x, t; w) = \sum_{(x^{(i)}, t^{(i)}) \in \mathcal{D}} -\log p(t|x)$
  - Combined with binary classification: $\ell(x, t; w) = \sum_{(x^{(i)}, t^{(i)}) \in \mathcal{D}} \log(1 + \exp(-t^{(i)} w^T x^{(i)}))$
  - Combined with softmax: $\ell(x, t; w) = \sum_{(x^{(i)}, t^{(i)}) \in \mathcal{D}} \left(-w_{t^{(i)}}^T x + \log \sum_{c \in \{0,\dots,C-1\}} \exp(w_c^T x)\right)$

```
torch.nn.CrossEntropyLoss(weight=None, size_average=None, ignore_index=-100, reduce=None,
    reduction='mean', label_smoothing=0.0)[source]
```

- Cross Entropy Loss:
  - Linear (SVM formulation): $\ell(x, t; w) = \frac{|W[1:]|}{2} + C \sum \max\left(0, 1 - t^{(i)} \cdot W x^{(i)}\right)^2$
  - Logistic: $\ell(x, t; w) = -t \log y - (1 - t) \log(1 - y)$

Optimizers and torch.optim

In standard gradient descent, the update rule is: $\mathbf{w}_{k+1} = \mathbf{w}_k - \alpha \nabla f(\mathbf{w}_k)$. In *gradient descent with momentum*, we introduce a *velocity term* v_k : $v_{k+1} = \beta v_k - \alpha \nabla f(\mathbf{w}_k)$ and $\mathbf{w}_{k+1} = \mathbf{w}_k + v_{k+1}$ where: α is the learning rate, $\beta \in [0, 1]$ is the momentum coefficient, and v_k is the velocity term.

The following are some useful optimizers provided by the torch.optim library including:

- Stochastic gradient descent

```
torch.optim.SGD(params, lr=0.001, momentum=0, dampening=0, weight_decay=0, nesterov=False, *,
    maximize=False, foreach=None, differentiable=False, fused=None)[source]
```

PyTorch datasets

Required functions for dataset class:

- *__init__*: The *__init__* method is the constructor for the new dataset.
- *__len__*: The *__len__* method overrides the *len()* function in Python to determine the length of the dataset.
- *__getitem__*: The *__getitem__* method overloads the use of brackets to index items in a dataset.

There are lots of cool dataloader attributes and methods including:

- *batch_size*: number of examples in each batch or call to the dataloader
- *shuffle*: Boolean option to shuffle dataset each pass or *epoch* through the dataset
- *sampler*: *Sampler* object that specifies how data will be extracted from the dataset. For example, the *SubsetRandomSampler* allows us to specify indices within the larger dataset to sample at random.

Other useful equations

- Gradient descent: $\mathbf{w} \leftarrow \mathbf{w} - \alpha \frac{\partial \mathcal{E}}{\partial \mathbf{w}}$
- Closed form solution for linear regression: $W = (X^T X)^{-1} X^T T$
- L2 Regularization with MSE: $L(w) = \|y - Xw\|^2 + \lambda \|w\|_2^2$, closed form linear regression solutions: $W = (X^T X + \lambda I_d)^{-1} X^T y$
- Support Vector Machines - Margins at $WX = 1$ and $WX = -1$, border at $WX = 0$. Margin width $= 2/|W|$

Sample Code

Here is a sample, two-dimensional logistic classifier code:

```
import numpy as np
import matplotlib.pyplot as plt
import torch
import torch.nn as nn
from torch.utils.data import Dataset
from torch.utils.data import DataLoader
from torch.utils.data import SubsetRandomSampler

class LogisticRegression(nn.Module):
    def __init__(self, N):
        super().__init__()
        self.w = nn.Parameter(torch.ones(N))
        self.b = nn.Parameter(torch.zeros(1))

    def forward(self, x):
        return 1/(1+torch.exp(-(self.w@x+self.b)))

class TwoClassDataset(Dataset):
    # don't forget the self identifier!
    def __init__(self, N, sigma):
        self.N = N # number of data points per class
        self.sigma = sigma # standard deviation of each class cluster
        self.plus_class = self.sigma*torch.randn(N, 2) + torch.tensor([-1, 1])
        self.negative_class = self.sigma*torch.randn(N, 2) + torch.tensor([1, -1])
        self.data = torch.cat((self.plus_class, self.negative_class), dim=0)
        self.labels = torch.cat((torch.ones(self.N), torch.zeros(self.N)))

    def __len__(self):
        return len(self.labels)

    def __getitem__(self, idx):
        x = self.data[idx]
        y = self.labels[idx]
        return x, y # return input and output pair

N = 100
sigma = 1.5
dataset = TwoClassDataset(N, sigma)
plus_data = dataset.plus_class
negative_data = dataset.negative_class

# create indices for each split of dataset
N_train = 60
N_val = 20
N_test = 20
indices = np.arange(len(dataset))
np.random.shuffle(indices)
train_indices = indices[:N_train]
val_indices = indices[N_train:N_train+N_val]
test_indices = indices[N_train+N_val:]

# create dataloader for each split
batch_size = 8
train_loader = DataLoader(dataset, batch_size=batch_size, sampler=SubsetRandomSampler(train_indices)
    )
val_loader = DataLoader(dataset, batch_size=batch_size, sampler=SubsetRandomSampler(val_indices))
test_loader = DataLoader(dataset, batch_size=batch_size, sampler=SubsetRandomSampler(test_indices))

criterion = nn.BCELoss(reduction='mean') # binary cross-entropy loss, use mean loss
logreg_model = LogisticRegression(2) # initialize model
optimizer = torch.optim.SGD(logreg_model.parameters()) # initialize optimizer

n_epoch = 200 # number of passes through the training dataset
loss_values, train_accuracies, val_accuracies = [], [], []
for n in range(n_epoch):
    epoch_loss, epoch_acc = 0, 0
    for x_batch, y_batch in train_loader:
        optimizer.zero_grad()
        predictions = logreg_model(x_batch.unsqueeze(-1)).squeeze(-1)
        loss = criterion(predictions, y_batch)
        loss.backward()
        optimizer.step()
```