Homework 3: Bayesian Methods and Neural Networks

Introduction

This homework is about Bayesian methods and Neural Networks. Section 2.9 in the textbook as well as reviewing MLE and MAP will be useful for Q1. Chapter 4 in the textbook will be useful for Q2.

Please type your solutions after the corresponding problems using this LaTeX template, and start each problem on a new page.

Please submit the **writeup PDF to the Gradescope assignment 'HW3'**. Remember to assign pages for each question. **All plots you submit must be included in your writeup PDF.** We will not be checking your code / source files except in special circumstances.

Please submit your **LaTeXfile and code files to the Gradescope assignment 'HW3 - Supplemental'**.

Problem 1 (Bayesian Methods)

This question helps to build your understanding of making predictions with a maximum-likelihood estimation (MLE), a maximum a posterior estimator (MAP), and a full posterior predictive.

Consider a one-dimensional random variable $x = \mu + \epsilon$, where it is known that $\epsilon \sim N(0, \sigma^2)$. Suppose we have a prior $\mu \sim N(0, \tau^2)$ on the mean. You observe iid data $\{x_i\}_{i=1}^n$ (denote the data as D).

We derive the distribution of $x|D$ for you.

The full posterior predictive is computed using:

$$p(x|D) = \int p(x, \mu|D)d\mu = \int p(x|\mu)p(\mu|D)d\mu$$

One can show that, in this case, the full posterior predictive distribution has a nice analytic form:

$$x|D \sim \mathcal{N}\left(\frac{\sum_{x_i \in D} x_i}{n + \frac{\sigma^2}{\tau^2}}, (\frac{n}{\sigma^2} + \frac{1}{\tau^2})^{-1} + \sigma^2\right) \tag{1}$$

1. Derive the distribution of $\mu|D$.
2. In many problems, it is often difficult to calculate the full posterior because we need to marginalize out the parameters as above (here, the parameter is μ). We can mitigate this problem by plugging in a point estimate of μ^* rather than a distribution.

   a) Derive the MLE estimate μ_{MLE}.

   b) Derive the MAP estimate μ_{MAP}.

   c) What is the relation between μ_{MAP} and the mean of $x|D$?

   d) For a fixed value of $\mu = \mu^*$, what is the distribution of $x|\mu^*$? Thus, what is the distribution of $x|\mu_{MLE}$ and $x|\mu_{MAP}$?

   e) Is the variance of $x|D$ greater or smaller than the variance of $x|\mu_{MLE}$? What is the limit of the variance of $x|D$ as n tends to infinity? Explain why this is intuitive.
3. Let us compare μ_{MLE} and μ_{MAP}. There are three cases to consider:

   a) Assume $\sum_{x_i \in D} x_i = 0$. What are the values of μ_{MLE} and μ_{MAP}?

   b) Assume $\sum_{x_i \in D} x_i > 0$. Is μ_{MLE} greater than μ_{MAP}?

   c) Assume $\sum_{x_i \in D} x_i < 0$. Is μ_{MLE} greater than μ_{MAP}?
4. Compute:

$$\lim_{n \to \infty} \frac{\mu_{MAP}}{\mu_{MLE}}$$

Solution:

1. Using Bayes' rule, we have

$$p(\mu|D) \propto p(D|\mu)p(\mu)$$

and as the distribution of $D|\mu$ we have the following:

$$\begin{aligned}
p(D|\mu) = \prod_i p(x_i|\mu) &= (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_i (x_i-\mu)^2\right) \\
&\propto_\mu \exp\left(-\frac{1}{2\sigma^2}\sum_i (x_i-\mu)^2\right) \\
&= \exp\left(-\frac{1}{2\sigma^2}\sum_i (x_i^2 - 2x_i\mu + \mu^2)\right) \\
&= \exp\left(-\frac{1}{2\sigma^2}\left(\sum_i x_i^2 - 2\mu\sum_i x_i + \sum_i \mu^2\right)\right) \\
&\propto_\mu \exp\left(-\frac{1}{2\sigma^2}\left(-2\mu\sum_i x_i + \sum_i \mu^2\right)\right) \\
&= \exp\left(-\frac{n}{2\sigma^2}\left(-2\mu\bar{x} + \mu^2\right)\right)
\end{aligned}$$

We also know that the prior for μ is the following, since $\mu \sim \mathcal{N}(0, \tau^2)$

$$p(\mu) \propto \exp\left(-\frac{1}{2\tau^2}(\mu-0)^2\right) = \exp\left(-\frac{1}{2\tau^2}\mu^2\right)$$

Multiplying these two to get $P(\mu|D)$ yields the following

$$\begin{aligned}
p(\mu|D) \propto p(D|\mu)p(\mu) &= \exp\left(-\frac{n}{2\sigma^2}\left(-2\mu\bar{x} + \mu^2\right)\right)\exp\left(-\frac{1}{2\tau^2}\mu^2\right) \\
&= \exp\left(-\frac{n}{2\sigma^2}\left(-2\mu\bar{x} + \mu^2\right) - \frac{1}{2\tau^2}\mu^2\right)
\end{aligned}$$

The product of two Gaussian is a Gaussian, so we know that the above equation can be expressed as one. In order to find what the parameters σ_{new}, μ_{new} are of this new distribution, we can deduce from the coefficients of μ^2 and μ above what they are.

$$\exp\left(-\frac{n}{2\sigma^2}\left(-2\mu\bar{x} + \mu^2\right) - \frac{1}{2\tau^2}\mu^2\right) \stackrel{\text{def}}{=} \exp\left(-\frac{1}{2\sigma_{new}^2}(\mu^2 - 2\mu\mu_{new} + \mu_{new}^2)\right)$$

First we find σ_{new}^2:

$$\begin{aligned}
&\implies -\frac{1}{2\sigma_{new}^2}\mu^2 = -\frac{1}{2}\left(\frac{1}{\tau^2} + \frac{n}{\sigma^2}\right) \\
&\implies \sigma_{new}^2 = \left(\frac{1}{\tau^2} + \frac{n}{\sigma^2}\right)^{-1}
\end{aligned}$$

and also μ_{new}^2:

$$\implies -\frac{1}{2\sigma_{new}^2}(-2\mu\mu_{new}) = \mu\left(\frac{n\bar{x}}{\sigma^2}\right)$$

$$\implies \frac{\mu_{new}}{\sigma_{new}^2} = \frac{n\bar{x}}{\sigma^2}$$

$$\implies \mu_{new} = \frac{\frac{n\bar{x}}{\sigma^2}}{\frac{1}{\tau^2}+\frac{n}{\sigma^2}} = \frac{n\tau^2\bar{x}}{\sigma^2+n\tau^2}$$

Giving us our distribution of $\mu|D$:

$$\mu|D \sim \mathcal{N}\left(\frac{n\tau^2\bar{x}}{\sigma^2+n\tau^2}, \left(\frac{1}{\tau^2}+\frac{n}{\sigma^2}\right)^{-1}\right)$$

2. a)

$$\mu_{MLE} = \underset{\mu}{\operatorname{argmax}}\ p(D|\mu)$$

We find this value by calculating the derivative w.r.t. μ of $\ln p(D|\mu)$ and setting it to zero. We use the simplification of $p(D|\mu)$ from the previous problem.

$$p(D|\mu) \propto_\mu \exp\left(-\frac{n}{2\sigma^2}\left(-2\mu\bar{x}+\mu^2\right)\right)$$

$$\ln p(D|\mu) \propto_\mu -\frac{n}{2\sigma^2}\left(-2\mu\bar{x}+\mu^2\right)$$

$$\frac{\partial}{\partial\mu}\ln p(D|\mu) = \frac{n\bar{x}}{\sigma^2}-\frac{n\mu}{\sigma^2} = 0$$

$$\implies \mu_{MLE} = \bar{x}$$

b)

$$\mu_{MAP} = \underset{\mu}{\operatorname{argmax}}\ p(\mu|D) = \underset{\mu}{\operatorname{argmax}}\ p(D|\mu)p(\mu)$$

Again, we take the derivative of $\ln p(D|\mu)p(\mu)$ w.r.t. μ and set it to zero. We use the simplification of $p(D|\mu)p(\mu)$ from the previous problem.

$$p(D|\mu)p(\mu) \propto_\mu \exp\left(-\frac{n}{2\sigma^2}\left(-2\mu\bar{x}+\mu^2\right)-\frac{1}{2\tau^2}\mu^2\right)$$

$$\ln p(D|\mu)p(\mu) \propto_\mu -\frac{n}{2\sigma^2}\left(-2\mu\bar{x}+\mu^2\right)-\frac{1}{2\tau^2}\mu^2$$

$$\frac{\partial}{\partial\mu}\ln p(D|\mu)p(\mu) = \frac{n\bar{x}}{\sigma^2}-\frac{n\mu}{\sigma^2}-\frac{\mu}{\tau^2} = 0$$

$$\implies \frac{n\bar{x}}{\sigma^2}-\frac{n\mu_{MAP}}{\sigma^2}-\frac{\mu_{MAP}}{\tau^2} = 0$$

$$\implies \tau^2 n\bar{x}-\tau^2 n\mu_{MAP}-\sigma^2\mu_{MAP} = 0$$

$$\implies (\tau^2 n+\sigma^2)\mu_{MAP} = \tau^2 n\bar{x}$$

$$\implies \mu_{MAP} = \frac{\tau^2 n\bar{x}}{\tau^2 n+\sigma^2}$$

c) μ_{MAP} is equal to the mean of $x|D$

d) Since $x = \mu + \epsilon$ and $\epsilon \sim \mathcal{N}(0, \sigma^2)$, given a fixed μ^* then

$$c|\mu_{MLE} = \mathcal{N}(\mu_{MLE}, \sigma^2)$$

and

$$c|\mu_{MAP} = \mathcal{N}(\mu_{MAP}, \sigma^2)$$

e) The variance of $x|D$ is greater than the variance of $x|\mu_{MLE}$. The variance of $x|D$ as n tends to infinity is σ^2.

This is intuitive since, as $n \to \infty$, $x|D$ is distributed the same way as if μ was held constant ($\mathcal{N}(\mu^*, \sigma^2)$), since an infinitely large data set, by the law of large numbers, informs us of what our mean is. This gives us a variance of σ^2. When n is finite, we have to account for the variance of our prior for μ, given that that value is not a given. This increases the variance for each our x's.

3. We have $\sum_{x_i \in D} x_i = n\bar{x}$

   a) If $n\bar{x} = 0$, then $\mu_{MLE} = 0$ and $\mu_{MAP} = 0$.

   b) If $n\bar{x} > 0$, then $\mu_{MLE} > \mu_{MAP}$.

   c) If $n\bar{x} > 0$, then $\mu_{MLE} < \mu_{MAP}$.

4. We have $\mu_{MAP} = \frac{\tau^2 n\bar{x}}{\tau^2 n + \sigma^2} = \frac{\bar{x}}{1 + \frac{\sigma^2}{\tau^2 n}}$.

$$\lim_{n\to\infty} \frac{\mu_{MAP}}{\mu_{MLE}} = \lim_{n\to\infty} \frac{\frac{\bar{x}}{1+\frac{\sigma^2}{\tau^2 n}}}{\bar{x}} = \lim_{n\to\infty} \frac{1}{1 + \frac{\sigma^2}{\tau^2 n}} = 1$$

Problem 2 (Bayesian Frequentist Reconciliation)

In this question, we connect the Bayesian version of regression with the frequentist view we have seen in the first week of class by showing how appropriate priors could correspond to regularization penalities in the frequentist world, and how the models can be different.

Suppose we have a $(p+1)$-dimensional labelled dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$. We can assume that y_i is generated by the following random process:

$$y_i = \mathbf{w}^\top \mathbf{x}_i + \epsilon_i$$

where all $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ are iid. Using matrix notation, we denote

$$\begin{aligned}
\mathbf{X} &= \begin{bmatrix} \mathbf{x}_1 & \dots & \mathbf{x}_N \end{bmatrix}^\top \in \mathbb{R}^{N \times p} \\
\mathbf{y} &= \begin{bmatrix} y_1 & \dots & y_N \end{bmatrix}^\top \in \mathbb{R}^N \\
\epsilon &= \begin{bmatrix} \epsilon_1 & \dots & \epsilon_N \end{bmatrix}^\top \in \mathbb{R}^N.
\end{aligned}$$

Then we can write have $\mathbf{y} = \mathbf{X}\mathbf{w} + \epsilon$. Now, we will suppose that $\mathbf{w}$ is random as well as our labels! We choose to impose the Laplacian prior $p(\mathbf{w}) = \frac{1}{2\tau} \exp\left(-\frac{\|\mathbf{w}-\mu\|_1}{\tau}\right)$, where $\|\mathbf{w}\|_1 = \sum_{i=1}^p |w_i|$ denotes the L^1 norm of $\mathbf{w}$, μ the location parameter, and τ is the scale factor.

1. Compute the posterior distribution $p(\mathbf{w}|\mathbf{X}, \mathbf{y})$ of $\mathbf{w}$ given the observed data $\mathbf{X}, \mathbf{y}$, up to a normalizing constant. You **do not** need to simplify the posterior to match a known distribution.
2. Determine the MAP estimate $\mathbf{w}_{\text{MAP}}$ of $\mathbf{w}$. You may leave the answer as the solution to an equation. How does this relate to regularization in the frequentist perspective? How does the scale factor τ relate to the corresponding regularization parameter λ? Provide intuition on the connection to regularization, using the prior imposed on $\mathbf{w}$.
3. Based on the previous question, how might we incorporate prior expert knowledge we may have for the problem? For instance, suppose we knew beforehand that $\mathbf{w}$ should be close to some vector $\mathbf{v}$ in value. How might we incorporate this in the model, and explain why this makes sense in both the Bayesian and frequentist viewpoints.
4. As τ decreases, what happens to the entries of the estimate $\mathbf{w}_{\text{MAP}}$? What happens in the limit as $\tau \to 0$?
5. Consider the point estimate $\mathbf{w}_{\text{mean}}$, the mean of the posterior $\mathbf{w}|\mathbf{X}, \mathbf{y}$. Further, assume that the model assumptions are correct. That is, $\mathbf{w}$ is indeed sampled from the posterior provided in subproblem 1, and that $y|\mathbf{x}, \mathbf{w} \sim \mathcal{N}(\mathbf{w}^T\mathbf{x}, \sigma^2)$. Suppose as well that the data generating processes for $\mathbf{x}, \mathbf{w}, y$ are all independent (note that $\mathbf{w}$ is random!). Between the models with estimates $\mathbf{w}_{\text{MAP}}$ and $\mathbf{w}_{\text{mean}}$, which model would have a lower expected test MSE, and why? Assume that the data generating distribution for $\mathbf{x}$ has mean zero, and that distinct features are independent and each have variance 1.[a]

[a] The unit variance assumption simplifies computation, and is also commonly used in practical applications.

Solution:

1. Using Bayes' rule, we have

$$p(\mathbf{w}|X, y) \propto_w p(\mathbf{y}|\mathbf{X}, \mathbf{w}) p(\mathbf{w})$$

$$= \left(\prod_i p(y_i|\mathbf{x}_i, \mathbf{w}) \right) p(\mathbf{w})$$

We are assuming that $y_i = \mathbf{w}^\top \mathbf{x}_i + \epsilon_i$ with $\epsilon \sim \mathcal{N}(0, \sigma^2)$, so we have $y_i \sim \mathcal{N}(\mathbf{w}^\top \mathbf{x}_i, \sigma^2)$. Plugging this in along with the distribution of our prior $p(\mathbf{w})$ defined in the problem prompt above, we have

$$p(\mathbf{w}|X, y) \propto_w \left(\prod_i (2\pi\sigma^2)^{-1/2} \exp\left(-\frac{1}{2\sigma^2}(y_i - \mathbf{w}^\top \mathbf{x}_i)^2 \right) \right) \left(\frac{1}{2\tau} \exp\left(-\frac{\|\mathbf{w} - \mu\|_1}{\tau} \right) \right)$$

$$\propto_w \left(\prod_i \exp\left(-\frac{1}{2\sigma^2}(y_i - \mathbf{w}^\top \mathbf{x}_i)^2 \right) \right) \exp\left(-\frac{\|\mathbf{w} - \mu\|_1}{\tau} \right)$$

$$= \exp\left(-\frac{1}{2\sigma^2} \sum_i (y_i - \mathbf{w}^\top \mathbf{x}_i)^2 \right) \exp\left(-\frac{\|\mathbf{w} - \mu\|_1}{\tau} \right)$$

$$p(\mathbf{w}|X, y) \propto_w \exp\left(-\frac{1}{2\sigma^2} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2 - \frac{1}{\tau} \|\mathbf{w} - \mu\|_1 \right)$$

2. The MAP estimate is defined as the following

$$\mathbf{w}_{MAP} = \underset{\mu}{\text{argmax}}\ p(\mathbf{w}|\mathbf{X}, \mathbf{y}) = \underset{\mu}{\text{argmax}}\ p(\mathbf{y}|\mathbf{X}, \mathbf{w}) p(\mathbf{w}) = \underset{\mu}{\text{argmax}}\ \ln p(\mathbf{y}|\mathbf{X}, \mathbf{w}) p(\mathbf{w})$$

$$= \underset{\mu}{\text{argmax}}\ -\frac{1}{2\sigma^2} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2 - \frac{1}{\tau} \|\mathbf{w} - \mu\|_1$$

$$= \underset{\mu}{\text{argmin}}\ \frac{1}{2\sigma^2} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2 + \frac{1}{\tau} \|\mathbf{w} - \mu\|_1$$

This estimate relates to regularization in the frequentist perspective in that it contains a 'regularization term'; the difference between $\mathbf{w}_{MAP}$ and $\mathbf{w}_{MLE}$ is that the MAP estimate, which incorporates knowledge on the value of $\mathbf{w}$ before the data is taken into account, has an added term ($\frac{1}{\tau}\|\mathbf{w} - \mu\|_1$) that brings the estimate closer to what the prior knowledge says it will be close to. This is exactly what regularization does, only that regularization (to the extent of what is covered in this class) tries to keep the weights close to zero, by some measure of the norm of the weights. The scale factor τ relates to the corresponding regularization parameter λ in a direct way; increasing/decreasing τ has the same effect on the resulting optimal weights as decreasing/increasing λ in the frequentist regularization term; $\lambda \sim \frac{1}{\tau}$.

3. Prior expert knowledge can be incorporated into a problem by specifying a prior distribution for the weights and solving for the MAP rather than the MLE. If what we know is that our weight vector $\mathbf{w}$ is going to be close to a vector $\mathbf{v}$ then we assume a prior $\mathbf{w} \sim \mathcal{N}(\mathbf{v}, \sigma^2)$ for some σ^2, or a prior $\mathbf{w} \sim \text{Laplace}(\mathbf{v}, b)$ for some b. In the equation to be optimized, σ^2 or b will play a similar role to that of $1/\lambda$ in a frequentist regularization term. This makes sense in the Bayesian sense, as we are specifying knowledge of a prior distribution whose mean we define, and this makes sense in the Frequentist sense, as we are penalizing weights that are far from the mean of this prior.

4. We have

$$\mathbf{w}_{MAP} = \underset{\mu}{\operatorname{argmin}} \frac{1}{2\sigma^2}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2 + \frac{1}{\tau}\|\mathbf{w} - \mu\|_1$$

   As τ decreases, the entries of the estimate change so that $\mathbf{w}_{MAP}$ is closer μ. As τ approaches infinity, $\mathbf{w}_{MAP}$ approaches μ.

5. $\mathbf{w}_{MAP}$ would have a lower expected MSE, given that [TODO]

Problem 3 (Neural Net Optimization)

In this problem, we will take a closer look at how gradients are calculated for backprop with a simple multi-layer perceptron (MLP). The MLP will consist of a first fully connected layer with a sigmoid activation, followed by a one-dimensional, second fully connected layer with a sigmoid activation to get a prediction for a binary classification problem. Assume bias has not been merged. Let:

- $\mathbf{W}_1$ be the weights of the first layer, $\mathbf{b}_1$ be the bias of the first layer.
- $\mathbf{W}_2$ be the weights of the second layer, $\mathbf{b}_2$ be the bias of the second layer.

The described architecture can be written mathematically as:

$$\hat{y} = \sigma(\mathbf{W}_2 \left[\sigma \left(\mathbf{W}_1 \mathbf{x} + \mathbf{b}_1\right)\right] + \mathbf{b}_2)$$

where $\hat{y}$ is a scalar output of the net when passing in the single datapoint $\mathbf{x}$ (represented as a column vector), the additions are element-wise additions, and the sigmoid is an element-wise sigmoid.

1. Let:
   - N be the number of datapoints we have
   - M be the dimensionality of the data
   - H be the size of the hidden dimension of the first layer. Here, hidden dimension is used to describe the dimension of the resulting value after going through the layer. Based on the problem description, the hidden dimension of the second layer is 1.

   Write out the dimensionality of each of the parameters, and of the intermediate variables:

   $$\mathbf{a}_1 = \mathbf{W}_1 \mathbf{x} + \mathbf{b}_1, \qquad \mathbf{z}_1 = \sigma(\mathbf{a}_1)$$
   $$a_2 = \mathbf{W}_2 \mathbf{z}_1 + \mathbf{b}_2, \qquad \hat{y} = z_2 = \sigma(a_2)$$

   and make sure they work with the mathematical operations described above.

2. We will derive the gradients for each of the parameters. The gradients can be used in gradient descent to find weights that improve our model's performance. For this question, assume there is only one datapoint $\mathbf{x}$, and that our loss is $L = -(y \log(\hat{y}) + (1 - y) \log(1 - \hat{y}))$. For all questions, the chain rule will be useful.
   (a) Find $\frac{\partial L}{\partial b_2}$.
   (b) Find $\frac{\partial L}{\partial W_2^h}$, where W_2^h represents the hth element of $\mathbf{W}_2$.
   (c) Find $\frac{\partial L}{\partial b_1^h}$, where b_1^h represents the hth element of $\mathbf{b}_1$. (*Hint: Note that only the hth element of $\mathbf{a}_1$ and $\mathbf{z}_1$ depend on b_1^h - this should help you with how to use the chain rule.)
   (d) Find $\frac{\partial L}{\partial W_1^{h,m}}$, where $W_1^{h,m}$ represents the element in row h, column m in $\mathbf{W}_1$.

Solution:

1. The dimension of $\mathbf{a}_1$ is H.

   The dimension of $\mathbf{z}_1$ is H.

   The dimension of z_1 is 1.

   The dimension of $\hat{y}$ is 1.

2. a) We have

$$L = -y \log(\hat{y}) - (1-y)\log(1-\hat{y})$$
$$\hat{y} = \sigma(a_2)$$
$$a_2 = \mathbf{W}_2\mathbf{z}_1 + \mathbf{b}_2$$

   and thus

$$\frac{\partial L}{\partial \hat{y}} = -\frac{y}{\hat{y}} + \frac{1-y}{1-\hat{y}}$$
$$\frac{\partial \hat{y}}{\partial a_2} = \sigma(a_2)(1-\sigma(a_2)) = \hat{y}(1-\hat{y})$$
$$\frac{\partial a_2}{\partial b_2} = 1$$

   Using the chain rule, this gives us

$$\frac{\partial L}{\partial b_2} = \frac{\partial L}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial a_2}\frac{\partial a_2}{\partial b_2} = \frac{\partial L}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial a_2}$$
$$= \left(-\frac{y}{\hat{y}} + \frac{1-y}{1-\hat{y}}\right)(\hat{y}(1-\hat{y}))$$
$$= -y\frac{1}{\hat{y}}\hat{y}(1-\hat{y}) + (1-y)\frac{1}{1-\hat{y}}\hat{y}(1-\hat{y})$$
$$= -y(1-\hat{y}) + (1-y)\hat{y}$$
$$= -y + y\hat{y} + \hat{y} - y\hat{y}$$
$$\boxed{\frac{\partial L}{\partial b_2} = \hat{y} - y}$$

   b) We introduce a new partial derivative:

$$\frac{\partial a_2}{\partial W_2^h} = z_1^h$$

   Using the chain rule, then, we have

$$\frac{\partial L}{\partial W_2^h} = \frac{\partial L}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial a_2}\frac{\partial a_2}{\partial W_2^h}$$

$$\boxed{\frac{\partial L}{\partial W_2^h} = (\hat{y} - y)(z_1^h)}$$

c) We have the following relationships between variables

$$L = -y \log(\hat{y}) - (1 - y) \log(1 - \hat{y})$$
$$\hat{y} = \sigma(a_2)$$
$$a_2 = \mathbf{W}_2 \mathbf{z}_1 + \mathbf{b}_2 = \sum_h (W_2^h z_1^h + b_2^h)$$
$$z_1^h = \sigma(a_1^h)$$
$$a_1^h = \mathbf{W}_1^h \mathbf{x} + b_1^h$$

,

thus we have the following partial derivatives

$$\frac{\partial L}{\partial \hat{y}} = -\frac{y}{\hat{y}} + \frac{1 - y}{1 - \hat{y}}$$
$$\frac{\partial \hat{y}}{\partial a_2} = \sigma(a_2)(1 - \sigma(a_2)) = \hat{y}(1 - \hat{y})$$
$$\frac{\partial a_2}{\partial z_1^h} = W_2^h$$
$$\frac{\partial z_1^h}{\partial a_1^h} = \sigma(a_1^h)(1 - \sigma(a_1^h)) = z_1^h(1 - z_1^h)$$
$$\frac{\partial a_1^h}{\partial b_1^h} = 1$$

Using the chain rule, then, we have

$$\frac{\partial L}{\partial b_1^h} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial a_2} \frac{\partial a_2}{\partial z_1^h} \frac{\partial z_1^h}{\partial a_1^h} \frac{\partial a_1^h}{\partial b_1^h}$$

$$\boxed{\frac{\partial L}{\partial b_1^h} = (\hat{y} - y)(W_2^h)(z_1^h(1 - z_1^h))}$$

d) As above, we have the following partial derivatives

$$\frac{\partial L}{\partial \hat{y}} = -\frac{y}{\hat{y}} + \frac{1 - y}{1 - \hat{y}}$$
$$\frac{\partial \hat{y}}{\partial a_2} = \sigma(a_2)(1 - \sigma(a_2)) = \hat{y}(1 - \hat{y})$$
$$\frac{\partial a_2}{\partial z_1^h} = W_2^h$$

$$\frac{\partial z_1^h}{\partial a_1^h} = \sigma(a_1^h)(1 - \sigma(a_1^h)) = z_1^h(1 - z_1^h)$$

and we are introducing a new one:

$$\frac{\partial a_1^h}{\partial W_1^{h,m}} = x_m$$

since $a_1 = \mathbf{W}_1\mathbf{x} + \mathbf{b}_1$ and the coefficient of $W_1^{h,m}$ in this equation is x_m .
Using the chain rule, then, we have

$$\frac{\partial L}{\partial b_1^h} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial a_2} \frac{\partial a_2}{\partial z_1^h} \frac{\partial z_1^h}{\partial a_1^h} \frac{\partial a_1^h}{\partial W_1^{h,m}}$$

$$\boxed{\frac{\partial L}{\partial W_1^{h,m}} = (\hat{y} - y)(W_2^h)(z_1^h(1 - z_1^h))(x_m)}$$

Problem 4 (Modern Deep Learning Tools: PyTorch)

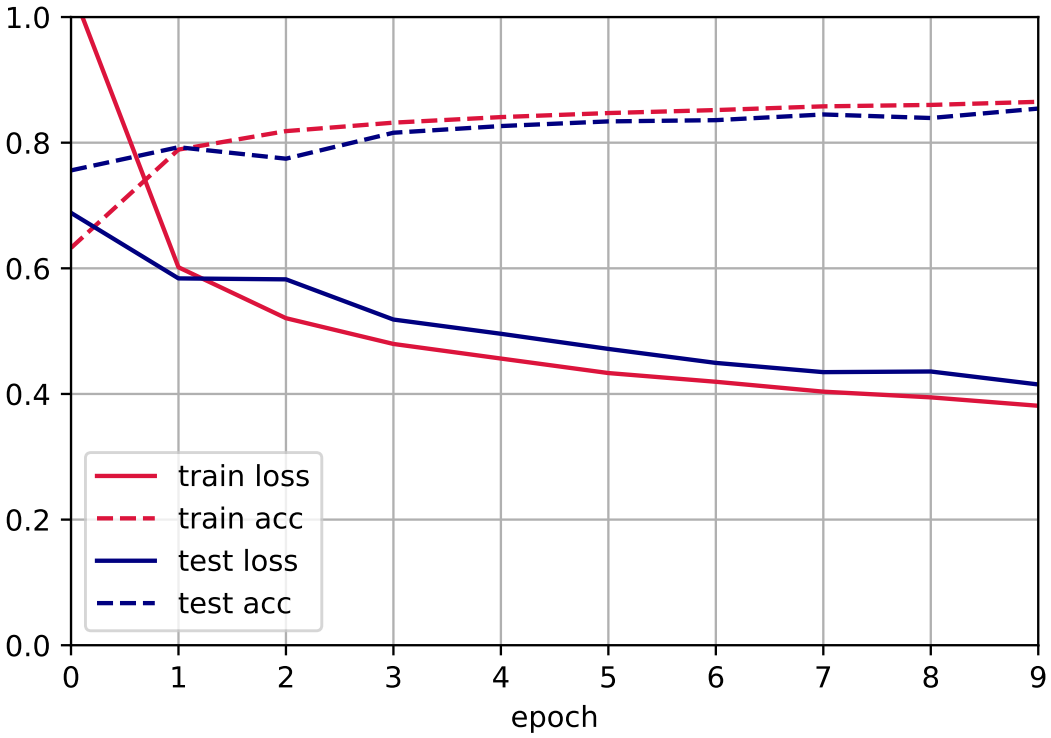
In this problem, you will learn how to use PyTorch. This machine learning library is massively popular and used heavily throughout industry and research. In `T3_P3.ipynb` you will implement an MLP for image classification from scratch. Copy and paste code solutions below and include a final graph of your training progress. Also submit your completed `T3_P3.ipynb` file.

You will recieve no points for code not included below.

You will recieve no points for code using built-in APIs from the `torch.nn` library.

Solution:

Plot:



Code:

```
n_inputs = 784 # dimension of flattened image (28 * 28)
n_hiddens = 256 # dimension of hidden layer
n_outputs = 10 # number of label possibilities

W1 = torch.nn.Parameter(0.01 * torch.randn(size=(n_inputs, n_hiddens)))
b1 = torch.nn.Parameter(torch.zeros(size=(n_hiddens, )))
W2 = torch.nn.Parameter(0.01 * torch.randn(size=(n_hiddens, n_outputs)))
b2 = torch.nn.Parameter(torch.zeros(size=(n_outputs, )))

def relu(x):
    return torch.max(torch.stack([x, torch.zeros_like(x)], dim=1), dim=1)[0]

def softmax(x):
    return torch.div(torch.exp(X), torch.sum(torch.exp(X), 1).view(-1, 1))

def net(X):
    X = torch.flatten(X, start_dim=1)
    H = relu(X @ W1 + b1)
    O = softmax(H @ W2 + b2)
    return O

def cross_entropy(y_hat, y):
    return -torch.log(y_hat)[range(len(y_hat)), y]

def sgd(params, lr=0.1):
    with torch.no_grad():
        for w in params:
            w -= lr * w.grad
            w.grad.zero_()

def train(net, params, train_iter, loss_func=cross_entropy, updater=sgd):
    for _ in range(epochs):
        for X, y in train_iter:
            y_hat = net(X)
            l = loss_func(y_hat, y).mean()
            l.backward()
            updater(params, lr)
```

Name

Rodney Lafuente Mercado

Collaborators and Resources

Whom did you work with, and did you use any resources beyond cs181-textbook and your notes?

I did not collaborate with anyone or use outside sources other than PyTorch documentation

Calibration

Approximately how long did this homework take you to complete (in hours)?

Easily more than 15