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:

1.1

We want to derive the distribution of $\mu|D$. We will do this using Bayes theorem, substituting the Gaussian PDF, and showing that the resulting distribution is Gaussian with a new mean and variance.

To begin, by Bayes theorem we have the following.

$$p(\mu|D) \propto p(D|\mu)p(\mu) = \left[\prod_{i=1}^{N} p(x_i|\mu)\right] p(\mu)$$

Since D is fixed, we can treat $p(D)$ (which would have been in the denominator) as a normalizing constant and so can ignore it. Now note that since we have $x = \mu + \epsilon$ and $\epsilon \sim \mathcal{N}(0, \sigma^2)$, then we know that $x \sim \mathcal{N}(\mu, \sigma^2)$, and from the question we have $\mu \sim \mathcal{N}(0, \tau^2)$. So we can substitute the PDFs of these distributions into the equation above.

$$p(\mu|D) \propto p(D|\mu)p(\mu) = \left[\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2}\right)\right] \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{1}{2}\frac{\mu^2}{\tau^2}\right)$$

For the rest of the equation, we will simplify this expression keeping only terms that rely on μ. This is because we assumed everything else is fixed, and so can ignore constant terms. Now let us simplify.

$$\begin{aligned}
p(\mu|D) &\propto \left[\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2}\right)\right] \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{1}{2}\frac{\mu^2}{\tau^2}\right) \\
&\propto \left[\prod_{i=1}^{N} \exp\left(-\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2}\right)\right] \exp\left(-\frac{1}{2}\frac{\mu^2}{\tau^2}\right) \\
&\propto \exp\left(-\frac{1}{2}\left[\frac{\mu^2}{\tau^2} + \sum_{i=1}^{N}\frac{(x_i-\mu)^2}{\sigma^2}\right]\right) \\
&\propto \exp\left(-\frac{1}{2}\left[\frac{\mu^2}{\tau^2} + \frac{1}{\sigma^2}\sum_{i=1}^{N} x_i^2 - 2\mu x_i + \mu^2\right]\right) \\
&\propto \exp\left(-\frac{1}{2}\left[\frac{\mu^2}{\tau^2} + \frac{1}{\sigma^2}\sum_{i=1}^{N} -2\mu x_i + \mu^2\right]\right) \\
&\propto \exp\left(-\frac{1}{2}\left(\frac{\mu^2}{\tau^2} - \frac{2\mu\sum_{i=1}^{N} x_i}{\sigma^2} + \frac{N\mu^2}{\sigma^2}\right)\right)
\end{aligned}$$

We can now simplify by completing the squares, again keeping only terms that rely on μ. Let us also

substitute the expression for the sample mean $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$.

$$\propto \exp\left(-\frac{1}{2}\left(\frac{\mu^2}{\tau^2} - \frac{2N\bar{x}\mu}{\sigma^2} + \frac{N\mu^2}{\sigma^2}\right)\right)$$
$$\propto \exp\left(-\frac{1}{2}\left(\left(\frac{1}{\tau^2} + \frac{N}{\sigma^2}\right)\mu^2 - \frac{2N\bar{x}}{\sigma^2}\mu\right)\right)$$
$$\propto \exp\left(-\frac{1}{2}\left(\frac{1}{\tau^2} + \frac{N}{\sigma^2}\right)\left(\mu^2 - \frac{2N\bar{x}}{\sigma^2\left(\frac{1}{\tau^2} + \frac{N}{\sigma^2}\right)}\mu\right)\right)$$
$$\propto \exp\left(-\frac{1}{2}\left(\frac{1}{\tau^2} + \frac{N}{\sigma^2}\right)\left(\mu - \frac{N\bar{x}}{\sigma^2\left(\frac{1}{\tau^2} + \frac{N}{\sigma^2}\right)}\right)^2\right)$$
$$\propto \exp\left(-\frac{1}{2}\left(\frac{1}{\tau^2} + \frac{N}{\sigma^2}\right)\left(\mu - \frac{N\bar{x}}{\frac{\sigma^2}{\tau^2} + N}\right)^2\right)$$

We can see that this matches the Gaussian PDF with the parameters for mean and variance below.

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

1.2(a)

We want to find the MLE estimate μ_{MLE}. Let us start with the expression for the likelihood.

$$L(\mu; D) = p(D|\mu) = p(\{x_i\}|\mu)$$

From the question, since $x = \mu + \epsilon$ and $\epsilon \sim \mathcal{N}(0, \sigma^2)$ then $x_i|\mu \sim \mathcal{N}(\mu, \sigma^2)$. We can use that fact here.

$$L(\mu; D) = \prod_{i=1}^{N} \mathcal{N}(x_i|\mu, \sigma^2)$$
$$= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(x_i - \mu)^2}{\sigma^2}\right)$$

Now we can take the log to get the log likelihood.

$$\ell(\mu; D) = -\frac{N}{2}\ln\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i - \mu)^2$$

Then we take the derivative with respect to μ, set the LHS to 0, and solve for μ.

$$\nabla_\mu \ell(\mu; D) = -\frac{1}{\sigma^2}\sum_{i=1}^{N}(x_i - \mu)(-1)$$
$$0 = \sum_{i=1}^{N}(x_i - \mu)$$
$$0 = -N\mu + \sum_{i=1}^{N} x_i$$
$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

Notice that this is the sample mean $\bar{x}$, and so $\boxed{\mu_{MLE} = \bar{x}}$.

1.2(b)

We want to find the MAP estimate μ_{MAP}. We know that μ_{MAP} is given by the following equation.

$$\mu_{MAP} = \text{argmax}_{\mu} p(\mu|D)$$

We find the MAP estimate by first applying Bayes theorem, then taking the gradient with respect to μ, setting to 0, and solving for μ. By Bayes we have the following.

$$p(\mu|D) = p(D|\mu)p(\mu) = \left[\prod_{i=1}^{N} p(x_i|\mu)\right] p(\mu)$$

We know from 1.1 and 1.2(a) that $x_i|\mu \sim \mathcal{N}(\mu, \sigma^2)$ and our prior $\mu \sim \mathcal{N}(0, \tau^2)$. So we can sub in the Gaussian PDFs.

$$p(\mu|D) = \left[\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2}\right)\right] \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{1}{2}\frac{\mu^2}{\tau^2}\right)$$

We can then take the log and simplify.

$$\ln(p(\mu|D)) = -\frac{N}{2}\ln\big(2\pi\sigma^2\big) - \frac{1}{2}\ln\big(2\pi\tau^2\big) - \frac{\mu^2}{2\tau^2} - \frac{1}{2}\sum_{i=1}^{N}\left(\frac{1}{\sigma^2}(x_i-\mu)^2\right)$$

Now we take the gradient with respect to μ and set the LHS to 0.

$$\begin{aligned}
\nabla_u \ln(p(\mu|D)) &= -\frac{\mu}{\tau^2} - \sum_{i=1}^{N}\frac{1}{\sigma^2}(x_i-\mu)(-1) \\
0 &= -\frac{\mu}{\tau^2} + \sum_{i=1}^{N}\frac{1}{\sigma^2}(x_i-\mu) \\
0 &= -\frac{1}{\tau^2}\mu - \frac{N}{\sigma^2}\mu + \frac{\sum_{i=1}^{N} x_i}{\sigma^2} \\
\left(\frac{N}{\sigma^2} + \frac{1}{\tau^2}\right)\mu &= \frac{\sum_{i=1}^{N} x_i}{\sigma^2}
\end{aligned}$$

We use the notation for the sample mean $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$ and finally solve for μ.

$$\mu = \frac{N\bar{x}}{\sigma^2(\frac{1}{\tau^2} + \frac{N}{\sigma^2})} = \frac{N\bar{x}}{\frac{\sigma^2}{\tau^2} + N}$$

And so we have found $\boxed{\mu_{MAP} = \frac{N\bar{x}}{\frac{\sigma^2}{\tau^2} + N}}$ as given by the expression above.

1.2(c)

We can see that μ_{MAP} and the mean of $x|D$ are the same. This makes sense: since $x|D$ is normally distributed and μ_{MAP} is the posterior mode, then the mean will be the same as the mode.

1.2(d)

Since $x = \mu + \epsilon$ and $\epsilon \sim \mathcal{N}(0, \sigma^2)$, then we must have the following distribution.

$$x|\mu^* \sim \mathcal{N}(\mu^*, \sigma^2)$$

And so for μ_{MLE} and μ_{MAP} we also have the following distributions.

$$x|\mu_{MLE} \sim \mathcal{N}(\mu_{MLE}, \sigma^2) \quad , \quad x|\mu_{MAP} \sim \mathcal{N}(\mu_{MAP}, \sigma^2)$$

Or equivalently by substituting our results from above:

$$x|\mu_{MLE} \sim \mathcal{N}(\bar{x}, \sigma^2) \ , \ x|\mu_{MAP} \sim \mathcal{N}\left(\frac{N\bar{x}}{\frac{\sigma^2}{\tau^2}+N}, \sigma^2\right)$$

1.2(e)

The variance of $x|D$ is greater than the variance of $x|\mu_{MLE}$. From 1.2(d) we know that the variance of $x|\mu_{MLE}$ is σ^2. Using the variance of $x|D$ we have the following.

$$(\frac{n}{\sigma^2} + \frac{1}{\tau^2})^{-1} + \sigma^2 > \sigma^2$$

The limit of the variance of $x|D$ as $n \to \infty$ is σ^2. This is intuitive because as size of the data increases we expect the data to better represent the ground truth variance and so the variance would converge to σ^2.

1.3

For all parts of this question we will use the following results from 1.2.

$$\mu_{MLE} = \bar{x} \ , \ \mu_{MAP} = \frac{N\bar{x}}{\frac{\sigma^2}{\tau^2}+N}$$

(a) Assuming $\sum_{x_i \in D} x_i = 0$ then $\boxed{\mu_{MLE} = 0 \text{ and } \mu_{MAP} = 0}$.

(b) Assuming $\sum_{x_i \in D} x_i > 0$ then $\boxed{\mu_{MLE} > \mu_{MAP}}$.

(c) Assuming $\sum_{x_i \in D} x_i < 0$ then $\boxed{\mu_{MLE} < \mu_{MAP}}$.

1.4

We want to compute the following expression.

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

We use our results from 1.2 and simplify.

$$\begin{aligned}
\lim_{N\to\infty} \frac{\mu_{MAP}}{\mu_{MLE}} &= \lim_{N\to\infty} \frac{\frac{N\bar{x}}{\sigma^2/\tau^2+N}}{\bar{x}} \\
&= \lim_{N\to\infty} \frac{N\bar{x}}{\bar{x}(\sigma^2/\tau^2+N)} \\
&= \lim_{N\to\infty} \frac{N}{\sigma^2/\tau^2+N} \\
&= 1
\end{aligned}$$

And so the value of the expression is $\boxed{1}$.

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 D-dimensional labelled dataset $\mathcal{D} = \{(y_i, \mathbf{x}_i)\}_{i=1}^N$. 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

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

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 providing the point estimate $\mathbf{w}_{\text{mean}}$ is based on the mean of the posterior $\mathbf{w}|\mathbf{X}, \mathbf{y}$. Provide an expression for the estimate $\mathbf{w}_{\text{mean}}$, up to a normalizing constant. Based on this expression, which model (original or Bayesian) would we expect to take longer to train? Further, **if** the model assumptions are correct (i.e. there indeed is a linear relationship and $\mathbf{w}$ is indeed sample from a Laplace distribution), which model would we expect to have a lower test MSE?

Solution:

2.1

We want to compute the posterior distribution $p(\mathbf{w}|\mathbf{X}, \mathbf{y})$ up to a normalizing constant. By Bayes we have the following.

$$\begin{aligned} p(\mathbf{w}|\mathbf{X}, \mathbf{y}) &= \frac{p(\mathbf{y}|\mathbf{w}, \mathbf{X})p(\mathbf{w}|\mathbf{X})}{p(\mathbf{y}|\mathbf{X})} \\ &\propto p(\mathbf{y}|\mathbf{w}, \mathbf{X})p(\mathbf{w}) \\ &= \left[\prod_{i=1}^{N} p(y_i|\mathbf{w}, \mathbf{x}_i)\right] p(\mathbf{w}) \end{aligned}$$

This assumes that $p(\mathbf{y}|\mathbf{X})$ is a normalizing constant and $p(\mathbf{w}|\mathbf{X}) = p(\mathbf{w})$ since $\mathbf{X}$ is fixed.

Since $\mathbf{y} = \mathbf{X}\mathbf{w} + \epsilon$ and $\epsilon \sim \mathcal{N}(0, \sigma^2)$, then we know that $y_i \sim \mathcal{N}(\mathbf{w}^T\mathbf{x}_i, \sigma^2)$. Plugging this into our equation above as well as our value of $p(\mathbf{w})$ given in the question, we get the following expression.

$$\begin{aligned} p(\mathbf{w}|\mathbf{X}, \mathbf{y}) &\propto \left[\prod_{i=1}^{N} p(y_i|\mathbf{w}, \mathbf{x}_i)\right] p(\mathbf{w}) \\ &= \left[\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(y_i - \mathbf{w}^T\mathbf{x}_i)^2}{\sigma^2}\right)\right] \left[\frac{1}{2\tau}\exp\left(-\frac{||\mathbf{w} - \mu||_1}{\tau}\right)\right] \end{aligned}$$

Now we remove normalizing constants and simplify.

$$\boxed{p(\mathbf{w}|\mathbf{X}, \mathbf{y}) \propto \exp\left(-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{N}(y_i - \mathbf{w}^T\mathbf{x}_i)^2\right] - \frac{1}{\tau}||\mathbf{w} - \mu||_1\right)}$$

And so we have computed the posterior distribution up to a normalizing constant.

2.2

First we want to determine the MAP estimate $\mathbf{w}_{MAP}$ of $\mathbf{w}$. We have the following expression for the MAP estimate.

$$\begin{aligned} \mathbf{w}_{MAP} &= \text{argmax}_{\mathbf{w}} p(\mathbf{w}|\mathbf{X}, \mathbf{y}) \\ &= \text{argmax}_{\mathbf{w}} p(\mathbf{y}|\mathbf{w}, \mathbf{X})p(\mathbf{w}) \end{aligned}$$

We use the distributions for $y_i|\mathbf{w}, \mathbf{X}$ and $\mathbf{w}$ from 2.1 and plug the PDFs into this expression.

$$p(\mathbf{w}|\mathbf{X}, \mathbf{y}) = \left[\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(y_i - \mathbf{w}^T\mathbf{x}_i)^2}{\sigma^2}\right)\right] \left[\frac{1}{2\tau}\exp\left(-\frac{||\mathbf{w} - \mu||_1}{\tau}\right)\right]$$

We take the negative log and remove normalizing constants. Taking the negative log means that the ultimate expression will be an argmin instead of an argmax.

$$\begin{aligned} -\ln(p(\mathbf{w}|\mathbf{X}, \mathbf{y})) &= \frac{N}{2}\ln\left(2\pi\sigma^2\right) + \frac{1}{2\sigma^2}\left[\sum_{i=1}^{N}(y_i - \mathbf{w}^T\mathbf{x}_i)^2\right] + N\ln(2\tau) + \frac{1}{\tau}||\mathbf{w} - \mu||_1 \\ &\propto \frac{1}{2\sigma^2}\left[\sum_{i=1}^{N}(y_i - \mathbf{w}^T\mathbf{x}_i)^2\right] + \frac{1}{\tau}||\mathbf{w} - \mu||_1 \end{aligned}$$

Plugging this into our original expression of the MAP, noting the change to argmin, we have the following expression.

$$\boxed{\mathbf{w}_{MAP} = \text{argmin}_{\mathbf{w}} \left(\frac{1}{2\sigma^2} \left[\sum_{i=1}^{N} (y_i - \mathbf{w}^T \mathbf{x}_i)^2 \right] + \frac{1}{\tau} ||\mathbf{w} - \mu||_1 \right)}$$

And so we have found an equation for the MAP estimate.

This relates to the frequentist perspective because we can see that the MAP estimate with a Laplace prior is almost identical to the least squares loss with a Lasso regularization. The regularization parameter $\lambda = \tau^{-1}$. This connection makes intuitive sense. If we place a prior on our weights w, we are effectively punishing potential weights that are unlikely given prior, just as a Lasso regularization punishes weights that are far from the given μ (which is usually 0). Even the effective of this punishment of driving some weights to μ is intuitive: due to the shape of the Laplace curve, it is more likely for many weights to be close to or exactly 0μ.

2.3

We might incorporate prior expert knowledge we may have for the problem by applying a prior on the variables for which we have expert knowledge about. For example, if we know beforehand that $\mathbf{w}$ should be close to some vector $\mathbf{v}$ in value, then we could apply a prior of some distribution on $\mathbf{w}$ with $\mu = \mathbf{v}$ as in question 2.2. This makes sense in both the Bayesian and frequentist perspectives because we are effectively "punishing" models that are far away from our believed value of $\mathbf{w}$: in the frequentist perspective we "punish" via the loss function, whereas in the Bayesian perspective we "punish" via our prior belief of the value of $\mathbf{w}$.

2.4

As τ decreases, the entries of the estimate $\mathbf{w}_{MAP}$ are driven to μ. By decreasing τ, we are effectively increasing the regularization punishment: we imply that our belief that $\mathbf{w}_{MAP}$ is close to μ is very strong, and so drive these weights to μ. As $\tau \to 0$, then we eventually drive all of our weights to μ.

2.5

We want to find which model $\mathbf{w}_{MAP}$ and $\mathbf{w}_{\text{mean}}$ would have lower expected test MSE and why. We will show this by starting with the MSE for some estimator $\hat{\mathbf{w}}$, reducing the expression, and concluding that the $\mathbf{w}_{\text{mean}}$ has a lower MSE.

We start with the MSE for some estimator $\hat{\mathbf{w}}$.

$$\mathbb{E}_{\mathbf{X},\mathbf{y}} \left[\sum_{i=1}^{N} (y_i - \hat{\mathbf{w}}^T \mathbf{x}_i)^2 \right]$$

We can simplifying using linearity and the fact that our individual $\mathbf{x}_i$ are identically distributed.

$$\begin{aligned} \mathbb{E}_{\mathbf{X},\mathbf{y}} \left[\sum_{i=1}^{N} (y_i - \hat{\mathbf{w}}^T \mathbf{x}_i)^2 \right] &= \sum_{i=1}^{N} \mathbb{E}_{\mathbf{X},\mathbf{y}} \left[(y_i - \hat{\mathbf{w}}^T \mathbf{x}_i)^2 \right] \\ &= N \left(\mathbb{E}_{\mathbf{x},y} \left[(y - \hat{\mathbf{w}}^T \mathbf{x})^2 \right] \right) \end{aligned}$$

Following the hint, let us add and subtract a term. Choose $\mathbf{w}^T \mathbf{x}$ as our term.

$$N \left(\mathbb{E}_{\mathbf{x},y} \left[(y - \hat{\mathbf{w}}^T \mathbf{x})^2 \right] \right) = N \left(\mathbb{E}_{\mathbf{x},y,\mathbf{w}|D} \left[(y - \mathbf{w}^T \mathbf{x} + \mathbf{w}^T \mathbf{x} - \hat{\mathbf{w}}^T \mathbf{x})^2 \right] \right)$$

By expanding via linearity we get the following expression with first, second, and third terms respectively.

$$N\left(\mathbb{E}_{\mathbf{x},y,\mathbf{w}|D}\left[(y-\mathbf{w}^T\mathbf{x})^2\right]+\mathbb{E}_{\mathbf{x},\mathbf{w}|D}\left[(\mathbf{w}^T\mathbf{x}-\hat{\mathbf{w}}^T\mathbf{x})^2\right]+2\mathbb{E}_{\mathbf{x},y,\mathbf{w}|D}\left[(y-\mathbf{w}^T\mathbf{x})(\mathbf{w}^T\mathbf{x}-\hat{\mathbf{w}}^T\mathbf{x})\right]\right) \tag{2}$$

First Term: Let us show that the first term from (2) is the squared noise ϵ by using the definition $y=\mathbf{w}^T\mathbf{x}+\epsilon$.

$$\begin{aligned}\mathbb{E}_{\mathbf{x},y,\mathbf{w}|D}\left[(y-\mathbf{w}^T\mathbf{x})^2\right]&=\mathbb{E}_{\mathbf{x},\mathbf{w}|D}\left[(\mathbf{w}^T\mathbf{x}+\epsilon-\mathbf{w}^T\mathbf{x})^2\right]\\&=\mathbb{E}\left[\epsilon^2\right]\end{aligned}$$

This cannot be reduced further and will not affect our optimal choice of $\hat{\mathbf{w}}$.

Third Term: Let us show that the third term from (2) is 0. By Adam's law we have the following.

$$\mathbb{E}_{\mathbf{x},y,\mathbf{w}|D}\left[(y-\mathbf{w}^T\mathbf{x})(\mathbf{w}^T\mathbf{x}-\hat{\mathbf{w}}^T\mathbf{x})\right]=\mathbb{E}_{\mathbf{x},\mathbf{w}|D}\left[\mathbb{E}_{y|\mathbf{x},\mathbf{w}}\left[(y-\mathbf{w}^T\mathbf{x})(\mathbf{w}^T\mathbf{x}-\hat{\mathbf{w}}^T\mathbf{x})\right]\right] \tag{3}$$

$$=\mathbb{E}_{\mathbf{x},\mathbf{w}|D}\left[(\mathbf{w}^T\mathbf{x}-\hat{\mathbf{w}}^T\mathbf{x})\mathbb{E}_{y|\mathbf{x},\mathbf{w}}\left[y-\mathbf{w}^T\mathbf{x}\right]\right] \tag{4}$$

Let us look at the inner-most expectation and show that it is zero.

$$\begin{aligned}\mathbb{E}_{y|\mathbf{x},\mathbf{w}}\left[y-\mathbf{w}^T\mathbf{x}\right]&=\mathbb{E}_{y|\mathbf{x},\mathbf{w}}\left[y\right]-\mathbb{E}_{y|\mathbf{x},\mathbf{w}}\left[\mathbf{w}^T\mathbf{x}\right]\\&=\mathbf{w}^T\mathbf{x}-\mathbf{w}^T\mathbf{x}\\&=0\end{aligned}$$

Plugging this back into (4) we have that the third term is 0.

Second Term: Let us simplify the second term from (2). We start by factoring out $\mathbf{x}$.

$$\mathbb{E}_{\mathbf{x},\mathbf{w}|D}\left[(\mathbf{w}^T\mathbf{x}-\hat{\mathbf{w}}^T\mathbf{x})^2\right]=\mathbb{E}_{\mathbf{x},\mathbf{w}|D}\left[((\mathbf{w}^T-\hat{\mathbf{w}}^T)\mathbf{x})^2\right]$$

Let $\mathbf{v}=\mathbf{w}-\hat{\mathbf{w}}$. We can then write this dot product as a sum of multiplications and expand into its components.

$$\begin{aligned}\mathbb{E}_{\mathbf{x},\mathbf{w}|D}\left[((\mathbf{w}^T-\hat{\mathbf{w}}^T)\mathbf{x})^2\right]&=\mathbb{E}_{\mathbf{x},\mathbf{w}|D}\left[(\mathbf{v}^T\mathbf{x})^2\right]\\&=\mathbb{E}_{\mathbf{x},\mathbf{w}|D}\left[\left(\sum_{i=1}^{p}v_ix_i\right)^2\right]\\&=\mathbb{E}_{\mathbf{x},\mathbf{w}|D}\left[\sum_{i=1}^{p}v_i^2x_i^2+2\sum_{\substack{i,j\in\{1,\ldots,p\},\\i<j}}v_iv_jx_ix_j\right]\end{aligned}$$

Using linearity we can simplify this to the following.

$$\sum_{i=1}^{p}\mathbb{E}_{\mathbf{x},\mathbf{w}|D}\left[v_i^2x_i^2\right]+2\sum_{\substack{i,j\in\{1,\ldots,p\},\\i<j}}\mathbb{E}_{\mathbf{x},\mathbf{w}|D}\left[v_iv_jx_ix_j\right] \tag{5}$$

Let us show that the second term of (5) is equal to 0. We use Adam's law.

$$\begin{aligned}\mathbb{E}_{\mathbf{x},\mathbf{w}|D}\left[v_iv_jx_ix_j\right]&=\mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_{\mathbf{w}|D}\left[v_iv_jx_ix_j|x_i,x_j\right]\right]\\&=\mathbb{E}_{\mathbf{x}}\left[x_ix_j\mathbb{E}_{\mathbf{w}|D}\left[v_iv_j\right]\right]\end{aligned}$$

The inner expectation is a constant with respect to x so we can bring it outside of the outer expectation. We then use the fact that each individual feature x_i are independent and have mean 0.

$$\begin{aligned}
\mathbb{E}_{\mathbf{x}}\left[x_i x_j \mathbb{E}_{\mathbf{w}|D}\left[v_i v_j\right]\right] &= \mathbb{E}_{\mathbf{w}|D}\left[v_i v_j\right] \mathbb{E}_{\mathbf{x}}\left[x_i x_j\right] \\
&= \mathbb{E}_{\mathbf{w}|D}\left[v_i v_j\right] \mathbb{E}_{\mathbf{x}}\left[x_i\right] \mathbb{E}_{\mathbf{x}}\left[x_j\right] \\
&= \mathbb{E}_{\mathbf{w}|D}\left[v_i v_j\right] (0)(0) \\
&= 0
\end{aligned}$$

So the second term of (5) is zero. Now let us work with the first term of (5). We again apply Adam's law and bring out the innermost expectation since it is constant with respect to $\mathbf{x}$.

$$\begin{aligned}
\mathbb{E}_{\mathbf{x},\mathbf{w}|D}\left[v_i^2 x_i^2\right] &= \mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_{\mathbf{w}|D}\left[v_i^2 x_i^2 | x_i\right]\right] \\
&= \mathbb{E}_{\mathbf{x}}\left[x_i^2 \mathbb{E}_{\mathbf{w}|D}\left[v_i^2\right]\right] \\
&= \mathbb{E}_{\mathbf{w}|D}\left[v_i^2\right] \mathbb{E}_{\mathbf{x}}\left[x_i^2\right]
\end{aligned}$$

We know from the question that $\text{var}(x_i) = 1$. Using the equation for variance as well as the fact that the mean of each x_i is 0 we have the following.

$$\begin{aligned}
\text{var}(x_i) &= \mathbb{E}_{\mathbf{x}}\left[x_i^2\right] + \mathbb{E}_{\mathbf{x}}\left[x_i\right]^2 \\
1 - \mathbb{E}_{\mathbf{x}}\left[x_i\right]^2 &= \mathbb{E}_{\mathbf{x}}\left[x_i^2\right] \\
1 &= \mathbb{E}_{\mathbf{x}}\left[x_i^2\right]
\end{aligned}$$

Thus the first term of (5) is the the following

$$\begin{aligned}
\mathbb{E}_{\mathbf{x},\mathbf{w}|D}\left[v_i^2 x_i^2\right] &= \mathbb{E}_{\mathbf{w}|D}\left[v_i^2\right] \mathbb{E}_{\mathbf{x}}\left[x_i^2\right] \\
&= \mathbb{E}_{\mathbf{w}|D}\left[v_i^2\right] \\
&= \mathbb{E}_{\mathbf{w}|D}\left[(w_i - \hat{w}_i)^2\right]
\end{aligned}$$

And so the second term of (2) can be simplified.

$$\mathbb{E}_{\mathbf{x},\mathbf{w}|D}\left[(\mathbf{w}^T\mathbf{x} - \hat{\mathbf{w}}^T\mathbf{x})^2\right] = \sum_{i=1}^{p} \mathbb{E}_{\mathbf{w}|D}\left[(w_i - \hat{w}_i)^2\right]$$

Conclusion: Note that this second term is the only term that we can control and that affects the optimal choice of $\hat{\mathbf{w}}$: the other two terms are equal to the noise or 0. Also note that this equation is equivalent to the sum of a MSE. By Theorem 6.1.4 on page 269 of the Stat 110 textbook, we know that any MSE of the form $\mathbb{E}[(X-c)^2]$ for some r.v. X is minimized when $c = \mu$ where μ is the mean of X. Therefore, we know that our MSE is minimized when $\hat{w}_i$ is equal to the mean of possible w_i. And so we can conclude that $\mathbf{w}_{\text{mean}}$ minimizes the test MSE and would have a lower expected test MSE than $\mathbf{w}_{\text{MAP}}$.

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:

3.1

We want to write out the dimensionality of each of the parameters and of the intermediate variables.

Parameter / Variable	Dimensions
$\mathbf{x}$	$M \times 1$
$\mathbf{W}_1$	$H \times M$
$\mathbf{b}_1$	$H \times 1$
$\mathbf{a}_1$	$H \times 1$
$\mathbf{z}_1$	$H \times 1$
$\mathbf{W}_2$	$1 \times H$
$\mathbf{b}_2$	1×1
a_2	1×1
z_2	1×1
$\hat{y}$	1×1

These all work with the mathematical operations described above.

3.2

For all parts of 3 we assume there is one datapoint $\mathbf{x}$ and the following loss.

$$L = -(y \log(\hat{y}) + (1-y)\log(1-\hat{y}))$$

3.2(a)

We want to find $\partial L/\partial b_2$. Using the chain rule, we have the following.

$$\frac{\partial L}{\partial b_2} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial a_2} \frac{\partial a_2}{\partial b_2}$$

Where $\hat{y} = \sigma(a_2)$ and $a_2 = \mathbf{W}_2\mathbf{z}_1 + \mathbf{b}_2$. We can substitute the respective partial derivatives and simplify.

$$\begin{aligned}\frac{\partial L}{\partial b_2} &= \left(-\frac{y}{\hat{y}} + \frac{(1-y)}{(1-\hat{y})}\right)(\hat{y}(1-\hat{y}))(1) \\ &= -y(1-\hat{y}) + \hat{y}(1-y) \\ &= \hat{y} - y\end{aligned}$$

And so we have found $\boxed{\partial L/\partial b_2 = \hat{y} - y}$.

3.2(b)

We want to find $\partial L/\partial W_2^h$ where W_2^h is the hth element of $\mathbf{W}_2$. By the chain rule we have the following expression.

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

Using our result from 3.2(a) and simplifying we have the following, where $\mathbf{z}_1^h$ is the hth element of $\mathbf{z}_1$.

$$\boxed{\frac{\partial L}{\partial W_2^h} = (\hat{y} - y)(\mathbf{z}_1^h)}$$

And so we have found $\partial L/\partial W_2^h$.

3.2(c)

We want to find $\partial L/\partial b_1^h$ where b_1^h is the hth element of $\mathbf{b}_1$. By the chain rule we have the following expression.

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

Where $z_1 = \sigma(a_1)$ and $a_1 = W_1 x + b_1$. We have the following partial derivatives.

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

Now we can sub these into our original equation.

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

And so we have found $\partial L/\partial b_1^h$.

3.2(d)

We want to find $\partial L/\partial W_1^{h,m}$, where $W_1^{h,m}$ is the element in row h and column m of $\mathbf{W}_1$. By the chain rule we have the following.

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

We have the following partial derivatives.

$$\frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial a_2} = \hat{y} - y \quad , \quad \frac{\partial a_2}{\partial z_1^m} = W_2^m \quad , \quad \frac{\partial z_1^m}{\partial a_1^m} = \sigma(a_1^m)(1 - \sigma(a_1^m)) \quad , \quad \frac{\partial a_1^m}{\partial W_1^{h,m}} = x^m$$

Now we can sub these into our original equation.

$$\boxed{\frac{\partial L}{\partial W_1^{h,m}} = (\hat{y} - y)(W_2^m)(\sigma(a_1^m)(1 - \sigma(a_1^m))(x^m)}$$

And so we have found $\partial L/\partial W_1^{h,m}$.

Problem 4 (Modern Deep Learning Tools: PyTorch)

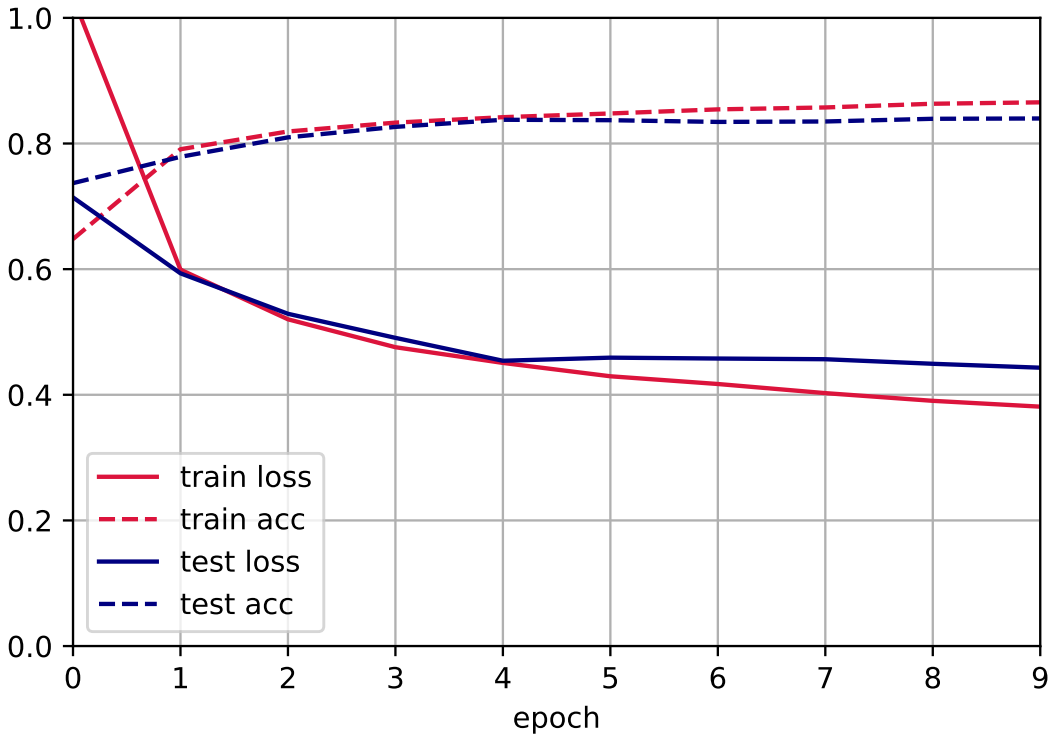
In this problem, you will learn how to use PyTorch. This machine learning library is massively popular and used heavily throughout industry and research. In `T3_P3.ipynb` you will implement an MLP for image classification from scratch. Copy and paste code solutions below and include a final graph of 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
n_hiddens = 256
n_outputs = 10

W1 = torch.randn(size=(n_hiddens, n_inputs), requires_grad=True)
b1 = torch.zeros(n_hiddens, requires_grad=True)
W2 = torch.randn(size=(n_outputs, n_hiddens), requires_grad=True)
b2 = torch.zeros(n_outputs, requires_grad=True)
with torch.no_grad():
    W1 *= 0.01
    W2 *= 0.01
```

```
def relu(x):
    return torch.clamp(x, min=0)



def softmax(X):
    Xexp = torch.exp(X)
    return torch.div(Xexp, torch.sum(Xexp, dim=1).unsqueeze(1))



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



def cross_entropy(y_hat, y):
    return -1. * torch.log(torch.gather(y_hat, 1, y.unsqueeze(1)))



def sgd(params, lr=0.1):
    with torch.no_grad():
        for i in range(len(params)):
            params[i] -= lr * params[i].grad
        for i in range(len(params)):
            params[i].grad.zero_()



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

Name

Rob Walker

Collaborators and Resources

Whom did you work with, and did you use any resources beyond cs181-textbook and your notes? James Kitch, Julian Schmitt, Luke Bailey (OH on Thursday night in Mather), https://pytorch.org/docs/stable/index.html

Calibration

Approximately how long did this homework take you to complete (in hours)? 25