CSE 546 - Homework 2

Nate Whybra

February 2025

Problem A1

(a) The model will **not necessarily** have a higher error if we remove the number of bathrooms feature. For instance, it seems likely that the number of bathrooms a house has would be correlated with its square footage (a bigger house probably has more bathrooms), and it seems obvious that larger houses should typically cost more. If square footage was also a feature, removing number of bathrooms might just mean that the square footage feature (and probably others) will be adjusted, and not necessarily improve the accuracy of the model.

(b) Suppose we are trying to train a model M with n weights, $\vec{w}$. Suppose we fix the regularization parameter $\mu > 0$, then training M with the L1 regularizer means that (1) $\|\vec{w}\|_1 \leq \mu$, and training M with the L2 regularizer means that (2) $\|\vec{w}\|_2 \leq \mu$. The region in $\mathbb{R}^n$ defined by (1) forms a polygonal region with pointy edges, so the weights are more likely to fall close to the coordinate axes (more zeros). The region defined by (2) is a sphere, so it is more likely the weights do not fall close to the coordinate axes (less zeros).

(c) Call $\|\vec{w}\|_* = \sum_{i=1}^{n} |w_i|^{0.5}$ the star-function. If we regularize our model with the star-function with parameter $\mu > 0$, then $\|\vec{w}\|_* \leq \mu$ defines a star-like region in $\mathbb{R}^n$ that is even pointier than the region formed with the L1 norm, so one advantage of using the star-norm is that it is likely to cause even more weights to be zero than even the L1 norm. A disadvantage is that the model may have a more difficult time generalizing, as the region of feasible weights has a relatively small volume (weights are overly biased).

(d) True. For instance, suppose we are trying to minimize the function $f(x) = x^2$ using gradient descent with an inital guess $x_0 \neq 0$ and learning rate η. The update rule is:

$$x_n = x_{n-1} - \eta \cdot \frac{df}{dx}|_{x=x_{n-1}} = x_{n-1} - \eta \cdot 2 \cdot x_{n-1} = (1 - 2\eta) \cdot x_{n-1}$$

This defines a simple recurrence relation with closed form solution $x_n = x_0 \cdot (1 - 2\eta)^n$. Our function $f(x)$ is minimized when $x = 0$, but x_n only converges to 0 when $|1 - 2\eta| < 1$, so choosing $\eta = 5$ for example would cause gradient descent to diverge.

(e) SGD works because it makes unbiased estimates of the true gradient of the loss function, ie. the mean of the loss gradient estimates from SGD approaches the true loss gradient.

(f) One advantage of SGD is that for large data sets, it has a smaller computational complexity than GD (uses 1 data point at a time to update loss estimate). One disadvantage of SGD is that the gradient estimate updates can have a large variance leading to slower convergence as compared to GD.

Problem A2

(a) To show that $f(x)$ is a norm, we must show that it satisfies (i), (ii), and (iii). Fix a vector $x \in \mathbb{R}^n$, then as each $|x_i| \geq 0$ with equality only if $x_i = 0$, then $f(x) \geq 0$ with equality only when each $x_i = 0$, which shows (i). Next, fix $a \in \mathbb{R}$, then we have:

$$f(ax) = \sum_{i=1}^{n} |ax_i| = \sum_{i=1}^{n} |a|\,|x_i| = |a| \sum_{i=1}^{n} |x_i| = |a|\, f(x)$$

Which shows (ii). For (iii), consider $x, y \in \mathbb{R}^n$. Then:

$$f(x+y) = \sum_{i=1}^{n} |x_i + y_i| \leq \sum_{i=1}^{n} |x_i| + |y_i| = \sum_{i=1}^{n} |x_i| + \sum_{i=1}^{n} |y_i| = f(x) + f(y)$$

Where the we used that $|x_i + y_i| \leq |x_i| + |y_i|$ (the triangle inequality in $\mathbb{R}$). To establish this, consider $a, b \in \mathbb{R}$. If $a, b > 0$ we have $a + b > 0$ so that:

$$|a + b| = a + b = |a| + |b|$$

If $a, b < 0$, then $-a, -b > 0$, so that:

$$|a + b| = |-1 \cdot (a + b)| = |-a - b| = -a - b = |a| + |b|$$

Now without loss of generality, assume $a > 0$ and $b < 0$. Then if $|b| > |a|$, that means $a + b < 0$ (or $-a - b > 0$) , so we have:

$$|a + b| = |-1 \cdot (a + b)| = -a - b \leq |a| + |b|$$

If $|a| > |b|$, that means $a + b > 0$, which means $|a + b| = a + b \leq |a| + |b|$. This establishes the triangle inequality in $\mathbb{R}$, so that our justification for (iii) is complete.

(b) Consider the vectors $x = (1, 2)^T$ and $y = (1, 0)^T$. Then:

$$g(x+y) = \left(|1+1|^{\frac{1}{2}} + |2+0|^{\frac{1}{2}}\right)^2 = (2\sqrt{2})^2 = 8$$

And:

$$g(x)+g(y) = \left(|1|^{\frac{1}{2}} + |2|^{\frac{1}{2}}\right)^2 + \left(|1|^{\frac{1}{2}} + |0|^{\frac{1}{2}}\right)^2 = (1+\sqrt{2})^2+1 = 4+2\sqrt{2} \approx 6.83$$

So we have found a pair of vectors x, y such that $g(x+y) > g(x) + g(y)$, which means g cannot be a norm.

Problem A3

(I) The diagram is not convex, as the points b and d cannot be connected with a straight line entirely contained in the region.

(II) The diagram is convex, and two points in the region can be connected with a straight line entirely contained in the region.

(III) The diagram is not convex, as the points a and d cannot be connected with a straight line entirely contained in the region.

Problem A4

(a) The function is convex on $[a, c]$. If we choose two points $x_1, x_2 \in [a, c]$, then the line connecting $f(x_1)$ and $f(x_2)$ is larger than f on $[x_1, x_2]$.

(b) The function is not convex on $[a, c]$. The straight line drawn between $f(a)$ and $f(b)$ is smaller than f on $[a, b]$.

(c) The function is not convex on $[a, d]$. The straight line drawn between $f(a)$ and $f(c)$ is smaller than f on $[a, c]$.

(d) The function is convex on $[c, d]$. If we choose two points $x_1, x_2 \in [c, d]$, then the line connecting $f(x_1)$ and $f(x_2)$ is larger than f on $[x_1, x_2]$.

Problem A5

(a)

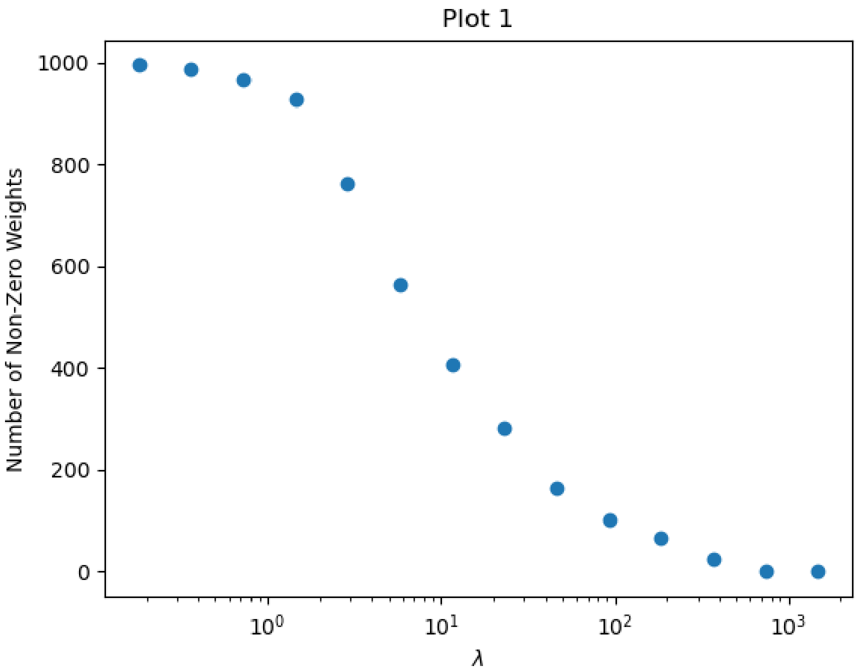

Figure 1: The number of non-zero model weights as a function of λ.

(b)

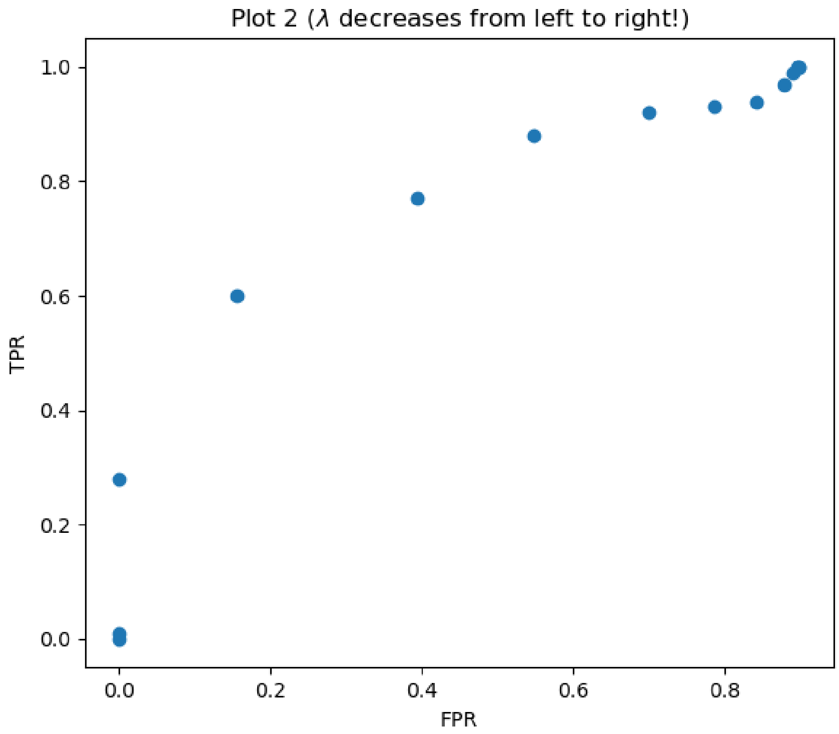

Figure 2: True Positive Rate vs. False Positive Rate

(c) In Plot 1, increasing λ makes more and more of the model weights 0, while decreasing λ makes less and less of the model weights 0. In Plot 2, increasing λ causes both the TPR and FPR to approach zero, but decreasing λ causes both the TPR and FPR to approach 1.

Problem A6

(a) (1) pctWPubAsst: The percentage of households with public assistance income largely depends on the policies of the current lawmakers. (2) PctPopUnderPov: The percentage of people under the poverty level. The poverty line is defined by the government and is consistently changed. (3) perCapInc: The per capita income. Inflation, changes in minimum wage, and economic cycles are all things that effect per capita income.

(b) (1) PctPopUnderPov: The percentage of people under the poverty level. It makes sense that people who are struggling are more likely to commit crimes, but violent crimes can also cause local communities to suffer. (2) NumKindsDrugsSeized: The number of different kinds of drugs seized. More drugs implying more violent crimes seems plausible, but it would also make sense if violent crimes implied more drugs. (3) PctLess9thGrade: The percentage of people 25 and over with less than a 9th grade education. It makes sense that less educated individuals might be more likely to commit crimes, but if you live in an area with a large density of crimes it can cause local schools to struggle and less people in the community make it through high-school.

(c)

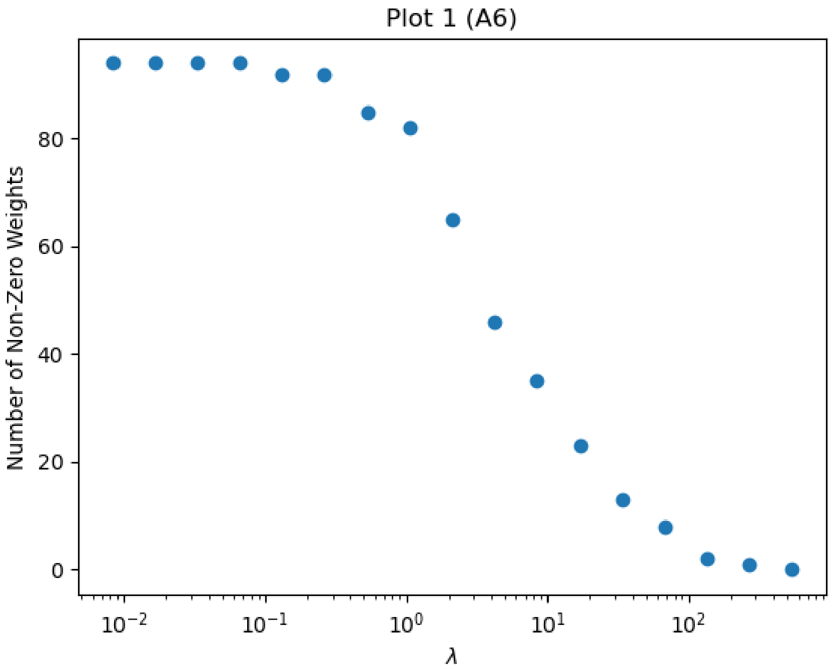

Figure 3: Plot 1

(d)

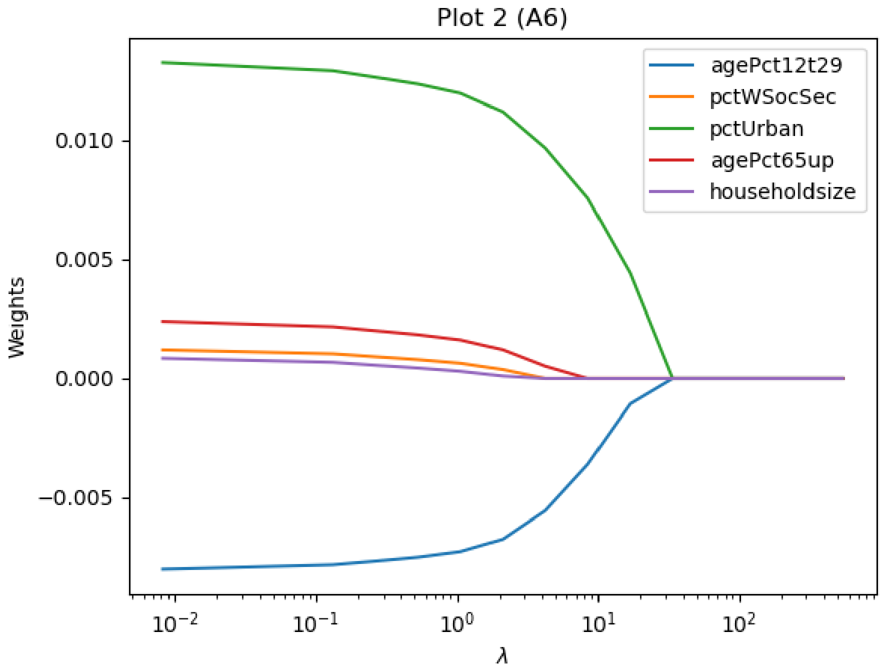

Figure 4: Plot 2

(e)

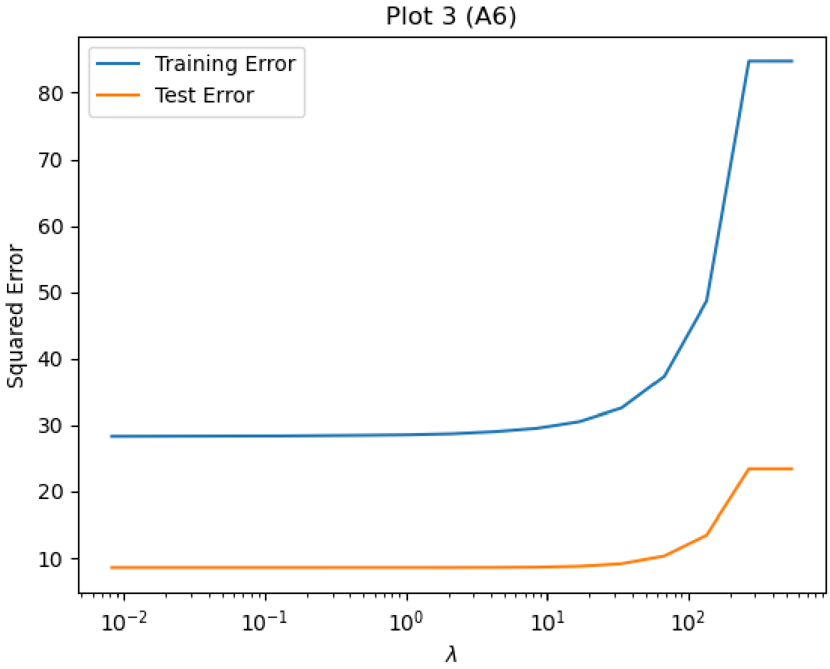

Figure 5: Plot 3

(f) The feature, PctIlleg: The percentage of kids born to never married, had the most positive weight in the model. The feature, PctKids2Par: The percentage of kids in family housing with two parents, had the most negative weight in the model. According to this model, violent crimes seem to be correlated with children not growing up in a traditional family environment (illegitmate children positively predict crime, but children raised with 2 parents negatively predict crime).

(g) The statistical flaw in this reasoning is that correlation does not imply causation. In this case, it's more likely that people aged 65 and up choose to live away from high crime areas, and less likely that they have the magical ability to scare criminals away. Just because seniors are negatively correlated with crime, they do not necessarily cause a reduction in crime.

Problem A7

(a) By expanding everything in terms of the components of w, we can write (with X an n x d matrix, where $x_i^T = x_{i,:}$):

$$J(w,b) = \frac{1}{n}\sum_{i=1}^{n} \log(1 + e^{-y_i(b + x_{i,1}w_1 + \cdots + x_{i,n}w_n)}) + \lambda(w_1^2 + \cdots + w_n^2)$$

This way, we have:

$$\begin{aligned}\frac{\partial J(w,b)}{\partial w_j} &= \frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial w_j}\log(1 + e^{-y_i(b + x_{i,1}w_1 + \cdots + x_{i,n}w_n)}) + \lambda\frac{\partial}{\partial w_j}(w_1^2 + \cdots + w_n^2) \\ &= \frac{1}{n}\sum_{i=1}^{n} \frac{-y_i x_{i,j} e^{-y_i(b + x_{i,1}w_1 + \cdots + x_{i,n}w_n)}}{1 + e^{-y_i(b + x_{i,1}w_1 + \cdots + x_{i,n}w_n)}} + 2\lambda w_j \\ &= \frac{1}{n}\sum_{i=1}^{n} \frac{-y_i x_{i,j} e^{-y_i(b + x_i^T w)}}{1 + e^{-y_i(b + x_i^T w)}} + 2\lambda w_j\end{aligned}$$

Now let $E_i = e^{-y_i(b + x_i^T w)}$, so that $\mu_i = \frac{1}{1+E_i}$ (where we drop the (w,b) for notational convenience). Solving this equation for E_i yields that $E_i = \frac{1-\mu_i}{\mu_i}$. So we can write the above as:

$$\begin{aligned}&= \frac{1}{n}\sum_{i=1}^{n} -y_i x_{i,j}\left(\frac{1-\mu_i}{\mu_i}\right)\mu_i + 2\lambda w_j \\ &= \frac{1}{n}\sum_{i=1}^{n} -y_i x_{i,j}(1-\mu_i) + 2\lambda w_j\end{aligned}$$

In matrix form, let $u, v \in \mathbb{R}^n$. Let u be the vector of $\mu_i(w, b)$ values for each i, and let v be such that $v_i = 1 - \mu_i$. Also, let $y \in \mathbb{R}^n$ be the target vector of y_i values. Then the above can be written:

$$\nabla_w J(w, b) = -\frac{1}{n} X^T (y \odot v) + 2\lambda w$$

Next, we have:

$$\begin{aligned}
\frac{\partial J(w,b)}{\partial b} &= \frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial b} \log(1 + e^{-y_i(b + x_{i,1}w_1 + \cdots + x_{i,n}w_n)}) + \lambda \frac{\partial}{\partial b}(w_1^2 + \cdots + w_n^2) \\
&= \frac{1}{n} \sum_{i=1}^{n} \frac{-y_i e^{-y_i(b + x_{i,1}w_1 + \cdots + x_{i,n}w_n)}}{1 + e^{-y_i(b + x_{i,1}w_1 + \cdots + x_{i,n}w_n)}} \\
&= \frac{1}{n} \sum_{i=1}^{n} \frac{-y_i e^{-y_i(b + x_i^T w)}}{1 + e^{-y_i(b + x_i^T w)}} \\
&= \frac{1}{n} \sum_{i=1}^{n} -y_i(1 - \mu_i) \\
&= -\frac{1}{n} y^T v
\end{aligned}$$

Now simply, as b is a scalar:

$$\nabla_b J(w, b) = -\frac{1}{n} y^T v$$

(b)

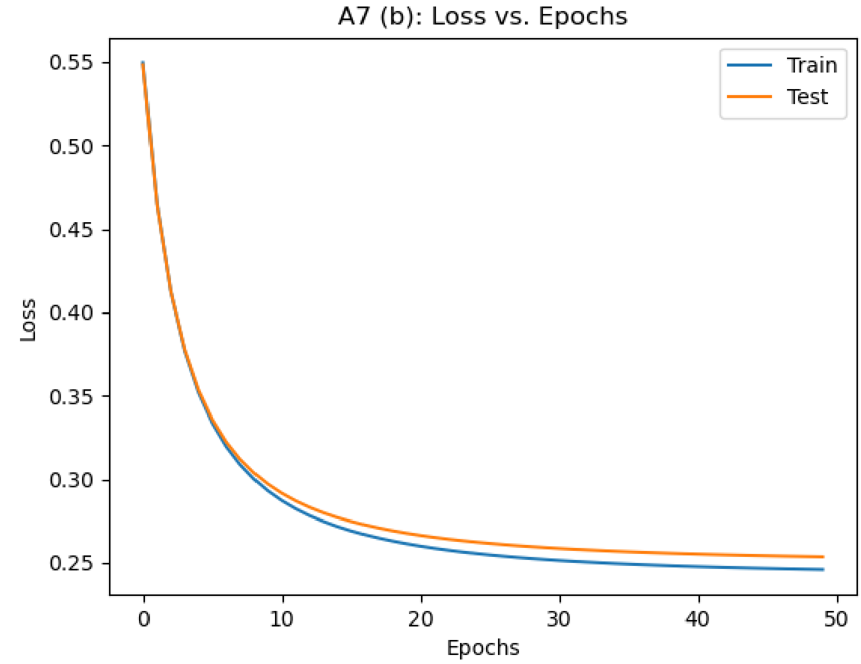

Figure 6: b(i)

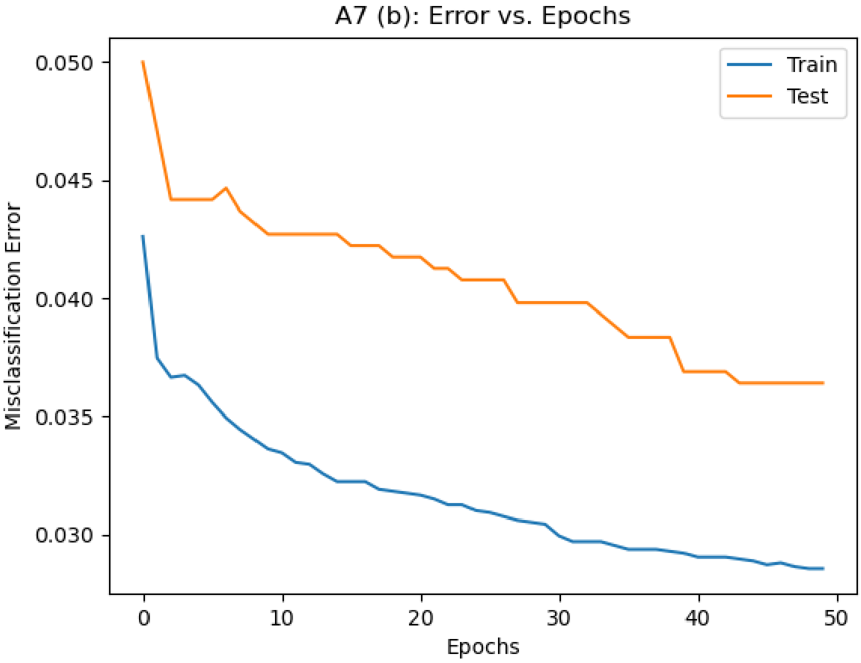

Figure 7: b(ii)

(c)

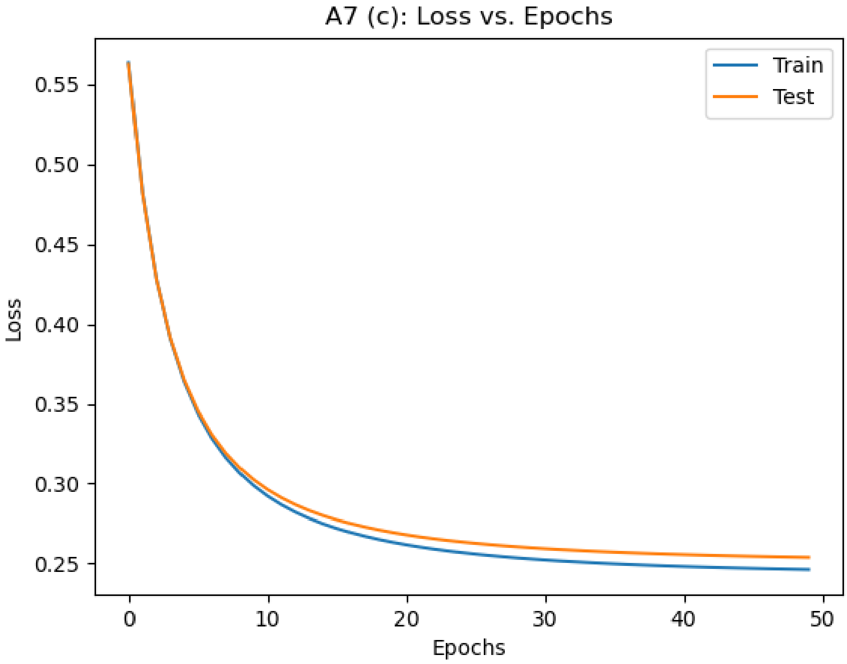

Figure 8: c(i)

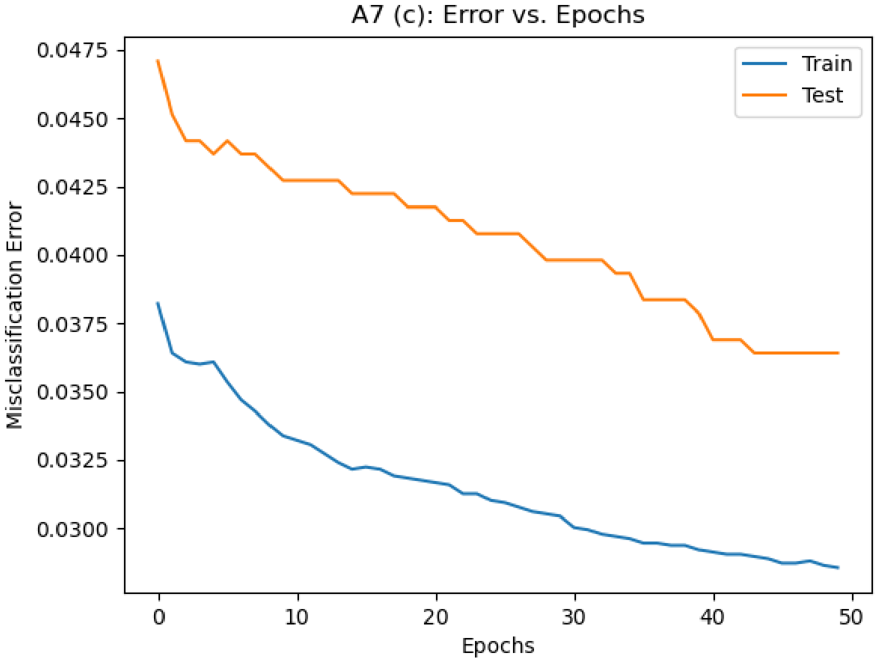

Figure 9: c(ii)

(d)

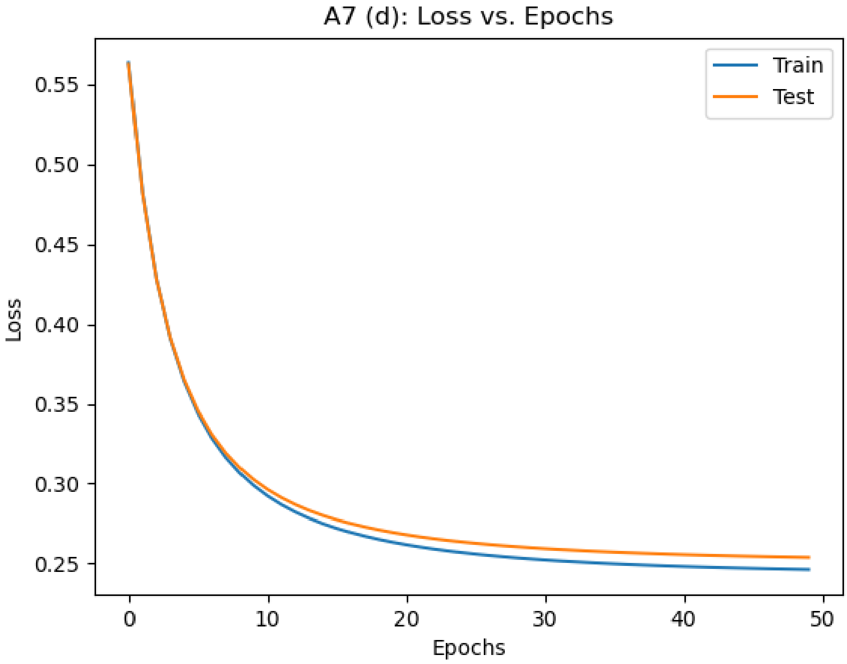

Figure 10: d(i)

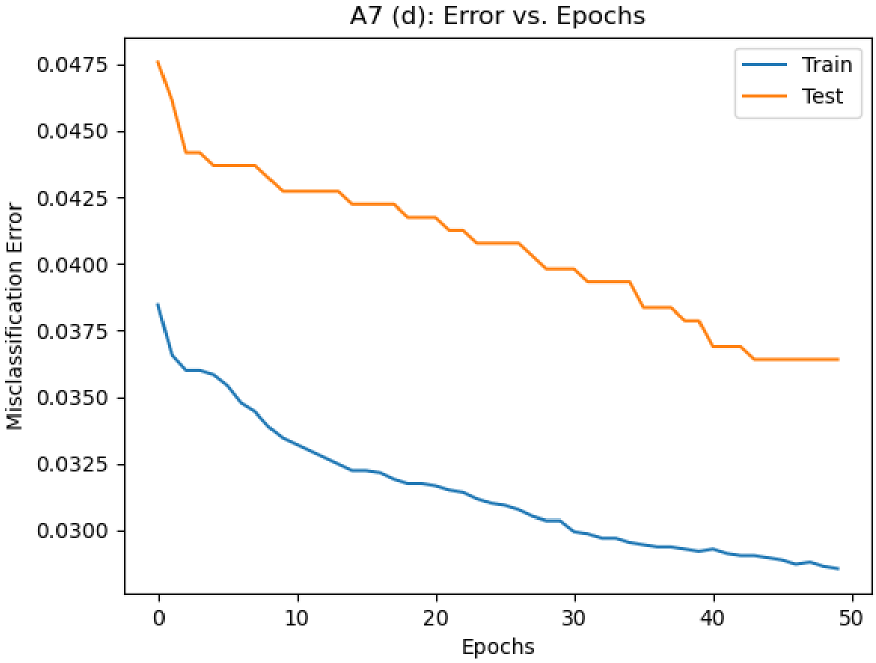

Figure 11: d(ii)

Problem B1

(a) Let $\lambda \in [0,1]$. Then:

$$f(\lambda x + (1-\lambda)y) = \|\lambda x + (1-\lambda)y\| \leq \|\lambda x\| + \|(1-\lambda)y\|$$

$$= |\lambda| \, \|x\| + |1-\lambda| \, \|y\|$$

$$= \lambda \, \|x\| + (1-\lambda) \, \|y\| = \lambda f(x) + (1-\lambda) f(y)$$

So f is convex.

(b) Let $A = \{x \in \mathbb{R}^n : \|x\| \leq 1\}$. Now take $x, y \in A$, and $\lambda \in [0,1]$. We have:

$$\|\lambda x + (1-\lambda)y\| \leq |\lambda| \, \|x\| + |1-\lambda| \, \|y\|$$

$$= \lambda \, \|x\| + (1-\lambda) \, \|y\|$$

$$\leq \lambda + (1-\lambda)$$

$$= 1$$

We have shown that for all $x, y \in A$, and $\lambda \in [0,1]$, $\lambda x + (1-\lambda)y \in A$, so A is a convex set.

(c) Looking at the diagram below, it is clear that G is not convex. If you tried drawing a line between any of the star tips for example, the line would exit the set G.

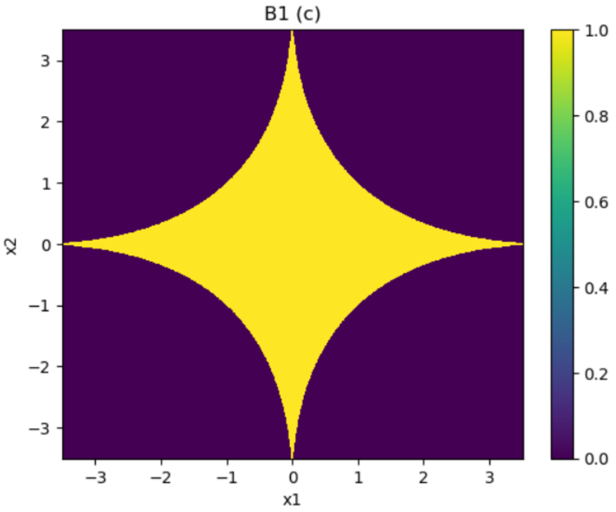

Figure 12: A diagram of the set $G = \{(x_1, x_2) : g(x_1, x_2) \leq 4\}$. Here, 1 means the coordinate is in G and 0 means the coordinate is not in G.

Problem B2

(a) Firstly, we know the least squares solution for $\hat{\beta}$. We have $\hat{\beta} = (X^T X)^{-1} X^T Y$. We also know that $Y = X\beta^* + \epsilon$, where $\epsilon \sim \mathcal{N}(0, I_d)$. Substituting for Y, we see:

$$\hat{\beta} = (X^T X)^{-1} X^T (X\beta^* + \epsilon)$$

$$= (X^T X)^{-1} (X^T X)\beta^* + (X^T X)^{-1} X^T \epsilon$$

$$= \beta^* + (X^T X)^{-1} X^T \epsilon$$

So by Proposition 1 in the confidence interval notes, with $A = (X^T X)^{-1} X^T$ we see:

$$\hat{\beta} \sim \mathcal{N}(A(0) + \beta^*, A I_d A^T)$$

$$= \mathcal{N}(\beta^*, AA^T)$$

$$= \mathcal{N}(\beta^*, (X^T X)^{-1} (X^T X)[(X^T X)^{-1}]^T)$$

$$= \mathcal{N}(\beta^*, [(X^T X)^{-1}]^T)$$

Looking element-wise now, for $1 \leq j \leq d$, we have:

$$\hat{\beta}_j \sim \mathcal{N}(\beta_j^*, [(X^T X)^{-1}]_{j,j}^T)$$

$$= \mathcal{N}(\beta_j^*, [(X^T X)^{-1}]_{j,j})$$

Since the individual variances are along the diagonal and the diagonal entries don't change after taking the transpose.

(b) With $\beta^* = 0$, we can directly apply the results of the argument beneath proposition 2 in the confidence interval notes with $\mu = 0$ and $\sigma_j^2 = [(X^T X)^{-1}]_{j,j}$, to say that:

$$1 - \delta \leq P\left(\left|\hat{\beta}_j\right| \leq \sqrt{2[(X^T X)^{-1}]_{j,j} \log\left(\frac{2}{\delta}\right)}\right)$$

Now, consider the probability that there is at least one j where the bound does not hold:

$$P\left(\exists j \in [d] : \left|\hat{\beta}_j\right| > \sqrt{2[(X^T X)^{-1}]_{j,j} \log\left(\frac{2}{\delta}\right)}\right)$$

The above is:

$$\leq \sum_{j=1}^{d} P\left(\left|\hat{\beta}_j\right| > \sqrt{2[(X^T X)^{-1}]_{j,j} \log\left(\frac{2}{\delta}\right)}\right)$$

$$\leq \sum_{j=1}^{d} \delta$$

$$= d\delta$$

So, the probability that all β_j's all satisfy the bound simultaneously is at least $1 - d\delta$, which is strictly smaller than $1 - \delta$ for $d > 1$. So if $d > 1$, we cannot say the bound holds for all the β_j's simultaneously with probability at least $1 - \delta$.

(c) 35 weights were outside the confidence interval.

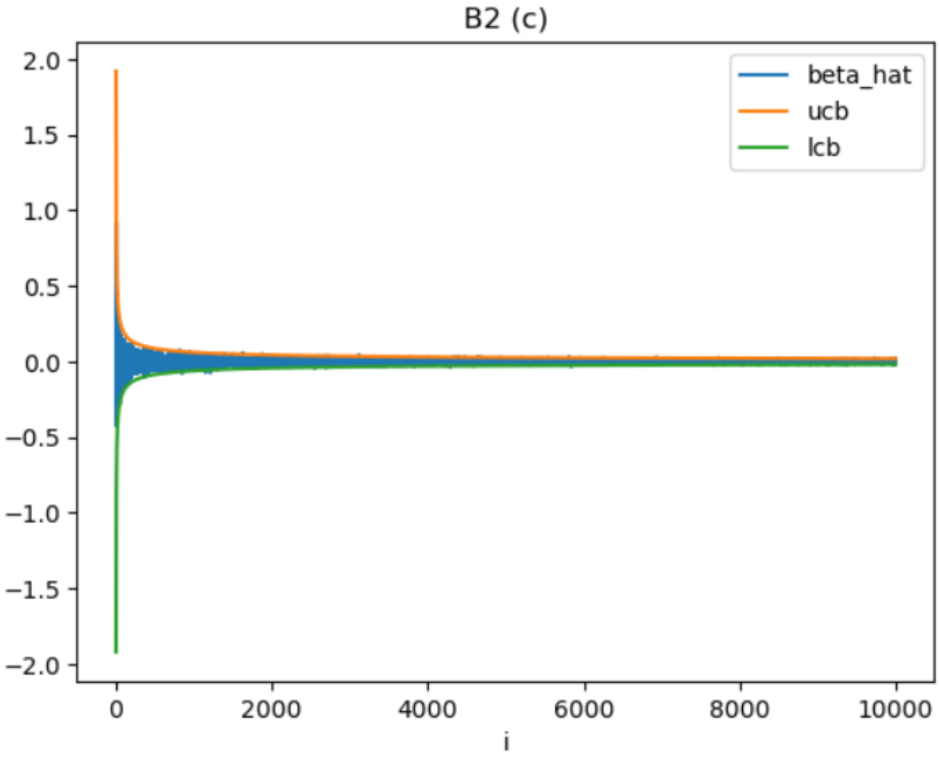

Figure 13: Weight estimates with confidence interval.