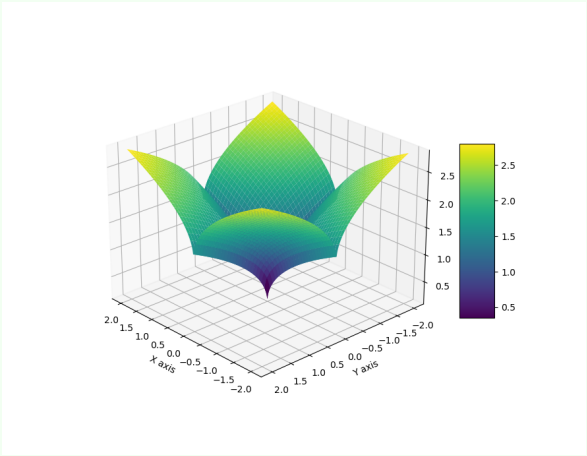
Homework 2

A1.

(a)

The L1 norm more likely to result in sparsity because it is it is linear w.r.t the magnitude of the weights compared to the L2 norm which is quadratic w.r.t weight magnitude giving a smoother LS penalty. This is why L1-ball has a pointy diamond shape, which means that small weights are likely to be push to zero and thus excluded from the model (giving sparsity). While the L2-ball is smooth and tends to find denser solutions.

(b)



Upside: L0.5 norm as shown in the graph is even pointier than L1 norm, which means that it can give even sparser results and select features more aggressively (filtering out even more small/zero features).
Downside: Also from the graph we can see that the L0.5 norm is not convex, which means that it is hard to optimize and find the global minimum.

(c)

True, when the step-size is too large, gradient decent can overshoot near the minimum and jump back and forth around the minimum and never converge.

(d)

Advantage SGD has over GD: SGD can converge significantly faster for large datasets since it updates the weights w.r.t. the gradient of a single randomly chosen sample, while GD requires calculating the gradient for the entire dataset.
Disadvantage of SGD has relative to GD: SGD is more susceptible to noise/variance in the chosen sample and the decent path might be more volatile and might converge to a local minimum instead of the global minimum.

(e)

Gradient descent is necessary for logistic regression because the loss function is convex but non-linear because of the sigmoid function and does not have a closed-form solution. While linear regression has convex function with a closed-form solution which makes gradient descent optional.

A2.

(a)

Non-negativity: $\forall x \in \mathbb{R}^n$, $f(x) \geq 0$ holds because absolute values are non-negative ($|x_i| \geq 0$), and the sum of non-negative values is also non-negative ($\sum_{i=1}^{n} |x_i| \geq 0$), and this sum is 0 if and only if all the elements of x are 0 (i.e. $x = 0$), thus $f(x) = 0$ iff $x = 0$ is also true.
Absolute scalability: $\forall x \in \mathbb{R}^n, \forall a \in \mathbb{R}$, $f(ax) = |a|f(x)$ holds because by the definition of absolute value $|ax_i| = |a||x_i|$ and therefore $f(ax) = \sum_{i=1}^{n} |ax_i| = \sum_{i=1}^{n} |a||x_i| = |a| \sum_{i=1}^{n} |x_i|$ which is $|a|f(x)$.
Triangle inequality: $\forall x, y \in \mathbb{R}^n$, $f(x+y) \leq f(x) + f(y)$ holds because:
for any $a, b \in \mathbb{R}$:
Case 1 when a, b has the same sign:
if $a, b \geq 0$ then $|a+b| = a+b = |a| + |b|$
if $a, b < 0$ then $|a+b| = -(a+b) = -a-b = |a| + |b|$
Case 2 when a, b has different signs:
w.l.o.g. assume $a \geq 0, b \leq 0$ then:
if $|a| \geq |b|$ then $a+b \geq 0$ and thus $|a+b| = a+b \leq a = |a| \leq |a| + |b|$
if $|a| < |b|$ then $a+b < 0$ and thus $|a+b| = -(a+b) = -a-b = |b| - |a| \leq |b| \leq |a| + |b|$
Thus $|a+b| \leq |a| + |b|$ holds for all $a, b \in \mathbb{R}$
Following this, we can see that $f(x+y) = \sum_{i=1}^{n} |x_i + y_i| \leq \sum_{i=1}^{n} |x_i| + |y_i| = f(x) + f(y)$

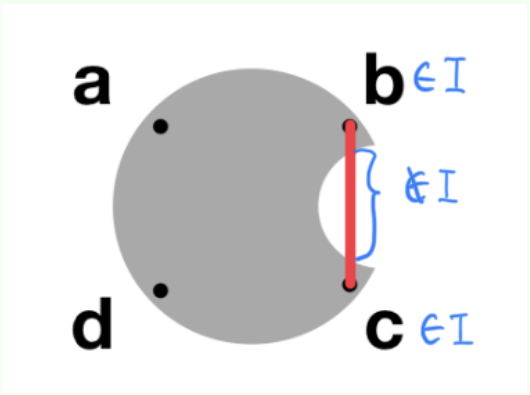
Therefore $f(x) = \sum_{i=1}^{n} |x_i|$ is a norm.

(b)

Assume we have $x, y \in \mathbb{R}^2$ where $x = [1, 0]$ and $y = [0, 1]$, then
$g(x+y) = (\sum_{i=1}^{2} |x_i + y_i|^{\frac{1}{2}})^2 = (|x_1 + y_1|^{\frac{1}{2}} + |x_2 + y_2|^{\frac{1}{2}})^2 = (|1+0|^{\frac{1}{2}} + |0+1|^{\frac{1}{2}})^2 = (1+1)^2 = 4 > 2 = 1^2 + 1^2 = (|x_1|^{\frac{1}{2}} + |x_2|^{\frac{1}{2}})^2 + (|y_1|^{\frac{1}{2}} + |y_2|^{\frac{1}{2}})^2 = (\sum_{i=1}^{2} |x_i|^{\frac{1}{2}})^2 + (\sum_{i=1}^{2} |y_i|^{\frac{1}{2}})^2 = g(x) + g(y)$
Therefore $g(x) = (\sum_{i=1}^{n} |x_i|^{\frac{1}{2}})^2$ is not a norm.

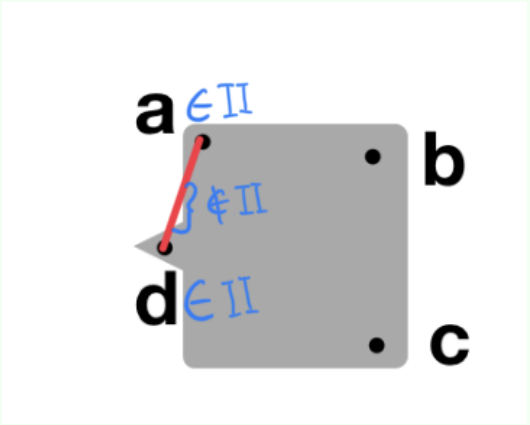
A3.

(I)



The set is not convex, because as show in the graph, $b, c \in I$ but the line segment between b and c (i.e. $\lambda b + (1-\lambda)c$) is not entirely in I (the segment marked blue is not in grey-shaded region).

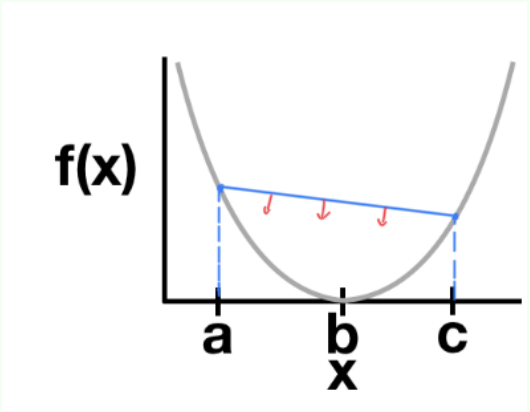
(II)



The set is not convex, because as show in the graph, $a, d \in II$ but the line segment between a and d (i.e. $\lambda a + (1-\lambda)d$) is not entirely in II (the segment marked blue is not in grey-shaded region).

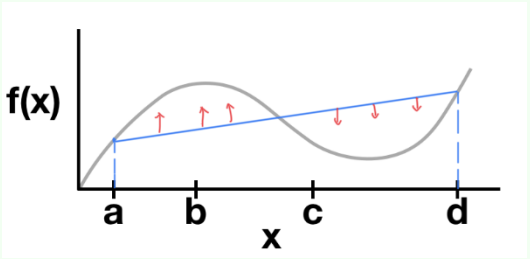
A4.

(a)



The function is convex on the interval [a,c], because as shown in the graph, the line segment between $(a, f(a))$ and $(c, f(c))$ (i.e. $\lambda f(a) + (1-\lambda) f(c)$) is above the function $f(x)$ for all $x \in [a, c]$ (i.e. $f(\lambda a + (1-\lambda)c) \leq \lambda f(a) + (1-\lambda) f(c)$).

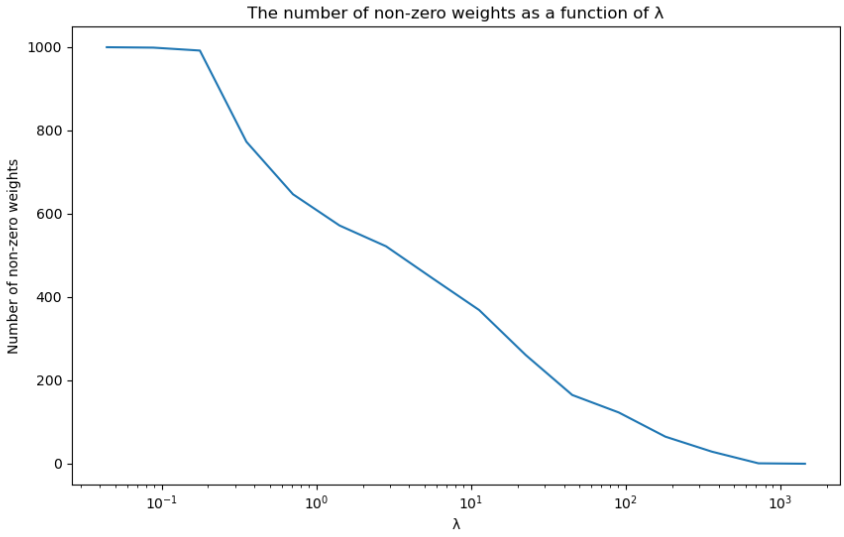
(b)



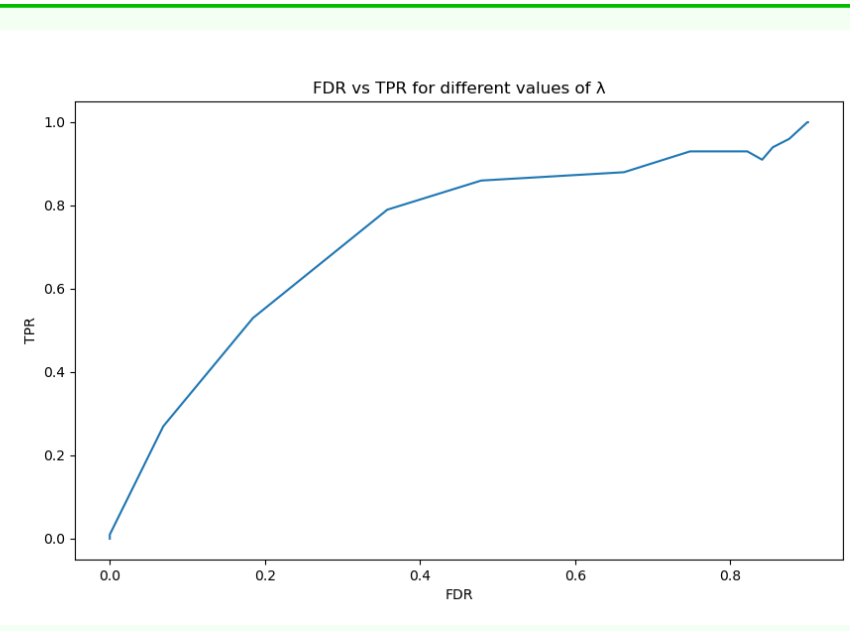
The function is not convex on the interval [a,d], because as shown in the graph the line segment between $(a, f(a))$ and $(b, f(b))$ (i.e. $\lambda f(a) + (1-\lambda) f(b)$) is below the function $f(x)$ for all $x \in [a, b]$ (i.e. $f(\lambda a + (1-\lambda)b) \geq \lambda f(a) + (1-\lambda) f(b)$), and $b \in [a, d]$.

A5.

(a)


The number of non-zero weights as a function of λ
Number of non-zero weights
1000
800
600
400
200
0
10^{-1}
10^{0}
10^{1}
10^{2}
10^{3}
λ

(b)


FDR vs TPR for different values of λ
TPR
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
FDR

(c)

In the part a plot, as λ decreases, the number of non-zero weights in the model increases (higher λ gives sparser results, lower λ decreases sparsity).
In the FDR vs TPR plot, as λ decreases (moving left to right in x-axis), the TPR increases (model identifies more features as relevant), and the FDR also increases (model includes more irrelevant features).

A6.

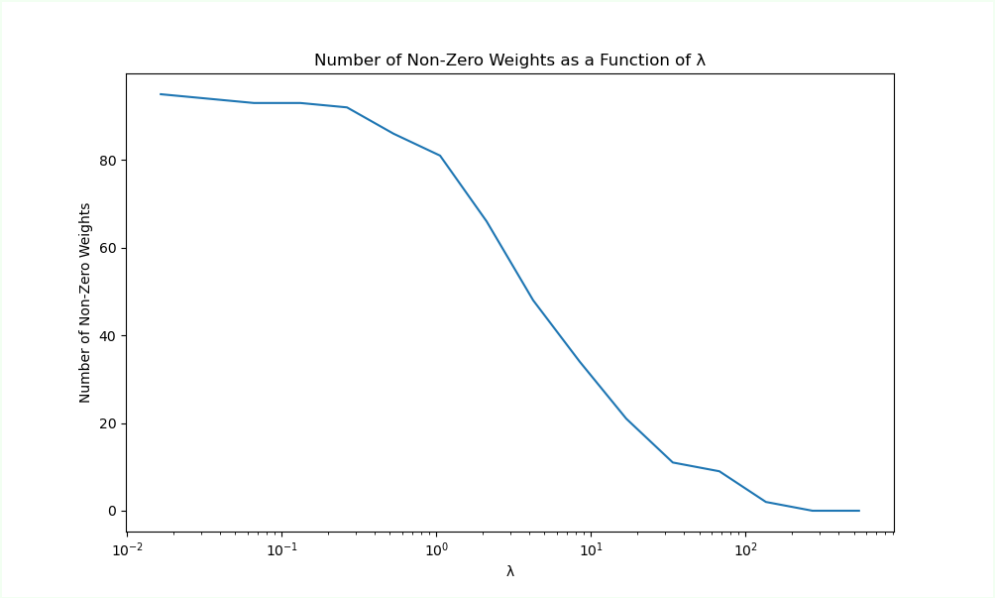
(a)

PctPopUnderPov: percentage of people under the poverty level could be influenced by minimum wage laws, tax policies, welfare programs, job creation plans made by the government.
OfficAssgnDrugUnits: number of officers assigned to special drug units could be influenced by the policy campaign of federal or state government (e.g. the war on drugs policy campaign).
PctPolicMinor: percent of police that are minority could be influenced by the government's diversity hiring policy in law enforcement agencies.
PctUsePubTrans: percent of people using public transit for commuting is highly influenced by government's decision on investments in public transit infrastructure.

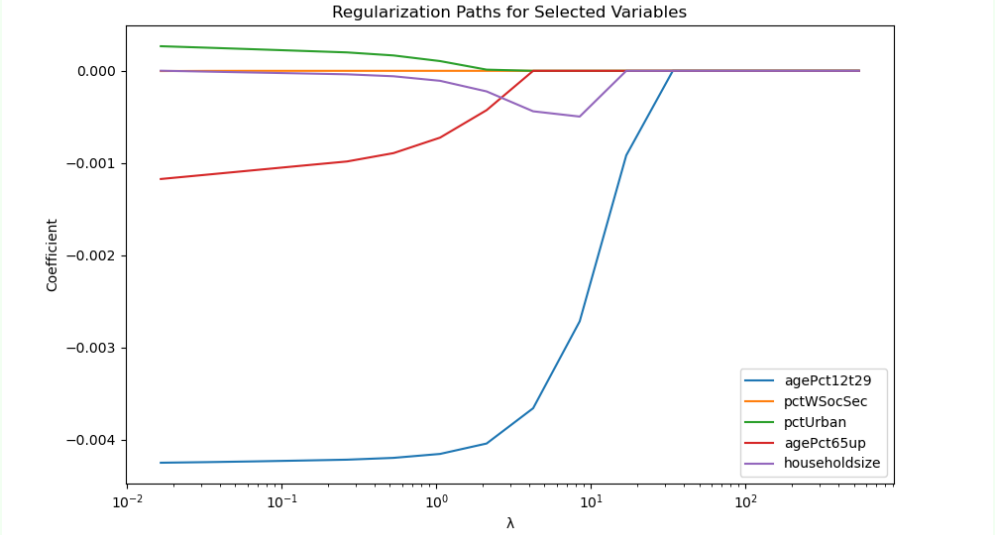
(b)

LemasSwornFT: a higher number of sworn full-time police officers could be a response to higher crime rates rather than a cause. As local decision makers may increase police presence in response to rising violence.
HousVacant: a higher number of vacant households could be a result of rising local crime rates, as people may move out of the area to avoid crime.
LemasTotalReq: total requests for police could rise in response to rising levels of violent crime.
RentLowQ: high number of lower quartile rent housing could be a result of higher crime rates which may lower the demand for housing in the area and thus lower the rent prices.

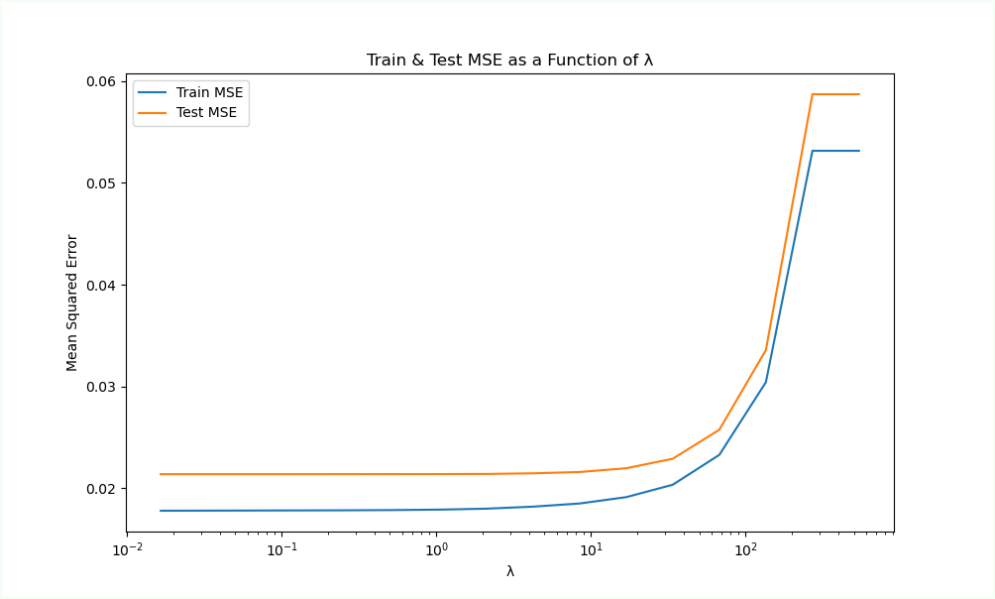
(c)



(d)


Regularization Paths for Selected Variables
Coefficient
λ
agePct12t29
pctWSocSec
pctUrban
agePct65up
householdsize

(e)


Train & Test MSE as a Function of λ
Mean Squared Error
λ
Train MSE
Test MSE

(f)

Most positive: PctIlleg (% of kids born to never married), Value: 0.06164620954468405
PctIlleg has the largest positive Lasso coefficient, which suggests that in our model, a higher percentage of children born to never married is correlated with the increase in violent crimes per capita.

Most negative: PctKids2Par (% of kids in family housing with two parents), Value: -0.04046680045005498
PPctKids2Par has the most negative Lasso coefficient, which suggests that in our model, communities with a higher percentage of children living with both parents are correlated with lower rates of violent crimes per capita.

(g)

Correlation $\neq$ causation. A large negative coefficient is only suggesting that higher values of agePct65up tend to show up with lower crime rates, and doesn't mean that increasing that feature will cause crime to decrease. People over 65 might have better financial autonomy to choose to live in safer neighborhoods, which could be a reason for the negative correlation.

A7.

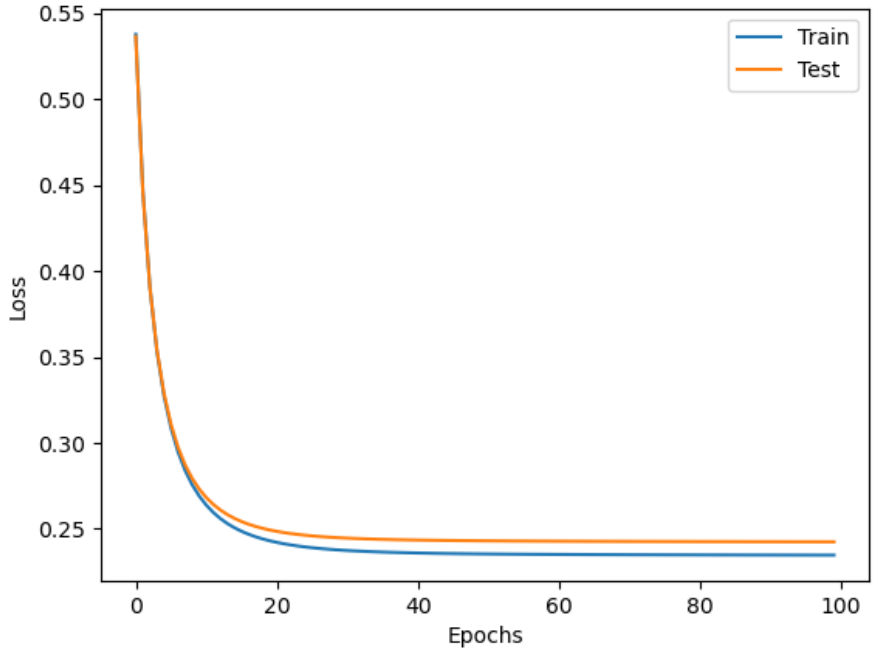
(a)

$$
\begin{aligned}
\nabla_w J(w,b) &= \frac{\partial}{\partial x}\Big(\frac{1}{n}\sum_{i=1}^{n}\log(1+\exp(-y_i(b+x_i^T w))) + \lambda||w||_2^2\Big) \\
&= \frac{1}{n}\sum_{i=1}^{n}\Big(\frac{\partial}{\partial x}(\log(1+\exp(-y_i(b+x_i^T w)))) + \frac{\partial}{\partial x}(\lambda w^\top w)\Big) \\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{-y_i x_i \exp(-y_i(b+x_i^T w))}{1+\exp(-y_i(b+x_i^T w))} + 2\lambda w && \text{(chain rule)} \\
&= \frac{1}{n}\sum_{i=1}^{n} -y_i x_i\Big(1 - \frac{1}{1+\exp(-y_i(b+x_i^T w))}\Big) + 2\lambda w \\
&= \frac{1}{n}\sum_{i=1}^{n} -y_i x_i(1-\mu_i(w,b)) + 2\lambda w && \text{(since } \mu_i(w,b) = \tfrac{1}{1+\exp(-y_i(b+x_i^T w))}\text{)} \\
&= \boxed{-\frac{1}{n}\sum_{i=1}^{n} y_i x_i(1-\mu_i(w,b)) + 2\lambda w}
\end{aligned}
$$

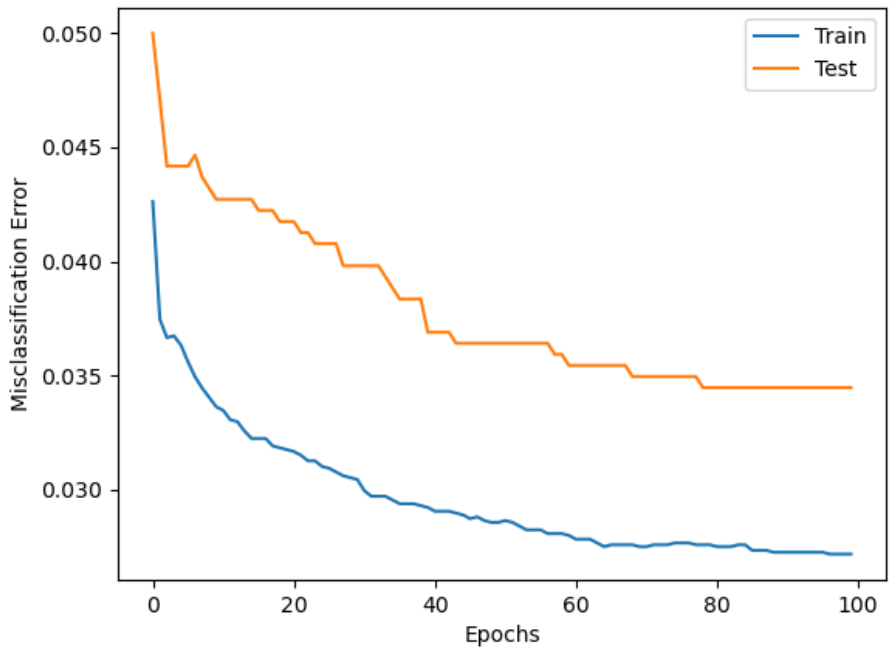
$$
\begin{aligned}
\nabla_b J(w,b) &= \frac{\partial}{\partial b}\Big(\frac{1}{n}\sum_{i=1}^{n}\log(1+\exp(-y_i(b+x_i^T w))) + \lambda||w||_2^2\Big) \\
&= \frac{1}{n}\sum_{i=1}^{n}\Big(\frac{\partial}{\partial b}(\log(1+\exp(-y_i(b+x_i^T w)))) + \frac{\partial}{\partial b}(\lambda w^\top w)\Big) \\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{-y_i \exp(-y_i(b+x_i^T w))}{1+\exp(-y_i(b+x_i^T w))} + 0 && \text{(chain rule)} \\
&= \frac{1}{n}\sum_{i=1}^{n} -y_i\Big(1 - \frac{1}{1+\exp(-y_i(b+x_i^T w))}\Big) \\
&= \frac{1}{n}\sum_{i=1}^{n} -y_i(1-\mu_i(w,b)) && \text{(since } \mu_i(w,b) = \tfrac{1}{1+\exp(-y_i(b+x_i^T w))}\text{)} \\
&= \boxed{-\frac{1}{n}\sum_{i=1}^{n} y_i(1-\mu_i(w,b))}
\end{aligned}
$$

(b)

(i)Learning rate = 1e-1, epochs = 100, $\lambda = 1e-1$

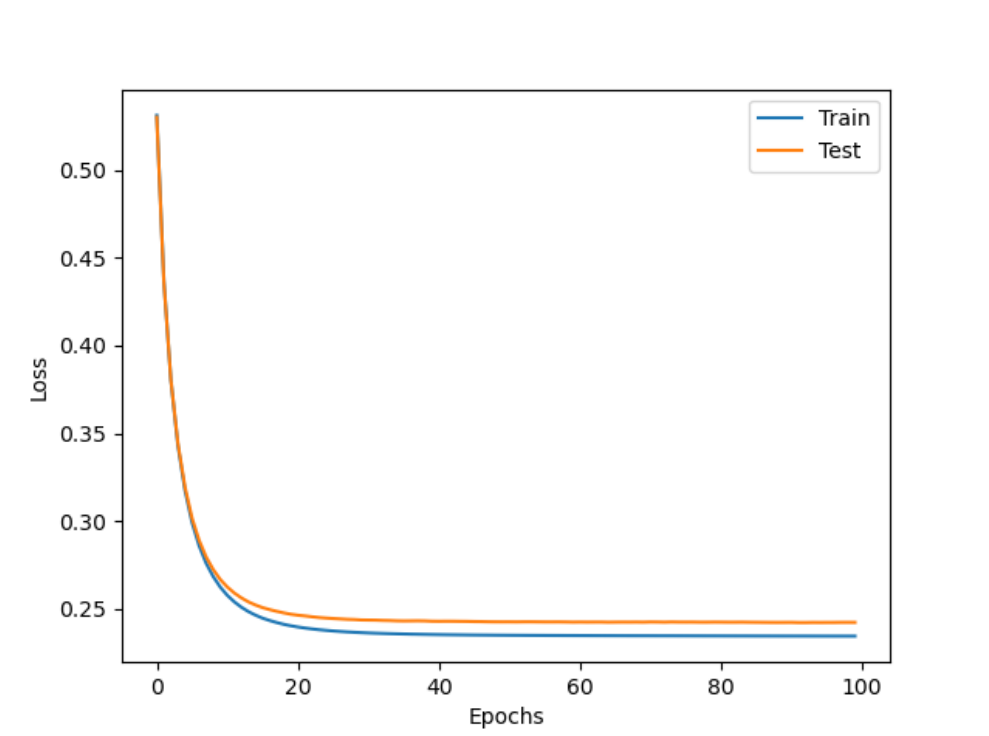


(ii)Learning rate = 1e-1, epochs = 100, $\lambda = 1e-1$

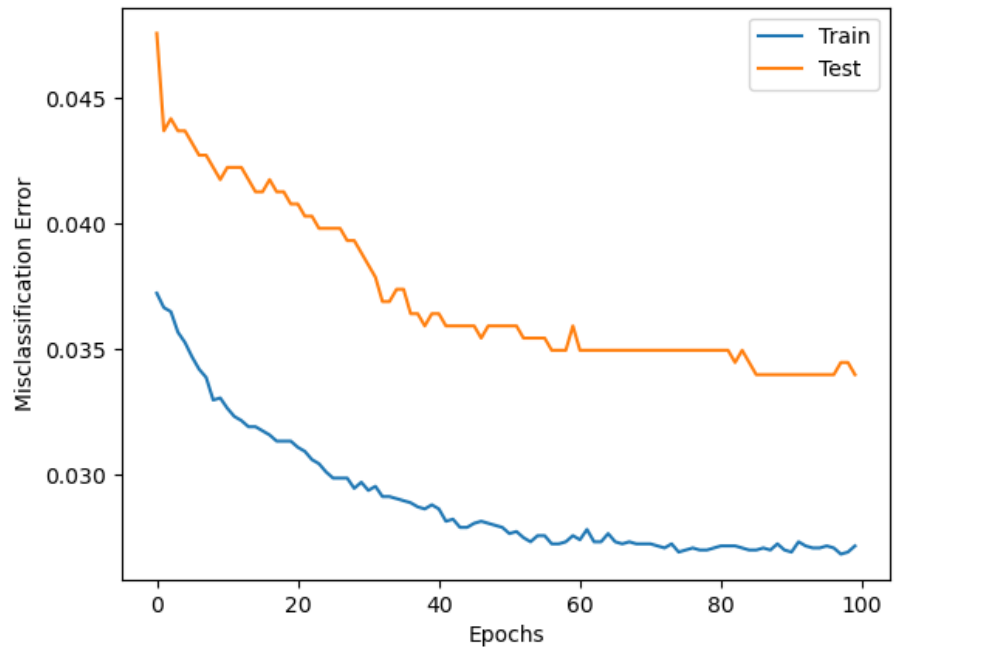


(c)

(i)Learning rate = 1e-5 (1e-4 times the learning rate of GD), epochs = 100, $\lambda = 1e-1$

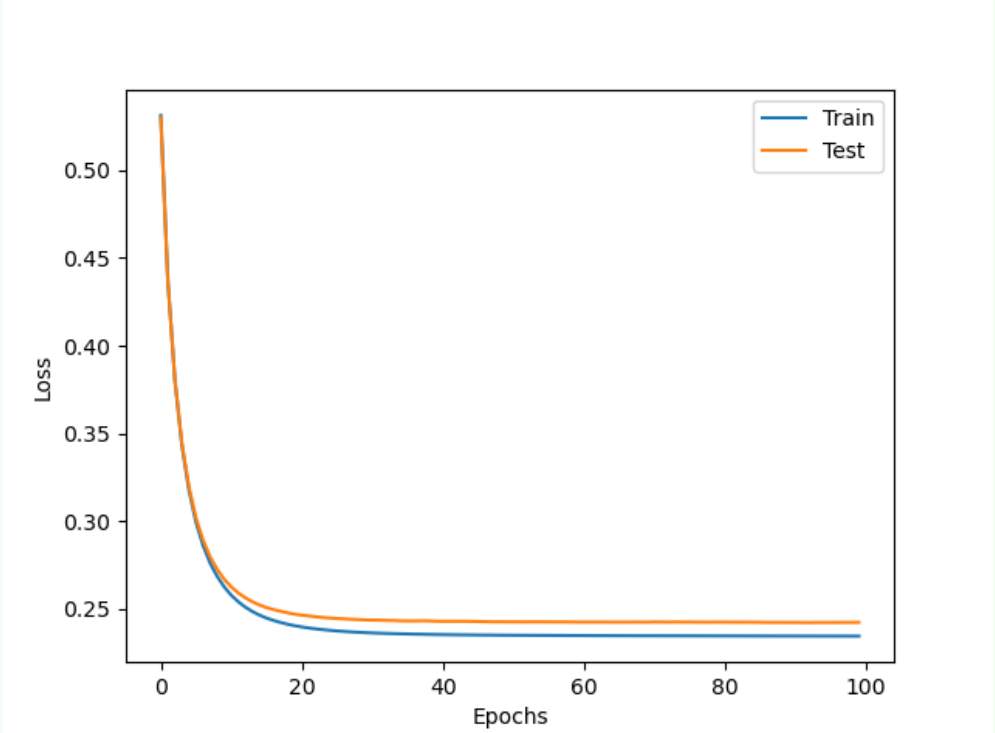


(ii)Learning rate = 1e-5 (1e-4 times the learning rate of GD), epochs = 100, $\lambda = 1e-1$

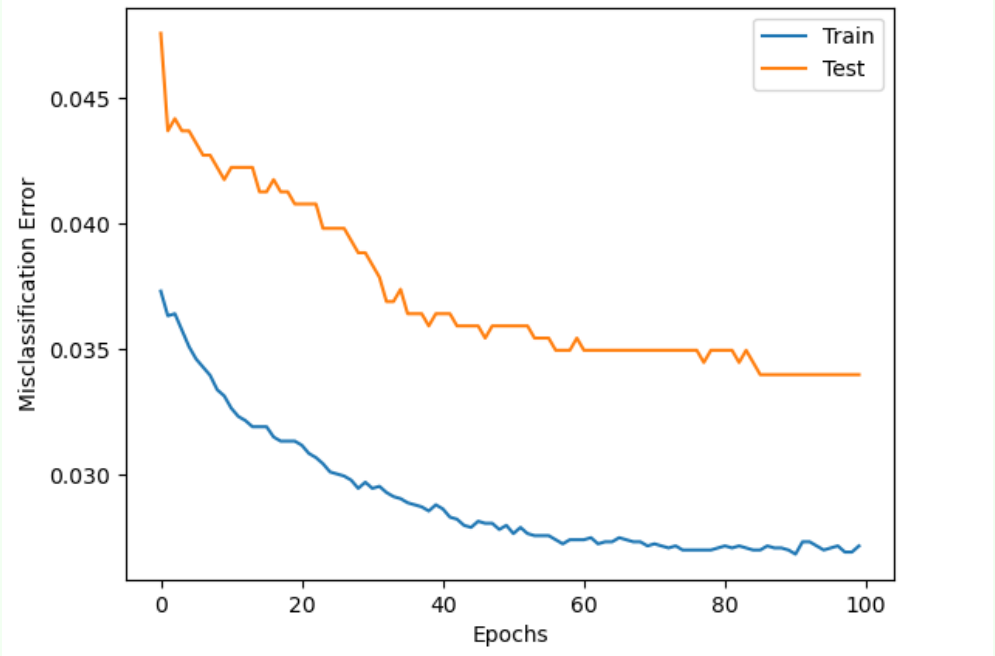


(d)

(i)Learning rate = 1e-3 (100 times the learning rate of batch size 1), epochs = 100, $\lambda = 1e-1$



(ii)Learning rate = 1e-3 (100 times the learning rate of batch size 1), epochs = 100, $\lambda = 1e-1$



A8.

(a)

20 hours