1.

a) The L_2 norm function on $\mathbb{R}^d$ space is not convex: FALSE

b) If f is a convex function, then its second derivative is non-negative at any point where it is continuous (assuming f is twice differentiable): TRUE

c) If f and g are both convex functions, then their combination f+g is also convex: TRUE

2. Fill in the blanks:

a) If $f : \mathbb{R} \to \mathbb{R}$ is convex and $a, b, x \in dom(f)$ with $a < x < b$, then

$$f(x) = \frac{b-x}{b-a} f(a) + \frac{x-a}{b-a} f(b)$$

b) A continuous function $f: \mathbb{R}^n \to \mathbb{R}$ is convex iff for every $x, y \in \mathbb{R}^n$

$$\int_0^1 f(x + \lambda(y-x))\, d\lambda \leq \frac{f(x)+f(y)}{2}$$

3.

a) $f(x) = x^2$ is 5-lipshitz over $\mathbb{R}$: FALSE

b) $f(x) = \log(1 + \exp(x))$ is 1-lipshitz over $\mathbb{R}$ is: TRUE

4. The maximum β for which the following are still smooth over $\mathbb{R}$ are:

a) $f(x) = x^2$: 2

b) $f(x) = \log(1 + \exp(x))$: $\frac{1}{4}$

5.

a) $\exists \beta$: f is β-smooth $\Rightarrow \exists \alpha$: f is α-Lipschitz: TRUE

Optimizers

Question 2. Suppose we have the following objective function

$$f(x) = \frac{1}{2}(x_1^2 + c x_2^2)$$

where $c > 0$

1. What is the optimal point?

$$\nabla f(x) = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2} \right) = (x_1, Cx_2)$$

Set $\nabla f(x) = \vec{0}$

$$x_1 = 0 \Rightarrow x_1^* = 0$$

$$Cx_2 = 0$$

since $C > 0 \Rightarrow x_2^* = 0$

since f is a sum of quadratic terms, f is convex and the optimal point is $(0,0)$. Hessian matrix is $H = \begin{bmatrix} 1 & 0 \\ 0 & C \end{bmatrix}$ which is positive definite, thus this point is the global maximum.

2. Suppose we apply Newton's method. Derive the following closed form expressions for the iterates $x^{(k)}$ and values $f(x^{(k)})$

In Newton's method

$$x^{(k+1)} = x^{(k)} - H_f^{-1} \nabla f(x^{(k)})$$

$$\nabla f(x) = \begin{bmatrix} x_1 \\ cx_2 \end{bmatrix}$$

$$H_f = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & c \end{bmatrix}$$

$$H_f^{-1} = \begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{c} \end{bmatrix}$$

The formula yields the following expression:

$$\begin{bmatrix} x_1^{(k+1)} \\ x_2^{(k+1)} \end{bmatrix} = \begin{bmatrix} x_1^{(k)} \\ x_2^{(k)} \end{bmatrix} - \begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{c} \end{bmatrix} \cdot \begin{bmatrix} x_1^{(k)} \\ c\,x_2^{(k)} \end{bmatrix}$$

$$= \begin{bmatrix} x_1^{(k)} \\ x_2^{(k)} \end{bmatrix} - \begin{bmatrix} x_1^{(k)} \\ x_2^{(k)} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\Rightarrow \vec{x}^{*} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$, Newton's method converges in 1 step.

3. How does the value converge?
Say in terms of how fast the objective function value changes.

The objective function decreases from $f(\vec{x}^{(0)})$ to 0 in a single iteration

4. What can you say about convergence when $c = 1$?

When $c = 1$, the Hessian matrix is the identity. Newton's method still converges in one iteration.

5. Suppose starting at point (1,0). For what learning rates will gradient descent converge?

Gradient descent:

$$\vec{x}^{k+1} = \vec{x}^{k} - \eta \nabla f(\vec{x}^{(k)})$$

$$\nabla f(\vec{x}^{(k)}) = \begin{bmatrix} x_1^{(k)} \\ c x_2^{(k)} \end{bmatrix}, \text{ then get}$$

$$\begin{bmatrix} x_1^{(k+1)} \\ x_2^{(k+1)} \end{bmatrix} = \begin{bmatrix} x_1^{(k)} \\ x_2^{(k)} \end{bmatrix} - \eta \begin{bmatrix} x_1^{(k)} \\ x_2^{(k)} \end{bmatrix}$$

This yields the following 2 equations

$$x_1^{(k+1)} = x_1^{(k)}(1-\eta)$$

$$x_2^{(k+1)} = x_2^{(k)}(1-\eta c)$$

Since we are starting at $x_1^{(0)} = 1$ and $x_2^{(0)} = 0$, this can be simplified to the general case

$$x_1^{(k)} = (1-\eta)^k$$

$$x_2^{(k)} = 0$$

For this to converge, we need that $x_1 \to 0$ as $k \to \infty$. This holds true if

$$|1-\eta| < 1$$

$$-1 < 1-\eta < 1$$

$$-2 < -\eta < 0$$

$$0 < \eta < 2$$

Gradient descent converges if learning rate η is $0 < \eta < 2$

Question 3

1.

Layer	Output Dimensions	Number of params
Input	32 x 32 x 3	0
Conv(3,12)	32 x 32 x 12	336
BatchNorm	32 x 32 x 12	24
ReLU	32 x 32 x 12	0
MaxPooling(2)	16 x 16 x 12	0
conv(3,8)	16 x 16 x 8	872
BatchNorm	16 x 16 x 8	16
ReLU	16 x 16 x 8	0
MaxPooling(2)	8 x 8 x 8	0
Reshape	512 x 1 x 0	0
FC(10)	10 x 1 x 0	5130

2. In the above step, how many parameters can I remove and still keep the output the same? Explain why?

We may remove 20 parameters, consisting of the 12 and 8 bias parameters from the 2 convolutional layers. This is because the batch normalization after the convolutional layer makes the biases redundant.

3.

Layer	Receptive Field
Conv(6, 1) with stride 3	6
MaxPooling(2) with stride 2	9
Conv(2, 1) with stride 1	15
MaxPooling(3) with stride 3	27
Conv(2, 1) with stride 1	45

Question 4

1. a) What should our risk be in this case?

The given, risk is

$$R(g) = \mathop{E}_{x,y \sim P(X,Y)} [\, l(g(x), y) \,]$$

However, we no longer have labels, so the new risk should be

$$R(g) = \mathop{E}_{x \sim p(x)} [\, p(y=+1|x) \cdot l(g(x), +1) + p(y=-1|x) \cdot l(g(x), -1) \,]$$

$$= \mathop{E}_{x \sim P(X)} [\, c(x) \cdot l(g(x), +1) + (1 - c(x)) \cdot l(g(x), -1) \,]$$

b) Given the risk we defined, how should empirical risk be defined here?

Given $\{x_i, c_i\}_{i=1}^{N}$

$$\hat{R} = \frac{1}{N} \sum_{i=1}^{N} \left[c_i \cdot l(g(x_i), +1) + (1 - c_i) \cdot l(g(x_i), -1) \right]$$

c) Why is the fdlowing training objective function not appropriate to minimize:

$$L = \sum_{i=1}^{N} [C_i \cdot \ell(g(x_i), +1) + (1 - C_i)\, \ell(-g(x_i), +1)]$$

The problem is the second term $(1 - C_i)\, \ell(-g(x_i), +1)$. For this loss to be appropriate, we would need that

$$(1 - C_i)\, \ell(-g(x_i), +1) = (1 - C_i)\, \ell(g(x_i), -1),$$

but it does not necessarily hold that

$$\ell(-g(x_i), +1) = \ell(g(x_i), -1)$$

2

a) The Risk of 0-1 loss is defined as

$$R = P(Y=1) \cdot P(g(x) \neq 1 \mid Y=1) +$$

$$P(Y=-1) \cdot P(g(x) \neq -1 \mid Y=-1)$$

$$= \theta \cdot P(g(x) \neq 1 \mid Y=1) + (1-\theta) \cdot P(g(x) \neq +1 \mid Y=-1)$$

$$= \theta \cdot P(g(x) = -1 \mid Y=1) + (1-\theta) \cdot P(g(x) = 1 \mid Y=-1)$$

$$= \theta \cdot \underset{x \sim p_+}{P}(g(x) = -1) + (1-\theta) \cdot \underset{x \sim p_-}{P}(g(x) = 1)$$

This can be written as

$$R = P(Y=1) \cdot \underset{x \sim p_+(x)}{E}\left[\mathbb{1}(g(x) \neq +1\right] +$$

$$P(Y=-1) \cdot \underset{x \sim p_-(x)}{E}\left[\mathbb{1}(g(x) \neq -1\right]$$

Suppose $P(Y=1)=1$, then our risk is

$$R(g) = \underset{x \sim p_+(x)}{E}\left[\mathbb{1}(g(x) \neq +1)\right]$$, this

gives us that $p^{tr} = p_+(x)$

we may then estimate R as

$$\hat{R}(g) = \frac{1}{|p^{tr}|} \sum_{i=1}^{|p^{tr}|} \mathbb{1}(g(x_i) \neq 1)$$

A similar argument holds when $P(Y=1)=0$.

These are the only scenarios where we may estimate R, when we know p^{tr} is only a single probability distribution, not a mix of many.

3.

a) Your friend tells you that instead of minimizing $\hat{R}(g)$, he/she minimizes

$$\tilde{R}(g) = |\hat{R}(g) - \varepsilon| + \varepsilon$$

where ε is some constant. Can you explain what they are attempting to do and how it works.

Instead of trying to minimize the discrepancy between $g(x)$ and y, this method attempts to make the discrepancy between $g(x)$ and y equal to ε. Smaller ε values will minimize the discrepancy

between $g(x)$ and y, like with $\hat{R}(g)$. With larger values of ε, the values of $g(x)$ and y will be increasingly dissimilar.

b) Your friend tells you their objective is better than yours. Can you justify their claim.

Their method has the potential to generalize to unseen data better, as ε can prevent you from over fitting.

Suppose you want to train a linear model $\hat{y} = ax$ and have the following data.

Train	Test
(1, 1.15)	(1,1)
(2, 2.30)	(2,2)

Training with MSE yields the fitted model : $\hat{y}_{(1)} = 1,15x$

with test error : $(0,15)^2 + (0,3)^2 = 0,1125$

Hawever, if you train with

$\min \tilde{R}(g) = \min | MSE - \varepsilon | + \varepsilon$

and set $\varepsilon = 0,1125$, you may yield the fitted model $y = x$

with test error : $(1-1)^2 + (2-2)^2 = 0$

c) Your friend claims their objective can be even more effective if ε is chosen more carefully. Explain how.

If ε is chosen to be the actual irreducible error, then $|R(g) - \varepsilon| + \varepsilon$ correctly reflects error due to bias and variance. Tuning ε can yield a better bias/variance balance. Decreasing ε can lead to lower bias but higher variance while increasing ε will do the opposide. Tune to find the best balance!