Ex 5.1

a) If f is convex, by definition:

$$f(tx_1+(1-t)x_2) \leqslant t\,f(x_1) - (1-t)f(x_2)$$

where $t \in (0,1)$
and $x_1, x_2 \in \mathbb{R}^n$

Given $g(x) = f(Ax+b)$, it will be convex when the function follows the same rule.

$$\Rightarrow g(tx_1 + (1-t)x_2)$$

$$= f(Atx_1 + A(1-t)x_2 + b)$$

$$= f(Atx_1 + Ax_2 - Atx_2 + b)$$

$$= f(Atx_1 + bt + Ax_2 - Atx_2 + b - bt)$$

$$= f(t(Ax_1+b) + Ax_2(1-t) + b(1-t))$$

$$= f(t(Ax_1+b) + (1-t)(Ax_2+b))$$

$$\leqslant tf(Ax_1+b) + (1-t)f(Ax_2+b)$$

$$\leqslant tg(x_1) + (1-t)g(x_2)$$

Therefore, this function is also convex.

b) Given m^* a local minimum of function over a convex set X, for any x_i, $x_i - m^*$ is a possible direction.

For any $x_i \in X$, we can write:

$$f(m^*) \leq f(m^* + \alpha(x_i - m^*)) \quad \text{— (1)}$$

Since f is convex,

$$f(m^* + \alpha(x_i - m^*)) = f(\alpha x_i + (1-\alpha) m^*)$$
$$\leq \alpha f(x_i) + (1-\alpha) f(m^*)$$

Therefore, we can write from eq. 1

$$f(m^*) \leq \alpha f(x_i) + (1-\alpha) f(m^*)$$

So, $$f(m^*) \leq f(x_i)$$

Since x_i is any arbitrary point in X, it is safe to say m^* is the global minimum.

Ex. 5.2 (a)

The second order Taylor expansion of f around x_k is

$$f(x_t+k) \approx f(x_t) + f'(x_t)k + \frac{1}{2} f''(x_t) k^2$$

In generalized terms when x is multidimen-sional:

$$f(x_t+k) \approx f(x_t) + \nabla f(x_t)k + \frac{1}{2} H(x_t)k^2$$

Here, x_{t+1} would be defined in a manner to minimize the above expansion in k.

$$\Rightarrow x_{t+1} = x_t + k \quad \text{——— } (1)$$

For the above expansion, the minimum can be found by setting its derivative to zero.

$$\Rightarrow \frac{d}{dk}\left(f(x_t) + \nabla f(x_t)k + \frac{1}{2} H(x_t)k^2\right) = 0$$

$$\Rightarrow \nabla f(x_t) + H(x_t)k = 0$$

$$\Rightarrow H(x_t)k = -\nabla f(x_t)$$

$$\Rightarrow k = -\nabla H(x_t)^{-1} \cdot \nabla f(x_t)$$

Plugging this into eq. 1:

$$x_{t+1} = x_t - \nabla H(x_t)^{-1} \cdot \nabla f(x_t)$$

In practice, a step size ($a \in [0,1]$) is included to prevent divergence. Hence Newton's method can be summed up as

$$x_{t+1} = x_t - a \nabla H(x_t)^{-1} \cdot \nabla f(x_t)$$

b) Newton's method is based on the assumption that the curve near the root is a straight line, that is if you very closer you will find it straight.

Since the method relies on quadratic convergence, it also follows the assumptions of it. So in many cases the failure to converge of Newton's method is due to the violation of quadratic assumption.

In single variable Newton's method, if the derivative of the function is zero, we cannot calculate using this method as the derivative is a denominator.

In the multivariate case, if the hessian matrix is not invertible, the method won't be able to conver

Ex 5.2 (c)

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

$$x_0 = \begin{bmatrix} 1 \\ -1 \end{bmatrix} \quad ; \quad a = 1$$

The update rule for Newton's Method -

$$x_{t+1} = x_t - a\,(H(f)|x_t)^{-1}\,\nabla f(x_t)$$

$$\frac{\partial f}{\partial x_1} = (1+\exp(-x_1^2-x_2^2))^{-2}.\exp(-x_1^2-x_2^2).2x_1$$

$$\frac{\partial^2 f}{\partial x_1^2} = (1+\exp(-x_1^2-x_2^2))^{-2}\frac{d}{dx_1}(2x_1.\exp(-x_1^2-x_2^2))$$
$$+ 2x_1.\exp(-x_1^2-x_2^2)[2(1+\exp(-x_1^2-x_2^2))^{-3}.\exp(-x_1^2-x_2^2).2x_1]$$
$$= (1+\exp(-x_1^2-x_2^2))^{-2}[2\exp(-x_1^2-x_2^2) - 4x_1^2\exp(-x_1^2-x_2^2)]$$
$$+ 2x_1.\exp(-x_1^2-x_2^2)[2(1+\exp(-x_1^2-x_2^2))^{-3}.\exp(-x_1^2-x_2^2).2x_1]$$

$$= \frac{2(1+\exp(-x_1^2-x_2^2)).\exp(-x_1^2-x_2^2)[1-2x_1^2] + 8x_1^2(\exp(-x_1^2-x_2^2))^2}{(1+\exp(-x_1^2-x_2^2))^3}$$

$$= \frac{2\exp(-x_1^2-x_2^2)[1-2x_1^2+\exp(-x_1^2-x_2^2)+6x_1^2\exp(-x_1^2-x_2^2)]}{(1+\exp(-x_1^2-x_2^2))^3}$$

$$\frac{\partial^2 f}{\partial x_1 \partial x_2} = (1+\exp(-x_1^2-x_2^2))^{-2}[-4x_1x_2\exp(-x_1^2-x_2^2)] + 2x_1\exp(-x_1^2-x_2^2)\left[4x_2(1+\exp(-x_1^2-x_2^2))^{-3}\exp(-x_1^2-x_2^2)\right]$$

$$= \frac{-4x_1x_2(1+\exp(-x_1^2-x_2^2))\exp(-x_1^2-x_2^2)+8x_1x_2\exp^2(-x_1^2-x_2^2)}{[1+\exp(-x_1^2-x_2^2)]^3}$$

$$= \frac{4x_1x_2\exp(-x_1^2-x_2^2)[\exp(-x_1^2-x_2^2)-1]}{(1+\exp(-x_1^2-x_2^2))^3}$$

Similarly we find $\frac{\partial^2 f}{\partial x_2^2}$ & $\frac{\partial^2 f}{\partial x_2 \partial x_1}$

$$\frac{\partial^2 f}{\partial x_2^2} = \frac{2\exp(-x_1^2-x_2^2)[1-2x_2^2+\exp(-x_1^2-x_2^2)- x_2^2\exp(-x_1^2-x_2^2)]}{(1+\exp(-x_1^2-x_2^2))^3}$$

$$\frac{\partial^2 f}{\partial x_2 \partial x_1} = \frac{4x_1x_2\exp(-x_1^2-x_2^2)[\exp(-x_1^2-x_2^2)-1]}{(1+\exp(-x_1^2-x_2^2))^3}$$

$$\boxed{f(x_0) = 0.8808}$$

$$\nabla f(x_0) = \begin{bmatrix} 0.2099 \\ -0.2099 \end{bmatrix}$$

Hesian Matrix for $x_0 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$

$$\Rightarrow H(f)|_{x_0} = \begin{bmatrix} -0.0097 & 0.31985 \\ 0.31985 & -0.097 \end{bmatrix}$$

$$(H(f)|x_0)^{-1} = \begin{bmatrix} 0.0949 & 3.1293 \\ 3.1293 & 0.0949 \end{bmatrix}$$

$$(H(f)|x_0)^{-1} \nabla f(x_0) = \begin{bmatrix} -0.6369 \\ 0.6369 \end{bmatrix}$$

$$\Rightarrow x_1 = \begin{bmatrix} 1 \\ -1 \end{bmatrix} - \begin{bmatrix} -0.6369 \\ 0.6369 \end{bmatrix} = \begin{bmatrix} 1.6369 \\ -1.6369 \end{bmatrix}$$

$$\boxed{f(x_1) = 0.9953}$$

$$\nabla(f(x_1)) = \begin{bmatrix} -0.0152 \\ 0.0152 \end{bmatrix}$$

$$H(f)|x_1 = \begin{bmatrix} -0.0397 & 0.0495 \\ 0.0495 & -0.0397 \end{bmatrix}$$

$$(H(f)|x_1)^{-1} = \begin{bmatrix} 45.4150 & 56.6258 \\ 566258 & 45.4150 \end{bmatrix} = \begin{bmatrix} 0.1704 \\ -0.1704 \end{bmatrix}$$

$$(H(f)|x_1)^{-1} \nabla f(x_1) = \begin{bmatrix} 0.1704 \\ -0.1704 \end{bmatrix}$$

$$x_2 = \begin{bmatrix} 1.6369 \\ -1.6369 \end{bmatrix} - \begin{bmatrix} 0.1704 \\ -0.1704 \end{bmatrix}$$

$$= \begin{bmatrix} 1.4665 \\ -1.4665 \end{bmatrix}$$

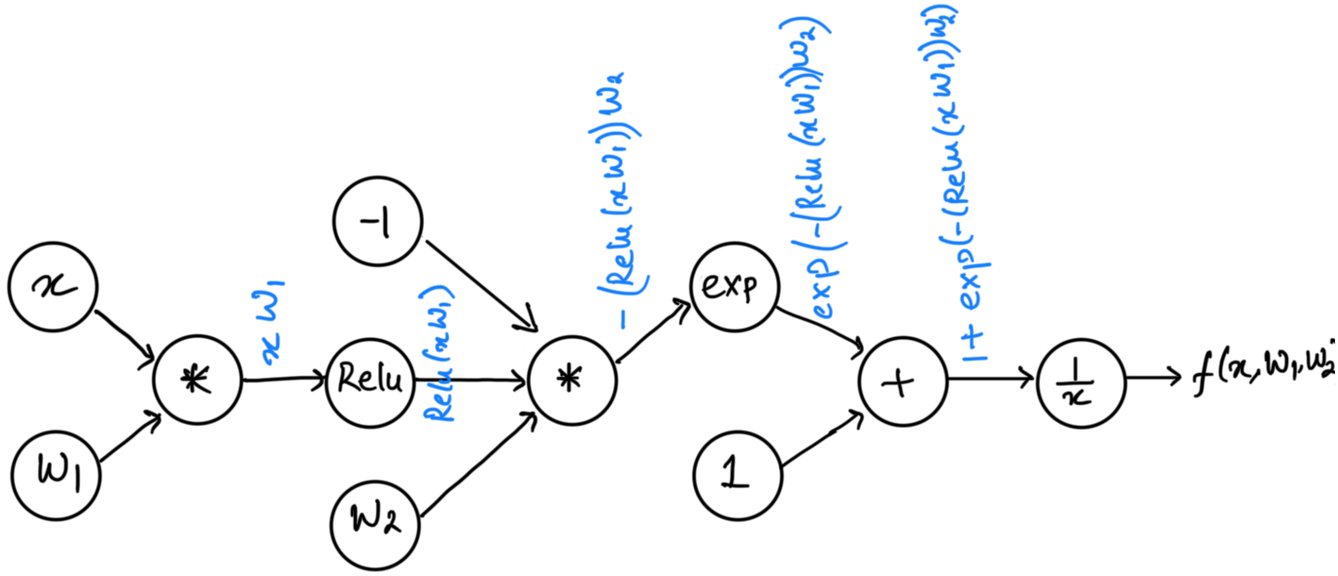
$$\boxed{f(x_2) = 0.9866}$$

In Gradient descent the function value reduces one step at a time. Whereas Newton's method is a more direct method, where we try to compute the roots for $f'(x) = 0$; which would cause direct convergence.

(d) Newton's method tends to find local maxima. This is majorly because it tends to find the roots of $f'(x) = 0$. These roots would then return the minimum value of $f(x)$. But these roots need not necessarily be global.

Ex 5.3

(a) Computation Graph



(b) $$xW_1 = \begin{bmatrix} 1 & -1 \end{bmatrix} \begin{bmatrix} 0.1 & -0.3 \\ 0.5 & -0.8 \end{bmatrix} = \begin{bmatrix} -0.4 & 0.5 \end{bmatrix}$$

$$\text{Relu}(xW_1) = \begin{bmatrix} 0 & 0.5 \end{bmatrix}$$

$$(\text{Relu}(xW_1))W_2 = \begin{bmatrix} 0 & 0.5 \end{bmatrix} \begin{bmatrix} -0.3 \\ -0.1 \end{bmatrix} = -0.5$$

$$f(x, W_1, W_2) = \frac{1}{1+\exp(-(-0.5))} = 0.37754$$

(c) Binary cross entropy loss (L)

$$= -y \log(f(x, W_1, W_2)) - (1-y)\log(1 - f(x, W, W_2))$$

$$= -1 \cdot \log(0.37754)$$

$$= 0.42303$$

Ex. 5.5

Given an sample $T = (x_i)_{i=1}^{n}$
and best possible estimator $= \eta$

$$f^n = \hat{f}(x)$$
$$f(x) = f$$

$$\text{Bias } f_n = \mathbb{E}_T[f_n] - f$$
$$\text{Var } f_n = \mathbb{E}_T[(f_n - \mathbb{E}_T f_n)^2]$$

Defining the noise term with the best estimator

$$Y = \eta(X) + \epsilon \quad \text{———} \quad ①$$
$$\epsilon = Y - \eta(X)$$

$$\text{Var}(\epsilon) = \mathbb{E}[\epsilon^2] = \mathbb{E}[(Y - \eta(X))^2] \quad \text{——} \quad ②$$

Expected error loss:

$$L(f_n) = \mathbb{E}[(Y - f_n(X))^2]$$
$$= \mathbb{E}_T[(f + \epsilon - f_n(x_i))^2] \qquad \text{[from eq. 1]}$$
$$= \mathbb{E}_T[(f - f_n(X_i))^2] + \mathbb{E}_T[\epsilon]^2 + 2\mathbb{E}_T[f - f_n(X_i)\epsilon]$$
$$= \mathbb{E}_T[(f - f_n(X_i))^2] + \mathbb{E}_T[\epsilon]^2 + 2\mathbb{E}_T[(f - f_n(X_i))] + \mathbb{E}_T[\epsilon]$$
$$= \mathbb{E}_T[(f - f_n(X_i))^2] + \mathbb{E}_T[\epsilon]^2 + 0$$

[since $\mathbb{E}_T[\epsilon] = 0$]

Analyzing the 1st term:

$$= \mathbb{E}_T\left[\left((f - \mathbb{E}_T[f_n(X_i)]) - (f_n(X_i) - \mathbb{E}_T[f_n(X_i)])\right)^2\right]$$

$$= \mathbb{E}_T\left[(\mathbb{E}_T[f_n(X_i)] - f)^2\right] + \mathbb{E}_T\left[(f_n(X_i) - \mathbb{E}_T[f_n(X_i)])^2\right]$$
$$- 2\mathbb{E}_T\left[(f - \mathbb{E}_T[f_n(X_i)])(f_n(X_i) - \mathbb{E}_T[f_n(X_i)])\right]$$

$$= (\mathbb{E}_T[f_n(X_i)] - f)^2 + \mathbb{E}_T\left[(f_n(X_i) - \mathbb{E}_T[f_n(X_i)])^2\right]$$
$$- 2\mathbb{E}_T\left[(f - \mathbb{E}_T[f_n(X_i)])(f_n(x_i) - \mathbb{E}_T[f_n(X_i)])\right]^*$$

[Since $f_n(X_i)] - f$ is just a constant the first term is reduced to $(\mathbb{E}_T[f_n(X_i)] - f)^2$]

* given linearity of expectation

$$= \underbrace{(\mathbb{E}_T[f_n(X_i)] - f)^2}_{(\text{Bias } f_n)^2} + \underbrace{\mathbb{E}_T\left[(f_n(X_i) - \mathbb{E}_T[f_n(X_i)])^2\right]}_{\text{Var } f_n}$$

Plugging this to Eq. 3 we get

$$= \mathbb{E}_T (\text{Bias } f_n)^2 + \mathbb{E}_T (\text{Var } f_n) + \mathbb{E}[\epsilon]^2$$

Considering all data points and replacing the value of $\mathbb{E}[\epsilon]^2$ from eq. 3 we can write:

$$\mathbb{E}_T L(f_n) = \mathbb{E}\left[(Y - \eta(X))^2\right] + \mathbb{E}_X[\text{Var } f_n(X)]$$
$$+ \mathbb{E}_X\left[(\text{Bias } f_n(X))^2\right]$$