Q2. (a)

The pseudoinverse ($A^+$) of a matrix ($A$) in linear algebra is a generalization of the inverse matrix. Pseudo-inverse is most frequently used to find the fit(best) solution to a system of linear equations that doesn't have a single solution. The most-well-known variety of matrix pseudoinverse is the Moore-Penrose inverse. Pseudo inverse of a matrix is given by

$$A^+ = (A^T A)^{-1} A^T$$

(i) Under-determined system of equations:

let $y = Mx$

where $M = n \times m$ matrix

A system is said to be under-determined if the number of variables in the system is greater than the no. of equations. In this case, the system may have an infinite no. of solution. We can pick one of these solutions by finding the smallest one, i.e, minimize $x$ subject to the equation

$$y = Mx$$

$$\Rightarrow \quad \|x\|^2 + \lambda^T (y - Mx)$$

$$\frac{\partial (\|x\|^2 + \lambda^T (y - Mx))}{dx} = 0$$

$$\Rightarrow 2x + 0 - M^T \lambda = 0$$

$$\Rightarrow 2Mx - MM^T\lambda = 0 \text{ (Multiplying L/s by } M)$$

We know $y = Mx$

$$\Rightarrow 2y - MM^T\lambda = 0$$

$$2y = MM^T\lambda$$

$$\Rightarrow \lambda = 2(MM^T)^{-1}y$$

$$\Rightarrow x = M^T(MM^T)^{-1}y$$

$\therefore$ $M^T(MM^T)^{-1}$ → is the pseudo-inverse for under determined system of equations.

(ii) Over-determined system of equations:
A system is termed overdetermined if it has a much higher number of equations than the no. of variables. In this case, the system may have many or no solutions. We can find a least-squares solution that minimizes the error $(y - Mx)$

$$\|y - Mx\|^2$$

$$(y - Mx)^T(y - Mx)$$

$$y^Ty - y^TMx - x^TM^Ty + x^TMx$$

differentiating w.r.t. $x$ we get

$$(-y^T M)^T - M^T y + 2M^T M x = 0$$

$$x = (M^T M)^{-1} M^T y$$

$\therefore (M^T M)^{-1} M^T$ is the pseudo-inverse for over-determined system of equations.

(b)

$$x_1 + 3x_2 = 17$$

$$5x_1 + 7x_2 = 19$$

$$11x_1 + 13x_2 = 23$$

$$\Rightarrow Mx = y$$

$$\begin{bmatrix} 1 & 3 \\ 5 & 7 \\ 11 & 13 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 17 \\ 19 \\ 23 \end{bmatrix}$$

we know pseudo Inverse of a matrix is given by (of a overdetermined system)

$$A^+ = (A^T A)^{-1} A^T$$

$$A^T A = \begin{bmatrix} 1 & 5 & 11 \\ 3 & 7 & 13 \end{bmatrix} \begin{bmatrix} 1 & 3 \\ 5 & 7 \\ 11 & 13 \end{bmatrix}$$

$$= \begin{bmatrix} 147 & 181 \\ 181 & 227 \end{bmatrix}$$

$$A^T = \begin{bmatrix} 1 & 5 & 11 \\ 3 & 7 & 13 \end{bmatrix} \qquad (A^T A)^{-1} = \begin{bmatrix} 0.37 & -0.29 \\ -0.29 & 0.24 \end{bmatrix}$$

$$\Rightarrow A^+ = \begin{bmatrix} -0.519 & -0.217 & 0.236 \\ 0.427 & 0.207 & -0.1315 \end{bmatrix}$$

Since it is a over-determined system of equations we know

$$x = A^+ y$$

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \underset{2\times 3}{\begin{bmatrix} -0.519 & -0.217 & 0.236 \\ 0.427 & 0.2039 & -0.1316 \end{bmatrix}} \underset{3\times 1}{\begin{bmatrix} 17 \\ 19 \\ 23 \end{bmatrix}}$$

$$\boxed{\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} -7.513 \\ 8.118 \end{bmatrix}}$$

1(c) (i) Hypothesis function:

$$h_\theta(x) = \theta_0 x_0 + \theta_1 x_1 + \cdots + \theta_n x_n$$

Minimizing least squares cost

$$J(\theta_{0\ldots n}) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2$$

$x^{(i)} \rightarrow i^{th}$ sample (from a set of m samples)

Now, hypothesis function:

$\theta \rightarrow$ vector $\begin{pmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_n \end{pmatrix} \in R^{n+1}$

$$\therefore \quad h_\theta(x) = \theta^T x$$

$$J(\theta) = \frac{1}{2m} (X\theta - y)^T (X\theta - y)$$

ignoring

$$J(\theta) = ((X\theta)^T - y^T)(X\theta - y)$$

$$= ((X\theta)^T - y^T)(X\theta - y) \quad \because (a-b)^T = a^T - b^T$$

$$= (X\theta)^T (X\theta) - (X\theta)^T y - y^T (X\theta) + y^T y$$

$$J(\theta) = \theta^T X^T X \theta - 2(X\theta)^T y + y^T y$$

Now, minimizing loss

$$\frac{\partial J(\theta)}{\partial \theta} = X^T X \frac{\partial \theta^T \theta}{\partial \theta} - 2X^T y \frac{\partial \theta^T}{\partial \theta} + 0 = 0$$

$$2X^T X \theta - 2X^T y = 0$$

$$X^T X \theta = X^T y \qquad \boxed{\Rightarrow \theta = (X^T X)^{-1} X^T y}$$

(ii) For closed form solution we need matrix computation, for a feature space of dimension space of 3, 4 or even 5 calculating the inverse of a matrix (5×5) is not that computationally heavy but for d 1000, million or billion dimensional data where dimension describes the features, calculating or computing the inverse of such a huge matrix is expensive as well as not judicious, this drawback is overcome by methods such as gradient descent which step wise or iteratively tries to reach a minima (local or global) reducing the computation overhead by a large factor.

Q3. (c)

we have

$$z = w_1 x_1 + w_2 x_2 + \cdots w_n x_n + w_0$$

let's define $g(z) = \tanh(z)$ be the predicted class by our model,

the log-loss, $l$, thus will be given as

$$l = -y \log g(z) - (1-y) \log(1-g(z))$$

where, $y \rightarrow$ ground truth label.

$$\frac{\partial l}{\partial w_1} = \frac{-y}{g(z)} \frac{\partial g(z)}{\partial w_1} - \frac{(1-y)}{1-g(z)} \cdot \frac{-\partial g(z)}{\partial w_1}$$

$$= \frac{-y}{g(z)} \cdot g'(z) \cdot \frac{\partial z}{\partial w_1} + \frac{(1-y)}{1-g(z)} \cdot g'(z) \cdot \frac{\partial z}{\partial w_1}$$

we know,

$$\frac{\partial z}{\partial w_1} = x_1$$

$$\frac{\partial l}{\partial w_1} = x_1 \left[ \frac{1-y}{1-g(z)} \cdot g'(z) - \frac{g'(z) \cdot y}{g(z)} \right]$$

$$= x_1 \left[ \frac{-y g'(z) + g(z) g'(z)}{g(z)(1-g(z))} \right.$$

$$= \left[ \frac{-y(1+g(z)) + g(z)(1+g(z))}{g(z)} \right] x_1$$

$$= \left[ \frac{-y - y\,g(z) + g(z) + g^2(z)}{g(z)} \right] x_1$$

$$= \left[ 1 + g(z) - y - \frac{y}{g(z)} \right] x_1$$

$$\boxed{\therefore\ w_{1\,new} = w_{1\,old} - \eta \left( 1 + g(z) - y - \frac{y}{g(z)} \right) x_1}$$

where $g(z) = \tanh(z)$

& $g'(z) = 1 - \tanh^2(z)$