Neural Networks and Deep Learning HW #3

1. Given $x \in \mathbb{R}^n$, $W \in \mathbb{R}^{m \times n}$ where $m < n$.

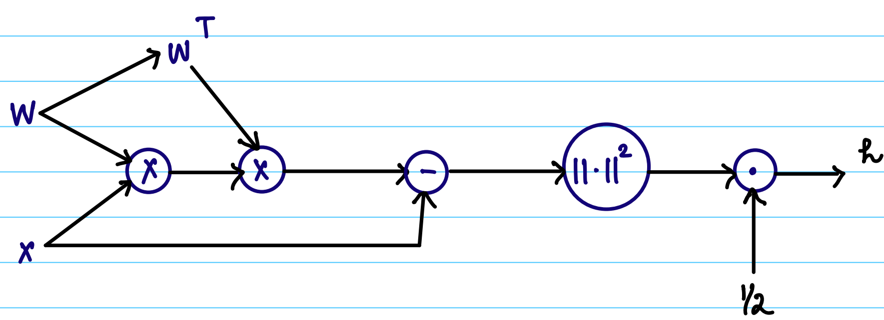
a) First let's consider the loss equation

$$L = \frac{1}{2} \| W^T W x - x \|^2$$

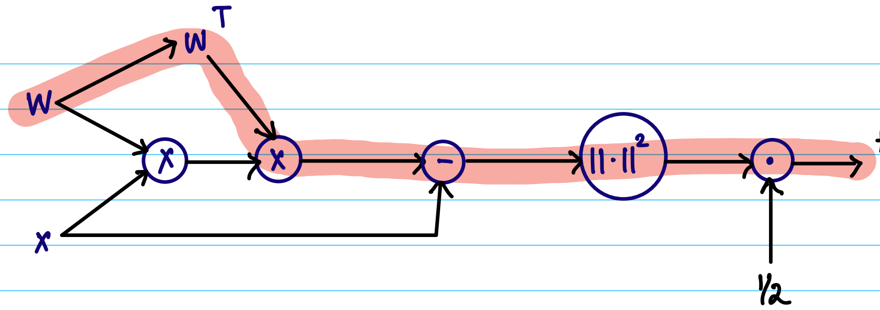
* We know that the loss L should be minimized, i.e $W^T W x - x$ must be minimized.

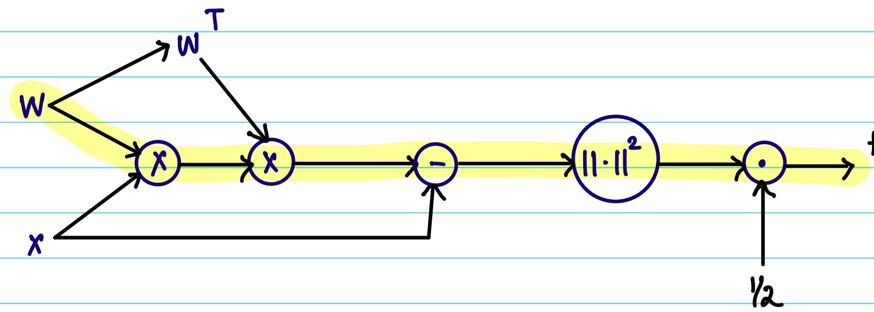
* Wx is the hidden representation and for the loss to be minimized, Wx will have to preserve information about x.

b) Computation graph:-

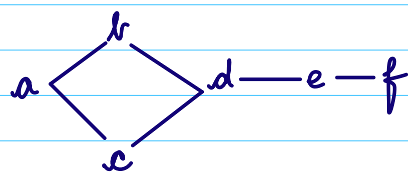


c) The two paths to W are highlighted below:





Since W contributes to two paths, the ∇_W^h will be expressed as a sum of derivatives along each path. Let's consider this case:



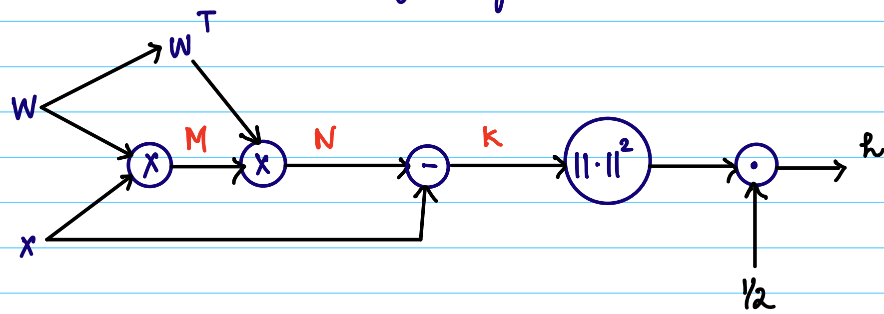
This graph represents the above scenario and consists of two paths to reach a.

$a \to b \to d \to e \to f$

$a \to c \to d \to e \to f$

Hence
$$\boxed{\begin{aligned} \partial f/\partial a &= \partial f/\partial e \cdot \partial e/\partial d \cdot \partial d/\partial b \cdot \partial b/\partial a \\ &\quad + \partial f/\partial e \cdot \partial e/\partial d \cdot \partial d/\partial c \cdot \partial c/\partial a \end{aligned}}$$

d) Let's consider the computation graph from 2) :



* $K = W^T W x - x$

where $W^T = n \times m$

$W = m \times n$

$x = n \times 1$

Hence $K = W^T W x - x \in \mathbb{R}^n$

* $N = W^T M$

$N = W^T \cdot W x$

where

$W^T = n \times m$

$Wx = m \times 1$

Hence $N = W^T M \in \mathbb{R}^n$

* $M = W x$

where $W = m \times n$

$x = n \times 1$

Hence $M = Wx \in \mathbb{R}^m$

* $\ell = \frac{1}{2} \|K\|^2$, $\partial L / \partial K = K \longrightarrow 0$

* $K = N - x$, $\partial L / \partial N = \partial K / \partial N \times \partial L / \partial K$ (Chain rule)

We know that $\partial L / \partial K = K$

$\partial K / \partial N = \partial (N - x) / \partial N = I$

Hence $\partial L / \partial N = K \longrightarrow 1$

We could also say that $\partial L / \partial K = \partial L / \partial N$ as $-$ operator distributes the gradient.

* $N = W^T M$ where $M = Wx \in \mathbb{R}^m$

$\partial h / \partial W^T = \partial h / \partial N \cdot \partial N / \partial W^T$

$\partial h / \partial N = K \in \mathbb{R}^n$

$\partial N / \partial W^T = \partial (W^T M) / \partial W^T = M^T = (Wx)^T \in \mathbb{R}^{1 \times m}$

Hence $\partial h / \partial W^T = K (Wx)^T \longrightarrow$ 2

* $\partial h / \partial M = \partial h / \partial N \cdot \partial N / \partial M$

$\partial h / \partial N = K \in \mathbb{R}^n$

We know that $M = Wx \in \mathbb{R}^m$, hence $\partial h / \partial M \in \mathbb{R}^m$

$\partial N / \partial M = \partial (W^T M) / \partial M = W \in \mathbb{R}^{m \times n}$

Hence $\partial h / \partial M = WK$

* $\partial h / \partial W = \partial h / \partial M \times \partial M / \partial W$

$\partial h / \partial W$ must be $\mathbb{R}^{m \times n}$

$\partial h / \partial M = WK \in \mathbb{R}^m$, $\partial M / \partial W = \partial (Wx) / \partial W = x^T \in \mathbb{R}^{1 \times n}$

$$\therefore \partial h / \partial W = WKX^T \in \mathbb{R}^{m \times n} \longrightarrow (3)$$

From (2) and (3), we know both contribute to $\nabla_W h$ i.e

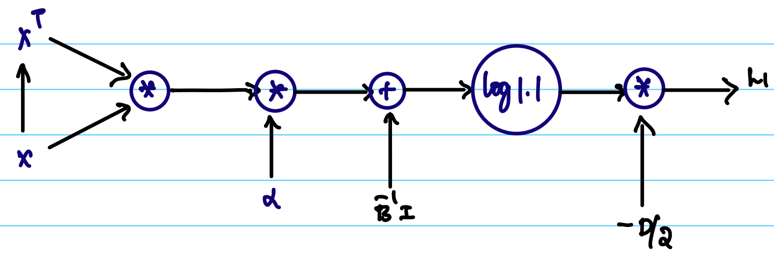
$$\nabla_W h = \text{Backprop of } W + \text{Backprop of } W^T$$

Hence

$$\boxed{\nabla_W h = WKX^T + K(WX)^T = WKX^T + WXK^T}$$

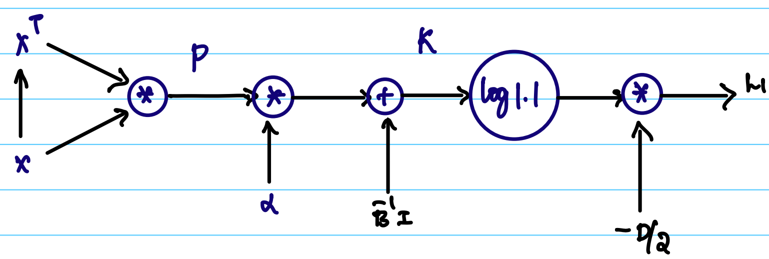
Question 2:

a) Computation Graph for L_1:



b) We need to compute $\partial L_1 / \partial x$

From the question, we know that $K = \alpha x x^T + \beta^{-1} I$



$L_1 = -D/2 \log(|K|)$

$\partial L_1 / \partial K = -D/2 \, (K^{-1})^T = -D/2 \, (K^T)^{-1}$ (From the Matrix cookbook)

Now, we propogate backwards encountering a ⊕ operation, we know ⊕ distributes the gradients and hence the gradient remains the same.

Now, we encounter a ⊛ operator, hence the gradient @ P would be:

$$\partial h_1/\partial P = -\alpha D/2 \cdot (K^T)^{-1}$$

$$\partial h_1/\partial xx^T = -\alpha D/2 \cdot (K^T)^{-1}$$

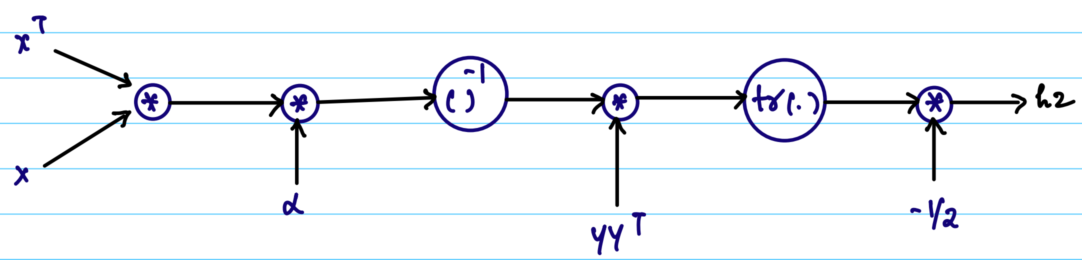
$$P = xx^T$$

$$\partial h_1/\partial x = \partial P/\partial x \cdot \partial h/\partial P$$

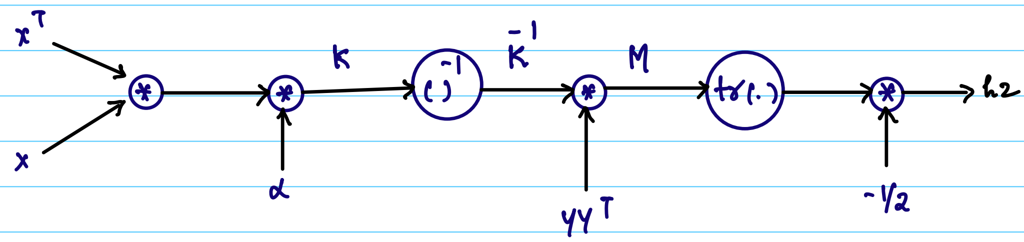
$$\partial P/\partial x = xx^T = 2x$$

$$\boxed{\therefore \partial h_1/\partial x = -\alpha D (K^T)^{-1} x}$$

c) Computational graph for h2:



d)



$$h_2 = -\tfrac{1}{2}\, tr\left((\alpha x x^T + B^{-1} I)^{-1} y y^T \right)$$

$$h_2 = -\tfrac{1}{2}\, tr(M)$$

$$\partial h_2 / \partial M = -\tfrac{1}{2} I$$ (From Matrix Cookbook)

$$\partial h_2 / \partial K^{-1} = -\tfrac{1}{2} \cdot y y^T$$

$$\partial h_2 / \partial K = \tfrac{1}{2} K^{-1} y y^T K^{-1}$$

Hence from the last part, we could say that:

$$\boxed{\partial h_2 / \partial x = \alpha K^{-1} y y^T K^{-1} x}$$

e) Hence $\partial h / \partial x = \partial h_1 / \partial x + \partial h_2 / \partial x$

$$\boxed{\partial h / \partial x = \alpha K^{-1} y y^T K^{-1} x - \alpha D K^{-1} x}$$