1.

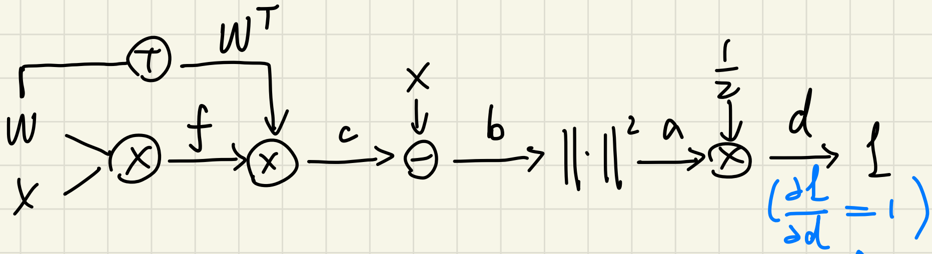
(a) If we use Wx in the loss function, the dimension of Wx is $(m \times n) \times (n \times 1) = (m \times 1)$, where $m < n$.

This means some infomation about x is lost.

However, if we use $W^T W x$ here, the dimension of $W^T W x$ is $(n \times m) \times (m \times n) \times (n \times 1) = (n \times 1)$, which is the same as $\dim(x)$.

In this way, the minimisation can find a W ought to preserve information about x.

(b)



assume

(c) $c = W^T f, \quad f = Wx$

Computing the local gradient,

$$\frac{\partial c}{\partial f} = (W^T)^T = W, \quad \frac{\partial l}{\partial f} = W\frac{\partial l}{\partial c}.$$

$$\left(\frac{\partial l}{\partial W}\right)_1 = \frac{\partial f}{\partial W}\cdot\frac{\partial l}{\partial f} = \frac{\partial l}{\partial f}\cdot x^T = W\frac{\partial l}{\partial c}x^T.$$

Now for the second path W^T:

$$\left(\frac{\partial l}{\partial W}\right)_2 = \left(\frac{\partial l}{\partial W^T}\right)^T = \left(\frac{\partial c}{\partial W^T}\cdot\frac{\partial l}{\partial c}\right)^T = \left(\frac{\partial l}{\partial c}\cdot W^T x^T\right)^T = Wx\left(\frac{\partial l}{\partial c}\right)^T$$

So, $\nabla_W l = \left(\frac{\partial l}{\partial W}\right)_1 + \left(\frac{\partial l}{\partial W}\right)_2 = W\frac{\partial l}{\partial c}x^T + Wx\left(\frac{\partial l}{\partial c}\right)^T$

(d) since ⊗ swithes the gradient,

$$\frac{\partial l}{\partial a} = \frac{1}{2}\frac{\partial l}{\partial d} = \frac{1}{2}.$$

Now, $a = \|b\|_2^2$

$$\frac{\partial a}{\partial b} = 2b \in \mathbb{R}^n. \quad \frac{\partial l}{\partial b} = b \in \mathbb{R}^n$$

since ⊖ passes the gradient,

$$\frac{\partial l}{\partial c} = \frac{\partial l}{\partial b} = b \quad \in \mathbb{R}^n$$

using the resule from question (c),

$$\nabla_W L = W \frac{\partial l}{\partial c} X^T + WX \left(\frac{\partial l}{\partial c}\right)^T$$

$$= W \cdot b \cdot X^T + WX \cdot b^T$$

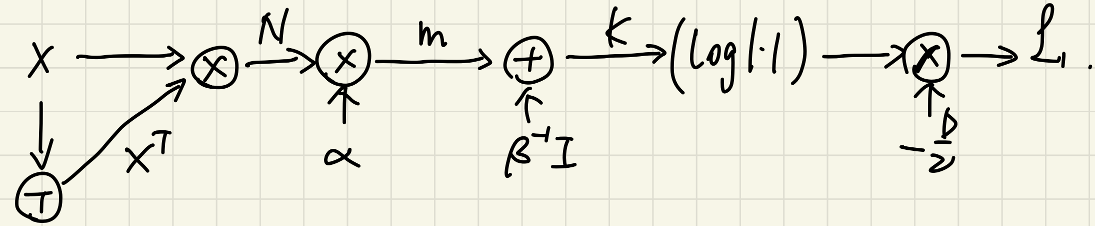
$b = W^T W X - X$,

so $\nabla_W L = W(W^T W X - X) X^T + WX(W^T W X - X)^T$

2. $\mathcal{L}_1 = -\frac{D}{2}\log|\alpha XX^T + \beta^{-1}I|$

$\mathcal{L}_2 = -\frac{1}{2}\mathrm{tr}((\alpha XX^T+\beta^{-1}I)^{-1}YY^T)$.

(a)



(b) $\mathcal{L}_1 = -\frac{D}{2}\log(\det(K))$

using equation 57 from cookbook,

$$\frac{\partial \mathcal{L}_1}{\partial K} = -\frac{D}{2}(K^{-1})^T = -\frac{D}{2}(K^T)^{-1}$$

since ⊕ passes the gradient,

$$\frac{\partial \mathcal{L}_1}{\partial m} = \frac{\partial \mathcal{L}_1}{\partial K} = -\frac{D}{2}(K^T)^{-1}$$

since ⊗ switches the gradient,

$$\frac{\partial \mathcal{L}_1}{\partial N} = -\frac{\alpha D}{2}(K^T)^{-1}$$

$N = XX^T$, using the hint,

$$\left(\frac{\partial \mathcal{L}_1}{\partial X}\right)_1 = \frac{\partial N}{\partial X}\cdot\frac{\partial \mathcal{L}_1}{\partial N} = \frac{\partial \mathcal{L}_1}{\partial N}(X^T)^T = -\frac{\alpha D}{2}(K^T)^{-1}X.$$

$$\left(\frac{\partial \mathcal{L}_1}{\partial X^T}\right)_2 = -\frac{\alpha D}{2}(K^{-1}X)^T$$

so $\frac{\partial \mathcal{L}_1}{\partial X} = \left(\frac{\partial \mathcal{L}_1}{\partial X}\right)_1 + \left(\frac{\partial \mathcal{L}_1}{\partial X^T}\right)_2^T = -\frac{\alpha D}{2}(K^T)^{-1}X - \frac{\alpha D}{2}K^{-1}X$

since $K = \alpha XX^T + \beta^{-1}I$, K is symmetric,

$$\frac{\partial \mathcal{L}_1}{\partial X} = -\alpha D K^{-1} X. \quad (\text{where } K = \alpha XX^T + \beta^{-1}I)$$

(c)

$$\begin{tikzcd}
X \arrow[r] \arrow[d] & \otimes \arrow[r, "N"] & \times \arrow[r, "m"] & + \arrow[r, "K"] & (\cdot)^{-1} \arrow[r] & \times \arrow[r] & \mathrm{tr} \arrow[dd, dash] \\
T \arrow[ur, "X^T"'] & & \alpha \arrow[u] & \beta^{-1}I \arrow[u] & & YY^T \arrow[u] & \\
& & & & \mathcal{L}_2 & \otimes \arrow[l] & {} \arrow[l] \\
& & & & & -\frac{1}{2} \arrow[u] &
\end{tikzcd}$$

(d) $\mathcal{L}_2 = -\frac{1}{2} tr(K^{-1}YY^T)$

using Equation 124 from cookbook,

$$\frac{\partial \mathcal{L}_2}{\partial K} = \frac{1}{2}(K^{-1}YY^TK^{-1})^T$$

since $\oplus$ passes the gradient,

$$\frac{\partial \mathcal{L}_2}{\partial m} = \frac{\partial \mathcal{L}_2}{\partial K}$$

since $\otimes$ switches the gradient,

$$\frac{\partial \mathcal{L}_2}{\partial N} = \frac{\alpha}{2}(K^{-1}YY^TK^{-1})^T$$

$$\frac{\partial \mathcal{L}_2}{\partial X} = \left(\frac{\partial \mathcal{L}_2}{\partial X}\right)_1 + \left(\frac{\partial \mathcal{L}_2}{\partial X^T}\right)_2^T = \frac{\partial \mathcal{L}_2}{\partial N} \cdot X + \left(X^T \frac{\partial \mathcal{L}_2}{\partial N}\right)^T$$

$$= \frac{\alpha}{2}(K^{-1}YY^TK^{-1})^T X + \frac{\alpha}{2}(K^{-1}YY^TK^{-1}) X$$

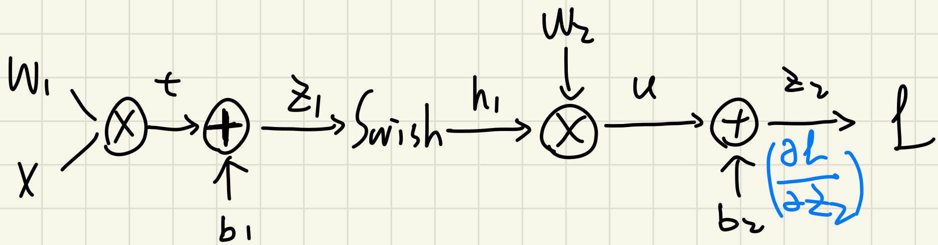
$$= \alpha K^{-1}YY^TK^{-1}X.$$

$$\frac{\partial \mathcal{L}_2}{\partial X} = \alpha K^{-1}YY^TK^{-1}X \quad (\text{where } K = \alpha XX^T + \beta^{-1}I)$$

(e) $$\frac{\partial \mathcal{L}}{\partial X} = \frac{\partial \mathcal{L}_1}{\partial X} + \frac{\partial \mathcal{L}_2}{\partial X} = \alpha\left(K^{-1}YY^TK^{-1}X - DK^{-1}X\right)$$

3.

(a)



(b) since $\oplus$ passes the gradient,

$$\nabla_{b_2} L = \frac{\partial L}{\partial b_2} = \frac{\partial L}{\partial z_2} . \quad \in \mathbb{R}^{C}$$

$$\frac{\partial L}{\partial u} = \frac{\partial L}{\partial z_2} .$$

Now, $u = W_2 h_1$.

$$\nabla_{W_2} L = \frac{\partial L}{\partial W_2} = \frac{\partial L}{\partial u} \cdot h_1^T = \frac{\partial L}{\partial z_2} \cdot h_1^T \quad \in \mathbb{R}^{C \times H} .$$

(c) $$\frac{\partial L}{\partial h_1} = \frac{\partial u}{\partial h_1} \cdot \frac{\partial L}{\partial u} = W_2^T \cdot \frac{\partial L}{\partial u} = W_2^T \frac{\partial L}{\partial z_2} .$$

Now, $h_1 = \text{swish}(z_1) = z_1 \sigma(z_1)$

$$\frac{\partial h_1}{\partial z_1} = z_1 \frac{\partial \sigma(z_1)}{\partial z_1} + \sigma(z_1)$$

$$= z_1 \sigma(z_1)[1 - \sigma(z_1)] + \sigma(z_1)$$

$$= \sigma(z_1)[z_1 + 1 - z_1 \sigma(z_1)]$$

$$\frac{\partial L}{\partial z_1} = \frac{\partial h_1}{\partial z_1} \cdot \frac{\partial L}{\partial h_1} .$$

Since ⊕ passes the gradient,

$$\nabla_{b_1} \mathcal{L} = \frac{\partial \mathcal{L}}{\partial b_1} = \frac{\partial \mathcal{L}}{\partial z_1} = \sigma(z_1)\left[z_1 + 1 - z_1 \sigma(z_1)\right] W_2^T \frac{\partial \mathcal{L}}{\partial z_2}.$$

$$\frac{\partial \mathcal{L}}{\partial t} = \frac{\partial \mathcal{L}}{\partial z_1}$$

Now, $t = W_1 X$

$$\nabla_{W_1} \mathcal{L} = \frac{\partial \mathcal{L}}{\partial W_1} = \frac{\partial \mathcal{L}}{\partial t} \cdot X^T = \sigma(z_1)\left[z_1 + 1 - z_1 \sigma(z_1)\right] W_2^T \frac{\partial \mathcal{L}}{\partial z_2} \cdot X^T$$