C247 HW3 Yiwen Zhang 805899489

1. (15 points) **Backpropagation for autoencoders.** In an autoencoder, we seek to reconstruct the original data after some operation that reduces the data's dimensionality. We may be interested in reducing the data's dimensionality to gain a more compact representation of the data.

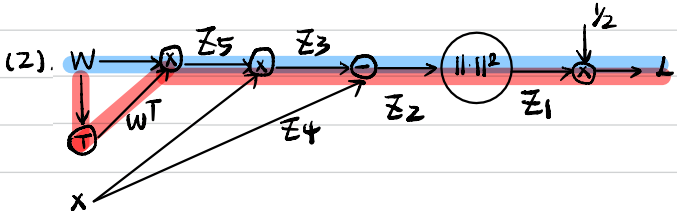
   For example, consider $\mathbf{x} \in \mathbb{R}^n$. Further, consider $\mathbf{W} \in \mathbb{R}^{m \times n}$ where $m < n$. Then $\mathbf{Wx}$ is of lower dimensionality than $\mathbf{x}$. One way to design $\mathbf{W}$ so that $\mathbf{Wx}$ still contains key features of $\mathbf{x}$ is to minimize the following expression

$$\mathcal{L} = \frac{1}{2} \left\| \mathbf{W}^T \mathbf{W} \mathbf{x} - \mathbf{x} \right\|^2$$

   with respect to $\mathbf{W}$. (To be complete, autoencoders also have a nonlinearity in each layer, i.e., the loss is $\frac{1}{2} \left\| f(\mathbf{W}^T f(\mathbf{Wx})) - x \right\|^2$. However, we'll work with the linear example.)

(1) Because if $\mathcal{L} \to 0$, then it means $W^TWx - x \to 0$. regard W^TW as a function $f(\cdot)$. Then $f(x) \to x$, then it means after transformation, $f(x)$ is still very close to original x, it means information of x has been preserved.

(2).



(3). There are W and W^T two paths to account for $\partial L / \partial W$.

$\partial L / \partial W = (\partial L / \partial W)_{\text{blue}} + (\partial L / \partial W^T)^T_{\text{pink}}$

(4). $L = \frac{1}{2} z_1$, $\frac{\partial L}{\partial z_1} = \frac{1}{2}$. $z_1 = \|z_2\|^2$, $\frac{\partial z_1}{\partial z_2} = 2z_2$, $\frac{\partial L}{\partial z_2} = \frac{\partial z_1}{\partial z_2} \cdot \frac{\partial L}{\partial z_1} = z_2$

"Add" passes gradient: $\frac{\partial L}{\partial z_3} = \frac{\partial L}{\partial z_2}$. $z_3 = z_5 \cdot x$. $\frac{\partial L}{\partial z_5} = \frac{\partial z_3}{\partial z_5} \cdot \frac{\partial L}{\partial z_3} = \frac{\partial L}{\partial z_3} \cdot x^T$

$\left(\frac{\partial L}{\partial W}\right)_1 = \frac{\partial z_5}{\partial W} \cdot \frac{\partial L}{\partial z_5} = W \cdot \frac{\partial L}{\partial z_5} = W \cdot z_2 \cdot x^T$, $\left(\frac{\partial L}{\partial W^T}\right)_2 = \frac{\partial z_5}{\partial W^T} \frac{\partial L}{\partial z_5} = \frac{\partial L}{\partial z_5} \cdot W^T = z_2 \cdot x^T \cdot W^T$.

$z_5 = W^TW$

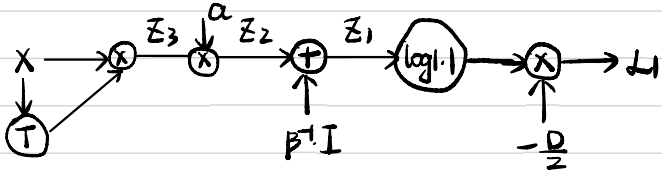
$$\Rightarrow \frac{\partial L}{\partial W} = \left(\frac{\partial L}{\partial W}\right)_1 + \left(\frac{\partial L}{\partial W^T}\right)^T = W \cdot z_2 \cdot x^T + Wx \cdot z_2^T = W\left((W^TWx - x)x^T + x(W^TWx - x)^T\right)$$

$z_2 = W^TWx - x$

2. $\mathcal{L} = -c - \frac{D}{2}\log|\mathbf{K}| - \frac{1}{2}\mathrm{tr}(\mathbf{K}^{-1}\mathbf{Y}\mathbf{Y}^T)\ \ \mathbf{K} = \alpha\mathbf{X}\mathbf{X}^T + \beta^{-1}\mathbf{I}$

$$\mathcal{L}_1 = -\frac{D}{2}\log|\alpha\mathbf{X}\mathbf{X}^T + \beta^{-1}\mathbf{I}|$$
$$\mathcal{L}_2 = -\frac{1}{2}\mathrm{tr}\left((\alpha\mathbf{X}\mathbf{X}^T + \beta^{-1}\mathbf{I})^{-1}\mathbf{Y}\mathbf{Y}^T\right)$$

(1). Draw Computation Graph for $\mathcal{L}_1$.



(2) Compute $\partial\mathcal{L}_1/\partial X$: $\mathcal{L}_1 = -\frac{D}{2}\log|\det z_1|$, $\frac{\partial\mathcal{L}}{\partial z_1} = -\frac{D}{2}(z_1^T)^T = -\frac{D}{2}(z_1^T)^{-1}$.

$\partial\mathcal{L}/\partial z_2 = \partial\mathcal{L}/\partial z_1$ $\partial\mathcal{L}/\partial z_3 = a\cdot\partial\mathcal{L}/\partial z_1 = -\frac{Da}{2}(z_1^T)^{-1}$

$z_3 = XX^T$, $\partial\mathcal{L}/\partial X = \partial z_3/\partial X\cdot\partial\mathcal{L}/\partial z_3 = \partial\mathcal{L}/\partial z_3\cdot(X^T)^T = \partial\mathcal{L}/\partial z_3\cdot X$

$\partial\mathcal{L}/\partial X^T = \partial z_3/\partial X^T\cdot\partial\mathcal{L}/\partial z_3 = X^T\cdot\partial\mathcal{L}/\partial z_3$

$\Rightarrow \partial\mathcal{L}/\partial X = (\partial\mathcal{L}/\partial X) + (\partial\mathcal{L}/\partial X^T)^T = \partial\mathcal{L}/\partial z_3\cdot X + (\partial\mathcal{L}/\partial z_3)^T\cdot X$

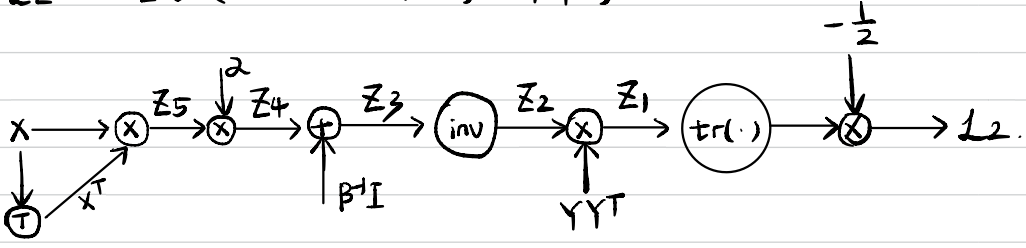
$= -\frac{Da}{2}(z_1^T)^{-1}\cdot X + (-\frac{Da}{2}\cdot(z_1^T)^T)^T\cdot X$

$z_1 = XX^T\cdot a + \beta^{-1}I$ is symmetric, $\partial\mathcal{L}_1/\partial X = -\frac{Da}{2}(z_1)^{-1}\cdot X - \frac{Da}{2}\cdot z_1^{-1}\cdot X = -D\cdot a\, z_1^{-1}\cdot X$

$= -D\cdot a\cdot(XX^T\cdot a + \beta^{-1}I)^{-1}\cdot X$

(3). Draw Computational Graph for L_2.

$$L_2 = -\frac{1}{2}\,tr\left((\alpha XX^T + \beta^{-1}I)^{-1}\cdot YY^T\right).$$



(4) Compute $\partial L_2/\partial X$.

$$L_2 = -\frac{1}{2}\,tr(Z_1).\quad \frac{\partial L_2}{\partial Z_1} = -1\cdot\frac{(Z_1)}{\partial Z_1} = -\frac{1}{2}I$$

$$Z_1 = Z_2(YY^T),\quad \partial L_2/\partial Z_2 = \partial Z_1/\partial Z_2\cdot \partial L_2/\partial Z_1 = -\tfrac{1}{2}I\cdot(YY^T)^T = -\tfrac{1}{2}YY^T$$

$$Z_2 = (Z_3)^{-1},\quad \partial Z_2/\partial Z_3 = -Z_3^{-T}\cdot \partial Z_2/\partial Z_3^{-1}\cdot Z_3^{-T} = -Z_3^{-T}\cdot I\cdot Z_3^{-T} = -(Z_3^{-T})^2$$

$$\frac{\partial L}{\partial Z_3} = \frac{\partial Z_2}{\partial Z_3}\cdot\frac{\partial L}{\partial Z_2} = +(Z_3^{-T})^2\cdot\tfrac{1}{2}YY^T$$

"Add" passes gradient: $\partial L_2/\partial Z_4 = \partial L_2/\partial Z_3$

$$Z_4 = \alpha\cdot Z_5,\quad \partial L_2/\partial Z_5 = \alpha\cdot\partial L_2/\partial Z_4 = \alpha\cdot\partial L_2/\partial Z_3 = \frac{\alpha}{2}\cdot(Z_3^{-T}\cdot Z_3^{-T})\cdot YY^T$$

$$Z_5 = XX^T,\quad \partial L_2/\partial X = \partial Z_5/\partial X\cdot\partial L_2/\partial Z_5 = \partial L_2/\partial Z_5\cdot(X^T)^T = \partial L_2/\partial Z_5\cdot X$$

$$\partial L_2/\partial X^T = \partial Z_5/\partial X^T\cdot\partial L_2/\partial Z_5 = X^T\cdot\partial L_2/\partial Z_5$$

$$\partial L_2/\partial X = (\partial L_2/\partial X)_1 + (\partial L_2/\partial X^T)^T = \partial L_2/\partial Z_5\cdot X + (\partial L_2/\partial Z_5)^T\cdot X$$

$$= \frac{\alpha}{2}\cdot(Z_3^{-T}\cdot Z_3^{-T})\cdot YY^T\cdot X + \left(\frac{\alpha}{2}(Z_3^{-T})^2\cdot YY^T\right)^T\cdot X$$

$$= \frac{\alpha}{2}\left((Z_3^{-T})^2 YY^T + YY^T\cdot(Z_3^{-1})^2\right)X$$

where $Z_3 = \alpha XX^T + \beta^{-1}\cdot I$

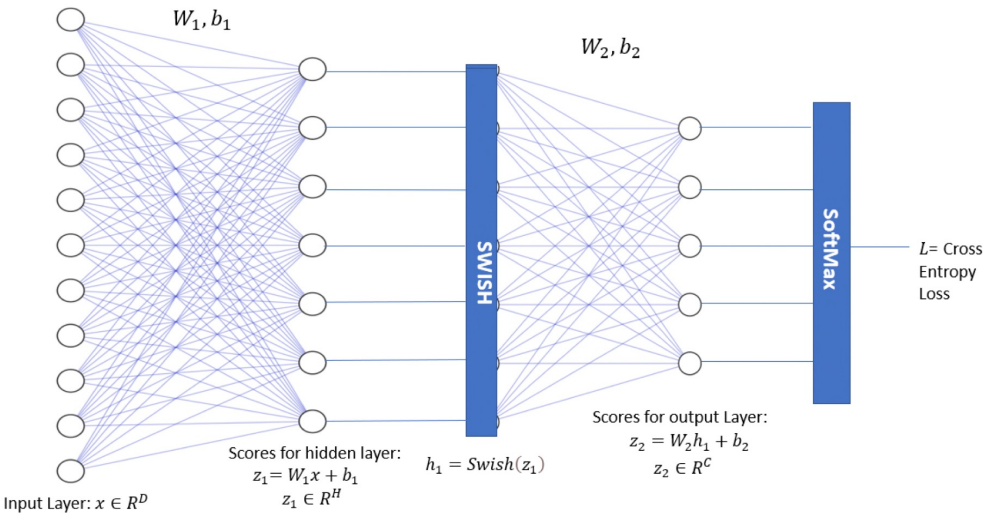
(e). $\frac{\partial L}{\partial X} = \frac{\partial}{\partial X}(-C + L_1 + L_2) = \frac{\partial L_1}{\partial X} + \frac{\partial L_2}{\partial X}$

$$= -D\cdot\alpha\cdot(XX^T\cdot\alpha + \beta^{-1}I)^{-1}X$$

$$+ \frac{\alpha}{2}\left((Z_3^{-T})^2 YY^T + YY^T\cdot(Z_3^{-1})^2\right)X$$

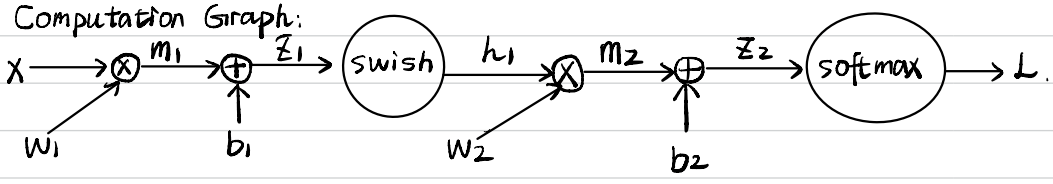
$$Z_3 = \alpha XX^T + \beta^{-1}I.$$

3



Swish(k) = k·σ(k)

(a). Computation Graph:



(b). m_1, m_2 defined as above.

"Add" passes gradient: $\frac{\partial L}{\partial m_2} = \frac{\partial L}{\partial z_2}$, $\frac{\partial L}{\partial b_2} = \frac{\partial L}{\partial z_2}$.

$m_2 = W_2 h_1$, $\frac{\partial L}{\partial W_2} = \frac{\partial m_2}{\partial W_2} \cdot \frac{\partial L}{\partial m_2} = \frac{\partial L}{\partial m_2} \cdot h_1^T = \frac{\partial L}{\partial z_2} \cdot h_1^T$

(c) $\frac{\partial L}{\partial h_1} = \frac{\partial m_2}{\partial h_1} \cdot \frac{\partial L}{\partial m_2} = W_2^T \cdot \frac{\partial L}{\partial z_2}$

$h_1 = swish(z_1)$, $\partial h_1 / \partial z_1 = \sigma(z_1) + z_1(1-\sigma(z_1))$

$= z_1 \cdot \sigma(z_1)$

$\frac{\partial L}{\partial z_1} = \frac{\partial h_1}{\partial z_1} \cdot \frac{\partial L}{\partial h_1} = diag\left(\sigma(z_1) + z_1(1-\sigma(z_1)\right) \cdot \frac{\partial L}{\partial h_1} = diag\left(\sigma(z_1) + z_1(1-\sigma(z_1))\right) \cdot W_2^T \cdot \frac{\partial L}{\partial z_2}$

"Add" passes gradient: $\frac{\partial}{\partial b_1} = \frac{\partial L}{\partial z_1}$, $\frac{\partial L}{\partial m_1} = \frac{\partial L}{\partial z_1}$.

$m_1 = W_1 X$, $\frac{\partial L}{\partial W_1} = \frac{\partial m_1}{\partial W_1} \cdot \frac{\partial L}{\partial m_1} = \frac{\partial L}{\partial m_1} \cdot X^T = \frac{\partial L}{\partial z_1} \cdot X^T$