HW3 Answer

Problem 1

The output is $(B, W', H', output\_channels)$, where

$$W' = \left\lfloor \frac{W + 2 * p_1 - k_1}{s_1} + 1 \right\rfloor$$
$$H' = \left\lfloor \frac{H + 2 * p_2 - k_2}{s_2} + 1 \right\rfloor$$

Note that you should give some explanation to get all points.

Problem 2

We use the gradient descent to update γ and β:

$$\gamma \leftarrow \gamma - \eta \frac{\partial l}{\partial \gamma}$$
$$\beta \leftarrow \beta - \eta \frac{\partial l}{\partial \beta}$$

where η is a learning rate and

$$\frac{\partial l}{\partial \gamma} = \sum_{i=1}^{m} \left(\frac{\partial l}{\partial y_i} \frac{\partial y_i}{\partial \gamma} \right) = \sum_{i=1}^{m} \frac{\partial l}{\partial y_i} \hat{x}_i$$
$$\frac{\partial l}{\partial \beta} = \sum_{i=1}^{m} \left(\frac{\partial l}{\partial y_i} \frac{\partial y_i}{\partial \beta} \right) = \sum_{i=1}^{m} \frac{\partial l}{\partial y_i}$$

Note that we sum from 1 to m because we are working with mini-batches.
Now, we derive some important term by chain rule:

$$
\begin{aligned}
\frac{\partial l}{\partial \hat{x}_i} &= \frac{\partial l}{\partial y_i}\frac{\partial y_i}{\partial \hat{x}_i} = \frac{\partial l}{\partial y_i}\gamma \\
\frac{\partial l}{\partial \sigma_B^2} &= \frac{\partial l}{\partial \hat{x}_i}\frac{\partial \hat{x}_i}{\partial \sigma_B^2} = -\frac{1}{2}\sum_{i=1}^{m}\frac{\partial l}{\partial \hat{x}_i}(x_i-\mu_B)(\sigma_B^2+\epsilon)^{-\frac{3}{2}}
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial l}{\partial \mu_B} &= \frac{\partial l}{\partial \hat{x}_i}\frac{\partial \hat{x}_i}{\partial \mu_B} \\
&= \sum_{i=1}^{m}\frac{\partial l}{\partial \hat{x}_i}\frac{\partial}{\partial \mu_B}(x_i-\mu_B)(\sigma_B^2+\epsilon)^{-\frac{1}{2}} \\
&= \sum_{i=1}^{m}\frac{\partial l}{\partial \hat{x}_i}\left[\frac{\partial(x_i-\mu_B)}{\partial \mu_B}(\sigma_B^2+\epsilon)^{-\frac{1}{2}} + (x_i-\mu_B)\frac{\partial(\sigma_B^2+\epsilon)^{-\frac{1}{2}}}{\partial \mu_B}\right] \\
&= \sum_{i=1}^{m}\frac{\partial l}{\partial \hat{x}_i}\left[\frac{-1}{\sqrt{\sigma_B^2+\epsilon}} + (x_i-\mu_B)\frac{\partial(\sigma_B^2+\epsilon)^{-\frac{1}{2}}}{\partial \mu_B}\right]
\end{aligned}
$$

We know $\sigma_B^2 = \frac{1}{m}\sum_{i=1}^{m}(x_i-\mu_B)^2$, so the second term in bracket is

$$
\begin{aligned}
(x_i-\mu_B)\frac{\partial(\sigma_B^2+\epsilon)^{-\frac{1}{2}}}{\partial \mu_B} &= (x_i-\mu_B)\frac{-1}{2}(\sigma_B^2+\epsilon)^{-\frac{3}{2}}\frac{\partial(\sigma_B^2+\epsilon)}{\partial \mu_B} \\
&= \frac{-1}{2}(x_i-\mu_B)(\sigma_B^2+\epsilon)^{-\frac{3}{2}}\frac{\partial\left(\frac{1}{m}\sum_{i=1}^{m}(x_i-\mu_B)^2+\epsilon\right)}{\partial \mu_B} \\
&= \frac{-1}{2}(x_i-\mu_B)(\sigma_B^2+\epsilon)^{-\frac{3}{2}}\left(\frac{-2}{m}\sum_{i=1}^{m}(x_i-\mu_B)\right)
\end{aligned}
$$

Hence,

$$
\begin{aligned}
\frac{\partial l}{\partial \mu_B} &= \sum_{i=1}^{m}\frac{\partial l}{\partial \hat{x}_i}\frac{-1}{\sqrt{\sigma_B^2+\epsilon}} + \underbrace{\sum_{i=1}^{m}\frac{\partial l}{\partial \hat{x}_i}\frac{-1}{2}(x_i-\mu_B)(\sigma_B^2+\epsilon)^{-\frac{3}{2}}}_{\frac{\partial l}{\partial \sigma_B^2}}\left(\frac{-2}{m}\sum_{i=1}^{m}(x_i-\mu_B)\right) \\
&= \sum_{i=1}^{m}\frac{\partial l}{\partial \hat{x}_i}\frac{-1}{\sqrt{\sigma_B^2+\epsilon}} + \frac{\partial l}{\partial \sigma_B^2}\left(\frac{-2}{m}\sum_{i=1}^{m}(x_i-\mu_B)\right)
\end{aligned}
$$

To derive $\frac{\partial l}{\partial x_i}$, we use the chain rule $\frac{\partial l}{\partial x_i} = \frac{\partial l}{\partial \hat{x}_i}\frac{\partial \hat{x}_i}{\partial x_i} + \frac{\partial l}{\partial \sigma_B^2}\frac{\partial \sigma_B^2}{\partial x_i} + \frac{\partial l}{\partial \mu_B}\frac{\partial \mu_B}{\partial x_i}$. Now, calculate the remain term:

$$
\begin{aligned}
\frac{\partial \hat{x}_i}{\partial x_i} &= \frac{1}{\sqrt{\sigma_B^2 + \epsilon}} \\
\frac{\partial \mu_B}{\partial x_i} &= \frac{1}{m} \\
\frac{\partial \sigma_B^2}{\partial x_i} &= \frac{2(x_i - \mu)}{m}
\end{aligned}
$$

That is,

$$
\frac{\partial l}{\partial x_i} = \frac{\partial l}{\partial \hat{x}_i}\frac{1}{\sqrt{\sigma_B^2 + \epsilon}} + \frac{\partial l}{\partial \sigma_B^2}\frac{2(x_i - \mu)}{m} + \frac{1}{m}\frac{\partial l}{\partial \mu_B}
$$

Problem 3

Note that we use Kronecker delta in our answer
(1)

$$
\begin{aligned}
\frac{\partial S_i}{\partial z_j} = \frac{\partial \frac{e^{z_i}}{\sum_{k=1}^{N} e^{z_k}}}{\partial z_j} &= \frac{\partial e^{z_i}}{\partial z_j}(\sum_{k=1}^{N} e^{z_k})^{-1} + e^{z_i}\frac{\partial}{\partial z_j}(\sum_{k=1}^{N} e^{z_k})^{-1} \\
&= e^{z_i}\delta_{ij}(\sum_{k=1}^{N} e^{z_k})^{-1} - e^{z_i}(\sum_{k=1}^{N} e^{z_k})^{-2} e^{z_j} \\
&= S_i\delta_{ij} - S_i S_j
\end{aligned}
$$

(2)

$$
\begin{aligned}
\frac{\partial L}{\partial z_i} &= -\sum_j y_j \frac{\partial}{\partial z_i} \log \hat{y}_j \\
&= -\sum_j y_j \frac{1}{\hat{y}_j} \frac{\partial \hat{y}_j}{\partial z_i} \\
&= -\sum_{j=i} \frac{y_j}{\hat{y}_j} \frac{\partial \hat{y}_j}{\partial z_i} - \sum_{j \neq i} \frac{y_j}{\hat{y}_j} \frac{\partial \hat{y}_j}{\partial z_i} \\
&= -\frac{y_i}{\hat{y}_i}(\hat{y}_i - \hat{y}_i \hat{y}_i) - \sum_{j \neq i} \frac{y_j}{\hat{y}_j}(-\hat{y}_i \hat{y}_j) \\
&= -y_i + y_i \hat{y}_i + \sum_{j \neq i} y_j \hat{y}_i \\
&= -y_i + \hat{y}_i \sum_j y_j = \hat{y}_i - y_i
\end{aligned}
$$

Problem 4

(1) $WLOG$(Without Loss of Generality), let $\mu = 0$. Since Σ is symmetric positive semi-definite matrix, we can perform eigen decomposition as follows:

$$\Sigma = UDU^T = \sum_{i=1}^{m} (d_i u_i u_i^T).$$

where U and U^T are orthogonal matrix.
Let $n = 2m$ and construct a set of points $x_1, \dots, x_m, \dots, x_{2m}$
where $x_i = \sqrt{d_i} u_i$ and $x_{m+i} = -\sqrt{d_i} u_i \; \forall \, 1 \le i \le m$. Then,

$$
\begin{aligned}
\frac{1}{n} \sum_{i=1}^{n} x_i = \mu = 0 \\
\frac{1}{n} \sum_{i=1}^{n} x_i x_i^T = \sum_{i=1}^{m} (d_i u_i u_i^T) = UDU^T = \Sigma
\end{aligned}
$$

Note that lots of students just use the covariance of eigen decomposition to construct $\{x_i\}_{i=1}^n$. However, It should satisfy the condition that $\frac{1}{n} \sum_{i=1}^n x_i = \mu$
(2)
Let $\phi_1, \cdots, \phi_k$ be the columns of Φ. Then

$$Trace(\Phi^T\Sigma\Phi) = \sum_{i=1}^{k}\phi_i^T\Sigma\phi_i = \sum_{i=1}^{k}\phi_i^T(\sum_{j=1}^{m}(d_ju_ju_j^T))\phi_i = \sum_{j=1}^{m}d_j\sum_{i=1}^{k}\langle u_j,\phi_i\rangle^2 = \sum_{j=1}^{m}c_jd_j$$

where $\langle\cdot,\cdot\rangle$ is standard inner product in Euclidean space, $c_j := \sum_{i=1}^{k}\langle u_j,\phi_i\rangle^2$ for each $j = 1,\cdots,m$ and $d_1 \geq d_2 \geq \cdots \geq d_m \geq 0$
Claim: $0 \leq c_j \leq 1$ and $\sum_{j=1}^{m}c_j = k$
Clearly, $c_j \geq 0$. Extending $\phi_1,\cdots,\phi_k$ to $\phi_1,\cdots,\phi_k,\phi_{k+1},\cdots,\phi_m$ for $\mathbb{R}^m$. Then, for each $j = 1,\cdots,m$

$$c_j = \sum_{i=1}^{k}\langle u_j,\phi_i\rangle^2 \leq \sum_{i=1}^{m}\langle u_j,\phi_i\rangle^2 = 1$$

Finally, since $u_1,\cdots,u_m$ is an orthonormal basis for $\mathbb{R}^m$,

$$\sum_{j=1}^{m}c_j = \sum_{j=1}^{m}\sum_{i=1}^{k}\langle u_j,\phi_i\rangle^2 = \sum_{i=1}^{k}\sum_{j=1}^{m}\langle u_j,\phi_i\rangle^2 = \sum_{i=1}^{k}\|\phi_i\|_2^2 = k$$

Hence, the minimum value of $\sum_{j=1}^{m}c_jd_j$ over all choice of $c_1,c_2,\cdots,c_m \in [0,1]$ with $\sum_{j=1}^{k}c_j = k$ is $d_{m-k+1},\cdots,d_m$. This is achieved when $c_1,\cdots,c_{m-k} = 0$ and $c_{m-k+1} = \cdots = c_m = 1$
在證明過程中，所有符號都要有定義，盡量不要直接使用上課的符號 ex. $\hat{x}^{\mathrm{PCA}}$