Homework Set 3

Introduction to Artificial Intelligence with Mathematics (MAS473)

Total Points = 50pts

1. (5pts) Consider the EM algorithm of a Gaussian mixture model

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\mu_k, \mathbf{\Sigma}_k)$$

in our lecture. Assume that $\Sigma_k = \epsilon I$ for all $k = 1, \cdots, K$. Letting $\epsilon \to 0$, prove that the limiting case is equivalent to the K-means clustering.

Solution. From the assumption, we can calculate in the same way as the general GMM model and we can see that the EM algorithm of the given model is

- (E-step) Using the current parameters $\theta = \{\mu_1, \cdots, \mu_K, \pi_1, \cdots, \pi_K\}$, compute

$$p(z_{ik} = 1|\mathbf{x}_i) = \frac{\pi_k \mathcal{N}(\mathbf{x}_i|\mu_k, \epsilon I)}{\sum_{l=1}^{K} \pi_l \mathcal{N}(\mathbf{x}_i|\mu_l, \epsilon I)}.$$

- (M-step) Update all parameters θ

$$\mu_k = \sum_{i=1}^{N} \frac{p(z_{ik} = 1|\mathbf{x}_i)}{\sum_{j=1}^{N} p(z_{jk} = 1|\mathbf{x}_j)} \mathbf{x}_i,$$
$$\pi_k = \frac{1}{N} \sum_{i=1}^{N} p(z_{ik} = 1|\mathbf{x}_i).$$

- Repeat E-step and M-step until convergence.

Letting $\epsilon \to 0$,

$$p(z_{ik} = 1|\mathbf{x}_i) = \frac{\pi_k \mathcal{N}(\mathbf{x}_i|\mu_k, \epsilon I)}{\sum_{l=1}^{K} \pi_l \mathcal{N}(\mathbf{x}_i|\mu_l, \epsilon I)} = \frac{\pi_k \cdot \exp(-\|\mathbf{x}_i - \mu_k\|^2/2\epsilon)}{\sum_{l=1}^{K} \pi_l \cdot \exp(-\|\mathbf{x}_i - \mu_l\|^2/2\epsilon)}$$

converges to $\frac{1}{|\mathcal{I}_i|}\mathbf{1}_{\{k \in \mathcal{I}_i\}}$ where $\mathcal{I}_i = \mathrm{argmin}_j \|\mathbf{x}_i - \mu_j\|^2$. By the further assumption :

> At any time, $\mu_1, \cdots, \mu_K$ satisfy that $\mathrm{argmin}_{j \in \{1, \cdots, K\}} \|\mathbf{x}_i - \mu_j\|^2$ has only 1 element for all $i = 1, \cdots, N$

E-step is equivalent to E-step of K-means clustering algorithm if we set $r_{ik} = \frac{1}{|\mathcal{I}_i|}\mathbf{1}_{\{k \in \mathcal{I}_i\}}$. Then the update rule for $\mu_1, \cdots, \mu_K$ in M-step is same as M-step of K-means clustering. $\pi_1, \cdots, \pi_K$ does not affect $p(z_{ik} = 1|\mathbf{x}_i)$ and μ_k as long as $\pi_1, \cdots, \pi_K > 0$.

2. (15pts) **(Latent Class Analysis)** Consider a D-dimensional random vector $\mathbf{x}_0 = [x_1, \cdots, x_D]$ where $x_i \sim \text{Bernoulli}(\mu_i)(i = 1, \cdots, D)$ are independent. Now we consider a mixture model of distributions of random vectors like $\mathbf{x}_0$. In other words, consider

$$p(\mathbf{x}|\mu, \pi) = \sum_{k=1}^{K} \pi_k p(\mathbf{x}|\mu_k)$$

where $\mu = [\mu_1, \cdots, \mu_K]$, $\mu_k = [\mu_{k1}, \cdots, \mu_{kD}] \in [0,1]^D$ for any $k = 1, \cdots, K$, $\pi = [\pi_1, \cdots, \pi_K] \in \triangle^{K-1} = \{[y_1, \cdots, y_K] \in \mathbb{R}^K : y_i \geq 0 \text{ for any } i = 1, \cdots, K \text{ and } \sum_{i=1}^{K} y_i = 1\}$ and

$$p(\mathbf{x}|\mu_k) = \prod_{i=1}^{D} \mu_{ki}^{x_i}(1-\mu_{ki})^{1-x_i}.$$

Assume $\mathcal{D} = \{\mathbf{x}_i : 1 \leq i \leq N\}$ is a given dataset. Using the similar argument as Gaussian mixture models, prove that the EM algorithm for this model is

- Initialization
- (E-step) Using the current parameters, compute

$$p(z_{ik} = 1|\mathbf{x}_i) = \frac{\pi_k p(\mathbf{x}_i|\mu_k)}{\sum_{l=1}^{K} \pi_l p(\mathbf{x}_i|\mu_l)}$$

- (M-step) Update all parameters

$$\mu_k = \frac{\sum_{i=1}^{N} p(z_{ik} = 1|\mathbf{x}_i)\mathbf{x}_i}{\sum_{i=1}^{N} p(z_{ik} = 1|\mathbf{x}_i)}$$
$$\pi_k = \frac{1}{N}\sum_{i=1}^{N} p(z_{ik} = 1|\mathbf{x}_i)$$

- Repeat E-step and M-step until convergence.

where $\mathbf{z}_i = [\mathbf{z}_{i1}, \cdots, \mathbf{z}_{iK}]$ is an one-hot vector which indicates that a group containing $\mathbf{x}_i$.

Solution. Note that

$$p(\mathbf{x}_i|\mathbf{z}_i = 1) = \prod_{j=1}^{D} \mu_{kj}^{\mathbf{x}_{ij}}(1-\mu_{kj})^{1-\mathbf{x}_{ij}},$$
$$p(\mathbf{x}_i|\mathbf{z}_i = 1) = \prod_{k=1}^{K}\left(\prod_{j=1}^{D} \mu_{kj}^{\mathbf{x}_{ij}}(1-\mu_{kj})^{1-\mathbf{x}_{ij}}\right)^{z_{ik}}.$$

Then

$$p(\mathbf{x}_i, \mathbf{z}_i) = p(\mathbf{x}_i|\mathbf{z}_i)p(\mathbf{z}_i) = \prod_{k=1}^{K}\left(\pi_k \prod_{j=1}^{D} \mu_{kj}^{\mathbf{x}_{ij}}(1-\mu_{kj})^{1-\mathbf{x}_{ij}}\right)^{z_{ik}}.$$

Moreover,

$$p(z_{ik} = 1|\mathbf{x}_i) = \frac{p(z_{ik} = 1, \mathbf{x}_i)}{p(\mathbf{x}_i)} = \frac{p(\mathbf{x}_i|z_{ik} = 1)p(z_{ik} = 1)}{\sum_{l=1}^{K} p(\mathbf{x}_i|z_{il} = 1)p(z_{il} = 1)} = \frac{\pi_k p(\mathbf{x}_i|\mu_k)}{\sum_{l=1}^{K} \pi_l p(\mathbf{x}_i|\mu_l)}.$$

From the log likelihood function of $\mathbf{x}$

$$J(\mu,\pi)=\sum_{i=1}^{N}\log p(\mathbf{x}_i|\mu,\pi)=\sum_{i=1}^{N}\log\left(\sum_{k=1}^{K}\pi_k\prod_{j=1}^{D}\mu_{kj}^{\mathbf{x}_{ij}}(1-\mu_{kj})^{1-\mathbf{x}_{ij}}\right),$$

we have

$$\begin{aligned}\nabla_{\mu_l}J(\mu,\pi)&=\sum_{i=1}^{N}\frac{\pi_l\prod_{j=1}^{D}\mu_{lj}^{\mathbf{x}_{ij}}(1-\mu_{lj})^{1-\mathbf{x}_{ij}}}{\sum_{k=1}^{K}\pi_k\prod_{j=1}^{D}\mu_{kj}^{\mathbf{x}_{ij}}(1-\mu_{kj})^{1-\mathbf{x}_{ij}}}\cdot((1-\mathbf{x}_i)\oslash\mu_l-(1-\mathbf{x}_i)\oslash(1-\mu_l))\\&=\sum_{i=1}^{N}\frac{\pi_l\prod_{j=1}^{D}\mu_{lj}^{\mathbf{x}_{ij}}(1-\mu_{lj})^{1-\mathbf{x}_{ij}}}{\sum_{k=1}^{K}\pi_k\prod_{j=1}^{D}\mu_{kj}^{\mathbf{x}_{ij}}(1-\mu_{kj})^{1-\mathbf{x}_{ij}}}\cdot((\mathbf{x}_i-\mu_l)\oslash(\mu_l(1-\mu_l)))\end{aligned}$$

From $\nabla_{\mu_l}J(\mu,\pi)=0$, we obtain

$$(\mu_l(1-\mu_l))\odot\nabla_{\mu_l}J(\mu,\pi)=\sum_{i=1}^{N}\frac{\pi_l\prod_{j=1}^{D}\mu_{lj}^{\mathbf{x}_{ij}}(1-\mu_{lj})^{1-\mathbf{x}_{ij}}}{\sum_{k=1}^{K}\pi_k\prod_{j=1}^{D}\mu_{kj}^{\mathbf{x}_{ij}}(1-\mu_{kj})^{1-\mathbf{x}_{ij}}}\cdot(\mathbf{x}_i-\mu_l)=0.$$

It then follows that

$$\mu_l=\frac{\sum_{i=1}^{N}p(z_{ik}=1|\mathbf{x}_i)\mathbf{x}_i}{\sum_{i=1}^{N}p(z_{ik}=1|\mathbf{x}_i)}.$$

To maximizing the log likelihood function with a constraint $\sum_{k=1}^{K}\pi_k=1$ w.r.t. $\{\pi_k:k=1,\cdots,K\}$, we introduce a Lagrange multiplier λ :

$$\mathcal{L}(\mu,\pi)=J(\mu,\pi)+\lambda\cdot\left(\sum_{k=1}^{K}\pi_k-1\right).$$

From $\frac{\partial}{\partial\pi}\mathcal{L}(\mu,\pi)=0$, i.e.

$$\sum_{i=1}^{N}\frac{\prod_{j=1}^{D}\mu_{lj}^{\mathbf{x}_{ij}}(1-\mu_{lj})^{1-\mathbf{x}_{ij}}}{\sum_{k=1}^{K}\pi_k\prod_{j=1}^{D}\mu_{kj}^{\mathbf{x}_{ij}}(1-\mu_{kj})^{1-\mathbf{x}_{ij}}}+\lambda=0,\ 1\le k\le K,$$

we get

$$\frac{1}{\pi_k}\sum_{i=1}^{N}\frac{\pi_k\prod_{j=1}^{D}\mu_{lj}^{\mathbf{x}_{ij}}(1-\mu_{lj})^{1-\mathbf{x}_{ij}}}{\sum_{k=1}^{K}\pi_k\prod_{j=1}^{D}\mu_{kj}^{\mathbf{x}_{ij}}(1-\mu_{kj})^{1-\mathbf{x}_{ij}}}+\lambda=0,\ 1\le k\le K.$$

Thus,

$$\pi_k=-\frac{1}{\lambda}\sum_{i=1}^{N}p(z_{ik}=1|\mathbf{x}_i).$$

From the constraint,

$$\sum_{k=1}^{K}=-\sum_{k=1}^{K}\frac{1}{\lambda}\sum_{i=1}^{N}p(z_{ik}=1|\mathbf{x}_i)=-\sum_{i=1}^{N}\sum_{k=1}^{K}\frac{1}{\lambda}p(z_{ik}=1|\mathbf{x}_i)=-N.$$

Therefore, we get

$$\pi_k=\frac{1}{N}\sum_{i=1}^{N}p(z_{ik}=1|\mathbf{x}_i).$$

3. (10pts) Prove that

$$\mathcal{N}(\mathbf{x}|\mathbf{a},A)\mathcal{N}(\mathbf{x}|\mathbf{b},B)=Z^{-1}\mathcal{N}(\mathbf{x}|\mathbf{c},C)$$

where

$\mathbf{c}=C(A^{-1}\mathbf{a}+B^{-1}\mathbf{b}), C=(A^{-1}+B^{-1})^{-1},$ and $Z^{-1}=(2\pi)^{-D/2}|A+B|^{-1/2}\exp\left(-\frac{1}{2}(\mathbf{a}-\mathbf{b})^\top(A+B)^{-1}(\mathbf{a}-\mathbf{b})\right).$

Also, prove that

$$(Z+UWV^\top)^{-1}=Z^{-1}-Z^{-1}U(W^{-1}+V^\top Z^{-1}U)^{-1}V^\top Z^{-1}$$

where $Z\in\mathbb{R}^{n\times n}, W\in\mathbb{R}^{m\times m}$, $U,V\in\mathbb{R}^{n\times m}$ with assumptions that the relevant inverses all exist.

Solution. First, we prove the second formula and then prove the first formula using the second one. The second formula is obtained by a directly calculation

$$\begin{aligned}
&(Z^{-1}-Z^{-1}U(W^{-1}+V^\top Z^{-1}U)^{-1}V^\top Z^{-1})(Z+UWV^\top)\\
&=I+Z^{-1}UWV^\top-Z^{-1}U(W^{-1}+V^\top Z^{-1}U)^{-1}V^\top Z^{-1}(Z+UWV^\top)\\
&=I+Z^{-1}UWV^\top-Z^{-1}U(W^{-1}+V^\top Z^{-1}U)^{-1}(V^\top+V^\top Z^{-1}UWV^\top)\\
&=I+Z^{-1}UWV^\top-Z^{-1}U(W^{-1}+V^\top Z^{-1}U)^{-1}(W^{-1}WV^\top+V^\top Z^{-1}UWV^\top)\\
&=I+Z^{-1}UWV^\top-Z^{-1}U(W^{-1}+V^\top Z^{-1}U)^{-1}(W^{-1}+V^\top Z^{-1}U)WV^\top\\
&=I+Z^{-1}UWV^\top-Z^{-1}UWV^\top=I.
\end{aligned}$$

Next,

$$\begin{aligned}
&\mathcal{N}(\mathbf{x}|\mathbf{a},A)\mathcal{N}(\mathbf{x}|\mathbf{b},B)\\
&=\frac{1}{(2\pi)^D}\cdot\frac{1}{|A|^{1/2}}\cdot\frac{1}{|B|^{1/2}}\exp\left(-\frac{1}{2}(\mathbf{x}-\mathbf{a})^\top A^{-1}(\mathbf{x}-\mathbf{a})-\frac{1}{2}(\mathbf{x}-\mathbf{b})^\top B^{-1}(\mathbf{x}-\mathbf{b})\right)\\
&=\frac{1}{(2\pi)^D}\cdot\frac{1}{|A|^{1/2}}\cdot\frac{1}{|B|^{1/2}}\exp\left(-\frac{1}{2}\mathbf{x}^\top(A^{-1}+B^{-1})\mathbf{x}+(A^{-1}\mathbf{a}+B^{-1}\mathbf{b})^\top\mathbf{x}-\frac{1}{2}\mathbf{a}^\top A^{-1}\mathbf{a}-\frac{1}{2}\mathbf{b}^\top B^{-1}\mathbf{b}\right)\\
&=\frac{1}{(2\pi)^D}\cdot\frac{1}{|A|^{1/2}}\cdot\frac{1}{|B|^{1/2}}\exp\left(-\frac{1}{2}(\mathbf{x}-(A^{-1}+B^{-1})^{-1}(A^{-1}\mathbf{a}+B^{-1}\mathbf{b}))^\top(A^{-1}+B^{-1})\cdot\right.\\
&\left.(\mathbf{x}-(A^{-1}+B^{-1})^{-1}(A^{-1}\mathbf{a}+B^{-1}\mathbf{b}))-\frac{1}{2}\mathbf{a}^\top A^{-1}\mathbf{a}-\frac{1}{2}\mathbf{b}^\top B^{-1}\mathbf{b}+\frac{1}{2}(A^{-1}\mathbf{a}+B^{-1}\mathbf{b})^\top(A^{-1}+B^{-1})^{-1}(A^{-1}\mathbf{a}+B^{-1}\mathbf{b})\right)\\
&=\frac{1}{(2\pi)^D}\cdot\frac{1}{|A|^{1/2}}\cdot\frac{1}{|B|^{1/2}}\exp\left(-\frac{1}{2}(\mathbf{x}-\mathbf{c})^\top C^{-1}(\mathbf{x}-\mathbf{c})-\frac{1}{2}\mathbf{a}^\top A^{-1}\mathbf{a}-\frac{1}{2}\mathbf{b}^\top B^{-1}\mathbf{b}\right.\\
&\qquad\left.+\frac{1}{2}(A^{-1}\mathbf{a}+B^{-1}\mathbf{b})^\top(A^{-1}+B^{-1})^{-1}(A^{-1}\mathbf{a}+B^{-1}\mathbf{b})\right)\\
&=\mathcal{N}(\mathbf{x}|\mathbf{c},C)\cdot\frac{1}{(2\pi)^{D/2}}\cdot\frac{|C|^{1/2}}{|A|^{1/2}|B|^{1/2}}\exp\left(\frac{1}{2}(A^{-1}\mathbf{a}+B^{-1}\mathbf{b})^\top(A^{-1}+B^{-1})^{-1}(A^{-1}\mathbf{a}+B^{-1}\mathbf{b})-\frac{1}{2}\mathbf{a}^\top A^{-1}\mathbf{a}-\frac{1}{2}\mathbf{b}^\top B^{-1}\mathbf{b}\right).
\end{aligned}$$

By the second formula,

$$(A^{-1}+B^{-1})^{-1}=A-A(A+B)^{-1}A=B-B(A+B)^{-1}B.$$

Thus

$$
\begin{aligned}
&\frac{1}{2}(A^{-1}\mathbf{a}+B^{-1}\mathbf{b})^\top(A^{-1}+B^{-1})^{-1}(A^{-1}\mathbf{a}+B^{-1}\mathbf{b})-\frac{1}{2}\mathbf{a}^\top A^{-1}\mathbf{a}-\frac{1}{2}\mathbf{b}^\top B^{-1}\mathbf{b}\\
&=\frac{1}{2}\mathbf{a}^\top A^{-1}(A-A(A+B)^{-1}A)A^{-1}\mathbf{a}+\frac{1}{2}\mathbf{b}^\top B^{-1}(B-B(A+B)^{-1}B)B^{-1}\mathbf{b}\\
&\quad+\mathbf{a}^\top A^{-1}(A^{-1}+B^{-1})^{-1}B^{-1}\mathbf{b}-\frac{1}{2}\mathbf{a}^\top A^{-1}\mathbf{a}-\frac{1}{2}\mathbf{b}^\top B^{-1}\mathbf{b}\\
&=-\frac{1}{2}\mathbf{a}^\top(A+B)^{-1}\mathbf{a}-\frac{1}{2}\mathbf{b}^\top(A+B)^{-1}\mathbf{b}+\mathbf{a}^\top(A+B)^{-1}\mathbf{b}\\
&=-\frac{1}{2}(\mathbf{a}-\mathbf{b})^\top(A+B)^{-1}(\mathbf{a}-\mathbf{b}).
\end{aligned}
$$

Also,

$$
\frac{|C|^{1/2}}{|A|^{1/2}|B|^{1/2}}=\frac{|A^{-1}+B^{-1}|^{-1/2}}{|A|^{1/2}|B|^{1/2}}=\frac{1}{|A(A^{-1}+B^{-1})B|^{1/2}}=\frac{1}{|A+B|^{1/2}}.
$$

We are done.

4. (10pts) **(Probabilistic PCA)** Let $\mathcal{D} = \{\mathbf{x}_i \in \mathbb{R}^D : 1 \leq i \leq N\}$, $\overline{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^N \mathbf{x}_i$ and

$$S = \frac{1}{N}\sum_{i=1}^N (\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^\top.$$

Introduce a latent random variable $\mathbf{z}$. Suppose

$$\mathbf{z} \sim \mathcal{N}(\mathbf{z}|0, I)$$

and

$$p(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}|W\mathbf{z} + \mu, \sigma^2 I)$$

where $W \in \mathbb{R}^{D\times M}$ and $\mu \in \mathbb{R}^D$.

(a) Find $p(\mathbf{x})$ and $p(\mathbf{z}|\mathbf{x})$.

(b) Prove that the maximum likelihood estimator of μ is

$$\mu_{ML} = \overline{\mathbf{x}}.$$

(c) Prove that the maximum likelihood estimator of W satisfies

$$S(WW^\top + \sigma^2 I)^{-1}W = W.$$

In fact, we can obtain the closed form of W_{ML} and σ^2_{ML}. If you are interested in this result, read "Probabilistic principal component analysis" written by M. E. Tipping and C. M. Bishop (1999).

Solution.

(a) First,

$$\begin{aligned}
p(\mathbf{x}) &= \int p(\mathbf{x}|\mathbf{z})p(\mathbf{z})d\mathbf{z} \\
&\propto \int \exp\left(-\frac{1}{2\sigma^2}(\mathbf{x} - W\mathbf{z} - \mu)^\top(\mathbf{x} - W\mathbf{z} - \mu) - \frac{1}{2}\mathbf{z}^\top\mathbf{z}\right)d\mathbf{z} \\
&= \int \exp\left(-\frac{1}{2}\mathbf{z}^\top\left(\frac{1}{\sigma^2}W^\top W + I\right)\mathbf{z} - \frac{1}{\sigma^2}(\mu - \mathbf{x})^\top W\mathbf{z} - \frac{1}{2\sigma^2}(\mu - \mathbf{x})^\top(\mu - \mathbf{x})\right)d\mathbf{z} \\
&= \int \exp\Bigg(-\frac{1}{2}\left(\mathbf{z} + \frac{1}{\sigma^2}\left(\frac{1}{\sigma^2}W^\top W + I\right)^{-1}W^\top(\mu - \mathbf{x})\right)^\top\left(\frac{1}{\sigma^2}W^\top W + I\right)\left(\mathbf{z} + \frac{1}{\sigma^2}\left(\frac{1}{\sigma^2}W^\top W + I\right)^{-1}W^\top(\mu - \mathbf{x})\right) + \\
&\qquad \frac{1}{2\sigma^4}(\mathbf{x} - \mu)^\top W\left(\frac{1}{\sigma^2}W^\top W + I\right)^{-1}W^\top(\mathbf{x} - \mu) - \frac{1}{2\sigma^2}(\mathbf{x} - \mu)^\top(\mathbf{x} - \mu)\Bigg)d\mathbf{z}.
\end{aligned}$$

Since $\frac{1}{\sigma^2}W^\top W + I$ is positive definite and is independent of $\mathbf{x}$,

$$\exp\left(-\frac{1}{2}\left(\mathbf{z} + \frac{1}{\sigma^2}\left(\frac{1}{\sigma^2}W^\top W + I\right)^{-1}W^\top(\mu - \mathbf{x})\right)^\top\left(\frac{1}{\sigma^2}W^\top W + I\right)\left(\mathbf{z} + \frac{1}{\sigma^2}\left(\frac{1}{\sigma^2}W^\top W + I\right)^{-1}W^\top(\mu - \mathbf{x})\right)\right)$$

is an unnormalized normal pdf with respect to $\mathbf{z}$ and the integration is also independent of $\mathbf{x}$ (Note that the integration is invariant to a translation). Thus,

$$\begin{aligned}
p(\mathbf{x}) &\propto \exp\left(\frac{1}{2\sigma^4}(\mathbf{x} - \mu)^\top W\left(\frac{1}{\sigma^2}W^\top W + I\right)^{-1}W^\top(\mathbf{x} - \mu) - \frac{1}{2\sigma^2}(\mathbf{x} - \mu)^\top(\mathbf{x} - \mu)\right) \\
&= \exp\left(\frac{1}{2\sigma^2}(\mathbf{x} - \mu)^\top\left(I - \frac{1}{\sigma^2}W\left(\frac{1}{\sigma^2}W^\top W + I\right)^{-1}W^\top\right)(\mathbf{x} - \mu)\right).
\end{aligned}$$

Therefore

$$p(\mathbf{x}) = \mathcal{N}\left(\mathbf{x}\,\middle|\,\mu, \sigma^2\left(I - \frac{1}{\sigma^2}W\left(\frac{1}{\sigma^2}W^\top W + I\right)^{-1}W^\top\right)^{-1}\right).$$

Note that

$$\sigma^2\left(I - \frac{1}{\sigma^2}W\left(\frac{1}{\sigma^2}W^\top W + I\right)^{-1}W^\top\right)^{-1} = \sigma^2\left(I + \frac{1}{\sigma^2}WW^\top\right) = \sigma^2 I + WW^\top$$

by the second formula in Problem 3. Next,

$$\begin{aligned}
p(\mathbf{z}|\mathbf{x}) &\propto p(\mathbf{x}|\mathbf{z})p(\mathbf{z}) \\
&\propto \exp\left(-\frac{1}{2\sigma^2}(\mathbf{x} - W\mathbf{z} - \mu)^\top(\mathbf{x} - W\mathbf{z} - \mu) - \frac{1}{2}\mathbf{z}^\top\mathbf{z}\right) \\
&\propto \exp\left(-\frac{1}{2}\left(\mathbf{z} + \frac{1}{\sigma^2}\left(\frac{1}{\sigma^2}W^\top W + I\right)^{-1}W^\top(\mu - \mathbf{x})\right)^\top\left(\frac{1}{\sigma^2}W^\top W + I\right)\left(\mathbf{z} + \frac{1}{\sigma^2}\left(\frac{1}{\sigma^2}W^\top W + I\right)^{-1}W^\top(\mu - \mathbf{x})\right)\right).
\end{aligned}$$

The detail calculation is already done above. Therefore,

$$p(\mathbf{z}|\mathbf{x}) = \mathcal{N}\left(\mathbf{z}\,\middle|\,\frac{1}{\sigma^2}\left(\frac{1}{\sigma^2}W^\top W + I\right)^{-1}W^\top(\mathbf{x} - \mu), \left(\frac{1}{\sigma^2}W^\top W + I\right)^{-1}\right).$$

(b) From (a), $p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\mu, \sigma^2 I + WW^\top)$. To maximize

$$\mathcal{L}(\mu, W) = \sum_{i=1}^N \log p(\mathbf{x}_i) = -\frac{ND}{2}\log(2\pi) - \frac{N}{2}\log|\sigma^2 I + WW^\top| - \frac{1}{2}\sum_{i=1}^N(\mathbf{x}_i - \mu)^\top(\sigma^2 I + WW^\top)^{-1}(\mathbf{x}_i - \mu),$$

differentiate w.r.t. μ :

$$\frac{\partial}{\partial\mu}\mathcal{L}(\mu, W) = -(\sigma^2 I + WW^\top)^{-1}\sum_{i=1}^N(\mathbf{x}_i - \mu).$$

Since $\sigma^2 I + WW^\top$ is positive definite,

$$\mu_{ML} = \frac{1}{N}\sum_{i=1}^N \mathbf{x}_i = \overline{\mathbf{x}}.$$

(c) Differentiate w.r.t. W:

$$\begin{aligned}
\frac{\partial}{\partial W}\mathcal{L}(\mu, W) &= \frac{\partial}{\partial W}\left(-\frac{N}{2}\log|\sigma^2 I + WW^\top| - \frac{1}{2}\sum_{i=1}^N \mathrm{tr}\left((\sigma^2 I + WW^\top)^{-1}(\mathbf{x}_i - \mu)(\mathbf{x}_i - \mu)^\top\right)\right) \\
&= \frac{\partial}{\partial W}\left(-\frac{N}{2}\log|\sigma^2 I + WW^\top| - \frac{N}{2}\sum_{i=1}^N \mathrm{tr}\left((\sigma^2 I + WW^\top)^{-1}S\right)\right).
\end{aligned}$$

Let $C = \sigma^2 I + WW^\top$. From Problem 5(c) in Homework 1,

$$\frac{\partial}{\partial C}\log|C| = C^{-1}.$$

Also,

$$\frac{\partial}{\partial C}\mathrm{tr}\left(C^{-1}S\right) = -C^{-1}SC^{-1}.$$

To prove this, we first observe that

$$CC^{-1} = I \;\Rightarrow\; \frac{\partial C}{\partial x}C^{-1} + C\frac{\partial C^{-1}}{\partial x} = 0 \;\Rightarrow\; \frac{\partial C^{-1}}{\partial x} = -C^{-1}\frac{\partial C}{\partial x}C^{-1}.$$

From Problem 5(a) in Homework 1,

$$\frac{\partial}{\partial C^{-1}}\operatorname{tr}\left(C^{-1}S\right) = S.$$

By the chain rule,

$$\frac{\partial}{\partial C_{ij}}\operatorname{tr}\left(C^{-1}S\right) = \left(-C^{-1}\frac{\partial C}{\partial C_{ij}}C^{-1}\right) : S$$

where $A : B = \operatorname{tr}(A^\top B) = \sum_{i,j} A_{ij}B_{ij}$. Then

$$\frac{\partial}{\partial C_{ij}}\operatorname{tr}\left(C^{-1}S\right) = -\operatorname{tr}\left(C^{-1}SC^{-1}\frac{\partial C}{\partial C_{ij}}\right) = -(C^{-1}SC^{-1})_{ij}$$

and so

$$\frac{\partial}{\partial C}\operatorname{tr}\left(C^{-1}S\right) = -C^{-1}SC^{-1}.$$

Now, back to the problem. By the matrix differentiation formula,

$$\frac{\partial}{\partial C}\left(-\frac{N}{2}\log|C| - \frac{N}{2}\sum_{i=1}^{N}\operatorname{tr}\left(C^{-1}S\right)\right) = -\frac{N}{2}\left(C^{-1} - C^{-1}SC^{-1}\right) =: P.$$

Also,

$$\frac{\partial}{\partial W_{ij}}C = \sum_{k=1}^{D} e_{ki}W_{kj} + \sum_{k=1}^{D} e_{ik}W_{kj}$$

where e_{mn} is a $D \times D$ matrix such that (m,n)-entry is one and other entries are zeros. By the chain rule,

$$\frac{\partial}{\partial W_{ij}}\left(-\frac{N}{2}\log|C| - \frac{N}{2}\sum_{i=1}^{N}\operatorname{tr}\left(C^{-1}S\right)\right) = P : \left(\sum_{k=1}^{D} e_{ki}W_{kj} + \sum_{k=1}^{D} e_{ik}W_{kj}\right)$$

where $A : B = \operatorname{tr}(A^\top B) = \sum_{i,j} A_{ij}B_{ij}$. Then

$$P : \left(\sum_{k=1}^{D} e_{ki}W_{kj} + \sum_{k=1}^{D} e_{ik}W_{kj}\right) = \sum_{k=1}^{D} P_{ki}W_{kj} + \sum_{k=1}^{D} P_{ik}W_{kj} = (P^\top W)_{ij} + (PW)_{ij}.$$

Thus,

$$\frac{\partial}{\partial W}\left(-\frac{N}{2}\log|C| - \frac{N}{2}\sum_{i=1}^{N}\operatorname{tr}\left(C^{-1}S\right)\right) = P^\top W + PW = 2PW = -N(C^{-1}W - C^{-1}SC^{-1}W)$$

since P is symmetric. Therefore,

$$W_{ML} = SC^{-1}W_{ML} = S(WW^\top + \sigma^2 I)^{-1}W_{ML}.$$