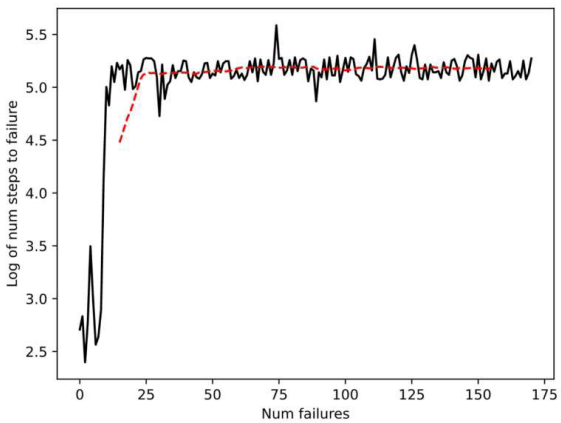
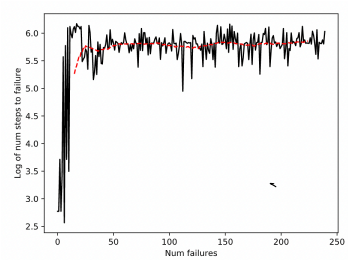
Question 1

i) It takes about 175 trials to converge

ii)




seed 1 — 240.




ii) As seed index changes to 1, 2, 3, the model has different convergance time than index at 0. In general, the time for the model to converge is longer at index = 1, 2, 3.
Since the random generator is completely depending on the random seed, it directly influences the initial value vector. Hence, we can imply that the initial value vector can affect the performance of our reinforced learning model.

# Question 2

(a) Given $P(x) > 0$, $Q(x) > 0$ $\forall P, Q$.

$$D_{KL}(P\|Q) = \sum_{x \in X} P(x) \log \frac{P(x)}{Q(x)}$$

let $f(x) = -\log(x)$ be the stictly convex function

since $X$ is a RV. we can apply the Jensen's Inequality:

$$D_{KL}(P\|Q) = E_{x\sim P}\left[-\log \frac{Q(x)}{P(x)}\right]$$

$$\geq -\log E_{x\sim p}\left[\frac{Q(x)}{P(x)}\right]$$

$$= -\log \sum_{x \in X} P(x) \cdot \frac{Q(x)}{P(x)}$$

$$= -\log \sum_{x \in X} Q(x) \qquad [Q(x) \text{ is distribution of } X]$$

$$= -\log 1$$

$$= 0.$$

Hence, $D_{KL}(P\|Q) \geq 0$ $\forall P, Q$.

"$\Rightarrow$"

if $P = Q$. we have $D_{KL}(P\|Q) = E_{x\sim p}\left[-\log \frac{Q(x)}{P(x)}\right] \geq -\log E_{x\sim p}\left[\frac{Q(x)}{P(x)}\right]$

RHS $= 0$ as proved in part i.

LHS $= E_{x\sim p}[-\log 1] = \sum_{x \in X} P(x) \cdot 0 = 0 = RHS.$

proved!

"$\Leftarrow$"

suppose $D_{KL}(P\|Q) = 0 = E_{x\sim p}\left[-\log \frac{Q(x)}{P(x)}\right]$

because we know $-\log E_{x\sim p}\left[\frac{Q(x)}{P(x)}\right] = 0$

Hence $E[f(x)] = f(E[x])$

which implies that $X = E(X)$ with probability 1.

So, the distribution $P, Q$ over $X$ must be the same.

$P = Q$.

∎

(b) $D_{KL}\left(P(X,Y) \| Q(X,Y)\right)$

$$= \sum_x \sum_y P(x,y) \log \frac{P(x,y)}{Q(x,y)}$$

$$= \sum_x \sum_y P(x,y) \log \frac{P(Y|X)P(X)}{Q(Y|X)Q(X)}$$

$$= \sum_x \sum_y P(x,y) \log \frac{P(X)}{Q(X)} + \sum_x \sum_y P(x,y) \log \frac{P(Y|X)}{Q(Y|X)}$$

The first part:

$$\sum_x \sum_y P(x,y) \log \frac{P(X)}{Q(X)} = \sum_x \sum_y P(x)P(y|x) \log \frac{P(X)}{Q(X)}$$

$$= \sum_x P(X) \sum_y P(y|x) \log \frac{P(X)}{Q(X)} \quad \text{since } y \text{ is irrelevant}$$

$$= \sum_x P(x) \log \frac{P(X)}{Q(X)} = D_{KL}\left(P(X) \| Q(X)\right)$$

The second part:

$$\sum_x \sum_y P(x,y) \log \frac{P(Y|X)}{Q(Y|X)} = \sum_x \sum_y P(y)P(x|y) \log \frac{P(Y|X)}{Q(Y|X)}$$

$$= \sum_y P(y) \left( \sum_x P(x|y) \log \frac{P(Y|X)}{Q(Y|X)} \right)$$

$$= D_{KL}\left(P(X|Y) \| Q(X|Y)\right)$$

Hence, the chain rule is proved.

(c) Expand $D_{KL}(\hat{P} \| P_\theta)$

$$= \sum_{x \in X} \hat{P}(x) \log \frac{\hat{P}(x)}{P_\theta(x)}$$

$$= \sum_{x \in X} \hat{P}(x) \log \hat{P}(x) - \sum_{x \in X} \hat{P}(x) \log P_\theta(x)$$

Given $\hat{P}(x)$ is a uniform distribution over $x \in X$.

so the first half of the equation is a constant term.

$$\operatorname{argmin}_\theta D_{KL}(\hat{P} \| P_\theta)$$

$$= \operatorname{argmin}_\theta - \sum_{x \in X} \hat{P}(x) \log P_\theta(x)$$

$$= \operatorname{argmax}_\theta \sum_{x \in X} \hat{P}(x) \log P_\theta(x)$$

since $\hat{P}(x)$ is constant for $\forall x \in X$.

$$= \operatorname{Argmax}_\theta \sum_{x \in X} \log P_\theta(x)$$

Say the $x \in X$ is defined as $\{x^{(i)}; i = 1, \cdots n\}$.

we thus show the equivalance.

Question : 3

part a :

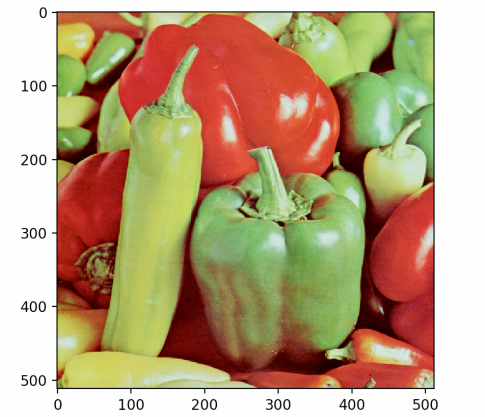
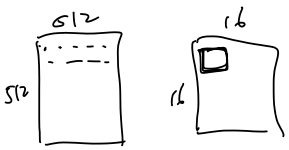



fig. large peppers (original)

fig. large peppers (generated)




3 x (8)

part b :

originally, each pixel in the image is stored using 24-bit ( $\log_2(256) = 8$ ; $3 \times 8 = 24$ )

After compressing using only 16 colors, each pixel only takes 4( $\log_2(16) = 4$ ) bits.

so the image compressed by a factor $\frac{4}{24} = \frac{1}{6}$ .

# Question 4

part a.

we want to prove the algorithm converges

⇔ proving $\ell_{\text{semi-sup}}(\theta)$ is monotonic increasing

i.e. let $\theta^{(t)}$ be the parameter at the end of $t$ EM-step, we show $\ell_{\text{semi-sup}}(\theta^{(t+1)}) \geq \ell_{\text{semi-sup}}(\theta^{(t)})$

$$\begin{aligned}
\ell_{\text{semi-sup}}\,\theta^{(t+1)} &= \ell_{\text{unsup}}\,\theta^{(t+1)} + \alpha\, \ell_{\text{sup}}\,\theta^{(t+1)} \\
&= \textstyle\sum_{i=1}^{n} \log \sum_{z^{(i)}} p(x^{(i)}, z^{(i)}; \theta^{(t+1)}) + \alpha\, \ell_{\text{sup}}\,\theta^{(t+1)} \\
&= \textstyle\sum_{i=1}^{n} \log p(x^{(i)}; \theta^{(t+1)}) + \alpha\, \ell_{\text{sup}}\,\theta^{(t+1)} \\
&\geq \textstyle\sum_{i=1}^{n} \text{ELBO}\left(x^{(i)}, Q_i^{(t)}, \theta^{(t+1)}\right) + \alpha\, \ell_{\text{sup}}\,\theta^{(t+1)} \qquad \text{[by Jensen's Inequality]} \\
&\geq \textstyle\sum_{i=1}^{n} \text{ELBO}\left(x^{(i)}, Q_i^{(t)}, \theta^{(t+1)}\right) + \alpha\, \ell_{\text{sup}}\,\theta^{(t)} \qquad \text{[The E-step]} \\
&= \ell_{\text{unsup}}(\theta^{(t)}) + \alpha\, \ell_{\text{sup}}(\theta^{(t)}) \\
&= \ell_{\text{semi-sup}}(\theta^{(t)})
\end{aligned}$$

since $\theta^{(t)}$ is the end of the last EM step. it satisfies that

$$\ell_{\text{unsup}}(\theta^{(t)}) = \textstyle\sum_{i=1}^{n} \text{ELBO}(x^{(i)}, Q_i^{(t)}, \theta^{(t+1)})$$

for $\forall t$, we show that $\ell_{\text{semi-sup}}(\theta^{(t+1)}) \geq \ell_{\text{sup}}(\theta^{(t)})$

$\ell(\theta)$ converges.

part b:

we need to reestimate $z^{(i)}$ for $i = 1, \cdots, n$

Calculate $w_j^{(i)} = Q_{(i)}(z^{(i)} = j) = P(z^{(i)} = j \mid x^{(i)}; \phi, \mu, \Sigma)$

$$= \frac{P(x^{(i)} \mid z^{(i)} = j; \mu, \Sigma)\, P(z^{(i)} = j; \phi)}{\sum_{l=1}^{K} P(x^{(i)} \mid z^{(i)} = l; \mu, \Sigma)\, P(z^{(i)} = l; \phi)}$$

$$= \frac{\frac{1}{(2\pi)^{d/2} |\Sigma_j|^{1/2}} \exp\left(-\frac{1}{2}(x^{(i)} - \mu_j)^T \Sigma_j^{-1} (x^{(i)} - \mu_j)\right) \cdot \phi_j}{\sum_{l=1}^{K} \frac{1}{(2\pi)^{d/2} |\Sigma_l|^{1/2}} \exp\left(-\frac{1}{2}(x^{(i)} - \mu_l)^T \Sigma_l^{-1} (x^{(i)} - \mu_l)\right) \cdot \phi_l}$$ (according to NOTE 11)

part c: will reestimate $\phi_j$, $\mu_j$, $\Sigma_j$ for $j = 1, \cdots, K$

say $Q_i(z^{(i)} = j)$ denotes the prob of $z^{(i)}$ taking the value $j$ under dist $Q_i$.

let $w_j^{(i)} = Q_i(z^{(i)} = j)$ and $\tilde{w}_j^{(i)} = 2 \cdot 1\{z^{(i)} = j\}$

the quantity we want to maximize is:

$$\sum_{i=1}^{n} \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{P(x^{(i)}, z^{(i)}; \phi, \mu, \Sigma)}{Q_i(z^{(i)})} + \sum_{i=1}^{\tilde{n}} \log\left(P(\tilde{x}^{(i)}, \tilde{z}^{(i)}; \phi, \mu, \Sigma)\right)$$

$$= \sum_{i=1}^{n} \sum_{l=1}^{K} w_l^{(i)} \log \frac{\frac{1}{(2\pi)^{d/2} |\Sigma_l|^{1/2}} \exp\left(-\frac{1}{2}(x^{(i)} - \mu_l)^T \Sigma_l^{-1} (x^{(i)} - \mu_l)\right) \cdot \phi_l}{w_l^{(i)}} + \sum_{i=1}^{\tilde{n}} \sum_{l=1}^{K} \tilde{w}_l^{(i)} \log \frac{\frac{1}{(2\pi)^{d/2} |\Sigma_l|^{1/2}} \exp\left(-\frac{1}{2}(x^{(i)} - \mu_l)^T \Sigma_l^{-1} (x^{(i)} - \mu_l)\right) \cdot \phi_l}{\tilde{w}_l^{(i)}}$$

① Maximize w.r.t. $\mu$. we attain the gradient:

$$-\nabla_{\mu_l} \sum_{i=1}^{n} w_l^{(i)} \frac{1}{2}(x^{(i)} - \mu_l)^T \Sigma_l^{-1} (x^{(i)} - \mu_l) + \sum_{i=1}^{\tilde{n}} \tilde{w}_l^{(i)} \frac{1}{2}(x^{(i)} - \mu_l)^T \Sigma_l^{-1} (x^{(i)} - \mu_l)$$

$$= \frac{1}{2} \sum_{i=1}^{n} w_l^{(i)} \nabla_{\mu_l} 2\mu_l^T \Sigma_l^{-1} x^{(i)} - \mu_l^T \Sigma_l^{-1} \mu_l + \left[\frac{1}{2} \sum_{i=1}^{\tilde{n}} w_l^{(i)} \nabla_{\mu_l} 2\mu_l^T \Sigma_l^{-1} x^{(i)} - \mu_l^T \Sigma_l^{-1} \mu_l\right]$$

$$= \sum_{i=1}^{n} w_l^{(i)} \left(\Sigma_l^{-1} x^{(i)} - \Sigma_l^{-1} \mu_l\right) + \sum_{i=1}^{\tilde{n}} w_l^{(i)} \left(\Sigma_l^{-1} x^{(i)} - \Sigma_l^{-1} \mu_l\right)$$

$$= \Sigma_l^{-1} \left(\sum_{i=1}^{n} w_l^{(i)} x^{(i)} + \sum_{i=1}^{\tilde{n}} w_l^{(i)} x^{(i)}\right) - \Sigma_l^{-1} \left(\sum_{i=1}^{n} w_l^{(i)} + \sum_{i=1}^{\tilde{n}} w_l^{(i)}\right) \mu_l = 0 .$$

So $$\mu_l = \frac{\sum_{i=1}^{n} w_l^{(i)} x^{(i)} + \sum_{i=1}^{\tilde{n}} w_l^{(i)} \tilde{x}^{(i)}}{\sum_{i=1}^{n} w_l^{(i)} + \sum_{i=1}^{\tilde{n}} w_l^{(i)}}$$

② Maximize w.r.t. $\phi$

construct the Lagrangian

$$\alpha(\phi) = \sum_{i=1}^{n} \sum_{j=1}^{k} w_j^{(i)} \log \phi_j + \sum_{i=1}^{\tilde{n}} \sum_{j=1}^{k} \tilde{w}_j^{(i)} \log \phi_j + \beta\left(\sum_{j=1}^{k} \phi_j - 1\right)$$

$$\frac{\partial \alpha(\phi)}{\partial \phi_j} = \sum_{i=1}^{n} w_j^{(i)} \frac{1}{\phi_j} + \sum_{i=1}^{\tilde{n}} \tilde{w}_j^{(i)} \frac{1}{\phi_j} + \beta = 0.$$

$$\phi_j = \frac{\sum_{i=1}^{n} w_j^{(i)} + \sum_{i=1}^{\tilde{n}} w_j^{(i)}}{-\beta}$$

using the constraint $\sum_j \phi_j = 1$, we easily find that

$$-\beta = \sum_{j=1}^{k} \sum_{i=1}^{n} w_j^{(i)} + \sum_{j=1}^{k} \sum_{i=1}^{\tilde{n}} \tilde{w}_j^{(i)} = n + \alpha\tilde{n}$$

following the Lecture note, we have

$$\phi_j = \frac{\sum_{i=1}^{n} w_j^{(i)} + \sum_{i=1}^{\tilde{n}} w_j^{(i)}}{n + \alpha\tilde{n}}$$

③ Maximize w.r.t. $\Sigma$, the gradient is:

$$\nabla_{\Sigma_j}\left(\sum_{i=1}^{n} w_j^{(i)} (\log|\Sigma_j| - (x^{(i)} - \mu_j)^T \Sigma_j^{-1} (x^{(i)} - \mu_j))\right) + \left(\sum_{i=1}^{\tilde{n}} \tilde{w}_j^{(i)} (\log|\Sigma_j| - (x^{(i)} - \mu_j)^T \Sigma_j^{-1} (x^{(i)} - \mu_j))\right)$$

$$= \left(\sum_{i=1}^{n} w_j^{(i)} + \sum_{i=1}^{\tilde{n}} \tilde{w}_j^{(i)}\right)\Sigma_j - \left(\sum_{i=1}^{n} w_j^{(i)} (x^{(i)} - \mu_j)(x^{(i)} - \mu_j)^T\right) + \sum_{i=1}^{\tilde{n}} \tilde{w}_j^{(i)} \left((x^{(i)} - \mu_j)(x^{(i)} - \mu_j)^T\right) = 0$$

$$\text{So, } \Sigma_j = \frac{\sum_{i=1}^{n} w_j^{(i)} (x^{(i)} - \mu_j)(x^{(i)} - \mu_j)^T + \sum_{i=1}^{\tilde{n}} \tilde{w}_j^{(i)} (x^{(i)} - \mu_j)(x^{(i)} - \mu_j)^T}{\sum_{i=1}^{n} w_j^{(i)} + \sum_{i=1}^{\tilde{n}} \tilde{w}_j^{(i)}}$$

part d:

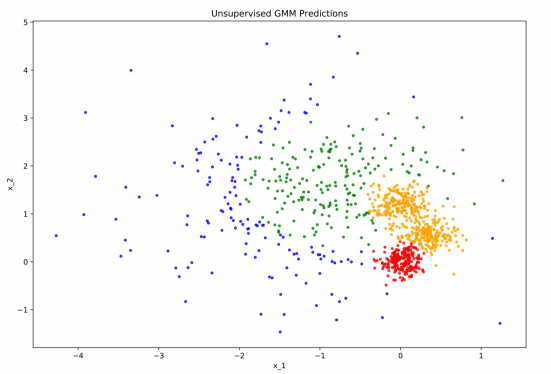



fig. unsupervised GMM

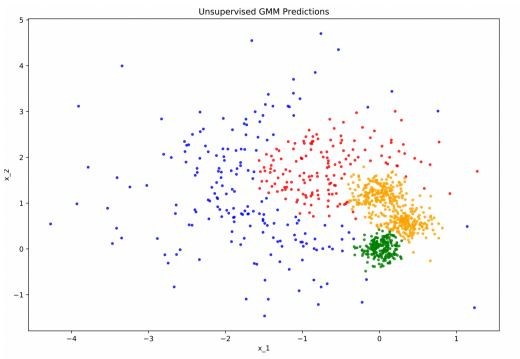



fig. unsupervised GMM

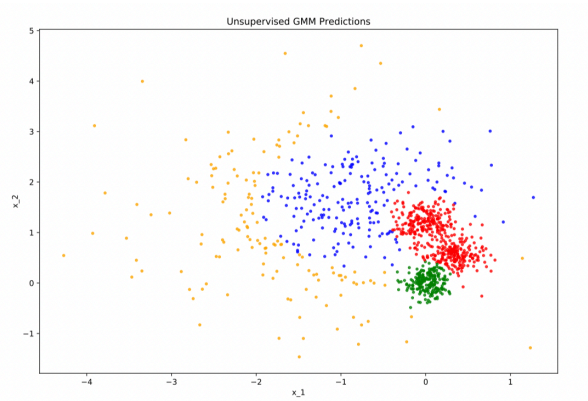



fig. unsupervised GMM

part e:

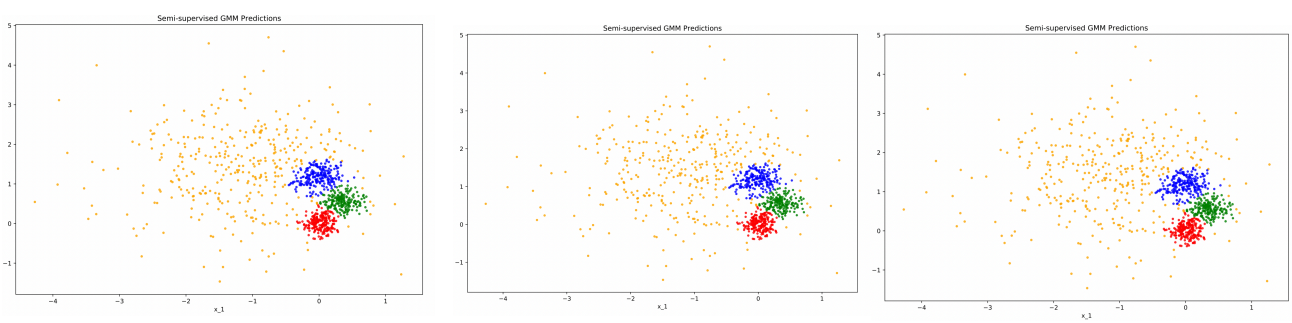



fig. semi-supervised GMM

f.

i) The semi-supervised model has a faster convergence rate. The semi-supervised model takes at average 25 iterations to converge, whereas unsupervised model takes at average 157 iterations

ii) with random initialization, unsupervised EM is less stable, which is obvious in 3 graphs attached at part d.
semi-supervised EM is more stable.

iii) overall, semi-supervised EM outperforms unsupervised EM. Since semi-supervised EM has additional matching pairs, it provides more accurate identification of clusters under random initialization and when the distribution of the training data is not quite obvious.

Question 5

show first principle component:

$$\arg\min_{u:\|u\|=1} \sum_{i=1}^{n} \| x^{(i)} - f_u(x^{(i)}) \|_2^2$$

$$= \arg\max_{u:\|u\|=1} \sum_{i=1}^{n} \| x^{(i)} \|^2 - \sum_{i=1}^{n} \| x^{(i)} - f_u(x^{(i)}) \|_2^2$$

$$= \arg\max_{u:\|u\|=1} \sum_{i=1}^{n} \left( \| x^{(i)} \|^2 - \| x^{(i)} - f_u(x^{(i)}) \|_2^2 \right)$$

$$= \arg\min_{u:\|u\|=1} \frac{1}{n} \sum_{i=1}^{n} \| f_u(x^{(i)}) \|_2^2$$

$$= \arg\min_{u:\|u\|=1} \frac{1}{n} \sum_{i=1}^{n} \left( (x^{(i)T} u)^2 \right)$$

$$= \arg\min_{u:\|u\|=1} u^T \left( \frac{1}{n} \sum_{i=1}^{n} (x^{(i)} x^{(i)T}) \right) u$$

Here we see that minimizing $\sum_{i=1}^{n} \| x^{(i)} - f_u(x^{(i)}) \|_2^2$ is equivalent to finding the max variance of projections onto the direction: $u: \frac{1}{n} \sum_{i=1}^{n} \| f_u(x^{(i)}) \|_2^2$, which is what first principle component of PCA is about: Maximizing the variance.

Question 6

part a:

Maximizing $l(W)$:

$$l(W) = \sum_{i=1}^{n} \log P_x(X^{(i)})$$

$$= \sum_{i=1}^{n} \log P_s(Wx^{(i)})|W|)$$

$$= \sum_{i=1}^{n} \log \left( \frac{1}{(2\pi)^{d/2}} \exp\left[ -\frac{1}{2} (Wx^{(i)})^T (Wx^{(i)}) \right) |W| \right)$$

$$= \sum_{i=1}^{n} \left[ -\frac{d}{2} \log(2\pi) - \frac{1}{2} x^{(i)T} W^T W x^{(i)} + \log|W| \right)$$

Then we take the gradient w.r.t. $W$

$$\nabla_W l(W) = \sum_{i=1}^{n} \left( -\frac{1}{2} \nabla_W X^{(i)T} W^T W X^{(i)} + \nabla_W \log|W| \right)$$

$$= \sum_{i=1}^{n} \left[ -W x^{(i)} x^{(i)T} + (W^{-1})^T \right)$$

$$= -W \left( \sum_{i=1}^{n} x^{(i)} x^{(i)T} \right) + n (W^{-1})^T$$

$$= -WX^TX + n(W^{-1})^T = 0$$

So $W^TW = \frac{1}{n}(X^TX)^{-1}$

Assuming RHS is invertible:

if there exists a valid solution $W$, then any $W' = W \times U$ where $U$ is an orthogonal matrix, $W'$ can be a valid solution too. This leads to a ambiguity, where $U$ cannot be determined by data $X$.

So, ICA cannot recover the original.

Part b:

let $x^{(i)}$ be an arbitrary example, the likelihood func is:

$$\begin{aligned} l_i(W) &= \sum_{j=1}^{d} \log P_s(w_j^T x^{(i)}) + \log|W| \\ &= \sum_{j=1}^{d} \log\left(\frac{1}{2}\exp(-|w_j^T x^{(i)}|)\right) + \log|W| \\ &= -d\log(2) - \sum_{j=1}^{d} |w_j^T x^{(i)}| + \log|W| \end{aligned}$$

Take the derivative:

$$\begin{aligned} \nabla_W l_i(W) &= -\sum_{j=1}^{d} \nabla_W |w_j^T x^{(i)}| + \nabla_W \log|W| \\ &= -\sum_{j=1}^{d} \operatorname{sign}(w_j^T x^{(i)}) \begin{bmatrix} x^{(i)T} \\ 0 \\ \vdots \\ 0 \end{bmatrix} \text{(row } j\text{)} + (W^{-1})^T \\ &= \begin{bmatrix} \operatorname{sign}(w_1^T x^{(i)}) \\ \vdots \\ \operatorname{sign}(w_d^T x^{(i)}) \end{bmatrix} x^{(i)T} + (W^{-1})^T \end{aligned}$$

Take the SGA update rule to the above case, we get:

$$W := W + \alpha\left(-\begin{bmatrix} \operatorname{sign}(w_1^T x^{(i)}) \\ \vdots \\ \operatorname{sign}(w_d^T x^{(i)}) \end{bmatrix} x^{(i)T} + (W^{-1})^T\right)$$

part c:

```
[[ 52.8352532   16.79619701  19.94171825 -10.19846303 -20.89757762]
 [ -9.9292747   -0.97875614  -4.67786427   8.04377382   1.7865852 ]
 [  8.31096507  -7.47675728  19.31500349  15.17429591 -14.32612384]
 [-14.66742843 -26.64517989   2.44081559  21.38210464  -8.4207738 ]
 [ -0.26929644  18.37414675   9.31198649   9.10287095  30.59463426]]
```

List. The unmixing matrix with Laplace