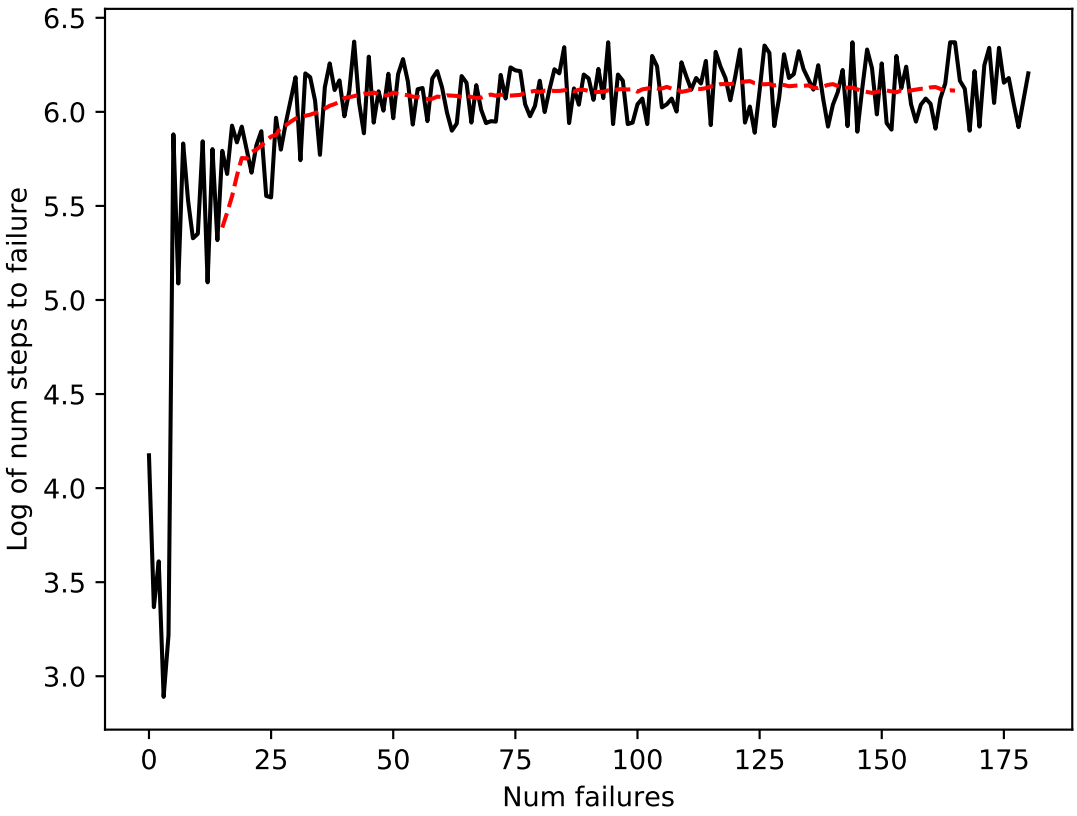
CS229 Problem Set 3

Tianyu Du

Sunday 11$^{\text{th}}$ August, 2019

1 Problem 1: Reinforcement Learning: The inverted pendulum

Answer `[INFO] Failure number 181`



2 Problem 2: KL Divergence and Maximum Likelihood

2.1 (a) Non-negativity

Proof. Show:

$$\forall P, Q, D_{KL}(P||Q) \geq 0 \tag{2.1}$$

Let P, Q characterize two arbitrary distributions over $\mathcal{X}$,

$$D_{KL}(P||Q) := \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{Q(x)} \tag{2.2}$$

$$= \sum_{x \in \mathcal{X}} P(x) \left(-\log \frac{Q(x)}{P(x)} \right) \tag{2.3}$$

$$= \mathbb{E}_{x \sim P} \left[-\log \frac{Q(x)}{P(x)} \right] \tag{2.4}$$

Since $-\log(\cdot)$ is a convex function, by Jensen's inequality

$$\mathbb{E}_{x \sim P} \left[-\log \frac{Q(x)}{P(x)} \right] \geq -\log \left(\mathbb{E}_{x \sim P} \left[\frac{Q(x)}{P(x)} \right] \right) \tag{2.5}$$

$$= -\log \left(\sum_{x \in \mathcal{X}} P(x) \frac{Q(x)}{P(x)} \right) \tag{2.6}$$

$$= -\log \left(\sum_{x \in \mathcal{X}} Q(x) \right) \tag{2.7}$$

$$= -\log(1) \because \ Q(\cdot) \text{ is a distribution} \tag{2.8}$$

$$= 0 \tag{2.9}$$

Therefore,

$$D_{KL}(P||Q) \geq 0 \tag{2.10}$$

Show

$$D_{KL}(P||Q) = 0 \iff P = Q \tag{2.11}$$

Suppose $P(x) = Q(x)$ for every $x \in \mathcal{X}$,

$$D_{KL}(P||Q) = \sum_{x \in \mathcal{X}} P(x) \log(1) \tag{2.12}$$

$$= \sum_{x \in \mathcal{X}} P(x) \ 0 = 0 \tag{2.13}$$

Show

$$D_{KL}(P||Q) = 0 \implies P = Q \tag{2.14}$$

Note that the second part of Jensen's inequality suggested that the equality, $D_{KL}(P||Q) = 0$ implies $-\log \frac{Q(x)}{P(x)}$ is constant over $\mathcal{X}$. Also, since $-\log(\cdot)$ is in fact injective, so it must be the case that $\frac{Q(x)}{P(x)}$ is constant over $\mathcal{X}$. Therefore, $\mathbb{E}_{x\sim P}\left[\frac{Q(x)}{P(x)}\right] = \frac{P(c)}{Q(c)}$ for every $c \in \mathcal{X}$. Further, because $\mathbb{E}_{x\sim P}\left[\frac{Q(x)}{P(x)}\right] = \sum_{x\in\mathcal{X}} P(x)\frac{Q(x)}{P(x)} = \sum_{x\in\mathcal{X}} Q(x)$, which equals one because Q is a probability distribution over $\mathcal{X}$. Therefore, for every $c \in \mathcal{X}$, $P(c) = Q(c)$. That's, $P = Q$.

■

2.2 (b) Chain rule for KL divergence

Proof.

$$D_{KL}(P(X)||Q(X)) + D_{KL}(P(Y|X)||Q(Y|X)) \tag{2.15}$$

$$= \sum_x P(x) \log \frac{P(x)}{Q(x)} + \sum_y P(y) \sum_x P(x|y) \log \frac{P(x|y)}{Q(x|y)} \tag{2.16}$$

$$= \sum_x \left(\sum_y P(x,y) \right) \log \frac{P(x)}{Q(x)} + \sum_y \sum_x P(y)P(x|y) \log \frac{P(x|y)}{Q(x|y)} \tag{2.17}$$

$$= \sum_{x,y} P(x,y) \log \frac{P(x)}{Q(x)} + \sum_{x,y} P(x,y) \log \frac{P(x|y)}{Q(x|y)} \tag{2.18}$$

$$= \sum_{x,y} P(x,y) \log \frac{P(x)}{Q(x)} + \log \frac{P(x|y)}{Q(x|y)} \tag{2.19}$$

$$= \sum_{x,y} P(x,y) \log \frac{P(x)P(x|y)}{Q(x)Q(x|y)} \tag{2.20}$$

$$= \sum_{x,y} P(x,y) \log \frac{P(x,y)}{Q(x,y)} \tag{2.21}$$

$$= D_{KL}(P(X,Y)||Q(X,Y)) \tag{2.22}$$

■

2.3 (c) KL and maximum likelihood

Proof. Let $\mathcal{M}$ denote the support of $\hat{P}$ (indeed, $\mathcal{M}$ is the set of distinct elements in dataset $\{x^{(i)}\}_{i=1}^n$), which is a subset of support of P_θ.

$$\begin{aligned}
\operatorname*{argmin}_\theta D_{KL}(\hat{P}||P_\theta) &= \operatorname*{argmin}_\theta \sum_{x\in\mathcal{M}} \hat{P}(x)\log\frac{\hat{P}(x)}{P_\theta(x)} && (2.23)\\
&= \operatorname*{argmin}_\theta \underbrace{\sum_{x\in\mathcal{M}} \hat{P}(x)\log\hat{P}(x)}_{\perp\theta} - \sum_{x\in\mathcal{M}} \hat{P}(x)\log P_\theta(x) && (2.24)\\
&= \operatorname*{argmin}_\theta -\sum_{x\in\mathcal{M}} \hat{P}(x)\log P_\theta(x) && (2.25)\\
&= \operatorname*{argmax}_\theta \sum_{x\in\mathcal{M}} \hat{P}(x)\log P_\theta(x) && (2.26)\\
&= \operatorname*{argmax}_\theta \sum_{x\in\mathcal{M}} \left(\frac{1}{n}\sum_{i=1}^n \mathbb{1}\{x^{(i)}=x\}\right)\log P_\theta(x) && (2.27)\\
&= \operatorname*{argmax}_\theta \frac{1}{n}\sum_{x\in\mathcal{M}} \left(\sum_{i=1}^n \mathbb{1}\{x^{(i)}=x\}\right)\log P_\theta(x) && (2.28)
\end{aligned}$$

For each $x\in\mathcal{M}$, $\sum_{i=1}^n \mathbb{1}\{x^{(i)}=x\}$ in fact counts the number of occurrences of x in dataset $\{x^{(i)}\}_{i=1}^n$. Therefore, the scope of summation above can be transformed from $\mathcal{M}$ to $\{x^{(i)}\}_{i=1}^n$, by dropping the occurrence counting multiplier $\sum_{i=1}^n \mathbb{1}\{x^{(i)}=x\}$.[1] Hence,

$$\operatorname*{argmax}_\theta \frac{1}{n}\sum_{x\in\mathcal{M}} \left(\sum_{i=1}^n \mathbb{1}\{x^{(i)}=x\}\right)\log P_\theta(x) = \frac{1}{n}\sum_{i=1}^n \log P_\theta(x^{(i)}) \tag{2.29}$$

Dropping the positive term $\frac{1}{n}$ does not affect the maximizer. Therefore,

$$\operatorname*{argmin}_\theta D_{KL}(\hat{P}||P_\theta) = \operatorname*{argmax}_\theta \sum_{i=1}^n \log P_\theta(x^{(i)}) \tag{2.30}$$

■

[1] This can be easily verified by writing out all terms in the summation.

3 Problem 3: K-means for Compression

3.1 (a) K-Means Compression Implementation

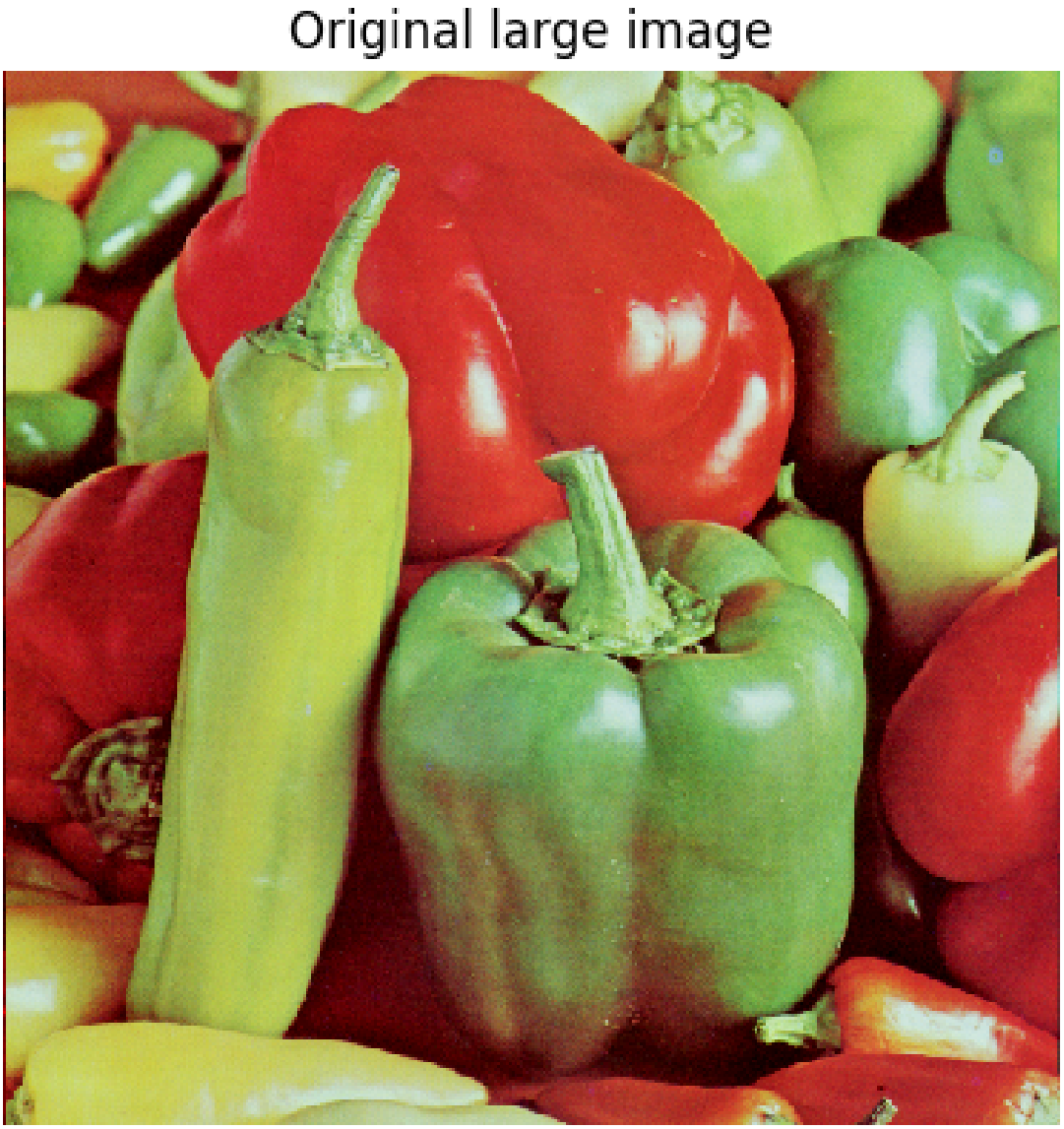


Figure 1: Original Large

Figure 2: Original Small

Figure 3: Compressed Large

3.2 (b) Compression Factor

Answer In the original 8-bit 3-channel representation, each pixel requires $3 \times 8 = 24$ bits to be fully characterized. In the compressed representation, each pixel can be mapped to one of 16 clusters indexed by 0 to 15. To represent which cluster one particular pixel belongs to, one would need 4-bits (from 0000 to 1111). Therefore, images are compressed by factor of $\frac{24}{4} = 6$.

4 Problem 4: Semi-supervised EM

4.1 (a) Convergence

Proof.

$$\ell_{\text{semi-sup}}(\theta^{(t+1)}) = \ell_{\text{unsup}}(\theta^{(t+1)}) + \alpha\ell_{\text{sup}}(\theta^{(t+1)}) \tag{4.1}$$

$$= \sum_{i=1}^{n} \log \left(\sum_{z^{(i)}} Q_i^{(t)} \frac{p(x^{(i)}, z^{(i)}; \theta^{(t+1)})}{Q_i^{(t)}} \right) + \alpha\ell_{\text{sup}}(\theta^{(t+1)}) \tag{4.2}$$

$$= \sum_{i=1}^{n} \log \mathbb{E}_{z^{(i)} \sim Q_i^{(t)}} \left[\frac{p(x^{(i)}, z^{(i)}; \theta^{(t+1)})}{Q_i^{(t)}} \right] + \alpha\ell_{\text{sup}}(\theta^{(t+1)}) \tag{4.3}$$

$$\geq \sum_{i=1}^{n} \mathbb{E}_{z^{(i)} \sim Q_i^{(t)}} \left[\log \frac{p(x^{(i)}, z^{(i)}; \theta^{(t+1)})}{Q_i^{(t)}} \right] + \alpha\ell_{\text{sup}}(\theta^{(t+1)}) \because \text{Jensen's Inequality} \tag{4.4}$$

$$= \sum_{i=1}^{n} \sum_{z^{(i)}} Q_i^{(t)} \log \left(\frac{p(x^{(i)}, z^{(i)}; \theta^{(t+1)})}{Q_i^{(t)}} \right) + \alpha\ell_{\text{sup}}(\theta^{(t+1)}) \tag{4.5}$$

$$\geq \sum_{i=1}^{n} \sum_{z^{(i)}} Q_i^{(t)} \log \left(\frac{p(x^{(i)}, z^{(i)}; \theta^{(t)})}{Q_i^{(t)}} \right) + \alpha\ell_{\text{sup}}(\theta^{(t)}) \because \text{M-step is maximizing w.r.t. } \theta \tag{4.6}$$

$$= \sum_{i=1}^{n} \log \left(\sum_{z^{(i)}} Q_i^{(t)} \frac{p(x^{(i)}, z^{(i)}; \theta^{(t)})}{Q_i^{(t)}} \right) + \alpha\ell_{\text{sup}}(\theta^{(t)}) \tag{4.7}$$

$$= \ell_{\text{semi-sup}}(\theta^{(t)}) \tag{4.8}$$

The last two steps were derive from the fact that $Q(\cdot)$ was specifically chosen in the E-step so that $\ell_{\text{semi-sup}}(\theta^{(t)})$ was equal to it's ELBO. ■

4.2 (b) Semi-supervised E-step

Answer $z^{(i)}$ for all unlabelled examples should be estimated. Specifically, posterior $p(z^{(i)}|x^{(i)};\mu,\Sigma,\phi)$ for all $i \in \{1,\cdots,n\}$ are estimated. Define $w_j^{(i)} = p(z^{(i)}=j|x^{(i)};\mu,\Sigma,\phi)$.

Proof.

$$
\begin{aligned}
w_j^{(i)} &:= p(z^{(i)}=j|x^{(i)};\mu,\Sigma,\phi) && (4.9)\\
&= \frac{p(x^{(i)}|z^{(i)}=j;\mu,\Sigma,\phi)p(z^{(i)}=j;\mu,\Sigma,\phi)}{p(x^{(i)};\mu,\Sigma,\phi)} && (4.10)\\
&= \frac{p(x^{(i)}|z^{(i)}=j;\mu,\Sigma,\phi)p(z^{(i)}=j;\mu,\Sigma,\phi)}{\sum_{\ell=1}^{k}\left\{p(x^{(i)}|z^{(i)}=\ell;\mu,\Sigma,\phi)p(z^{(i)}=\ell;\mu,\Sigma,\phi)\right\}} && (4.11)\\
&= \frac{\frac{1}{(2\pi)^{d/2}|\Sigma_j|^{1/2}}\exp\left[-\frac{1}{2}(x^{(i)}-\mu_j)^T\Sigma_j^{-1}(x^{(i)}-\mu_j)\right]\phi_j}{\sum_{\ell=1}^{k}\left\{\frac{1}{(2\pi)^{d/2}|\Sigma_\ell|^{1/2}}\exp\left[-\frac{1}{2}(x^{(i)}-\mu_\ell)^T\Sigma_\ell^{-1}(x^{(i)}-\mu_\ell)\right]\phi_\ell\right\}} && (4.12)
\end{aligned}
$$

■

4.3 (c) Semi-supervised M-step

4.3.1 Choosing $\mu_\ell^{(t+1)}$

Answer Let $\Theta := \{\mu_\ell, \Sigma_\ell, \phi_\ell\}_{\ell=1}^k$. The first order condition is for optimal μ_ℓ is:

Proof.

$$\nabla_{\mu_\ell} \ell_{\text{semi-sup}}(\Theta) = \nabla_{\mu_\ell} \sum_{i=1}^{n}\sum_{j=1}^{k} w_j^{(i)} \log\left(\frac{p(x^{(i)}, z^{(i)}; \Theta)}{w_j^{(i)}}\right) + \alpha \sum_{i=1}^{\tilde{n}} \log\Big(p(\tilde{x}^{(i)}, \tilde{z}^{(i)}; \Theta)\Big) \tag{4.13}$$

$$= \nabla_{\mu_\ell} \sum_{i=1}^{n} w_\ell^{(i)} \log\left(\frac{p(x^{(i)}, z^{(i)}; \Theta)}{w_\ell^{(i)}}\right) + \alpha \sum_{i=1}^{\tilde{n}} \mathbb{1}\{\tilde{z}^{(i)} = \ell\} \log\Big(p(\tilde{x}^{(i)}, \tilde{z}^{(i)}; \Theta)\Big) \tag{4.14}$$

$$= \nabla_{\mu_\ell} \sum_{i=1}^{n} w_\ell^{(i)} \log\Big(p(x^{(i)}, z^{(i)}; \Theta)\Big) + \alpha \sum_{i=1}^{\tilde{n}} \mathbb{1}\{\tilde{z}^{(i)} = \ell\} \log\Big(p(\tilde{x}^{(i)}, \tilde{z}^{(i)}; \Theta)\Big) \tag{4.15}$$

$$= \nabla_{\mu_\ell} \sum_{i=1}^{n} w_\ell^{(i)} \log\Big(p(x^{(i)}|z^{(i)}; \Theta)p(z^{(i)}; \Theta)\Big) + \alpha \sum_{i=1}^{\tilde{n}} \mathbb{1}\{\tilde{z}^{(i)} = \ell\} \log\Big(p(\tilde{x}^{(i)}|\tilde{z}^{(i)}; \Theta)p(\tilde{z}^{(i)}; \Theta)\Big) \tag{4.16}$$

$$= \nabla_{\mu_\ell} \sum_{i=1}^{n} w_\ell^{(i)} \Big\{\log\Big(p(x^{(i)}|z^{(i)}; \Theta)\Big)\Big\} + \alpha \sum_{i=1}^{\tilde{n}} \mathbb{1}\{\tilde{z}^{(i)} = \ell\} \log\Big(p(\tilde{x}^{(i)}|\tilde{z}^{(i)}; \Theta)\Big) \tag{4.17}$$

$$= \nabla_{\mu_\ell} \sum_{i=1}^{n} w_\ell^{(i)} \left\{-\frac{1}{2}(x^{(i)} - \mu_\ell)^T \Sigma_\ell^{-1}(x^{(i)} - \mu_\ell)\right\} \tag{4.18}$$

$$+ \alpha \sum_{i=1}^{\tilde{n}} \mathbb{1}\{\tilde{z}^{(i)} = \ell\} \left\{-\frac{1}{2}(\tilde{x}^{(i)} - \mu_\ell)^T \Sigma_\ell^{-1}(\tilde{x}^{(i)} - \mu_\ell)\right\} \tag{4.19}$$

$$= \sum_{i=1}^{n} w_\ell^{(i)} \Big\{x^{(i)T}\Sigma_\ell^{-1} - \mu_\ell^T \Sigma_\ell^{-1}\Big\} + \alpha \sum_{i=1}^{\tilde{n}} \mathbb{1}\{\tilde{z}^{(i)} = \ell\} \Big\{\tilde{x}^{(i)T}\Sigma_\ell^{-1} - \mu_\ell^T \Sigma_\ell^{-1}\Big\} \tag{4.20}$$

$$= 0 \tag{4.21}$$

By right multiplying Σ_ℓ^{-1},

$$\sum_{i=1}^{n} w_\ell^{(i)} \Big\{x^{(i)} - \mu_\ell\Big\} + \alpha \sum_{i=1}^{\tilde{n}} \mathbb{1}\{\tilde{z}^{(i)} = \ell\} \Big\{\tilde{x}^{(i)} - \mu_\ell\Big\} = 0 \tag{4.22}$$

$$\implies \mu_\ell = \frac{\sum_{i=1}^{n} w_\ell^{(i)} x^{(i)} + \alpha \sum_{i=1}^{\tilde{n}} \mathbb{1}\{\tilde{z}^{(i)} = \ell\}\tilde{x}^{(i)}}{\sum_{i=1}^{n} w_\ell^{(i)} + \alpha \sum_{i=1}^{\tilde{n}} \mathbb{1}\{\tilde{z}^{(i)} = \ell\}} \tag{4.23}$$

■

4.3.2 Choosing Σ_ℓ

Proof. Let $S_\ell := \Sigma_\ell^{-1}$. From lecture we know that the first order condition of optimizing Σ_ℓ is the same as finding the first order condition for S_ℓ.

$$\nabla_{S_\ell}\ \ell_{\text{semi-sup}}(\Theta) = \nabla_{S_\ell}\ \sum_{i=1}^{n}\sum_{j=1}^{k} w_j^{(i)} \log\left(\frac{p(x^{(i)}, z^{(i)}; \Theta)}{w_j^{(i)}}\right) + \alpha \sum_{i=1}^{\tilde{n}} \log\Big(p(\tilde{x}^{(i)}, \tilde{z}^{(i)}; \Theta)\Big) \tag{4.24}$$

$$= \nabla_{S_\ell}\ \sum_{i=1}^{n} w_\ell^{(i)} \log\left(\frac{p(x^{(i)}, z^{(i)}; \Theta)}{w_\ell^{(i)}}\right) + \alpha \sum_{i=1}^{\tilde{n}} \mathbb{1}\{\tilde{z}^{(i)} = \ell\} \log\Big(p(\tilde{x}^{(i)}, \tilde{z}^{(i)}; \Theta)\Big) \tag{4.25}$$

$$= \nabla_{S_\ell} \sum_{i=1}^{n} w_\ell^{(i)} \left(\log(|\Sigma_\ell|^{-1}) - \frac{1}{2}(x^{(i)} - \mu_\ell)^T S_\ell (x^{(i)} - \mu_\ell)\right) \tag{4.26}$$

$$+ \alpha \sum_{i=1}^{\tilde{n}} \mathbb{1}\{\tilde{z}^{(i)} = \ell\} \left(\log(|\Sigma_\ell|^{-1}) - \frac{1}{2}(\tilde{x}^{(i)} - \mu_\ell)^T S_\ell (\tilde{x}^{(i)} - \mu_\ell)\right) \tag{4.27}$$

$$= \nabla_{S_\ell} \sum_{i=1}^{n} w_\ell^{(i)} \left(\log(|S_\ell|) - \frac{1}{2}(x^{(i)} - \mu_\ell)^T S_\ell (x^{(i)} - \mu_\ell)\right) \tag{4.28}$$

$$+ \alpha \sum_{i=1}^{\tilde{n}} \mathbb{1}\{\tilde{z}^{(i)} = \ell\} \left(\log(|S_\ell|) - \frac{1}{2}(\tilde{x}^{(i)} - \mu_\ell)^T S_\ell (\tilde{x}^{(i)} - \mu_\ell)\right) \tag{4.29}$$

$$= \sum_{i=1}^{n} w_\ell^{(i)} S_\ell^{-T} - w_\ell^{(i)} (x^{(i)} - \mu_\ell)(x^{(i)} - \mu_\ell)^T + \alpha \sum_{i=1}^{\tilde{n}} \mathbb{1}\{\tilde{z}^{(i)} = \ell\} \left(S_\ell^{-T} - (\tilde{x}^{(i)} - u_\ell)(\tilde{x}^{(i)} - u_\ell)^T\right) \tag{4.30}$$

$$= 0 \tag{4.31}$$

Since Σ_ℓ is symmetric, so $S_\ell^{-T} = \Sigma_\ell$. Above first order condition implies

$$\Sigma_\ell \left(\sum_{i=1}^{n} w_\ell^{(i)} + \alpha \sum_{i=1}^{\tilde{n}} \mathbb{1}\{\tilde{z}^{(i)} = \ell\}\right) = \sum_{i=1}^{n} w_\ell^{(i)} (x^{(i)} - \mu_\ell)(x^{(i)} - \mu_\ell)^T + \alpha \sum_{i=1}^{\tilde{n}} \mathbb{1}\{\tilde{z}^{(i)} = \ell\} (\tilde{x}^{(i)} - u_\ell)(\tilde{x}^{(i)} - u_\ell)^T \tag{4.32}$$

$$\implies \Sigma_\ell = \frac{\sum_{i=1}^{n} w_\ell^{(i)} (x^{(i)} - \mu_\ell)(x^{(i)} - \mu_\ell)^T + \alpha \sum_{i=1}^{\tilde{n}} \mathbb{1}\{\tilde{z}^{(i)} = \ell\} (\tilde{x}^{(i)} - u_\ell)(\tilde{x}^{(i)} - u_\ell)^T}{\sum_{i=1}^{n} w_\ell^{(i)} + \alpha \sum_{i=1}^{\tilde{n}} \mathbb{1}\{\tilde{z}^{(i)} = \ell\}} \tag{4.33}$$

■

4.3.3 Choosing ϕ_ℓ

Proof.

$$\nabla_{\phi_\ell} \ell_{\text{semi-sup}}(\Theta) = \nabla_{\phi_\ell} \sum_{i=1}^{n} \sum_{\ell=1}^{k} w_\ell^{(i)} \log \left(\frac{p(x^{(i)}|z^{(i)} = \ell; \Theta) p(z^{(i)} = \ell; \Theta)}{p(x^{(i)}|z^{(i)} = \ell; \Theta)} \right) + \alpha \sum_{i=1}^{\tilde{n}} \log \Big(p(\tilde{x}^{(i)}|\tilde{z}^{(i)}; \Theta) p(\tilde{z}; \Theta) \Big) \tag{4.34}$$

$$= \nabla_{\phi_\ell} \sum_{i=1}^{n} \sum_{\ell=1}^{k} w_\ell^{(i)} \log \Big(p(z^{(i)} = \ell; \Theta) \Big) + \alpha \sum_{i=1}^{\tilde{n}} \mathbb{1}\{\tilde{z}^{(i)} = \ell\} \log(\phi_\ell) \tag{4.35}$$

$$= \nabla_{\phi_\ell} \sum_{i=1}^{n} \sum_{\ell=1}^{k} w_\ell^{(i)} \log(\phi_\ell) + \alpha \sum_{i=1}^{\tilde{n}} \mathbb{1}\{\tilde{z}^{(i)} = \ell\} \log(\phi_\ell) \tag{4.36}$$

$$= \nabla_{\phi_\ell} \sum_{i=1}^{n} w_\ell^{(i)} \log(\phi_\ell) + \alpha \sum_{i=1}^{\tilde{n}} \mathbb{1}\{\tilde{z}^{(i)} = \ell\} \log(\phi_\ell) \tag{4.37}$$

Using the constraint that $\sum_{\ell=1}^{k} \phi_\ell = 1$, the Lagrangian can be constructed as

$$\mathcal{L}(\cdot) = \ell_{\text{semi-sup}}(\phi, \cdot) + \lambda \left(1 - \sum_{\ell=1}^{k} \phi_\ell \right) \tag{4.38}$$

Solving the stationary point for $\mathcal{L}(\cdot)$ gives

$$\frac{\partial \mathcal{L}(\phi_\ell, \cdot)}{\partial \phi_\ell} = \nabla_{\phi_\ell} \ell_{\text{semi-sup}} - \lambda \tag{4.39}$$

$$= \frac{1}{\phi_\ell} \left(\sum_{i=1}^{n} w_\ell^{(i)} + \alpha \sum_{i=1}^{\tilde{n}} \mathbb{1}\{\tilde{z}^{(i)} = \ell\} \right) - \lambda = 0 \tag{4.40}$$

$$\implies \frac{1}{\lambda} \left(\sum_{i=1}^{n} w_\ell^{(i)} + \alpha \sum_{i=1}^{\tilde{n}} \mathbb{1}\{\tilde{z}^{(i)} = \ell\} \right) = \phi_\ell \tag{4.41}$$

Then, $\sum_{\ell=1}^{k} \phi_\ell = 1$ implies

$$\sum_{\ell=1}^{k} \frac{1}{\lambda} \left(\sum_{i=1}^{n} w_\ell^{(i)} + \alpha \sum_{i=1}^{\tilde{n}} \mathbb{1}\{\tilde{z}^{(i)} = \ell\} \right) = 1 \tag{4.42}$$

$$\implies \sum_{\ell=1}^{k} \left(\sum_{i=1}^{n} w_\ell^{(i)} + \alpha \sum_{i=1}^{\tilde{n}} \mathbb{1}\{\tilde{z}^{(i)} = \ell\} \right) = \lambda \tag{4.43}$$

Therefore,

$$\phi_\ell = \frac{\sum_{i=1}^{n} w_\ell^{(i)} + \alpha \sum_{i=1}^{\tilde{n}} \mathbb{1}\{\tilde{z}^{(i)} = \ell\}}{\sum_{j=1}^{k} \left(\sum_{i=1}^{n} w_j^{(i)} + \alpha \sum_{i=1}^{\tilde{n}} \mathbb{1}\{\tilde{z}^{(i)} = j\} \right)} \tag{4.44}$$

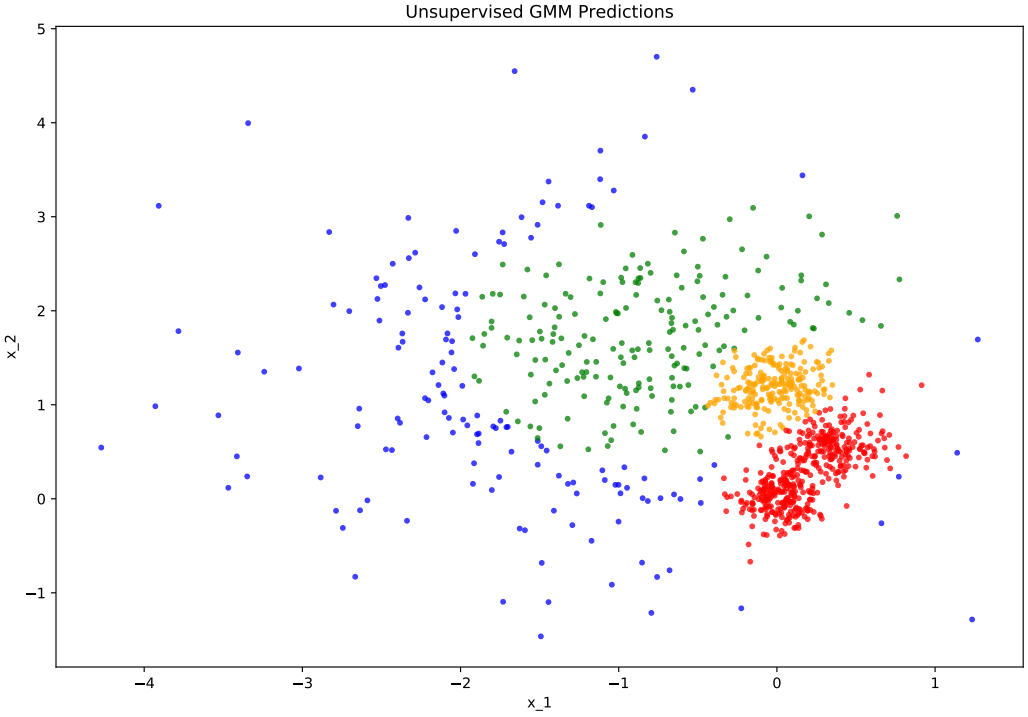
$$= \frac{\sum_{i=1}^{n} w_\ell^{(i)} + \alpha \sum_{i=1}^{\tilde{n}} \mathbb{1}\{\tilde{z}^{(i)} = \ell\}}{\sum_{j=1}^{k} \sum_{i=1}^{n} w_j^{(i)} + \alpha \sum_{j=1}^{k} \sum_{i=1}^{\tilde{n}} \mathbb{1}\{\tilde{z}^{(i)} = j\}} \tag{4.45}$$

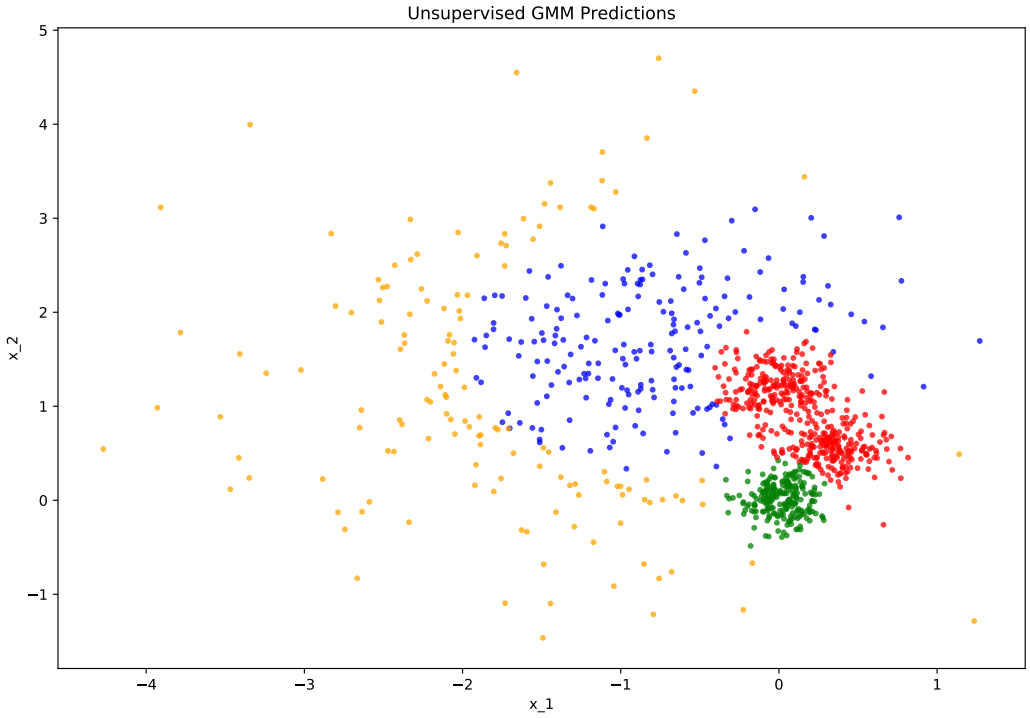
$$= \frac{\sum_{i=1}^{n} w_\ell^{(i)} + \alpha \sum_{i=1}^{\tilde{n}} \mathbb{1}\{\tilde{z}^{(i)} = \ell\}}{\sum_{i=1}^{n} \sum_{j=1}^{k} w_j^{(i)} + \alpha \sum_{i=1}^{\tilde{n}} \sum_{j=1}^{k} \mathbb{1}\{\tilde{z}^{(i)} = j\}} \tag{4.46}$$

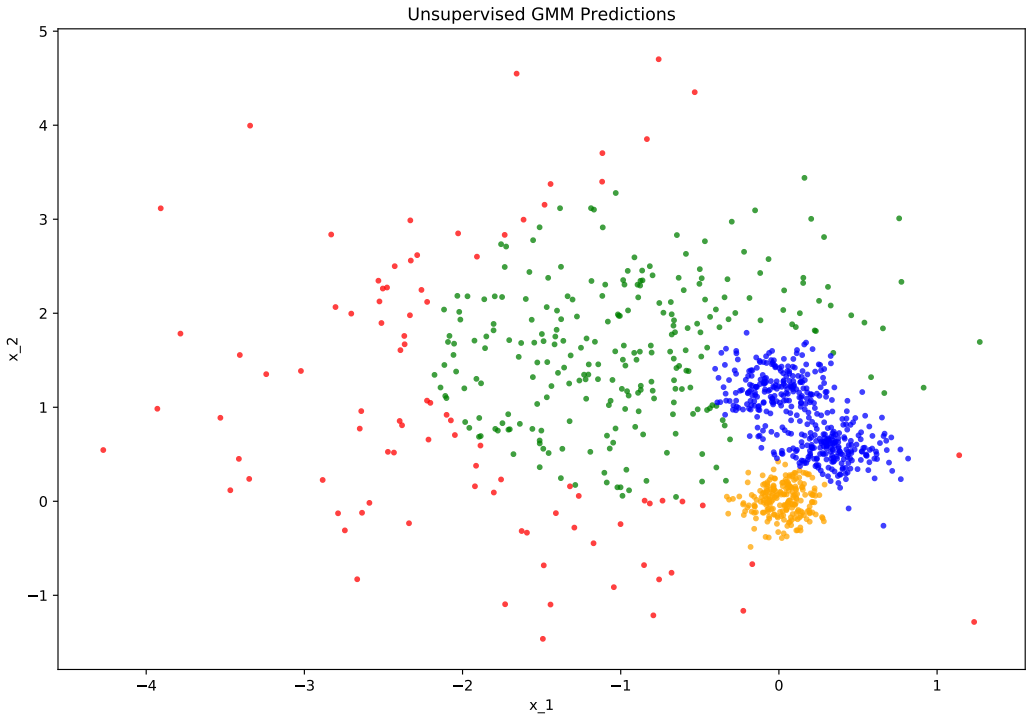
$$= \frac{\sum_{i=1}^{n} w_\ell^{(i)} + \alpha \sum_{i=1}^{\tilde{n}} \mathbb{1}\{\tilde{z}^{(i)} = \ell\}}{n + \alpha \tilde{n}} \tag{4.47}$$

■

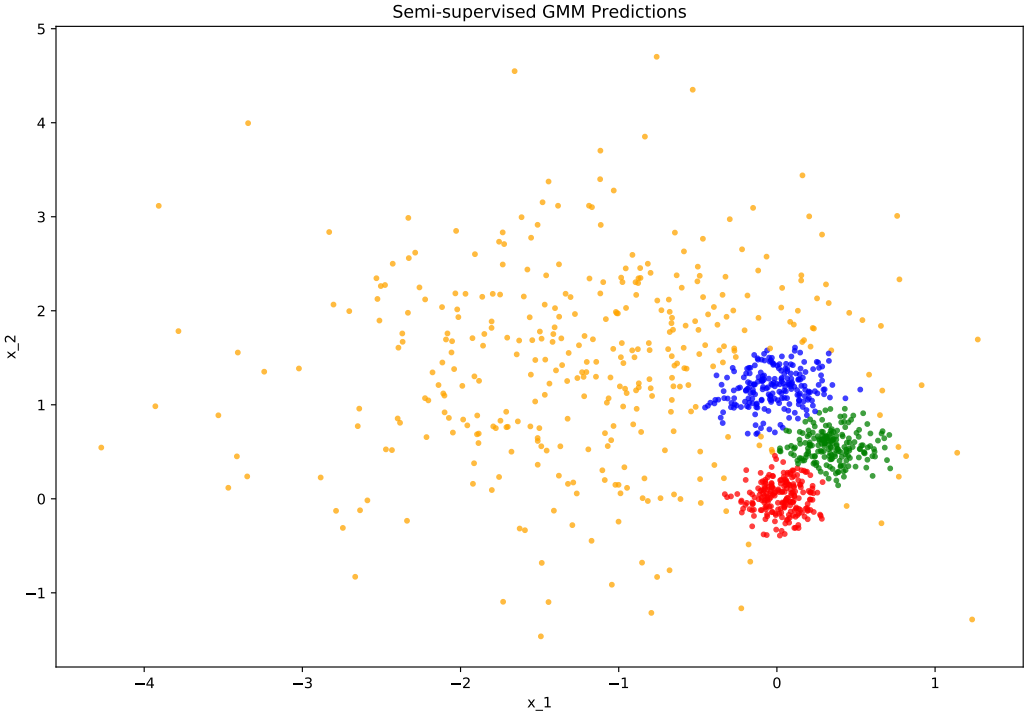
4.4 (d) Classical (Unsupervised) EM Implememtation

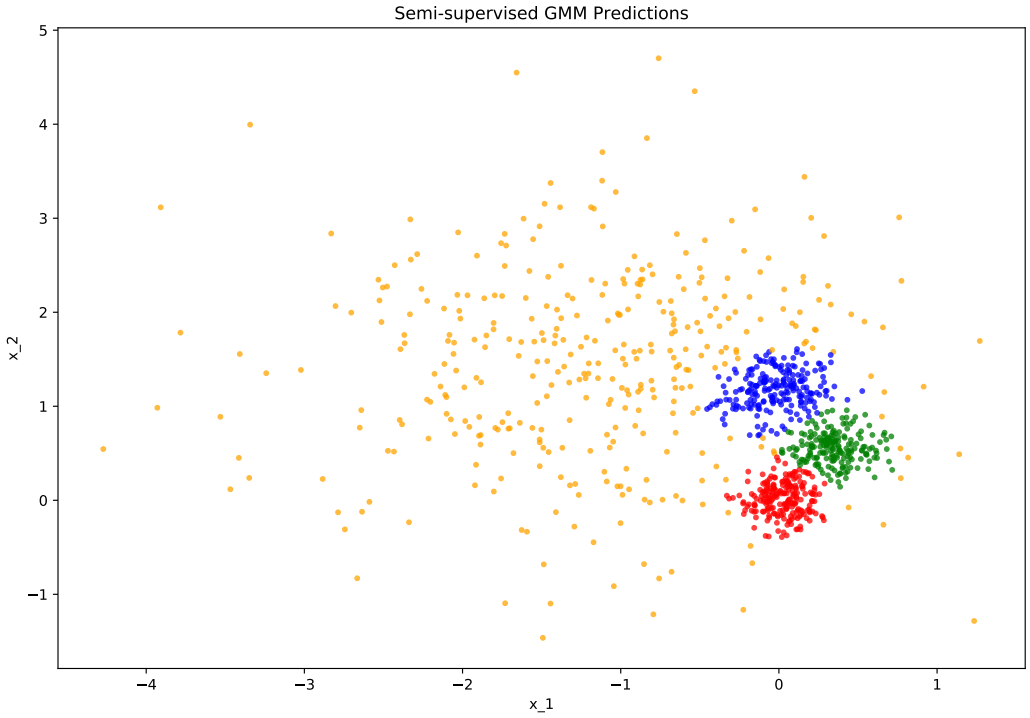


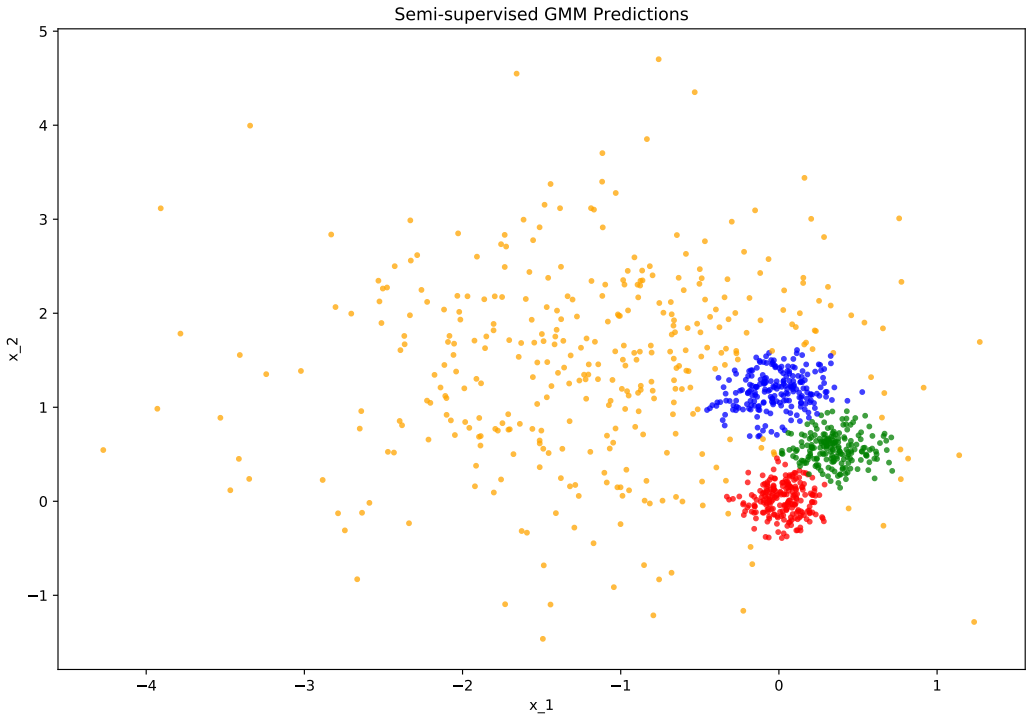

Unsupervised GMM Predictions
x_2
x_1
5
4
3
2
1
0
−1
−4
−3
−2
−1
0
1


Unsupervised GMM Predictions
x_2
x_1
5
4
3
2
1
0
−1
−4
−3
−2
−1
0
1

4.5 (e) Semi-supervised EM Implementation




Semi-supervised GMM Predictions
x_2
x_1
5
4
3
2
1
0
−1
−4
−3
−2
−1
0
1


Semi-supervised GMM Predictions
x_2
x_1
5
4
3
2
1
0
−1
−4
−3
−2
−1
0
1

4.6 (f) Comparison

Answer

1. The unsupervised GMM took more iterations (more than 100 iterations on average) to converge than the semi-supervised GMM took (around 25 iterations on average).

2. Semi-supervised GMM was more stable than the unsupervised version, clusters were almost exactly (by looking at the clustering result, there were still couple of points changed) in all three experiments conducted. However, the clustering result for unsupervised GMM varied among three experiments.

3. Given that the dataset was sampled from a mixture of three low-variance Gaussian distributions with one additional large variance Gaussian, the semi-supervised GMM successfully separated the the entire dataset into three low-variance groups and a fourth high-variance group. However, the unsupervised GMM clustered the dataset into two low-variance and two high-variance groups instead. Overall, the semi-supervised GMM provided higher quality results.

5 PCA

Proof. Note that given a subspace $\mathcal{V} = \{\alpha u : \alpha \in \mathbb{R}\} \subset \mathbb{R}^d$ characterized by some unit vector u. Let $x \in \mathbb{R}^d$, $f_u(x)$ is defined to be the a vector in $\mathcal{V}$ closest to x. Indeed, $f_u(x)$ would be exactly the *projection* of x on u, which is given by

$$f_u(x) = \frac{\langle x, u\rangle}{||u||_2^2} u = \langle x, u\rangle u \tag{5.1}$$

because u is a unit vector. Therefore,

$$\begin{aligned}
\underset{u:||u||_2^2=1}{\operatorname{argmin}} \sum_{i=1}^{n} ||x^{(i)} - f_u(x^{(i)})||_2^2 &= \underset{u:||u||=1}{\operatorname{argmin}} \sum_{i=1}^{n} ||x^{(i)} - \langle x^{(i)}, u\rangle u||_2^2 && (5.2)\\
&= \underset{u:||u||_2^2=1}{\operatorname{argmin}} \sum_{i=1}^{n} \left\langle x^{(i)} - \langle x^{(i)}, u\rangle u, x^{(i)} - \langle x^{(i)}, u\rangle u \right\rangle && (5.3)\\
&= \underset{u:||u||_2^2=1}{\operatorname{argmin}} \sum_{i=1}^{n} \left\langle x^{(i)}, x^{(i)} - \langle x^{(i)}, u\rangle u \right\rangle - \left\langle \langle x^{(i)}, u\rangle u, x^{(i)} - \langle x^{(i)}, u\rangle u \right\rangle && (5.4)\\
&= \underset{u:||u||_2^2=1}{\operatorname{argmin}} \sum_{i=1}^{n} \langle x^{(i)}, x^{(i)}\rangle - 2\left\langle x^{(i)}, \langle x^{(i)}, u\rangle u \right\rangle + \left\langle \langle x^{(i)}, u\rangle u, \langle x^{(i)}, u\rangle u \right\rangle && (5.5)\\
&= \underset{u:||u||_2^2=1}{\operatorname{argmin}} \sum_{i=1}^{n} \underbrace{||x^{(i)}||_2^2}_{\perp u} - 2\langle x^{(i)}, u\rangle\langle x^{(i)}, u\rangle + \langle x^{(i)}, u\rangle^2 ||u||_2^2 && (5.6)\\
&= \underset{u:||u||_2^2=1}{\operatorname{argmin}} \sum_{i=1}^{n} -2\langle x^{(i)}, u\rangle^2 + \langle x^{(i)}, u\rangle^2 && (5.7)\\
&= \underset{u:||u||_2^2=1}{\operatorname{argmax}} \sum_{i=1}^{n} \langle x^{(i)}, u\rangle^2 && (5.8)\\
&= \underset{u:||u||_2^2=1}{\operatorname{argmax}} \sum_{i=1}^{n} u^T x^{(i)} x^{(i)T} u && (5.9)\\
&= \underset{u:||u||_2^2=1}{\operatorname{argmax}}\, u^T \left(\sum_{i=1}^{n} x^{(i)} x^{(i)T} \right) u && (5.10)\\
&= \underset{u:||u||_2^2=1}{\operatorname{argmax}}\, u^T \left(\frac{1}{n} \sum_{i=1}^{n} x^{(i)} x^{(i)T} \right) u && (5.11)
\end{aligned}$$

where the last maximization is exactly the same as the maximization for choosing principal component on page 5 of the lecture notes for PCA. ■

6 Problem 6: Independent Components Analysis

6.1 (a) Gaussian source

Proof. Let $\phi(z)$ denote the PDF for $\mathcal{N}(0,1)$.

$$\nabla_W \ell(W) = \nabla_W \sum_{i=1}^{n} \left(\log |W| + \sum_{j=1}^{d} \log \phi(w_j^T x^{(i)}) \right) \tag{6.1}$$

$$= \sum_{i=1}^{n} \left(W^{-T} + \nabla_W \sum_{j=1}^{d} \log \phi(w_j^T x^{(i)}) \right) \tag{6.2}$$

$$= \sum_{i=1}^{n} \left(W^{-T} + \nabla_W \sum_{j=1}^{d} \left(-\frac{1}{2}(w_j^T x^{(i)})^2 \right) \right) \tag{6.3}$$

where the last step was derived by explicitly expanding $\phi(z)$ and dropping terms independent from W. Note that, for each w_j,

$$\nabla_{w_j} \sum_{i=1}^{n} \sum_{j=1}^{d} \left(-\frac{1}{2}(w_j^T x^{(i)})^2 \right) = -(w_j^T x^{(i)}) x^{(i)T} \tag{6.4}$$

Gather results,

$$\nabla_W \ell(W) = nW^{-T} - \underbrace{\sum_{i=1}^{n} \begin{pmatrix} (w_1^T x^{(i)}) x^{(i)T} \\ (w_2^T x^{(i)}) x^{(i)T} \\ \vdots \\ (w_j^T x^{(i)}) x^{(i)T} \end{pmatrix}}_{:=\Omega \in \mathbb{R}^{d \times d}} \tag{6.5}$$

Claim: $\Omega = WX^TX$. This can be easily verified by considering the (a,b) component of Ω. $\Omega_{a,b} = \sum_{i=1}^{n} (w_a^T x^{(i)}) x_b^{(i)} = \sum_{i=1}^{n} \sum_{j=1}^{d} W_{a,j} x_j^{(i)} x_b^{(i)}$, which is exactly the same as the (a,b) component of WX^TX. Setting the gradient equal zero for the first order condition:

$$\nabla_W \ell(W) = nW^{-T} - WX^TX = 0 \tag{6.6}$$

$$\implies W^TW = n\left(X^TX\right)^{-1} \tag{6.7}$$

Let $P \in \mathbb{R}^{d \times d}$ be a rotation matrix, and note that the transpose and inverse of rotation matrices are the same. Suppose W^* solves above first order condition, then PW^* also solves the first order

condition:

$$(PW^*)^T(PW^*) = W^{*T}P^TPW^* \tag{6.8}$$
$$= W^{*T}IW^* \tag{6.9}$$
$$= W^{*T}W^* \tag{6.10}$$
$$= n\left(X^TX\right)^{-1} \tag{6.11}$$

That is, any valid unmixing matrix is still valid after arbitrary rotations, as result, one cannot distinguish the actual unmixing matrix among infinitely many rotational copies of it. ■

6.2 (b) Laplace Source

Proof. For training instance $x^{(i)}$:

$$\nabla_W \ell(W) = \nabla_W \log|W| + \nabla_W \sum_{j=1}^{d} \underbrace{\log(\frac{1}{2})}_{\perp W} - |w_j^T x^{(i)}| \tag{6.12}$$

$$= W^{-T} - \begin{pmatrix} \text{sign}(w_1^T x^{(i)}) x^{(i)T} \\ \text{sign}(w_2^T x^{(i)}) x^{(i)T} \\ \vdots \\ \text{sign}(w_d^T x^{(i)}) x^{(i)T} \end{pmatrix} \tag{6.13}$$

$$= W^{-T} - \begin{pmatrix} \text{sign}(w_1^T x^{(i)}) \\ \text{sign}(w_2^T x^{(i)}) \\ \vdots \\ \text{sign}(w_d^T x^{(i)}) \end{pmatrix} x^{(i)T} \tag{6.14}$$

Therefore, the stochastic gradient ascending for likelihood can be written as

$$W := W + \alpha \left(W^{-T} - \begin{pmatrix} \text{sign}(w_1^T x^{(i)}) \\ \text{sign}(w_2^T x^{(i)}) \\ \vdots \\ \text{sign}(w_d^T x^{(i)}) \end{pmatrix} x^{(i)T} \right) \tag{6.15}$$

■

6.3 (c) Cocktail Party Problem

Answer The unmixing matrix W:

```
[[ 52.81523328  16.78831566  19.94349962 -10.1923758  -20.88602262]
 [ -9.91331065  -0.9700641   -4.66262202   8.0267521    1.77817318]
 [  8.31337278  -7.46981144  19.31167865  15.17714755 -14.324928  ]
 [-14.66345468 -26.63906716   2.44484521  21.37869701  -8.41625584]
 [ -0.27410635  18.37656455   9.31495731   9.10572431  30.59639414]]
```