Homework Set 3 of CS229 class.

3 a) Done.

b) [redacted] Previously each pixel was represented by $8 \times 3 = 24$ bits. Now, after compression, only 4 bits are required. This gives a compression ratio of $24/4 = \underline{6}$.

4 a) Regarding the derivation of the E-M steps:

E-step is obviously right per definition: we're just estimating the pdf of z, to softly classify the datapoints.

M-step: $\log p(x^i, z^i; \theta) = \log Q(z^i) \frac{p(x^i, z^i; \theta)}{Q(z^i)} \geq Q(z^i) \log \frac{p(x^i, z^i; \theta)}{Q(z^i)}$
per Jensen's inequality, since "log" is a concave function. Continuing, maximizing l_{sup} is straightforward relative to l_{unsup}, so for l_{sup} we don't need to bother with any lower-bounds.

$$\Rightarrow \theta^{t+1} := \underset{\theta}{\operatorname{argmax}} \left[\sum_{i=1}^{n} \left(\sum_{z^i} Q_i(z^i; \theta) \log \frac{p(x^i, z^i; \theta)}{Q_i(z^i; \theta)} \right) + \alpha \sum_{i=1}^{\tilde{n}} \log p(\tilde{x}^i, \tilde{z}^i; \theta) \right]$$

And then to the E-step again, and back...

Have this after $t+1$ E-steps, t M-steps.

4a2) $\ell_{semisup}(\theta^{t+1}) = \sum_{i=1}^{n} \log \sum_{z^i} p(x^i, z^i; \theta^{t+1}) + \ell_{sup}(\theta^{t+1})$

$= \sum_{i=1}^{n} \log \sum_{z^i} Q^{t+1}(z^i) \frac{p(x^i, z^i; \theta^{t+1})}{Q^{t+1}(z^i)} + \ell_{sup}(\theta^{t+1})$

← Jensen's + log is concave.

$\geq \sum_{i=1}^{n} \sum_{z^i} Q^{t+1}(z^i) \log \frac{p(x^i, z^i; \theta^{t+1})}{Q^{t+1}(z^i)} + \ell_{sup}(\theta^{t+1})$

θ^{t+1} comes from argmaxing $\ell(\theta^t)$ wrt θ^{t+1}.

$\geq \sum_{i=1}^{n} \sum_{z^i} Q^{t}(z^i) \log \frac{p(x^i, z^i; \theta^{t})}{Q^{t}(z^i)} + \ell_{sup}(\theta^{t})$

Since $Q^t(z^i)$ was set to $p(z^i | x^i; \theta^t)$,

$= \sum_{i=1}^{n} \log \sum_{z^i} p(x^i, z^i; \theta^t) + \ell_{sup}(\theta^t) = \ell_{semisup}(\theta^t)$. qed.

making $\sum_{z^i} Q^t(z^i) \log \frac{p(x^i, z^i; \theta^t)}{Q^t(z^i)} = \log p(x^i; \theta^t) = \log \sum_{z^i} p(x^i, z^i; \theta^t)$

4b) Clearly, all z^i for $i = \{1, .., n\}$ (unlabelled cases) need to be estimated. The $\tilde{z}^i$'s are all known.

$Q_i^t(z^i) := p(z^i | x^i; \theta^t) = \frac{p(x^i, z^i; \theta^t)}{p(x^i; \theta^t)} = \frac{\overbrace{p(x^i | z^i; \theta^t)}^{\text{Gaussian}} \overbrace{p(z^i; \theta^t)}^{\text{Multinomial}}}{\sum_{z^i} p(x^i | z^i; \theta^t) p(z^i; \theta^t)}$

$\Leftrightarrow$

$p(z^i = j | x^i; \theta^t) = \frac{p(x^i | z^i = j; \theta^t) p(z^i = j; \phi)}{\sum_{j=1}^{k} p(x^i | z^i = j; \theta^t) p(z^i = j; \phi)} = \frac{e^{-\frac{1}{2}(x - \mu_j)^T \Sigma_j^{-1} (x - \mu_j)} \cdot \phi_j / (\sqrt{2\pi}^d |\Sigma_j|^{1/2})}{\sum_{i=1}^{k} e^{-\frac{1}{2}(x - \mu_i)^T \Sigma_i^{-1} (x - \mu_i)} \cdot \phi_i / (\sqrt{2\pi}^d |\Sigma_i|^{1/2})}$

dxd = k, dxk, kx1

c) $\theta^{t+1} = \{\mu^{t+1}, \Sigma^{t+1}, \phi^{t+1}\}$ needs to be re-estimated, while Q^t is kept constant (since otherwise we'd be [illegible] directly, not iteratively, finding the maximum of $\ell_{semi-sup}$).

$\Rightarrow \nabla_\mu \ell(\theta) = \nabla_\mu \left[\sum_{i=1}^{n} \sum_{z^i} Q_i^t \log \frac{p(x^i, z^i; \theta)}{Q_i^t} + \alpha \sum_{i=n}^{n+\tilde{n}} \log (p(x^i, z^i; \theta)) \right]$

$= \nabla_\mu \left[\sum_{i=1}^{n} \sum_{z^i} Q_i^t \log p(x^i | z^i; \theta) p(z^i; \theta) + \alpha \sum_{i=n}^{n+\tilde{n}} \log \left(p(x^i | z^i; \theta) p(z^i; \theta) \right) \right]$

$= Q$

$$= \nabla_\mu \left[\sum_{i=1}^{n} \sum_{z^i} Q_i^t \log \frac{\ell^{-\frac{1}{2}(x^i-\mu_{z^i})^T \Sigma_{z^i}^{-1}(x^i-\mu_{z^i})}}{\sqrt{2\pi}^d |\Sigma_{z^i}|^{1/2}} + \alpha \sum_{i=n}^{\tilde{n}+n} \log \frac{\ell^{-\frac{1}{2}(x^i-\mu_{\tilde{z}^i})^T \tilde{\Sigma}_{\tilde{z}^i}^{-1}(x^i-\mu_{\tilde{z}^i})}}{\sqrt{2\pi}^d |\tilde{\Sigma}_{\tilde{z}^i}|^{1/2}} p(\tilde{z}^i;\theta) \right]$$

$\delta_{\tilde{z}^i,\cdot}$

$$\Leftrightarrow \quad 0 = \nabla_\mu \left[\sum_{i=1}^{n} \sum_{z^i} Q_i^t (x^i-\mu_{z^i})^T \Sigma_{z^i}^{-1} (x^i-\mu_{z^i}) + \alpha \sum_{i=n}^{n+\tilde{n}} \overbrace{p(\tilde{z}^i;\theta)}^{\text{delta-distrib}} \cdot (x^i-\mu_{\tilde{z}^i})^T \tilde{\Sigma}_{\tilde{z}^i}^{-1} (x^i-\mu_{\tilde{z}^i}) \right]$$

$$\Leftrightarrow \quad 0 = \nabla_{\mu_\ell} \left[\sum_{i=1}^{n} Q_i^t(z^i=\ell;\theta^t) \cdot (x^i-\mu_\ell)^T \Sigma_\ell^{-1} (x^i-\mu_\ell) + \alpha \sum_{i=n}^{n+\tilde{n}} \delta_{\tilde{z}^i=\ell} (x^i-\mu_\ell)^T \tilde{\Sigma}_\ell^{-1} (x^i-\mu_\ell) \right]$$

$$\Leftrightarrow \quad 0 = \left[\sum_{i=1}^{n} Q_i^t(z^i=\ell) \Sigma_\ell^{-1} (x^i-\mu_\ell) + \alpha \sum_{i=n}^{n+\tilde{n}} \delta(\tilde{z}^i=\ell) \tilde{\Sigma}_\ell^{-1} (x^i-\mu_\ell) \right]$$

(left-multiply w/ Σ_ℓ)

$$\Leftrightarrow \quad \mu_\ell \left[\sum_{i=1}^{n} Q_i^t(z^i=\ell) + \alpha \sum_{i=1}^{\tilde{n}} \delta(\tilde{z}^i=\ell) \right] = \sum_{i=1}^{n} Q^t(z^i=\ell) x^i + \sum_{i=1}^{n+\tilde{n}} \delta(\tilde{z}^i=\ell) x^i$$

$$\Leftrightarrow \quad \underline{\mu_\ell = \frac{\sum_{i=1}^{n} Q^t(z^i=\ell) x^i + \sum_{i=n}^{n+\tilde{n}} \delta(\tilde{z}^i=\ell) x^i}{\sum_{i=1}^{n} Q_i^t(z^i=\ell) + \sum_{i=n}^{n+\tilde{n}} \alpha \delta(\tilde{z}^i=\ell)}} = \frac{\sum_{i=1}^{n+\tilde{n}} w^t(z^i=\ell) x^i}{\sum_{i=1}^{n+\tilde{n}} w^t(z^i=\ell)}$$

As for Σ: $\nabla_{\Sigma_\ell} \left[\sum_{i=1}^{n} \frac{Q_i^t(z^i=\ell;\theta^t)(x^i-\mu_\ell)^T \Sigma_\ell (x^i-\mu_\ell) + \alpha \sum_{i=1}^{\tilde{n}} \delta(\tilde{z}^i=\ell)(x^i-\mu_\ell)^T \tilde{\Sigma}_\ell (x^i-\mu_\ell)}{|\Sigma_\ell|^{1/2}} \right]$

with $w^t(z^i=\ell) = \begin{cases} Q^t(z^i=\ell), & i \le n \\ \alpha \delta(\tilde{z}^i=\ell), & i > \tilde{n} \end{cases}$

As for Σ: $\nabla_\Sigma \left[\sum_{i=1}^{n+\tilde{n}} \sum_{z^i} w_i^t \log \frac{\ell^{-1/2}(x^i-\mu_{z^i})^T \Sigma_{z^i}^{-1}(x^i-\mu_{z^i})}{\sqrt{2\pi}^d |\Sigma_{z^i}|^{1/2}} \right] = 0$

$\frac{-1}{2} \log |\Sigma_{z^i}|$

$$\Leftrightarrow \quad \nabla_{\Sigma_\ell} \left[\sum_{i=1}^{n+\tilde{n}} w_i^t(z^i=\ell) \left(\frac{(x^i-\mu_\ell)^T \Sigma_\ell^{-1}(x^i-\mu_\ell)}{} - \log |\Sigma_\ell^{-1}| \right) \right] = 0$$

$$= \sum_{i=1}^{n+\tilde{n}} w_i^t(z^i=\ell) \left(-\Sigma_\ell^{-1} (x^i-\mu_\ell)(x^i-\mu_\ell)^T \Sigma_\ell^{-1} + \frac{|\Sigma_\ell| \Sigma_\ell^{-1}}{|\Sigma_\ell|} \right)$$

$$= \sum_{i=1}^{n+\tilde{n}} w^t(z^i=\ell)\left(-(x^i-\mu_\ell)(x^i-\mu_\ell)^T + \Sigma_\ell\right)$$

$$\Leftrightarrow \underline{\Sigma_\ell = \frac{\sum_{i=1}^{n+\tilde{n}} w^t(z^i=\ell)(x^i-\mu_\ell)(x^i-\mu_\ell)^T}{\sum_{i=1}^{n+\tilde{n}} w^t(z^i=\ell)}}$$

Finally, for ϕ: Here we need to add a Lagrange multiplier to $\ell(\theta)$ that enforces $\sum_{i=1}^{k}\phi_i = 1$.

$$\Rightarrow \nabla_\phi\left[\ell(\theta) + \lambda\left(\sum_{i=1}^{k}\phi_i - 1\right)\right] = 0 \Leftrightarrow \nabla_{\phi_\ell}\left[\sum_{i=1}^{n+\tilde{n}} w^t(z^i=\ell)\log\left(p(x^i|z^i=\ell;\theta)\right)\right]$$

$$\#\theta - \lambda\left(\sum_{i=1}^{k}\phi_i - 1\right)\Big] = 0$$

$$\Leftrightarrow \sum_{i=1}^{n+\tilde{n}} w^t(z^i=\ell)\frac{1}{\phi_\ell} = \lambda \Leftrightarrow \underline{\frac{1}{\lambda}\sum_{i=1}^{n+\tilde{n}} w^t(z^i=\ell) = \phi_\ell}. \text{ Now,}$$

$$\sum_{\ell=1}^{k}\phi_\ell = 1 = \frac{1}{\lambda}\sum_{i=1}^{n+\tilde{n}}\sum_{\ell=1}^{k} w^t(z^i=\ell) \Rightarrow \underline{\lambda = \sum_i\sum_\ell w^t(z^i=\ell)}$$

$$\Rightarrow \underline{\phi_\ell = \frac{\sum_{i=1}^{n+\tilde{n}} w^t(z^i=\ell)}{\sum_{\delta=1}^{n+\tilde{n}}\sum_{\ell=1}^{k} w^t(z^i=\ell)}} = \underline{\frac{\sum_{i=1}^{n+\tilde{n}} w^t(z^i=\ell)}{n+\alpha\tilde{n}}}$$

Note that w_i^t is a conditional probability $p(z^i|x^i;\theta^t)$ that is computed during E-Step for $i=1,2,\ldots,n$, and set to some label-distribution

4d) Done.

4e) Done.

4f)

i) Convergence is much faster for the semi-supervised one.

ii) Stability is present for semi sup., the unsup. one is ~~unstable~~ rather unstable.

iii) Overall quality is much better in the semi-supervised case.

5) The v which minimizes $\|x - v\|^2$ (given $v = \alpha u$) is $f_u(x) = \frac{x^T u}{\underbrace{u^T u}_{=1}} u = \underline{\underline{(x^T u)\, u}}$.

Plugging this result into the latter cost function gives:

$$L = \left| x^i - \underbrace{(u^T x^i)}_{= u(u^T x^i)} u \right|_2^2 = ((\mathbb{1} - uu^T)x^i)^T (\mathbb{1} - uu^T) x^i = x^{iT} (\mathbb{1} - uu^T)^2 x^i$$

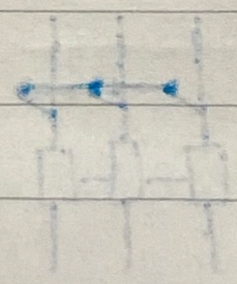
$$= x^{iT} (\mathbb{1} - uu^T) x^i = x^{iT} x^i - u^T x^i x^{iT} u = \underbrace{tr \Sigma'}_{\text{sum of } X \text{ from } \Sigma' = x^T x} - u^T \Sigma' u$$

$$= tr\left(\underbrace{VS^2V^T}\right) - u^T V S^2 V^T u = tr\, S^2 - u^T [S_1 V_1, S_2 V_2 \dots][S_1 V_1, S_2 V_2 \dots]^T u$$

$$= \underbrace{tr\, S^2}_{S_1^2 + S_2^2 + \dots,\ S_1 \geq S_2 \geq \dots} - \left(S_1^2 (u^T V_1)^2 + S_2^2 (u^T V_2)^2 + \dots\right) = S_2^2 + S_3^2 + \dots$$

because the minimum of L is found when $u^T V_1 = 1 \Rightarrow \underline{\underline{u = V_1}}$.

Therefore, the first PC of the data corresponds to the axis of maximum variation.



Homework Problem 6

a) If $g'(s) = \frac{e^{-s^2/2}}{\sqrt{2\pi}}$, then $\log g' = -s^2/2 + \dots$, which in turn implies

$$\ell(W) = \sum_{i=1}^{n}\left(\log|W| - \frac{1}{2}\sum_{j=1}^{d}(w_j^T x^i)^2\right) = n\log|W| - \frac{1}{2}\sum_{i=1}^{n} x^{i^T} W^T W x^i$$

$$= n\log|W| - \frac{1}{2}\mathrm{tr}(\underset{d\times m}{X^T} W^T W \underset{m\times n}{X}).$$

$$\Rightarrow \frac{\partial \ell}{\partial W} = \frac{n}{|W|}|W|W^{T^{-1}} - \frac{1}{2}2WXX^T = nW^{T^{-1}} - WXX^T = 0$$

$\Rightarrow n\mathbb{I} = W^TWXX^T \Rightarrow \underline{W^TW = n(XX^T)^{-1}}$. This expression shows that there is no way to compute a unique W, because if $W = W_0$ is a solution, then $W = AW_0$ for any unitary A would also be a solution. ~~There is an [illegible] in the [illegible] matrix [illegible] unique solution is possible [illegible] be Gaussian distributed like this for a solution to be found.~~

There is therefore a rotational invariance that exists for W ~~which~~ for Gaussian independent source terms which precludes a unique solution to the problem.

b) $\ell(W) = n\log|W| + \sum_{i,j}\log\frac{1}{2}e^{-|s_j^i|} \sim n\log|W| + \sum_{i,j} -|w_j^T x^i|$

$= n\log|W| - \sum_i |Wx^i|. \Longrightarrow \frac{\partial \ell}{\partial W} = \frac{n}{|W|}|W|W^{-T} - \sum_i \mathrm{sgn}(Wx^i)x^i = 0$

$\Rightarrow W := W + \alpha\frac{\partial \ell}{\partial W} = W + \alpha(nW^{-T} - \sum_i \mathrm{sgn}(Wx^i)x^i)$ in batch gradient descents, and

$W := W + \alpha(W^{-T} - \mathrm{sgn}(Wx^i)x^i)$ for single-update (stochastic) gradient descents.

c) Ok.

CS229 Problem set #3

1. $a \in \{"L", "R"\}$, $S = \{x, \dot{x}, \theta, \dot{\theta}\}$, $\gamma = .995$ are all known beforehand. Reward $R = R(S)$ and P_{sa} need to be learned. Eg through:

$$P_{sa}(s') = \frac{\#_{s,a \to s'}}{\#_{s,a}}, \quad R(S) = \frac{\sum_i R_i @ s \cdot \#_i @ s}{\sum_i \#_i @ s}.$$

Then Bellman's equations are to be solved to find the policy π^* which optimally selects actions at states:

$$\pi^*: V^*(s) = R(s) + \gamma \max_{a \in A} \sum_{s' \in S} P_{sa}(s') V^*(s')$$

$$\Rightarrow \pi^*(s) = \operatorname*{argmax}_{a \in A} \sum_{s' \in S} P_{sa}(s') V^*(s')$$

8:20 Answers to questions:

- 286 in my case
- Done.
- Seed "1" gives convergence at 259, seed 2 @ 135, seed 3 @ 377.

$\Rightarrow$ Algorithm is quite sensitive to initial conditions, likely due to the nature of the underlying physical system (chaotic?).