Stanford CS 229, Public Course, Problem Set 4

Dylan Price

March 6, 2017

1

a)

Given training set $\{(x^{(1)}, y^{(1)}, z^{(1)}), ..., (x^{(m)}, y^{(m)}, z^{(m)})\}$, write the log-likelihood of the parameters and derive the maximum likelihood estimate for ϕ, θ_0, and θ_1.

First we write the likelihood for a single point $(x^{(i)}, y^{(i)}, z^{(i)})$:

$$\mathscr{L}(x^{(i)}, y^{(i)}, z^{(i)}) = g(\phi^T x^{(i)})^{z^{(i)}} (1 - g(\phi^T x^{(i)}))^{1-z^{(i)}} \frac{1}{\sqrt{2\pi}\sigma} \exp \frac{-(y^{(i)} - \theta_{z^{(i)}}^T x^{(i)})^2}{2\sigma^2}$$

Then the log-likelihood of the parameters is:

$$\begin{aligned}
\ell(\phi, \theta_{z^{(i)}}) &= \log \prod_{i=1}^{m} L(x^{(i)}, y^{(i)}, z^{(i)}) \\
&= \sum_{i=1}^{m} \log \left[g(\phi^T x^{(i)})^{z^{(i)}} (1 - g(\phi^T x^{(i)}))^{1-z^{(i)}} \right] + \log \left[\frac{1}{\sqrt{2\pi}\sigma} \exp \frac{-(y^{(i)} - \theta_{z^{(i)}}^T x^{(i)})^2}{2\sigma^2} \right] \\
&= \sum_{i=1}^{m} z^{(i)} \log g(\phi^T x^{(i)}) + (1 - z^{(i)}) \log(1 - g(\phi^T x^{(i)})) + \log \left[\frac{1}{\sqrt{2\pi}\sigma} \exp \frac{-(y^{(i)} - \theta_{z^{(i)}}^T x^{(i)})^2}{2\sigma^2} \right] \\
&= \sum_{i=1}^{m} z^{(i)} \log g(\phi^T x^{(i)}) + (1 - z^{(i)}) \log(1 - g(\phi^T x^{(i)})) + \log(\frac{1}{\sqrt{2\pi}\sigma}) + \frac{-(y^{(i)} - \theta_{z^{(i)}}^T x^{(i)})^2}{2\sigma^2} \\
&= \sum_{i=1}^{m} z^{(i)} \log g(\phi^T x^{(i)}) + (1 - z^{(i)}) \log(1 - g(\phi^T x^{(i)})) + \log(\frac{1}{\sqrt{2\pi}\sigma}) \\
&\qquad - 1\{z^{(i)} = 0\} \frac{(y^{(i)} - \theta_0^T x^{(i)})^2}{2\sigma^2} - 1\{z^{(i)} = 1\} \frac{(y^{(i)} - \theta_1^T x^{(i)})^2}{2\sigma^2}
\end{aligned}$$

Now take the gradient with respect to θ_0 and set equal to 0:

$$\begin{aligned}
\nabla_{\theta_0}\ell(\phi,\theta_0,\theta_1) &= \nabla_{\theta_0}\sum_{i=1}^{m} -1\{z^{(i)}=0\}\frac{(y^{(i)}-\theta_0^T x^{(i)})^2}{2\sigma^2} \\
&= -\sum_{i=1}^{m} 1\{z^{(i)}=0\}\frac{-2y^{(i)}x^{(i)}+2x^{(i)}\theta_0^T x^{(i)}}{2\sigma^2} \\
\text{set} &= 0 \\
0 &= \sum_{i=1}^{m} 1\{z^{(i)}=0\}(-y^{(i)}x^{(i)}+x^{(i)}\theta_0^T x^{(i)}) \\
\sum_{i=1}^{m} 1\{z^{(i)}=0\}x^{(i)}\theta_0^T x^{(i)} &= \sum_{i=1}^{m} 1\{z^{(i)}=0\}y^{(i)}x^{(i)}
\end{aligned}$$

Let $X_{z=0}$ and $\vec{y}_{z=0}$ be equal to the design matrices except that all entries where $z=1$ are 0. Then the above is equivalent to:

$$\begin{aligned}
X_{z=0}^T X_{z=0}\theta_0 &= X_{z=0}^T \vec{y}_{z=0} \\
\theta_0 &= (X_{z=0}^T X_{z=0})^{-1} X_{z=0}^T \vec{y}_{z=0} \\
&\text{and similarly,} \\
\theta_1 &= (X_{z=1}^T X_{z=1})^{-1} X_{z=1}^T \vec{y}_{z=1}
\end{aligned}$$

Now find the gradient and hessian of $\ell(\phi,\theta_0,\theta_1)$ with respect to ϕ:

$$\begin{aligned}
\nabla_{\phi}\ell(\phi,\theta_0,\theta_1) &= \nabla_{\phi}\sum_{i=1}^{m} z^{(i)}\log g(\phi^T x^{(i)}) + (1-z^{(i)})\log(1-g(\phi^T x^{(i)})) \\
&= \sum_{i=1}^{m} \frac{z^{(i)}}{g(\phi^T x^{(i)})}x^{(i)}g(\phi^T x^{(i)})(1-g(\phi^T x^{(i)})) + \frac{1-z^{(i)}}{1-g(\phi^T x^{(i)})} - x^{(i)}g(\phi^T x^{(i)})(1-g(\phi^T x^{(i)})) \\
&= \sum_{i=1}^{m} \left[z^{(i)}(1-g(\phi^T x^{(i)})) - (1-z^{(i)})g(\phi^T x^{(i)})\right]x^{(i)} \\
&= \sum_{i=1}^{m} \left[z^{(i)} - z^{(i)}g(\phi^T x^{(i)}) - g(\phi^T x^{(i)}) + z^{(i)}g(\phi^T x^{(i)})\right]x^{(i)} \\
&= \sum_{i=1}^{m} \left[z^{(i)} - g(\phi^T x^{(i)})\right]x^{(i)} \\
\nabla_{\phi}^2\ell(\phi,\theta_0,\theta_1) &= \sum_{i=1}^{m} -g(\phi^T x^{(i)})(1-g(\phi^T x^{(i)}))x^{(i)}x^{(i)T}
\end{aligned}$$

b)

$\ell(\phi,\theta_0,\theta_1)$ is the same as part (a):

$$\ell(\phi,\theta_0,\theta_1) = \sum_{i=1}^{m}\log p(y^{(i)}|x^{(i)},z^{(i)})p(z^{(i)}|x^{(i)})$$

E-step

We want $Q_i(z^{(i)})$ to be proportional to $p(y^{(i)}|x^{(i)}, z^{(i)})p(z^{(i)}|x^{(i)})$ so that Jensen's inequality holds with equality. Therefore let $Q_i(z^{(i)}) = p(y^{(i)}|x^{(i)}, z^{(i)})p(z^{(i)}|x^{(i)})$. During the E-step, we calculate $Q_i(z^{(i)} = 0)$ and $Q_i(z^{(i)} = 1)$ for all i, using the formulas given in the problem statement and our current estimate of the parameters.

M-step

$$\phi, \theta_0, \theta_1 := \arg\max_{\phi,\theta_0,\theta_1} \sum_{i=1}^{m} \sum_{z^{(i)}=0}^{1} Q_i(z^{(i)}) \log \frac{p(y^{(i)}|x^{(i)}, z^{(i)})p(z^{(i)}|x^{(i)})}{Q_i(z^{(i)})}$$

Find the closed-form solution for θ_0 by taking the gradient with respect to θ_0 and setting equal to 0.

$$\nabla_{\theta_0} \sum_{i=1}^{m} \sum_{z^{(i)}=0}^{1} Q_i(z^{(i)}) \log \frac{p(y^{(i)}|x^{(i)}, z^{(i)}; \theta_{z^{(i)}})p(z^{(i)}|x^{(i)}; \phi)}{Q_i(z^{(i)})}$$

Let $1\{\text{guess } z^{(i)} = 0\}$ be equivalent to $1\{Q_i(z^{(i)} = 0) > Q_i(z^{(i)} = 1)\}$

$$= \nabla_{\theta_0} \sum_{i=1}^{m} 1\{\text{guess } z^{(i)} = 0\} Q_i(z^{(i)} = 0) \log \frac{p(y^{(i)}|x^{(i)}, z^{(i)} = 0; \theta_0)p(z^{(i)} = 0|x^{(i)}; \phi)}{Q_i(z^{(i)} = 0)}$$

$$= \nabla_{\theta_0} \sum_{i=1}^{m} 1\{\text{guess } z^{(i)} = 0\} Q_i(z^{(i)} = 0) \big[\log p(y^{(i)}|x^{(i)}, z^{(i)} = 0; \theta_0) + \log p(z^{(i)} = 0|x^{(i)}; \phi) - \log Q_i(z^{(i)} = 0) \big]$$

$$= \nabla_{\theta_0} \sum_{i=1}^{m} 1\{\text{guess } z^{(i)} = 0\} Q_i(z^{(i)} = 0) \log p(y^{(i)}|x^{(i)}, z^{(i)} = 0; \theta_0)$$

$$= \nabla_{\theta_0} \sum_{i=1}^{m} 1\{\text{guess } z^{(i)} = 0\} Q_i(z^{(i)} = 0) \log \Big(\frac{1}{\sqrt{2\pi}\sigma} \exp \frac{-(y^{(i)} - \theta_0^T x^{(i)})^2}{2\sigma^2} \Big)$$

$$= \nabla_{\theta_0} \sum_{i=1}^{m} 1\{\text{guess } z^{(i)} = 0\} Q_i(z^{(i)} = 0) \log \frac{1}{\sqrt{2\pi}\sigma} - Q_i(z^{(i)} = 0) \frac{(y^{(i)} - \theta_0^T x^{(i)})^2}{2\sigma^2}$$

$$= \nabla_{\theta_0} \sum_{i=1}^{m} -1\{\text{guess } z^{(i)} = 0\} Q_i(z^{(i)} = 0) \frac{(y^{(i)} - \theta_0^T x^{(i)})^2}{2\sigma^2}$$

$$= \sum_{i=1}^{m} -1\{\text{guess } z^{(i)} = 0\} Q_i(z^{(i)} = 0) \frac{-y^{(i)} x^{(i)} + x^{(i)} \theta_0^T x^{(i)}}{2\sigma^2}$$

Set $= 0$

$$0 = \sum_{i=1}^{m} -1\{\text{guess } z^{(i)} = 0\} Q_i(z^{(i)} = 0) \frac{-y^{(i)} x^{(i)} + x^{(i)} \theta_0^T x^{(i)}}{2\sigma^2}$$

$$\sum_{i=1}^{m} 1\{\text{guess } z^{(i)} = 0\} \frac{Q_i(z^{(i)} = 0)}{2\sigma^2} x^{(i)} \theta_0^T x^{(i)} = \sum_{i=1}^{m} 1\{\text{guess } z^{(i)} = 0\} \frac{Q_i(z^{(i)} = 0)}{2\sigma^2} y^{(i)} x^{(i)}$$

Let X_0 and $\vec{y}_0$ be the design matrices where the elements are non-zero if $Q_i(z^{(i)} = 0) > Q_i(z^{(i)} = 1)$.

Let $Q = \text{diag}(\frac{Q_0(z^{(0)} = 0)}{2\sigma^2}, \frac{Q_1(z^{(1)} = 0)}{2\sigma^2}, ..., \frac{Q_m(z^{(m)} = 0)}{2\sigma^2})$.

Then the above can be written as:

$$
\begin{aligned}
X_0^T Q X_0 \theta_0 &= X_0^T Q \vec{y}_0 \\
\theta_0 &= (X_0^T Q X_0)^{-1} X_0^T Q \vec{y}_0 \\
&\text{and similarly,} \\
\theta_1 &= (X_1^T Q X_1)^{-1} X_1^T Q \vec{y}_1
\end{aligned}
$$

Find the gradient and hessian for ϕ in order to numerically optimize the parameter.

$$
\begin{aligned}
&\nabla_\phi \sum_{i=1}^m \sum_{z^{(i)}=0}^1 Q_i(z^{(i)}) \log \frac{p(y^{(i)}|x^{(i)}, z^{(i)}; \theta_{z^{(i)}}) p(z^{(i)}|x^{(i)}; \phi)}{Q_i(z^{(i)})} \\
&= \nabla_\phi \sum_{i=1}^m \sum_{z^{(i)}=0}^1 Q_i(z^{(i)}) \big[\log p(y^{(i)}|x^{(i)}, z^{(i)}; \theta_{z^{(i)}}) + \log p(z^{(i)}|x^{(i)}; \phi) - \log Q_i(z^{(i)}) \big] \\
&= \nabla_\phi \sum_{i=1}^m \sum_{z^{(i)}=0}^1 Q_i(z^{(i)}) \log p(z^{(i)}|x^{(i)}; \phi) \\
&= \nabla_\phi \sum_{i=1}^m \sum_{z^{(i)}=0}^1 Q_i(z^{(i)}) \log \left[g(\phi^T x^{(i)})^{z^{(i)}} (1 - g(\phi^T x^{(i)}))^{1-z^{(i)}} \right] \\
&= \nabla_\phi \sum_{i=1}^m \sum_{z^{(i)}=0}^1 Q_i(z^{(i)}) \big[z^{(i)} \log g(\phi^T x^{(i)}) + (1 - z^{(i)}) \log(1 - g(\phi^T x^{(i)})) \big] \\
&= \nabla_\phi \sum_{i=1}^m Q_i(z^{(i)} = 1) \log g(\phi^T x^{(i)}) + Q_i(z^{(i)} = 0) \log(1 - g(\phi^T x^{(i)})) \\
&= \nabla_\phi \sum_{i=1}^m Q_i(z^{(i)} = 1) \log g(\phi^T x^{(i)}) + (1 - Q_i(z^{(i)} = 1)) \log(1 - g(\phi^T x^{(i)})) \\
&\text{similarly to part (a) this becomes} \\
&= \sum_{i=1}^m (Q_i(z^{(i)} = 1) - g(\phi^T x^{(i)})) x^{(i)} \\
&\text{and} \\
\nabla_\phi^2 &= \sum_{i=1}^m -g(\phi^T x^{(i)})(1 - g(\phi^T x^{(i)})) x^{(i)} x^{(i)^T}
\end{aligned}
$$

2

a)

First determine the joint distribution over (x, z).

$(x, z) \sim \mathcal{N}(\mu_{xz}, \Sigma_{xz})$
$\mu_{xz} = \begin{bmatrix} \mu_x \\ \mu_z \end{bmatrix} \Sigma_{xz} = \begin{bmatrix} \Sigma_{xx} & \Sigma_{xz} \\ \Sigma_{zx} & \Sigma_{zz} \end{bmatrix}$

We know $x|z \sim \mathcal{N}(Uz, \sigma^2 I)$, so we can define x as $x = Uz + \epsilon$ where $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$.
Then $E[x] = E[Uz + \epsilon] = UE[z] + E[\epsilon] = 0$. Therefore $\mu_x = 0$.

We know $z \sim \mathcal{N}(0, I)$, therefore $\mu_z = 0$.

$$\begin{aligned}
\Sigma_{xx} &= E[(x - E[x])(x - E[x])^T] \\
&= E[xx^T] \\
&= E[(Uz + \epsilon)(Uz + \epsilon)^T] \\
&= E[(Uz + \epsilon)(z^T U^T + \epsilon^T)] \\
&= E[Uzz^T U^T + \epsilon z^T U^T + Uz\epsilon^T + \epsilon\epsilon^T] \\
&= UE[zz^T]U^T + E[\epsilon]E[z^T]U^T + UE[z]E[\epsilon^T] + E[\epsilon\epsilon^T] \\
&= UIU^T + \sigma^2 I \\
&= UU^T + \sigma^2 I
\end{aligned}$$

$$\begin{aligned}
\Sigma_{xz} &= E[(x - E[x])(z - E[z])^T] \\
&= E[xz^T] \\
&= E[(Uz + \epsilon)z^T] \\
&= E[Uzz^T + \epsilon z^T] \\
&= UE[zz^T] + E[\epsilon z^T] \\
&= UI + E[\epsilon]E[z^T] \\
&= U
\end{aligned}$$

$$\begin{aligned}
\Sigma_{zx} &= \Sigma_{xz}^T = U^T \\
\Sigma_{zz} &= I
\end{aligned}$$

So, $(x, z) \sim \mathcal{N}(\mu_{xz}, \Sigma_{xz})$ where

$$\mu_{xz} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \; \Sigma_{xz} = \begin{bmatrix} UU^T + \sigma^2 I & U \\ U^T & I \end{bmatrix}$$

Now determine the conditional distribution of $z|x$.

$z|x \sim \mathcal{N}(\mu_{z|x}, \Sigma_{z|x})$

By the rules for manipulating Gaussians from the lecture notes:

$$\begin{aligned}
\mu_{z|x} &= \mu_z + \Sigma_{zx}\Sigma_{xx}^{-1}(x - \mu_x) \\
&= 0 + U^T(UU^T + \sigma^2 I)^{-1}x \\
&= U^T x(\sigma^2 I + U^T U)^{-1} && \text{(by the identity in the problemset)}
\end{aligned}$$

$$\begin{aligned}
\Sigma_{z|x} &= \Sigma_{zz} - \Sigma_{zx}\Sigma_{xx}^{-1}\Sigma_{xz} \\
&= I - U^T(UU^T + \sigma^2 I)^{-1}U \\
&= I - U^T U(\sigma^2 I + U^T U)^{-1} && \text{(by the identity in the problemset)}
\end{aligned}$$

b)

E-step
Compute $Q_i(z^{(i)}) = p(z^{(i)}|x^{(i)};U)$ for all i,
where $Q_i(z^{(i)}) \sim \mathcal{N}\big(U^T x^{(i)}(\sigma^2 I + U^T U)^{-1}, I - U^T U(\sigma^2 I + U^T U)^{-1}\big)$

M-step
$U := \arg\max_U \sum_{i=1}^{m} \int_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; U)}{Q_i(z^{(i)})} dz^{(i)}$
where $p(x^{(i)}, z^{(i)}; U) \sim \mathcal{N}(\mu_{xz}, \Sigma_{xz})$

To compute the above, take the gradient with respect to U and set equal to 0.

$$
\begin{aligned}
&\nabla_U \sum_{i=1}^{m} \int_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; U)}{Q_i(z^{(i)})} dz^{(i)} \\
&= \nabla_U \sum_{i=1}^{m} E_{z^{(i)} \sim Q_i}\big[\log p(x^{(i)}, z^{(i)}; U) - \log Q_i(z^{(i)})\big] \\
&= \nabla_U \sum_{i=1}^{m} E_{z^{(i)} \sim Q_i}\big[\log p(x^{(i)}|z^{(i)}; U) p(z^{(i)}) - \log Q_i(z^{(i)})\big] \\
&= \nabla_U \sum_{i=1}^{m} E_{z^{(i)} \sim Q_i}\big[\log p(x^{(i)}|z^{(i)}; U) + \log p(z^{(i)}) - \log Q_i(z^{(i)})\big] \\
&= \nabla_U \sum_{i=1}^{m} E_{z^{(i)} \sim Q_i}\big[\log p(x^{(i)}|z^{(i)}; U)\big]
\end{aligned}
$$

Now substitute in $p(x^{(i)}|z^{(i)};U)$.

$$
\begin{aligned}
&= \nabla_U \sum_{i=1}^{m} E_{z^{(i)}\sim Q_i}\big[\log \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma_{x^{(i)}|z^{(i)}}|^{\frac{1}{2}}} \exp\big(-\frac{1}{2}(x^{(i)}-\mu_{x^{(i)}|z^{(i)}})^T\Sigma^{-1}_{x^{(i)}|z^{(i)}}(x^{(i)}-\mu_{x^{(i)}|z^{(i)}})\big)\big] \\
&= \nabla_U \sum_{i=1}^{m} E_{z^{(i)}\sim Q_i}\big[\log \frac{1}{(2\pi)^{\frac{n}{2}}|\sigma^2 I|^{\frac{1}{2}}} \exp\big(-\frac{1}{2}(x^{(i)}-Uz^{(i)})^T(\sigma^2 I)^{-1}(x^{(i)}-Uz^{(i)})\big)\big] \\
&= \nabla_U \sum_{i=1}^{m} E_{z^{(i)}\sim Q_i}\big[\log \frac{1}{(2\pi)^{\frac{n}{2}}|\sigma^2 I|^{\frac{1}{2}}} - \frac{1}{2}(x^{(i)}-Uz^{(i)})^T(\sigma^2 I)^{-1}(x^{(i)}-Uz^{(i)})\big] \\
&= \nabla_U \sum_{i=1}^{m} E_{z^{(i)}\sim Q_i}\big[-\frac{1}{2}(x^{(i)}-Uz^{(i)})^T(\sigma^2 I)^{-1}(x^{(i)}-Uz^{(i)})\big] \\
&= \nabla_U \sum_{i=1}^{m} E_{z^{(i)}\sim Q_i}\big[-\frac{1}{2}({x^{(i)}}^T-{z^{(i)}}^TU^T)(\frac{1}{\sigma^2}I)(x^{(i)}-Uz^{(i)})\big] \\
&= \nabla_U \sum_{i=1}^{m} E_{z^{(i)}\sim Q_i}\big[-\frac{1}{2}({x^{(i)}}^T-{z^{(i)}}^TU^T)(\frac{1}{\sigma^2}x^{(i)}-\frac{1}{\sigma^2}Uz^{(i)})\big] \\
&= \nabla_U \sum_{i=1}^{m} E_{z^{(i)}\sim Q_i}\big[-\frac{1}{2}(\frac{1}{\sigma^2}{x^{(i)}}^Tx^{(i)}-\frac{1}{\sigma^2}{x^{(i)}}^TUz^{(i)}-\frac{1}{\sigma^2}{z^{(i)}}^TU^Tx^{(i)}+\frac{1}{\sigma^2}{z^{(i)}}^TU^TUz^{(i)})\big] \\
&= \nabla_U \sum_{i=1}^{m} E_{z^{(i)}\sim Q_i}\big[-\frac{1}{2\sigma^2}({x^{(i)}}^Tx^{(i)}-{x^{(i)}}^TUz^{(i)}-{z^{(i)}}^TU^Tx^{(i)}+{z^{(i)}}^TU^TUz^{(i)})\big] \\
&= \nabla_U \sum_{i=1}^{m} E_{z^{(i)}\sim Q_i}\big[-\frac{1}{2\sigma^2}({x^{(i)}}^Tx^{(i)}-2{z^{(i)}}^TU^Tx^{(i)}+{z^{(i)}}^TU^TUz^{(i)})\big] \\
&= \sum_{i=1}^{m} E_{z^{(i)}\sim Q_i}\big[-\frac{1}{2\sigma^2}\nabla_U({x^{(i)}}^Tx^{(i)}-2{z^{(i)}}^TU^Tx^{(i)}+{z^{(i)}}^TU^TUz^{(i)})\big] \\
&= \sum_{i=1}^{m} E_{z^{(i)}\sim Q_i}\big[-\frac{1}{2\sigma^2}(-2x^{(i)}{z^{(i)}}^T+2Uz^{(i)}{z^{(i)}}^T)\big] \\
&= \frac{1}{\sigma^2}\sum_{i=1}^{m} E_{z^{(i)}\sim Q_i}\big[x^{(i)}{z^{(i)}}^T-Uz^{(i)}{z^{(i)}}^T\big]
\end{aligned}
$$

Set equal to 0.

$$
\begin{aligned}
0 &= \frac{1}{\sigma^2}\sum_{i=1}^{m} E_{z^{(i)}\sim Q_i}\big[x^{(i)}{z^{(i)}}^T\big] - E_{z^{(i)}\sim Q_i}\big[Uz^{(i)}{z^{(i)}}^T\big] \\
\sum_{i=1}^{m} x^{(i)}E_{z^{(i)}\sim Q_i}\big[{z^{(i)}}^T\big] &= U\sum_{i=1}^{m} E_{z^{(i)}\sim Q_i}\big[z^{(i)}{z^{(i)}}^T\big] \\
U &= \big(\sum_{i=1}^{m} x^{(i)}E_{z^{(i)}\sim Q_i}\big[{z^{(i)}}^T\big]\big)\big(\sum_{i=1}^{m} E_{z^{(i)}\sim Q_i}\big[z^{(i)}{z^{(i)}}^T\big]\big)^{-1}
\end{aligned}
$$

Since $Q_i(z^{(i)}) = p(z^{(i)}|x^{(i)};U)$,
$E_{z^{(i)}\sim Q_i}\big[{z^{(i)}}^T\big] = \mu^T_{z^{(i)}|x^{(i)}}$
and
$E_{z^{(i)}\sim Q_i}\big[z^{(i)}{z^{(i)}}^T\big] = \Sigma_{z^{(i)}|x^{(i)}} + \mu_{z^{(i)}|x^{(i)}}\mu^T_{z^{(i)}|x^{(i)}}$

So our update computation for U is:
$U = (\sum_{i=1}^{m} x^{(i)} \mu_{z^{(i)}|x^{(i)}})(\sum_{i=1}^{m} \Sigma_{z^{(i)}|x^{(i)}} + \mu_{z^{(i)}|x^{(i)}} \mu_{z^{(i)}|x^{(i)}}^T)^{-1}$

c)

First we show that the E-step can be expressed as $w = \frac{XU}{U^TU}$.
Recall that

$$\begin{aligned} Q(z^{(i)}) &\sim \mathcal{N}\big(\mu_{z^{(i)}|x^{(i)}}, \Sigma_{z^{(i)}|x^{(i)}}\big) \\ &\sim \mathcal{N}\big(U^T x^{(i)}(\sigma^2 I + U^T U)^{-1}, I - U^T U(\sigma^2 I + U^T U)\big) \end{aligned}$$

As $\sigma^2 \to 0$,

$$\begin{aligned} \mu_{z^{(i)}|x^{(i)}} &= U^T x^{(i)} (U^T U)^{-1} \\ \Sigma_{z^{(i)}|x^{(i)}} &= I - U^T U (U^T U)^{-1} \\ &= I - U^T U U^{-1} U^{-T} \\ &= I - I \\ &= 0 \end{aligned}$$

Therefore $Q_i(z^{(i)}) = \mu_{z^{(i)}|x^{(i)}} = \frac{U^T x^{(i)}}{U^T U}$.
Define $w \in \mathbb{R}^m$ such that $w_i = Q_i(z^{(i)}) = \mu_{z^{(i)}|x^{(i)}}$. Then $w = \frac{XU}{U^TU}$.

Now we show that the M-step can be expressed as $U = \frac{X^T w}{w^T w}$.
Recall that

$$\begin{aligned} U &= (\sum_{i=1}^{m} x^{(i)} \mu_{z^{(i)}|x^{(i)}})(\sum_{i=1}^{m} \Sigma_{z^{(i)}|x^{(i)}} + \mu_{z^{(i)}|x^{(i)}} \mu_{z^{(i)}|x^{(i)}}^T)^{-1} \\ &= (\sum_{i=1}^{m} x^{(i)} w)(\sum_{i=1}^{m} \Sigma_{z^{(i)}|x^{(i)}} + w_i w_i^T)^{-1} \end{aligned}$$

As $\sigma^2 \to 0$,

$$\begin{aligned} U &= (\sum_{i=1}^{m} x^{(i)} w)(\sum_{i=1}^{m} w_i w_i^T)^{-1} \\ &= (X^T w)(w^T w)^{-1} \\ &= \frac{X^T w}{w^T w} \end{aligned}$$

If the algorithm has converged, then U will converge to some U^* and w will converge to some w^*, and we will have that
$w^* = \frac{XU^*}{U^{*T}U^*}$ and $U^* = \frac{X^T w^*}{w^{*T} w^*}$

Plugging w^* into the equation for U^*, we have

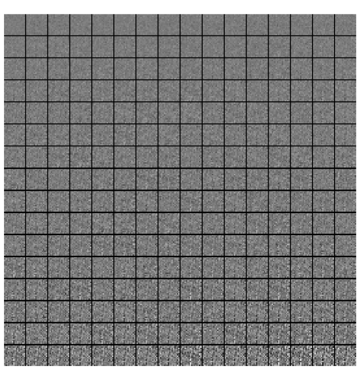
$$\begin{aligned}
U^* &= \frac{X^T \frac{XU^*}{U^{*T}U^*}}{(\frac{XU^*}{U^{*T}U^*})^T(\frac{XU^*}{U^{*T}U^*})} \\
&= X^T XU^*(U^{*T}U^*)^{-1}\big[(XU^*(U^{*T}U^*)^{-1})^T(XU^*(U^{*T}U^*)^{-1})\big]^{-1} \\
&\text{let } a = U^{*T}U^* \text{ (where } a \in \mathbb{R} \text{ since we are assuming } U \in \mathbb{R}^n \text{ in the problem statement)} \\
&= X^T XU^*(\frac{1}{a})\big[(XU^*(\frac{1}{a}))^T(XU^*(\frac{1}{a}))\big]^{-1} \\
&= \frac{1}{a}X^T XU^*\big[\frac{1}{a^2}U^{*T}X^T XU^*\big]^{-1} \\
&= \frac{1}{a}X^T XU^* a^2\big[U^{*T}X^T XU^*\big]^{-1} \\
&= aX^T XU^*\big[U^{*T}X^T XU^*\big]^{-1} \\
&\text{But } U^{*T}X^T XU^* \in \mathbb{R}, \text{ so} \\
&= X^T XU^*\frac{a}{U^{*T}X^T XU^*} \\
&= X^T XU^*\frac{U^{*T}U^*}{U^{*T}X^T XU^*} \\
&= \frac{1}{m}X^T XU^*\frac{mU^{*T}U^*}{U^{*T}X^T XU^*}
\end{aligned}$$

So we have

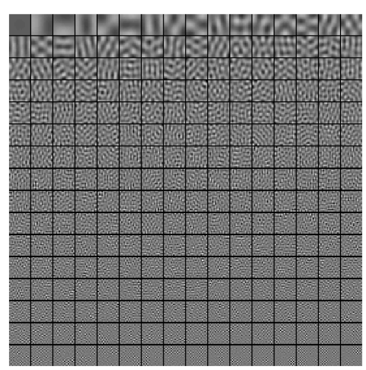
$$\begin{aligned}
U^* &= \frac{1}{m}X^T XU^*\frac{mU^{*T}U^*}{U^{*T}X^T XU^*} \\
(\frac{U^{*T}X^T XU^*}{mU^{*T}U^*})U^* &= (\frac{1}{m}X^T X)U^* \\
\text{Let } \lambda &= \frac{U^{*T}X^T XU^*}{mU^{*T}U^*} \\
\lambda U^* &= \Sigma U^*
\end{aligned}$$

3

See q3/ folder for code.



(a) ICA unmixing matrix



(b) PCA principal components matrix

The ICA unmixing matrix separates the input data into its original sources, while the PCA principal components matrix represents the principal dimensions the data lies in.

4

a)

Let $s \in \mathbb{S}$ be any state in $\mathbb{S}$.
Let $s' \sim P_{s\pi(s)}$ be a random variable drawn from the $P_{s\pi(s)}$ distribution.
Since $V_1(s) \leq V_2(s) \quad \forall s \in \mathbb{S}, \quad V_1(s') \leq V_2(s')$, and we have

$$\begin{aligned}
V_1(s') &\leq V_2(s') \\
E_{s' \sim P_{s\pi(s)}}[V_1(s')] &\leq E_{s' \sim P_{s\pi(s)}}[V_2(s')] \\
\gamma E_{s' \sim P_{s\pi(s)}}[V_1(s')] &\leq \gamma E_{s' \sim P_{s\pi(s)}}[V_2(s')] \\
R(s) + \gamma E_{s' \sim P_{s\pi(s)}}[V_1(s')] &\leq R(s) + \gamma E_{s' \sim P_{s\pi(s)}}[V_2(s')] \\
B(V_1)(s) &\leq B(V_2)(s) \quad \forall s \in \mathbb{S}
\end{aligned}$$

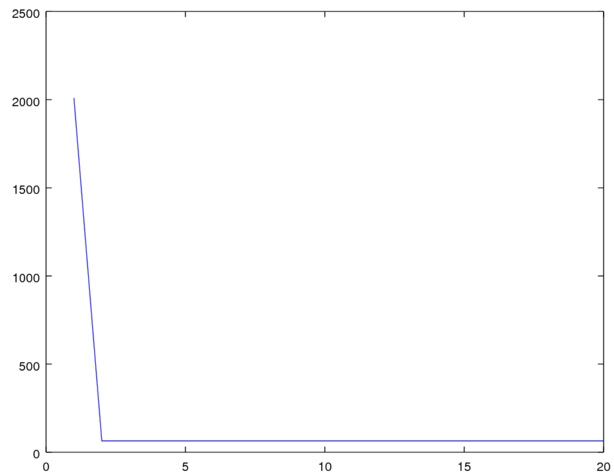
b)

$$\begin{aligned}
||B^\pi(v) - V^\pi||_\infty &= ||R(s) + \gamma \sum_{s' \in \mathbb{S}} P_{s\pi(s)}(s')V(s') - (R(s) + \gamma \sum_{s' \in \mathbb{S}} P_{s\pi(s)}(s')V^\pi(s'))||_\infty \\
&= ||\gamma \sum_{s' \in \mathbb{S}} P_{s\pi(s)}(s')V(s') - P_{s\pi(s)}(s')V^\pi(s')||_\infty \\
&= ||\gamma \sum_{s' \in \mathbb{S}} P_{s\pi(s)}(s')(V(s') - V^\pi(s'))||_\infty \\
&= \max_{s' \in \mathbb{S}} \big[\gamma \sum_{s' \in \mathbb{S}} P_{s\pi(s)}(s')(V(s') - V^\pi(s'))\big] \\
&= \gamma \max_{s' \in \mathbb{S}} \big[\sum_{s' \in \mathbb{S}} P_{s\pi(s)}(s')(V(s') - V^\pi(s'))\big] \\
&= \gamma \max_{s' \in \mathbb{S}} \big[E_{s' \sim P_{s\pi(s)}}[V(s') - V^\pi(s')]\big] \\
&\leq \gamma \max_{s' \in \mathbb{S}} \big[\max_{s' \in \mathbb{S}}[V(s') - V^\pi(s')]\big] \\
&= \gamma \max_{s' \in \mathbb{S}} \big[V(s') - V^\pi(s')\big] \\
&= \gamma||V - V^\pi||_\infty
\end{aligned}$$

Therefore $||B^\pi(v) - V^\pi||_\infty \leq \gamma||V - V^\pi||_\infty$.

5

See q5/ folder for code.



x-axis is approximate episode number, y-axis is number of steps before the car reached the top of the hill. Each value plotted represents an average over 10 runs and 500 consecutive episodes. In total there were 10 runs of 10000 episodes each.