Problem Set #3

Omer Hazon

November 15, 2017

1 A Simple Neural Network

(a)

Defining the delta vectors:

$$\delta^{[2]} = \nabla_{w_0^{[2]}} J = \nabla_o (o-y)^2 \circ \sigma'(z^{[2]}) = 2(o-y) \circ o(1-o) = 2(o-y)o(1-o)$$

$$\nabla_{w^{[2]}} J = \delta^{[2]} a^{[1]T} = \delta^{[2]} h^T = 2(o-y)o(1-o)h^T$$

$$\delta^{[1]} = \nabla_{w_0^{[1]}} J = \left(w^{[2]T}\delta^{[2]}\right) \circ \sigma'(z^{[1]}) = \left(w^{[2]T} 2(o-y)o(1-o)\right) \circ h \circ (1-h) =$$

$$= 2(o-y)o(1-o)w^{[2]T} \circ h \circ (1-h)$$

$$\nabla_{w^{[1]}} J = \delta^{[1]} a^{[0]T} = \left(2(o-y)o(1-o)w^{[2]T} \circ h \circ (1-h)\right) x^T$$

The (1,2) component of the above is:

$$\left[\left(2(o-y)o(1-o)w^{[2]T} \circ h \circ (1-h)\right) x^T\right]_{1,2} = \left(2(o-y)o(1-o)w^{[2]T} \circ h \circ (1-h)\right)_1 x_2 =$$

$$= \left(2(o-y)o(1-o)w_1^{[2]T} h_1(1-h_1)\right) x_2 = 2(o-y)o(1-o)w_1^{[2]} h_1(1-h_1)x_2$$

And so the full gradient descent step is:

$$w_{1,2}^{[1]} \leftarrow w_{1,2}^{[1]} - \frac{\alpha}{m} \sum_{i=1}^{m} 2(o^{(i)} - y^{(i)})o^{(i)}(1-o^{(i)})w_1^{[2]} h_1^{(i)}(1-h_1^{(i)})x_2^{(i)}$$

The value of $h_1^{(i)}$ in terms of x and weigths is $h_1^{(i)} = \sigma(w_{:,1}^{[1]T} x^{(i)} + w_{0,1}^{[1]})$

$$w_{1,2}^{[1]} \leftarrow w_{1,2}^{[1]} - 2\frac{\alpha}{m} \sum_{i=1}^{m} (o^{(i)} - y^{(i)})o^{(i)}(1-o^{(i)})w_1^{[2]} \sigma(w_{:,1}^{[1]T} x^{(i)} + w_{0,1}^{[1]}) \left(1 - \sigma(w_{:,1}^{[1]T} x^{(i)} + w_{0,1}^{[1]})\right) x_2^{(i)}$$

(b)

The first layer comprises of three linear separations. These three linear separations can then be combined to distinguish between the inside and outside of the triangle.

The form of the linear separation equation is $w_0 + w_1 x_1 + w_2 x_2 = 0$.

The three vertices of the triangle are approximately at the points defined by the rows of the following matrix, found by closely analyzing the pixels in the given figure:

```
0.31978    0.37050
0.31978    3.95197
3.73442    0.37050
```

A line between two points $(x_1, y_1), (x_2, y_2)$ is given by the equation:

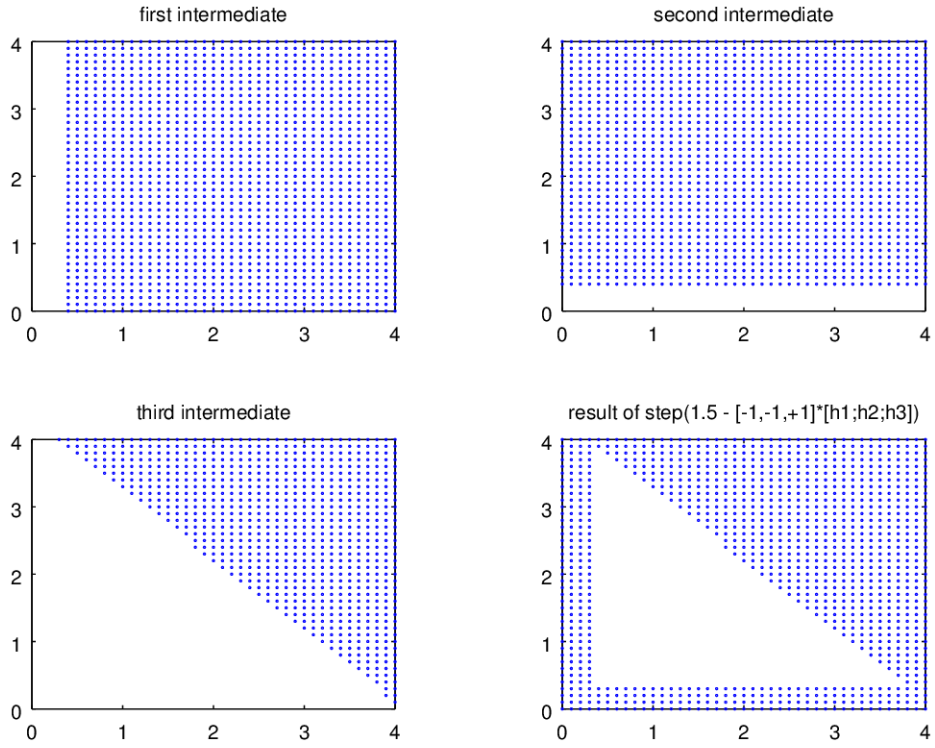
$$w_0 + w_1 x + w_2 y = (x_1 y_2 - y_1 x_2) + (y_1 - y_2)x + (x_2 - x_1)y = 0$$

Applying this formula to each pair of points, gives three sets of weights:

```
w0: −1.145296  w1:  3.581475  w2:  0.000000
w0: −1.265113  w1:  0.000000  w2:  3.414634
w0: −14.639836 w1:  3.581475  w2:  3.414634
```

And by visualizing the inequalities created by the edges of the triangle, the following second layer weights were chosen:

```
bias: 1.5, weights: (−1, −1, +1)
```



In the language of the problem, the weights assigned to the network are:

$$w_{0,1}^{[1]} = -1.145296, w_{1,1}^{[1]} = 3.581475, w_{2,1}^{[1]} = 0$$

$$w_{0,2}^{[1]} = -1.265113, w_{1,2}^{[1]} = 0, w_{2,2}^{[1]} = 3.414634$$

$$w_{0,3}^{[1]} = -14.639836, w_{1,3}^{[1]} = 3.581475, w_{2,3}^{[1]} = 3.414634$$

$$w_0^{[2]} = 1.5, w_1^{[2]} = -1, w_2^{[2]} = -1, w_3^{[2]} = 1$$

(c)

Such weights do not exist, since in that case the output amounts to a step function over a linear function of the input (with bias). This is identical to just a single layer, and represents a linear separation of the data. The data is not linearly seperable and therefore such weights do not exist.

$$f\left(w^{[2]}\left(w^{[1]}x + w_0^{[1]}\right) + w_0^{[2]}\right) = f\left(\left(w^{[2]}w^{[1]}\right)x + \left(w^{[2]}w_0^{[1]} + w_0^{[2]}\right)\right)$$

2 EM for MAP estimation

We wish to maximize:

$$p(\theta)\prod_{i=1}^{m} p(x^{(i)} \mid \theta) = p(\theta)\left(\prod_{i=1}^{m}\sum_{z^{(i)}} p(x^{(i)}, z^{(i)} \mid \theta)\right)$$

Taking the log:

$$\log p(\theta) + \sum_{i=1}^{m} \log p(x^{(i)} \mid \theta) = \log p(\theta) + \sum_{i=1}^{m} \log\left(\sum_{z^{(i)}} p(x^{(i)}, z^{(i)} \mid \theta)\right) =$$

$$\log p(\theta) + \sum_{i=1}^{m} \log\left(\sum_{z^{(i)}} Q_i(z^{(i)})\frac{p(x^{(i)}, z^{(i)} \mid \theta)}{Q_i(z^{(i)})}\right) \geq \log p(\theta) + \sum_{i=1}^{m}\sum_{z^{(i)}} Q_i(z^{(i)}) \log\left(\frac{p(x^{(i)}, z^{(i)} \mid \theta)}{Q_i(z^{(i)})}\right)$$

By the concavity of the log function and Jensen's inequality. The expectation of the log over the distribution $Q_i(z^{(i)})$ is lesser or equal to the log of the expectation.

Let the E-step consist of setting the Q's such that Jensen's inequality holds with equality, that is, that

$$Q_i(z^{(i)}) \propto p(x^{(i)}, z^{(i)} \mid \theta) \rightarrow^{\text{implies}} Q_i(z^{(i)}) = \frac{p(x^{(i)}, z^{(i)} \mid \theta)}{\sum_z p(x^{(i)}, z \mid \theta)} = \frac{p(x^{(i)}, z^{(i)} \mid \theta)}{p(x^{(i)} \mid \theta)} = p(z^{(i)} \mid x^{(i)}, \theta)$$

$$Q_i(z^{(i)}) \leftarrow p(z^{(i)} \mid x^{(i)}, \theta)$$

And let the M-step consist of maximizing the right hand side of the inequality, which starts out being equal with the initial value of θ but ends up being less than the left hand side.

$$\theta \leftarrow \arg\max_{\theta}\left[\log p(\theta) + \sum_{i=1}^{m}\sum_{z^{(i)}} Q_i(z^{(i)}) \log\left(\frac{p(x^{(i)}, z^{(i)} \mid \theta)}{Q_i(z^{(i)})}\right)\right]$$

This amounts to maximizing a linear combination of $\log p(\theta)$ and $\log p(x, z \mid \theta)$ with various inputs for x and z. These are assumed to be concave in θ and therefore the maximization of their linear combination is tractable.

To prove that $p(\theta)\prod_{i=1}^{m} p(x^{(i)} \mid \theta) == p(\theta)\left(\prod_{i=1}^{m}\sum_{z^{(i)}} p(x^{(i)}, z^{(i)} \mid \theta)\right)$ monotonically increases with each update of θ, we look at its log, $\log p(\theta) + \sum_{i=1}^{m} \log\left(\sum_{z^{(i)}} p(x^{(i)}, z^{(i)} \mid \theta)\right)$ and use Jensen's inequality from above.

Given any value of θ from step t of the EM-MAP algorithm, that is, the value $\theta^{(t)}$, we arrive at the value of the log-probability as:

$$\log p(\theta^{(t)}) + \sum_{i=1}^{m} \log\left(\sum_{z^{(i)}} p(x^{(i)}, z^{(i)} \mid \theta^{(t)})\right)$$

Applying EM-MAP, we arrive at

$$Q_i^{(t)}(z^{(i)}) \leftarrow p(z^{(i)} \mid x^{(i)}, \theta^{(t)})$$

And then to

$$\theta^{(t+1)} = \arg\max_{\theta}\left[\log p(\theta) + \sum_{i=1}^{m}\sum_{z^{(i)}} p(z^{(i)} \mid x^{(i)}, \theta^{(t)}) \log\left(\frac{p(x^{(i)}, z^{(i)} \mid \theta)}{p(z^{(i)} \mid x^{(i)}, \theta^{(t)})}\right)\right]$$

It is known that by design,

$$\log p(\theta^{(t)}) + \sum_{i=1}^{m}\sum_{z^{(i)}} p(z^{(i)} \mid x^{(i)}, \theta^{(t)}) \log\left(\frac{p(x^{(i)}, z^{(i)} \mid \theta^{(t)})}{p(z^{(i)} \mid x^{(i)}, \theta^{(t)})}\right) = \log p(\theta^{(t)}) + \sum_{i=1}^{m} \log \sum_{z^{(i)}} p(x^{(i)}, z^{(i)} \mid \theta^{(t)})$$

And due to the maximization step,

$$\log p(\theta^{(t+1)}) + \sum_{i=1}^{m}\sum_{z^{(i)}} p(z^{(i)} \mid x^{(i)}, \theta^{(t)}) \log\left(\frac{p(x^{(i)}, z^{(i)} \mid \theta^{(t+1)})}{p(z^{(i)} \mid x^{(i)}, \theta^{(t)})}\right)$$

$$\geq \log p(\theta^{(t)}) + \sum_{i=1}^{m}\sum_{z^{(i)}} p(z^{(i)} \mid x^{(i)}, \theta^{(t)}) \log\left(\frac{p(x^{(i)}, z^{(i)} \mid \theta^{(t)})}{p(z^{(i)} \mid x^{(i)}, \theta^{(t)})}\right)$$

By the equality above,

$$\log p(\theta^{(t+1)}) + \sum_{i=1}^{m}\sum_{z^{(i)}} p(z^{(i)} \mid x^{(i)}, \theta^{(t)}) \log\left(\frac{p(x^{(i)}, z^{(i)} \mid \theta^{(t+1)})}{p(z^{(i)} \mid x^{(i)}, \theta^{(t)})}\right) \geq \log p(\theta^{(t)}) + \sum_{i=1}^{m} \log \sum_{z^{(i)}} p(x^{(i)}, z^{(i)} \mid \theta^{(t)})$$

And by Jensen's inequality,

$$\log p(\theta^{(t+1)}) + \sum_{i=1}^{m} \log \sum_{z^{(i)}} p(x^{(i)}, z^{(i)} \mid \theta^{(t+1)}) \geq \log p(\theta^{(t+1)}) + \sum_{i=1}^{m}\sum_{z^{(i)}} p(z^{(i)} \mid x^{(i)}, \theta^{(t)}) \log\left(\frac{p(x^{(i)}, z^{(i)} \mid \theta^{(t+1)})}{p(z^{(i)} \mid x^{(i)}, \theta^{(t)})}\right)$$

Therefore, putting together the two inequalities,

$$\log p(\theta^{(t+1)}) + \sum_{i=1}^{m} \log \sum_{z^{(i)}} p(x^{(i)}, z^{(i)} \mid \theta^{(t+1)}) \geq \log p(\theta^{(t)}) + \sum_{i=1}^{m} \log \sum_{z^{(i)}} p(x^{(i)}, z^{(i)} \mid \theta^{(t)})$$

$$\log p(\theta^{(t+1)}) + \sum_{i=1}^{m} \log p(x^{(i)} \mid \theta^{(t+1)}) \geq \log p(\theta^{(t)}) + \sum_{i=1}^{m} \log p(x^{(i)} \mid \theta^{(t)})$$

$$p(\theta^{(t+1)}) \prod_{i=1}^{m} p(x^{(i)} \mid \theta^{(t+1)}) \geq p(\theta^{(t)}) \prod_{i=1}^{m} p(x^{(i)} \mid \theta^{(t)})$$

And the desired quantity increases monotonically upon application of an EM-MAP step.

3 EM Application

(a)

(i)

$$p(y^{(pr)}, z^{(pr)}, x^{(pr)}) = p(x^{(pr)} \mid y^{(pr)}, z^{(pr)})p(y^{(pr)})p(z^{(pr)}) =$$

$$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x^{(pr)} - y^{(pr)} - z^{(pr)})^2}{2\sigma^2}\right) \frac{1}{\sqrt{2\pi\sigma_p^2}} \exp\left(-\frac{(y^{(pr)} - \mu_p)^2}{2\sigma_p^2}\right) \frac{1}{\sqrt{2\pi\tau_r^2}} \exp\left(-\frac{(z^{(pr)} - \nu_r)^2}{2\tau_r^2}\right)$$

If you assume that the total distribution has the form of a multivariate gaussian, then it is possible to reason about the mean and covariance matrix. The means of $y^{(pr)}$ and $z^{(pr)}$ will be the same as before, μ_p and ν_r, respectively, and the mean of $x^{(pr)}$ will remain $\mu_p + \nu_r$. The diagonals of the covariance matrix will be the same as the variance of the univariate distributions for $y^{(pr)}$ and $z^{(pr)}$, being σ_p^2 and τ_r^2 respectively, while for $x^{(pr)}$ since we are interested in the uncoditional probabilities, we cannot state that the variance is σ^2. Instead view $x^{(pr)}$ as the sum of the random variables $y^{(pr)} + z^{(pr)} + \epsilon^{(pr)}$ where $\epsilon^{(pr)} \sim N(0, \sigma^2)$ is uncorrelated. Then the variance of $x^{(pr)}$ will be the sum of the variances $\sigma_p^2 + \tau_r^2 + \sigma^2$. As for the off-diagonal terms, $y^{(pr)}$ and $z^{(pr)}$ are independent and do not covary, while for the off-diagonal terms involving $x^{(pr)}$ the covariance will come from the term in the random-variable-sum that is not independent from the covariant, that is, the covariance of $x^{(pr)}$ with $y^{(pr)}$ is identical to the covariance of $y^{(pr)}$ and $y^{(pr)}$, or σ_p^2, and similarly for the covariance between $x^{(pr)}$ and $z^{(pr)}$, it will be τ_r^2.

In summary, the mean vector is (using the ordering $[x, y, z]^T$):

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_p + \nu_r \\ \mu_p \\ \nu_r \end{bmatrix}$$

And the covariance matrix is:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma^2 + \sigma_p^2 + \tau_r^2 & \sigma_p^2 & \tau_r^2 \\ \sigma_p^2 & \sigma_p^2 & 0 \\ \tau_r^2 & 0 & \tau_r^2 \end{bmatrix}$$

(ii)

Taking the conditional over $x^{(pr)}$ to get the mean vector and covariance matrix of $Q_{pr}(y^{(pr)}, z^{(pr)}) = p(y^{(pr)}, z^{(pr)} \mid x^{(pr)})$

$$Q_{pr}(y^{(pr)}, z^{(pr)}) = \det(2\pi\bar{\Sigma})^{-1/2} \exp\left(-\frac{1}{2}(\begin{bmatrix} y^{(pr)} \\ z^{(pr)} \end{bmatrix} - \bar{\mu})^T \bar{\Sigma}^{-1} (\begin{bmatrix} y^{(pr)} \\ z^{(pr)} \end{bmatrix} - \bar{\mu})\right)$$

With $\bar{\mu}, \bar{\Sigma}$ defined by:

$$\bar{\mu} := \mu_{y^{(pr)}, z^{(pr)} | x^{(pr)}} = \mu_{y^{(pr)}, z^{(pr)}} + \Sigma_{\left(y^{(pr)}, z^{(pr)}\right), x^{(pr)}} \Sigma^{-1}_{x^{(pr)}, x^{(pr)}} \left(x^{(pr)} - \mu_{x^{(pr)}}\right)$$

$$= \begin{bmatrix} \mu_p \\ \nu_r \end{bmatrix} + \begin{bmatrix} \sigma_p^2 \\ \tau_r^2 \end{bmatrix} \frac{1}{\sigma^2 + \sigma_p^2 + \tau_r^2} \left(x^{(pr)} - \mu_p - \nu_r\right)$$

$$\bar{\Sigma} := \Sigma_{\left(y^{(pr)}, z^{(pr)}\right) | x^{(pr)}} = \Sigma_{\left(y^{(pr)}, z^{(pr)}\right), \left(y^{(pr)}, z^{(pr)}\right)} - \Sigma_{\left(y^{(pr)}, z^{(pr)}\right), x^{(pr)}} \Sigma^{-1}_{x^{(pr)}, x^{(pr)}} \Sigma_{x^{(pr)}, \left(y^{(pr)}, z^{(pr)}\right)}$$

$$= \begin{bmatrix} \sigma_p^2 & 0 \\ 0 & \tau_r^2 \end{bmatrix} - \begin{bmatrix} \sigma_p^2 \\ \tau_r^2 \end{bmatrix} \frac{1}{\sigma^2 + \sigma_p^2 + \tau_r^2} \begin{bmatrix} \sigma_p^2 & \tau_r^2 \end{bmatrix}$$

(b)

$$\arg \max_{\mu_p,\nu_r,\sigma_p^2,\tau_r^2} \sum_p \sum_r \mathbb{E}_{(y^{(pr)},z^{(pr)})\sim Q_{pr}} \left[\log \left(\frac{p(y^{(pr)}, z^{(pr)}, x^{(pr)}; \mu_p, \nu_r, \sigma_p^2, \tau_r^2)}{Q_{pr}(y^{(pr)}, z^{(pr)})} \right) \right] =$$

$$= \arg \max_{\mu_p,\nu_r,\sigma_p^2,\tau_r^2} \sum_p \sum_r \mathbb{E}_{(y^{(pr)},z^{(pr)})\sim Q_{pr}} \log \left(p(y^{(pr)}, z^{(pr)}, x^{(pr)}; \mu_p, \nu_r, \sigma_p^2, \tau_r^2) \right) =$$

Where

$$p(y^{(pr)}, z^{(pr)}, x^{(pr)}; \mu_p, \nu_r, \sigma_p^2, \tau_r^2) = \det(2\pi\boldsymbol{\Sigma})^{-\frac{1}{2}} \exp\left(-\frac{1}{2} \left(\begin{bmatrix} x^{(pr)} \\ y^{(pr)} \\ z^{(pr)} \end{bmatrix} - \boldsymbol{\mu} \right)^T \boldsymbol{\Sigma}^{-1} \left(\begin{bmatrix} x^{(pr)} \\ y^{(pr)} \\ z^{(pr)} \end{bmatrix} - \boldsymbol{\mu} \right) \right)$$

The log of which is:

$$-\frac{3}{2}\log(2\pi) - \frac{1}{2}\log\left(\sigma_p^2\tau_r^2\sigma^2\right) - \frac{1}{2}\left(\frac{1}{\sigma^2}\left({z^{(pr)}}^2 + 2y^{(pr)}z^{(pr)} - 2x^{(pr)}z^{(pr)} + {y^{(pr)}}^2 - 2x^{(pr)}y^{(pr)} + {x^{(pr)}}^2 \right) \right) +$$

$$-\frac{1}{2}\left(\frac{1}{\tau_r^2}\left({z^{(pr)}}^2 - 2\nu_r z^{(pr)} + \nu_r^2 \right) + \frac{1}{\sigma_p^2}\left({y^{(pr)}}^2 - 2\mu_p y^{(pr)} + \mu_p^2 \right) \right)$$

The only parts relevant to the argmax operation are:

$$-\frac{1}{2}\left(\log\left(\sigma_p^2\tau_r^2\right) + \frac{1}{\tau_r^2}\left(z^{(pr)} - \nu_r \right)^2 + \frac{1}{\sigma_p^2}\left(y^{(pr)} - \mu_p \right)^2 \right)$$

Returning the the main expression:

$$\arg \max_{\mu_p,\nu_r,\sigma_p^2,\tau_r^2} \sum_p \sum_r \mathbb{E}_{(y^{(pr)},z^{(pr)})\sim Q_{pr}} -\frac{1}{2}\left(\log\left(\sigma_p^2\tau_r^2\right) + \frac{1}{\tau_r^2}\left(z^{(pr)} - \nu_r \right)^2 + \frac{1}{\sigma_p^2}\left(y^{(pr)} - \mu_p \right)^2 \right) =$$

The expectation of the $\frac{1}{\tau_r^2}\left(z^{(pr)} - \nu_r\right)^2$ term, using the mean and covariance matrix from (a), and using the fact that $\mathbb{E}((x-a)^2) = \mathbb{E}(x^2 - 2ax - a^2) = \mathbb{E}x^2 - 2a\mathbb{E}x - a^2 = \sigma_x^2 + \mu_x^2 - 2a\mu_x - a^2 = \sigma_x^2 + (\mu_x - a)^2$, is:

$$\frac{1}{\tau_r^2}\left(\tilde{\tau}_r^2 - \frac{\tilde{\tau}_r^4}{\sigma^2 + \tilde{\sigma}_p^2 + \tilde{\tau}_r^2} + \left(\tilde{\nu}_r + \frac{\tilde{\tau}_r^2(x^{(pr)} - \tilde{\nu}_r - \tilde{\mu}_p)}{\sigma^2 + \tilde{\sigma}_p^2 + \tilde{\tau}_r^2} - \nu_r \right)^2 \right)$$

Similiarly the expectation of the $\frac{1}{\sigma_p^2}\left(y^{(pr)} - \mu_p\right)^2$ term is:

$$\frac{1}{\sigma_p^2}\left(\tilde{\sigma}_p^2 - \frac{\tilde{\sigma}_p^4}{\sigma^2 + \tilde{\sigma}_p^2 + \tilde{\tau}_r^2} + \left(\tilde{\mu}_p + \frac{\tilde{\sigma}_p^2(x^{(pr)} - \tilde{\nu}_r - \tilde{\mu}_p)}{\sigma^2 + \tilde{\sigma}_p^2 + \tilde{\tau}_r^2} - \mu_p \right)^2 \right) =$$

The variables with a tilde denote those determined in the E-step, that are not subject to maximization. Taking these into account, and removing the $-\frac{1}{2}$ factor, and changing max to min:

$$\arg \min_{\mu_p,\nu_r,\sigma_p^2,\tau_r^2} \sum_p \sum_r \log\sigma_p^2 + \frac{1}{\sigma_p^2}\left(\tilde{\sigma}_p^2 - \frac{\tilde{\sigma}_p^4}{\sigma^2 + \tilde{\sigma}_p^2 + \tilde{\tau}_r^2} + \left(\tilde{\mu}_p + \frac{\tilde{\sigma}_p^2(x^{(pr)} - \tilde{\nu}_r - \tilde{\mu}_p)}{\sigma^2 + \tilde{\sigma}_p^2 + \tilde{\tau}_r^2} - \mu_p \right)^2 \right)$$

$$+ \log\tau_r^2 + \frac{1}{\tau_r^2}\left(\tilde{\tau}_r^2 - \frac{\tilde{\tau}_r^4}{\sigma^2 + \tilde{\sigma}_p^2 + \tilde{\tau}_r^2} + \left(\tilde{\nu}_r + \frac{\tilde{\tau}_r^2(x^{(pr)} - \tilde{\nu}_r - \tilde{\mu}_p)}{\sigma^2 + \tilde{\sigma}_p^2 + \tilde{\tau}_r^2} - \nu_r \right)^2 \right)$$

Partial derivative with respect to $\mu_{p'}$, set to 0:

$$\sum_p \sum_r \frac{1}{\sigma_p^2} \left(-2 \left(\tilde{\mu}_p + \frac{\tilde{\sigma}_p^2 (x^{(pr)} - \tilde{\nu}_r - \tilde{\mu}_p)}{\sigma^2 + \tilde{\sigma}_p^2 + \tilde{\tau}_r^2} - \mu_p \right) \mathbf{1}\{p' = p\} \right) = 0$$

$$\sum_r \sum_p \frac{1}{\sigma_p^2} \left(\tilde{\mu}_p + \frac{\tilde{\sigma}_p^2 (x^{(pr)} - \tilde{\nu}_r - \tilde{\mu}_p)}{\sigma^2 + \tilde{\sigma}_p^2 + \tilde{\tau}_r^2} - \mu_p \right) \mathbf{1}\{p' = p\} = 0$$

$$\frac{1}{\sigma_{p'}^2} \sum_r \left(\tilde{\mu}_{p'} + \frac{\tilde{\sigma}_{p'}^2 (x^{(p'r)} - \tilde{\nu}_r - \tilde{\mu}_{p'})}{\sigma^2 + \tilde{\sigma}_{p'}^2 + \tilde{\tau}_r^2} - \mu_{p'} \right) = 0$$

$$\sum_r \left(\tilde{\mu}_{p'} + \frac{\tilde{\sigma}_{p'}^2 (x^{(p'r)} - \tilde{\nu}_r - \tilde{\mu}_{p'})}{\sigma^2 + \tilde{\sigma}_{p'}^2 + \tilde{\tau}_r^2} - \mu_{p'} \right) = 0$$

$$\mu_{p'} = \frac{1}{R} \sum_r \left(\tilde{\mu}_{p'} + \frac{\tilde{\sigma}_{p'}^2 (x^{(p'r)} - \tilde{\nu}_r - \tilde{\mu}_{p'})}{\sigma^2 + \tilde{\sigma}_{p'}^2 + \tilde{\tau}_r^2} \right)$$

And reindexing from p' to p:

$$\mu_p = \frac{1}{R} \sum_{r=1}^{R} \left(\tilde{\mu}_p + \frac{\tilde{\sigma}_p^2 (x^{(pr)} - \tilde{\nu}_r - \tilde{\mu}_p)}{\sigma^2 + \tilde{\sigma}_p^2 + \tilde{\tau}_r^2} \right)$$

$$\mu_p = \tilde{\mu}_p + \frac{1}{R} \sum_{r=1}^{R} \left(\frac{\tilde{\sigma}_p^2 (x^{(pr)} - \tilde{\nu}_r - \tilde{\mu}_p)}{\sigma^2 + \tilde{\sigma}_p^2 + \tilde{\tau}_r^2} \right)$$

Partial derivative with respect to $\nu_{r'}$, set to 0:

$$\sum_p \sum_r \frac{1}{\tau_r^2} \left(-2 \left(\tilde{\nu}_r + \frac{\tilde{\tau}_r^2 (x^{(pr)} - \tilde{\nu}_r - \tilde{\mu}_p)}{\sigma^2 + \tilde{\sigma}_p^2 + \tilde{\tau}_r^2} - \nu_r \right) \mathbf{1}\{r' = r\} \right) = 0$$

$$\sum_p \sum_r \frac{1}{\tau_r^2} \left(\tilde{\nu}_r + \frac{\tilde{\tau}_r^2 (x^{(pr)} - \tilde{\nu}_r - \tilde{\mu}_p)}{\sigma^2 + \tilde{\sigma}_p^2 + \tilde{\tau}_r^2} - \nu_r \right) \mathbf{1}\{r' = r\} = 0$$

$$\frac{1}{\tau_{r'}^2} \sum_p \left(\tilde{\nu}_{r'} + \frac{\tilde{\tau}_{r'}^2 (x^{(pr')} - \tilde{\nu}_{r'} - \tilde{\mu}_p)}{\sigma^2 + \tilde{\sigma}_p^2 + \tilde{\tau}_{r'}^2} - \nu_{r'} \right) = 0$$

$$\sum_p \left(\tilde{\nu}_{r'} + \frac{\tilde{\tau}_{r'}^2 (x^{(pr')} - \tilde{\nu}_{r'} - \tilde{\mu}_p)}{\sigma^2 + \tilde{\sigma}_p^2 + \tilde{\tau}_{r'}^2} - \nu_{r'} \right) = 0$$

$$\nu_{r'} = \frac{1}{P} \sum_{p=1}^{P} \left(\tilde{\nu}_{r'} + \frac{\tilde{\tau}_{r'}^2 (x^{(pr')} - \tilde{\nu}_{r'} - \tilde{\mu}_p)}{\sigma^2 + \tilde{\sigma}_p^2 + \tilde{\tau}_{r'}^2} \right)$$

And reindexing from r' to r:

$$\nu_r = \frac{1}{P}\sum_{p=1}^{P}\left(\tilde{\nu}_r + \frac{\tilde{\tau}_r^2(x^{(pr)} - \tilde{\nu}_r - \tilde{\mu}_p)}{\sigma^2 + \tilde{\sigma}_p^2 + \tilde{\tau}_r^2}\right)$$

$$\nu_r = \tilde{\nu}_r + \frac{1}{P}\sum_{p=1}^{P}\left(\frac{\tilde{\tau}_r^2(x^{(pr)} - \tilde{\nu}_r - \tilde{\mu}_p)}{\sigma^2 + \tilde{\sigma}_p^2 + \tilde{\tau}_r^2}\right)$$

Partial derivative with respect to $\sigma_{p'}^2$, set to 0:

$$\sum_p \sum_r \mathbf{1}\{p' = p\}\left[\frac{1}{\sigma_p^2} - \frac{1}{\sigma_p^4}\left(\tilde{\sigma}_p^2 - \frac{\tilde{\sigma}_p^4}{\sigma^2 + \tilde{\sigma}_p^2 + \tilde{\tau}_r^2} + \left(\tilde{\mu}_p + \frac{\tilde{\sigma}_p^2(x^{(pr)} - \tilde{\nu}_r - \tilde{\mu}_p)}{\sigma^2 + \tilde{\sigma}_p^2 + \tilde{\tau}_r^2} - \mu_p\right)^2\right)\right] = 0$$

$$\sum_r \left(\frac{1}{\sigma_{p'}^2} - \frac{1}{\sigma_{p'}^4}\left(\tilde{\sigma}_{p'}^2 - \frac{\tilde{\sigma}_{p'}^4}{\sigma^2 + \tilde{\sigma}_{p'}^2 + \tilde{\tau}_r^2} + \left(\tilde{\mu}_{p'} + \frac{\tilde{\sigma}_{p'}^2(x^{(p'r)} - \tilde{\nu}_r - \tilde{\mu}_{p'})}{\sigma^2 + \tilde{\sigma}_{p'}^2 + \tilde{\tau}_r^2} - \mu_{p'}\right)^2\right)\right) = 0$$

$$\sum_r \left(\sigma_{p'}^2 - \left(\tilde{\sigma}_{p'}^2 - \frac{\tilde{\sigma}_{p'}^4}{\sigma^2 + \tilde{\sigma}_{p'}^2 + \tilde{\tau}_r^2} + \left(\tilde{\mu}_{p'} + \frac{\tilde{\sigma}_{p'}^2(x^{(p'r)} - \tilde{\nu}_r - \tilde{\mu}_{p'})}{\sigma^2 + \tilde{\sigma}_{p'}^2 + \tilde{\tau}_r^2} - \mu_{p'}\right)^2\right)\right) = 0$$

$$\sigma_{p'}^2 = \frac{1}{R}\sum_{r=1}^{R}\left(\tilde{\sigma}_{p'}^2 - \frac{\tilde{\sigma}_{p'}^4}{\sigma^2 + \tilde{\sigma}_{p'}^2 + \tilde{\tau}_r^2} + \left(\tilde{\mu}_{p'} + \frac{\tilde{\sigma}_{p'}^2(x^{(p'r)} - \tilde{\nu}_r - \tilde{\mu}_{p'})}{\sigma^2 + \tilde{\sigma}_{p'}^2 + \tilde{\tau}_r^2} - \mu_{p'}\right)^2\right)$$

And reindexing from p' to p:

$$\sigma_p^2 = \frac{1}{R}\sum_{r=1}^{R}\left(\tilde{\sigma}_p^2 - \frac{\tilde{\sigma}_p^4}{\sigma^2 + \tilde{\sigma}_p^2 + \tilde{\tau}_r^2} + \left(\tilde{\mu}_p + \frac{\tilde{\sigma}_p^2(x^{(pr)} - \tilde{\nu}_r - \tilde{\mu}_p)}{\sigma^2 + \tilde{\sigma}_p^2 + \tilde{\tau}_r^2} - \mu_p\right)^2\right)$$

$$\sigma_p^2 = \tilde{\sigma}_p^2 + \frac{1}{R}\sum_{r=1}^{R}\left(-\frac{\tilde{\sigma}_p^4}{\sigma^2 + \tilde{\sigma}_p^2 + \tilde{\tau}_r^2} + \left(\frac{\tilde{\sigma}_p^2(x^{(pr)} - \tilde{\nu}_r - \tilde{\mu}_p)}{\sigma^2 + \tilde{\sigma}_p^2 + \tilde{\tau}_r^2} - (\mu_p - \tilde{\mu}_p)\right)^2\right)$$

Where μ_p is the new value derived before.
Partial derivative with respect to $\tau_{r'}^2$, set to 0:

$$\sum_p \sum_r \mathbf{1}\{r' = r\}\left[\frac{1}{\tau_r^2} - \frac{1}{\tau_r^4}\left(\tilde{\tau}_r^2 - \frac{\tilde{\tau}_r^4}{\sigma^2 + \tilde{\sigma}_p^2 + \tilde{\tau}_r^2} + \left(\tilde{\nu}_r + \frac{\tilde{\tau}_r^2(x^{(pr)} - \tilde{\nu}_r - \tilde{\mu}_p)}{\sigma^2 + \tilde{\sigma}_p^2 + \tilde{\tau}_r^2} - \nu_r\right)^2\right)\right] = 0$$

$$\sum_p \left(\frac{1}{\tau_{r'}^2} - \frac{1}{\tau_{r'}^4}\left(\tilde{\tau}_{r'}^2 - \frac{\tilde{\tau}_{r'}^4}{\sigma^2 + \tilde{\sigma}_p^2 + \tilde{\tau}_{r'}^2} + \left(\tilde{\nu}_{r'} + \frac{\tilde{\tau}_{r'}^2(x^{(pr')} - \tilde{\nu}_{r'} - \tilde{\mu}_p)}{\sigma^2 + \tilde{\sigma}_p^2 + \tilde{\tau}_{r'}^2} - \nu_{r'}\right)^2\right)\right) = 0$$

$$\sum_p \left(\tau_{r'}^2 - \left(\tilde{\tau}_{r'}^2 - \frac{\tilde{\tau}_{r'}^4}{\sigma^2 + \tilde{\sigma}_p^2 + \tilde{\tau}_{r'}^2} + \left(\tilde{\nu}_{r'} + \frac{\tilde{\tau}_{r'}^2(x^{(pr')} - \tilde{\nu}_{r'} - \tilde{\mu}_p)}{\sigma^2 + \tilde{\sigma}_p^2 + \tilde{\tau}_{r'}^2} - \nu_{r'}\right)^2\right)\right) = 0$$

$$\tau_{r'}^2 = \frac{1}{P}\sum_{p=1}^{P}\left(\tilde{\tau}_{r'}^2 - \frac{\tilde{\tau}_{r'}^4}{\sigma^2+\tilde{\sigma}_p^2+\tilde{\tau}_{r'}^2} + \left(\tilde{\nu}_{r'} + \frac{\tilde{\tau}_{r'}^2(x^{(pr')} - \tilde{\nu}_{r'} - \tilde{\mu}_p)}{\sigma^2+\tilde{\sigma}_p^2+\tilde{\tau}_{r'}^2} - \nu_{r'}\right)^2\right)$$

And reindexing from r' to r:

$$\tau_r^2 = \frac{1}{P}\sum_{p=1}^{P}\left(\tilde{\tau}_r^2 - \frac{\tilde{\tau}_r^4}{\sigma^2+\tilde{\sigma}_p^2+\tilde{\tau}_r^2} + \left(\tilde{\nu}_r + \frac{\tilde{\tau}_r^2(x^{(pr)} - \tilde{\nu}_r - \tilde{\mu}_p)}{\sigma^2+\tilde{\sigma}_p^2+\tilde{\tau}_r^2} - \nu_r\right)^2\right)$$

$$\tau_r^2 = \tilde{\tau}_r^2 + \frac{1}{P}\sum_{p=1}^{P}\left(-\frac{\tilde{\tau}_r^4}{\sigma^2+\tilde{\sigma}_p^2+\tilde{\tau}_r^2} + \left(\frac{\tilde{\tau}_r^2(x^{(pr)} - \tilde{\nu}_r - \tilde{\mu}_p)}{\sigma^2+\tilde{\sigma}_p^2+\tilde{\tau}_r^2} - (\nu_r - \tilde{\nu}_r)\right)^2\right)$$

Where ν_r is the new value derived before.
Rewritten in terms of updates to the parameters:

$$\Delta\mu_p \leftarrow \frac{1}{R}\sum_{r=1}^{R}\left(\frac{\sigma_p^2(x^{(pr)} - \nu_r - \mu_p)}{\sigma^2+\sigma_p^2+\tau_r^2}\right)$$

$$\Delta\nu_r \leftarrow \frac{1}{P}\sum_{p=1}^{P}\left(\frac{\tau_r^2(x^{(pr)} - \nu_r - \mu_p)}{\sigma^2+\sigma_p^2+\tau_r^2}\right)$$

$$\sigma_p^{2^{(\text{new})}} \leftarrow \sigma_p^2 + \frac{1}{R}\sum_{r=1}^{R}\left(-\frac{\sigma_p^4}{\sigma^2+\sigma_p^2+\tau_r^2} + \left(\frac{\sigma_p^2(x^{(pr)} - \nu_r - \mu_p)}{\sigma^2+\sigma_p^2+\tau_r^2} - \Delta\mu_p\right)^2\right)$$

$$\tau_r^{2^{(\text{new})}} \leftarrow \tau_r^2 + \frac{1}{P}\sum_{p=1}^{P}\left(-\frac{\tau_r^4}{\sigma^2+\sigma_p^2+\tau_r^2} + \left(\frac{\tau_r^2(x^{(pr)} - \nu_r - \mu_p)}{\sigma^2+\sigma_p^2+\tau_r^2} - \Delta\nu_r\right)^2\right)$$

$$\mu_p^{(\text{new})} \leftarrow \mu_p + \Delta\mu_p$$

$$\nu_r^{(\text{new})} \leftarrow \nu_r + \Delta\nu_r$$

4 KL divergence and Maximum Likelihood

(a)

$$0 = -\log 1 = -\log\sum_x Q(x) = -\log\sum_x P(x)\frac{Q(x)}{P(x)} = -\log\mathbb{E}_P\frac{Q(x)}{P(x)} \le \mathbb{E}_P\left(-\log\frac{Q(x)}{P(x)}\right) =$$

$$= -\sum_x P(x)\log\frac{Q(x)}{P(x)} = \sum_x P(x)\log\frac{P(x)}{Q(x)} = KL(P||Q)$$

The inequality is due to $-\log(x)$ being a convex function and the use of Jensen's inequality, resulting in $KL(P||Q) \ge 0$ for any P,Q.

Also by Jensen's inequality, is that equality holds if and only if the variable in question, in this case, $Q(x)/P(x)$ is a constant.

$$\forall x,\ Q(x)/P(x) = c$$

$$Q(x) = cP(x)$$

$$1 = \sum_x Q(x) = \sum_x cP(x) = c\sum_x P(x) = c$$

$$\text{therefore } \forall x,\ Q(x) = P(x) \text{ iff. the equality holds}$$

When the equality holds, $KL(P||Q) = 0$, and this will occur if and only if P=Q.

(b)

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

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

$$= \sum_x P(x) \log \frac{P(x)}{Q(x)} \left(\sum_y P(y \mid x) \right) + \sum_x P(x) \left(\sum_y P(y \mid x) \log \frac{P(y \mid x)}{Q(y \mid x)} \right) =$$

$$= \sum_x P(x) \log \frac{P(x)}{Q(x)} + \sum_x P(x) \left(\sum_y P(y \mid x) \log \frac{P(y \mid x)}{Q(y \mid x)} \right) =$$

$$= KL(P(X)||Q(X)) + KL(P(Y \mid X)||Q(Y \mid X))$$

(c)

$$\arg\min_\theta KL(\hat{P}||P_\theta) = \arg\min_\theta \sum_x \hat{P}(x) \log \frac{\hat{P}(x)}{P_\theta(x)} =$$

$$= \arg\min_\theta \sum_x \left(\hat{P}(x) \log \hat{P}(x) - \hat{P}(x) \log P_\theta(x) \right) = \arg\min_\theta \sum_x \left(-\hat{P}(x) \log P_\theta(x) \right) =$$

$$= \arg\max_\theta \sum_x \hat{P}(x) \log P_\theta(x) = \arg\max_\theta \sum_x \frac{1}{m} \sum_{i=1}^m \mathbf{1}\{x^{(i)} = x\} P_\theta(x) =$$

$$= \arg\max_\theta \frac{1}{m} \sum_{i=1}^m \left(\sum_x \mathbf{1}\{x^{(i)} = x\} P_\theta(x) \right) = \arg\max_\theta \frac{1}{m} \sum_{i=1}^m P_\theta(x^{(i)}) =$$

$$= \arg\max_\theta \sum_{i=1}^m P_\theta(x^{(i)})$$

5 K-means for compression

(a)

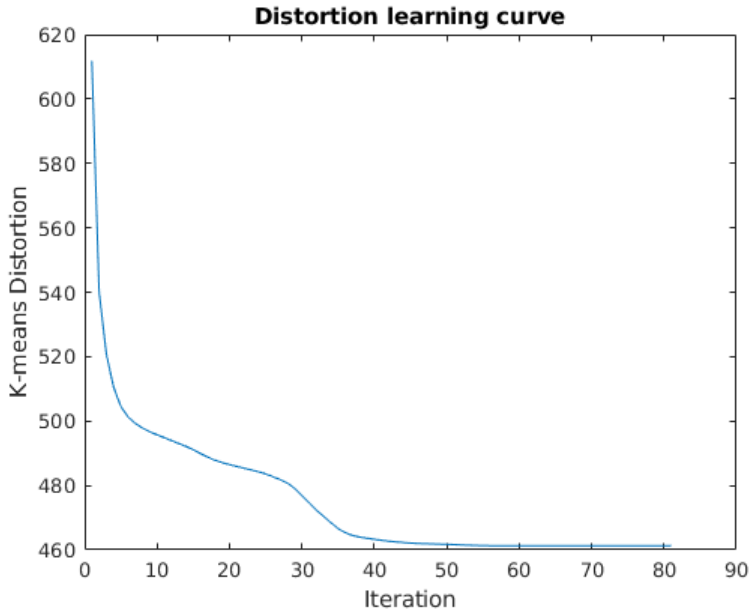
Entering:

```
A = double(imread('mandrill-large.tiff'));
imshow(uint8(round(A)));
```

Resulted in displaying the image:

(b)

K-means with K=16 was implemented and followed the following learning curve:



Convergence was reached in 81 iterations (time to convergence may vary due to random initialization).

(c)

This is the original image.

This is the image after applying k-means to reduce the amount of colors in it. The image lost some of its crispness, and compression artifacts can be seen in the colors of the eyes, and in the yellowish region on the bottom left that turned into a solid color.

```
A_large = double(imread('mandrill-large.tiff'));
%imshow(uint8(round(A_large)));


%%
A_small = double(imread('mandrill-small.tiff'));
%imshow(uint8(round(A_small)));
```

```
%%
n_channels = 3;
C = reshape(A_small,[],n_channels);
n_pix = size(C,1);

K = 16;
mu = C(randi(n_pix,K,1),:);
class = zeros(n_pix,1);
T = 1000;
J = zeros(T,1);
for iter = 1:T
    class_old = class(:);
    for i = 1:n_pix
        min_dist = Inf;
        for j = 1:K
            d = norm(C(i,:) - mu(j,:));
            if d < min_dist
                min_dist = d;
                class(i) = j;
            end
        end
    end
    for j = 1:K
        mu(j,:) = mean(C(class==j,:));
    end
    J(iter) = mean(sum((C - mu(class,:)).^2,2));
    if all(class_old == class)
        disp('Converged');
        break;
    end
end
J = J(1:iter);
%%
[Nx, Ny, n_channels] = size(A_large);
A_compressed = zeros(size(A_large));
for x = 1:Nx
    for y = 1:Ny
        color = reshape(A_large(x,y,:),1,[]);
        min_dist = Inf;
        for j = 1:K
            d = norm(color - mu(j,:));
            if d < min_dist
                min_dist = d;
                A_compressed(x,y,:) = mu(j,:);
            end
        end
    end
end
imshow(uint8(round(A_compressed)));
imwrite(uint8(round(A_compressed)), 'kmeans_image.tiff');
```

(d)

If the original image had N pixels, each represented by 24 bits (8 bits per channel, 3 channels), then the total size is $24N$ bits. The compressed image with 16 colors can be represented by a 4 bit number for each pixel, where the numbers from 0 to 15 each represent one of the colors, resulting in $4N$ bits. The values of the corresponding colors must be encoded too, 16×24 bits, but if N is large then this can be neglected in the ratio. The approximate factor is $\frac{24N}{4N} = 6$. Compression by a factor of 6.