1.

$$p(x|\pi_k, \mu_k, \Sigma) = \sum_{k=1}^{K} \pi_k N(x|\mu_k, \Sigma)$$

$$\Theta = \{\pi_k, \mu_k, \Sigma\}$$

Finding log likelihood:

$$p(X|\pi, \mu, \Sigma) = \prod_{n=1}^{N} \sum_{k=1}^{K} \pi_k N(x_n|\mu_k, \Sigma)$$

$$ln\big(p(X|\pi, \mu, \Sigma)\big) = \sum_{n=1}^{N} ln\left(\sum_{k=1}^{K} \pi_k N(x_n|\mu_k, \Sigma)\right)$$

$$= \sum_{n=1}^{N} ln\left(\sum_{k=1}^{K} \pi_k N(x_n|\mu_k, \Sigma)\right)$$

$$= \sum_{n=1}^{N} ln\left(\sum_{k=1}^{K} \pi_k \left(\frac{1}{\sqrt{2\pi^K|\Sigma|}}\right) \exp(-0.5(x_n - \mu_k)^T \Sigma^{-1}(x_n - \mu_k))\right)$$

$$= \sum_{n=1}^{N} ln\left(\left(\frac{1}{\sqrt{2\pi^K|\Sigma|}}\right)\left(\sum_{k=1}^{K} \pi_k \exp(-0.5(x_n - \mu_k)^T \Sigma^{-1}(x_n - \mu_k))\right)\right)$$

$$= \sum_{n=1}^{N} ln\left(\sum_{k=1}^{K} \pi_k \exp(-0.5(x_n - \mu_k)^T \Sigma^{-1}(x_n - \mu_k))\right) - \sum_{n=1}^{N} ln\left(\sqrt{2\pi^K|\Sigma|}\right)$$

Finding responsibility:

$$\gamma(z_k) = p(z_k = 1|x) = \frac{p(z_k = 1|x)p(x|z_k = 1)}{p(x)}$$

$$= \frac{\pi_k \left(\frac{1}{\sqrt{2\pi^K|\Sigma|}}\right) \exp(-0.5(x_n - \mu_k)^T \Sigma^{-1}(x_n - \mu_k))}{\sum_{i=1}^{K} \pi_k \left(\frac{1}{\sqrt{2\pi^K|\Sigma|}}\right) \exp(-0.5(x_n - \mu_i)^T \Sigma^{-1}(x_n - \mu_i))}$$

$$= \frac{\pi_k \exp(-0.5(x_n - \mu_k)^T \Sigma^{-1}(x_n - \mu_k))}{\sum_{i=1}^{K} \pi_i \exp(-0.5(x_n - \mu_i)^T \Sigma^{-1}(x_n - \mu_i))}$$

Taking derivatives:

$$= \sum_{n=1}^{N} ln\left(\sum_{k=1}^{K} \pi_{nk} \exp\left(-\frac{(x_n - \mu_{nk})^2}{2\Sigma^2}\right)\right) - \sum_{n=1}^{N} ln\left(\Sigma\sqrt{2\pi}\right)$$

$$\frac{\partial ln\big(p(X|\pi,\mu,\Sigma)\big)}{\partial \mu_k} = \sum_{n=1}^{N} \frac{\pi_k \exp(-0.5(x_n - \mu_k)^T \Sigma^{-1}(x_n - \mu_k))}{\sum_{i=1}^{K} \pi_i \exp(-0.5(x_n - \mu_i)^T \Sigma^{-1}(x_n - \mu_i))} (-\Sigma^{-1}(x_n - \mu_k))$$

$$0 = \sum_{n=1}^{N} \frac{\pi_k \exp(-0.5(x_n - \mu_k)^T \Sigma^{-1}(x_n - \mu_k))}{\sum_{i=1}^{K} \pi_i \exp(-0.5(x_n - \mu_i)^T \Sigma^{-1}(x_n - \mu_i))} (-\Sigma^{-1}(x_n - \mu_k))$$

$$0 = \sum_{n=1}^{N} \frac{\pi_k \left(\frac{1}{\sqrt{2\pi^K |\Sigma|}}\right) \exp\big(-0.5(x_n - \mu_k)^T \Sigma^{-1}(x_n - \mu_k)\big)}{\sum_{i=1}^{K} \pi_i \left(\frac{1}{\sqrt{2\pi^K |\Sigma|}}\right) \exp\big(-0.5(x_n - \mu_i)^T \Sigma^{-1}(x_n - \mu_i)\big)} (x_n - \mu_k)$$

$$0 = \sum_{n=1}^{N} \frac{\pi_k \left(\frac{1}{\sqrt{2\pi^K |\Sigma|}}\right) \exp(-0.5(x_n - \mu_k)^T \Sigma^{-1}(x_n - \mu_k))}{\sum_{i=1}^{K} \pi_i \left(\frac{1}{\sqrt{2\pi^K |\Sigma|}}\right) \exp(-0.5(x_n - \mu_i)^T \Sigma^{-1}(x_n - \mu_i))} (x_n - \mu_k)$$

$$\mu_k = \frac{\sum_{n=1}^{N} \gamma(z_{nk})\, x_n}{\sum_{n=1}^{N} \gamma(z_{nk})}$$

$$\frac{\partial ln\big(p(X|\pi,\mu,\Sigma)\big)}{\partial \Sigma}$$
$$= \sum_{n=1}^{N} \left(\frac{\sum_{i=1}^{K} \pi_i \exp\big(-0.5(x_n - \mu_i)^T \Sigma^{-1}(x_n - \mu_i)\big)\,(-0.5\Sigma^{-T}(x_n - \mu_i)^T (x_n - \mu_i)\Sigma^{-T})}{\sum_{k=1}^{K} \pi_k \exp\big(-0.5(x_n - \mu_k)^T \Sigma^{-1}(x_n - \mu_k)\big)}\right)$$
$$- \frac{N(2\pi^K)^{0.5} 0.5 |\Sigma^{0.5}| \Sigma^{-T}}{\sqrt{2\pi^K |\Sigma|}}$$

$$\frac{\partial ln\big(p(X|\pi,\mu,\Sigma)\big)}{\partial \Sigma}$$

$$= \sum_{n=1}^{N}\left(\frac{\sum_{i=1}^{K}\pi_i\left(\frac{1}{\sqrt{2\pi^K|\Sigma|}}\right)\exp\big(-0.5(x_n-\mu_i)^T\Sigma^{-1}(x_n-\mu_i)\big)(-0.5\Sigma^{-T}(x_n-\mu_i)^T(x_n-\mu_i)\Sigma^{-T})}{\sum_{k=1}^{K}\pi_k\left(\frac{1}{\sqrt{2\pi^K|\Sigma|}}\right)\exp\big(-0.5(x_n-\mu_k)^T\Sigma^{-1}(x_n-\mu_k)\big)}\right.$$

$$\left.-\frac{(2\pi^K)^{0.5}0.5|\Sigma^{0.5}|\Sigma^{-T}}{\sqrt{2\pi^K|\Sigma|}}\right)$$

$$0 = \sum_{n=1}^{N}\left(\left(\sum_{k=1}^{K}\gamma(z_{nk})(-0.5\Sigma^{-T}(x_n-\mu_k)^T(x_n-\mu_k)\Sigma^{-T})\right)\right)-\frac{N(2\pi^K)^{0.5}0.5|\Sigma^{0.5}|\Sigma^{-T}}{\sqrt{2\pi^K|\Sigma|}}$$

$$0 = \sum_{n=1}^{N}\left(\left(\sum_{k=1}^{K}\gamma(z_{nk})(-(x_n-\mu_k)^T(x_n-\mu_k)\Sigma^{-T})\right)\right)-\frac{N(2\pi^K)^{0.5}|\Sigma^{0.5}|}{\sqrt{2\pi^K|\Sigma|}}$$

$$0 = \sum_{n=1}^{N}\left(\left(\sum_{k=1}^{K}\gamma(z_{nk})(-(x_n-\mu_k)^T(x_n-\mu_k)\Sigma^{-T})\right)\right)-\frac{N(2\pi^K)^{0.5}|\Sigma^{0.5}|}{(2\pi^K)^{0.5}|\Sigma|^{0.5}}$$

Suppose there is some matrix A such that $A^2 = \Sigma$. $\det(AA)$
$= \det(A)\det(A)$. *Then* $|\Sigma^{0.5}| = \det(A)$ *and* $|\Sigma|^{0.5} = \det(AA)^{0.5}$
$= (\det(A)\det(A))^{0.5} = \det(A)$, *so they are equal.*

$$0 = \sum_{n=1}^{N}\left(\left(\sum_{k=1}^{K}\gamma(z_{nk})(-(x_n-\mu_k)^T(x_n-\mu_k)\Sigma^{-T})\right)\right)-N$$

$$0 = \sum_{n=1}^{N}\left(\left(\sum_{k=1}^{K}\gamma(z_{nk})(-(x_n-\mu_k)^T(x_n-\mu_k))\right)\right)-{}^{N}/_{\Sigma^{-T}}$$

$${}^{N}/_{\Sigma^{-T}} = \sum_{n=1}^{N}\left(\left(\sum_{k=1}^{K}\gamma(z_{nk})(-(x_n-\mu_k)^T(x_n-\mu_k))\right)\right)$$

$$\Sigma = \left(\frac{N}{\sum_{n=1}^{N}\left(\left(\sum_{k=1}^{K}\gamma(z_{nk})(-(x_n-\mu_k)^T(x_n-\mu_k))\right)\right)}\right)^{-T}$$

As per the textbook lambda is added to the equation as a constraint

$$\frac{\partial ln\big(p(X|\pi,\mu,\Sigma)\big) - \lambda(\sum_{k=1}^{K}\pi_k - 1)}{\partial \pi_k} = \sum_{n=1}^{N}\frac{\exp(-0.5(x_n-\mu_k)^T\Sigma^{-1}(x_n-\mu_k))}{\sum_{i=1}^{K}\pi_i\exp(-0.5(x_n-\mu_i)^T\Sigma^{-1}(x_n-\mu_i))} - \lambda$$

$$0 = \sum_{n=1}^{N}\frac{\exp(-0.5(x_n-\mu_k)^T\Sigma^{-1}(x_n-\mu_k))}{\sum_{i=1}^{K}\pi_i\exp(-0.5(x_n-\mu_i)^T\Sigma^{-1}(x_n-\mu_i))} - \lambda$$

$$0 = \sum_{n=1}^{N}\frac{\pi_k\exp(-0.5(x_n-\mu_k)^T\Sigma^{-1}(x_n-\mu_k))}{\sum_{i=1}^{K}\pi_i\exp(-0.5(x_n-\mu_i)^T\Sigma^{-1}(x_n-\mu_i))} - \lambda\pi_k$$

$$0 = \sum_{k=1}^{K}\left(\sum_{n=1}^{N}\frac{\pi_k\exp(-0.5(x_n-\mu_k)^T\Sigma^{-1}(x_n-\mu_k))}{\sum_{i=1}^{K}\pi_i\exp(-0.5(x_n-\mu_i)^T\Sigma^{-1}(x_n-\mu_i))} - \lambda\pi_k\right)$$

Because all π sums to 1, the sum of $\lambda\pi_k = \lambda$

$$0 = N - \lambda$$

$$\lambda = -N$$

$$\frac{\partial ln\big(p(X|\pi,\mu,\Sigma)\big) - \lambda(\sum_{k=1}^{K}\pi_k - 1)}{\partial \pi_k} = \sum_{n=1}^{N}\frac{\exp(-0.5(x_n-\mu_k)^T\Sigma^{-1}(x_n-\mu_k))}{\sum_{i=1}^{K}\pi_i\exp(-0.5(x_n-\mu_i)^T\Sigma^{-1}(x_n-\mu_i))} - N$$

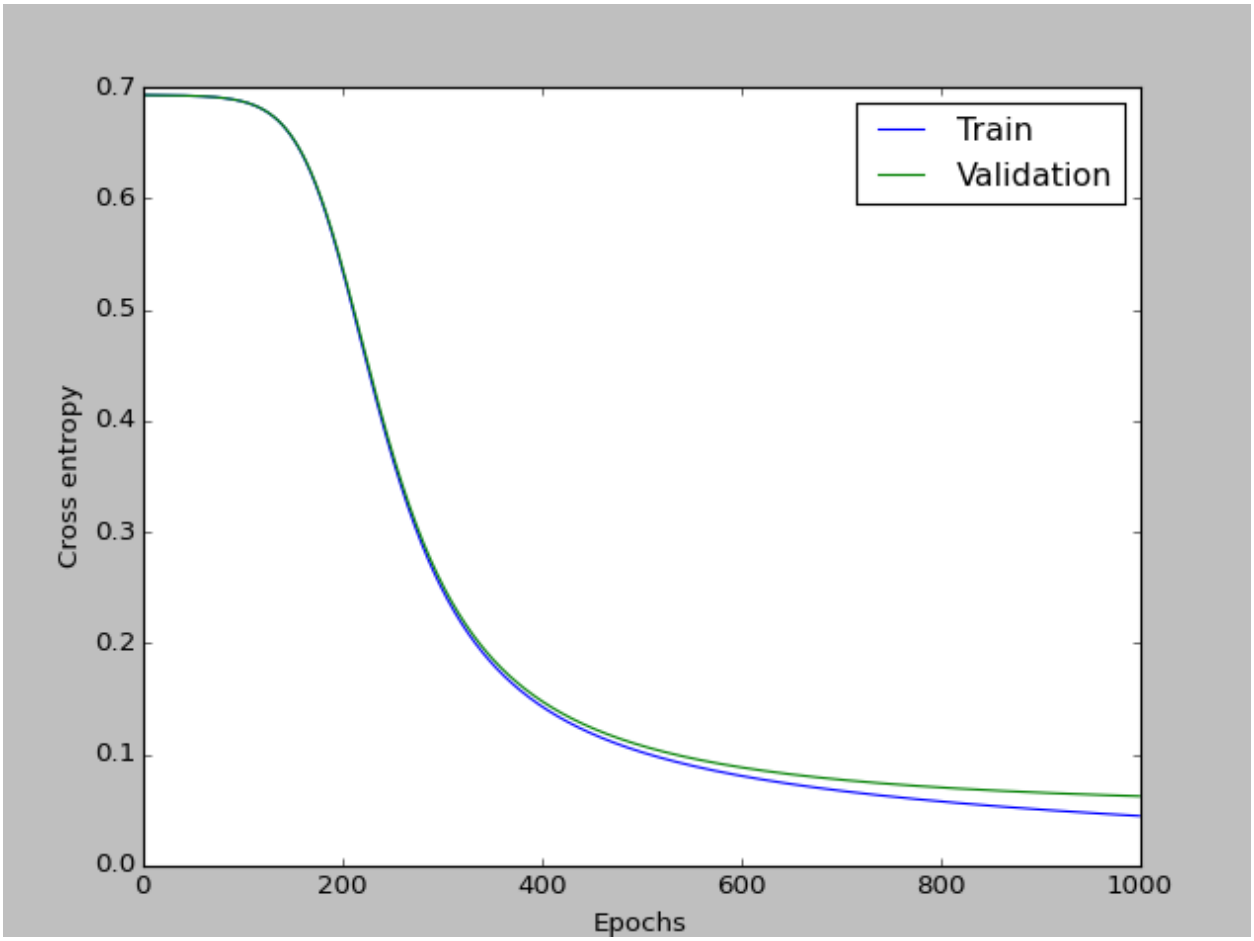
$$N\pi_k = \sum_{n=1}^{N}\frac{\pi_k\exp(-0.5(x_n-\mu_k)^T\Sigma^{-1}(x_n-\mu_k))}{\sum_{i=1}^{K}\pi_i\exp(-0.5(x_n-\mu_i)^T\Sigma^{-1}(x_n-\mu_i))}$$

$$N\pi_k = \sum_{n=1}^{N}\frac{\pi_k\left(\frac{1}{\sqrt{2\pi^K|\Sigma|}}\right)\exp(-0.5(x_n-\mu_k)^T\Sigma^{-1}(x_n-\mu_k))}{\sum_{i=1}^{K}\pi_i\left(\frac{1}{\sqrt{2\pi^K|\Sigma|}}\right)\exp(-0.5(x_n-\mu_i)^T\Sigma^{-1}(x_n-\mu_i))}$$

$$N\pi_k = \sum_{n=1}^{N}\gamma(z_{nk})$$

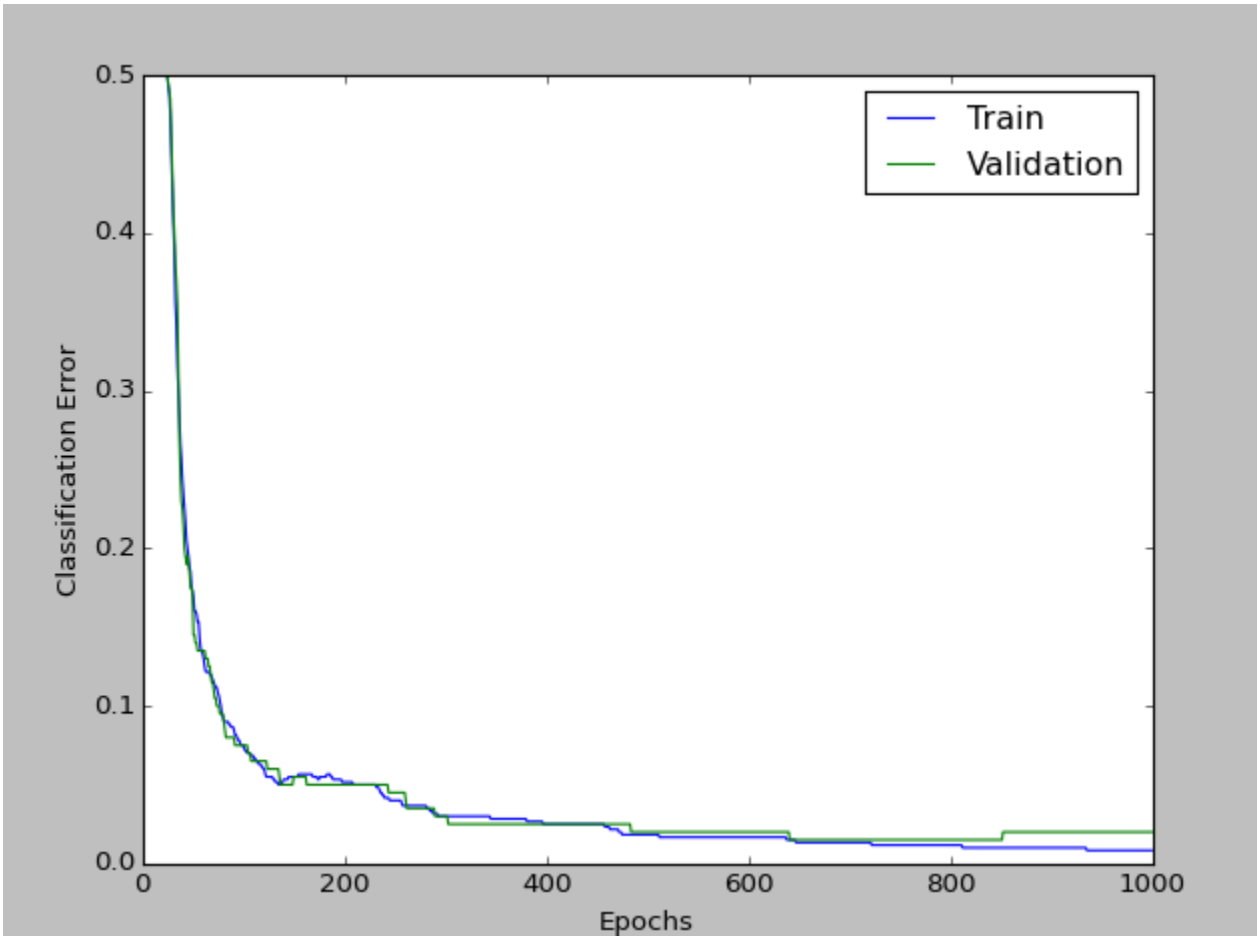
$$\pi_k = \frac{\sum_{n=1}^{N} \gamma(z_{nk})}{N}$$

2.1



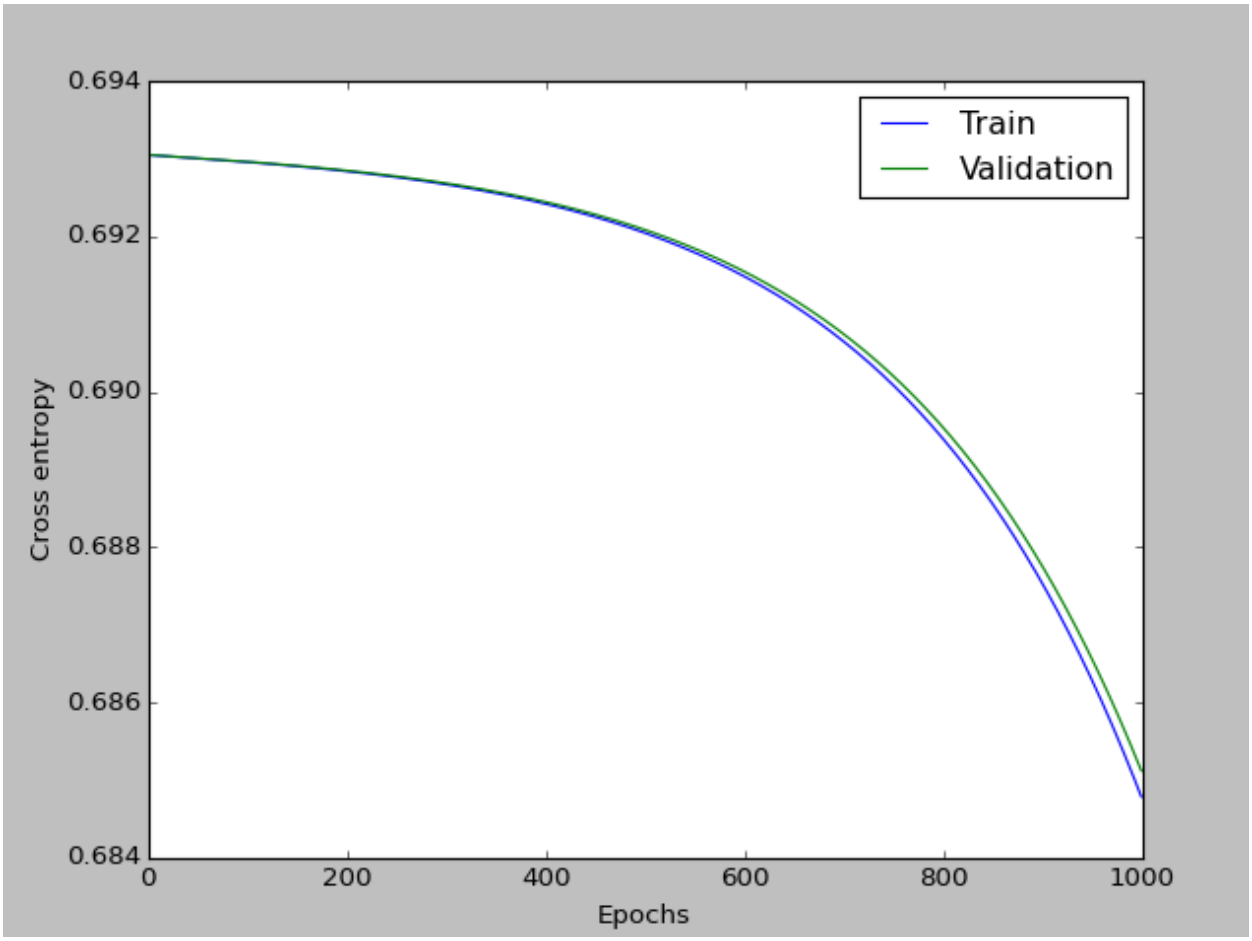
Training and validation cross entropy follow roughly the same trajectory.

2.2

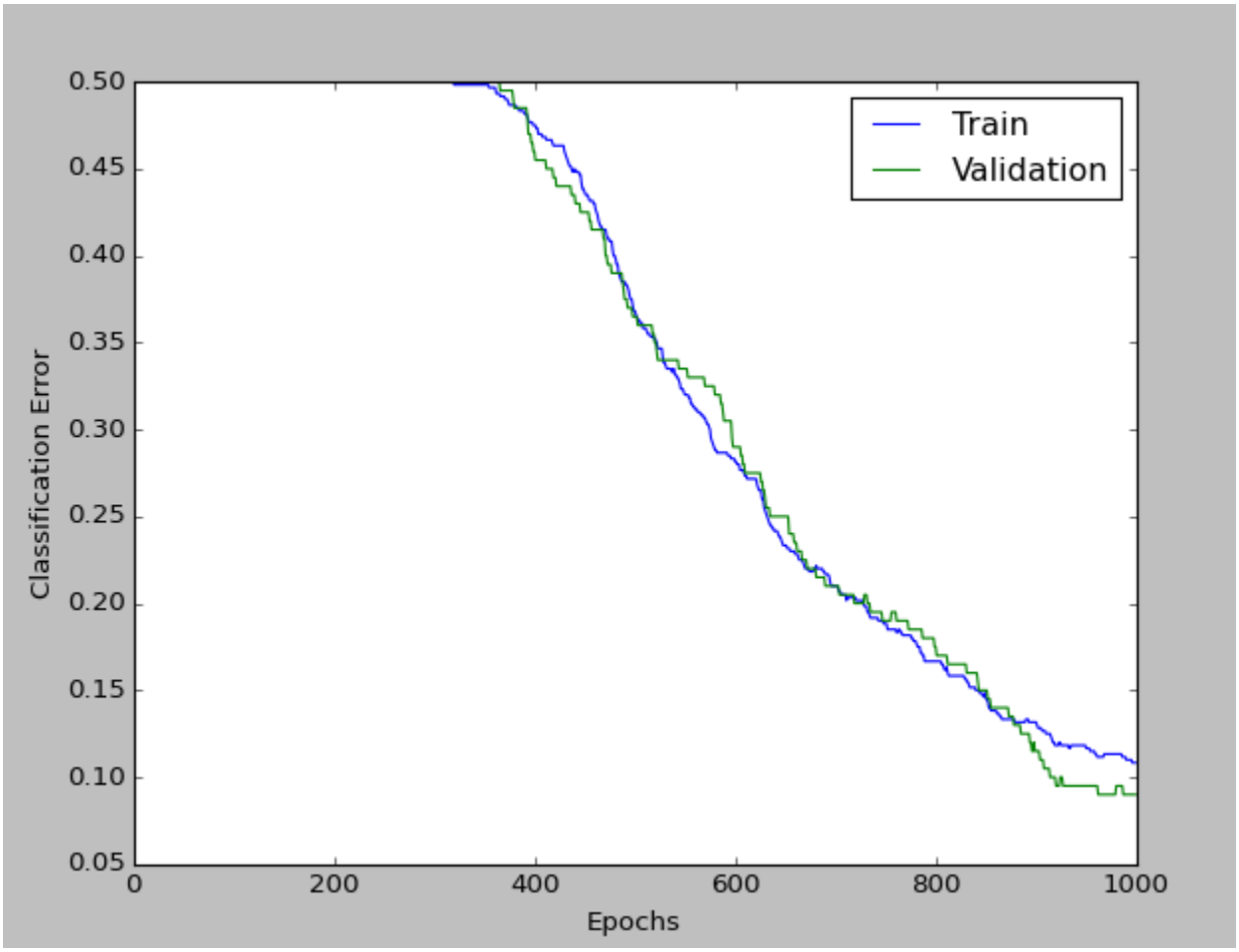

Train
Validation
Classification Error
Epochs
0.0
0.1
0.2
0.3
0.4
0.5
0
200
400
600
800
1000

2.3

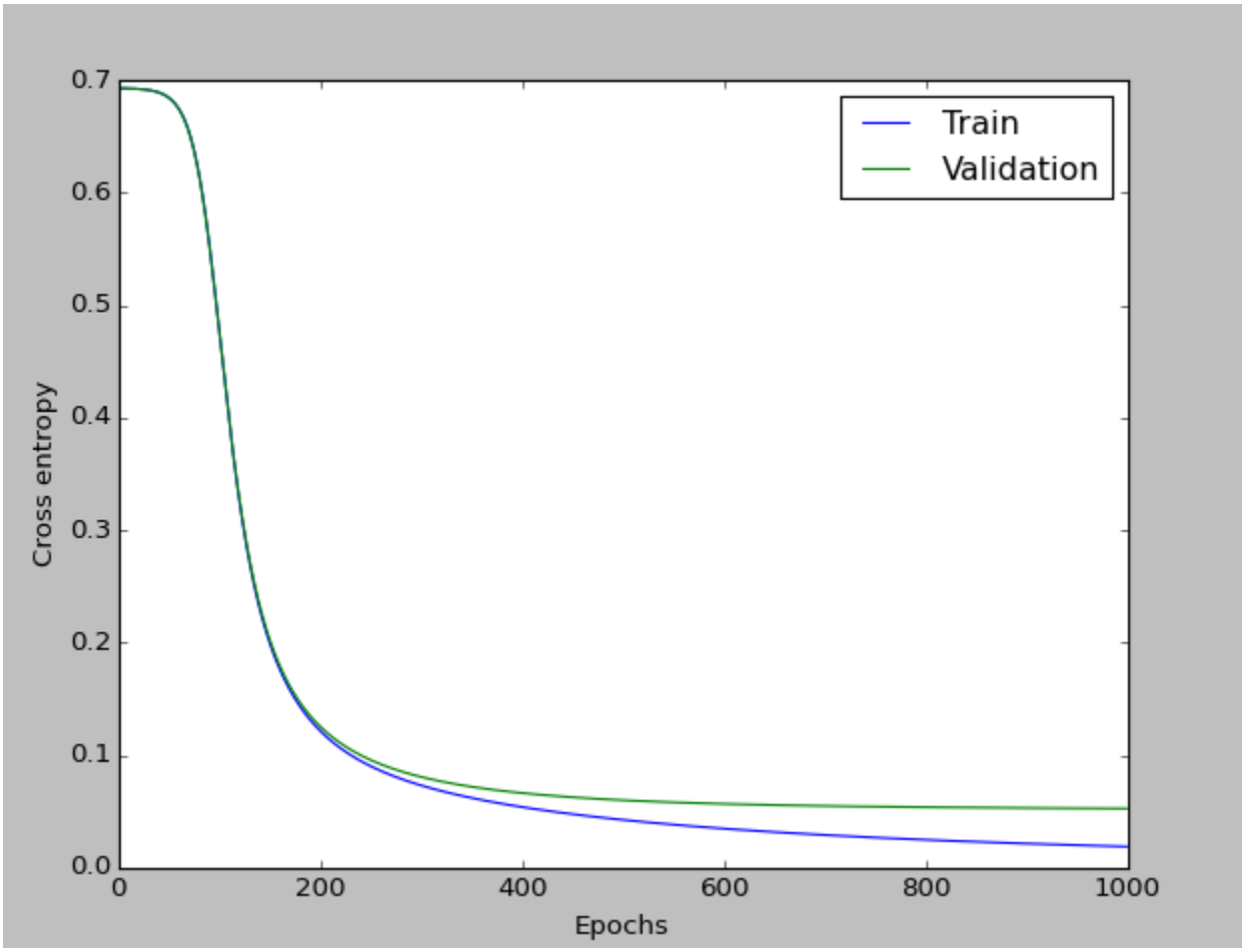
$$\epsilon = 0.01, momentum = 0$$



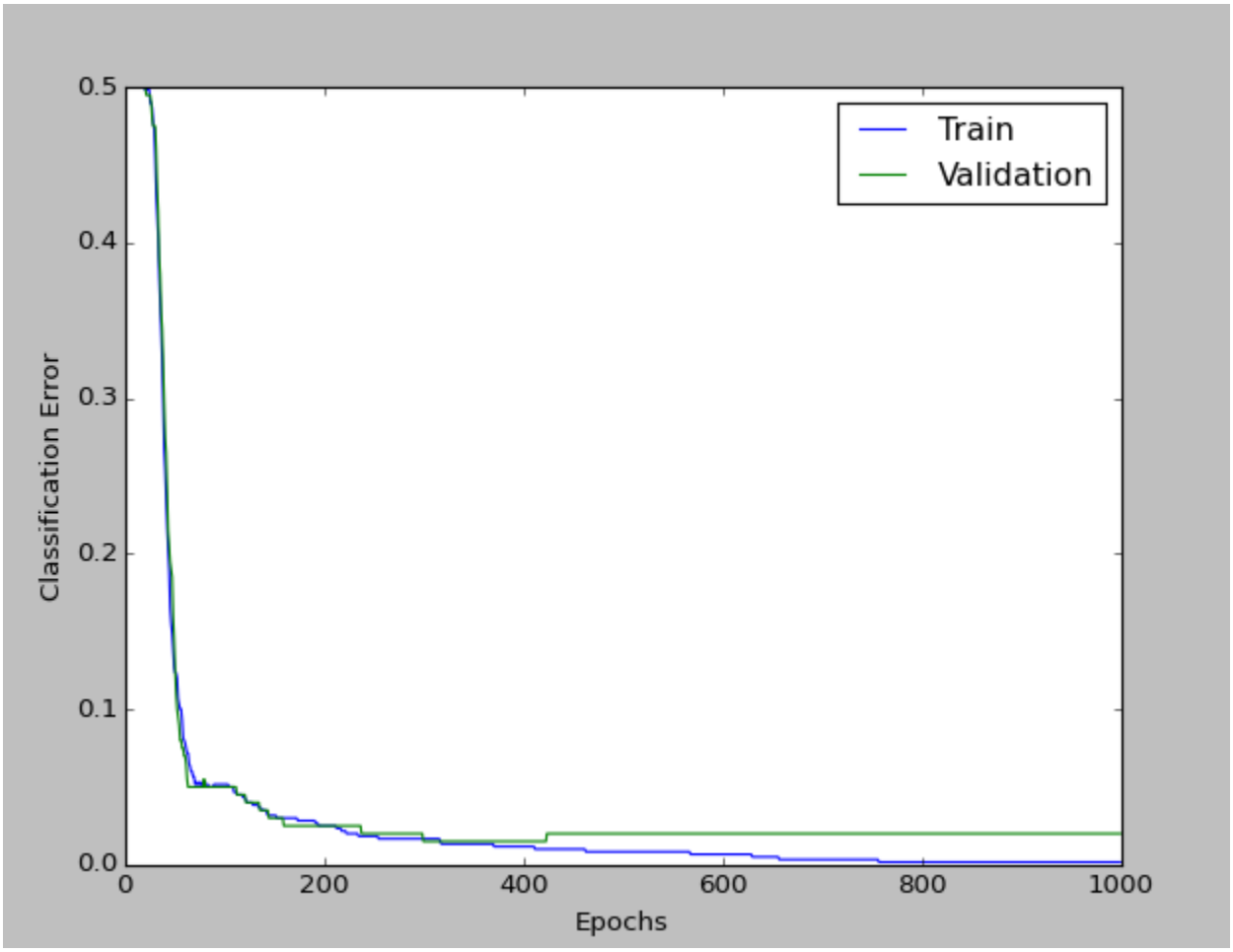
$\epsilon = 0.01, momentum = 0$



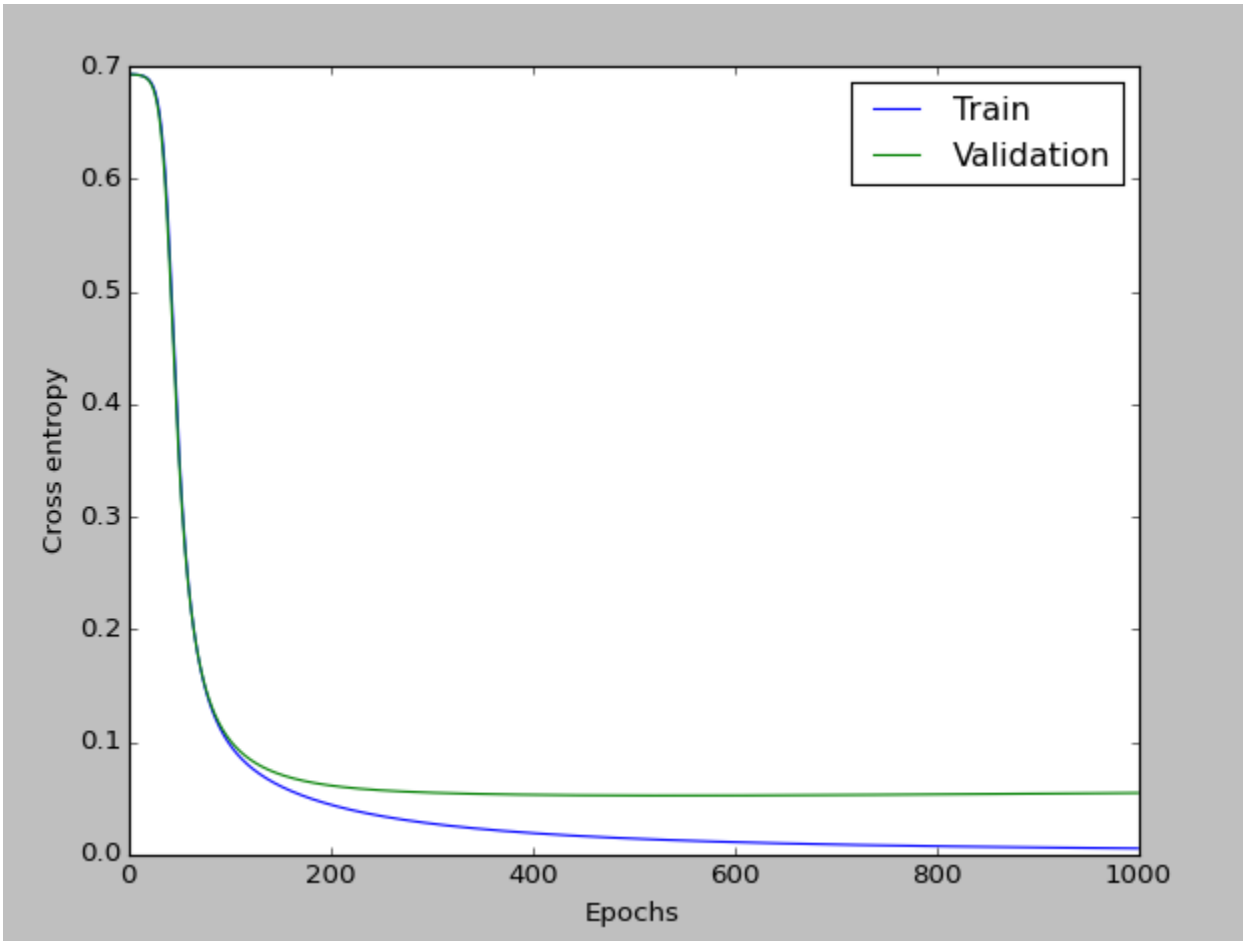
$\epsilon = 0.2, momentum = 0$



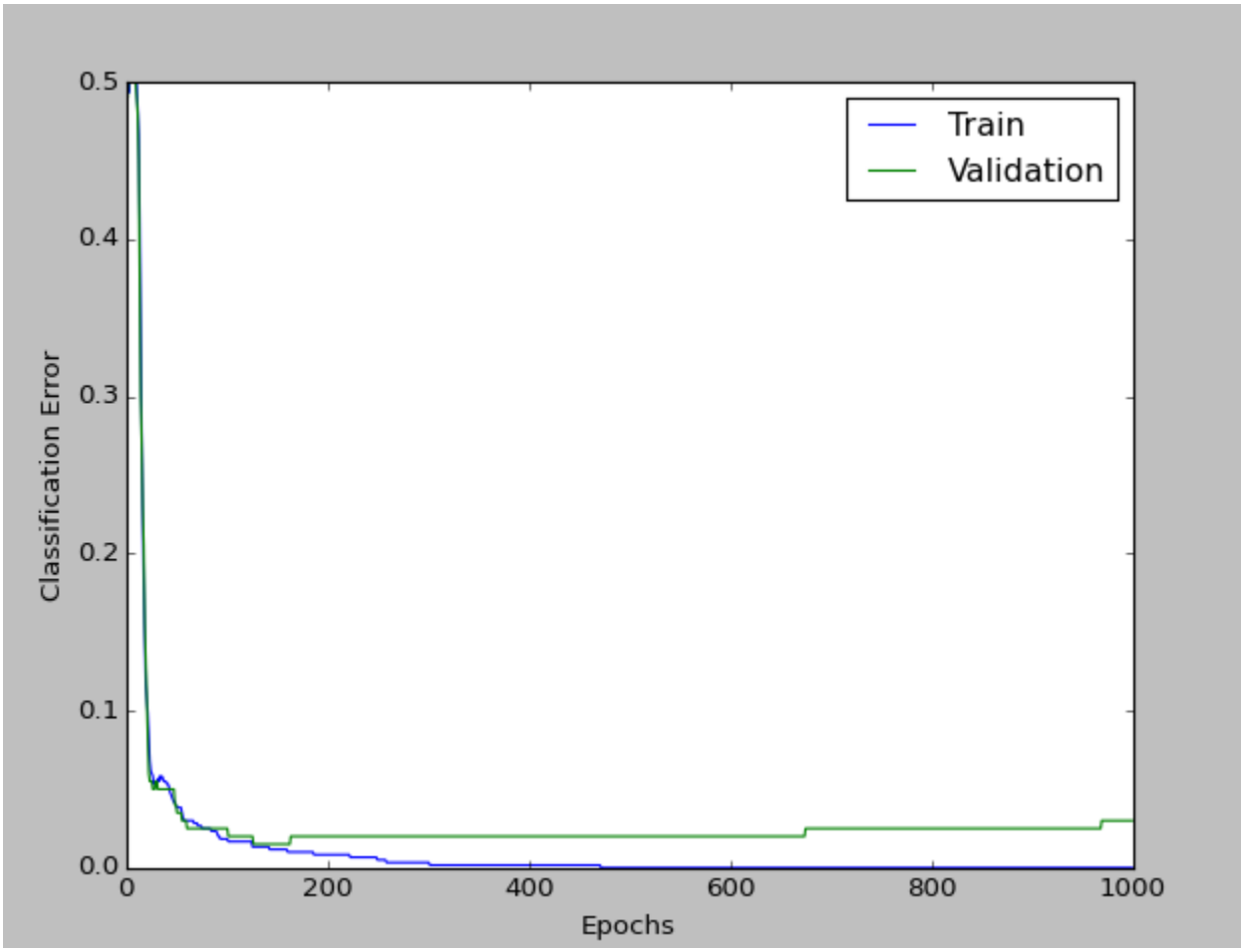
$\epsilon = 0.2, momentum = 0$



$\epsilon = 0.5, momentum = 0$

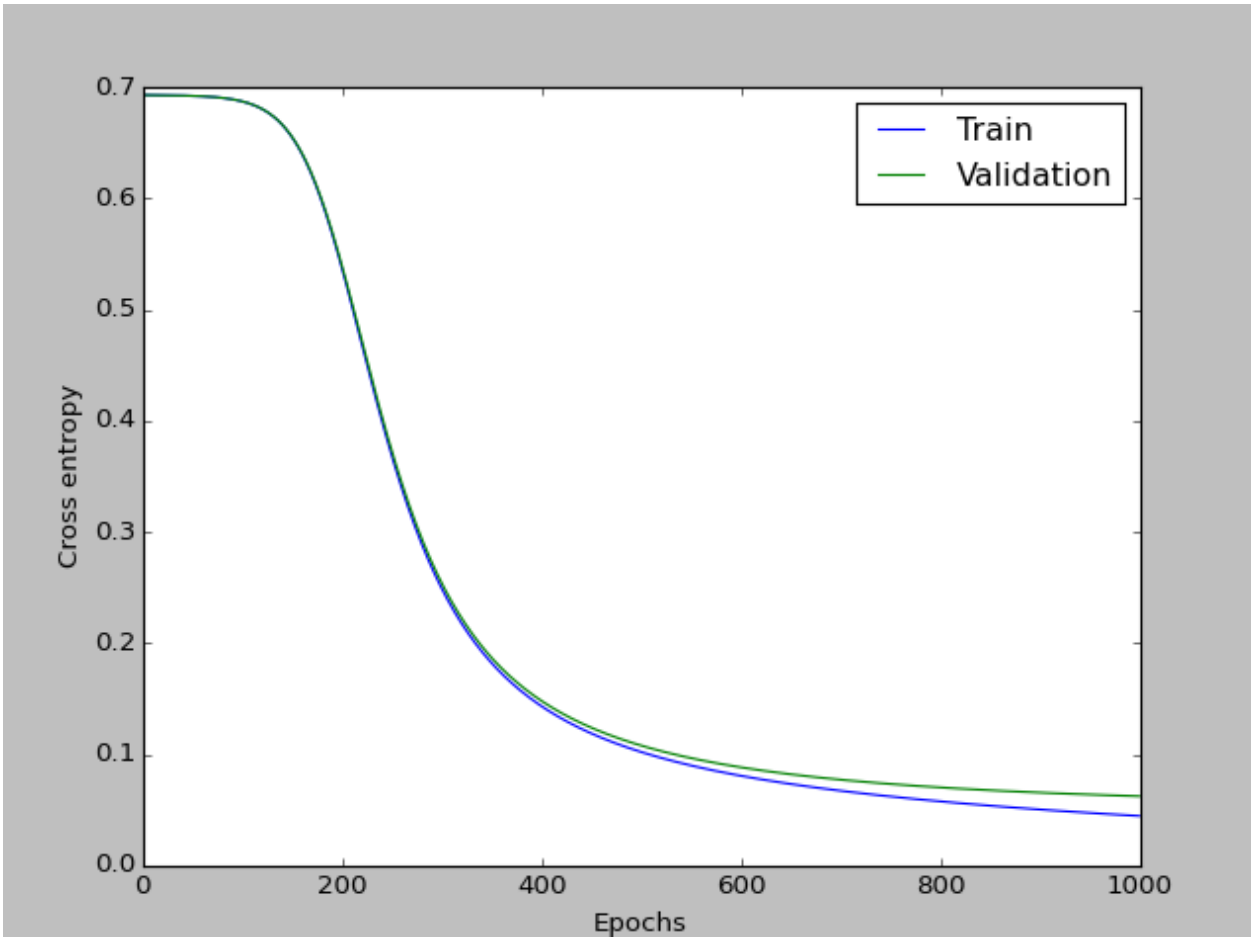


$\epsilon = 0.5, momentum = 0$

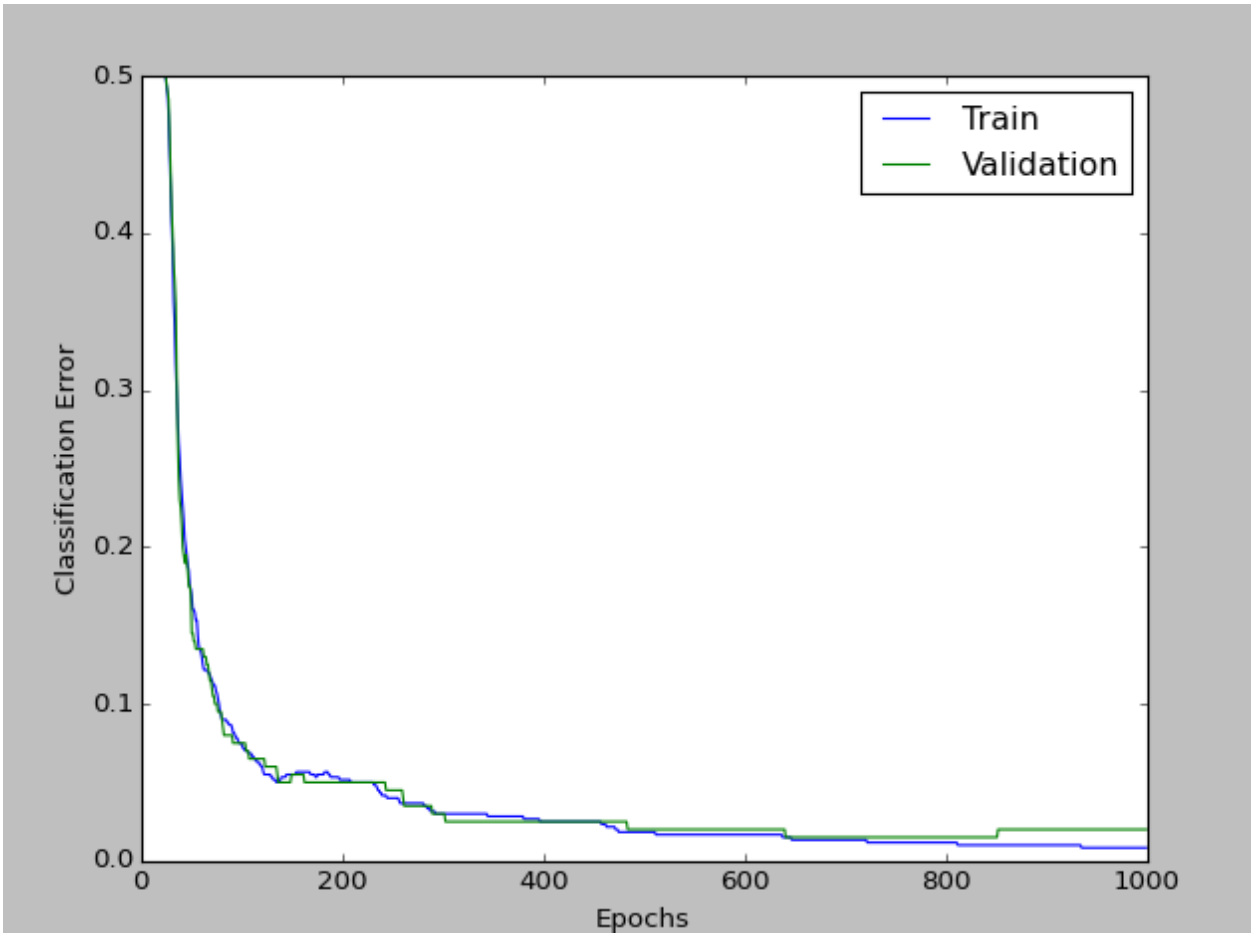


Lower learning rates slow down the process of convergence, whereas higher learning rate speed it up. However, it looks like a risk associated with that is less generalizability.

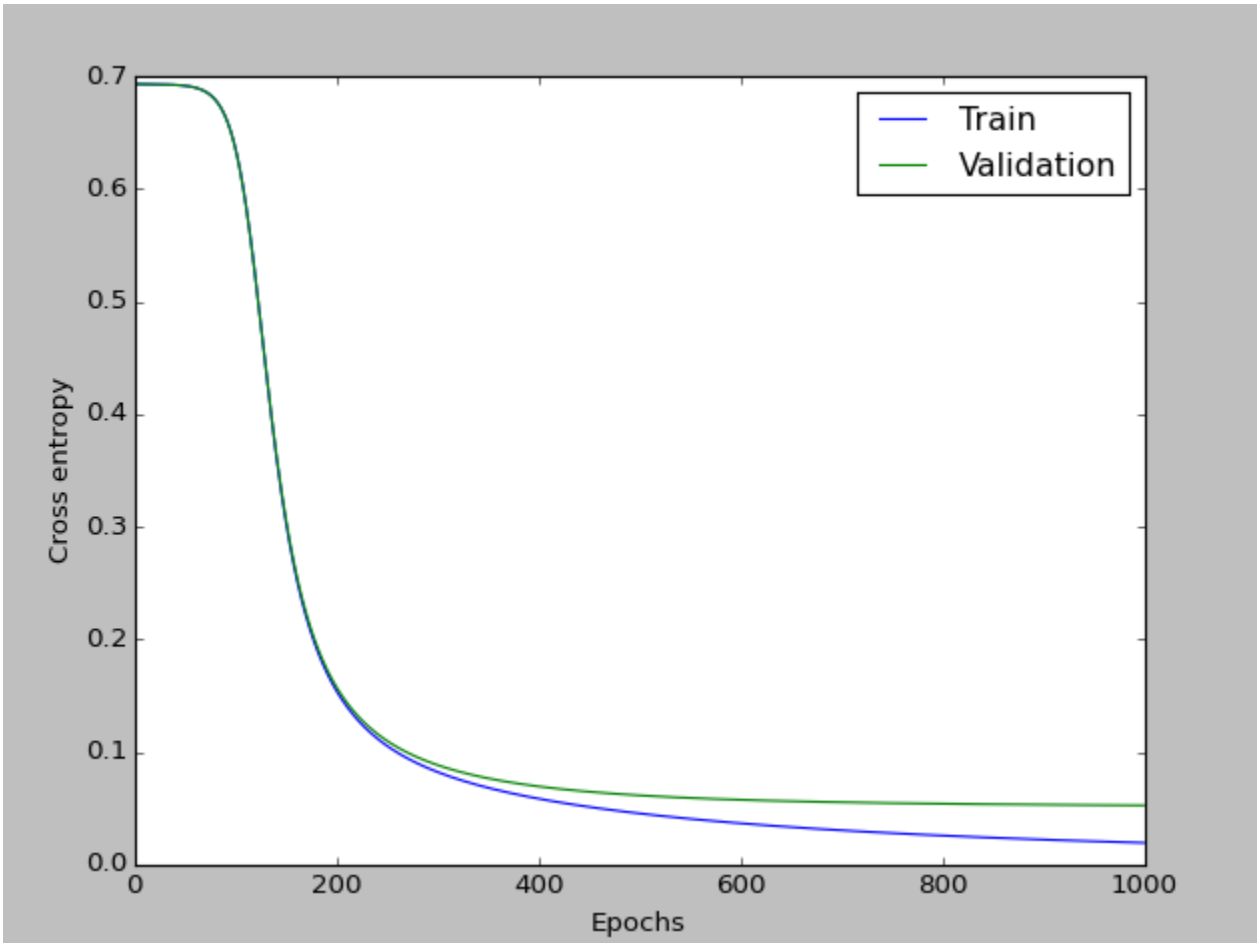
$\epsilon = 0.1, momentum = 0$



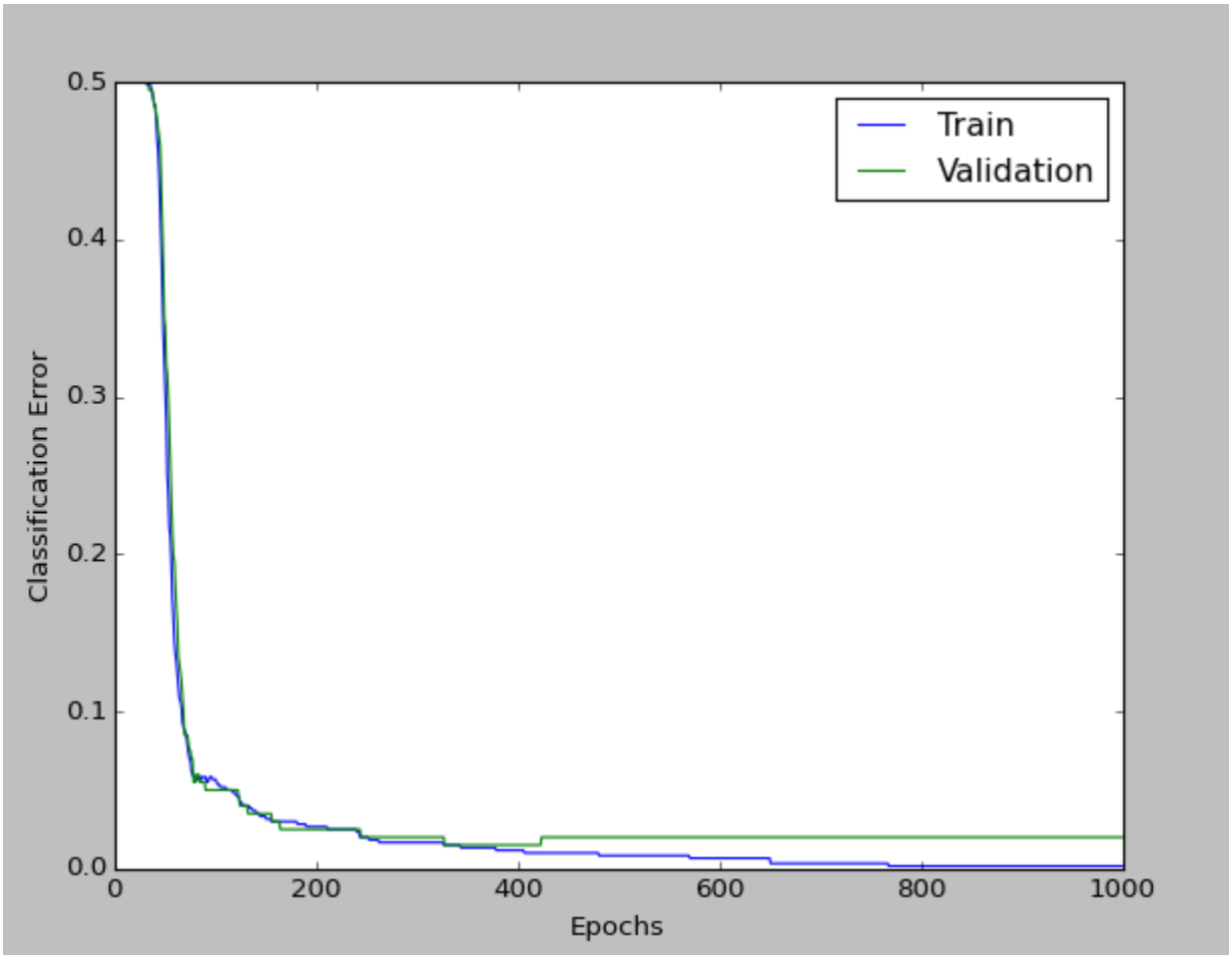
$\epsilon = 0.1, momentum = 0$



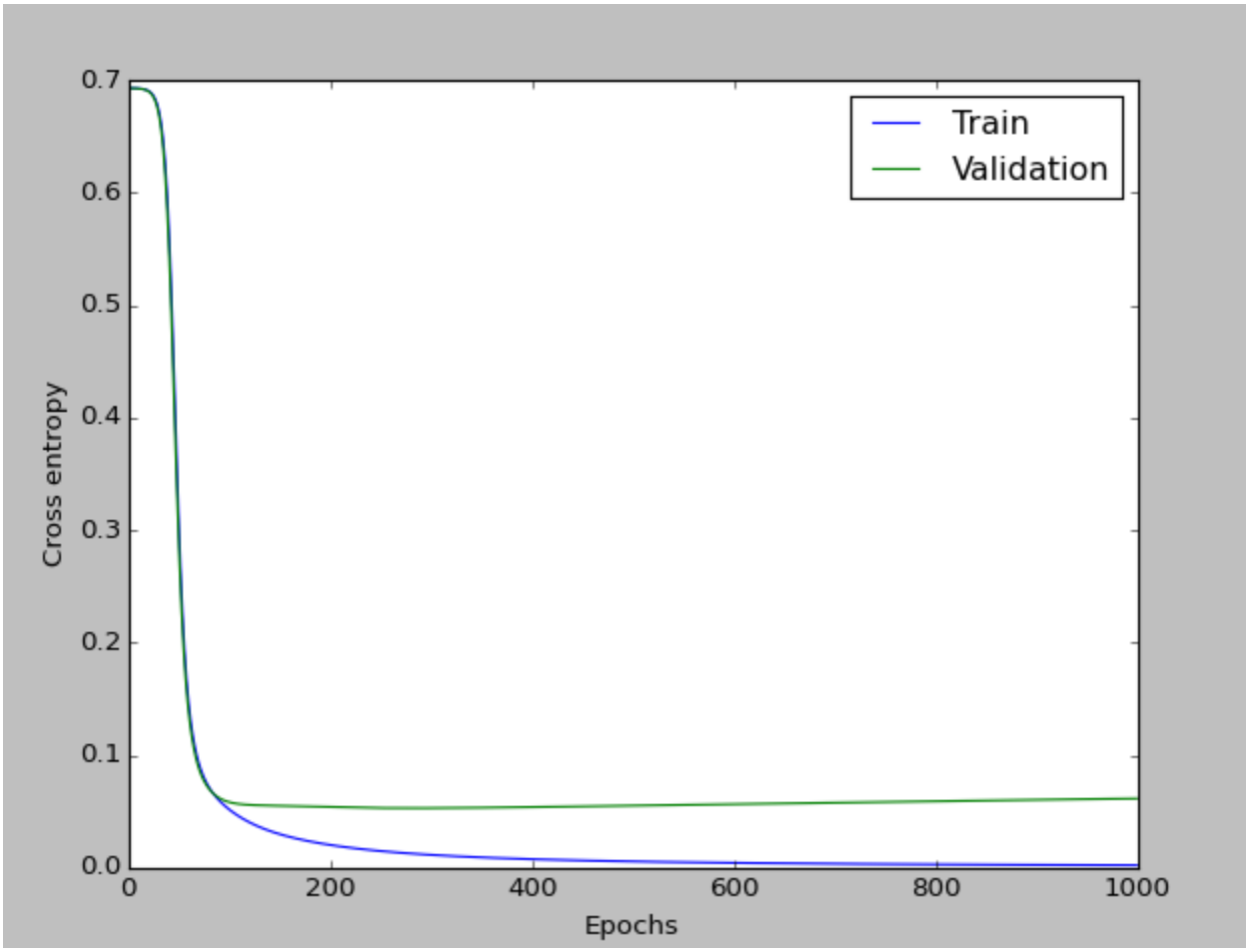
$\epsilon = 0.1, momentum = .5$



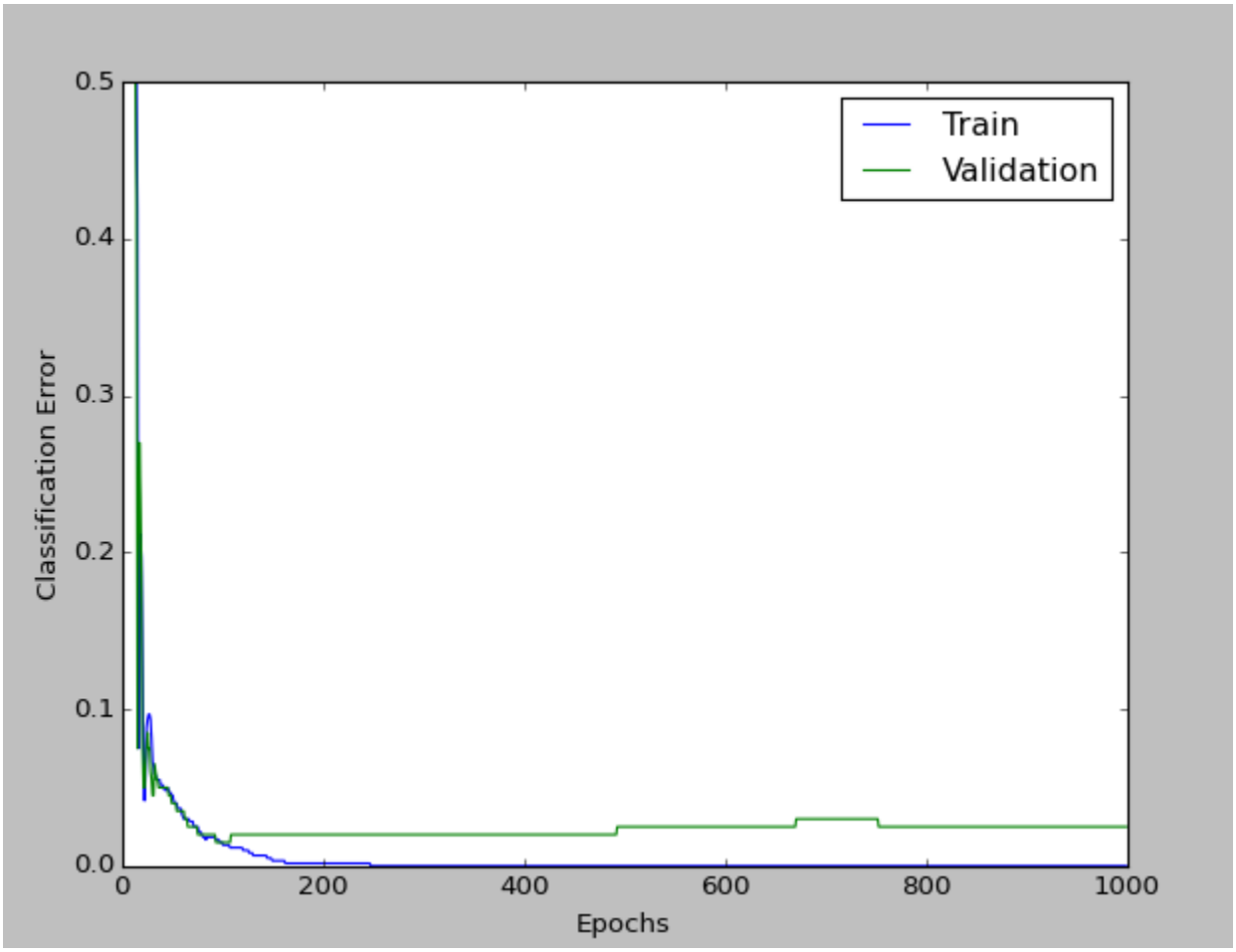
$\epsilon = 0.1, momentum = 0.5$



$\epsilon = 0.1, momentum = 0.9$



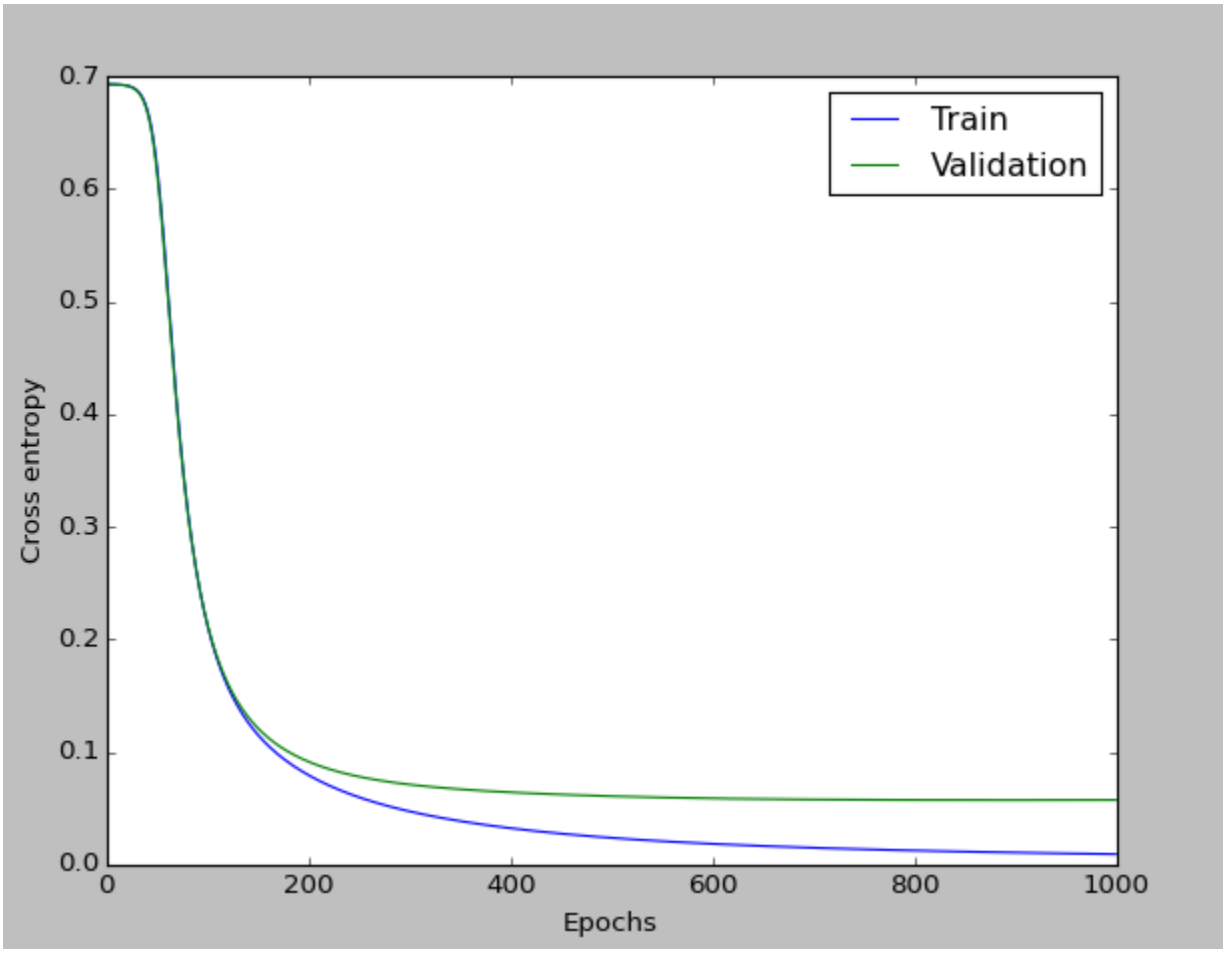
$\epsilon = 0.1, momentum = 0.9$



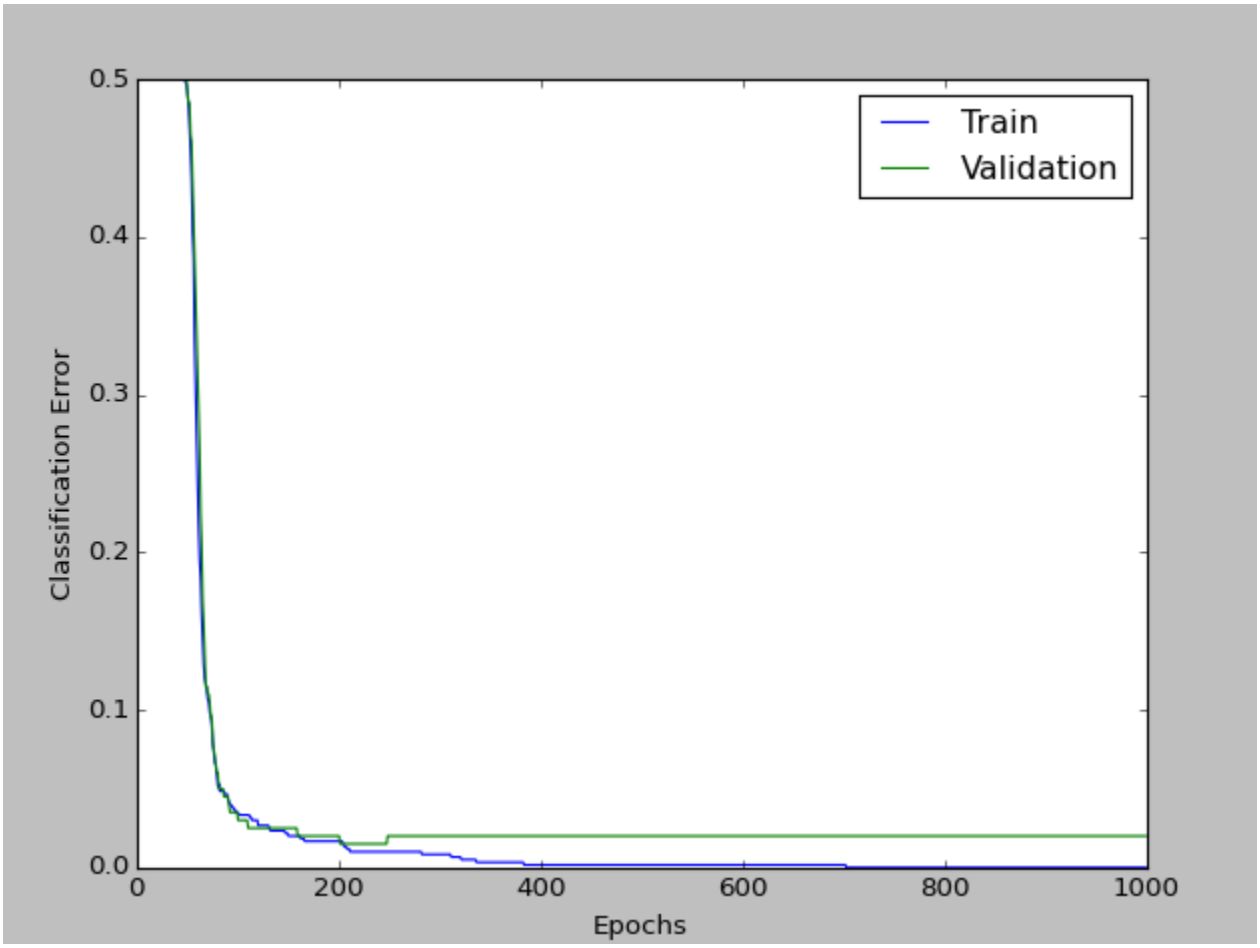
Greater momentum has a similar effect as a higher learning rate (faster convergence), but the decrease in generalizability seem seven greater. Therefore, the best parameters would be low, such as 0.1 for both.

2.4

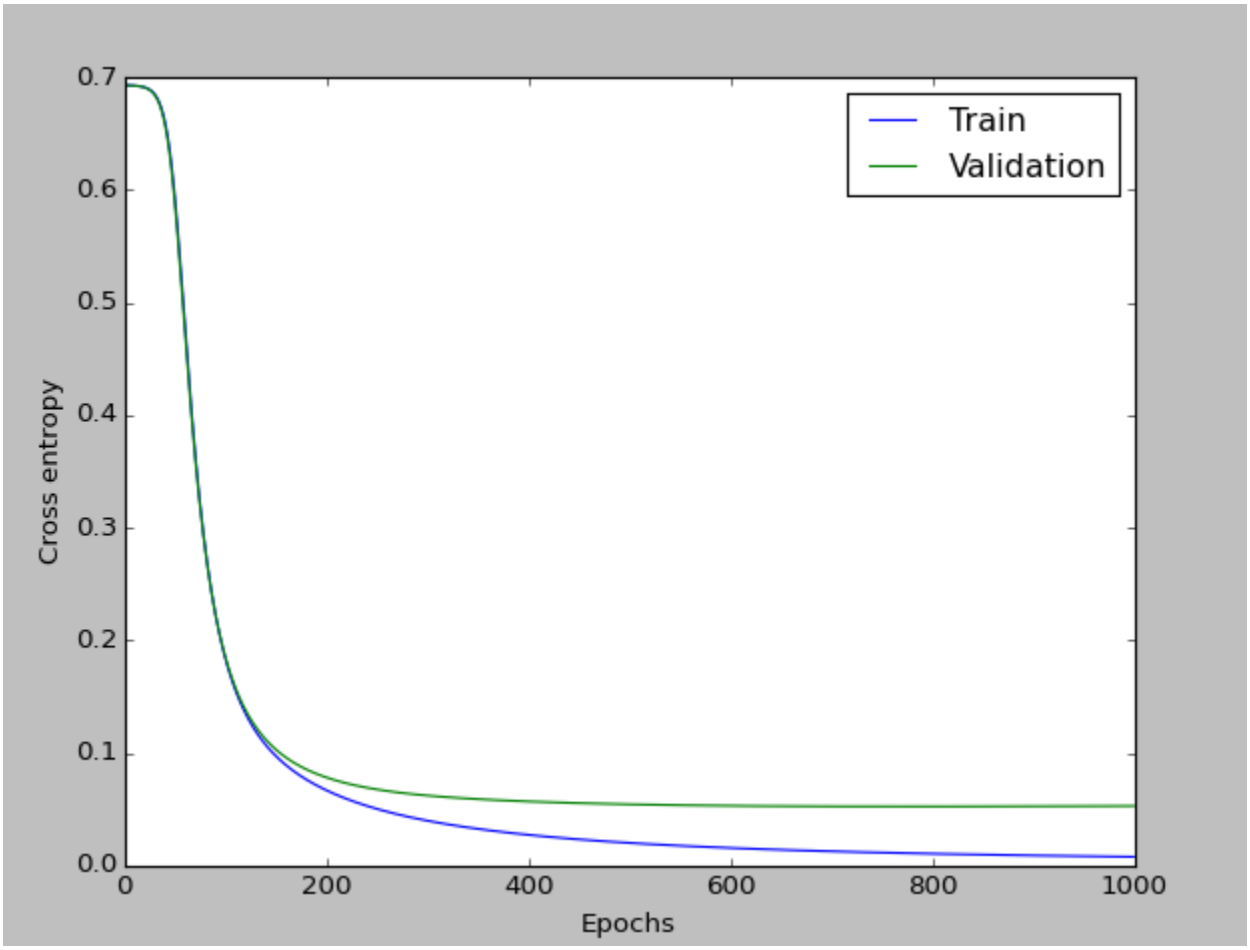
2 hidden units



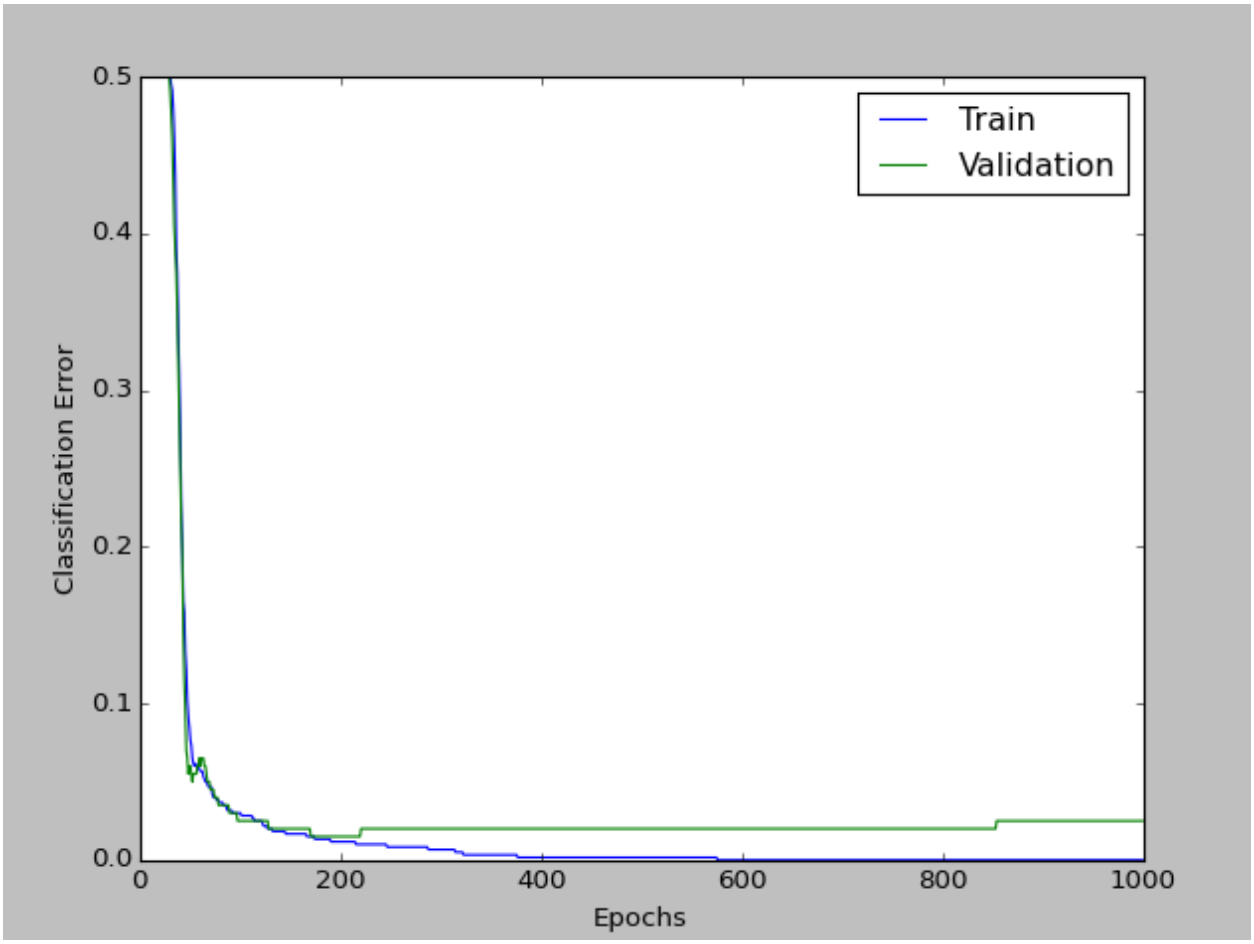
2 hidden units



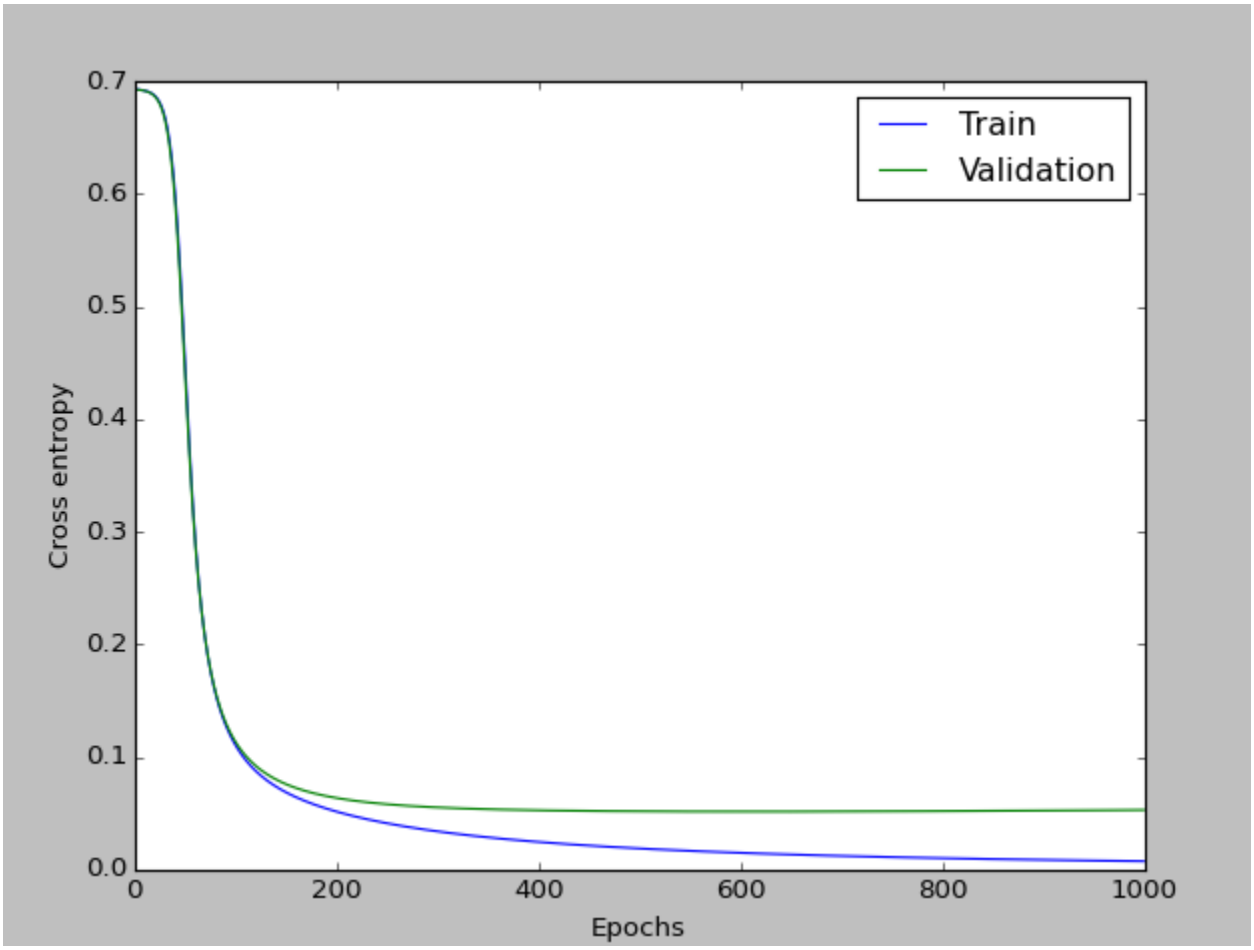
5 hidden units



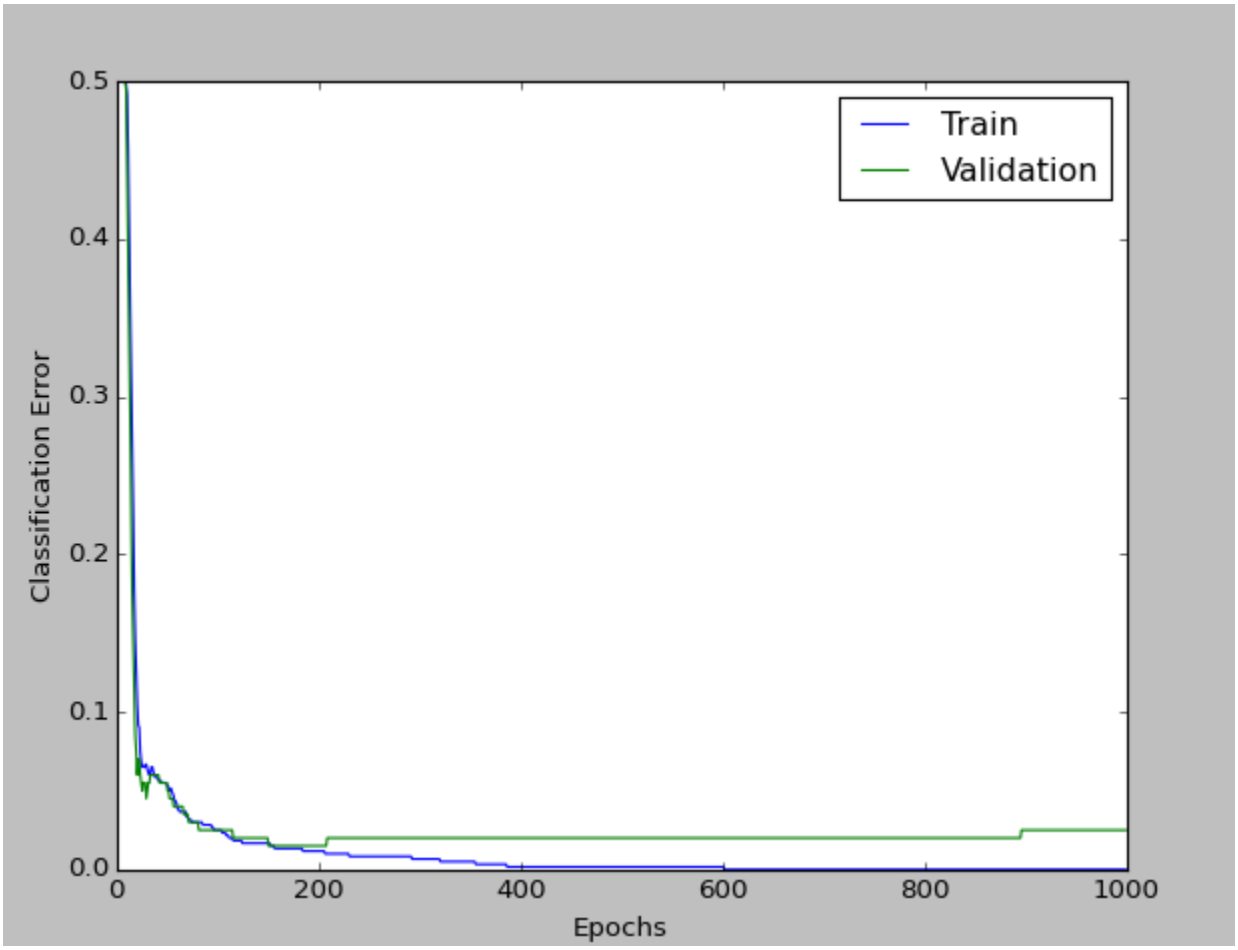
5 hidden units



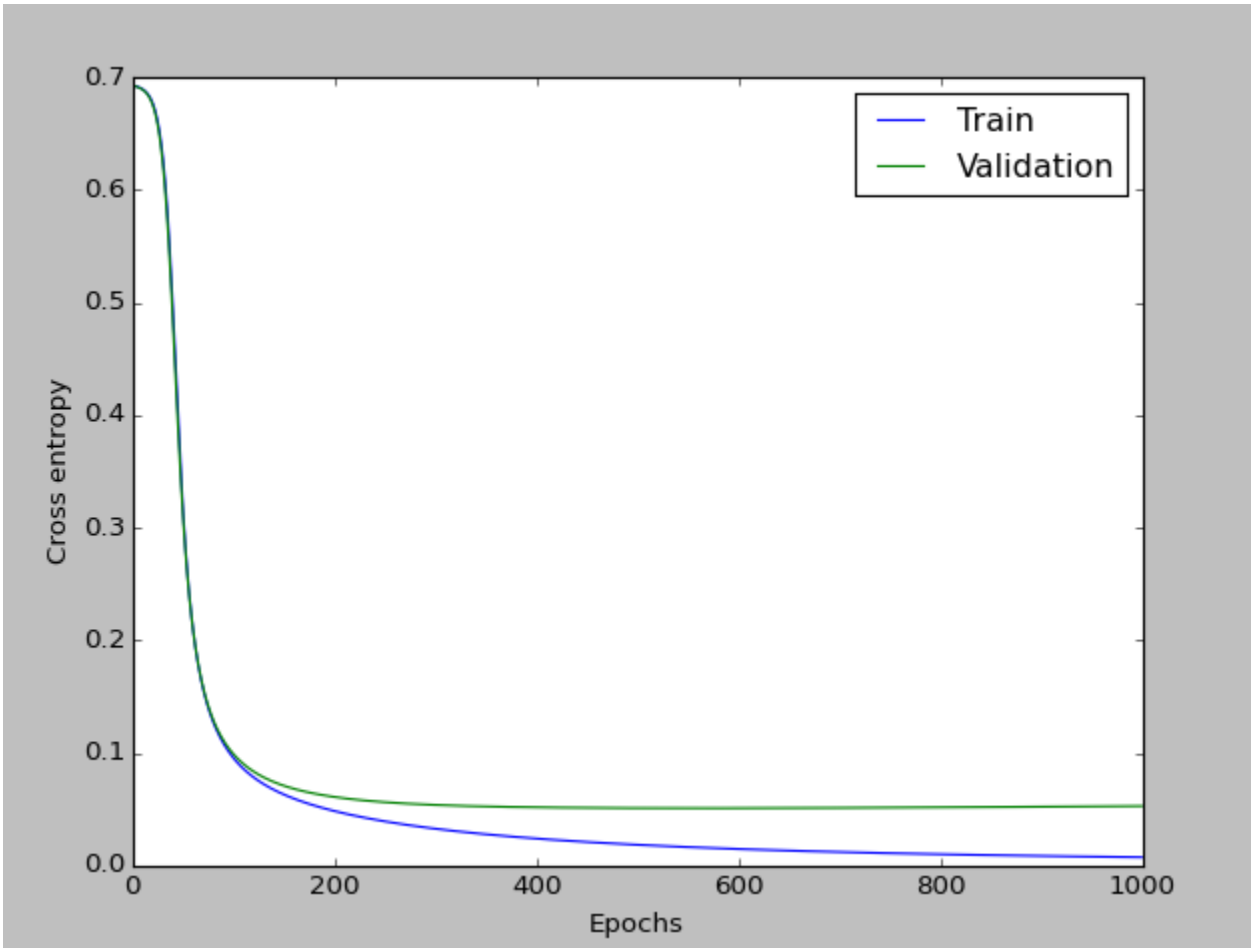
30 hidden units



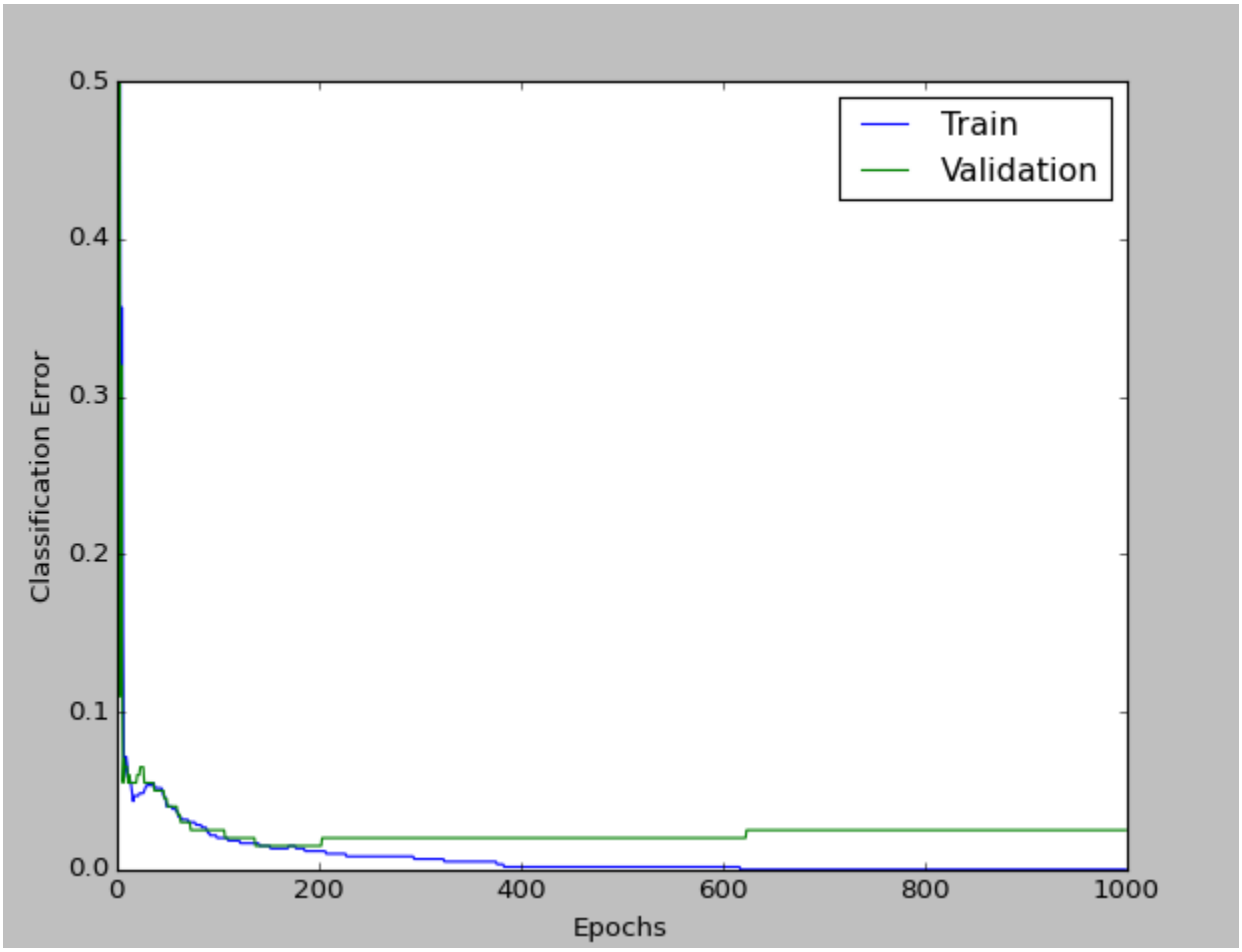
30 hidden units



100 hidden units

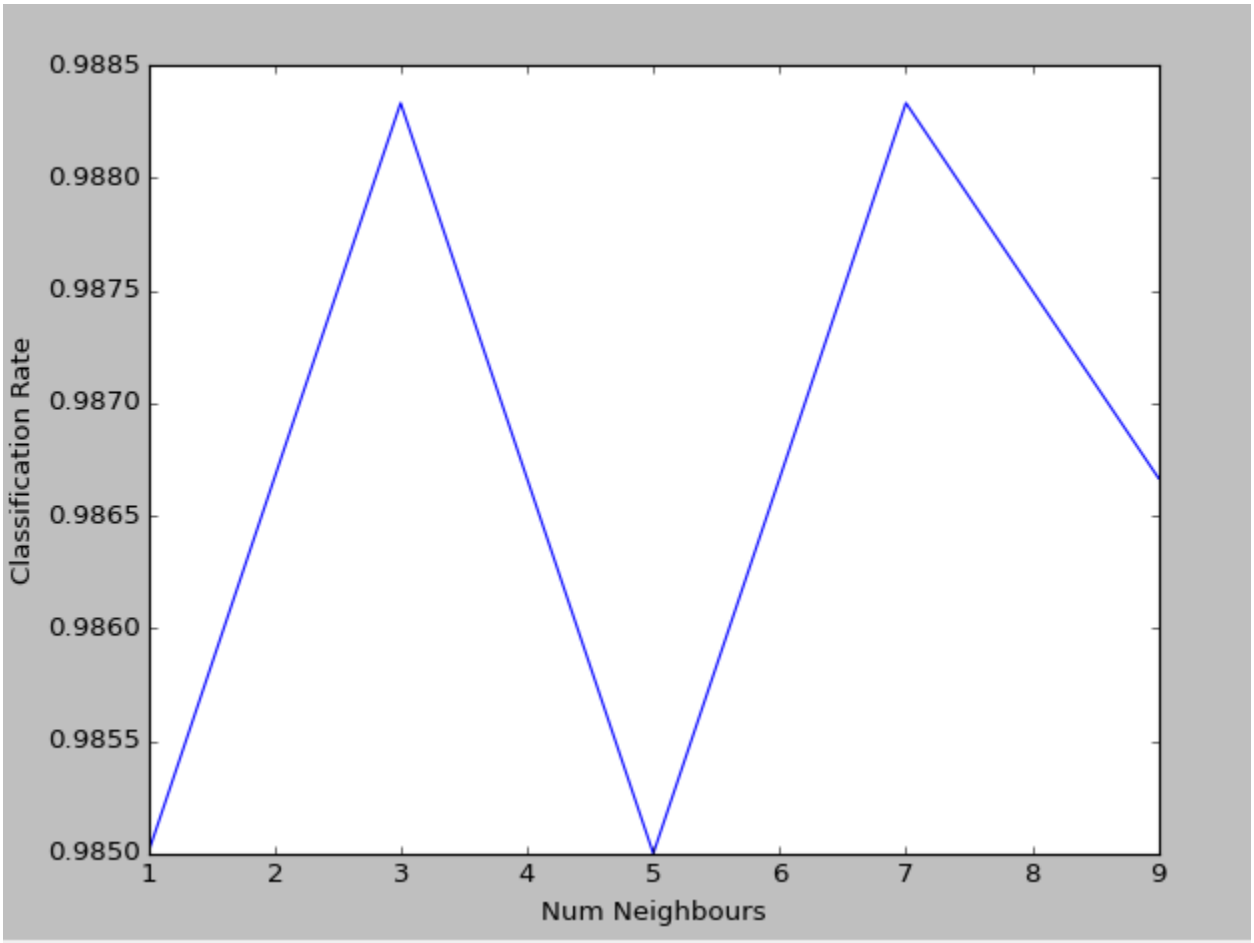


100 hidden units



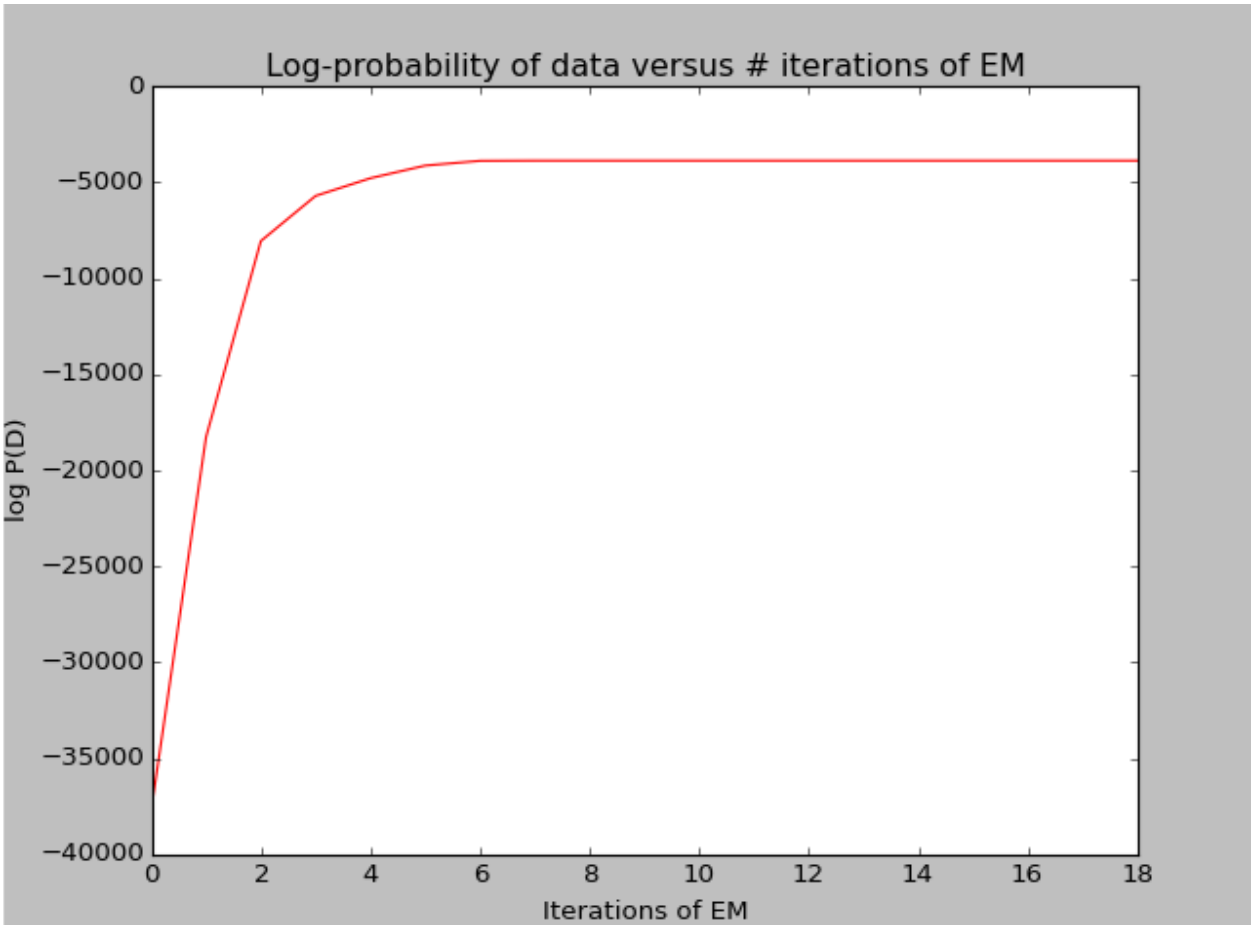
Surprisingly, there doesn't seem to be a large effect due to more hidden units, other than a slightly faster start. If enough hidden units were added, however, overfitting would start to be a problem.

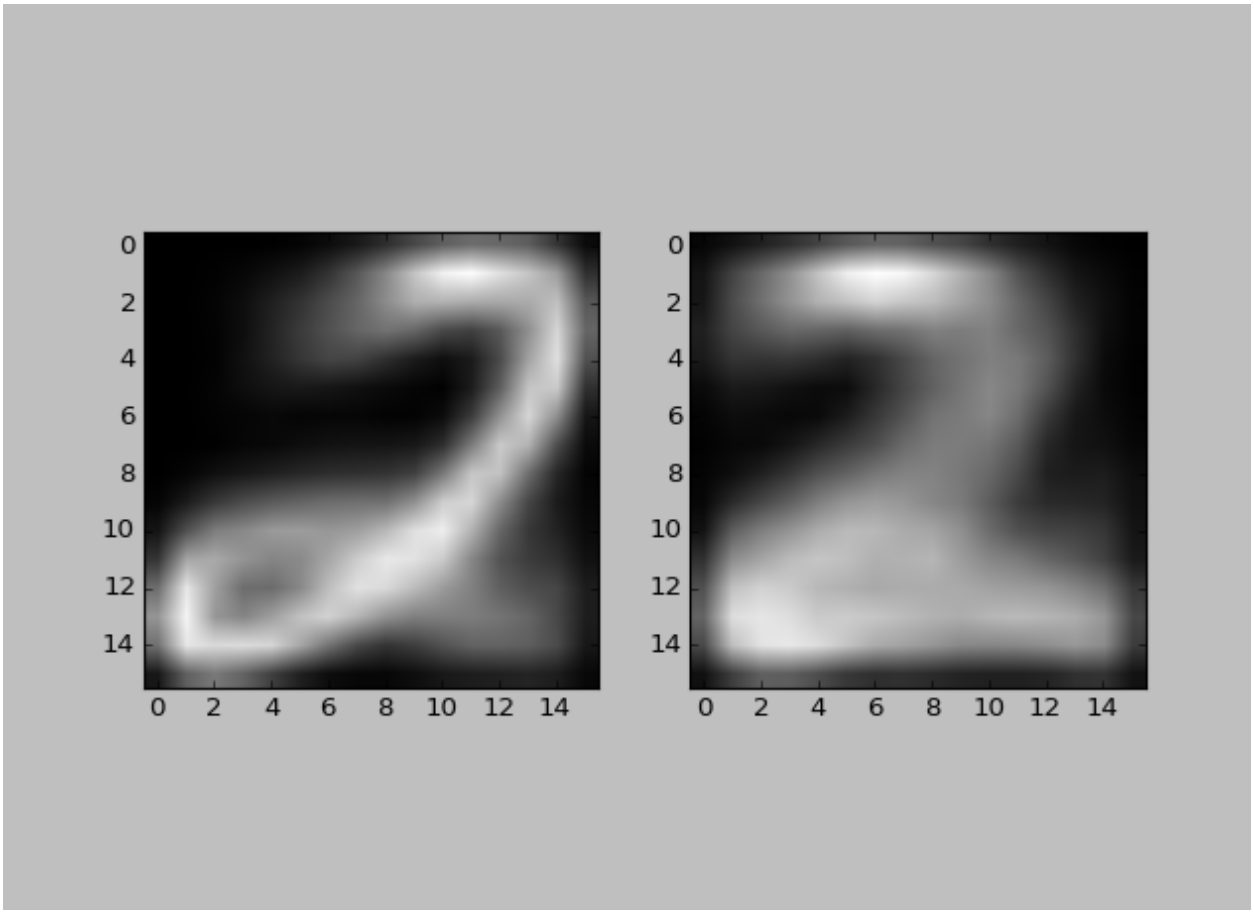
2.5

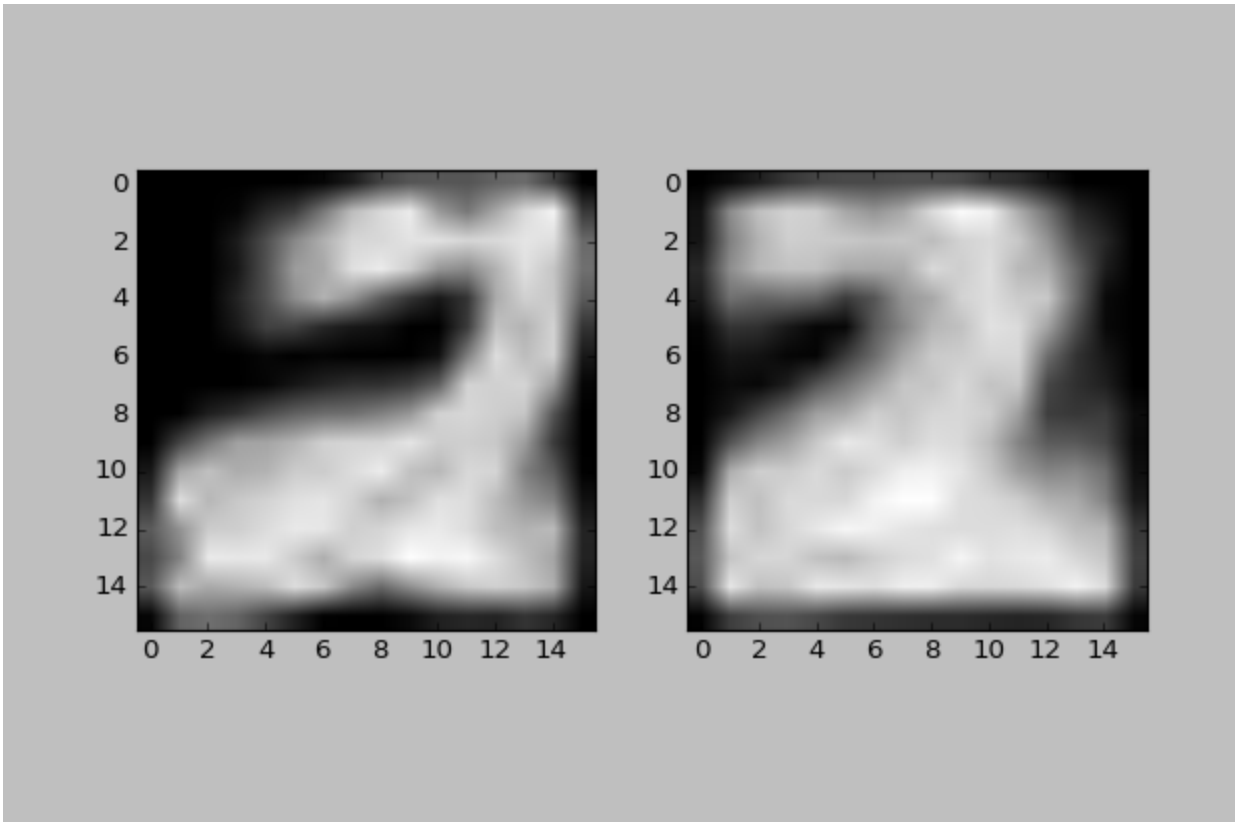


Surprisingly, k nearest neighbours actually seems to do better than this implementation of the neural network.

3.2

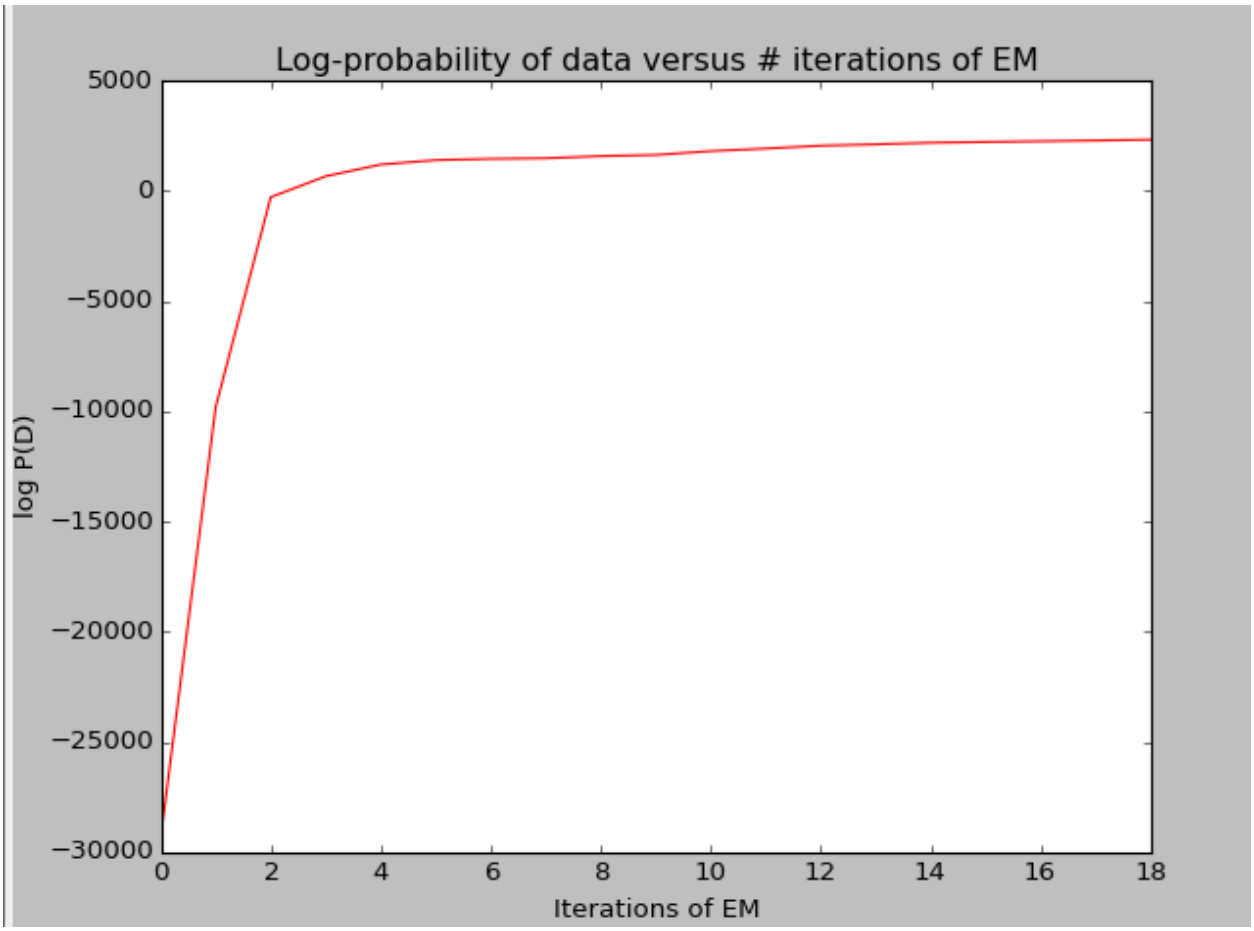

Log-probability of data versus # iterations of EM
log P(D)
0
−5000
−10000
−15000
−20000
−25000
−30000
−35000
−40000
0
2
4
6
8
10
12
14
16
18
Iterations of EM

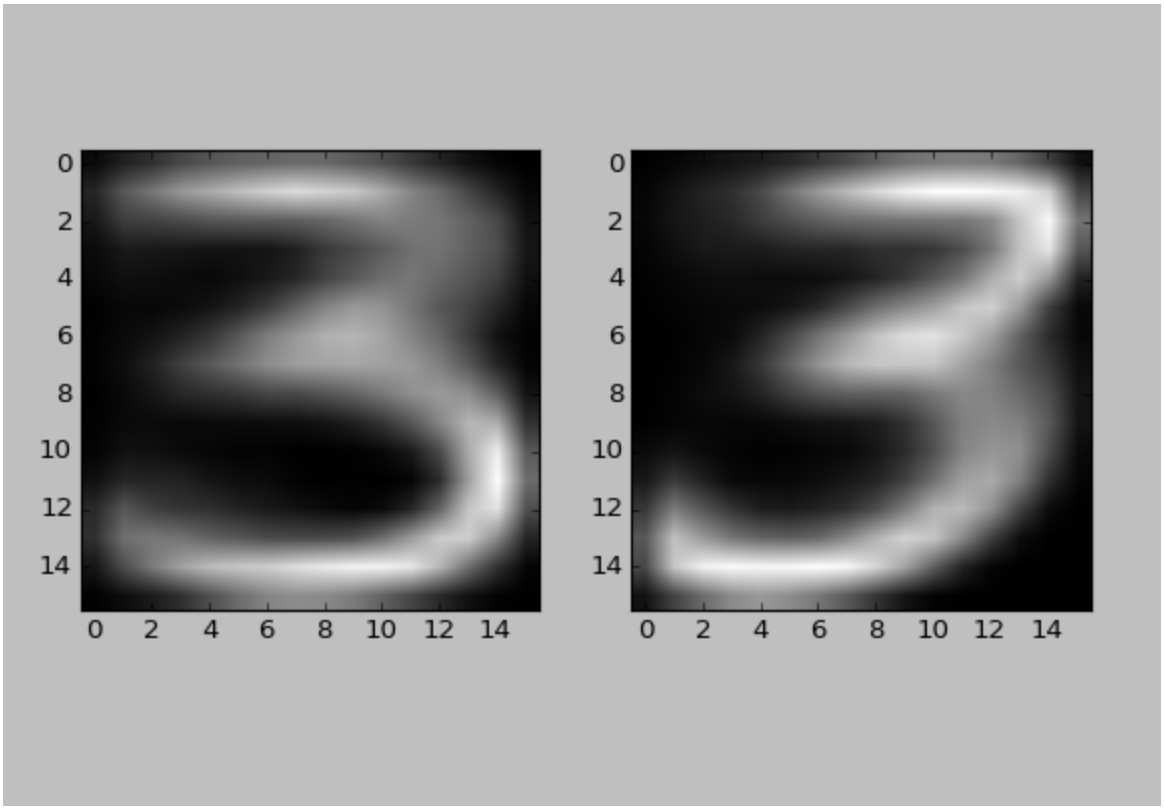

0
2
4
6
8
10
12
14
0 2 4 6 8 10 12 14
0
2
4
6
8
10
12
14
0 2 4 6 8 10 12 14

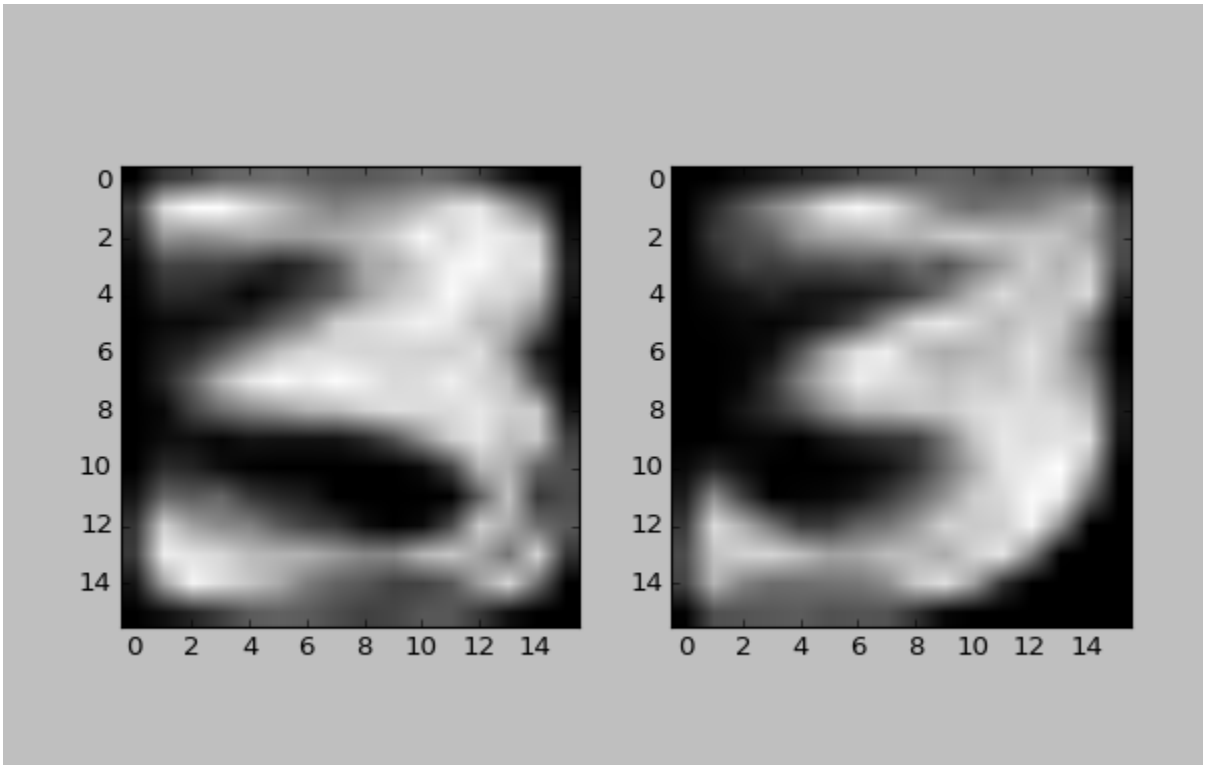


Mixing: array([[0.50664608], [0.49335392]])

logProb -3868.40634


Log-probability of data versus # iterations of EM
5000
0
−5000
−10000
−15000
−20000
−25000
−30000
log P(D)
0
2
4
6
8
10
12
14
16
18
Iterations of EM

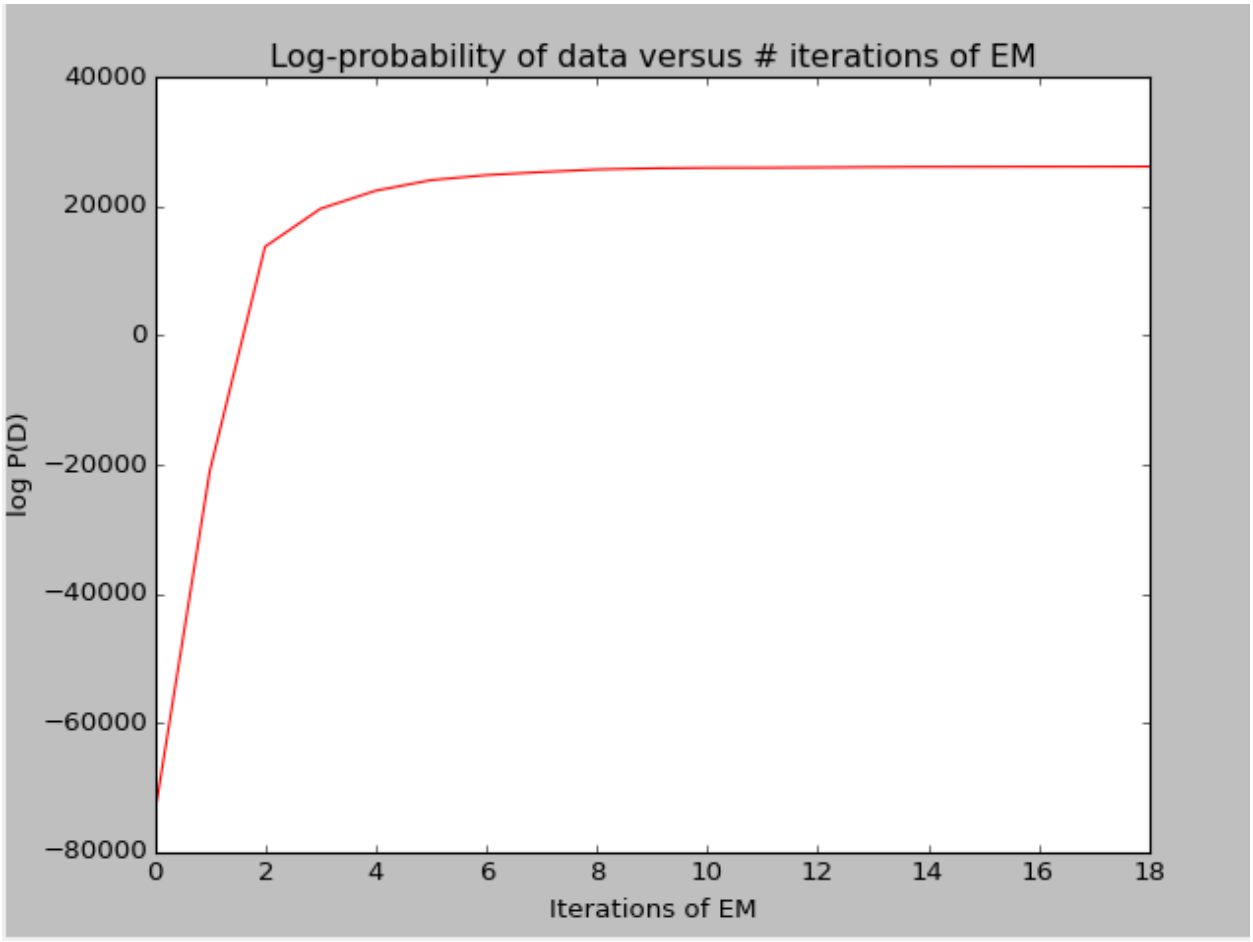


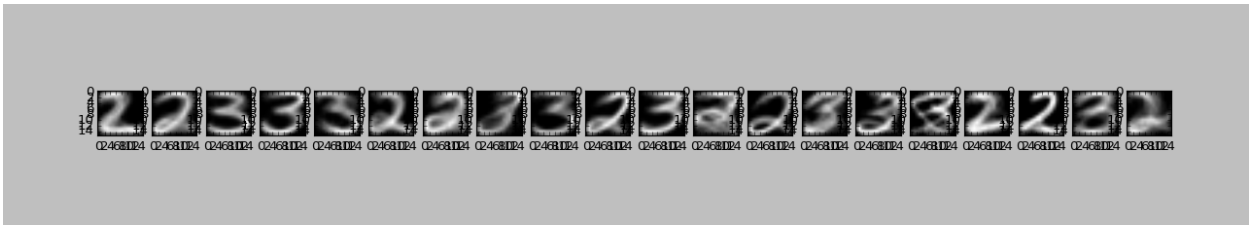


Mixing: array([[0.46351236],[0.53648764]])

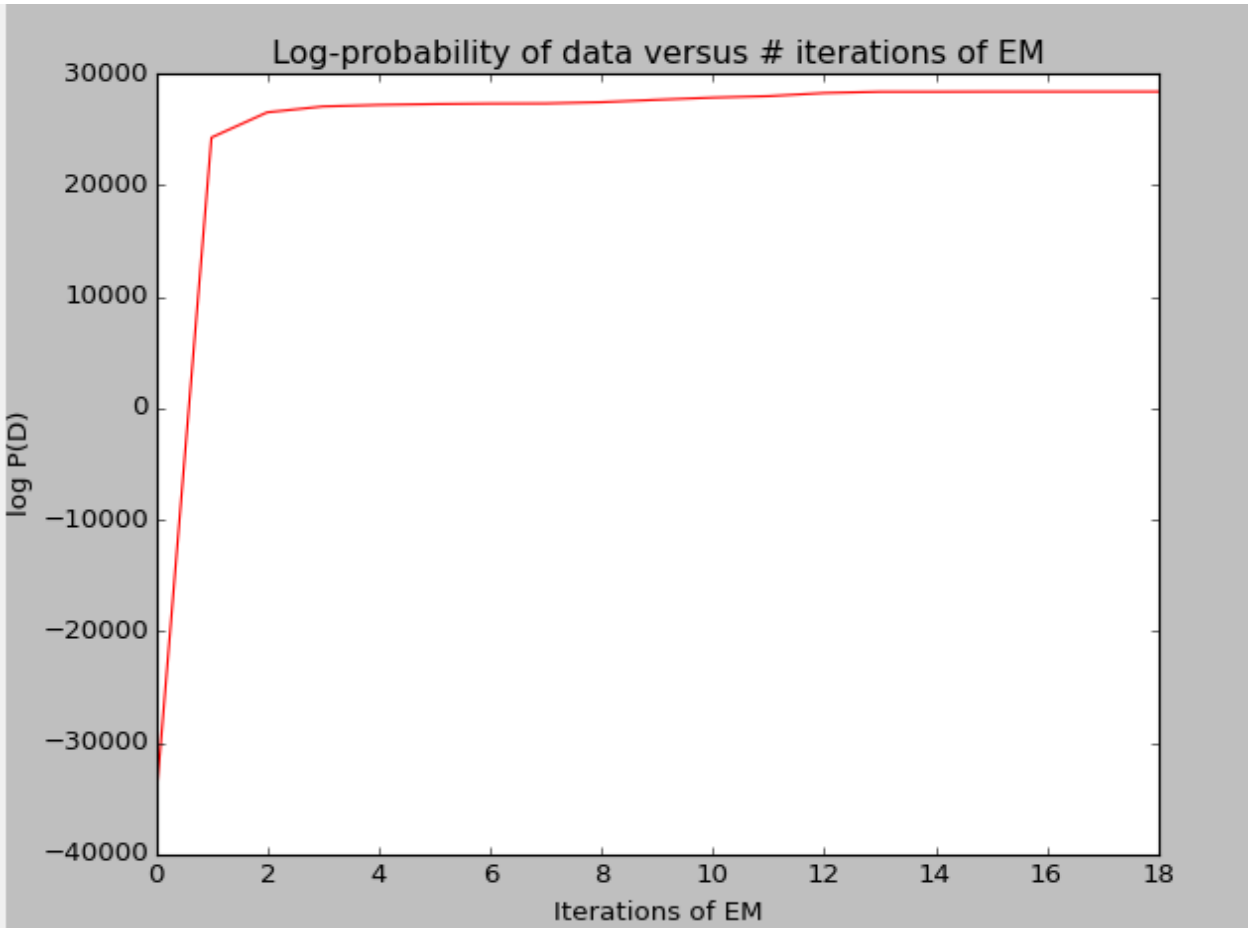
logProb 2350.27751

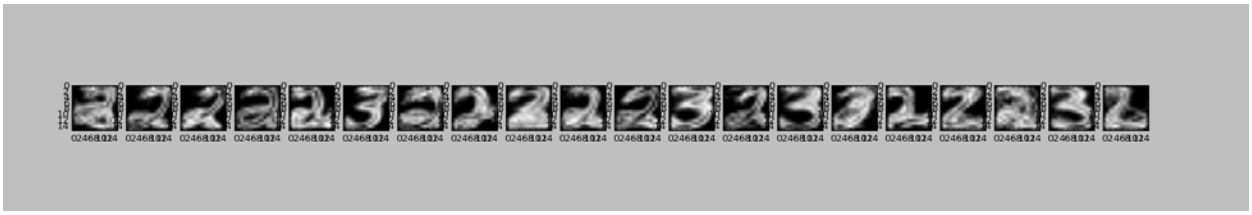
3.3





logProb 26203.57393

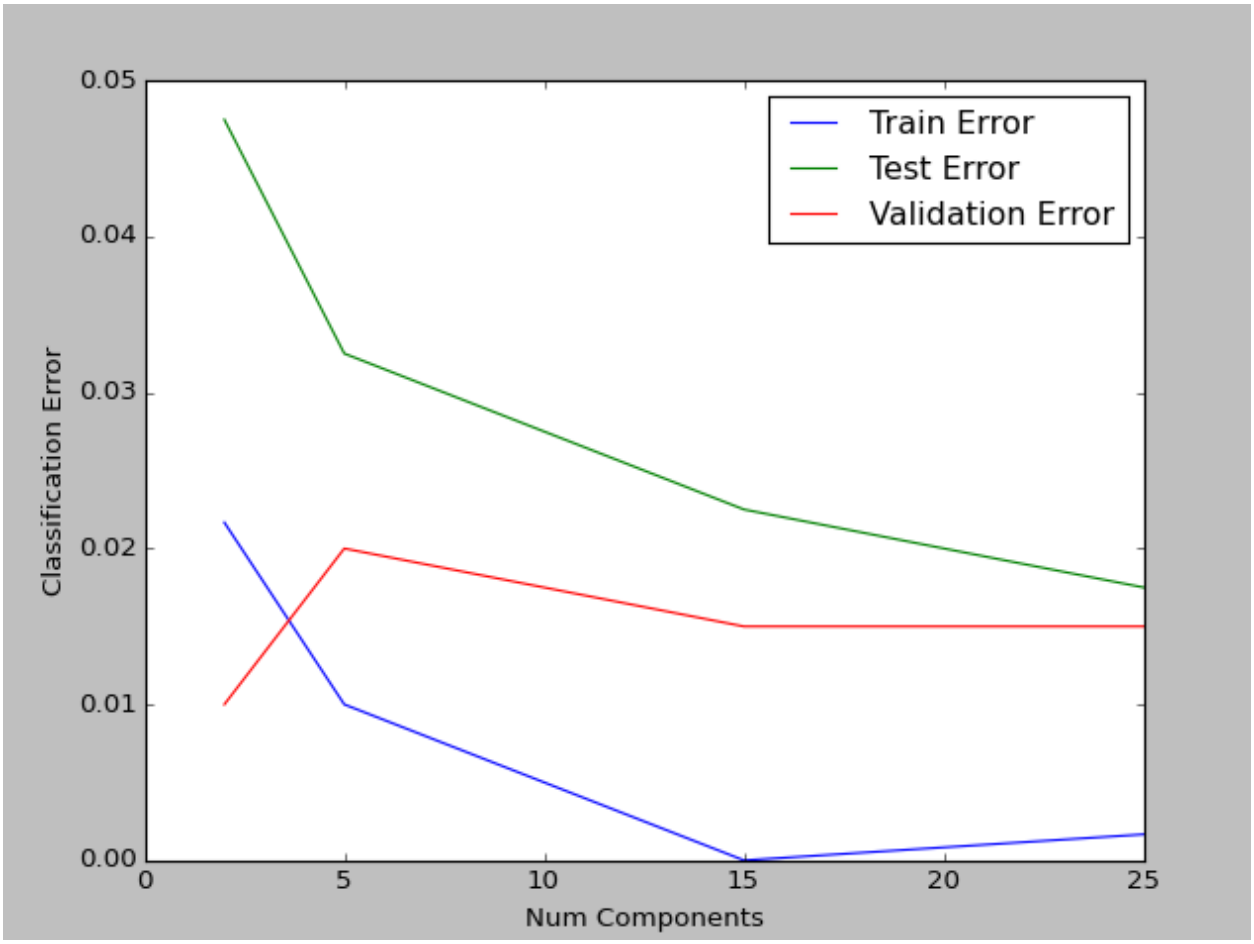




logProb 28405.81939

Compared to non-k-means initialization, convergence is achieved much faster, and the log likelihood is much higher.

3.4



1. As the number of clusters increases, the fewer points each cluster has to account for, meaning each one can be more accurate. If enough clusters are used there could be one cluster per point, leading to 100% classification rate, but overfitting and a lack of generalizability.
2. Despite starting with a very high error rate when 2 components are used (compared to the training and validation set; absolutely it is not that high), as more components are added the error rate decreases, as would be expected.
3. I would use 15 components, because although in this run the test error rate decreased as number of components increased, it is within the realm of possibility that that 25 is too many components, and lack of generalizability could set in.