I checked the soln's.

1)

a) $p(x_1) = N(\mu_1, \Sigma_{11}) = N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}6 & 8\\8 & 13\end{pmatrix}\right)$

b) $p(x_2) = N(\mu_2, \Sigma_{22}) = N(5, 14)$

c) $p(x_1|x_2) = N(\mu_{1|2}, \Sigma_{1|2})$

$$= N\left(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right)$$

$$= N\left(\begin{pmatrix}0\\0\end{pmatrix} + \frac{1}{14}\begin{pmatrix}5\\11\end{pmatrix}(x_2 - 5), \begin{pmatrix}6 & 8\\8 & 13\end{pmatrix} - \frac{1}{14}\begin{pmatrix}5\\11\end{pmatrix}(5 \quad 11)\right)$$

$$= N\left(\frac{x_2 - 5}{14}\begin{pmatrix}5\\11\end{pmatrix}, \frac{1}{14}\begin{pmatrix}59 & 57\\57 & 61\end{pmatrix}\right)$$

d) $p(x_2|x_1) = N(\mu_{2|1}, \Sigma_{2|1})$

$$= N\left(\mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(x_1 - \mu_1), \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right)$$

$$= N\left(5 + (5 \quad 11)\begin{pmatrix}6 & 8\\8 & 13\end{pmatrix}^{-1}\left(x_1 - \begin{pmatrix}0\\0\end{pmatrix}\right), 14 - (5 \quad 11)\begin{pmatrix}6 & 8\\8 & 13\end{pmatrix}^{-1}\begin{pmatrix}5\\11\end{pmatrix}\right)$$

$$= N\left(5 + (5 \quad 11)\frac{1}{14}\begin{pmatrix}13 & -6\\-8 & 6\end{pmatrix}x_1, 14 - (5 \quad 11)\begin{pmatrix}13 & -6\\-8 & 6\end{pmatrix}\begin{pmatrix}5\\11\end{pmatrix}\frac{1}{14}\right)$$

$$= N\left(5 + \frac{1}{14}(-23 \quad 26)x_1, \frac{25}{14}\right)$$

2)

a) $p(y=1|x;\theta) = \sigma\theta^T x$

$$nll(\theta) = -\sum_i y_i \log\sigma(\theta^T x) + (1-y_i)\log(1-\sigma(\theta^T x)) + \lambda/2 \|\theta\|_2^2$$

then we have

$$\nabla_\theta l = \sum_i y_i(1-\sigma(\theta^T x))x - (1-y_i)\,\sigma(\theta^T x)\,x + \lambda\theta$$

$$= \sum_i [y_i - \sigma(\theta^T x_i)]^T x_i + \lambda\theta = x^T(\sigma(X\theta)-y) + \lambda\theta$$

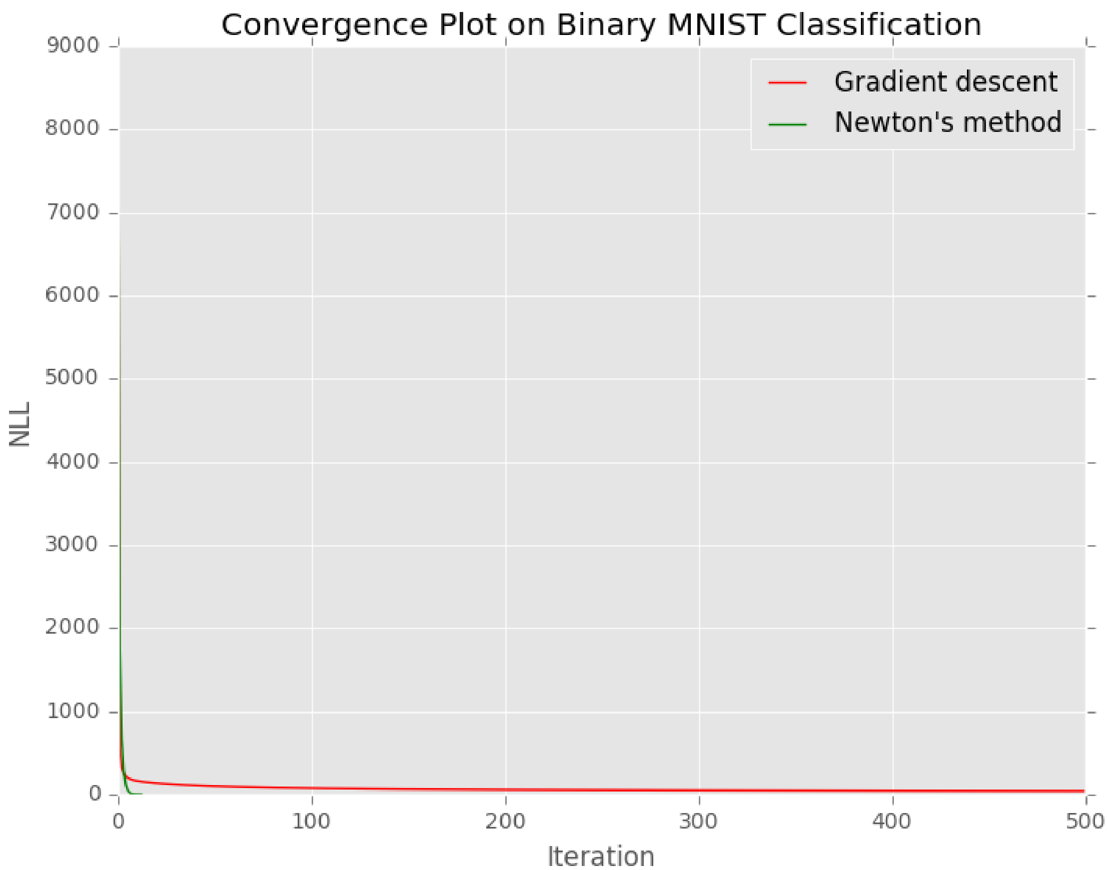
$$\nabla^2 l = \sum_i \nabla_\theta \sigma(\theta^T x) x_i^T + \lambda I = x^T diag[\sigma(X\theta)(1-\sigma(X\theta))]X + \lambda I$$

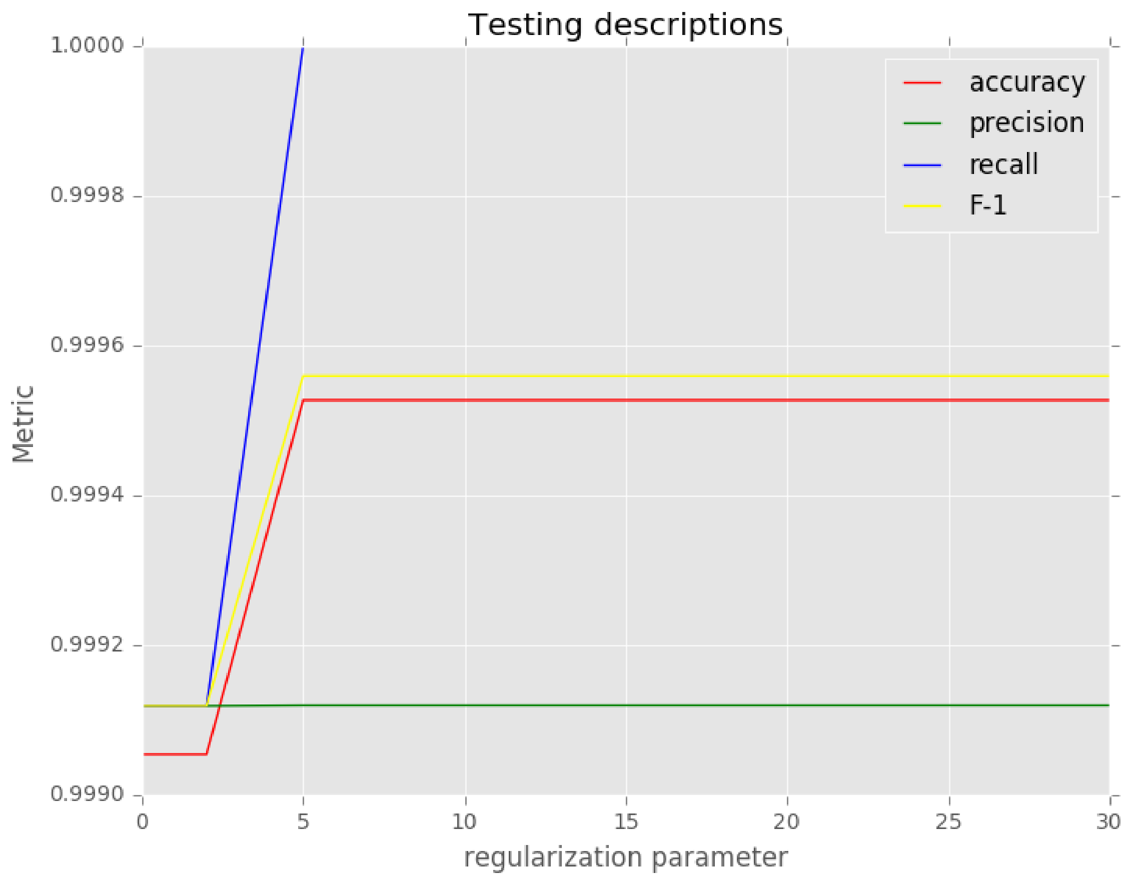
b) $\mathbb{P}(y=c|x,w) = \dfrac{exp(w_c^T x)}{\sum exp(w_i^T x)}$

$$nll(w) = -\log \prod_i \prod_c \mu_{ic}^{y_{ic}} - \lambda tr(w^T w)$$

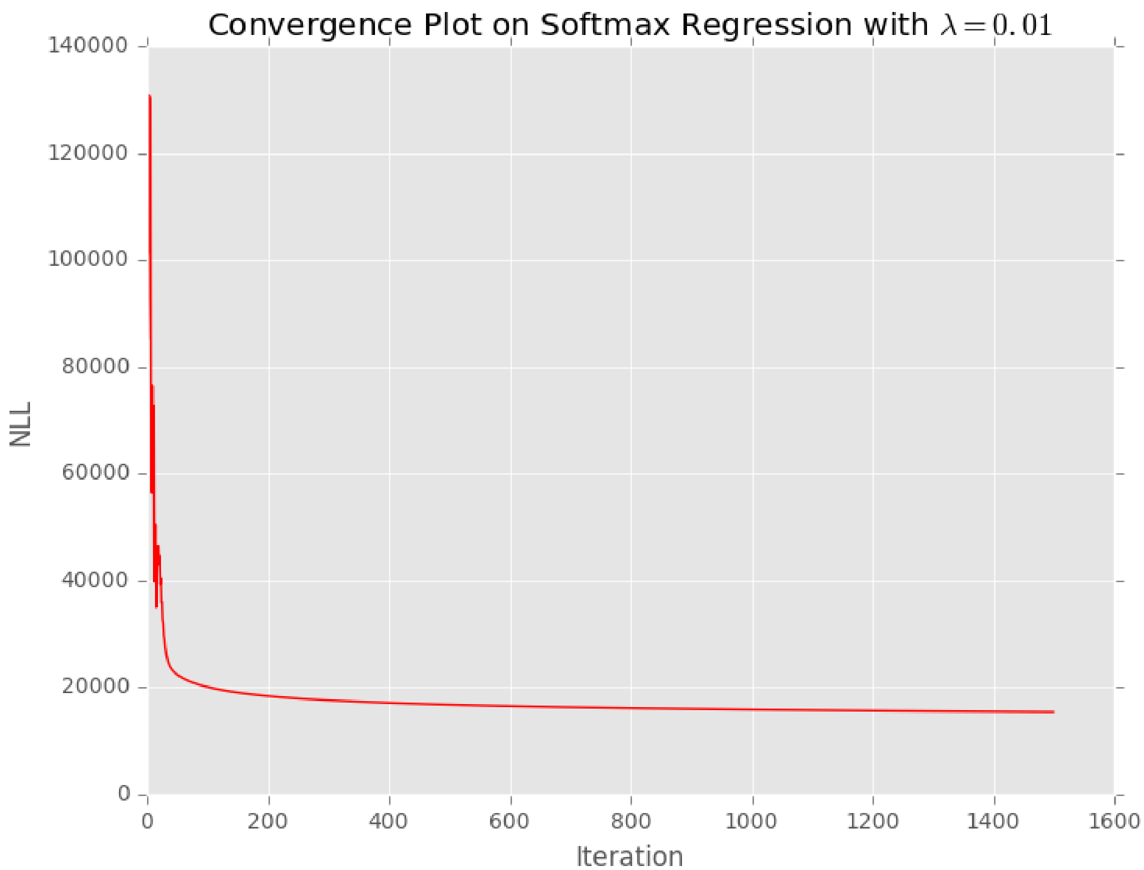
$$= \sum_i \sum_c y_{ic} \log \mu_{ic} + \lambda tr(w^T w)$$

then $\nabla_w nll = x^T(\mu - y) + \lambda w$

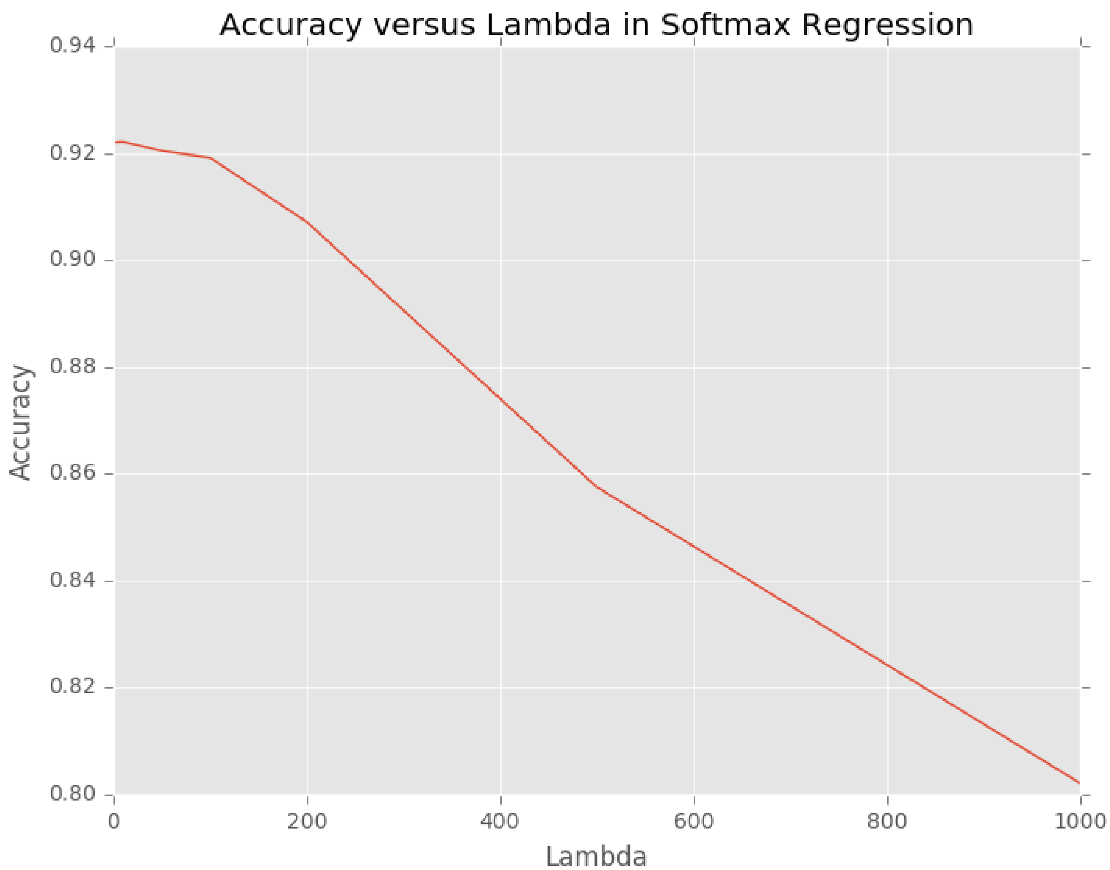

Convergence Plot on Binary MNIST Classification
Gradient descent
Newton's method
NLL
Iteration
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
0
100
200
300
400
500



Testing descriptions
accuracy
precision
recall
F-1
Metric
regularization parameter
1.0000
0.9998
0.9996
0.9994
0.9992
0.9990
0
5
10
15
20
25
30



Convergence Plot on Softmax Regression with $\lambda = 0.01$
NLL
Iteration
140000
120000
100000
80000
60000
40000
20000
0
0
200
400
600
800
1000
1200
1400
1600


Accuracy versus Lambda in Softmax Regression
Accuracy
Lambda
0.94
0.92
0.90
0.88
0.86
0.84
0.82
0.80
0
200
400
600
800
1000