Oscar Scholin
189 HW 4

1. a) By definition the marginal probability for $\vec{x}_1$ is $N(\vec{\mu}_1, \Sigma_{11})$, where $\vec{\mu}_1$ is the vector of means for each dimension of $\vec{x}_1$, and Σ_{11} represents the covariance matrix for $\vec{x}_1$, and N is the normal distribution.

$$\therefore\ p(\vec{x}_1) = N(\mu_1, \Sigma_{11}) = N\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}6 & 8\\3 & 13\end{bmatrix}\right)$$

* Assuming $\vec{x} = (\vec{x}_1, \vec{x}_2)$ is jointly Gaussian.

b) Similarily for $\vec{x}_2 = x_2$, which is 1-dimensional:

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

c) The conditional distribution $p(\vec{x}_1 | \vec{x}_2) = N(\mu_{1|2}, \Sigma_{1|2})$.

We know
$$\mu_{1|2} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(\vec{x}_2 - \mu_2)$$
$$= \begin{bmatrix}0\\0\end{bmatrix} + \begin{bmatrix}5\\11\end{bmatrix} 14^{-1} (\vec{x}_2 - 5)$$
$$= \frac{1}{14}\begin{bmatrix}5\\11\end{bmatrix}(\vec{x}_2 - 5)$$

$$\Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} = \begin{bmatrix}6 & 8\\8 & 13\end{bmatrix} - \frac{1}{14}\begin{bmatrix}5\\11\end{bmatrix}[5 \quad 11]$$
$$= \begin{bmatrix}6 & 8\\8 & 13\end{bmatrix} - \frac{1}{14}\begin{bmatrix}25 & 55\\55 & 25\end{bmatrix}$$
$$= \begin{bmatrix}\frac{59}{14} & \frac{57}{14}\\ \frac{57}{14} & \frac{157}{14}\end{bmatrix}$$

$$\therefore\ p(\vec{x}_1|\vec{x}_2) = N\left(\frac{1}{14}\begin{bmatrix}5\\11\end{bmatrix}(\vec{x}_2 - 5), \begin{bmatrix}\frac{59}{14} & \frac{57}{14}\\ \frac{57}{14} & \frac{157}{14}\end{bmatrix}\right).$$

d) We want to find $p(\vec{x}_2 | \vec{x}_1)$:

$$\mu_{2|1} = \mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(\vec{x}_1 - \mu_1) = 5 + [5 \quad 11]\underbrace{\begin{bmatrix}6 & 8\\8 & 13\end{bmatrix}^{-1}}\vec{x}_1$$

$1\times 2 \cdot 2\times 2$
$= 1 \times 2$

$$= \frac{1}{6\cdot 13 - 64}\begin{bmatrix}13 & -8\\-8 & 6\end{bmatrix} = \frac{1}{14}\begin{bmatrix}13 & -8\\-8 & 6\end{bmatrix}$$

$$= 5 + \frac{1}{14}\left[5\cdot 13 + 11\cdot -8 \,,\; 5\cdot -8 + 11\cdot 6 \right] \vec{x}_1$$

$$= 5 + \left[-\frac{23}{14}, \frac{13}{7} \right] \vec{x}_1$$

$$\Sigma_{2|1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} = 14 - [5 \;\; 11] \begin{bmatrix} 6 & 8 \\ 8 & 13 \end{bmatrix}^{-1} \begin{bmatrix} 5 \\ 11 \end{bmatrix}$$

$$= 14 - [5 \;\; 11] \frac{1}{14} \begin{bmatrix} 13 & -8 \\ -8 & 6 \end{bmatrix} \begin{bmatrix} 5 \\ 11 \end{bmatrix}$$

$$= 14 - \left[-\frac{23}{14}, \frac{13}{7} \right] \begin{bmatrix} 5 \\ 11 \end{bmatrix}$$

$$= 14 - \left(\frac{-115 + 286}{14} \right)$$

$$= 14 - \frac{171}{14} = \frac{25}{14}$$

$$\boxed{\therefore\; p(\vec{x}_2 \mid \vec{x}_1) = N\left(5 + \left[-\frac{23}{14}, \frac{13}{7} \right] \vec{x}_1 \,,\; \frac{25}{14} \right)}$$

2. Fun with MNIST.

a) L_2 regularized logistic regression has loss function:

$$L(\hat{y}, y) = -y \ln \hat{y} - (1-y) \ln (1-\hat{y})$$

$$L_{reg} = L(\hat{y}, y) + \lambda \underbrace{\|\theta\|^2}_{\text{norm of weights}}$$

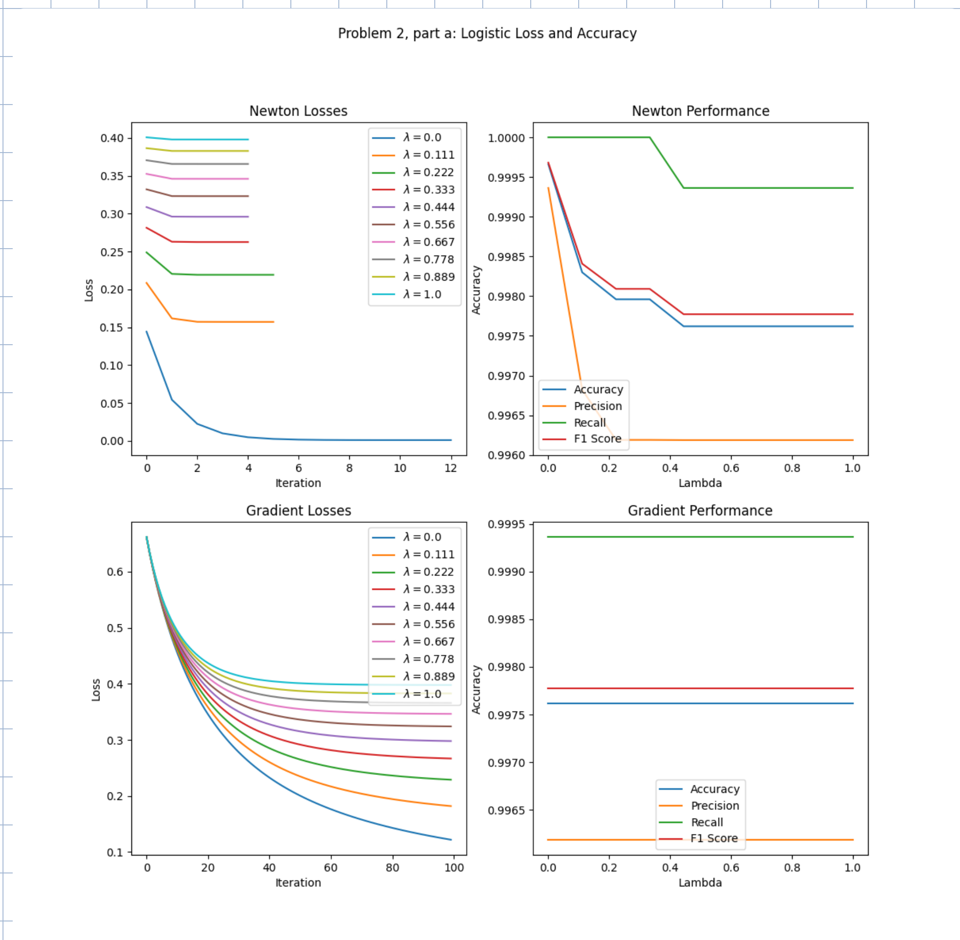
reg. param.

Newton's method is: $x_{n+1} = x_n - (\nabla^2 L)^{-1} \nabla L$

Gradient descent is: $x_{n+1} = x_n - \alpha \nabla f(x_n)$

Note, $\nabla L = X^T(\hat{y} - y) + \lambda \theta$ by the chain rule since $\hat{y} = \sigma(X\theta)$

$$\nabla^2 L = X^T \operatorname{diag}(\hat{y}(1-\hat{y})) X + \lambda I$$



Best acc is > 0.9995 on test.

Best acc is ~ 0.9976 on test.

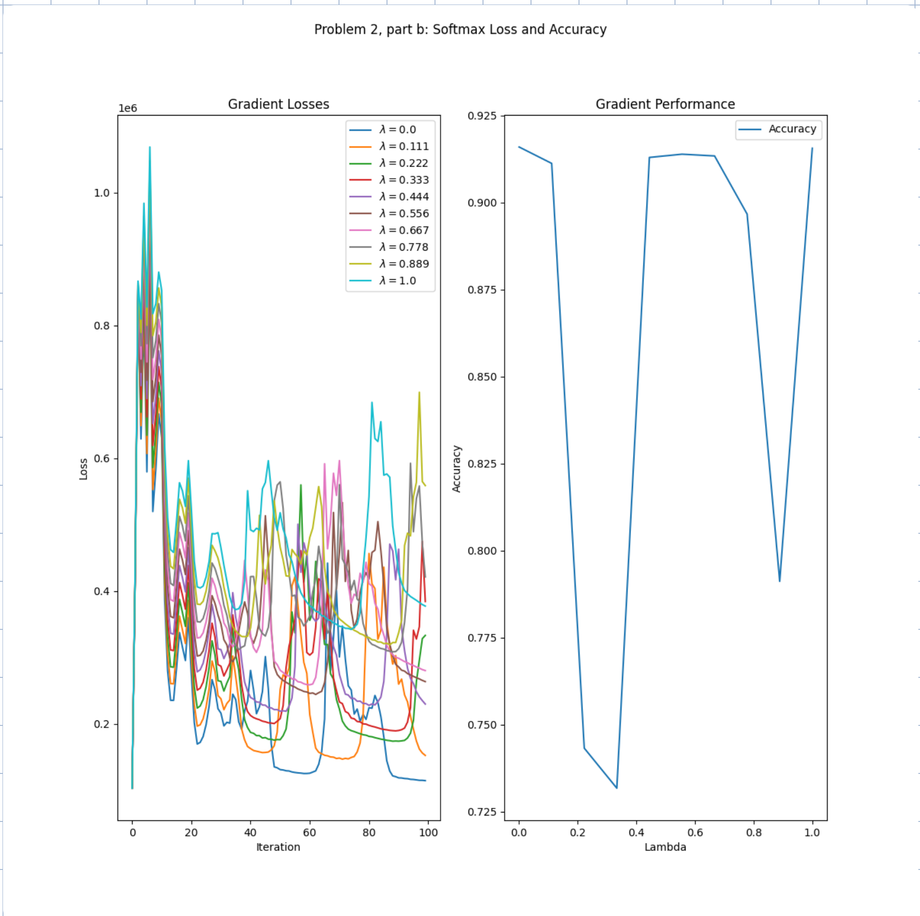
We see that not only is Newton method more accurate in the end, but also it converges much quicker.

b) Here we implement softmax regression, which has loss function

$$L(y,\hat{y}) = -\sum_{i=1}^{c} y_i \ln \hat{y}_i \quad , \text{ where } y \text{ is a 1-hot encoded vector}$$

$$\hat{y}_i = \text{softmax}(y_i) = \frac{e^{(y_i - \max(y_i))}}{\sum_{j=1}^{c} e^{y_j}}$$

$$\nabla L(y,\hat{y}) = X^T(\hat{y} - y) + \lambda \theta \quad \text{by chain rule.}$$



Best accuracy is ~0.92 on test.