

**1 (Conditioning a Gaussian)** Note that from Murphy page 113. “Equation 4.69 is of such importance in this book that we have put a box around it, so you can easily find it.” That equation is important. Read through the proof of the result. Suppose we have a distribution over random variables  $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)$  that is jointly Gaussian with parameters

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \quad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix},$$

where

$$\mu_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad \mu_2 = 5, \quad \boldsymbol{\Sigma}_{11} = \begin{bmatrix} 6 & 8 \\ 8 & 13 \end{bmatrix}, \quad \boldsymbol{\Sigma}_{21}^\top = \boldsymbol{\Sigma}_{12} = \begin{bmatrix} 5 \\ 11 \end{bmatrix}, \quad \boldsymbol{\Sigma}_{22} = [14].$$

Compute

- (a) The marginal distribution  $p(\mathbf{x}_1)$ .
- (b) The marginal distribution  $p(\mathbf{x}_2)$ .
- (c) The conditional distribution  $p(\mathbf{x}_1|\mathbf{x}_2)$
- (d) The conditional distribution  $p(\mathbf{x}_2|\mathbf{x}_1)$

(a)

By Equation 4.69, we have that:

$$p(\mathbf{x}_1) = \mathcal{N}(\mathbf{x}_1 | \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}) = \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 6 & 8 \\ 8 & 13 \end{bmatrix}\right).$$

(b)

By Equation 4.69, we have that:

$$p(\mathbf{x}_2) = \mathcal{N}(\mathbf{x}_2 | \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22}) = \mathcal{N}(5, [14]).$$

(c)

By Equation 4.69, we have that:

$$p(\mathbf{x}_1 | \mathbf{x}_2) = \mathcal{N}(\mathbf{x}_1 | \boldsymbol{\mu}_{1|2}, \boldsymbol{\Sigma}_{1|2}).$$

First, we will calculate  $\mu_{1|2}$ . We again use Equation 4.69 to see that:

$$\begin{aligned}\mu_{1|2} &= \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{x}_2 - \mu_2), \text{ where } \Sigma_{22}^{-1} = \begin{bmatrix} \frac{1}{14} \end{bmatrix}, \\ &= \begin{bmatrix} 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 5 \\ 11 \end{bmatrix} \begin{bmatrix} \frac{1}{14} \end{bmatrix} (\mathbf{x}_2 - 5) \\ &= \begin{bmatrix} \frac{5}{14} \\ \frac{11}{14} \end{bmatrix} (\mathbf{x}_2 - 5).\end{aligned}$$

Next, we will calculate  $\Sigma_{1|2}$  using Equation 4.69:

$$\begin{aligned}\Sigma_{1|2} &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \\ &= \begin{bmatrix} 6 & 8 \\ 8 & 13 \end{bmatrix} - \begin{bmatrix} 5 \\ 11 \end{bmatrix} \begin{bmatrix} \frac{1}{14} \end{bmatrix} \begin{bmatrix} 5 & 11 \end{bmatrix} \\ &= \begin{bmatrix} 6 & 8 \\ 8 & 13 \end{bmatrix} - \begin{bmatrix} \frac{5}{14} \\ \frac{11}{14} \end{bmatrix} \begin{bmatrix} 5 & 11 \end{bmatrix} \\ &= \begin{bmatrix} 6 & 8 \\ 8 & 13 \end{bmatrix} - \begin{bmatrix} \frac{25}{14} & \frac{55}{14} \\ \frac{55}{14} & \frac{121}{14} \end{bmatrix} \\ &= \begin{bmatrix} \frac{59}{14} & \frac{57}{14} \\ \frac{57}{14} & \frac{61}{14} \end{bmatrix}.\end{aligned}$$

Thus, we see that:

$$p(\mathbf{x}_1 | \mathbf{x}_2) = \mathcal{N}(\mathbf{x}_1 | \mu_{1|2}, \Sigma_{1|2}) = \mathcal{N}\left(\begin{bmatrix} \frac{5}{14} \\ \frac{11}{14} \end{bmatrix} (\mathbf{x}_2 - 5), \begin{bmatrix} \frac{59}{14} & \frac{57}{14} \\ \frac{57}{14} & \frac{61}{14} \end{bmatrix}\right).$$

(d)

By Equation 4.69, we have that:

$$p(\mathbf{x}_2 | \mathbf{x}_1) = \mathcal{N}(\mathbf{x}_2 | \mu_{2|1}, \Sigma_{2|1}).$$

First, we will calculate  $\mu_{2|1}$ . We again use Equation 4.69 to see that:

$$\begin{aligned}\mu_{2|1} &= \mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(\mathbf{x}_1 - \mu_1), \text{ where } \Sigma_{11}^{-1} = \frac{1}{|\Sigma_{11}|} \begin{bmatrix} 13 & -8 \\ -8 & 13 \end{bmatrix} = \begin{bmatrix} \frac{13}{14} & -\frac{4}{7} \\ -\frac{4}{7} & \frac{3}{7} \end{bmatrix}, \\ &= 5 + \begin{bmatrix} 5 & 11 \end{bmatrix} \begin{bmatrix} \frac{13}{14} & -\frac{4}{7} \\ -\frac{4}{7} & \frac{3}{7} \end{bmatrix} (\mathbf{x}_1 - \begin{bmatrix} 0 \\ 0 \end{bmatrix}) \\ &= 5 + \begin{bmatrix} -\frac{23}{14} & \frac{13}{7} \end{bmatrix} (\mathbf{x}_1).\end{aligned}$$

Next, we will calculate  $\Sigma_{2|1}$  using Equation 4.69:

$$\begin{aligned}
\Sigma_{2|1} &= \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} \\
&= [14] - [5 \ 11] \begin{bmatrix} \frac{13}{14} & -\frac{4}{7} \\ -\frac{4}{7} & \frac{3}{7} \end{bmatrix} \begin{bmatrix} 5 \\ 11 \end{bmatrix} \\
&= [14] - [\frac{171}{14}] \\
&= [\frac{25}{24}].
\end{aligned}$$

Thus, we see that:

$$p(\mathbf{x}_2 \mid \mathbf{x}_1) = \mathcal{N}(\mathbf{x}_2 \mid \mu_{2|1}, \Sigma_{2|1}) = \mathcal{N}(5 + [-\frac{23}{14} \ \frac{13}{7}] \mathbf{x}_1, [\frac{25}{24}]).$$

■

**2 (MNIST)** In this problem, we will use the MNIST dataset, a classic in the deep learning literature as a toy dataset to test algorithms on, to set up a model for logistic regression and softmax regression. In the starter code, we have already parsed the data for you. However, you might need internet connection to access the data and therefore successfully run the starter code.

The problem is this: we have images of handwritten digits with  $28 \times 28$  pixels in each image, as well as the label of which digit  $0 \leq \text{label} \leq 9$  the written digit corresponds to. Given a new image of a handwritten digit, we want to be able to predict which digit it is. The format of the data is `label, pix-11, pix-12, pix-13, ...` where `pix-ij` is the pixel in the  $i$ th row and  $j$ th column.

- (a) (**logistic**) Restrict the dataset to only the digits with a label of 0 or 1. Implement L2 regularized logistic regression as a model to compute  $\mathbb{P}(y = 1|\mathbf{x})$  for a different value of the regularization parameter  $\lambda$ . Plot the learning curve (objective vs. iteration) when using Newton's Method *and* gradient descent. Plot the accuracy, precision ( $p = \mathbb{P}(y = 1|\hat{y} = 1)$ ), recall ( $r = \mathbb{P}(\hat{y} = 1|y = 1)$ ), and F1-score ( $F1 = 2pr/(p + r)$ ) for different values of  $\lambda$  (try at least 10 different values including  $\lambda = 0$ ) on the test set and report the value of  $\lambda$  which maximizes the accuracy on the test set. What is your accuracy on the test set for this model? Your accuracy should definitely be over 90%.
- (b) (**softmax**) Now we will use the whole dataset and predict the label of each digit using L2 regularized softmax regression (multinomial logistic regression). Implement this using gradient descent, and plot the accuracy on the test set for different values of  $\lambda$ , the regularization parameter. Report the test accuracy for the optimal value of  $\lambda$  as well as its learning curve. Your accuracy should be over 90%.

(a)

In Homework 2, we derived the gradient of the negative log-likelihood for logistic regression, which was:

$$\begin{aligned}\nabla_{\mathbf{w}} NLL(\mathbf{w}) &= \mathbf{X}^T(\boldsymbol{\mu} - \mathbf{y}), \text{ so} \\ \nabla_{\boldsymbol{\theta}} NLL(\boldsymbol{\theta}) &= \mathbf{X}^T(\sigma(\mathbf{X}\boldsymbol{\theta}) - \mathbf{y}).\end{aligned}$$

We must add the weight term, which Equation 8.31 gives us is  $\mathbf{g}'(\mathbf{w}) = \mathbf{g}(\mathbf{w}) + \lambda \mathbf{w}$ , so

$$\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}) = \mathbf{X}^T(\sigma(\mathbf{X}\boldsymbol{\theta}) - \mathbf{y}) + \lambda \boldsymbol{\theta}.$$

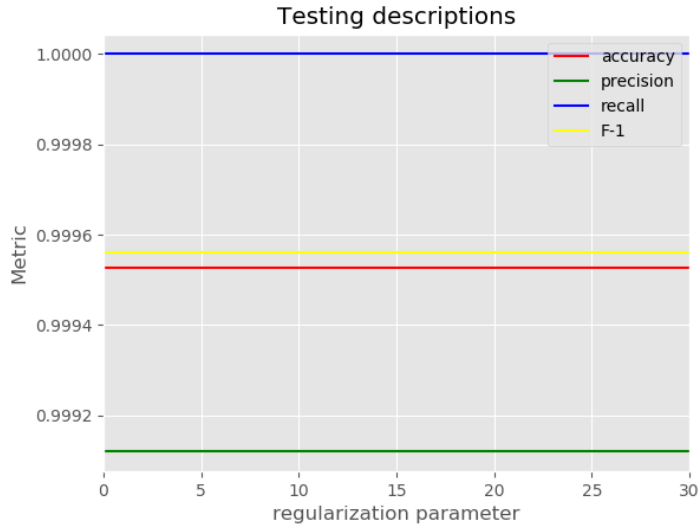
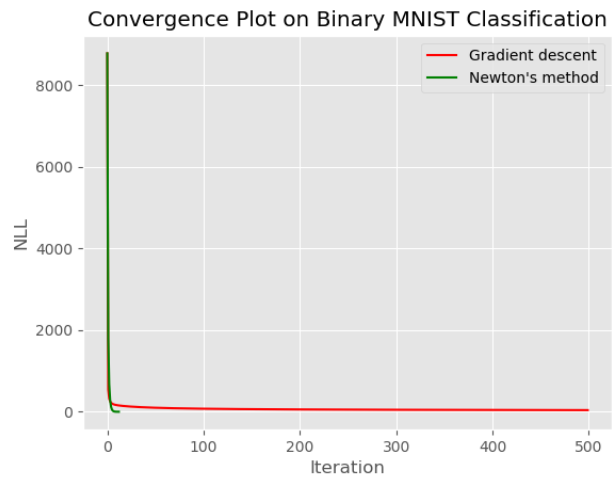


Figure 1: I realize that the lines should not be horizontal, as we would expect changing the regularization parameter to have an affect on the output. This means there is a bug in my code, but I was not able to find it.

(b)

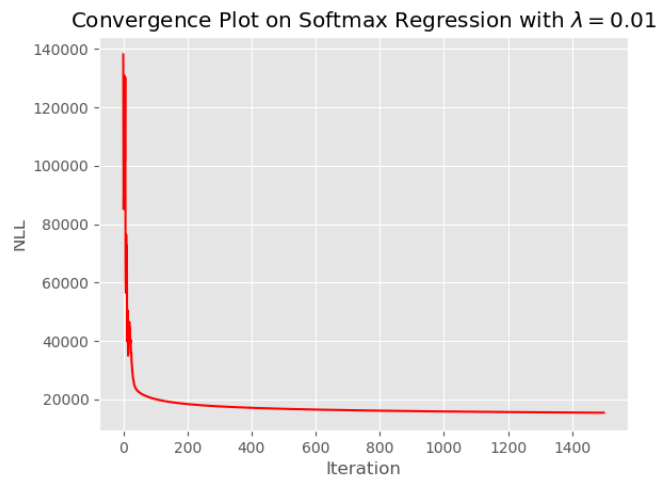


Figure 2: The NLL on the provided graph increases as the number of iterations increases, but on this graph, the NLL decreases. I was not able to find the bug that caused this issue.

