Math189R SP19
Homework 2
Monday, Feb 11, 2019

Feel free to work with other students, but make sure you write up the homework and code on your own (no copying homework *or* code; no pair programming). Feel free to ask students or instructors for help debugging code or whatever else, though.

The starter files can be found under the Resource tab on course website. The graphs for problem 3 generated by the sample solution could be found in the corresponding zipfile. These graphs only serve as references to your implementation. You should generate your own graphs for submission. Please print out all the graphs generated by your own code and submit them together with the written part, and make sure you upload the code to your Github repository.

1 (**Murphy 8.3**) Gradient and Hessian of the log-likelihood for logistic regression.
(a) Let $\sigma(x) = \frac{1}{1+e^{-x}}$ be the sigmoid function. Show that

$$\sigma'(x) = \sigma(x)\left[1 - \sigma(x)\right].$$

(b) Using the previous result and the chain rule of calculus, derive an expression for the gradient of the log likelihood for logistic regression.
(c) The Hessian can be written as $\mathbf{H} = \mathbf{X}^\top \mathbf{S} \mathbf{X}$ where $\mathbf{S} = \text{diag}(\mu_1(1-\mu_1), \ldots, \mu_n(1-\mu_n))$. Derive this and show that $\mathbf{H} \succeq 0$ ($A \succeq 0$ means that A is positive semidefinite).

Hint: Use the **negative** log-likelihood of logistic regression for this problem.

a. We know $\sigma(x) = (1+e^{-x})^{-1}$

Then, $\sigma'(x) = \frac{-1}{(1+e^{-x})^2} \cdot -e^{-x} = \frac{e^{-x}}{(1+e^{-x})^2} = \frac{1+e^{-x}-1}{(1+e^{-x})^2} = \frac{1}{1+e^{-x}} - \frac{1}{(1+e^{-x})^2} = \sigma(x) - \sigma(x)^2$

Factoring out $\sigma(x)$ from the above expression yields:

$$\sigma'(x) = \sigma(x)[1 - \sigma(x)]$$

b. The negative log likelihood for logistic regression, $l(\theta)$, is given by the expression:

$$l(\boldsymbol{\theta}) = -\Sigma_i^N y^{(i)} \log[\sigma(\boldsymbol{\theta}^T \mathbf{X}^{(i)})] + (1 - y^{(i)}) \log[1 - \sigma(\boldsymbol{\theta}^T \mathbf{X}^{(i)})]$$

Then, $\nabla l(\boldsymbol{\theta})$ is given by

$$\nabla l(\boldsymbol{\theta}) = -\Sigma_i^N \frac{y^{(i)}}{\sigma(\boldsymbol{\theta}^T \mathbf{X}^{(i)})} \cdot \sigma(\boldsymbol{\theta}^T \mathbf{X}^{(i)})[1 - \sigma(\boldsymbol{\theta}^T \mathbf{X}^{(i)})] \cdot \mathbf{X}^{(i)} + (1 - y^{(i)}) \cdot \frac{-\sigma(\boldsymbol{\theta}^T \mathbf{X}^{(i)})[1 - \sigma(\boldsymbol{\theta}^T \mathbf{X}^{(i)})]}{1 - \sigma(\boldsymbol{\theta}^T \mathbf{X}^{(i)})} \cdot \mathbf{X}^{(i)}$$

$$= -\Sigma_i^N y^{(i)} \cdot [1 - \sigma(\boldsymbol{\theta}^T \mathbf{X}^{(i)})] \cdot \mathbf{X}^{(i)} + (y^{(i)} - 1)\sigma(\boldsymbol{\theta}^T \mathbf{X}^{(i)}) \cdot \mathbf{X}^{(i)}$$

$$= \Sigma_i^N [\sigma(\boldsymbol{\theta}^T \mathbf{X}^{(i)}) - y^{(i)}] \cdot \mathbf{X}^{(i)}$$

Note here that θ and $\mathbf{X}^i$ are vectors. The gradient of individual elements of θ, is given by:

$$\frac{\partial}{\partial \theta_j} l(\theta) = \Sigma_i^N [\sigma(\boldsymbol{\theta}^T \mathbf{X}^{(i)}) - y^{(i)}] \cdot X_j^{(i)}$$

We can also do away with the summation by expressing the computation of the gradient with matrix multiplication:

$$\nabla l(\boldsymbol{\theta}) = \mathbf{X}^T (\sigma(\mathbf{X}\boldsymbol{\theta}) - \mathbf{y})$$

c. We know that $H = \frac{d}{d\boldsymbol{\theta}} \nabla l(\boldsymbol{\theta})^T$. From this we find:

$$H = \frac{d}{d\boldsymbol{\theta}} (\Sigma_i^N [\sigma(\boldsymbol{\theta}^T \mathbf{X}^{(i)}) - y^{(i)}] \cdot \mathbf{X}^{(i)} T) = \Sigma_i^N \sigma(\boldsymbol{\theta}^T \mathbf{X}^{(i)}) [1 - \sigma(\boldsymbol{\theta}^T \mathbf{X}^{(i)})] \cdot \mathbf{X}^{(i)} \cdot \mathbf{X}^{(i)T}$$

Let $\sigma(\boldsymbol{\theta}^T \mathbf{X}^{(i)}) = \mu_i$. Then,

$$\mathbf{H} = \Sigma_i^N \mu_i [1 - \mu_i] \cdot \mathbf{X}^{(i)} \cdot \mathbf{X}^{(i)T}$$

The first multiplication term in the summation can be rolled into a matrix multiplication of the form $\mathbf{X}^T \cdot \mathbf{S}$ where $\mathbf{S} = \text{diag}(\mu_1(1 - \mu_1), \ldots, \mu_n(1 - \mu_n))$, as we are effectively multiplying the columns of $\mathbf{X}^T$, which are the $\mathbf{X}^{(i)}$ vectors by the diagonal of $\mathbf{S}$. Now we have to multiple by $\mathbf{X}$ to take into account the $\mathbf{X}^{(i)T}$ term. This yields the result:

$$\mathbf{H} = \mathbf{X}^T \mathbf{S} \mathbf{X}$$

We will now prove that $\mathbf{H}$ is positive semidefinite.

First note that $\mathbf{S}$ is symmetric by its definition as a diagonal matrix. Also note that since $\mathbf{S}$ is a diagonal matrix, it is diagonally dominant, and the real parts of its eigenvalues arse positive.[1]. Since we are not dealing with the complex domain here, we may assume that the eigenvalues of $\mathbf{S}$ are positive.

Since $\mathbf{S}$ is real-symmetric and has positive real eigenvalues, the $\mathbf{S}$ is positive definite. [2]

Therefore, $\mathbf{w}^T \mathbf{S} \mathbf{w} > 0 \ \forall \ \mathbf{w} \in R^n$

We now need to show that $\mathbf{z}^T H \mathbf{z} > 0 \ \forall \ \mathbf{z} \in R^m$, where m is the size of H.

[1] Theorem taken from http://mathworld.wolfram.com/DiagonallyDominantMatrix.html

[2] Theorem taken from https://yutsumura.com/positive-definite-real-symmetric-matrix-and-its-eigenvalues/ #b_Prove_that_if_eigenvalues_of_a_real_symmetric_matrix_A_are_all_positive_then_A_is_ positive-definite

Let $\mathbf{z} \in \mathbf{R}^m$ be some arbitrary vector. Then,

$$\mathbf{z}^T\mathbf{H}\mathbf{z} = \mathbf{z}^T\mathbf{X}^T S\mathbf{X}\mathbf{z} = (\mathbf{X}\mathbf{z})^T\mathbf{S}(\mathbf{X}\mathbf{z})$$

Note that $\mathbf{X}\mathbf{z} \in R^n$. Since S i positive definite, we have:

$$(\mathbf{X}\mathbf{z})^T\mathbf{S}(\mathbf{X}\mathbf{z}) > 0$$

Thus proven. ■

2 (**Murphy 2.11**) Derive the normalization constant (Z) for a one dimensional zero-mean Gaussian

$$\mathbb{P}(x;\sigma^2) = \frac{1}{Z}\exp\left(-\frac{x^2}{2\sigma^2}\right)$$

such that $\mathbb{P}(x;\sigma^2)$ becomes a valid density.

In order for the density to be valid, $\int_{-\infty}^{\infty}\frac{1}{Z}\text{exp}(-\frac{x^2}{2\sigma^2}) = 1$. Let $I = \int_{-\infty}^{\infty}\frac{1}{Z}\text{exp}(-\frac{x^2}{2\sigma^2})$ Then,

$$I^2 == \frac{1}{Z^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} exp(-\frac{x^2}{2\sigma^2})exp(-\frac{y^2}{2\sigma^2})\cdot dx \cdot dy = \frac{1}{Z^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} exp(-\frac{x^2+y^2}{2\sigma^2})\cdot dy\cdot dx$$

We can change the integral to polar coordinates, which yields:

$$I^2 = \frac{1}{Z^2}\int_0^{2\pi}\int_0^{\infty} exp(\frac{-r^2}{2\sigma^2})\cdot r\cdot dr\cdot d\theta$$

Let $t = \frac{r^2}{2\sigma^2}$. Then $dr = \frac{\sigma^2\cdot dt}{r}$. Plugging this substitution into the integral yields:

$$I^2 = \frac{\sigma^2}{Z^2}\int_0^{2\pi}\int_0^{\infty} exp(-t)\cdot dt\cdot d\theta = \frac{\sigma^2}{Z^2}\int_0^{2\pi} d\theta = \frac{2\pi\sigma^2}{Z^2}$$

Since $I^2 = 1$, we get:

$$Z = \sqrt{2\pi}\sigma$$

■

3 (**regression**). In this problem, we will use the online news popularity dataset to set up a model for linear 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.

We split the csv file into a training and test set with the first two thirds of the data in the training set and the rest for testing. Of the testing data, we split the first half into a 'validation set' (used to optimize hyperparameters while leaving your testing data pristine) and the remaining half as your test set. We will use this data for the remainder of the problem. The goal of this data is to predict the **log** number of shares a news article will have given the other features.

(a) (**math**) Show that the maximum a posteriori problem for linear regression with a zero-mean Gaussian prior $\mathbb{P}(\mathbf{w}) = \prod_j \mathcal{N}(w_j|0, \tau^2)$ on the weights,

$$\arg\max_{\mathbf{w}} \sum_{i=1}^{N} \log \mathcal{N}(y_i|w_0 + \mathbf{w}^\top \mathbf{x}_i, \sigma^2) + \sum_{j=1}^{D} \log \mathcal{N}(w_j|0, \tau^2)$$

is equivalent to the ridge regression problem

$$\arg\min \frac{1}{N} \sum_{i=1}^{N} (y_i - (w_0 + \mathbf{w}^\top \mathbf{x}_i))^2 + \lambda ||\mathbf{w}||_2^2$$

with $\lambda = \sigma^2/\tau^2$.

(b) (**math**) Find a closed form solution $\mathbf{x}^\star$ to the ridge regression problem:

$$\text{minimize: } ||A\mathbf{x} - \mathbf{b}||_2^2 + ||\Gamma\mathbf{x}||_2^2.$$

(c) (**implementation**) Attempt to predict the log shares using ridge regression from the previous problem solution. Make sure you include a bias term and *don't regularize the bias term*. Find the optimal regularization parameter λ from the validation set. Plot both λ versus the validation RMSE (you should have tried at least 150 parameter settings randomly chosen between 0.0 and 150.0 because the dataset is small) and λ versus $||\boldsymbol{\theta}^\star||_2$ where $\boldsymbol{\theta}$ is your weight vector. What is the final RMSE on the test set with the optimal $\lambda^\star$?

(continued on the following pages)

■

3 (continued)

(d) (**math**) Consider regularized linear regression where we pull the basis term out of the feature vectors. That is, instead of computing $\hat{\mathbf{y}} = \boldsymbol{\theta}^\top \mathbf{x}$ with $\mathbf{x}_0 = 1$, we compute $\hat{\mathbf{y}} = \boldsymbol{\theta}^\top \mathbf{x} + b$. This corresponds to solving the optimization problem

$$\text{minimize: } ||A\mathbf{x} + b\mathbf{1} - \mathbf{y}||_2^2 + ||\Gamma\mathbf{x}||_2^2.$$

Solve for the optimal $\mathbf{x}^\star$ explicitly. Use this close form to compute the bias term for the previous problem (with the same regularization strategy). Make sure it is the same.

(e) (**implementation**) We can also compute the solution to the least squares problem using gradient descent. Consider the same bias-relocated objective

$$\text{minimize: } f = ||A\mathbf{x} + b\mathbf{1} - \mathbf{y}||_2^2 + ||\Gamma\mathbf{x}||_2^2.$$

Compute the gradients and run gradient descent. Plot the ℓ_2 norm between the optimal $(\mathbf{x}^\star, b^\star)$ vector you computed in closed form and the iterates generated by gradient descent. Hint: your plot should move down and to the left and approach zero as the number of iterations increases. If it doesn't, try decreasing the learning rate.

a. From above, we have:

$$\arg\max_{\mathbf{w}} \sum_{i=1}^{N} \log \mathcal{N}(y_i | w_0 + \mathbf{w}^\top \mathbf{x}_i, \sigma^2) + \sum_{j=1}^{D} \log \mathcal{N}(w_j | 0, \tau^2)$$

We will proceed by simplifying this expression to derive the ridge regression result. The above expression yields:

$$\arg\max_{\mathbf{w}} \sum_{i=1}^{N} \log(\frac{1}{\sqrt{2\pi}\sigma} \exp{(-\frac{(y_i - (w_0 + \mathbf{w}^T\mathbf{x}_i))^2}{2\sigma^2})} + \sum_{j=1}^{D} \log(\frac{\exp{(-\frac{w_j^2}{2\tau^2})}}{\sqrt{2\pi}\tau}))$$

$$= \arg\max_{\mathbf{w}} \sum_{i=1}^{N} \log(\frac{1}{\sqrt{2\pi}\sigma}) - \frac{(y_i - (w_0 + \mathbf{w}^T\mathbf{x}_i))^2}{2\sigma^2} + \sum_{j=1}^{D} \log(\frac{1}{\sqrt{2\pi}\tau}) - \frac{w_j^2}{2\tau^2}$$

Note that the $\frac{1}{\sqrt{2\pi}\sigma}$ and the $\frac{1}{\sqrt{2\pi}\tau}$ terms can be removed from this optimisation expression, and they will yield only constants outside the summations which will have no bearing on the optimal $\mathbf{w}$. Also note that maximising some function $f(\mathbf{w})$ is equivalent to minimising $-f(\mathbf{w})$. The above expression is therefore equivalent to:

$$\arg\min_{\mathbf{w}} \sum_{i=1}^{N} \frac{(y_i - (w_0 + \mathbf{w}^T\mathbf{x}_i))^2}{2\sigma^2} + \sum_{j=1}^{D} \frac{w_j^2}{2\tau^2}$$

$$= \arg\min_{\mathbf{w}} \sum_{i=1}^{N} \frac{(y_i - (w_0 + \mathbf{w}^T \mathbf{x}_i))^2}{2\lambda\tau^2} + \frac{1}{2\tau^2}||\mathbf{w}||_2^2$$

$$= \arg\min_{\mathbf{w}} \frac{1}{2\lambda\tau^2} \sum_{i=1}^{N} (y_i - (w_0 + \mathbf{w}^T \mathbf{x}_i))^2 + \lambda||\mathbf{w}||_2^2$$

Since multiplying or dividing by constants will not affect the optimal value of **w**, the above expression is equivalent to:

$$= \arg\min_{\mathbf{w}} \frac{1}{N} \sum_{i=1}^{N} (y_i - (w_0 + \mathbf{w}^T \mathbf{x}_i))^2 + \lambda||\mathbf{w}||_2^2$$

b. Set $J(\mathbf{x}) = ||A\mathbf{x} - \mathbf{b}||_2^2 + ||\Gamma\mathbf{x}||_2^2$. We will proceed by minimising J by taking its gradient and setting it equal to 0. Note that minimising $J(\mathbf{x})$ is equivalent to minimising $(A\mathbf{x} - \mathbf{b})(A\mathbf{x} - \mathbf{b}) + (\Gamma\mathbf{x})(\Gamma\mathbf{x})$. We therefore get:

$$2 \cdot A^T(A\mathbf{x}^\star - \mathbf{b}) + 2 \cdot \Gamma^2\mathbf{x}^\star = 0$$

Then,

$$A^T A\mathbf{x}^\star + \Gamma^2\mathbf{x}^\star = A^T\mathbf{b}$$

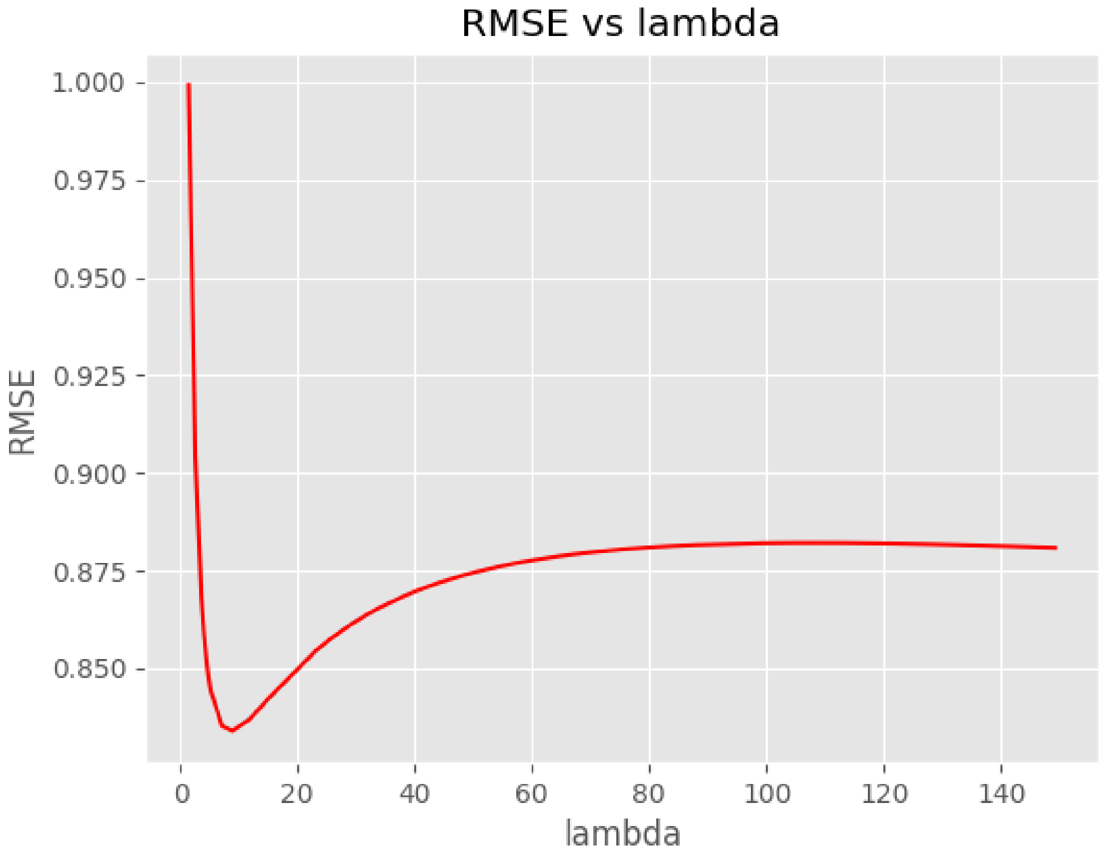
Which results in

$$\mathbf{x}^\star = (A^T A + \Gamma^2 I_D)^{-1} A^T\mathbf{b}$$

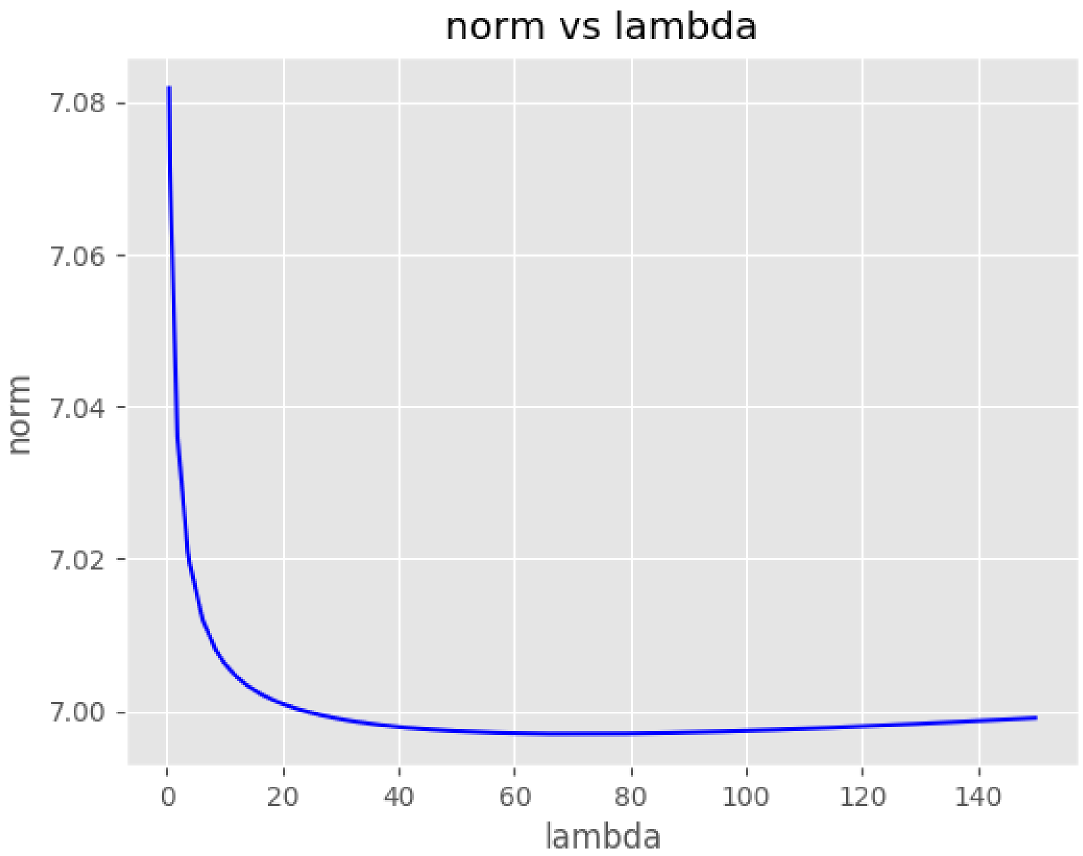
Where D is the number of columns in A.

c. My final RMSE on the test set with optimal λ was 0.8628. The plots are given below:

RMSE vs λ:



$\boldsymbol{\theta}$ norm vs λ



d. I had to consult the solutions to answer this problem. At first I did not understand what it was asking for. The solution gave me an idea on where and how to start. The solution also guided me in choosing the order of operations for my matrix multiplication.

Set $J(\mathbf{x}) = ||A\mathbf{x} + b\mathbf{1} - \mathbf{y}||_2^2 + ||\Gamma\mathbf{x}||_2^2$. This expression is equivalent to:

$$J(\mathbf{x}, b) = (A\mathbf{x} + b\mathbf{1} - \mathbf{y})^T(A\mathbf{x} + b\mathbf{1} - \mathbf{y}) + (\Gamma\mathbf{x})^T(\Gamma\mathbf{x})$$

$$= (\mathbf{x}^T A^T + b\mathbf{1}^T - \mathbf{y}^T)(A\mathbf{x} + b\mathbf{1} - \mathbf{y}) + \mathbf{x}^T\Gamma^T\Gamma\mathbf{x}$$

$$= \mathbf{x}^T A^T A\mathbf{x} + \mathbf{x}^T A^T b\mathbf{1} - \mathbf{x}^T A^T\mathbf{y} + b\mathbf{1}^T A\mathbf{x} + b^2\mathbf{1}^T\mathbf{1} - b1^T\mathbf{y} - \mathbf{y}^T A\mathbf{x} - b\mathbf{y}^T\mathbf{1} + \mathbf{y}^T\mathbf{y} + \mathbf{x}^T\Gamma^T\Gamma\mathbf{x}$$

$$= \mathbf{x}^T A^T A\mathbf{x} + 2b\mathbf{1}^T A\mathbf{x} - 2\mathbf{y}^T A\mathbf{x} + b^2 n - 2b\mathbf{1}^T\mathbf{y} + \mathbf{y}^T\mathbf{y} + \mathbf{x}^T\Gamma^T\Gamma\mathbf{x}$$

In order to find b, we will take the gradient of J with respect to b and set it equal to 0. This yields:

$$\mathbf{2}\mathbf{1}^T A\mathbf{x} + 2b^\star n - \mathbf{2}\mathbf{1}^T\mathbf{y} = 0$$

Therefore,

$$b^\star = \frac{\mathbf{1}^T\mathbf{y} - \mathbf{1}^T A\mathbf{x}}{n}$$

Now we will try to find $\mathbf{x}^\star$. We will take the gradient of J with respect to $\mathbf{x}$ and set it equal to 0.

$$2\mathbf{x}^\star A^T A + 2bA^T\mathbf{1} - 2A^T\mathbf{y} + 2\mathbf{x}^\star\Gamma^T\Gamma$$

. Substituting the value of $_\star$ into the above yields:

$$2A^T A\mathbf{x}^\star + \frac{\mathbf{1}^T\mathbf{y}A^T\mathbf{1} - \mathbf{1}^T A\mathbf{x}^\star A^T\mathbf{1}}{n} - 2A^T\mathbf{y} + \Gamma^T\Gamma 2\mathbf{x}^\star = 0$$

Then,

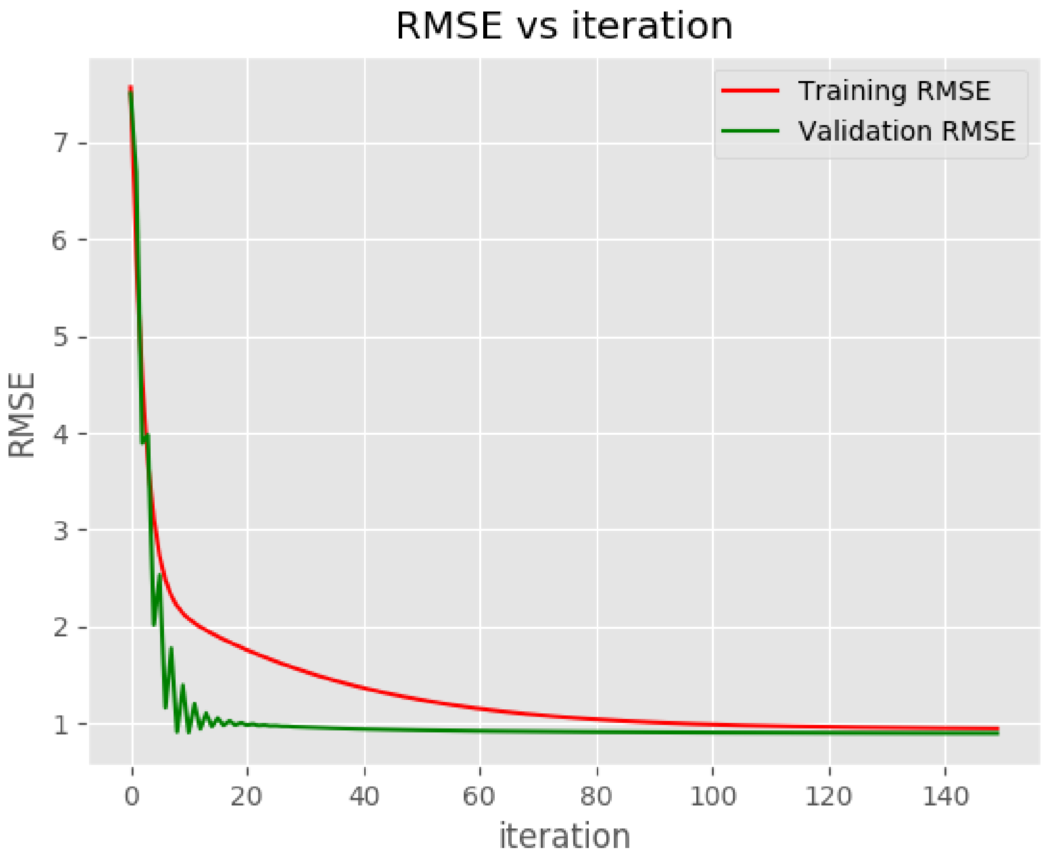
$$(A^T A + \Gamma^T\Gamma - \frac{A^T\mathbf{1}\mathbf{1}^T A}{n})\mathbf{x}^\star = A^T\mathbf{y} - \frac{A^T\mathbf{1}\mathbf{1}^T\mathbf{y}}{n}$$

So,

$$\mathbf{x}^\star = (A^T A + \Gamma^T\Gamma - \frac{A^T\mathbf{1}\mathbf{1}^T A}{n})^{-1}(A^T\mathbf{y} - \frac{A^T\mathbf{1}\mathbf{1}^T\mathbf{y}}{n})$$

The difference in bias and weights is of the order of 10^{-11} and is negligible. We get the same result as before!

e. The convergence is plotted below:



The difference in values is given below:
Difference in bias is 1.5388E-01
Difference in weights is 8.0420E-01