Xingyao Chen
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) First rewrite $\sigma(x)$

$$\begin{aligned}\sigma(x) =& \frac{1}{1+e^{-x}} \\ =& \frac{e^x}{e^x} \cdot \frac{1}{1+e^{-x}} \\ =& \frac{e^x}{e^x+1}\end{aligned}$$

Take the derivative

$$\begin{aligned}
\frac{d}{dx}\sigma(x) =& \frac{d}{dx}(1+e^{-x})^{-1} \\
=& (-1)(1+e^{-x})^{-2}(-e^{-x}) \\
=& \frac{e^{-x}}{(1+e^{-x})(1+e^{-x})} \\
=& \frac{e^{-x}}{(1+e^{-x})}\sigma(x) \\
=& \frac{e^{x}}{e^{x}}\frac{e^{-x}}{(1+e^{-x})}\sigma(x) \\
=& \frac{1}{e^{x}+1}\sigma(x) \\
=& \frac{e^{x}+1-e^{x}}{e^{x}+1}\sigma(x) \\
=& \left[\frac{e^{x}+1}{e^{x}+1} - \frac{e^{x}}{e^{x}+1}\right]\sigma(x) \\
=& [1-\sigma(x)]\sigma(x) \\
=& \sigma(x)[1-\sigma(x)]
\end{aligned}$$

(b) We want to optimize the parameter $\boldsymbol{\theta}$ for the function:

$$h_\theta(\mathbf{x}) = \frac{1}{1+e^{-\boldsymbol{\theta}^T\mathbf{x}}} = g(\boldsymbol{\theta}^T\mathbf{x})$$

Recall that for a Bernoulli distribution, the probability of one data point y, given that $X = \mathbf{x}$ and ` is

$$Pr(y|\mathbf{x};\boldsymbol{\theta}) = g(\boldsymbol{\theta}^T\mathbf{x})^y(1-g(\boldsymbol{\theta}^T\mathbf{x}))^{(1-y)}$$

As derived in class, the log likelihood of $\boldsymbol{\theta}$ over a total of n examples for a sigmoid function is

$$\ell(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta}) = \log\prod_{i=1}^{n}(Pr(y^{(i)}|\mathbf{x}^{(i)};\theta))$$

Expand and simplify

$$\begin{aligned}
&\ell(\boldsymbol{\theta}) \\
=& \sum_{i=1}^{n}\log(Pr(y^{(i)}|\mathbf{x}^{(i)};\theta)) \\
=& \sum_{i=1}^{n}\log\left[g(\boldsymbol{\theta}^T\mathbf{x}^{(i)})^{y^{(i)}}(1-g(\boldsymbol{\theta}^T\mathbf{x}^{(i)}))^{(1-y^{(i)})}\right] \\
=& \sum_{i=1}^{n}\left[y^{(i)}\log h_\theta(\boldsymbol{\theta}^T\mathbf{x}^{(i)}) + (1-y^{(i)})\log(1-h_\theta(\boldsymbol{\theta}^T\mathbf{x}^{(i)}))\right]
\end{aligned}$$

We want to maximize the log likelihood, so we define the cost function $J(\boldsymbol{\theta})$ such that

$$argmax(\ell(\theta)) = argmin(J(\theta))$$

Therefore

$$J(\theta) = -\ell(\theta)$$

And we find the global minimum of $J(\theta)$ by taking its gradient. Consider 1 element of the its gradient:

$$\begin{aligned}
&\frac{\partial}{\partial\theta_j}J(\boldsymbol{\theta}) \\
&= -\frac{\partial}{\partial\theta_{(j)}}\sum_{i=1}^{n}\left[y^{(i)}\log g(\boldsymbol{\theta}^T\mathbf{x}^{(i)}) + (1-y^{(i)})\log(1-g(\boldsymbol{\theta}^T\mathbf{x}^{(i)}))\right] \\
&= -\sum_{i=1}^{n}\left[\frac{y^{(i)}}{h_\theta(\boldsymbol{\theta}^T\mathbf{x}^{(i)})}g(\boldsymbol{\theta}^T\mathbf{x}^{(i)})[1-g(\boldsymbol{\theta}^T\mathbf{x}^{(j)})] + \frac{1-y^{(i)}}{1-g(\boldsymbol{\theta}^T\mathbf{x}^{(i)})}(-1)g(\boldsymbol{\theta}^T\mathbf{x}^{(i)})[1-g(\boldsymbol{\theta}^T\mathbf{x}^{(i)})]\right]\frac{\partial}{\partial\theta_{(j)}}\boldsymbol{\theta}x^{(i)} \\
&= -\sum_{i=1}^{n}\left[y^{(i)}[1-g(\boldsymbol{\theta}^T\mathbf{x}^{(i)})] + [y^{(i)}-1]g(\boldsymbol{\theta}^T\mathbf{x}^{(i)})\right]x_j^{(i)} \\
&= -\sum_{i=1}^{n}[y^{(i)} - h_\theta(\boldsymbol{\theta}^T\mathbf{x}^{(i)})]x_j^{(i)}
\end{aligned}$$

Let $\boldsymbol{\mu}$ be defined as a vector were each $\mu_i = h_\theta(\mathbf{x}^i)$ Therefore the gradient is

$$\begin{aligned}
g(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}}J(\boldsymbol{\theta}) &= \sum_{i=1}^{n}[h_\theta(\mathbf{x}^{(i)}) - y^{(i)}]\mathbf{x}^{(i)} \\
&= \boldsymbol{X}^T[\boldsymbol{\mu} - \boldsymbol{y}]
\end{aligned}$$

(c) Let $\mu_i = h_\theta(\boldsymbol{\theta}^T\mathbf{x}^{(i)})$

$$\begin{aligned}
\boldsymbol{H} = \nabla_{\boldsymbol{\theta}}g(\boldsymbol{\theta})^T &= \sum_{i=1}^{n}\left[\frac{\partial}{\partial\theta_i}\mu_i\right]\mathbf{x}^{(i)T} \\
&= \sum_{i=1}^{n}\mu_i(1-\mu_i)\mathbf{x}^{(i)T}\mathbf{x}^{(i)} \\
&= \boldsymbol{X}^T\boldsymbol{S}\boldsymbol{X}
\end{aligned}$$

Notice that $0 < \mu_i < 1$. Therefore for any vector $\mathbf{v}$, it's the case that

$$\mathbf{v}^T\boldsymbol{H}\mathbf{v} = \sum_{i=1}^{n}\mu_i(1-\mu_i)(\mathbf{v}^T\mathbf{x}^{(i)})^2 \geq 0$$

Therefore, $\boldsymbol{H} \succeq 0$.

■

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 $\mathbb{P}(x;\sigma^2)$ to become a valid density, we must have

$$\int_{\infty}^{\infty} \mathbb{P}(x;\sigma^2)\,dx = 1$$

Solve for Z

$$\int_{-\infty}^{\infty} \mathbb{P}(x;\sigma^2)\,dx = \int_{-\infty}^{\infty} \frac{1}{Z}\exp(-\frac{x^2}{2\sigma^2})\,dx$$

Let

$$I = \int_{-\infty}^{\infty} \exp\left(\frac{-x^2}{2\sigma^2}\right)\,dx$$

also let

$$I = \int_{-\infty}^{\infty} \exp\left(\frac{-y^2}{2\sigma^2}\right)\,dy$$

Then,

$$I^2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\left(\frac{-x^2-y^2}{2\sigma^2}\right)\,dx\,dy$$

Convert to polar coordinates, letting $r^2 = x^2 + y^2$

$$\begin{aligned} I^2 =& \int_0^{2\pi}\int_0^{\infty} r\exp\left(\frac{-r^2}{2\sigma^2}\right)\,dr\,d\theta \\ =& 2\pi\int_0^{\infty} r\exp\left(\frac{-r^2}{2\sigma^2}\right)\,dr \end{aligned}$$

Let $u = r^2$, and $du = 2r\,dr$

$$\begin{aligned}
I^2 =& 2\pi \int_0^\infty r \exp\left(\frac{-u}{2\sigma^2}\right) \frac{1}{2r}\, du \\
=& \pi \int_0^\infty \exp\left(\frac{-u}{2\sigma^2}\right)\, du \\
=& \pi 2\sigma^2 \left[-\exp\left(\frac{-u}{2\sigma^2}\right) \Big|_0^\infty \right] \\
=& \pi 2\sigma^2 \left[-\exp(-\infty) + \exp(0) \right] \\
=& 2\pi\sigma^2 \\
I =& \sqrt{2\pi}\sigma
\end{aligned}$$

Therefore,

$$\begin{aligned}
\int_\infty^\infty \mathbb{P}(x;\sigma^2)\, dx =& 1 = \frac{1}{Z}\sqrt{2\pi}\sigma \\
Z =& \boxed{\sqrt{2\pi}\sigma}
\end{aligned}$$

■

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) We know that the PDF for a normal distribution is

$$\begin{aligned}
\mathbb{P}(y|\omega_0 + \omega x; \sigma^2) &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{[y - (\omega_0 + \mathbf{w}^T\mathbf{x})]^2}{2\sigma^2}\right) \\
\mathbb{P}(\omega; \tau^2) &= \frac{1}{\sqrt{2\pi}\tau} \exp\left(-\frac{\omega^2}{2\tau^2}\right)
\end{aligned}$$

Let $\ell(\mathbf{w}) = \sum_{i=1}^N \log \mathcal{N}(y_i|w_0 + \mathbf{w}^\top \mathbf{x}_i, \sigma^2) + \sum_{j=1}^D \log \mathcal{N}(w_j|0, \tau^2)$, simplify the equation:
For the first piece:

$$\begin{aligned}
\sum_{i=1}^N \log \mathcal{N}(y_i|w_0 + \mathbf{w}^\top \mathbf{x}_i, \sigma^2) &= \sum_{i=1}^N \log\left[\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{[y_i - (\omega_0 + \mathbf{w}^T\mathbf{x}_i)]^2}{2\sigma^2}\right)\right] \\
&= \sum_{i=1}^N \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{[y_i - (\omega_0 + \mathbf{w}^T\mathbf{x}_i)]^2}{2\sigma^2} \\
&= -\frac{N}{2\sigma^2} \log \frac{1}{\sqrt{2\pi}} \sum_{i=1}^N [y_i - (\omega_0 + \mathbf{w}^T x_i]^2
\end{aligned}$$

For the second peice:

$$\sum_{j=1}^{D} \log \mathcal{N}(w_j|0,\tau^2) = \sum_{j=1}^{D} \log \left[\frac{1}{\sqrt{2\pi}\tau} \exp\left(-\frac{\omega_j^2}{2\tau^2} \right) \right]$$
$$= -\frac{N}{2\tau^2} \log \frac{1}{\sqrt{2\pi}\tau} \sum_{j=1}^{D} \omega_j^2$$

Therefore,

$$\ell(\mathbf{w}) = -\frac{N}{2\sigma^2} \log \frac{1}{\sqrt{2\pi}\sigma} \sum_{i=1}^{N} [y_i - (\omega_0 + \mathbf{w}^T x_i]^2 - \frac{N}{2\tau^2} \log \frac{1}{\sqrt{2\pi}\tau} \sum_{j=1}^{D} \omega_j^2$$

Since the terms $\log \frac{1}{\sqrt{2\pi}\sigma}$ and $\log \frac{1}{\sqrt{2\pi}\tau}$ are constant, it does not impact the optimal value of w_0 and $\mathbf{w}$, therefore we are left with

$$\ell(\mathbf{w}) = -\frac{N}{2\sigma^2} \sum_{i=1}^{N} [y_i - (\omega_0 + \mathbf{w}^T x_i]^2 - \frac{N}{2\tau^2} \sum_{j=1}^{D} \omega_j^2$$

Multiplying by positive constants also do not impact the optimal value of w_0 and $\mathbf{w}$, so we can do the following

$$\arg\max \ell(\mathbf{w}) = \arg\max \frac{2\sigma^2}{N} \left[-\frac{N}{2\sigma^2} \sum_{i=1}^{N} [y_i - (\omega_0 + \mathbf{w}^T x_i]^2 - \frac{N}{2\tau^2} \sum_{j=1}^{D} \omega_j^2 \right]$$
$$= \arg\max - \sum_{i=1}^{N} [y_i - (\omega_0 + \mathbf{w}^T x_i]^2 + \frac{\sigma^2}{\tau^2} \sum_{j=1}^{D} \omega_j^2$$
$$= \arg\max -\frac{1}{N} \sum_{i=1}^{N} [y_i - (\omega_0 + \mathbf{w}^T x_i]^2 + \frac{\sigma^2}{\tau^2} \sum_{j=1}^{D} \omega_j^2$$

Since $\sum_{j=1}^{D} \omega_j^2$ is the sum of squares, we can write it as the L2 norm of $\mathbf{w}$, we can also negate the function and take the arg min, since $\arg\max f(x) = \arg\min -f(x)$. Therefore, we can rewrite the function as

$$\arg\max \ell(\mathbf{w}) = \arg\min \frac{1}{N} \sum_{i=1}^{N} [y_i - (\omega_0 + \mathbf{w}^T x_i]^2 + \lambda ||\mathbf{w}||_2^2$$

Where $\lambda = \frac{\sigma^2}{\tau^2}$

(b) Let $f = ||A\mathbf{x} - \mathbf{b}||_2^2 + ||\Gamma\mathbf{x}||_2^2$ Expand f

$$\begin{aligned} f =& ||A\mathbf{x} - \mathbf{b}||_2^2 + ||\Gamma\mathbf{x}||_2^2 \\ =& (A\mathbf{x} - \mathbf{b})^T(A\mathbf{x} - \mathbf{b}) + (\Gamma\mathbf{x})^T(\Gamma\mathbf{x}) \\ =& \left[(A\mathbf{x} - \mathbf{b})^T(A\mathbf{x} - \mathbf{b}) + (\Gamma\mathbf{x})^T(\Gamma\mathbf{x})\right] \\ =& \left[(\mathbf{x}^T A^T - \mathbf{b}^T)(A\mathbf{x} - \mathbf{b}) + \mathbf{x}^T\Gamma^T\Gamma\mathbf{x}\right] \end{aligned}$$

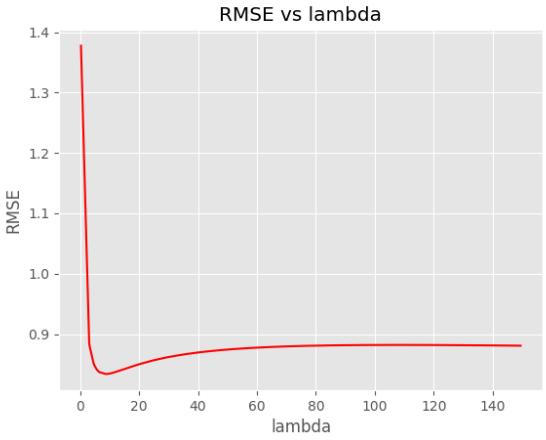
Take th gradient

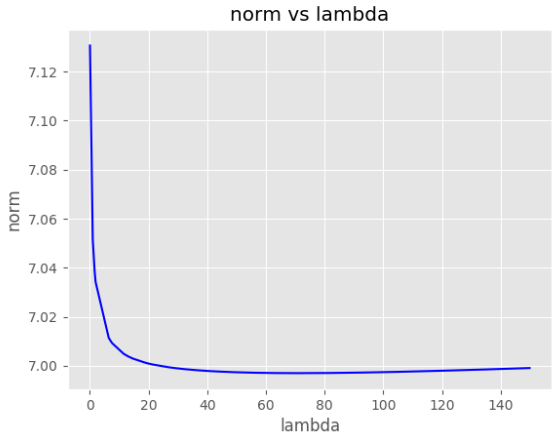
$$\begin{aligned} \nabla_{\mathbf{x}} f =& \nabla_{\mathbf{x}} \left[(\mathbf{x}^T A^T - \mathbf{b}^T)(A\mathbf{x} - \mathbf{b}) + \mathbf{x}^T\Gamma^T\Gamma\mathbf{x}\right] \\ =& 2A^T A\mathbf{x} - 2A^T\mathbf{b} - 2\Gamma^T\Gamma\mathbf{x} \end{aligned}$$

Set the gradient to 0 and solve for $\mathbf{x}*$, which gives the closed form solution. Let $\lambda = \sqrt{\lambda}\boldsymbol{I}$

$$\begin{aligned} 0 =& 2A^T A\mathbf{x} * -2A^T\mathbf{b} - 2\Gamma^T\Gamma\mathbf{x}* \\ \mathbf{x}* =& (A^T A + \lambda\boldsymbol{I})^{-1}A^T\mathbf{b} \end{aligned}$$

(c) The optimal regularization parameter $\lambda*$ is 8.0475. The RMSE on the validation set with the optimal regularization parameter is 0.8342. The RMSE on the test set with the optimal regularization parameter is 0.8628. The plots are shown below





(d) First expand f

$$\begin{aligned} f =& (A\mathbf{x} + b\mathbf{1} - \mathbf{y})^T (A\mathbf{x} + b\mathbf{1} - \mathbf{y}) + (\Gamma\mathbf{x})^T (\Gamma\mathbf{x}) \\ =& (\mathbf{x}^T A^T + b\mathbf{1}^T - \mathbf{y}^T)(A\mathbf{x} + b\mathbf{1}\mathbf{y}) + \mathbf{x}^T \Gamma^T \Gamma\mathbf{x} \\ =& \mathbf{x}^T A^T A\mathbf{x} + 2b\mathbf{1}^T A\mathbf{x} - 2\mathbf{y}^T A\mathbf{x} - 2b\mathbf{1}^T\mathbf{y} + b^2 n + \mathbf{y}^T\mathbf{y} + \mathbf{x}^T\Gamma^T\Gamma\mathbf{x} \end{aligned}$$

Find the gradient of f with respect to $\mathbf{x}$ and b. Optimize by setting the gradients to zero.

$$\nabla_{\mathbf{x}} f = 2A^T A\mathbf{x} + 2bA^T\mathbf{1} - 2A^T\mathbf{y} + 2\Gamma^T\Gamma\mathbf{x} = 0 \tag{1}$$

$$\nabla_b f = 2\mathbf{1}^T A\mathbf{x} 2\mathbf{1}^T\mathbf{y} + 2bn = 0 \tag{2}$$

Now solving for the optimal b, $b*$ gives us

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

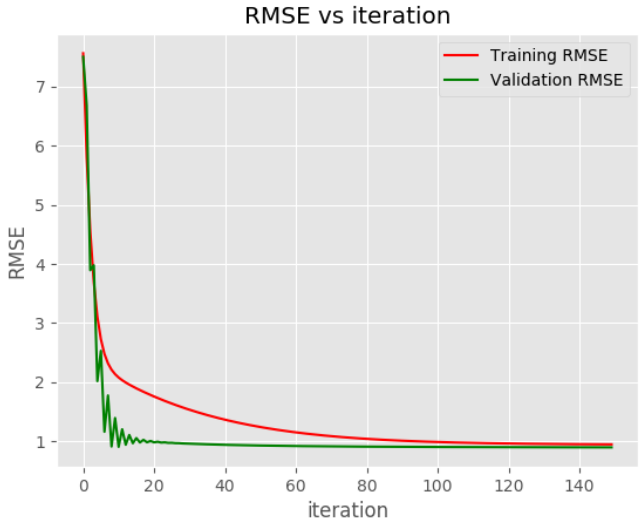
Let $\boldsymbol{I}$ be the identity matrix, and $\mathbf{1}$ be the vector of all ones.

Plug the $b*$ back in to equation (1) to solve for $\mathbf{x}*$, we find

$$\begin{aligned} 0 =& (A^T A + \Gamma^T\Gamma)\mathbf{x} + \left(\frac{\mathbf{1}^T(\mathbf{y} - A\mathbf{x})}{n}\right) A^T\mathbf{1} - A^T\mathbf{y} \\ 0 =& (A^T A + \Gamma^T\Gamma)\mathbf{x} + 1A^T\mathbf{1}\mathbf{1}^T\mathbf{y} - \mathbf{1}A^T\mathbf{1}\mathbf{1}^T A\mathbf{x}A^T\mathbf{y} \\ \left[A^T A + \Gamma^T\Gamma\frac{1}{n}A^T\mathbf{1}\mathbf{1}^T A\right]\mathbf{x} =& A^T\mathbf{y} - \frac{1}{n}A^T\mathbf{1}\mathbf{1}^T\mathbf{y} \\ \left[A\left(\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)A + \Gamma^T\Gamma\right]\mathbf{x} =& A^T\left(\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)\mathbf{y} \\ x* =& \left[A^T\left(\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)A + \Gamma^T\Gamma\right]^{-1} A^T\left(\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)\mathbf{y} \end{aligned}$$

We computed the bias term using given starter code, and compared it to our answer from part (c). We see that the differences are negligibly small, so the results are essentially the same:
Difference in bias is 4.2654E-10
Difference in weights is 5.6292E-10

(e) Comparing the results to bias, and weights obtained from part (c) and (d) we get:
Difference in bias is 1.5387E-01
Difference in weights is 7.9879E-01
The convergence plot is shown below.



I looked at the solution key while doing problem 3. ■