Math189R SP19
Homework 1
Monday, Feb 4, 2017

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 code for problem 2 part c and d can be found under the Resource tab on course website.

Note: You need to create a Github account for submission of the coding part of the homework. Please create a repository on Github to hold all your code and include your Github account username as part of the answer to problem 2.

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) Given $\sigma(x) = \frac{1}{1+e^{-x}}$:

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

thus we find $\sigma(x) = \frac{1}{1+e^{-x}}$ as desire.

(b) Given that the equation for negative log likelihood is:

$$\begin{aligned}\ell(\theta) &= -log(L(\theta)) \\ &= -\sum_{i=1}^{m} y_i log(h(x_i)) + (1-y_i)log(1-h(x_i))\end{aligned}$$

Note that $h(x_i) = \sigma(\theta^T x_i)$ in this function, thus:

$$\sigma'(\theta^T x_i) = \sigma(\theta^T x_i)(1-\sigma(\theta^T x_i))$$

We can then take the gradient of the log likelihood equation:

$$\begin{aligned}\nabla_\theta \ell(\theta) &= -\nabla_\theta \sum_{i=1}^{m} y_i log(\sigma(\theta^T x_i)) + \nabla_\theta (1-y_i) log(1-\sigma(\theta^T x_i)) \\ &= -\sum_{i=1}^{m} y_i \frac{1}{\sigma(\theta^T x_i)}(\sigma'(\theta^T x_i)) + (1-y_i)\frac{1}{\sigma(1-\theta^T x_i)}(-\sigma'(\theta^T x_i)) \\ &= -\sum_{i=1}^{m} y_i \frac{\sigma(\theta^T x_i)(1-\sigma(\theta^T x_i))}{\sigma(\theta^T x_i)} x_i + (1-y_i)\frac{-\sigma(\theta^T x_i)(1-\sigma(\theta^T x_i))}{1-\sigma(\theta^T x_i)} x_i \\ &= -\sum_{i=1}^{m} y_i (1-\sigma(\theta^T x_i)) x_i + (1-y_i)(-\sigma(\theta^T x_i)) x_i \\ &= -\sum_{i=1}^{m} y_i x_i - y_i \sigma(\theta^T x_i) x_i - \sigma(\theta^T x_i) x_i + y_i \sigma(\theta^T x_i) x_i \\ &= -\sum_{i=1}^{m} y_i x_i - \sigma(\theta^T x_i) x_i \\ &= \sum_{i=1}^{m} (\sigma(\theta^T x_i) - y_i) x_i\end{aligned}$$

Substituted in $\mu_i = \sigma(\theta^T x_i)$:

$$\begin{aligned}\nabla_\theta \ell(\theta) &= \sum_{i=1}^{m} (\mu_i - y_i) x_i \\ &= (\mu - y) X^T\end{aligned}$$

(c) We want to prove that the Hessian Matrix is positive semi-definite, or in other words, the matrix's eigenvalues are non negative. Using equation from previous

parts, find Hessian Matrix:

$$\begin{aligned}
H_\theta &= \nabla_\theta(\nabla_\theta \ell(\theta))^T \\
&= \nabla_\theta(X^T(\mu - y))^T \\
&= \nabla_\theta(X^T\mu - X^T y)^T \\
&= \nabla_\theta(\mu^T X - y^T X) \\
&= \nabla_\theta \mu^T X - \nabla_\theta y^T X \\
&= \nabla_\theta \mu^T X \\
&= \nabla_\theta(\sigma(\theta^T x_i))^T X \\
&= X^T diag(\sigma(\theta^T x_i)(1 - \sigma(\theta^T x_i)))X
\end{aligned}$$

Substituting the equation, $S = diag(\mu_1(1-\mu_1), ..., \mu_n(1-\mu_n))$, we find the following:

$$H = X^T S X$$

Next, given the previously found equation for the hessian matrix, the resulting matrix is semi-definite if and only if the diagonal is greater than zero. Since $0 \leq \sigma(f) \leq 1$ for arbitrary input f, we know that $diag(\sigma(\theta^T x_i)(1 - \sigma(\theta^T x_i))) \geq 0$, therefore the Hessian Matrix is proven to be semi-definite, as desired.

■

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.

The integral of probability integral is 1, therefor we can find the normalization constant by:

$$\begin{aligned}
1 &= \int_{\text{Re}} \mathbb{P}(x;\sigma^2)dx \\
1 &= \int_{\text{Re}} \frac{1}{Z}\exp\left(-\frac{x^2}{2\sigma^2}\right)dx \\
Z &= \int_{\text{Re}} \exp\left(-\frac{x^2}{2\sigma^2}\right)dx
\end{aligned}$$

To compute this constant we will consider taking the square:

$$\begin{aligned}
Z^2 &= \int_{\text{Re}} \exp\left(-\frac{x^2}{2\sigma^2}\right)dx \int_{\text{Re}} \exp\left(-\frac{y^2}{2\sigma^2}\right)dy \\
&= \int_{\text{Re}}\int_{\text{Re}} \exp\left(-\frac{x^2+y^2}{2\sigma^2}\right)dxdy
\end{aligned}$$

We can change the equation from cartesian to polar and get,
note that $\frac{d}{dr}\exp\left(-\frac{r^2}{2\sigma^2}\right) = -\frac{r}{\sigma^2}\exp\left(-\frac{r^2}{2\sigma^2}\right)$:

$$\begin{aligned}
Z^2 &= \int_0^{2\pi}\int_0^{\infty} r\exp\left(-\frac{r^2}{2\sigma^2}\right)d\theta dr \\
&= 2\pi\int_0^{\infty} r\exp\left(-\frac{r^2}{2\sigma^2}\right)dr \\
&= -2\pi\sigma^2\int_0^{\infty} -\frac{r}{\sigma^2}\exp\left(-\frac{r^2}{2\sigma^2}\right)dr \\
&= -2\pi\sigma^2\left[\exp\left(-\frac{r^2}{2\sigma^2}\right)\right]_0^{\infty} \\
&= 2\pi\sigma^2
\end{aligned}$$

Thus we can find that:

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

■

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) start with the given equation:

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

since we know that the probability is described by the Gaussian equation, the equation can be rewritten as follows:

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

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

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

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

by taking advantage of the properties of argmax, we can ignore all constants and

scalars in the equation and negate and negatives by changing argmax to argmin:

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

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

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

now we can substitute in $\lambda = \frac{\sigma^2}{\tau^2}$ and rewrite the equation as:

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

therefore finding the ridge regression problem as desired

(b) the close form solution $\mathbf{x}^\star$ is defined as where the gradient of $f = ||A\mathbf{x} - \mathbf{b}||_2^2 + ||\Gamma\mathbf{x}||_2^2$ is 0, so we can manipulate the equation as follows:

$$\begin{aligned}
0 &= \nabla_x f \\
0 &= \nabla_x ||A\mathbf{x} - \mathbf{b}||_2^2 + ||\Gamma\mathbf{x}||_2^2 \\
0 &= \nabla_x \left[(A\mathbf{x} - \mathbf{b})^\top (A\mathbf{x} - \mathbf{b}) + (\Gamma\mathbf{x})^\top (\Gamma\mathbf{x})\right] \\
0 &= \nabla_x \left[(\mathbf{x}^\top A^\top - \mathbf{b}^\top)(A\mathbf{x} - \mathbf{b}) + (\mathbf{x}^\top \Gamma^\top)(\Gamma\mathbf{x})\right] \\
0 &= \nabla_x \left(\mathbf{x}^\top A^\top A\mathbf{x} - \mathbf{x}^\top A^\top \mathbf{b} - \mathbf{b}^\top A\mathbf{x} + \mathbf{b}^\top \mathbf{b} + \mathbf{x}^\top \Gamma^\top \Gamma\mathbf{x}\right) \\
0 &= \nabla_x \left(\mathbf{x}^\top A^\top A\mathbf{x} - 2\mathbf{x}^\top A^\top \mathbf{b} + \mathbf{b}^\top \mathbf{b} + \mathbf{x}^\top \Gamma^\top \Gamma\mathbf{x}\right) \\
0 &= 2A^\top A\mathbf{x} - 2A^\top \mathbf{b} + 2\Gamma^\top \Gamma\mathbf{x} \\
0 &= A^\top A\mathbf{x} - A^\top \mathbf{b} + \Gamma^\top \Gamma\mathbf{x} \\
A^\top A\mathbf{x} + \Gamma^\top \Gamma\mathbf{x} &= A^\top \mathbf{b} \\
\mathbf{x} &= \frac{A^\top \mathbf{b}}{A^\top A + \Gamma^\top \Gamma}
\end{aligned}$$

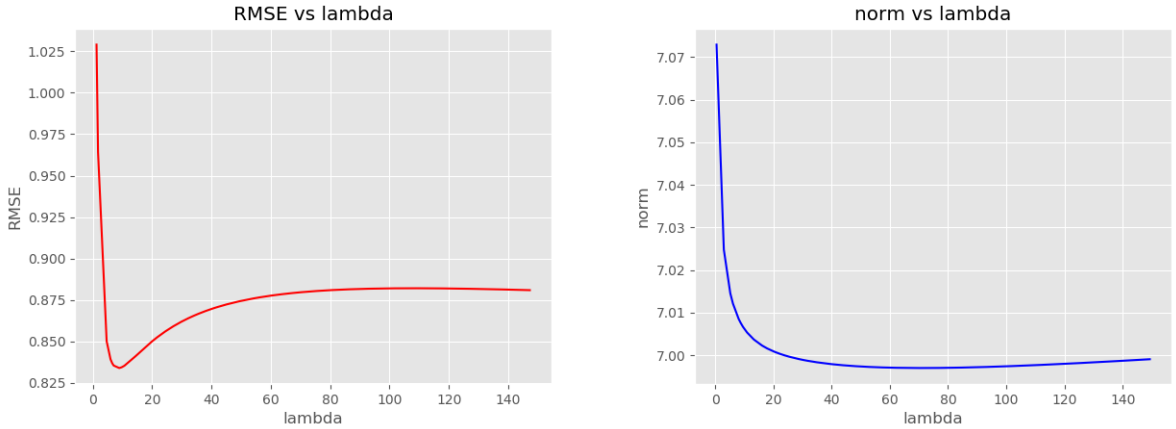
thus we find:

$$\mathbf{x}^* = \frac{A^\top \mathbf{b}}{A^\top A + \Gamma^\top \Gamma}$$

alternatively, if we define $\Gamma = \sqrt{\lambda} I$:

$$\mathbf{x}^* = \frac{A^\top \mathbf{b}}{A^\top A + \sqrt{\lambda} I^\top \sqrt{\lambda} I}$$
$$\mathbf{x}^* = \frac{A^\top \mathbf{b}}{A^\top A + \lambda I}$$

(c) We have the following plots



The optimal λ^* is 8.9284.Given this parameter, we find the associated validation set RMSE to be 0.8341 and the final RMSE on the test set is 0.8628.

(d) to find the gradient of $f = ||A\mathbf{x} + b\mathbf{1} - \mathbf{y}||_2^2 + ||\Gamma\mathbf{x}||_2^2$ is 0, we can manipulate the equation as follows:

$$\begin{aligned}
f &= ||A\mathbf{x} + b\mathbf{1} - \mathbf{y}||_2^2 + ||\Gamma\mathbf{x}||_2^2 \\
&= \left[(A\mathbf{x} + b\mathbf{1} - \mathbf{y})^\top (A\mathbf{x} + b\mathbf{1} - \mathbf{y}) + (\Gamma\mathbf{x})^\top (\Gamma\mathbf{x})\right] \\
&= (\mathbf{x}^\top A^\top + b\mathbf{1}^\top - \mathbf{y}^\top)(A\mathbf{x} + b\mathbf{1} - \mathbf{y}) + (\mathbf{x}^\top \Gamma^\top)(\Gamma\mathbf{x}) \\
&= \mathbf{x}^\top A^\top A\mathbf{x} + b\mathbf{1}^\top A\mathbf{x} - \mathbf{y}^\top A\mathbf{x} + \mathbf{x}^\top A^\top b\mathbf{1} + b\mathbf{1}^\top b\mathbf{1} - \mathbf{y}^\top b\mathbf{1} \\
&\quad - \mathbf{x}^\top A^\top \mathbf{y} - b\mathbf{1}^\top \mathbf{y} + \mathbf{y}^\top \mathbf{y} + \mathbf{x}^\top \Gamma^\top \Gamma\mathbf{x} \\
&= \mathbf{x}^\top A^\top A\mathbf{x} + 2b\mathbf{1}^\top A\mathbf{x} - 2\mathbf{y}^\top A\mathbf{x} + b^2 n - 2\mathbf{y}^\top b\mathbf{1} + \mathbf{y}^\top \mathbf{y} + \mathbf{x}^\top \Gamma^\top \Gamma\mathbf{x} \\
\nabla_x f &= 2A^\top A\mathbf{x} + 2b\mathbf{1}^\top A - 2\mathbf{y}^\top A + 2\Gamma^\top \Gamma\mathbf{x} \\
\nabla_b f &= 2\mathbf{1}^\top A\mathbf{x} + 2bn - 2\mathbf{y}^\top \mathbf{1}
\end{aligned}$$

to find the specific solution $\mathbf{b}^*$ we set the gradient with respect to b equal to 0:

$$\begin{aligned} 0 &= \nabla_b f \mathbf{1} \\ 0 &= 2\mathbf{1}^\top A\mathbf{x} + 2bn - 2\mathbf{y}^\top \mathbf{1} \\ 2bn &= 2\mathbf{y}^\top \mathbf{1} - 2\mathbf{1}^\top A\mathbf{x} \\ \mathbf{b}^* &= \frac{\mathbf{y}^\top \mathbf{1} - \mathbf{1}^\top A\mathbf{x}}{n} \end{aligned}$$

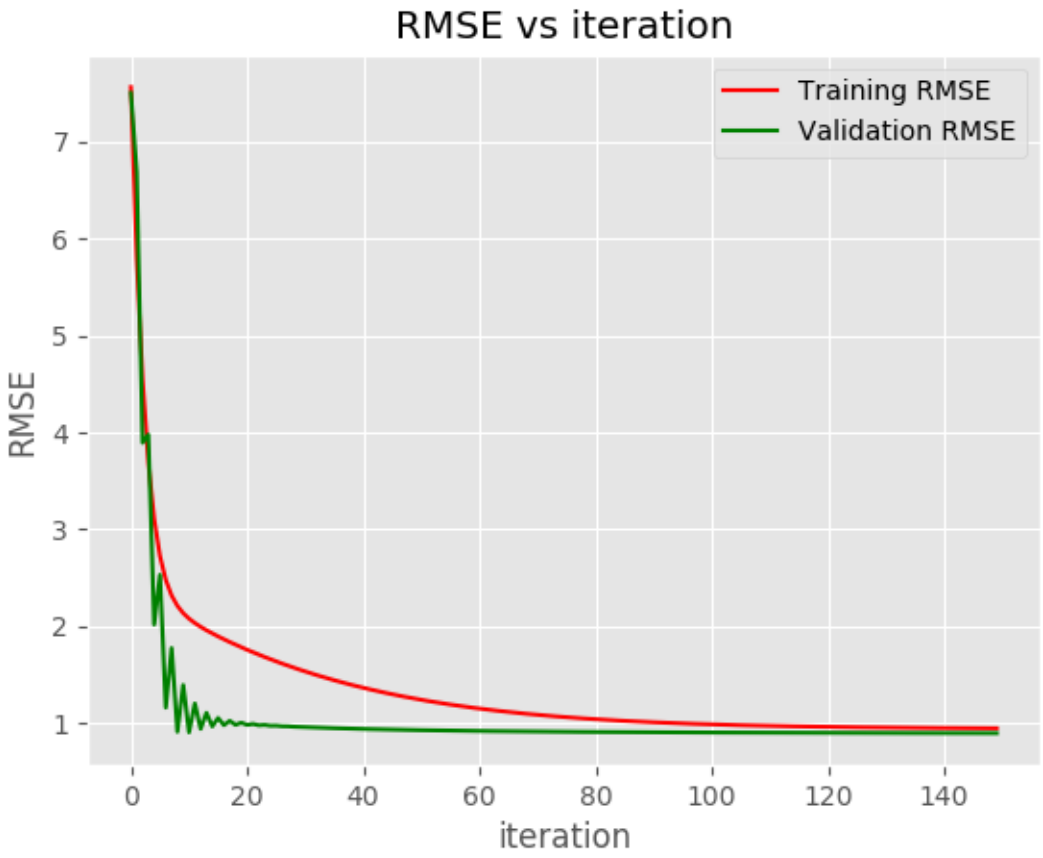
using the specific solution $\mathbf{b}^*$, we can now find $\mathbf{x}^*$ by setting the gradient with respect to x equal to 0:

$$\begin{aligned} 0 &= \nabla_b x \mathbf{1} \\ 0 &= 2A^\top A\mathbf{x} + 2b\mathbf{1}^\top A - 2\mathbf{y}^\top A + 2\Gamma^\top \Gamma\mathbf{x} \\ 0 &= A^\top A\mathbf{x} + \frac{\mathbf{y}^\top \mathbf{1} - \mathbf{1}^\top A\mathbf{x}}{n}\mathbf{1}^\top A - \mathbf{y}^\top A + \Gamma^\top \Gamma\mathbf{x} \\ 0 &= A^\top A\mathbf{x} + \frac{\mathbf{y}^\top \mathbf{1}\mathbf{1}^\top A}{n} - \frac{\mathbf{1}^\top A\mathbf{x}\mathbf{1}^\top A}{n} - \mathbf{y}^\top A + \Gamma^\top \Gamma\mathbf{x} \\ 0 &= A^\top A\mathbf{x} + \frac{\mathbf{y}^\top \mathbf{1}\mathbf{1}^\top A}{n} - \frac{A^\top \mathbf{1}\mathbf{1}^\top A}{n}\mathbf{x} - \mathbf{y}^\top A + \Gamma^\top \Gamma\mathbf{x} \\ \frac{A^\top \mathbf{1}\mathbf{1}^\top A}{n}\mathbf{x} - A^\top A\mathbf{x} - \Gamma^\top \Gamma\mathbf{x} &= \frac{\mathbf{y}^\top \mathbf{1}\mathbf{1}^\top A}{n} - \mathbf{y}^\top A \\ \left[\frac{A^\top \mathbf{1}\mathbf{1}^\top A}{n} - A^\top A - \Gamma^\top \Gamma\right]\mathbf{x} &= \frac{\mathbf{y}^\top \mathbf{1}\mathbf{1}^\top A}{n} - \mathbf{y}^\top A \\ \mathbf{x}^* &= \left[\frac{A^\top \mathbf{1}\mathbf{1}^\top A}{n} - A^\top A - \Gamma^\top \Gamma\right]^{-1}\left[\frac{\mathbf{y}^\top \mathbf{1}\mathbf{1}^\top A}{n} - \mathbf{y}^\top A\right] \end{aligned}$$

this can be rearrange to match the solution (for grading purposes):

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

as a result: Difference in bias is 4.8540E-10 Difference in weights is 6.5626E-10



(e)

Difference in bias is 1.5389E-01
Difference in weights is 8.1438E-01

■