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 $\mathbf{S} = \mathrm{diag}(\mu_1(1-\mu_1), \dots, \mu_n(1-\mu_n))$. 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) $\sigma(x) = \frac{1}{1+e^{-x}}$, so:

$$\sigma'(x) = \frac{0-(1)(-e^{-x})}{(1+e^{-x})^2} = \frac{(1+e^{-x})-1}{(1+e^{-x})^2} = \frac{1}{1+e^{-x}}\left(\frac{1+e^{-x}}{1+e^{-x}} - \frac{1}{1+e^{-x}}\right) = \sigma(x)\left[1-\sigma(x)\right]$$

(b) The log likelihood function is:

$$l(\theta) = log(L(\theta)) = \sum_{i=1}^{n}[y^{(i)}\log(h_\theta(x^{(i)})) + (1-y^{(i)})\log(1-h_\theta(x^{(i)}))]$$

$$= \sum_{i=1}^{n}[y^{(i)}\log(\sigma(\theta^T x^{(i)})) + (1-y^{(i)})\log(1-\sigma(\theta^T x^{(i)}))]$$

The gradient is:

$$\nabla l(\theta) = \left(\frac{\partial l}{\partial \theta_1}, \frac{\partial l}{\partial \theta_2}, ..., \frac{\partial l}{\partial \theta_n}\right)$$

And each one of these derivatives is given by:

$$\frac{\partial l}{\partial \theta_j} = \sum_{i=1}^{n}[y^{(i)}\frac{1}{\sigma(\theta^T x^{(i)})}(\sigma'(\theta^T x^{(i)}))(x_j^{(i)}) + (1-y^{(i)})\frac{1}{1-\sigma(\theta^T x^{(i)})}(-\sigma'(\theta^T x^{(i)}))(x_j^{(i)})]$$

$$= \sum_{i=1}^{n} [y^{(i)} \frac{\sigma(\theta^T x^{(i)})}{\sigma(\theta^T x^{(i)})} (1 - \sigma(\theta^T x^{(i)}))(x_j^{(i)}) - (1 - y^{(i)}) \frac{\sigma(\theta^T x^{(i)})}{1 - \sigma(\theta^T x^{(i)})} (1 - \sigma(\theta^T x^{(i)}))(x_j^{(i)})]$$

$$= \sum_{i=1}^{n} [y^{(i)} (1 - \sigma(\theta^T x^{(i)}))(x_j^{(i)}) - (1 - y^{(i)})(\sigma(\theta^T x^{(i)}))(x_j^{(i)})]$$

$$= \sum_{i=1}^{n} [(y^{(i)} - y^{(i)}(\sigma(\theta^T x^{(i)})) + y^{(i)}(\sigma(\theta^T x^{(i)})) - \sigma(\theta^T x^{(i)}))(x_j^{(i)})]$$

$$= \sum_{i=1}^{n} [(y^{(i)} - \sigma(\theta^T x^{(i)}))(x_j^{(i)})]$$

When we put these elements back in the gradient we get that $\nabla l(\theta) =$

$$\langle \sum_{i=1}^{n} [(y^{(i)} - \sigma(\theta^T x^{(i)}))(x_1^{(i)})], \sum_{i=1}^{n} [(y^{(i)} - \sigma(\theta^T x^{(i)}))(x_2^{(i)})], ..., \sum_{i=1}^{n} [(y^{(i)} - \sigma(\theta^T x^{(i)}))(x_n^{(i)})] \rangle$$

(c) We know the jth element in our $\nabla l(\theta)$ is $\sum_{i=1}^{n} [(y^{(i)} - y^{(i)}(\sigma(\theta^T x^{(i)})) + y^{(i)}(\sigma(\theta^T x^{(i)})) - \sigma(\theta^T x^{(i)}))(x_j^{(i)})]$. From this we differentiate with respect to k to get the element in the jth row and kth column of the Hessian.

$$H_{j,k} = \sum_{i=1}^{n} [0 - (\sigma'(\theta^T x^{(i)}) x_k^{(i)}] x_j^{(i)} = \sum_{i=1}^{n} -(\sigma(\theta^T x^{(i)})[1 - \sigma(\theta^T x^{(i)})]) x_k^{(i)} x_j^{(i)}$$

We will use the negative log likelihood of logistic regression here which eliminates the negative, and we will now let $\mu_i = \sigma(\theta^T x^{(i)})$. This allows us to simplify the above expression to:

$$\sum_{i=1}^{n} (\mu_i [1 - \mu_i]) x_k^{(i)} x_j^{(i)}$$

Now consider the matrices **X** and **S**

$$\mathbf{X} = \begin{bmatrix} x_1^{(1)} & x_2^{(1)} & ... & x_n^{(1)} \\ x_1^{(2)} & x_2^{(2)} & ... & x_n^{(2)} \\ \vdots & \vdots & \vdots & \vdots \\ x_1^{(n)} & x_2^{(n)} & ... & x_n^{(n)} \end{bmatrix} \mathbf{S} = \begin{bmatrix} \mu_1[1 - \mu_1] & 0 & ... & 0 \\ 0 & \mu_2[1 - \mu_2] & ... & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & ... & \mu_n[1 - \mu_n] \end{bmatrix}$$

We see that $\mathbf{X}^T \mathbf{S} \mathbf{X}$ is the following:

$$\begin{bmatrix} x_1^{(1)} & x_1^{(2)} & ... & x_1^{(n)} \\ x_2^{(1)} & x_2^{(2)} & ... & x_2^{(n)} \\ \vdots & \vdots & \vdots & \vdots \\ x_n^{(1)} & x_n^{(2)} & ... & x_n^{(n)} \end{bmatrix} \begin{bmatrix} \mu_1[1 - \mu_1] & 0 & ... & 0 \\ 0 & \mu_2[1 - \mu_2] & ... & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & ... & \mu_n[1 - \mu_n] \end{bmatrix} \begin{bmatrix} x_1^{(1)} & x_2^{(1)} & ... & x_n^{(1)} \\ x_1^{(2)} & x_2^{(2)} & ... & x_n^{(2)} \\ \vdots & \vdots & \vdots & \vdots \\ x_1^{(n)} & x_2^{(n)} & ... & x_n^{(n)} \end{bmatrix}$$

$$= \begin{bmatrix} x_1^{(1)}\mu_1[1-\mu_1] & x_1^{(2)}\mu_2[1-\mu_2] & ... & x_1^{(n)}\mu_n[1-\mu_n] \\ x_2^{(1)}\mu_1[1-\mu_1] & x_2^{(2)}\mu_2[1-\mu_2] & ... & x_2^{(n)}\mu_n[1-\mu_n] \\ \vdots & \vdots & \vdots & \vdots \\ x_n^{(1)}\mu_1[1-\mu_1] & x_n^{(2)}\mu_2[1-\mu_2] & ... & x_n^{(n)}\mu_n[1-\mu_n] \end{bmatrix} \begin{bmatrix} x_1^{(1)} & x_2^{(1)} & ... & x_n^{(1)} \\ x_1^{(2)} & x_2^{(2)} & ... & x_n^{(2)} \\ \vdots & \vdots & \vdots & \vdots \\ x_1^{(n)} & x_2^{(n)} & ... & x_n^{(n)} \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{i=1}^n(\mu_i[1-\mu_i])x_1^{(i)}x_1^{(i)} & \sum_{i=1}^n(\mu_i[1-\mu_i])x_1^{(i)}x_2^{(i)} & ... & \sum_{i=1}^n(\mu_i[1-\mu_i])x_1^{(i)}x_n^{(i)} \\ \sum_{i=1}^n(\mu_i[1-\mu_i])x_2^{(i)}x_1^{(i)} & \sum_{i=1}^n(\mu_i[1-\mu_i])x_2^{(i)}x_2^{(i)} & ... & \sum_{i=1}^n(\mu_i[1-\mu_i])x_2^{(i)}x_n^{(i)} \\ \vdots & \vdots & \vdots & \vdots \\ \sum_{i=1}^n(\mu_i[1-\mu_i])x_n^{(i)}x_1^{(i)} & \sum_{i=1}^n(\mu_i[1-\mu_i])x_n^{(i)}x_2^{(i)} & ... & \sum_{i=1}^n(\mu_i[1-\mu_i])x_n^{(i)}x_n^{(i)} \end{bmatrix}$$

As we can see this is the Hessian, so the Hessian can be written as $\mathbf{H} = \mathbf{X}^\top \mathbf{S}\mathbf{X}$.

$\mathbf{H}$ is positive semidefinite if and only if $u^T\mathbf{H}u \geq 0$ for all real valued $n \times 1$ vectors u. Consider:

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

$$\begin{bmatrix} u_1 & ... & u_n \end{bmatrix} \begin{bmatrix} \sum_{i=1}^n(\mu_i[1-\mu_i])x_1^{(i)}x_1^{(i)} & ... & \sum_{i=1}^n(\mu_i[1-\mu_i])x_1^{(i)}x_n^{(i)} \\ \vdots & \vdots & \vdots \\ \sum_{i=1}^n(\mu_i[1-\mu_i])x_n^{(i)}x_1^{(i)} & ... & \sum_{i=1}^n(\mu_i[1-\mu_i])x_n^{(i)}x_n^{(i)} \end{bmatrix} \begin{bmatrix} u_1 \\ \vdots \\ u_n \end{bmatrix}$$

$$= \begin{bmatrix} u_1 & ... & u_n \end{bmatrix} \begin{bmatrix} \sum_{j=1}^n u_j \sum_{i=1}^n(\mu_i[1-\mu_i])x_j^{(i)}x_1^{(i)} \\ \vdots \\ \sum_{j=1}^n u_j \sum_{i=1}^n(\mu_i[1-\mu_i])x_j^{(i)}x_n^{(i)} \end{bmatrix}$$

$$= \sum_{k=1}^n u_k \sum_{j=1}^n u_j \sum_{i=1}^n(\mu_i[1-\mu_i])x_j^{(i)}x_k^{(i)}$$

■

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 normalization constant Z will be the value such that.

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

To calculate this integral we'll square both sides:

$$(\int_{-\infty}^{\infty}\frac{1}{Z}\exp\left(-\frac{x^2}{2\sigma^2}\right)dx)^2 = 1^2$$

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

Since y is a constant with respect to x and vice versa we can merge these two integrals like so:

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

$$\frac{1}{Z^2}(\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\exp\left(-\frac{y^2}{2\sigma^2}\right)\exp\left(-\frac{x^2}{2\sigma^2}\right)dydx) = 1$$

$$\frac{1}{Z^2}(\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\exp\left(-\frac{y^2+x^2}{2\sigma^2}\right)dydx) = 1$$

Now we can convert this integral to polar with $x^2 + y^2 = r^2$:

$$\frac{1}{Z^2}(\int_{0}^{\infty}\int_{0}^{2\pi}(\exp\left(-\frac{r^2}{2\sigma^2}\right)r)d\theta dr) = 1$$

$$\frac{2\pi}{Z^2}(\int_{0}^{\infty}(\exp\left(-\frac{r^2}{2\sigma^2}\right)r)dr) = 1 \qquad u = -\frac{r^2}{2\sigma^2},\ du = -\frac{r}{\sigma^2}\,dr$$

$$-\frac{2\pi\sigma^2}{Z^2}(\int_{0}^{-\infty}\exp u)dr) = 1$$

$$-\frac{2\pi\sigma^2}{Z^2}(\exp(-\infty) - \exp(0)) = 1$$

$$\frac{2\pi\sigma^2}{Z^2} = 1$$

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

■

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)

(a) We start with the given equation:

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

Plugging into the zero mean Gaussian equation we get:

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

We can turn this into a min by removing the negative, and we can also remove the constant factor $\log(1/\sqrt{2\pi}\sigma)$ or $\log(1/\sqrt{2\pi}\tau)$:

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

We substitute in $\sum_{j=1}^{D} w_j^2 = ||\mathbf{w}||_2^2$:

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

We can multiply by a constant factor:

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

And substitute in λ to get our final result:

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

(b) We are trying to find a closed form solution to

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

Letting $\mathbf{a_i}$ be the ith row vector of the matrix A this is:

$$\text{minimize: } (A\mathbf{x} - \mathbf{b})^T (A\mathbf{x} - \mathbf{b}) + (\Gamma\mathbf{x})^T (\Gamma\mathbf{x})$$

To do this we find the critical points by finding the gradient and set to 0:

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

Distributing our transpose we get:

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

Distributing the matrix multiplication:

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

Using gradient rules and matrix rules:

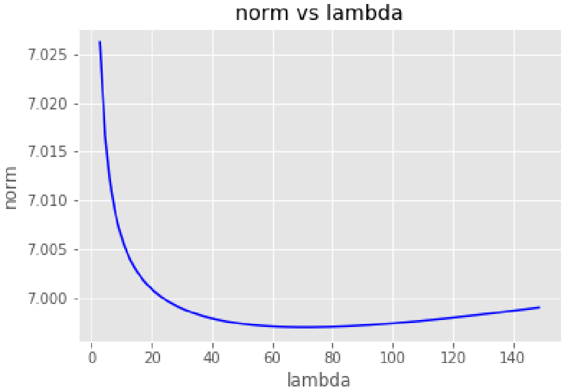
$$2A^TA\mathbf{x} - A^T\mathbf{b} - A^T\mathbf{b} + 2\Gamma^T\Gamma\mathbf{x} = 0$$

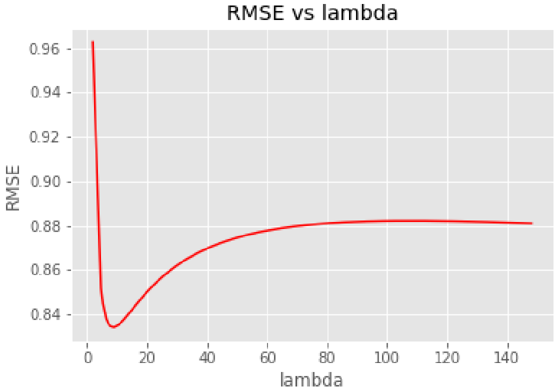
$$A^TA\mathbf{x} - A^T\mathbf{b} + \Gamma^T\Gamma\mathbf{x} = 0$$
$$A^TA\mathbf{x} + \Gamma^T\Gamma\mathbf{x} = A^T\mathbf{b}$$
$$(A^TA + \Gamma^T\Gamma)\mathbf{x} = A^T\mathbf{b}$$
$$\mathbf{x} = (A^TA + \Gamma^T\Gamma)^{-1}A^T\mathbf{b}$$

Let $\Gamma = \sqrt{\lambda}\mathbf{I}$, so $\Gamma^T\Gamma = \lambda\mathbf{I}$, so:

$$\mathbf{x} = (A^TA + \lambda\mathbf{I})^{-1}A^T\mathbf{b}$$

(c) After running the code, I have found that the optimal λ from the validation set was 8.8960, and the RMSE on the test set with the optimal regularization parameter is 0.8628.





■

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.

(d) We are trying to find a closed form solution to

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

Letting $\mathbf{a_i}$ be the ith row vector of the matrix A this is:

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

To do this we find the critical points by finding the gradient with respect to $\mathbf{x}$ and b and set to 0, but first lets expand:

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

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

Taking the gradient of each term with respect to b this will be 0:

$$\mathbf{1}^TA\mathbf{x} + \mathbf{x}^TA^T\mathbf{1} + 2\mathbf{1}^Tb\mathbf{1} - \mathbf{y}^T\mathbf{1} - \mathbf{1}^T\mathbf{y} = 0$$

$$2\mathbf{1}^Tb\mathbf{1} = \mathbf{y}^T\mathbf{1} + \mathbf{1}^T\mathbf{y} - \mathbf{1}^TA\mathbf{x} - \mathbf{x}^TA^T\mathbf{1}$$

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

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

Taking the gradient of each term with respect to **x** this will be 0:

$$2A^TA\mathbf{x} + A^Tb\mathbf{1} + A^Tb\mathbf{1} - A^T\mathbf{y} - A^T\mathbf{y} + 2\Gamma^T\Gamma\mathbf{x} = 0$$

$$A^TA\mathbf{x} + A^Tb\mathbf{1} - A^T\mathbf{y} + \Gamma^T\Gamma\mathbf{x} = 0$$

Now we plug in the optimal b we found:

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

$$nA^TA\mathbf{x} + \mathbf{y}^T\mathbf{1}A^T\mathbf{1} - \mathbf{x}^TA^T\mathbf{1}A^T\mathbf{1} - nA^T\mathbf{y} + n\Gamma^T\Gamma\mathbf{x} = 0$$

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

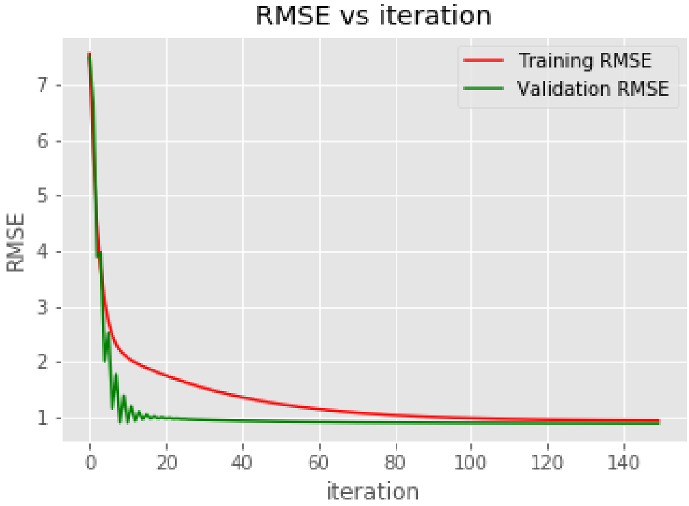
$$(nA^TA - \mathbf{1}^TA\mathbf{1}^TA + n\Gamma^T\Gamma)\mathbf{x} = nA^T\mathbf{y} - \mathbf{y}^T\mathbf{1}A^T\mathbf{1}$$

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

Important note: I had to copy most of the solution code for both (d) and (e) coding parts. After I attempted on my own for quite some time, I was out of time and using the solution code to figure out what was wrong with mine, because even though I think my solution to the math part of (d) is equivalent to the solution given, they are too different for me to use the code to fix my code.

The code told me that:
Difference in bias is 4.3279E-10
Difference in weights is 5.7041E-10

(d) The plot I obtained was the following:



Difference in bias is 1.5387E-01
Difference in weights is 7.9950E-01

■