

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

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

(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  $\mathbf{S} = \text{diag}(\mu_1(1 - \mu_1), \dots, \mu_n(1 - \mu_n))$ . 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)**

Taking the derivative of  $\sigma(x)$ , we have:

$$\begin{aligned}\sigma'(x) &= \frac{d}{dx} \left( \frac{1}{1 + e^{-x}} \right) \quad (\text{definition}) \\ &= \frac{d}{du} \left( \frac{1}{u} \right) \left( \frac{du}{dx} \right), \text{ where } u(x) = 1 + e^{-x} \quad (\text{Chain rule}) \\ &= \left( \frac{-1}{u^2} \right) (-e^{-x}), \text{ where } u(x) = 1 + e^{-x} \\ &= \left( \frac{-1}{(1 + e^{-x})^2} \right) (-e^{-x}) \\ &= \frac{-e^{-x}}{(1 + e^{-x})^2}.\end{aligned}$$

Now for the right side of the equation:

$$\begin{aligned}
 \sigma(x)[1 - \sigma(x)] &= \frac{1}{1 + e^{-x}} \left[ 1 - \frac{1}{1 + e^{-x}} \right] \\
 &= \frac{1}{1 + e^{-x}} - \frac{1}{(1 + e^{-x})^2} \\
 &= \frac{(1 + e^{-x})^2 - (1 + e^{-x})}{(1 + e^{-x})^3} \\
 &= \frac{(1 + e^{-x}) - 1}{(1 + e^{-x})^2} \\
 &= \frac{e^{-x}}{(1 + e^{-x})^2}.
 \end{aligned}$$

We see that  $\sigma'(x) = \frac{-e^{-x}}{(1+e^{-x})^2} = \sigma(x)[1 - \sigma(x)]$ . ■

**(b)**

By Equation 8.2, the negative log-likelihood for logistic regression is given by:

$$NLL(\mathbf{w}) = - \sum_{i=1}^N [y_i \log u_i + (1 - y_i) \log(1 - u_i)], \text{ where } u_i(\mathbf{x}) = \text{sigm}(\mathbf{w}^T \mathbf{x}) = \sigma(\mathbf{w}^T \mathbf{x}).$$

The gradient of the negative log-likelihood for logistic regression is given by:

$$\begin{aligned}
\mathbf{g} &= \frac{d}{d\mathbf{w}}(NLL(\mathbf{w})) \\
&= \nabla_{\mathbf{w}}(NLL(\mathbf{w})) \\
&= -\nabla_{\mathbf{w}}\left(\sum_{i=1}^N [y_i \log \sigma(\mathbf{w}^T \mathbf{x}_i) + (1 - y_i) \log(1 - \sigma(\mathbf{w}^T \mathbf{x}_i))]\right) \\
&= -\nabla_{\mathbf{w}}\left[\sum_{i=1}^N y_i \log \sigma(\mathbf{w}^T \mathbf{x}_i) + \sum_{i=1}^N (1 - y_i) \log(1 - \sigma(\mathbf{w}^T \mathbf{x}_i))\right] \\
&= -\sum_{i=1}^N y_i \frac{d\sigma}{dx} \frac{d}{d\sigma}(\log \sigma) - \sum_{i=1}^N (1 - y_i) \frac{d\sigma}{dx} \frac{d}{d\sigma}(\log(1 - \sigma)) \\
&= -\sum_{i=1}^N y_i \sigma'(\mathbf{w}^T \mathbf{x}_i) \mathbf{x}_i \left(\frac{1}{\sigma(\mathbf{w}^T \mathbf{x}_i)}\right) - \sum_{i=1}^N (1 - y_i) \sigma'(\mathbf{w}^T \mathbf{x}_i) \mathbf{x}_i \frac{-1}{1 - \sigma(\mathbf{w}^T \mathbf{x}_i)} \\
&= -\sum_{i=1}^N y_i \sigma(\mathbf{w}^T \mathbf{x}_i) [1 - \sigma(\mathbf{w}^T \mathbf{x}_i)] \mathbf{x}_i \left(\frac{1}{\sigma(\mathbf{w}^T \mathbf{x}_i)}\right) - \sum_{i=1}^N (1 - y_i) \sigma(\mathbf{w}^T \mathbf{x}_i) [1 - \sigma(\mathbf{w}^T \mathbf{x}_i)] \mathbf{x}_i \frac{-1}{1 - \sigma(\mathbf{w}^T \mathbf{x}_i)} \\
&= -\sum_{i=1}^N y_i [1 - \sigma(\mathbf{w}^T \mathbf{x}_i)] \mathbf{x}_i + \sum_{i=1}^N (1 - y_i) \sigma(\mathbf{w}^T \mathbf{x}_i) \mathbf{x}_i \\
&= -\sum_{i=1}^N [y_i (1 - \sigma(\mathbf{w}^T \mathbf{x}_i)) \mathbf{x}_i + (1 - y_i) \sigma(\mathbf{w}^T \mathbf{x}_i) \mathbf{x}_i] \\
&= -\sum_{i=1}^N [y_i \mathbf{x}_i - y_i \sigma(\mathbf{w}^T \mathbf{x}_i) \mathbf{x}_i + \sigma(\mathbf{w}^T \mathbf{x}_i) \mathbf{x}_i - y_i \sigma(\mathbf{w}^T \mathbf{x}_i) \mathbf{x}_i] \\
&= -\sum_{i=1}^N [y_i \mathbf{x}_i - \sigma(\mathbf{w}^T \mathbf{x}_i) \mathbf{x}_i] \\
&= \sum_{i=1}^N [\sigma(\mathbf{w}^T \mathbf{x}_i) \mathbf{x}_i - y_i \mathbf{x}_i] \\
&= \sum_{i=1}^N [\mu_i - y_i] \mathbf{x}_i \\
&= \mathbf{X}^T (\boldsymbol{\mu} - \mathbf{y}) \text{ (Equation 8.5).}
\end{aligned}$$

■

(c)

The Hessian is given by:

$$\begin{aligned}
&= \frac{d}{d\mathbf{w}}(\mathbf{g}(\mathbf{w})^T) \\
&= \nabla_{\mathbf{w}}(\mathbf{g}(\mathbf{w})^T) \\
&= \nabla_{\mathbf{w}}(\mathbf{X}^T(\boldsymbol{\mu} - \mathbf{y}))^T \\
&= \nabla_{\mathbf{w}}((\boldsymbol{\mu} - \mathbf{y})^T \mathbf{X}) \quad (\text{by the property of the matrix transpose of a product}) \\
&= \nabla_{\mathbf{w}}(\boldsymbol{\mu}^T \mathbf{X} - \mathbf{y}^T \mathbf{X}) \quad (\text{by the property of the matrix transpose of a difference}) \\
&= \nabla_{\mathbf{w}}(\boldsymbol{\mu}^T \mathbf{X}) - \nabla_{\mathbf{w}}(\mathbf{y}^T \mathbf{X}) \\
&= \nabla_{\mathbf{w}}(\boldsymbol{\mu}^T \mathbf{X}) \quad (\text{since } \mathbf{y} \text{ is not dependent on } \mathbf{w}) \\
&= \nabla_{\mathbf{w}} \sum_i^N \sigma(\mathbf{w}^T \mathbf{x}_i)^T \mathbf{x}_i \\
&= \sum_i^N \sigma'(\mathbf{w}^T \mathbf{x}_i)^T \mathbf{x}_i \mathbf{x}_i \\
&= \mathbf{X}^T \text{diag}(\mu_1(1 - \mu_1), \dots, \mu_n(1 - \mu_n)) \mathbf{X} \\
&= \mathbf{X}^T \mathbf{S} \mathbf{X}.
\end{aligned}$$

Since  $\mathbf{S}$  is a diagonal matrix, the eigenvalues are the values on the diagonal. Since all of the values on the diagonal are positive, we see that  $\mathbf{S}$  satisfies the definition of being a positive semi-definite matrix. ■

(I did not have time to fully explain the last three steps in the derivation of the Hessian matrix, and I did not have time to prove that the values in the diagonal are all positive).

**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.

For  $\mathbb{P}(x; \sigma^2)$  to be a valid density, then  $\mathbb{P}(x; \sigma^2)$  must integrate to 1 (Murphy 2.2.5). Thus, we know that:

$$\begin{aligned} 1 &= \int_{-\infty}^{\infty} \mathbb{P}(x; \sigma^2) dx \\ &= \int_{-\infty}^{\infty} \frac{1}{Z} e^{-\frac{x^2}{2\sigma^2}} dx, \text{ and squaring the equation:} \\ 1 &= \int_{-\infty}^{\infty} \frac{1}{Z} e^{-\frac{x^2}{2\sigma^2}} dx \int_{-\infty}^{\infty} \frac{1}{Z} e^{-\frac{y^2}{2\sigma^2}} dy \\ &= \frac{1}{Z^2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-\frac{x^2+y^2}{2\sigma^2}} dx dy \\ &= \frac{1}{Z^2} \int_0^{\infty} \int_0^{2\pi} e^{-\frac{r^2}{2\sigma^2}} r dr d\theta \quad (\text{converting to polar}) \\ &= \frac{2\pi}{Z^2} \int_0^{\infty} e^{-\frac{r^2}{2\sigma^2}} r dr \\ &= \frac{2\pi}{Z^2} \sigma^2, \text{ so we know that} \\ Z^2 &= 2\pi\sigma^2 \\ Z &= \sqrt{2\pi\sigma^2} \\ Z &= \sigma\sqrt{2\pi}. \end{aligned}$$

■

\*Note: I learned how to solve this integral from an MIT OpenCourseWare lecture.

**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  $\mathbf{x}^*$  to the ridge regression problem:

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

- (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  $\lambda$  from the validation set. Plot both  $\lambda$  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  $\lambda$  versus  $\|\boldsymbol{\theta}^*\|_2$  where  $\boldsymbol{\theta}$  is your weight vector. What is the final RMSE on the test set with the optimal  $\lambda^*$ ?

(continued on the following pages)

**3 (continued)**

- (d) **(math)** Consider regularized linear regression where we pull the bias term out of the feature vectors. That is, instead of computing  $\hat{\mathbf{y}} = \boldsymbol{\theta}^\top \mathbf{x}$  with  $x_0 = 1$ , we compute  $\hat{\mathbf{y}} = \boldsymbol{\theta}^\top \mathbf{x} + b$ . This corresponds to solving the optimization problem

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

Solve for the optimal  $\mathbf{x}^*$  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

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

Compute the gradients and run gradient descent. Plot the  $\ell_2$  norm between the optimal  $(\mathbf{x}^*, b^*)$  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.

\*Note: The use of "log" below refers to the natural logarithm.

**(a)**

The Gaussian distribution is defined as:

$$\mathcal{N}(x|\mu, \sigma^2) \triangleq \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \quad (\text{Murphy, Equation 2.43}).$$

Beginning with the maximum a posteriori problem for linear regression, we have:

$$\begin{aligned}
& \arg \max_{\mathbf{w}} \sum_{i=1}^N \log \mathcal{N}(y_i | w_0 + \mathbf{w}^T \mathbf{x}_i, \sigma^2) + \sum_{j=1}^D \log \mathcal{N}(w_j | 0, \tau^2) \\
&= \arg \max_{\mathbf{w}} \sum_{i=1}^N \log \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i - (w_0 + \mathbf{w}^T \mathbf{x}_i))^2} + \sum_{j=1}^D \log \frac{1}{\sqrt{2\pi\tau^2}} e^{-\frac{1}{2\tau^2}w_j^2} \quad (\text{Equation 2.43}) \\
&= \arg \max_{\mathbf{w}} \sum_{i=1}^N [\log \frac{1}{\sqrt{2\pi\sigma^2}} + \log e^{-\frac{1}{2\sigma^2}(y_i - (w_0 + \mathbf{w}^T \mathbf{x}_i))^2}] + \sum_{j=1}^D [\log \frac{1}{\sqrt{2\pi\tau^2}} + \log e^{-\frac{1}{2\tau^2}w_j^2}] \\
&= \arg \max_{\mathbf{w}} \sum_{i=1}^N [\log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2}(y_i - (w_0 + \mathbf{w}^T \mathbf{x}_i))^2] + \sum_{j=1}^D [\log \frac{1}{\sqrt{2\pi\tau^2}} - \frac{1}{2\tau^2}w_j^2] \\
&= \arg \max_{\mathbf{w}} N \log \frac{1}{\sqrt{2\pi\sigma^2}} + D \log \frac{1}{\sqrt{2\pi\tau^2}} - \frac{1}{2\sigma^2} \sum_{i=1}^N (y_i - (w_0 + \mathbf{w}^T \mathbf{x}_i))^2 + \sum_{j=1}^D \frac{\sigma^2}{\tau^2} w_j^2 \\
&= \arg \max_{\mathbf{w}} -[\sum_{i=1}^N (y_i - (w_0 + \mathbf{w}^T \mathbf{x}_i))^2 + \sum_{j=1}^D \frac{\sigma^2}{\tau^2} w_j^2] \\
&\quad (\text{the maximization of a function is unaffected by constant terms or scalars}) \\
&= \arg \min_{\mathbf{w}} \sum_{i=1}^N (y_i - (w_0 + \mathbf{w}^T \mathbf{x}_i))^2 + \sum_{j=1}^D \frac{\sigma^2}{\tau^2} w_j^2 \\
&\quad (\text{maximization is equivalent to the minimization of the negative of the function}) \\
&= \arg \min_{\mathbf{w}} \sum_{i=1}^N (y_i - (w_0 + \mathbf{w}^T \mathbf{x}_i))^2 + \sum_{j=1}^D \lambda w_j^2 \quad (\text{since } \lambda = \frac{\sigma^2}{\tau^2}) \\
&= \arg \min_{\mathbf{w}} \sum_{i=1}^N (y_i - (w_0 + \mathbf{w}^T \mathbf{x}_i))^2 + \lambda \sum_{j=1}^D w_j^2 \\
&= \arg \min_{\mathbf{w}} \frac{1}{N} \sum_{i=1}^N (y_i - (w_0 + \mathbf{w}^T \mathbf{x}_i))^2 + \lambda \sum_{j=1}^D w_j^2 \quad (\text{scalar does not effect minimization of function}) \\
&= \arg \min_{\mathbf{w}} \sum_{i=1}^N (y_i - (w_0 + \mathbf{w}^T \mathbf{x}_i))^2 + \lambda \|\mathbf{w}\|_2^2 \quad (\text{Murphy, Equation 7.32}) \\
&\quad (\text{since } \|\mathbf{w}\|_2^2 = \sum_{j=1}^D w_j^2).
\end{aligned}$$

■

(b)

First, we will find a critical point by taking the gradient of the function with respect to  $\mathbf{x}$



and setting it equal to zero:

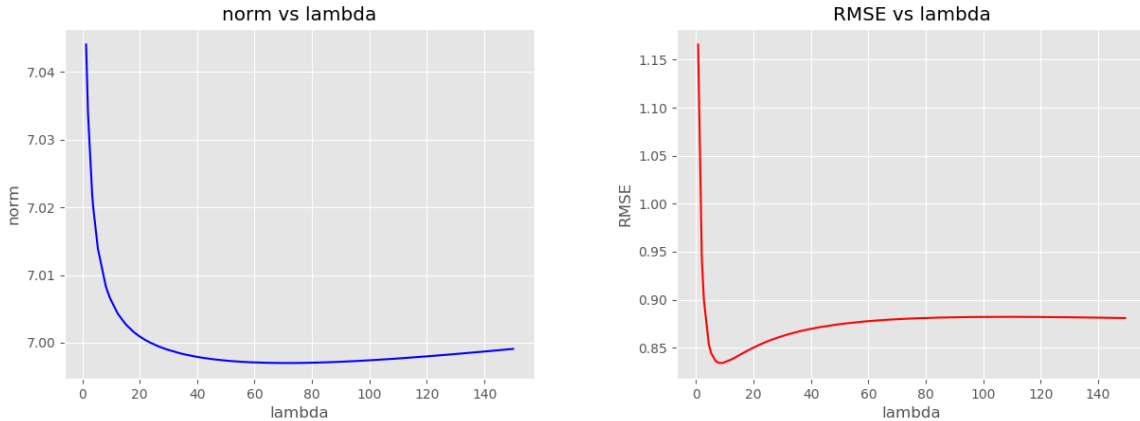
$$\begin{aligned}
& \nabla_{\mathbf{x}}[||A\mathbf{x} - \mathbf{b}||_2^2 + ||\Gamma\mathbf{x}||_2^2] \\
&= \nabla_{\mathbf{x}}[(A\mathbf{x} - \mathbf{b})^T(A\mathbf{x} - \mathbf{b}) + (\Gamma\mathbf{x})^T(\Gamma\mathbf{x})] \quad (\text{since } ||\mathbf{w}||_2^2 = \mathbf{w}^T\mathbf{w} \text{ by Equation 7.32}) \\
&= \nabla_{\mathbf{x}}[(\mathbf{x}^T A^T - \mathbf{b}^T)(A\mathbf{x} - \mathbf{b}) + (\mathbf{x}^T \Gamma^T)(\Gamma\mathbf{x})] \quad (\text{distributive property of matrix transpose}) \\
&= \nabla_{\mathbf{x}}[\mathbf{x}^T A^T A\mathbf{x} - \mathbf{x}^T A^T \mathbf{b} - \mathbf{b}^T A\mathbf{x} - \mathbf{b}^T \mathbf{b} + \mathbf{x}^T \Gamma^T \Gamma \mathbf{x}] \\
&= (A^T A\mathbf{x} + \mathbf{x}^T A^T A) - A^T \mathbf{b} - \mathbf{b}^T A - 0 + (\Gamma^T \Gamma \mathbf{x} + \mathbf{x}^T \Gamma^T \Gamma) \quad (\text{Product rule}).
\end{aligned}$$

Setting the gradient of the function equal to 0, we can solve for  $\mathbf{x}$ :

$$\begin{aligned}
0 &= A^T A\mathbf{x} + \mathbf{x}^T A^T A - A^T \mathbf{b} - \mathbf{b}^T A + \Gamma^T \Gamma \mathbf{x} + \mathbf{x}^T \Gamma^T \Gamma \\
A^T \mathbf{b} + \mathbf{b}^T &= A^T A\mathbf{x} + \mathbf{x}^T A^T A + \Gamma^T \Gamma \mathbf{x} + \mathbf{x}^T \Gamma^T \Gamma \\
2A^T \mathbf{b} &= 2A^T A\mathbf{x} + 2\Gamma^T \Gamma \mathbf{x} \\
A^T \mathbf{b} &= A^T A\mathbf{x} + \Gamma^T \Gamma \mathbf{x} \\
A^T \mathbf{b} &= (A^T A + \Gamma^T \Gamma)\mathbf{x}, \text{ so} \\
\mathbf{x} &= (A^T A + \Gamma^T \Gamma)^{-1} A^T \mathbf{b}, \text{ thus} \\
\mathbf{x}^* &= (A^T A + \Gamma^T \Gamma)^{-1} A^T \mathbf{b}.
\end{aligned}$$

■

(c) The optimal regularization parameter is 8.4952. The RMSE on the validation set with the optimal regularization parameter is 0.8340. The RMSE on the test set with the optimal regularization parameter is 0.8628.



(d)

\*Note: I used the posted solutions to check my matrix algebra in this part.

As in part (b), we will use Equation 7.32 to minimize the function:

$$\begin{aligned}
& ||A\mathbf{x} + b\mathbf{1} - \mathbf{y}||_2^2 + ||\Gamma\mathbf{x}||_2^2 \\
&= (A\mathbf{x} + b\mathbf{1} - \mathbf{y})^T (A\mathbf{x} + b\mathbf{1} - \mathbf{y}) + (\Gamma\mathbf{x})^T (\Gamma\mathbf{x}) \\
&= (\mathbf{x}^T A^T + b\mathbf{1}^T - \mathbf{y}^T) (A\mathbf{x} + b\mathbf{1} - \mathbf{y}) + \mathbf{x}^T \Gamma^T \Gamma \mathbf{x} \\
&= \mathbf{x}^T A^T A \mathbf{x} + \mathbf{x}^T A^T b\mathbf{1} - \mathbf{x}^T A^T \mathbf{y} + b\mathbf{1}^T A \mathbf{x} + b\mathbf{1}^T b\mathbf{1} - b\mathbf{1}^T \mathbf{y} - \mathbf{y}^T A \mathbf{x} - \mathbf{y}^T b\mathbf{1} + \mathbf{y}^T \mathbf{y} + \mathbf{x}^T \Gamma^T \Gamma \mathbf{x} \\
&= \mathbf{x}^T A^T A \mathbf{x} + 2b\mathbf{1}^T A \mathbf{x} - 2\mathbf{y}^T A \mathbf{x} - 2b\mathbf{1}^T \mathbf{y} + b^2 n + \mathbf{y}^T \mathbf{y} + \mathbf{x}^T \Gamma^T \Gamma \mathbf{x}.
\end{aligned}$$

We will use the following rule when taking the gradient of the function above:

$$\frac{d(x^T a)}{dx} = \frac{d(a^T x)}{dx} = a^T.$$

Taking the gradient of this function with respect to  $b$ , we have:

$$\begin{aligned}
& \nabla_b [\mathbf{x}^T A^T A \mathbf{x} + 2b\mathbf{1}^T A \mathbf{x} - 2\mathbf{y}^T A \mathbf{x} - 2b\mathbf{1}^T \mathbf{y} + b^2 n + \mathbf{y}^T \mathbf{y} + \mathbf{x}^T \Gamma^T \Gamma \mathbf{x}] \\
&= 2\mathbf{1}^T A \mathbf{x} - 2\mathbf{1}^T \mathbf{y} + 2bn. \quad \text{Setting the gradient equal to zero, we have:} \\
0 &= 2\mathbf{1}^T A \mathbf{x} - 2\mathbf{1}^T \mathbf{y} + 2bn, \quad \text{so} \\
2bn &= 2\mathbf{1}^T \mathbf{y} - 2\mathbf{1}^T A \mathbf{x} \\
b &= \frac{2\mathbf{1}^T \mathbf{y} - 2\mathbf{1}^T A \mathbf{x}}{2n} \\
b &= \frac{2\mathbf{1}^T (\mathbf{y} - A \mathbf{x})}{2n}, \quad \text{so} \\
b^* &= \frac{\mathbf{1}^T (\mathbf{y} - A \mathbf{x})}{n}.
\end{aligned}$$

Now we can take the gradient of the function with respect to  $x$ , and we can then substitute

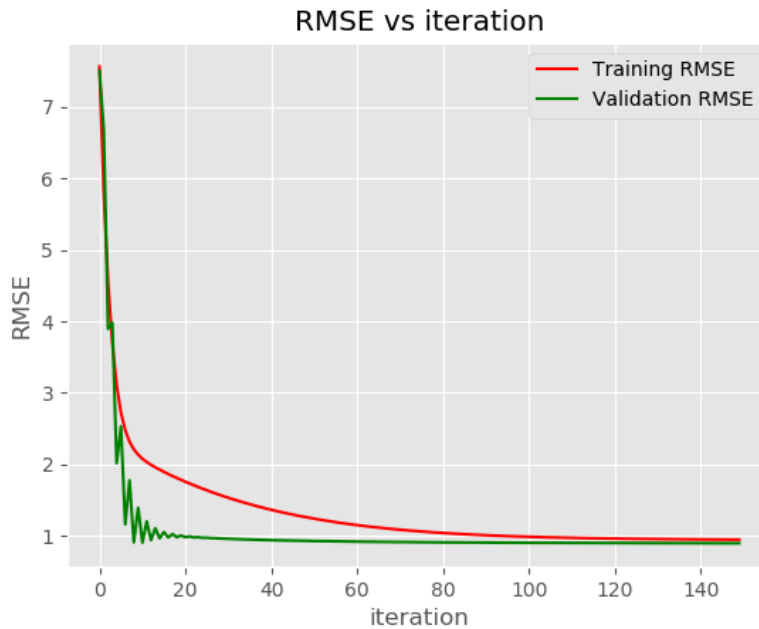
the value for  $b^*$  into the equation:

$$\begin{aligned}
& \nabla_{\mathbf{x}}[\mathbf{x}^T A^T A \mathbf{x} + 2b \mathbf{1}^T A \mathbf{x} - 2\mathbf{y}^T A \mathbf{x} - 2b \mathbf{1}^T \mathbf{y} + b^2 n + \mathbf{y}^T \mathbf{y} + \mathbf{x}^T \Gamma^T \Gamma \mathbf{x}] \\
& = 2A^T A \mathbf{x} + 2A^T b \mathbf{1} - 2A^T \mathbf{y} + 2\Gamma^T \Gamma \mathbf{x}. \quad \text{Setting the gradient equal to zero, we have:} \\
& 0 = 2A^T A \mathbf{x} + 2A^T b \mathbf{1} - 2A^T \mathbf{y} + 2\Gamma^T \Gamma \mathbf{x} \\
& 0 = 2A^T A \mathbf{x} + 2A^T \left( \frac{\mathbf{1}^T (\mathbf{y} - A \mathbf{x})}{n} \right) \mathbf{1} - 2A^T \mathbf{y} + 2\Gamma^T \Gamma \mathbf{x} \\
& 0 = A^T A \mathbf{x} + A^T \mathbf{1}^T \mathbf{y} \frac{1}{n} - A^T \mathbf{1}^T A \mathbf{x} \frac{1}{n} - A^T \mathbf{y} + \Gamma^T \Gamma \mathbf{x} \\
& A^T \mathbf{y} - A^T \mathbf{1}^T \mathbf{y} \frac{1}{n} = A^T A \mathbf{x} - A^T \mathbf{1}^T A \mathbf{x} \frac{1}{n} + \Gamma^T \Gamma \mathbf{x} \\
& A^T \mathbf{y} - A^T \mathbf{1}^T \mathbf{y} \frac{1}{n} = (A^T A - A^T \mathbf{1}^T A \frac{1}{n} + \Gamma^T \Gamma) \mathbf{x} \\
& A^T \mathbf{y} - A^T \mathbf{1}^T \mathbf{y} \frac{1}{n} = (A^T A - A^T \mathbf{1}^T A \frac{1}{n} + \Gamma^T \Gamma) \mathbf{x}, \quad \text{thus} \\
& \mathbf{x} = (A^T A - A^T \mathbf{1}^T A \frac{1}{n} + \Gamma^T \Gamma)^{-1} (A^T \mathbf{y} - A^T \mathbf{1}^T \mathbf{y} \frac{1}{n}).
\end{aligned}$$

■

Time elapsed for training: 26.63 seconds. Difference in bias is 4.2715E-10. Difference in weights is 5.6147E-10.

(e)



Difference in bias is 1.5385E-01. Difference in weights is 7.9228E-01.