Thitaree Tanprasert (Mint)
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) Given $\sigma(x) = \frac{1}{1+e^{-x}}$, we can use the formula for derivative of quotient, then simply as follow:

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

(b) The nagative log-likehood of logistic regression is

$$\mathcal{L}(\boldsymbol{\theta}) = -\sum_i y_i \log \sigma(\boldsymbol{\theta}^\top \mathbf{x}_i) + (1-y_i)\log(1-\sigma(\boldsymbol{\theta}^\top \mathbf{x}_i))$$

We take the gradient with respect to $\boldsymbol{\theta}$ using the chain rule as follow:

$$\nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}) = -\sum_i y_i \frac{1}{\sigma(\boldsymbol{\theta}^\top \mathbf{x}_i)} \sigma'(\boldsymbol{\theta}^\top \mathbf{x}_i) + (1-y_i)\frac{1}{1-\sigma(\boldsymbol{\theta}^\top \mathbf{x}_i)}(-\sigma'(\boldsymbol{\theta}^\top \mathbf{x}_i))$$

Then, we use result from part(a) to replace $\sigma'(\boldsymbol{\theta}^\top \mathbf{x}_i)$ and use the chain rule as follow:

$$\begin{aligned}
\nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}) &= -\sum_i y_i \frac{\sigma(\boldsymbol{\theta}^\top \mathbf{x}_i)}{\sigma(\boldsymbol{\theta}^\top \mathbf{x}_i)}(1-\sigma(\boldsymbol{\theta}^\top \mathbf{x}_i))\mathbf{x}_i + (1-y_i)1-\sigma(\boldsymbol{\theta}^\top \mathbf{x}_i)1-\sigma(\boldsymbol{\theta}^\top \mathbf{x}_i)(-\sigma(\boldsymbol{\theta}^\top \mathbf{x}_i))\mathbf{x}_i \\
&= -\sum_i y_i(1-\sigma(\boldsymbol{\theta}^\top \mathbf{x}_i))\mathbf{x}_i + (1-y_i)(-\sigma(\boldsymbol{\theta}^\top \mathbf{x}_i))\mathbf{x}_i \\
&= -\sum_i y_i \mathbf{x}_i - y_i \sigma(\boldsymbol{\theta}^\top \mathbf{x}_i)\mathbf{x}_i - \sigma(\boldsymbol{\theta}^\top \mathbf{x}_i)\mathbf{x}_i + y_i \sigma(\boldsymbol{\theta}^\top \mathbf{x}_i)\mathbf{x}_i \\
&= \sum_i (\sigma(\boldsymbol{\theta}^\top \mathbf{x}_i) - y_i)\mathbf{x}_i
\end{aligned}$$

If we let $\mu_i = \sigma(\boldsymbol{\theta}^\top \mathbf{x}_i)$, the gradient becomes

$$\nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}) = \sum_i (\mu_i - y_i)\mathbf{x}_i$$

Let $\boldsymbol{\mu}$ be the vector of all μ_i, $\mathbf{y}$ be the vector of all y_i and $\mathbf{X}$ be a matrix where each row corresponds to $\mathbf{x}_i$. We can simply the equation above as

$$\nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}) = \mathbf{X}^\top(\boldsymbol{\mu} - \mathbf{y})$$

(c) The formula for Hessian matrix is $H_{\boldsymbol{\theta}}(\mathcal{L}(\boldsymbol{\theta})) = \nabla_{\boldsymbol{\theta}}(\nabla_{\boldsymbol{\theta}}(\mathcal{L}(\boldsymbol{\theta})))^\top$. We can plug the result from part (b) into this equation as follow:

$$\begin{aligned}
H_{\boldsymbol{\theta}}(\mathcal{L}(\boldsymbol{\theta})) &= \nabla_{\boldsymbol{\theta}}[\mathbf{X}^\top(\boldsymbol{\mu} - \mathbf{y})]^\top \\
&= \nabla_{\boldsymbol{\theta}}(\boldsymbol{\mu} - \mathbf{y})^\top \mathbf{X} \\
&= \nabla_{\boldsymbol{\theta}}(\boldsymbol{\mu}^\top \mathbf{X} - \mathbf{y}^\top \mathbf{X})
\end{aligned}$$

Since $\mathbf{y}$ is independent of $\boldsymbol{\theta}$, it disappears when we take the gradient with respect to $\boldsymbol{\theta}$. And then, we can replace $\boldsymbol{\mu}$ with the definition we assign it in part (b):

$$\begin{aligned}
H_{\boldsymbol{\theta}}(\mathcal{L}(\boldsymbol{\theta})) &= \nabla_{\boldsymbol{\theta}} \boldsymbol{\mu}^\top \mathbf{X} \\
&= \nabla_{\boldsymbol{\theta}} \sigma(\mathbf{X}\boldsymbol{\theta})^\top \mathbf{X} \\
&= \nabla_{\boldsymbol{\theta}} \mathbf{X}^\top \sigma(\boldsymbol{\theta})^\top \mathbf{X} \\
&= \mathbf{X}^\top \nabla_{\boldsymbol{\theta}} \sigma(\boldsymbol{\theta})^\top \mathbf{X}
\end{aligned}$$

And then, we replace derivative of σ with the result from part (a):

$$H_{\boldsymbol{\theta}}(\mathcal{L}(\boldsymbol{\theta})) = \mathbf{X}^\top \text{ↄↄð}(\boldsymbol{\mu}(1-\boldsymbol{\mu}))\mathbf{X} = \mathbf{X}^\top \mathbf{S}\mathbf{X}$$

Finally, to show that H is positive semidefinite, we need to show that its eigenvalue is positive. But for a diagonal matrix, we only need to consider the diagonal elements, which, in this case, are $\text{diag}(\mu_1(1-\mu_1), \ldots, \mu_n(1-\mu_n))$. Since $\mu_i = \sigma(\boldsymbol{\theta}^\top \mathbf{x}_i)(1 - \sigma(\boldsymbol{\theta}^\top \mathbf{x}_i))$ and σ are probabilities, so we know that they are positive value less than 1. Therefore, all μ_i must be positive, and so the eigenvalue is positive, which means that H is positive semidefinite.

■

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.

We know that a valid density means that the probability integrates to 1. So, we can get an equation for Z as follow:

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

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

Following the steps in the textbook, we take the sequare of Z as follow:

$$Z^2 = \int_{\mathbb{R}}\int_{\mathbb{R}} \exp\left(-\frac{(x^2+y^2)}{2\sigma^2}\right)dxdy$$

Now, we will represent x and y in polar coordinates, where $x = r\cos\theta$ and $y = r\sin\theta$. Consequently, we have $dxdy = rdrd\theta$. Therefore, our equation becomes

$$Z^2 = \int_0^{\infty}\int_0^{2\pi} \exp\left(-\frac{r^2(\cos^2\theta+\sin^2\theta)}{2\sigma^2}\right)rdrd\theta$$

Since, $\cos^2\theta + \sin^2\theta = 1$, we get

$$Z^2 = \int_0^{\infty}\int_0^{2\pi} \exp\left(-\frac{r^2}{2\sigma^2}\right)rdrd\theta$$

We first integrate the $d\theta$,

$$Z^2 = 2\pi\int_0^{\infty} \exp\left(-\frac{r^2}{2\sigma^2}\right)rdr$$

Then, we observe that if $u = \exp\left(\frac{-r^2}{2\sigma^2}\right)$, then $du/dr = \frac{-1}{\sigma^2}r\exp\left(\frac{-r^2}{2\sigma^2}\right)$. Therefore, our equation becomes

$$\begin{aligned}
Z^2 &= 2\pi\int_0^{\infty}(-\sigma^2)(\frac{-1}{\sigma^2})\exp\left(-\frac{r^2}{2\sigma^2}\right)rdr \\
&= 2\pi(-\sigma^2)[\exp\left(\frac{-r^2}{2r^2}\right)]\Big|_0^{\infty} \\
&= 2\pi(-\sigma^2)(0-1) \\
&= 2\pi\sigma^2
\end{aligned}$$

Therefore, we get $Z = \sigma\sqrt{2\pi}$.

■

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) First, we plug in the formula for probability distribution for general variables y, x, σ, which is $\mathcal{N}(y|x, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y-\boldsymbol{\theta}^\top \mathbf{x})^2}{2\sigma^2}\right)$ into the given maximization equation to get:

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

Then, we expand the log function into the equation:

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

We combine the like terms $\sqrt{2\pi}\sigma$:

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

Now, we will make two observations. The first observation is that maximizing a problem is equivalent to minimizing the negative of the original problem. The second observation is that, in the maximization process, we don't have to consider the constant terms. We can remove or add the constant terms without effecting the result. Therefore, we can modify the equation by removing the term $(N+D)\sqrt{2\pi}\sigma$

and multiply a constant term $2\sigma^2$:

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

Then, we plug in $\lambda=\sigma^2/\tau^2$:

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

Finally, we plug in $||\mathbf{w}||_2^2=\sum_{j=1}^{D}w_j^2$ to get the desired result:

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

(b) Let $f=||A\mathbf{x}-\mathbf{b}||_2^2+||\Gamma\mathbf{x}||_2^2$. We want to find x that makes $\nabla_x f=0$. First, we expand the ridge regression problem by using the definition $||\mathbf{x}||_2^2=\mathbf{x}^\top\mathbf{x}$:

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

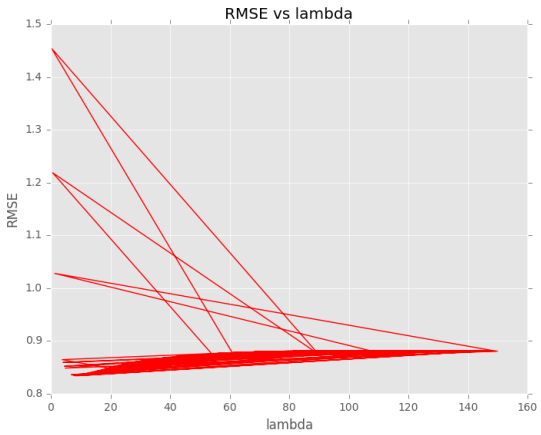
Now, we set this equation = 0 and solve for $\mathbf{x}^*$:

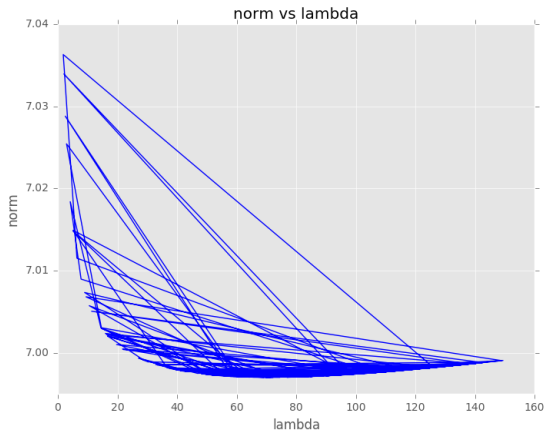
$$\begin{aligned}2A^\top A\mathbf{x}^*-2A^\top\mathbf{b}+2\Gamma^\top\Gamma\mathbf{x}^*&=0\\A^\top A\mathbf{x}^*+\Gamma^\top\Gamma\mathbf{x}^*&=A^\top\mathbf{b}\\(A^\top A+\Gamma^\top\Gamma)\mathbf{x}^*&=A^\top\mathbf{b}\\\mathbf{x}^*&=(A^\top A+\Gamma^\top\Gamma)^{-1}A^\top\mathbf{b}\end{aligned}$$

Finally, we can replace Γ using the regularization parameter λ: $\Gamma=\sqrt{\lambda}\mathbf{I}$

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

(c) From running the implementation, we get the following results:
- The optimal regularization parameter is 8.4517.
- 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) Similarly to part (b), we have to find $\mathbf{x}^*$ and $\mathbf{b}^*$ which makes the gradient taken with respect to $\mathbf{x}$ and with respect to $\mathbf{b}$ 0. Let $f = ||A\mathbf{x} - b\mathbf{1} - \mathbf{y}||_2^2 + ||\Gamma\mathbf{x}||_2^2$. Firstly, we will expand and simplify f:

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

Now, we can take the gradient with respect to $\mathbf{x}$ and b and set them equal to 0. First, we will take the gradient with respect to b and solve for b^*:

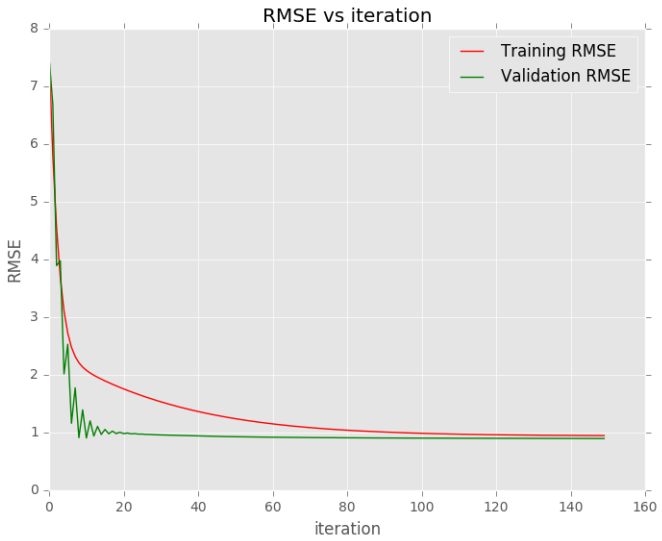
$$\begin{aligned}
\nabla_b f &= 2\mathbf{1}^\top A\mathbf{x} - 2\mathbf{1}^\top\mathbf{y} + 2bm \\
0 &= 2\mathbf{1}^\top A\mathbf{x} - 2\mathbf{1}^\top\mathbf{y} + 2b^* m \\
b^* &= (2\mathbf{1}^\top A\mathbf{x} - 2\mathbf{1}^\top\mathbf{y})/2m \\
&= \mathbf{1}^\top(\mathbf{y} - A\mathbf{x})/m
\end{aligned}$$

Now, we will use this value to solve for x^*:

$$\begin{aligned}
\nabla_x f &= 2A^\top A\mathbf{x} + 2bA^\top \mathbf{1} - 2A^\top \mathbf{y} + 2\Gamma^\top \Gamma \mathbf{x} \\
&= 2A^\top A\mathbf{x} + 2[\mathbf{1}^\top(\mathbf{y} - A\mathbf{x})/m]A^\top \mathbf{1} - 2A^\top \mathbf{y} + 2\Gamma^\top \Gamma \mathbf{x} \\
&= 2A^\top A\mathbf{x} + \frac{2}{m}\mathbf{1}^\top \mathbf{y} A^\top \mathbf{1} - \frac{2}{m}\mathbf{1}^\top A\mathbf{x} A^\top \mathbf{1} - 2A^\top \mathbf{y} + 2\Gamma^\top \Gamma \mathbf{x} \\
&= 2A^\top A\mathbf{x} + \frac{2}{m}A^\top \mathbf{1}\mathbf{1}^\top \mathbf{y} - \frac{2}{m}A^\top \mathbf{1}\mathbf{1}^\top A\mathbf{x} - 2A^\top \mathbf{y} + 2\Gamma^\top \Gamma \mathbf{x} \\
0 &= 2A^\top A\mathbf{x}^* + \frac{2}{m}A^\top \mathbf{1}\mathbf{1}^\top \mathbf{y} - \frac{2}{m}A^\top \mathbf{1}\mathbf{1}^\top A\mathbf{x}^* - 2A^\top \mathbf{y} + 2\Gamma^\top \Gamma \mathbf{x}^* \\
&= A^\top A\mathbf{x}^* + \frac{1}{m}A^\top \mathbf{1}\mathbf{1}^\top \mathbf{y} - \frac{1}{m}A^\top \mathbf{1}\mathbf{1}^\top A\mathbf{x}^* - A^\top \mathbf{y} + \Gamma^\top \Gamma \mathbf{x}^* \\
A^\top \mathbf{y} - \frac{1}{m}A^\top \mathbf{1}\mathbf{1}^\top \mathbf{y} &= A^\top A\mathbf{x}^* + \frac{1}{m}A^\top \mathbf{1}\mathbf{1}^\top A\mathbf{x}^* + \Gamma^\top \Gamma \mathbf{x}^* \\
&= (A^\top A + \frac{1}{n}A^\top \mathbf{1}\mathbf{1}^\top A + \Gamma^\top \Gamma)\mathbf{x}^* \\
A^\top (\mathbf{I} - \frac{1}{m}\mathbf{1}\mathbf{1}^\top)\mathbf{y} &= [A^\top (\mathbf{I} + \frac{1}{m}\mathbf{1}\mathbf{1}^\top)A + \Gamma^\top \Gamma]\mathbf{x}^* \\
\mathbf{x}^* &= [A^\top (\mathbf{I} + \frac{1}{m}\mathbf{1}\mathbf{1}^\top)A + \Gamma^\top \Gamma]^{-1} A^\top (\mathbf{I} - \frac{1}{m}\mathbf{1}\mathbf{1}^\top)\mathbf{y}
\end{aligned}$$

Using the closed form to compute the bias term, we get the following values from Python implementation:
- Difference in bias is 4.4007E-10
- Difference in weights is 5.8277E-10

(e) The implementation produces the following results:
- Difference in bias is 1.5387E-01
- Difference in weights is 8.0196E-01



■