Colin Adams
Math189R SU20
Homework 2
June 2020

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) Note that

$$1-\sigma(x) = \frac{e^{-x}}{1+e^{-x}} \quad \text{and} \quad \sigma'(x) = \frac{e^{-x}}{(1+e^{-x})^2},$$

which makes it easy to see that

$$\sigma'(x) = \frac{e^{-x}}{(1+e^{-x})^2} = \frac{e^{-x}}{1+e^{-x}}\frac{1}{1+e^{-x}} = \sigma(x)(1-\sigma(x)).$$

(b) Okay, so if we use the negative log-likelihood of the logistic regression, we have

$$\text{NLL} = -\sum_i \left[ y_i \ln \sigma(\mathbf{w}\cdot\mathbf{x}_i) + (1-y_i)\ln(1-\sigma(\mathbf{w}\cdot\mathbf{x}_i))\right]$$

and if we take the gradient with respect to $\mathbf{w}$ is given by

$$\begin{aligned}
\nabla_{\mathbf{w}}\text{NLL} &= -\sum_i \nabla_{\mathbf{w}}\big[y_i \ln\sigma(\mathbf{w}\cdot\mathbf{x}_i) + (1-y_i)\ln(1-\sigma(\mathbf{w}\cdot\mathbf{x}_i))\big] \\
&= -\sum_i \left[y_i \frac{\sigma'(\mathbf{w}\cdot\mathbf{x}_i)}{\sigma(\mathbf{w}\cdot\mathbf{x}_i)}\mathbf{x}_i - (1-y_i)\frac{\sigma'(\mathbf{w}\cdot\mathbf{x}_i)}{1-\sigma(\mathbf{w}\cdot\mathbf{x}_i)}\mathbf{x}_i\right] \\
&= -\sum_i \big[y_i(1-\sigma(\mathbf{w}\cdot\mathbf{x}_i))\mathbf{x}_i - (1-y_i)\sigma(\mathbf{w}\cdot\mathbf{x}_i)\mathbf{x}_i\big] \qquad \text{(result of (a))} \\
&= \sum_i \big[\sigma(\mathbf{w}\cdot\mathbf{x}_i) - y_i\big]\mathbf{x}_i
\end{aligned}$$

which can be rewritten in a more compact matrix form as

$$\nabla_{\mathbf{w}}\text{NLL} = \mathbf{X}^T(\boldsymbol{\mu}-\mathbf{y}) \quad \text{where} \quad \boldsymbol{\mu} = \sigma(\mathbf{X}\mathbf{w}).$$

(c) The Hessian $\mathbf{H}$ is the gradient of the gradient (transposed) and so

$$\begin{aligned}
\mathbf{H} &= \nabla_{\mathbf{w}}\big(\mathbf{X}^T(\boldsymbol{\mu}-\mathbf{y})\big)^T \\
&= \nabla_{\mathbf{w}}\big((\boldsymbol{\mu}-\mathbf{y})^T\mathbf{X}\big) \\
&= \left(\sum_i \nabla_{\mathbf{w}}[\sigma(\mathbf{w}\cdot\mathbf{x}_i)]\mathbf{x}_i\right)\mathbf{X} \\
&= \left(\sum_i \sigma(\mathbf{w}\cdot\mathbf{x}_i)(1-\sigma(\mathbf{w}\cdot\mathbf{x}_i))\mathbf{x}_i\mathbf{x}_i^T\right)\mathbf{X} \\
&= \left(\sum_i \mu_i(1-\mu_i)\mathbf{x}_i\mathbf{x}_i^T\right)\mathbf{X} \\
&= \mathbf{X}^T\text{diag}(\mu_i(1-\mu_i))\mathbf{X} \\
&= \mathbf{X}^T\mathbf{S}\mathbf{X} \quad \text{where} \quad \mathbf{S} \equiv \text{diag}(\mu_i(1-\mu_i))
\end{aligned}$$

which is what we were hoping to get. Phew. Now we need to show that $\mathbf{H}$ is positive semidefinite, which means $\langle\mathbf{v}|\mathbf{H}\mathbf{v}\rangle \geq 0$ for any nonzero $\mathbf{v}$. A way to do this is to look at the components. So

$$H_{ij} = x_i^2 x_j^2 \mu_i(1-\mu_i)\delta_{ij} \quad \text{where } \delta_{ij} \text{ is the Kronecker delta function}$$

and so—using the Einstein summation convention—we find $\langle\mathbf{v}|\mathbf{H}\mathbf{v}\rangle$ then

$$\begin{aligned}
\langle\mathbf{v}|\mathbf{H}\mathbf{v}\rangle &= v_i H_{ij} v_j \\
&= v_i(x_j x_i \mu_i(1-\mu_i) v_j \delta_{ij} x_i x_j) v_j \\
&= v_i v_j x_i^2 x_j^2 \mu_i(1-\mu_i)\delta_{ij} \\
&= v_i^2 x_i^4 \mu_i(1-\mu_i)
\end{aligned}$$

and since $\sum_i (v_i x_i^2)^2 \geq 0$ for all $v_i, x_i \in \mathbb{R}$, we only need to really care about the $\mu_i(1-\mu_i)$ term. Since $0 < \mu_i < 1$ then $\mu_i(1-\mu_i)$ is also greater than zero. Therefore, $\langle\mathbf{v}|\mathbf{H}\mathbf{v}\rangle \geq 0$ and so $\mathbf{H}$ is positive, semi-definite, i.e. $\mathbf{H} \succeq 0$.

■

**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.

I'm just going to call $\mathbb{P}$ with the easier-to-type $p$ if that's okay. Well, we need to use the fact that

$$\int p(x)\,\mathrm{d}x = 1.$$

In our case

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

so

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

Unfortuntately, this is a hard integral to solve, so we might as well make it twice as hard. If we let

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

then

$$\begin{aligned} I^2 &= \int_{-\infty}^{\infty}\exp\left(-\frac{x^2}{2\sigma^2}\right)\mathrm{d}x\int_{-\infty}^{\infty}\exp\left(-\frac{y^2}{2\sigma^2}\right)\mathrm{d}y \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\exp\left(-\frac{x^2+y^2}{2\sigma^2}\right)\mathrm{d}x\,\mathrm{d}y. \end{aligned}$$

We can evaluate this using polar coordinates, so

$$\begin{aligned} I^2 &= \int_0^{2\pi}\int_0^{\infty}\exp\left(-r^2/2\sigma^2\right)r\,\mathrm{d}r\,\mathrm{d}\theta \\ &= 2\pi\sigma^2\int_{-\infty}^{0}\exp(u)\,\mathrm{d}u \qquad \text{(using a } u\text{-sub. with } u = -r^2/2\sigma^2.\text{)} \\ &= 2\pi\sigma^2 \end{aligned}$$

which implies $I = \sqrt{2\pi}\sigma = 1/Z$. And, finally, $Z = 1/\sqrt{2\pi}\sigma$. Special shout out to Prof. Saeta for showing us how to solve this like three years ago now. His unparalleled sass allowed me to remember exactly what he did to solve this.

■

**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 logshares 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 $||\boldsymbol{\theta}^\star||_2$ 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 $\ell_2$ 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) Well, if we are looking for $\mathbf{w}$ such that

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

is maximized, let's plug in the normal distributions and ignore any constant terms, then we can do a little rearranging to see

$$\begin{aligned}
\text{answer} &= \arg\max_{\mathbf{w}} \sum_{i=1}^{N} \log \mathcal{N}(y_i | w_0 + \mathbf{w}^\top \mathbf{x}_i, \sigma^2) + \sum_{j=1}^{D} \log \mathcal{N}(w_j | 0, \tau^2) \\
&= \arg\max_{\mathbf{w}} \sum_{i}^{N} \frac{-\big(y_i - (w_0 + \mathbf{w} \cdot \mathbf{x}_i)\big)^2}{2\sigma^2} - \sum_{i}^{D} -\frac{w_i^2}{2\tau^2} \\
&= \arg\min_{\mathbf{w}} \sum_{i}^{N} \frac{\big(y_i - (w_0 + \mathbf{w} \cdot \mathbf{x}_i)\big)^2}{2\sigma^2} + \sum_{i}^{D} \frac{w_i^2}{2\tau^2} \\
&= \arg\min_{\mathbf{w}} \sum_{i}^{N} \frac{\big(y_i - (w_0 + \mathbf{w} \cdot \mathbf{x}_i)\big)^2}{2\sigma^2} + \frac{1}{2\tau^2} \|\mathbf{w}\|^2.
\end{aligned}$$

Since we can multiply the function by some constant without affecting where the maximum $\mathbf{w}$ occurs, we can get

$$\text{answer} = \arg\min_{\mathbf{w}} \sum_{i}^{N} \big(y_i - (w_0 + \mathbf{w} \cdot \mathbf{x}_i)\big)^2 + \frac{\sigma^2}{\tau^2} \|\mathbf{w}\|^2.$$

We can also just scale the first term without changing where $\mathbf{w}$ maximizes the function, so

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

as desired. I'm sorry for the clunky notation.

(b) Well, I guess we should probably rewrite $\|\mathbf{Ax}-\mathbf{b}\|^2$ and $\|\Gamma\mathbf{x}\|^2$. Since

$$\|\Gamma\mathbf{x}\|^2 = \mathbf{x}^T\Gamma^T\Gamma\mathbf{x} \quad \text{and} \quad \|\mathbf{Ax}-\mathbf{b}\|^2 = (\mathbf{Ax}-\mathbf{b})^T(\mathbf{Ax}-\mathbf{b})$$
$$= \mathbf{x}^T\mathbf{A}^T\mathbf{Ax} - 2\mathbf{x}^T\mathbf{A}^T\mathbf{b} + \mathbf{b}^T\mathbf{b}$$

then we can take the gradient with respect to $\mathbf{x}$ and get

$$\frac{\mathrm{d}}{\mathrm{d}\mathbf{x}}\left(\|\mathbf{Ax}-\mathbf{b}\|^2 + \|\Gamma\mathbf{x}\|^2\right) = \frac{\mathrm{d}}{\mathrm{d}\mathbf{x}}\left(\mathbf{x}^T\mathbf{A}^T\mathbf{Ax} - 2\mathbf{x}^T\mathbf{A}^T\mathbf{b} + \mathbf{b}^T\mathbf{b}\right) + \frac{\mathrm{d}}{\mathrm{d}\mathbf{x}}\left(\mathbf{x}^T\Gamma^T\Gamma\mathbf{x}\right)$$
$$= 2\left(\mathbf{A}^T\mathbf{Ax} - \mathbf{A}^T\mathbf{b} + \Gamma^T\Gamma\mathbf{x}\right).$$

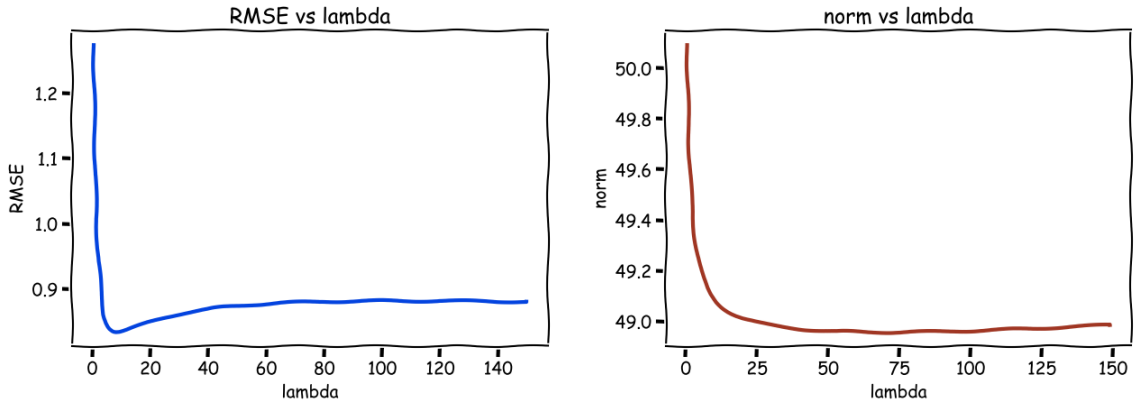
We should probably set this to zero to find a critical point (presumably a minimum since there's no bound to how poorly you can do) and solve for $\mathbf{x}$ so

$$\mathbf{A}^T\mathbf{Ax} - \mathbf{A}^T\mathbf{b} + \Gamma^T\Gamma\mathbf{x} = 0$$
$$\Rightarrow \left(\mathbf{A}^T\mathbf{A} + \Gamma^T\Gamma\right)\mathbf{x} = \mathbf{A}^T\mathbf{b}$$

And thus,

$$\mathbf{x} = \left(\mathbf{A}^T\mathbf{A} + \Gamma^T\Gamma\right)^{-1}\mathbf{A}^T\mathbf{b}.$$

(c) We plot $\lambda$ vs. the root mean square error as well as $\|\boldsymbol{\theta}\|$ below.




The optimal regularization value was $\lambda = 8.4493$ with a root mean square error of 0.8340 on the validation set and 0.8628 on the test set.

(d) Well, we're looking to minimize

$$||A\mathbf{x} + b\mathbf{1} - \mathbf{y}||_2^2 + ||\Gamma\mathbf{x}||_2^2 = \mathbf{x}^T\mathbf{A}^T\mathbf{Ax} - 2\mathbf{x}^T\mathbf{A}^T\mathbf{b} + 2\mathbf{b}^T\mathbf{Ax} - 2\mathbf{b}^T\mathbf{y} + \mathbf{b}^T\mathbf{b} + \mathbf{x}^T\Gamma^T\Gamma\mathbf{x}$$

where $\mathbf{b} = b\mathbf{1}$ so we should probably take a gradient

$$\frac{\mathrm{d}}{\mathrm{d}\mathbf{x}}\left(||A\mathbf{x} + b\mathbf{1} - \mathbf{y}||_2^2 + ||\Gamma\mathbf{x}||_2^2\right) = \frac{\mathrm{d}}{\mathrm{d}\mathbf{x}}\left(\mathbf{x}^T\mathbf{A}^T\mathbf{Ax} - 2\mathbf{x}^T\mathbf{A}^T\mathbf{b} + 2\mathbf{bAx} - 2\mathbf{b}^T\mathbf{y} + \mathbf{b}^T\mathbf{b} + \mathbf{x}^T\Gamma^T\Gamma\mathbf{x} + 2\mathbf{y}^T\mathbf{Ax} + \mathbf{y}^T\mathbf{y}\right)$$
$$= 2\left(\mathbf{A}^T\mathbf{Ax} - \mathbf{A}^T\mathbf{b} + \Gamma^T\Gamma\mathbf{x} + \mathbf{A}^T\mathbf{y}\right)$$

which we set equal to zero and solve for $\mathbf{x}$, so

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

which isn't super helpful, I suppose without $\mathbf{b}$. Sad. I guess, doing the same thing of taking the gradient of the norm terms with respect to $\mathbf{b}$—or really $b$ where $\mathbf{b} = b\mathbf{1}$—gives us

$$\frac{\mathrm{d}}{\mathrm{d}b}\left(\mathbf{x}^T\mathbf{A}^T\mathbf{A}\mathbf{x} - 2\mathbf{x}^T\mathbf{A}^T\mathbf{b} + 2\mathbf{b}^T\mathbf{A}\mathbf{x} - 2\mathbf{b}^T\mathbf{y} + \mathbf{b}^T\mathbf{b} + \mathbf{x}^T\Gamma^T\Gamma\mathbf{x}\right) = 2\left(b\mathbf{1}^T\mathbf{1} - \mathbf{1}^T\mathbf{y} + \mathbf{1}^T\mathbf{A}\mathbf{x}\right)$$

which we set to zero to find

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

which is disappointing because I need to plug it into our $\mathbf{x}$ equation and resolve for $\mathbf{x}$. FML. So

$$\mathbf{x} = (\mathbf{A}^T\mathbf{A}+\Gamma^T\Gamma)^{-1}\mathbf{A}^T\left(\frac{1}{n}\mathbf{1}^T(\mathbf{y}-\mathbf{A}\mathbf{x})\mathbf{1}-\mathbf{y}\right),$$

which can be rearranged to give

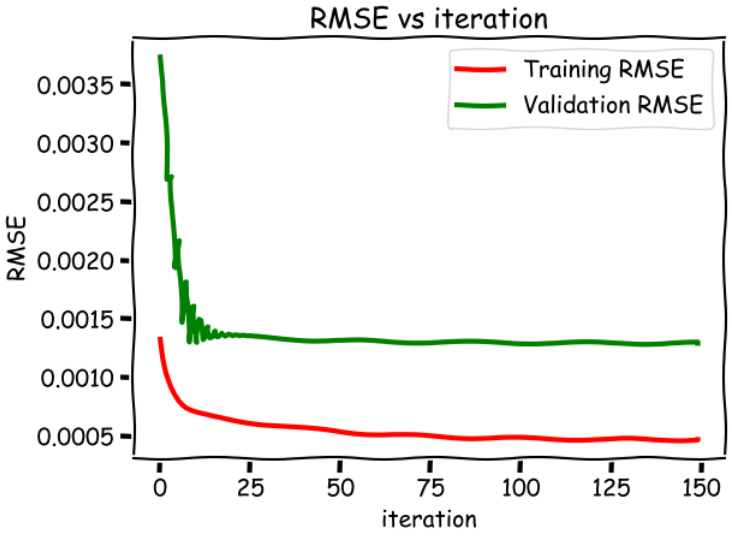
$$\begin{aligned}
(\mathbf{A}^T\mathbf{A}+\Gamma^T\Gamma)\mathbf{x} &= \mathbf{A}^T\left(\frac{1}{n}\mathbf{1}^T(\mathbf{y}-\mathbf{A}\mathbf{x})\mathbf{1}-\mathbf{y}\right) \\
&= \mathbf{A}^T\left(\frac{1}{n}\mathbf{1}\mathbf{1}^T\mathbf{y} - \frac{1}{n}\mathbf{1}\mathbf{1}^T\mathbf{A}\mathbf{x} - \mathbf{y}\right)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

which tells us that

$$\mathbf{x} = \left[\mathbf{A}^T\left(\mathbf{I}-\frac{1}{n}\mathbf{1}\mathbf{1}^T\right)\mathbf{A}+\Gamma^T\Gamma\right]^{-1}\mathbf{A}^T\left(\mathbf{I}-\frac{1}{n}\mathbf{1}\mathbf{1}^T\right)\mathbf{A}\mathbf{y}$$

which is a horribly disgusting—but analytic—answer.

(e) Here is our convergence plot:




The difference between the norms is 080378 and the difference between the biases is 0.15388.

■