HW #2.

3. (a) We know $\max_w P(w|D) = \max_w P(D|w)P(w)$

Which in log form is:

$$\arg_w \max \sum_{i=1}^{N} \log \mathcal{N}(y_i | w_0 + w^T x_i, \sigma^2) + \sum_{j=1}^{D} \log \mathcal{N}(w_j | 0, \tau^2)$$

Applying probability distribution $\mathcal{N}(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$

$$= \arg_w \max \sum_{i=1}^{N} \log \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(\frac{-(y_i - w_0 - w^T x_i)^2}{2\sigma^2}\right) + \sum_{j=1}^{D} \log \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(\frac{w_j^2}{2\tau^2}\right)$$

using property of logs, we get:

$$= \arg_w \max - \left((N+D) \log \sqrt{2\pi}\,\sigma + \sum_{i=1}^{N} \frac{(y_i - w_0 - w^T x_i)^2}{2\sigma^2} + \sum_{j=1}^{D} \left(\frac{w_j^2}{2\tau^2} \right) \right)$$

Since our constant $-(N+D)\log\sqrt{2\pi}\,\sigma$ doesn't affect or change our optimal value. So we can scale by $2\sigma^2$ without changing our optimal value.

Thus, our equivalent optimization problem is:

$$\arg_w \min \sum_{i=1}^{N} (y_i - w_0 - w^T x_i)^2 + \frac{\sigma^2}{\tau^2} \sum_{j=1}^{D} w_j^2$$

$$= \arg_w \min \sum_{i=1}^{N} (y_i - w_0 - w^T x_i)^2 + \lambda \|w\|_2^2$$

(b) We want to find gradient of f wrt x and set it to 0;

$$\nabla_x f = \nabla_x \left[(Ax-b)^T (Ax-b) + (\Gamma x)^T (\Gamma x) \right]$$

$$= \nabla_x \left[x^T A^T A x - 2x^T A^T b + b^T b + x^T \Gamma^T \Gamma x \right]$$

$$= 2A^T A x - 2A^T b + 2\Gamma^T \Gamma x = 0$$

$$(A^T A + \Gamma^T \Gamma) x = A^T b$$

$$\therefore \; x^* = (A^T A + \Gamma^T \Gamma)^{-1} A^T b$$

(c) Check images $\lambda^* = 8.7418$, Validation set RMSE $= 0.8340$, test set RMSE $= 0.8628$

(d) If we expand f, we get

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

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

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

At optimality, we have $\nabla_x f = 0$

So,

$$\nabla_x f = 2A^T A x + 2bA^T \mathbf{1} - 2A^T y + 2\Gamma^T \Gamma x = 0 \quad \ldots (*)$$

$$\nabla_b f = 2\mathbf{1}^T A x - 2\mathbf{1}^T y + 2bn = 0$$

$$\therefore \; b^* = \frac{\mathbf{1}^T (y - Ax)}{n}$$

On plugging b^* in $(*)$, we get:

$$(A^T A + \Gamma^T\Gamma)x + \left(\frac{\mathbf{1}^T(y - Ax)}{n}\right) A^T\mathbf{1} - A^T y = 0$$

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

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

$y \in \mathbb{R}^n$

From code:

Difference in bias = 4.2822E-10

weights = 5.6328E-10

(e) Check image

Difference in bias = 1.5387E-01

weights = 7.9860E-01