Henry Trinh

1. Gradient + Hessian of log-likelihood regression:

a. Let $\sigma(x) = \frac{1}{1+e^{-x}}$ for a sigmoid function. Show that $\sigma'(x) = \sigma(x)[1-\sigma($

We start with $\sigma(x) = (1+e^{-x})^{-1}$. Then, the derivative is $\sigma'(x) = e^{-x}(1+e^{-x})^{-2}$. By substitution of $\sigma(x)$, we see that $\sigma'(x) = \sigma(x)e^{-x}(1+e^{-x})^{-1}$, or $\sigma'(x) = \sigma(x)\frac{e^{-x}}{1+e^{-x}}$. By further simplification, $\sigma'(x) = \sigma(x)[e^{-x}+1]$. Then, $\sigma'(x) = \sigma(x)\sigma^{-1}(x)$. Hence, we see that $\sigma'(x) = \sigma(x)[1-\sigma(x)]$, as desired.

b. Derive expression for the gradient of the log-likelihood for logistic regression.

Per the provided hint, we use the negative log-likelihood of logistic regression, which Murphey defines as follows:

$$NLL(w) = -\sum_{i=1}^{n}[y_i \cdot \log u_i + (1-y_i)\log(1-u_i)] \quad \text{(page 246)}$$

Let $\mu_i = \sigma(\theta^T x_i)$ to apply the sigmoid function to the linear combination, $\theta^T x_i$, where x_i represents the transpose of row i of Matrix X. By substitution, we see that:

$$NLL(\theta) = -\sum_{i=1}^{n} y_i \log\sigma(\theta^T x_i) + (1-y_i)\log(1-\sigma(\theta^T x_i))$$

Then, we take the gradient w.r.t. θ: {requires chain rule}

$$\nabla_\theta NLL(\theta) = -\sum_{i=1}^{N} y_i \cdot \frac{1}{\sigma(\theta^T x_i)}\sigma'(\theta^T x_i) + (1-y_i)\frac{1}{1-\sigma(\theta^T x_i)}[-\sigma'(\theta^T x_i)]$$

* Using derivation from Part (a)

$$= -\sum_{i=1}^{N} y_i \cdot \frac{\sigma(\theta^T x_i)[1-\sigma(\theta^T x_i)]}{\sigma(\theta^T x_i)} + (1-y_i)\left[\frac{-\sigma(\theta^T x_i)[1-\sigma(\theta^T x_i)]}{1-\sigma(\theta^T x_i)}\right]$$

$$= -\sum_{i=1}^{N} y_i(1-\sigma(\theta^T x_i))x_i - (1-y_i)(\sigma(\theta^T x_i))x_i$$

Then, by expansion

$$= -\sum_{i=1}^{N} y_i x_i - \underline{y_i\sigma(\theta^T x_i)x_i} - \sigma(\theta^T x_i)x_i + \underline{y_i\sigma(\theta^T x_i)x_i}$$

$$= -\sum_{i=1}^{N} x_i(y_i - \sigma(\theta^T x_i)) = -\sum_{i=1}^{N} x_i(y_i - u_i)$$

$$= \boxed{X^T(u-y)}$$

c. Hessian is written as $H = X^T S X$, where $S = \mathrm{diag}(\mu_1(1-\mu_1), \ldots, \mu_n(1-\mu_n))$. Derive and show that $H \succeq 0$ {H is positive semi-definite}

To compute the Hessian, we compute the second derivative, which is simple as we have computed the gradient, $\nabla_\theta NLL(\theta)$ in Part (b):

Hence, $H_\theta = \nabla_\theta [\nabla_\theta NLL(\theta)]^T = \nabla_\theta [X^T(\mu - y)]^T = \nabla_\theta(\mu^T X - y^T X)$

$= \nabla_\theta \mu^T X = \nabla_\theta \sigma(X\theta)^T X = X^T \cdot \mathrm{diag}(\mu(1-\mu)) X$

$= \boxed{X^T S X}$, as desired. Now, $H_\theta \succeq 0$ if $S \succeq 0$:

We then need to show that $\mu_i(1-\mu_i) = \sigma(\theta^T x_i)(1-\sigma(\theta^T x_i)) \geq 0$ to show that H is positive semi-definite. Since the sigmoid function must be between 0 and 1 for logistic regression, we thus see that $\sigma(1-\sigma) \geq 0$. $\therefore \boxed{H \succeq 0}$

2. Derive the normalization constant, Z, for a one-dimensional zero-mean Gaussian: $P(X;\sigma^2) = \frac{1}{Z}\exp\left(\frac{x^2}{2\sigma^2}\right)$, such that $P(x;\sigma^2)$ is a valid density.

Hence, $Z = \int_a^b \exp\left(\frac{-x^2}{2\sigma^2}\right)dx$, where $a = -\infty$ and $b = \infty$, as the integral of $P(X,\sigma^2)$ is 1. To compute Z, we consider its square,

$$\Rightarrow Z^2 = \int_a^b \int_a^b \exp\left(-\frac{x^2+y^2}{2\sigma^2}\right)dx\,dy$$

Now, using polar coordinates, let $x = r\cos\theta$, $y = r\sin\theta$, $dx\,dy = r\,dr\,d\theta$, and since $\sin^2\theta + \cos^2\theta = 1$, we see that:

$$Z^2 = \int_0^{2\pi}\int_0^{\infty} x\exp\left(\frac{-x^2}{2\sigma^2}\right)dr\,d\phi, \text{ where } \phi = \arctan\left(\frac{\sin\theta}{\cos\theta}\right)$$

$$= \int_0^{2\pi}\left[-\sigma^2\exp\left(\frac{-x^2}{2\sigma^2}\right)x\,dr\,d\theta\right]_0^{\infty} d\phi$$

$$= \int_0^{2\pi}\left[(-\sigma^2\cdot 0) - (-\sigma^2 \cdot 1)\right]d\phi$$

$$= \int_0^{2\pi}\sigma^2\,d\phi \Rightarrow \sigma^2 2\pi$$

$$\therefore\ Z = \sqrt{2\pi\sigma^2} = \boxed{\sigma\sqrt{2\pi}}$$

3 a. Show that the maximum a posteriori problem for linear regression with a zero-mean Gaussian prior $P(w) = \prod_j N(w_j | \emptyset, \tau^2)$ on the weights:

$$\underset{w}{\operatorname{argmax}} \sum_{i=1}^{N} \log N(y_i | w_0 + w^T x_i, \sigma^2) + \sum_{j=1}^{D} \log N(w_j | \emptyset, \tau^2) \quad (1)$$

is equivalent to the ridge regression problem

$$\operatorname{argmin} \frac{1}{N} \sum_{i=1}^{N} (y_i - w_0 + w^T x_i))^2 + \lambda \|w\|_2^2, \text{ where } \lambda = \frac{\sigma^2}{\tau^2}$$

We first apply the Gaussian Distribution $N(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$, thereby yielding the following result from (1):

$$\underset{w}{\operatorname{argmax}} \left\{ \sum_{i=1}^{N} \log\left[\frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(y_i - w_0 - w^T x_i)^2}{2\sigma^2}\right)\right] + \sum_{j=1}^{D} \log\left[\frac{1}{\sqrt{2\pi}\tau} \exp\left(\frac{-w_j^2}{2\tau^2}\right)\right] \right\} \quad (2)$$

By the power rule of logarithms, we simplify (2):

$$\underset{w}{\operatorname{argmax}} \left\{ \sum_{i=1}^{N} \left[-\frac{(y_i - w_0 - w^T x_i)^2}{2\sigma^2} - \log\sqrt{2\pi}\sigma\right] + \sum_{j=1}^{D} \left[\frac{-w_j^2}{2\tau^2} - \log\sqrt{2\pi}\tau\right] \right\} \quad (3)$$

Hence, by simplification, our objective becomes:

$$\underset{w}{\operatorname{argmax}} \left\{ -(N+D)\log\sqrt{2\pi}\sigma + \sum_{i=1}^{N} \left[\frac{-(y_i - w_0 - w^T x_i)^2}{2\sigma^2}\right] + \sum_{j=1}^{D} \left[\frac{-w_j^2}{2\tau^2}\right] \right\}, \quad (4)$$

where $-(N-D)\log\sqrt{2\pi}\sigma$ does not influence the value of w^* that maximizes our expression. Then, we can neglect the constant and scale the problem by $2\sigma^2$ without affecting the optimal solution, w^*:

$$\text{Hence, } \underset{w}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} (y_i - w_0 - w^T x_i)^2 + \frac{\sigma^2}{\tau^2} \sum_{j=1}^{D} w_j^2 \right. \quad (5)$$

and, by substitution of $\lambda = \frac{\sigma^2}{\tau^2}$, we get:

$$\boxed{\underset{w}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} (y_i - w_0 - w^T x_i)^2 + \lambda \|w\|_2^2 \right\}} \quad (6)$$

//

b. Find the closed form solution x^* to the ridge regression problem

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

To do so, we calculate the gradient of f w.r.t x, and set it to $\emptyset$:

$$\begin{aligned}
\nabla_x f &= \nabla_x ((Ax - b)^T (Ax - b) + (\Gamma x)^T (\Gamma x)) \\
&= \nabla_x ((x^T A^T - b^T)(Ax - b) + x^T \Gamma^T \Gamma x \\
&= \nabla_x (x^T A^T A x - 2x^T A^T b + b^T b + x^T \Gamma^T \Gamma x) \\
&= 2A^T A x - 2A^T b + 2\Gamma^T \Gamma x
\end{aligned}$$

Then, let $\nabla_x f = \emptyset$: $(A^T A + \Gamma^T \Gamma)x = A^T b$.

Hence, the closed-form solution is: $\boxed{x^* = (A^T A + \Gamma^T \Gamma)^{-1} A^T b}$

For simplification, let $\Gamma = \sqrt{\lambda} I$, such that our objective for ridge regression is $f = \|Ax - b\|^2 + \lambda x^T x$.

$$\therefore \boxed{x^* = (A^T A + \lambda I)^{-1} A^T b}$$

//

c. See hw2pr3.py

```
==> The optimal regularization parameter is  8.9527.
==> The RMSE on the validation set with the optimal regularization parameter is  0.8341.
==> The RMSE on the test set with the optimal regularization parameter is  0.8628.
```

d. Instead of computing $\hat{y} = \theta^T x$ with $x_0 = 1$, compute $\hat{y} = \theta^T x + b$.
Hence, solve the following optimization problem:

$$\text{minimize}: \|Ax + b\mathbb{1} - y\|_2^2 + \|\Gamma x\|_2^2 \quad (1)$$

Solve for x^* explicitly, using the closed-form to compute the bias term.

Expanding the objective function, we get:

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

We then find the gradient of f and set it to $\emptyset$:

$$\nabla_x f = 2A^TAx + 2bA^T\mathbb{1} - 2A^Ty + 2\Gamma^T\Gamma x = \emptyset \quad (2)$$

$$\therefore \nabla_b f = 2\cdot\vec{\mathbb{1}}^T Ax - 2\cdot\vec{\mathbb{1}}^T y + 2bn = \emptyset$$

Then, we solve for b^*:

$$b^* = \frac{\vec{\mathbb{1}}^T(y - Ax)}{n} \quad (3)$$

In other words, when the model predicts a flat line where $x = \emptyset$, the optimal bias term, b^*, is the average of the outputs, y, as expected.
Now, we substitute (3) into (2) to solve for x^*, resulting in:

$$\begin{aligned} \Rightarrow \emptyset &= (A^TA + \Gamma^T\Gamma)\vec{x} + \left(\frac{\vec{\mathbb{1}}^T(\vec{y} - A\vec{x})}{n}\right)A^T\vec{\mathbb{1}} - A^T\vec{y} \\ &= (A^TA + \Gamma^T\Gamma)\vec{x} + \tfrac{1}{n}A^T\vec{\mathbb{1}}\vec{\mathbb{1}}^T\vec{y} - \tfrac{1}{n}A^T\vec{\mathbb{1}}\vec{\mathbb{1}}^TA\vec{x} - A^T\vec{y} \\ &= A^TA + \Gamma^T\Gamma - \tfrac{1}{n}A^T\vec{\mathbb{1}}\vec{\mathbb{1}}^TA\vec{x} \\ &= A^T\vec{y} - \tfrac{1}{n}A^T\vec{\mathbb{1}}\vec{\mathbb{1}}^T\vec{y} \\ &= A^T(I - \tfrac{1}{n}\vec{\mathbb{1}}\vec{\mathbb{1}}^T)A + \Gamma^T\Gamma\vec{x} \quad \{I = \text{identity matrix}\} \\ &= A^T(I - \tfrac{1}{n}\vec{\mathbb{1}}\vec{\mathbb{1}}^T)\vec{y} \end{aligned}$$

$\therefore$ The closed-form solution for x^* becomes:

$$\Rightarrow \boxed{x^* = \left[A^T(I - \tfrac{1}{n}\vec{\mathbb{1}}\vec{\mathbb{1}}^T)A + \Gamma^T\Gamma\right]^{-1}A^T(I - \tfrac{1}{n}\vec{\mathbb{1}}\vec{\mathbb{1}}^T)y} \;//$$

The difference in bias is 2.1643E-11.
weights is 2.7774E-11 } near-negligible errors!

e. See hw2pr3.py

```
==> Difference in bias is  1.5386E-01
==> Difference in weights is  7.9626E-01
```