I was aided by the solutions guide

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

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

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

$$= \frac{e^{-x}}{1+e^{-x}}\left(\frac{1}{1+e^{-x}}\right)$$

$$= \frac{e^{-x}}{(1+e^{-x})^2} = \frac{e^{-x}}{(1+e^{-x})}\left(\frac{1}{1+e^{-x}}\right)$$

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

$$= \sigma(x)\,\frac{1+e^{-x}-1}{(1+e^{-x})}$$

$$= \sigma(x) \quad -\frac{1}{1+e^{-x}}$$

$$= \sigma(x)\,[1-\sigma(x)]$$

as desired.

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

Given negative log likelihood for logistic regression is:

$$nll(\theta) = -\sum_i y_i \log\sigma(\theta^T x_i) + (1-y_i)\log(1-\sigma(\theta^T x_i))$$

$$\nabla_\theta nll(\theta) = -\sum_i y_i \frac{1}{\sigma(\theta^T x_i)}\sigma'(\theta^T x_i) + (1-y_i)\frac{1}{1-\sigma(\theta^T x_i)}(-\sigma'(\theta^T x_i))$$

$$= -\sum_i y_i(1-\sigma(\theta^T x_i))x_i - (1-y_i)\sigma(\theta^T x_i)x_i$$

$$= -\sum_i y_i x_i - y_i\sigma(\theta^T x_i)x_i - \sigma(\theta^T x_i)x_i + y_i\sigma(\theta^T x_i)x_i$$

$$= \sum_i(\sigma(\theta^T x_i) - y_i)x_i$$

$$= \sum_i(\mu_i - y_i)x_i$$

$$= X^T(\mu - y)$$

c. $H = X^T S X$

$S = \mathrm{diag}(\mu_1(1-\mu_1), \dots, \mu_n(1-\mu_n))$

Derive & show $H \geq 0$

$$H_\theta = \nabla_\theta(\nabla_\theta nll(\theta))^T$$

$$= \nabla_\theta[X^T(\mu-y)]^T$$

$$= \nabla_\theta(X\mu^T - \underbrace{X y^T}_{=0})$$

from part b

$$= \nabla_\theta(X\mu^T) = \nabla_\theta\,\sigma(X\theta^T)X$$

$$= X^T \mathrm{diag}(\mu(1-\mu))X$$

$$= X^T S X \quad \text{as desired.}$$

Showing that H_θ is positive semi-definite is = showing $S = \mathrm{diag}(\mu(1-\mu))$ is positive semi-definite.

We just need to show

$$\mu_i(1-\mu_i) = \sigma(\theta^T x_i)(1-\sigma(\theta^T x_i)) \geq 0$$

$0 < \sigma(\cdot) < 1$ implies

$$\sigma(\cdot)(1-\sigma(\cdot)) \geq 0.$$

$\therefore$ H is positive semi-definite

2. $P(x;\sigma^2) = \frac{1}{z} e\left(\frac{-x^2}{2\sigma^2}\right)$

$$= \frac{1}{z} \int e^{\left(\frac{-x^2}{2\sigma^2}\right)} dx = 1$$

$$Z = \int e^{\left(\frac{-x^2}{2\sigma^2}\right)} dx = 1$$

$$Z^2 = \int e^{\left(\frac{-x^2}{2\sigma^2}\right)} dx \int e^{\left(\frac{-y^2}{2\sigma^2}\right)} dy$$

$$= \iint e^{\left(\frac{-x^2}{2\sigma^2}\right)} e^{\frac{-y^2}{2\sigma^2}} dx dy$$

$$= \iint_{\mathbb{R}^2} e^{\left(-\frac{x^2+y^2}{2\sigma^2}\right)} dx dy$$

$$= \int_0^{\infty} \int_0^{2\pi} e^{\left(-\frac{r^2}{2\sigma^2}\right)} r \, d\theta \, dr$$

$$= 2\pi \int_0^{\infty} e^{\left(\frac{-r^2}{2\sigma^2}\right)} r \, dr$$

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

$$= -2\pi\sigma^2 e\left(\frac{-r^2}{2\sigma^2}\right)\Bigg|_0^{\infty}$$

$$= -2\pi\sigma^2 (0 - 1)$$

$$Z^2 = 2\pi\sigma^2$$

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

3a. Show

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

is equivalent to the ridge regression problem

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

w/ $\lambda = \sigma^2 / T^2$

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

So,

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

$$= \arg\max_w \sum_{i=1}^{N} \left(-\frac{(y_i - w_0 - w^T x_i)^2}{2\sigma^2} - \log \sqrt{2\pi}\sigma \right)$$

$$+ \sum_{j=1}^{D} \left(-\frac{w_j^2}{2T^2} - \log \sqrt{2\pi}\sigma \right)$$

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

Maxing function equivalent to minimizing negative

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

$\lambda = \sigma^2 / T^2$

so $\arg\min_w \sum_{i=1}^{N} (y_i - w_0 - w^T x_i)^2 + \lambda \sum_{j=1}^{D} w_j^2$

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

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

sum of squared distance between values predicted by Ax & observed b

$$\min: \|Ax-b\|_2^2 + \|\Gamma x\|_2^2$$

penalizes large values in parameter vector x for avoid over fitting

$$J(x) = \|Ax-b\|_2^2 + \|\Gamma x\|_2^2$$

$$= (Ax-b)^T (Ax-b) + (\Gamma x)^T (\Gamma x)$$

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

$$\nabla_x [(Ax-b)^T (Ax-b)] = 2A^T(Ax-b)$$

$$\nabla_x [(\Gamma x)^T (\Gamma x)] = 2\Gamma^T \Gamma x$$

Combining, we get

$$\nabla_x J(x) = 2A^T(Ax-b) + 2\Gamma^T\Gamma x$$

$$2A^T Ax + 2\Gamma^T \Gamma x = 2A^T b$$

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

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

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

d. $\min f = \|Ax - b1 - y\|_2^2 + \|\Gamma x\|_2^2$

$= (Ax + b1 - y)^T (Ax + b1 - y) + (\Gamma x)^T (\Gamma x)$

$= (x^T A^T + b1^T - y^T)(Ax + b1 - y) + x^T \Gamma^T \Gamma x$

$= x^T A^T A x + 2b1^T Ax - 2y^T Ax - 2b1^T y + b^2 n + y^T y + x^T \Gamma^T \Gamma x$

$\nabla_x f = 2A^T Ax + 2bA^T 1 - 2A^T y + 2\Gamma^T \Gamma x = 0$

$\nabla_b f = 21^T Ax - 21^T y + 2bn = 0$

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

Plugging back in:

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

$(A^T A + \Gamma^T \Gamma)x + \frac{1}{n} A^T 11^T y - \frac{1}{n} A^T 11^T Ax - A^T y = 0$

$\left[A^T A + \Gamma^T \Gamma - \frac{1}{n} A^T 11^T A\right] x = A^T y - \frac{1}{n} A^T 11^T y$

$= A^T \left(I - \frac{1}{n} 11^T\right) y$

$$x^* = \left[A^T \left(I - \frac{1}{n} 11^T\right) A + \Gamma^T \Gamma\right]^{-1} A^T \left(I - \frac{1}{n} 11^T\right) y$$