Amelia Zhao
Math 189 HW#2

1.) a.) $\sigma_x = \frac{1}{1+e^{-x}}$

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

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

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

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

b.) $NLL(\theta) = -\sum_{i=1}^{N}[y_i \log \mu_i + (1-y_i)\log(1-\mu_i)]$ 1.9 $\mu(x) = \sigma(\theta^T \vec{x})$

$= -\sum_{i=1}^{N}[y_i \log(\sigma(\theta^T \vec{x}_i)) + (1-y_i)\log(1-\sigma(\theta^T \vec{x}_i))]$

$\nabla_\theta NLL(\theta) = -\sum_{i=1}^{N} y_i \frac{1}{\sigma\theta^T \vec{x}_i}(\nabla_\theta \sigma)(\theta^T \vec{x}_i) + (1-y_i)\frac{1}{1-\sigma(\theta^T \vec{x}_i)}(\nabla_\theta \sigma)(\theta^T \vec{x}_i)$

$= -\sum_{i=1}^{N} y_i \underbrace{\frac{1}{\sigma\theta^T \vec{x}_i}(\sigma(1-\sigma))\theta^T \vec{x}_i(\vec{x}_i)}_{(1-\sigma(\theta^T \vec{x}_i))} + (1-y_i)\underbrace{\frac{1}{1-\sigma\theta^T \vec{x}_i}(\sigma(1-\sigma\ (\theta^T \vec{x}_i))(-\vec{x}_i)}_{\sigma(\theta^T \vec{x}_i)}$

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

$= -\sum_{i=1}^{N} y_i \vec{x}_i - \cancel{y_i \sigma(\theta^T \vec{x}_i)\vec{x}_i} - \sigma(\theta^T \vec{x}_i)\vec{x}_i + \cancel{y_i \sigma(\theta^T \vec{x}_i)\vec{x}_i}$

$= -\sum_{i=1}^{N}(y_i - \underbrace{\sigma(\theta^T \vec{x}_i)}_{\mu_i})\vec{x}_i = \sum_{i=1}^{N}(\mu_i - y_i)\vec{x}_i$

$= \boxed{X^T(\vec{\mu} - \vec{y})}$ This can't be set to zero to find critical pt (MLE)

c.) Hessian → Hessian = full matrix of 2nd partial derivatives s.t. $(H_f)_{i,j} = \frac{\partial^2 f}{\partial x_i \partial x_j}$

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

$= \sum_{i=1}^{N} \nabla_\theta(x_i^T(\vec{\mu}_i - \vec{y}_i))^T = \sum_{i=1}^{N} \nabla_\theta((\vec{\mu}_i - \vec{y}_i)^T \vec{x}_i) = \sum_{i=1}^{N}(\nabla_\theta(\vec{\mu}_i - \vec{y}_i)^T \vec{x}_i)$

$= \sum_{i=1}^{N}(\nabla_\theta \sigma(\theta^T \vec{x}_i)^T)\vec{x}_i = \sum_{i=1}^{N}(\nabla_\theta \sigma(\vec{x}_i^T \theta))\vec{x}_i$

$= \sum_{i=1}^{N} \vec{x}_i^T \sigma(\theta^T \vec{x})(1-\sigma(\theta^T \vec{x}))\vec{x}_i = X^T \text{diag}(\mu(1-\mu))X$

$= \boxed{X^T S X}$ $S \equiv \text{diag}(\mu(1-\mu))$

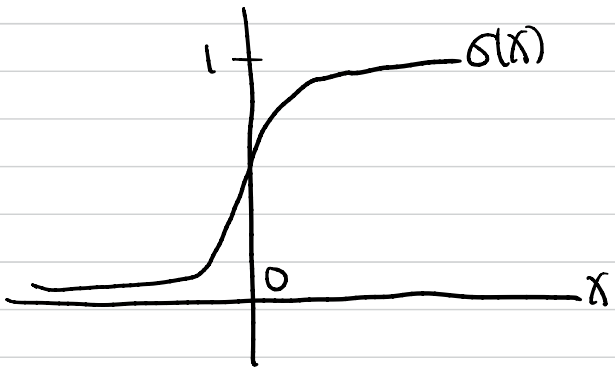
quadratic form of summation uses diagonal matrix by definition

H_θ is positive semidefinite if $x^T H_\theta x \geq 0 \quad \forall x \in \mathbb{R}^n$

$x^T H_\theta x = x^T X^T S X x = x^T X^T S (Xx) = x^T X^T S^{1/2} S^{1/2} (Xx) = \|S^{1/2} X x\|_2^2$

Since $\|Xx\|_2^2 \geq 0$, we now check if $\|S^{1/2}\|_2^2 \geq 0$. This is true if all values of $\mu_i(1-\mu_i) \geq 0$ because S is a diagonal matrix.
(positive semidefinite if matrix is Hermitian + all eigenvalues are real & non-negative)

$\mu_i(1-\mu_i) = \sigma(\theta^T \vec{x}_i)(1-\sigma(\theta^T \vec{x}_i))$



$\forall x, \; 0 < \sigma(x) < 1.$
Thus, μ_i must be between 0 & 1.
So, H_θ must be positive semidefinite.

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

$\int_{-\infty}^{\infty} \mathbb{P}(x;\sigma^2)\,dx = 1$

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

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

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

switch to polar coordinates $x = r\cos\theta$, $y = r\sin\theta$

$Z^2 = \int_{\theta=0}^{2\pi}\int_{r=0}^{\infty} \exp\left(\frac{-r^2}{2\sigma^2}\right) r\,dr\,d\theta$

$u = e^{-r^2/2\sigma^2}$

$du = e^{-r^2/2\sigma^2}\left(\frac{-1}{2\sigma^2}(2r)\right)dr$

$= 2\pi\int_{r=0}^{\infty} \exp\left(\frac{-r^2}{2\sigma^2}\right) r\,dr$

$= 2\pi\left(-\sigma^2 \exp\frac{-r^2}{2\sigma^2}\right)\Big|_{r=0}^{\infty}$

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

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

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

3.) a.) $P(w) = \prod_j N(\omega_j|0,\tau^2)$

$$\underset{W}{\text{argmax}} \underbrace{\sum_{i=1}^{N} \log N(y_i|\omega_0 + W^T x_i, \sigma^2)}_{x_1,\ldots,x_N} + \underbrace{\sum_{j=1}^{D} \log N(\omega_j|0,\tau^2)}_{\text{Data}}$$

$$\theta_{MAP} = \underset{\theta}{\text{argmax}}\, p(\theta|D) = \text{argmax}\, \frac{p(D|\theta)\,p(\theta)}{\underbrace{p(D)}_{\text{ignore w/ argmax}}}$$

$$= \text{argmax}\,(\log p(D|\theta) + \log p(\theta))$$

$$= \underset{W}{\text{argmax}} \sum_{i=1}^{N} \log\left(\frac{1}{\sqrt{2\pi}\sigma}\right)\exp\left(\frac{-(y_i-\omega_0+W^Tx_i)^2}{2\sigma^2}\right) + \sum_{j=1}^{D} \log\frac{1}{\sqrt{2\pi}\tau}\exp\left(\frac{-\omega_j^2}{2\tau^2}\right)$$

$$= \underset{W}{\text{argmax}} \sum_{i=1}^{N} \cancel{\log\left(\frac{1}{\sqrt{2\pi}\sigma}\right)} + \left(\frac{-(y_i-\omega_0+W^Tx_i)^2}{2\sigma^2}\right) + \sum_{j=1}^{D} \cancel{\log\left(\frac{1}{\sqrt{2\pi}\tau}\right)} + \frac{-\omega_j^2}{2\tau^2}$$

remove constants from argmax

$$= \underset{W}{\text{argmax}} \sum_{i=1}^{N} \frac{1}{2\sigma^2}(-(y_i-\omega_0+W^Tx_i)^2) + \sum_{j=1}^{D} \frac{1}{2\tau^2}(-\omega_j)^2$$

scale by $2\sigma^2$

$$= \underset{W}{\text{argmax}} \sum_{i=1}^{N} -(y_i-\omega_0+W^Tx_i)^2 + \sum_{j=1}^{D} \frac{\sigma^2}{\tau^2}(-\omega_j)^2$$

this expression is equivalent to argmin of its negative:

$$= \underset{W}{\text{argmin}} \sum_{i=1}^{N} (y_i-\omega_0+W^Tx_i)^2 + \frac{\sigma^2}{\tau^2}\underbrace{\sum_{j=1}^{D}\omega_j^2}_{\|W\|_2^2}$$

Define $\frac{\sigma^2}{\tau^2} \equiv \lambda$

$$= \underset{W}{\text{argmin}} \sum_{i=1}^{N} (y_i-\omega_0+W^Tx_i)^2 + \lambda\|W\|_2^2$$

We can scale by a factor of $\frac{1}{N}$ (will not change argmin)

$$= \boxed{\underset{W}{\text{argmin}} \frac{1}{N}\sum_{i=1}^{N} (y_i-\omega_0+W^Tx_i)^2 + \lambda\|W\|_2^2}$$

↓ turns it into average squared error

b.) minimize: $\|Ax-b\|_2^2 + \|\Gamma x\|_2^2$

$\nabla_x \|Ax-b\|_2^2 + \|\Gamma x\|_2^2 = 0$

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

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

$= \nabla_x [x^TA^TAx - x^TA^Tb - \underbrace{b^TAx}_{(x^TA^Tb)^T} + b^Tb + x^T\Gamma^T\Gamma x]$

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

$(2A^TA + 2\Gamma^T\Gamma)x = 2A^Tb$

$$\boxed{x = (A^TA + \Gamma^T\Gamma)^{-1}A^Tb}$$

d.) minimize: $\|Ax + b1 - y\|_2^2 + \|\Gamma x\|_2^2$ $\quad 1 =$ ones matrix

$\nabla_x [(Ax+b1-y)^T(Ax+b1-y) + (\Gamma x)^T(\Gamma x)]$

$= \nabla_x [(x^TA^T + b1^T - y^T)(Ax+b1-y) + (x^T\Gamma^T\Gamma x)]$

$= \nabla_x [x^T\cancel{A^TA}x + b1^TAx - y^TAx + x^TA^Tb1 + \underbrace{b^2n}_{b1^Tb1} - y^Tb1$

$- x^TA^Ty - b1^Ty + y^Ty + x^T\Gamma^T\Gamma x]$

$= \nabla_x [x^TA^TAx + 2b1^TAx - 2y^TAx - 2b1^Ty + b^2n + y^Ty + x^T\Gamma^T\Gamma x]$

$= 2A^TAx + 2b1^TA - 2y^TA + 2\Gamma^T\Gamma x = 0$

Also find ∇_b and set equal to 0 to solve for optimal b.

$\nabla_b[f] = 21^TAx - 21^Ty + 2bn = 0$

$b = \dfrac{1^TAx - 1^Ty}{-n} = \boxed{\dfrac{-1^T(Ax-y)}{n}}$

solve for x:

$(2A^TA + 2\Gamma^T\Gamma)x + 2\left(\dfrac{-1^T(Ax-y)}{n}\right)1^TA - 2y^TA = 0$

$(A^TA + \Gamma^T\Gamma - \underbrace{\frac{1}{n}1^TA(1^TA)}_{\frac{1}{n}1^TA^T1A})x = -\dfrac{1^Ty}{n}(1^TA) + y^TA$

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