1. $J(\theta) = -\frac{1}{m}\sum_{i=1}^{m} \log(h_\theta(y^i x^i))$

$$\frac{\partial}{\partial\theta_j} J(\theta) = -\frac{1}{m}\sum_{i=1}^{m} \frac{1}{\cancel{h_\theta(y^i x^i)}} \cdot \cancel{h_\theta(y^i x^i)}(1-h_\theta(y^i x^i)) \cdot y^i x_j^i$$

$$= -\frac{1}{m}\sum_{i=1}^{m} (1-h_\theta(y^i x^i)) \cdot y^i x_j^i$$

$$\therefore H_{j,k} = \frac{\partial^2}{\partial\theta_j \partial\theta_k} J(\theta) = \frac{1}{m}\sum_{i=1}^{m} h_\theta(y^i x^i) \cdot (1-h_\theta(y^i x^i)) \cdot y^i x_k^i \cdot y^i x_j^i$$

$$\because \sum_i \sum_j z_i x_i x_j z_j = \sum_i z_i x_i \cdot \sum_j x_j z_j = (x^T z)(x^T z) = (x^T z)^2$$

$$z^T H z = \sum_k \sum_j \frac{1}{m}\sum_{i=1}^{m} h_\theta(y^i x^i) \cdot (1-h_\theta(y^i x^i)) \cdot y^i x_k^i \cdot y^i x_j^i \cdot z_j \cdot z_k$$

$$= \frac{1}{m}\sum_{i=1}^{m} h_\theta(y^i x^i) \cdot (1-h_\theta(y^i x^i)) \cdot \sum_j (y^i x_j^i z_j) \cdot \sum_k (y^i x_k^i z_k)$$

$$= \frac{1}{m}\sum_{i=1}^{m} h_\theta(y^i x^i) \cdot (1-h_\theta(y^i x^i)) \cdot (z^T y^i x^i)^2$$

$\because h_\theta(y^i x^i) \in (0,1)$, $(1-h_\theta(y^i x^i)) \in (0,1)$, $(z^T y^i x^i)^2 \geq 0$.

$\therefore z^T H z \geq 0 \Rightarrow H \geq 0$

2.

(a) $p(y;\lambda) = \dfrac{e^{-\lambda}\lambda^y}{y!}$ (Poisson Distribution)

Exponential family: $p(y;\eta) = b(y) \cdot \exp(\eta^T \cdot T(y) - a(\eta))$

$b(y) = \dfrac{1}{y!}$

$e^{-\lambda}\lambda^y = e^{-\lambda} \cdot e^{\ln \lambda^y} = e^{\ln\lambda^y - \lambda}$

$\therefore \eta^T \cdot T(y) - a(\eta) = \ln \lambda^y - \lambda = y\ln\lambda - \lambda$.

$\therefore T(y) = y \quad a(\eta) = \lambda = e^\eta$

$\eta = \ln\lambda$

(b) $g(\eta) = \frac{\partial}{\partial \eta} a(\eta) = \frac{\partial}{\partial \eta} e^{\eta} = e^{\eta}$

or $g(\eta) = E(y;\eta) = \lambda = e^{\eta}$

(c) $\frac{\partial}{\partial \theta_j} \log p(y^{(i)} | x^{(i)}; \theta) = \frac{\partial}{\partial \theta_j} \eta^T y - e^{\eta}$

$\eta = \theta^T x$

$= x_j^{(i)} y^{(i)} - e^{\theta^T x^{(i)}} \cdot x_j^{(i)} = x_j^{(i)} (y^{(i)} - e^{\theta^T x^{(i)}})$

stochastic gradient descent: $\theta_j := \theta_j + \alpha \cdot x_j^{(i)} (y^{(i)} - e^{\theta^T x^{(i)}})$.

(d) $p(y;\eta) = b(y) \cdot \exp(\eta^T T(y) - a(\eta))$

$= b(y) \exp(\eta^T y - a(\eta))$.

$\frac{\partial}{\partial \theta_i} \log p(\vec{y} | X; \theta) = \frac{\partial}{\partial \theta_i} (\eta^T y^i - a(\eta))$

$$= x^i y^i - h(x) \cdot x^i$$

$$= x^i (y^i - h(x))$$

stochastic gradient descent: $\theta_i := \theta_i - \alpha(h(x) - y^{(i)}) x^{(i)}$

3.

(a) $P(y=1 \mid x; \phi, \Sigma, \mu_{-1}, \mu_1) = \dfrac{P(x \mid y=1) \cdot P(y=1)}{P(x)}$

$$= \frac{\cancel{\frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}}} \exp\left(-\frac{1}{2}(x-\mu_1)^T \Sigma^{-1} (x-\mu_1)\right) \cdot \phi}{\cancel{\frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}}} \left[\phi \exp\left(-\frac{1}{2}(x-\mu_1)^T \Sigma^{-1} (x-\mu_1)\right) + (1-\phi) \exp\left(-\frac{1}{2}(x-\mu_{-1})^T \Sigma^{-1} (x-\mu_{-1})\right)\right.}$$

$$= \frac{1}{1 + \frac{1-\phi}{\phi} \cdot \exp\left(-\frac{1}{2}(x-\mu_{-1})^T \Sigma^{-1} (x-\mu_{-1}) + \frac{1}{2}(x-\mu_1)^T \Sigma^{-1} (x-\mu_1)\right)}$$

$$= \frac{1}{1+\exp\left(\ln\left(\frac{1-\phi}{\phi}\right)+\frac{1}{2}\Sigma^{-1}\left((x-\mu_1)^T(x-\mu_1)-(x-\mu_{-1})^T(x-\mu_{-1})\right)\right)}$$

$$= \frac{1}{1+\exp\left(\ln\left(\frac{1-\phi}{\phi}\right)+\frac{1}{2}\Sigma^{-1}\left(\cancel{x^T x}-2\mu_1 x+\mu_1^2-\cancel{x^T x}+2\mu_{-1}x-\mu_{-1}^2\right)\right)}$$

$$= \frac{1}{1+\exp\left(\ln\left(\frac{1-\phi}{\phi}\right)+\Sigma^{-1}(\mu_{-1}-\mu_1)x+\frac{1}{2}\Sigma^{-1}(\mu_1^2-\mu_{-1}^2)\right)}$$

$$\therefore\ \theta_0 = -\ln\left(\frac{1-\phi}{\phi}\right)+\frac{1}{2}\left(\mu_{-1}^T\Sigma^{-1}\mu_{-1}-\mu_1^T\Sigma^{-1}\mu_1\right)$$

$$\theta = \Sigma^{-1}(\mu_1-\mu_{-1})$$

$$\Rightarrow p(y|x;\phi,\Sigma,\mu_{-1},\mu_1) = \frac{1}{1+\exp(-y(\theta^T x+\theta_0))}$$

(b) & (c)

$$l(\phi, \mu_{-1}, \mu_1, \Sigma) = \log \prod_{i=1}^{m} p(x^{(i)}, y^{(i)}; \phi, \mu_{-1}, \mu_1, \Sigma)$$

$$= \log \prod_{i=1}^{m} p(x^{(i)} | y^{(i)}; \phi, \mu_{-1}, \mu_1, \Sigma) \cdot p(y^{(i)}; \phi)$$

$$= \sum_{i=1}^{m} \log p(x^{(i)} | y^{(i)}) + \sum_{i=1}^{m} \log p(y^{(i)}; \phi)$$

$$= \sum_{i=1}^{m} \log \left(\frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x^i - \mu_{y^i})^T \Sigma^{-1} (x^i - \mu_{y^i})\right)\right) + \sum_{i=1}^{m} \log\left(\phi^{y_i} (1-\phi)^{(1-y_i)}\right)$$

$$= \sum_{i=1}^{m} \left[\log\left(\frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}}\right) - \frac{1}{2}(x^i - \mu_{y^i})^T \Sigma^{-1} (x^i - \mu_{y^i}) + y^i \log \phi + (1-y^i) \log(1-\phi)\right]$$

$$\approx \sum_{i=1}^{m} \left(-\frac{1}{2} \log |\Sigma| - \frac{1}{2}(x^i - \mu_{y^i})^T \Sigma^{-1} (x^i - \mu_{y^i}) + y^i \log \phi + (1-y^i) \log(1-\phi)\right)$$

$$\frac{\partial l}{\partial \phi} = \sum_{i=1}^{m} \left(y^i \cdot \frac{1}{\phi} - (1-y^i) \frac{1}{1-\phi}\right) = \sum_{i=1}^{m} \frac{I(y^i = 1)}{\phi} + \frac{m - \sum_{i=1}^{m} I(y^i = 1)}{1-\phi} = 0$$

$$\Rightarrow \phi = \frac{\sum_{i=1}^{m} I\{y^i = 1\}}{m}$$

$$\nabla_{\mu_{-1}} l = \sum_{i;y=-1} \left[\Sigma^{-1} x^i - \Sigma^{-1} \mu_{-1} \right] = 0$$

$$\Rightarrow \mu_{-1} = \frac{\sum_{i=1}^{m} I\{y=-1\} \cdot x^{(i)}}{\sum_{i=1}^{m} I\{y=-1\}}$$

$$\nabla_{\mu_1} l = \sum_{i;y=1} \left[\Sigma^{-1} x^i - \Sigma^{-1} \mu_1 \right] = 0.$$

$$\Rightarrow \mu_1 = \frac{\sum_{i=1}^{m} I\{y=1\} \cdot x^{(i)}}{\sum_{i=1}^{m} I\{y=1\}}$$

let $s = \Sigma^{-1}$. $\nabla_s l = \sum_{i=1}^{m} -\frac{1}{2} \cdot s \cdot (-1) s^{-2} - \frac{1}{2} (x^i - \mu_{y^i}) (x^i - \mu_{y^i})^T$

$$= \sum_{i=1}^{m} \frac{1}{2} s^{-1} - \frac{1}{2} (x^i - \mu_{y^i}) (x^i - \mu_{y^i})^T = 0.$$

$$\Rightarrow \Sigma = \frac{1}{m}\sum_{i=1}^{m}(x^i - \mu_{y^i})(x^i - \mu_{y^i})^T$$

4. (a) Newton Method: $\theta^{(t+1)} := \theta^{(t)} - H^{-1}\nabla_\theta l$

$$\therefore z^{(t+1)} := z^{(t)} - H_{g|z}^{-1} \cdot \nabla_{g|z}$$

$$:= z^{(t)} - H_{f|Az}^{-1} \cdot \nabla_{f|Az}$$

$$H_{f|Az} = A^T \cdot H_{f|Az} \cdot A \Rightarrow H^{-1}_{f|Az} = A^{-1} H_{f|Az}^{-1} A^{T-1}$$

$$\nabla_{f|Az} = A^T \cdot \nabla_{f|Az}$$

$$\Rightarrow z^{(t+1)} := z^{(t)} - (A^{-1} H_{f|Az}^{-1} A^{T-1})(A^T \cdot \nabla_{f|Az})$$

$$:= z^{(t)} - A^{-1} Hf|_{Az}^{-1} \cdot \nabla f|_{Az}$$

$$:= z^{(t)} - A^{-1} Hf|_{x}^{-1} \cdot \nabla f|_{x}$$

$$:= A^{-1} (x - Hf|_{x}^{-1} \cdot \nabla f|_{x})$$

(b) gradient descent: $\theta := \theta - \alpha \frac{\partial}{\partial \theta} J(\theta)$.

$$\therefore z := z - \alpha \frac{df}{dz}$$

$$:= z - \alpha \frac{df}{dx} \cdot \frac{dx}{dz}$$

$$:= \lambda x - \frac{\alpha}{\lambda} (2 \lambda x) = \lambda x - \alpha (2x)$$

$$\neq \lambda (x - \alpha \frac{df}{dx}) = \lambda (x - \alpha . (2x))$$

5.

a. (i) $J(\theta) = \frac{1}{2}\sum_{i=1}^{m} w^{(i)} (\theta^T x^{(i)} - y^{(i)})^2$

$$= \frac{1}{2}\sum_{i=1}^{m} (\theta^T x^{(i)} - y^{(i)}) \cdot w^{(i)} \cdot (\theta^T x^{(i)} - y^{(i)})$$

$$= \frac{1}{2}(X\theta - y)^T W (X\theta - y)$$

where $W = \text{diag}(w^i)$.

(ii) $\nabla_\theta J(\theta) = X^T W X\theta - X^T W Y = 0.$

$$\Rightarrow \quad X^T W X \theta = X^T W Y$$

$$\theta = X^T W Y \cdot (X^T)^{-1} W^{-1} X^{-1} = (X^T W X)^{-1} X^T W Y$$

(iii) $$\ell(\theta) = \log \prod_{i=1}^{m} p(y^{(i)} \mid x^{(i)}; \theta)$$

$$= \log \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\,\sigma^{(i)}} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2(\sigma^{(i)})^2}\right)$$

$$= \sum_{i=1}^{m} \log \frac{1}{\sqrt{2\pi}\,\sigma^{(i)}} + \sum_{i=1}^{m} -\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2(\sigma^{(i)})^2}$$

Maximize = Minimize $\displaystyle\sum_{i=1}^{m} \frac{(y^{(i)} - \theta^T x^{(i)})^2}{2(\sigma^{(i)})^2}$

$\because$ $J(\theta) = \frac{1}{2} \sum_{i=1}^{m} w^{(i)} (\theta^T x^{(i)} - y^{(i)})^2$

$\therefore$ $w^{(i)} = \dfrac{1}{\sigma^{(i)2}}$