Solution to CS229 Problem Set 1

Son Nguyen

5/10/2019

Problem 1

(a)

The first derivative of $J(\theta)$ with respect to $\theta_j$is

$$\frac{\partial}{\partial\theta_j}J(\theta) = \frac{1}{m}\sum_{i=1}^{m}\frac{-y^{(i)}x_j^{(i)}e^{-y^{(i)}\theta^T x^{(i)}}}{1+e^{-y^{(i)}\theta^T x^{(i)}}} \tag{1}$$

Hence, the second derivative of $J(\theta)$ with respect to θ_j and θ_k is

$$\begin{aligned}\frac{\partial^2}{\partial\theta_j\theta_k}J(\theta) &= \frac{\partial}{\partial\theta_j}\frac{\partial}{\partial\theta_k}J(\theta) \\ &= \frac{1}{m}\sum_{i=1}^{m}\frac{x_j^{(i)}x_k^{(i)}\left(y^{(i)}e^{-y^{(i)}\theta^T x^{(i)}}\right)^2}{\left(1+e^{-y^{(i)}\theta^T x^{(i)}}\right)^2}\end{aligned} \tag{2}$$

Therefore, the hessian matrix of $J(\theta)$ is

$$H_{jk} = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{y^{(i)}e^{-y^{(i)}\theta^T x^{(i)}}}{1+e^{-y^{(i)}\theta^T x^{(i)}}}\right)^2 x_j^{(i)}x_k^{(i)} \tag{3}$$

with $H \in \mathbb{R}^{n\times n}$ (n is the number of features of x).
Consider an arbitrary vector $z \in \mathbb{R}^{n\times 1}$. Note that,

$$\begin{aligned}\sum_i\sum_j z_i x_i z_j x_j &= \left(\sum_i z_i x_i\right)\left(\sum_j x_j z_j\right) \\ &= \left(x^T z\right)^2 \\ &\geq 0\end{aligned} \tag{4}$$

Hence we have,

$$z^T H z = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{y^{(i)}e^{-y^{(i)}\theta^T x^{(i)}}}{1+e^{-y^{(i)}\theta^T x^{(i)}}}\right)^2\left(\sum_i x_i z_i\right)\left(\sum_j x_j z_j\right) \tag{5}$$

Since $\frac{1}{m}\sum_{i=1}^{m}\left(\frac{y^{(i)}e^{-y^{(i)}\theta^T x^{(i)}}}{1+e^{-y^{(i)}\theta^T x^{(i)}}}\right)^2 \geq 0$ and $\sum_i\sum_j z_i x_i z_j x_j \geq 0$, $z^T H z \geq 0$

Problem 2

(a)

The common form of exponential family is

$$p\left(y;\eta\right) = b\left(y\right) exp\left(\eta^T T\left(y\right) - a\left(\eta\right)\right) \tag{6}$$

Transform the original Poisson distribution:

$$\begin{aligned} p\left(y;\lambda\right) &= \frac{e^{-\lambda}\lambda^y}{y!} \\ &= \frac{1}{y!}e^{-\lambda}e^{y\ln\lambda} \\ &= \frac{1}{y!}exp\left(y\ln\lambda - \lambda\right) \end{aligned} \tag{7}$$

Hence, Poisson distribution has the form of exponential family with:

$$b\left(y\right) = \frac{1}{y!} \tag{8}$$

$$\eta = \ln\lambda \tag{9}$$

$$T\left(y\right) = y \tag{10}$$

$$a\left(\eta\right) = -e^{\eta} \tag{11}$$

(b)

Since a Poisson random variable with parameter λ has mean λ,

$$\begin{aligned} h_\theta\left(x\right) &= E[y|x;\theta] \\ &= \lambda \\ &= e^{\eta} \\ &= e^{\theta^T x} \end{aligned} \tag{12}$$

(the last equality follows from the third assumption used when constructing a GLM: the natural parameter η and the inputs x are related linearly)

(c)

Log-likelihood of a training example $(x^{(i)}, y^{(i)})$:

$$\begin{aligned} l(\theta) &= \log p(y^{(i)}|x^{(i)};\theta) \\ &= -e^{\theta^T x^{(i)}} + y^{(i)}(\theta^T x^{(i)}) - \log(y!) \end{aligned} \tag{13}$$

First derivative of log-likelihood with respect to θ_j of a training example $(x^{(i)}, y^{(i)})$:

$$\begin{aligned} \frac{\partial}{\partial \theta_j} l(\theta) &= -x_j^{(i)} e^{\theta^T x} + y^{(i)} x_j^{(i)} \\ &= x_j^{(i)} \left(y^{(i)} - e^{\theta^T x} \right) \end{aligned} \tag{14}$$

Hence, the stochastic gradient ascent rule with learning-rate α for learning using a GLM model with Poisson responses y and the canonical response function is:

repeat until convergence:
for $i = 1$ to m:
for $j = 1$ to n:

$$\theta_j = \theta_j - \alpha \left(y^{(i)} - e^{\theta^T x} \right) x_j^{(i)} \tag{15}$$

(d)

From the general formula of exponential family, the first derivative of log-likelihood of $p(y;\eta)$ is

$$\begin{aligned} \frac{\partial}{\partial \theta_i} l(\theta) &= \frac{\partial}{\partial \theta_i} \log \left(b(y) exp(\eta^T T(y) - a(\eta)) \right) \\ &= \frac{\partial}{\partial \theta_i} \left(\log(b(y)) + \left(\eta^T T(y) - a(\eta) \right) \right) \\ &= \frac{\partial}{\partial \theta_i} \left((\theta^T x) y - a(\theta^T x) \right) \\ &= x_i y - a(\theta^T x)' x_i \\ &= x_i \left(y - a(\theta^T x)' \right) \end{aligned} \tag{16}$$

Let $a(\theta^T x)'$ be $h(x)$, the stochastic ascent rule with learning-rate α on the log-likelihood is

$$\theta_i = \theta_i - \alpha(h(x) - y)x_i \qquad (Q.E.D) \tag{17}$$

Problem 3

(a)

Consider $p(y=1|x)$

$$p(y=1|x) = \frac{p(x|y=1)p(y=1)}{p(x)}$$
$$= \frac{p(x|y=1)p(y=1)}{p(x|y=1)p(y=1)+p(x|y=-1)p(y=-1)}$$
$$= \frac{1}{1+\frac{p(x|y=-1)p(y=-1)}{p(x|y=1)p(y=1)}}$$
$$= \frac{1}{1+exp\Big(-\frac{1}{2}\left((x-\mu_{-1})^T\Sigma^{-1}(x-\mu_{-1})-(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)\right)+\ln\left(\frac{1-\phi}{\phi}\right)\Big)} \tag{18}$$

Note that $(x-\mu)^T\Sigma^{-1}(x-\mu) = x^T\Sigma^{-1}x - 2\mu^T\Sigma^{-1}x + \mu^T\Sigma^{-1}\mu$. Hence the equation (19) becomes:

$$p(y=1|x) = \frac{1}{1+exp\left(-(\mu_1-\mu_{-1})^T\Sigma^{-1}x-\frac{1}{2}(\mu_1-\mu_{-1})^T\Sigma^{-1}(\mu_1-\mu_{-1})-\ln\left(\frac{\phi}{1-\phi}\right)\right)}$$
$$= \frac{1}{1+exp\left((-1)\left((\mu_1-\mu_{-1})^T\Sigma^{-1}x+\frac{1}{2}(\mu_1-\mu_{-1})^T\Sigma^{-1}(\mu_1-\mu_{-1})+\ln\left(\frac{\phi}{1-\phi}\right)\right)\right)} \tag{19}$$

Similarly,

$$p(y=-1|x) = \frac{1}{1+exp\left((1)\left((\mu_1-\mu_{-1})^T\Sigma^{-1}x+\frac{1}{2}(\mu_1-\mu_{-1})^T\Sigma^{-1}(\mu_1-\mu_{-1})+\ln\left(\frac{\phi}{1-\phi}\right)\right)\right)} \tag{20}$$

Therefore, with

$$\theta = (\Sigma^{-1})^T(\mu_1-\mu{-1}) \tag{21}$$
$$\theta_0 = \frac{1}{2}(\mu_1-\mu_{-1})^T\Sigma^{-1}(\mu_1-\mu_{-1})+\ln\left(\frac{\phi}{1-\phi}\right) \tag{22}$$

the posterior distribution of the label at x takes the form of a logistic function (*Q.E.D*)

(b)

$$
\begin{aligned}
l(\phi, \mu_{-1}, \mu_1, \Sigma) &= \log \prod_{i=1}^{m} p(x^{(i)}|y^{(i)})p(y^{(i)}) \\
&= \sum_{i=1}^{m} \log \left(p(x^{(i)}|y^{(i)})p(y^{(i)}) \right) \\
&= \sum_{i=1}^{m} \left[\log \left(p(x^{(i)}|y^{(i)}) \right) + \log \left(p(y^{(i)}) \right) \right]
\end{aligned} \tag{23}
$$

Given the assumption that the dimension of $x^{(i)}$ is 1, $x^{(i)} \in \mathbb{R}$ and $\Sigma \in \mathbb{R}$
Note that:

$$
\begin{aligned}
\log \left(p(x^{(i)}|y^{(i)}) \right) &= - \left[\log \left((2\pi|\Sigma|)^{1/2} \right) + \frac{1}{2}(x^{(i)} - \mu_{y^{(i)}})^T \Sigma^{-1} (x^{(i)} - \mu_{y^{(i)}}) \right] \\
&= - \left[\log \left((2\pi|\Sigma|)^{1/2} \right) + \frac{1}{2} \left((x^{(i)})^T \Sigma^{-1} x^{(i)} - 2\mu_{y^{(i)}}^T \Sigma^{-1} x^{(i)} + \mu_{y^{(i)}}^T \Sigma^{-1} \mu_{y^{(i)}} \right) \right] \\
&= - \left[\log \left(\sqrt{2\pi\Sigma} \right) + \frac{1}{2\Sigma}(x^{(i)})^2 - \frac{\mu_{y^{(i)}}}{\Sigma} x^{(i)} + \frac{\mu_{y^{(i)}}^2}{2\Sigma} \right]
\end{aligned}
$$

$\log p(y^{(i)}) = \frac{1}{2}[(1-y)\log(1-\phi) + (1+y)\log(\phi)]$
First derivatives of $l(\phi, \mu_{-1}, \mu_1, \Sigma)$ with respect to ϕ, μ_{-1}, μ_1, Σ are

$$
\begin{aligned}
\frac{\partial l}{\partial \phi} &= \sum_{i=1}^{m} \frac{1}{2} \left(\frac{1+y^{(i)}}{\phi} - \frac{1-y^{(i)}}{1-\phi} \right) \\
&= \frac{1}{2} \sum_{i=1}^{m} \frac{(1+y^{(i)})(1-\phi) - (1-y^{(i)})\phi}{\phi(1-\phi)}
\end{aligned}
$$

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

$$
\frac{\partial l_i}{\partial \mu_1} = \sum_{i=1}^{m} \frac{x^{(i)} - \mu_1}{\Sigma} \{y^{(i)} = 1\}
$$

$$
\frac{\partial l_i}{\partial \Sigma} = \sum_{i=1}^{m} \left[\frac{1}{2\Sigma^2} \left(x^{(i)} - \mu_{y^{(i)}} \right)^2 - \frac{1}{2\Sigma} \right]
$$

Set all first derivatives to 0, we have the conditions of parameters ϕ, μ_{-1}, μ_1, Σ such that $l(\phi, \mu_{-1}, \mu_1, \Sigma)$ is maximum:

$$
\frac{\partial l}{\partial \phi} = 0 \Leftrightarrow \sum_{i=1}^{m} \left[(1+y^{(i)})(1-\phi) - (1-y^{(i)})\phi \right] = 0 \Leftrightarrow \phi = \sum_{i=1}^{m} \frac{1+y^{(i)}}{2m} = \frac{1}{m} \sum_{i=1}^{m} 1\{y^{(i)}\}
$$

$$
\frac{\partial l}{\partial \mu_{-1}} = 0 \Leftrightarrow \sum_{i=1}^{m} (x^{(i)} - \mu_{-1})\{y^{(i)} = -1\} = 0 \Leftrightarrow \mu_{-1} = \frac{\sum_{i=1}^{m} 1\{y^{(i)} = -1\} x^{(i)}}{\sum_{i=1}^{m} 1\{y^{(i)=-1}\}}
$$

$$\frac{\partial l}{\partial \mu_1} = 0 \Leftrightarrow \sum_{i=1}^{m} (x^{(i)} - \mu_1)\{y^{(i)} = -1\} = 0 \Leftrightarrow \mu_1 = \frac{\sum_{i=1}^{m} 1\{y^{(i)} = 1\} x^{(i)}}{\sum_{i=1}^{m} 1\{y^{(i)=1}\}}$$

$$\frac{\partial l}{\partial \Sigma} = 0 \Leftrightarrow \sum_{i=1}^{m} \Sigma = \sum_{i=1}^{m} \left(x^{(i)} - \mu_{y^{(i)}}\right)^2 \Leftrightarrow \Sigma = \frac{1}{m} \sum_{i=1}^{m} \left(x^{(i)} - \mu_{y^{(i)}}\right) \left(x^{(i)} - \mu_{y^{(i)}}\right)^T$$

(Q.E.D)

(c)

Note that the change in value of n (the dimension of $x^{(i)}$) does not affect MLE of ϕ. Then,

$$\phi = \frac{1}{m} \sum_{i=1}^{m} 1\{y^{(i)}\}$$

We have:

$$\begin{aligned}
\log\left(p(x^{(i)}|y^{(i)})\right) &= -\left[\log\left((2\pi)^{n/2}|\Sigma|^{1/2}\right) + \frac{1}{2}(x^{(i)} - \mu_{y^{(i)}})^T \Sigma^{-1} (x^{(i)} - \mu_{y^{(i)}})\right] \\
&= -\left[\log\left((2\pi)^{n/2}|\Sigma|^{1/2}\right) + \frac{1}{2}\left((x^{(i)})^T \Sigma^{-1} x^{(i)} - 2\mu_{y^{(i)}}^T \Sigma^{-1} x^{(i)} + \mu_{y^{(i)}}^T \Sigma^{-1} \mu_{y^{(i)}}\right)\right]
\end{aligned}$$

Gradients of $l(\phi, \mu_{-1}, \mu_1, \Sigma)$ with respect to $\mu_{-1} \in \mathbb{R}^{n\times 1}$, $\mu_1 \in \mathbb{R}^{n\times 1}$ are

$$\nabla_{\mu_{-1}} l = \sum_{i=1}^{m} \Sigma^{-1} \left(x^{(i)} - \mu_{-1}\right) \{y^{(i)} = -1\}$$

$$\nabla_{\mu_1} l = \sum_{i=1}^{m} \Sigma^{-1} \left(x^{(i)} - \mu_1\right) \{y^{(i)} = 1\}$$

Set each gradient to 0, we have:

$$\mu_{-1} = \frac{\sum_{i=1}^{m} 1\{y^{(i)} = -1\} x^{(i)}}{\sum_{i=1}^{m} 1\{y^{(i)} = -1\}} \tag{24}$$

$$\mu_1 = \frac{\sum_{i=1}^{m} 1\{y^{(i)} = 1\} x^{(i)}}{\sum_{i=1}^{m} 1\{y^{(i)} = 1\}} \tag{25}$$

Note that,

$$\nabla_{\Sigma^{-1}} \log|\Sigma| = \nabla_{\Sigma^{-1}} \left(-\log|\Sigma^{-1}|\right) = -\Sigma$$

$$\nabla_{\Sigma^{-1}} (x^{(i)} - \mu_{y^{(i)}})^T \Sigma^{-1} (x^{(i)} - \mu_{y^{(i)}}) = (x^{(i)} - \mu_{y^{(i)}})(x^{(i)} - \mu_{y^{(i)}})^T$$

Hence, gradient of $l(\phi, \mu_{-1}, \mu_1, \Sigma)$ with respect to $\Sigma^{-1} \in \mathbb{R}^{n\times n}$ is (NEED TO CHECK BACK)

$$\begin{aligned}
\nabla_{\Sigma^{-1}} l &= -\left[\sum_{i=1}^{m} \nabla_{\Sigma^{-1}} \log\left((2\pi)^{n/2}|\Sigma|^{1/2}\right) + \frac{1}{2}\sum_{i=1}^{m} \nabla_{\Sigma^{-1}} \left((x^{(i)} - \mu_{y^{(i)}})^T \Sigma^{-1} (x^{(i)} - \mu_{y^{(i)}})\right)\right] \\
&= \frac{1}{2}\left[\sum_{i=1}^{m} \Sigma - \sum_{i=1}^{m} (x^{(i)} - \mu_{y^{(i)}})(x^{(i)} - \mu_{y^{(i)}})^T\right]
\end{aligned}$$

Set the gradient to 0, we have:

$$\Sigma = \frac{1}{m}\sum_{i=1}^{m}(x^{(i)} - \mu_{y^{(i)}})(x^{(i)} - \mu_{y^{(i)}})^T \tag{26}$$

(24), (25), (26) $\Rightarrow$ Q.E.D

Problem 4

(a)

Let H_z be the hessian of function $g(z)$ and H_x be the hessian of the function $f(x)$. Note that, if $z = A^{-1}x$ then $g(z) = f(A(A^{-1}x)) = f(AA^{-1}x) = f(x)$. First of all, here are some useful matrix calculus identities:

$$\begin{aligned}
\boldsymbol{\nabla}_{A^{-1}x} f(x) &= \boldsymbol{\nabla}_{A^{-1}x} f(AA^{-1}x) \\
&= \left(\frac{\partial f(AA^{-1}x)}{\partial (A^{-1}x)_i}\right)_i \\
&= A\left(\frac{\partial f(A^{-1}x)}{\partial (A^{-1}x)_i}\right) \\
&= A\left(\frac{\partial f(x)}{\partial x}\right)_i
\end{aligned} \tag{27}$$

$$\begin{aligned}
H_{A^{-1}x} &= \left(\frac{\partial^2 f(x)}{\partial (A^{-1}x)_i \partial (A^{-1}x)_j}\right)_{ij} \\
&= \frac{\partial}{\partial (A^{-1}x)_i}\left(\frac{\partial f(AA^{-1}x)}{\partial (A^{-1}x)_j}\right) \\
&= \frac{\partial}{\partial (A^{-1}x)_i}\left(A\frac{\partial f(A^{-1}x)}{\partial (A^{-1}x)_j}\right) \\
&= A\frac{\partial}{\partial (A^{-1}x)_i}\left(\frac{\partial f(A^{-1}x)}{\partial (A^{-1}x)_j}\right) \\
&= A\frac{\partial}{\partial (A^{-1}x)_i}\frac{\partial f(x)}{\partial x_j} \\
&= A\frac{\partial}{\partial x_j}\left(\frac{\partial f(x)}{\partial (A^{-1}x)_i}\right) \\
&= A\frac{\partial}{\partial x_j}\left(\frac{\partial f(AA^{-1}x)}{\partial (A^{-1}x)_i}\right) \\
&= A\frac{\partial}{\partial x_j}\left(A\frac{\partial f(A^{-1}x)}{\partial (A^{-1}x)_i}\right) \\
&= A^2\frac{\partial^2 f(x)}{\partial x_i \partial x_j}
\end{aligned} \tag{28}$$

Note that Newton's method is invariant to linear reparameterizations for the base case 0. Suppose it is true for i. We are going to prove that it is also true for $i+1$.
Using Newton's method, we have:

$$\begin{aligned}
z^{(i+1)} &= z^{(i)} - H_{z^{(i)}}^{-1} \boldsymbol{\nabla}_{z^{(i)}} g(z^{(i)}) \\
&= \left(A^{-1}x^{(i)}\right) - H_{A^{-1}x^{(i)}}^{-1} \boldsymbol{\nabla}_{A^{-1}x^{(i)}} f(A(A^{-1}x^{(i)})) \\
&= \left(A^{-1}x^{(i)}\right) - H_{A^{-1}x^{(i)}}^{-1} \boldsymbol{\nabla}_{A^{-1}x^{(i)}} f(x^{(i)}))
\end{aligned}$$

From (27),

$$\boldsymbol{\nabla}_{A^{-1}x^{(i)}} = A \boldsymbol{\nabla}_{x^{(i)}} f(x^{(i)}) \tag{29}$$

From (28),

$$H_{A^{-1}x^{(i)}}^{-1} = A^{-2} H_{x^{(i)}}^{-1} \tag{30}$$

Also,

$$x^{(i+1)} = x^{(i)} - H_{x^{(i)}}^{-1} \boldsymbol{\nabla}_{x^{(i)}} f(x^{(i)}) \tag{31}$$

Hence,

$$\begin{aligned}
z^{(i+1)} &= \left(A^{-1}x^{(i)}\right) - (A^{-2}) H_{x^{(i)}}^{-1} A \boldsymbol{\nabla}_{x^{(i)}} f(x^{(i)}) \\
&= \left(A^{-1}x^{(i)}\right) - A^{-1} H_{x^{(i)}}^{-1} \boldsymbol{\nabla}_{x^{(i)}} f(x^{(i)}) \\
&= A^{-1}(x^{(i)} - H_{x^{(i)}}^{-1} \boldsymbol{\nabla}_{x^{(i)}} f(x^{(i)}) \\
&= A^{-1} x^{(i+1)}
\end{aligned}$$

By mathematical induction, Newton's method is indeed invariant to linear reparameterizations.

(b)

Using gradient descent with learning rate α, we have:

$$\begin{aligned}
z^{(i+1)} &= z^{(i)} - \alpha \boldsymbol{\nabla}_{z^{(i)}} g(z^{(i)}) \\
&= \left(A^{-1}x^{(i)}\right) - \alpha A \boldsymbol{\nabla}_{x^{(i)}} f(x^{(i)})
\end{aligned}$$

Hence, $z^{(i+1)} \neq A^{-1}x^{(i+1)}$. Gradient descent is not invariant to linear reparameterizations.

Problem 5

(a)

i. We have:

$$
\begin{aligned}
(X\theta - y)^T W (X\theta - y) &= \left[(X\theta - y)_1 W_{11} \quad (X\theta - y)_2 W_{22} \quad \dots \quad (X\theta - y)_n W_{nn}\right](X\theta - y) \\
&= \sum_{i=1}^{n} W_{ii}(X\theta - y)_i^2 \\
&= \sum_{i=1}^{n} W_{ii}(\theta^T x^{(i)} - y^{(i)})^2
\end{aligned}
$$

Hence, $J(\theta)$ can be written as $(X\theta - y)^T W (X\theta - y)$ with diagonal matrix W:

$$
\begin{bmatrix}
\frac{1}{2}w^{(1)} & 0 & 0 & 0 \\
0 & \frac{1}{2}w^{(2)} & 0 & 0 \\
0 & 0 & \dots & 0 \\
0 & 0 & 0 & \frac{1}{2}w^{(n)}
\end{bmatrix}
$$

ii. We have;

$$
\begin{aligned}
\nabla_\theta J(\theta) &= \nabla_\theta \left[(X\theta - y)^T W (X\theta - y)\right] \\
&= \nabla_\theta \left[\theta^T X^T W (X\theta - y) - y^T W (X\theta - y)\right] \\
&= \nabla_\theta \left(\theta^T X^T W X\theta - \theta^T X^T W y - y^T W X\theta + y^T W y\right)
\end{aligned}
\tag{32}
$$

Since trace of a real number is that real number,

$$
\begin{aligned}
\nabla_\theta J(\theta) &= \nabla_\theta tr\left(\theta^T X^T W X\theta - \theta^T X^T W y - y^T W X\theta + y^T W y\right) \\
&= \nabla_\theta \left[tr(\theta^T X^T W X\theta) - tr(\theta^T X^T W y) - tr(y^T W X\theta) + tr(y^T W y)\right]
\end{aligned}
$$

Since $tr(A) = tr(A^T)$,

$$
\begin{aligned}
\nabla_\theta J(\theta) &= \nabla_\theta \left[tr(\theta^T X^T W X\theta) - 2tr(y^T W X\theta) + tr(y^T W y)\right] \\
&= \nabla_\theta tr(\theta^T X^T W X\theta) - 2\nabla_\theta tr(y^T W X\theta) + \nabla_\theta tr(y^T W y) \\
&= \nabla_\theta tr(\theta^T X^T W X\theta) - 2\nabla_\theta tr(y^T W X\theta)
\end{aligned}
$$

Note two useful matrix calculus identities:

$$
\nabla_{A^T} f(A) = (\nabla_A f(A))^T \tag{33}
$$

$$
\nabla_A tr(ABA^T C) = CAB + C^T A B^T \tag{34}
$$

From the above two identities, we have:

$$
\nabla_{A^T} tr(ABA^T C) = B^T A^T C^T + BA^T C \tag{35}
$$

Let $X\theta^T$ be A, X^TWX be B, I be C, we have:

$$\begin{aligned}\nabla_\theta tr(\theta^T X^T W X\theta) &= X^T W^T X\theta I^T + X^T W X\theta I \\ &= X^T W^T X\theta + X^T W X\theta\end{aligned} \tag{36}$$

Note an useful matrix calculus identity:

$$\nabla_A tr(AB) = B^T \tag{37}$$

Since $tr(A) = tr(A^T)$, $\nabla_\theta tr(y^T W X\theta) = \nabla_\theta tr(\theta^T X^T W^T y)$. From (33),

$$\nabla_\theta tr(\theta^T X^T W^T y) = (\nabla_{\theta^T} tr(\theta^T X^T W^T y))^T \tag{38}$$

Let θ^T be A, $X^T W^T y$ be B, we have:

$$(\nabla_{\theta^T} tr(\theta^T X^T W^T y))^T = X^T W^T y \tag{39}$$

From (36) and (39),

$$\nabla_\theta J(\theta) = X^T W^T X\theta + X^T W X\theta - 2X^T W^T y$$

Note that since W is a diagonal matrix, $W^T = W$. Let $\nabla_\theta J(\theta)$ be 0 and solve for θ:

$$\begin{aligned}X^T W^T X\theta + X^T W X\theta - 2X^T W^T y = 0 \\ X^T W X\theta + X^T W X\theta &= 2X^T W y \\ 2X^T W X\theta &= 2X^T W y \\ \theta &= (X^T W X)^{-1} X^T W y\end{aligned}$$

iii. The log-likelihood is

$$\begin{aligned}l(\theta) &= \log \prod_{i=1}^{m} p(y^{(i)}|x^{(i)};\theta) \\ &= \sum_{i=1}^{m} \log p(y^{(i)}|x^{(i)};\theta) \\ &= -\sum_{i=1}^{m} \left[\log(\sqrt{2\pi}\sigma^{(i)}) + \frac{(y^{(i)} - \theta^T x^{(i)})^2}{2(\sigma^{(i)})^2}\right]\end{aligned}$$

Hence, finding the maximum likelihood estimate of θ reduces to minimizing

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

which is equivalent to solving a weighted linear regression problem with

$$w^{(i)} = \frac{-1}{2(\sigma^{(i)})^2}$$