CS229 Machine Learning Problem Set #1 Solution

Roy C.K. Chan

Problem 1(a). Let $\theta \in \mathbb{R}^n$ be a n-dimensional vector.

$$\begin{aligned} h_\theta(x) &= g(\theta^T x) \\ \implies \nabla h_\theta(x) &= g'(\theta^T x) \cdot x \,. \end{aligned} \tag{1}$$

It can be easily show that g satisfies

$$g'(z) = g(z)(1 - g(z)) \,.$$

Hence, using (1), we have

$$\nabla h_\theta(x) = h_\theta(x)(1 - h_\theta(x)) \cdot x \,. \tag{2}$$

So that, by (2),

$$\begin{aligned} J(\theta) &= -\frac{1}{m} \sum_{i=1}^{m} \log(h_\theta(y^{(i)} x^{(i)})) \\ \implies \nabla J(\theta) &= -\frac{1}{m} \sum_{i=1}^{m} \frac{1}{h_\theta(y^{(i)} x^{(i)})} \nabla h_\theta(y^{(i)} x^{(i)}) \\ &= -\frac{1}{m} \sum_{i=1}^{m} (1 - h_\theta(y^{(i)} x^{(i)})) \cdot y^{(i)} x^{(i)} \,, \\ i.e., \quad \frac{\partial J}{\partial \theta_k} &= -\frac{1}{m} \sum_{i=1}^{m} (1 - h_\theta(y^{(i)} x^{(i)})) \cdot y^{(i)} x_k^{(i)} \,, \end{aligned}$$

for $k = 1, 2, ..., n$. Therefore, using (2) again, for $1 \le j, k \le n$,

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

where the second line follows from the fact that $y^{(i)} \in \{-1, 1\}$. Simplifying, we have

$$H = (H_{j,k})_{1 \le j,k \le n} = \frac{1}{m} \sum_{i=1}^{m} h_\theta(y^{(i)} x^{(i)})(1 - h_\theta(y^{(i)} x^{(i)})) \cdot x^{(i)} x^{(i)T} \,.$$

For any $z \in \mathbb{R}^n$,

$$\begin{aligned} z^T H z &= \frac{1}{m}\sum_{i=1}^{m} h_\theta(y^{(i)}x^{(i)})(1-h_\theta(y^{(i)}x^{(i)}))\cdot z^T x^{(i)} x^{(i)T} z \\ &= \frac{1}{m}\sum_{i=1}^{m} h_\theta(y^{(i)}x^{(i)})(1-h_\theta(y^{(i)}x^{(i)}))\cdot z^T x^{(i)} (z^T x^{(i)})^T \\ &= \frac{1}{m}\sum_{i=1}^{m} h_\theta(y^{(i)}x^{(i)})(1-h_\theta(y^{(i)}x^{(i)}))\cdot (z^T x^{(i)})^2 \\ &\geq 0\,, \end{aligned}$$

since $0 \leq h_\theta(y^{(i)}x^{(i)}) \leq 1,\ (z^T x^{(i)})^2 \geq 0$.

Problem 1(b)(c). See the files "CostFunction.py" and "Q1.ipynb".

Problem 2(a). Re-write the probability mass function of Poisson distribution (parametrized by λ) as follows:

$$p(y;\lambda) = \frac{1}{y!}\exp(y\log(\lambda)-\lambda)\,.$$

By Fisher-Neyman factorization theorem, $T(y)=y$ is sufficient for λ, and we have

$$\begin{aligned} b(y) &= \frac{1}{y!}\,, \\ \eta &= \log(\lambda)\,, \\ T(y) &= y\,, \\ a(\eta) &= \lambda \\ &= \exp(\eta)\,. \end{aligned}$$

Problem 2(b). With $y|x;\lambda \sim Poi(\lambda)$, the canonical response function is

$$\begin{aligned} g(\eta) &= \mathbb{E}[y;\eta] \\ &= \lambda \\ &= \exp(\eta)\,. \end{aligned}$$

Problem 2(c). Using part (a),

$$\begin{aligned} l(\theta) &= \log p(y^{(i)}|x^{(i)};\theta) \\ &= \log\Big(\frac{1}{y^{(i)}!}\exp\big(y^{(i)}\log(\lambda)-\lambda\big)\Big) \\ &= \log\Big(\frac{1}{y^{(i)}!}\exp\big(y^{(i)}\log(\exp(\theta^T x^{(i)}))-\exp(\theta^T x^{(i)})\big)\Big) \\ &= \log\frac{1}{y^{(i)}!} + \big(y^{(i)}\theta^T x^{(i)} - \exp(\theta^T x^{(i)})\big) \\ \Longrightarrow \frac{\partial l}{\partial \theta_j} &= y^{(i)}\cdot x_j^{(i)} - \exp(\theta^T x^{(i)})\cdot x_j^{(i)} = \big(y^{(i)}-\exp(\theta^T x^{(i)})\big)x_j^{(i)} \end{aligned}$$

Therefore, the stochastic gradient ascent rule for learning using a GLM with Poisson responses y is

$$\theta_j := \theta_j + \alpha\big(y^{(i)} - \exp(\theta^T x^{(i)})\big)x_j^{(i)}\,.$$

Problem 2(d). Now, for a GLM with a response variable $y \in \mathbb{R}^K$ ($K \geq 1$) from some exponential family and $T(y) = y$, we have

$$p(y;\eta) = b(y)\exp(\eta^T y - \alpha(\eta))\,.$$

where $\eta \in \mathbb{R}^K$, $b : \mathbb{R}^K \to \mathbb{R}$ and $\alpha : \mathbb{R}^K \to \mathbb{R}$. Therefore, the log-likelihood of a single training example $(x^{(i)}, y^{(i)})$ is given by

$$l(\Theta) \;=\; \log b(y^{(i)}) + \eta^{(i)T} y^{(i)} - \alpha(\eta^{(i)})$$

where $\Theta \in \mathbb{R}^{n\times K}$ is the weight matrix, and $\eta^{(i)} = \Theta^T x^{(i)}$, where $i = 1, \cdots, m$. By chain rule (of matrix calculus), for $j = 1, \cdots, n$ and $k = 1, \cdots, K$,

$$\begin{aligned}
\frac{\partial l}{\partial \Theta_{jk}} &= tr\Big(\big(\frac{\partial(\eta^{(i)T}y^{(i)})}{\partial \eta^{(i)}}\big)^T \frac{\partial \eta^{(i)}}{\partial \Theta_{jk}}\Big) - tr\Big(\big(\frac{\partial(\alpha(\eta^{(i)}))}{\partial \eta^{(i)}}\big)^T \frac{\partial \eta^{(i)}}{\partial \Theta_{jk}}\Big) \\
&= tr\big(y^{(i)T}\frac{\partial \eta^{(i)}}{\partial \Theta_{jk}}\big) - tr\big(\nabla\alpha(\eta^{(i)})^T \frac{\partial \eta^{(i)}}{\partial \Theta_{jk}}\big)\,. \qquad (3)
\end{aligned}$$

From the definition of $\eta^{(i)} = \Theta^T x^{(i)}$, it can be easily shown that

$$\frac{\partial \eta^{(i)}}{\partial \Theta_{jk}} = \begin{bmatrix} 0 \\ \vdots \\ x_j^{(i)} \\ \vdots \\ 0 \end{bmatrix}, \qquad (4)$$

which is a K-dimensional column vector with all elements being zeros, except that the k^{th} element is $x_j^{(i)}$. From (3), (4), we have

$$\begin{aligned}
\frac{\partial l}{\partial \Theta_{jk}} &= y_k^{(i)} \cdot x_j^{(i)} - (\nabla\alpha(\eta^{(i)}))_k \cdot x_j^{(i)} = (y_k^{(i)} - (\nabla\alpha(\eta^{(i)}))_k)x_j^{(i)} \\
\Leftrightarrow \quad \frac{\partial l}{\partial \theta_j} &= (y^{(i)} - \nabla\alpha(\eta^{(i)}))x_j^{(i)}\,, \qquad (5)
\end{aligned}$$

where (5) comes from vectorization and θ_j denotes the j^{th} row of Θ (as a column vector). It remains to consider $\nabla\alpha(\cdot)$. Note that, by properties of probability, $p(y;\eta)$ sums/integrates over y to 1, i.e.,

$$\begin{aligned}
\int_y p(y;\eta)\,dy &= 1 \\
\Longrightarrow \int_y b(y)\exp(\eta^T y - \alpha(\eta))\,dy &= 1 \\
\Longrightarrow \int_y b(y)\exp(\eta^T y)\,dy &= \exp(\alpha(\eta))\,.
\end{aligned}$$

Differentiate both sides w.r.t. η, by Leibniz integral rule,

$$\begin{aligned}\int_y b(y)\exp(\eta^T y)\cdot y\,dy &= \exp(\alpha(\eta))\cdot\nabla\alpha(\eta)\\ \implies \nabla\alpha(\eta) &= \int_y y\cdot b(y)\exp(\eta^T y-\alpha(\eta))\,dy\\ &= \mathbb{E}[y;\eta]\\ &= h(x)\,, \end{aligned}\tag{6}$$

where the second last equality comes from the definition of expectation of a random variable. Hence, by (5) and (6), the update rule is given by

$$\theta_j := \theta_j - \alpha(h(x^{(i)})-y^{(i)})x_j^{(i)}\,.$$

Problem 3(a). Let Θ be the collection of $\phi, \Sigma, \mu_{-1}, \mu_1$. Consider the cases when $y=1$ and $y=-1$.

$$\begin{aligned}&p(y=1|x;\Theta)\\ =\;& \frac{p(x|y=1;\Theta)\cdot p(y=1;\Theta)}{p(x|y=1;\Theta)\cdot p(y=1;\Theta)+p(x|y=-1;\Theta)\cdot p(y=-1;\Theta)}\\ =\;& \frac{\frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}}\exp\big(-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)\big)\cdot\phi}{\frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}}\exp(-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1))\cdot\phi+\frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}}\exp(-\frac{1}{2}(x-\mu_{-1})^T\Sigma^{-1}(x-\mu_{-1}))\cdot(1-\phi)}\\ =\;& \frac{1}{1+\exp\big(-\frac{1}{2}(x-\mu_{-1})^T\Sigma^{-1}(x-\mu_{-1})+\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)\big)\cdot\frac{1-\phi}{\phi}}\,. \end{aligned}\tag{7}$$

In the denominator, the two qradratic forms have the same square matrix Σ and the qradratic terms cancel out each other, so the exponent is linear in x. Moreover, the factor $\frac{1-\phi}{\phi}$ can be written as $\exp\big(\log(\frac{1-\phi}{\phi})\big)$ so that it is included in the exponent as part of the bias term.

Below, I show some (unnecessary) calculations to find the explicit expressions of θ and θ_0. Consider the denominator of (7), after the above-mentioned operations,

$$\begin{aligned}&\exp\Big(-\tfrac{1}{2}(x-\mu_{-1})^T\Sigma^{-1}(x-\mu_{-1})+\tfrac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)\Big)\cdot\tfrac{1-\phi}{\phi}\\ =\;& \exp\Big(-\frac{1}{2}(-x^T\Sigma^{-1}\mu_{-1}-\mu_{-1}^T\Sigma^{-1}x+\mu_{-1}^T\Sigma^{-1}\mu_{-1}+x^T\Sigma^{-1}\mu_1+\mu_1^T\Sigma^{-1}x-\mu_1^T\Sigma^{-1}\mu_1)+\log(\frac{1-\phi}{\phi})\Big)\\ =\;& \exp\Big(-\frac{1}{2}(-\mu_{-1}^T\Sigma^{-1}x-\mu_{-1}^T\Sigma^{-1}x+\mu_{-1}^T\Sigma^{-1}\mu_{-1}+\mu_1^T\Sigma^{-1}x+\mu_1^T\Sigma^{-1}x-\mu_1^T\Sigma^{-1}\mu_1)+\log(\frac{1-\phi}{\phi})\Big)\\ =\;& \exp\Big(-1\cdot\big((\mu_1-\mu_{-1})^T\Sigma^{-1}x+(\frac{1}{2}\mu_{-1}^T\Sigma^{-1}\mu_{-1}-\frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1-\log\big(\frac{1-\phi}{\phi}\big))\big)\Big)\,. \end{aligned}$$

We choose

$$\theta = \Sigma^{-1}(\mu_1-\mu_{-1})\,,\tag{8}$$

$$\theta_0 = \frac{1}{2}\mu_{-1}^T\Sigma^{-1}\mu_{-1}-\frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1-\log\big(\frac{1-\phi}{\phi}\big)\,,\tag{9}$$

therefore, we have

$$\begin{aligned} p(y=1|x;\Theta) &= \frac{1}{1+\exp(-1\cdot(\theta^T x+\theta_0))}\,, \\ p(y=-1|x;\Theta) &= 1-p(y=1|x;\Theta) \\ &= 1-\frac{1}{1+\exp(-1\cdot(\theta^T x+\theta_0))} \\ &= \frac{\exp(-(\theta^T x+\theta_0))}{1+\exp(-(\theta^T x+\theta_0))} \\ &= \frac{1}{1+\exp(\theta^T x+\theta_0)} \\ &= \frac{1}{1+\exp(-(-1)\cdot(\theta^T x+\theta_0))}\,. \end{aligned}$$

Problem 3(b)(c). Write the log-likelihood as follows.

$$\begin{aligned} & l(\phi,\mu_{-1},\mu_1,\Sigma) \\ = \;& \log\prod_{i=1}^m p(x^{(i)}|y^{(i)};\mu_{-1},\mu_1,\Sigma)p(y^{(i)};\phi) \\ = \;& -\sum_{i=1}^m \log(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}} - \frac{1}{2}\sum_{i=1}^m (x^{(i)}-\mu_{y^{(i)}})^T\Sigma^{-1}(x^{(i)}-\mu_{y^{(i)}}) + \sum_{i=1}^m \log(\phi^{1\{(y^{(i)}=1)\}}(1-\phi)^{1-1\{(y^{(i)}=1)\}}) \end{aligned}$$

First, consider ϕ. Note that only the third term of the log-likelihood is relevant.

$$\begin{aligned} \frac{\partial l}{\partial\phi} &= \frac{\partial}{\partial\phi}\sum_{i=1}^m \log p(y^{(i)};\phi) \\ &= \frac{\partial}{\partial\phi}\sum_{i=1}^m 1\{(y^{(i)}=1)\}\log(\phi) + (1-1\{(y^{(i)}=1)\})\log(1-\phi) \\ &= \frac{\sum_{i=1}^m 1\{(y^{(i)}=1)\}}{\phi} - \frac{m-\sum_{i=1}^m 1\{(y^{(i)}=1)\}}{1-\phi}\,. \end{aligned}$$

Setting this to zero and solving for ϕ gives the desired MLE of ϕ.

Second, consider μ_1. Write $\mu_{y^{(i)}} = \mu_1 1\{y^{(i)} = 1\} + \mu_{-1} 1\{y^{(i)} = -1\}$ and note that only the second term of the log-likelihood is relevant. Expand the expression and consider those terms involving μ_1,

$$
\begin{aligned}
&\nabla_{\mu_1} l \\
&= -\frac{1}{2}\nabla_{\mu_1}\Big(\sum_{i=1}^{m}\big(-x^{(i)T}\Sigma^{-1}\mu_1 1\{y^{(i)}=1\} - \mu_1^T 1\{y^{(i)}=1\}\Sigma^{-1}x^{(i)} + \mu_1^T 1\{y^{(i)}=1\}\Sigma^{-1}\mu_1 1\{y^{(i)}=1\} \\
&\quad + \mu_1^T 1\{y^{(i)}=1\}\mu_{-1}1\{y^{(i)}=-1\} + \mu_{-1}^T\mu_1 1\{y^{(i)}=1\}1\{y^{(i)}=-1\}\big)\Big) \\
&= -\frac{1}{2}\nabla_{\mu_1}\Big(\sum_{i=1}^{m}\big(-2x^{(i)T}\Sigma^{-1}\mu_1 + \mu_1^T\Sigma^{-1}\mu_1\big)1\{y^{(i)}=1\}\Big) \\
&= -\frac{1}{2}\sum_{i=1}^{m}(-2\Sigma^{-1}x^{(i)} + 2\Sigma^{-1}\mu_1)1\{y^{(i)}=1\} \\
&= \sum_{i=1}^{m}\Sigma^{-1}(x^{(i)}-\mu_1)1\{y^{(i)}=1\}\,.
\end{aligned}
$$

Setting this to zero and solving for μ_1 gives the desired MLE of μ_1. Similar calculations can be used to derive the MLE of μ_{-1}.

Lastly, consider Σ and let $\Sigma_1 = \Sigma^{-1}$. Note that only the first two terms of log-likelihood are relevant. Using the following two facts,

$$
\begin{aligned}
\nabla_X |X| &= |X|(X^{-1})^T\,, &(10)\\
\nabla_X a^T X a &= \nabla tr(a^T X a) = \nabla tr(aa^T X) = aa^T\,, &(11)
\end{aligned}
$$

we have

$$
\begin{aligned}
\nabla_{\Sigma_1} l &= \nabla_{\Sigma_1}\sum_{i=1}^{m}\Big(\frac{1}{2}\log(|\Sigma_1|) - \frac{1}{2}(x^{(i)}-\mu_{y^{(i)}})^T\Sigma_1(x^{(i)}-\mu_{y^{(i)}})\Big) \\
&= \sum_{i=1}^{m}\Big(\frac{1}{2|\Sigma_1|}|\Sigma_1|(\Sigma_1^{-1})^T - \frac{1}{2}(x^{(i)}-\mu_{y^{(i)}})(x^{(i)}-\mu_{y^{(i)}})^T\Big) \\
&= \frac{1}{2}\sum_{i=1}^{m}\Big(\Sigma^T - (x^{(i)}-\mu_{y^{(i)}})(x^{(i)}-\mu_{y^{(i)}})^T\Big) \\
&= \frac{1}{2}\sum_{i=1}^{m}\Big(\Sigma - (x^{(i)}-\mu_{y^{(i)}})(x^{(i)}-\mu_{y^{(i)}})^T\Big)
\end{aligned}
$$

Setting this to zero and solving for Σ gives the desired MLE of Σ.

Problem 4(a). With $g(z) = f(Az)$, we now express its gradient $\nabla_z g(z)$ and Hessian $\nabla_z^2 g(z)$ in terms of $\nabla_x f(\cdot)$ and $\nabla_x^2 f(\cdot)$. By chain rule,

$$\begin{aligned}
\frac{\partial g(z)}{\partial z_j} &= \sum_{k=1}^{n} \frac{\partial f(Az)}{\partial (Az)_k} \cdot \frac{\partial (Az)_k}{\partial z_j} \\
&= \sum_{k=1}^{n} \frac{\partial f(Az)}{\partial (Az)_k} A_{kj} \\
&= \sum_{k=1}^{n} (\nabla_x f(Az))_k A_{kj} \\
&= A_j^T \nabla_x f(Az)\,,
\end{aligned} \tag{12}$$

where A_j denotes the j^{th} column of A, and $\nabla_x f(Az)$ means $\nabla_x f(\cdot)$ evaluated at Az. Therefore,

$$\nabla_z g(z) = A^T \nabla_x f(Az)\,. \tag{13}$$

For Hessian $\nabla_z^2 g(z)$, from (12),

$$\begin{aligned}
\frac{\partial^2 g(z)}{\partial z_i \partial z_j} &= \frac{\partial}{\partial z_i} \sum_{k=1}^{n} \frac{\partial f(Az)}{\partial (Az)_k} A_{kj} \\
&= \sum_{k=1}^{n} \sum_{l=1}^{n} \frac{\partial^2 f(Az)}{\partial (Az)_l \partial (Az)_k} A_{li} A_{kj}\,,
\end{aligned}$$

thus,

$$\nabla_z^2 g(z) = A^T \nabla_x^2 f(Az) A\,. \tag{14}$$

We now prove the linear invariance of Newton's method by induction. First, $z^{(0)} = 0 = A^{-1}x^{(0)}$. Then, as the induction hypothesis, we assume

$$z^{(i)} = A^{-1}x^{(i)}\,, \tag{15}$$

for some i. By the update rule of Newton's method, we have

$$x^{(i+1)} = x^{(i)} - \left(\nabla_x^2 f(x^{(i)})\right)^{-1} \nabla_x f(x^{(i)})\,. \tag{16}$$

Therefore, by (13), (14), (15), (16) and the update rule of Newton's method, we have

$$\begin{aligned}
z^{(i+1)} &= z^{(i)} - \left(\nabla_z^2 g(z^{(i)})\right)^{-1} \nabla_z g(z^{(i)}) \\
&= A^{-1}x^{(i)} - \left(A^T \nabla_x^2 f(Az^{(i)}) A\right)^{-1} A^T \nabla_x f(Az^{(i)}) \\
&= A^{-1}x^{(i)} - A^{-1}\left(\nabla_x^2 f(x^{(i)})\right)^{-1} (A^T)^{-1} A^T \nabla_x f(x^{(i)}) \\
&= A^{-1}\left(x^{(i)} - \left(\nabla_x^2 f(x^{(i)})\right)^{-1} \nabla_x f(x^{(i)})\right) \\
&= A^{-1}x^{(i+1)}\,.
\end{aligned}$$

Problem 4(b). No. Consider the following counter-example:

$$\begin{aligned} f(x) &= (2x-1)^2\,, \\ A &= \frac{1}{2}\,, \\ A^{-1} &= 2\,, \\ g(z) = f(\frac{1}{2}z) &= (z-1)^2\,, \end{aligned}$$

we have $f'(x) = 4(2x-1)$ and $g'(z) = 2(z-1)$. With $x^{(0)} = z^{(0)} = 0$,

$$\begin{aligned} x^{(1)} &= 0 - \alpha(-4) = 4\alpha\,, \\ z^{(1)} &= 0 - \alpha(-2) = 2\alpha\,, \end{aligned}$$

where $\alpha > 0$ is the learning rate. But obviously, $A^{-1}x^{(1)} = 8\alpha \neq 2\alpha = z^{(1)}$.

Problem 5(a)(i). Let $W_{ii} = \frac{1}{2}w^{(i)}, W_{ij} = 0$ for $i \neq j$. Note that $X\theta - y = [\theta^T x^{(i)} - y^{(i)}]_{1\leq i\leq m}$ is a m-dimensional column vector. So it is easy to verify that

$$\begin{aligned} J(\theta) &= \frac{1}{2}\sum_{i=1}^{m} w^{(i)}(\theta^T x^{(i)} - y^{(i)})^2 \\ &= \sum_{i=1}^{m}(\theta^T x^{(i)} - y^{(i)}) \cdot \frac{1}{2}w^{(i)} \cdot (\theta^T x^{(i)} - y^{(i)}) \\ &= (X\theta - y)^T W (X\theta - y)\,. \end{aligned}$$

Problem 5(a)(ii). We only need to consider all the "weights" in the weight matrix W are non-negative; otherwise, we can choose suitable θ such that $J(\theta) = -\infty$.

The method stated in the problem set is more straightforward, but this solution makes use of the results shown in the class directly and can be served as an alternative. Define the "square root" of W, denoted by $\sqrt{W}$, where $(\sqrt{W})_{ij} = \sqrt{W_{ij}}$ for all i, j. Note that $W = \sqrt{W}\sqrt{W}$ and $\sqrt{W}^T = \sqrt{W}$. Re-write the cost function as follows:

$$\begin{aligned} J(\theta) &= (X\theta - y)^T \sqrt{W}\sqrt{W}(X\theta - y) \\ &= (X\theta - y)^T \sqrt{W}^T \sqrt{W}(X\theta - y) \\ &= (\sqrt{W}X\theta - \sqrt{W}y)^T(\sqrt{W}X\theta - \sqrt{W}y)\,, \end{aligned}$$

which is the same form as we saw in the class, except that X is replaced by $\sqrt{W}X$ and y is replaced by $\sqrt{W}y$. Therefore, using the closed form shown in the class, the new value of θ in this weighted setting is given by

$$\begin{aligned} \theta &= \big((\sqrt{W}X)^T\sqrt{W}X\big)^{-1}(\sqrt{W}X)^T\sqrt{W}y \\ &= \big(X^T\sqrt{W}^T\sqrt{W}X\big)^{-1}X^T\sqrt{W}^T\sqrt{W}y \\ &= (X^TWX)^{-1}X^TWy\,. \end{aligned}$$

Problem 5(a)(iii). Maximum likelihood estimate of θ is obtained by

$$\begin{aligned}
\arg\max_{\theta} \log \prod_{i=1}^{m} p(y^{(i)}|x^{(i)};\theta) &= \arg\max_{\theta} \sum_{i=1}^{m} \log p(y^{(i)}|x^{(i)};\theta) \\
&= \arg\max_{\theta} \sum_{i=1}^{m} \Big(\log \frac{1}{\sqrt{2\pi}\sigma^{(i)}} - \frac{(y^{(i)} - \theta^T x^{(i)})^2}{2(\sigma^{(i)})^2} \Big) \\
&= \arg\min_{\theta} \sum_{i=1}^{m} \frac{(y^{(i)} - \theta^T x^{(i)})^2}{2(\sigma^{(i)})^2} \\
&= \arg\min_{\theta} \frac{1}{2} \sum_{i=1}^{m} \frac{1}{(\sigma^{(i)})^2} (y^{(i)} - \theta^T x^{(i)})^2 \,.
\end{aligned}$$

Choosing $w^{(i)} = \frac{1}{(\sigma^{(i)})^2}$ reduces the problem to weighted linear regression.

Problem 5(b)(i)(ii)(iii). See the files "Q5(b).ipynb" and "LOESS.py".

Problem 5(c)(i)(ii)(iii). See the files "Q5(c).ipynb" and "LOESS.py".