Homework 1

Solution 1. If a multivariate function is convex, its Hessian Matrix is positive semi-definite.

$$
\begin{aligned}
y &= \frac{1}{1+e^{-(w^T x+b)}} \\
\frac{dy}{dw} &= \frac{-xe^{-(w^T x+b)}}{1+e^{-(w^T x+b)}} = x(y-y^2) \\
\frac{d}{dw^T}(\frac{dy}{dw}) &= x(1-2y)(\frac{dy}{dw})^T = xx^T y(y-1)(1-2y)
\end{aligned}
$$

Because xx^T is congruent to identity matrix, xx^T is positive definite matrix. The range of y is (0, 1), when y ϵ (0.5, 1), y(y-1)(1-2y)<0, leading to $\frac{d}{dw^T}(\frac{dy}{dw})$ not positive definite. So $y = \frac{1}{1+e^{-(w^T x+b)}}$ is non-convex.

$$
\begin{aligned}
l(\boldsymbol{w}) &= \sum_{i=1}^{m}(-y_i(\boldsymbol{w^T x_i}+b)+\ln(1+e^{\boldsymbol{w^T x_i}+b})) \\
\frac{d}{dw^T}(\frac{dl}{dw}) &= \sum_{i=1}^{m} xx^T p_i(x|w)(1-p_i(x|w))
\end{aligned}
$$

Obviously, $p_i \epsilon(0,1)$, $p_i(x|\boldsymbol{w})(1-p_i(x|\boldsymbol{w})) \geq 0$, so $l(\boldsymbol{w}) = \sum_{i=1}^{m}(-y_i(\boldsymbol{w^T x_i}+b)+\ln(1+e^{\boldsymbol{w^T x_i}+b}))$ is a convex function.

Solution 2. We can rewrite $\sum_{j=1}^{M}|w_j|^q \leq \eta$ as

$$
\frac{1}{2}(\sum_{j=1}^{M}|w_j|^q - \eta) \leq 0
$$

where we have incorporated the $\frac{1}{2}$ scaling factor for convenience. Clearly this does not affect the constraint. We can combine this with $E_D(w) = \frac{1}{2}\sum_{n=1}^{N}\{t_n - w^T\phi(x_n)\}^2$ to obtain the Lagrangian function

$$
L(\boldsymbol{w},\lambda) = \frac{1}{2}\sum_{n=1}^{N}\{t_n - \boldsymbol{w^T\phi}(x_n)\}^2 + \frac{\lambda}{2}(\sum_{j=1}^{M}|w_j|^q - \eta)
$$

and by comparing this with $\frac{1}{2}\sum_{n=1}^{N}\{t_n - \boldsymbol{w^T}\boldsymbol{\phi}(x_n)\}^2 + \frac{\lambda}{2}\sum_{j=1}^{M}|w_j|^q$ we see immediately that they are identical in their dependence on $\boldsymbol{w}$.
Now suppose we choose a specific value of $\lambda > 0$ and minimize $\frac{1}{2}\sum_{n=1}^{N}\{t_n - \boldsymbol{w^T}\boldsymbol{\phi}(x_n)\}^2 + \frac{\lambda}{2}\sum_{j=1}^{M}|w_j|^q$. Denoting the resulting value of $\boldsymbol{w}$ by $\boldsymbol{w}^*(\lambda)$, and using the KKT condition, we see that the value of η is given by

$$\eta = \sum_{j=1}^{M} |w_j^*(\lambda)|^q.$$

Solution 3. If we define $\mathbf{R} = diag(r_1, ..., r_N)$ to be a diagonal matrix containing the weighting coefficients, then we can write the weighted sum-of-squares cost function in the form

$$E_D(\mathbf{w}) = \frac{1}{2}(\mathbf{t} - \mathbf{\Phi w})^T \mathbf{R}(\mathbf{t} - \mathbf{\Phi w}).$$

Setting the derivative with respect to $\mathbf{w}$ to zero, and re-arranging, then gives

$$\mathbf{w}^* = (\mathbf{\Phi}^T \mathbf{R} \mathbf{\Phi})^{\mathbf{-1}} \mathbf{\Phi}^T \mathbf{R} \mathbf{t}$$

which reduces to the standard solution for the case $\mathbf{R} = \mathbf{I}$.

Solution 4. Let $\mu_{ic} = P(y_i = c|\mathbf{x}_i, \mathbf{W}), y_{ic} = \mathbf{1}\{y_i = c\}$.
(a)

$$l(W) = log \prod_{i=1}^{n}\prod_{c=1}^{C} \mu_{ic}^{y_i c} = \sum_{i=1}^{n}\sum_{c=1}^{C} y_{ic} log \mu_{ic} = \sum_{i=1}^{n}\left(\sum_{c=1}^{C} y_{ic}\mathbf{w}_c^T\mathbf{x}_i - log\sum_{c'=1}^{C} exp(\mathbf{w}_{c'}^T\mathbf{x}_i)\right)$$

(b)

$$\begin{aligned}
g_c(W) &= \frac{\partial}{\partial \mathbf{w}_c}\sum_{i=1}^{n}\left(\sum_{c=1}^{C} y_{ic}\mathbf{w}_c^T\mathbf{x}_i - log\sum_{c'=1}^{C} exp(\mathbf{w}_{c'}^T\mathbf{x}_i)\right) \\
&= \sum_{i=1}^{n}\left(\frac{\partial}{\partial \mathbf{w}_c}\sum_{c=1}^{C} y_{ic}\mathbf{w}_c^T\mathbf{x}_i - \frac{\partial}{\partial \mathbf{w}_c} log\sum_{c'=1}^{C} exp(\mathbf{w}_{c'}^T\mathbf{x}_i)\right) \\
&= \sum_{i=1}^{n}\left(y_{ic}\mathbf{x}_i - \frac{\frac{\partial}{\partial \mathbf{w}_c}\sum_{c'=1}^{C} exp(\mathbf{w}_{c'}^T\mathbf{x}_i)}{\sum_{c'=1}^{C} exp(\mathbf{w}_{c'}^T\mathbf{x}_i)}\right) \\
&= \sum_{i=1}^{n}\left(y_{ic}\mathbf{x}_i - \frac{exp(\mathbf{w}_c^T\mathbf{x}_i)\mathbf{x}_i}{\sum_{c'=1}^{C} exp(\mathbf{w}_{c'}^T\mathbf{x}_i)}\right) \\
&= \sum_{i=1}^{n}(y_{ic} - \mu_{ic})\mathbf{x}_i
\end{aligned}$$

(c) $\delta_{cc'}$ denotes the Dirac delta function and is equal to one if $c = c'$ and zero otherwise.

$$
\begin{aligned}
H_{c,c'}(W) &= \frac{\partial}{\partial \mathbf{w}_c} g_{c'}(W) \\
&= \frac{\partial}{\partial \mathbf{w}_c} \sum_{i=1}^{n} \left(y_{ic'} \mathbf{x}_i - \frac{exp(\mathbf{w}_{c'}^T \mathbf{x}_i)\mathbf{x}_i}{\sum_{c''=1}^{C} exp(\mathbf{w}_{c''}^T \mathbf{x}_i)} \right) \\
&= -\sum_{i=1}^{n} \frac{\partial}{\partial \mathbf{w}_c} \frac{exp(\mathbf{w}_{c'}^T \mathbf{x}_i)\mathbf{x}_i}{\sum_{c''=1}^{C} exp(\mathbf{w}_{c''}^T \mathbf{x}_i)} \\
&= -\sum_{i=1}^{n} \frac{\delta_{cc'} exp(\mathbf{w}_{c'}^T \mathbf{x}_i)\mathbf{x}_i \mathbf{x}_i^T \sum_{c''=1}^{C} exp(\mathbf{w}_{c''}^T \mathbf{x}_i) - exp(\mathbf{w}_{c'}^T \mathbf{x}_i)\mathbf{x}_i \mathbf{x}_i^T exp(\mathbf{w}_{c}^T \mathbf{x}_i)}{\left(\sum_{c''=1}^{C} exp(\mathbf{w}_{c''}^T \mathbf{x}_i)\right)^2} \\
&= -\sum_{i=1}^{n} (\delta_{cc'} \mu_{ic} - \mu_{ic'} \mu_{ic}) \mathbf{x}_i \mathbf{x}_i^T \\
&= \sum_{i=1}^{n} \mu_{ic} (\mu_{ic'} - \delta_{cc'}) \mathbf{x}_i \mathbf{x}_i^T
\end{aligned}
$$