1. Maximiser $Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(q)})$ par rapport à $\boldsymbol{\alpha}$ revient à maximiser la fonction

$$\begin{aligned} Q_\alpha(\boldsymbol{\theta};\boldsymbol{\theta}^{(q)}) &= \sum_{t=1}^n \tau_t(\boldsymbol{\theta}^{(q)}) \ln \mathcal{N}(y_t;\boldsymbol{\alpha}^\top \boldsymbol{x}_t,\sigma^2) \\ &= \sum_{t=1}^n \tau_t(\boldsymbol{\theta}^{(q)}) \ln \left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_t-\boldsymbol{\alpha}^\top \boldsymbol{x}_t)^2\right)\right] \\ &= \sum_{t=1}^n \tau_t(\boldsymbol{\theta}^{(q)}) \ln \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2}\sum_{t=1}^n \tau_t(\boldsymbol{\theta}^{(q)})(y_t-\boldsymbol{\alpha}^\top \boldsymbol{x}_t)^2. \end{aligned}$$

   En dérivant par rapport à $\boldsymbol{\alpha}$ on obtient la dérivée suivante

$$\begin{aligned} \frac{\partial Q_\alpha(\boldsymbol{\theta};\boldsymbol{\theta}^{(q)})}{\partial \boldsymbol{\alpha}} &= -\frac{1}{2\sigma^2}\sum_{t=1}^n \tau_t(\boldsymbol{\theta}^{(q)}) \times 2 \times (-\boldsymbol{x}_t) \times (y_t - \boldsymbol{x}_t^\top \boldsymbol{\alpha}). \\ &= \frac{1}{\sigma^2}\sum_{t=1}^n \tau_t(\boldsymbol{\theta}^{(q)}) \boldsymbol{x}_t (y_t - \boldsymbol{x}_t^\top \boldsymbol{\alpha}) \\ &\propto \sum_{t=1}^n \tau_t(\boldsymbol{\theta}^{(q)}) y_t \boldsymbol{x}_t - \sum_{t=1}^n \tau_t(\boldsymbol{\theta}^{(q)}) \boldsymbol{x}_t \boldsymbol{x}_t^\top \boldsymbol{\alpha}. \end{aligned}$$

   et en annulant cette dérivée pour $\boldsymbol{\alpha}$ on trouve :

$$\Big[\sum_{t=1}^n \tau_t(\boldsymbol{\theta}^{(q)}) \boldsymbol{x}_t \boldsymbol{x}_t^\top\Big]\boldsymbol{\alpha} = \sum_{t=1}^n \tau_t(\boldsymbol{\theta}^{(q)}) y_t \boldsymbol{x}_t,$$

   et l'équation de mise à jour est donc donnée par

$$\boldsymbol{\alpha}^{(q+1)} = \Big[\sum_{t=1}^n \tau_t(\boldsymbol{\theta}^{(q)}) \boldsymbol{x}_t \boldsymbol{x}_t^\top\Big]^{-1} \sum_{t=1}^n \tau_t(\boldsymbol{\theta}^{(q)}) y_t \boldsymbol{x}_t. \quad (1)$$

2. De la même façon, en maximisant, par rapport à $\boldsymbol{\beta}$, la fonction

$$Q_{\boldsymbol{\beta}}(\boldsymbol{\theta};\boldsymbol{\theta}^{(q)}) = \sum_{t=1}^n (1-\tau_t(\boldsymbol{\theta}^{(q)})) \ln \mathcal{N}(y_t;\boldsymbol{x}_t^\top \boldsymbol{\beta},\sigma^2)$$

   on obtient

$$\boldsymbol{\beta}^{(q+1)} = \Big[\sum_{t=1}^n (1-\tau_t(\boldsymbol{\theta}^{(q)})) \boldsymbol{x}_t \boldsymbol{x}_t^\top\Big]^{-1} \sum_{t=1}^n (1-\tau_t(\boldsymbol{\theta}^{(q)})) y_t \boldsymbol{x}_t. \quad (2)$$

3. Maximiser $Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(q)})$ par rapport à σ^2 revient à maximiser la fonction

$$\begin{aligned} Q_{\sigma^2}(\boldsymbol{\theta};\boldsymbol{\theta}^{(q)}) &= \sum_{t=1}^n \Big\{\tau_t(\boldsymbol{\theta}^{(q)}) \ln \mathcal{N}(y_t;\boldsymbol{\alpha}^\top \boldsymbol{x}_t,\sigma^2) + (1-\tau_t(\boldsymbol{\theta}^{(q)})) \ln \mathcal{N}(y_t;\boldsymbol{\beta}^\top \boldsymbol{x}_t,\sigma^2)\Big\} \\ &= \sum_{t=1}^n \tau_t(\boldsymbol{\theta}^{(q)}) \ln \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2}\sum_{t=1}^n \tau_t(\boldsymbol{\theta}^{(q)})(y_t-\boldsymbol{x}_t^\top \boldsymbol{\alpha})^2 \\ &\quad + \sum_{t=1}^n (1-\tau_t(\boldsymbol{\theta}^{(q)})) \ln \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2}\sum_{t=1}^n (1-\tau_t(\boldsymbol{\theta}^{(q)}))(y_t-\boldsymbol{x}_t^\top \boldsymbol{\beta})^2. \\ &= n \ln \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2}\sum_{t=1}^n \tau_t(\boldsymbol{\theta}^{(q)})(y_t-\boldsymbol{x}_t^\top \boldsymbol{\alpha})^2 - \frac{1}{2\sigma^2}\sum_{t=1}^n (1-\tau_t(\boldsymbol{\theta}^{(q)}))(y_t-\boldsymbol{x}_t^\top \boldsymbol{\beta})^2. \end{aligned}$$

En dérivant par rapport à σ^2 on obtient

$$\frac{\partial Q_{\sigma^2}(\boldsymbol{\theta};\boldsymbol{\theta}^{(q)})}{\partial \sigma^2} = -n\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{t=1}^{n}\tau_t(\boldsymbol{\theta}^{(q)})(y_t - \boldsymbol{x}_t^\top\boldsymbol{\alpha})^2 + \frac{1}{2\sigma^4}\sum_{t=1}^{n}(1-\tau_t(\boldsymbol{\theta}^{(q)}))(y_t - \boldsymbol{x}_t^\top\boldsymbol{\beta})^2.$$

et en annulant cette dérivée (et en multipliant ses termes par $2\sigma^4$) on trouve

$$n\sigma^2 = \sum_{t=1}^{n}\tau_t(\boldsymbol{\theta}^{(q)})(y_t - \boldsymbol{x}_t^\top\boldsymbol{\alpha})^2 + \sum_{t=1}^{n}(1-\tau_t(\boldsymbol{\theta}^{(q)}))(y_t - \boldsymbol{x}_t^\top\boldsymbol{\beta})^2.$$

soit, l'équation de mise à jour suivante pour σ^2

$$\sigma^{2(q+1)} = \frac{1}{n}\sum_{t=1}^{n}\left\{\tau_t(\boldsymbol{\theta}^{(q)})(y_t - \boldsymbol{x}_t^\top\boldsymbol{\alpha}^{(q+1)})^2 + (1-\tau_t(\boldsymbol{\theta}^{(q)}))(y_t - \boldsymbol{x}_t^\top\boldsymbol{\beta}^{(q+1)})^2\right\} \tag{3}$$

où $\boldsymbol{\alpha}^{(q+1)}$ et $\boldsymbol{\beta}^{(q+1)}$ sont les mises à jour respectives de $\boldsymbol{\alpha}$ et $\boldsymbol{\beta}$ données par (1) et (2), respectivement.

4. La mise à jour du vecteur paramètre $\boldsymbol{w}$ des poids logistiques s'effectue en maximisant $Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(q)})$ par rapport à $\boldsymbol{w}$ par l'algorithme de Newton-Raphson qui consiste à partir d'un modèle initial de paramètres $\boldsymbol{w}^{(0)}$ et à mettre à jour, à chaque itération m, le vecteur paramètre $\boldsymbol{w}$ selon l'équation de mise à jour suivante :

$$\boldsymbol{w}^{(q+1)} = \boldsymbol{w}^{(q)} - \left(\frac{\partial^2 Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(q)})}{\partial\boldsymbol{w}\partial\boldsymbol{w}^\top}\right)^{-1}_{\boldsymbol{w}=\boldsymbol{w}^{(q)}}\left(\frac{\partial Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(q)})}{\partial\boldsymbol{w}}\right)_{\boldsymbol{w}=\boldsymbol{w}^{(q)}} \tag{4}$$

Soit $Q_{\boldsymbol{w}}(\boldsymbol{\theta};\boldsymbol{\theta}^{(q)})$ la partie de $Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(q)})$ qui dépend de $\boldsymbol{w}$, on a

$$\begin{aligned}
Q_{\boldsymbol{w}}(\boldsymbol{\theta};\boldsymbol{\theta}^{(q)}) &= \sum_{t=1}^{n}\left\{\tau_t(\boldsymbol{\theta}^{(q)})\ln\pi(t;\boldsymbol{w}) + (1-\tau_t(\boldsymbol{\theta}^{(q)}))\ln\left[(1-\pi(t;\boldsymbol{w}))\right]\right\}\\
&= \sum_{t=1}^{n}\left\{\tau_t(\boldsymbol{\theta}^{(q)})\ln\frac{\exp(\boldsymbol{w}^\top\boldsymbol{x}_t)}{1+\exp(\boldsymbol{w}^\top\boldsymbol{x}_t)} + (1-\tau_t(\boldsymbol{\theta}^{(q)}))\ln\left(1-\frac{\exp(\boldsymbol{w}^\top\boldsymbol{x}_t)}{1+\exp(\boldsymbol{w}^\top\boldsymbol{x}_t)}\right)\right\}\\
&= \sum_{t=1}^{n}\left\{\tau_t(\boldsymbol{\theta}^{(q)})\ln\frac{\exp(\boldsymbol{w}^\top\boldsymbol{x}_t)}{1+\exp(\boldsymbol{w}^\top\boldsymbol{x}_t)} + (1-\tau_t(\boldsymbol{\theta}^{(q)}))\ln\frac{1}{1+\exp(\boldsymbol{w}^\top\boldsymbol{x}_t)}\right\}\\
&= \sum_{t=1}^{n}\left\{\tau_t(\boldsymbol{\theta}^{(q)})\boldsymbol{w}^\top\boldsymbol{x}_t - \ln\left(1+\exp(\boldsymbol{w}^\top\boldsymbol{x}_t)\right)\right\}.
\end{aligned}$$

Le vecteur du gradient de $Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(q)})$ par rapport à $\boldsymbol{w}$ est donné par

$$\begin{aligned}
\frac{\partial Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(q)})}{\partial\boldsymbol{w}} &= \frac{\partial Q_{\boldsymbol{w}}(\boldsymbol{\theta};\boldsymbol{\theta}^{(q)})}{\partial\boldsymbol{w}} = \frac{\partial\sum_{t=1}^{n}\left\{\tau_t(\boldsymbol{\theta}^{(q)})\boldsymbol{w}^\top\boldsymbol{x}_t - \ln\left(1+\exp(\boldsymbol{w}^\top\boldsymbol{x}_t)\right)\right\}}{\partial\boldsymbol{w}}\\
&= \sum_{i=1}^{n}\tau_t(\boldsymbol{\theta}^{(q)})\boldsymbol{x}_t - \frac{\boldsymbol{x}_t\exp(\boldsymbol{w}^\top\boldsymbol{x}_t)}{1+\exp(\boldsymbol{w}^\top\boldsymbol{x}_t)}\\
&= \sum_{i=1}^{n}\boldsymbol{x}_t\left(\tau_t(\boldsymbol{\theta}^{(q)}) - \frac{\exp(\boldsymbol{w}^\top\boldsymbol{x}_t)}{1+\exp(\boldsymbol{w}^\top\boldsymbol{x}_t)}\right)\\
&= \sum_{i=1}^{n}\boldsymbol{x}_t(\tau_t(\boldsymbol{\theta}^{(q)}) - \pi(\boldsymbol{x}_i;\boldsymbol{\theta})).
\end{aligned} \tag{5}$$

La matrice hessienne de $Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(q)})$ est donnée par

$$\begin{aligned}
\frac{\partial^2 Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(q)})}{\partial \boldsymbol{w}\partial \boldsymbol{w}^\top} &= \frac{\sum_{i=1}^n \boldsymbol{x}_t \left(\tau_t(\boldsymbol{\theta}^{(q)}) - \frac{\exp(\boldsymbol{w}^\top \boldsymbol{x}_t)}{1+\exp(\boldsymbol{w}^\top \boldsymbol{x}_t)}\right)}{\partial \boldsymbol{w}^\top} \\
&= -\sum_{i=1}^n \boldsymbol{x}_t \frac{\boldsymbol{x}_t^\top \exp(\boldsymbol{w}^\top \boldsymbol{x}_t)}{(1+\exp(\boldsymbol{w}^\top \boldsymbol{x}_t))^2} \\
&= -\sum_{i=1}^n \boldsymbol{x}_t \boldsymbol{x}_t^\top \frac{\exp(\boldsymbol{w}^\top \boldsymbol{x}_t)}{1+\exp(\boldsymbol{w}^\top \boldsymbol{x}_t)} \frac{1}{1+\exp(\boldsymbol{w}^\top \boldsymbol{x}_t)} \\
&= -\sum_{i=1}^n \boldsymbol{x}_t \boldsymbol{x}_t^\top \pi(\boldsymbol{x}_i;\boldsymbol{\theta})(1-\pi(\boldsymbol{x}_i;\boldsymbol{\theta})) \qquad (6)
\end{aligned}$$

L'algorithme de Newton-Raphson est donc défini par la mise à jour itérative de $\boldsymbol{w}$ selon l'expression suivante :

$$\begin{aligned}
\boldsymbol{w}^{(q+1)} &= \boldsymbol{w}^{(q)} - \left(\frac{\partial^2 L(\boldsymbol{\theta})}{\partial \boldsymbol{w}\partial \boldsymbol{w}^\top}\right)^{-1}_{\boldsymbol{w}=\boldsymbol{w}^{(q)}} \left(\frac{\partial^2 L(\boldsymbol{\theta})}{\partial \boldsymbol{w}\partial \boldsymbol{w}^\top}\right)_{\boldsymbol{w}=\boldsymbol{w}^{(q)}} \\
&= \boldsymbol{w}^{(q)} + \left(\sum_{t=1}^n \boldsymbol{x}_t \boldsymbol{x}_t^\top \pi(\boldsymbol{x}_t;\boldsymbol{w}^{(q)})(1-\pi(\boldsymbol{x}_t;\boldsymbol{w}^{(q)}))\right)^{-1} \sum_{t=1}^n \boldsymbol{x}_t(\tau_t(\boldsymbol{\theta}^{(q)}) - \pi(\boldsymbol{x}_t;\boldsymbol{w}^{(q)})) \quad (7)
\end{aligned}$$

soit sous forme matricielle

$$\begin{aligned}
\boldsymbol{w}^{(q+1)} &= \boldsymbol{w}^{(q)} + \left(\mathbf{X}^\top \mathbf{W}^{(q)} \mathbf{X}\right)^{-1} \mathbf{X}^\top (\boldsymbol{\tau} - \boldsymbol{p}^{(q)}) \\
&= (\mathbf{X}^\top \mathbf{W}^{(q)} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W}^{(q)} \boldsymbol{z} \qquad (8)
\end{aligned}$$

avec $\mathbf{X}$ la matrice de dimension $n \times 2$ construire en concaténant les prédicteurs $\boldsymbol{x}_t$ pour $t = 1, \ldots, n$: $\mathbf{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^\top$, où :

- $\mathbf{X}$ est la matrice de dimensions $n \times 2$ de lignes $(1, t)$ pour $t = 1, \ldots, n$;
- $\boldsymbol{\tau}$ est le vecteur colonne de dimension $n \times 1$ des probabilités a posteriori $\tau_t(\boldsymbol{\theta})$: $\boldsymbol{\tau} = (\tau_1(\boldsymbol{\theta}), \ldots, \tau_n(\boldsymbol{\theta}))^\top$;
- $\boldsymbol{p}$ est le vecteur colonne de dimension $n \times 1$ des probabilités : $\boldsymbol{p} = (\pi(\boldsymbol{x}_1;\boldsymbol{\theta}), \ldots, \pi(\boldsymbol{x}_n;\boldsymbol{\theta}))^\top$;
- $\mathbf{W}$ est la matrice diagonale de dimension $n \times n$ d'éléments $\pi(\boldsymbol{x}_1;\boldsymbol{\theta})\,(1-\pi(\boldsymbol{x}_n;\boldsymbol{\theta}))$;
- $\boldsymbol{z} = \mathbf{X}\boldsymbol{w}^{(q)} + (\mathbf{W}^{(q)})^{-1}(\boldsymbol{y} - \boldsymbol{\tau}^{(q)})$: la réponse courante.
- D'où l'appellation IRLS (Iteratively reweighted least squares) moindres carrés pondérés itératifs. Noter que cet algorithme est en réalité itéré à chaque itération de l'algorithme EM, mais l'énoncé ici pose le problème de façon plus simple pour alléger les notations ; l'algorithme EM dans ce cas correspond à un algorithme EM généralisé (GEM) plutôt qu'à un algorithme EM au sens strict, puisque à l'étape E on augmente la fonction Q plutôt que de la maximiser.