Solution 1: Logistic Regression Basics

a) As we would expect, the two formulations are equivalent (up to reparameterization). In order to see this, consider the softmax function components:

$$\pi_1(\mathbf{x} \mid \boldsymbol{\theta}) = \frac{\exp(\boldsymbol{\theta}_1^\top \mathbf{x})}{\exp(\boldsymbol{\theta}_1^\top \mathbf{x}) + \exp(\boldsymbol{\theta}_2^\top \mathbf{x})}$$
$$\pi_2(\mathbf{x} \mid \boldsymbol{\theta}) = \frac{\exp(\boldsymbol{\theta}_2^\top \mathbf{x})}{\exp(\boldsymbol{\theta}_1^\top \mathbf{x}) + \exp(\boldsymbol{\theta}_2^\top \mathbf{x})},$$

where $\pi_1(\mathbf{x} \mid \boldsymbol{\theta}) + \pi_2(\mathbf{x} \mid \boldsymbol{\theta}) = 1$.

$$\begin{aligned} \Longrightarrow \pi_1(\mathbf{x} \mid \boldsymbol{\theta}) &= \frac{1}{\frac{\exp(\boldsymbol{\theta}_1^\top \mathbf{x}) + \exp(\boldsymbol{\theta}_2^\top \mathbf{x})}{\exp(\boldsymbol{\theta}_1^\top \mathbf{x})}} \\ &= \frac{1}{1 + \exp(\boldsymbol{\theta}_2^\top \mathbf{x} - \boldsymbol{\theta}_1^\top \mathbf{x})} \\ &= \frac{1}{1 + \exp(-\boldsymbol{\theta}^\top \mathbf{x})} \\ &= \pi(\mathbf{x} \mid \boldsymbol{\theta}), \end{aligned}$$

i.e., the binary-case logistic function, if we set $\boldsymbol{\theta} := \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2$.

b) The joint likelihood is easy to compute with the *iid* assumption we are willing to make. It is simply given by the product over all individual likelihoods:

$$\mathcal{L}(\boldsymbol{\theta}) = \prod_{i=1}^{n} \prod_{j=1}^{g} \pi_j \left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right)^{\mathbb{I}(y^{(i)}=j)}.$$

c) Right now, $\mathcal{L}(\boldsymbol{\theta})$ does not look anything like an empirical risk function. However, we will arrive there by some simple transformations you might recall from the first exercise sheet:

- First we convert our *maximum* likelihood problem into an empirical risk *minimization* problem:

$$\arg\max_{\boldsymbol{\theta} \in \Theta} \mathcal{L}(\boldsymbol{\theta}) = \arg\min_{\boldsymbol{\theta} \in \Theta} -\mathcal{L}(\boldsymbol{\theta}).$$

- Then we get rid of the (outer) product over all observations, which we would like to turn into a sum. This is achieved by taking the log, a strictly monotonic transformation that has no effect on the optimizer (recall that $\log(a \cdot b) = \log a + \log b$):

$$\arg\min_{\boldsymbol{\theta} \in \Theta} \prod_{i=1}^{n} -\mathcal{L}_i(\boldsymbol{\theta}) = \arg\min_{\boldsymbol{\theta} \in \Theta} \sum_{i=1}^{n} -\ell_i(\boldsymbol{\theta}).$$

  The inner product over all classes also becomes a sum in this new formulation (before, we wanted all probability functions but the one corresponding to the true class to become 1 factors, now we want them to become 0 summands):

$$\arg\min_{\boldsymbol{\theta} \in \Theta} \sum_{i=1}^{n} -\ell_i(\boldsymbol{\theta}) = \arg\min_{\boldsymbol{\theta} \in \Theta} \sum_{i=1}^{n} - \left(\sum_{j=1}^{g} \mathbb{I}(y^{(i)} = j) \log \pi_j \left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right) \right)$$

- And we have already found an expression that is conformal with the empirical risk minimization principle:

$$\hat{\boldsymbol{\theta}}_{\text{MLE}} = \hat{\boldsymbol{\theta}}_{\text{ERM}} = \arg\min_{\boldsymbol{\theta}\in\Theta} \underbrace{\sum_{i=1}^{n} \underbrace{-\left(\sum_{j=1}^{g} \mathbb{I}(y^{(i)} = j) \log \pi_j \left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right)\right)}_{L\left(y^{(i)}, f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right)\right)}}_{\mathcal{R}_{\text{emp}}(\boldsymbol{\theta})}$$

As the above transformations are universally applicable, we can always use the negative log-likelihood (NLL) as a loss function in empirical risk minimization (not every loss function, however, has a corresponding likelihood formulation).

d) The k-th discriminant function has the following form:

$$\pi_k(\mathbf{x} \mid \boldsymbol{\theta}) = \frac{\exp(\boldsymbol{\theta}_k^\top \mathbf{x})}{\sum\limits_{j=1}^{g} \exp(\boldsymbol{\theta}_j^\top \mathbf{x})} \quad \in [0, 1],$$

and $\sum_{k=1}^{g} \pi_k(\mathbf{x} \mid \boldsymbol{\theta}) = 1$. This sum-one constraint means that one set of parameters is actually redundant: if we know the first $g-1$ discriminant functions, the g-th one is fully specified. Therefore, we set $\hat{\boldsymbol{\theta}}_g = \mathbf{0}$ and compute $\mathbb{P}(\hat{y} = g \mid \mathbf{x}, \boldsymbol{\theta}) = 1 - \sum_{k=1}^{g-1} \hat{\pi}_k(\mathbf{x} \mid \boldsymbol{\theta})$.

The highest of the thus estimated posterior class probabilities then determines the actual class label prediction:

$$\hat{y} = \arg\max_{k\in\{1,\dots,g\}} \hat{\pi}_k(\mathbf{x} \mid \boldsymbol{\theta}).$$

e) In order to state the hypothesis space for the multiclass case we can define a length-g vector of class-individual probability functions that results from applying the softmax function $\mathcal{S}$:

$$[\pi_k(\mathbf{x} \mid \boldsymbol{\theta})]_{k=1,2,\dots,g} = \begin{pmatrix}\pi_1(\mathbf{x} \mid \boldsymbol{\theta}) & \pi_2(\mathbf{x} \mid \boldsymbol{\theta}) & \dots & \pi_g(\mathbf{x} \mid \boldsymbol{\theta})\end{pmatrix}^\top =: \mathcal{S}(\mathbf{x} \mid \boldsymbol{\theta}) \ \in [0, 1]^g.$$

Our parameters are now matrix-valued, where every class-individual parameter vector is of length p, such that

$$\boldsymbol{\theta} = \begin{pmatrix}\boldsymbol{\theta}_1 & \boldsymbol{\theta}_2 & \dots & \boldsymbol{\theta}_g\end{pmatrix} = \left(\begin{pmatrix}\theta_{1,1} \\ \vdots \\ \theta_{1,p}\end{pmatrix} \quad \begin{pmatrix}\theta_{2,1} \\ \vdots \\ \theta_{2,p}\end{pmatrix} \quad \dots \quad \begin{pmatrix}\theta_{g,1} \\ \vdots \\ \theta_{g,p}\end{pmatrix}\right) \quad \in \Theta = \mathbb{R}^{p\times g}$$

Then we have for our hypothesis space:

$$\mathcal{H} = \left\{ \mathcal{S}_{\boldsymbol{\theta}} : \mathcal{X} \to [0, 1]^g \;\middle|\; \mathcal{S}(\mathbf{x} \mid \boldsymbol{\theta}) = \left[\frac{\exp(\boldsymbol{\theta}_k^\top \mathbf{x})}{\sum\limits_{j=1}^{g} \exp(\boldsymbol{\theta}_j^\top \mathbf{x})}\right]_{k=1,2,\dots,g}, \ \boldsymbol{\theta} \in \mathbb{R}^{p\times g} \right\}$$

Solution 2: Decision Boundaries & Thresholds in Logistic Regression

a) We evaluate

$$\begin{aligned}
\hat{y} = 1 \Leftrightarrow \hat{\pi}(\mathbf{x}) &= \frac{1}{1 + \exp(-\hat{\boldsymbol{\theta}}^\top \mathbf{x})} \geq \alpha \\
&\Leftrightarrow \ 1 + \exp(-\hat{\boldsymbol{\theta}}^\top \mathbf{x}) \leq \frac{1}{\alpha} \\
&\Leftrightarrow \ \exp(-\hat{\boldsymbol{\theta}}^\top \mathbf{x}) \leq \frac{1}{\alpha} - 1 \\
&\Leftrightarrow \ -\hat{\boldsymbol{\theta}}^\top \mathbf{x} \leq \log\left(\frac{1}{\alpha} - 1\right) \\
&\Leftrightarrow \ \hat{\boldsymbol{\theta}}^\top \mathbf{x} \geq -\log\left(\frac{1}{\alpha} - 1\right).
\end{aligned}$$

$\hat{\boldsymbol{\theta}}^\top \mathbf{x} = -\log\left(\frac{1}{\alpha} - 1\right)$ is the equation of a linear hyperplane comprised of all linear combinations $\hat{\boldsymbol{\theta}}^\top \mathbf{x}$ that are equal to $-\log\left(\frac{1}{\alpha} - 1\right)$. The inequality therefore describes the decision rule for setting $\hat{y}$ equal to 1 by taking all points that lie on or above this hyperplane.

b) We observe

- in plot (1): the logistic function runs parallel to the x_2 axis, so it is the same for every value of x_2. In other words, x_2 does not contribute anything to the class discrimination and its associated parameter $\hat{\theta}_2$ is equal to 0.
- in plot (2): both dimensions affect the logistic function – to equal degree in this case, meaning x_1 and x_2 are equally important. If $\hat{\theta}_1$ were larger than $\hat{\theta}_2$ or vice versa the hypersurface would be more tilted towards the respective axis. Furthermore, due to $\hat{\theta}_1$ and $\hat{\theta}_2$ being positive, $\hat{\pi}(\mathbf{x})$ increases with higher values for x_1 and x_2.
- in plot (3): this is the same situation as in plot (2) but the logistic function is steeper, which is due to $\hat{\theta}_1, \hat{\theta}_2$ having larger absolute values. We therefore get a sharper separation between classes (fewer predicted probability values close to 0.5, so we are overall more confident in our decision). As in plot (2), the increasing probability of $\hat{y} = 1$ for higher values of x_1 and x_2 indicates positive values for $\hat{\theta}_1$ and $\hat{\theta}_2$.
- in plot (4): this is the same situation as in plot (1). The different values for α represent different thresholds: a high value (leftmost line) means we only assign class 1 if the estimated class-1 probability is large. Conversely, a low value (rightmost line) signifies we are ready to predict class 1 at a low threshold – in effect, this is the same as the previous scenario, only the class labels are flipped. The mid line corresponds to the common case $\alpha = 0.5$ where we assign class 1 as soon as the predicted probability is more than 50%.

c) We make use of our results from a):

$$\begin{aligned}
\hat{y} = 1 \Leftrightarrow \hat{\boldsymbol{\theta}}^\top \mathbf{x} &\geq -\log\left(\frac{1}{\alpha} - 1\right) \\
\Leftrightarrow \hat{\boldsymbol{\theta}}^\top \mathbf{x} &\geq -\log\left(\frac{1}{0.5} - 1\right) \\
\Leftrightarrow \hat{\boldsymbol{\theta}}^\top \mathbf{x} &\geq -\log 1 \\
\Leftrightarrow \hat{\boldsymbol{\theta}}^\top \mathbf{x} &\geq 0.
\end{aligned}$$

The 0.5 threshold therefore leads to the coordinate hyperplane and divides the input space into the positive "1" halfspace where $\hat{\boldsymbol{\theta}}^\top \mathbf{x} \geq 0$ and the "0" halfspace where $\hat{\boldsymbol{\theta}}^\top \mathbf{x} < 0$.

d) When the threshold $\alpha = 0.5$ is chosen, the losses of misclassified observations, i.e., $L(\hat{y} = 0 \mid y = 1)$ and $L(\hat{y} = 1 \mid y = 0)$, are treated equally, which is often the intuitive thing to do. It means $\alpha = 0.5$ is a sensible threshold if we do not wish to avoid one type of misclassification more than the other. If, however, we need to be cautious to only predict class 1 if we are very confident (for example, when the decision triggers a costly therapy), it would make sense to set the threshold considerably higher.