Exericse 8.6 Elementary properties of $\ell_2$ regularized logistic regression

(Source: Jaakkola.). Consider minimizing

$$J(\mathbf{w}) = -\ell(\mathbf{w}, D_{train}) + \lambda \|\mathbf{w}\|_2^2$$

where

$$\ell(\mathbf{w}, D_{train}) = \frac{1}{|D|} \sum_{i \in D} \log \sigma(y_i \mathbf{x}_i^T \mathbf{w})$$

is the average log-likelihood on data set $D$, for $y_i \in \{-1, +1\}$ [and σ() is the logistic function].

Answer the following true / false questions.

a. $J(\mathbf{w})$ has multiple locally optimal solutions:

   False: this loss function is convex.

b. Let $\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} J(\mathbf{w})$ be a global optimum. $\hat{\mathbf{w}}$ is sparse (has many zero entries):

   False: since $\ell_2$ regularization penalizes larger weights more heavily, it does not promote sparsity [it does not encourage reducing a subset of the weights to zero].

c. If the training data is linearly separable, then some weights $w_j$ might become infinite if λ=0:

   True: the closer we get to a step function, the lower our log loss on the training data. Nota Bene: I would strongly encourage the use of regularization, to help avoid overfitting. It's not common to see super large weights in a well-behaved model.

d. $\ell(\mathbf{w}, D_{train})$ always increases as we increase λ:

   False: we expect the log-likelihood of the training data to decrease as we reduce the flexibility of the model (by increasing lambda). The model will eventually exhibit high bias for both the training and testing data, as the weights approach 0. Note: we want to maximize the log-likelihood of the data, which is equivalent to minimizing the negative log-likelihood of the data.

e. $\ell(\mathbf{w}, D_{test})$ always increases as we increase λ:

   False: we initially expect the log-likelihood of the testing data to increase (as we reduce overfitting); but we also expect the log-likelihood of the data to eventually decrease as we reduce the flexibility of the model. The model will eventually exhibit high bias for both the training and testing data, as the weights approach 0.