Model Answer Exercise 3 - April 30th
Rashka's paper

The article by (Rashka et al., 2018) discusses a number of approaches for optimizing the choice of hyperparameters. It is recommended that you read the paper but skip Paragraphs 1.7 (Confidence Intervals via Normal Approximations) and 2.4 (The Bootstrap Method and Empirical Confidence Intervals).

Questions about the text:

1. **Please explain what is stratified sampling - or stratification - and why it is important.**
   When a dataset is partitioned into a training set and a test (or validation) set, and when this is performed randomly, it may happen that the proportion of each class is quite different in the test (or validation) set as compared to the training set. This is especially true with small datasets. If there are just two classes the training set and test set can be unbalanced in opposite directions (ie if the training set has too much of class 1 data, the Test Set will have too much of class 0 data). Stratification is an approach which produces a test set with the same proportion of data in each class as in the training set (and hence as in the complete dataset). Unbalance between datasets is less of a concern in the case of large and well balanced initial datasets.

2. **Explain what is pessimistic and optimistic bias and how it relates to the size of the training and test set.**
   Pessimistic bias: the algorithm could learn a better model if it was given more data; by splitting off a portion of the dataset for testing in order to estimate the generalization performance, we withhold valuable data from the training set. To address this issue we may recalculate the network using both training and test data once we have checked that the model trained only on the training set fits the test (or validation) set well.
   We absolutely do not want to train and evaluate a model on the same training dataset, since it would introduce a very optimistic bias due to overfitting. In other words, we would not be able to tell whether the model simply memorized the training data, or whether it learnt a genuine relationship between input and output. This optimism bias can be characterized by the difference between the training and test accuracy.

3. **What are the three goals of performance estimation?**

a. Estimate the generalization accuracy, that is the predictive performance of a model on future (unseen) data. This is for a fixed value of the hyper-parameters.
b. Increase the predictive performance by tweaking the hyper-parameters of the learning algorithm and selecting the best-performing hyper-parameters combination.

c. Identify not only the best hyper-parameters but more generally the deep learning algorithm - or model - that is best-suited for the problem at hand: in other words, we want to compare different algorithms and select the best performing one.

4. **Why do we need a validation set?**
   The training set cannot be used to optimize the hyper-parameters, as we need to use data that have not yet been seen. A validation set must be used. The hyper-parameters which lead to the best neural network for predicting the validation set are chosen. The results on the test set should not be used to optimize the hyper-parameters.

5. **What is the difference between holdout method, 2-fold cross-validation, and repeated holdout methods?**
   With the holdout method there is half of the data used as training set and half as test (or validation) set. With 2-fold cross-validation, we do as with the holdout method but then the two sets are swapped, training set becomes test set and test set becomes training set. With the repeated holdout method (also called Monte Carlo cross-validation), we create the test set by randomly drawing half of the data. Then we repeat again by randomly drawing half of the data. We do this a number of times and calculate the average performance of the algorithm.