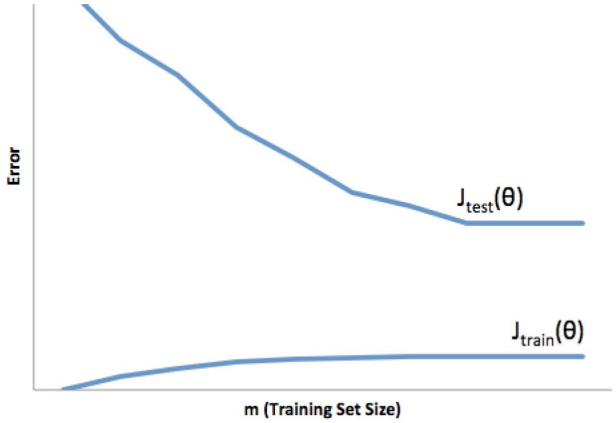
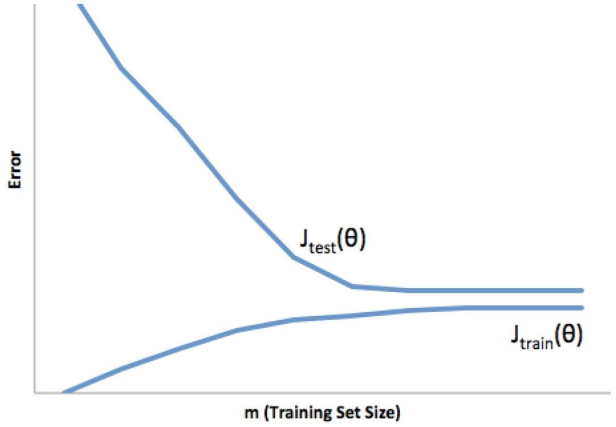
HMC CS 158
Quiz 6.5: Applying ML

1. You train a learning algorithm and find that it has unacceptably high error on the test set. You plot the learning curve and obtain the figure below. Is the algorithm suffering from high bias, high variance, or neither?



2. You train a learning algorithm and find that it has unacceptably high error on the test set. You plot the learning curve and obtain the figure below. Is the algorithm suffering from high bias, high variance, or neither?



This quiz is adapted from course material by Andrew Ng (Stanford).

3. Suppose you have implemented regularized logistic regression to classify what object is in an image (i.e. to do object recognition). However, when you test your hypothesis on a new set of images, you find that it makes unacceptably large errors with its predictions on the new images. However, your hypothesis performs **well** (has low error) on the training set. Which of the following are promising steps to take? Check all that apply.

   (a) Try decreasing the regularization parameter λ.

   (b) Try evaluating the hypothesis on a cross-validation set rather than the test set.

   (c) Get more training examples.

   (d) Try using a smaller set of features.

4. Suppose you have implemented regularized logistic regression to predict what items customers will purchase on a web shopping site. However, when you test your hypothesis on a new set of customers, you find that if makes large errors in its predictions. Furthermore, the hypothesis performs **poorly** on the training set. Which of the following might be promising steps to take? Check all that apply.

   (a) Try to obtain and use additional features.

   (b) Try adding polynomial features.

   (c) Use fewer training examples.

   (d) Try evaluating the hypothesis on a cross-validation set rather than the test set.

5. Which of the following statements are true? Check all that apply.

   (a) It is okay to use data from the test set to choose the regularization parameter λ but not the model parameters $\boldsymbol{\theta}$.

   (b) Suppose you are using linear regression to predict housing prices, and your dataset comes sorted in order of increasing sizes of houses. It is then important to randomly shuffle the dataset before splitting into training, validation, and test sets, so that we do not have all the smallest houses going into the training set, and all the largest houses going into the test set.

   (c) Suppose you are training a regularized linear regression model. The recommended way to choose what value of regularization parameter λ to use is to choose the value of λ which gives the lowest **training set** error.

   (d) A typical split of a dataset into training, validation, and test sets might be 60% training set, 20% validation set, and 20% test set.

   (e) The performance of a learning algorithm on the training set will typically be better than its performance of the test set.

   (f) Suppose you are training a logistic regression classifier using polynomial features and want to select what degree polynomial to use. After training the classifier on the entire training set, you decide to use a subset of the training examples as a validation set. This will work just as well as having a validation set that is separate (disjoint) from the training set.

6. Which of the following statements are true? Check all that apply.

   (a) If the training and test errors are about the same, adding more features will **not** help improve the results.

   (b) When debugging learning algorithms, it is useful to plot a learning curve to understand if there is a high bias or high variance problem.

   (c) A model with more parameters is more prone to overfitting and typically has higher variance.

   (d) If a learning algorithm is suffering from high variance, adding more training examples is likely to improve the test error.

   (e) If a learning algorithm is suffering from high bias, only adding more training examples may **not** improve the test error significantly.