Bias vs. Variance

Problem 1:
Suppose we have a classification problem with a binary response $Y$ and a $p$-dimensional predictor variable $X = (X_1, \ldots, X_p)$. Logistic regression is fitted to a set of $n$ samples. Then, logistic regression is fitted again to the same observations, where we include one additional predictor, such that:

$$X = (X_1, \ldots, X_p, X_{p+1}).$$

Explain how the training error, test error, and coefficients change in each of the following cases:

(a) $X_{p+1} = X_1 + 2X_p$.
(b) $X_{p+1}$ is a random variable independent of $Y$.

Note: We use logistic regression in this problem because it is simple to reason with. We should get the same results with a neural net.

Problem 2:
Assume we have a set of data from patients who have visited UPMC hospital during the year 2011. A set of features (e.g., temperature, height) have been also extracted for each patient. Our goal is to decide whether a new visiting patient has any of diabetes, heart disease, or Alzheimer (a patient can have one or more of these diseases). We have decided to use a neural network to solve this problem. We have two choices: either to train a separate neural network for each of the diseases or to train a single neural network with one output neuron for each disease, but with a shared hidden layer. Which method do you prefer? Justify your answer.
Problem 3:

Suppose you have regression data generated by a polynomial of degree 3. Characterize the bias-variance of the estimates of the following models on the data with respect to the true model by circling the appropriate entry.

<table>
<thead>
<tr>
<th>Model</th>
<th>Bias</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear regression</td>
<td>low/high</td>
<td>low/high</td>
</tr>
<tr>
<td>Polynomial regression with degree 3</td>
<td>low/high</td>
<td>low/high</td>
</tr>
<tr>
<td>Polynomial regression with degree 10</td>
<td>low/high</td>
<td>low/high</td>
</tr>
</tbody>
</table>

For each part below indicate whether we would generally expect the performance of a flexible statistical learning method to be better or worse than an inflexible method. Justify your answer.

(i) The sample size n is extremely large, and the number of features p is small.
(ii) The number of predictors p is extremely large, and the number of observations n is small.
(iii) The relationship between the predictors and response is highly non-linear.
(iv) The variance of the error terms, i.e. $\sigma^2 = \text{Var}()$ is extremely high.

Problem 4:

Applying the methodology seen in class, you split your dataset in 3 sub-datasets, train 10 models on the training set and then choose the model on the validation set. We achieve good performance on the validation set but then the testing error is terrible. What could be the problem? Give 3 reasons

Problem 5:

Assume we try to estimate a quantity, let’s call it $\mu$. We know that Y is random variable whose distribution is the Normal distribution centered on $\mu$ and with standard deviation $\sigma$.

Imagine we can sample (do random draws) Y. This gives $Y_1', Y_2', ..., Y_n$

Question 1: How would you estimate $\mu$?
Question 2: The problem of the previous estimator is that it depends a lot on the standard deviation $\sigma$. What can we do to avoid that problem?
Question 3: If the cost of sampling is high, we need another way to get a better estimate of $\mu$. Can you think of another measure to estimate $\mu$? How is the bias/variance of this new estimator?
Problem 6

The following graph plots the error distribution of three different models. Comment on each model in terms of bias-variance: (Credit: Ramesh Johari)