Machine Learning Mechods
P160B124

Bias and Variance Trade Off

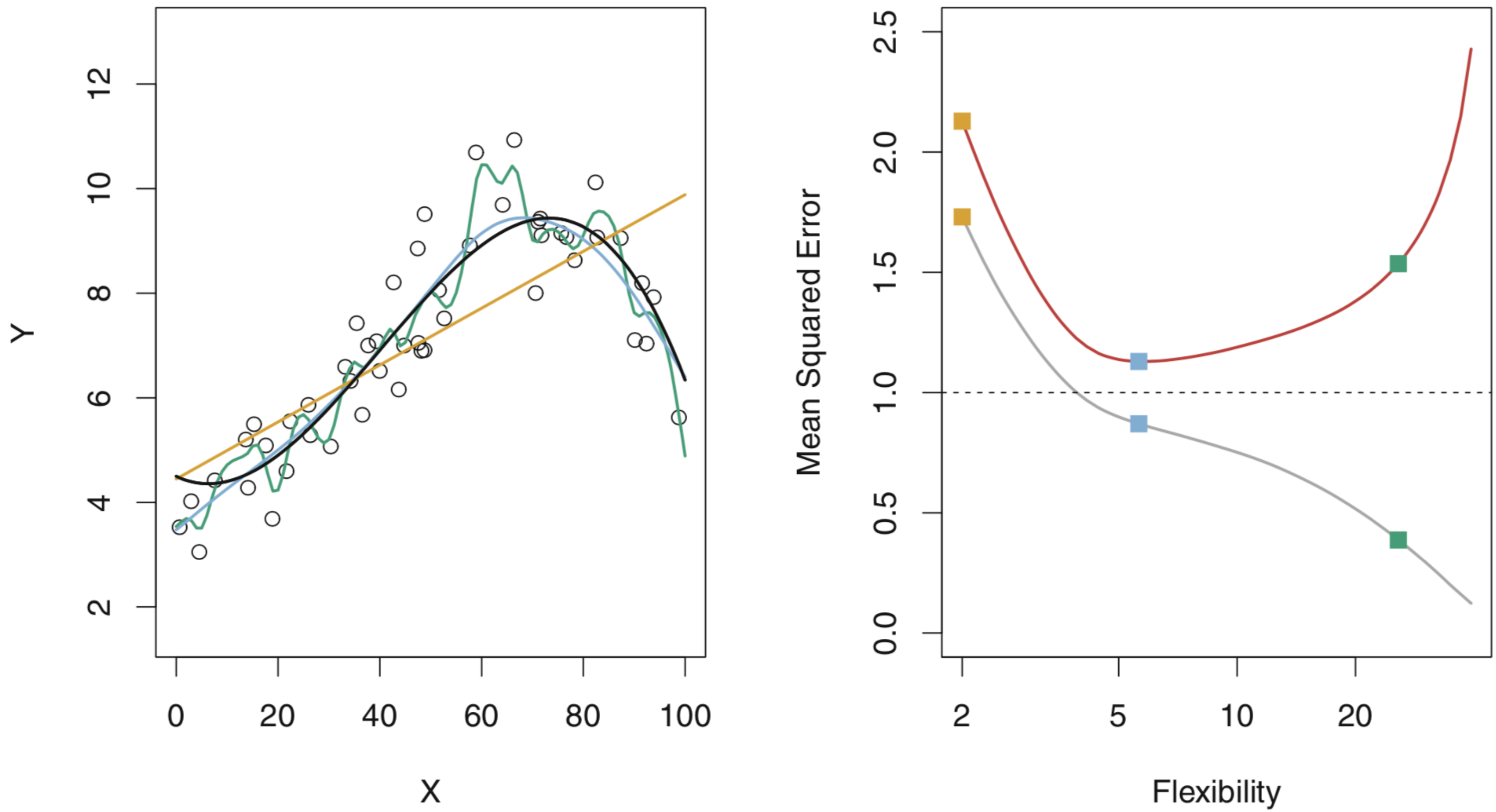
assoc. prof. dr. Tomas Iešmantas
tomas.iesmantas@ktu.lt
Room 319

Training set vs Testing set

- Should we trust training accuracy of a model? Never. But why?
- Training accuracy overestimates the true accuracy for entire population.
- What to do? Use independent (i.e. unseen) data set for accuracy estimation.
- True accuracy is never known (cannot have all population). Testing set allows to estimate approximately the true accuracy.

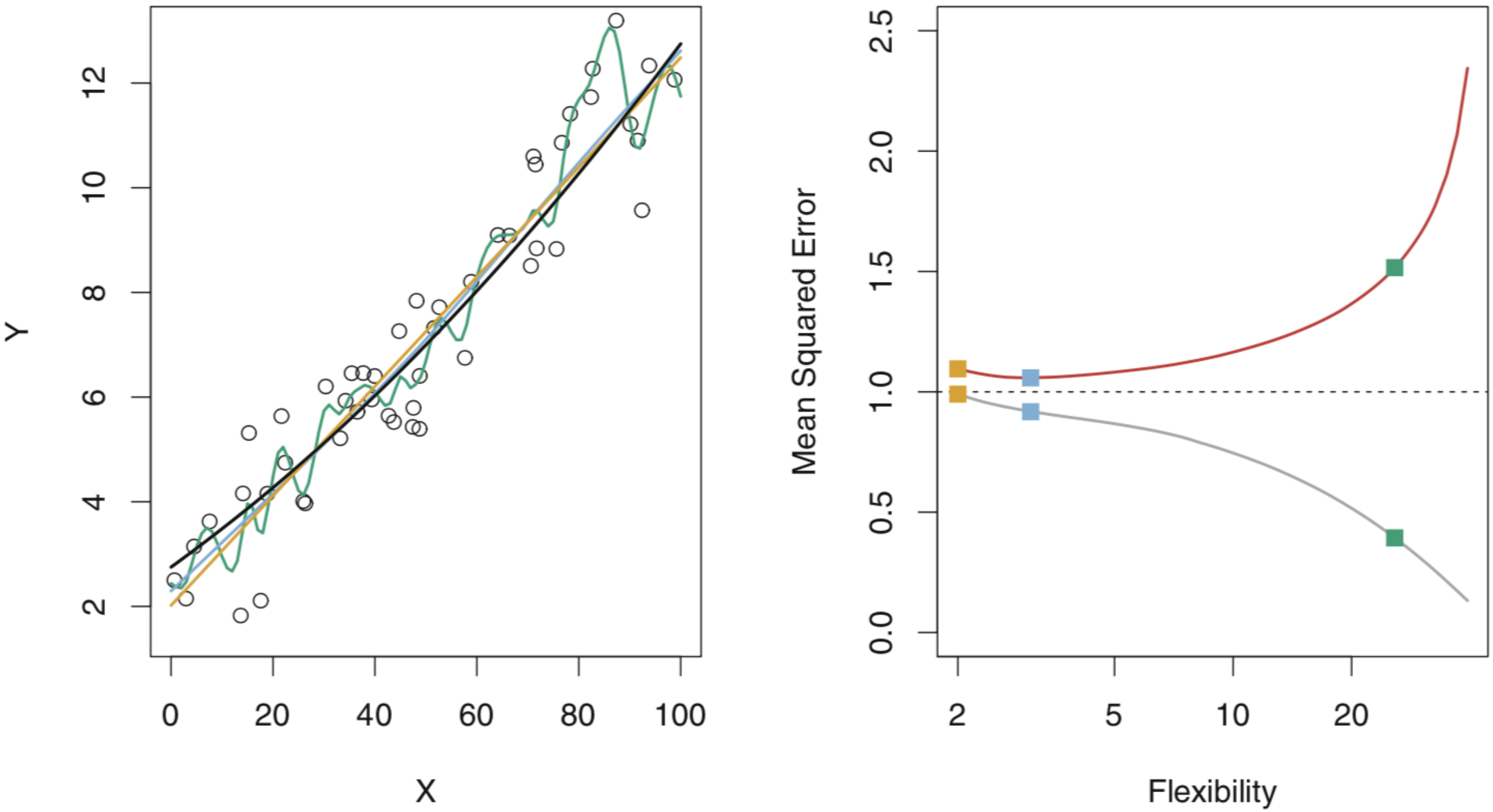
Example

- More flexibility – better fit to training set;



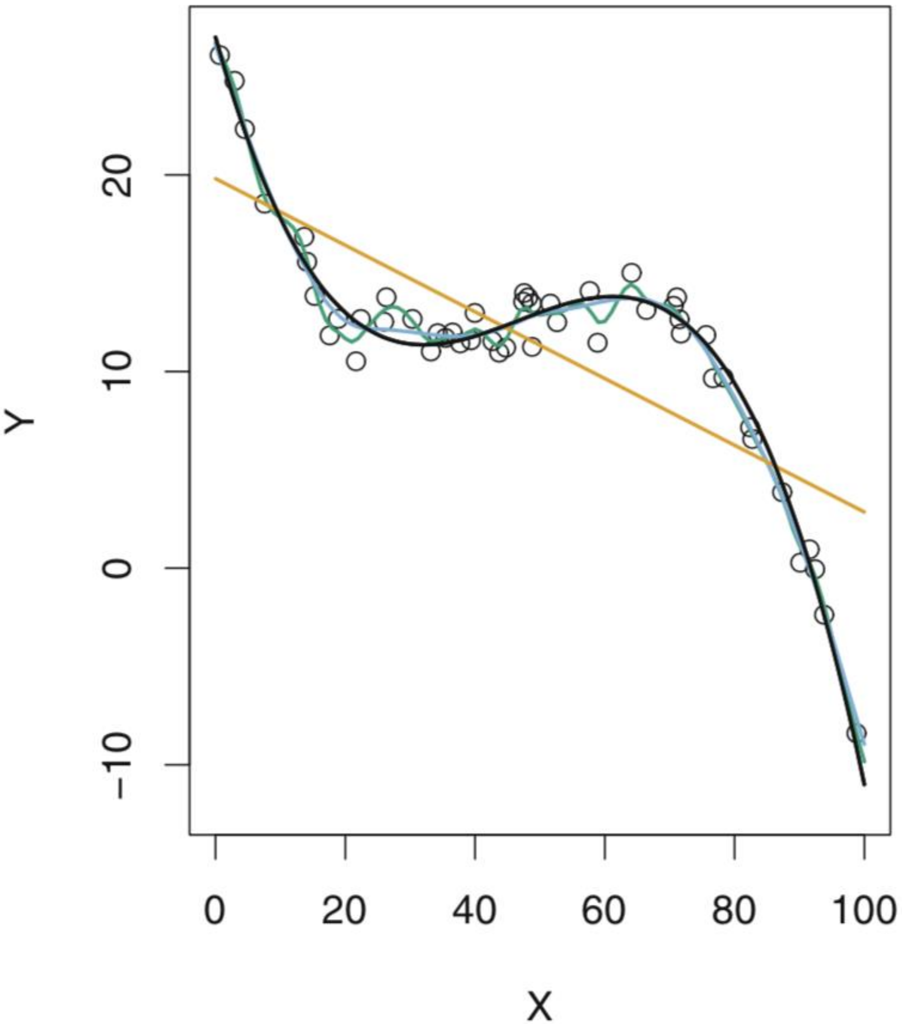
Example

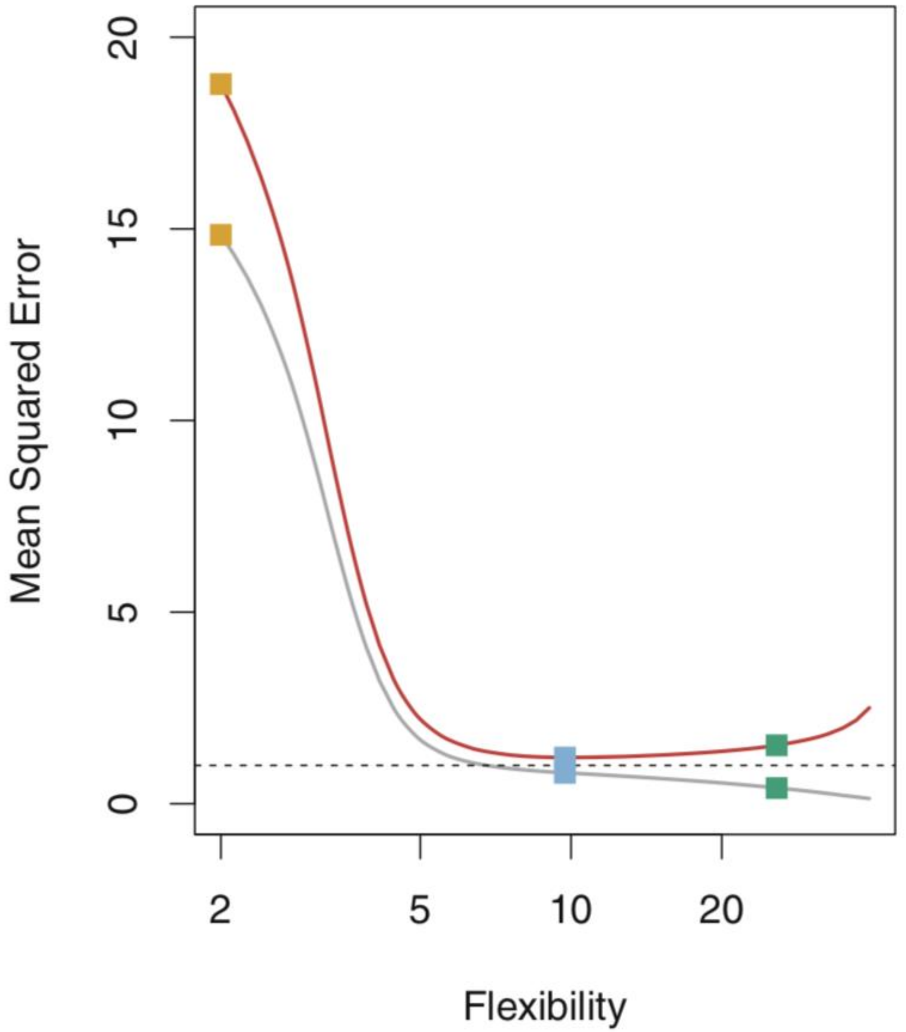
- Increasing model flexibility makes test set error increase (after some complexity threshold)



Example

- Increasing model flexibility makes test set error increase (after some complexity threshold)





Bias vs. Variance

- The **U-shape** observed in the test MSE curves turns out to be **the result of two competing properties** of machine learning methods.
- It can be showed, that MSE for test value x_0 can be decomposed into three quantities: the variance of $\hat{f}(x_0)$, the squared bias of $\hat{f}(x_0)$, and the variance of the error term ϵ.

$$E\left(y_0 - \hat{f}(x_0)\right)^2$$
$$= Var\left(\hat{f}(x_0)\right) + \left[Bias\left(\hat{f}(x_0)\right)\right]^2 + Var(\epsilon)$$

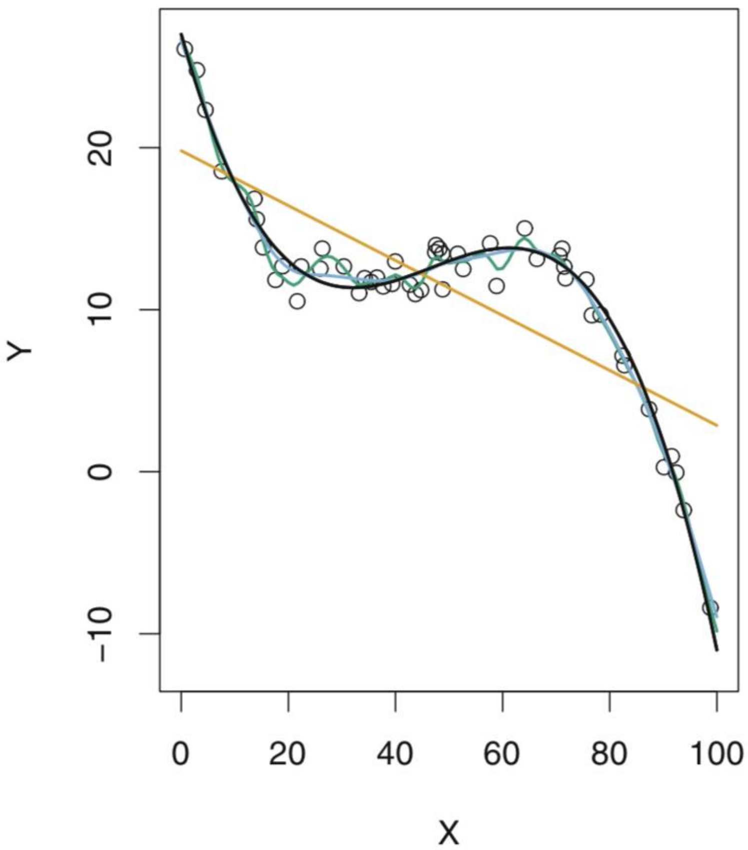
Bias vs. Variance

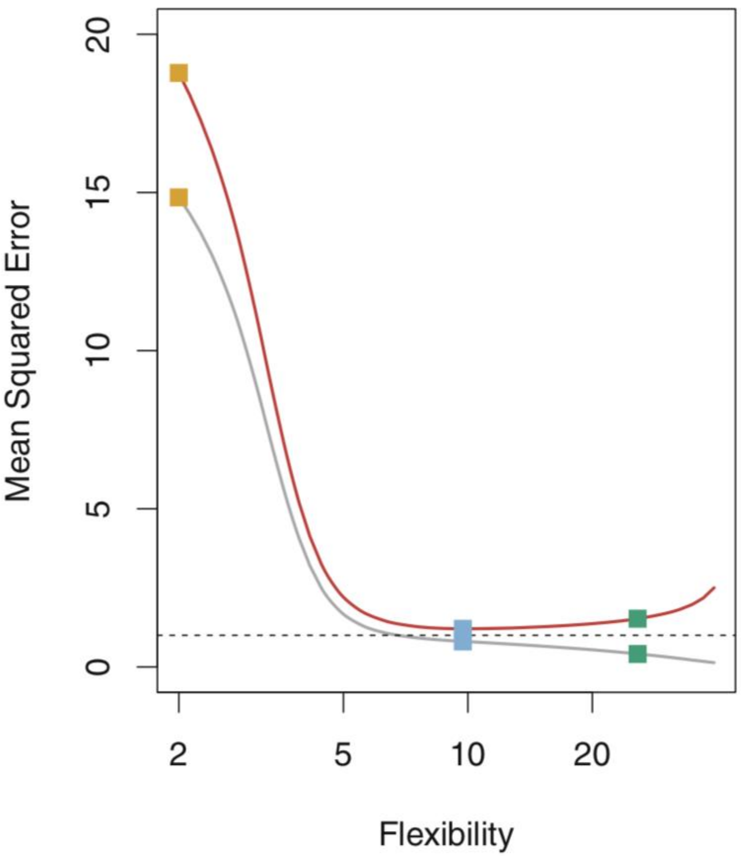
- $E\left(y_0 - \hat{f}(x_0)\right)^2$ defines the **expected test MSE** and refers to the average test MSE that we would obtain if we repeatedly estimated f using a large number of training sets, and tested each at x_0.

- To minimize test MSE, select a machine learning method that simultaneously achieves **low variance and low bias**.
- Expected test MSE cannot be smaller than $Var(\epsilon)$

Bias vs. Variance

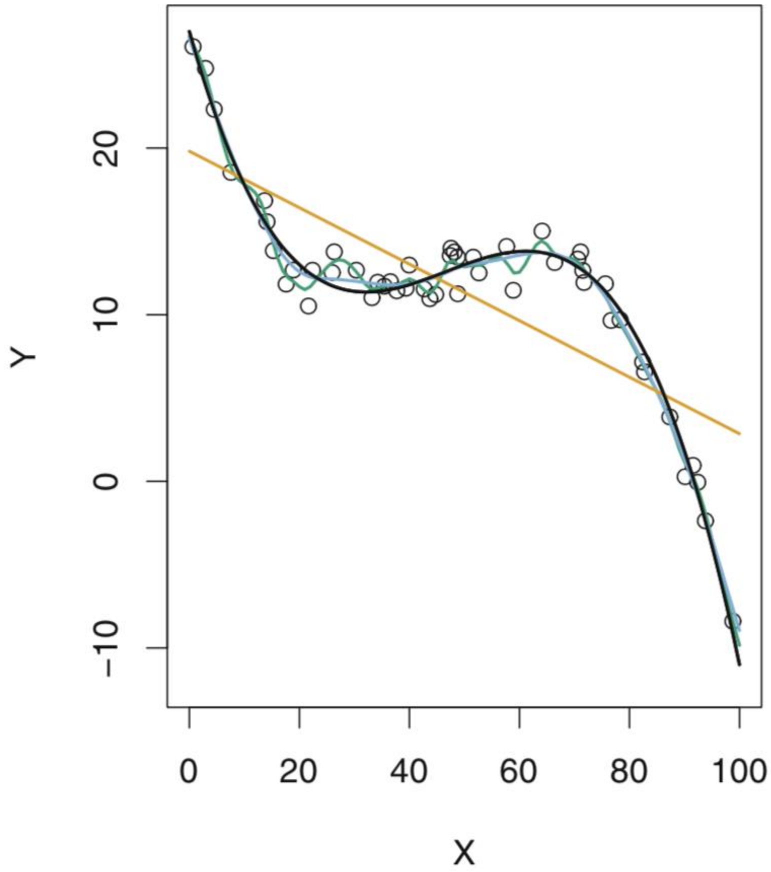
- **What is variance** of $\hat{f}(x_0)$? *The amount by which $\hat{f}$ would change if we estimated it using a different training data set.* f should not vary too much between training sets. More flexibility, more variance.

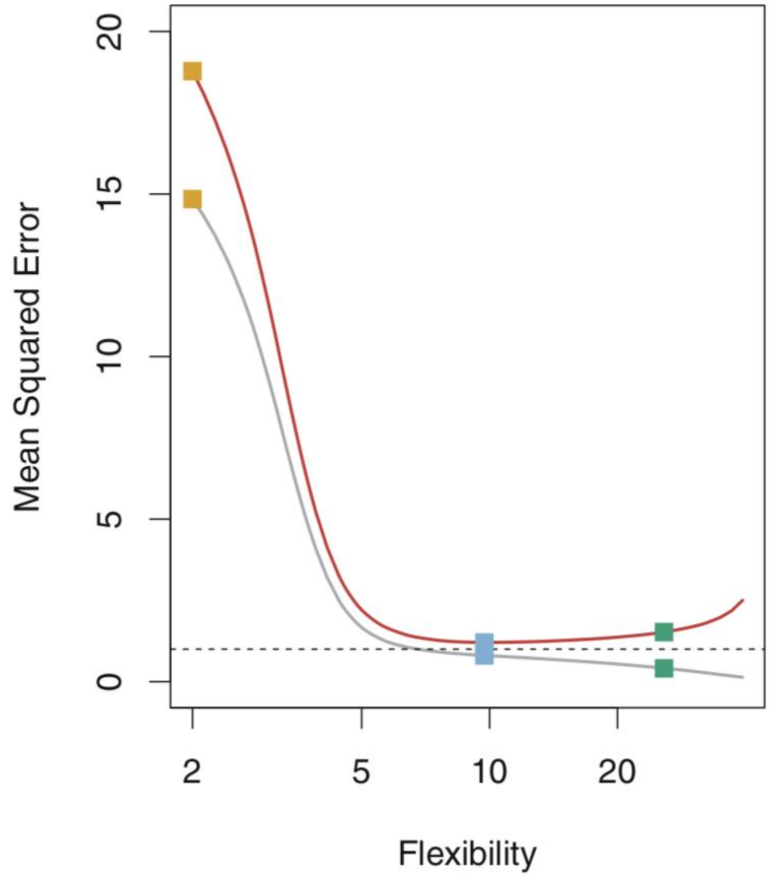




Bias vs. Variance

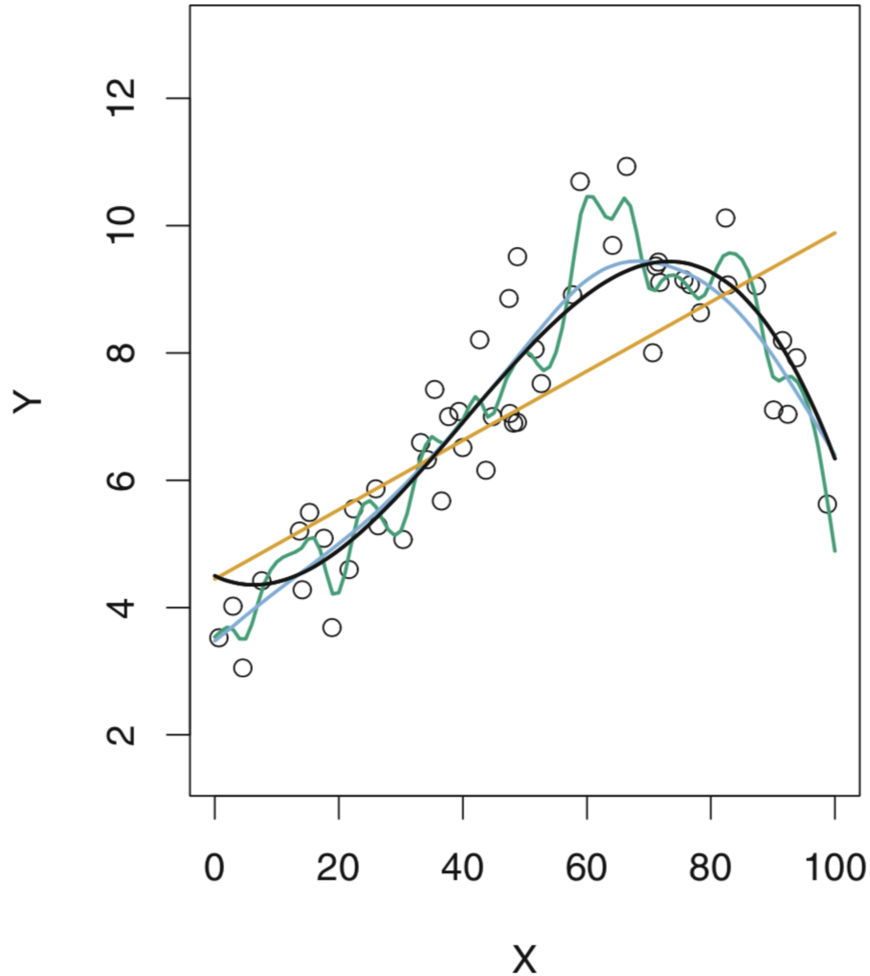
- **What is bias** of $\hat{f}(x_0)$? *The error that is introduced by approximating a real-life problem, which may be extremely complicated, by a much simpler model*. Less flexibility, more bias.

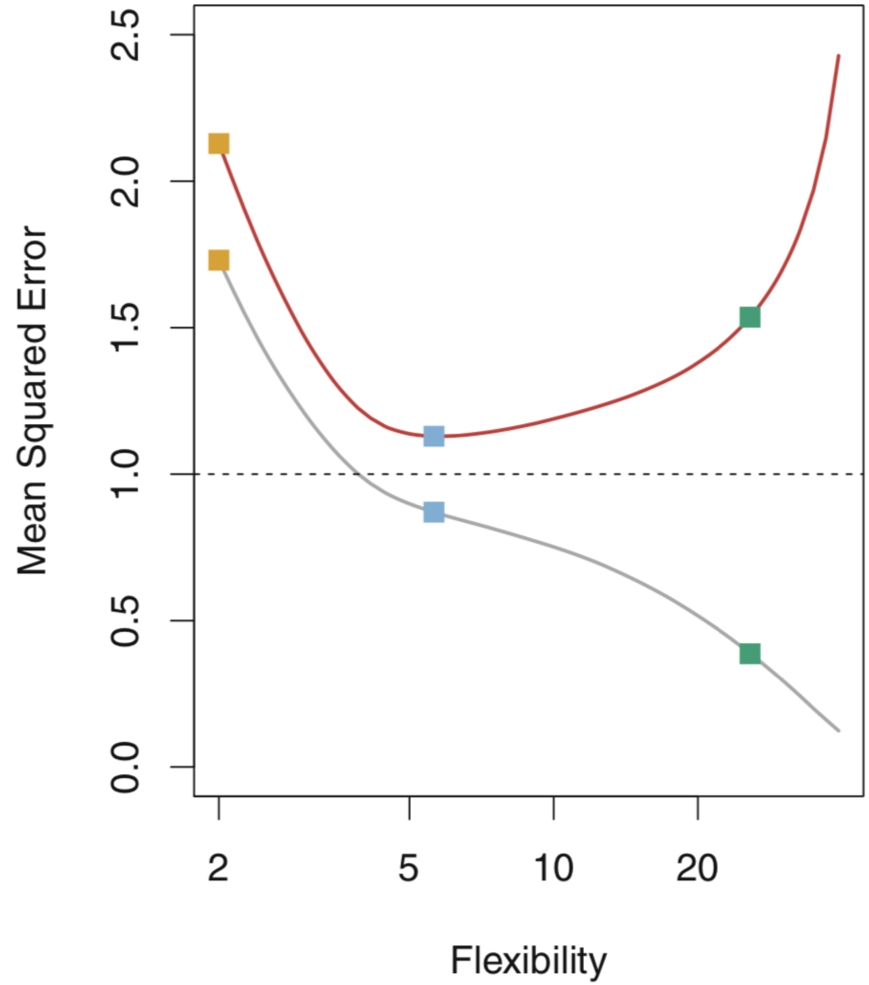




Bias vs. Variance

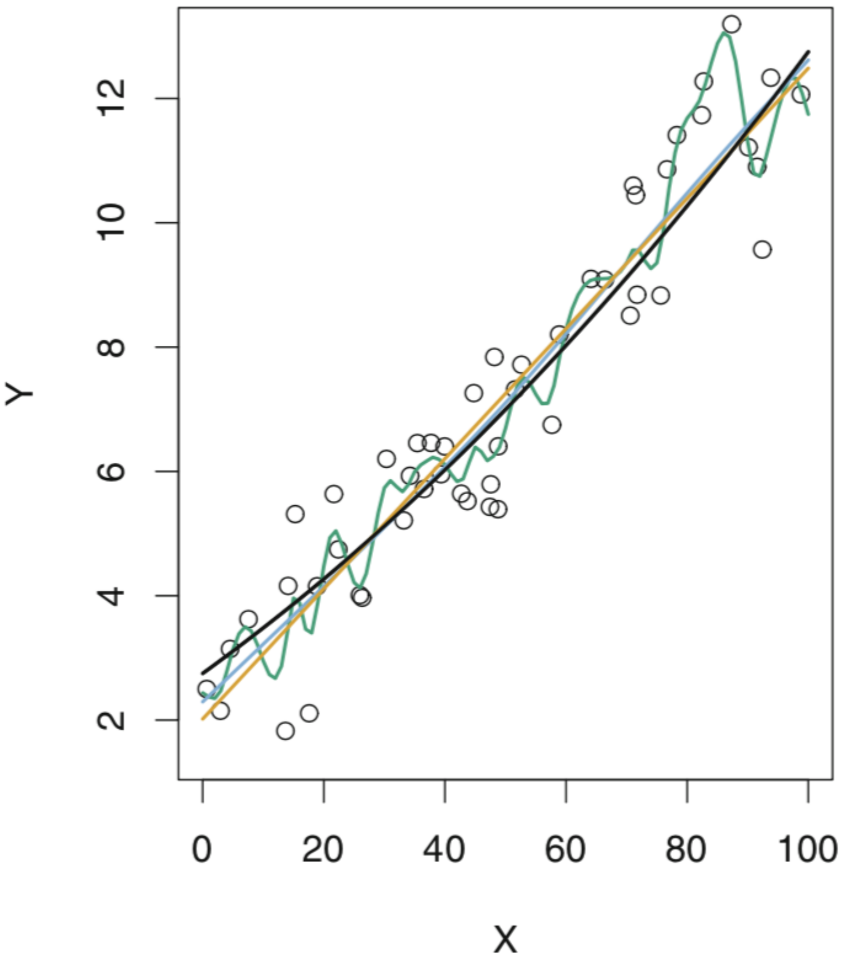
- In this example, true relationship is highly nonlinear and linear regression will always be inaccurate. **Hence, high bias**.

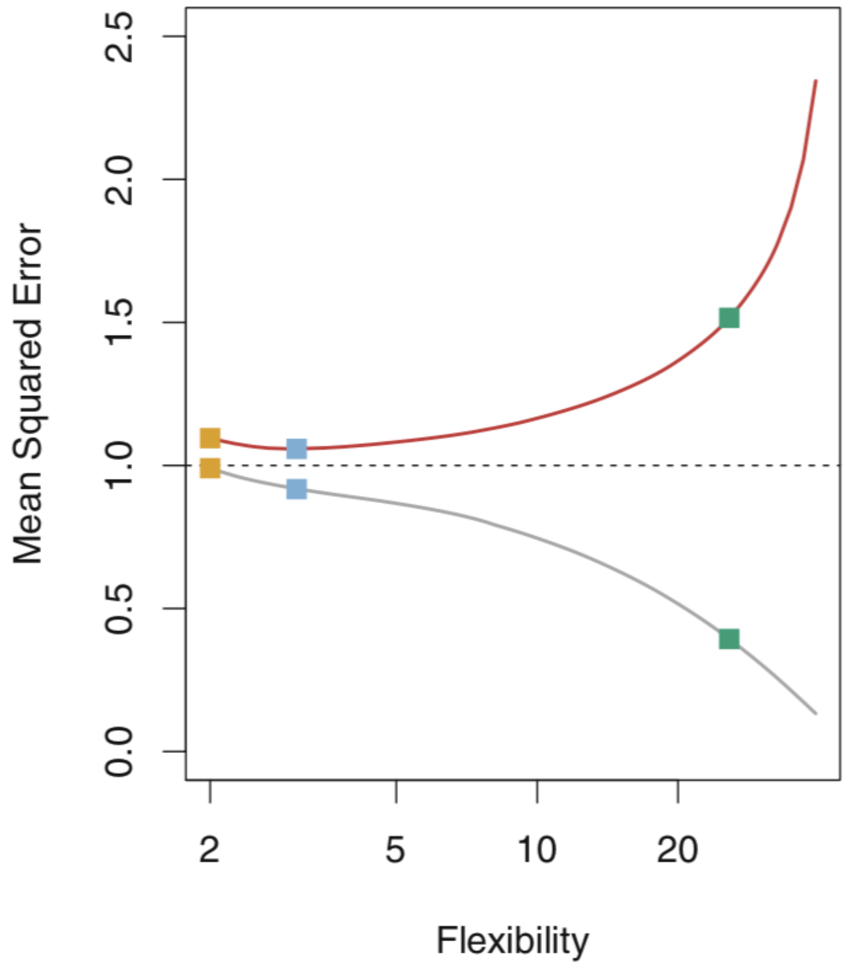




Bias vs. Variance

- In this example, true relationship is approximately linear and linear regression provides quite accurate fit. **Hence, low bias.**



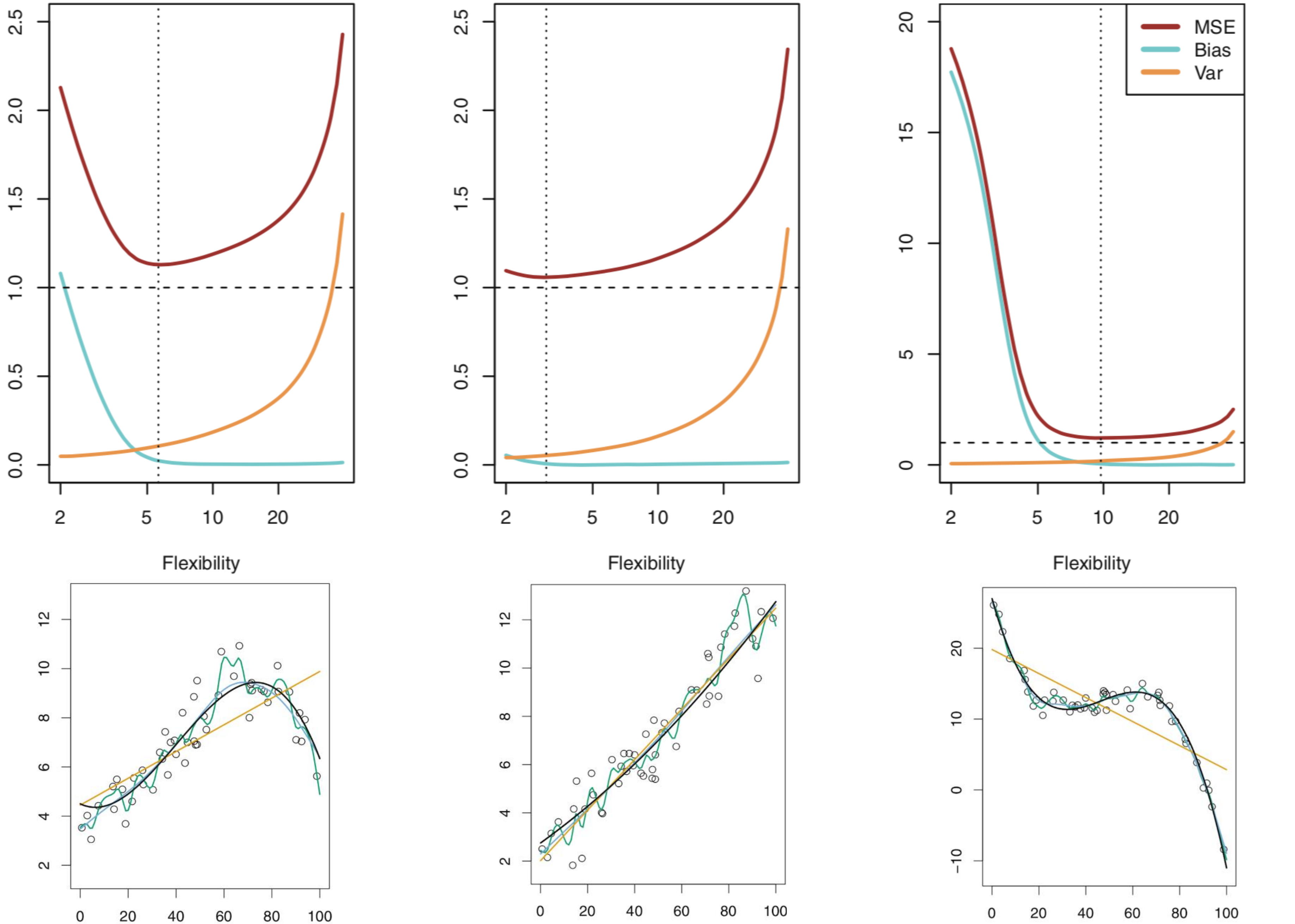


Bias vs. Variance

- ***As a general rule***, as we use more flexible methods, the variance will increase and the bias will decrease (at different speed);
- The relative rate of change of these two quantities determines whether the test MSE increases or decreases;

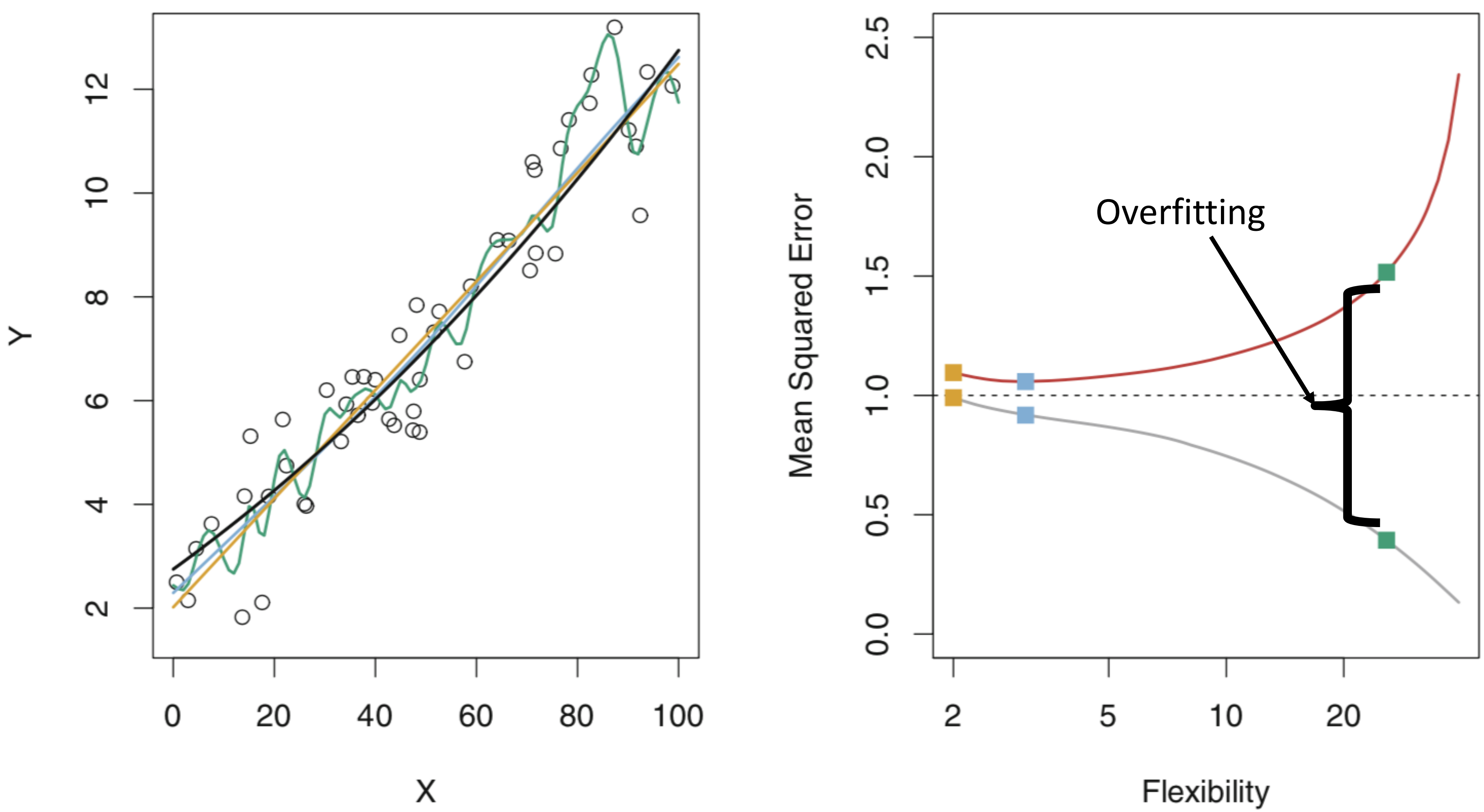
$$E\left(y_0 - \hat{f}(x_0)\right)^2$$
$$= Var\left(\hat{f}(x_0)\right) + \left[Bias\left(\hat{f}(x_0)\right)\right]^2 + Var(\epsilon)$$

Bias – Variance plots



Overfitting

- **Overfitting** informally is *a situation when test error is significantly higher as compared to the training error*.



Overfitting

- **Overfitting** informally is *a situation when test error is significantly higher as compared to the training error*.
- This occurs partly due to the high flexibility of the model. It's like learning not only the true signal, but also remembering the noise in the data.
- That's why we always test on separate test set: to make sure that our model well approximates the true signal and not imitates the noise.

Classification setting: error

- Consider a training data set $\{(x_1, y_1), \cdots, (x_N, y_N)\}$, where y is a categorical variable.
- We seek to estimate a classifier f.
- Error rate of the estimated classifier $\hat{f}$ can be expressed as

$$\frac{1}{N}\sum_{i=1}^{N} I(\hat{y}_i \neq y_i)$$

- A good classifier is the one for which this test error is the smallest (obvious!). But which one is the best?

Classification setting: Bayes classifier

- Test error is minimized (on average) by a classifier that ***assigns each observation to the most likely class, given its predictor values, i.e.*** assign a test observation x_0 to the class j, for which $P(y = j|x = x_0)$ is largest.
- This classifier (or a rule) is known as **Bayes classifier** (Bayes rule).
- In two class problem, Bayes classifier corresponds to the rule: predicting class 1 if $P(y = 1|x = x_0) > 0.5$.
- **The Bayes classifier produces the lowest possible test error rate, called the Bayes error rate.**

Classification setting: Bayes error

- The Bayes classifier produces the lowest possible test error rate, called the **Bayes error rate**.
- In the simulated data example, Bayes error rate is 0.1304 – and no other classifier can achieve a smaller error rate.
- *Bayes error rate is analogous to the irreducible error in the regression setting* $Var(\epsilon)$

Classification setting: Bayes error

- *Bayes error rate is analogous to the irreducible error in the regression setting* $Var(\epsilon)$
- **Theorem**: No other classification rule gives better results than Bayes classifier.

- Why then bother with other classifiers if Bayes is the best possible one?

- Is Logistic regression a Bayes classifier?

Vocabulary

- Training (testing) accuracy – apmokymo (testavimo) tikslumas;
- Training (testing) error – apmokymo (testavimo) paklaida;
- Bias – poslinkis;
- Variance – variacija;
- Overfitting – persimokymas;
- Bayes classifier – Bajeso klasifikatorius.