Decision Tree Learning

Dr. Xiaowei Huang

https://cgi.csc.liv.ac.uk/~xiaowei/

Briefing about First Assignment

Decision Tree up to now,

- Decision tree representation

- A general top-down algorithm
- How to do splitting on numeric features
- Occam’s razor

- Entropy and information gain
- Types of decision-tree splits

Today’s Topics

- Stopping criteria of decision trees
- Accuracy of decision trees
- Pruning and validation dataset

- Overfitting

Step (3): Stopping criteria

Stopping criteria

- We should form a leaf when
  - all of the given subset of instances are of the same class
  - we've exhausted all of the candidate splits

Accuracy of Decision Tree

Definition of Accuracy and Error

- Given a set D of samples and a trained model M, the accuracy is the percentage of correctly labeled samples. That is,

$$Accuracy(D, M) = \frac{|\{M(x) = l_x \mid x \in D\}|}{|D|}$$

Where l_x is the true label of sample x and $M(x)$ gives the predicted label of x by M

- Error is a dual concept of accuracy.

But, what is D?

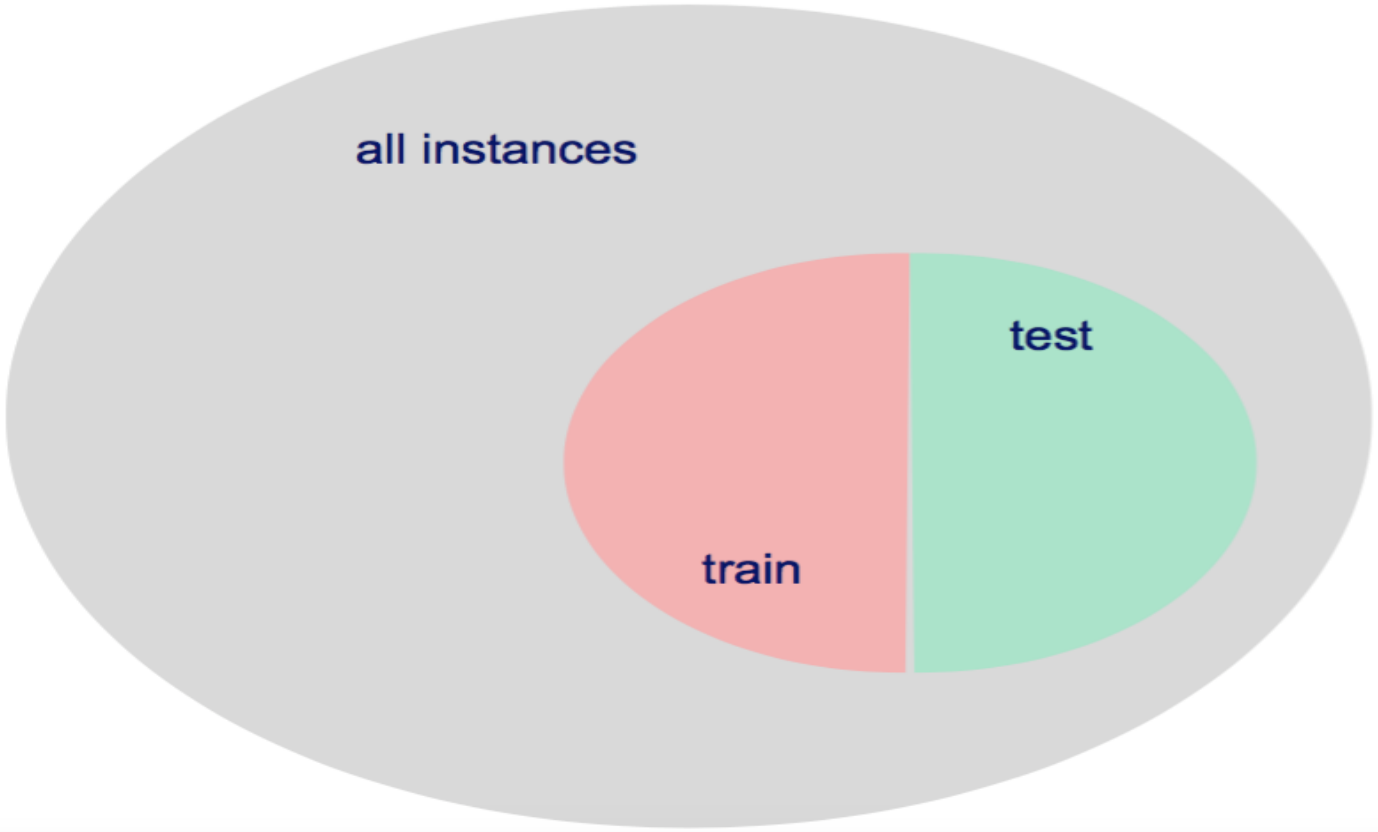
$$Error(D, M) = 1 - Accuracy(D, M)$$

How can we assess the accuracy of a tree?

- Can we just calculate the fraction of training instances that are correctly classified?
- Consider a problem domain in which instances are assigned labels at random with P(Y = t) = 0.5
  - how accurate would a learned decision tree be on previously unseen instances?
    - Can never reach 1.0.
  - how accurate would it be on its training set?
    - Can be arbitrarily close to, or reach, 1.0 if model can be very large.

How can we assess the accuracy of a tree?

- to get an unbiased estimate of a learned model's accuracy, we must use a set of instances that are held-aside during learning
- this is called a *test set*



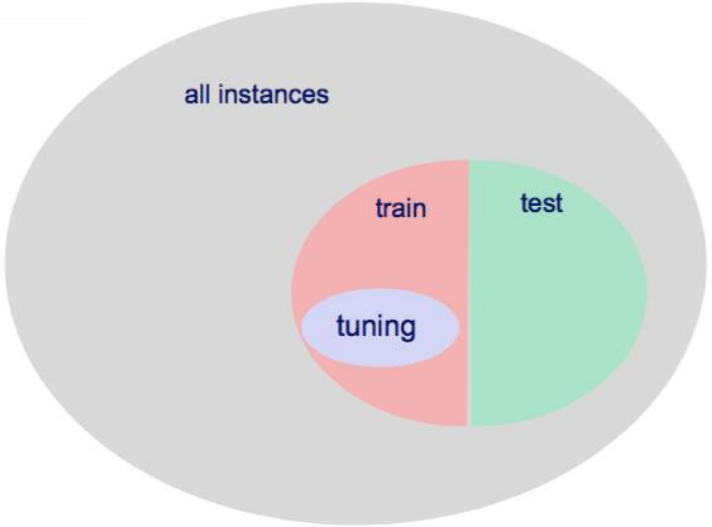
Pruning and Validation Dataset

Stopping criteria

- We should form a leaf when
  - all of the given subset of instances are of the same class
  - we've exhausted all of the candidate splits
- Is there a reason to stop earlier, or to prune back the tree?

Pruning in C4.5

- split given data into training and *validation* (*tuning*) sets
- a *validation set* (a.k.a. *tuning set*) is a subset of the training set that is held aside
  - not used for primary training process (e.g. tree growing)
  - but used to select among models (e.g. trees pruned to varying degrees)

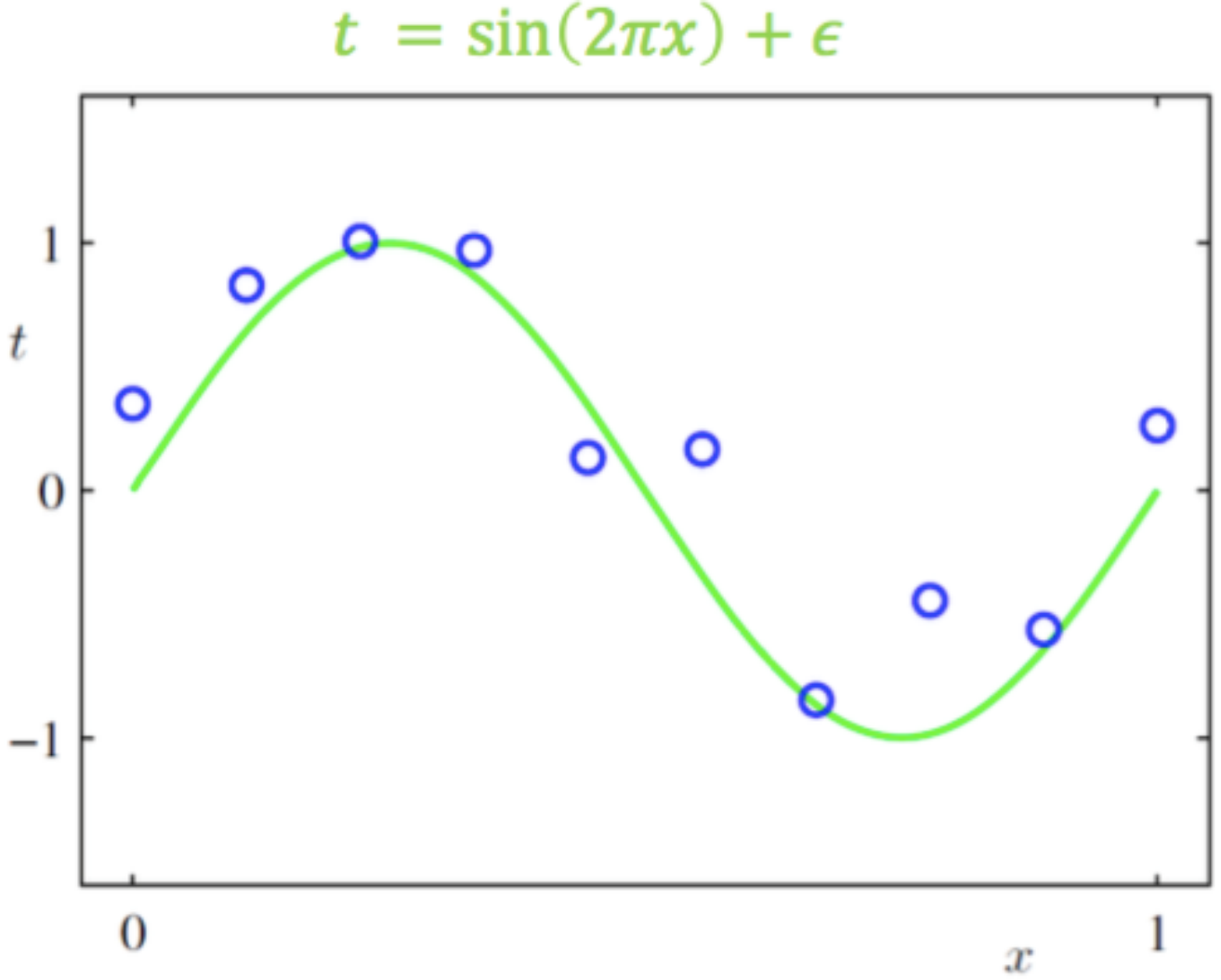


Pruning in C4.5

- Split given data into training and *validation* (*tuning*) sets
- Grow a complete tree
- do until further pruning is harmful
  - evaluate impact on tuning-set accuracy of pruning each node
  - greedily remove the one that least reduces tuning-set accuracy

Overfitting

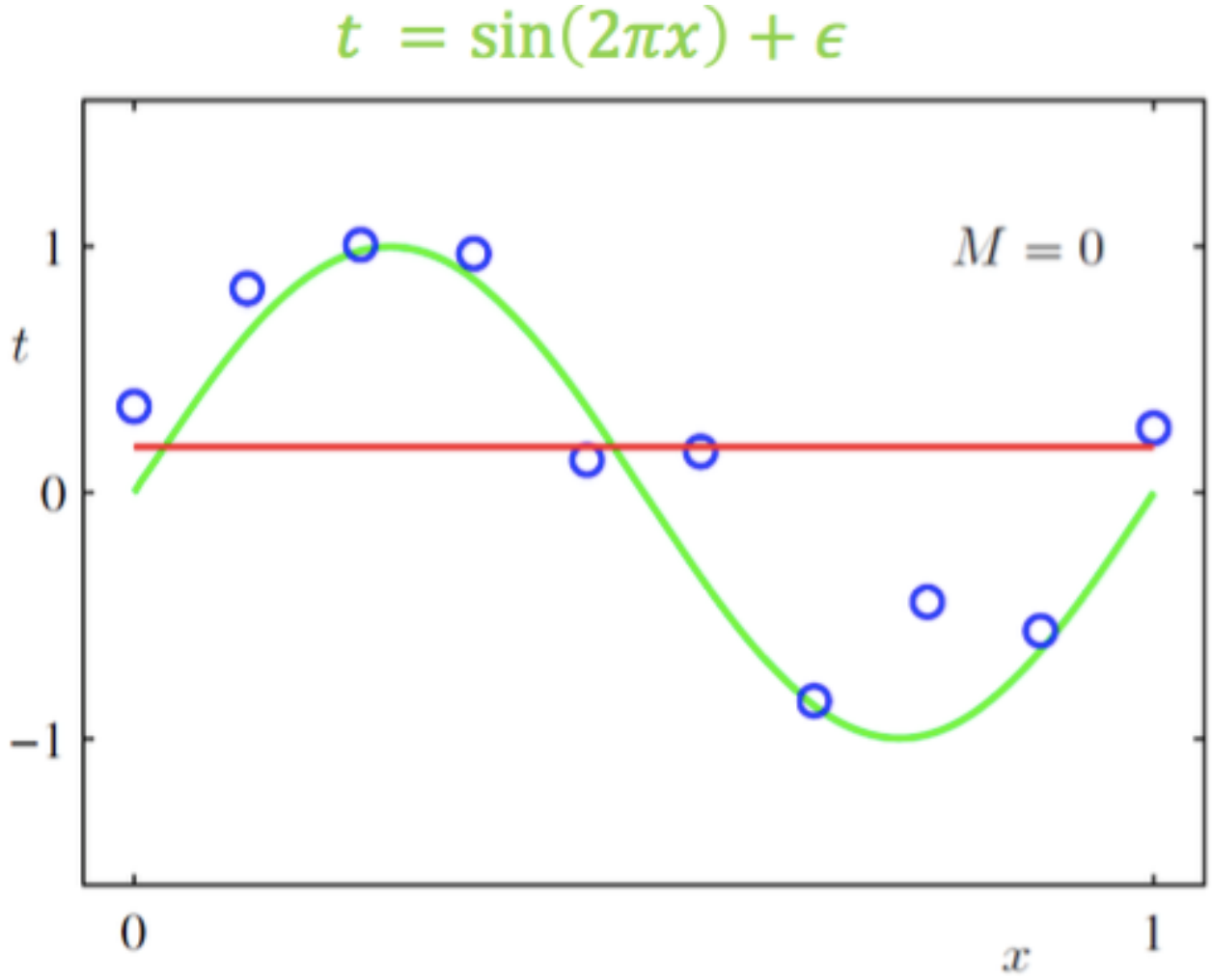
Example 3: regression using polynomial



Regression using polynomial of degree M

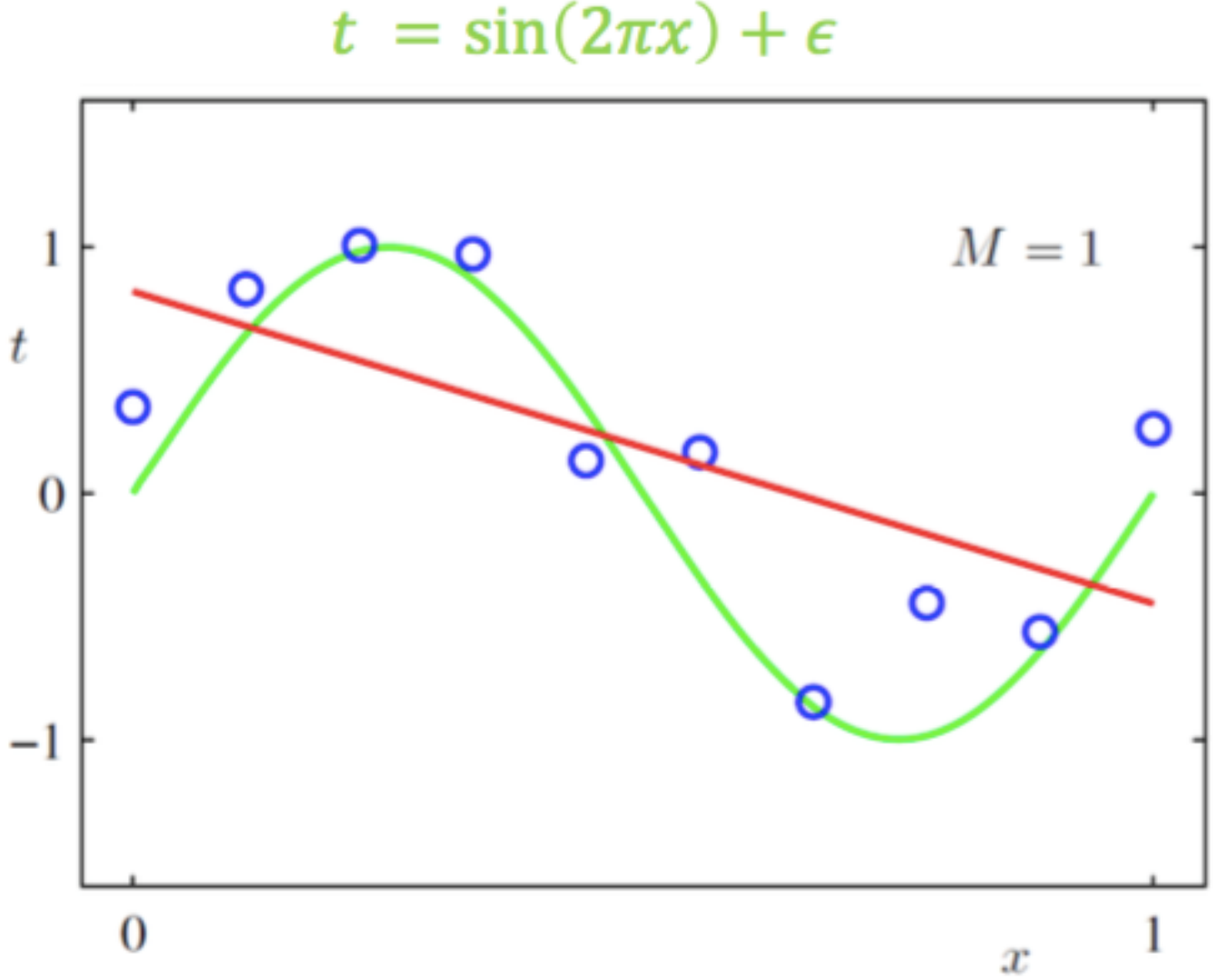
$y = w_M x^M + w_{M-1} x^{M-1} + \ldots + w_1 x + w_0$

Example 3: regression using polynomial



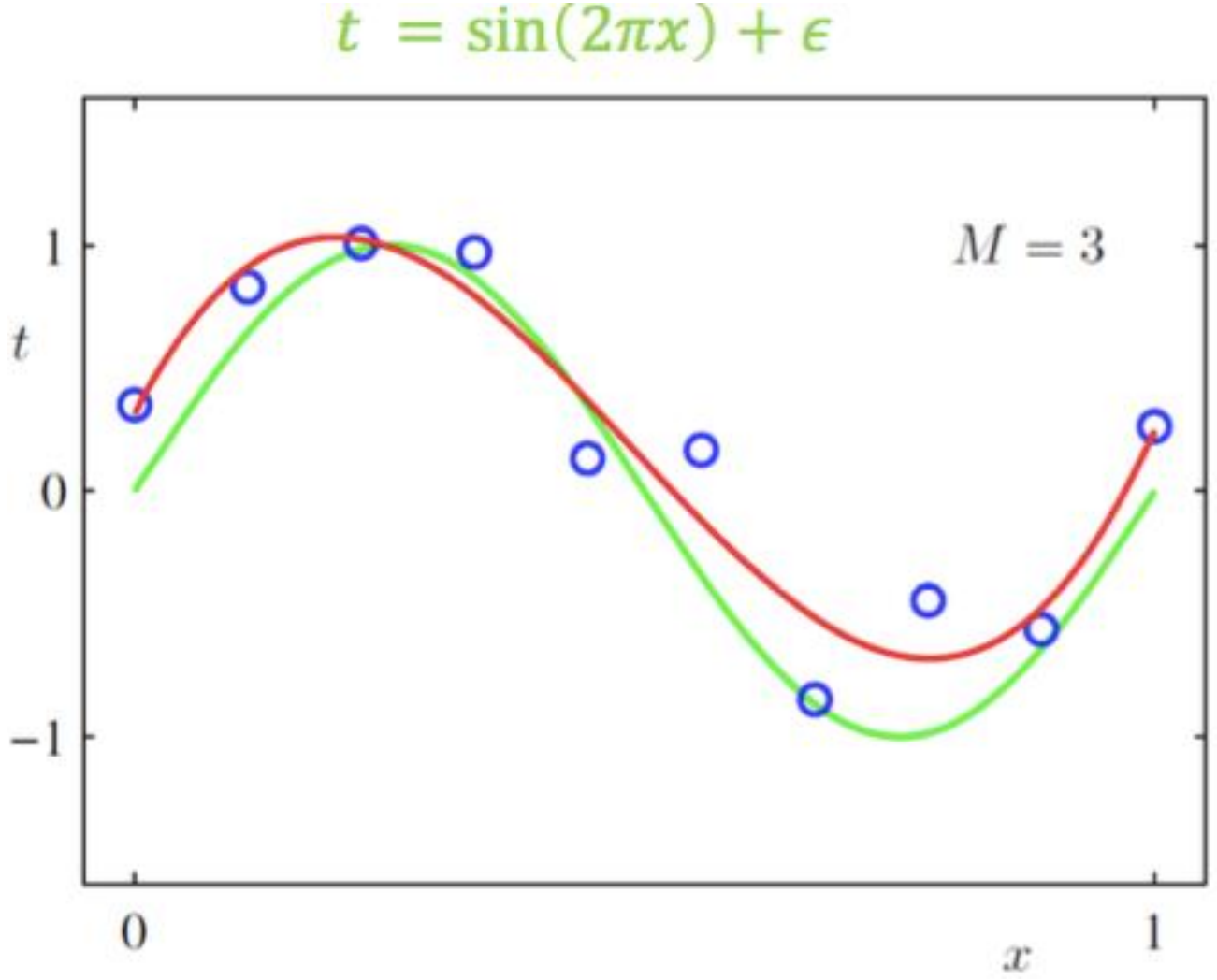
y = c

Example 3: regression using polynomial



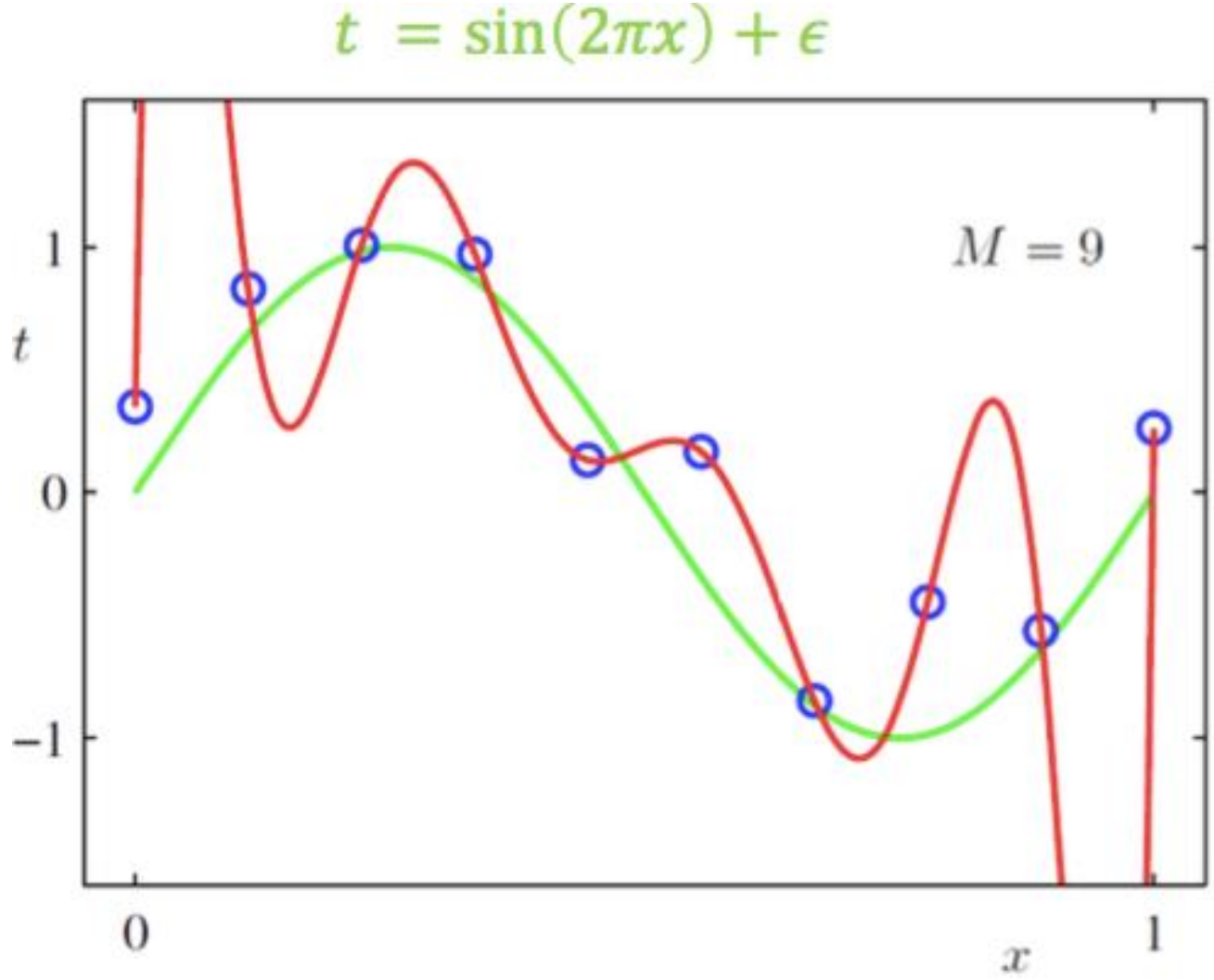
$y = w_1 x + w_0$

Example 3: regression using polynomial



$y = w_3x^3 + w_2x^2 + w_1x + w_0$

Example 3: regression using polynomial



$y = w_9x^9 + w_8x^8 + \ldots + w_1x + w_0$

Overfits, why?