# MACHINE LEARNING

## CHAPTER 2: LINEAR MODELS FOR REGRESSION

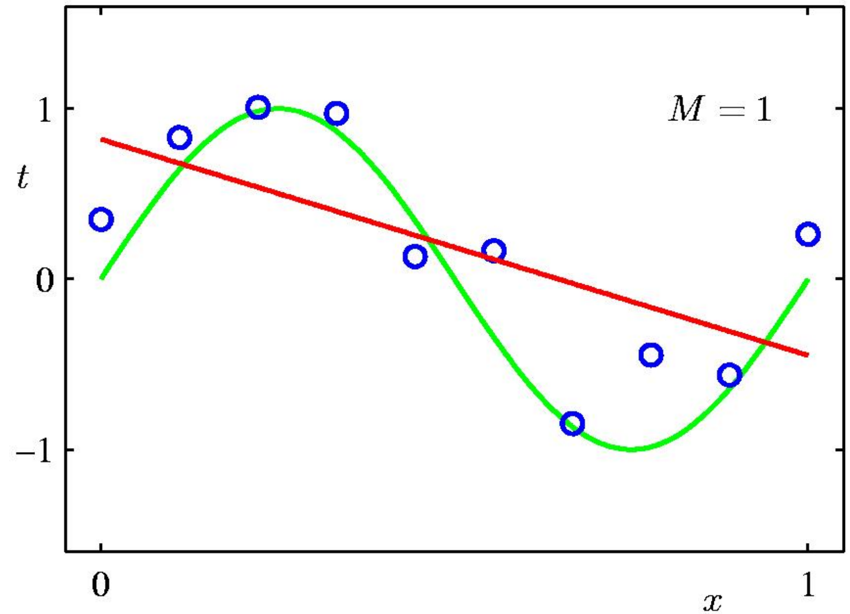
# LINEAR MODELS FOR REGRESSION

1. The concept of regression
2. Maximum Likelihood and Least Square
3. Over-fitting and Regularization
4. The Bias-Variance Trade-off
5. Bayesian Linear Regression
6. Sparse regression

# The Bias-Variance Decomposition

Model too "simple" → does not fit the data well

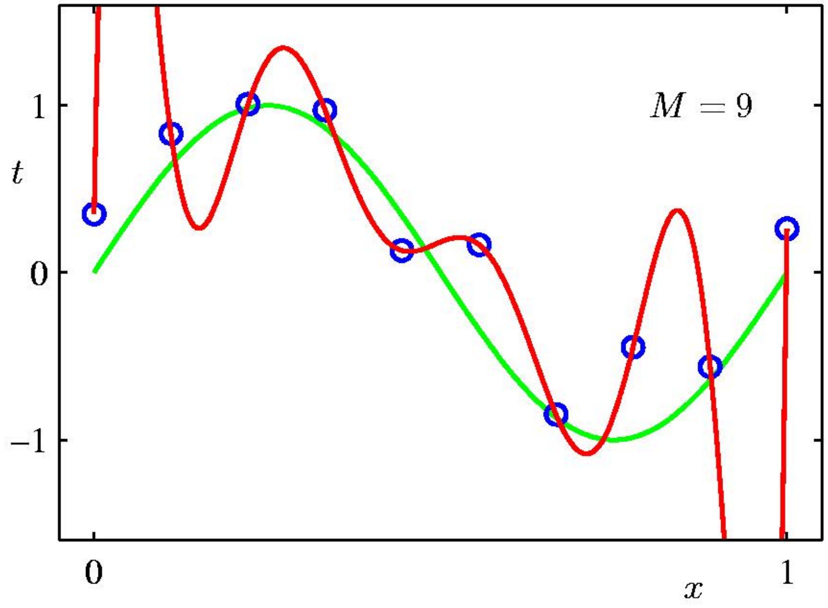
- A biased solution



# The Bias-Variance Decomposition

Model too complex $\longrightarrow$ small changes to the data, solution changes a lot

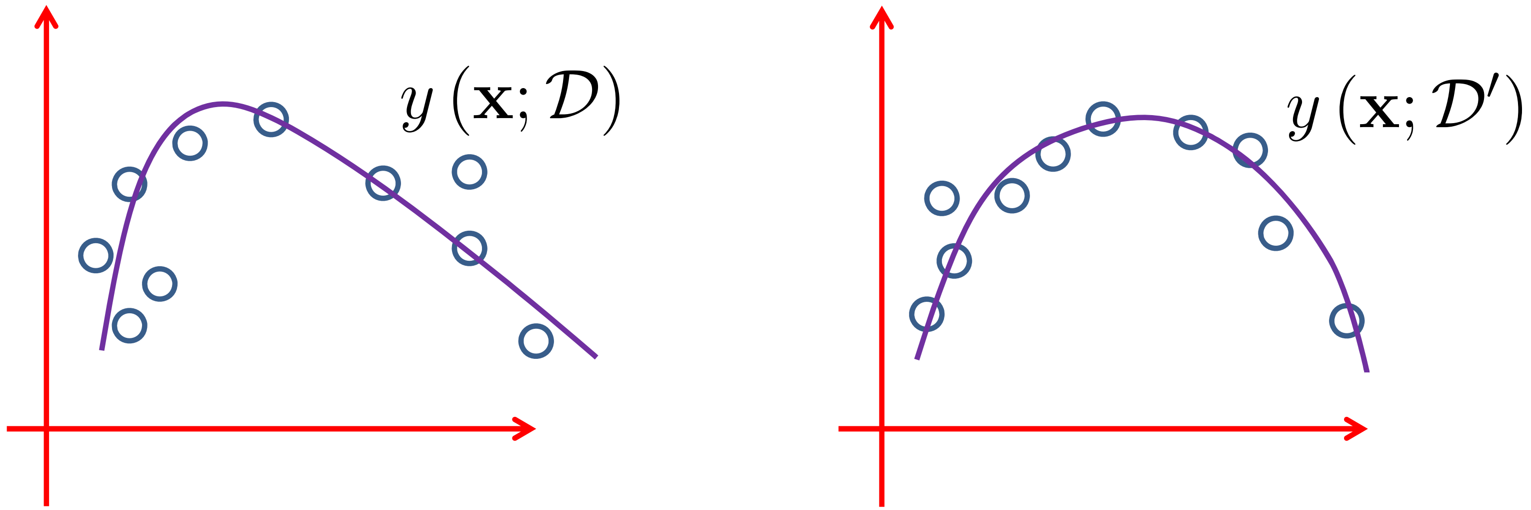
- A high variance solution



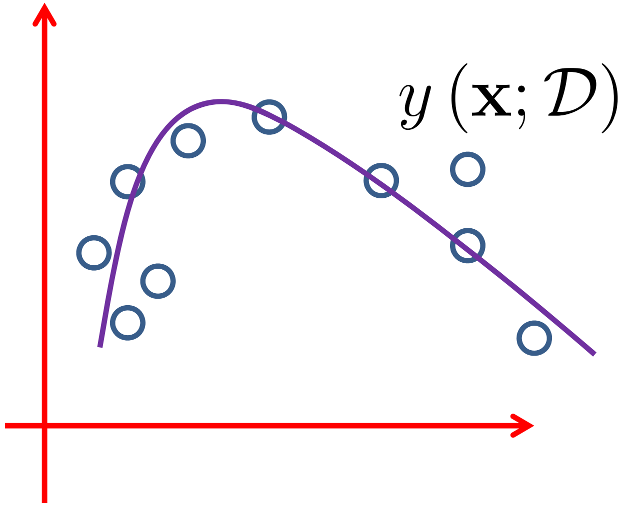
# The Bias-Variance Decomposition

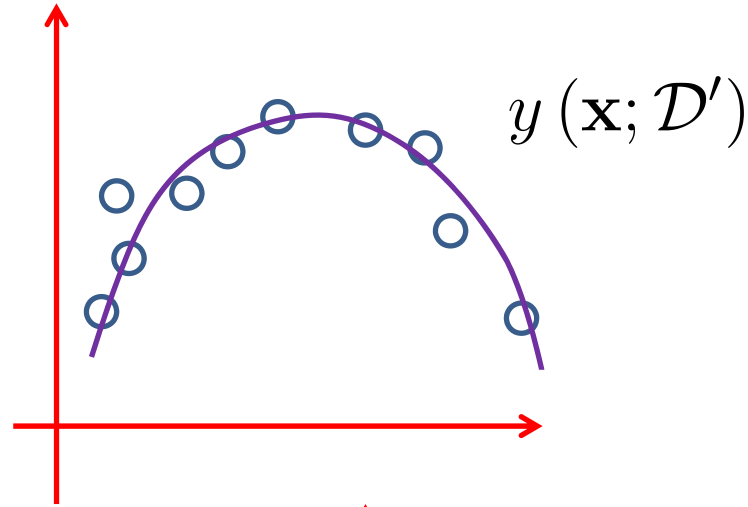
Given dataset $\mathcal{D}$ with N samples, learn function $y\left(\mathbf{x};\mathcal{D}\right)$

If you sample a different dataset $\mathcal{D}'$ , you will learn different $y\left(\mathbf{x};\mathcal{D}'\right)$



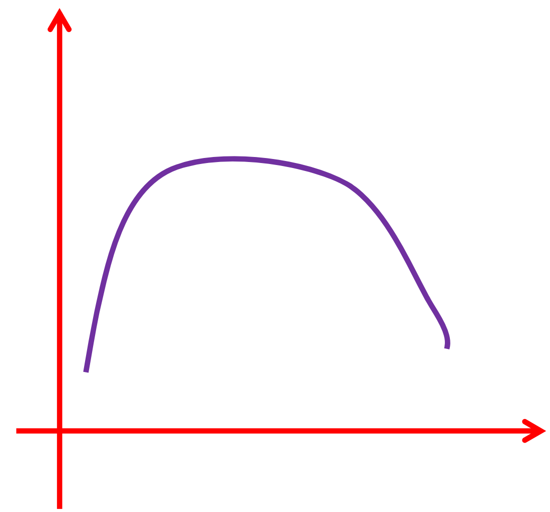
# The Bias-Variance Decomposition




Expected hypothesis: $E_{\mathcal{D}}\left[y(\mathbf{x}; \mathcal{D})\right]$

Take the average over the ensemble of data sets.



# The Bias-Variance Decomposition

Bias: difference between what you expected to learn and the truth

- Measure how well you expected to represent true solution
- Decreased with more complex model

$$(\text{bias})^2 = \int \{\underbrace{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})]}_{\text{average learned model}} - \underbrace{h(\mathbf{x})}_{\text{truth}}\}^2 \underbrace{p(\mathbf{x})}_{\text{distribution of input}} \, \mathrm{d}\mathbf{x}$$

# The Bias-Variance Decomposition

Variance: difference between what you expected to learn and what you learn from a particular dataset

- Measure how sensitive learner is to specific dataset
- Decreases with simple model

$$\text{variance} \quad = \quad \int \mathbb{E}_{\mathcal{D}}\left[\{y(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})]\}^2\right] p(\mathbf{x})\,\mathrm{d}\mathbf{x}$$

# The Bias-Variance Decomposition

Choice of hypothesis class introduces learning bias

- More complex class $\longrightarrow$ less bias
- More complex class $\longrightarrow$ more variance

# The Bias-Variance Decomposition

Consider simple regression problem:

$$t = \underbrace{h(\mathbf{x})}_{\text{True function}} + \underbrace{\epsilon}_{\text{noise } \mathcal{N}(\mathbf{0}, \sigma^2)}$$

Collect some data, and learn a function $y(\mathbf{x}; \mathcal{D})$

What are sources of prediction error?

# The Bias-Variance Decomposition

The expected square error over fixed size training sets $\mathcal{D}$ drawn from $p(\mathbf{x}, t)$ can be expressed as

$$\boxed{\mathbb{E}[L] = \int\int \{y(\mathbf{x};\mathcal{D}) - t\}^2 p(\mathbf{x}, t)\, d\mathbf{x}dt}$$

# The Bias-Variance Decomposition

Recall the *expected squared loss,*

$$\mathbb{E}\left[L\right] = \int \left\{y\left(\mathbf{x},\mathcal{D}\right) - h\left(\mathbf{x}\right)\right\}^2 p\left(\mathbf{x}\right) d\mathbf{x} + \underbrace{\iint \{h(\mathbf{x}) - t\}^2 p(\mathbf{x},t)\,\mathrm{d}\mathbf{x}\,\mathrm{d}t}$$

where

The second term of $\mathbb{E}\left[L\right]$ corresponds to the noise inherent in the random variable $\mathrm{w}$.

The first term of $\mathbb{E}\left[L\right]$ corresponds to the squared bias and variance.

# The Bias-Variance Decomposition

Suppose we were given multiple data sets, each of size N. Any particular data set, $\mathcal{D}$, will give a particular function $y(\mathbf{x}; \mathcal{D})$. We then have

$$\begin{aligned}
&\{y(\mathbf{x};\mathcal{D}) - h(\mathbf{x})\}^2 \\
&\quad = \quad \{y(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] + \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] - h(\mathbf{x})\}^2 \\
&\quad = \quad \{y(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})]\}^2 + \{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] - h(\mathbf{x})\}^2 \\
&\qquad\quad +2\{y(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})]\}\{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] - h(\mathbf{x})\}.
\end{aligned}$$

# The Bias-Variance Decomposition

Taking the expectation over $\mathcal{D}$ yields

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}}\left[\{y(\mathbf{x};\mathcal{D}) - h(\mathbf{x})\}^2\right] \\
&\quad = \underbrace{\{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] - h(\mathbf{x})\}^2}_{(\text{bias})^2} + \underbrace{\mathbb{E}_{\mathcal{D}}\left[\{y(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})]\}^2\right]}_{\text{variance}}.
\end{aligned}$$

# The Bias-Variance Decomposition

Thus we can write

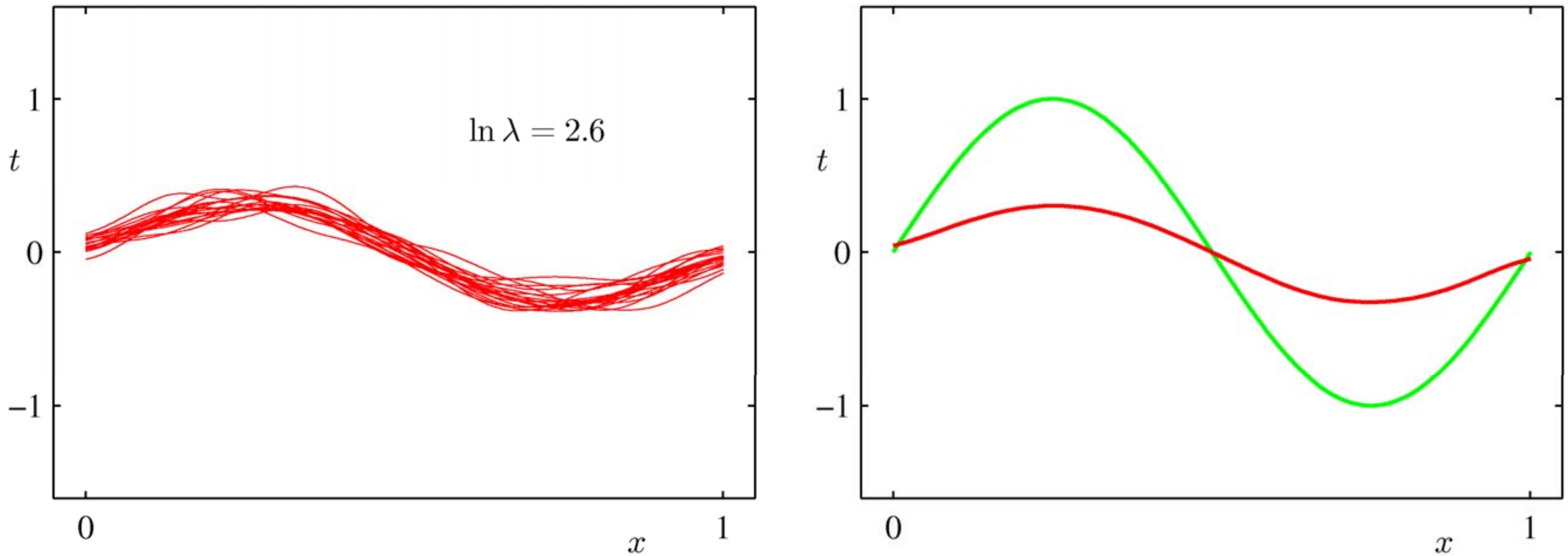
$$\text{expected loss} = (\text{bias})^2 + \text{variance} + \text{noise}$$

where

$$\begin{aligned}
(\text{bias})^2 &= \int \{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] - h(\mathbf{x})\}^2 p(\mathbf{x})\,\mathrm{d}\mathbf{x} \\
\text{variance} &= \int \mathbb{E}_{\mathcal{D}}\left[\{y(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})]\}^2\right] p(\mathbf{x})\,\mathrm{d}\mathbf{x} \\
\text{noise} &= \iint \{h(\mathbf{x}) - t\}^2 p(\mathbf{x},t)\,\mathrm{d}\mathbf{x}\,\mathrm{d}t
\end{aligned}$$

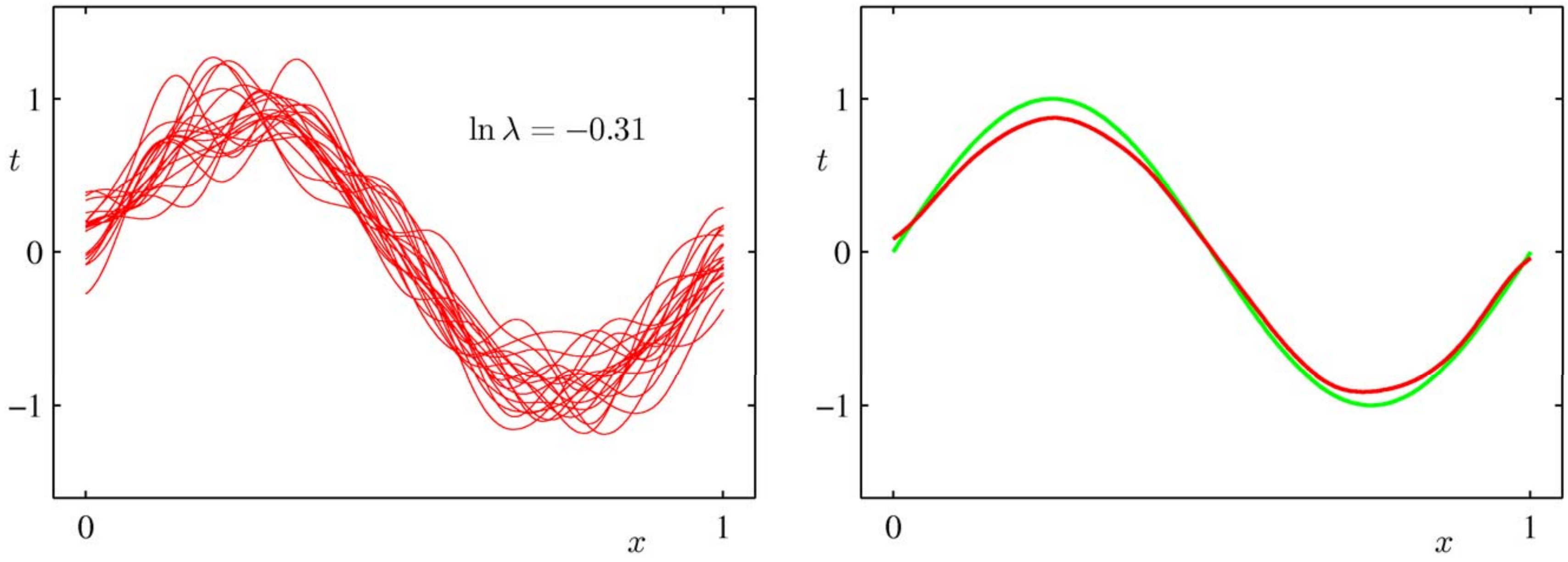
# The Bias-Variance Decomposition

Example: 25 data sets from the sinusoidal, varying the degree of regularization, $\lambda$.



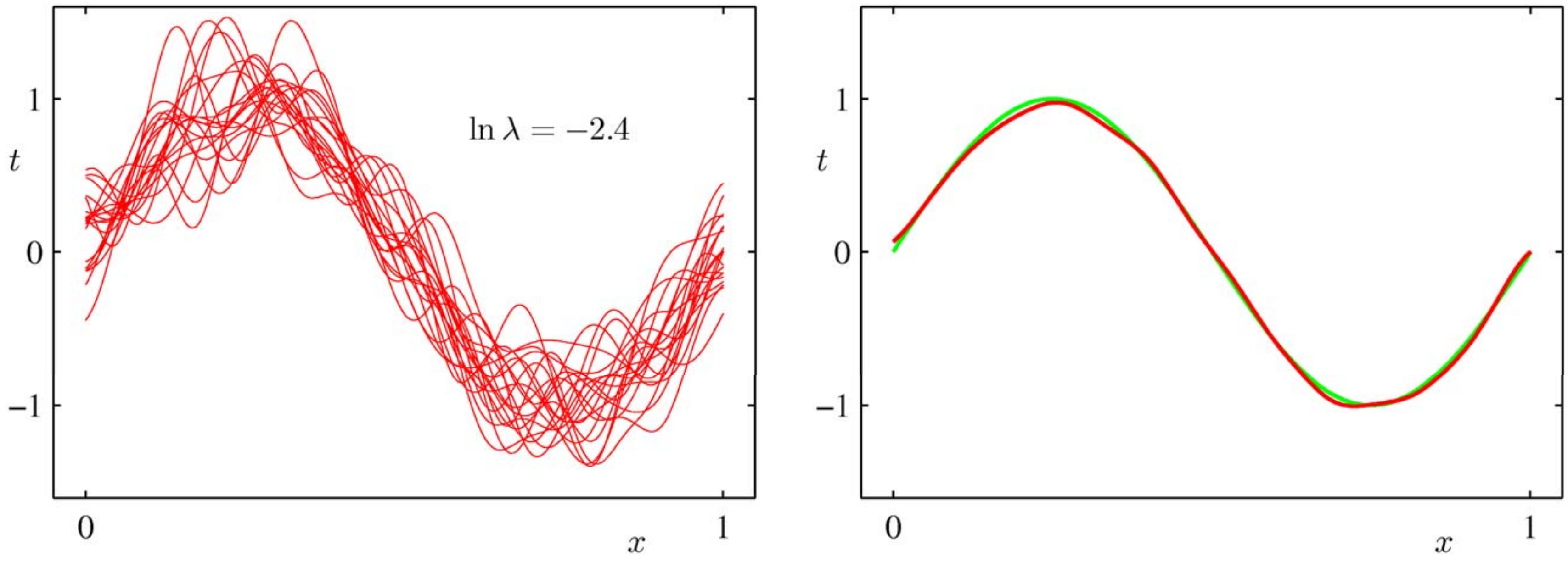
# The Bias-Variance Decomposition

Example: 25 data sets from the sinusoidal, varying the degree of regularization, $\lambda$.



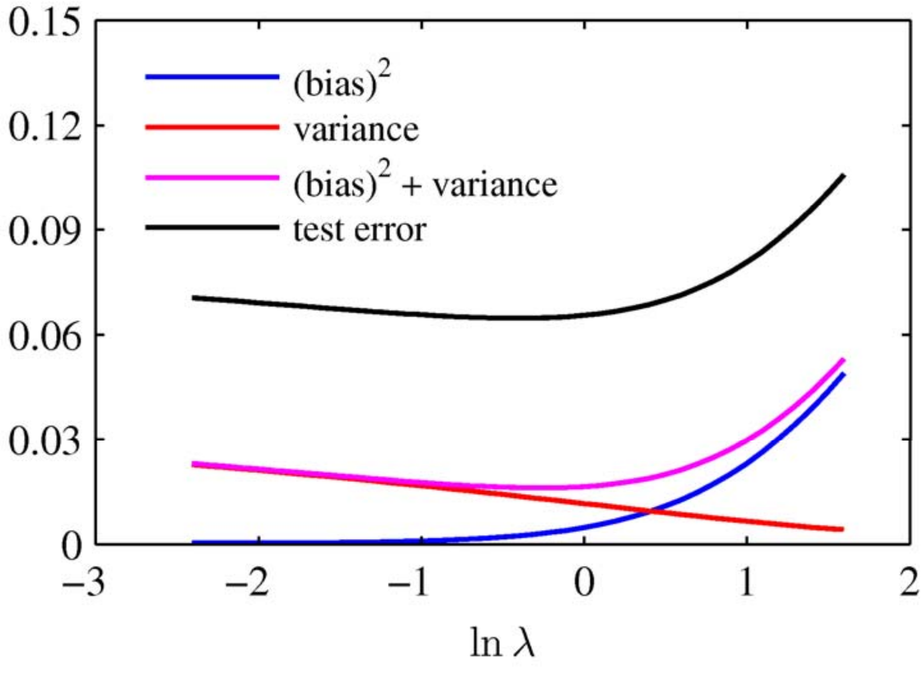
# The Bias-Variance Decomposition

Example: 25 data sets from the sinusoidal, varying the degree of regularization, $\lambda$.



# The Bias-Variance Trade-off

From these plots, we note that an over-regularized model (large $\lambda$) will have a high bias, while an under-regularized model (small $\lambda$) will have a high variance.



# The Bias-Variance Trade-off

Training set error

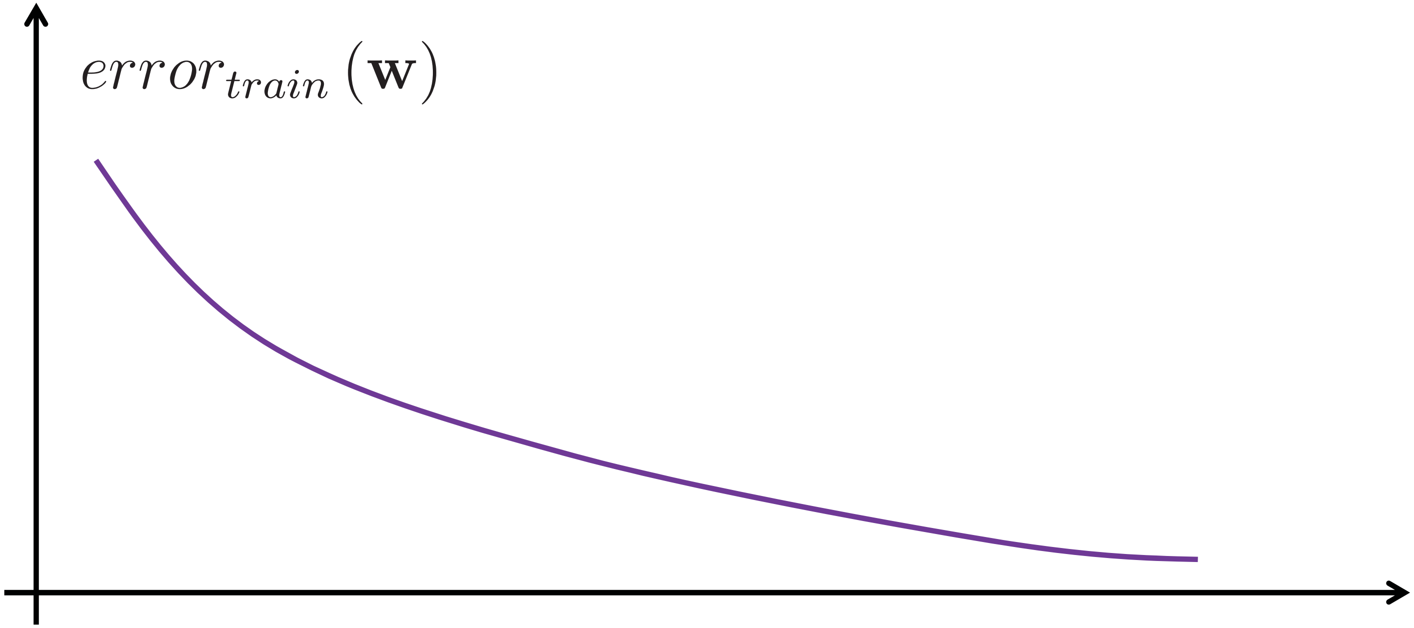
- Given a dataset (Training data)
- Choose a loss function

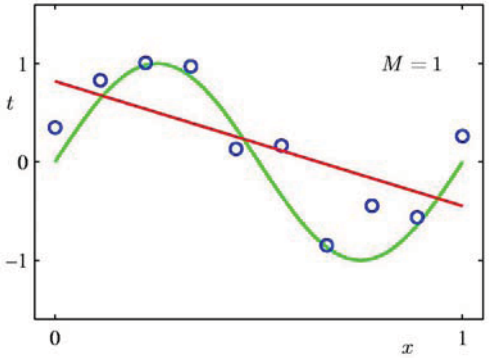
  e.g. squared error for regression

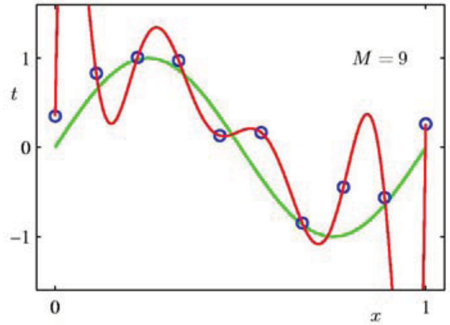
Training set error: For a particular set of parameters, loss function on training data:

$$error_{train}(\mathbf{w}) = \frac{1}{N_{train}} \sum_{j=1}^{N_{train}} \left( t_j - \sum_i w_i \phi_i(\mathbf{x}_j) \right)^2$$

# The Bias-Variance Decomposition



# The Bias-Variance Decomposition

Prediction error

- Training set error can be poor measure of "quality " of solution
- Prediction error: we really care about error over all possible input points, not just training data:

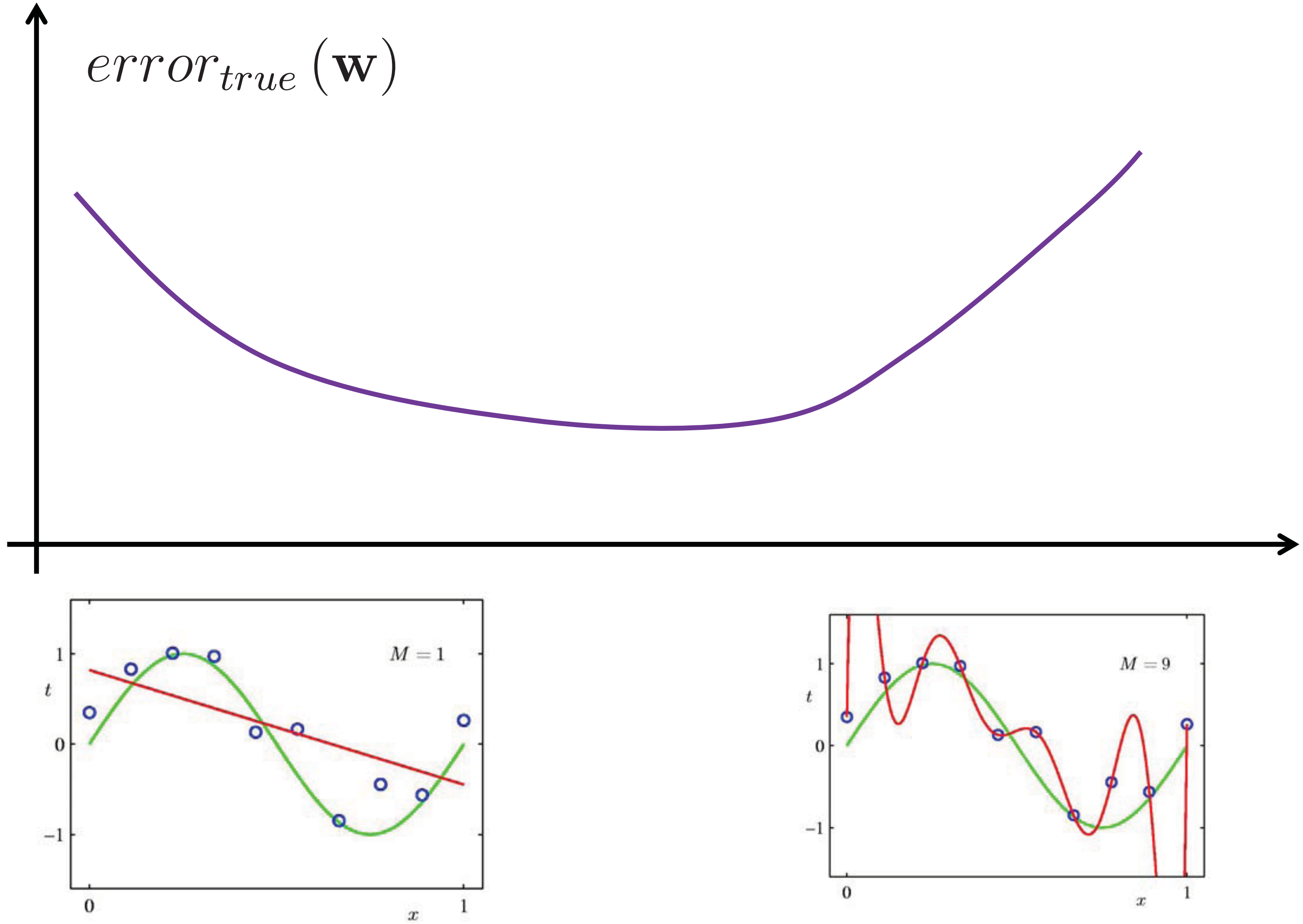
$$error_{true}\left(\mathbf{w}\right) = \mathbb{E}_{\mathbf{x}}\left[\left(t - \sum_{i} w_i \phi_i\left(\mathbf{x}\right)\right)^2\right]$$

# The Bias-Variance Decomposition

$$error_{true}\left(\mathbf{w}\right) = \mathbb{E}_{\mathbf{x}}\left[\left(t - \sum_i w_i \phi_i\left(\mathbf{x}\right)\right)^2\right]$$

$$= \int \left(t - \sum_i w_i \phi_i\left(\mathbf{x}\right)\right)^2 p\left(\mathbf{x}\right) d\mathbf{x}$$

# The Bias-Variance Decomposition



# The Bias-Variance Decomposition

Computing prediction error

- ❑ Hard integral
- ❑ May not $t$ for every $\boldsymbol{x}$

$$error_{true} = \int \left( t - \sum_i w_i \phi_i \left( \mathbf{x} \right) \right)^2 p \left( \mathbf{x} \right) d\mathbf{x}$$

# The Bias-Variance Decomposition

Computing prediction error

Monte Carlo integration (sampling approximation)

Sample a set of i.i.d. points $\{\mathbf{x}_1, ..., \mathbf{x}_M\}$ from $p(\mathbf{x})$

Approximate integral with sample average

$$error_{true}(\mathbf{w}) \approx \frac{1}{M} \sum_{j=1}^{M} \left( t_j - \sum_i w_i \phi_i(\mathbf{x}_j) \right)^2$$

# The Bias-Variance Decomposition

Sampling approximation of prediction error:

$$error_{true}\left(\mathbf{w}\right) \approx \frac{1}{M}\sum_{j=1}^{M}\left(t_j - \sum_i w_i \phi_i\left(\mathbf{x}_j\right)\right)^2$$

Training error:

$$error_{train}\left(\mathbf{w}\right) = \frac{1}{N_{train}}\sum_{j=1}^{N_{train}}\left(t_j - \sum_i w_i \phi_i\left(\mathbf{x}_j\right)\right)^2$$

Very similar equations!

Why is training set a bad measure of prediction error?

# The Bias-Variance Decomposition

Why is training set a bad measure of prediction error?

Training error good estimate for a single **w**, but you optimized **w** with respect to the training error, and found **w** that is good for this set of samples.

**Training error is a biased estimate of prediction error.**

# The Bias-Variance Decomposition

Test error

Given a dataset, randomly split it into two parts:

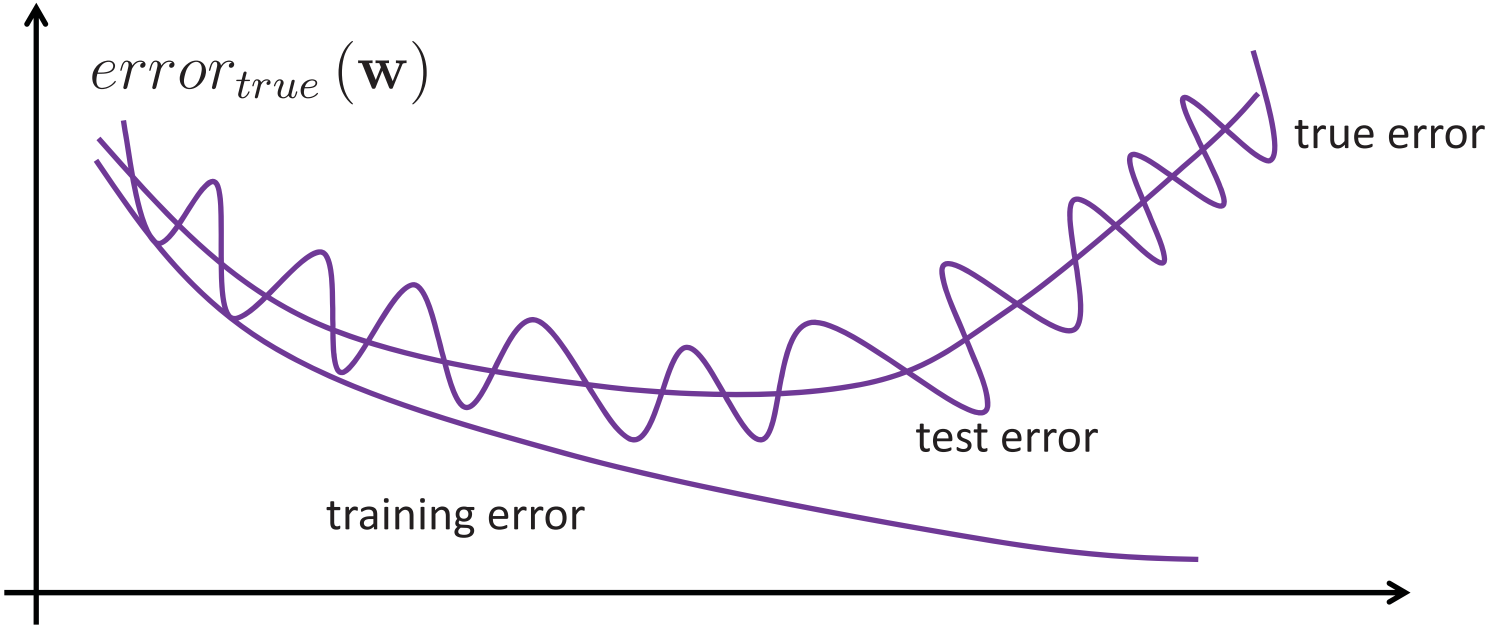
- Training data $\{\mathbf{x}_1, \ldots, \mathbf{x}_{Ntrain}\}$
- Test data $\{\mathbf{x}_1, \ldots, \mathbf{x}_{Ntest}\}$

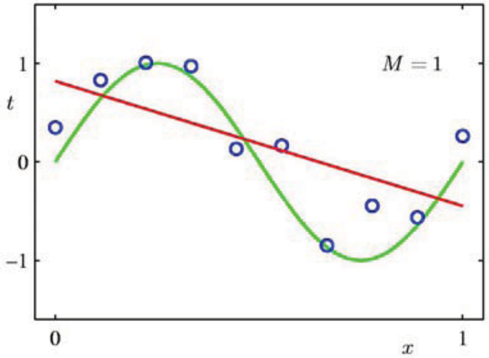
Using training data to optimize parameters **w**

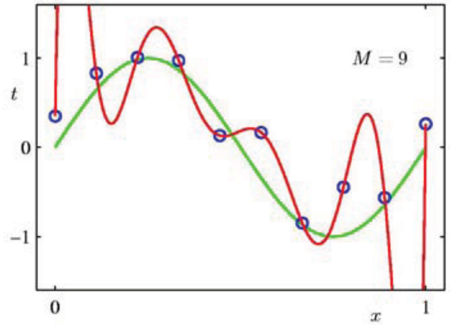
**Test set error:** for the final solution $\boldsymbol{w}^*$, evaluate the error using:

$$error_{test}(\mathbf{w}) = \frac{1}{N_{trest}} \sum_{j=1}^{N_{test}} \left( t_j - \sum_i w_i \phi_i(\mathbf{x}_j) \right)^2$$

# The Bias-Variance Decomposition



# The Bias-Variance Decomposition

Overfitting: a learning algorithm overfits the training data if it outputs a solution **w** when there exists another solution **w'** such that:

$$error_{train}(\mathbf{w}) < error_{train}(\mathbf{w}')$$

and $error_{true}(\mathbf{w}') < error_{true}(\mathbf{w})$