Bias Variance Trade-Off, Regularization, Early Stopping, Dropout

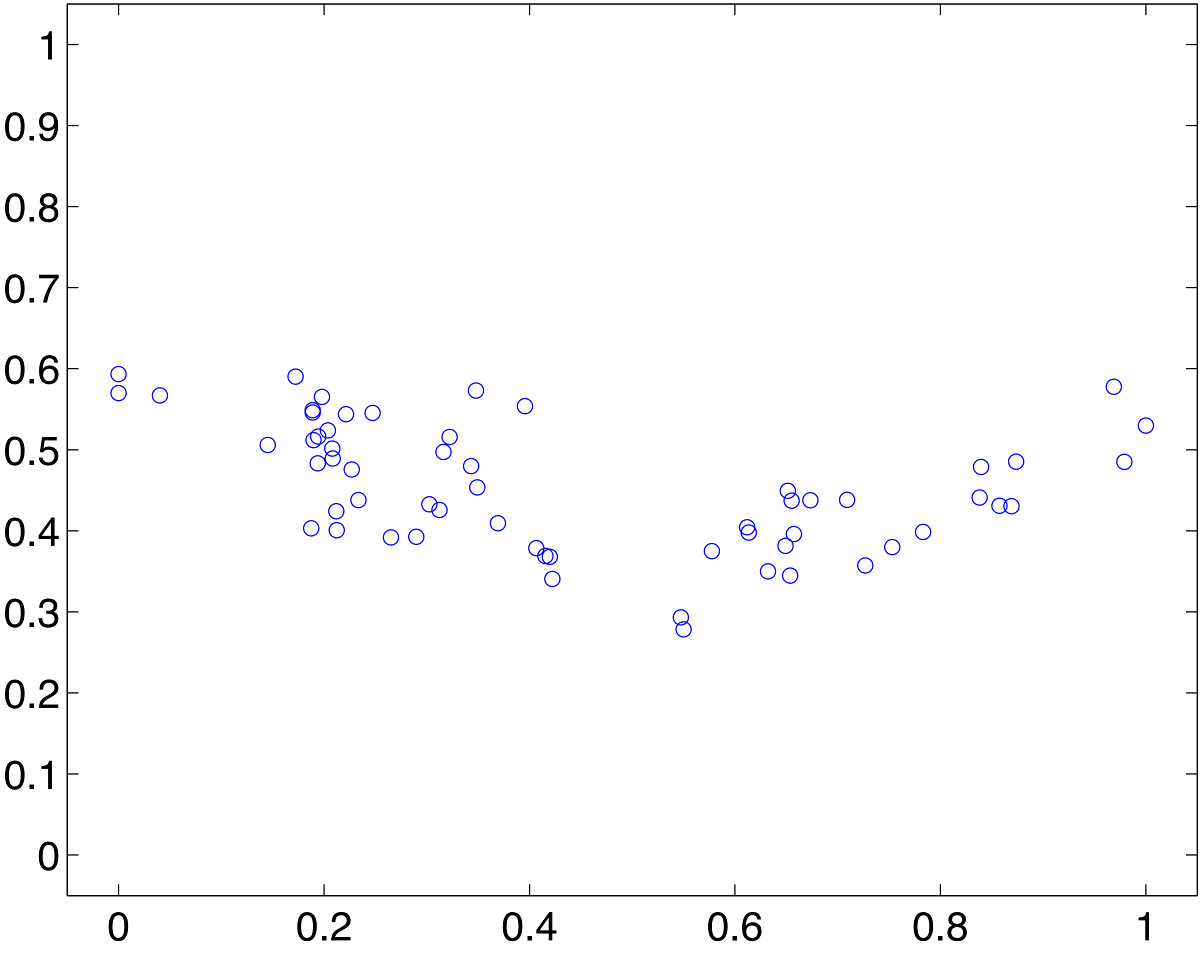
Naresh Manwani

Machine Learning Lab, IIIT Hyderabad

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY

HYDERABAD

Example1: We want to fit a curve for the following data !



Example1: continue

- Here we want to fit a polynomial of degree p as follows.

$$y = w_0 + w_1 x + w_2 x^2 + \ldots + w_p x^p$$

- Training data = $\{(x_1, y_1), \ldots, (x_N, y_N)\}$

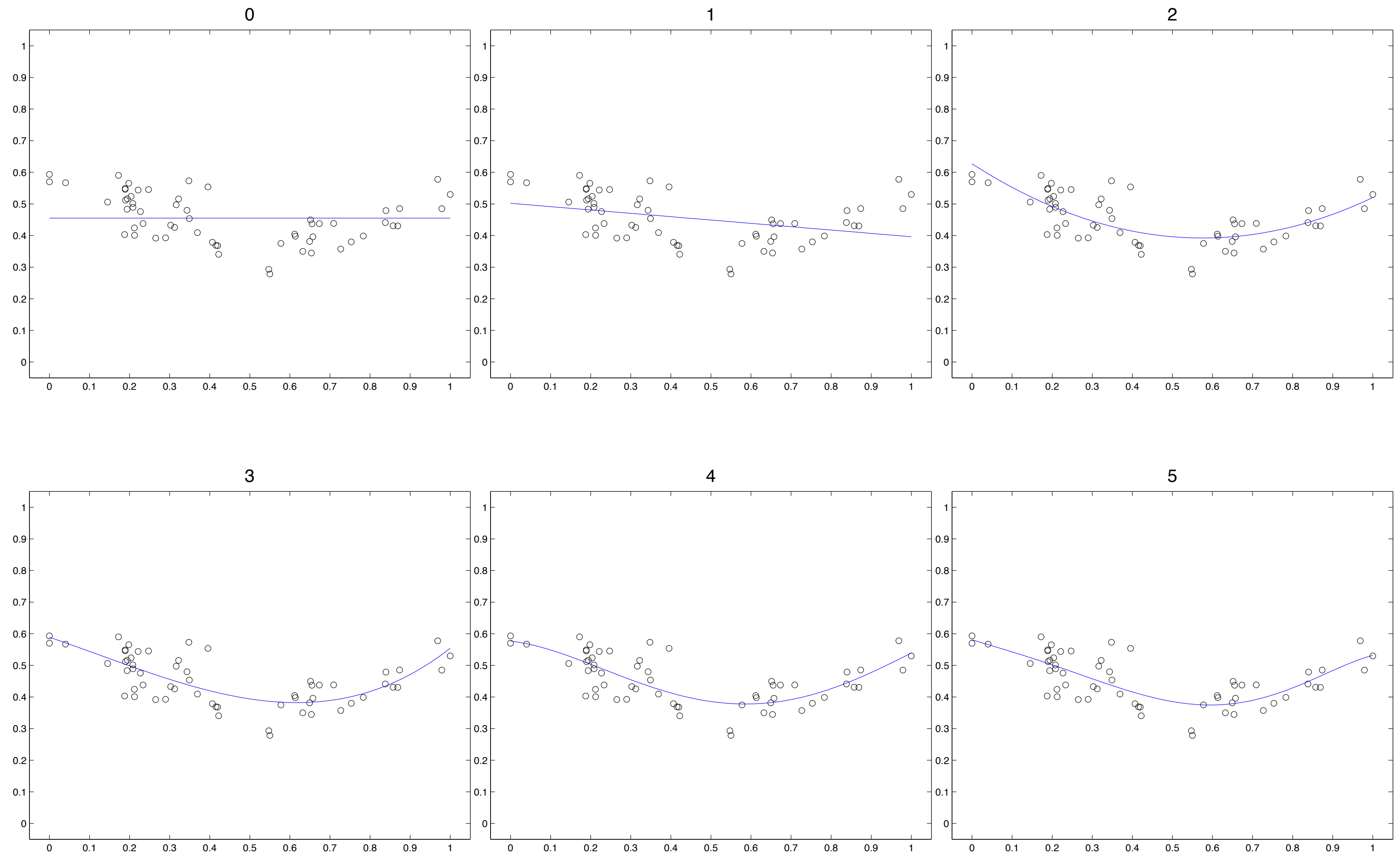
- Test data = $\{(x_{N+1}, y_{N+1}), \ldots, (x_M, y_M)\}$
- Objective function:

$$\text{Training Error} = \frac{1}{2} \sum_{i=1}^{N} (w_0 + w_1 x_i + w_2 x_i^2 + \ldots + w_p x_i^p - y_i)^2$$

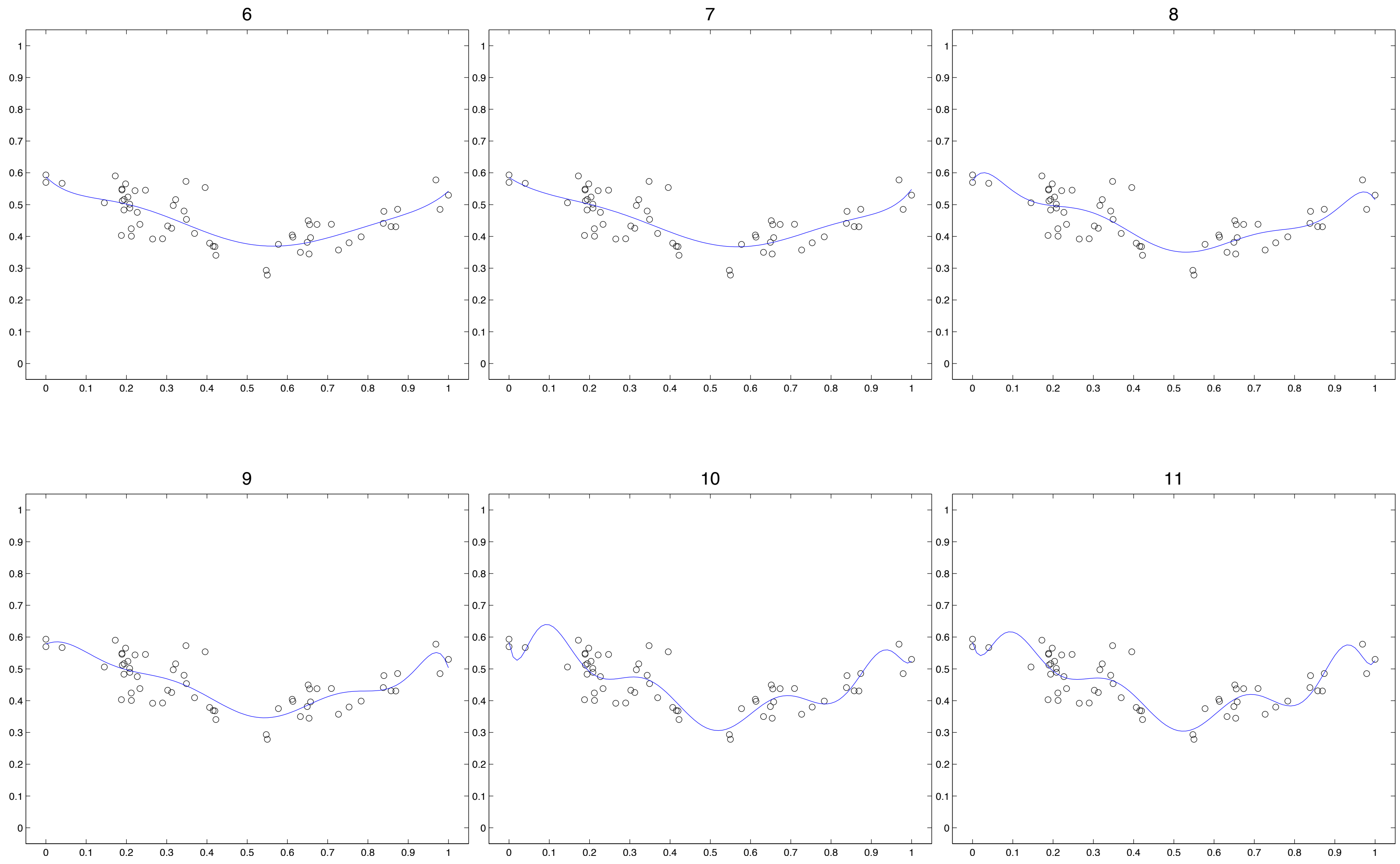
Performance on unseen data

$$\text{Test Error} = \frac{1}{2}\sum_{i=N+1}^{M}(w_0 + w_1 x_i + w_2 x_i^2 + \ldots + w_p x_i^p - y_i)^2$$

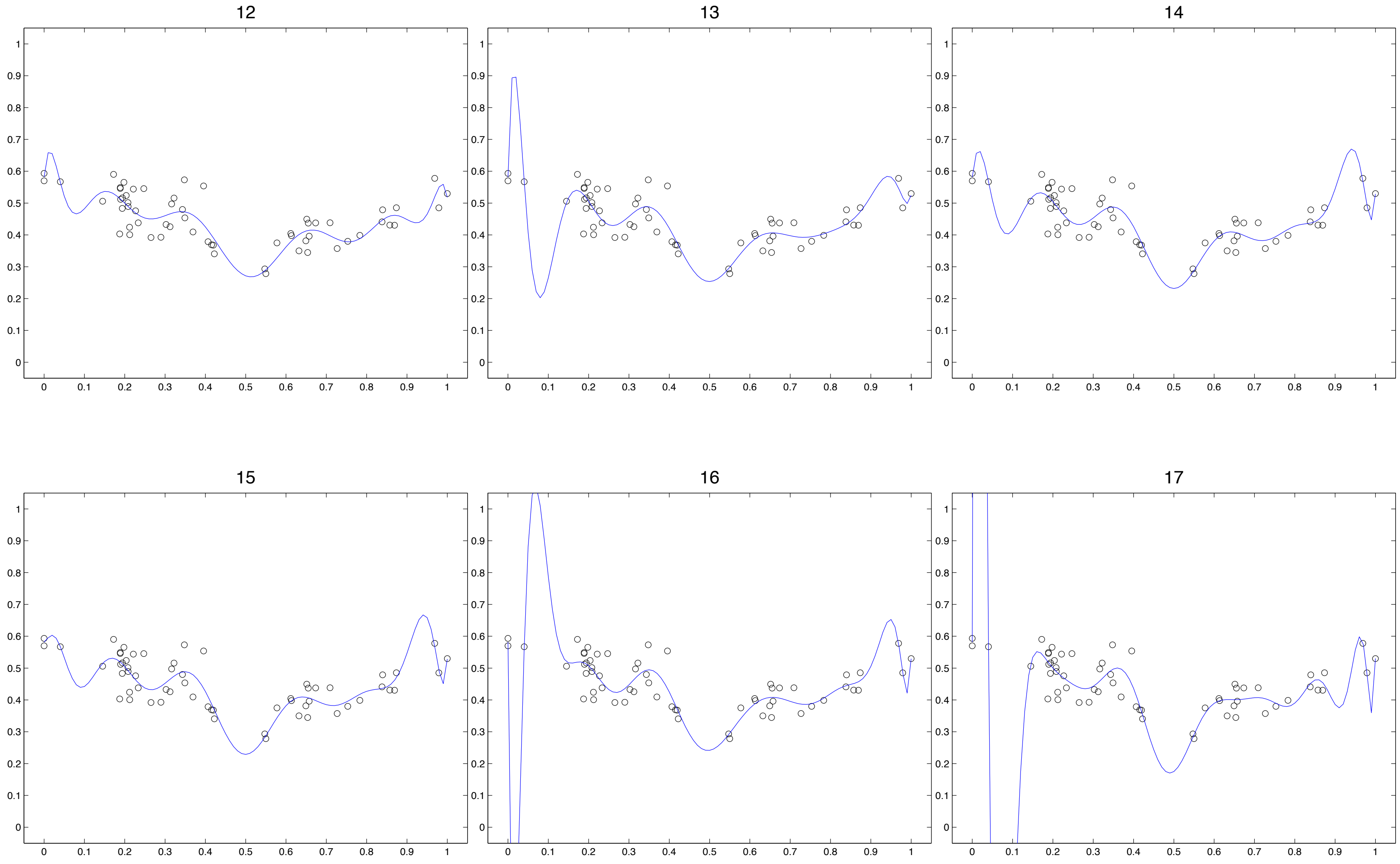
Example1: Fitted curve for p=0,1,2,3,4,5



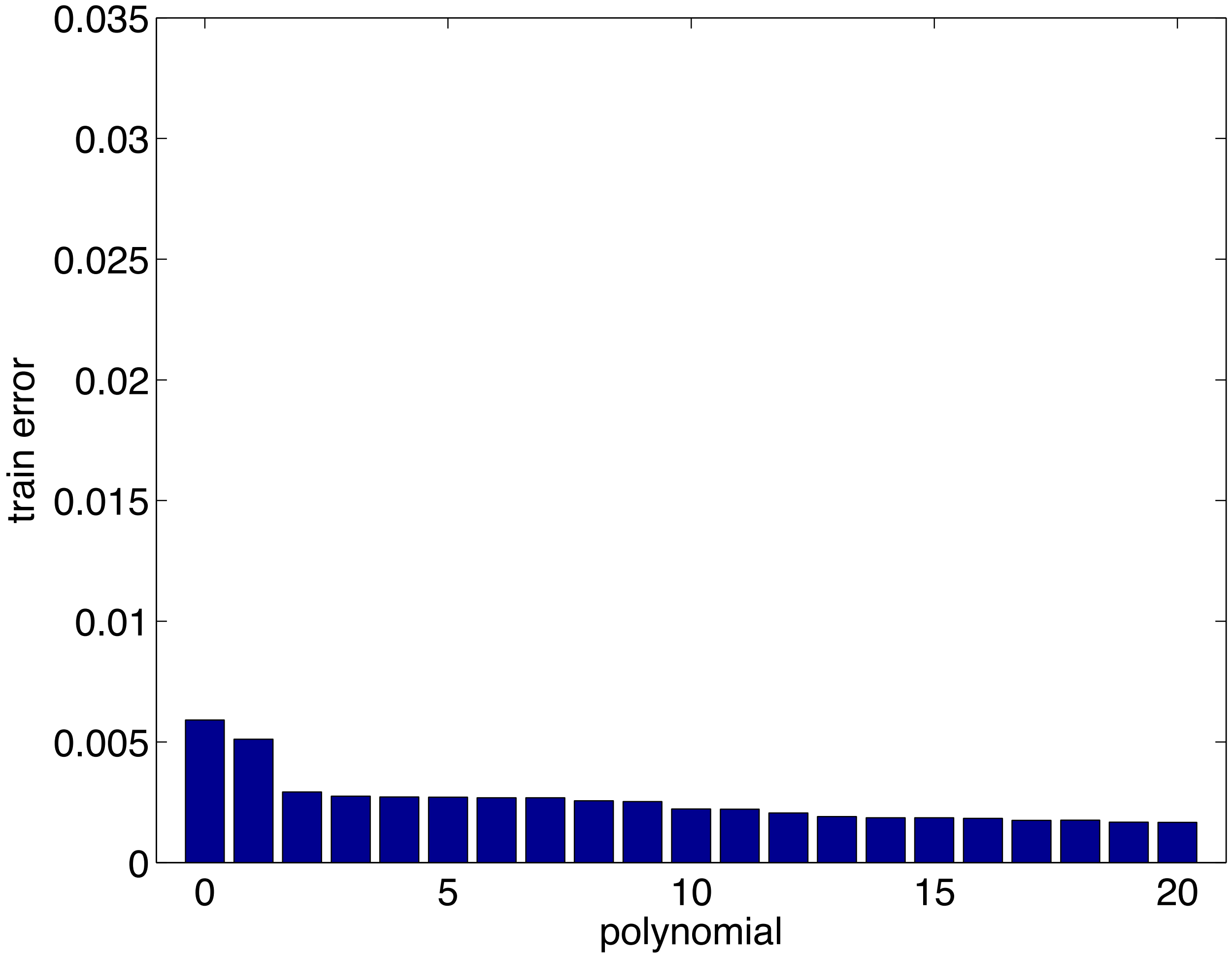
Example1: Fitted curve for p=6,7,8,9,10,11



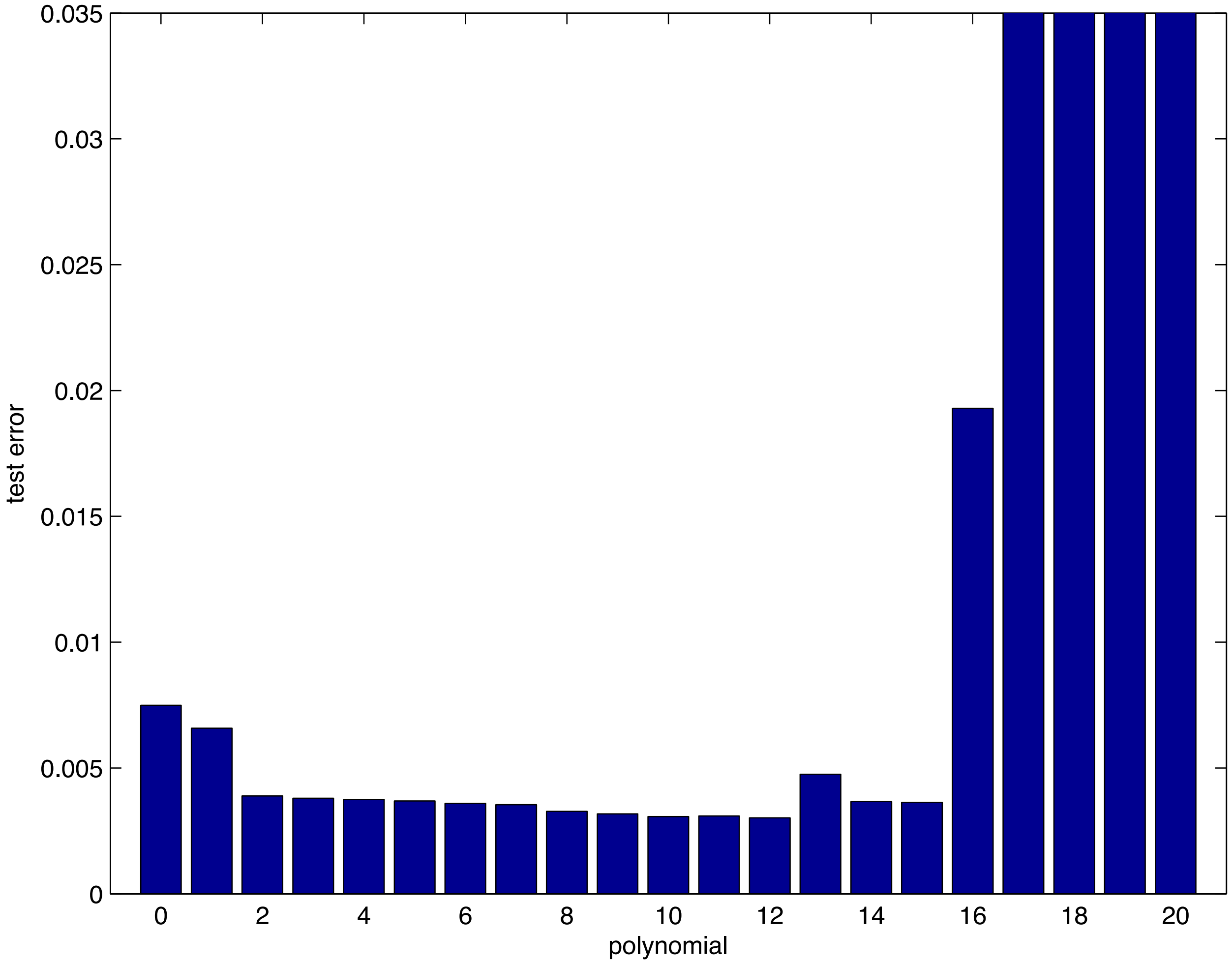
Example1: Fitted curve for p=12,13,14,15,16,17



Example1: Training Error



Example1: Test Error



Bias Variance Tradeoff

- For very low **p**, the model is very simple, and so can't capture the full complexities of the data. It "underfits" the data. This is called **bias**.
- For very high **p**, the model is complex, and so tends to "overfit" to spurious properties of the data. This is called **variance**.

Formalizing Bias and Variance

Given data set

- $\mathcal{D} = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)\}$

And model built from data set,

- $f(\boldsymbol{x}; \mathcal{D})$

We can evaluate the effectiveness of the model using mean squared error:

- $\text{MSE} = \boldsymbol{E}_{p(\boldsymbol{x}, y, \mathcal{D})}\left[\left(y - f(\boldsymbol{x}; \mathcal{D})\right)^2\right]$
- with constant $|\mathcal{D}| = N$

$$\mathbb{E}[L] = \iint L(t, y(\mathbf{x}))p(\mathbf{x}, t)\,\mathrm{d}\mathbf{x}\,\mathrm{d}t. \tag{1.86}$$

A common choice of loss function in regression problems is the squared loss given by $L(t, y(\mathbf{x})) = \{y(\mathbf{x}) - t\}^2$. In this case, the expected loss can be written

$$\mathbb{E}[L] = \iint \{y(\mathbf{x}) - t\}^2 p(\mathbf{x}, t)\,\mathrm{d}\mathbf{x}\,\mathrm{d}t. \tag{1.87}$$

Our goal is to choose $y(\mathbf{x})$ so as to minimize $\mathbb{E}[L]$. If we assume a completely flexible function $y(\mathbf{x})$, we can do this formally using the calculus of variations to give

$$\frac{\delta\mathbb{E}[L]}{\delta y(\mathbf{x})} = 2\int \{y(\mathbf{x}) - t\}p(\mathbf{x}, t)\,\mathrm{d}t = 0. \tag{1.88}$$

Solving for $y(\mathbf{x})$, and using the sum and product rules of probability, we obtain

$$y(\mathbf{x}) = \frac{\displaystyle\int tp(\mathbf{x}, t)\,\mathrm{d}t}{p(\mathbf{x})} = \int tp(t|\mathbf{x})\,\mathrm{d}t = \mathbb{E}_t[t|\mathbf{x}] \tag{1.89}$$

$$
\begin{aligned}
\{y(\mathbf{x}) - t\}^2 &= \{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}] + \mathbb{E}[t|\mathbf{x}] - t\}^2 \\
&= \{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\}^2 + 2\{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\}\{\mathbb{E}[t|\mathbf{x}] - t\} + \{\mathbb{E}[t|\mathbf{x}] - t\}^2
\end{aligned}
$$

The Bias-Variance Decomposition (1)

- Recall the *expected squared loss*,

$$\mathbb{E}[L] = \int \{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\}^2 \, p(\mathbf{x}) \, \mathrm{d}\mathbf{x} + \int \mathrm{var}\,[t|\mathbf{x}] \, p(\mathbf{x}) \, \mathrm{d}\mathbf{x}$$

Lets denote, for simplicity:

$$h(\mathbf{x}) = \mathbb{E}[t|\mathbf{x}] = \int t p(t|\mathbf{x}) \, \mathrm{d}t.$$

- We said that the second term corresponds to the noise inherent in the random variable t.

- What about the first term?

The Bias-Variance Decomposition (2)

- Suppose we were given multiple data sets, each of size N.
- Any particular data set, *D*, will give a particular function y(x; *D*).
- Consider the error in the estimation:

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

The Bias-Variance Decomposition (3)

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

- Taking the expectation over D yields:

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

The Bias-Variance Decomposition (4)

- **Thus we can write**

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

- **where**

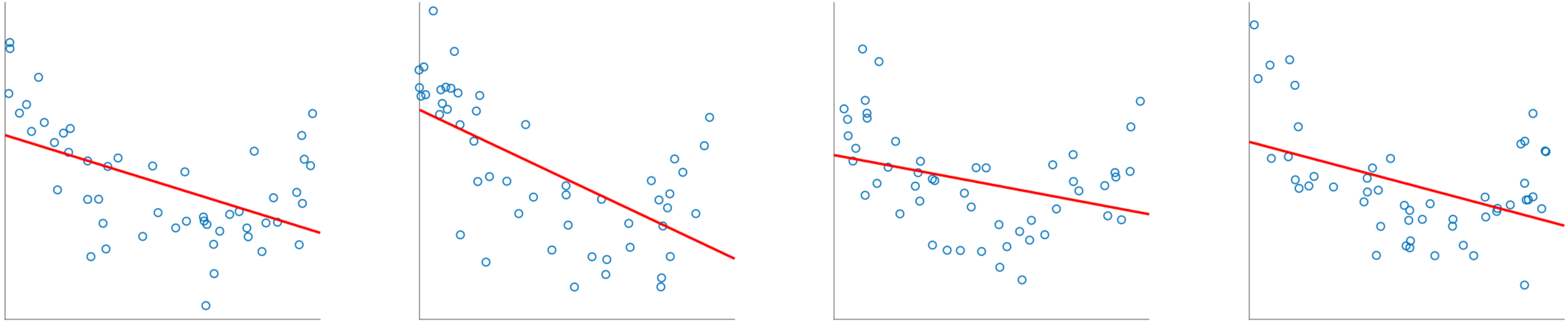
$$\begin{aligned}
(\text{bias})^2 &= \int \{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] - h(\mathbf{x})\}^2 p(\mathbf{x})\,\mathrm{d}\mathbf{x} \\
\text{variance} &= \int \mathbb{E}_{\mathcal{D}}\left[\{y(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})]\}^2\right]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

- **Bias** measures how much the prediction (averaged over all data sets) differs from the desired regression function.

- **Variance** measures how much the predictions for individual data sets vary around their average.

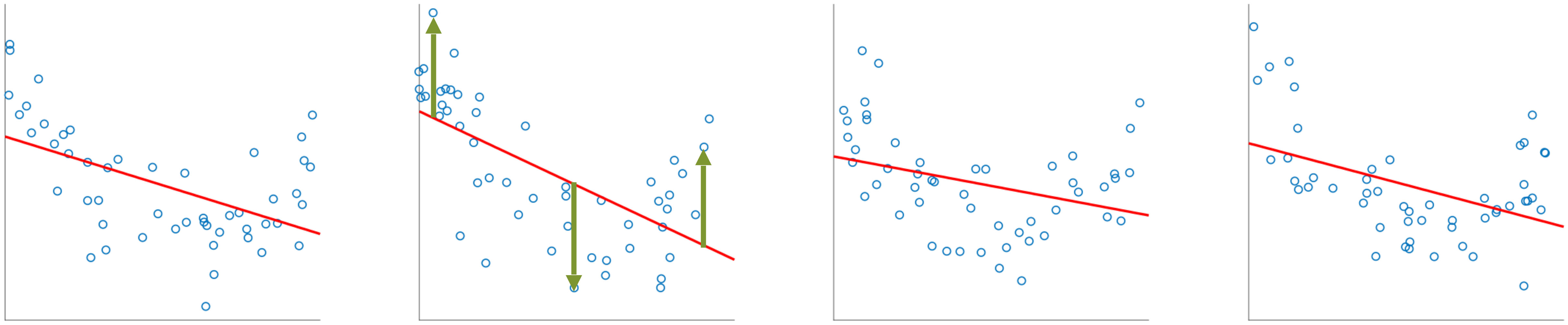
- There is a trade-off between bias and variance
- As we increase **model complexity**,
- bias decreases (a better fit to data) and
- variance increases (fit varies more with data)

Example2: Bias



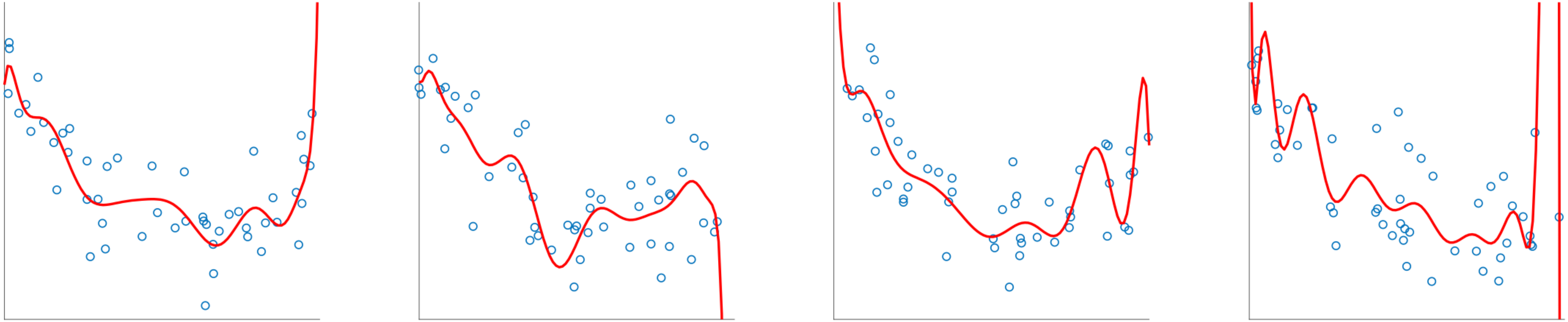
Linear model learnt on different training samples. Regardless of training sample, or size of training sample, model will produce consistent errors

Example2: Bias



Linear model learnt on different training samples. Regardless of training sample, or size of training sample, model will produce consistent errors

Example2: Variance



Keeping the degree p very high. Different samples of training data yield different model fits

$$
\begin{aligned}
\mathrm{MSE}_{\boldsymbol{x}} &= \boldsymbol{E}_{\mathcal{D}|\boldsymbol{x}}\left[\left(y - f(\boldsymbol{x};\mathcal{D})\right)^{2}\right] \\
&= \left(\boldsymbol{E}_{\mathcal{D}}[f(\boldsymbol{x};\mathcal{D})] - \boldsymbol{E}[y|\boldsymbol{x}]\right)^{2} \\
&\quad + \boldsymbol{E}_{\mathcal{D}}\left[\left(f(\boldsymbol{x};\mathcal{D}) - \boldsymbol{E}_{\mathcal{D}}[f(\boldsymbol{x};\mathcal{D})]\right)^{2}\right] \\
&\quad + \boldsymbol{E}\left[\left(y - \boldsymbol{E}[y|\boldsymbol{x}]\right)^{2}\right]
\end{aligned}
$$

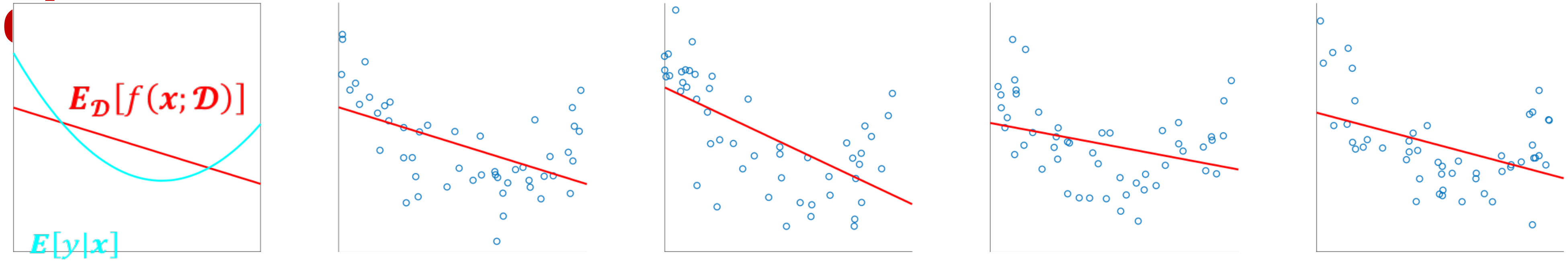
$$\mathrm{MSE}_x = \boldsymbol{E}_{\mathcal{D}|x}\left[\left(y - f(x;\mathcal{D})\right)^2\right]$$

$$= \left(\boldsymbol{E}_{\mathcal{D}}[f(x;\mathcal{D})] - \boldsymbol{E}[y|x]\right)^2$$

$$+ \boldsymbol{E}_{\mathcal{D}}\left[\left(f(x;\mathcal{D}) - \boldsymbol{E}_{\mathcal{D}}[f(x;\mathcal{D})]\right)^2\right]$$

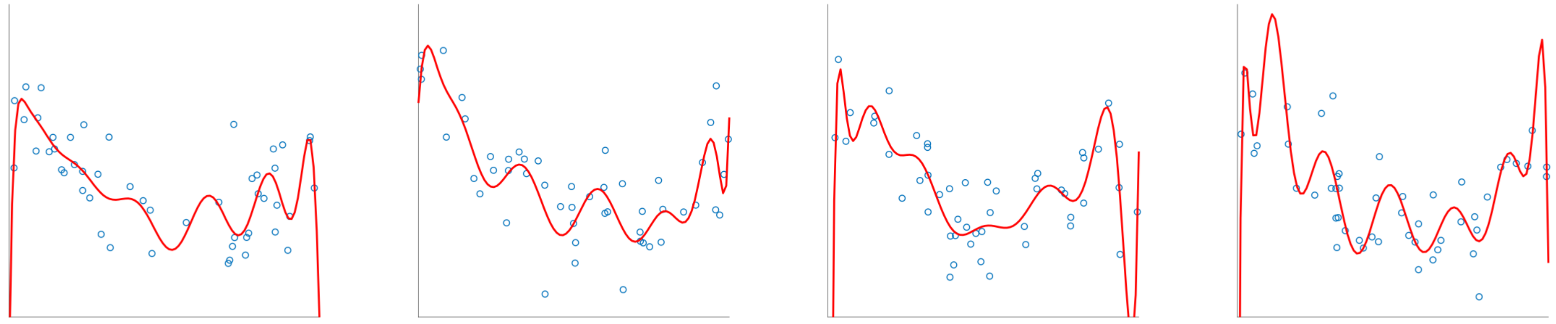
$$+ \boldsymbol{E}\left[\left(y - \boldsymbol{E}[y|x]\right)^2\right]$$

bias: difference between average model prediction (across data sets) and the targ



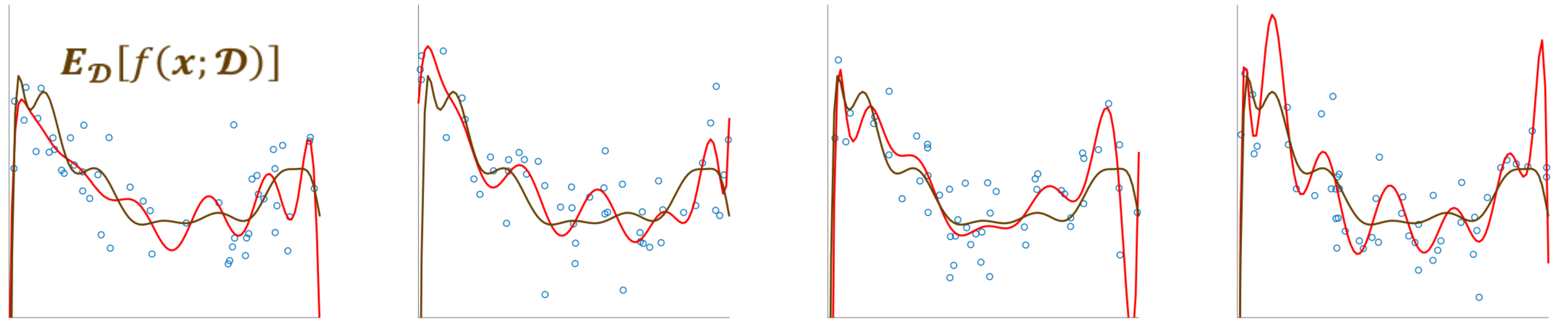
$$
\begin{aligned}
\mathrm{MSE}_x &= \boldsymbol{E}_{\mathcal{D}|x}\left[\left(y - f(\boldsymbol{x};\mathcal{D})\right)^2\right] \\
&= \left(\boldsymbol{E}_{\mathcal{D}}[f(\boldsymbol{x};\mathcal{D})] - \boldsymbol{E}[y|\boldsymbol{x}]\right)^2 \\
&\quad + \boldsymbol{E}_{\mathcal{D}}\left[\left(f(\boldsymbol{x};\mathcal{D}) - \boldsymbol{E}_{\mathcal{D}}[f(\boldsymbol{x};\mathcal{D})]\right)^2\right] \\
&\quad + \boldsymbol{E}\left[\left(y - \boldsymbol{E}[y|\boldsymbol{x}]\right)^2\right]
\end{aligned}
$$

variance of models (across data sets) for a given point



$$
\begin{aligned}
\mathrm{MSE}_x &= \boldsymbol{E}_{\mathcal{D}|\boldsymbol{x}}\left[\left(y - f(\boldsymbol{x};\mathcal{D})\right)^2\right] \\
&= \left(\boldsymbol{E}_{\mathcal{D}}[f(\boldsymbol{x};\mathcal{D})] - \boldsymbol{E}[y|\boldsymbol{x}]\right)^2 \\
&\quad + \boldsymbol{E}_{\mathcal{D}}\left[\left(f(\boldsymbol{x};\mathcal{D}) - \boldsymbol{E}_{\mathcal{D}}[f(\boldsymbol{x};\mathcal{D})]\right)^2\right] \\
&\quad + \boldsymbol{E}\left[\left(y - \boldsymbol{E}[y|\boldsymbol{x}]\right)^2\right]
\end{aligned}
$$

variance of models (across data sets) for a given point



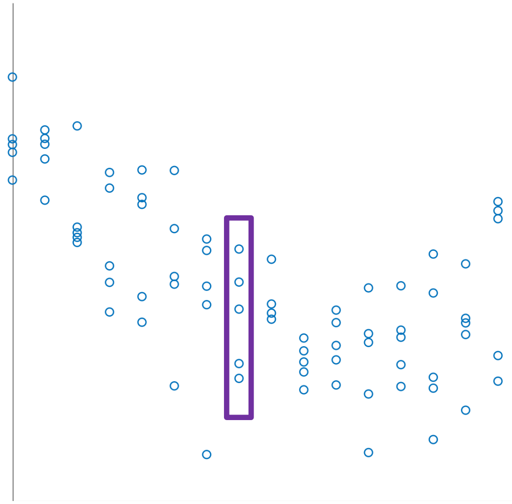
$$\mathrm{MSE}_x = \boldsymbol{E}_{\mathcal{D}|x}\left[\left(y - f(\boldsymbol{x};\mathcal{D})\right)^2\right]$$

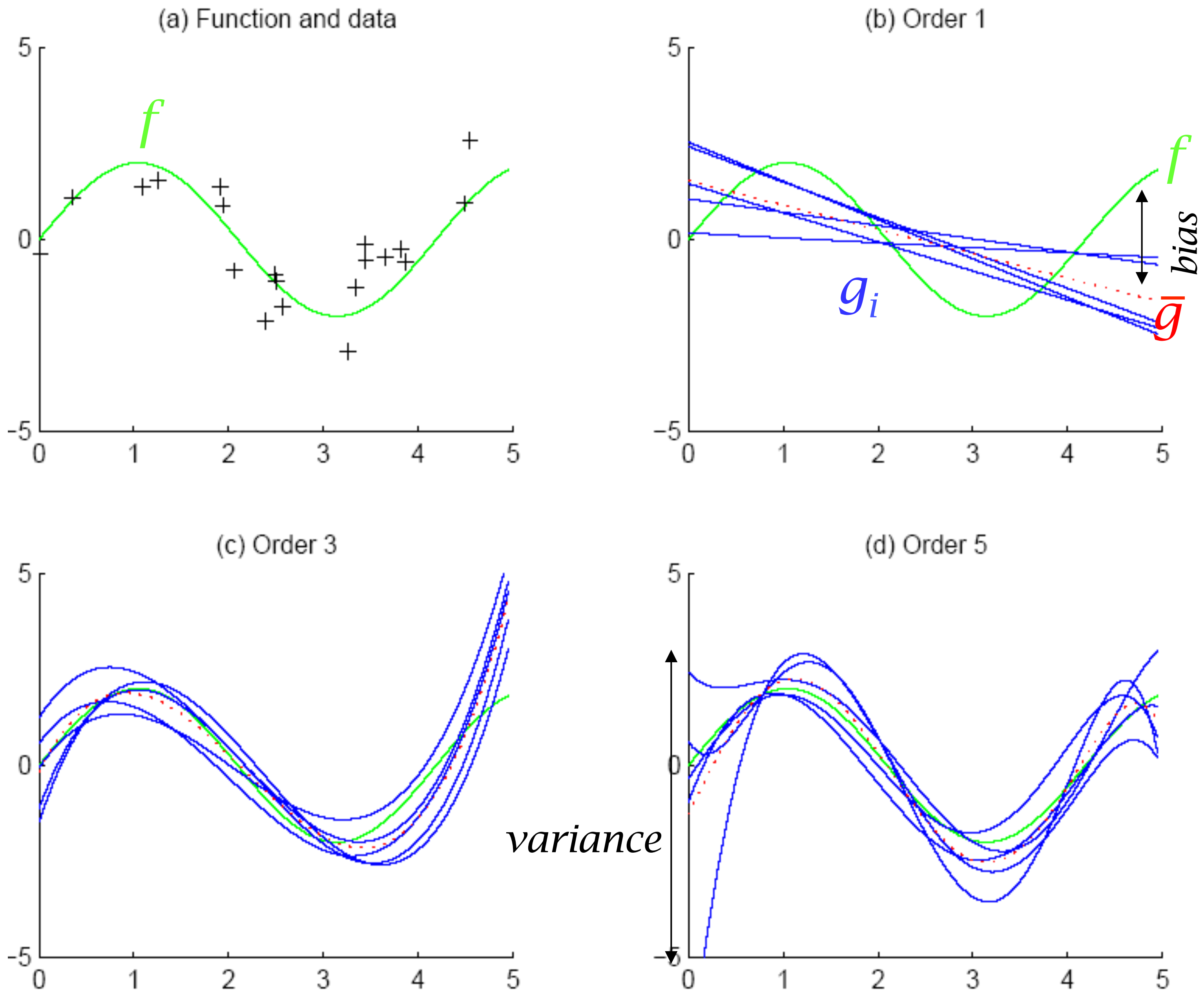
$$= \left(\boldsymbol{E}_{\mathcal{D}}[f(\boldsymbol{x};\mathcal{D})] - \boldsymbol{E}[y|\boldsymbol{x}]\right)^2$$

$$+ \boldsymbol{E}_{\mathcal{D}}\left[\left(f(\boldsymbol{x};\mathcal{D}) - \boldsymbol{E}_{\mathcal{D}}[f(\boldsymbol{x};\mathcal{D})]\right)^2\right]$$

$$+ \boldsymbol{E}\left[\left(y - \boldsymbol{E}[y|\boldsymbol{x}]\right)^2\right]$$

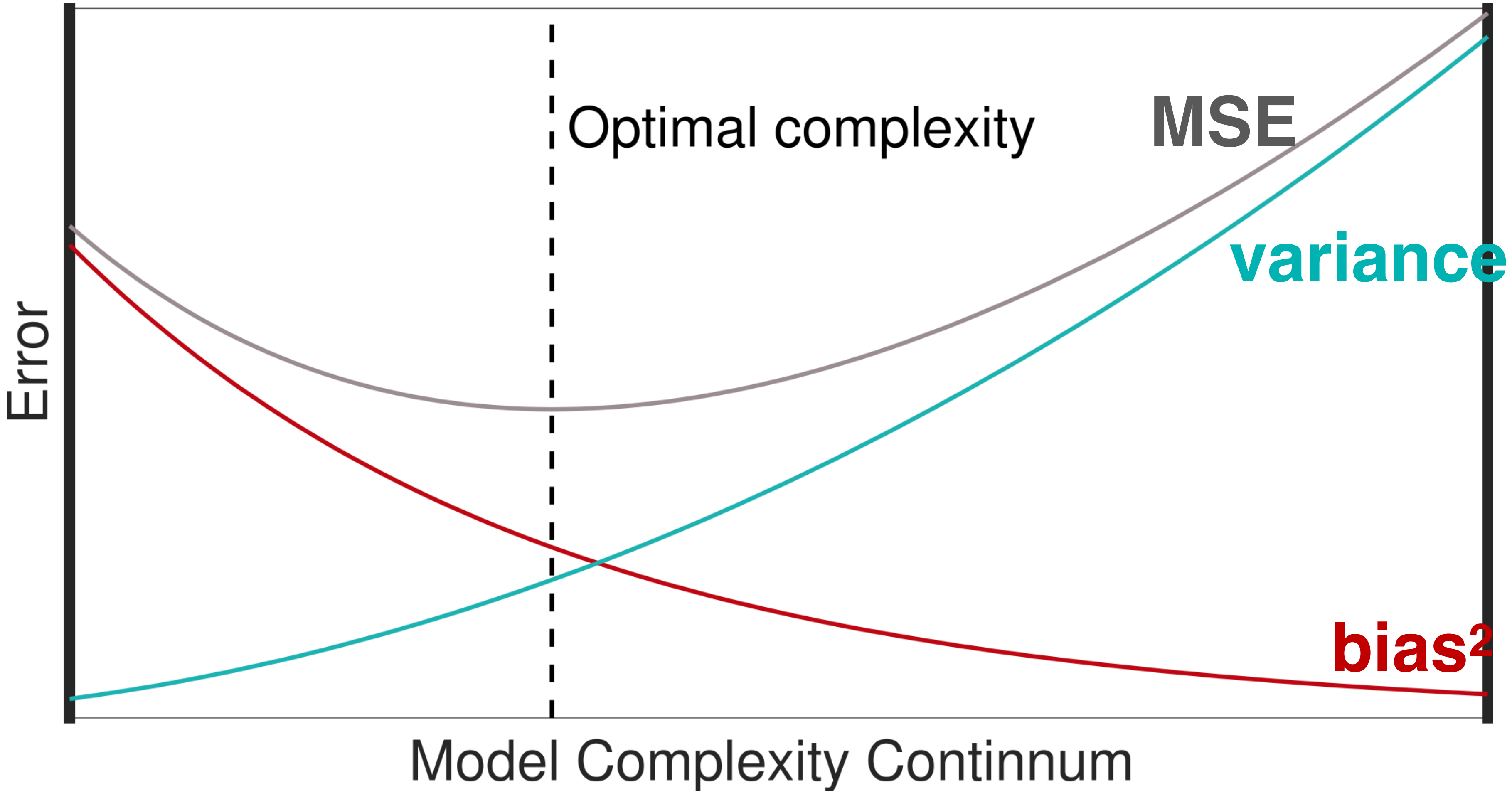
intrinsic noise in data set



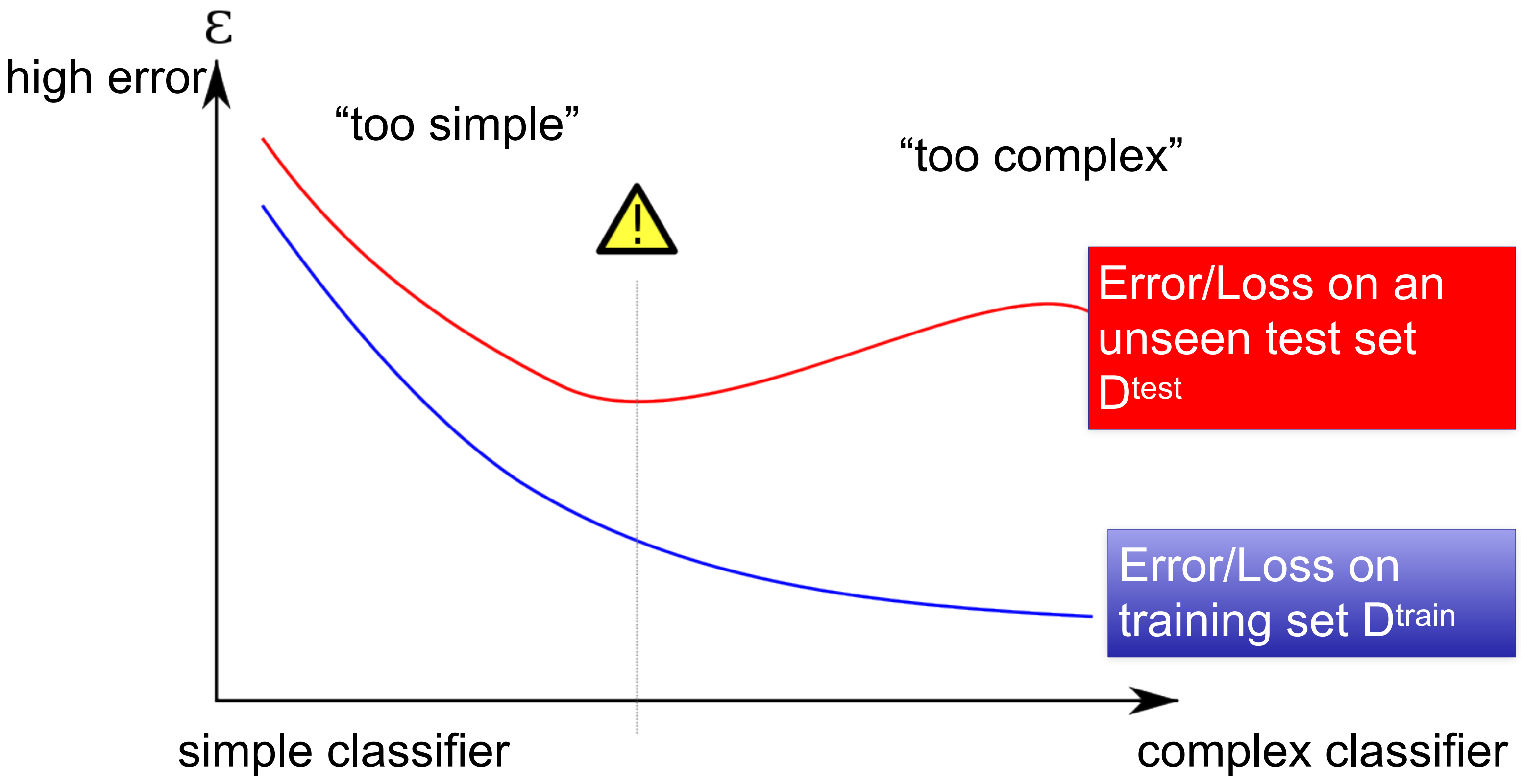


(a) Function and data
f
(b) Order 1
f
bias
g_i
ḡ
(c) Order 3
(d) Order 5
variance

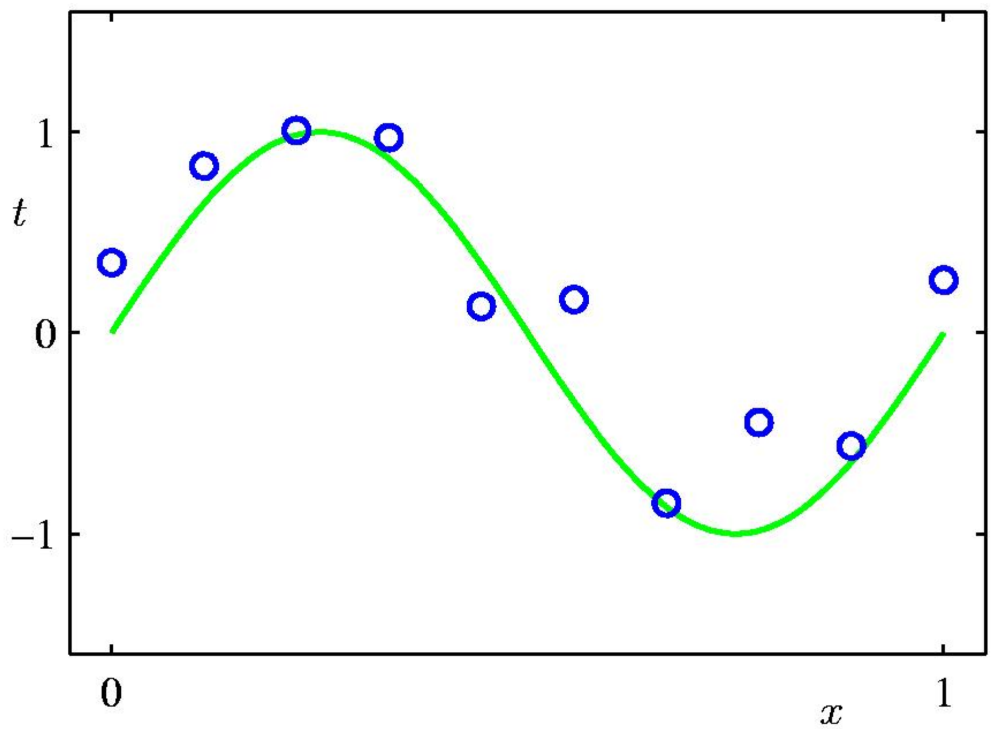
Bias-Variance Trade Off



Bias/Variance is a Way to Understand Overfitting and Underfitting

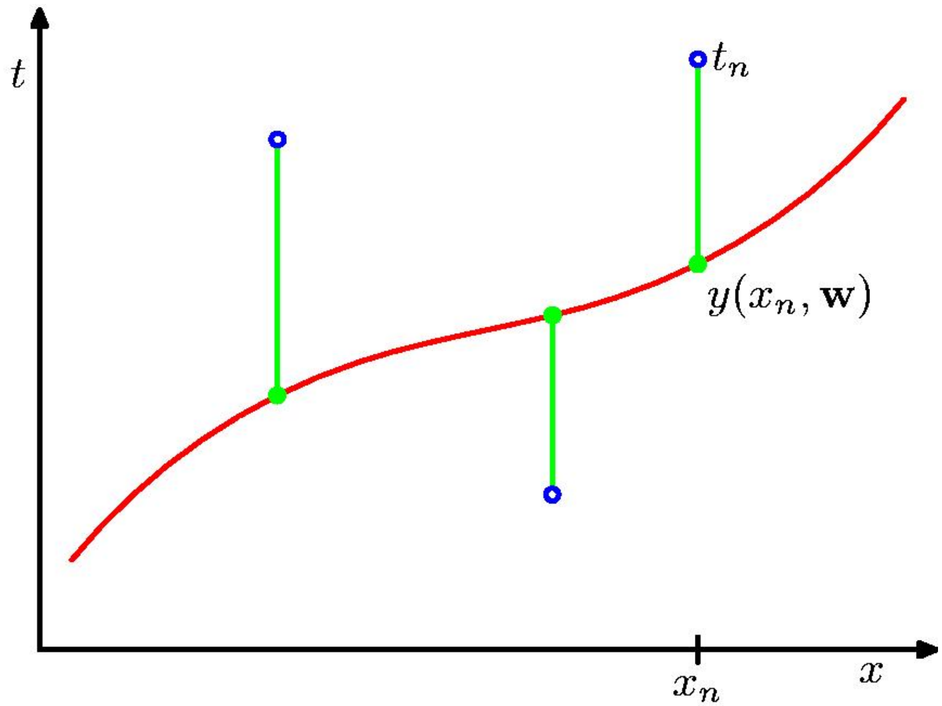


Polynomial Curve Fitting



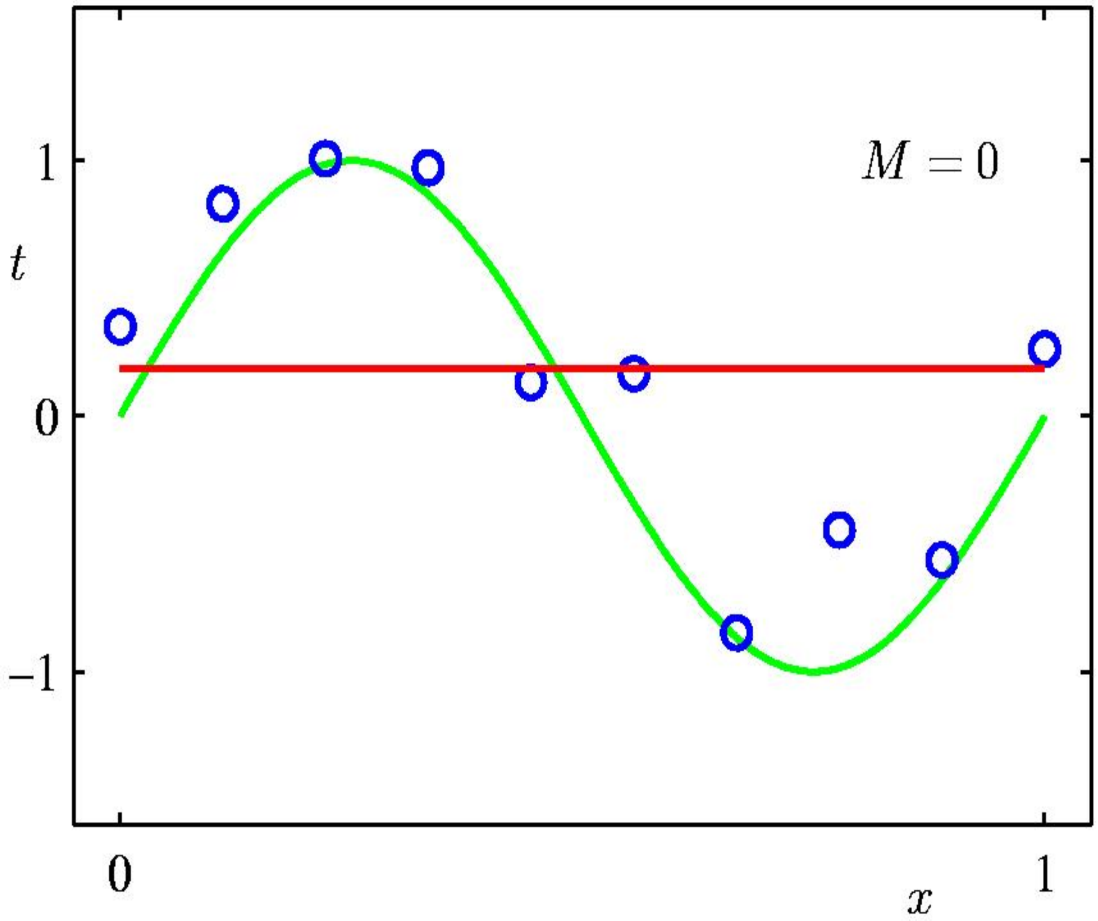
$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

Sum-of-Squares Error Function

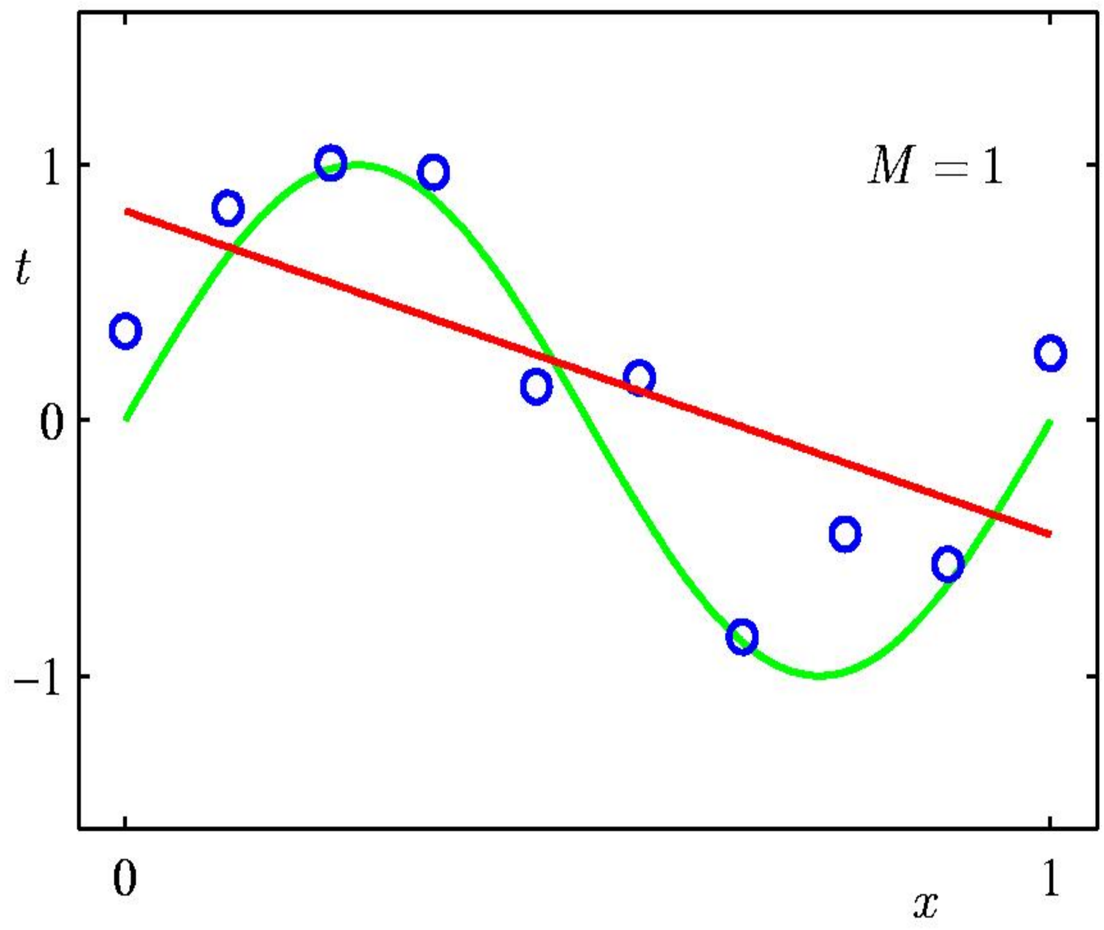


$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2$$

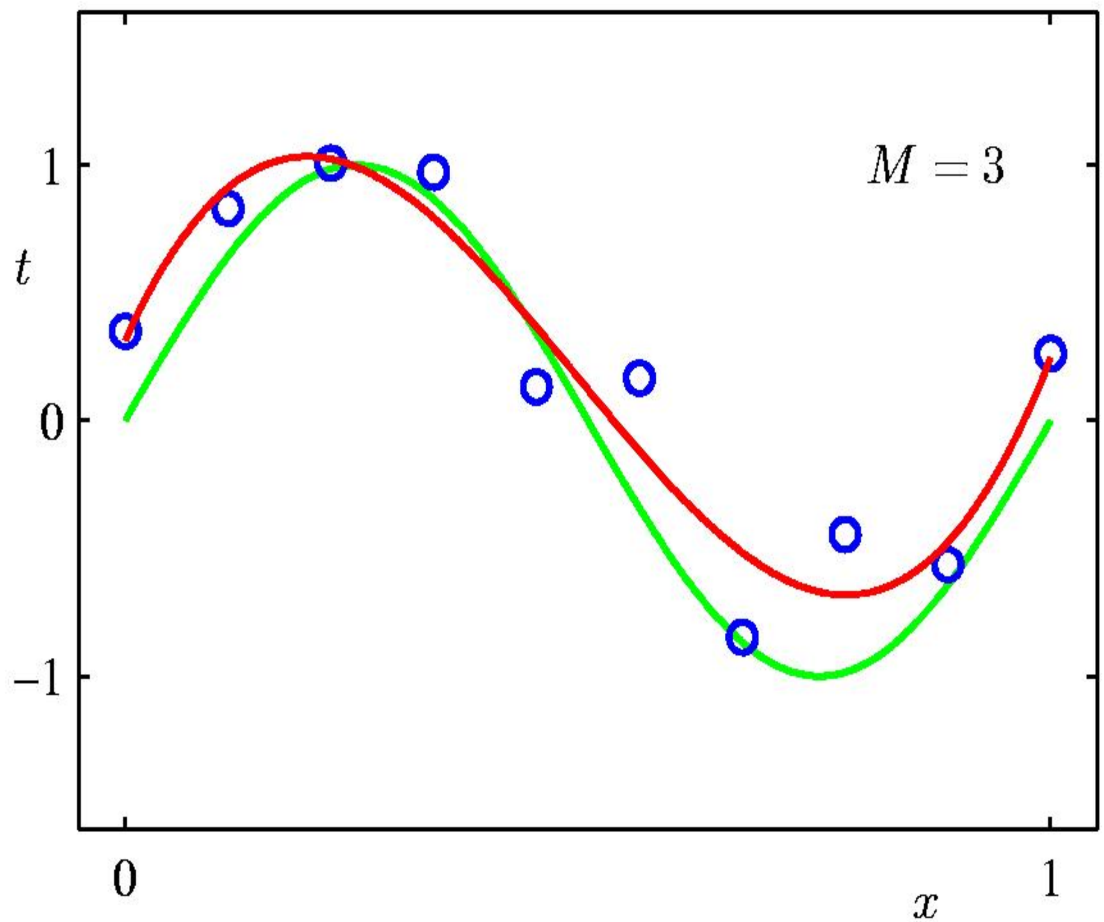
0^{th} Order Polynomial



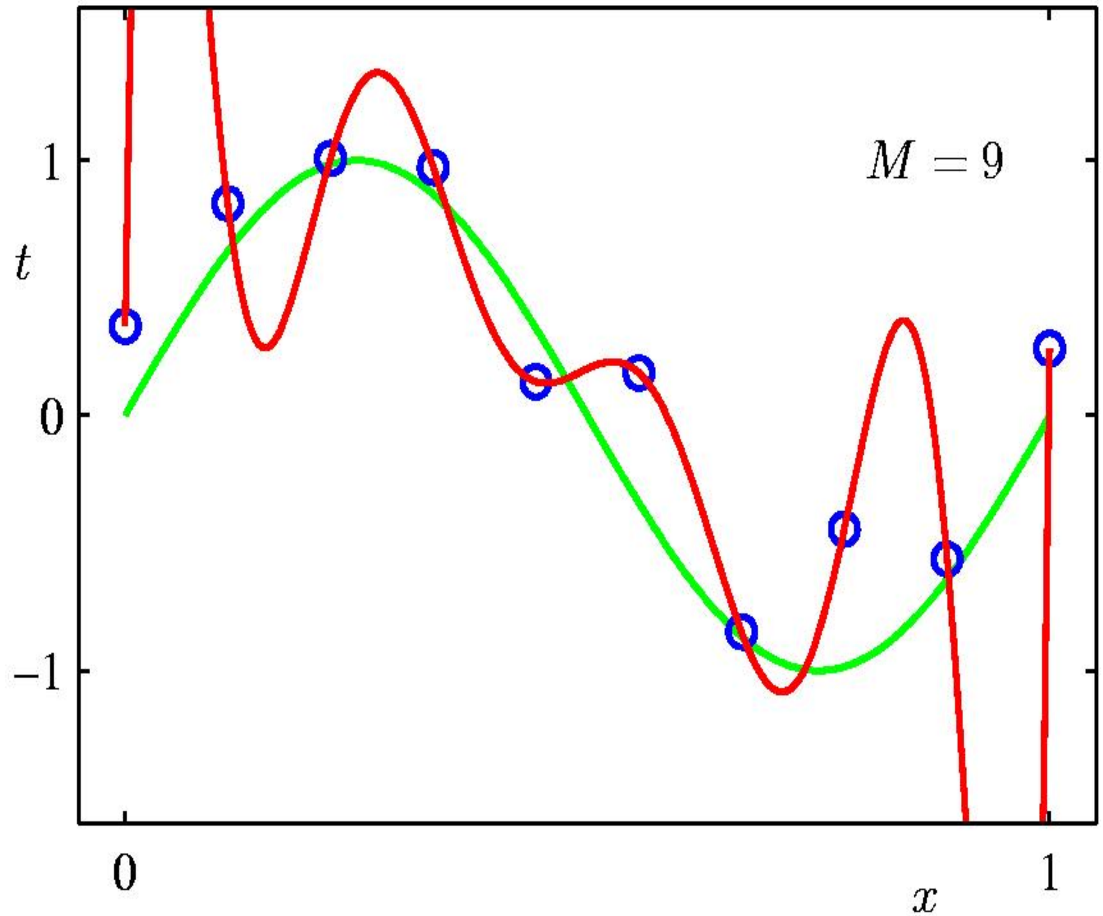
1^{st} Order Polynomial



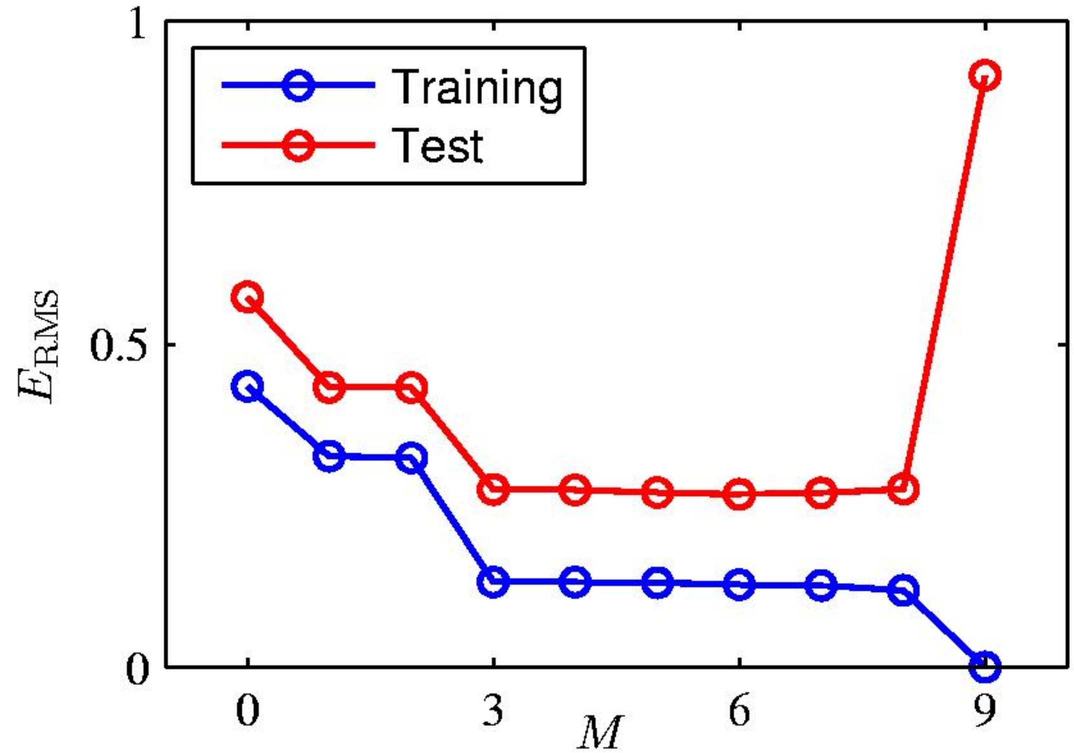
3^{rd} Order Polynomial



9^{th} Order Polynomial



Over-fitting



Root-Mean-Square (RMS) Error: $E_{\mathrm{RMS}} = \sqrt{2E(\mathbf{w}^{\star})/N}$

Polynomial Coefficients

	$M = 0$	$M = 1$	$M = 3$	$M = 9$
$w_0^\star$	0.19	0.82	0.31	0.35
$w_1^\star$		-1.27	7.99	232.37
$w_2^\star$			-25.43	-5321.83
$w_3^\star$			17.37	48568.31
$w_4^\star$				-231639.30
$w_5^\star$				640042.26
$w_6^\star$				-1061800.52
$w_7^\star$				1042400.18
$w_8^\star$				-557682.99
$w_9^\star$				125201.43