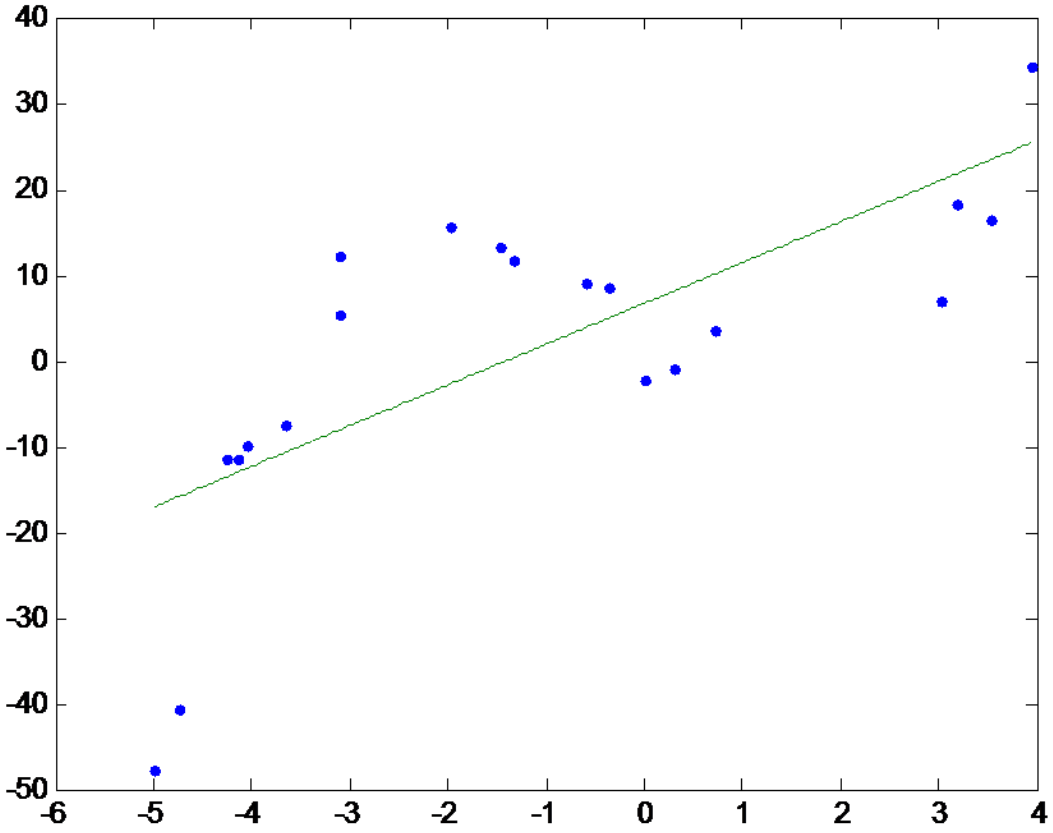
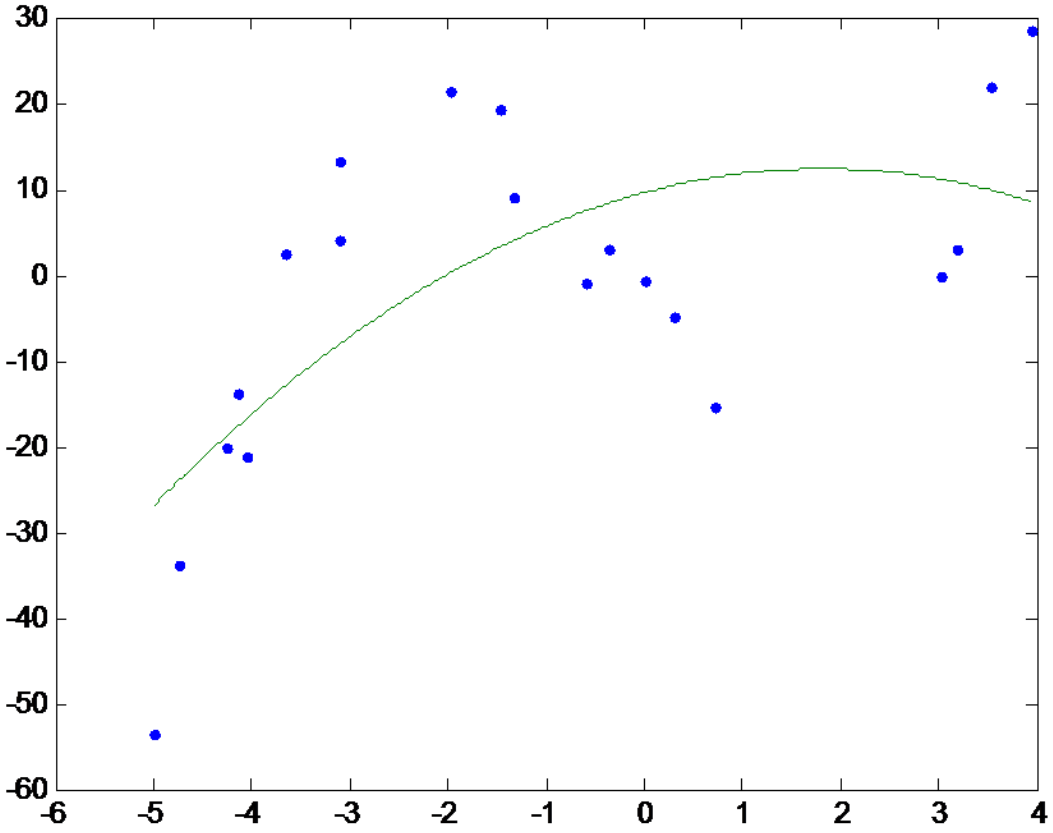
Regression

- Linear least squares
  - increasing the dimensionality
  - fitting polynomials to data
  - over fitting
  - regularization
- Support vector regression

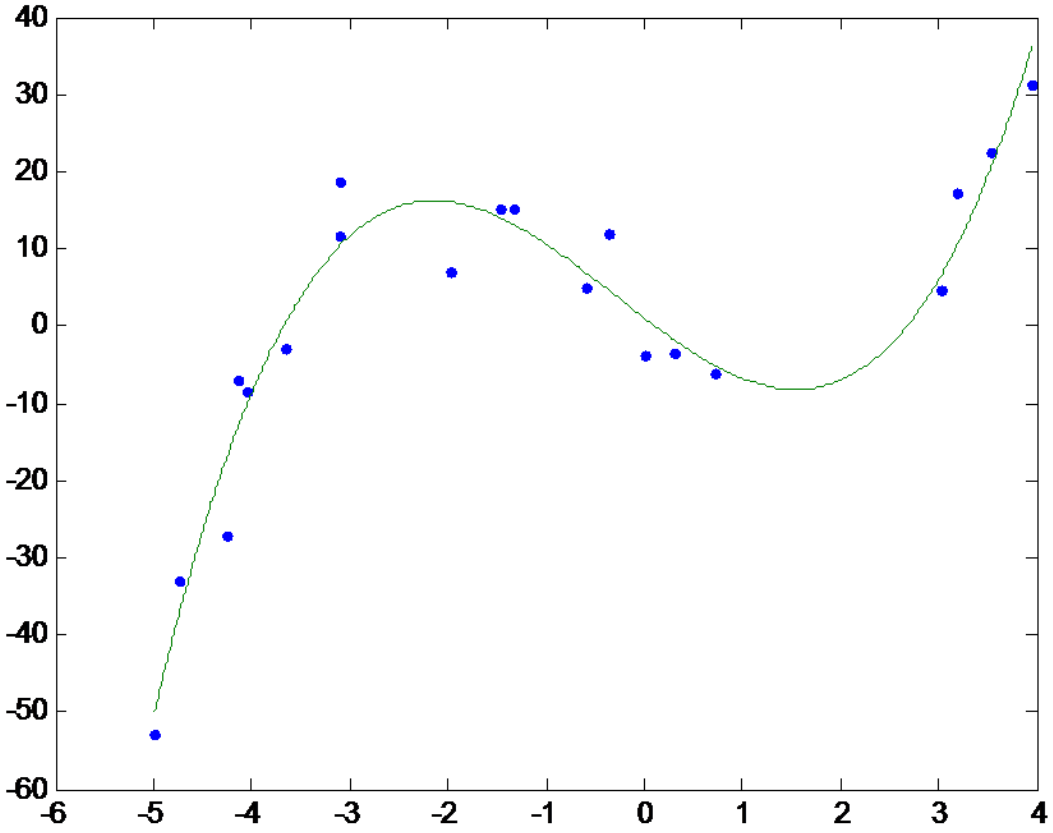
Fitting a degree-1 polynomial



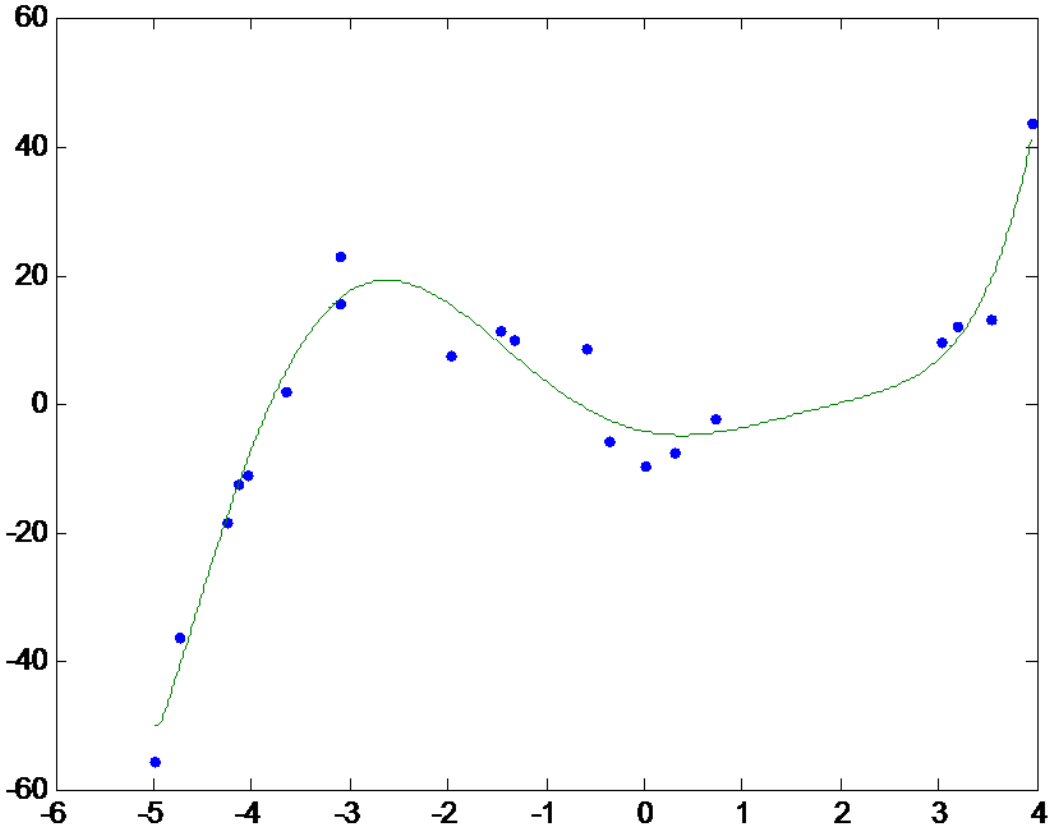
Fitting a degree-2 polynomial



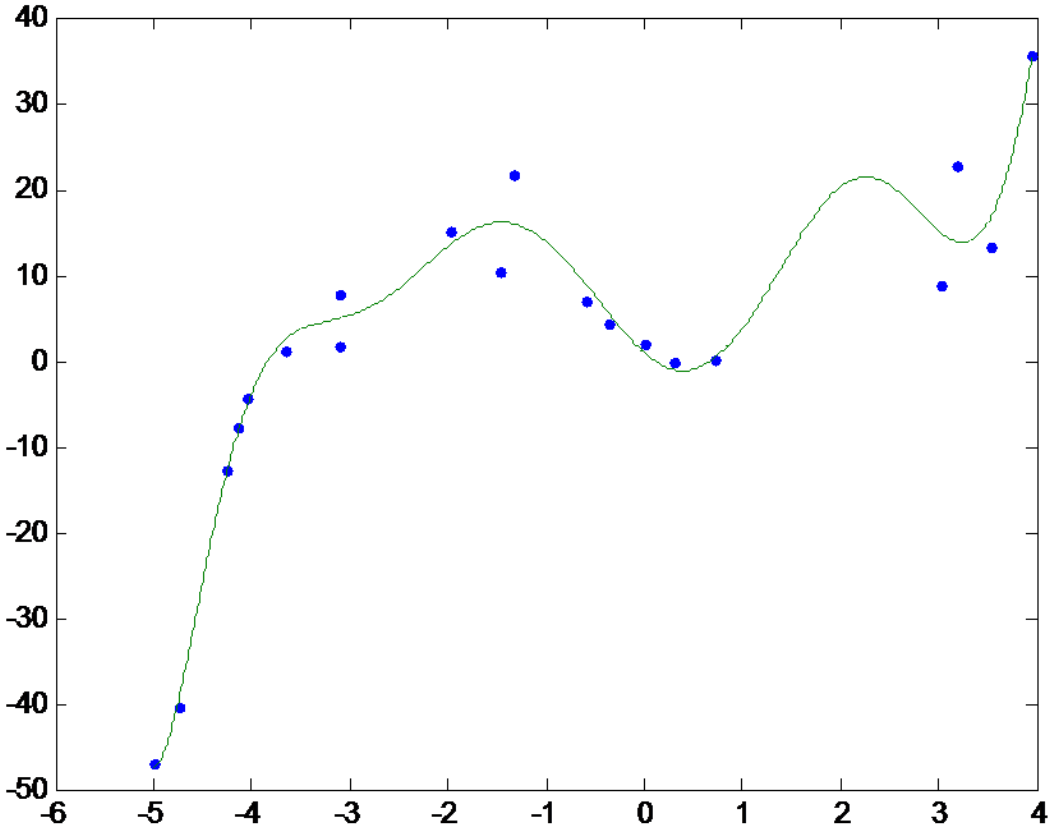
Fitting a degree-3 polynomial



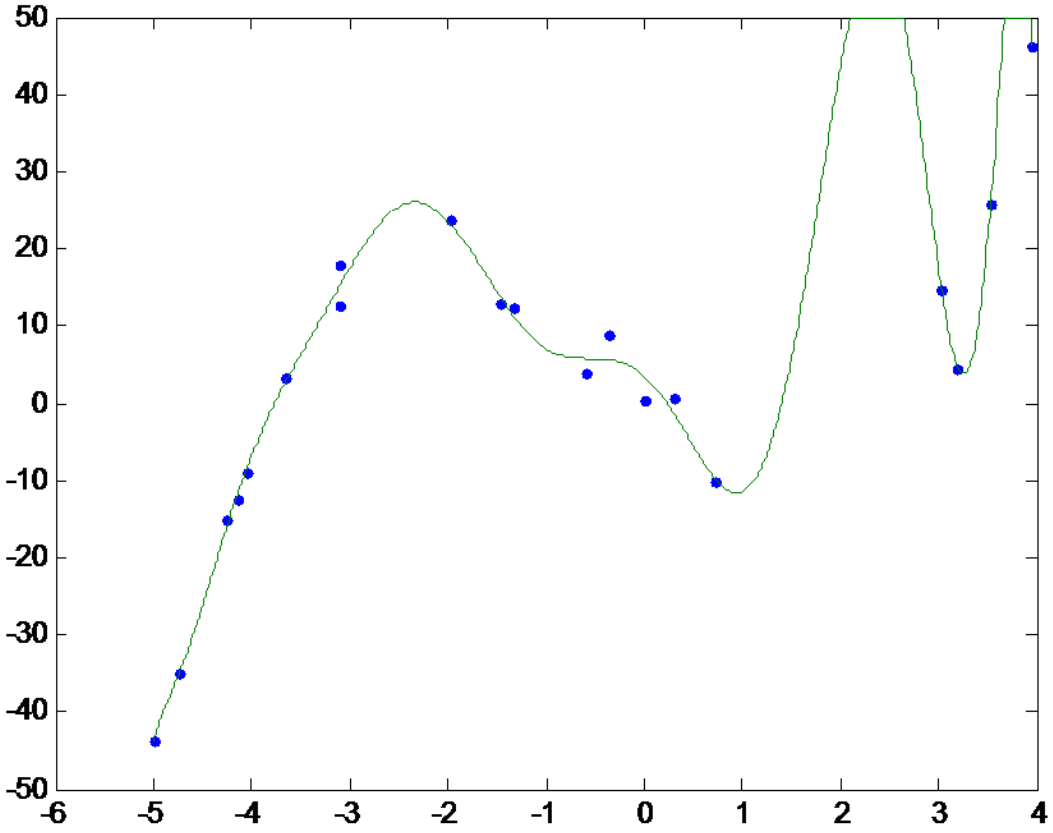
Fitting a degree-6 polynomial



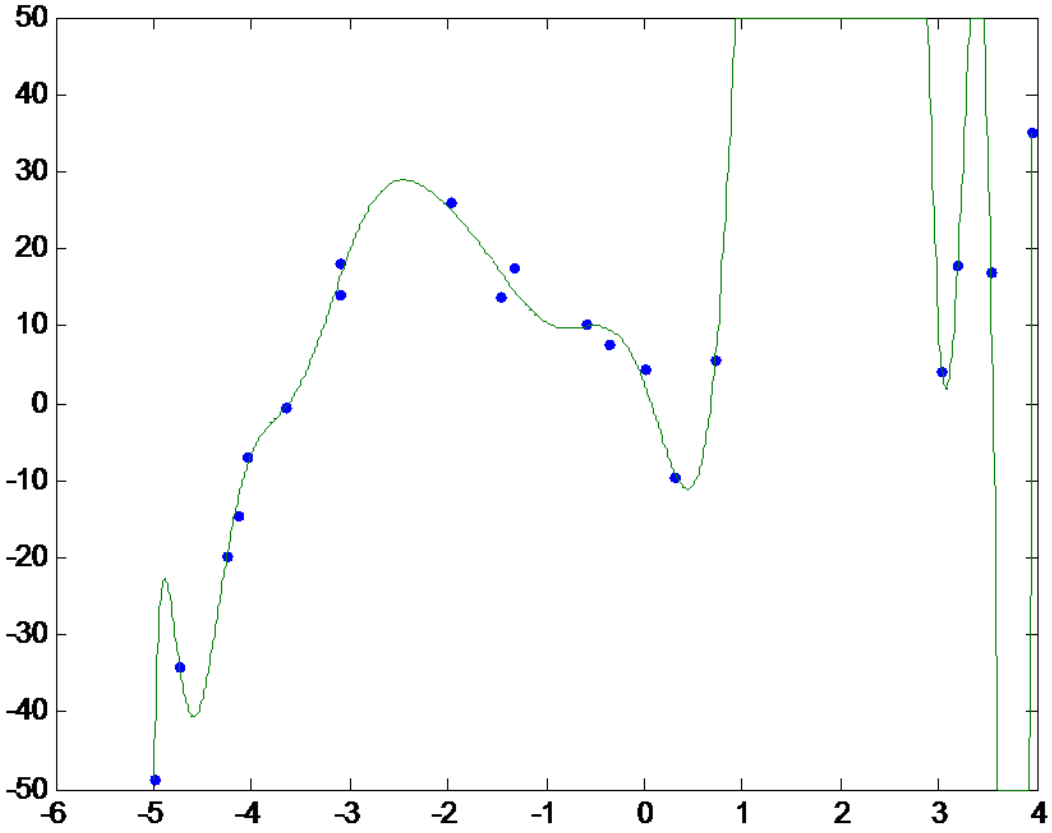
Fitting a degree-9 polynomial



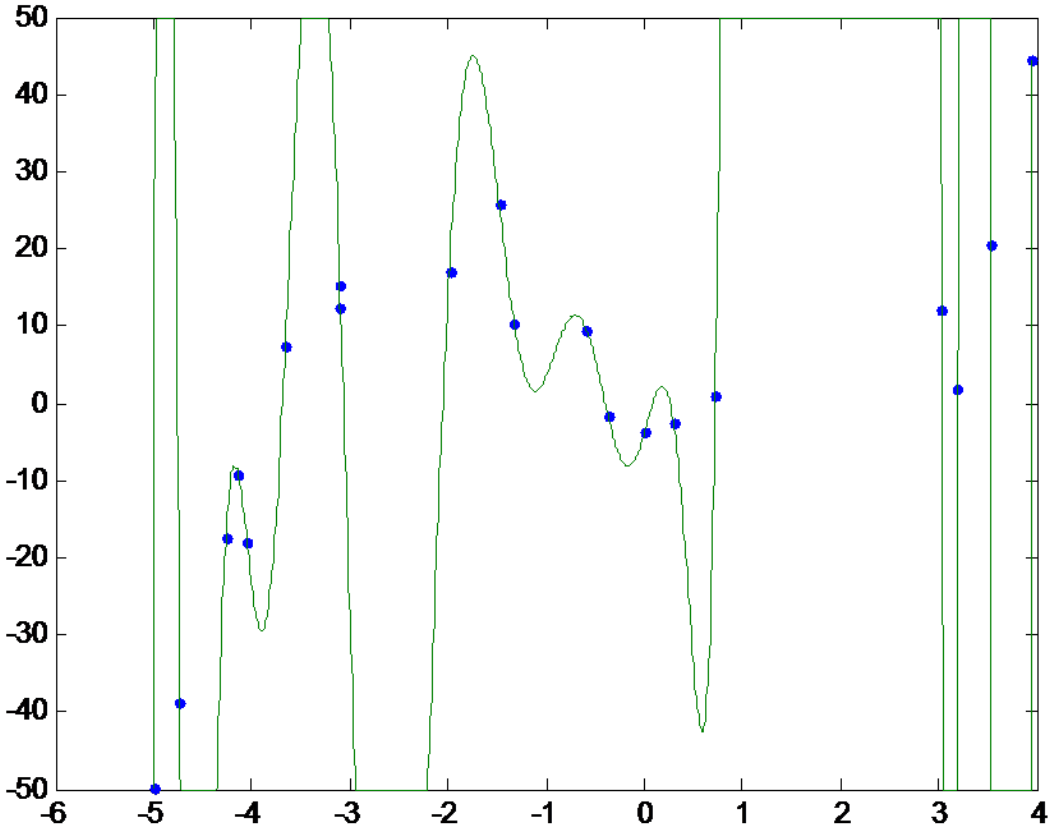
Fitting a degree-12 polynomial



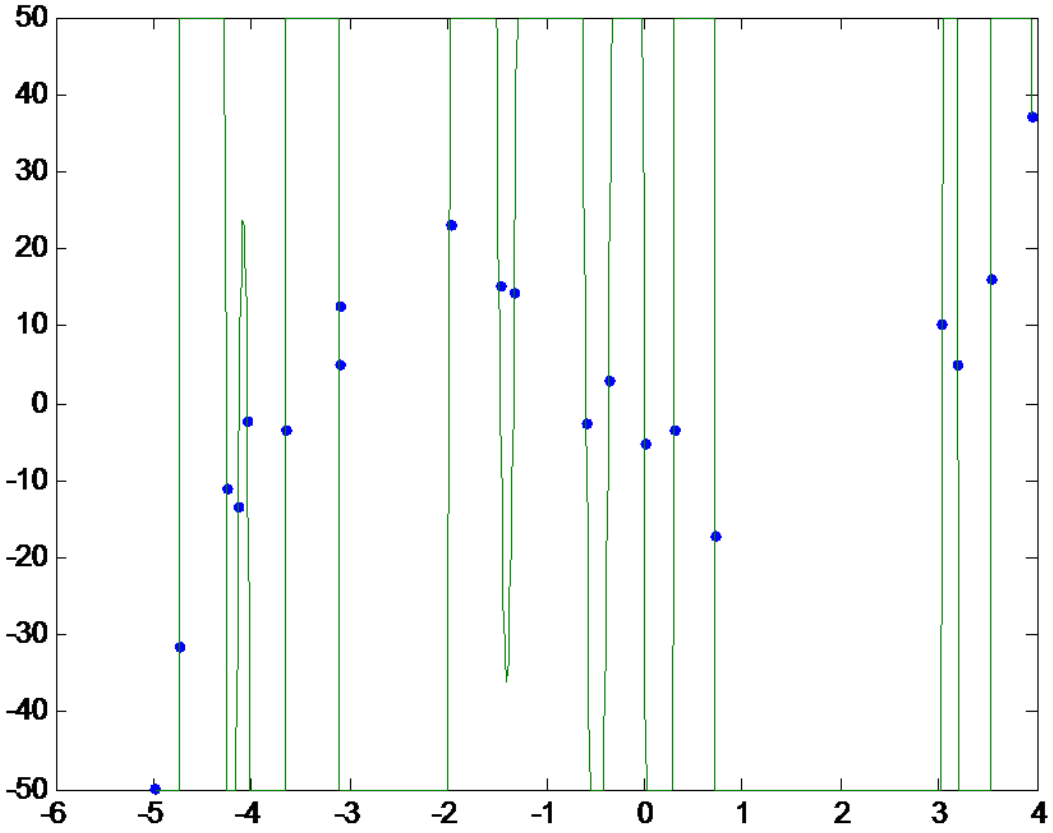
Fitting a degree-15 polynomial



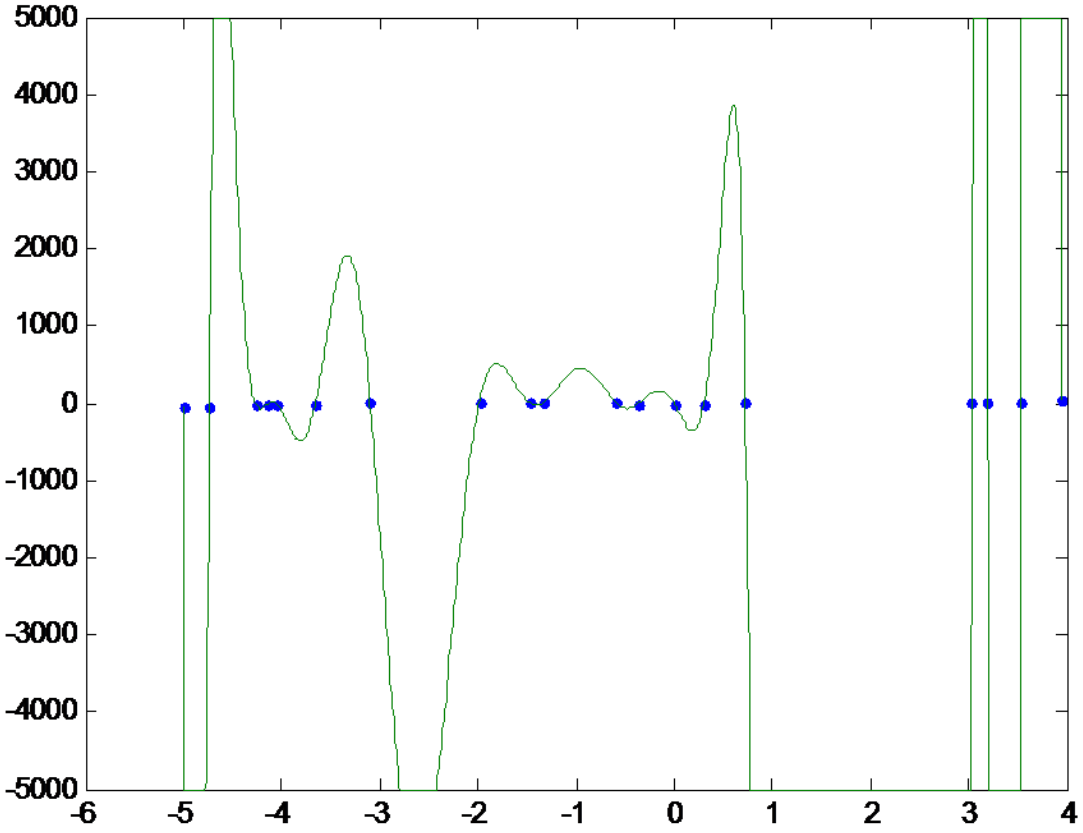
Fitting a degree-18 polynomial



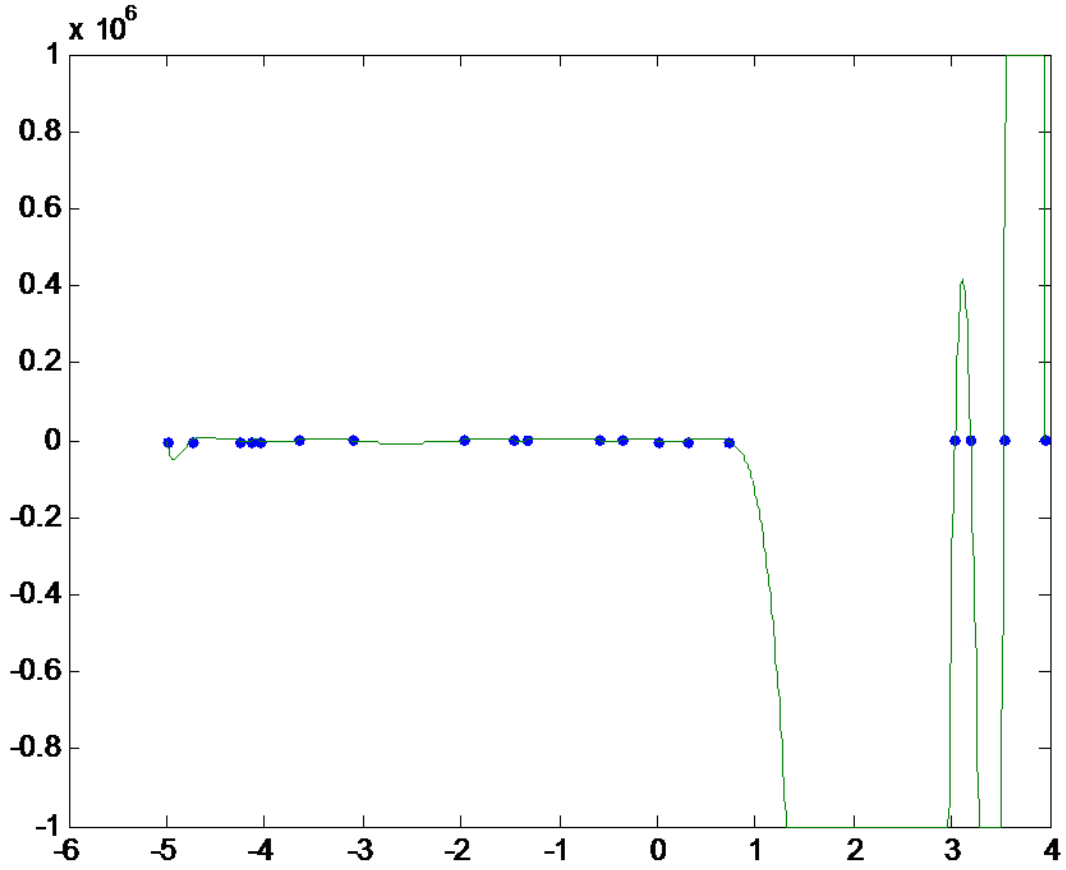
Fitting a degree-19 polynomial



Fitting a degree-19 polynomial



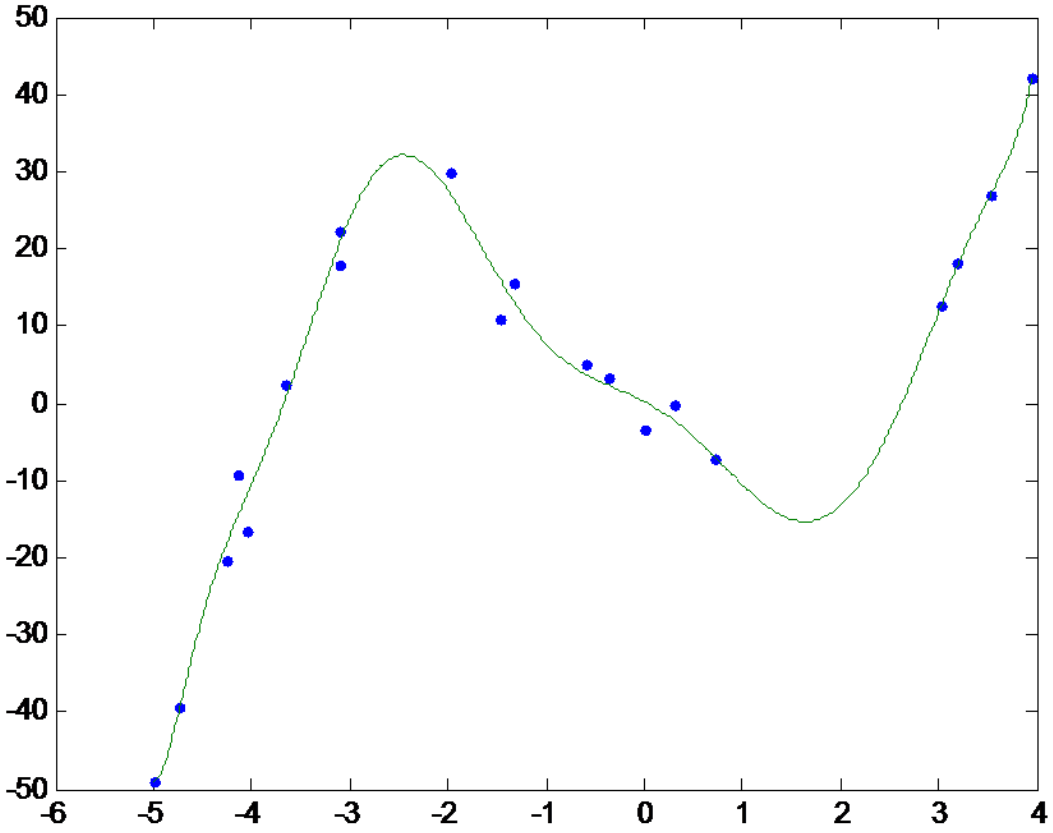
Fitting a degree-19 polynomial



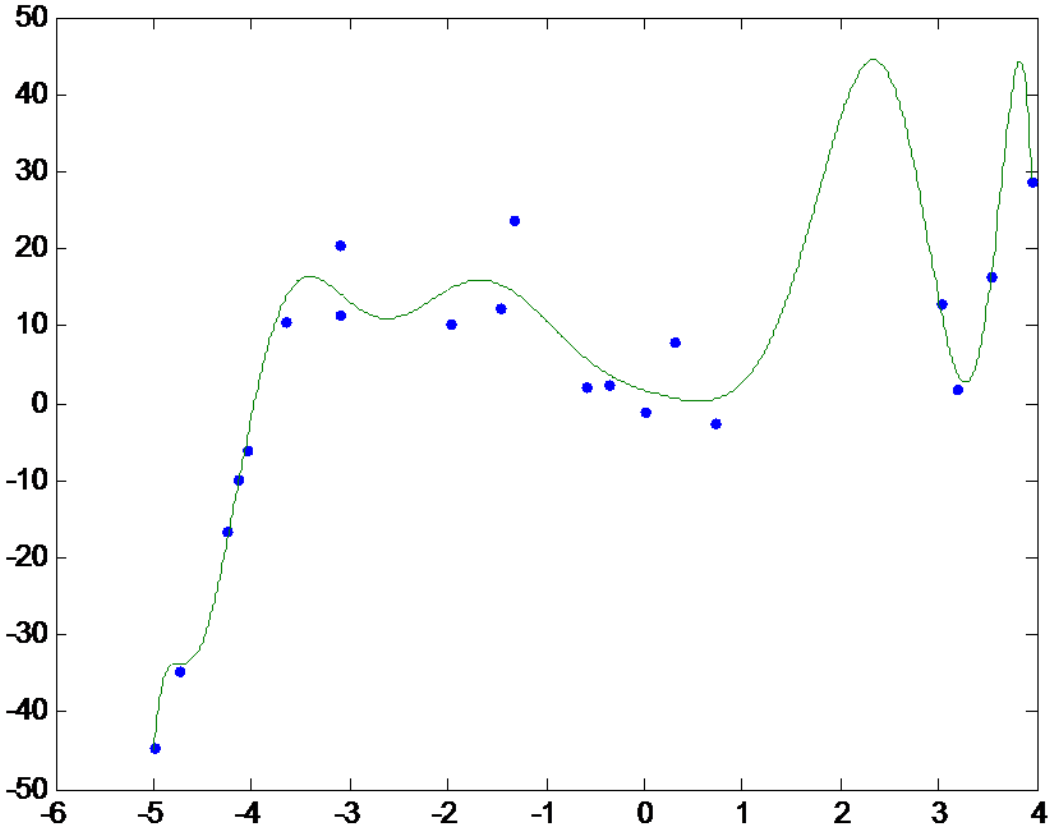
Over Fitting

- What we have just seen:
  - Over fitting leads to overly-complex models and poor generalization.
  - Over fitting fits the noise not the underlying trend.
- What we will see next:
  - Over fitted functions can be very sensitive to slight changes in the training data.
  - Regularization can eliminate over fitting.

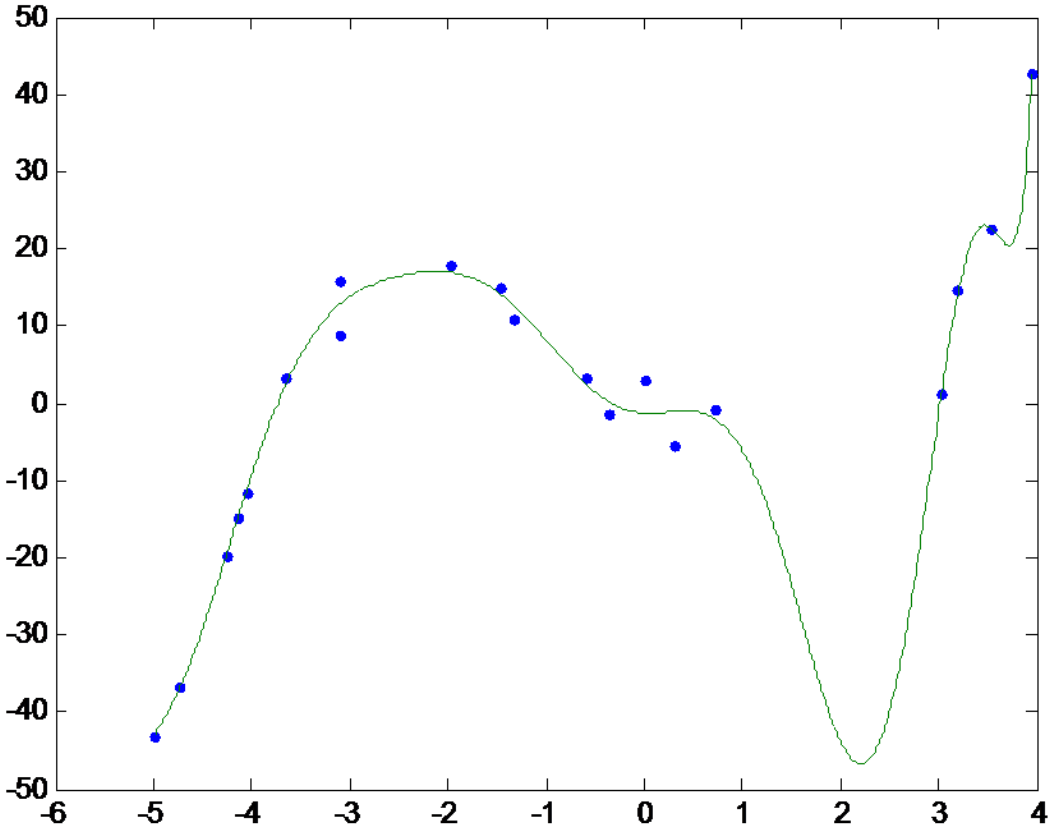
Fitting a degree-12 polynomial: data sample 1



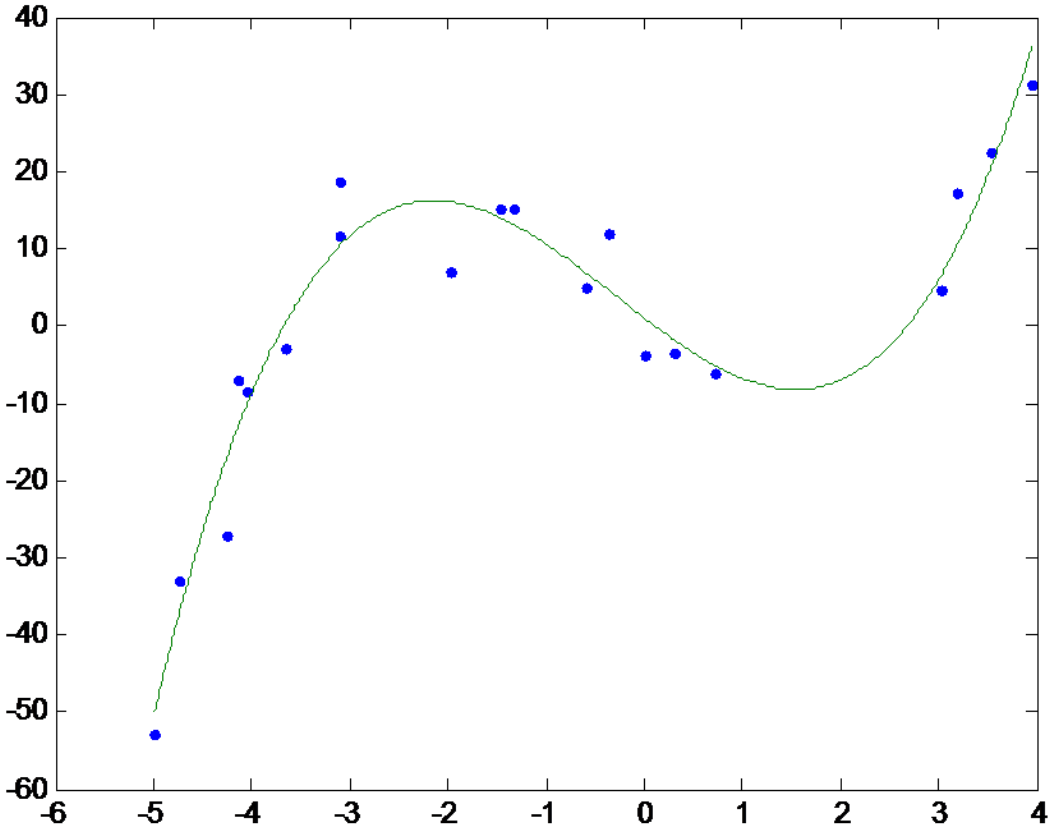
Fitting a degree-12 polynomial: data sample 2



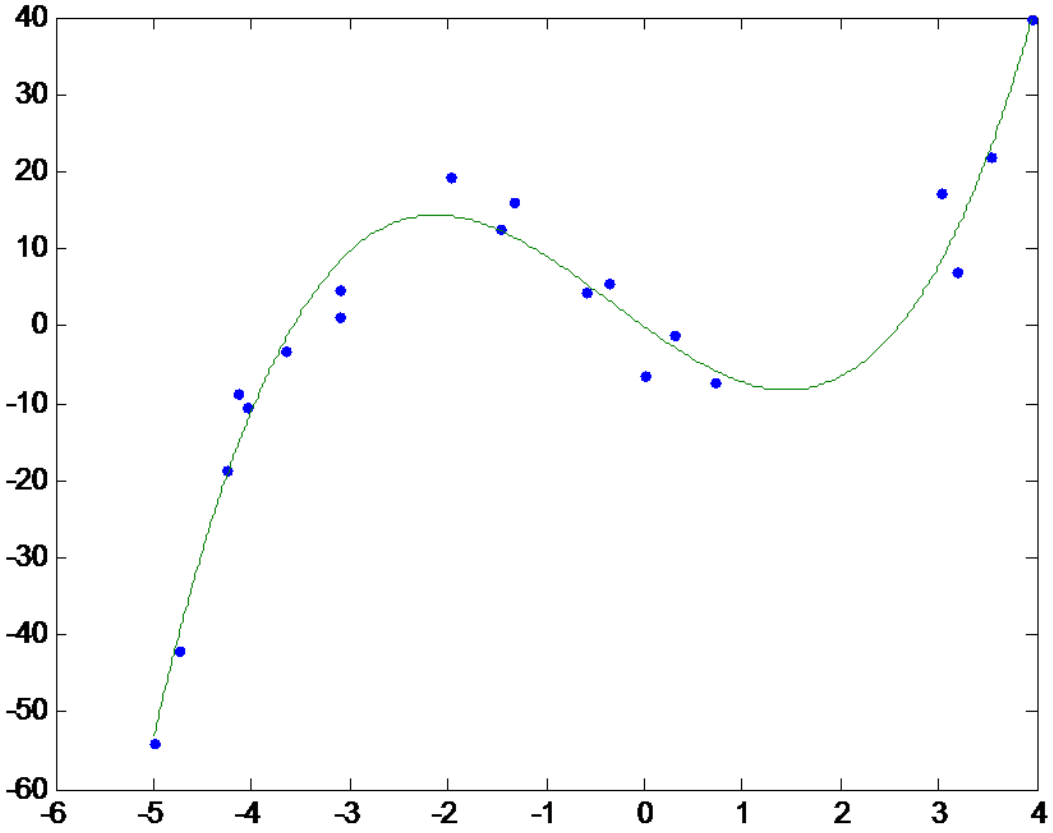
Fitting a degree-12 polynomial: data sample 3



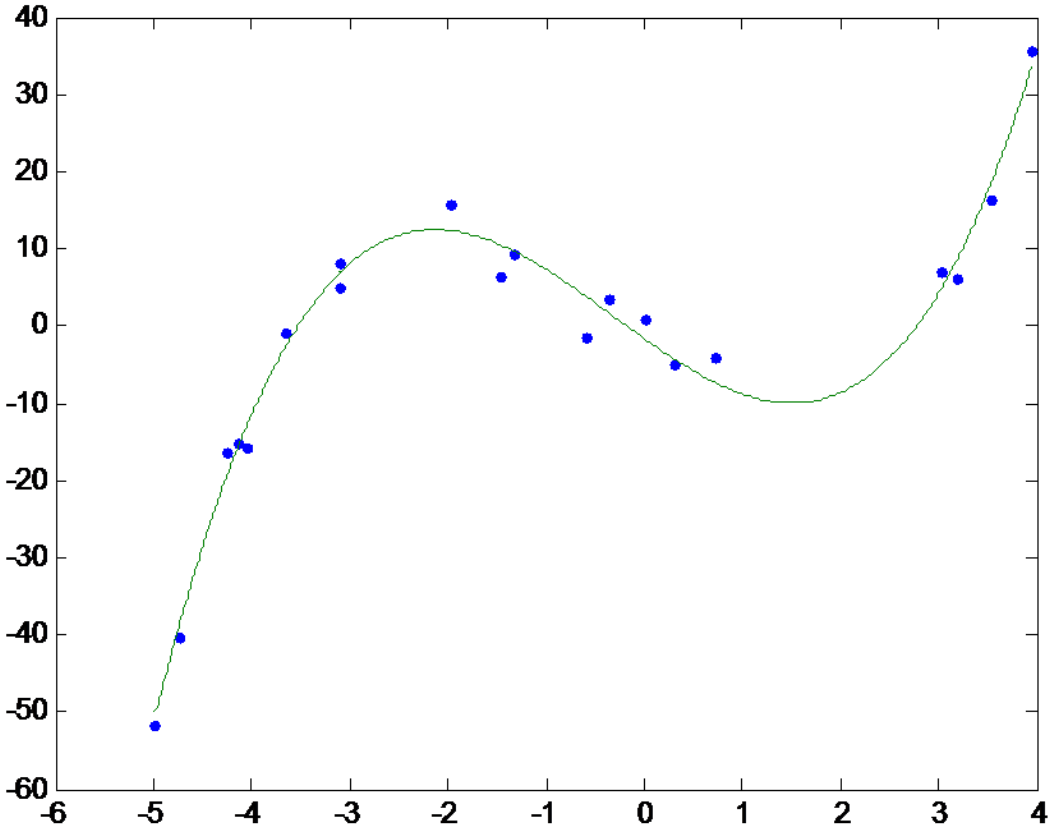
Fitting a degree-3 polynomial: data sample 1



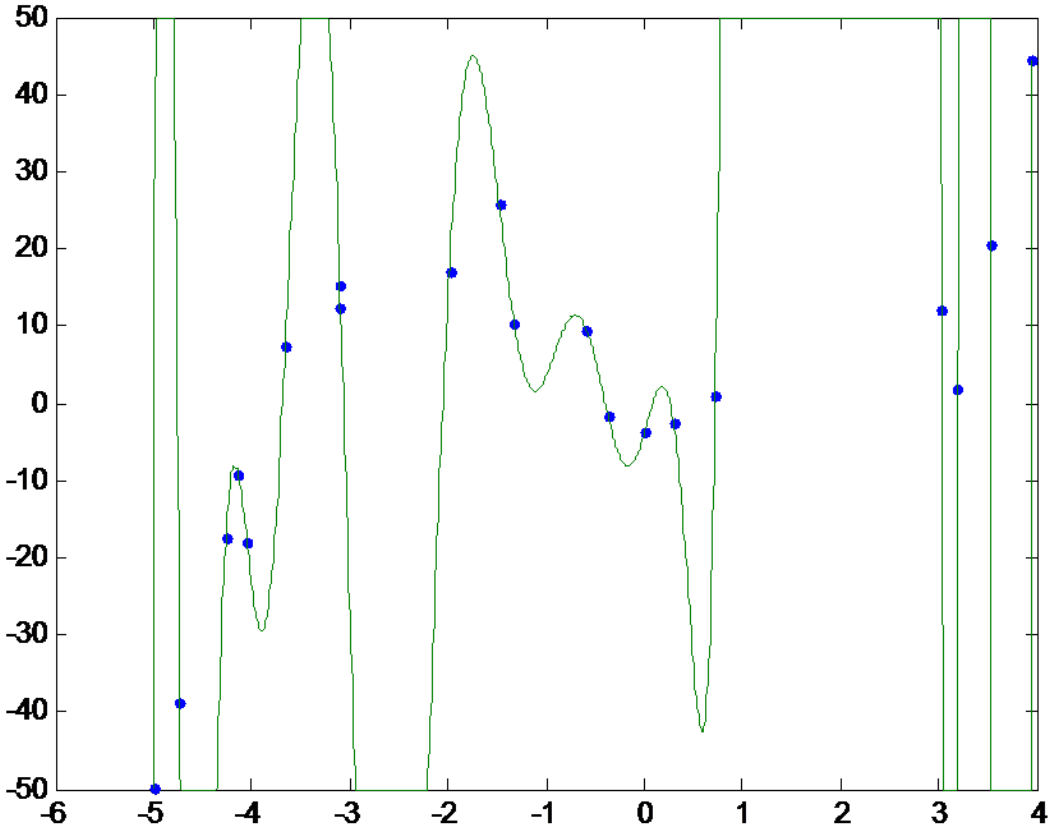
Fitting a degree-3 polynomial: data sample 2



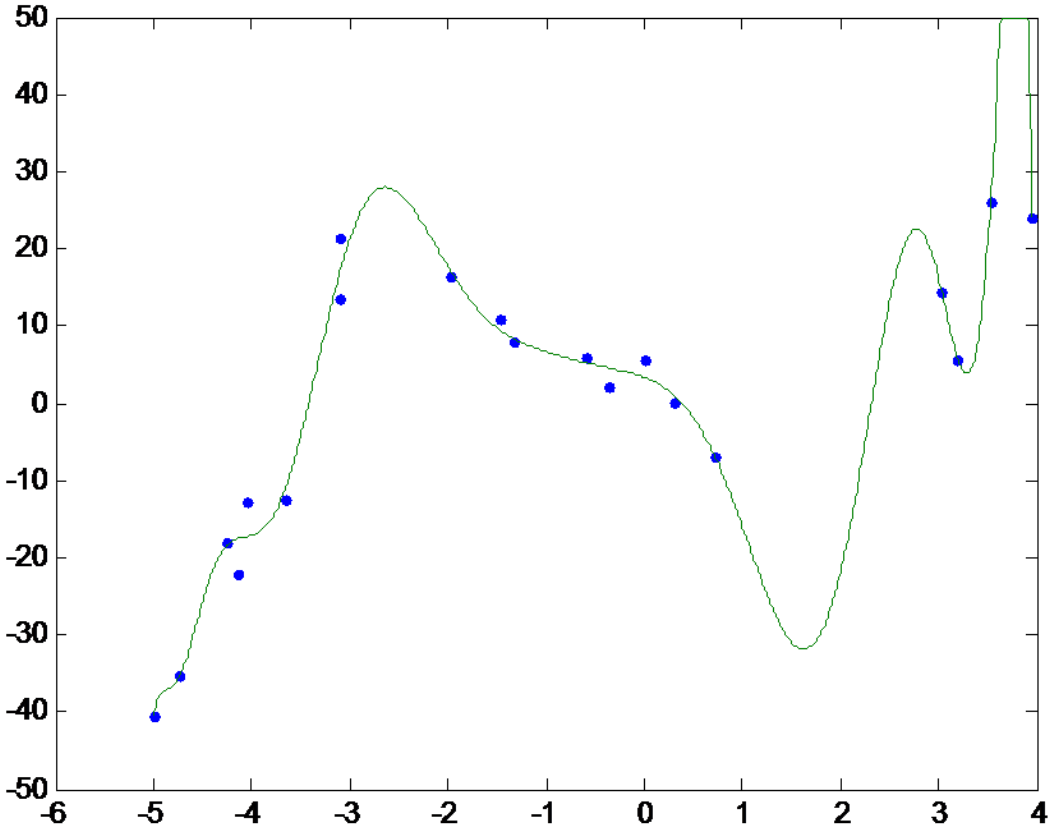
Fitting a degree-3 polynomial: data sample 3



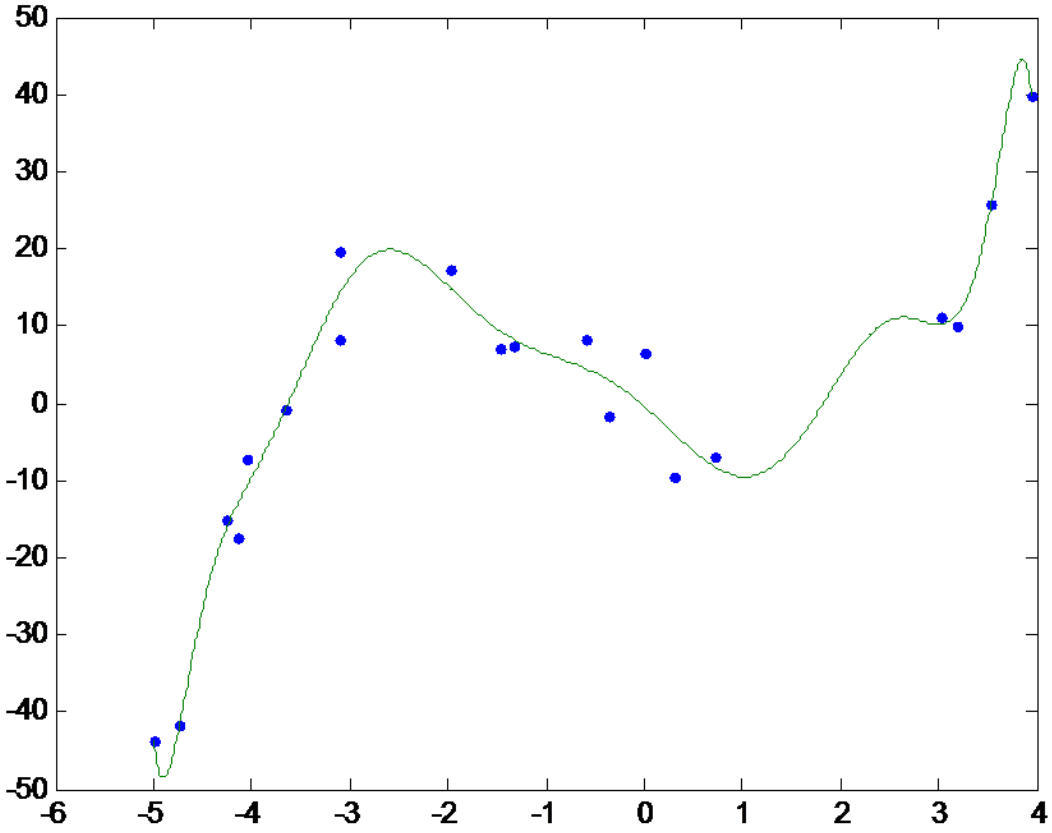
Regularization: degree=18, lambda=0



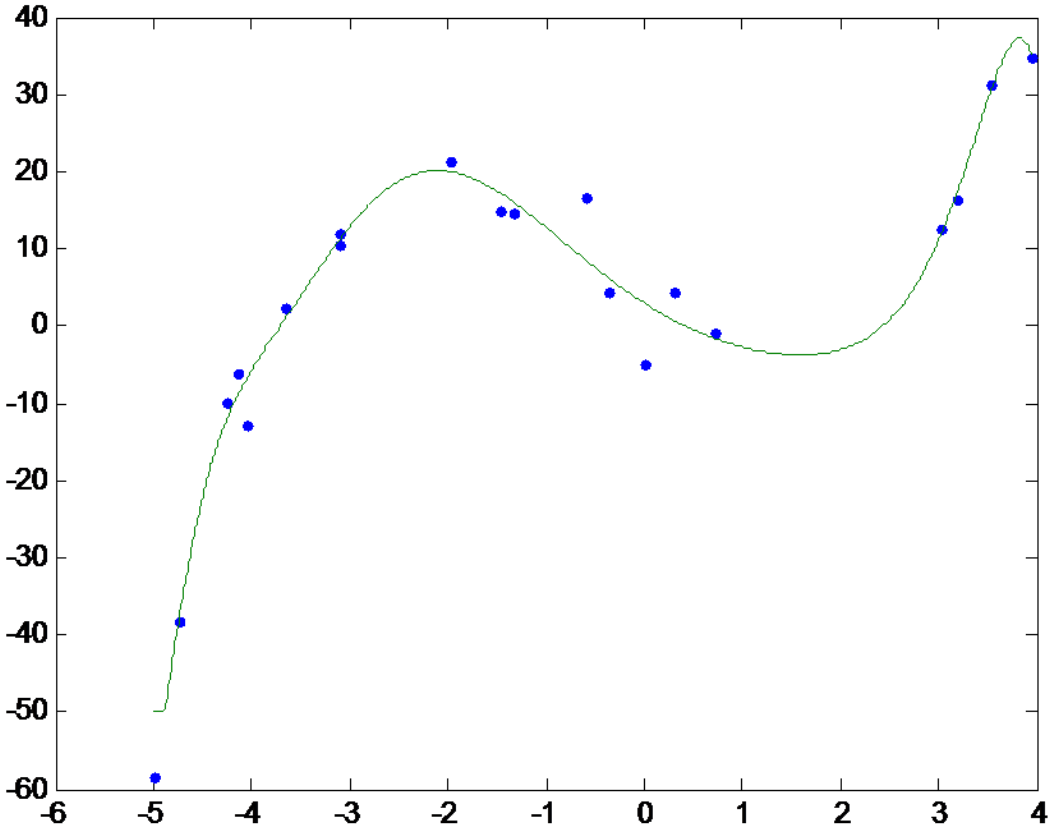
Regularization: degree=18, lambda=0.000000001



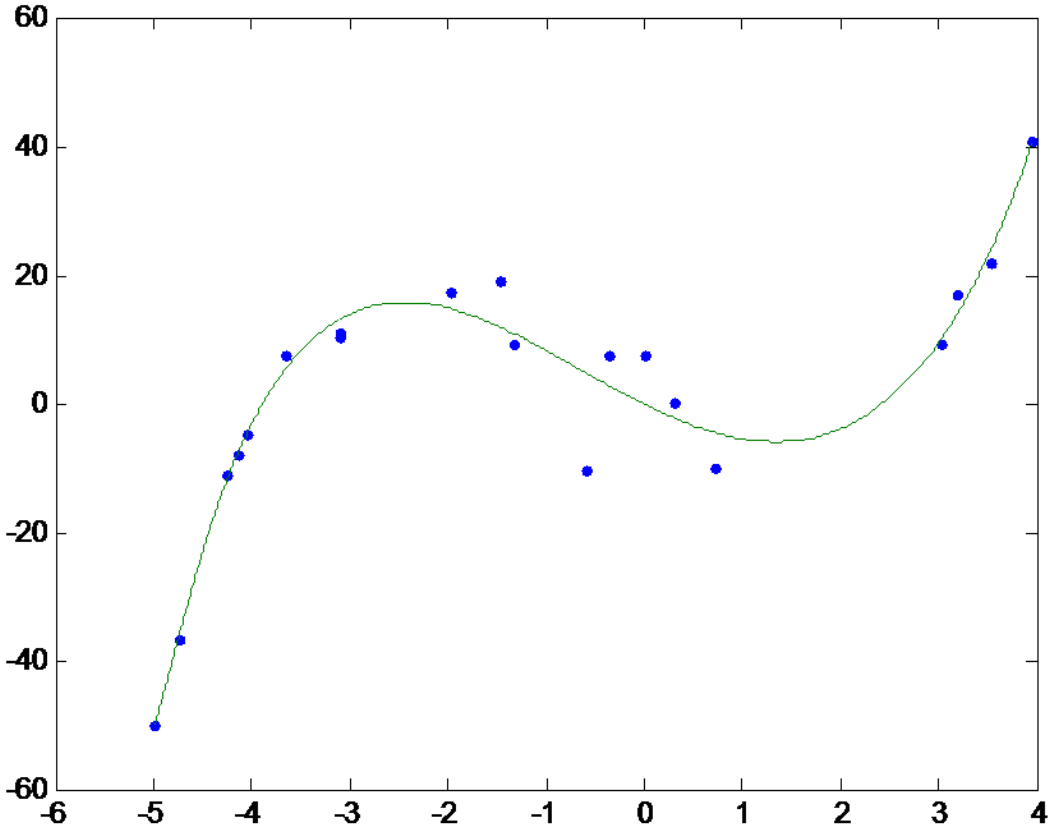
Regularization: degree=18, lambda=0.000001



Regularization: degree=18, lambda=0.0001



Regularization: degree=18, lambda=0.01



Cost Functions

- Minimizing squared error means that points far from the regression line have the most influence.
- The solution is therefore very sensitive to outliers and anomalies.
- Other cost functions can reduce this sensitivity.
- Sometimes it is convenient to formulate cost functions in terms of slack variables.

Support Vector Regression

- Regularization
- Primal problem
- Dual problem
- KKT conditions
- Support vectors
- *v*-SV regression

SV Regression: ε-Insensitive Loss

Goal: generalize SV pattern recognition to regression, preserving the following properties:

- formulate the algorithm for the linear case, and then use kernel trick
- sparse representation of the solution in terms of SVs

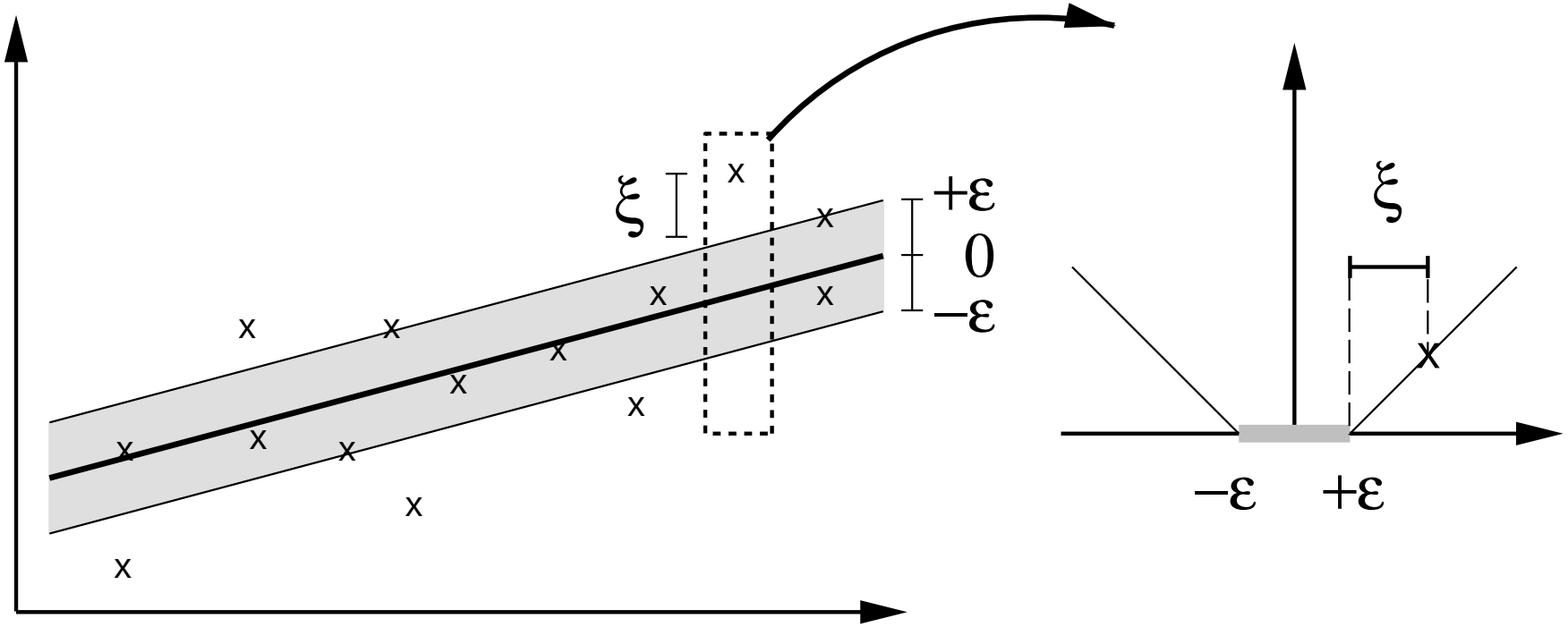
ε-Insensitive Loss:

$$|y - f(\mathbf{x})|_\varepsilon := \max\{0, |y - f(\mathbf{x})| - \varepsilon\}$$

Estimate a linear regression $f(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle + b$ by minimizing

$$\frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{m}\sum_{i=1}^{m} |y_i - f(\mathbf{x}_i)|_\varepsilon.$$

ε-SV Regression Estimation



Formulation as an Optimization Problem

Estimate a linear regression

$$f(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle + b$$

with precision ε by minimizing

$$\begin{aligned} &\text{minimize} && \tau(\mathbf{w}, \boldsymbol{\xi}, \boldsymbol{\xi}^*) = \frac{1}{2}\|\mathbf{w}\|^2 + C \sum_{i=1}^{m} (\xi_i + \xi_i^*) \\ &\text{subject to} && (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - y_i \leq \varepsilon + \xi_i \\ & && y_i - (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \leq \varepsilon + \xi_i^* \\ & && \xi_i, \xi_i^* \geq 0 \end{aligned}$$

for all $i = 1, \ldots, m$.

Dual Problem, In Terms of Kernels

For $C > 0, \varepsilon \geq 0$ chosen a priori,

$$\text{maximize} \quad W(\boldsymbol{\alpha}, \boldsymbol{\alpha}^*) = -\varepsilon \sum_{i=1}^{m} (\alpha_i^* + \alpha_i) + \sum_{i=1}^{m} (\alpha_i^* - \alpha_i) y_i$$
$$- \frac{1}{2} \sum_{i,j=1}^{m} (\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j) k(\mathbf{x}_i, \mathbf{x}_j)$$

$$\text{subject to} \quad 0 \leq \alpha_i, \alpha_i^* \leq C, \ i = 1, \ldots, m, \ \text{and} \ \sum_{i=1}^{m} (\alpha_i - \alpha_i^*) = 0.$$

The regression estimate takes the form

$$f(\mathbf{x}) = \sum\nolimits_{i=1}^{m} (\alpha_i^* - \alpha_i) k(\mathbf{x}_i, \mathbf{x}) + b,$$

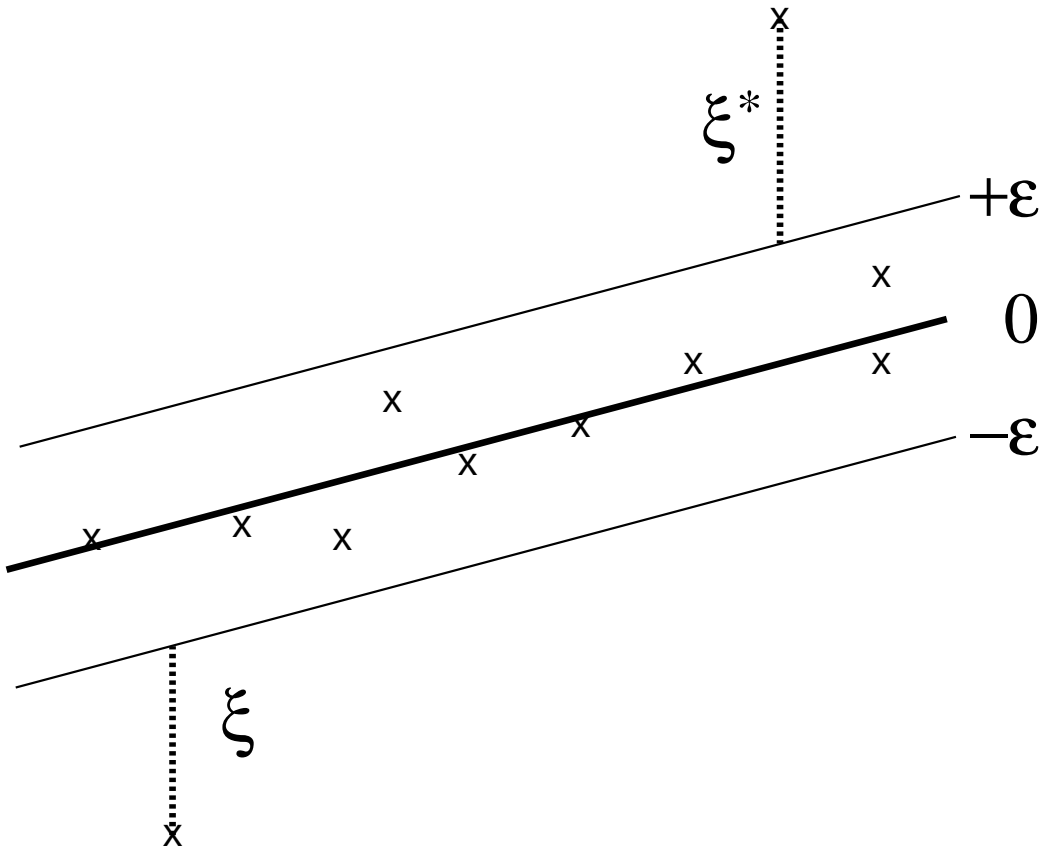
ν-SV Regression

Again, use ν to eliminate another parameter:
Estimate ε from the data s.t. the ν-property holds.

Primal problem: for $0 \leq \nu \leq 1$, minimize

$$\tau(\mathbf{w}, \varepsilon) = \frac{1}{2}\|\mathbf{w}\|^2 + C\left(\nu\varepsilon + 1/m \sum_{i=1}^{m} |y_i - f(\mathbf{x}_i)|_\varepsilon\right)$$

A Graphical Proof of the ν-Property



Cost function: $\frac{1}{2C}\|\mathbf{w}\|^2 + \nu\varepsilon + \frac{1}{m}\sum_{i=1}^{m}(\xi_i + \xi_i^*)$

The ν-Property

Proposition 3 *Assume $\varepsilon > 0$. The following statements hold:*

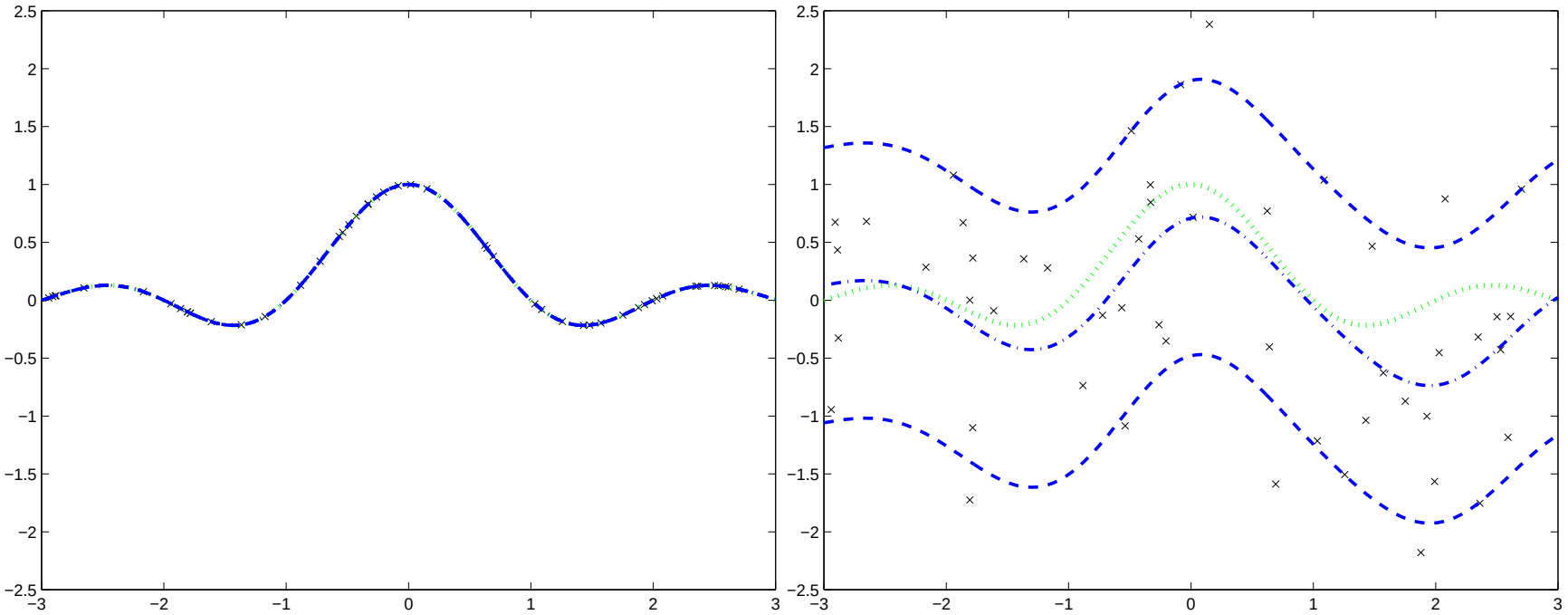
(i) *ν is an upper bound on the fraction of errors.*

(ii) *ν is a lower bound on the fraction of SVs.*

(iii) *Suppose the data were generated iid from a 'well-behaved'* distribution $\mathrm{P}(\mathbf{x}, y)$. With probability 1, asymptotically, ν equals both the fraction of SVs and the fraction of errors.*

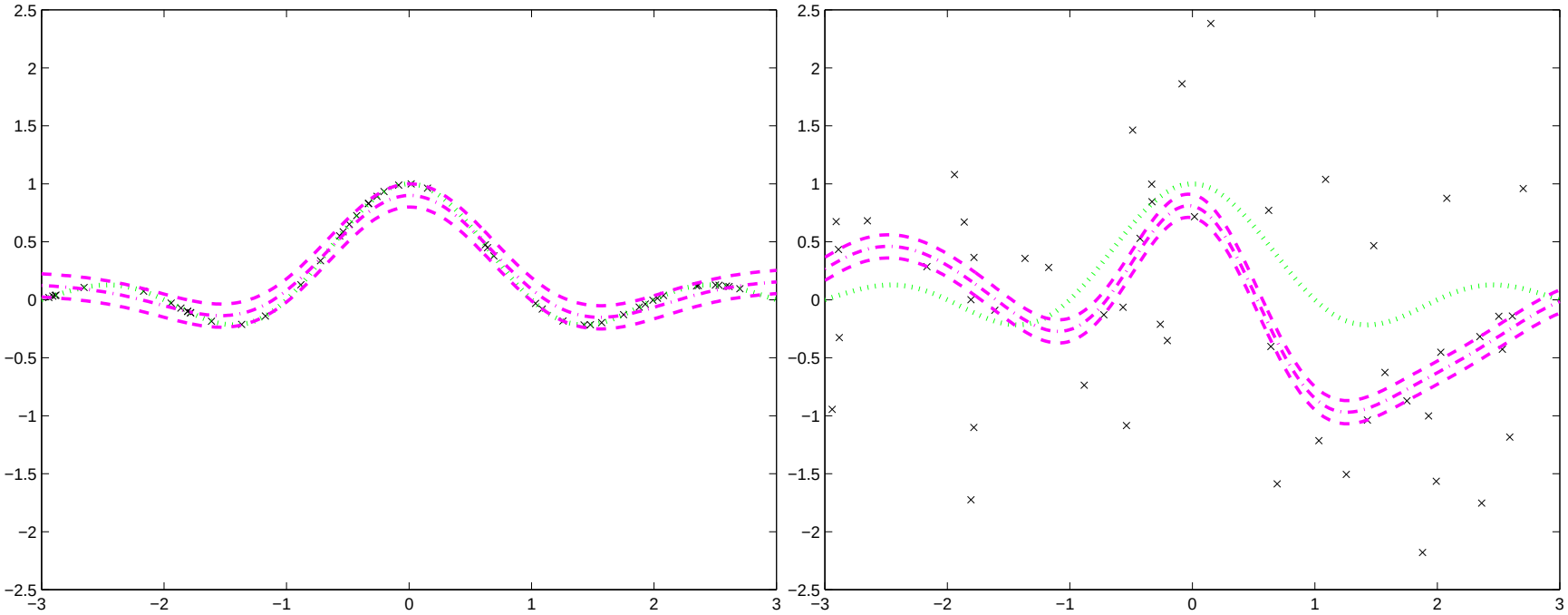
* Essentially, $\mathrm{P}(\mathbf{x}, y) = \mathrm{P}(\mathbf{x})\mathrm{P}(y|\mathbf{x})$ with $\mathrm{P}(y|\mathbf{x})$ continuous (some details omitted).

ν-SV-Regression: Automatic Tube Tuning



Identical machine parameters ($\nu = 0.2$), but different amounts of noise in the data.

ε-SV-Regression, Run on the Same Data



Identical machine parameters ($\varepsilon = 0.2$), but different amounts of noise in the data.

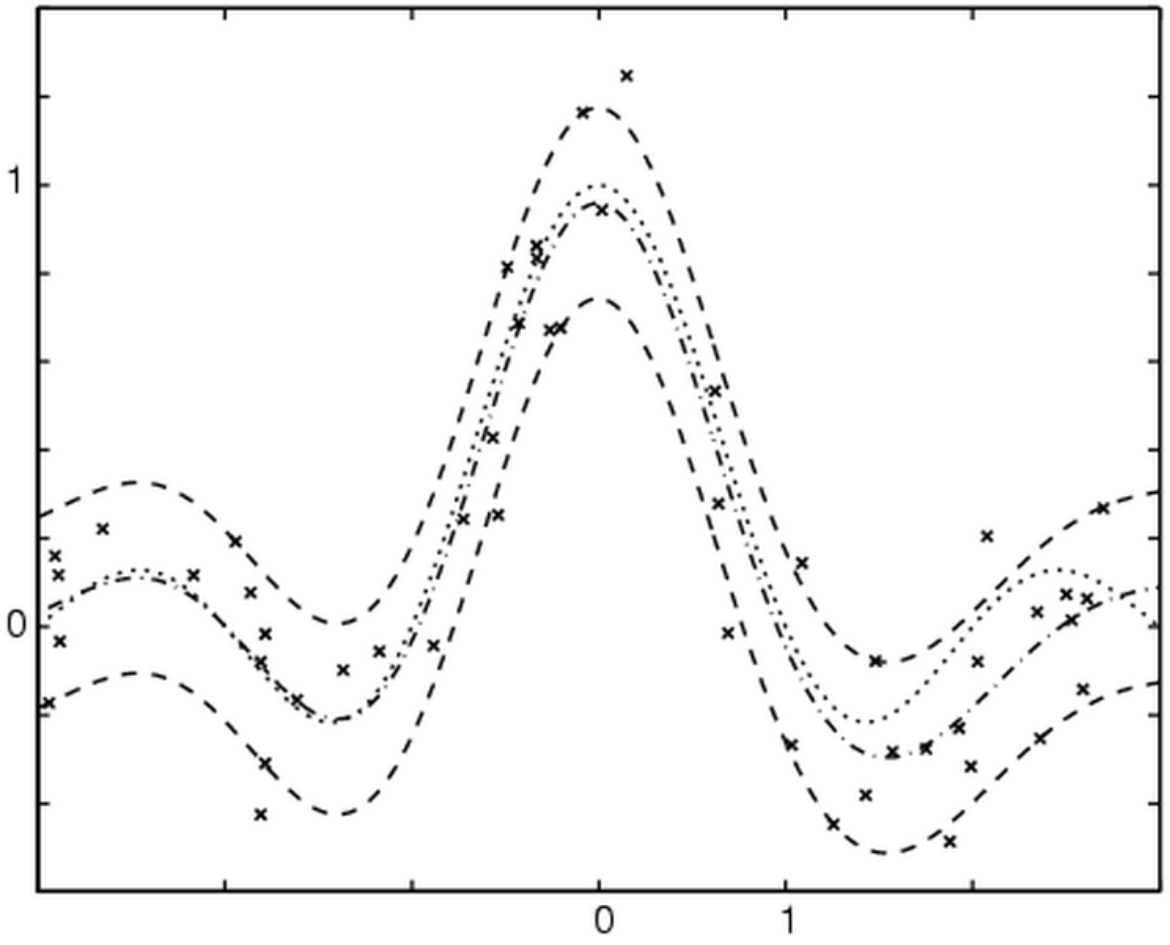
Toy Examples: Estimating a Noisy Sinc Function

$\nu = 0.2$

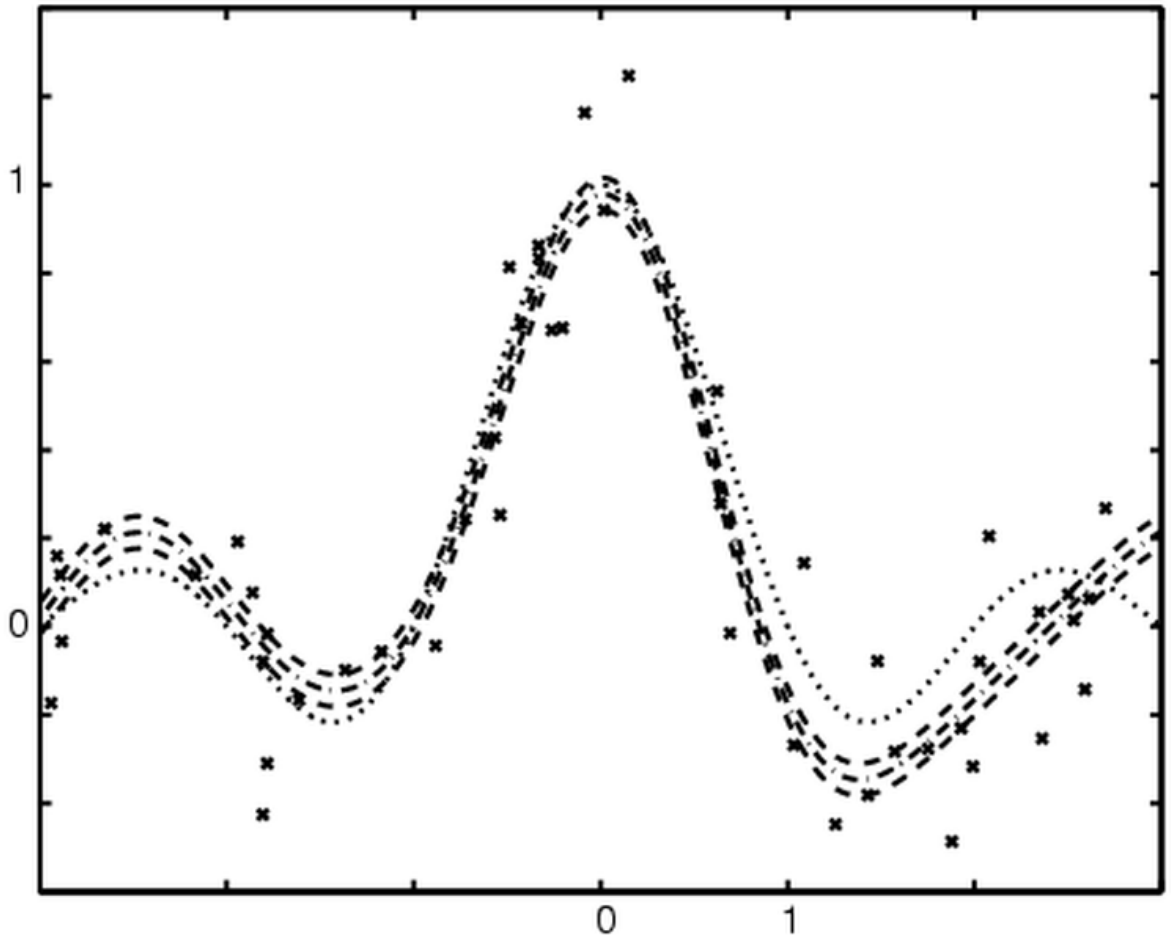
m	10	50	100	200	500	1000	1500	2000
ε	0.27	0.22	0.23	0.25	0.26	0.26	0.26	0.26
fraction of errors	0.00	0.10	0.14	0.18	0.19	0.20	0.20	0.20
fraction of SVs	0.40	0.28	0.24	0.23	0.21	0.21	0.20	0.20

- automatically computed ε largely independent of m
- asymptotics consistent with theorem

v-SV regression with *v*=0.2



v-SV regression with *v*=0.8



Robustness of SV Regression

Proposition. Using SVR with $|.|_\varepsilon$, local movements of target values of points outside the tube do not change the estimated regression.

Proof.

1. Shift y_i locally $\longrightarrow$ $(\mathbf{x}_i, y_i)$ still outside the tube $\longrightarrow$ original dual solution $\boldsymbol{\alpha}^{(*)}$ still feasible ($\alpha_i^{(*)} = C$, since *all* points outside the tube are at the upper bound).
2. The primal solution, with ξ_i transformed according to the movement, is also feasible.
3. The KKT conditions are still satisfied, as still $\alpha_i^{(*)} = C$. Thus [5, e.g.], $\boldsymbol{\alpha}^{(*)}$ is still the optimal solution.