Supervised Learning

CS536, Spring 2015

Partially based on slides by Mehryar Mohri, NYU

Today's Topics

- **Supervised Learning**:
  - Classification
  - Regression
- Input, Output, Examples
- Hypothesis, hypothesis class
- Error: true & empirical
- Version spaces
- VC-dimension
- PAC learning
- Model selection
- Generalization, underfitting, overfitting
- Train, test, validation

Learning a Class from Examples

- Class C of a "family car"
  - Prediction: Is car *x* a family car?
  - Knowledge extraction: What do people expect from a family car?
- Output:

  Positive (+) and negative (–) examples
- Input representation:

  x_1: price, x_2 : engine power

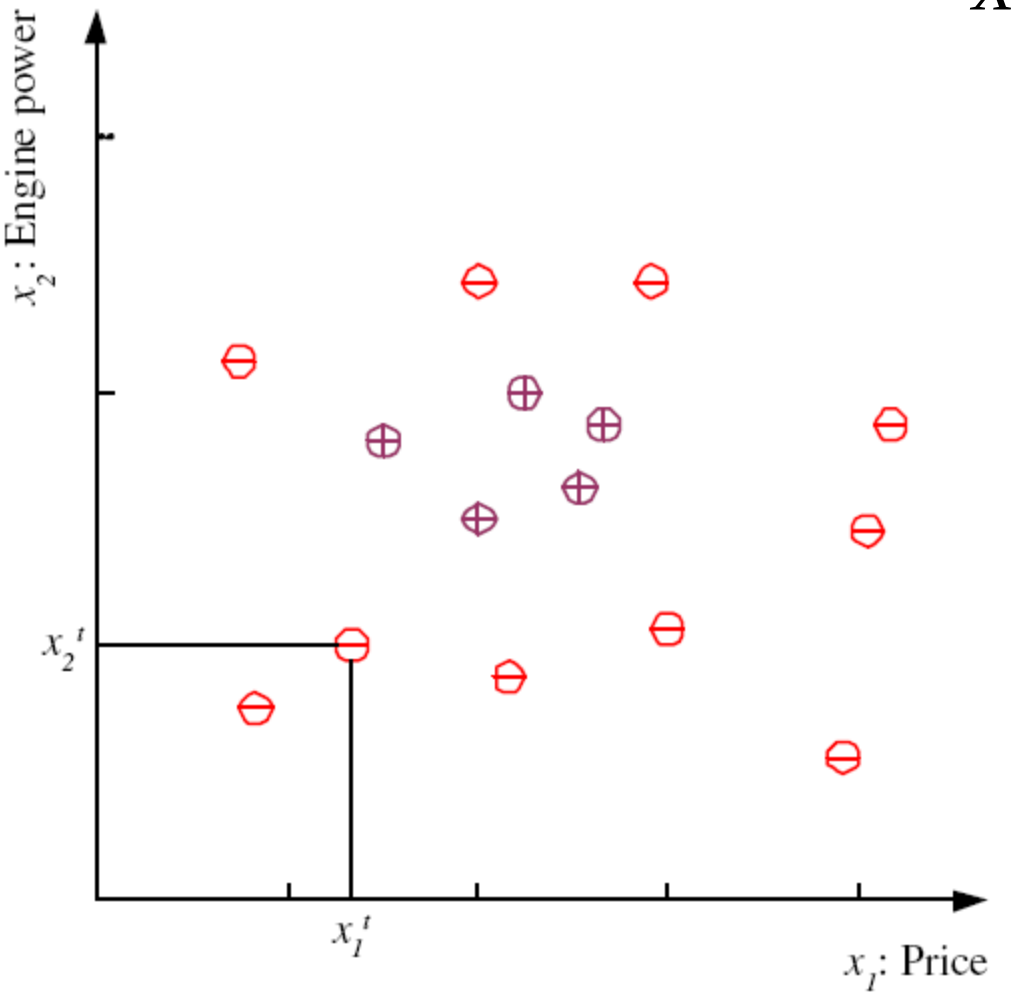
Data set $\mathcal{X}$

features | labels

$$X = \left\{ \left(x^t, y^t \right) \right\}_{t=1}^{N}$$

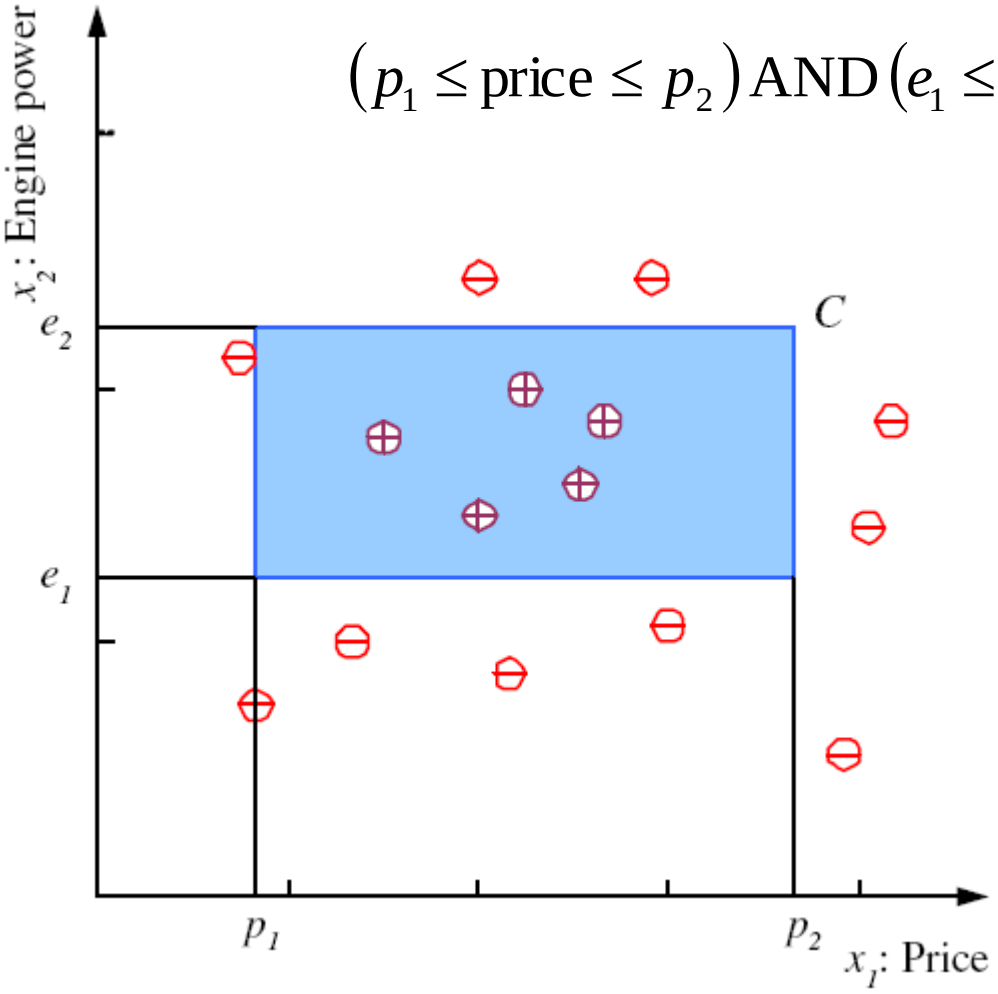
$$y = \begin{cases} 1 \text{ if } x \text{ is positive} \\ 0 \text{ if } x \text{ is negative} \end{cases}$$

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$



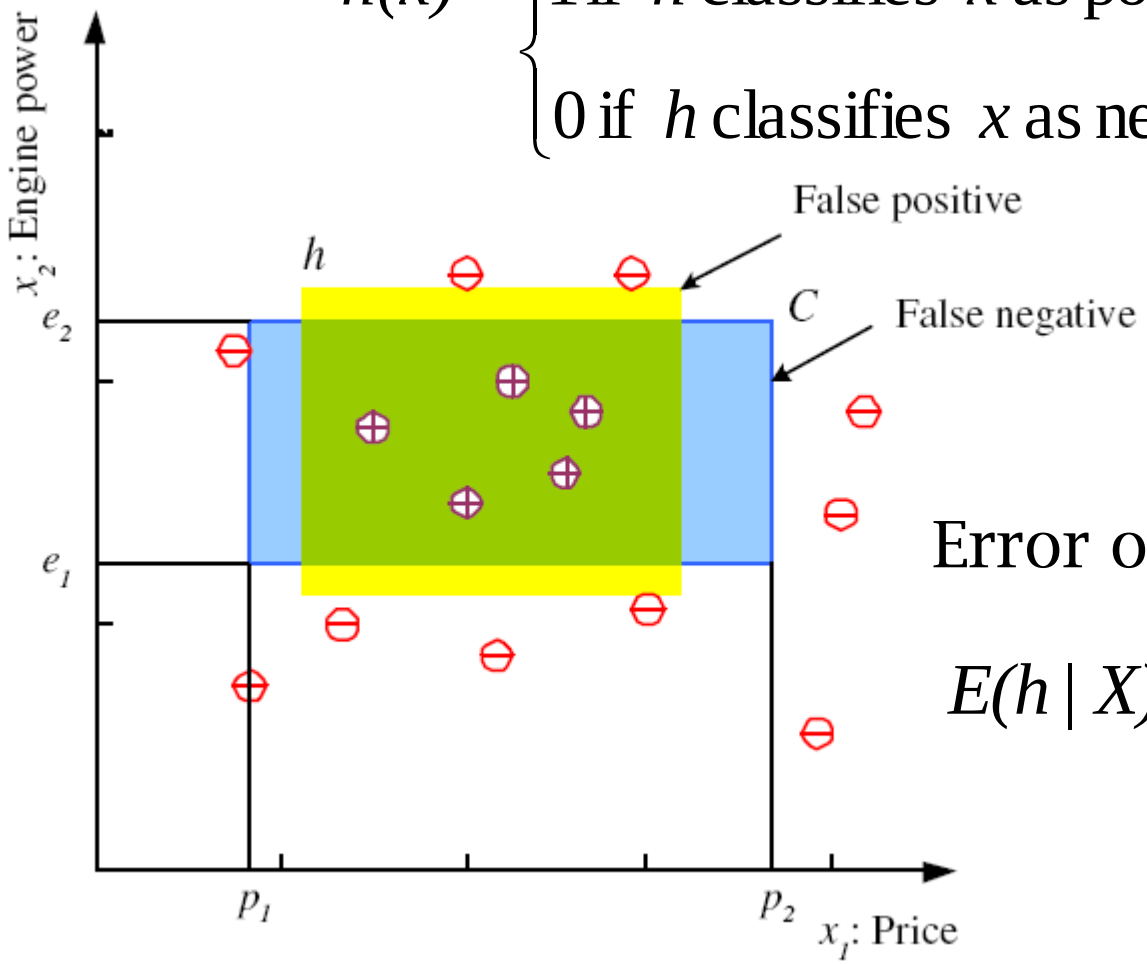
Class C

$$(p_1 \le \text{price} \le p_2) \text{ AND } (e_1 \le \text{engine power} \le e_2)$$



Hypothesis class $\mathcal{H}$

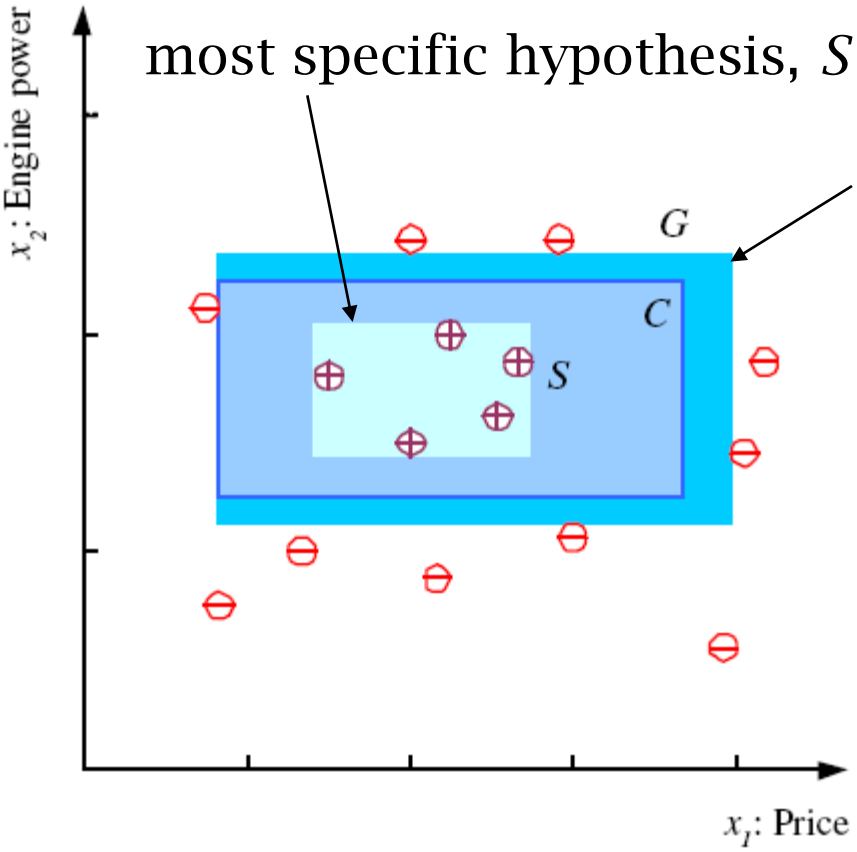
$$h(x) = \begin{cases} 1 \text{ if } h \text{ classifies } x \text{ as positive} \\ 0 \text{ if } h \text{ classifies } x \text{ as negative} \end{cases}$$



Error of h on $\mathcal{H}$

$$E(h \mid X) = \sum_{t=1}^{N} 1\left(h\left(x^t\right) \neq y^t\right)$$

S, G, and the Version Space



$h \in \mathcal{H}$, between S and G is consistent

and make up the version space

(Mitchell, 1997)

Supervised Learning Setup

- **Data (x,y)**
  - Comes from some unknown "true" density $(x, y) \sim P$
- **Training data**
  - An i.i.d. sample of fixed size N from P

$$\mathcal{D} = \{(x_1, y_1), \ldots, (x_N, y_N)\} = \{(x_i, y_i)\}_{i=1}^{N}$$

- **Empirical density**
  - Density of the training sample D

$$\hat{P} = \frac{1}{N} \sum_{i=1}^{N} \delta(x - x_i)\delta(y - y_i)$$

- **Generalization error** of hypothesis h

$$E[h] = \mathbb{E}_P \left[y \neq h(x) \right] = \sum_{x,y} I[y \neq h(x)] P(x, y)$$

- **Training error** of hypothesis h

$$\hat{E}[h|\mathcal{D}] = \mathbb{E}_{\hat{P}} \left[y \neq h(x) \right] = \sum_{x,y} I[y \neq h(x)] \hat{P}(x, y) = \frac{1}{N} \sum_{i=1}^{N} I[y_i \neq h(x_i)]$$

Learning Objective

- **Error minimization**:

Lowest possible error

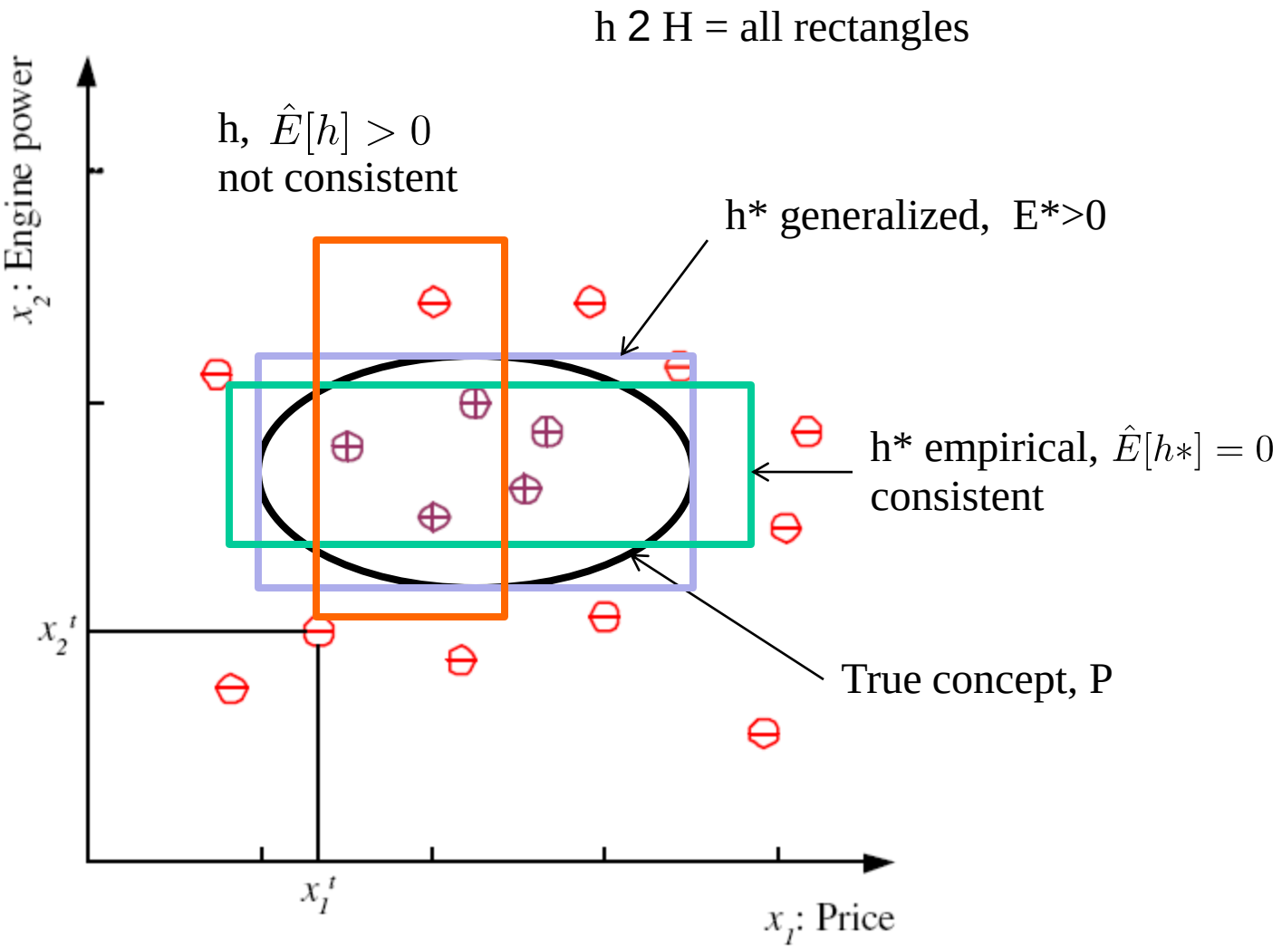
$$h^* = \arg\min_{h \in \mathcal{H}} E[h] \qquad E^* = E[h^*] \geq 0$$

Problem: P is not known, so cannot compute h*

- **Empirical error minimization**:

$$\begin{aligned} h^* &= \arg\min_{h \in \mathcal{H}} \hat{E}[h|\mathcal{D}] \\ &= \arg\min_{h \in \mathcal{H}} \frac{1}{N} \sum_{i=1}^{N} I[y_i \neq h(x_i)] \end{aligned}$$

  - Problem:
    - *H* may be too complex
    - *D* may not be large enough

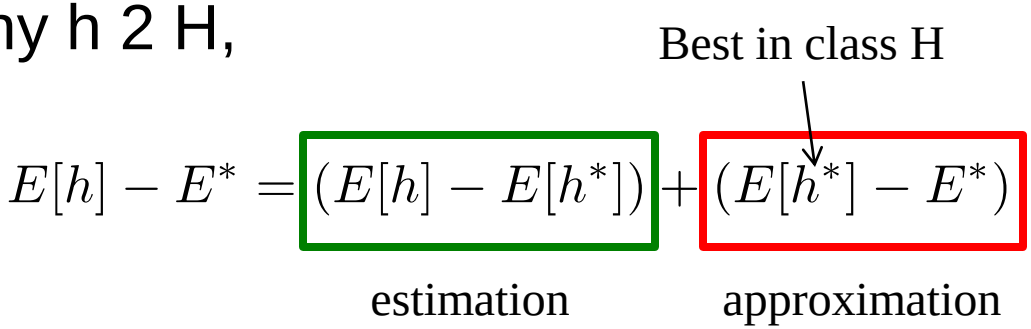

h 2 H = all rectangles
h, $\hat{E}[h] > 0$
not consistent
h* generalized, E*>0
h* empirical, $\hat{E}[h*] = 0$
consistent
True concept, P
x_2: Engine power
x_2^t
x_1^t
x_1: Price

Questions

- How do we estimate the error of a learning algorithm?
- How do we give performance guarantees for a learning algorithm?

Estimation and Approximation

- For any h 2 H,



- **Approximation**: not a random variable, depends on H

- **Estimation**: term we can hope to bound

Generalization Bounds

- **Definition**: given small ², upper bound on

$$Pr\left[|E(h) - \hat{E}(h)| > \epsilon\right]$$

- Bound on estimation error for hypothesis h_0 given by empirical error minimization

$$\begin{aligned} E(h_0) - E(h^*) &= E(h_0) - \hat{E}(h_0) + \hat{E}(h_0) - E(h^*) \\ &\leq E(h_0) - \hat{E}(h_0) + \hat{E}(h^*) - E(h^*) \\ &\leq 2 \max_{h \in \mathcal{H}} |E(h) - \hat{E}(h)| \end{aligned}$$

Probability Inequalities

- Boole's inequality (union bound):

$$P([_{k=1}^{N} E_k) \leq \min(\square_{k=1}^{N} P(E_k), 1)$$

- Bonferroni inequality (intersection bound):

$$P(\cap_{k=1}^{N} E_k) \geq \max(\square_{k=1}^{N} P(E_k)-N+1, 0)$$
$$= \max(1 - \square_{k=1}^{N} P(E^c_k), 0)$$

Basic Probability Tools

- **Union bound** $Pr[A \cup B] \leq Pr[A] + Pr[B]$
- **Inversion** $Pr[X \geq \epsilon] \leq f(\epsilon) \Rightarrow \forall \delta > 0,\ Pr[X \leq f^{-1}(\delta)] \geq 1 - \delta$
- **Jensen's inequality** $f \text{ convex} \Rightarrow f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]$
- **Expectation** $X \geq 0 \rightarrow \mathbb{E}[X] = \int_0^\infty Pr[X > t]dt$
- **Markov inequality** $X \geq 0, \epsilon > 0 \Rightarrow Pr[X \geq \epsilon] \leq \frac{E[X]}{\epsilon}$
- **Chebyshev's inequality** $Pr[|X - \mathbb{E}[X]| \geq \epsilon] \leq \frac{\sigma_X^2}{\epsilon^2}$

Hoeffding's Inequality

- **Theorem**: Let $X_1, \ldots, X_m$ be indep. rand. variables with the same expectation μ and X_i 2 [a, b], (a < b). Then, for any ² > 0 , the following inequalities hold:

$$Pr\left[\left(\mu - \frac{1}{m}\sum_{i=1}^{m} X_i\right) > \epsilon\right] \leq e^{-\frac{2m\epsilon^2}{(b-a)^2}}$$

$$Pr\left[\left(\frac{1}{m}\sum_{i=1}^{m} X_i - \mu\right) > \epsilon\right] \leq e^{-\frac{2m\epsilon^2}{(b-a)^2}}$$

McDiarmid's Inequality

- Theorem: let $X_1, \ldots, X_m$ be independent random variables taking values in U and f: U^m ! R a function verifying for all i 2 [1,m],

$$\max_{x_1,\ldots,x_m,x_i'} |f(x_1,\ldots,x_i,\ldots,x_m) - f(x_1,\ldots,x_i',\ldots,x_m)| \leq c_i.$$

Then, for all ² > 0

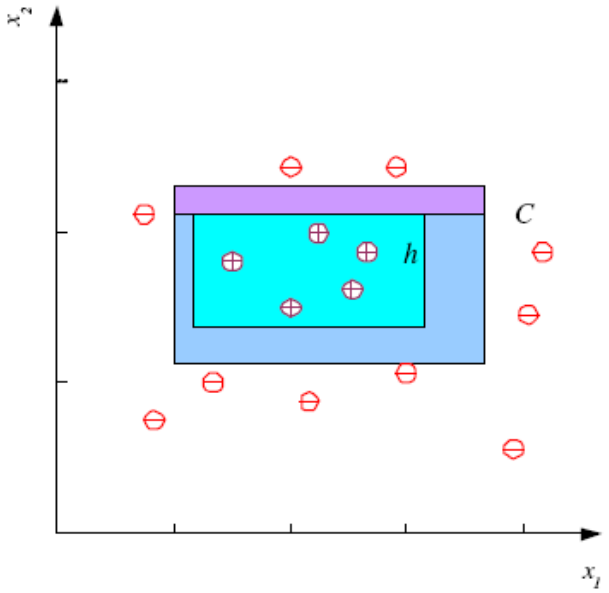
$$Pr\left[|f(X_1,\ldots,X_m) - \mathbb{E}[f(X_1,\ldots,X_m)]| > \epsilon\right] \leq 2e^{-\frac{2\epsilon^2}{\sum_{i=1}^m c_i}}$$

Probably Approximately Correct (PAC) Learning

- How many training examples *N* should we have, such that with probability at least 1 – δ, *h* has error at most ε ?
  (Blumer et al., 1989)

- Each strip is at most ε/4
- Pr that we miss a strip 1– ε/4
- Pr that *N* instances miss a strip $(1 - \varepsilon/4)^N$
- Pr that *N* instances miss 4 strips $4(1 - \varepsilon/4)^N$
- $4(1 - \varepsilon/4)^N \leq \delta$ and $(1 - x) \leq \exp(-x)$
- $4\exp(-\varepsilon N/4) \leq \delta$ and $N \geq (4/\varepsilon)\log(4/\delta)$

- E.g., ² = 0.01, ± = 0.05,
  N ¸ 1753



Things to Consider

- Set of axis parallel triangles = Infinite hypothesis set
  - Yet simple proof
- But does it apply to other classes?
- Geometric properties:
  - Key in this proof
  - Non-trivial in general, eg non-concentric circles

Learning Bound for Finite H – Consistent Case

- **Thm**: let H be a finite set of functions from X to {0, 1} and L an algorithm that for any target concept c 2 H and sample D returns a consistent hypothesis h_D: $^{E}(h_D) = 0$. Then, for any 0 < ± < ½, with probability at least 1-±,

$$E[h_D] \leq \frac{1}{N}\left(\log|H| + \log\frac{1}{\delta}\right).$$

- Proof:

$$Pr\left[h \text{ consistent} | E[h] < \epsilon\right] \leq (1-\epsilon)^N$$

$$\begin{aligned}
&Pr\left[\exists h \in H : h \text{ consistent} \wedge E[h] > \epsilon\right] \\
&= Pr\left[(h_1 \text{ consistent} \wedge E[h_1] > \epsilon) \vee \ldots \vee h_{|H|} \text{ consistent} \wedge E[h_{|H|}] > \epsilon)\right] \\
&\leq \sum_{h \in H} Pr\left[h \text{ consistent} \wedge E[h] > \epsilon\right] \\
&\leq \sum_{h \in H} Pr\left[h \text{ consistent} | E[h] > \epsilon\right] \\
&\leq \sum_{h \in H} (1-\epsilon)^N = |H|(1-\epsilon)^N \leq |H| e^{-N\epsilon}
\end{aligned}$$

Example: Boolean Conjunction Learning

- Domain: binary vectors of length n, $(x_1,\ldots,x_n)$
- Class label each domain point as +-1
- Goal: learn mapping h: $(x_1,\ldots,x_n)$! {+1,-1}

x_1	x_2	x_3	x_4	x_5	x_6	y
0	1	1	0	1	1	+
0	1	1	1	1	1	+
0	0	1	1	0	1	-
0	1	1	1	1	1	+
1	0	0	1	1	0	-
0	1	0	0	1	1	+
0	1	?	?	1	1	?

$$\overline{x}_1 \wedge x_2 \wedge x_5 \wedge x_6$$

Learning Conjunctions

- Algorithm: Choose h consistent with D

  Start with $x_1\bar{x}_1x_2\bar{x}_2\ldots x_n\bar{x}_n$ and eliminate inconsistent literals. Set consistent literals to either x_i or $\bar{x}_i$

- |H| = k^n, k – number of literals, n – number of features

- Therefore, for e.g., k=3, n=10, ± = 0.02, ² = 0.1

$$N \geq \frac{1}{\epsilon}\left(n\log 3 + \log\frac{1}{\delta}\right) = 149$$

- Complexity of learning: polynomial, cost per data point is O(n)

Inconsistent Case

- No h2 H is consistent with D.
- This is a typical case in practice.
- Even inconsistent hypotheses with small errors can be useful.

- Analysis: needs Hoeffding's inequality.

$$
\begin{aligned}
Pr\left[\left(E[h]-\hat{E}[h]\right)>\epsilon\right] &\leq e^{-2N\epsilon^2} \\
Pr\left[\left(\hat{E}[h]-E[h]\right)>\epsilon\right] &\leq e^{-2N\epsilon^2} \\
Pr\left[|E[h]-\hat{E}[h]|>\epsilon\right] &\leq 2e^{-2N\epsilon^2}
\end{aligned}
$$

Generalization Bound – Finite H

- **Thm**: Let H be a finite hypothesis set. Then, for any 0 < ± <1/2, with probability at least 1-±

$$\forall h \in H,\ E[h] \leq \hat{E}[h|D] + \sqrt{\frac{\log|H| + \log\frac{2}{\delta}}{2N}}$$

- Proof:

$$\begin{aligned} & Pr\left[\max_{h\in H}\left|E[h] - \hat{E}[h]\right| > \epsilon\right] \\ = \quad & Pr\left[\left|E[h_1] - \hat{E}[h_1]\right| > \epsilon \ldots \left|E[h_{|H|}] - \hat{E}[h_{|H|}]\right| > \epsilon\right] \\ \leq \quad & \sum_{h\in H} Pr\left[\left|E[h] - \hat{E}[h]\right| > \epsilon\right] \\ \leq \quad & 2|H|e^{-2N\epsilon^2} \end{aligned}$$

Remarks

- For a finite hypothesis set, with high probability

$$\forall h \in H,\ E[h] \leq \hat{E}[h|D] + \sqrt{\frac{\log |H|}{N}}$$

- Error bound in $O(1/N^{1/2})$, so quadratic worse
- $\log_2|H|$ can be seen as the **number of bits** needed to encode H

- **Occam's Razor principle** (theologian William of Occam):"plurality should not be posited without necessity".

Occam's Razor

- Principle formulated by controversial theologian William of Occam:"plurality should not be posited without necessity", rephrased as "the simplest explanation is best";
  - invoked in a variety of contexts, e.g., syntax. Kolmogorov complexity can be viewed as the corresponding framework in information theory.
  - here, to minimize true error, choose the most parsimonious explanation (smallest |H|).
  - we will see later other applications of this principle.

Learning Bounds for Infinite H: Rademacher Complexity

- G is a family of functions mapping from Z to [a,b]
- Samples D = $\{z_1,\ldots,z_N\}$
- $¾_i$ (Rademacher variables): iid rv's taking values in {-1,+1}

- Empirical Rademacher complexity of G

$$\hat{\mathcal{R}}_D(G) = \mathbb{E}_\sigma \left[\sup_{g \in G} \frac{1}{N} \sum_{i=1}^{N} \sigma_i g(z_i) \right]$$

Correlation with random noise

- Rademacher complexity of G

$$\mathcal{R}_N(G) = \mathbb{E}_{D \sim P^N} \left[\hat{\mathcal{R}}_D(G) \right]$$

Rademacher Complexity Bound

- Thm: Let G be a family of functions mapping from Z to [0,1]. Then, for any 0<±<1/2, with probability 1-±, the following holds for all g 2 G:

$$E[g(z)] \leq \frac{1}{N}\sum_{i=1}^{N} g(z_i) + 2\mathcal{R}_N(G) + \sqrt{\frac{\log\frac{1}{\delta}}{2N}}$$

$$E[g(z)] \leq \frac{1}{N}\sum_{i=1}^{N} g(z_i) + 2\hat{\mathcal{R}}_D(G) + 3\sqrt{\frac{\log\frac{2}{\delta}}{2N}}$$

(Koltchinskii and Panchenko, 2002)

Generalization Bounds Rademacher

- Thm: Let H be a family of functions taking values in {-1,+1}. Then, for any ±>0, with probability at least 1-±, for any h 2 H,

$$E[h] \leq \hat{E}[h] + \mathcal{R}_N(H) + \sqrt{\frac{\log \frac{1}{\delta}}{2N}}$$

$$E[h] \leq \hat{E}[h] + 2\hat{\mathcal{R}}_D(H) + 3\sqrt{\frac{\log \frac{2}{\delta}}{2N}}$$

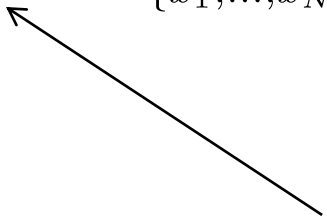
Remarks

- First bounds are distribution dependent, second are data dependent
- How to compute Rademacher complexity? Need to do Empirical Error Minimization, computationally hard.
- Alternative: combinatorial measures.

Growth Functions

- Def: the growth function ¦H: N ! N, for a hypothesis set H is defined by

$$\forall N \in \mathbb{N}, \Pi_H(N) = \max_{\{x_1,\ldots,x_N\} \subseteq X} |\{(h(x_1), \ldots, h(x_N)) : h \in H\}|$$



Max. number of ways to classify N points using H

Massart's Lemma

(Massart, 2000)

Theorem: Let $A \subseteq \mathbb{R}^m$ be a finite set, with $R = \max_{x \in A} \|x\|_2$, then, the following holds:!

$$\mathop{\mathbb{E}}_{\sigma}\left[\frac{1}{m}\sup_{x \in A}\left|\sum_{i=1}^{m}\sigma_i x_i\right|\right] \leq \frac{R\sqrt{2\log|A|}}{m}.$$

Proof:

$$\exp\left(t\mathop{\mathbb{E}}_{\sigma}\left[\sup_{x \in A}\sum_{i=1}^{m}\sigma_i x_i\right]\right) \leq \mathop{\mathbb{E}}_{\sigma}\left(\exp\left[t\sup_{x \in A}\sum_{i=1}^{m}\sigma_i x_i\right]\right) \quad \text{(Jensen's ineq.)}$$

$$= \mathop{\mathbb{E}}_{\sigma}\left(\sup_{x \in A}\exp\left[t\sum_{i=1}^{m}\sigma_i x_i\right]\right)$$

$$\leq \sum_{x \in A}\mathop{\mathbb{E}}_{\sigma}\left(\exp\left[t\sum_{i=1}^{m}\sigma_i x_i\right]\right) = \sum_{x \in A}\prod_{i=1}^{m}\mathop{\mathbb{E}}_{\sigma}\left(\exp\left[t\sigma_i x_i\right]\right)$$

$$\text{(Hoeffding's ineq.)} \leq \sum_{x \in A}\left(\exp\left[\frac{\sum_{i=1}^{m}t^2(2|x_i|)^2}{8}\right]\right) \leq |A|e^{\frac{t^2R^2}{2}}.$$

Taking the log yields:!

$$\mathop{\mathbb{E}}_{\sigma}\left[\sup_{x \in A}\left|\sum_{i=1}^{m}\sigma_i x_i\right|\right] \leq \frac{\log|A|}{t} + \frac{tR^2}{2}.$$

Minimizing the bound by choosing $t = \frac{\sqrt{2\log|A|}}{R}$ gives

$$\mathop{\mathbb{E}}_{\sigma}\left[\sup_{x \in A}\left|\sum_{i=1}^{m}\sigma_i x_i\right|\right] \leq R\sqrt{2\log|A|}.$$

Growth Function Bound on R-Complexity

Corollary: Let G be a family of functions taking values in $\{-1, +1\}$, then the following holds:!

$$R_m(G) \leq \sqrt{\frac{2 \log \Pi_G(m)}{m}}.$$

Proof:

$$\begin{aligned}
\hat{R}_S(G) &= \mathop{\mathrm{E}}_{\sigma}\left[\sup_{g \in G} \frac{1}{m} \begin{bmatrix} \sigma_1 \\ \vdots \\ \sigma_m \end{bmatrix} \cdot \begin{bmatrix} g(z_1) \\ \vdots \\ g(z_m) \end{bmatrix}\right] \\
&\leq \frac{\sqrt{m}\sqrt{2 \log |\{(g(z_1), \ldots, g(z_m)) : g \in G\}|}}{m} \quad \text{(Massart's Lemma)} \\
&\leq \frac{\sqrt{m}\sqrt{2 \log \Pi_G(m)}}{m} = \sqrt{\frac{2 \log \Pi_G(m)}{m}}.
\end{aligned}$$

Generalization Bound – Growth Function

Corollary: Let H be a family of functions taking values in $\{-1, +1\}$. Then, for any $\delta > 0$, with probability at least $1 - \delta$, for any $h \in H$,

$$R(h) \leq \widehat{R}(h) + \sqrt{\frac{2 \log \Pi_H(m)}{m}} + \sqrt{\frac{\log \frac{1}{\delta}}{2m}}.$$

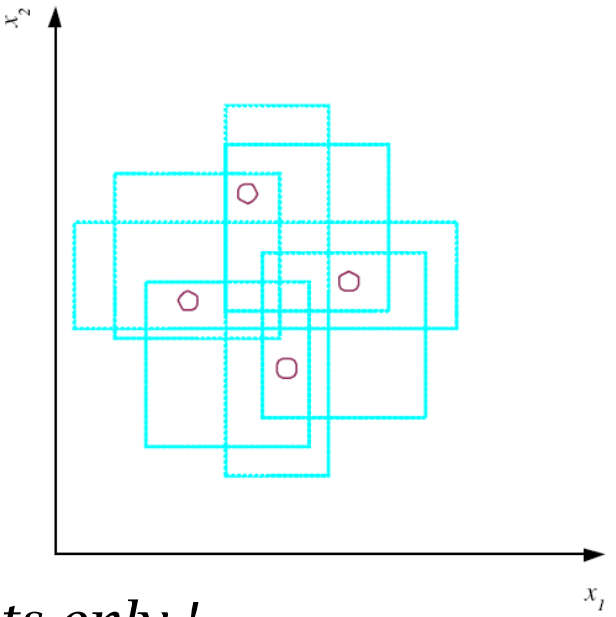
But, how do we compute the growth function? Relationship with the VC-dimension (Vapnik-Chervonenkis dimension).

VC Dimension

- N points can be labelled in 2^m ways as +/–
- $\mathcal{H}$ shatters m if there exists $h \in \mathcal{H}$ consistent with any such labelling:

$\text{VC}(\mathcal{H}) = m$

$$\text{VCdim}(H) = \max\{m : \Pi_H(m) = 2^m\}.$$

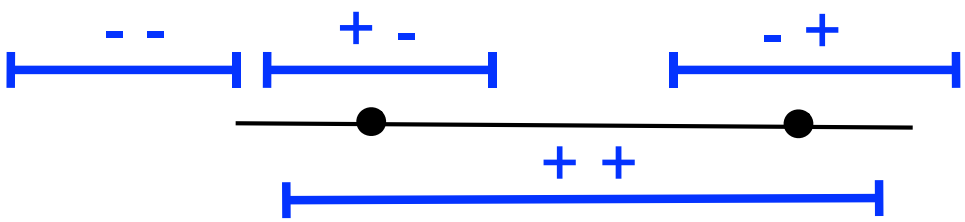


An axis-aligned rectangle shatters 4 points only !

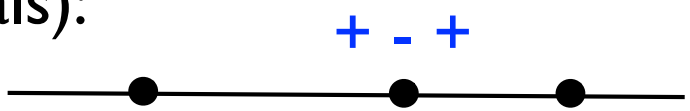
Example: Intervals on Real Line

Observations:

- Any set of two points can be shattered by four intervals



- No set of three points can be shattered since the following dichotomy "+ - +" is not realizable (by definition of intervals):

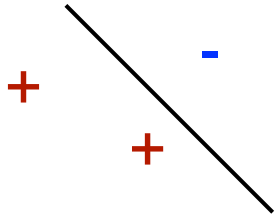


- Thus, $\mathrm{VCdim}(\text{intervals in } \mathbb{R}) = 2$.

Hyperplanes

Observations:

- Any three non-collinear points can be shattered:



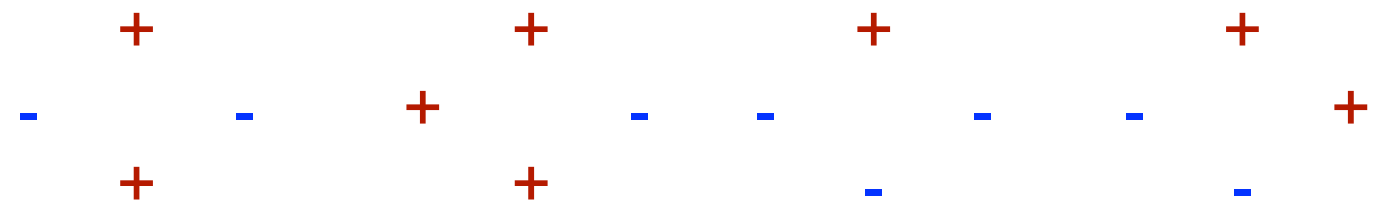
- Unrealizable dichotomies for four points:

- Thus, $\mathrm{VCdim}(\text{hyperplanes in } \mathbb{R}^d) = d+1.$

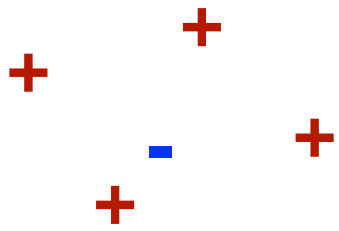
Axis-Aligned Rectangles

Observations:

- The following four points can be shattered:



- No set of five points can be shattered: label negatively the point that is not near the sides.

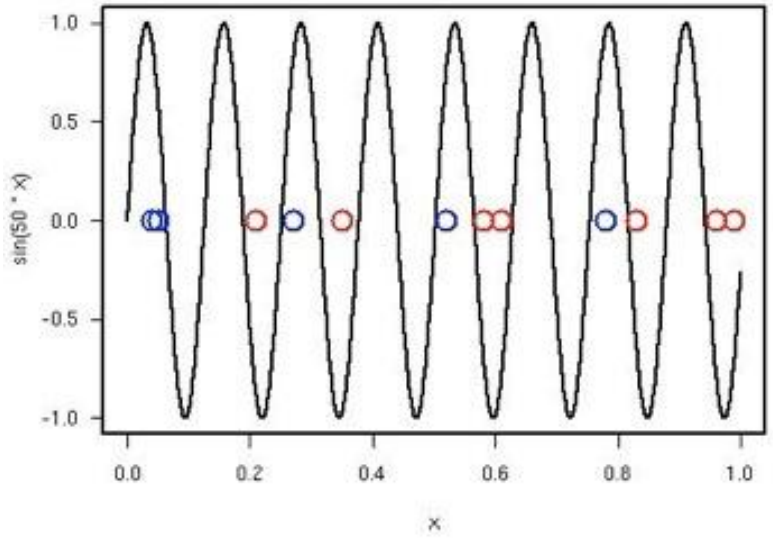


- Thus, $\mathrm{VCdim}(\text{axis-aligned rectangles}) = 4$.

Sine Function

Observations:

- Any finite set of points can be shattered by the hypothesis set $\{t \mapsto \sin(\omega t) : \omega \in \mathbb{R}\}$.
- Thus, $\mathrm{VCdim}(\text{sine functions}) = +\infty$.



Sauer's Lemma

(Vapnik & Chervonenkis, 1968-1971; Sauer, 1972)

Theorem: let H be a hypothesis set with $\mathrm{VCdim}(H) = d$ then, for all $m \in \mathbb{N}$,

$$\Pi_H(m) \leq \sum_{i=0}^{d} \binom{m}{i}.$$

Corollary: let H be a hypothesis set with $\mathrm{VCdim}(H) = d$ then, for all $m \geq d$,!

$$\Pi_H(m) \leq \left(\frac{em}{d}\right)^d = O(m^d).$$

Remarkable property of growth function:

- either $\mathrm{VCdim}(H) = d < +\infty$ and $\Pi_H(m) = O(m^d)$
- or $\mathrm{VCdim}(H) = +\infty$ and $\Pi_H(m) = 2^m$.

Generalization Bound – VC Dimension

Corollary: Let H be a family of functions taking values in $\{-1, +1\}$ with VC dimension d. Then, for any $\delta > 0$, with probability at least $1-\delta$, for any $h \in H$,

$$R(h) \leq \widehat{R}(h) + \sqrt{\frac{2d \log \frac{em}{d}}{m}} + \sqrt{\frac{\log \frac{1}{\delta}}{2m}}.$$

(Vapnik & Chervonenkis, 1971; Vapnik, 1982)

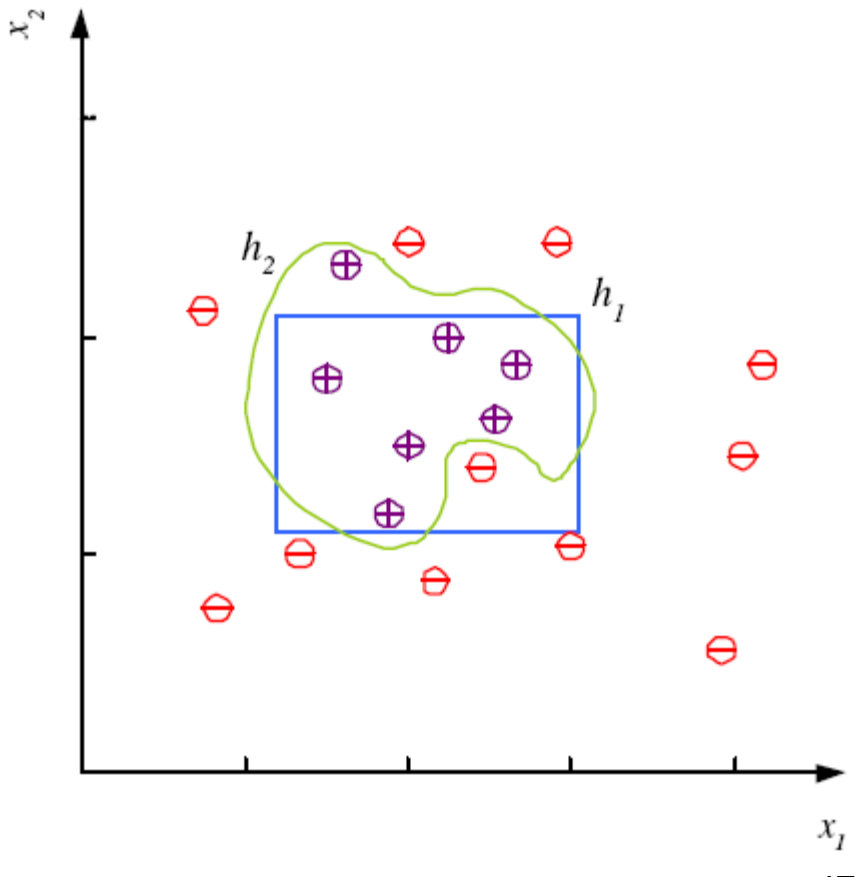
Theorem: Let H be a family of functions taking values in $\{-1, +1\}$ with VC dimension d. Then, for any $\delta > 0$, with probability at least $1-\delta$, for any $h \in H$,!

$$R(h) \leq \widehat{R}(h) + \sqrt{\frac{8d \log \frac{2em}{d} + 8 \log \frac{4}{\delta}}{m}}.$$

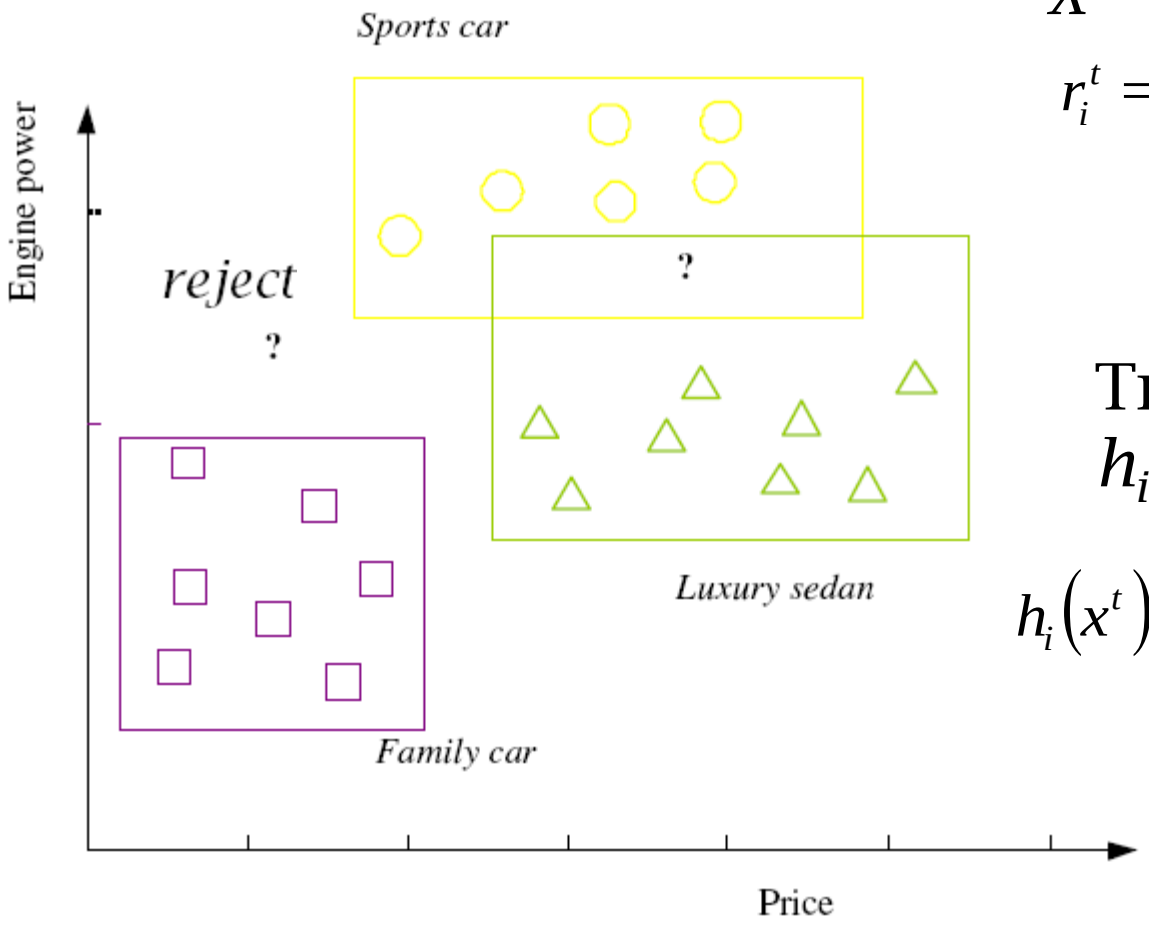
Noise and Model Complexity

Use the simpler one because

- Simpler to use (lower computational complexity)
- Easier to train (lower space complexity)
- Easier to explain (more interpretable)
- Generalizes better (lower variance - Occam's razor)



Multiple Classes, C_i i=1,...,K



$$X = \left\{x^t, r^t\right\}_{t=1}^{N}$$

$$r_i^t = \begin{cases} 1 \text{ if } x^t \in C_i \\ 0 \text{ if } x^t \in C_j,\ j \neq i \end{cases}$$

Train hypotheses $h_i(\boldsymbol{x}),\ i=1,...,K$:

$$h_i\left(x^t\right) = \begin{cases} 1 \text{ if } x^t \in C_i \\ 0 \text{ if } x^t \in C_j,\ j \neq i \end{cases}$$

Model Selection & Generalization

- Learning is an ill-posed problem; data is not sufficient to find a unique solution
- The need for inductive bias, assumptions about $\mathcal{H}$
- Generalization: How well a model performs on new data
- Overfitting: $\mathcal{H}$ more complex than C or f
- Underfitting: $\mathcal{H}$ less complex than C or f

Dimensions of a Supervised Learner

1. Model: $g(x|\theta)$
2. Loss function & training error: $R(\theta|X) = \sum_t L(r^t, g(x^t|\theta))$
3. Optimization procedure: $\theta^* = \arg\min_\theta R(\theta|X)$
4. Generalization error: $\mathbb{E}_{(x,r)\sim D}\left[L(r, g(x|\theta))\right]$

Triple Trade-Off

- There is a trade-off between three factors (Dietterich, 2003):
  - Complexity of $\mathcal{H}$, c ($\mathcal{H}$),
  - Training set size, N,
  - Generalization error, E, on new data

- As N *grows*, R *drops*
- As $c(\mathcal{H})$ grows, first E *grows*, and then E *drops*

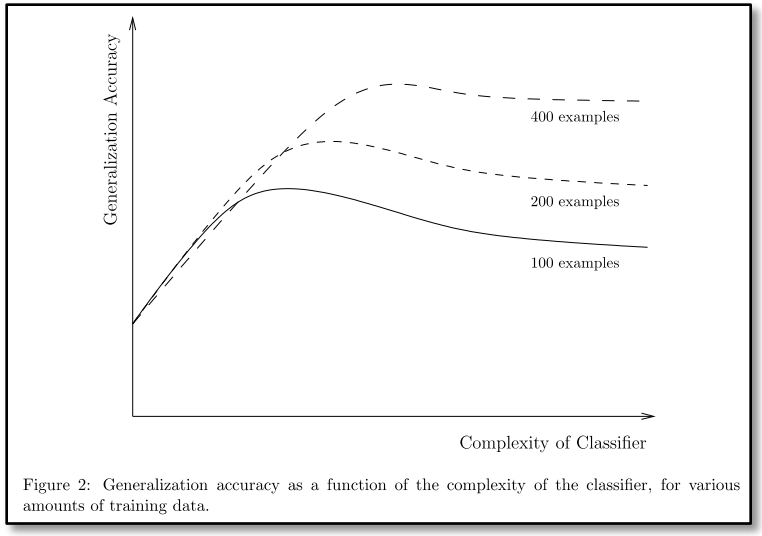


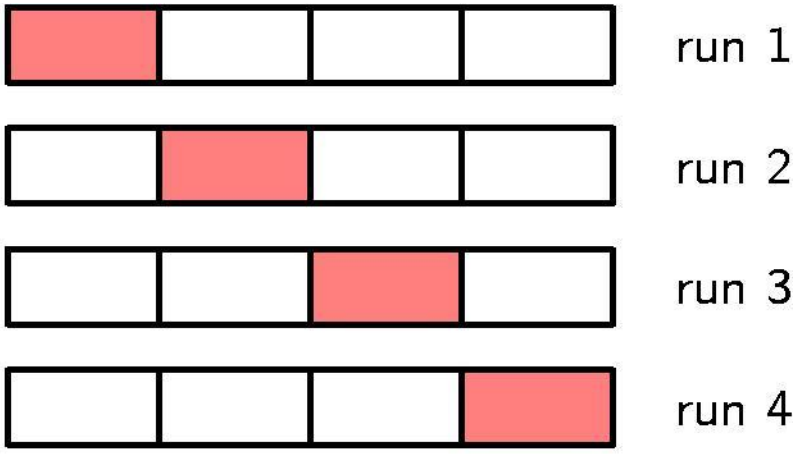
Figure 2: Generalization accuracy as a function of the complexity of the classifier, for various amounts of training data.

Dietterich, T. G. (2003). Machine Learning In Nature Encyclopedia of Cognitive Science, London: Macmillan, 2003

http://web.engr.oregonstate.edu/~tgd/projects/tutorials.html

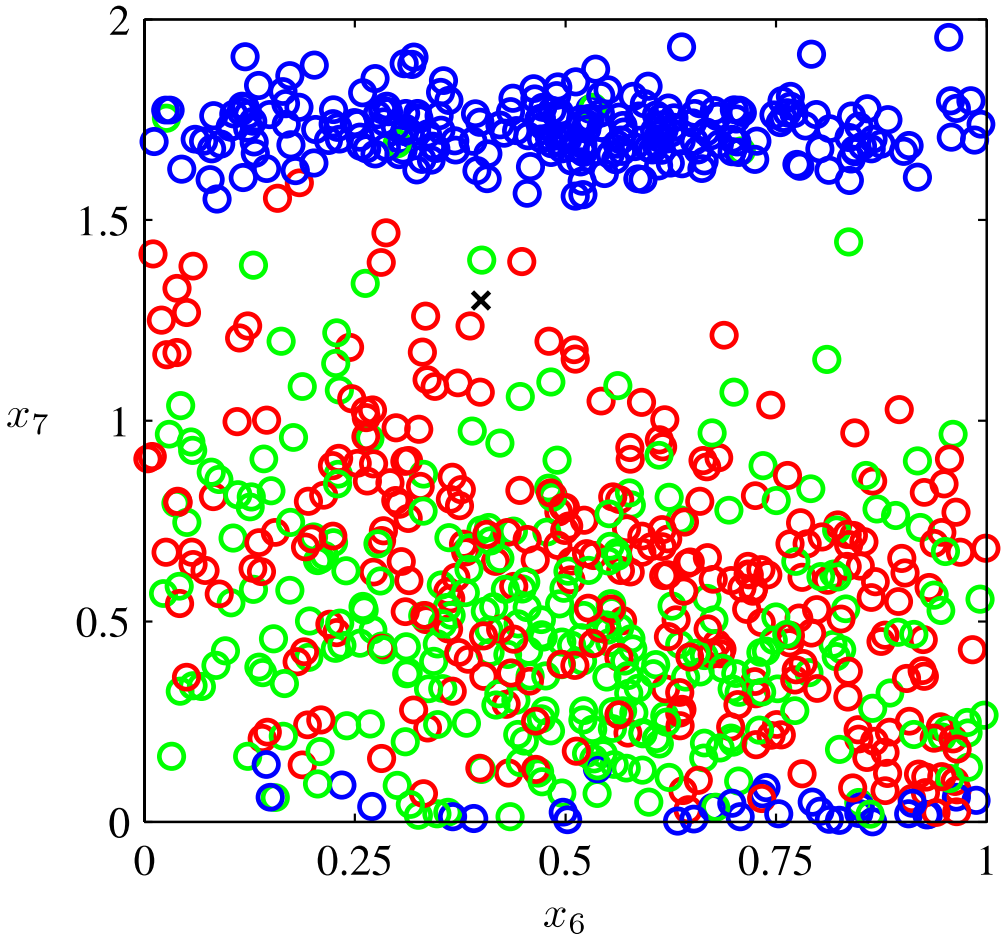
Cross-Validation

- To estimate generalization error, we need data unseen during training. We split the data as
  - Training set (50%)
    Validation set (25%)
  - Test (publication) set (25%)
- Resampling when there is few data



Oil Pipeline Data

x = 12-dimensional input vector consisting of measurements taken with gamma ray densitometers that measure the attenuation of gamma rays passing along narrow beams through the pipe.

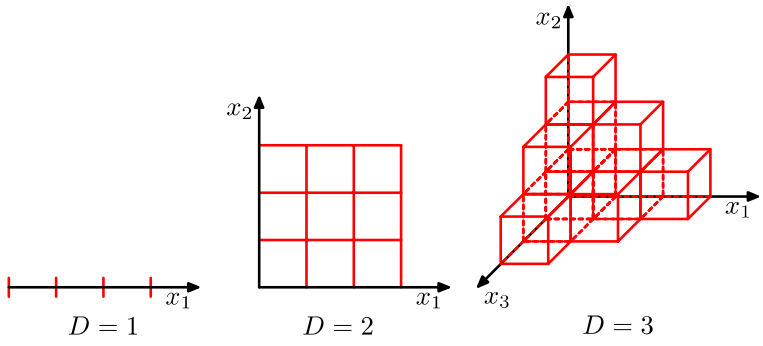


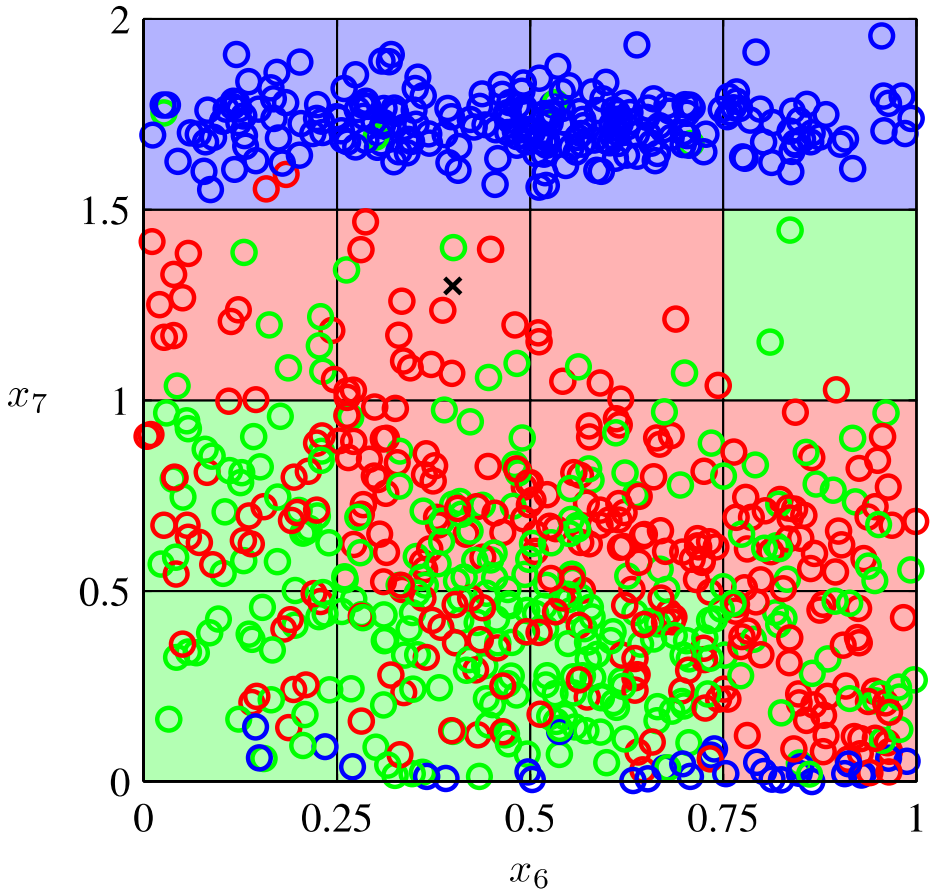
Scatter plot of the oil flow data for input variables x_6 and x_7, in which red denotes the 'homogenous' class, green denotes the 'annular' class, and blue denotes the 'laminar' class. Our goal is to classify the new test point denoted by '$\times$'.

Oil Pipeline (cont'd)

- Simple classification approach: regular cell partitioning

- Problem: what happens as dim(x) grows?

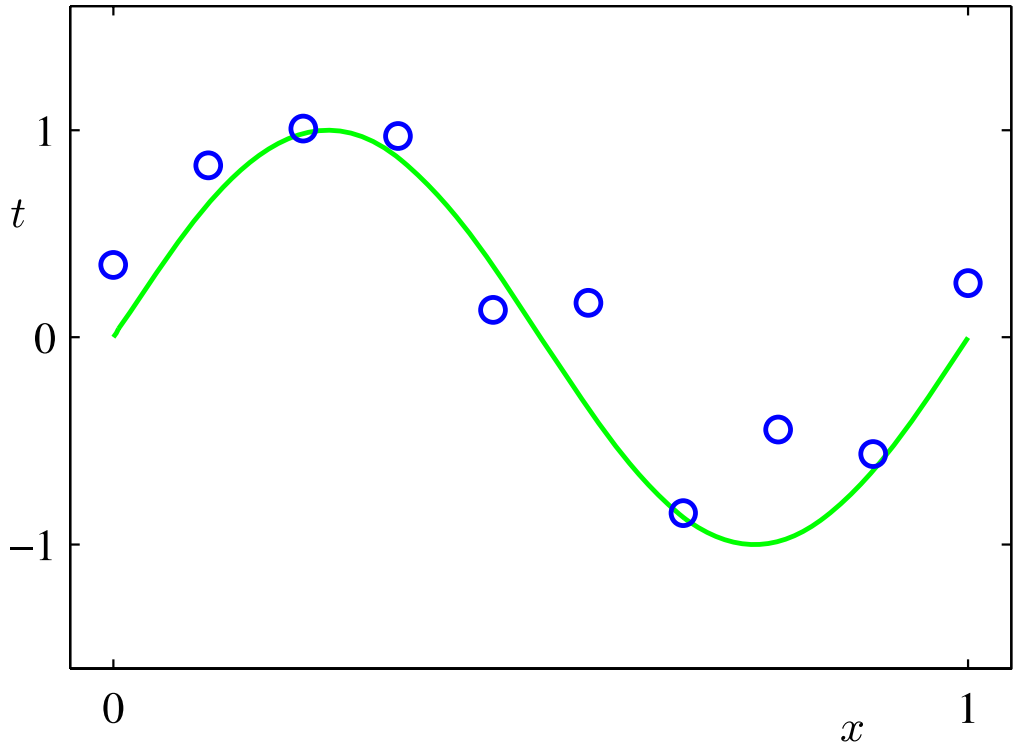




Example: polynomial curve fitting

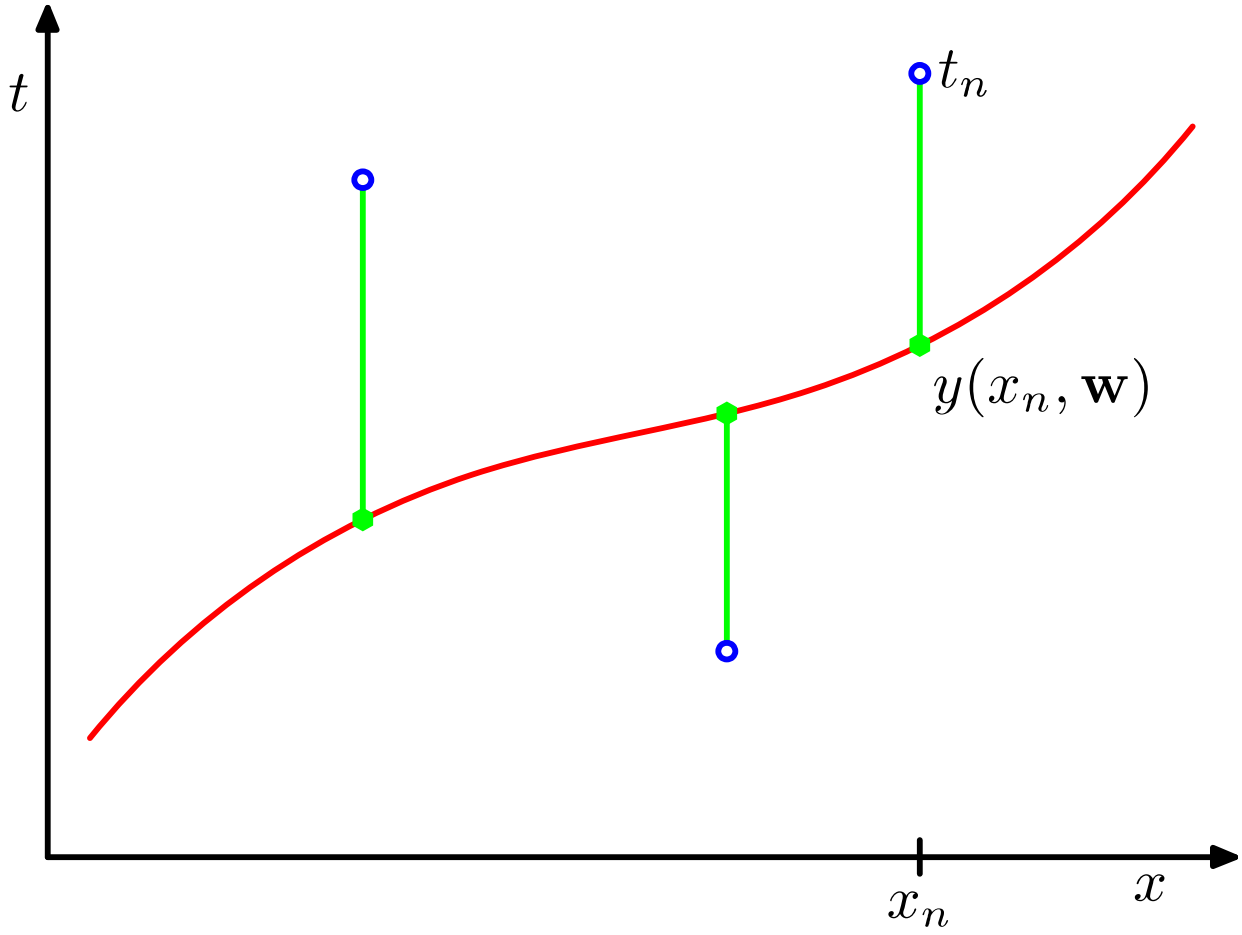
C = sinusoid, t(x) = sin(x)

D = { (x,t) }, |D| = N = 10



H = polynomials of order M $g(\mathbf{x}\,|\,\mathbf{w}) = \sum_{i=0}^{M} w_i x^i = \mathbf{w}^T\mathbf{x}$

Sum of Squares Error

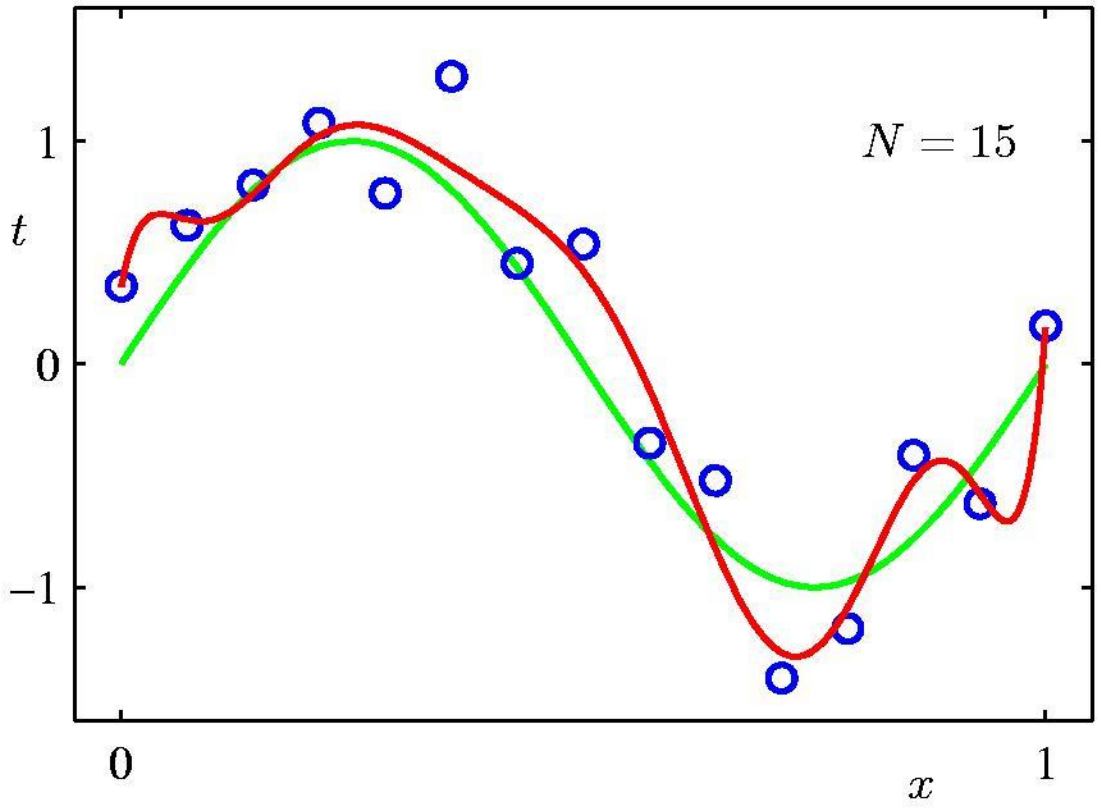


L = squared loss

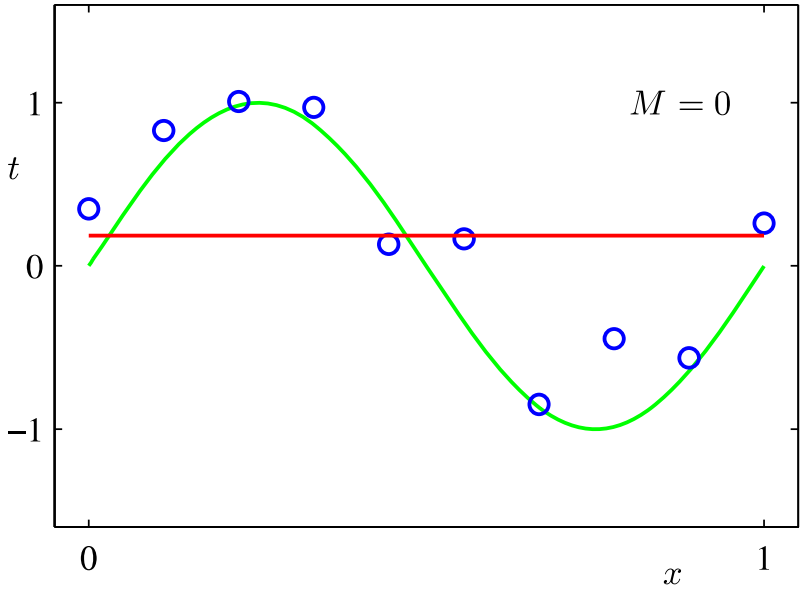
$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\left(\mathbf{w}^T\mathbf{x}_n - t_n\right)^2$$

Data Set Size: N = 15

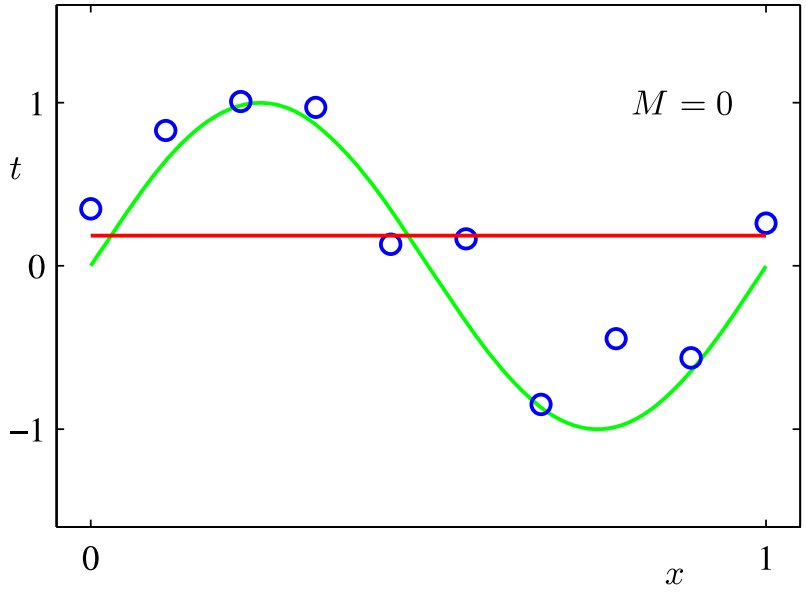
9th Order Polynomial

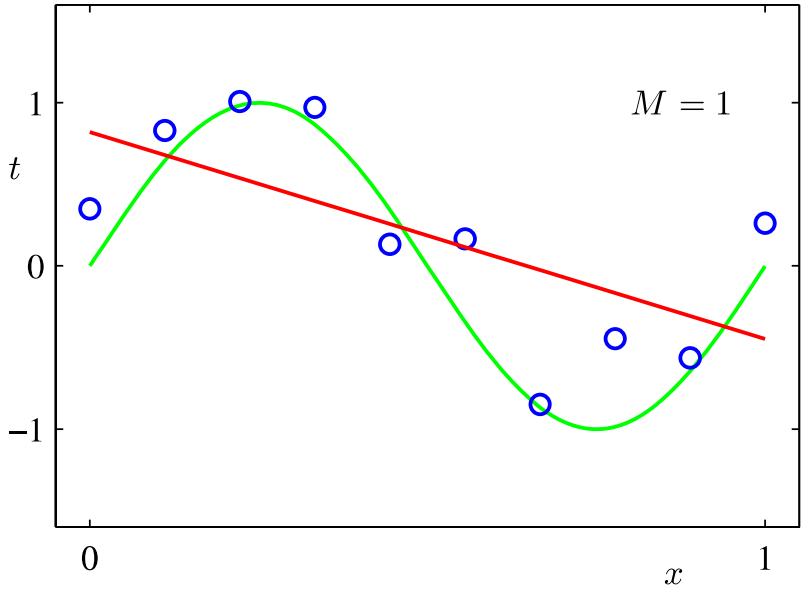


0-th order polynomial

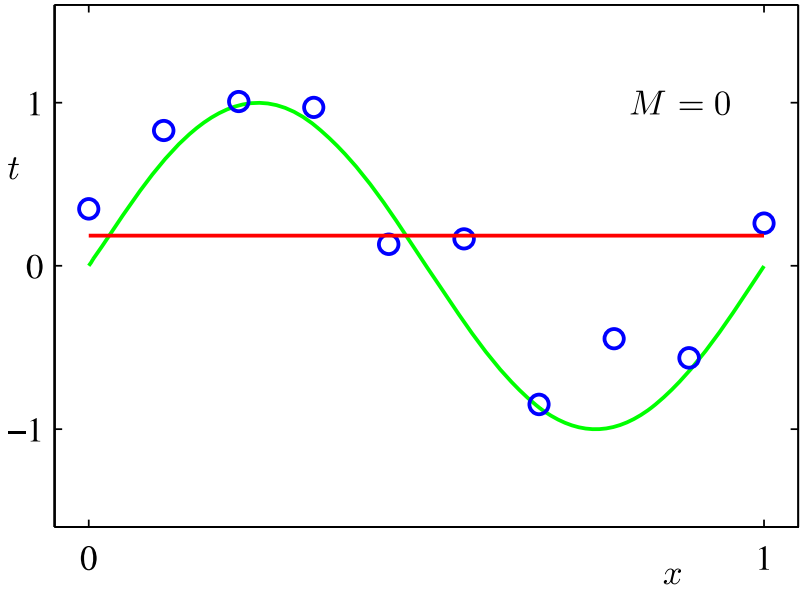


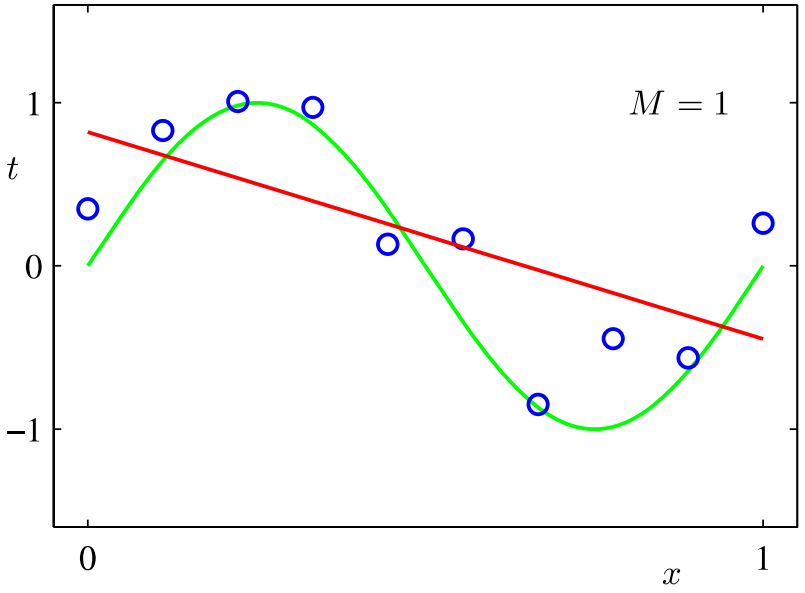
1-st order polynomial

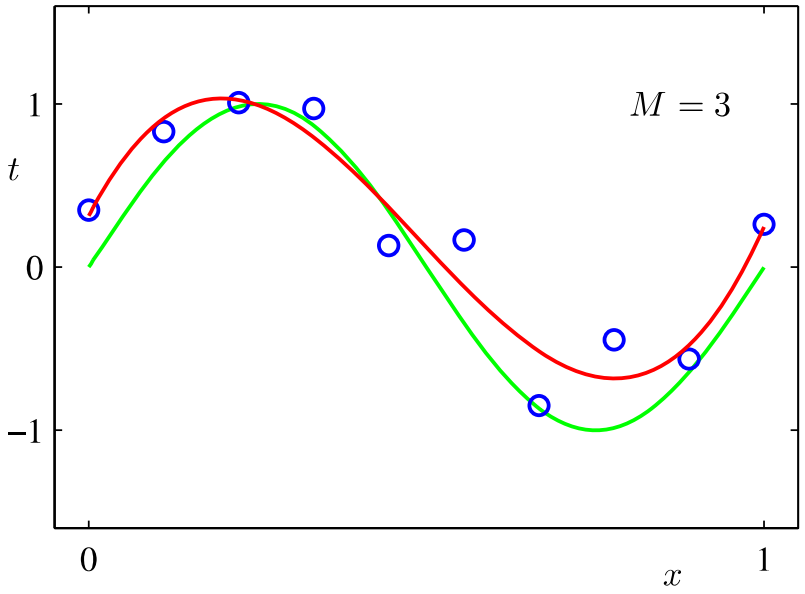




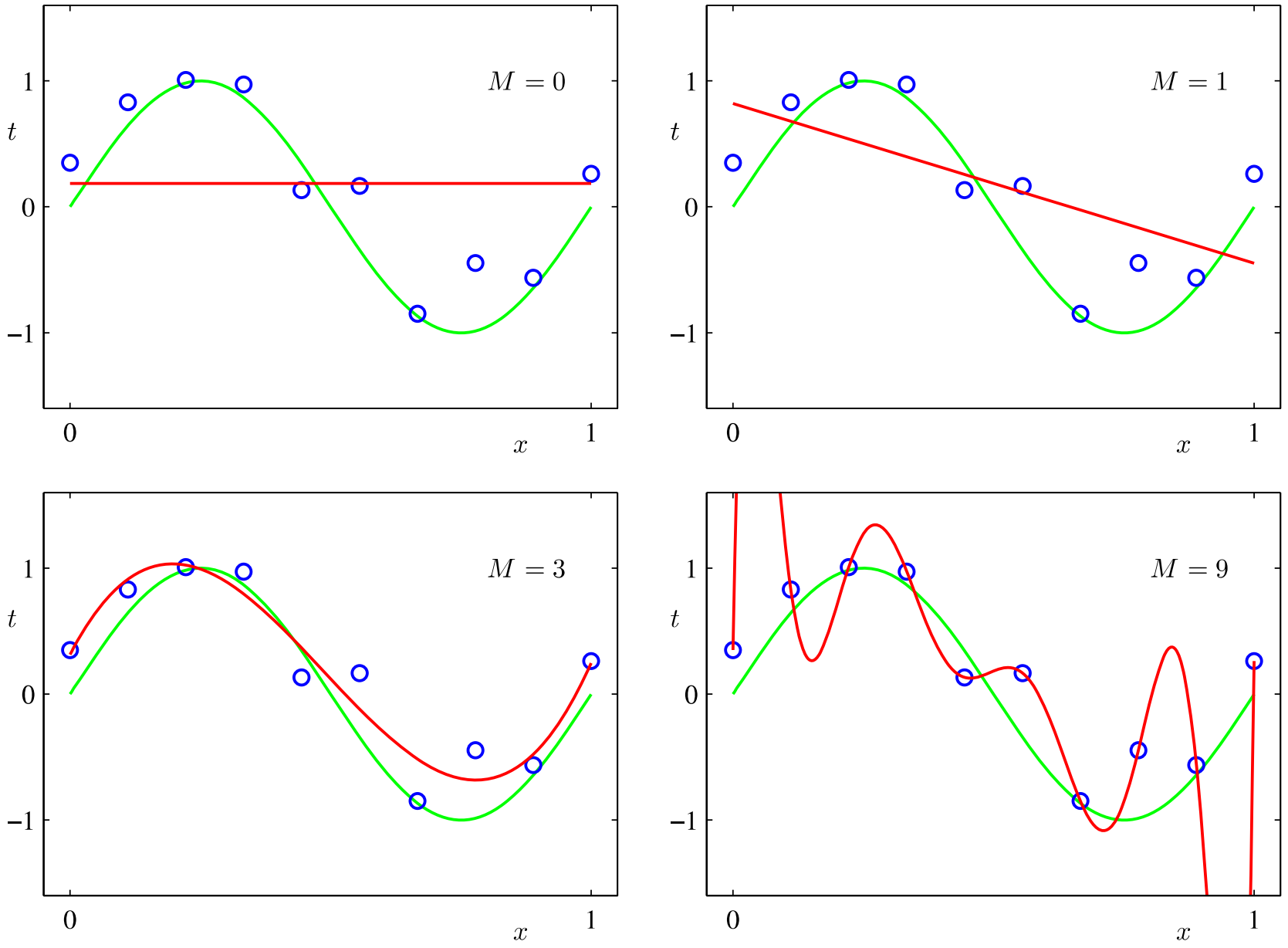
3-rd order polynomial



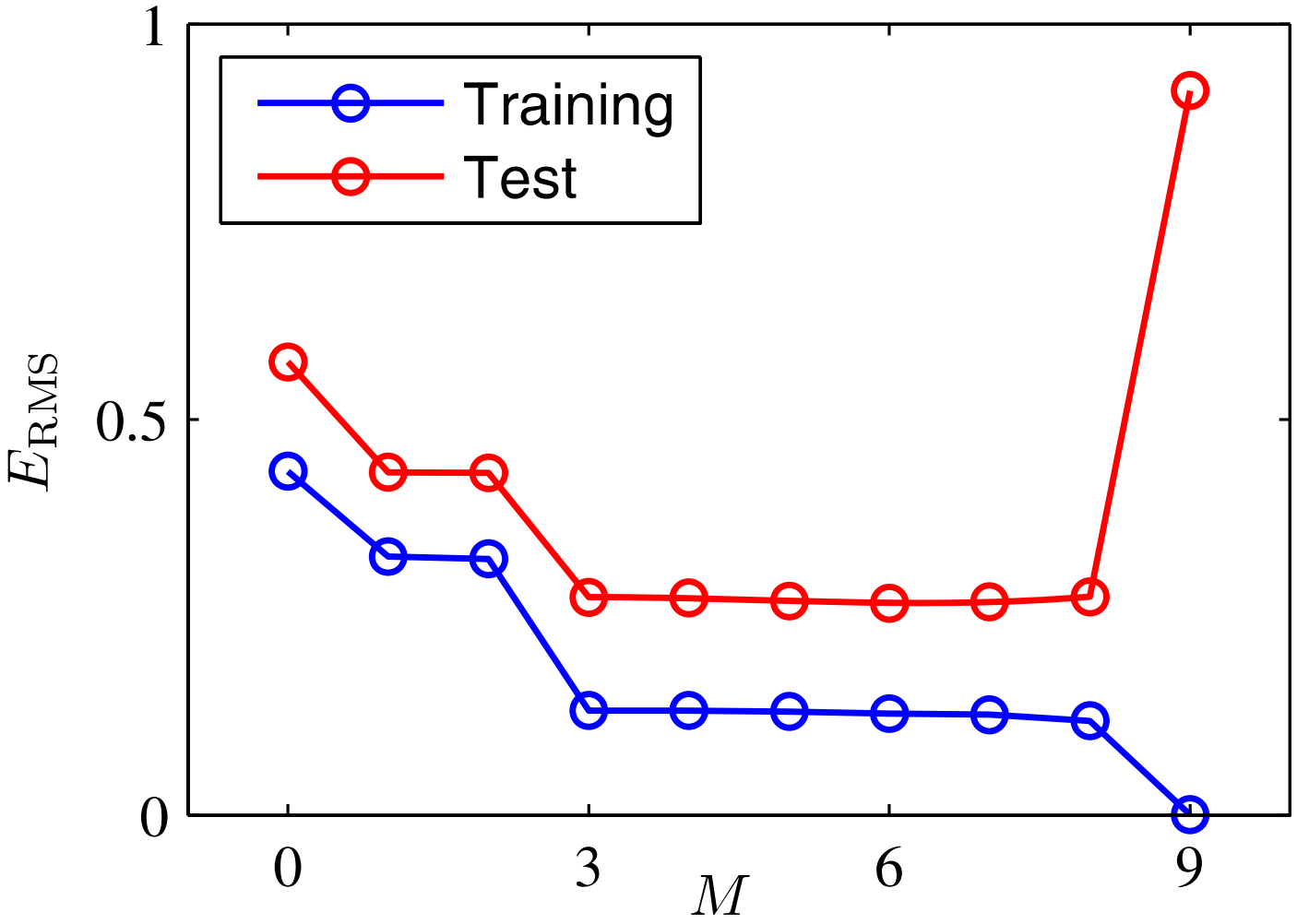




9-th order polynomial



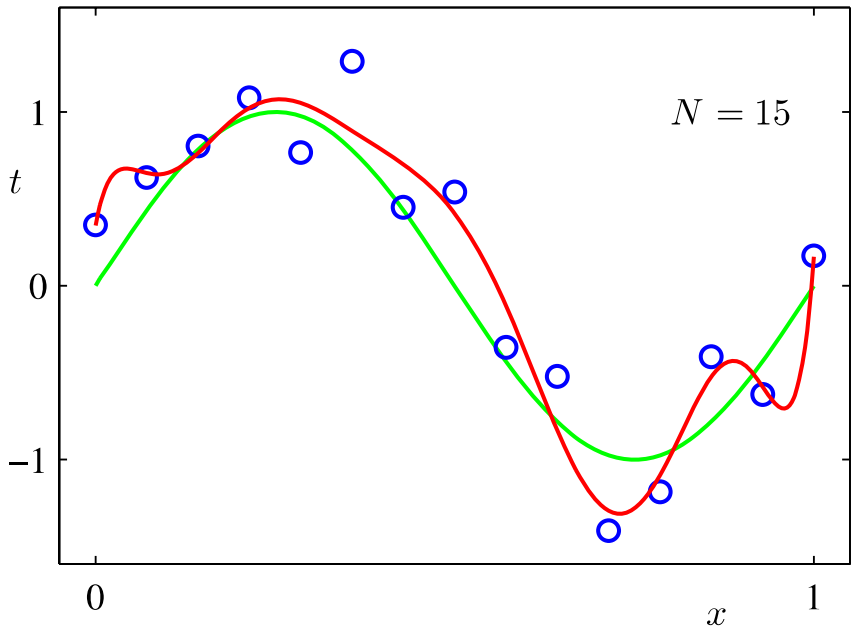
Over-fitting



$$E_{RMS} = \sqrt{2E(w^*)/N}$$

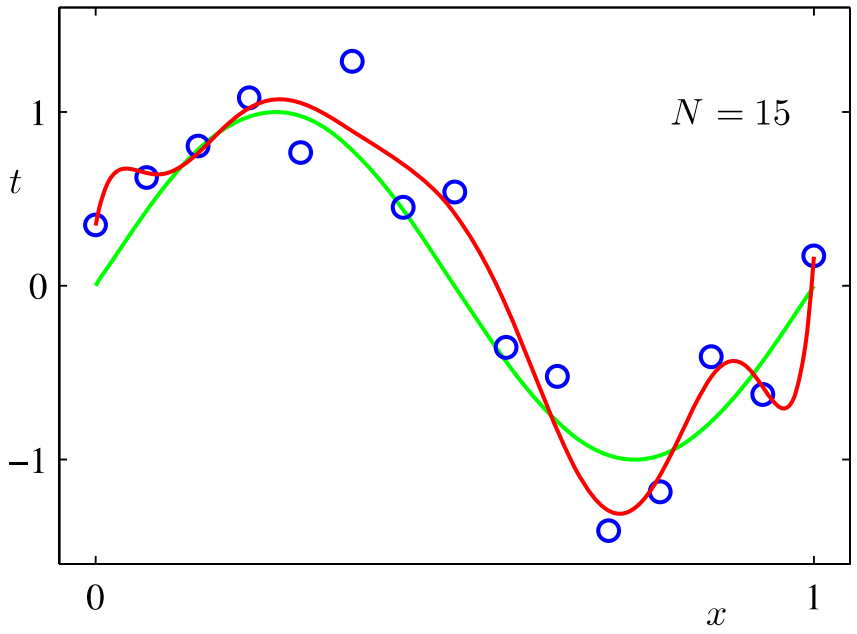
Data size N=15, M=9

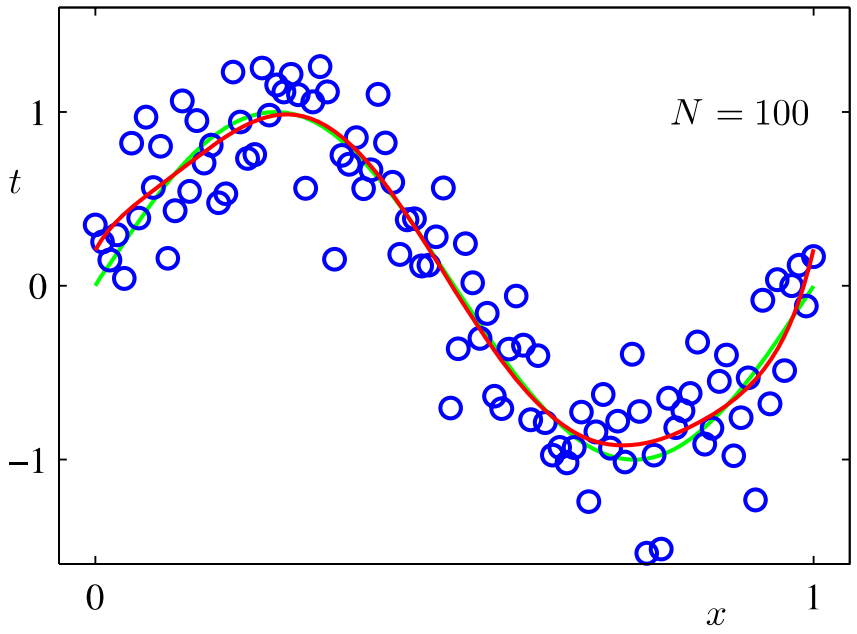
9th order polynomial



Data size N=100, M=9

9th order polynomial





Polynomial coefficients

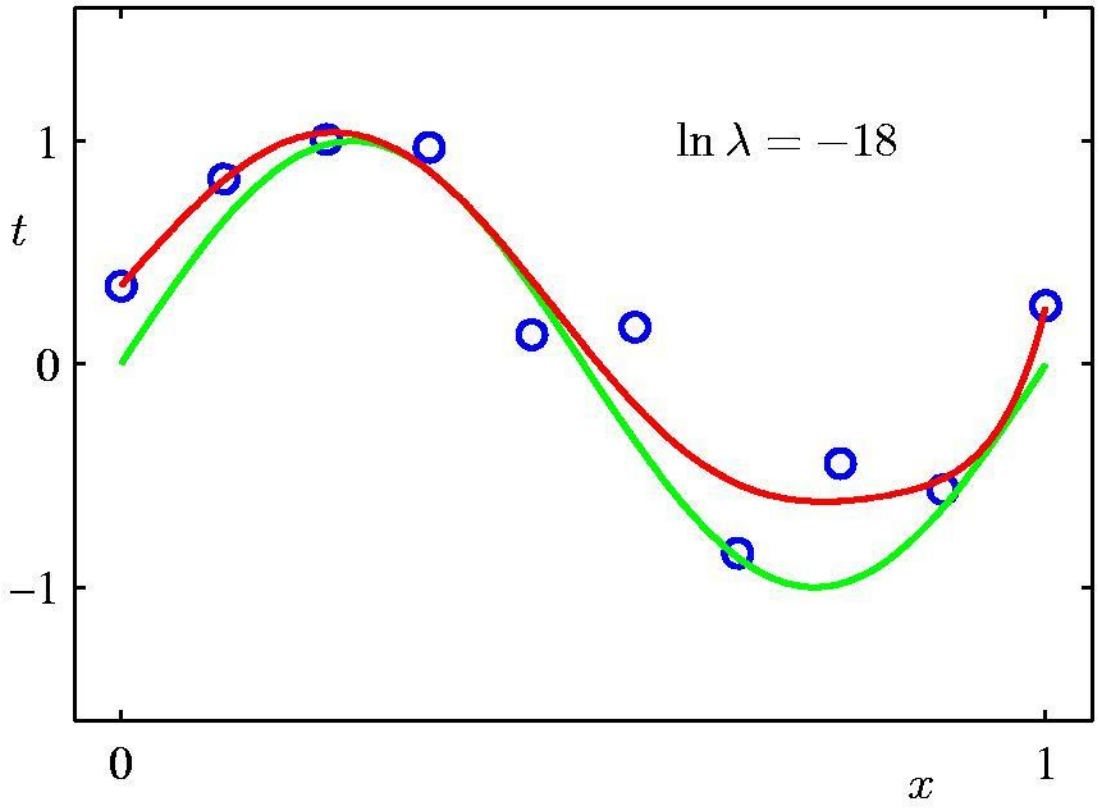
	$M = 0$	$M = 1$	$M = 6$	$M = 9$
$w_0^\star$	0.19	0.82	0.31	0.35
$w_1^\star$		-1.27	7.99	232.37
$w_2^\star$			-25.43	-5321.83
$w_3^\star$			17.37	48568.31
$w_4^\star$				-231639.30
$w_5^\star$				640042.26
$w_6^\star$				-1061800.52
$w_7^\star$				1042400.18
$w_8^\star$				-557682.99
$w_9^\star$				125201.43

Regularization

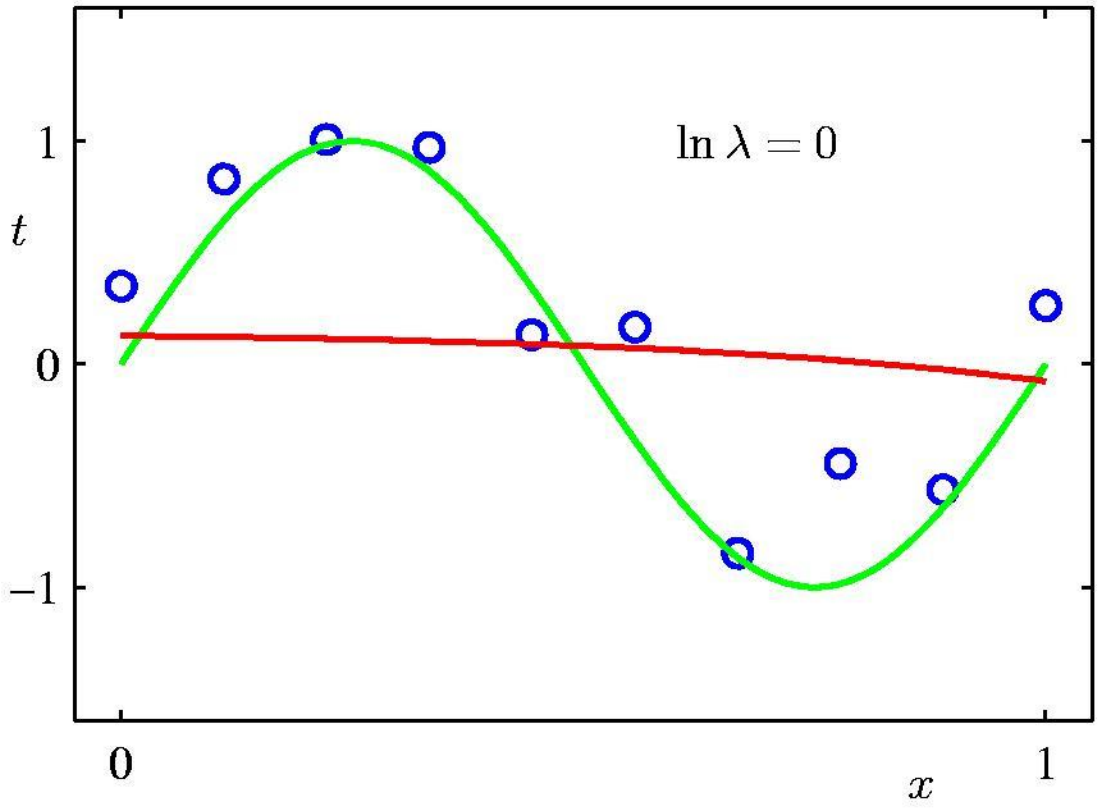
Penalize large coefficient values

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

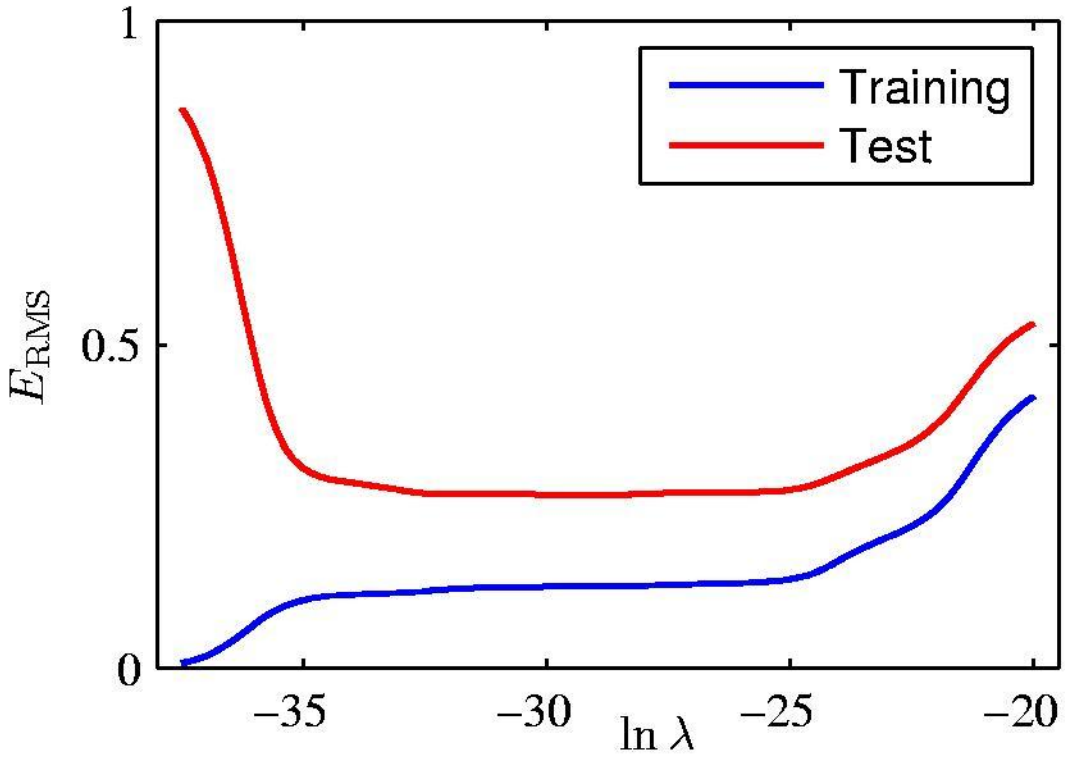
Regularization: $\ln \lambda = -18$



Regularization: $\ln \lambda = 0$



Regularization: E_{RMS} vs. $\ln \lambda$



Polynomial Coefficients

	$\ln \lambda = -\infty$	$\ln \lambda = -18$	$\ln \lambda = 0$
$w_0^\star$	0.35	0.35	0.13
$w_1^\star$	232.37	4.74	-0.05
$w_2^\star$	-5321.83	-0.77	-0.06
$w_3^\star$	48568.31	-31.97	-0.05
$w_4^\star$	-231639.30	-3.89	-0.03
$w_5^\star$	640042.26	55.28	-0.02
$w_6^\star$	-1061800.52	41.32	-0.01
$w_7^\star$	1042400.18	-45.95	-0.00
$w_8^\star$	-557682.99	-91.53	0.00
$w_9^\star$	125201.43	72.68	0.01

References

- Anselm Blumer,A. Ehrenfeucht, David Haussler, and Manfred K.Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM (JACM)*,Volume 36, Issue 4, 1989.
- Michael Kearns and Umesh Vazirani. *An Introduction to Computational Learning Theory*, MIT Press, 1994.
- Leslie G.Valiant. *A Theory of the Learnable*, Communications of the ACM 27(11):1134–1142 (1984).
- Martin Anthony, Peter L. Bartlett. *Neural network learning: theoretical foundations*. Cambridge University Press. 1999.
- Anselm Blumer,A. Ehrenfeucht, David Haussler, and Manfred K.Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM (JACM)*,Volume 36, Issue 4, 1989.
- A. Ehrenfeucht, David Haussler, Michael Kearns, Leslie Valiant. A general lower bound on the number of examples needed for learning. Proceedings of *1st COLT*. pp. 139-154, 1988.
- Koltchinskii,Vladimir and Panchenko, Dmitry. Empirical margin distributions and bounding the generalization error of combined classifiers.The Annals of Statistics, 30(1), 2002.
- Pascal Massart. Some applications of concentration inequalities to statistics. Annales de la Faculte des Sciences de Toulouse, IX:245–303, 2000.
- N. Sauer. On the density of families of sets. *Journal of Combinatorial Theory* (A), 13:145-147, 1972.
- Vladimir N. Vapnik. *Estimation of Dependences Based on Empirical Data*. Springer, 1982.
- Vladimir N. Vapnik. *The Nature of Statistical Learning Theory.* Springer, 1995.
- Vladimir N. Vapnik. *Statistical Learning Theory*.Wiley-Interscience, New York, 1998.
- Vladimir N. Vapnik and Alexey Chervonenkis. *Theory of Pattern Recognition*. Nauka, Moscow (in Russian). 1974.
- Vladimir N. Vapnik and Alexey Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Prob. and its Appl.*, vol. 16, no. 2, pp. 264-280, 1971.