EECS 545: Machine Learning

Lecture 3. Linear Regression (part 2)

Honglak Lee

1/17/2020

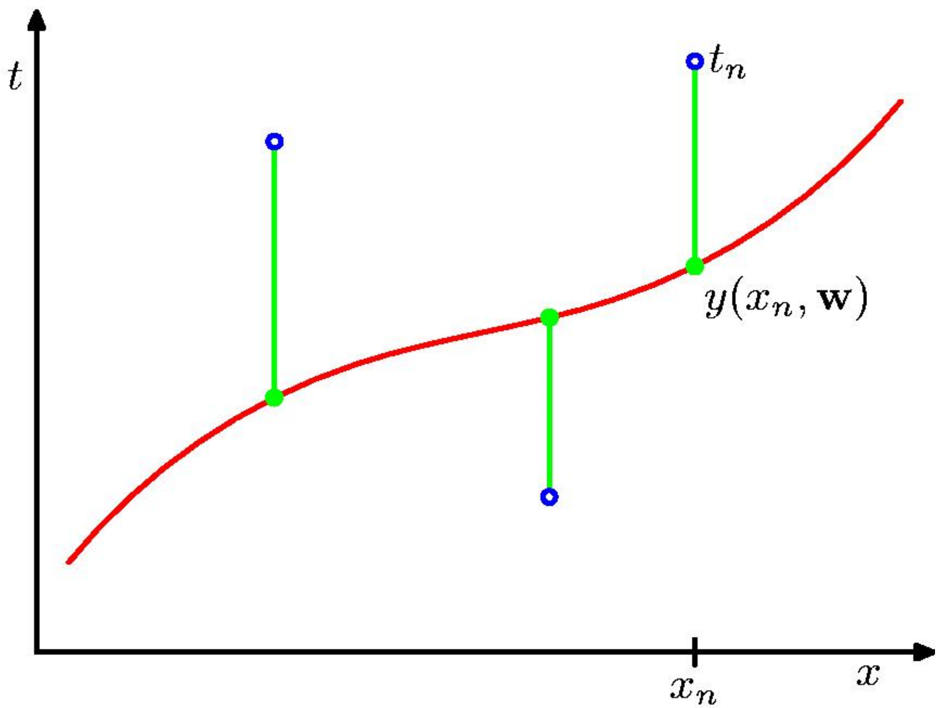
Announcements

- HW1 due: 1/28 (Tue) 11:55pm
- One more review session: Probability
- Office hours (Jan 20 -):
  Tue, 3:00-5:00pm, BBB2901
  Wed, 6:00-8:00pm, BBB3941
  Thu, 2:00-4:00pm, BBB2901
- Office hours today:
  Fri(Jan 17), 6:30-8:30pm, BBB1637

Outline

- Linear regression review
- Regularized linear regression
- Review on probability
- Maximum likelihood interpretation of linear regression
- Locally-weighted linear regression

Regression, sum of square error



Linear

$$h(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{j=1}^{M-1} w_j \phi_j(\mathbf{x})$$

$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \left\{ h(\mathbf{x}^{(n)}, \mathbf{w}) - y^{(n)} \right\}^2$$

We want to find w that minimizes E(**w**) over the training data.

Least squares problem

- Objective function

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}(\sum_{j=0}^{M-1} w_j\phi_j(\mathbf{x}^{(n)}) - y^{(n)})^2$$

M: dimension of feature, N: number of data

$\phi_j(\mathbf{x}^{(n)})$: j-th feature of data

- Two ways to find w that minimizes E(w)

1. Gradient descent
2. Closed-form solution

Least squares problem

- Objective function

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}(\sum_{j=0}^{M-1} w_j\phi_j(\mathbf{x}^{(n)}) - y^{(n)})^2$$

1. Gradient Descent

$$\frac{\partial E(\mathbf{w})}{\partial w_j} = \sum_{n=1}^{N}(\sum_{j=0}^{M-1} w_j\phi_j(\mathbf{x}^{(n)}) - y^{(n)})\phi_j(\mathbf{x}^{(n)})$$

$$\nabla_{\mathbf{w}}E(\mathbf{w}) = \sum_{n=1}^{N}(\mathbf{w}^T\phi(\mathbf{x}^{(n)}) - y^{(n)})\phi(\mathbf{x}^{(n)})$$

$$\mathbf{w} := \mathbf{w} - \eta\nabla_{\mathbf{w}}E(\mathbf{w})$$

2. Closed-form solution

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}(\sum_{j=0}^{M-1} w_j\phi_j(\mathbf{x}^{(n)}) - y^{(n)})^2$$

$$= \frac{1}{2}\mathbf{w}^T\Phi^T\Phi\mathbf{w} - \mathbf{w}^T\Phi^T\mathbf{y} + \frac{1}{2}\mathbf{y}^T\mathbf{y}$$

Recap

w: M by 1 $\mathbf{w} = [w_0, w_1, ..., w_{M-1}]^T$

y : N by 1 $\mathbf{y} = [y^{(1)}, y^{(2)}, ..., y^{(N)}]^T$

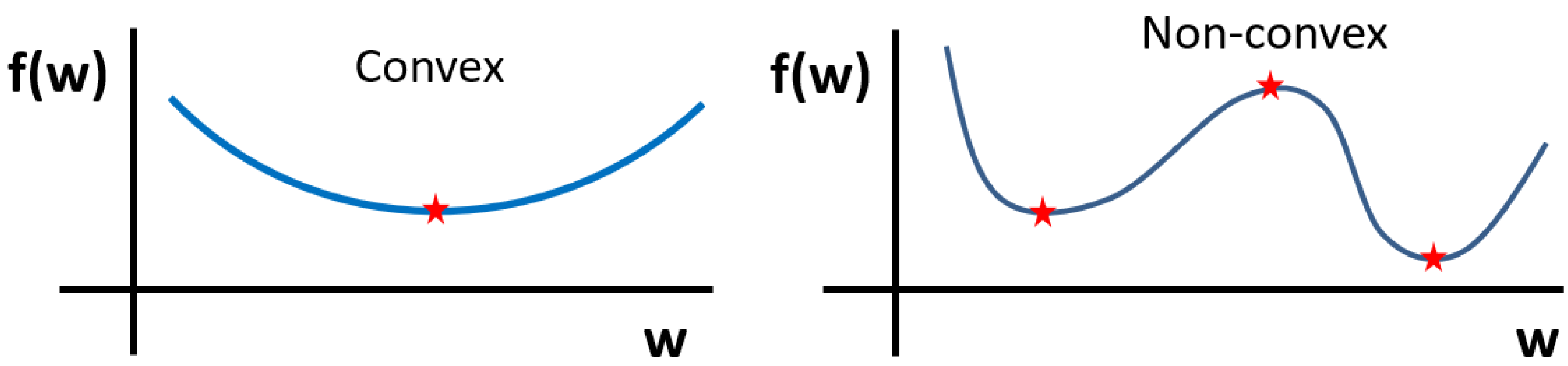
Φ: N by M

$$\Phi = \begin{pmatrix} \phi_0(\mathbf{x}^{(1)}) & \phi_1(\mathbf{x}^{(1)}) & \ldots & \phi_{M-1}(\mathbf{x}^{(1)}) \\ \phi_0(\mathbf{x}^{(2)}) & \phi_1(\mathbf{x}^{(2)}) & \ldots & \phi_{M-1}(\mathbf{x}^{(2)}) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(\mathbf{x}^{(N)}) & \phi_1(\mathbf{x}^{(N)}) & \ldots & \phi_{M-1}(\mathbf{x}^{(N)}) \end{pmatrix}$$

Useful trick: Matrix Calculus

- Compute gradient and set gradient to 0
  (condition for optimal solution)

$$\nabla_{\mathbf{w}} E(\mathbf{w}) = 0$$



- Need to compute the first derivative in matrix form

Gradient via matrix calculus

- Compute gradient and set to zero

$$\begin{aligned}\nabla_{\mathbf{w}} E(\mathbf{w}) &= \nabla_{\mathbf{w}} \left(\frac{1}{2}\mathbf{w}^T \Phi^T \Phi \mathbf{w} - \mathbf{w}^T \Phi^T \mathbf{y} + \frac{1}{2}\mathbf{y}^T \mathbf{y} \right) \\ &= \Phi^T \Phi \mathbf{w} - \Phi^T \mathbf{y} \\ &= 0\end{aligned}$$

- Solve the resulting equation (normal equation)

$$\Phi^T \Phi \mathbf{w} = \Phi^T \mathbf{y}$$

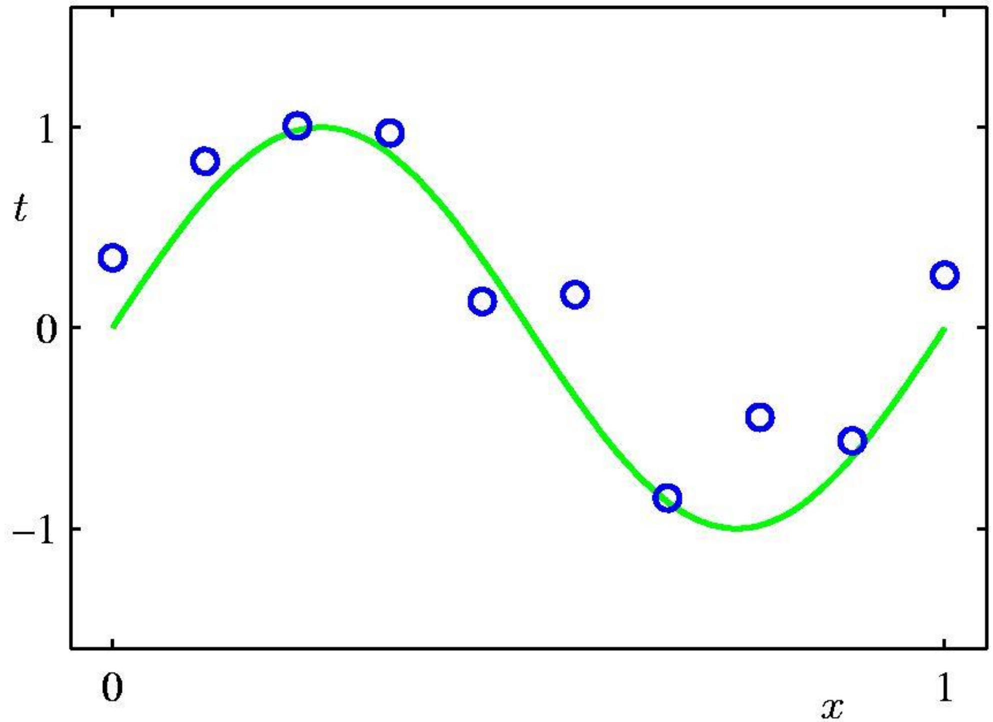
$$\mathbf{w}_{ML} = (\Phi^T \Phi)^{-1} \Phi^T \mathbf{y}$$

This is the *Moore-Penrose pseudo-inverse*: $\mathbf{\Phi}^{\dagger} = (\mathbf{\Phi}^T \mathbf{\Phi})^{-1} \mathbf{\Phi}^T$

applied to: $\Phi \mathbf{w} \approx \mathbf{y}$

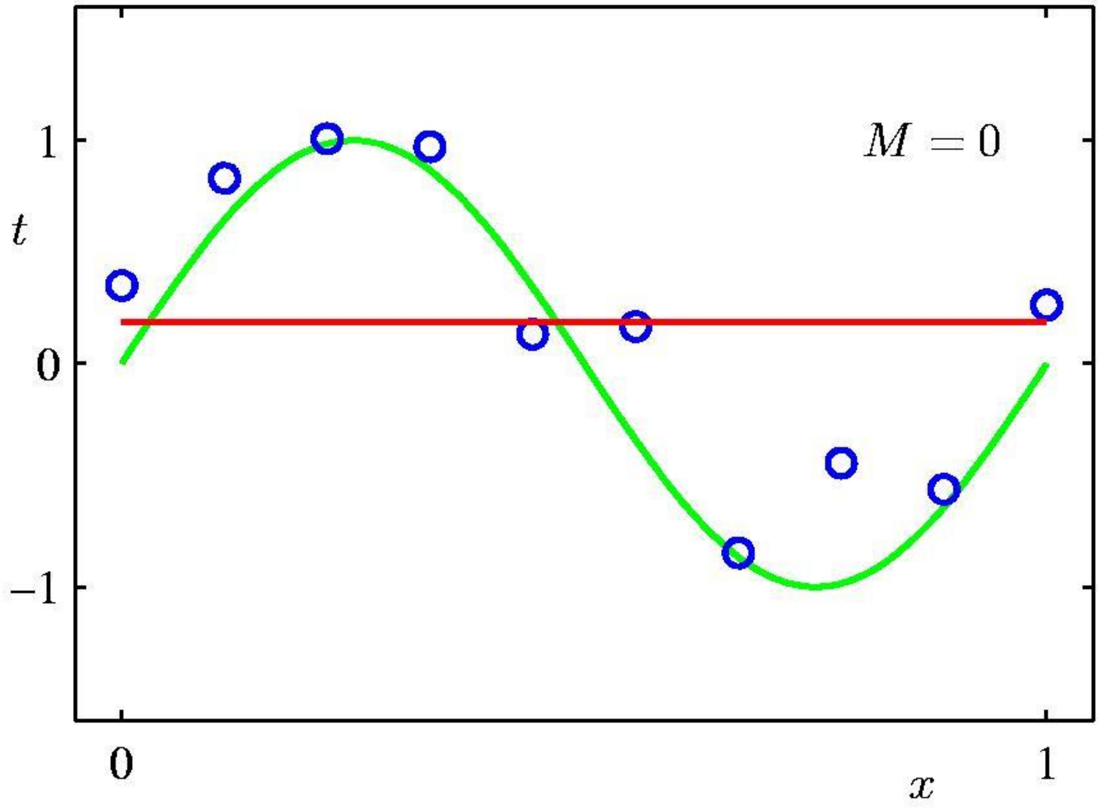
Back to curve-fitting examples

Polynomial Curve Fitting

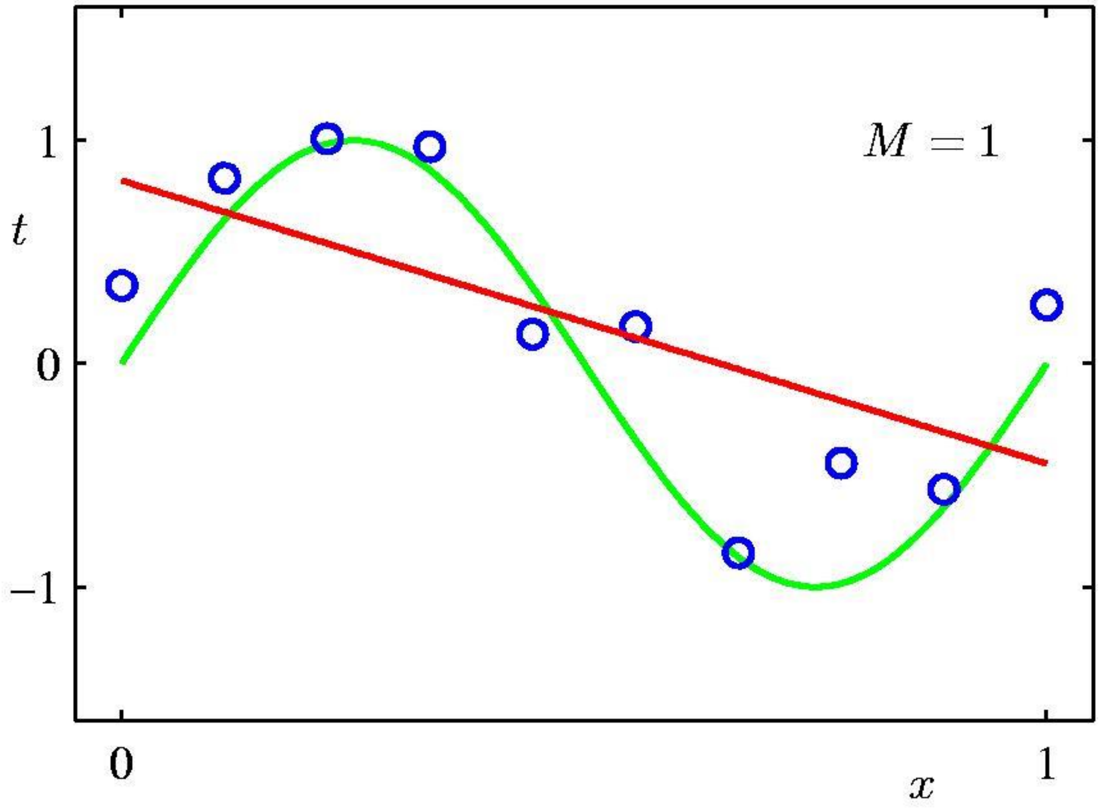


$$h(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_{M-1} x^{M-1} = \sum_{j=0}^{M-1} w_j x^j$$

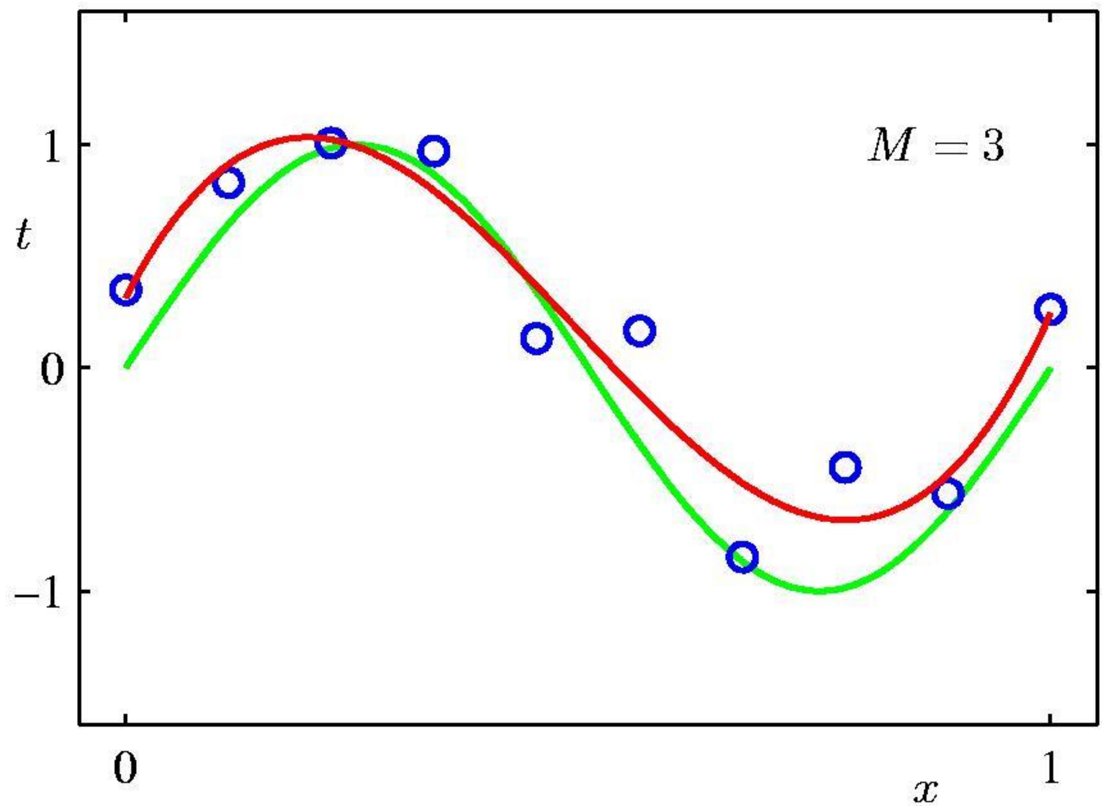
0^{th} Order Polynomial



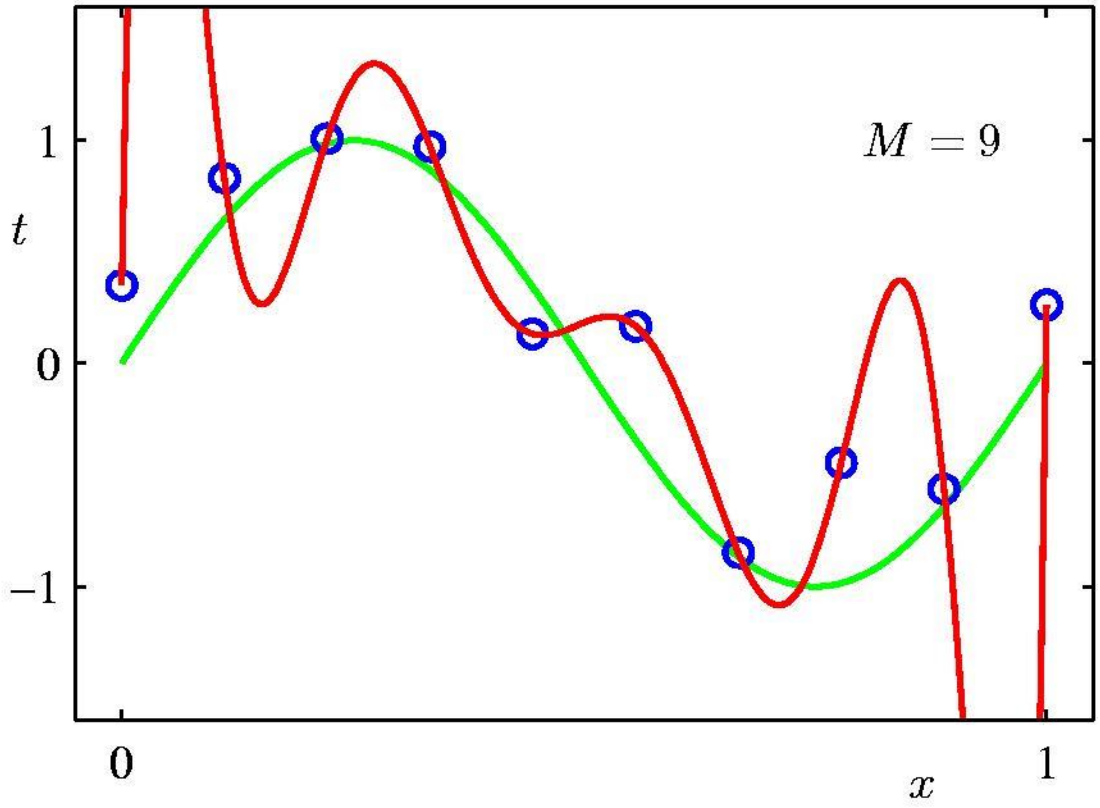
1st Order Polynomial



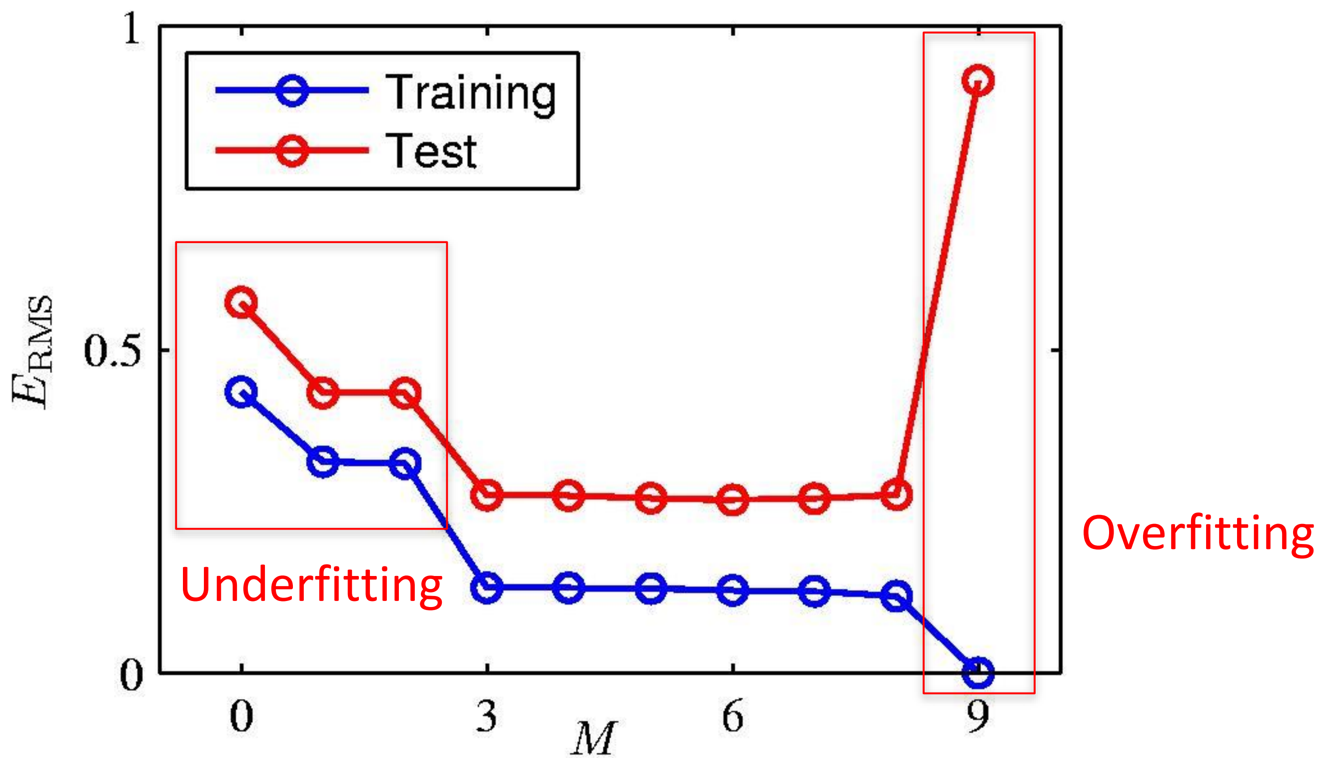
3rd Order Polynomial



9^{th} Order Polynomial



Over-fitting

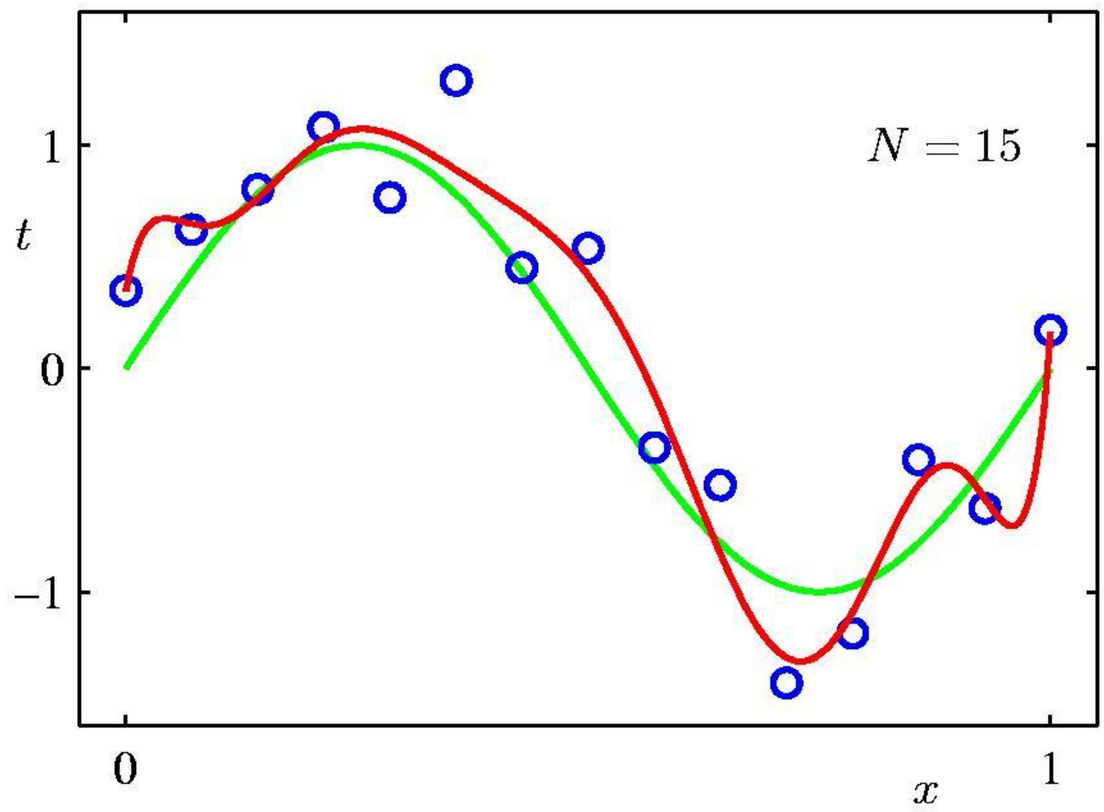


Root-Mean-Square (RMS) Error: $E_{\mathrm{RMS}} = \sqrt{2E(\mathbf{w}^\star)/N}$

Q: How do we resolve the over-fitting problem?

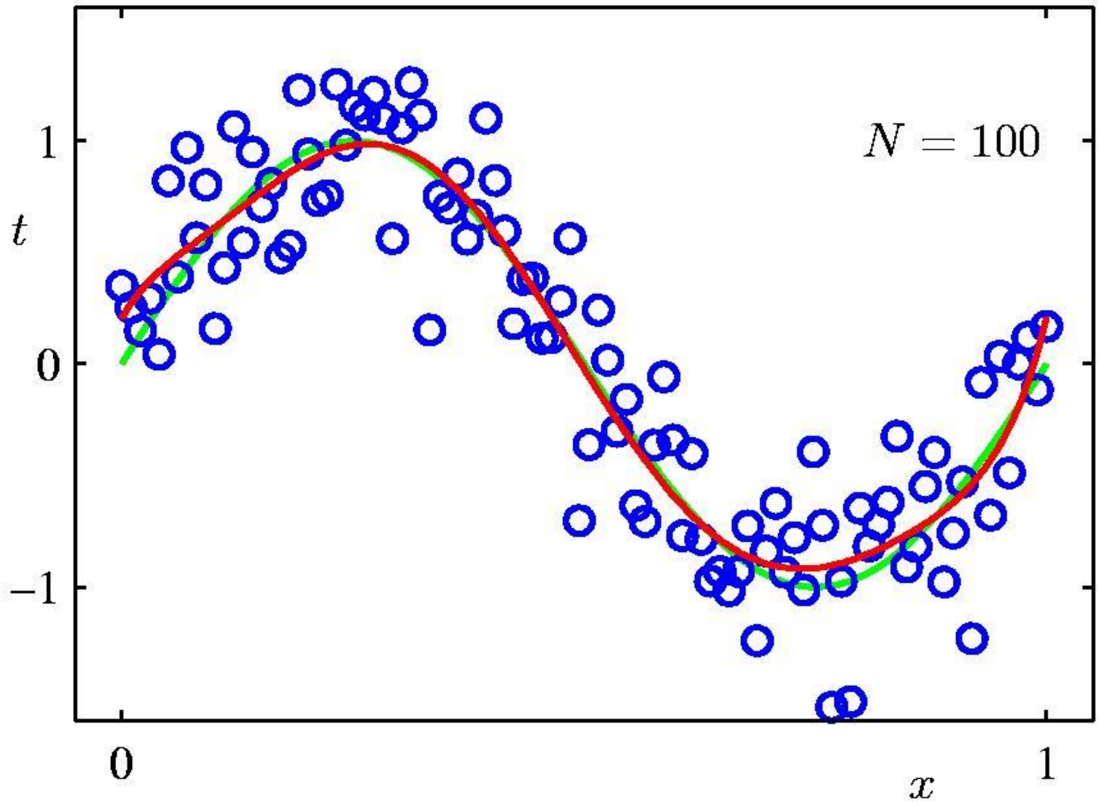
Data Set Size: $N = 15$

9th Order Polynomial



Data Set Size: $N = 100$

9th Order Polynomial



Increasing data set size can help

Q. How do we choose the degree of polynomial?

Rule of thumb

- If you have a small number of data, then use low order polynomial (small number of features).
  - Otherwise, your model will overfit

- As you obtain more data, you can gradually increase the order of the polynomial (more features).
  - However, your model is still limited by the finite amount of the data available (i.e., the optimal model for finite data cannot be infinite dimensional polynomial).

- Controlling model complexity: **regularization**

Regularized Linear Regression

Back to 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

Underfitting

Good

Overfitting;
Coefficients are large!

$$h(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_{M-1} x^{M-1}$$

Regularized Least Squares (1)

- Consider the error function:

$$E_D(\mathbf{w}) + \lambda E_W(\mathbf{w})$$

Data term + Regularization term

λ is called the regularization coefficient.

- With the sum-of-squares error function and a quadratic regularizer, we get

Penalize large coefficient values

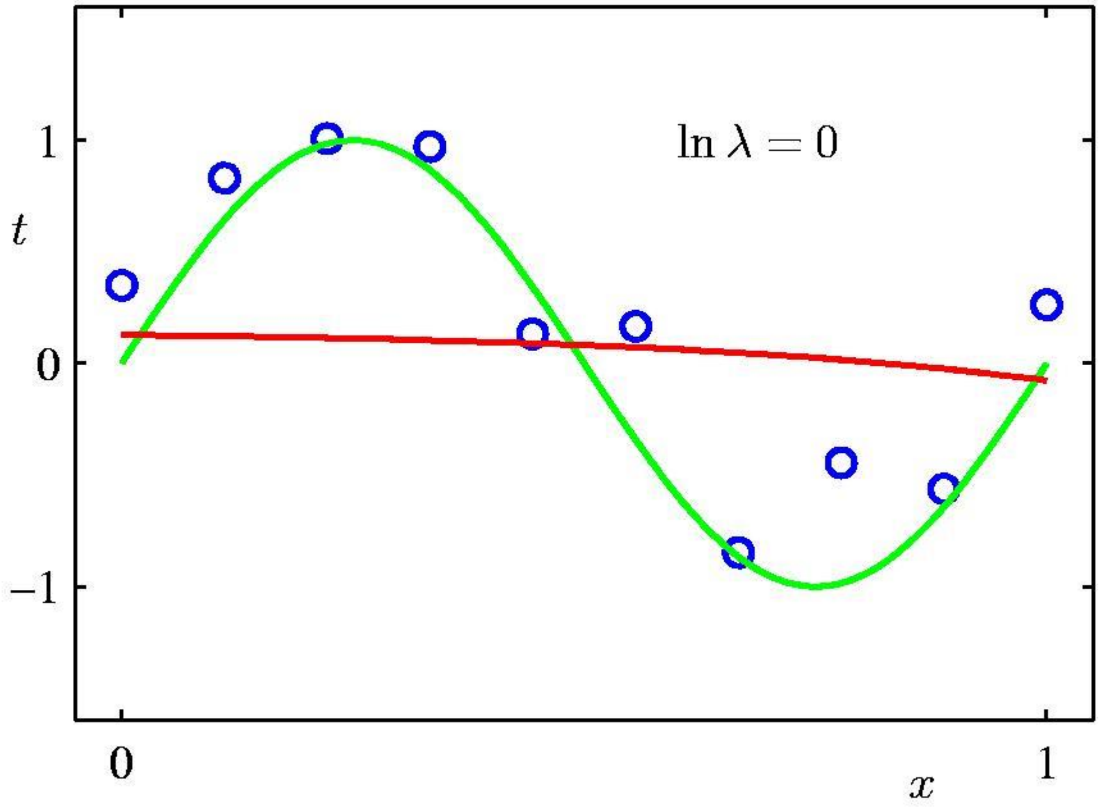
$$\widetilde{E}(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}(\mathbf{w}^T\phi(\mathbf{x}^{(n)}) - y^{(n)})^2 + \frac{\lambda}{2}||\mathbf{w}||^2$$

New objective function

Definition (L2): $||\mathbf{w}||_2^2 = \sum_{j=0}^{M-1} w_j^2$

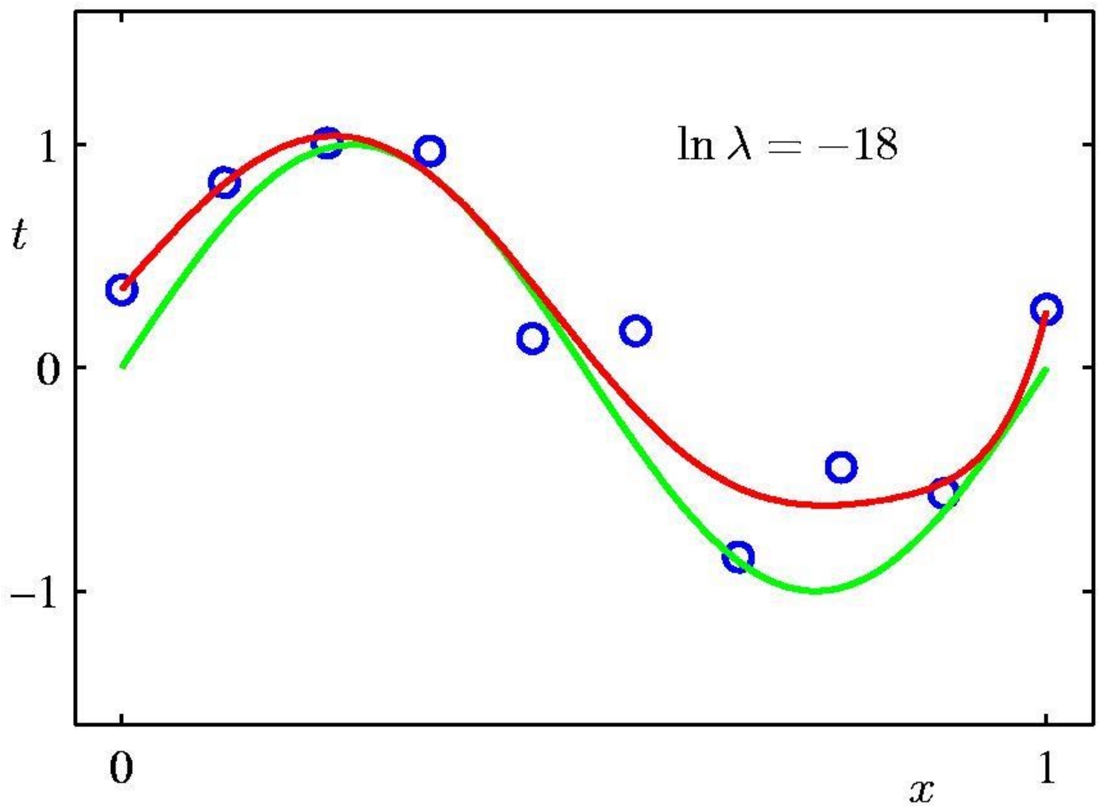
- Effect of λ

L2 Regularization: $\ln \lambda = 0$



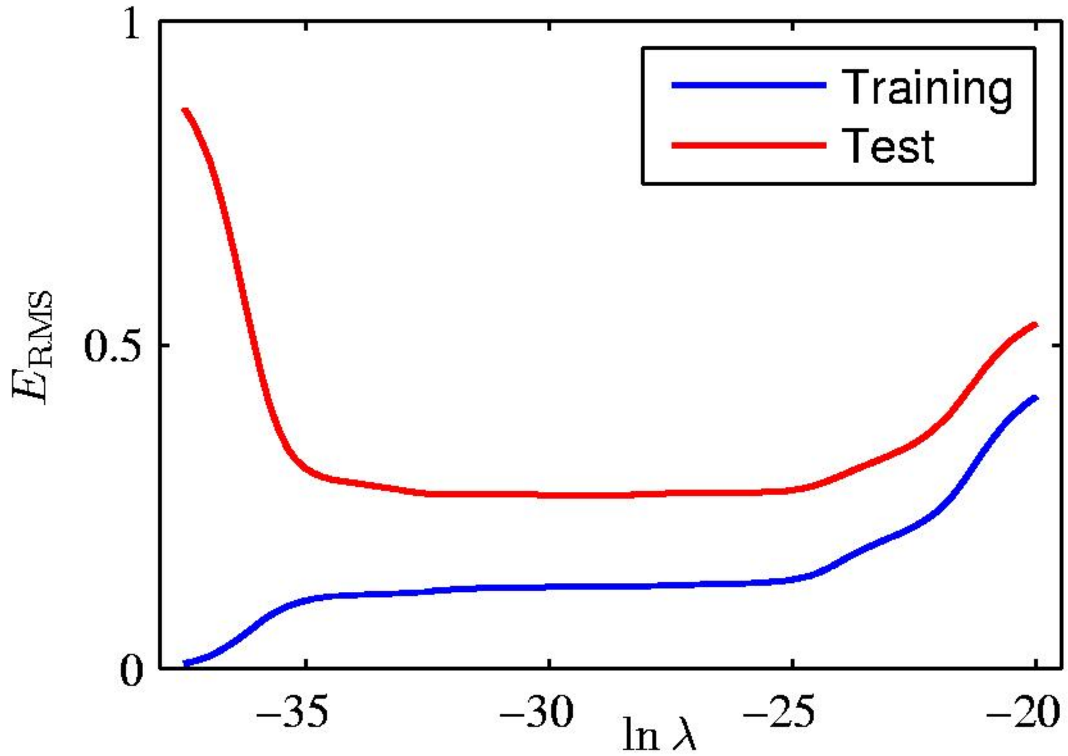
$M = 9$ $$\widetilde{E}(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}(\mathbf{w}^T\phi(\mathbf{x}^{(n)}) - y^{(n)})^2 + \frac{\lambda}{2}||\mathbf{w}||^2$$

L2 Regularization: $\ln\lambda = -18$



$M = 9$ $$\widetilde{E}(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}(\mathbf{w}^T\phi(\mathbf{x}^{(n)}) - y^{(n)})^2 + \frac{\lambda}{2}||\mathbf{w}||^2$$

L2 Regularization: E_{RMS} vs. $\ln \lambda$



Larger regualrization

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

NOTE: For simplicity of presentation, we divided the data into training set and test set. However, it's **not** legitimate to find the optimal hyperparameter based on the test set. We will talk about legitimate ways of doing this when we cover model selection and cross-validation.

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

Regularized Least Squares (1)

- Consider the error function:

$$E_D(\mathbf{w}) + \lambda E_W(\mathbf{w})$$

Data term + Regularization term

λ is called the regularization coefficient.

- With the sum-of-squares error function and a quadratic regularizer, we get

Penalize large coefficient values

$$\widetilde{E}(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}(\mathbf{w}^T\phi(\mathbf{x}^{(n)}) - y^{(n)})^2 + \frac{\lambda}{2}||\mathbf{w}||^2$$

- Closed-form solution:

$$\mathbf{w}_{ML} = (\lambda\mathbf{I} + \Phi^T\Phi)^{-1}\Phi^T\mathbf{y}$$

Derivation

Objective function

$$\widetilde{E}(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}(\mathbf{w}^T\phi(\mathbf{x}^{(n)}) - y^{(n)})^2 + \frac{\lambda}{2}||\mathbf{w}||^2$$

$$= \frac{1}{2}\mathbf{w}^T\Phi^T\Phi\mathbf{w} - \mathbf{w}^T\Phi^T\mathbf{y} + \frac{1}{2}\mathbf{y}^T\mathbf{y} + \frac{\lambda}{2}\mathbf{w}^T\mathbf{w}$$

Compute gradient and set it zero:

$$\begin{aligned}\nabla_{\mathbf{w}}E(\mathbf{w}) &= \nabla_{\mathbf{w}}[\frac{1}{2}\mathbf{w}^T\Phi^T\Phi\mathbf{w} - \mathbf{w}^T\Phi^T\mathbf{y} + \frac{1}{2}\mathbf{y}^T\mathbf{y} + \frac{\lambda}{2}\mathbf{w}^T\mathbf{w}] \\ &= \Phi^T\Phi\mathbf{w} - \Phi^T\mathbf{y} + \lambda\mathbf{w} \\ &= (\lambda\mathbf{I} + \Phi^T\Phi)\mathbf{w} - \Phi^T\mathbf{y} \\ &= 0\end{aligned}$$

$$\mathbf{w}_{ML} = (\Phi^T\Phi)^{-1}\Phi^T\mathbf{y}$$

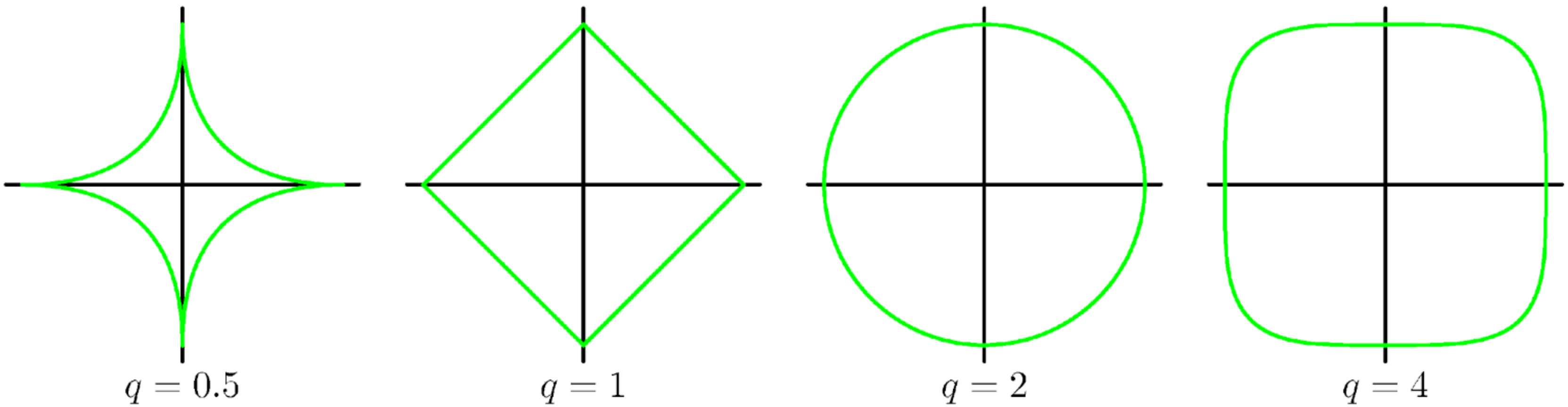
vs Ordinary Least Square

Therefore, we get: $\mathbf{w}_{ML} = (\lambda\mathbf{I} + \Phi^T\Phi)^{-1}\Phi^T\mathbf{y}$

Regularized Least Squares (2)

- With a more general regularizer, we have

$$\frac{1}{2}\sum_{n=1}^{N}(\mathbf{w}^T\phi(\mathbf{x}^{(n)}) - y^{(n)})^2 + \frac{\lambda}{2}\sum_{j=1}^{M}|w_j|^q$$



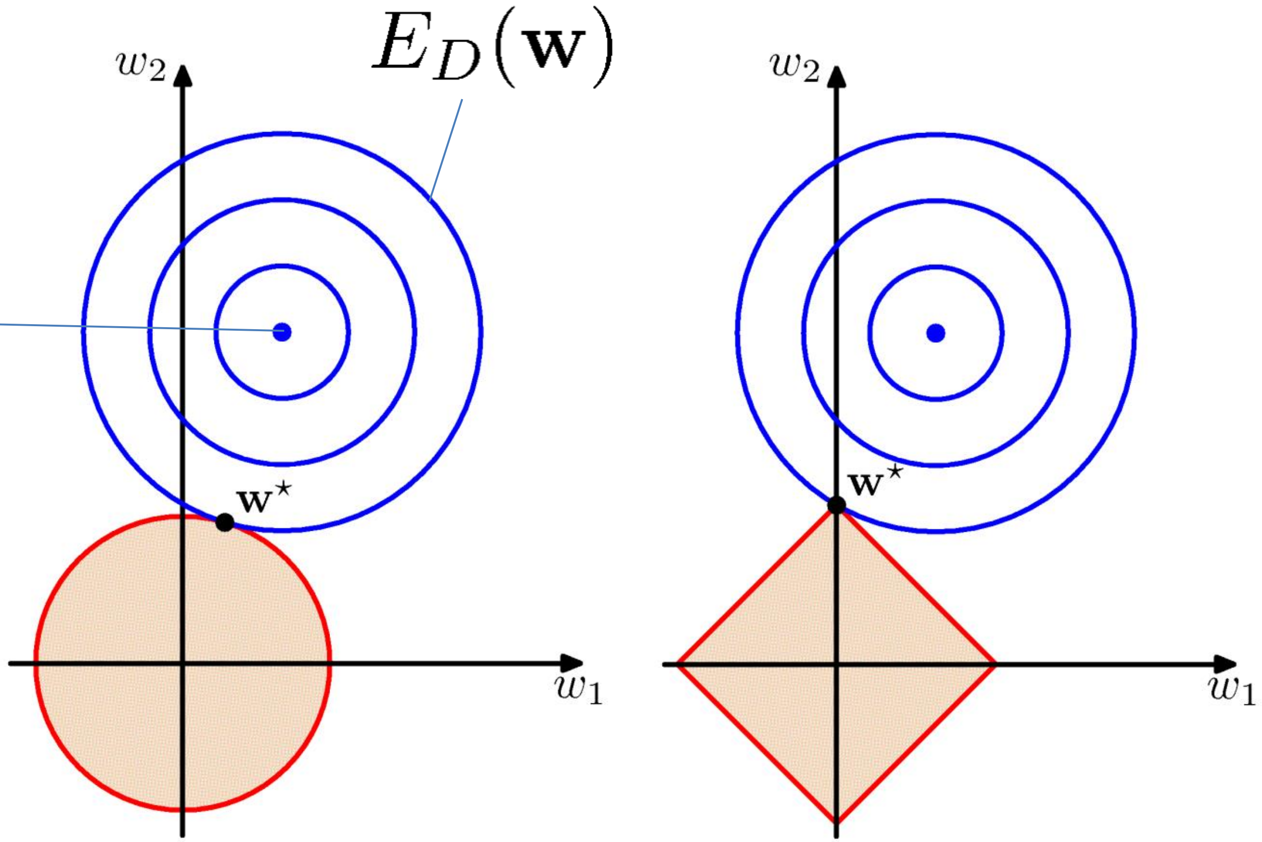
Lasso
"L1 regularization"

Quadratic
"L2 regularization"

Regularized Least Squares (3)

- Lasso tends to generate sparser solutions than a quadratic regularizer.

Ordinary least square solution



Regularization as constraints:

$$\sum_j w_j^2 \leq C \qquad \sum_j |w_j| \leq C$$

Summary: Regularized Linear Regression

- Simple modification of linear regression
- Regularization controls the tradeoff between “fitting error” and “complexity”
  - Small regularization results in complex models (but with risk of overfitting)
  - Large regularization results in simple models (but with risk of underfitting)

- It is important to find an optimal regularization that balances between the two.

Maximum Likelihood interpretation of least squares regression

Review on probability

Probability: Terminology

- Experiment: Procedure that yields an outcome
  - E.g., Tossing a coin three times:
    - Outcome: HHH in one trial, HTH in another trial, etc.
- Sample space: Set of all possible outcomes in the experiment, denoted as Ω (or S)
  - E.g., for the above example:
    - Ω = {HHH, HHT, HTH, HTT, THH, THT, TTH, TTT}
- Event: subset of the sample space Ω (i.e., an event is a set consisting of individual outcomes)
  - Event space: Collection of all events, called $\mathcal{F}$ (aka σ-algebra)
  - E.g., Event that # of heads is an even number.
    - E = {HHT, HTH, THH, TTT}
- Probability measure: function (mapping) from events to probability levels. I.e., $P: \mathcal{F} \rightarrow [0,1]$ (see next slide)
  - Probability that # of heads is an even number: 4/8 = 1/2.
- Probability space: $(\Omega, \mathcal{F}, P)$

Law of Total Probability

- $P(A) \geq 0, \forall A \in \mathcal{F}$
- $P(\Omega) = 1$

- Law of total probability

$$P(A) = P(A \cap B) + P(A \cap B^C)$$

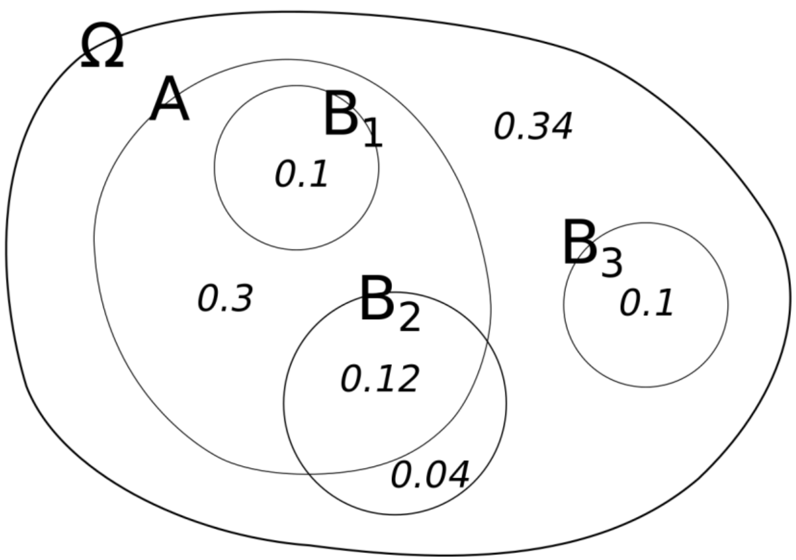
$$P(A) = \sum_i P(A \cap B_i)$$ Discrete B_i

$$P(A) = \int P(A \cap B_i) dB_i$$ Continuous B_i

Conditional Probability

For events $A, B \in \mathcal{F}$ with $P(B) > 0$, we may write the **conditional probability of A given B**:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$



From Wikipedia

$P(A|B_1) = 1$

$P(A|B_2) = 0.12 \div (0.12 + 0.04) = 0.75$

$\mathrm{P}(A|B_3) = 0$ (disjoint)

P(A) (The unconditional probability)

= 0.30 + 0.10 + 0.12 = 0.52

Bayes' Rule

Using the chain rule we may see:

$$P(A|B)P(B) = P(A \cap B) = P(B|A)P(A)$$

Rearranging this yields **Bayes' rule**:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

Often this is written as:

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum_i P(A|B_i)P(B_i)}$$

Where B_i are a partition of Ω (note the bottom is just the law of total probability).

Likelihood Functions

- Why is Bayes' so useful in learning? Allows us to compute the posterior of w given data D:

$$p(w|D) = \frac{p(D|w)p(w)}{p(D)}$$

Prior

Posterior

- Bayes' rule in words: posterior ∝ likelihood × prior

$$p(\mathbf{w}|D) \propto p(D|\mathbf{w})p(\mathbf{w})$$

- The likelihood function, $p(D|w)$, is evaluated for observed data D as a function of w. It expresses how probable the observed data set is for various parameter settings w.

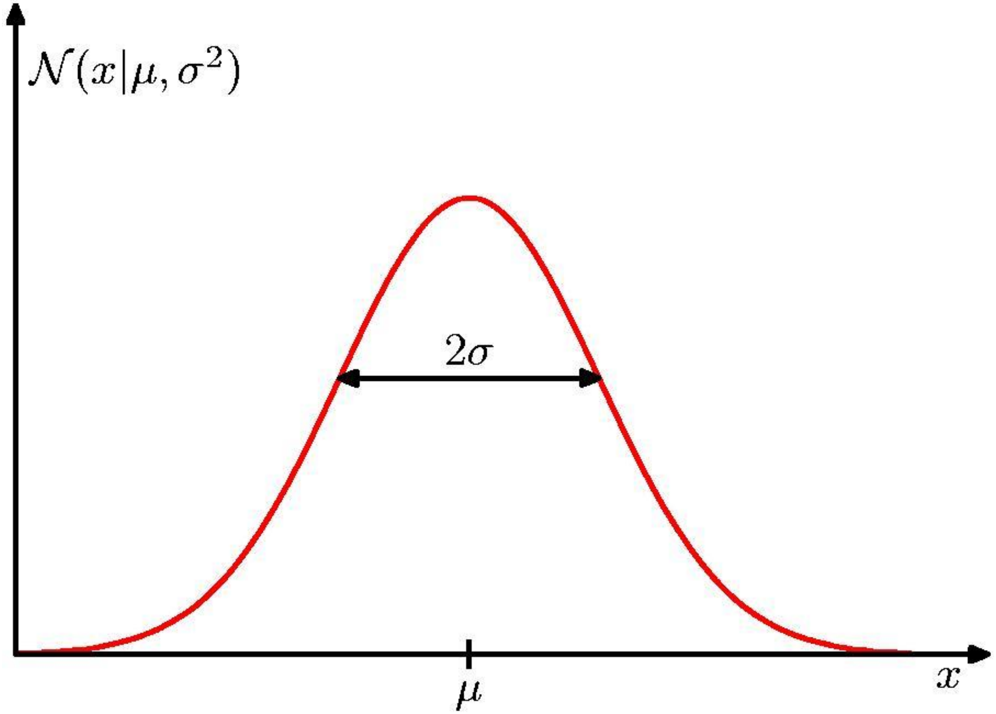
Maximum Likelihood Estimation (MLE)

- Maximum likelihood:
  - choose parameter setting w that maximizes likelihood function $p(D|w)$.
  - Choose the value of w that maximizes the probability of observed data.

- Cf. MAP (Maximum a posteriori) estimation
  - Equivalent to maximizing $P(w|D) \propto P(D|w)P(w)$
  - Can compute this using Bayes rule!
  - This will be covered in later lectures

The Gaussian Distribution

$$\mathcal{N}\left(x|\mu,\sigma^2\right) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

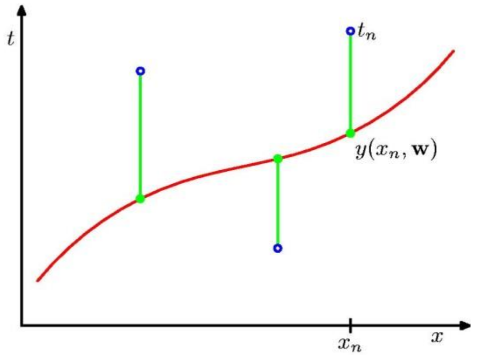


$$\mathcal{N}(x|\mu,\sigma^2) > 0$$

$$\int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu,\sigma^2\right) \,\mathrm{d}x = 1$$

Maximum Likelihood interpretation of least squares regression

MLE for Linear Regression



- Assume a stochastic model:

$$y^{(n)} = \mathbf{w}^T \phi(\mathbf{x}^{(n)}) + \epsilon \ \text{ where } \epsilon \sim \mathcal{N}(0, \beta^{-1})$$

- This gives a likelihood function:

$$p(y^{(n)} | \phi(\mathbf{x}^{(n)}), \mathbf{w}, \beta) = \mathcal{N}(y^{(n)} | \mathbf{w}^T \phi(\mathbf{x}^{(n)}), \beta^{-1})$$

- With input matrix $\mathbf{\Phi}$ and output matrix $\mathbf{y}$, the data likelihood is:

$$p(\mathbf{y} | \mathbf{\Phi}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(y^{(n)} | \mathbf{w}^T \phi(\mathbf{x}^{(n)}), \beta^{-1})$$

Log-likelihood

- Given data likelihood (prev. slide)

$$p(\mathbf{y}|\mathbf{\Phi}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(y^{(n)}|\mathbf{w}^T\phi(\mathbf{x}^{(n)}), \beta^{-1})$$

- Log likelihood:

$$\log p(\mathbf{y}|\mathbf{\Phi}, \mathbf{w}, \beta) = N \log \beta - \frac{N}{2} \log 2\pi - \beta E_D(\mathbf{w})$$

Where $E_D(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} (\mathbf{w}^T\phi(\mathbf{x}^{(n)}) - y^{(n)})^2$

- Derivation?

Derivation of log-likelihood of p

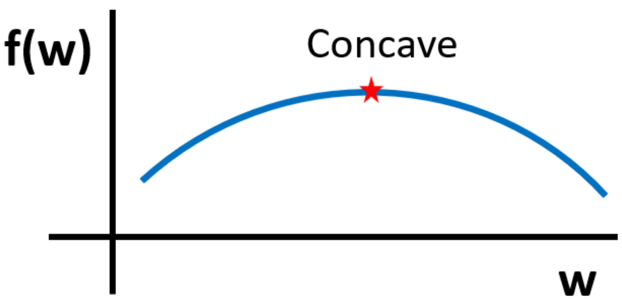
From
$$p(y^{(n)}|\phi(\mathbf{x}^{(n)}), \mathbf{w}, \beta) = \mathcal{N}(y^{(n)}|\mathbf{w}^T\phi(\mathbf{x}^{(n)}), \beta^{-1})$$
$$= \frac{\beta}{\sqrt{2\pi}} \exp(-\frac{\beta}{2}||y^{(n)} - \mathbf{w}^T\phi(\mathbf{x}^{(n)})||^2)$$

Derive:
$$\log p(y^{(1)}, y^{(2)}, ..., y^{(N)}|\boldsymbol{\Phi}, \mathbf{w}, \beta)$$
$$= \log \prod_{n=1}^{N} \mathcal{N}(y^{(n)}|\mathbf{w}^T\phi(\mathbf{x}^{(n)}), \beta^{-1})$$
$$= \sum_{n=1}^{N} \log \left(\frac{\beta}{\sqrt{2\pi}} \exp(-\frac{\beta}{2}||y^{(n)} - \mathbf{w}^T\phi(\mathbf{x}^{(n)})||^2) \right)$$
$$= \sum_{n=1}^{N} \left(\log\beta - \frac{1}{2}\log 2\pi - \frac{\beta}{2}||y^{(n)} - \mathbf{w}^T\phi(\mathbf{x}^{(n)})||^2 \right)$$
$$= N\log\beta - \frac{N}{2}\log 2\pi - \sum_{n=1}^{N} \frac{\beta}{2}||y^{(n)} - \mathbf{w}^T\phi(\mathbf{x}^{(n)})||^2$$

Maximum likelihood estimation (MLE)

- Let's maximize the log-likelihood!
- Set the gradient of log-likelihood = 0 (Why?)

$$\nabla_{\mathbf{w}} \log p(\mathbf{y}|\boldsymbol{\Phi}, \mathbf{w}, \beta) = \nabla_{\mathbf{w}} \left(\underbrace{N \log \beta - \frac{N}{2} \log 2\pi}_{\text{Constant}} - \sum_{n=1}^{N} \frac{\beta}{2} ||y^{(n)} - \mathbf{w}^T \phi(\mathbf{x}^{(n)})||^2 \right)$$



$$= \beta \sum_{n=1}^{N} (y^{(n)} - \underbrace{\mathbf{w}^T \phi(\mathbf{x}^{(n)})}_{\text{Scalar}}) \phi(\mathbf{x}^{(n)})$$

$$= \beta \left(\sum_{n=1}^{N} y^{(n)} \phi(\mathbf{x}^{(n)}) - \phi(\mathbf{x}^{(n)}) \phi(\mathbf{x}^{(n)})^T \mathbf{w} \right) = 0$$

- In matrix form, $\beta(\Phi^T \mathbf{y} - \Phi^T \Phi \mathbf{w}) = 0$

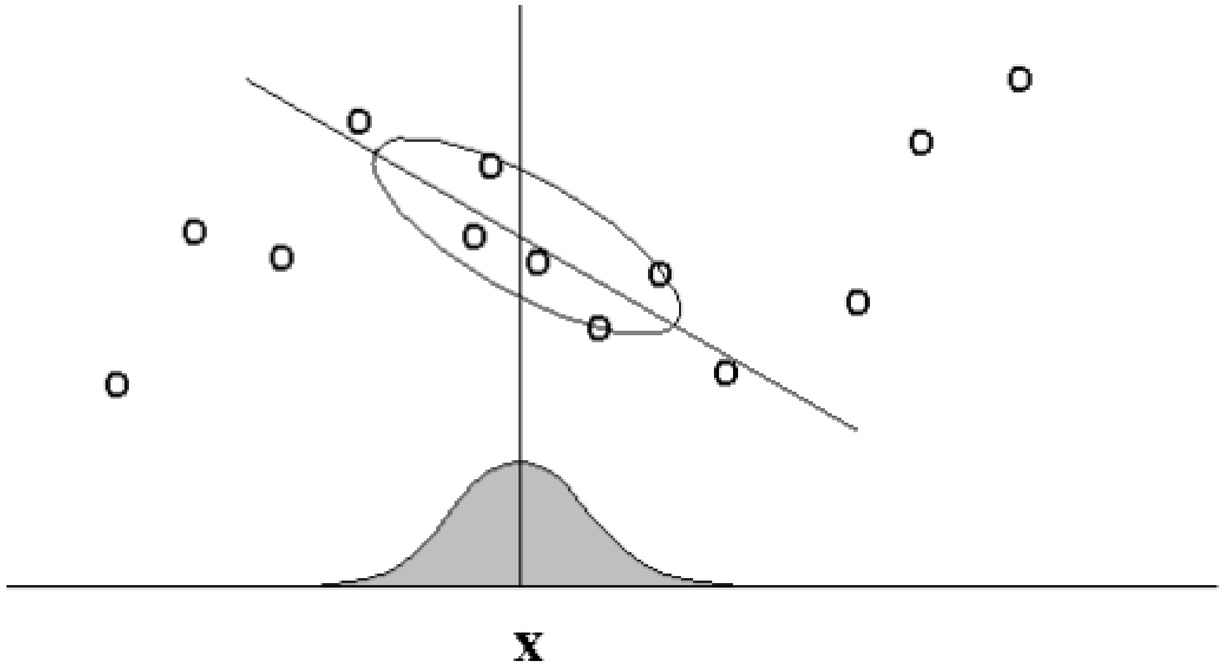
$$\mathbf{w}_{ML} = (\Phi^T \Phi)^{-1} \Phi^T \mathbf{y}$$

- MLE solution is equivalent to OLS solution!

Locally-weighted Linear Regression

Locally weighted linear regression

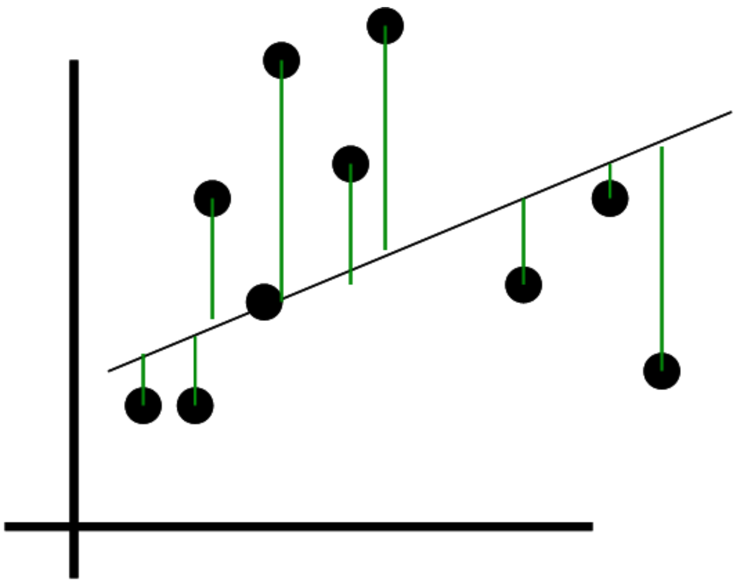
- Main idea: When predicting f(x), give high weights for “neighbors” of x.



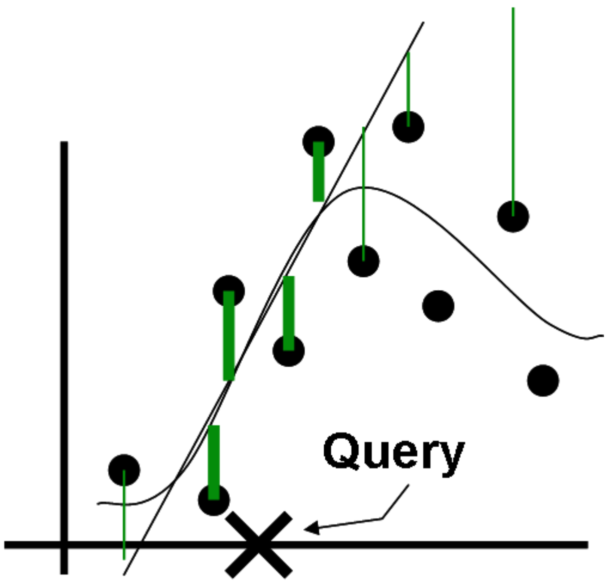
In locally weighted regression, points are weighted by proximity to the current **x** in question using a kernel. A regression is then computed using the weighted points.

Slide credit: William Cohen

Regular linear regression vs. locally weighted linear regression



Regular linear regression



Locally weighted linear regression

$$\sum_{n=1}^{N}(\mathbf{w}^T\phi(\mathbf{x}^{(n)}) - y^{(n)})^2$$

$$\sum_{n=1}^{N} r^{(n)}(\mathbf{w}^T\phi(\mathbf{x}^{(n)}) - y^{(n)})^2$$

Linear regression vs. Locally-weighted Linear Regression

- A new observation $\mathbf{x}$, training set $\{(\mathbf{x}^{(n)}, y^{(n)})\}_{n=1}^{N}$
- Linear regression
  - 1. Fit **w** to minimize $\sum_{n=1}^{N}(\mathbf{w}^T\phi(\mathbf{x}^{(n)}) - y^{(n)})^2$
  - 2. Predict: $\mathbf{w}^T\phi(\mathbf{x})$
- Locally-weighted linear regression
  - 1. Fit **w** to minimize $\sum_{n=1}^{N} r^{(n)}(\mathbf{w}^T\phi(\mathbf{x}^{(n)}) - y^{(n)})^2$ For every x
  - 2. Predict: $\mathbf{w}^T\phi(\mathbf{x})$

Linear regression vs. Locally-weighted Linear Regression

- Locally-weighted linear regression
  - 1. Fit **w** to minimize $$\sum_{n=1}^{N} r^{(n)}(\mathbf{w}^T\phi(\mathbf{x}^{(n)}) - y^{(n)})^2$$
  - 2. Predict: $\mathbf{w}^T\phi(\mathbf{x})$

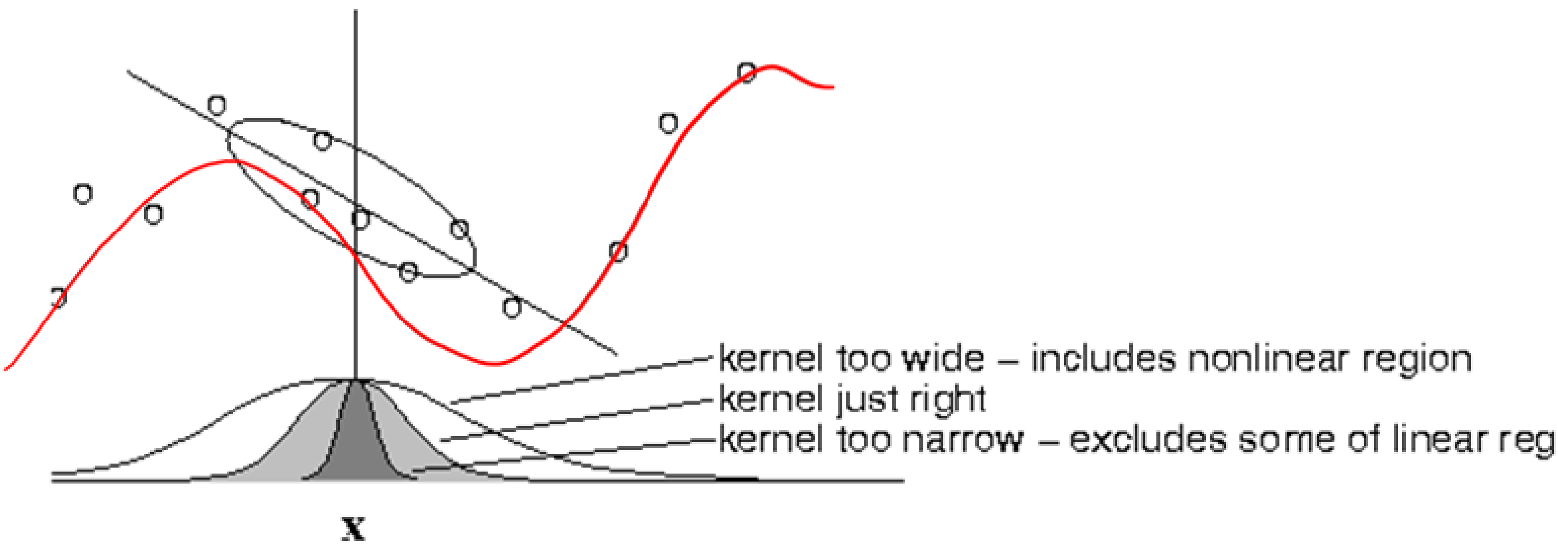
- Remarks:

  "Gaussian Kernel" τ: "kernel width"

  - Standard choice: $r^{(n)} = \exp\left(-\frac{\left\|\phi(\mathbf{x}^{(n)}) - \phi(\mathbf{x})\right\|^2}{2\tau^2}\right)$
  - Note that $r^{(n)}$ depends on x (query point), and you solve linear regression for each query point x.
  - The problem can be formulated as a modified version of least squares problem (HW#1)

Locally weighted linear regression

- Choice of kernel width τ matters
  - Requires hyper-parameter tuning



The estimator is minimized when kernel includes as many training points as can be accommodated by the model. Too large a kernel includes points that degrade the fit; too small a kernel neglects points that increase confidence in the fit.

Slide credit: William Cohen