Linear Regression (Part II)

nanopoulos@ismll.de

Outline

- Generalize the linear regression
- Simple multiple regression
- Regularization
- Bias vs. Variance tradeoff
- Model selection

2

Linear regression in D dimensions

From D = 1 $$y(\mathrm{x}, \mathrm{w}) = w_0 + w_1 x_1$$

To D > 1 and linear on **x** (multiple regression) $$y(\mathrm{x}, \mathrm{w}) = w_0 + w_1 x_1 + \ldots + w_D x_D$$

To D > 1 and any set of nonlinear functions on **x** $$y(\mathrm{x}, \mathrm{w}) = \sum_{j=0}^{M-1} w_j \phi_j(\mathrm{x}) = \mathrm{w}^{\mathrm{T}} \phi(\mathrm{x})$$

All are linear on **w**
Linear regression

3

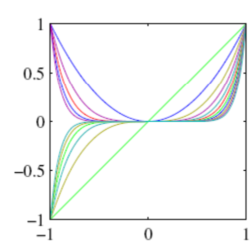
Linear basis functions models

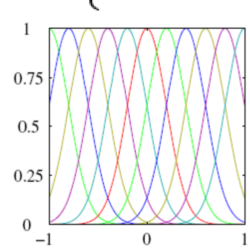
$$y(\mathrm{x}, \mathrm{w}) = \sum_{j=0}^{M-1} w_j \phi_j(\mathrm{x}) = \mathrm{w}^{\mathrm{T}} \phi(\mathrm{x})$$

$$\phi_j(x) = x$$

$$\phi_j(x) = x^j$$

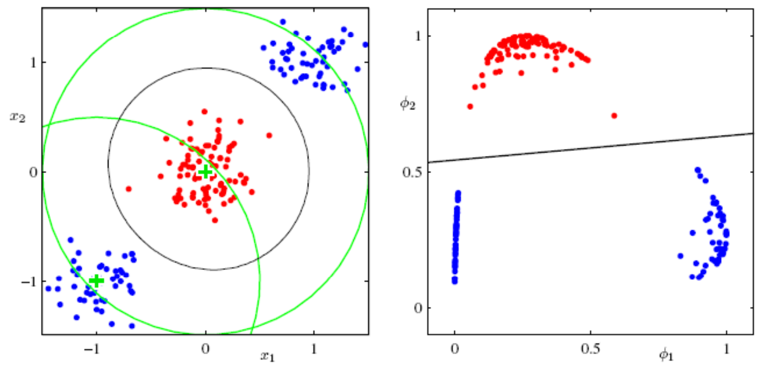
$$\phi_j(x) = \exp\left\{-\frac{(x-\mu_j)^2}{2s^2}\right\}$$





4

Example of basis functions transformation



Maximum likelihood and least squares

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

Assuming normal distribution

$$p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n|\mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n), \beta^{-1})$$

The likelihood to observe N target values **t** from a set of **X** D dimensional predictors

$$\begin{aligned} \ln p(\mathbf{t}|\mathbf{w}, \beta) &= \sum_{n=1}^{N} \ln \mathcal{N}(t_n|\mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n), \beta^{-1}) \\ &= \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi) - \beta E_D(\mathbf{w}) \end{aligned}$$

Maximize ln of likelihood…

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

is the same as minimizing E

Solving MLE

$$\nabla \ln p(\mathbf{t}|\mathbf{w}, \beta) = \sum_{n=1}^{N} \{t_n - \mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n)\} \phi(\mathbf{x}_n)^{\mathrm{T}}$$

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

Find optimal **w**...

$$\mathbf{w}_{\mathrm{ML}} = \left(\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\right)^{-1} \mathbf{\Phi}^{\mathrm{T}}\mathbf{t}$$

based on Moore-Penrose pseudo-inverse

$$\mathbf{\Phi} = \begin{pmatrix} \phi_0(\mathbf{x}_1) & \phi_1(\mathbf{x}_1) & \cdots & \phi_{M-1}(\mathbf{x}_1) \\ \phi_0(\mathbf{x}_2) & \phi_1(\mathbf{x}_2) & \cdots & \phi_{M-1}(\mathbf{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(\mathbf{x}_N) & \phi_1(\mathbf{x}_N) & \cdots & \phi_{M-1}(\mathbf{x}_N) \end{pmatrix}$$

$$\frac{1}{\beta_{\mathrm{ML}}} = \frac{1}{N} \sum_{n=1}^{N} \{t_n - \mathbf{w}_{\mathrm{ML}}^{\mathrm{T}} \phi(\mathbf{x}_n)\}^2$$

Outline

- Generalize the linear regression
- Simple multiple regression
- Regularization
- Bias vs. Variance tradeoff
- Model selection

The case of simple multiple regression

$$Y = w_0 + \sum_{i=1}^{p} w_i X_i$$
$$= \langle \mathbf{w}, X \rangle$$

where

$$\mathbf{w} := \begin{pmatrix} w_0 \\ w_1 \\ \vdots \\ w_D \end{pmatrix}, \quad X := \begin{pmatrix} 1 \\ X_1 \\ \vdots \\ X_p \end{pmatrix},$$

Thus, the intercept is handled like any other parameter, for the artificial constant variable $X_0 \equiv 1$.

The case of simple multiple regression

For the whole dataset $(x_1, y_1), \ldots, (x_n, y_n)$:

$$\mathbf{Y} = \mathbf{X}\,\mathbf{w} + \epsilon$$

where

$$\mathbf{Y} := \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{X} := \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} x_{1,1} & x_{1,2} & \ldots & x_{1,p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n,1} & x_{n,2} & \ldots & x_{n,p} \end{pmatrix}$$

Least squares estimate

Least squares estimates $\mathbf{w}$ minimize

$$||\mathbf{Y} - \hat{\mathbf{Y}}||^2 = ||\mathbf{Y} - \mathbf{X}\mathbf{w}||^2$$

The least squares estimates $\mathbf{w}$ are computed via

$$\mathbf{X}^T\mathbf{X}\mathbf{w} = \mathbf{X}^T\mathbf{Y}$$

Proof:

$$||\mathbf{Y} - \mathbf{X}\mathbf{w}||^2 = \langle \mathbf{Y} - \mathbf{X}\mathbf{w}, \mathbf{Y} - \mathbf{X}\mathbf{w} \rangle$$

$$\frac{\partial(\ldots)}{\partial \mathbf{w}} = 2\langle -\mathbf{X}, \mathbf{Y} - \mathbf{X}\mathbf{w} \rangle = -2(\mathbf{X}^T\mathbf{Y} - \mathbf{X}^T\mathbf{X}\mathbf{w}) \stackrel{!}{=} 0$$

11

Least squares estimate

Solve the $p \times p$ system of linear equations

$$X^T X\, \mathbf{w} = X^T Y$$

i.e., $Ax = b$ (with $A := X^T X, b = X^T Y, x = \mathbf{w}$).

There are several numerical methods available:

1. Gaussian elimination
2. Cholesky decomposition
3. QR decomposition

12

Example

Given is the following data:

x_1	x_2	y
1	2	3
2	3	2
4	1	7
5	5	1

Predict a y value for $x_1 = 3, x_2 = 4$.

Solution

x_1	x_2	y
1	2	3
2	3	2
4	1	7
5	5	1

$$X = \begin{pmatrix} 1 & 1 & 2 \\ 1 & 2 & 3 \\ 1 & 4 & 1 \\ 1 & 5 & 5 \end{pmatrix}, \quad Y = \begin{pmatrix} 3 \\ 2 \\ 7 \\ 1 \end{pmatrix}$$

$$X^T X = \begin{pmatrix} 4 & 12 & 11 \\ 12 & 46 & 37 \\ 11 & 37 & 39 \end{pmatrix}, \quad X^T Y = \begin{pmatrix} 13 \\ 40 \\ 24 \end{pmatrix}$$

Solution (cont.)

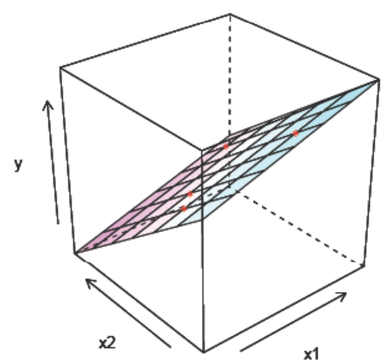
$$\left(\begin{array}{ccc|c} 4 & 12 & 11 & 13 \\ 12 & 46 & 37 & 40 \\ 11 & 37 & 39 & 24 \end{array}\right) \sim \left(\begin{array}{ccc|c} 4 & 12 & 11 & 13 \\ 0 & 10 & 4 & 1 \\ 0 & 16 & 35 & -47 \end{array}\right) \sim \left(\begin{array}{ccc|c} 4 & 12 & 11 & 13 \\ 0 & 10 & 4 & 1 \\ 0 & 0 & 143 & -243 \end{array}\right)$$

$$\sim \left(\begin{array}{ccc|c} 4 & 12 & 11 & 13 \\ 0 & 1430 & 0 & 1115 \\ 0 & 0 & 143 & -243 \end{array}\right) \sim \left(\begin{array}{ccc|c} 286 & 0 & 0 & 1597 \\ 0 & 1430 & 0 & 1115 \\ 0 & 0 & 143 & -243 \end{array}\right)$$

i.e.,

$$\mathbf{w} = \begin{pmatrix} 1597/286 \\ 1115/1430 \\ -243/143 \end{pmatrix} \approx \begin{pmatrix} 5.583 \\ 0.779 \\ -1.699 \end{pmatrix}$$

What it looks like?



Outline

- Generalize the linear regression
- Simple multiple regression
- Regularization
- Bias vs. Variance tradeoff
- Model selection

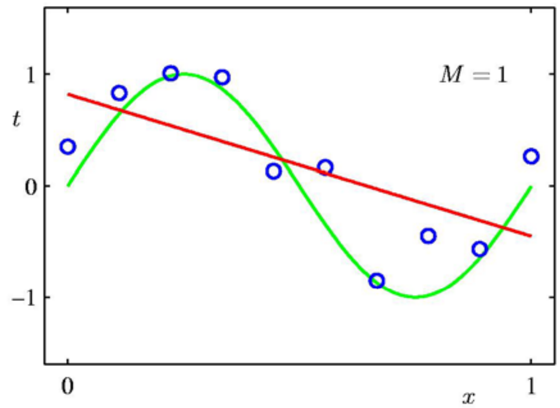
Regularization

Recall the case of

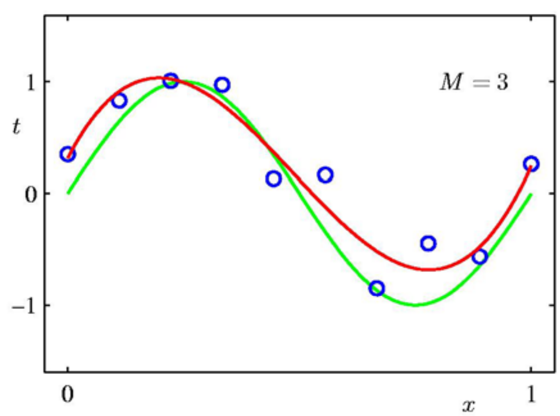
- a single predictor
- and polynomial basis function $\phi_j(x) = x^j$

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

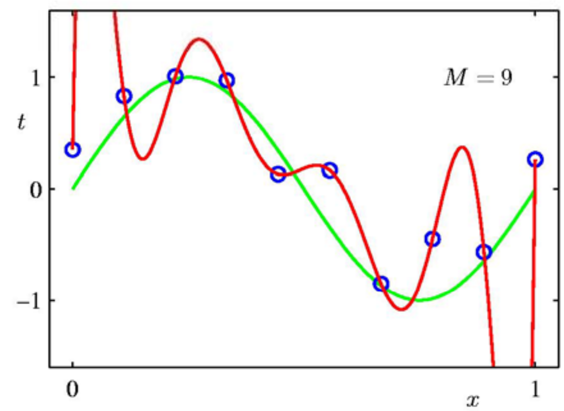
1st Order Polynomial



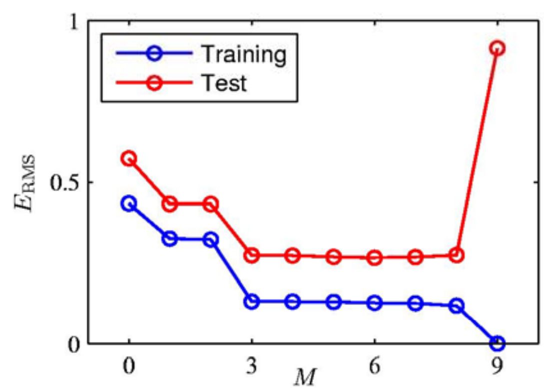
3rd Order Polynomial



9th Order Polynomial



Over-fitting



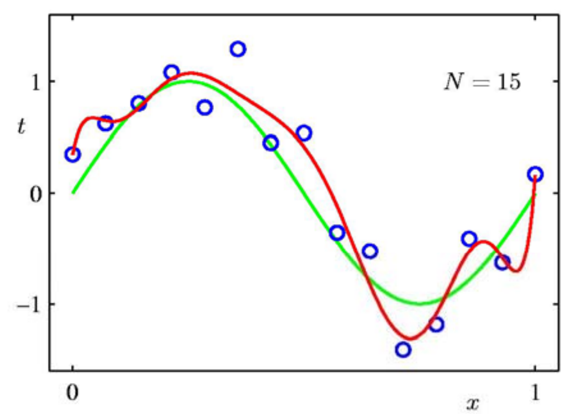
Root-Mean-Square (RMS) Error: $E_{RMS} = \sqrt{2E(\mathbf{w}^*)/N}$

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

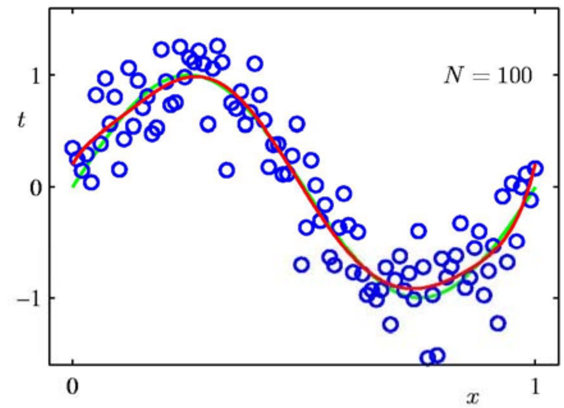
Data Set Size: $N = 15$

9^{th} Order Polynomial



Data Set Size: $N = 100$

9th Order Polynomial

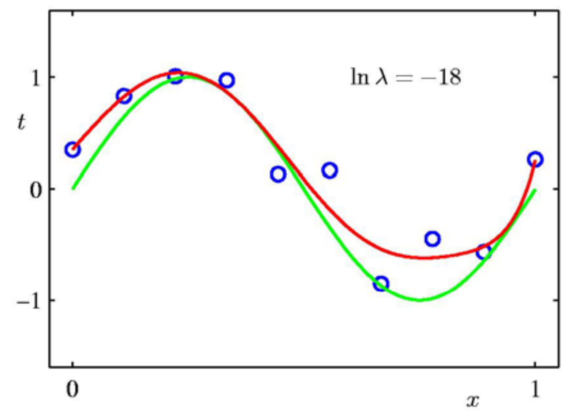


Regularization

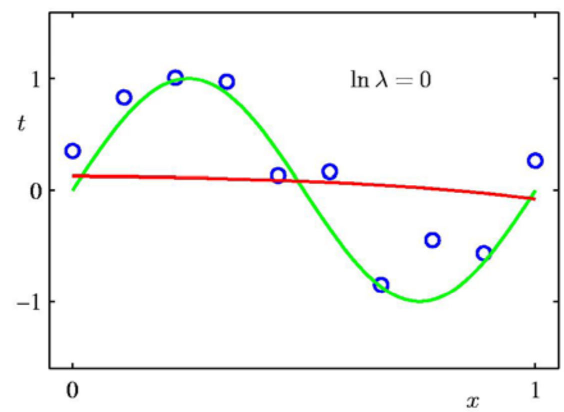
Penalize large coefficient values

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

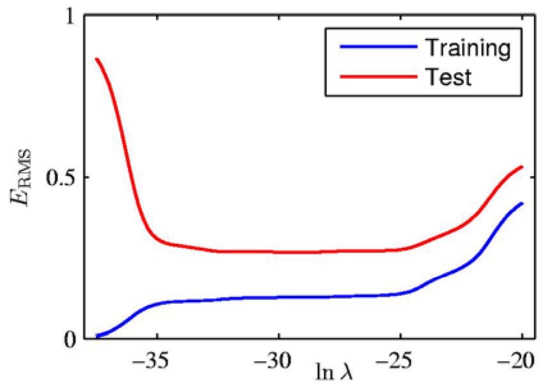
Regularization: $\ln \lambda = -18$



Regularization: $\ln \lambda = 0$



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



Polynomial Coefficients

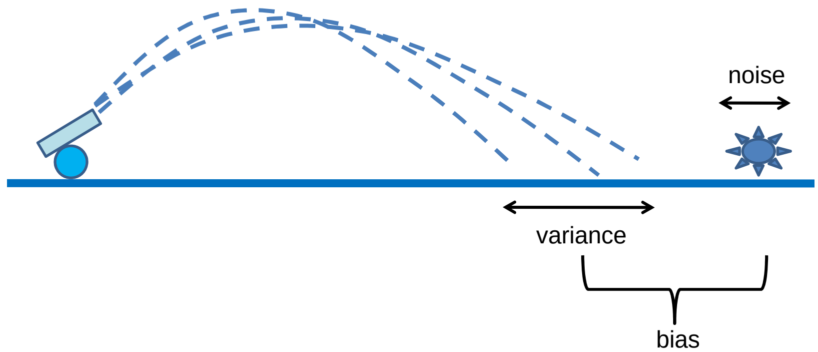
	$\ln \lambda = -\infty$	$\ln \lambda = -18$	$\ln \lambda = 0$
$w_0^\star$	0.35	0.35	0.13
$w_1^\star$	232.37	4.74	-0.05
$w_2^\star$	-5321.83	-0.77	-0.06
$w_3^\star$	48568.31	-31.97	-0.05
$w_4^\star$	-231639.30	-3.89	-0.03
$w_5^\star$	640042.26	55.28	-0.02
$w_6^\star$	-1061800.52	41.32	-0.01
$w_7^\star$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

Outline

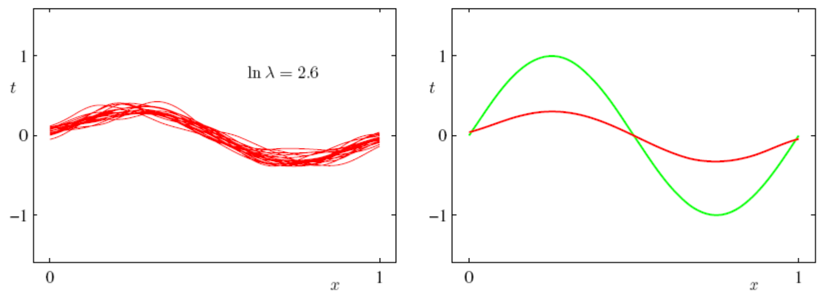
- Generalize the linear regression
- Simple multiple regression
- Regularization
- Bias vs. Variance tradeoff
- Model selection

Bias vs. Variance

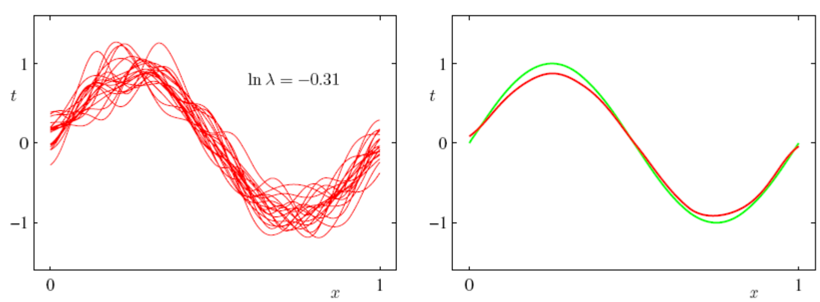
$$\text{expected loss} = (\text{bias})^2 + \text{variance} + \text{noise}$$



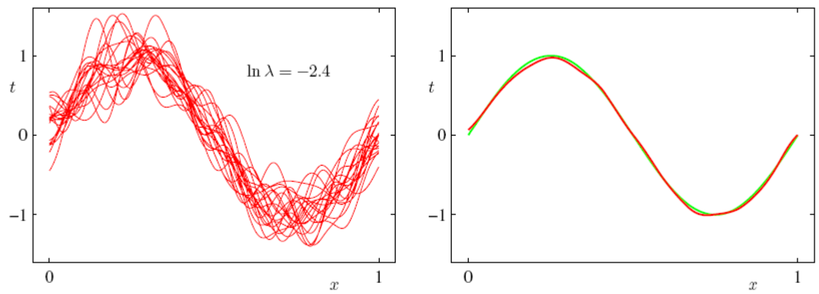
How this applies to regression?



How this applies to regression?



How this applies to regression?



Outline

- Generalize the linear regression
- Simple multiple regression
- Regularization
- Bias vs. Variance tradeoff
- Model selection

Model Selection

M (degree of polynomial) corresponds to the complexity of the model

The bigger M, the bigger the overfitting

Regularization (λ) helps by controlling the complexity

But how to find good combinations for M and λ?

Information criteria

$p(D|\mathbf{w}_{\mathrm{ML}})$ is the likelihood

M is the number of parameters

- AIC (Akaike Information Criterion)
  maximize $p(D|\mathbf{w}_{\mathrm{ML}})$ - M

Tends to favor overly simply models