University of Amsterdam
Informatics Institute

AMLAB
Amsterdam Machine Learning Lab

Machine Learning 1

Lecture 3.5 - Supervised Learning
Regularized Least Squares

Erik Bekkers

(Bishop 3.1.4)

Slide credits: Patrick Forré and Rianne van den Berg

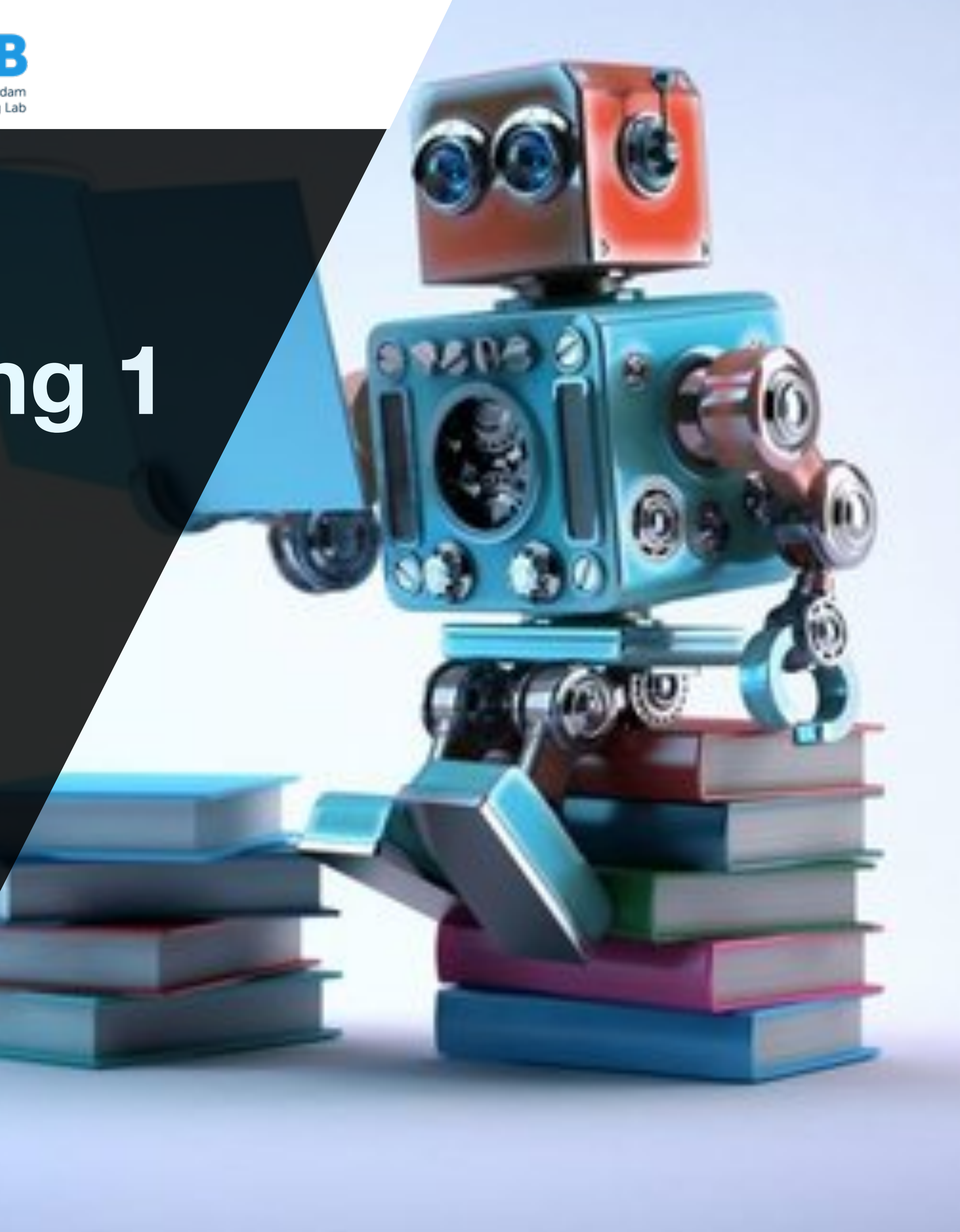


Image credit: Kirillm | Getty Images

Example: Overfitting and Underfitting

$$t = \sin(2\pi x) + \varepsilon \qquad \varepsilon \sim \mathcal{N}(0, \beta^{-1}) \qquad y(x, \mathbf{w}) = w_0 + \sum_{i=1}^{M} w_i x^i$$

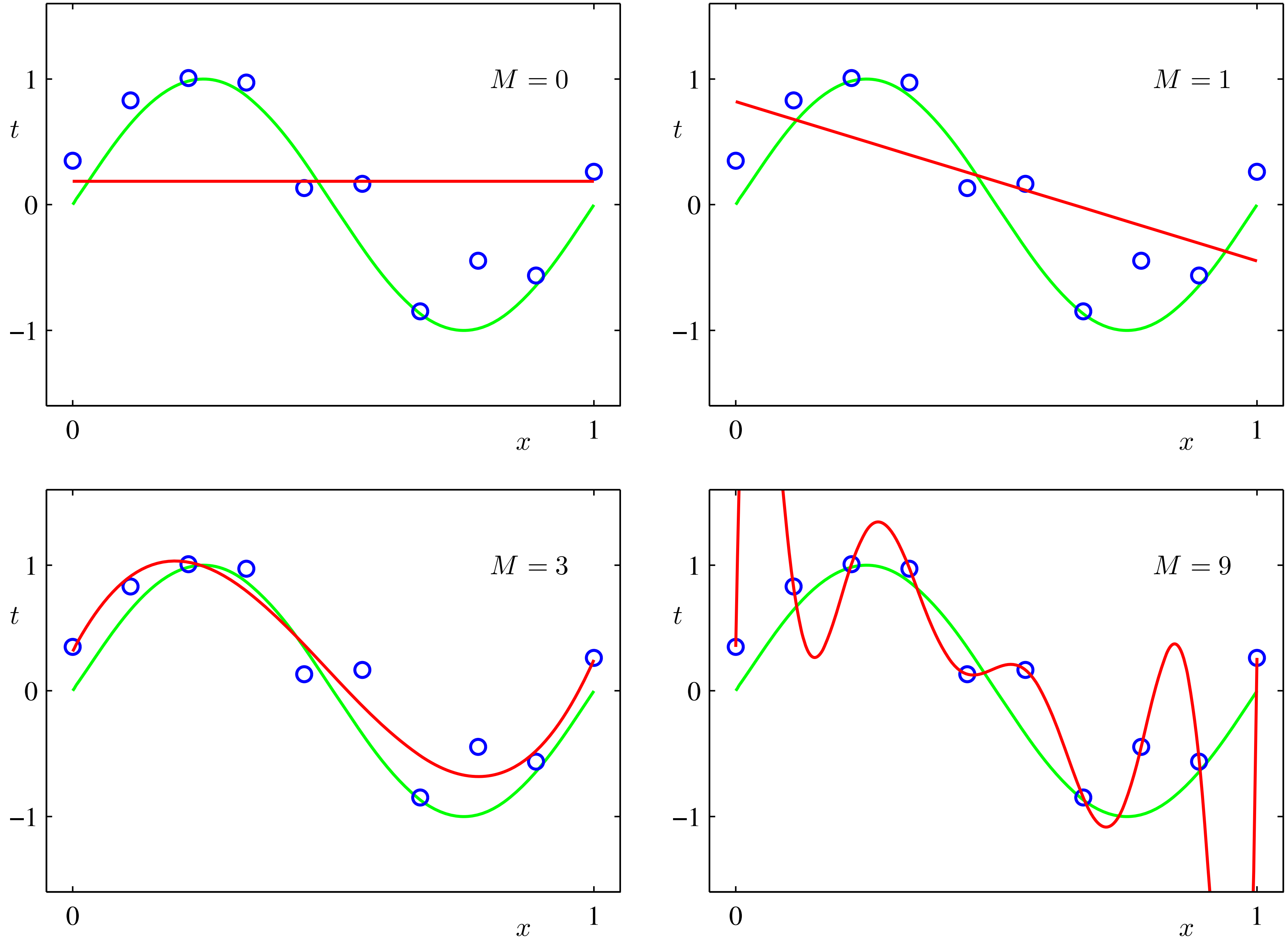


Figure: Fits of different polynomials (Bishop 1.4)

Example: Overfitting and Underfitting

	$M=0$	$M=1$	$M=6$	$M=9$
$w_0^\star$	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

prevent large values

overfitting

Table: Polynomial coefficients (Bishop 1.1)

Regularized Least Squares

- Instead of manually constraining the number of parameters for small datasets, add penalty term for large parameter values:

$$\tilde{E}(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{N}\{t_i - y(\mathbf{x}_i, \mathbf{w})\}^2 + \frac{\lambda}{2}\sum_{i=1}^{M-1}(w_i)^2$$

ridge regression

- The bias term w_0 is not always included in regularization

↳ its role is to allow for offsets

↳ doesn't add to "model complexity"

Regularized Least Squares

$$\tilde{E}(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{N}\{t_i - y(\mathbf{x}_i, \mathbf{w})\}^2 + \frac{1}{2}\lambda \mathbf{w}^T\mathbf{w}$$

- Note: equivalent to Maximum A Posteriori (MAP) approach for estimating **w** with Gaussian prior:

$$c\, e^{-\frac{\alpha}{2} \underline{w}^T \underline{w}}$$

$$p(\mathbf{w}\,|\,\mathbf{X}, \mathbf{t}, \alpha) = \frac{p(\mathbf{t}\,|\,\mathbf{X}, \mathbf{w})p(\mathbf{w}\,|\,\alpha)}{p(\mathbf{t}\,|\,\mathbf{X}, \alpha)} \qquad p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \mathbb{1}\alpha^{-1})$$

$$\mathbf{w}_{\text{MAP}} = \arg\min_{\mathbf{w}} -\log p(\mathbf{w}|\mathbf{X}, \mathbf{t}, \alpha) = \arg\min_{\mathbf{w}} -\log p(\mathbf{t}|\mathbf{X}, \mathbf{w}) - \log p(\mathbf{w}|\alpha)$$

$$= \arg\min_{\underline{w}} \frac{\beta}{2}\sum_{i=1}^{N}(t_i - y(\underline{x}_i, \underline{w}))^2 + \frac{\alpha}{2}\underline{w}^T\underline{w}$$

$(\beta > 0)$

$$= \arg\min_{\underline{w}} \frac{1}{2}\sum_{i=1}^{N}(t_i - y(\underline{x}_i, \underline{w}))^2 + \frac{1}{2}\frac{\alpha}{\beta}\underline{w}^T\underline{w}$$

same $\lambda = \frac{\alpha}{\beta}$

Example: Regularized Polynomial Regression

no regularization

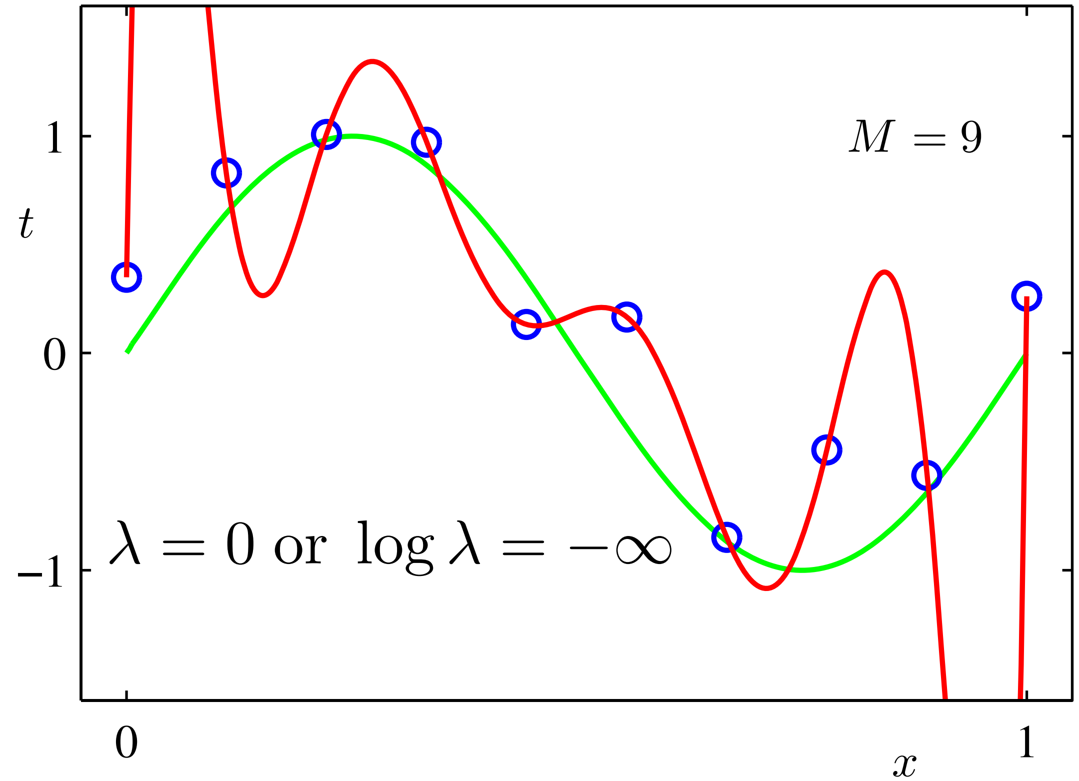


Figure: polynomial regression (Bishop 1.4)

suppress weights just enough

too much regularization

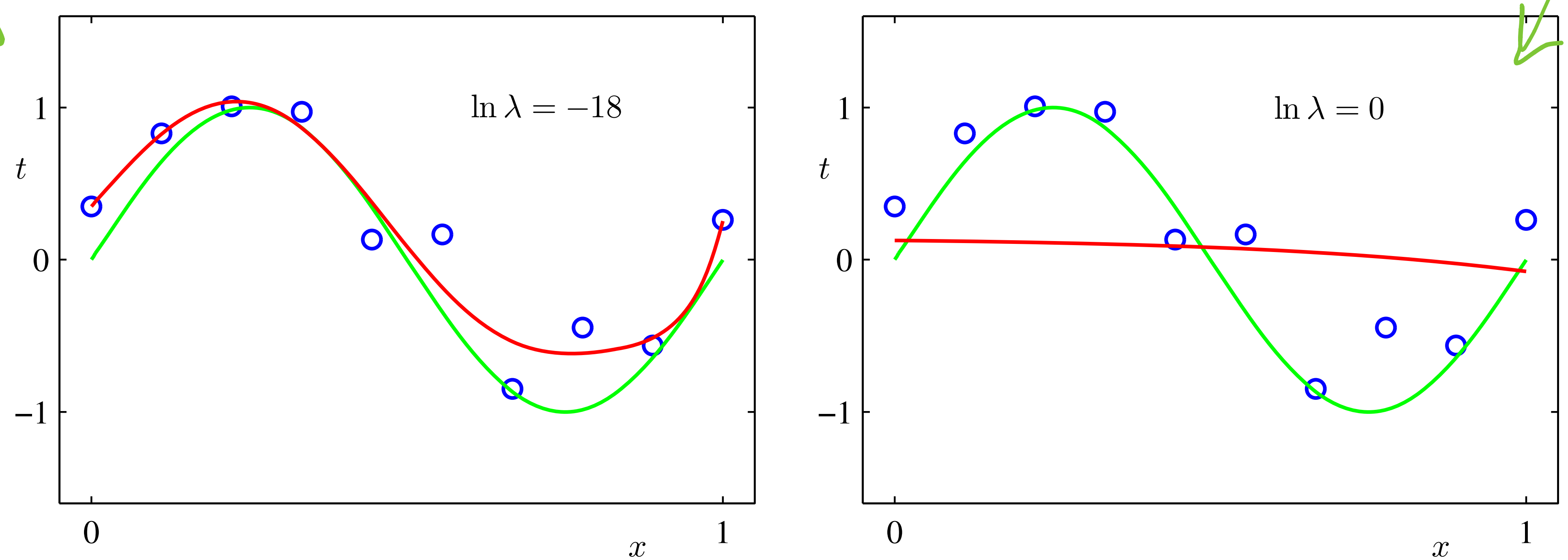


Figure: Regularized polynomial regression (Bishop 1.7)

Example: Regularized Polynomial Regression

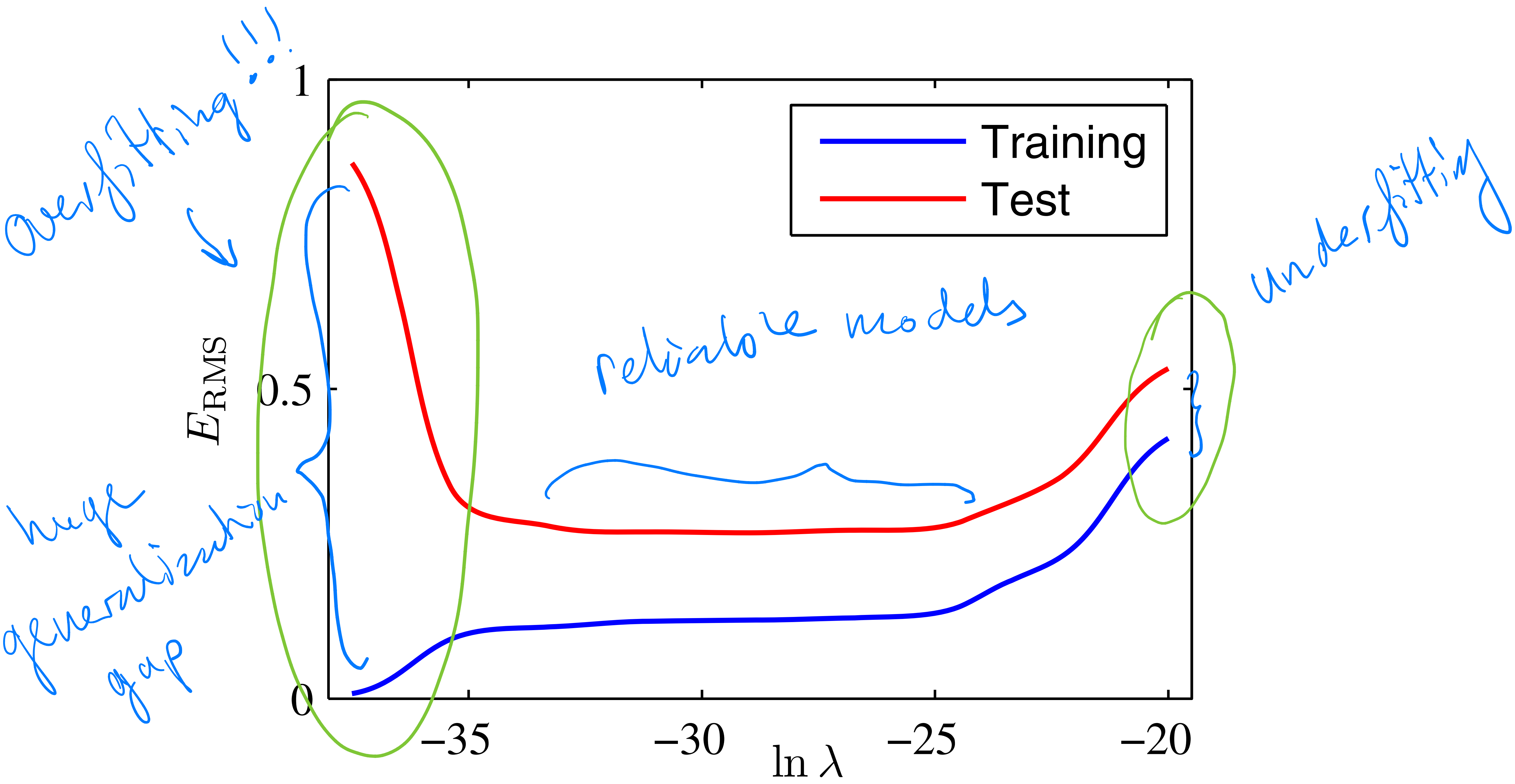


Figure: train and test errors for regularized M=9 polynomial regression (Bishop 1.8)

Regularized Least Squares (II)

- Weight decay : $\tilde{E}(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{N}\{t_i - \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i)\}^2 + \frac{\lambda}{2}\sum_{i=1}^{M}|w_i|^2$
- More general :

$$\tilde{E}(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{N}\{t_i - \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i)\}^2 + \frac{\lambda}{2}\sum_{i=1}^{M}|w_i|^q$$

- q = 1: Lasso

$\rightarrow$ $\underline{w}$ sparse

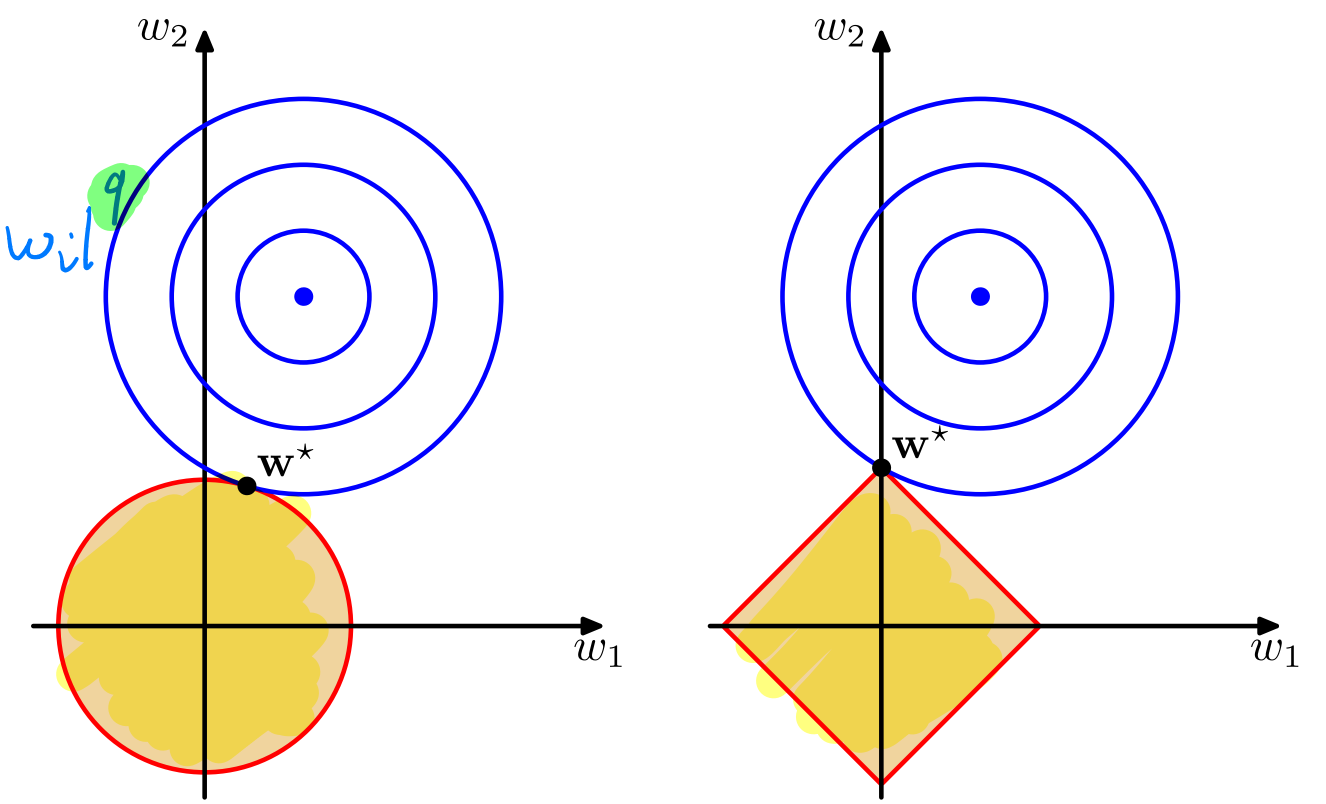


Figure: regularization as constrained optimization (Bishop 3.4)

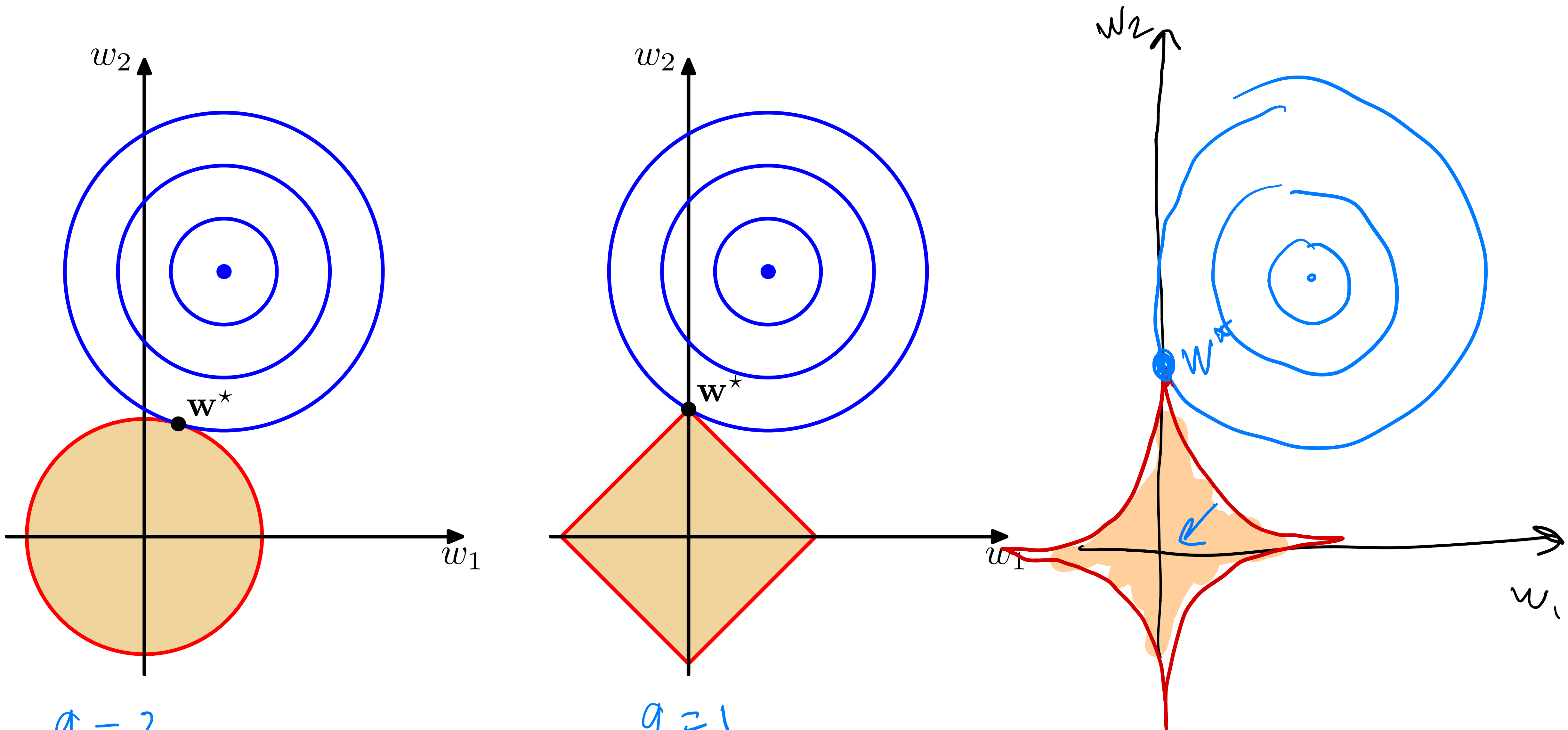
- Equivalent to minimizing

Bishop App G

$$\frac{1}{2}\sum_{i=1}^{N}\{t_i - \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i)\}^2 \quad \text{with} \quad \sum_{j=1}^{M}|w_j|^q \leq \eta$$

Regularized Least Squares: sparse weights

$$\frac{1}{2}\sum_{i=1}^{N}\{t_i - \mathbf{w}^T\phi(\mathbf{x}_i)\}^2 \quad \text{with} \quad \sum_{j=1}^{M}|w_j|^q \leq \eta$$



$q = 2$ $\quad$ $q = 1$ $\quad$ $q = \frac{1}{2}$

Figure: regularization as constrained optimization (Bishop 3.4)

Example: Prostate specific antigen predediction

q=2 (Ridge regression)

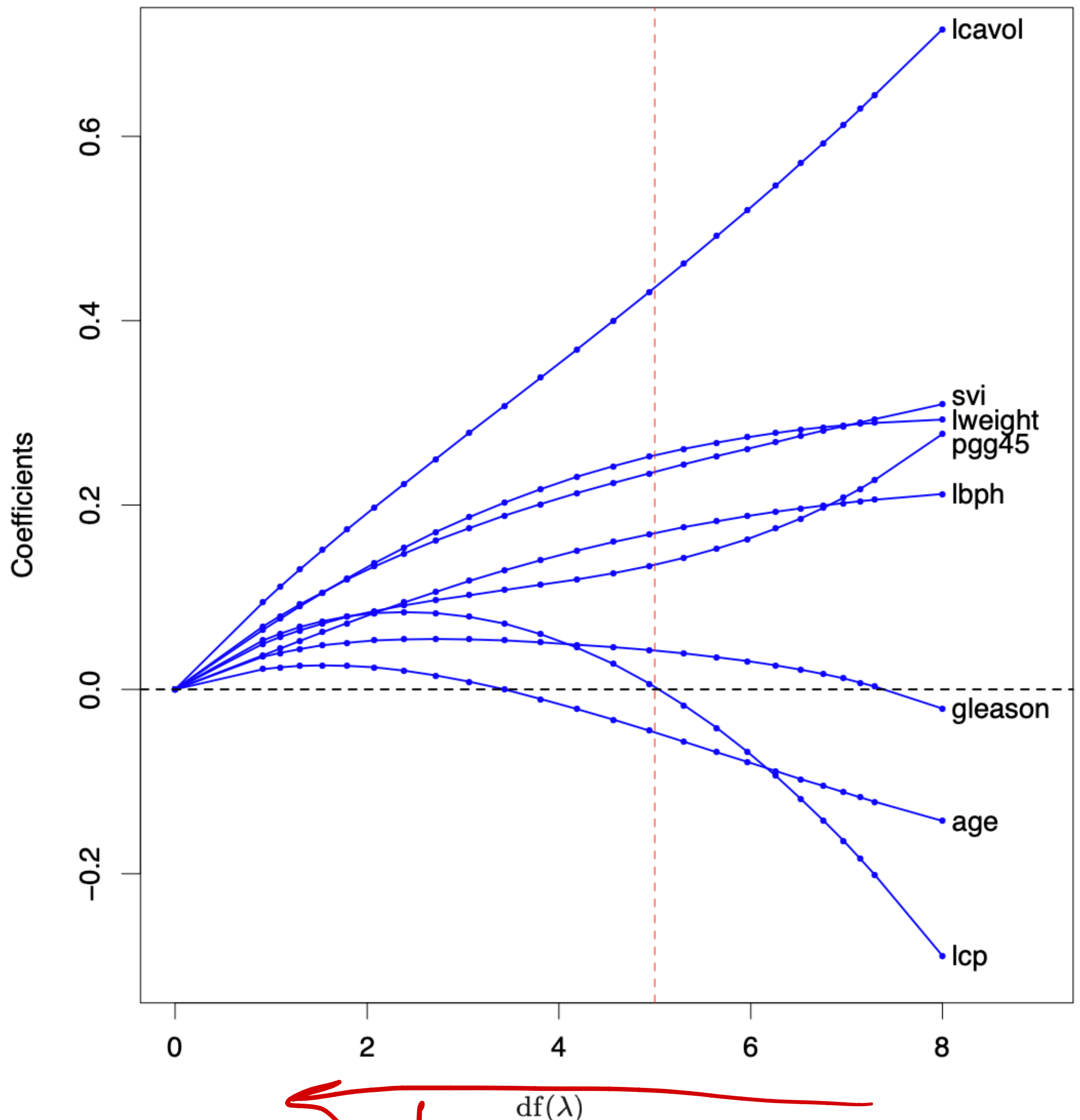


FIGURE 3.8. Profiles of ridge coefficients for the prostate cancer example, as the tuning parameter λ is varied. Coefficients are plotted versus $\mathrm{df}(\lambda)$, the effective degrees of freedom. A vertical line is drawn at $\mathrm{df} = 5.0$, the value chosen by cross-validation.

q=1 (Lasso regression)

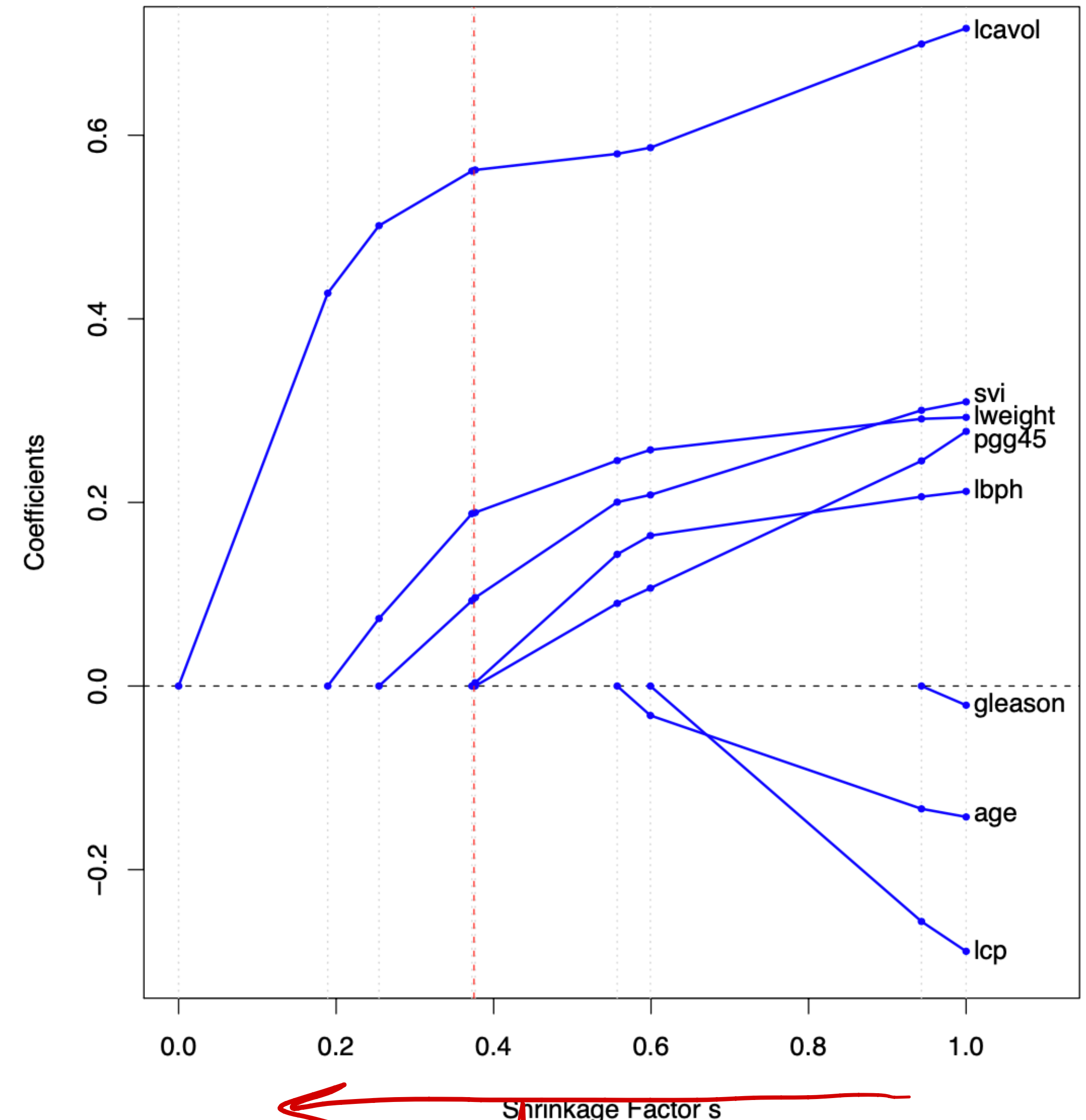


FIGURE 3.10. Profiles of lasso coefficients, as the tuning parameter t is varied. Coefficients are plotted versus $s = t/\sum_1^p |\hat{\beta}_j|$. A vertical line is drawn at $s = 0.36$, the value chosen by cross-validation. Compare Figure 3.8 on page 65; the lasso profiles hit zero, while those for ridge do not. The profiles are piece-wise linear, and so are computed only at the points displayed; see Section 3.4.4 for details.

Figures from the Elements of Statistical Learning (ESL - Hastie et al.)