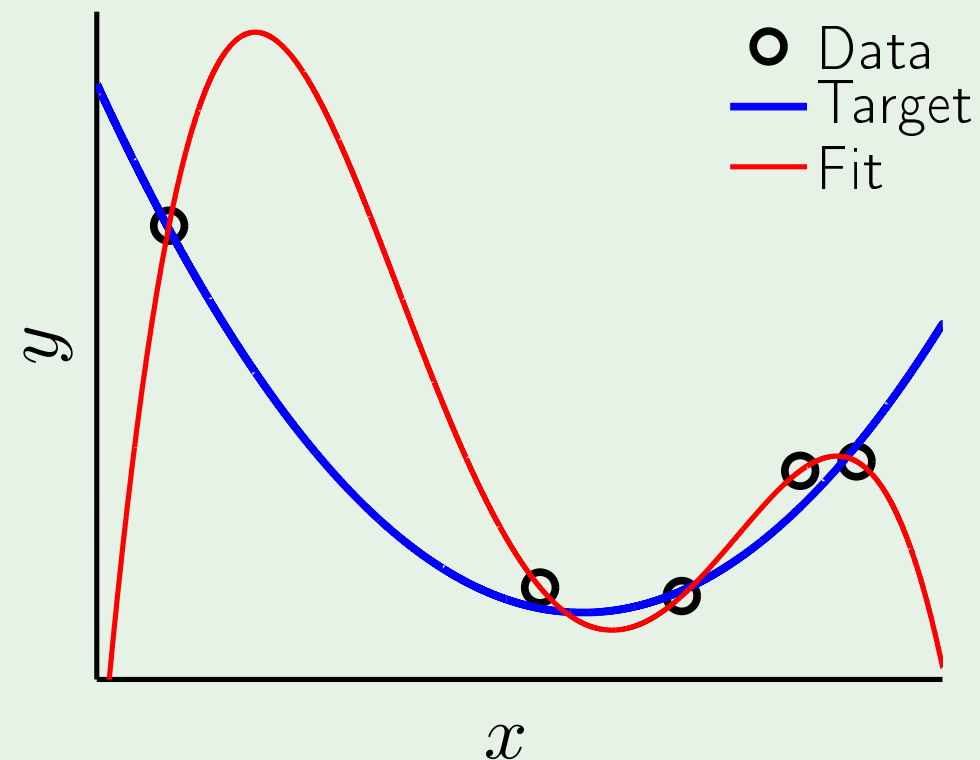


Review of Lecture 11

- Overfitting

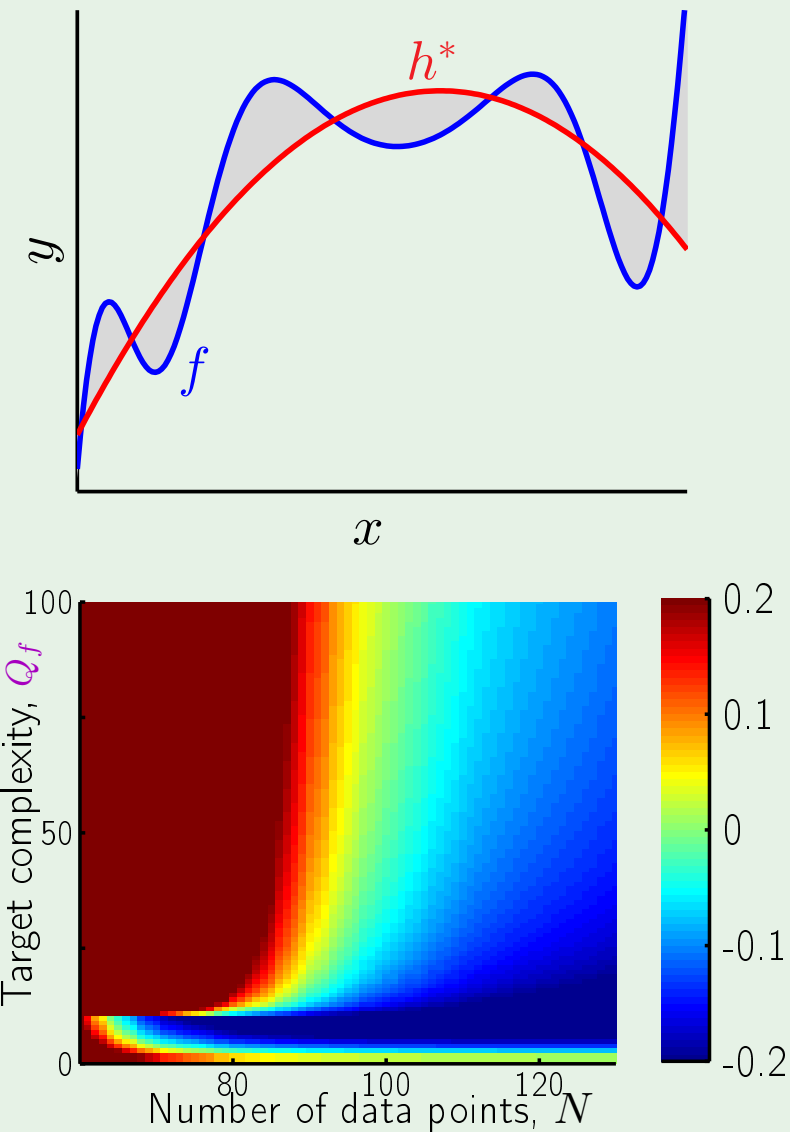
Fitting the data more than is warranted



VC allows it; doesn't predict it

Fitting the noise, stochastic/deterministic

- Deterministic noise



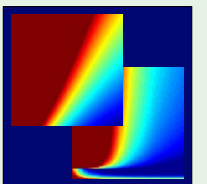
Learning From Data

Yaser S. Abu-Mostafa
California Institute of Technology

Lecture 12: Regularization



Sponsored by Caltech's Provost Office, E&AS Division, and IST • Thursday, May 10, 2012



Outline

- Regularization - informal
- Regularization - formal
- Weight decay
- Choosing a regularizer

Two approaches to regularization

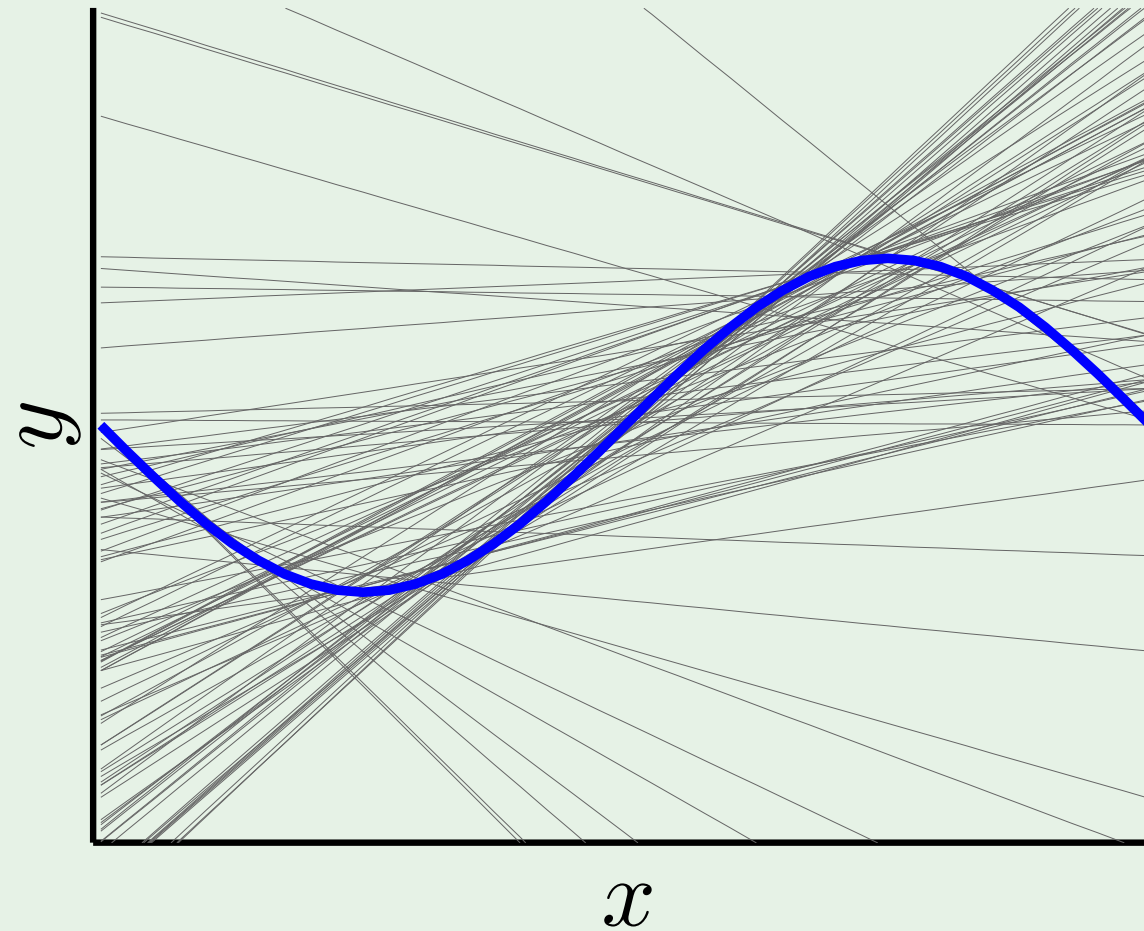
Mathematical:

Ill-posed problems in function approximation

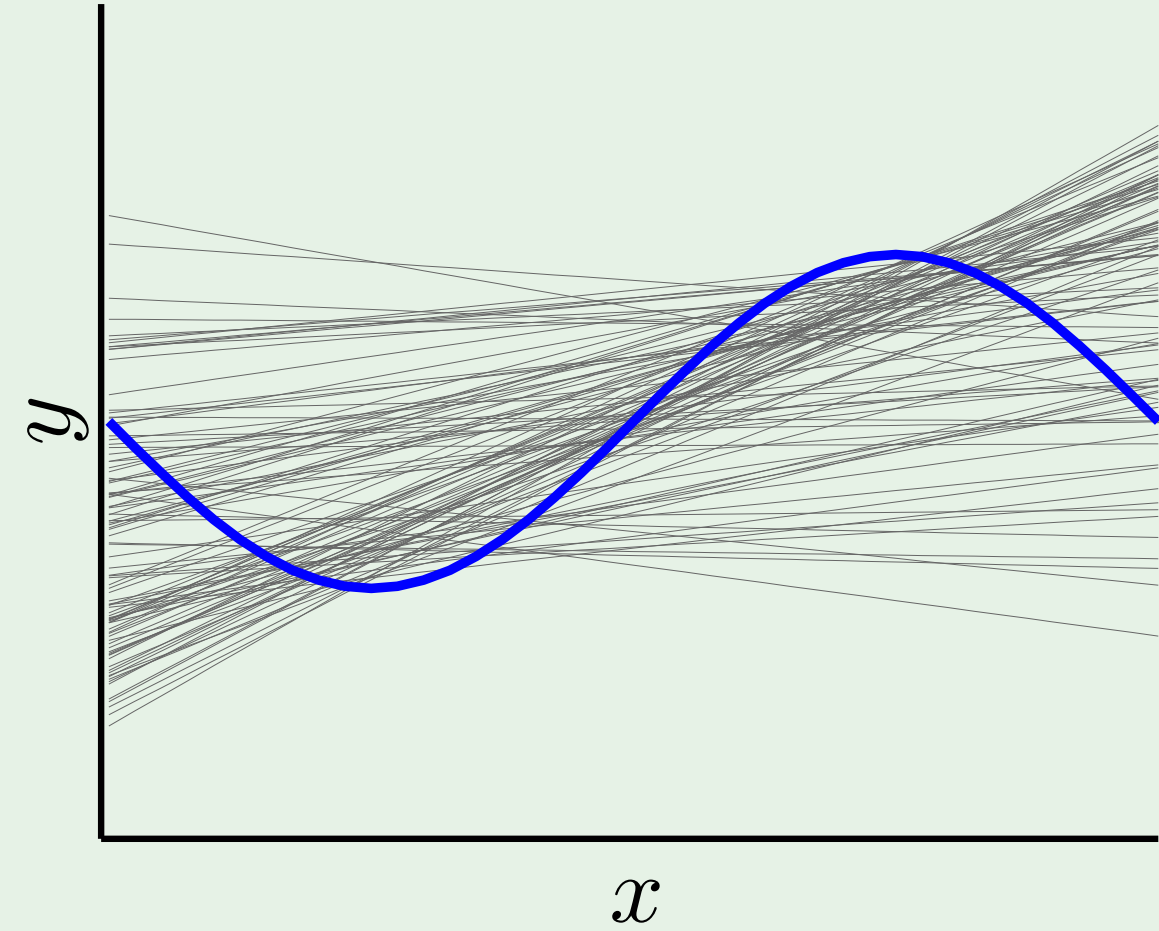
Heuristic:

Handicapping the minimization of E_{in}

A familiar example



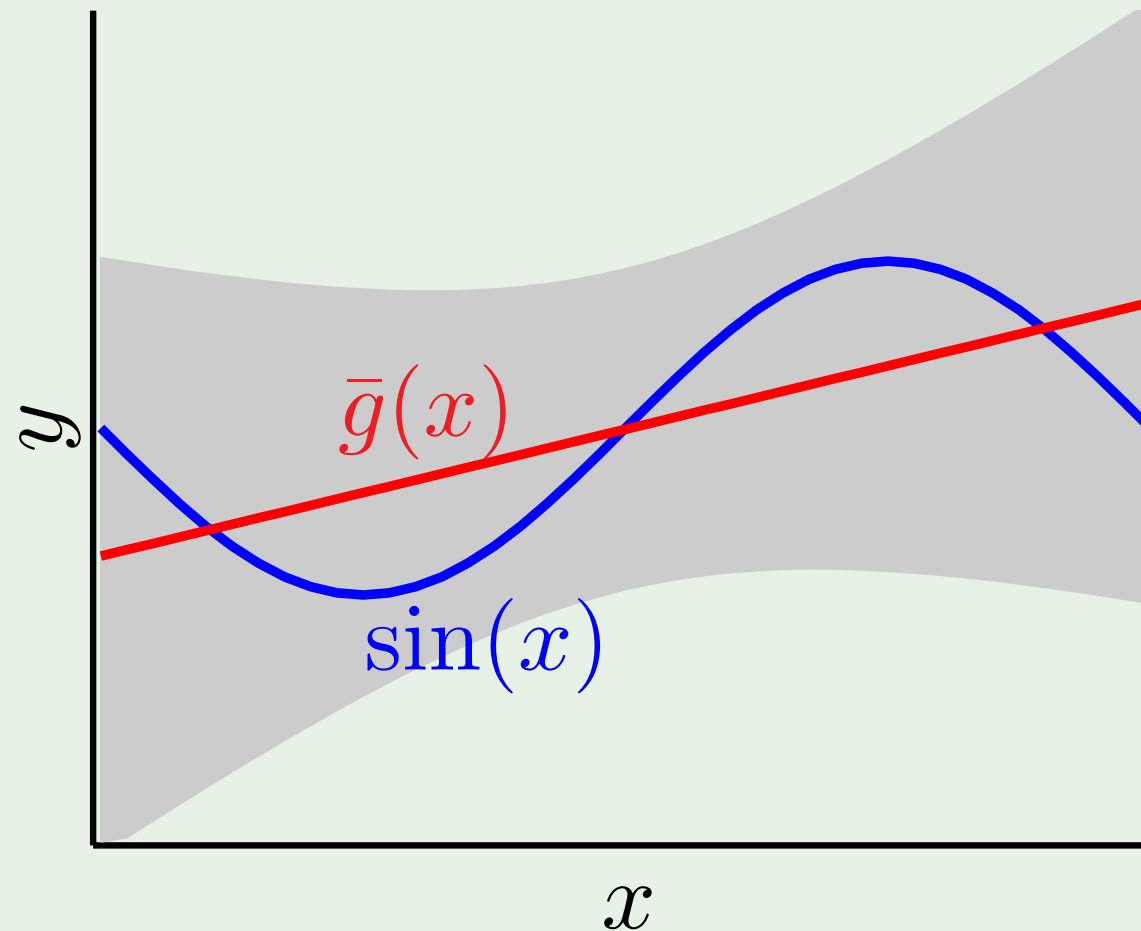
without regularization



with regularization

and the winner is ...

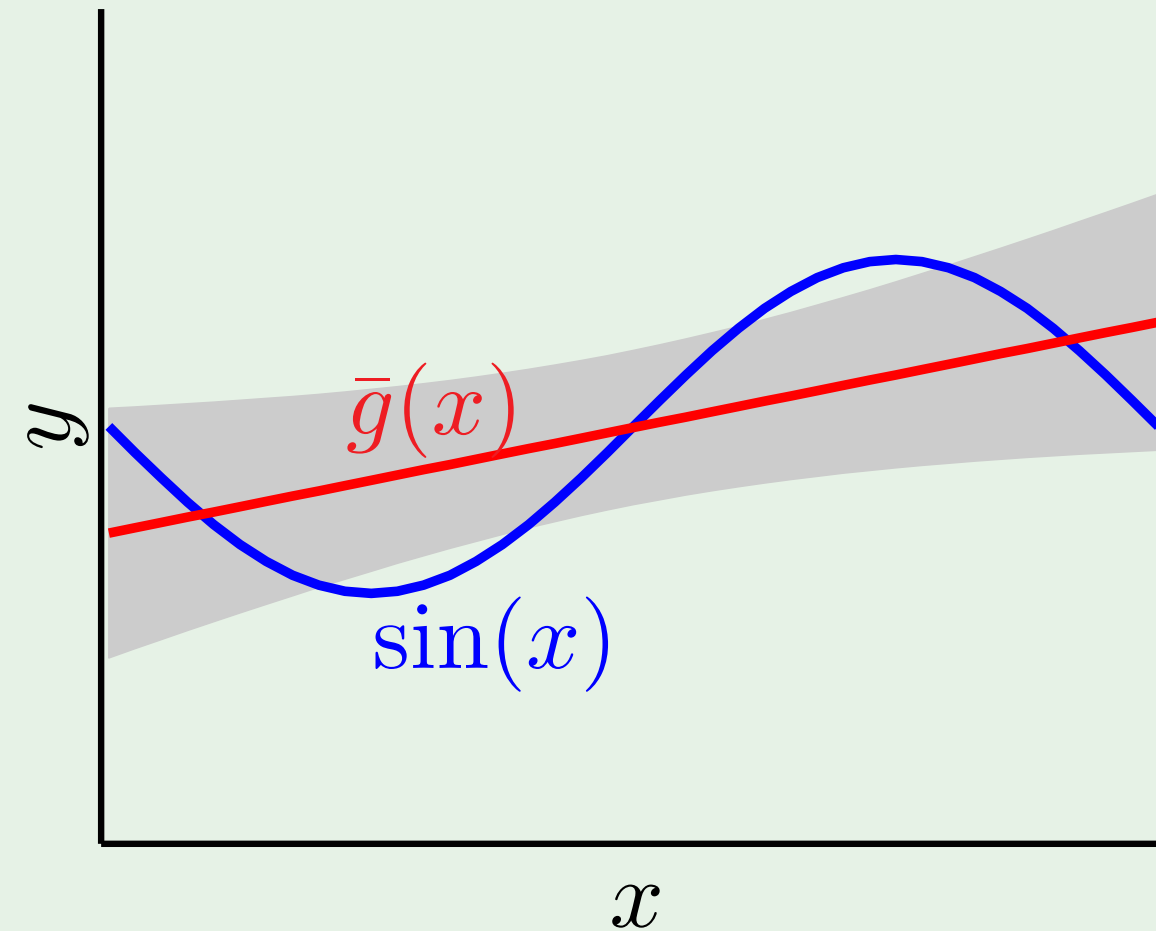
without regularization



bias = **0.21**

var = **1.69**

with regularization



bias = **0.23**

var = **0.33**

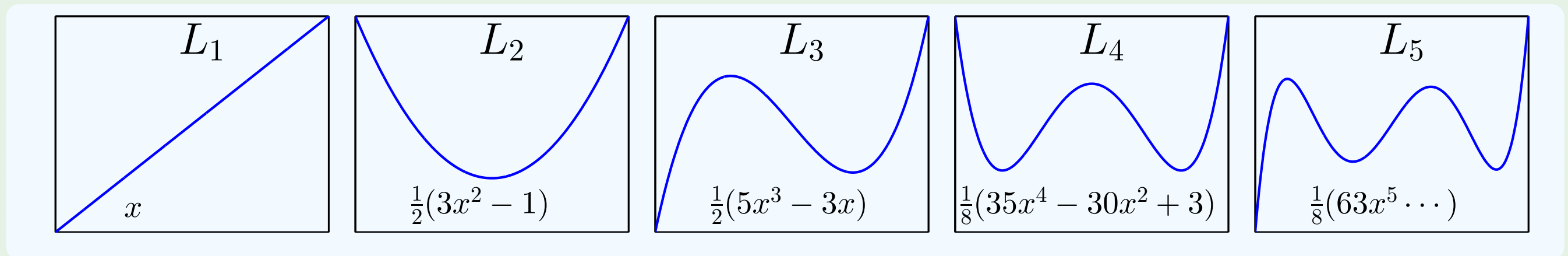
The polynomial model

\mathcal{H}_Q : polynomials of order Q

linear regression in \mathcal{Z} space

$$\mathbf{z} = \begin{bmatrix} 1 \\ L_1(x) \\ \vdots \\ L_Q(x) \end{bmatrix} \quad \mathcal{H}_Q = \left\{ \sum_{q=0}^Q w_q L_q(x) \right\}$$

Legendre polynomials:



Unconstrained solution

Given $(x_1, y_1), \dots, (x_N, y_N) \longrightarrow (\mathbf{z}_1, y_1), \dots, (\mathbf{z}_N, y_N)$

$$\text{Minimize } E_{\text{in}}(\mathbf{w}) = \frac{1}{N} \sum_{n=1}^N (\mathbf{w}^\top \mathbf{z}_n - y_n)^2$$

$$\text{Minimize } \frac{1}{N} (\mathbf{Z}\mathbf{w} - \mathbf{y})^\top (\mathbf{Z}\mathbf{w} - \mathbf{y})$$

$$\mathbf{w}_{\text{lin}} = (\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top \mathbf{y}$$

Constraining the weights

Hard constraint: \mathcal{H}_2 is constrained version of \mathcal{H}_{10} with $w_q = 0$ for $q > 2$

Softer version: $\sum_{q=0}^Q w_q^2 \leq C$ “soft-order” constraint

Minimize $\frac{1}{N} (\mathbf{Z}\mathbf{w} - \mathbf{y})^\top (\mathbf{Z}\mathbf{w} - \mathbf{y})$

subject to: $\mathbf{w}^\top \mathbf{w} \leq C$

Solution: \mathbf{w}_{reg} instead of \mathbf{w}_{lin}

Solving for \mathbf{w}_{reg}

$$\text{Minimize } E_{\text{in}}(\mathbf{w}) = \frac{1}{N} (\mathbf{Z}\mathbf{w} - \mathbf{y})^\top (\mathbf{Z}\mathbf{w} - \mathbf{y})$$

$$\text{subject to: } \mathbf{w}^\top \mathbf{w} \leq C$$

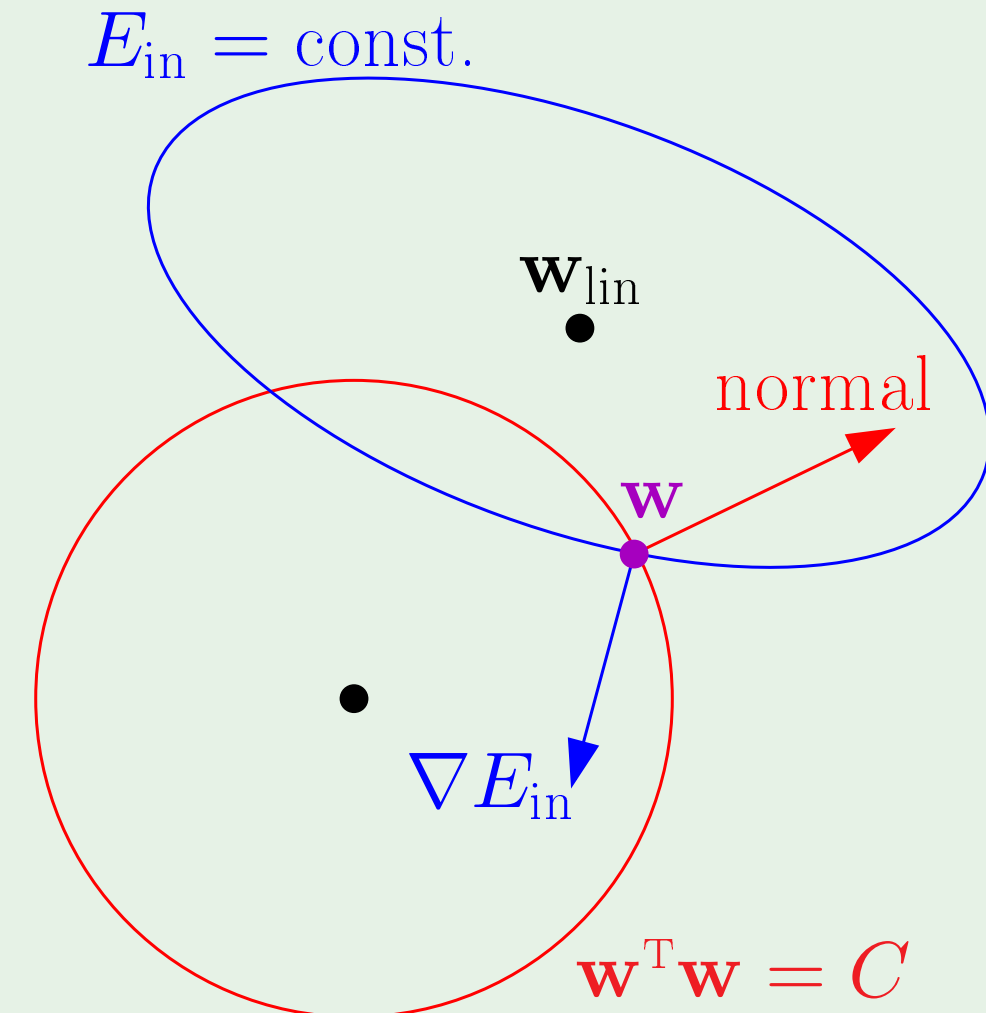
$$\nabla E_{\text{in}}(\mathbf{w}_{\text{reg}}) \propto -\mathbf{w}_{\text{reg}}$$

$$= -2\frac{\lambda}{N}\mathbf{w}_{\text{reg}}$$

$$\nabla E_{\text{in}}(\mathbf{w}_{\text{reg}}) + 2\frac{\lambda}{N}\mathbf{w}_{\text{reg}} = \mathbf{0}$$

$$\text{Minimize } E_{\text{in}}(\mathbf{w}) + \frac{\lambda}{N}\mathbf{w}^\top \mathbf{w}$$

$$\boxed{C \uparrow \quad \lambda \downarrow}$$



Augmented error

Minimizing $E_{\text{aug}}(\mathbf{w}) = E_{\text{in}}(\mathbf{w}) + \frac{\lambda}{N} \mathbf{w}^\top \mathbf{w}$

$$= \frac{1}{N} (\mathbf{Z}\mathbf{w} - \mathbf{y})^\top (\mathbf{Z}\mathbf{w} - \mathbf{y}) + \frac{\lambda}{N} \mathbf{w}^\top \mathbf{w} \quad \text{unconditionally}$$

— solves —

Minimizing $E_{\text{in}}(\mathbf{w}) = \frac{1}{N} (\mathbf{Z}\mathbf{w} - \mathbf{y})^\top (\mathbf{Z}\mathbf{w} - \mathbf{y})$

subject to: $\mathbf{w}^\top \mathbf{w} \leq C$ \longleftarrow VC formulation

The solution

Minimize $E_{\text{aug}}(\mathbf{w}) = E_{\text{in}}(\mathbf{w}) + \frac{\lambda}{N} \mathbf{w}^T \mathbf{w}$

$$= \frac{1}{N} \left((Z\mathbf{w} - \mathbf{y})^T (Z\mathbf{w} - \mathbf{y}) + \lambda \mathbf{w}^T \mathbf{w} \right)$$

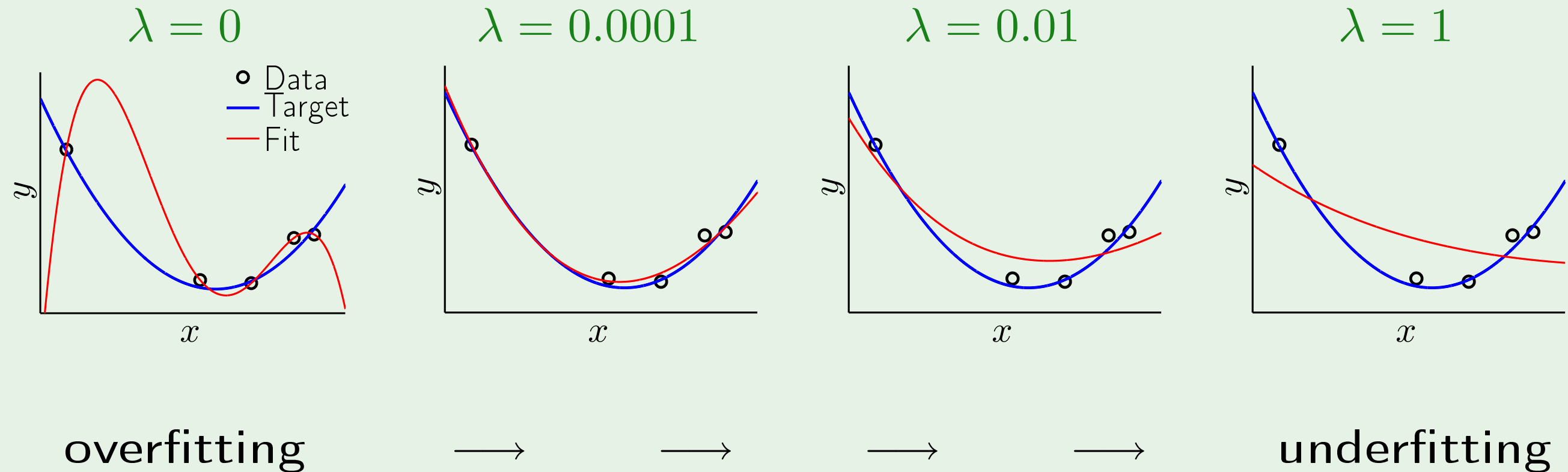
$$\nabla E_{\text{aug}}(\mathbf{w}) = \mathbf{0} \quad \implies \quad Z^T(Z\mathbf{w} - \mathbf{y}) + \lambda \mathbf{w} = \mathbf{0}$$

$$\boxed{\mathbf{w}_{\text{reg}} = (Z^T Z + \lambda I)^{-1} Z^T \mathbf{y}} \quad (\text{with regularization})$$

as opposed to $\mathbf{w}_{\text{lin}} = (Z^T Z)^{-1} Z^T \mathbf{y}$ (without regularization)

The result

Minimizing $E_{\text{in}}(\mathbf{w}) + \frac{\lambda}{N} \mathbf{w}^T \mathbf{w}$ for different λ 's:



Weight 'decay'

Minimizing $E_{\text{in}}(\mathbf{w}) + \frac{\lambda}{N} \mathbf{w}^T \mathbf{w}$ is called weight *decay*. Why?

Gradient descent:

$$\mathbf{w}(t+1) = \mathbf{w}(t) - \eta \nabla E_{\text{in}}(\mathbf{w}(t)) - 2\eta \frac{\lambda}{N} \mathbf{w}(t)$$

$$= \mathbf{w}(t) \left(1 - 2\eta \frac{\lambda}{N}\right) - \eta \nabla E_{\text{in}}(\mathbf{w}(t))$$

Applies in neural networks:

$$\mathbf{w}^T \mathbf{w} = \sum_{l=1}^L \sum_{i=0}^{d^{(l-1)}} \sum_{j=1}^{d^{(l)}} \left(w_{ij}^{(l)}\right)^2$$

Variations of weight decay

Emphasis of certain weights:

$$\sum_{q=0}^Q \gamma_q w_q^2$$

Examples:

$$\gamma_q = 2^q \implies \text{low-order fit}$$

$$\gamma_q = 2^{-q} \implies \text{high-order fit}$$

Neural networks: different layers get different γ 's

Tikhonov regularizer: $\mathbf{w}^\top \mathbf{\Gamma}^\top \mathbf{\Gamma} \mathbf{w}$

Even weight growth!

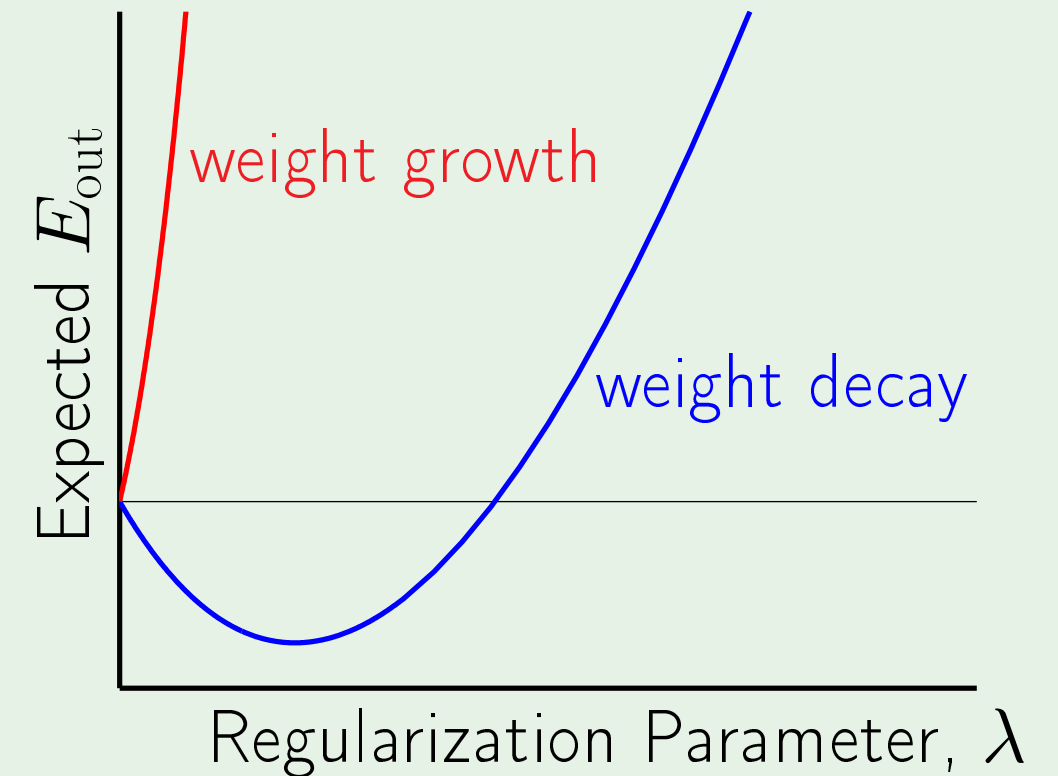
We 'constrain' the weights to be large - bad!

Practical rule:

stochastic noise is 'high-frequency'

deterministic noise is also non-smooth

⇒ constrain learning towards smoother hypotheses



General form of augmented error

Calling the regularizer $\Omega = \Omega(h)$, we minimize

$$E_{\text{aug}}(h) = E_{\text{in}}(h) + \frac{\lambda}{N} \Omega(h)$$

Rings a bell?

↓ ↓

$$E_{\text{out}}(h) \leq E_{\text{in}}(h) + \Omega(\mathcal{H})$$

E_{aug} is better than E_{in} as a proxy for E_{out}

Outline

- Regularization - informal
- Regularization - formal
- Weight decay
- Choosing a regularizer

The perfect regularizer Ω

Constraint in the 'direction' of the target function (going in circles 😊)

Guiding principle:

Direction of **smoother** or “simpler”

Chose a bad Ω ?

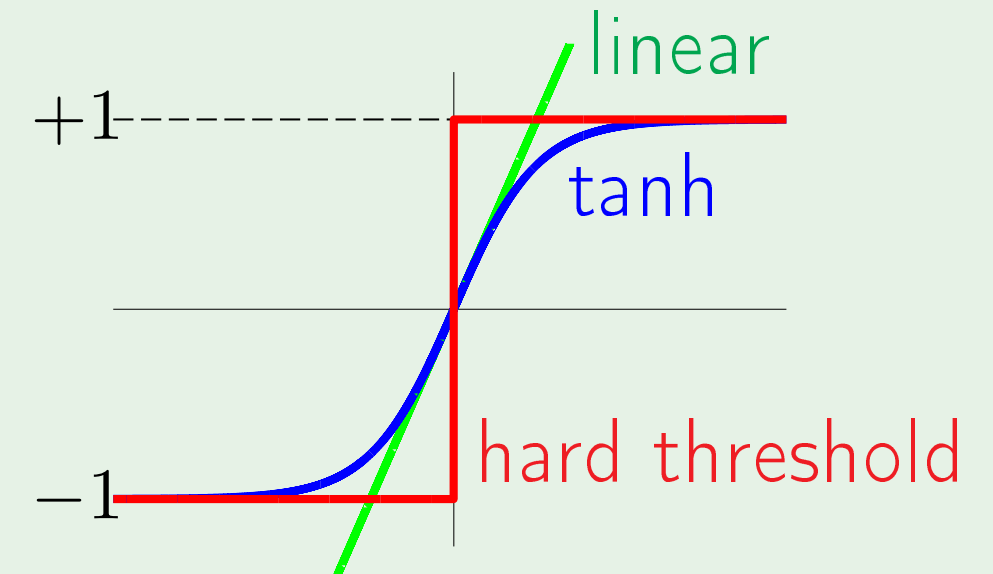
We still have λ !

Neural-network regularizers

Weight decay: From linear to logical

Weight elimination:

Fewer weights \implies smaller VC dimension



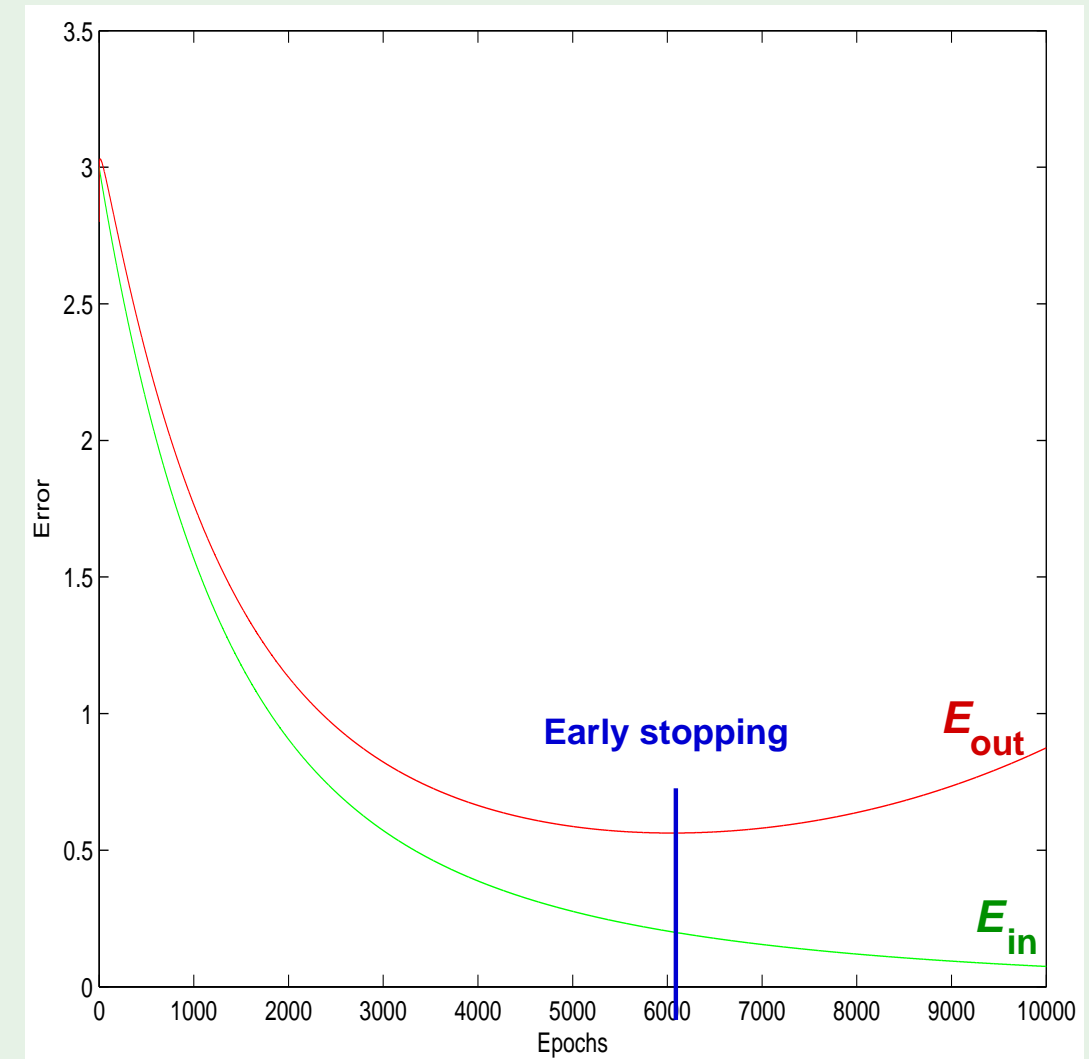
Soft weight elimination:

$$\Omega(\mathbf{w}) = \sum_{i,j,l} \frac{\left(w_{ij}^{(l)}\right)^2}{\beta^2 + \left(w_{ij}^{(l)}\right)^2}$$

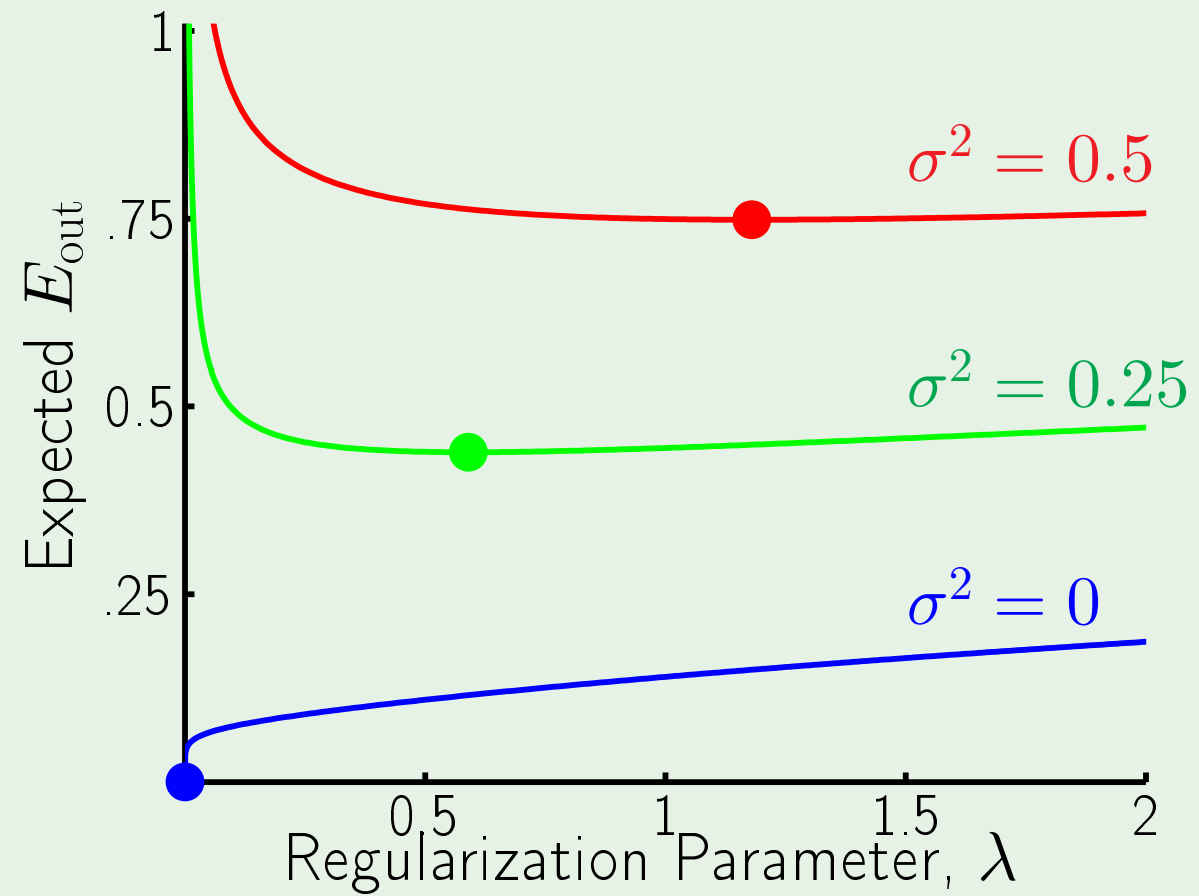
Early stopping as a regularizer

Regularization through the optimizer!

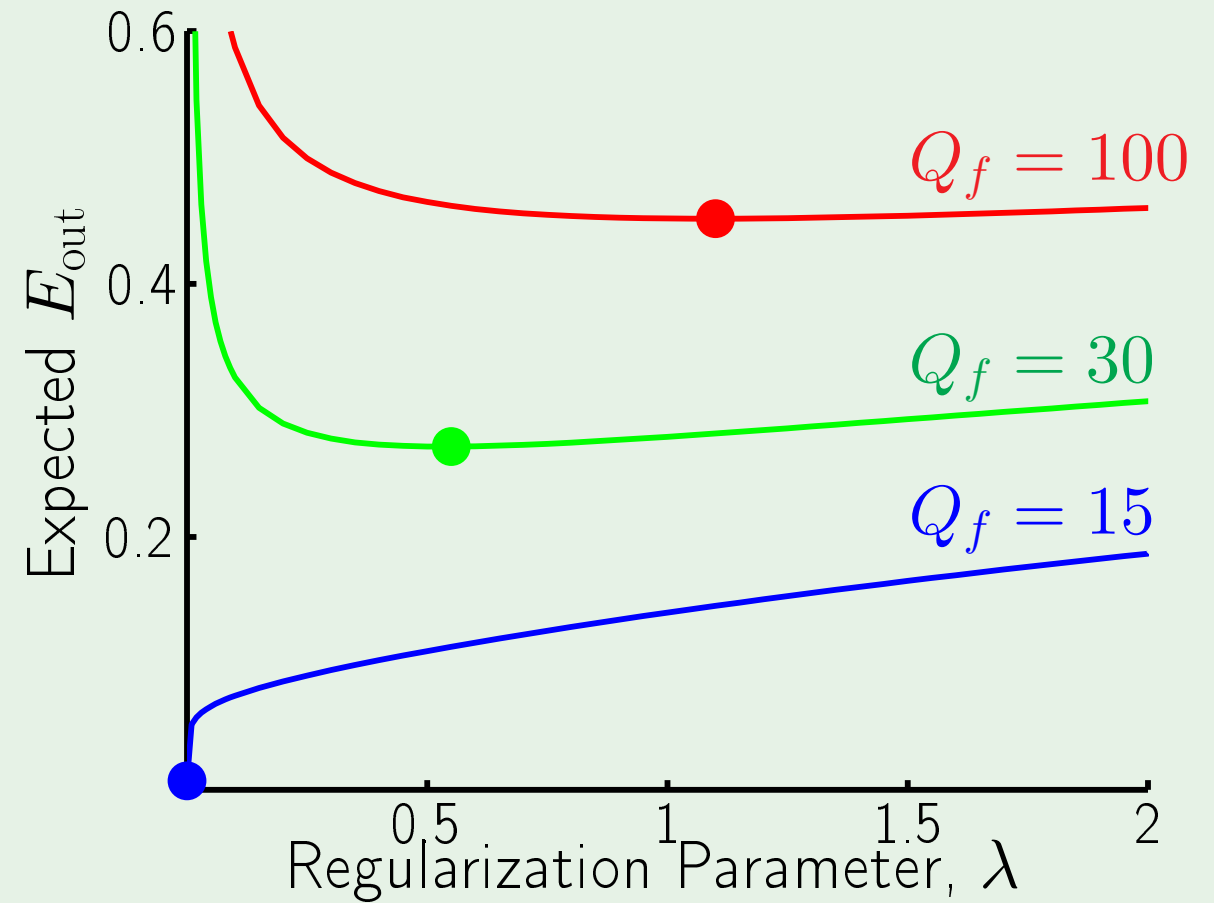
When to stop? **validation**



The optimal λ



Stochastic noise



Deterministic noise