Review of Lecture 11

Fitting the noise, stochastic/deterministic

- **Overfitting**

  Fitting the data more than is warranted
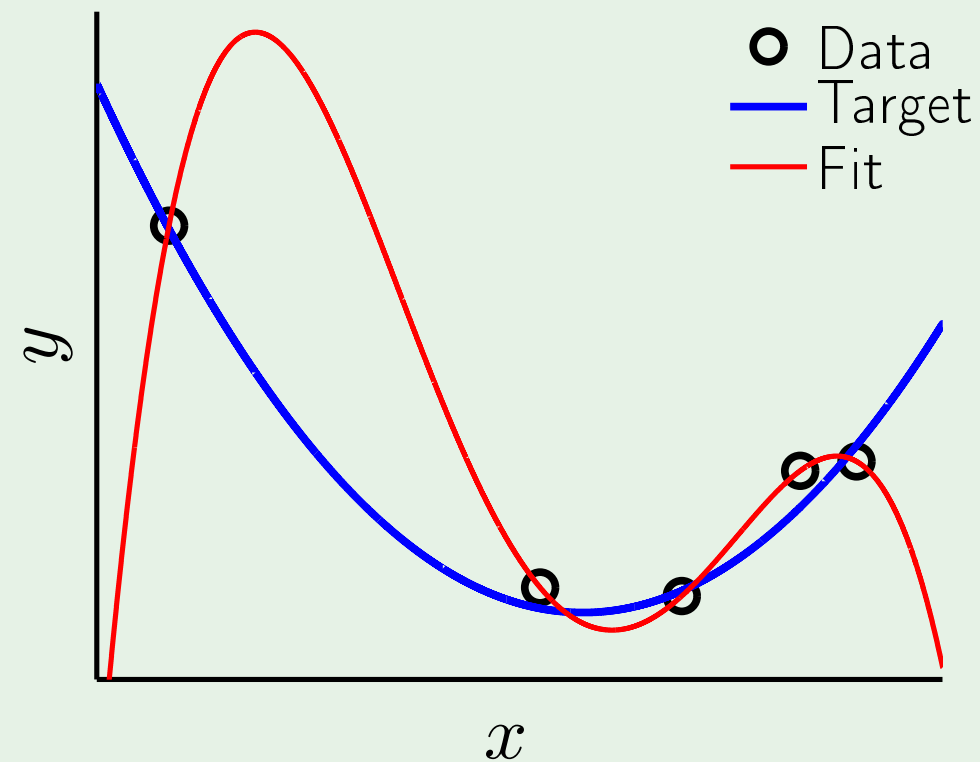
  VC allows it; doesn't predict it

- **Deterministic noise**
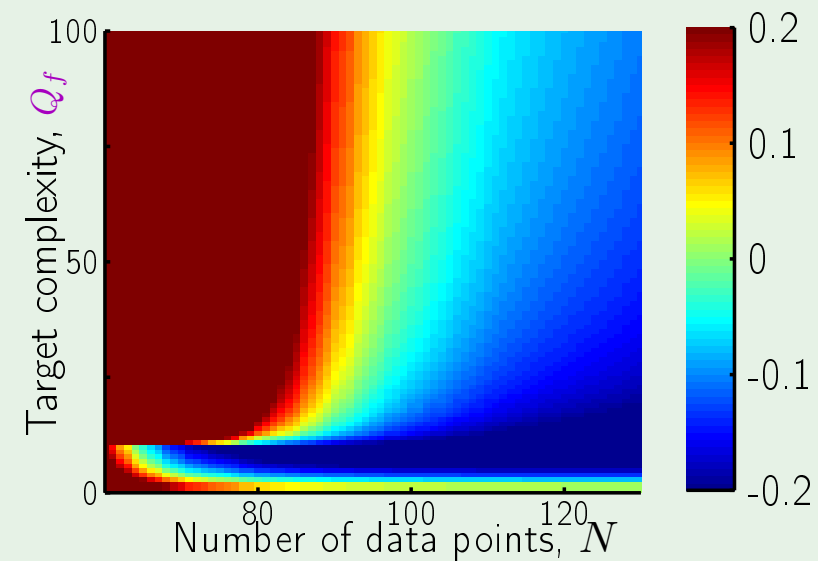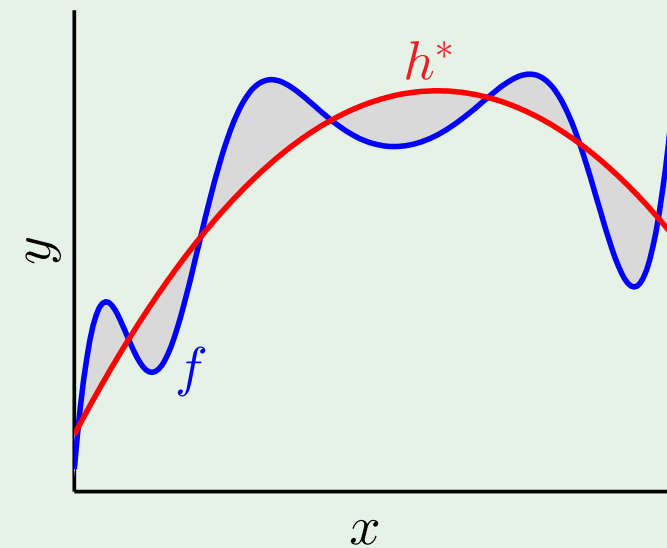
Learning From Data

Yaser S. Abu-Mostafa
California Institute of Technology

Lecture 12: **Regularization**
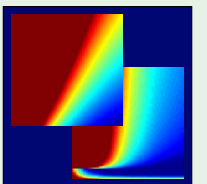
Outline

- Regularization - informal

- Regularization - formal

- Weight decay

- Choosing a regularizer

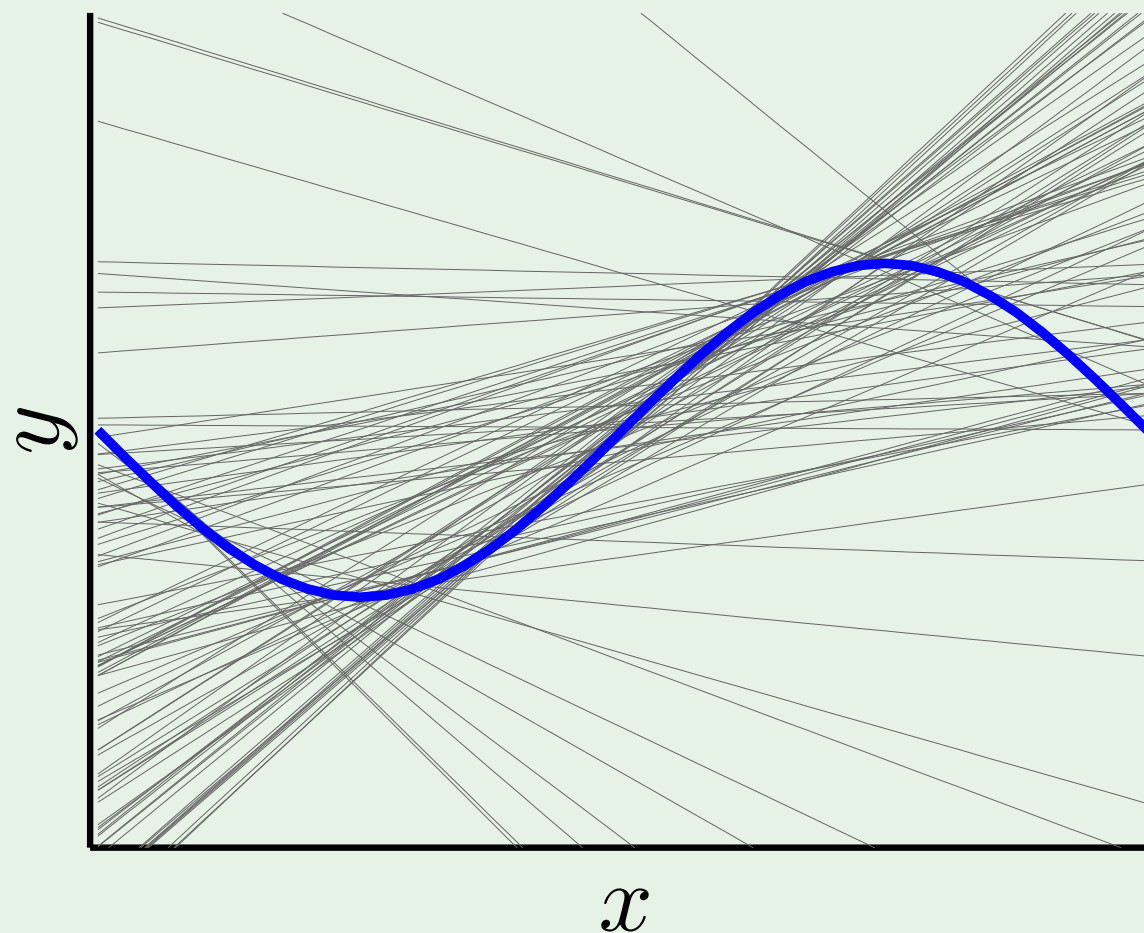# Two approaches to regularization

Mathematical:

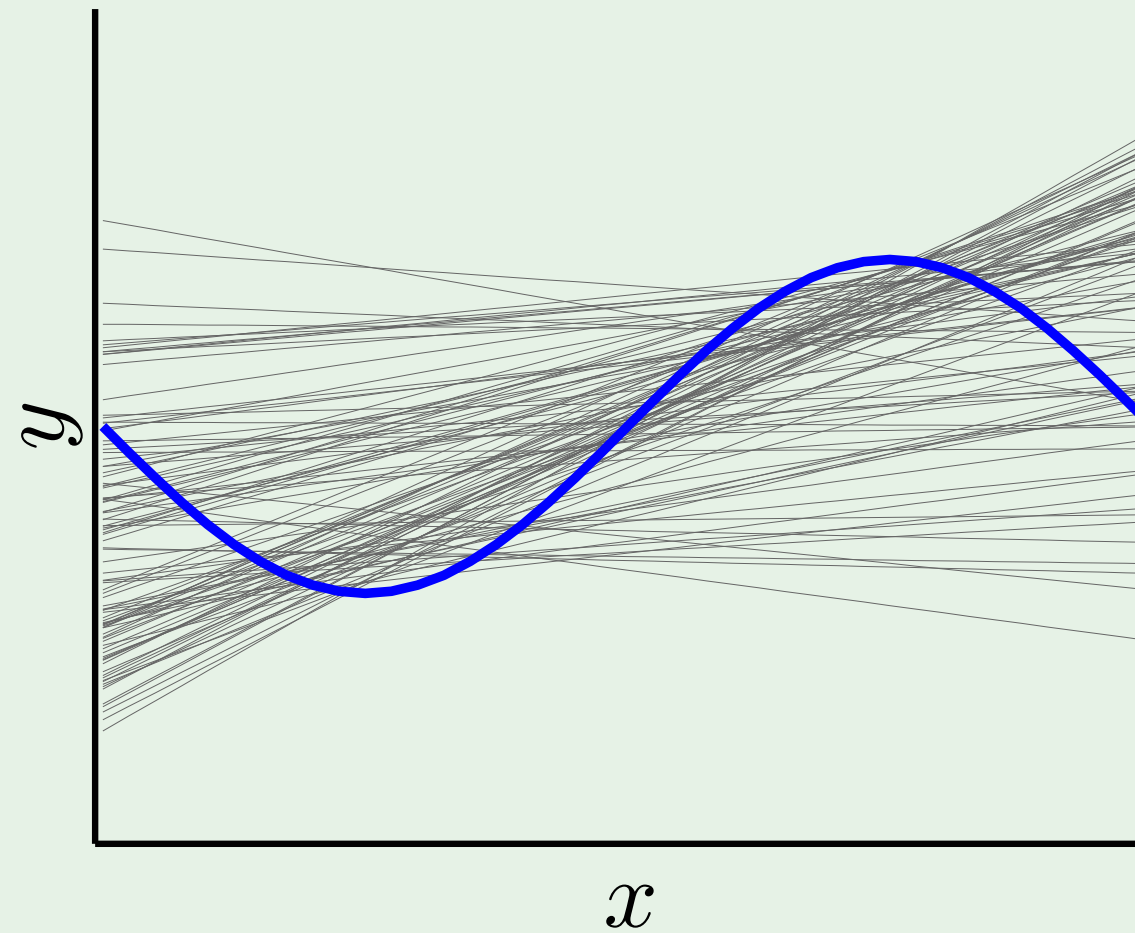Ill-posed problems in function approximation

Heuristic:

Handicapping the minimization of E_{in}

A familiar example

without regularization

with regularization

and the winner is ...

without regularization
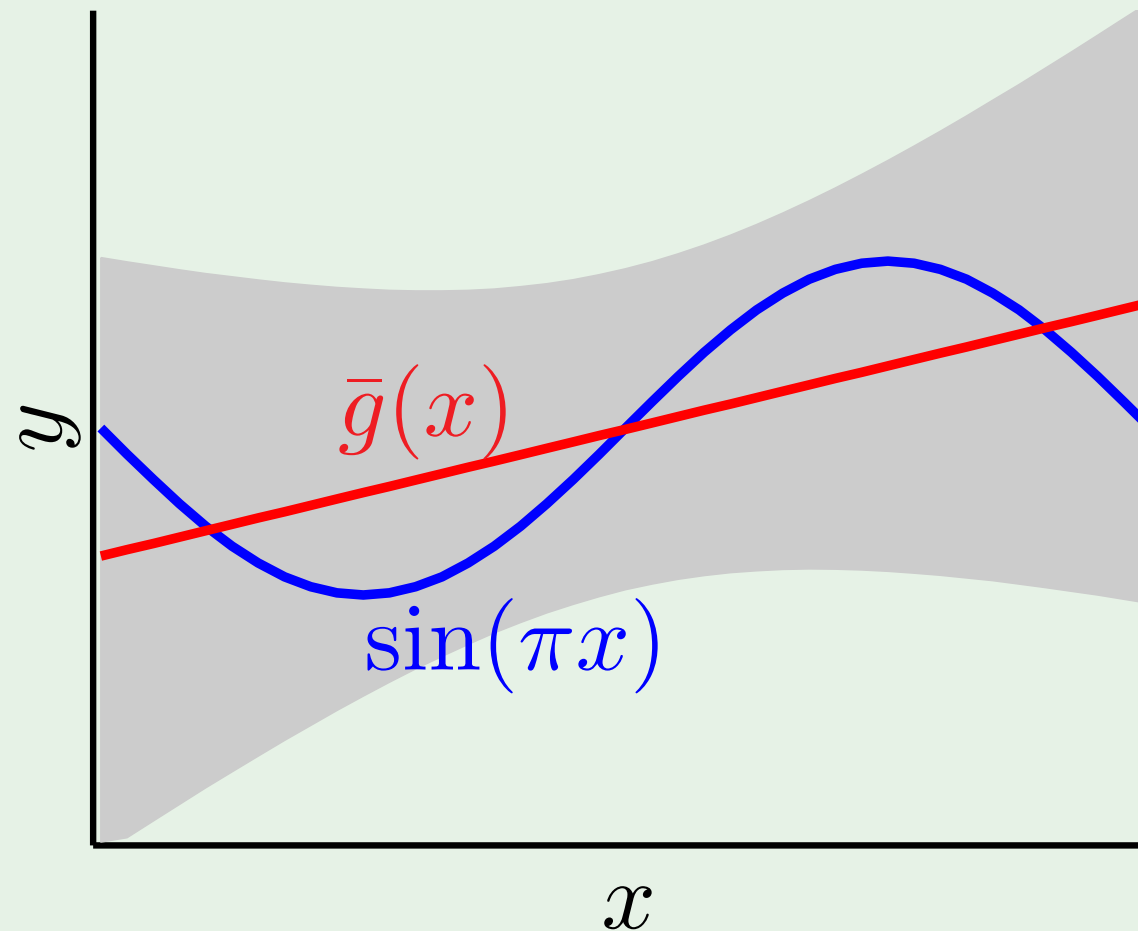
with regularization

bias = **0.21**   var = **1.69**
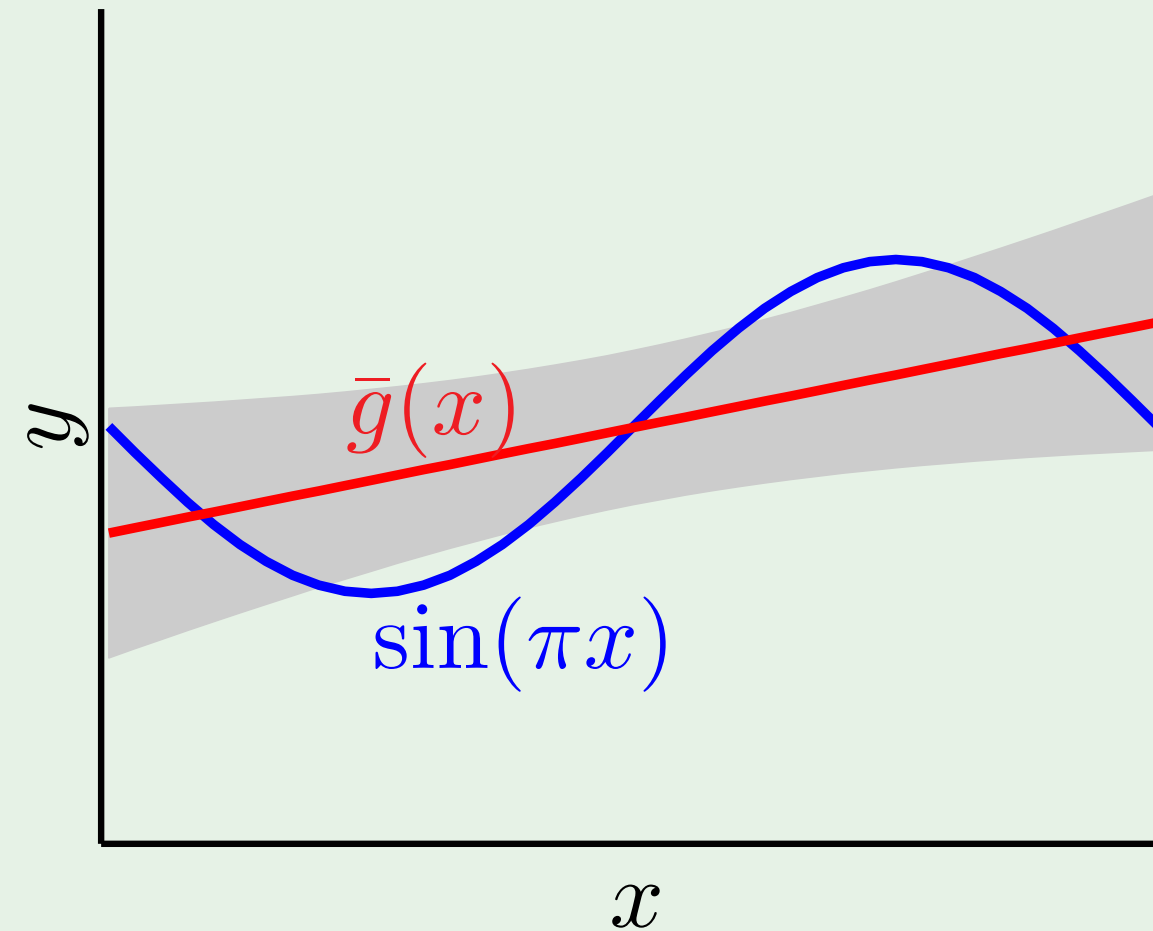
bias = **0.23**   var = **0.33**

The polynomial model

$\mathcal{H}_Q$: polynomials of order Q linear regression in $\mathcal{Z}$ space

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

Legendre polynomials:

L_1	L_2	L_3	L_4	L_5
x	$\frac{1}{2}(3x^2 - 1)$	$\frac{1}{2}(5x^3 - 3x)$	$\frac{1}{8}(35x^4 - 30x^2 + 3)$	$\frac{1}{8}(63x^5 \cdots)$
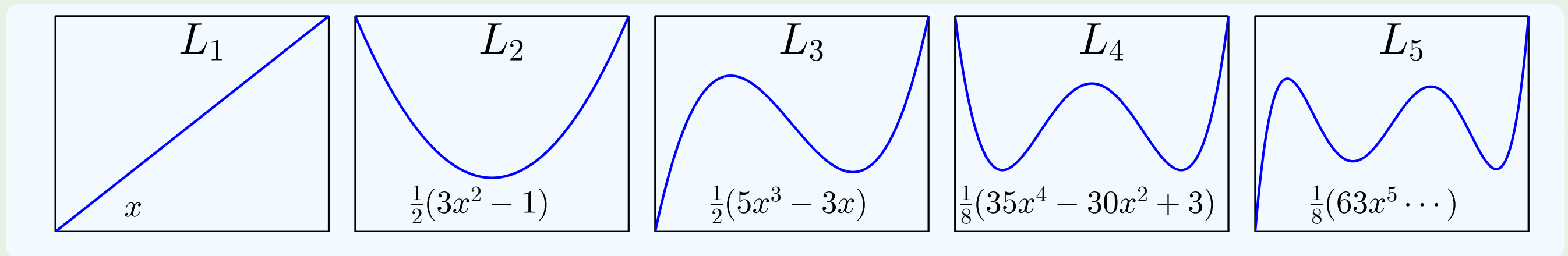
Unconstrained solution

Given $\quad (x_1, y_1), \cdots, (x_N, y_n) \quad \longrightarrow \quad (\mathbf{z}_1, y_1), \cdots, (\mathbf{z}_N, y_n)$

Minimize $\quad E_{\text{in}}(\mathbf{w}) = \dfrac{1}{N} \sum_{n=1}^{N} (\mathbf{w}^{\mathsf{T}} \mathbf{z}_n - y_n)^2$

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

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

Constraining the weights

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

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

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

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

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

Solving for $\mathbf{w}_{\text{reg}}$

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

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

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

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

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

Minimize $\quad E_{\text{in}}(\mathbf{w}) + \frac{\lambda}{N}\mathbf{w}^\top\mathbf{w}$ $\qquad \boxed{C \uparrow \quad \lambda \downarrow}$

$E_{\text{in}} = \text{const.}$

$\mathbf{w}_{\text{lin}}$

normal

$\mathbf{w}$

∇E_{in}

$\mathbf{w}^\top\mathbf{w} = C$
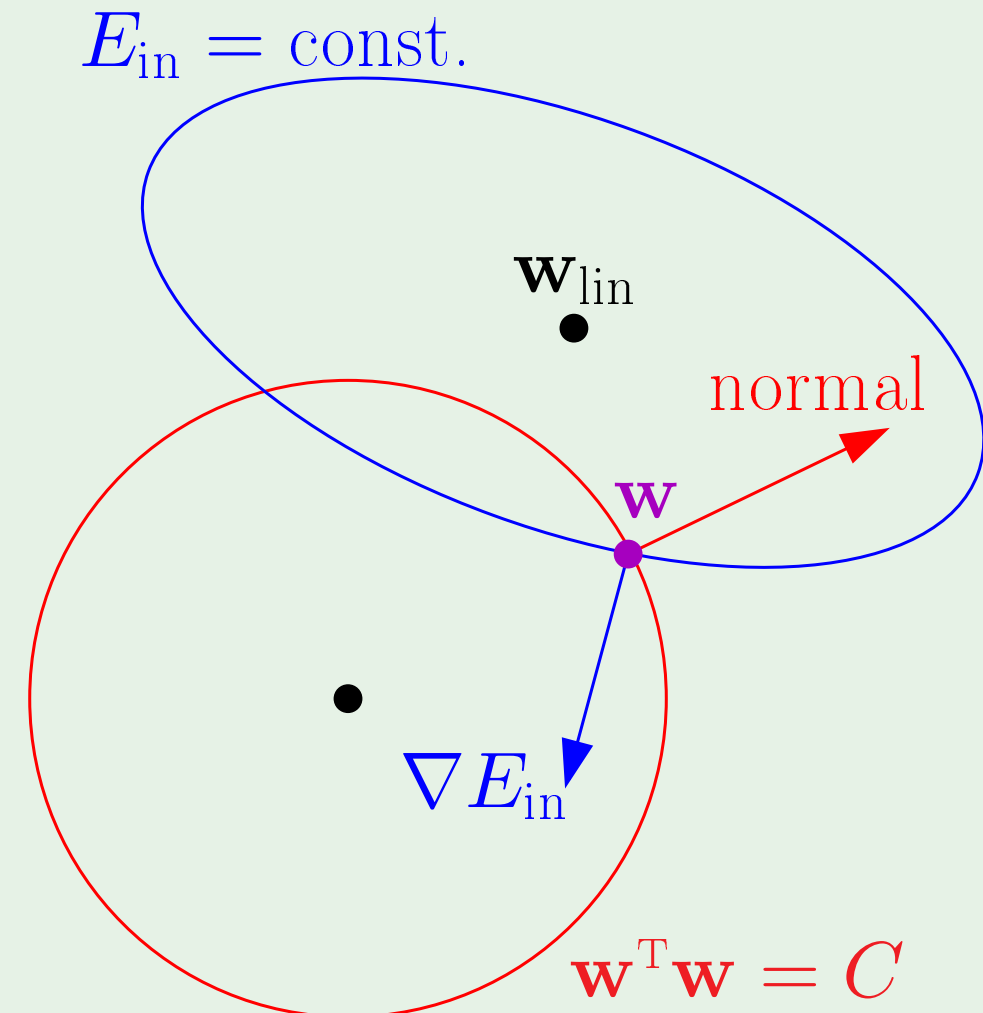
Augmented error

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

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

$$- \text{ solves } -$$

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

$$\text{subject to: } \quad \mathbf{w}^{\intercal}\mathbf{w} \leq C \qquad\qquad \longleftarrow \text{ VC formulation}$$

The solution

Minimize
$$E_{\text{aug}}(\mathbf{w}) \;=\; E_{\text{in}}(\mathbf{w}) + \tfrac{\lambda}{N}\mathbf{w}^{\!\top}\mathbf{w}$$

$$= \;\frac{1}{N}\,\Big((\mathbf{Z}\mathbf{w} - \mathbf{y})^{\!\top}(\mathbf{Z}\mathbf{w} - \mathbf{y}) \;+\; \lambda\,\mathbf{w}^{\!\top}\mathbf{w}\Big)$$
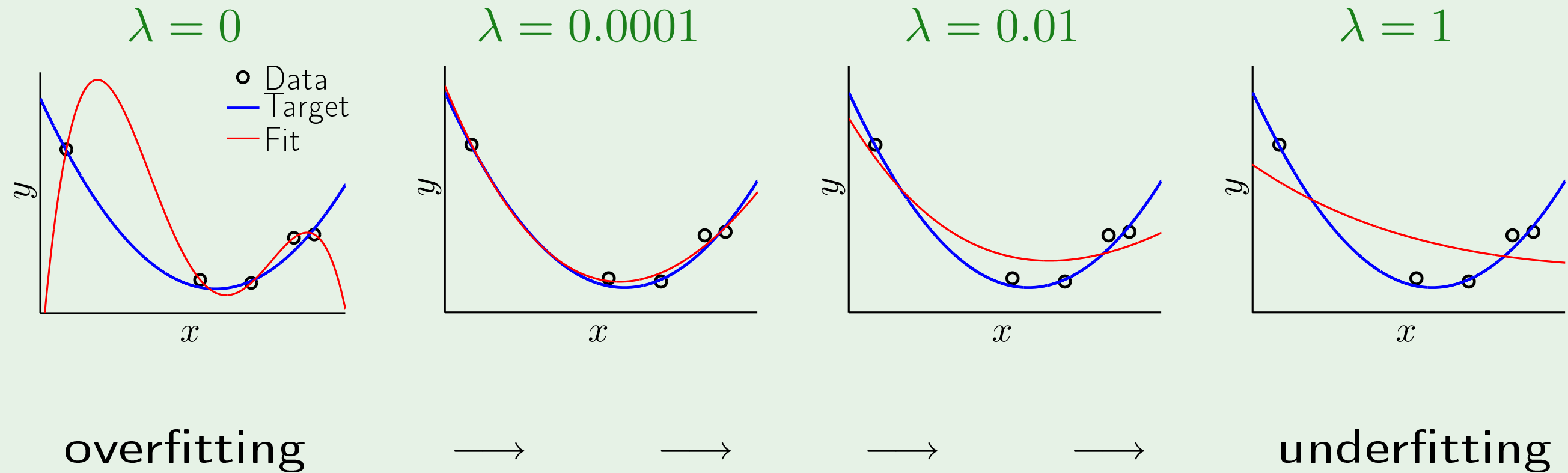
$$\nabla E_{\text{aug}}(\mathbf{w}) \;=\; \mathbf{0} \qquad \Longrightarrow \qquad \mathbf{Z}^{\!\top}(\mathbf{Z}\mathbf{w} - \mathbf{y}) + \lambda\mathbf{w} = \mathbf{0}$$

$$\boxed{\;\mathbf{w}_{\text{reg}} \;=\; (\mathbf{Z}^{\!\top}\mathbf{Z} + \lambda\mathbf{I})^{-1}\,\mathbf{Z}^{\!\top}\mathbf{y}\;} \qquad \text{(with regularization)}$$

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

The result

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

$\lambda = 0 \qquad\qquad \lambda = 0.0001 \qquad\qquad \lambda = 0.01 \qquad\qquad \lambda = 1$

overfitting $\quad\longrightarrow\qquad\longrightarrow\qquad\longrightarrow\qquad\longrightarrow\quad$ **underfitting**

Weight 'decay'

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

Gradient descent:

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

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

Applies in neural networks:

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

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: $\quad \mathbf{w}^\mathsf{T} \Gamma^\mathsf{T} \Gamma \mathbf{w}$

Even weight growth!

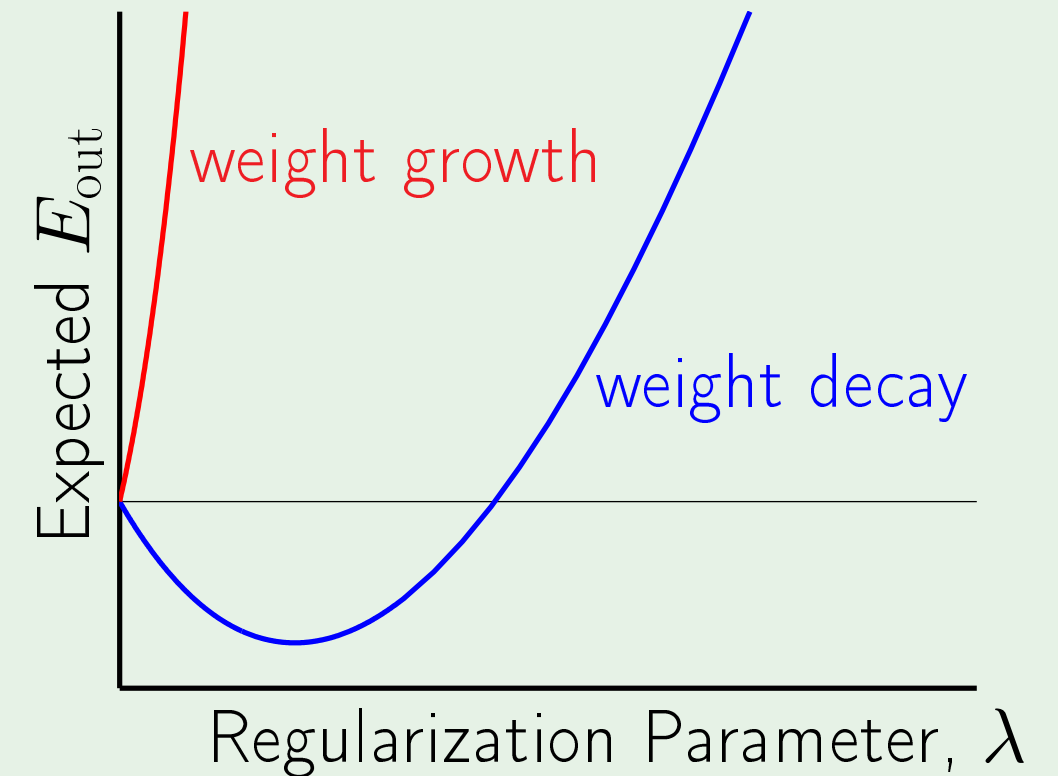We 'constrain' the weights to be large - bad!

Practical rule:

    stochastic noise is 'high-frequency'

    deterministic noise is also non-smooth

$\Longrightarrow$ constrain learning towards smoother hypotheses

General form of augmented error

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

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

Rings a bell? $\qquad\qquad \downarrow\downarrow$

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

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:
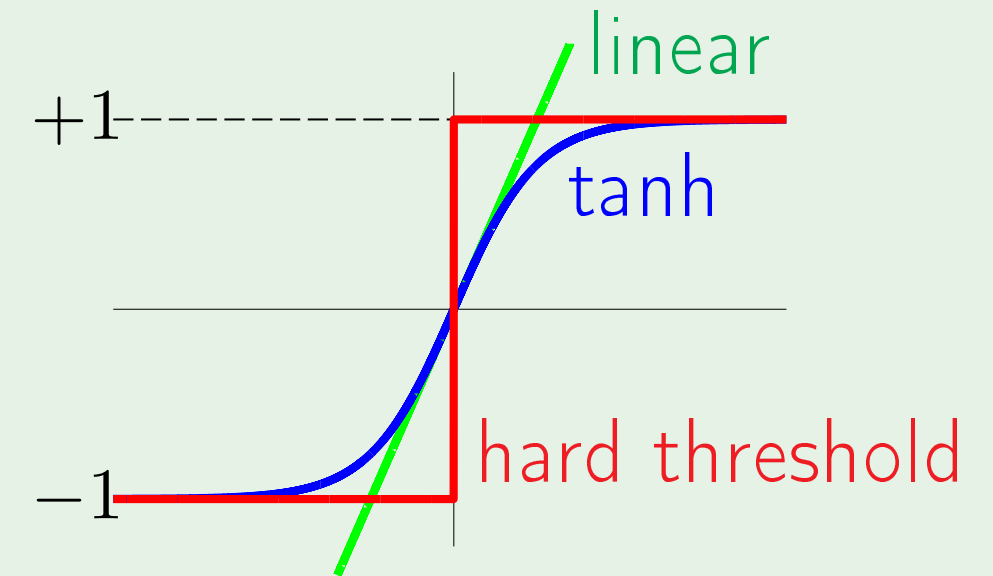
Direction of **smoother** or "simpler"

Chose a bad Ω?

We still have λ!

Neural-network regularizers

Weight decay: From linear to logical

Weight elimination:

Fewer weights $\Longrightarrow$ smaller VC dimension

linear

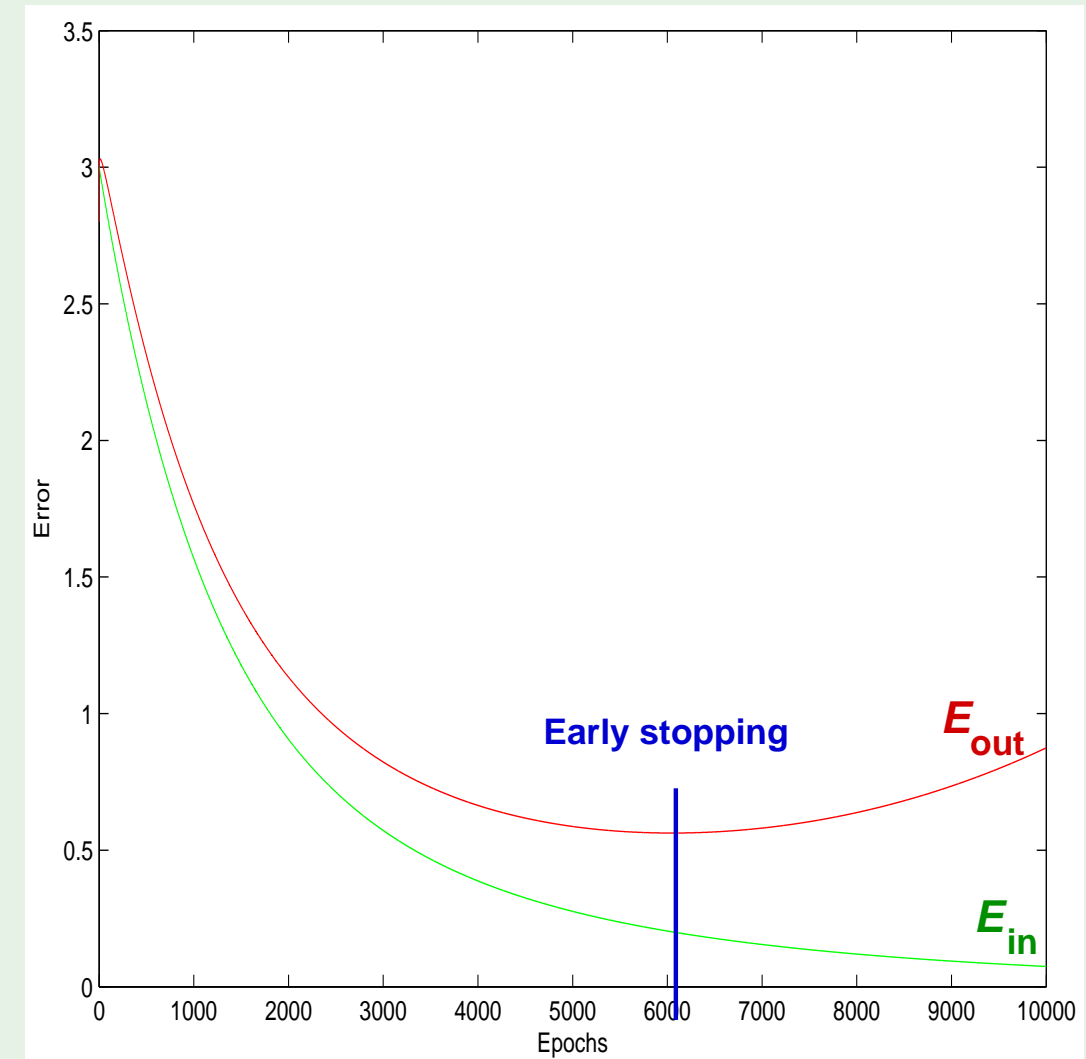$+1$

tanh

-1

hard threshold

Soft weight elimination:

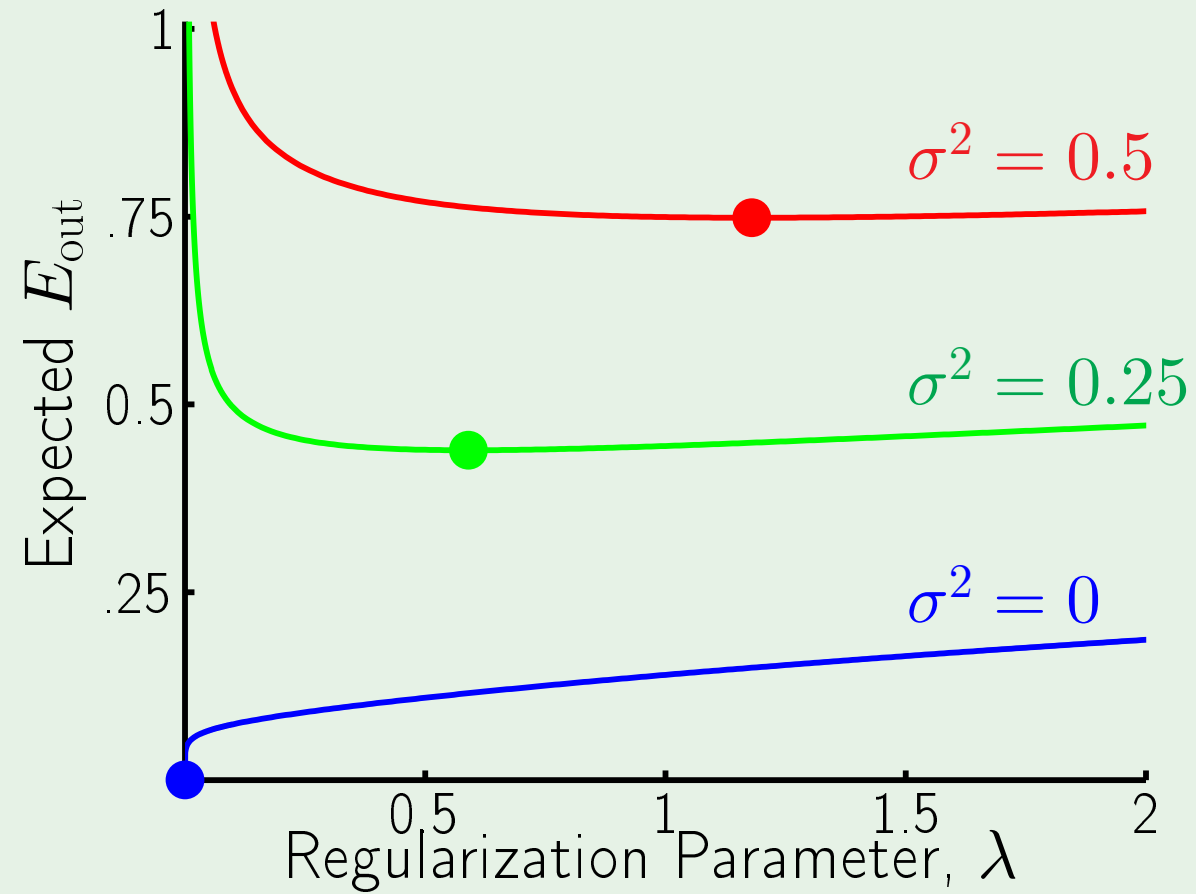$$\Omega(\mathbf{w}) = \sum_{i,j,l} \frac{\left(w_{ij}^{(l)}\right)^2}{\beta^2 + \left(w_{ij}^{(l)}\right)^2}$$

Early stopping as a regularizer
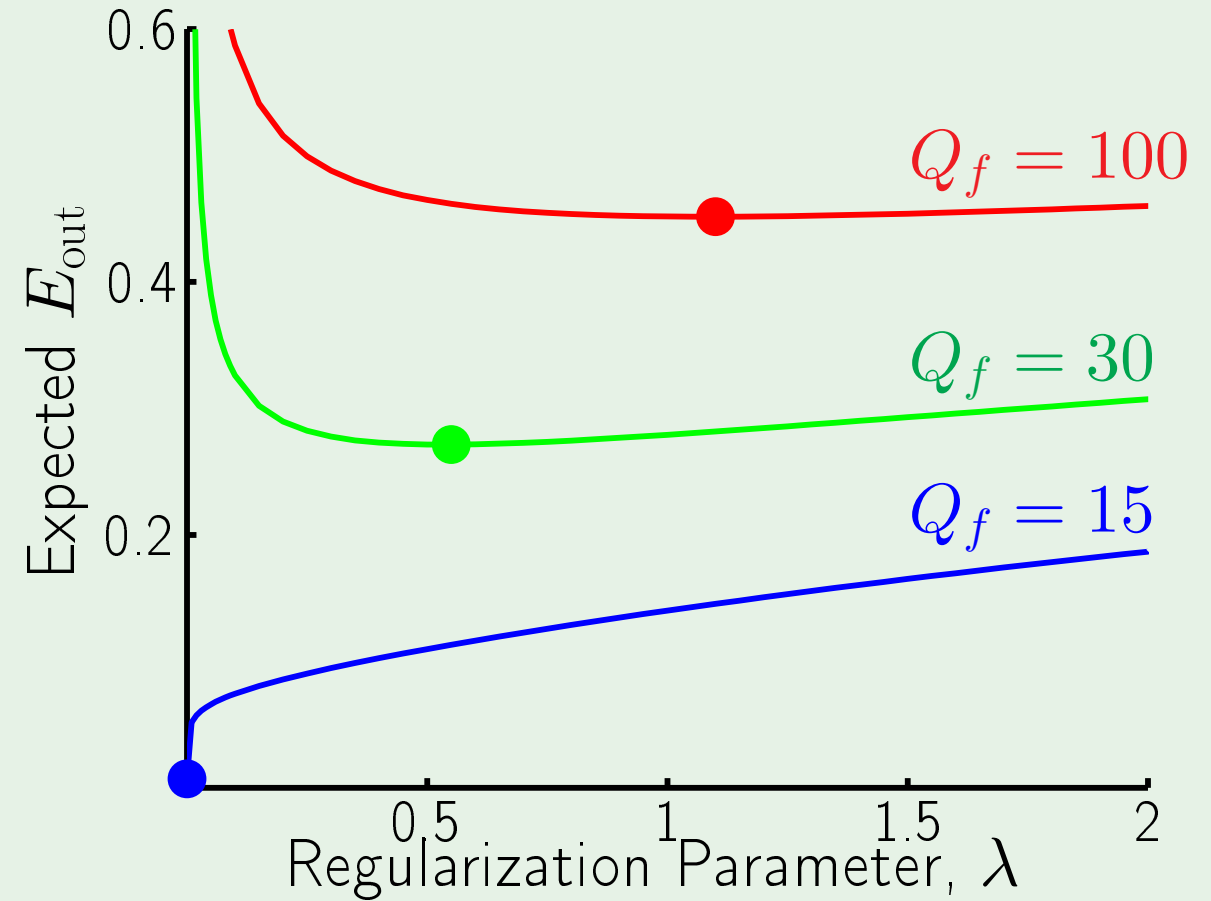
Regularization through the optimizer!

When to stop?   **validation**

The optimal λ

Stochastic noise

Deterministic noise