Outline

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

Two approaches to regularization

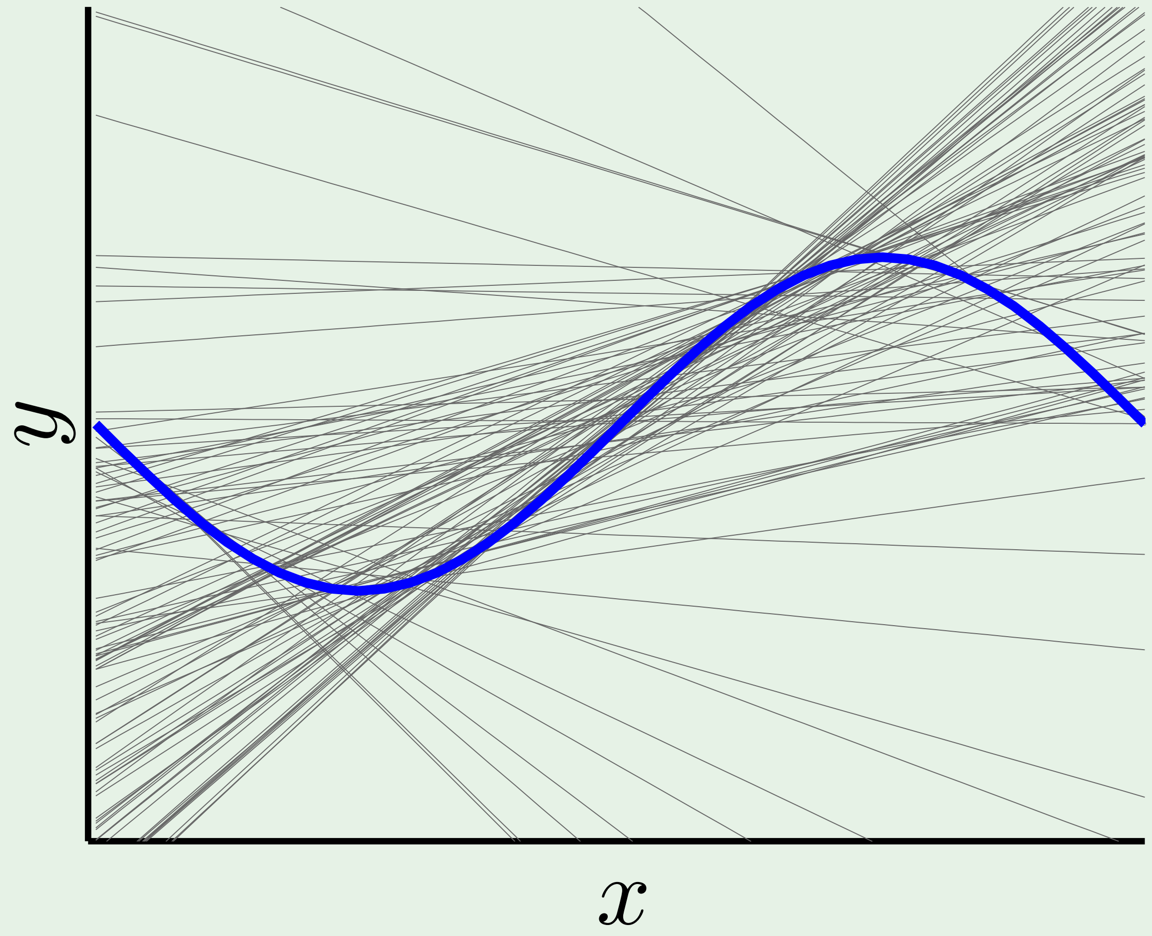
Mathematical:

Ill-posed problems in function approximation

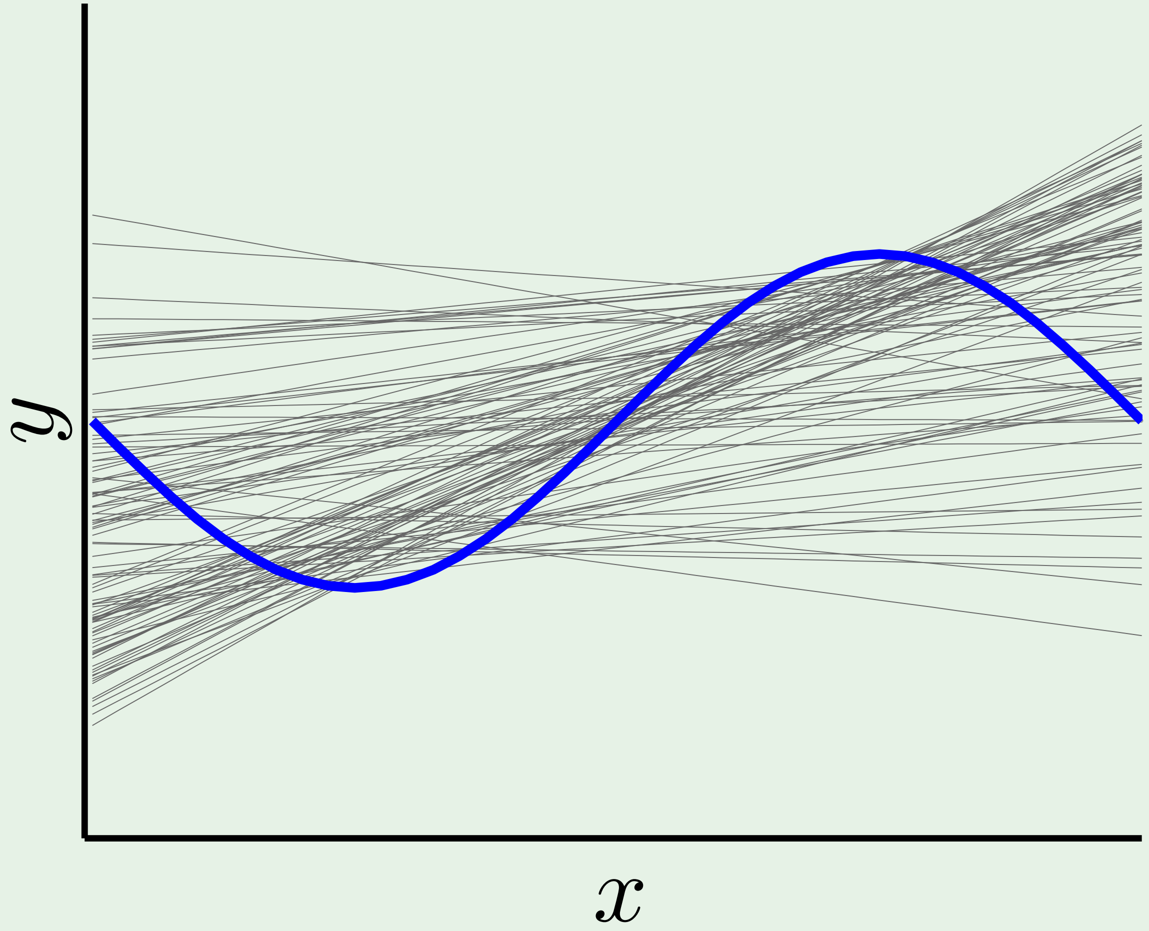
Heuristic:

Handicapping the minimization of E_{in}

A familiar example



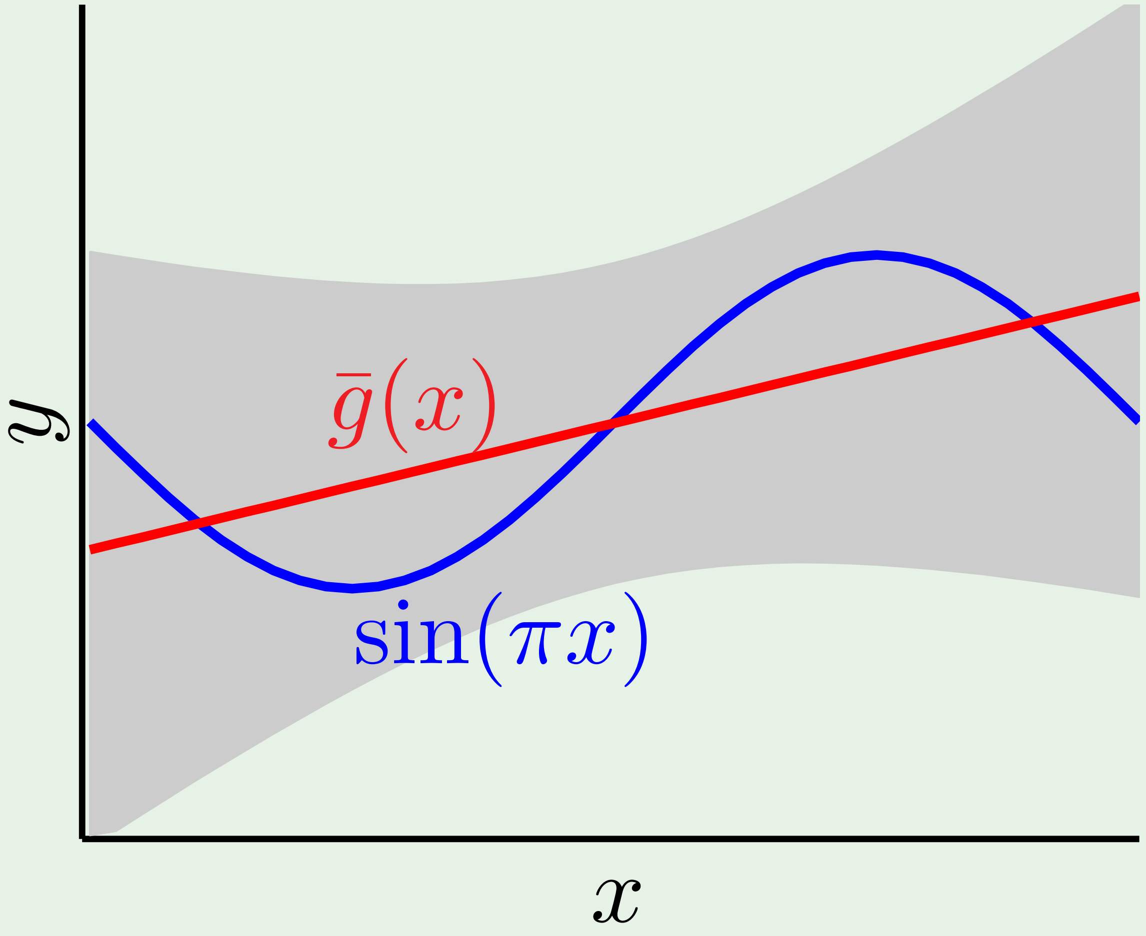
without regularization



with regularization

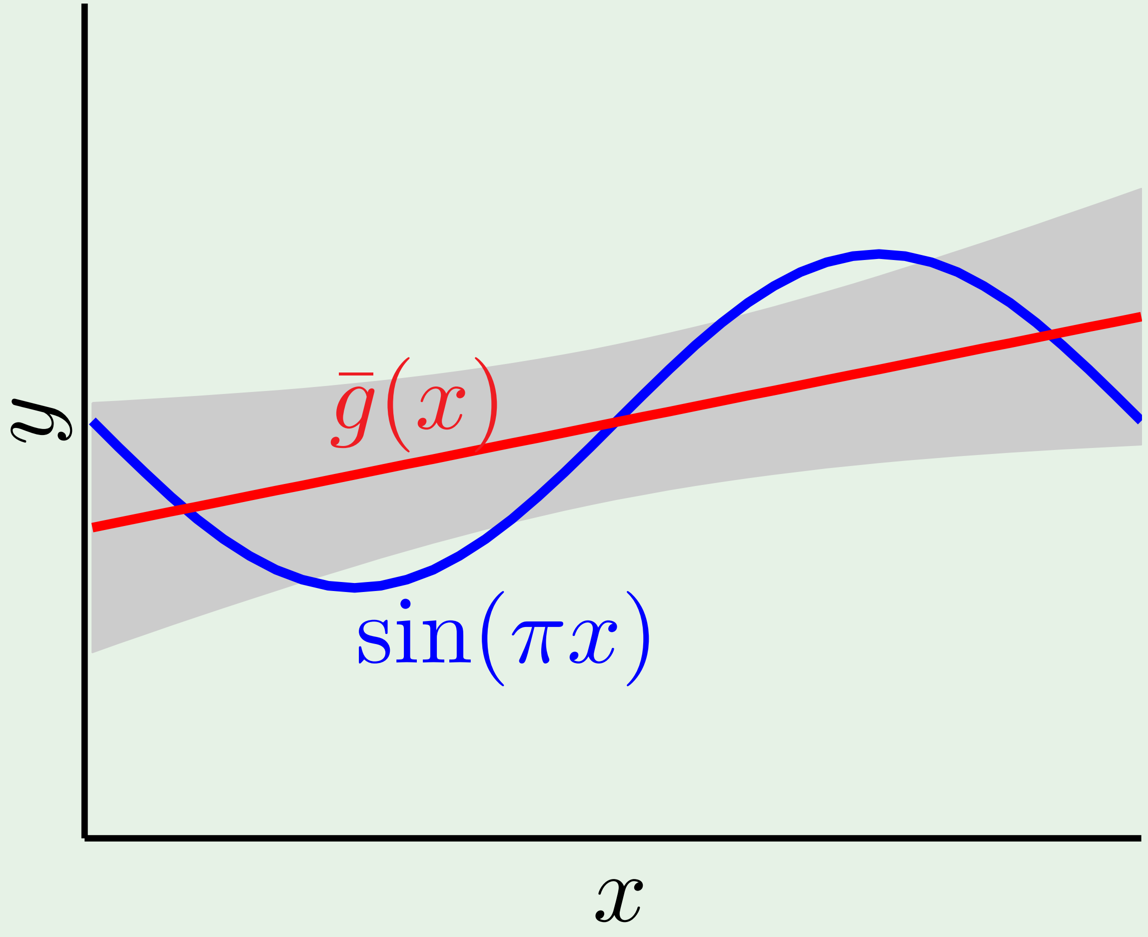
and the winner is ...

without regularization



bias $=$ **0.21** var $=$ **1.69**

with regularization



bias $=$ **0.23** var $=$ **0.33**

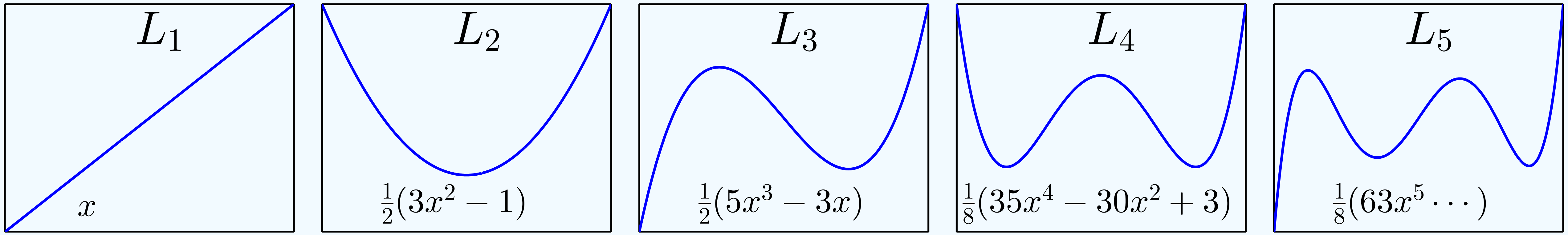
The polynomial model

$\mathcal{H}_{\mathrm{Q}}$: polynomials of order Q

linear regression in $\mathcal{Z}$ space

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

Legendre polynomials:



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

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

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

Constraining the weights

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

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

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

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

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

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

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

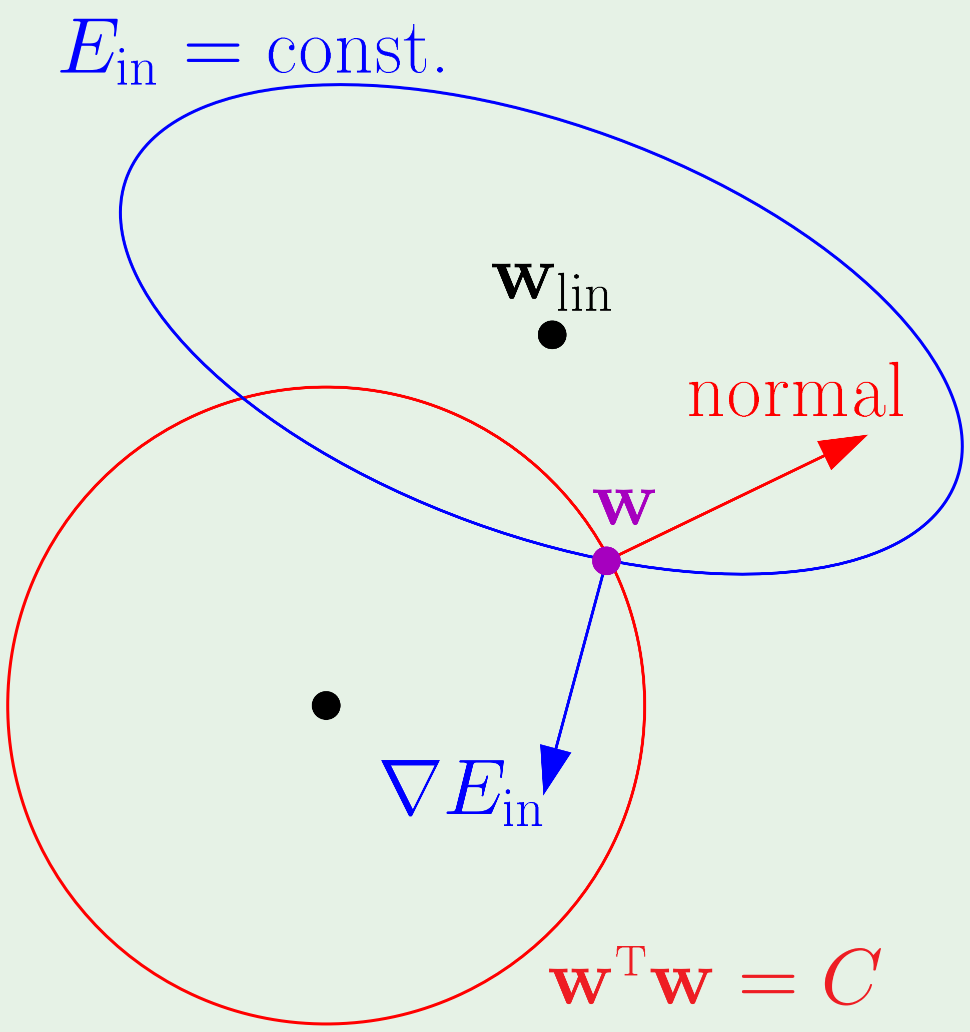
subject to: $\quad \mathbf{w}^{\mathsf{T}}\mathbf{w} \leq C$

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

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

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

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



Augmented error

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

$= \frac{1}{N}(\mathrm{Z}\mathbf{w} - \mathbf{y})^{\mathsf{T}}(\mathrm{Z}\mathbf{w} - \mathbf{y}) + \frac{\lambda}{N}\mathbf{w}^{\mathsf{T}}\mathbf{w}$ unconditionally

– solves –

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

subject to: $\mathbf{w}^{\mathsf{T}}\mathbf{w} \le C$ $\longleftarrow$ VC formulation

The solution

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

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

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

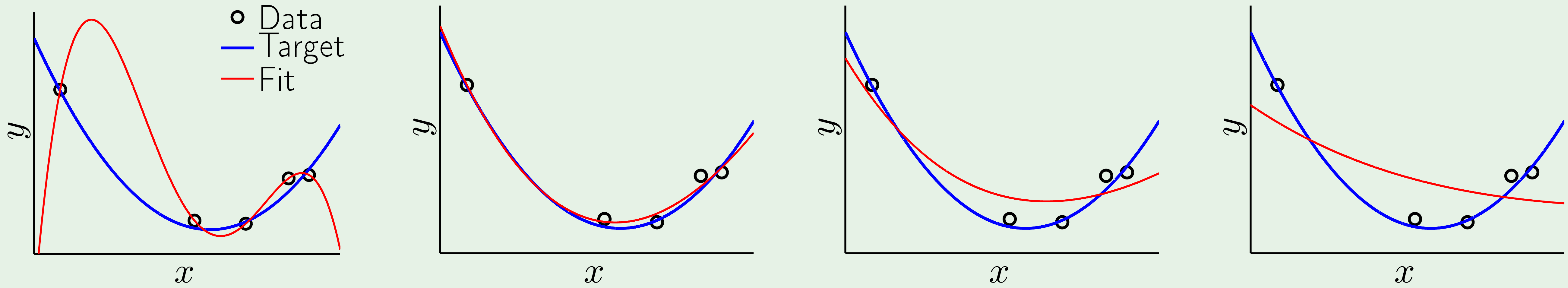
$$\boxed{\mathbf{w}_{\mathrm{reg}} = (\mathrm{Z}^{\mathsf{T}}\mathrm{Z} + \lambda\mathrm{I})^{-1}\,\mathrm{Z}^{\mathsf{T}}\mathbf{y}}$$ (with regularization)

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

The result

Minimizing $\quad E_{\text{in}}(\mathbf{w}) + \frac{\lambda}{N}\,\mathbf{w}^{\intercal}\mathbf{w} \quad$ for different λ's:

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



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

Weight 'decay'

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

Gradient descent:

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

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

Applies in neural networks:

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

Variations of weight decay

Emphasis of certain weights: $$\sum_{q=0}^{Q} \gamma_q \, w_q^2$$

Examples:

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

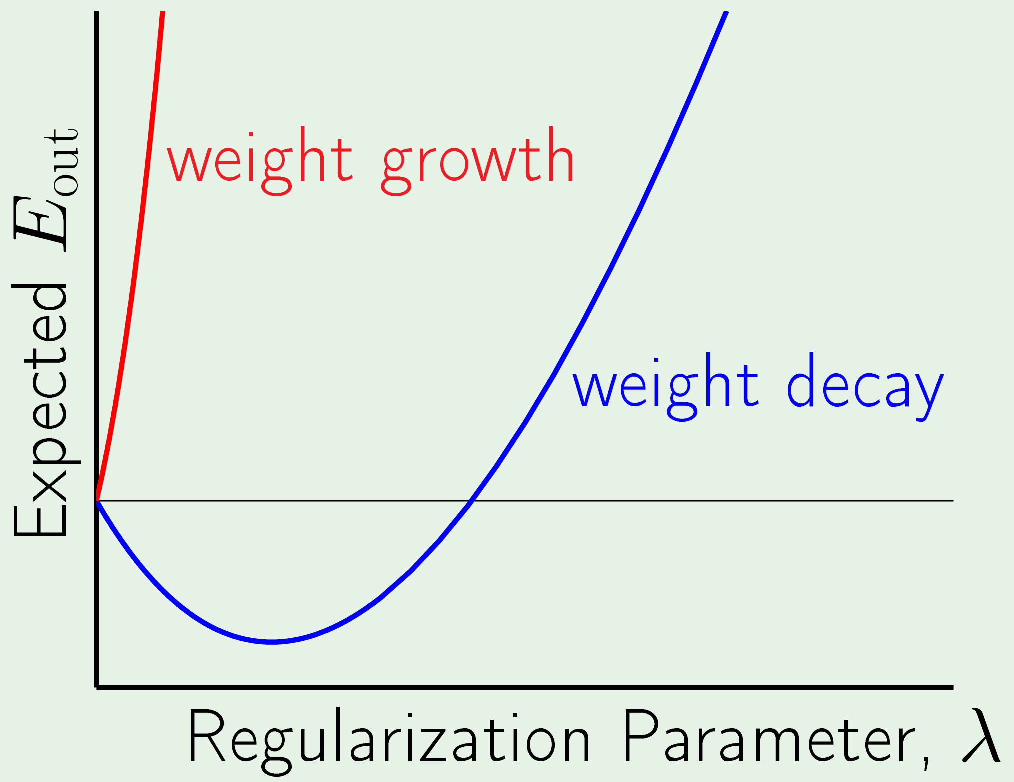
$$\gamma_q = 2^{-q} \implies \text{high-order fit}$$

Neural networks: different layers get different γ's

Tikhonov regularizer: $\mathbf{w}^{\mathsf{T}}\Gamma^{\mathsf{T}}\Gamma\mathbf{w}$

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

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

Rings a bell? $\downarrow\downarrow$

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

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:

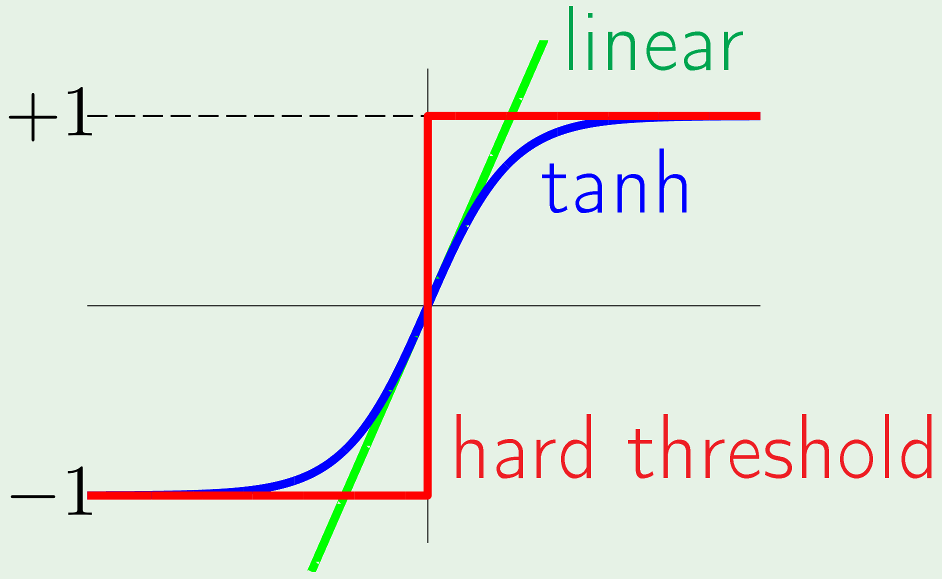
Direction of **smoother** or "simpler"

Chose a bad Ω?

We still have λ!

Neural-network regularizers

Weight decay: From linear to logical



Weight elimination:

Fewer weights $\Longrightarrow$ smaller VC dimension

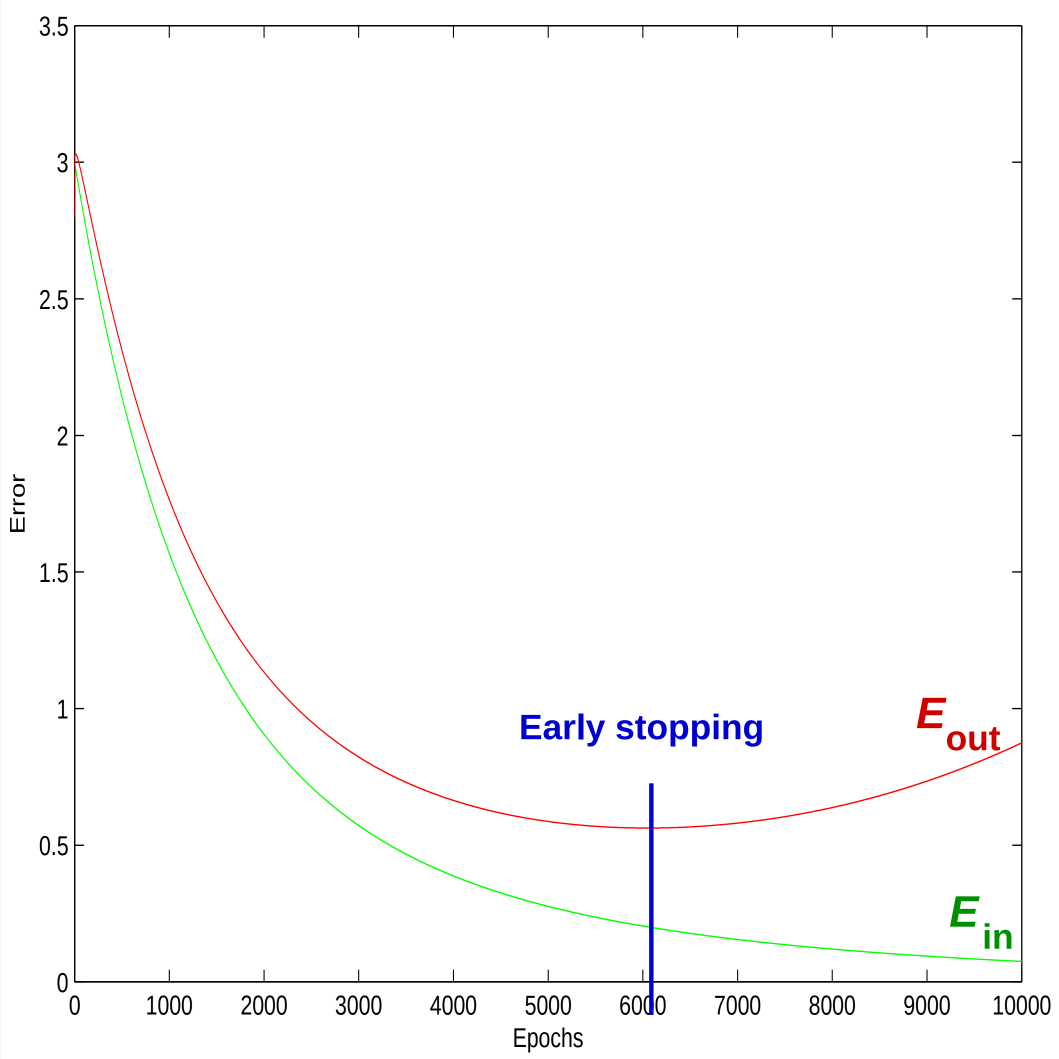
Soft weight elimination:

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

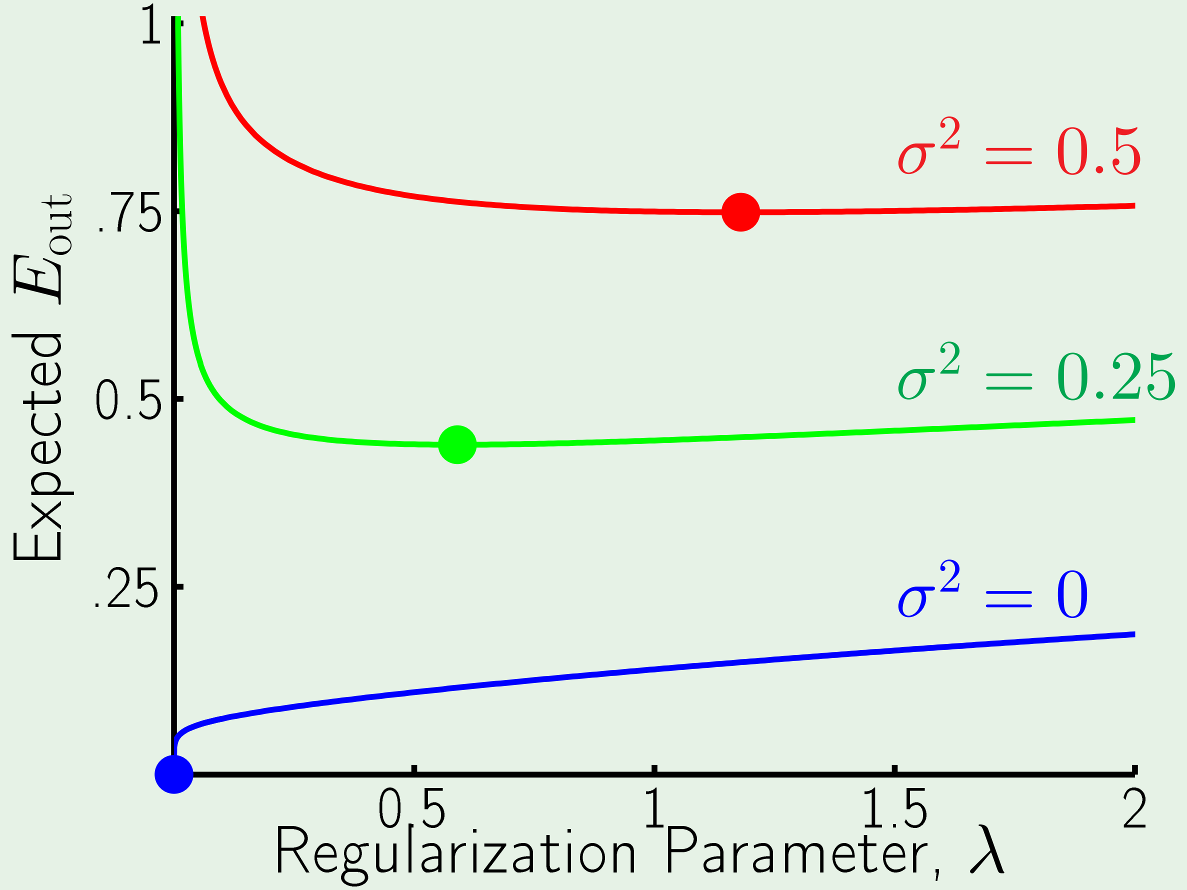
Early stopping as a regularizer

Regularization through the optimizer!

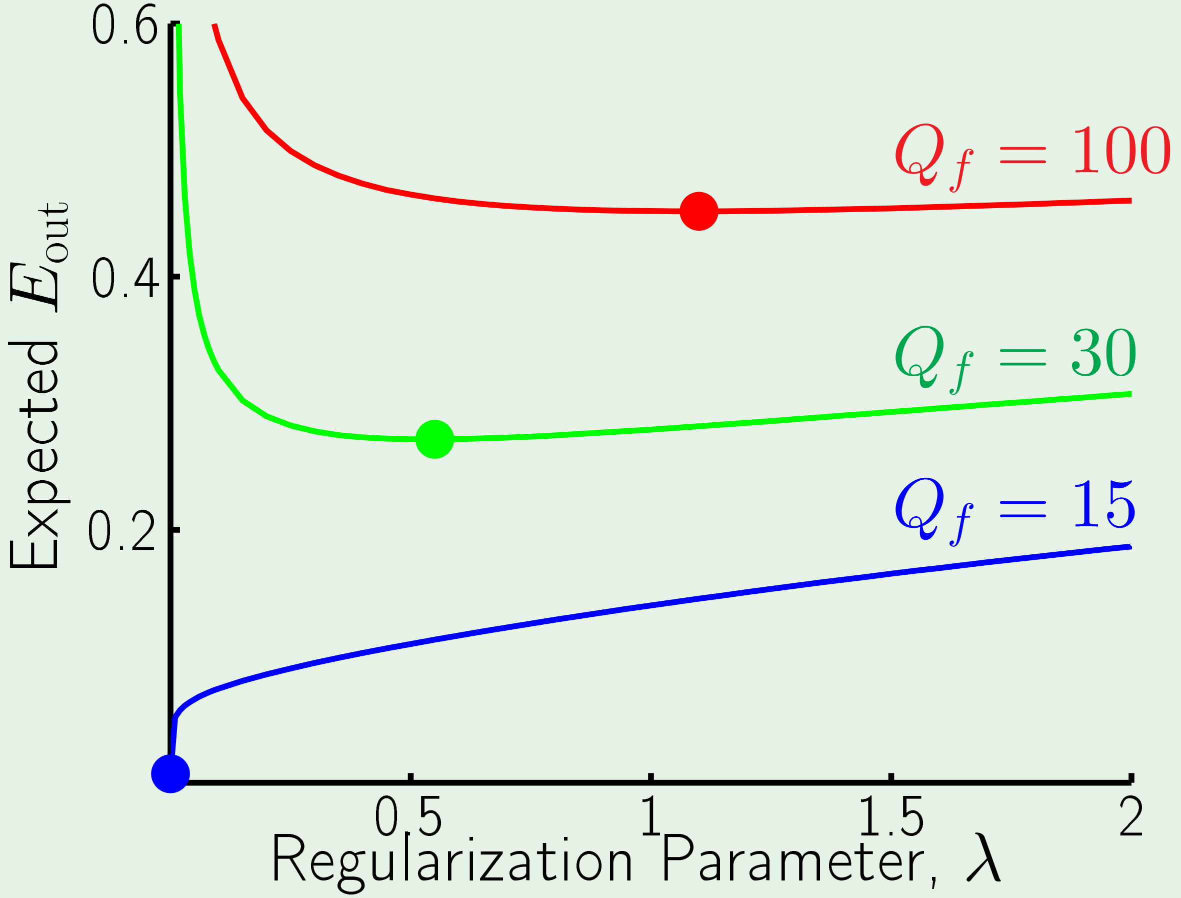
When to stop? **validation**



The optimal λ



Stochastic noise



Deterministic noise