Machine Learning from Data

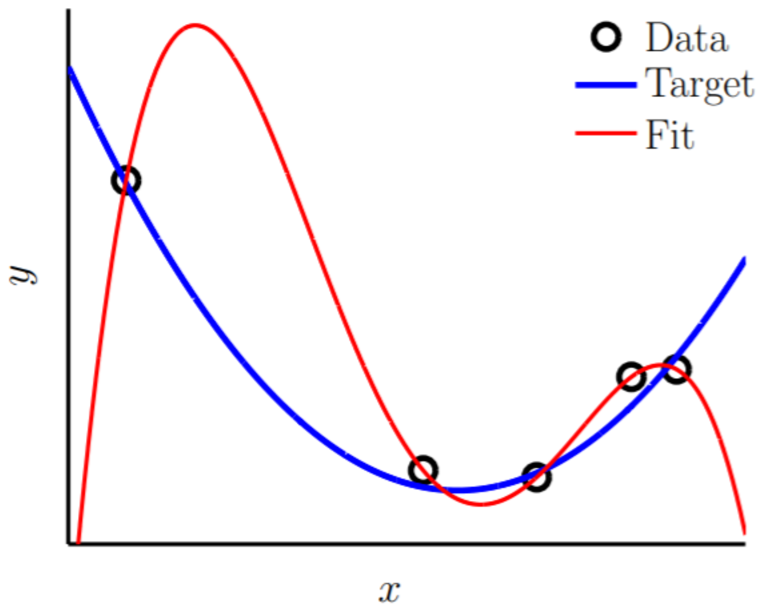
Lecture 12: Spring 2021

Today’s Lecture

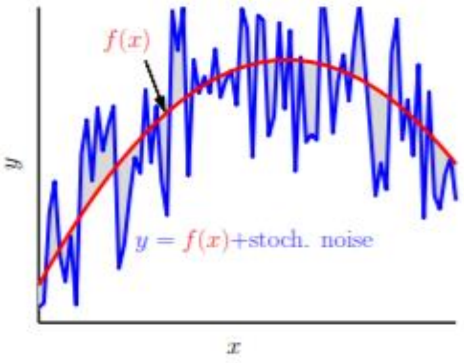
- Regularization
- Constraining the Model
- Augmented Error

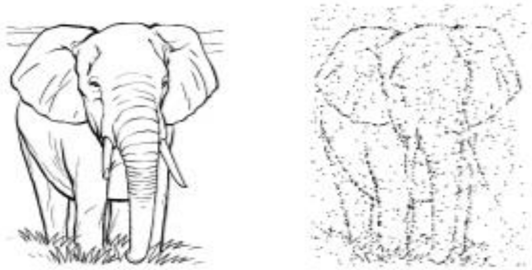
Overfit (Recap)

Fitting the data more than is warranted

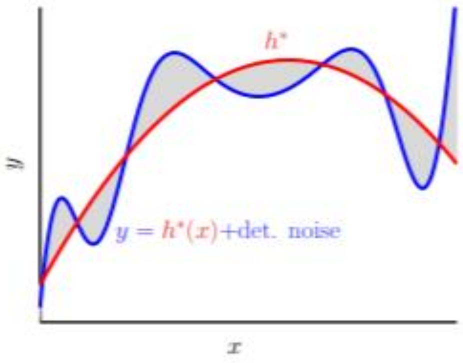


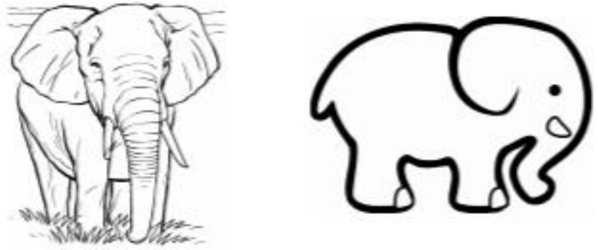
Stochastic Noise





Deterministic Noise





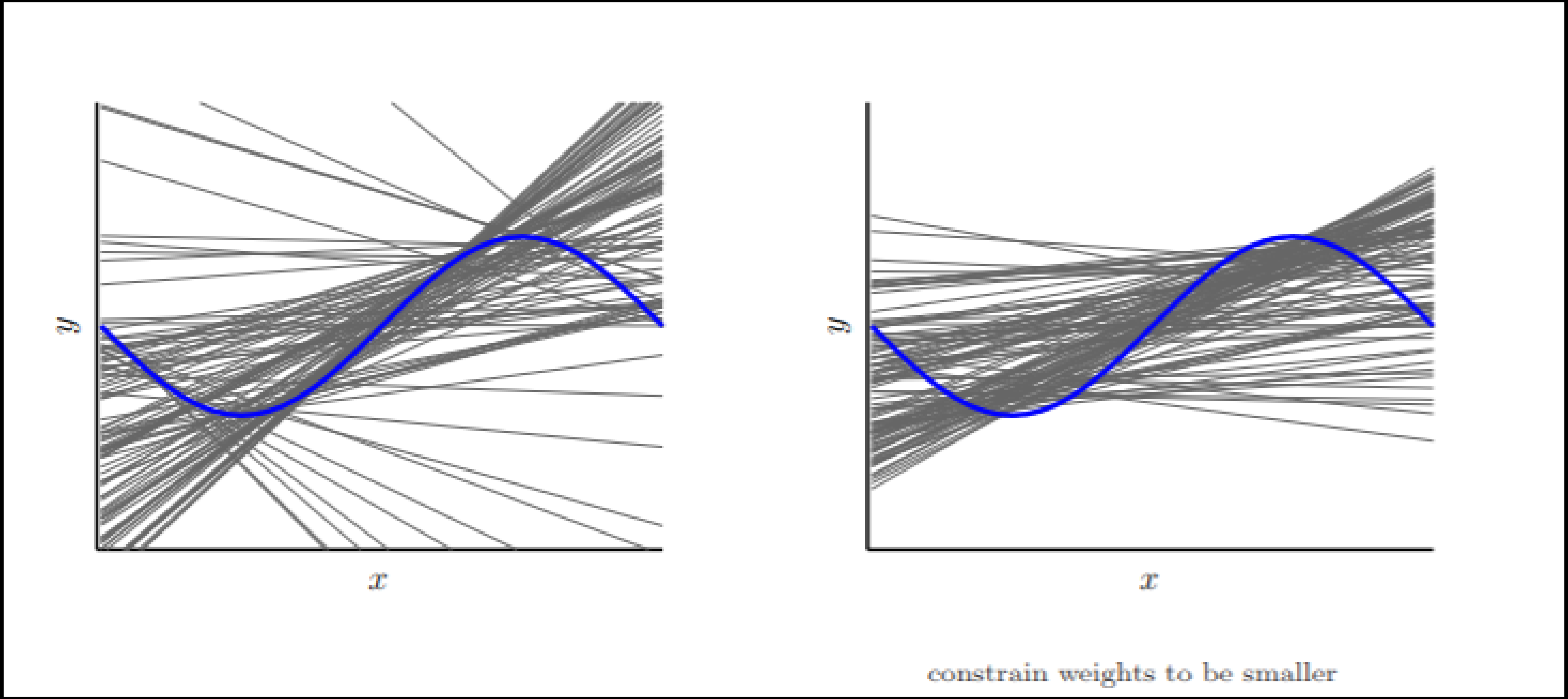
Stochastic and Deterministic Noise Hurt Learning

Human: Good at extracting the simple pattern, ignoring the noise and complications.

Computer: Pays equal attention to all pixels. Needs help simplifying → (features, regularization).

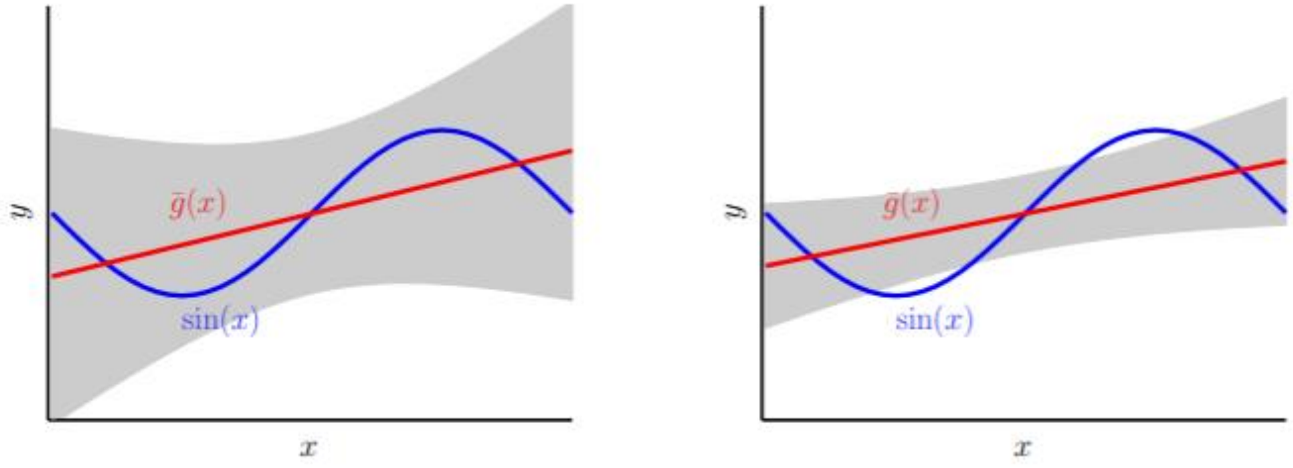
What is Regularization?

- A cure for our tendency to fit noise, hence improve out-of-sample Error.
- It works by constraining the model so that we cannot fit noise.
- Side effects: If we cannot fit noise maybe we cannot fit the actual signal (f)



Constraining the Model

Bias Variance



no regularization

bias $= 0.21$
var $= 1.69$

regularization

bias $= 0.23$ ← side effect
var $= 0.33$ ← treatment

(Constant model had bias=0.5 and var=0.25.)

Mathematics of Regularization

•

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

Standard Polynomial

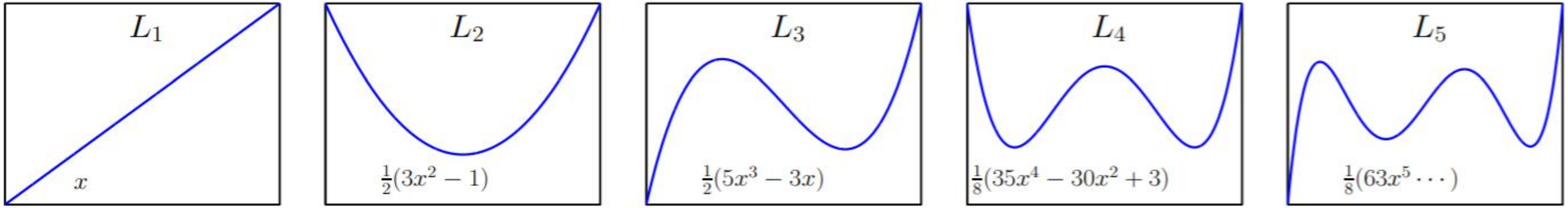
$$\mathbf{z} = \begin{bmatrix} 1 \\ x \\ x^2 \\ \vdots \\ x^{\mathrm{Q}} \end{bmatrix} \qquad \begin{aligned} h(x) &= \mathbf{w}^{\mathrm{T}}\mathbf{z}(x) \\ &= w_0 + w_1 x + \cdots + w_{\mathrm{Q}} x^{\mathrm{Q}} \end{aligned}$$

Legendre Polynomial

we're using linear regression

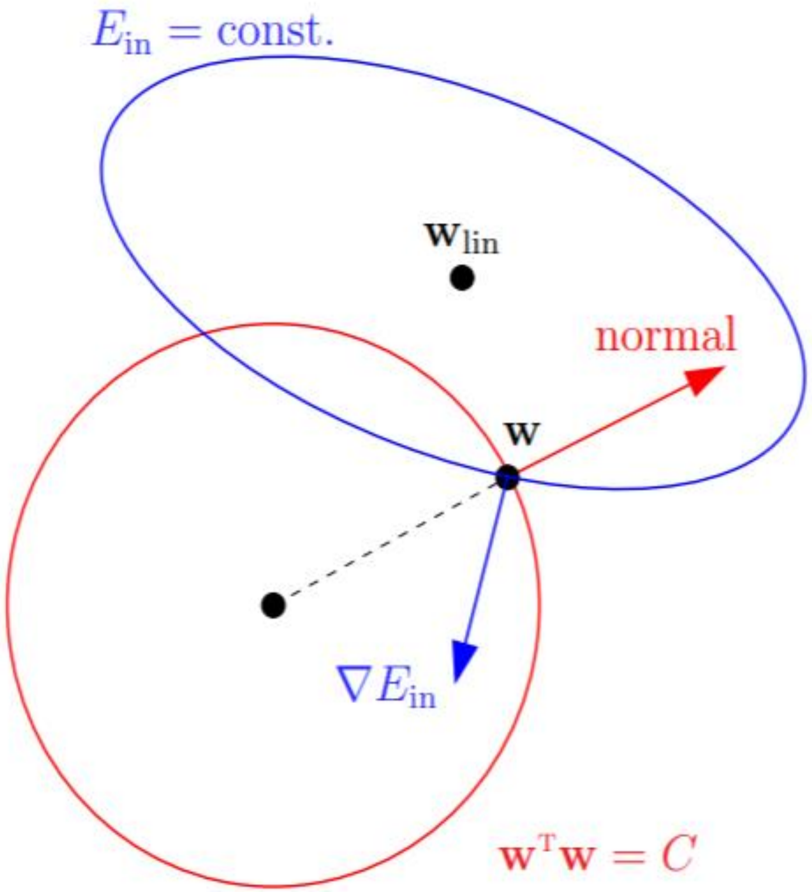
$$\mathbf{z} = \begin{bmatrix} 1 \\ L_1(x) \\ L_2(x) \\ \vdots \\ L_{\mathrm{Q}}(x) \end{bmatrix} \qquad \begin{aligned} h(x) &= \mathbf{w}^{\mathrm{T}}\mathbf{z}(x) \\ &= w_0 + w_1 L_1(x) + \cdots + w_{\mathrm{Q}} L_{\mathrm{Q}}(x) \end{aligned}$$

allows us to treat the weights 'independently'



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

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



Observations:

1. Optimal $\mathbf{w}$ tries to get as 'close' to $\mathbf{w}_{\text{lin}}$ as possible.

   Optimal $\mathbf{w}$ will use full budget and be on the surface $\mathbf{w}^{\mathrm{T}}\mathbf{w} = C$.

2. Surface $\mathbf{w}^{\mathrm{T}}\mathbf{w} = C$, at optimal $\mathbf{w}$, should be perpindicular to ∇E_{in}.

   Otherwise can move along the surface and decrease E_{in}.

3. Normal to surface $\mathbf{w}^{\mathrm{T}}\mathbf{w} = C$ is the vector $\mathbf{w}$.

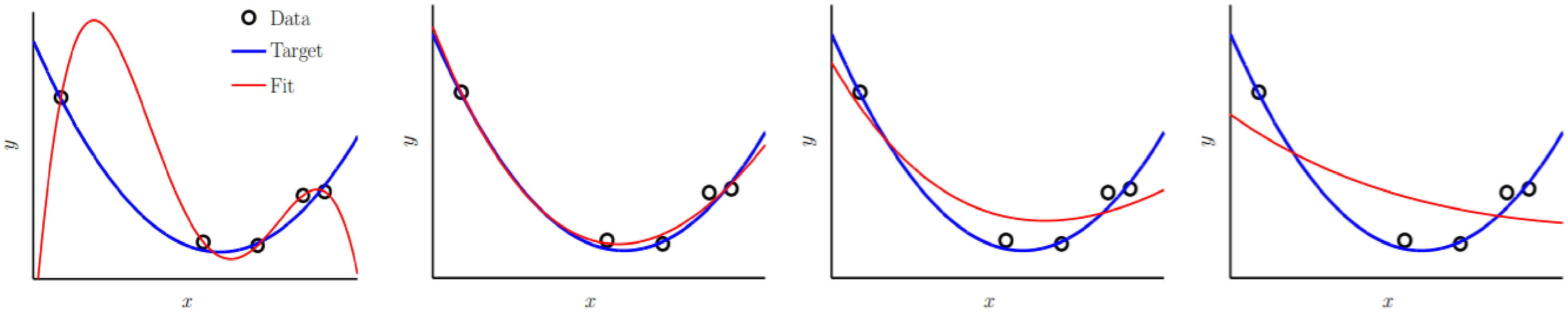
4. Surface is $\perp$ ∇E_{in}; surface is $\perp$ normal.

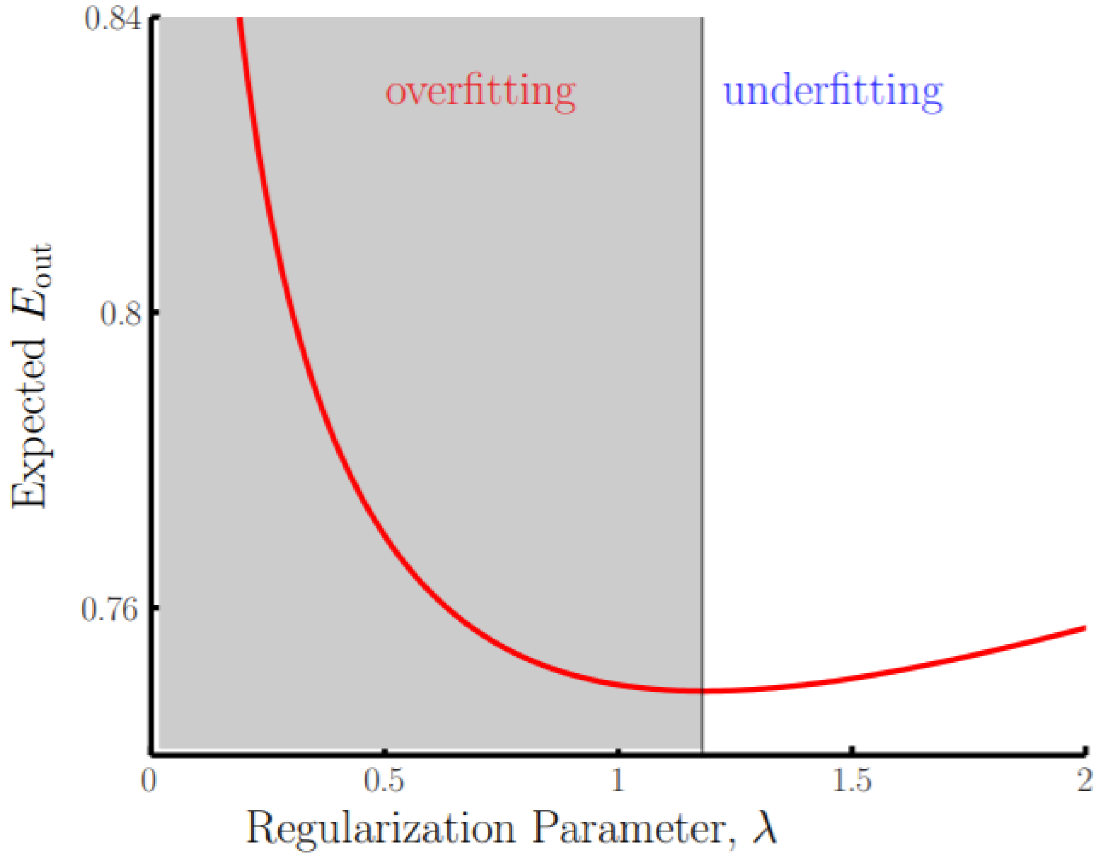
   ∇E_{in} is parallel to normal (but in opposite direction).

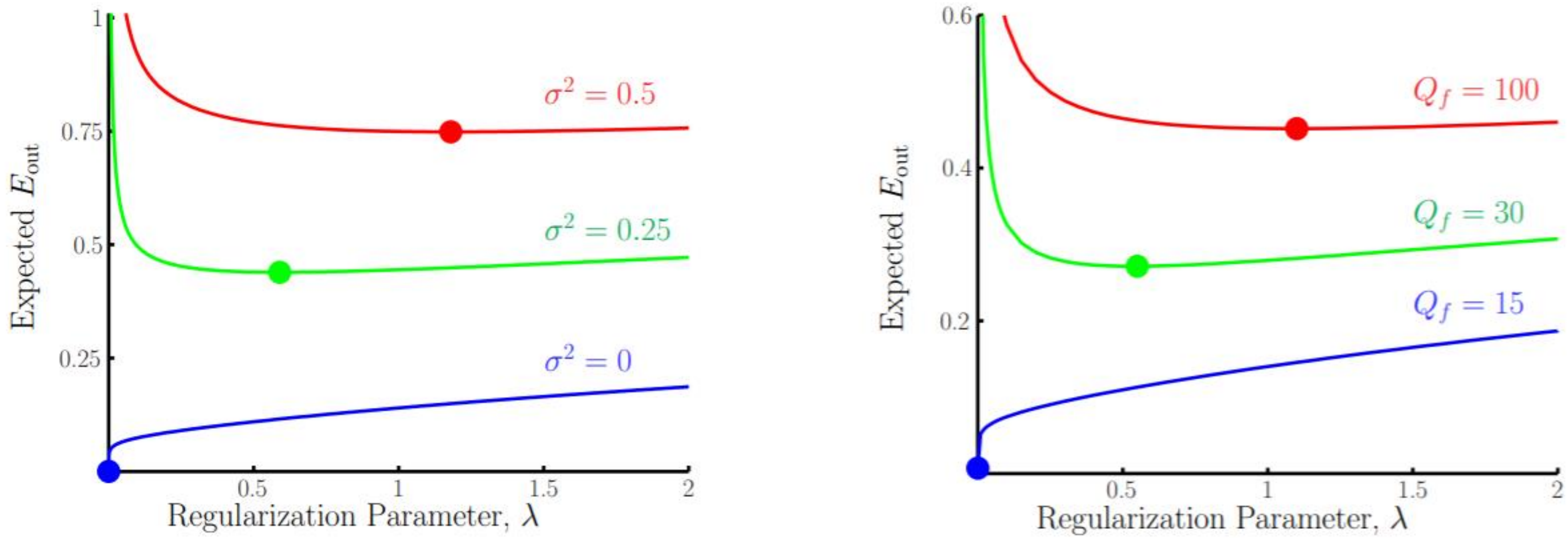
Regularization In Action

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

$\underline{\lambda = 0}$ $\quad$ $\underline{\lambda = 0.0001}$ $\quad$ $\underline{\lambda = 0.01}$ $\quad$ $\underline{\lambda = 1}$

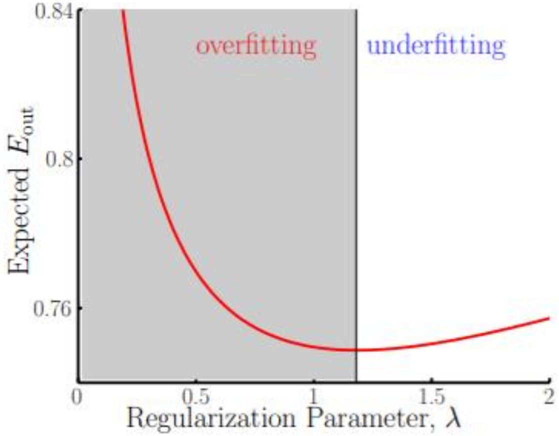



0.84
0.8
0.76
Expected E_{out}
overfitting
underfitting
0
0.5
1
1.5
2
Regularization Parameter, λ



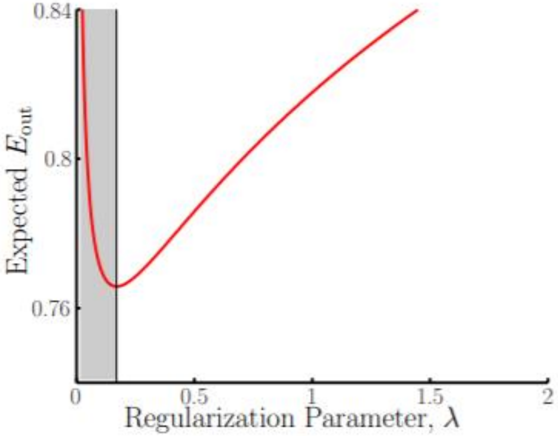
Expected E_{out}
Regularization Parameter, λ
$\sigma^2 = 0.5$
$\sigma^2 = 0.25$
$\sigma^2 = 0$
Expected E_{out}
Regularization Parameter, λ
$Q_f = 100$
$Q_f = 30$
$Q_f = 15$

Uniform Weight Decay



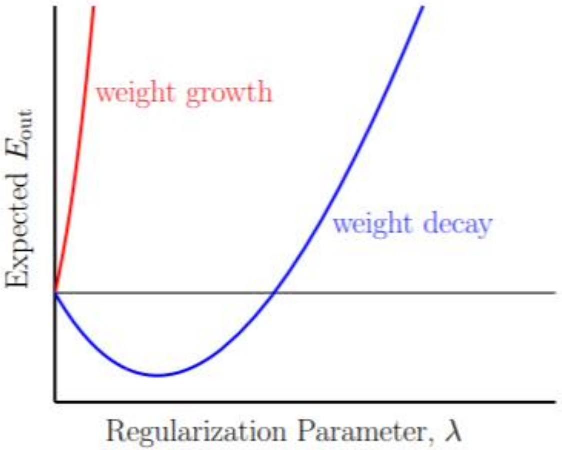
$$\sum_{q=0}^{Q} w_q^2$$

Low Order Fit



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

Weight Growth!



$$\sum_{q=0}^{Q} \frac{1}{w_q^2}$$

Choosing a Regularizer – A Practitioner's Guide

The perfect regularizer:

constrain in the 'direction' of the target function.

target function is <u>unknown</u> (going around in circles ☺).

The guiding principle:

constrain in the 'direction' of **smoother** (usually simpler) hypotheses

hurts your ability to fit the 'high frequency' noise

smoother and simpler $\xrightarrow{\text{usually means}}$ weight decay not weight growth.

Stochastic noise $\longrightarrow$ nothing you can do about that.

Good features $\longrightarrow$ helps to reduce deterministic noise.

Regularization:

Helps to combat what noise remains, especially when N is small.

Typical modus operandi: sacrifice a little bias for a huge improvement in var.

VC angle: you are using a smaller $\mathcal{H}$ without sacrificing too much E_{in}

this was $\mathbf{w}^{\mathrm{T}}\mathbf{w}$

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

$$\updownarrow$$

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

this was $O\left(\sqrt{\frac{d_{\text{vc}}}{N}\ln N}\right)$

E_{aug} can beat E_{in} as a proxy for E_{out}.

depends on choice of λ

Thanks!