Machine Learning from Data

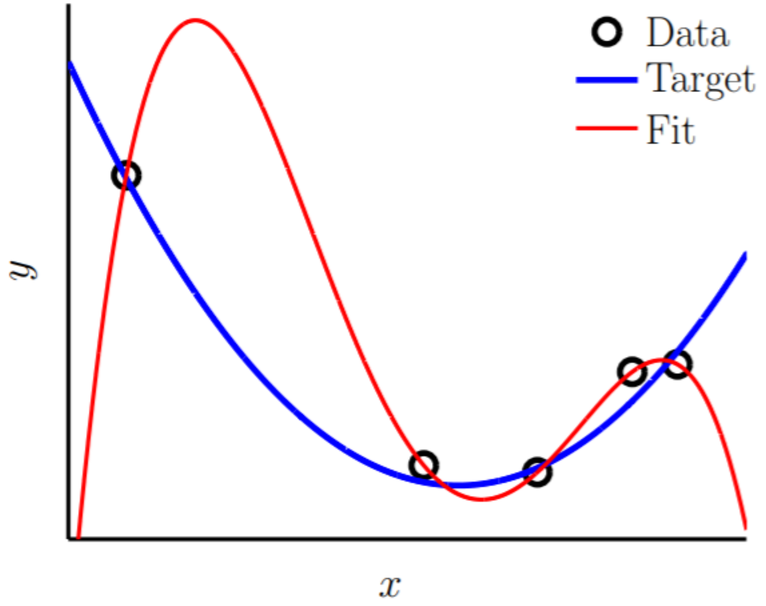
Lecture 12: Spring 2021

Today's Lecture

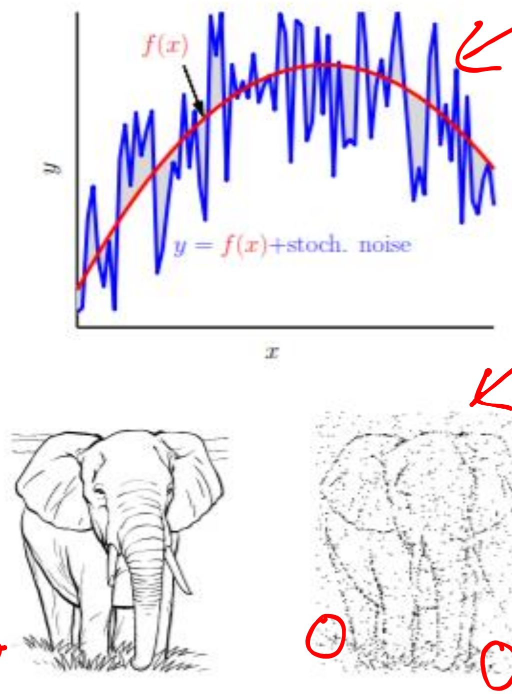
- Regularization
- Constraining the Model
- Augmented Error

Overfit (Recap)

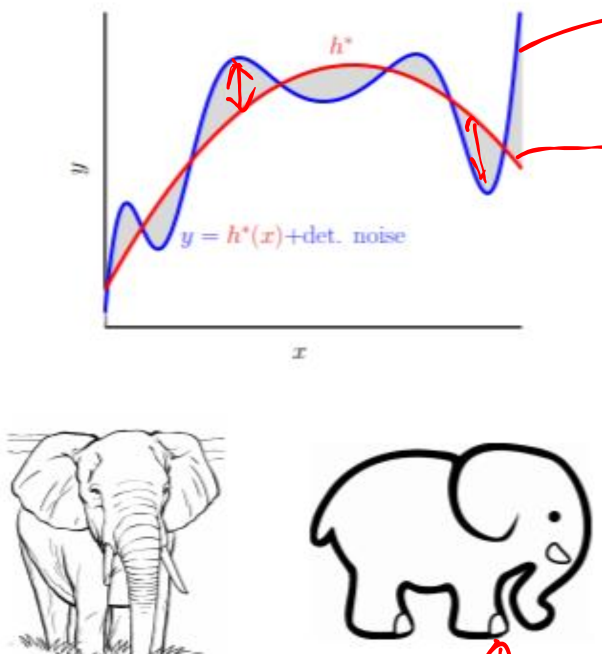
Fitting the data more than is warranted



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).

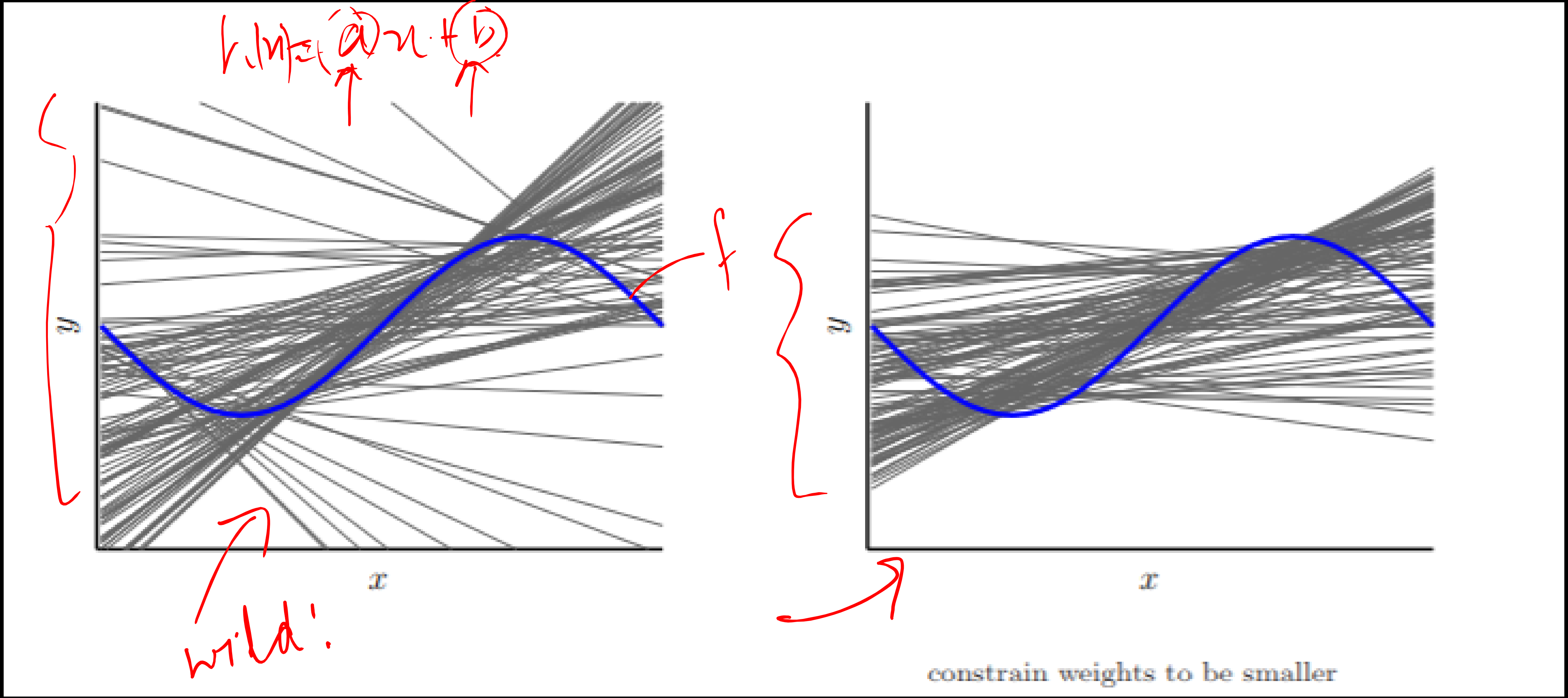
What is Regularization?

Overfitting → disease

- A cure for our tendency to fit noise, hence improve out-of-sample Error. E_{out}

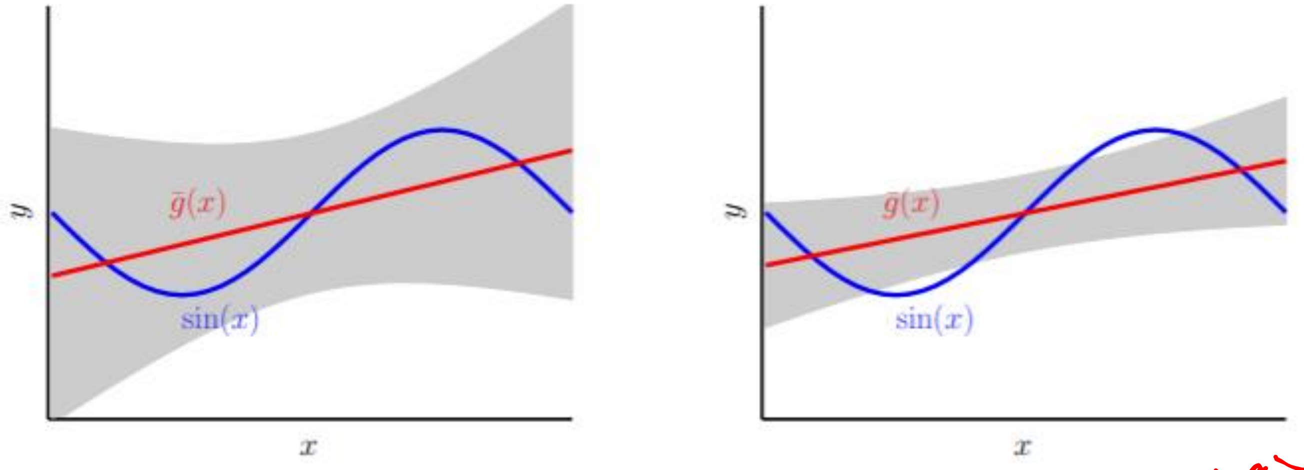
Medicine

- 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	regularization	
bias = 0.21	bias = 0.23	← side effect
var = 1.69	var = 0.33	← treatment

paying the price

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

$\boxed{h_0(x) = b}$

$h(x) = ax + b$

Mathematics of Regularization

- Constraint the model

Linear Models → Constraints → Hard, Soft

VC Inequality :

$$E_{out}(g) \leq E_{in}(g) + \underline{\underline{\Omega(H)}}$$

$\Omega(H)$ → fitting with a simple H

$\Omega(H)$ → $\Omega(g)$

tighter

$(x_1, y_1)\ (x_2\, y_2) \ldots\ (x_N\, y_N) \xrightarrow{\Phi} (z_1, y_1)\ (z_2\, y_2) \ldots (z_N\, y_N)$

Data matrix $= X$

Target vector $= Y$

Transformed data matrix $= Z$

Find $\tilde{w} \rightarrow Z$ space

Min. $(Z\tilde{w} - y)^T (Z\tilde{w} - y)$

$Z\tilde{w} = y$

If $Z \rightarrow$ was invertible

$\tilde{w} = Z^{+} y$

$(Z^{+} = (Z^T Z)^{-1} Z^T)$

Legendre's Polynomials.

$$\begin{bmatrix} 1 \\ x \end{bmatrix} \longrightarrow \begin{bmatrix} 1 \\ L_1(x) \\ L_2(x) \\ \vdots \\ L_Q(x) \end{bmatrix}$$

$L_1(x)$ → 1st order

$L_Q(x)$ → Qth order ~~poly~~

x^3, x^5

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

Standard Polynomial

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

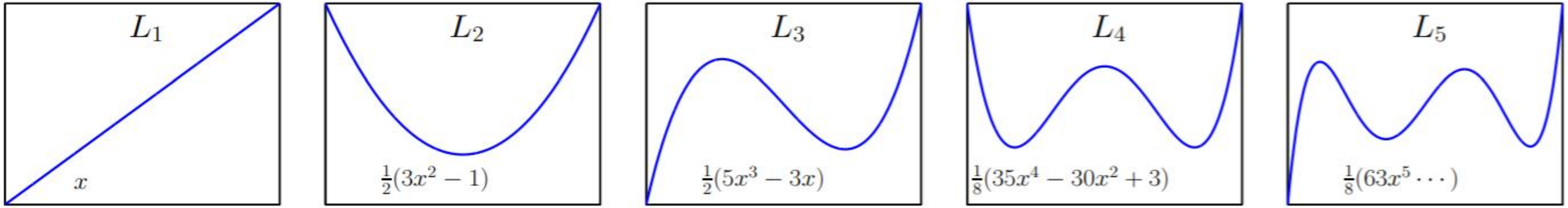
Legendre Polynomial

we're using linear regression

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

allows us to treat the weights 'independently'

orthogonal.



$$\mathcal{H}_2 = \{ h(x) \mid h(x) = w_0 + w_1 l_1(x) + w_2 l_2(x), \ w \in \mathbb{R}^3 \}$$

$$\mathcal{H}_{10} = \{ h(x) \mid h(x) = w_0 + w_1 l_1(x) + w_2 l_2(x) \dots w_{10} l_{10}(x), \ w \in \mathbb{R}^{11} \}$$

$$\mathcal{H}_Q = \sum_{q=0}^{Q} w_q l_q(x)$$

$$\underline{\underline{\mathcal{H}_2}} \subset \mathcal{H}_{10}$$

if $E_{in}(\mathcal{H}_2) \simeq E_{in}(\mathcal{H}_{10})$ ✓

VC Theory.

constrain

$$\mathcal{H}_2 = \{ h(x) \mid h(x) = w_0 + w_1 l_1(x) + w_2 l_2(x) + \dots \dots w_{10} l_{10}(x),\ w \in \mathbb{R}^{11} \}$$

s.t. $\to w_3 = w_4 \dots = w_{10} = 0$

Hard-order constraint.

Budget $= C$

Soft order constraint.

$$\mathcal{H}_C = \{ h(x) \mid h(x) = w_0 + w_1 l_1(x) + \dots \dots w_{10} l_{10}(x),\ w \in \mathbb{R}^{11} \}$$

$$\sum_{q=0}^{10} w_q^2 \le C$$

$\mathcal{H}_C \subset \mathcal{H}_{10}$

$C \to \infty$

$\to \mathcal{H}_{10}$

$\to \mathcal{H}_{10}$

C explore

$\to \mathcal{H}_2$

$\tilde{w}$ $\quad$ x $\quad$ $w^T z$ $\quad$ $w^T w \le C$ ✓

$$E_{in}(w) = \frac{1}{N}(Zw - y)^T (Zw - y)$$

$$w_{lin} = Z^{\dagger} y = (Z^T Z)^{-1} Z^T y$$

$$\min_{w} \underline{E_{in}(w)} \quad \text{subject to } w^T w \le C$$

Case 1: $w_{lin}^T w_{lin} \le C$

$w_{reg} = w_{lin}$

Case 2: $w_{lin}^T w_{lin} > C$

E_{in} is convex

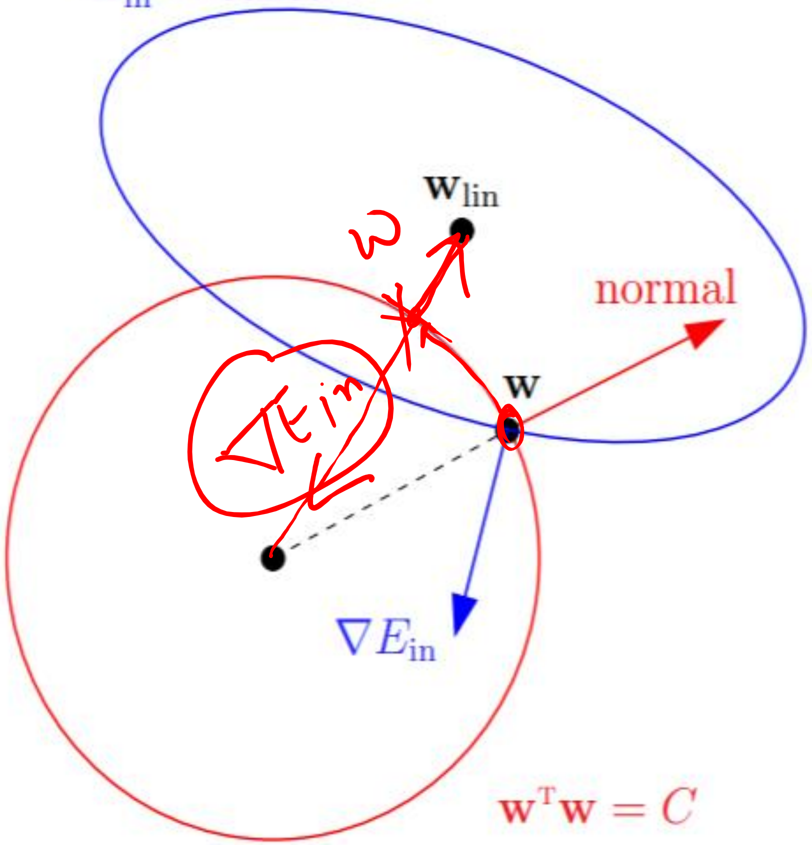
quadratic

feasible set

$$\nabla E_{in} \propto -w$$

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

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



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 $\boxed{\mathbf{w}^{\mathsf{T}}\mathbf{w} = C}$.

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

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

3. Normal to surface $\mathbf{w}^{\mathsf{T}}\mathbf{w} = C$ 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).

Necessary condition for optimality:

$\nabla E_{in} \propto -w$ lagrange's multiplier >0

or $\nabla E_{in} = -2\lambda_C w$ mathematical convenience

$\nabla E_{in} + \underbrace{2\lambda_C w}_{\text{gradient of } \lambda_C w^T w} = 0$

$\nabla (E_{in} + \lambda_C w^T w) = 0$ — condition.

$$\boxed{\min \; E_{in}(w) + \lambda_C w^T w}$$

Constrained optimization

$\downarrow$

Unconstrained optimization

$\lambda_C \uparrow \longrightarrow C$ must go down

Fitting under a constraint (Budget)

$\Downarrow$

Unconstrained penalized fitting (penalty parameter λ)

$$\underline{\min}\ E_{aug}(w) = E_{in}(w) + \underbrace{\frac{\lambda}{N} \underline{\underline{w^T w}}}$$

penalty term

$\downarrow$

complexity of hypothesis.

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

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

$w^T w$ — Regularizer (we pick) }

$\lambda \longrightarrow$ Regularization parameter } knob control

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

Weight Decay $\qquad E_{aug}. \qquad w_{reg}.$

$\lambda (w^T w) \longrightarrow$ weight decay.

$$\min \; E_{in}(w) + \lambda/_N \, w^T w$$

$$\min \; \frac{1}{N}(Zw-y)^T(Zw-y) + \frac{\lambda}{N} w^T w$$

$$\min \; w^T Z^T Z w - 2 w^T Z^T y + y^T y + \lambda \underline{w^T w}$$

$$\min \; w^T (Z^T Z + \lambda I) w - 2 w^T Z^T y + y^T y$$

gradient $\longrightarrow 0$

$$2\underbrace{(\underbrace{Z^TZ}_{\text{+ve semi definite}} + \lambda I)}_{\text{Invertible}}w - 2Z^Ty = 0$$

+ve definite

Solve for w

$$w_{reg} = \underbrace{(Z^TZ + \lambda I)^{-1}Z^Ty}_{\text{Regularized linear regression feature transform.}}$$

$\lambda \rightarrow$ penalizing large weights.

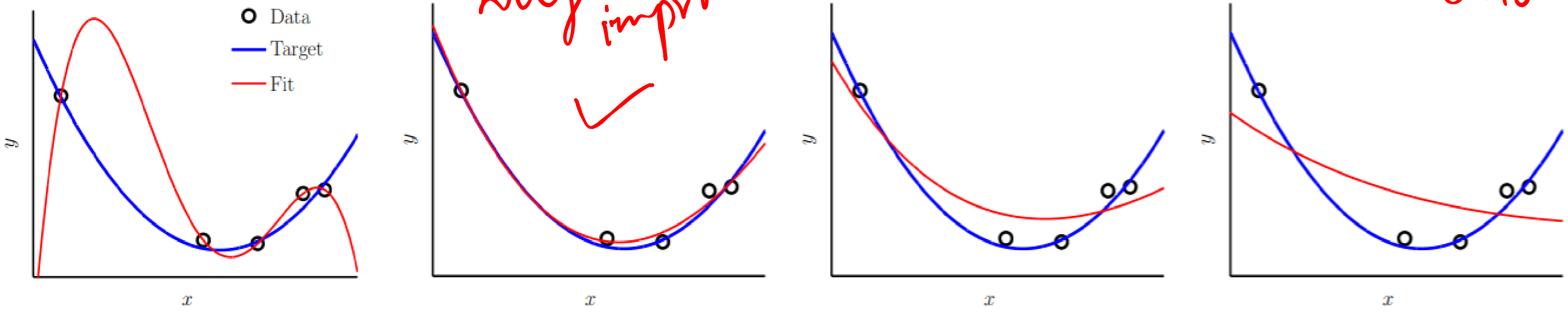
Regularization In Action

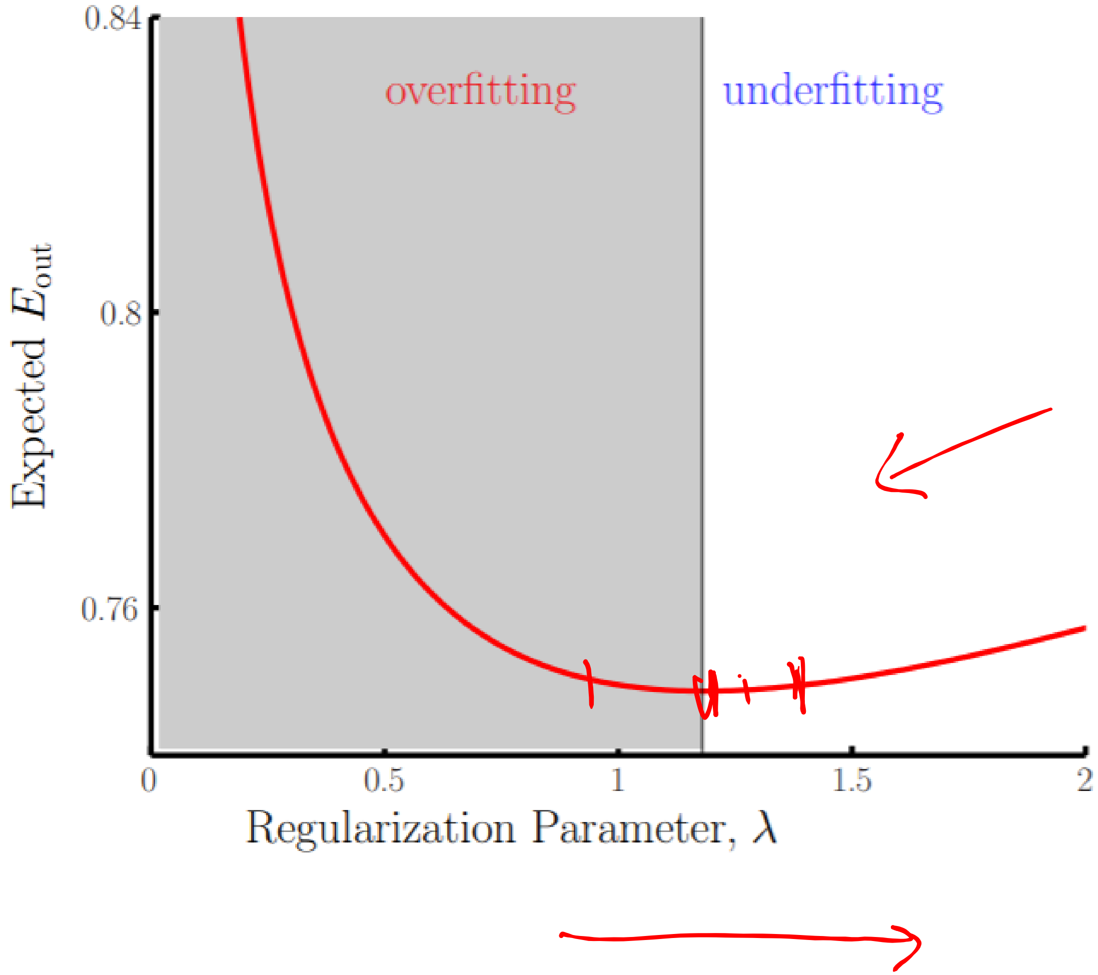
Minimizing $E_{\text{in}}(\mathbf{w}) + \frac{\lambda}{N}\mathbf{w}^{\mathsf{T}}\mathbf{w}$ with different λ's

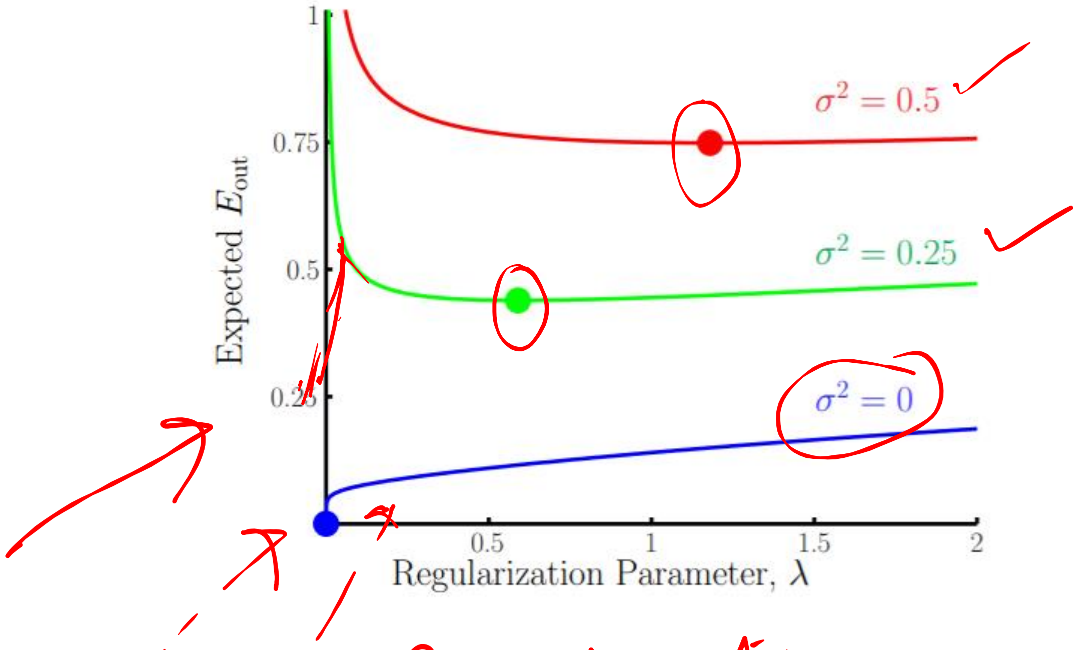
$\lambda = 0$ $\lambda = 0.0001$ $\lambda = 0.01$ $\lambda = 1$

Huge improvement!

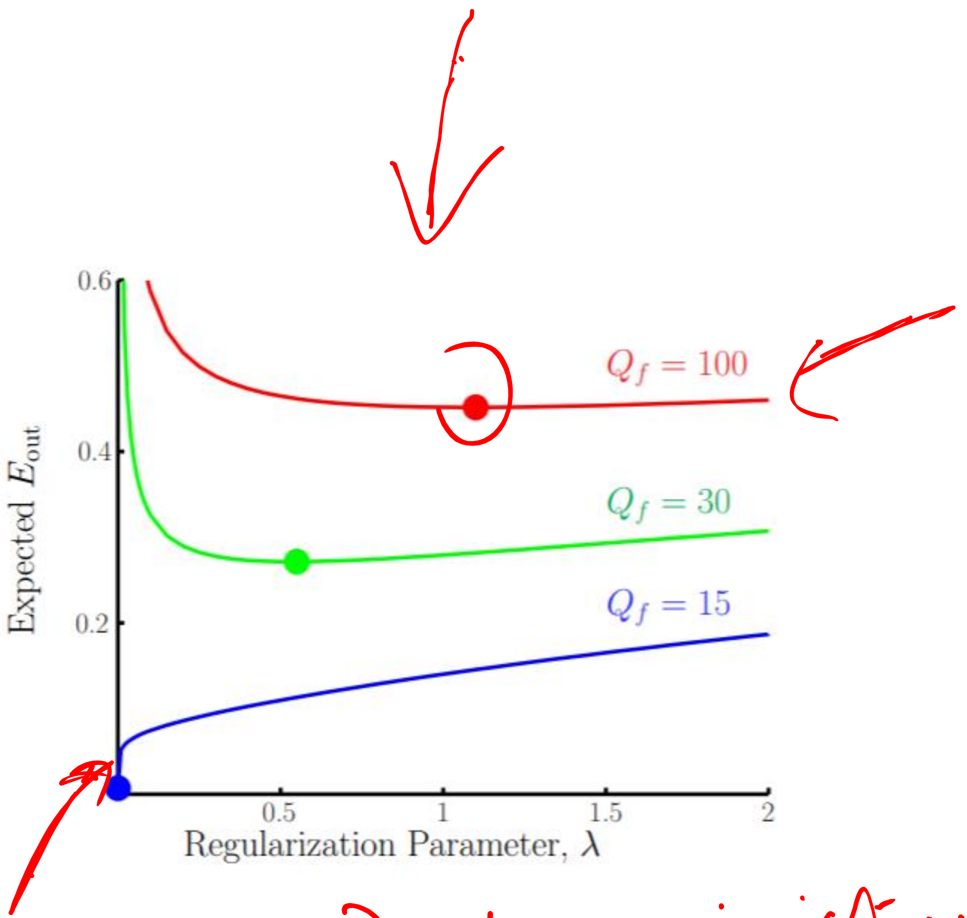
underfitting.




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

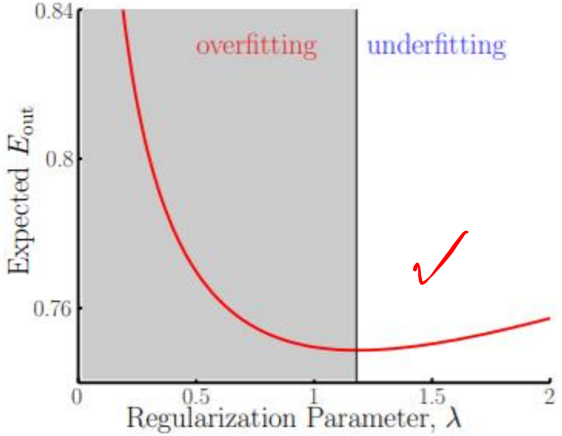


Stochastic



Deterministic

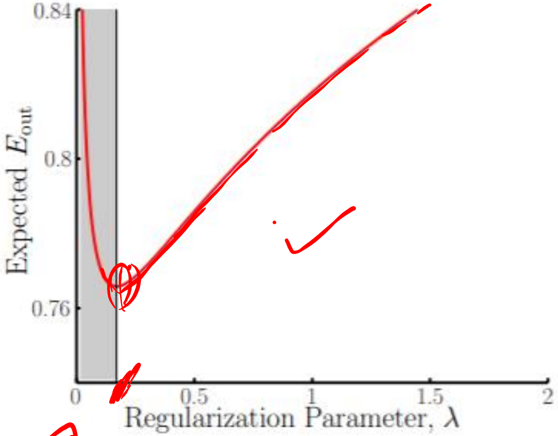
Uniform Weight Decay



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

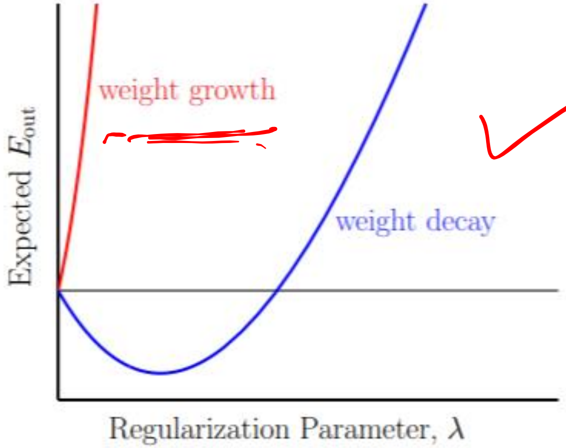
Ω_{unit}

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 unknown (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 $\mathbf{w}^{\mathsf{T}}\mathbf{w}$

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

$$\updownarrow$$

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

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!