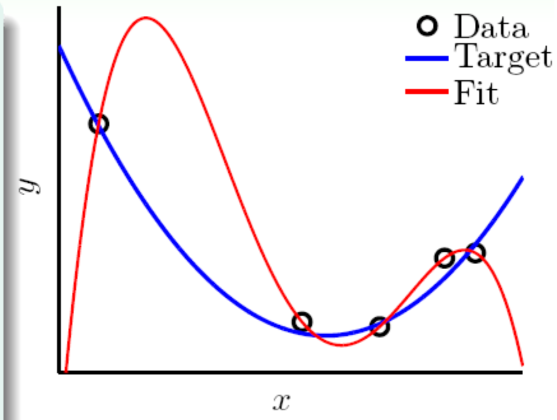
Machine Learning

Lecture 12
Regularization

Chen-Kuo Chiang (江 振 國)
ckchiang@cs.ccu.edu.tw
中正大學 資訊工程學系

Bad Generalization

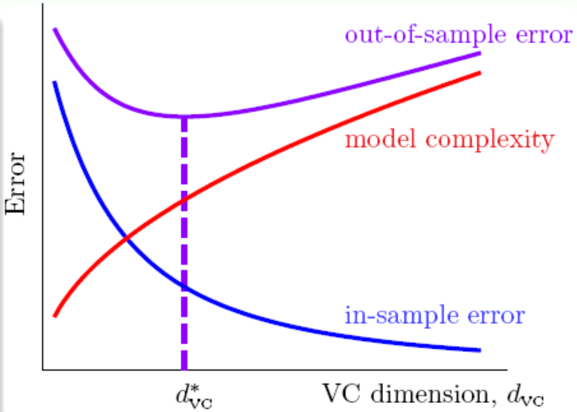
- regression for $x \in \mathbb{R}$ with $N = 5$ examples
- target $f(x)$ = 2nd order polynomial
- label $y_n = f(x_n) +$ very small noise
- linear regression in $\mathcal{Z}$-space + $\mathbf{\Phi}$ = 4th order polynomial
- unique solution passing all examples $\Longrightarrow E_{\text{in}}(g) = 0$
- $E_{\text{out}}(g)$ **huge**



bad generalization: low E_{in}, high E_{out}

Bad Generalization and Overfitting

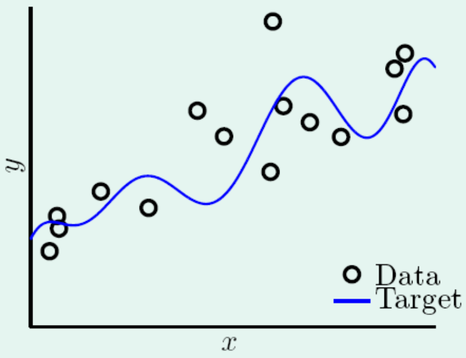
- take $d_{VC} = 1126$ for learning: bad generalization —(E_{out} - E_{in}) large
- switch from $d_{VC} = d^*_{VC}$ to $d_{VC} = 1126$: **overfitting** —$E_{in} \downarrow$, $E_{out} \uparrow$
- switch from $d_{VC} = d^*_{VC}$ to $d_{VC} = 1$: **underfitting** —$E_{in} \uparrow$, $E_{out} \uparrow$



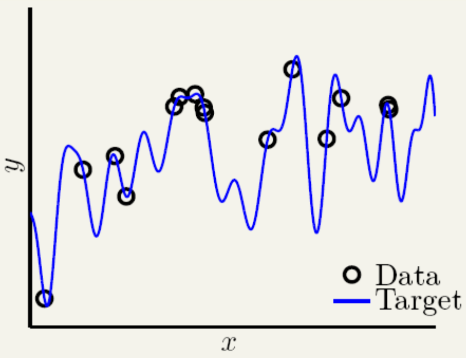
bad generalization: low E_{in}, high E_{out};
overfitting: low**er** E_{in}, high**er** E_{out}

Case Study (1/2)

10-th order target function + noise



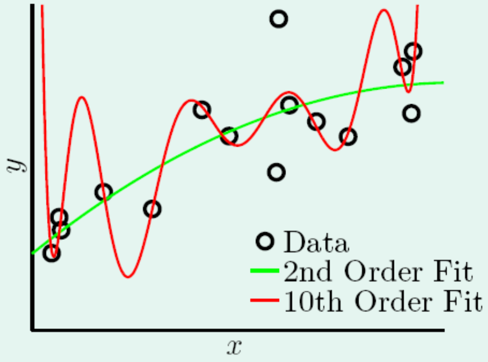
50-th order target function noiselessly



overfitting from best $g_2 \in \mathcal{H}_2$ to best $g_{10} \in \mathcal{H}_{10}$?

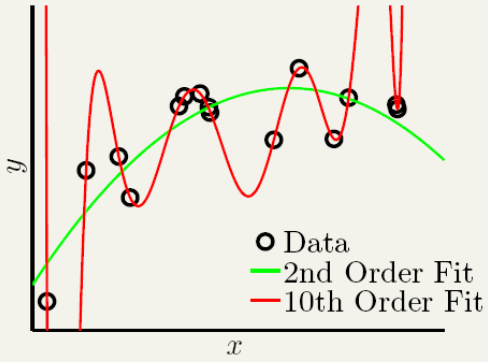
Case Study (2/2)

10-th order target function + noise



	$g_2 \in \mathcal{H}_2$	$g_{10} \in \mathcal{H}_{10}$
E_{in}	0.050	0.034
E_{out}	0.127	**9.00**

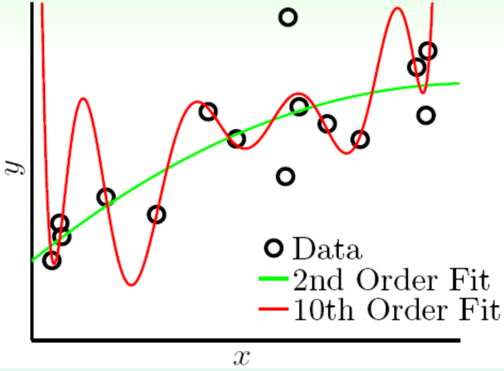
50-th order target function noiselessly

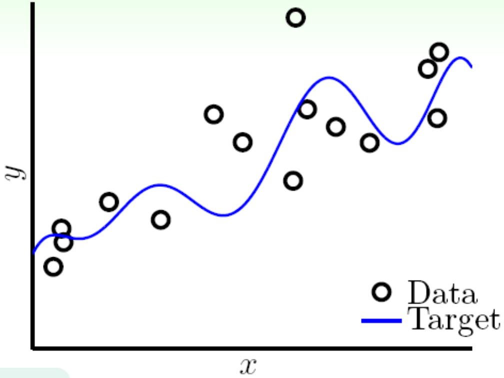


	$g_2 \in \mathcal{H}_2$	$g_{10} \in \mathcal{H}_{10}$
E_{in}	0.029	0.00001
E_{out}	0.120	**7680**

overfitting from g_2 to g_{10}? **both yes!**

Irony of Two Learners

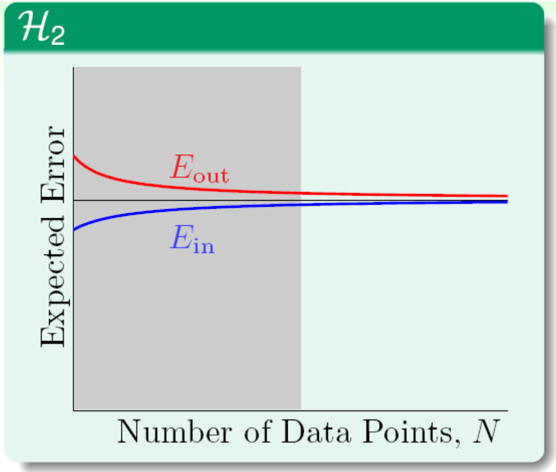


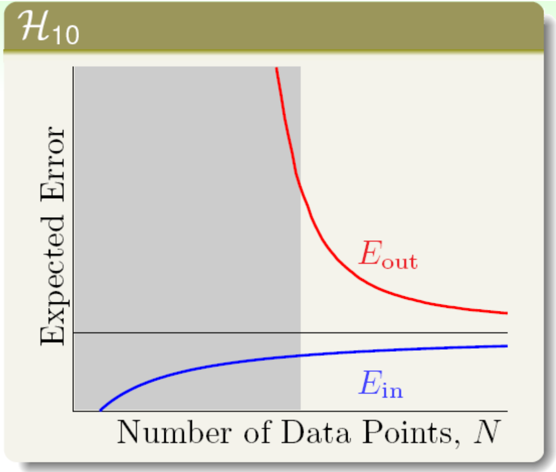


- learner *Overfit*: pick $g_{10} \in \mathcal{H}_{10}$
- learner *Restrict*: pick $g_2 \in \mathcal{H}_2$
- when both **know that target = 10th**
  —*R* 'gives up' ability to fit

but *R* **wins in** E_{out} a lot!
philosophy: concession for **advantage**? :-)

Learning Curves Revisited

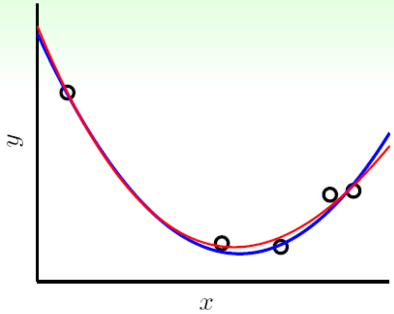




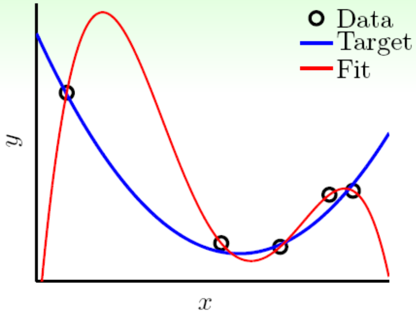
- $\mathcal{H}_{10}$: lower $\overline{E_{\text{out}}}$ when $N \to \infty$,
  but much larger generalization error for small N
- gray area: O overfits! ($\overline{E_{\text{in}}} \downarrow$, $\overline{E_{\text{out}}} \uparrow$)

R always **wins in** $\overline{E_{\text{out}}}$ if N small!

Regularization: The Magic

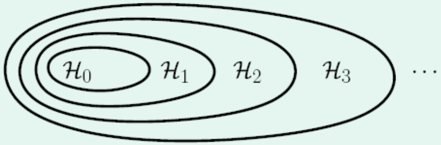


'regularized fit' $\Longleftarrow$



overfit

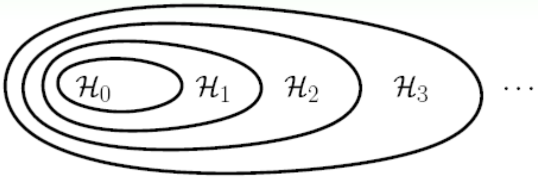
- idea: 'step back' from $\mathcal{H}_{10}$ to $\mathcal{H}_2$



- name history: function approximation for **ill-posed problems**

how to step back?

Stepping Back as Constraint



Q-th order polynomial transform for $x \in \mathbb{R}$:

$$\mathbf{\Phi}_Q(x) = (1, x, x^2, \ldots, x^Q)$$

+ linear regression, denote $\tilde{\mathbf{w}}$ by $\mathbf{w}$

hypothesis $\mathbf{w}$ in $\mathcal{H}_{10}$: $\quad w_0 + w_1 x + w_2 x^2 + w_3 x^3 + \ldots + w_{10} x^{10}$

hypothesis $\mathbf{w}$ in $\mathcal{H}_2$: $\quad w_0 + w_1 x + w_2 x^2$

that is, $\mathcal{H}_2 = \mathcal{H}_{10}$ AND 'constraint that $w_3 = w_4 = \ldots = w_{10} = 0$'

step back = **constraint**

Regression with Constraint

$\mathcal{H}_{10} \equiv \left\{ \mathbf{w} \in \mathbb{R}^{10+1} \right\}$

regression with $\mathcal{H}_{10}$:

$$\min_{\mathbf{w} \in \mathbb{R}^{10+1}} E_{\text{in}}(\mathbf{w})$$

$\mathcal{H}_2 \equiv \left\{ \mathbf{w} \in \mathbb{R}^{10+1} \right.$ while $\left. w_3 = w_4 = \ldots = w_{10} = 0 \right\}$

regression with $\mathcal{H}_2$:

$$\min_{\mathbf{w} \in \mathbb{R}^{10+1}} E_{\text{in}}(\mathbf{w})$$
$$\text{s.t.} \quad w_3 = w_4 = \ldots = w_{10} = 0$$

step back = constrained optimization of E_{in}

why don't you just use $\mathbf{w} \in \mathbb{R}^{2+1}$? :-)

Regression with Looser Constraint

$$\mathcal{H}_2 \equiv \left\{ \mathbf{w} \in \mathbb{R}^{10+1} \right.$$
$$\left. \text{while } w_3 = \ldots = w_{10} = 0 \right\}$$

regression with $\mathcal{H}_2$:

$$\min_{\mathbf{w} \in \mathbb{R}^{10+1}} \quad E_{\text{in}}(\mathbf{w})$$
$$\text{s.t.} \quad w_3 = \ldots = w_{10} = 0$$

$$\mathcal{H}'_2 \equiv \left\{ \mathbf{w} \in \mathbb{R}^{10+1} \right.$$
$$\left. \text{while } \geq 8 \text{ of } w_q = 0 \right\}$$

regression with $\mathcal{H}'_2$:

$$\min_{\mathbf{w} \in \mathbb{R}^{10+1}} \quad E_{\text{in}}(\mathbf{w})$$
$$\text{s.t.} \quad \sum_{q=0}^{10} [\![w_q \neq 0]\!] \leq 3$$

- more flexible than $\mathcal{H}_2$: $\quad \mathcal{H}_2 \subset \mathcal{H}'_2$
- less risky than $\mathcal{H}_{10}$: $\quad \mathcal{H}'_2 \subset \mathcal{H}_{10}$

bad news for sparse hypothesis set $\mathcal{H}'_2$:
NP-hard to solve :-(

Regression with Softer Constraint

$$\mathcal{H}'_2 \equiv \left\{ \mathbf{w} \in \mathbb{R}^{10+1} \right.$$
$$\left. \text{while } \geq 8 \text{ of } w_q = 0 \right\}$$

regression with $\mathcal{H}'_2$:

$$\min_{\mathbf{w} \in \mathbb{R}^{10+1}} E_{\text{in}}(\mathbf{w}) \text{ s.t. } \sum_{q=0}^{10} [\![w_q \neq 0]\!] \leq 3$$

$$\mathcal{H}(C) \equiv \left\{ \mathbf{w} \in \mathbb{R}^{10+1} \right.$$
$$\left. \text{while } \|\mathbf{w}\|^2 \leq C \right\}$$

regression with $\mathcal{H}(C)$:

$$\min_{\mathbf{w} \in \mathbb{R}^{10+1}} E_{\text{in}}(\mathbf{w}) \text{ s.t. } \sum_{q=0}^{10} w_q^2 \leq C$$

- $\mathcal{H}(C)$: overlaps but not exactly the same as $\mathcal{H}'_2$
- soft and smooth structure over $C \geq 0$:
  $\mathcal{H}(0) \subset \mathcal{H}(1.126) \subset \ldots \subset \mathcal{H}(1126) \subset \ldots \subset \mathcal{H}(\infty) = \mathcal{H}_{10}$

regularized hypothesis $\mathbf{w}_{\text{REG}}$:
optimal solution from
regularized hypothesis set $\mathcal{H}(C)$

Fun Time

For $Q \geq 1$, which of the following hypothesis (weight vector $\mathbf{w} \in \mathbb{R}^{Q+1}$) is not in the regularized hypothesis set $\mathcal{H}(1)$?

1. $\mathbf{w}^T = [0, 0, \ldots, 0]$
2. $\mathbf{w}^T = [1, 0, \ldots, 0]$
3. $\mathbf{w}^T = [1, 1, \ldots, 1]$
4. $\mathbf{w}^T = \left[\sqrt{\frac{1}{Q+1}}, \sqrt{\frac{1}{Q+1}}, \ldots, \sqrt{\frac{1}{Q+1}}\right]$

Reference Answer: 3

The squared length of $\mathbf{w}$ in 3 is $Q+1$, which is not ≤ 1.

Matrix Form of Regularized Regression Problem

$$\min_{\mathbf{w}\in\mathbb{R}^{Q+1}} \quad E_{\text{in}}(\mathbf{w}) = \frac{1}{N}\underbrace{\sum_{n=1}^{N}(\mathbf{w}^T\mathbf{z}_n - y_n)^2}_{(\mathrm{Z}\mathbf{w}-\mathbf{y})^T(\mathrm{Z}\mathbf{w}-\mathbf{y})}$$

$$\text{s.t.} \quad \underbrace{\sum_{q=0}^{Q} w_q^2}_{\mathbf{w}^T\mathbf{w}} \leq C$$

- $\sum_n \cdots$ = $(\mathrm{Z}\mathbf{w} - \mathbf{y})^T(\mathrm{Z}\mathbf{w} - \mathbf{y})$, **remember? :-)**
- $\mathbf{w}^T\mathbf{w} \leq C$: feasible $\mathbf{w}$ within a radius-$\sqrt{C}$ hypersphere

how to solve
constrained optimization problem?

Augmented Error

- if oracle tells you $\lambda > 0$, then

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

$$\frac{2}{N}\left(\mathrm{Z}^T\mathrm{Z}\mathbf{w}_{\text{REG}} - \mathrm{Z}^T\mathbf{y}\right) + \frac{2\lambda}{N}\boxed{\mathbf{w}_{\text{REG}}} = \mathbf{0}$$

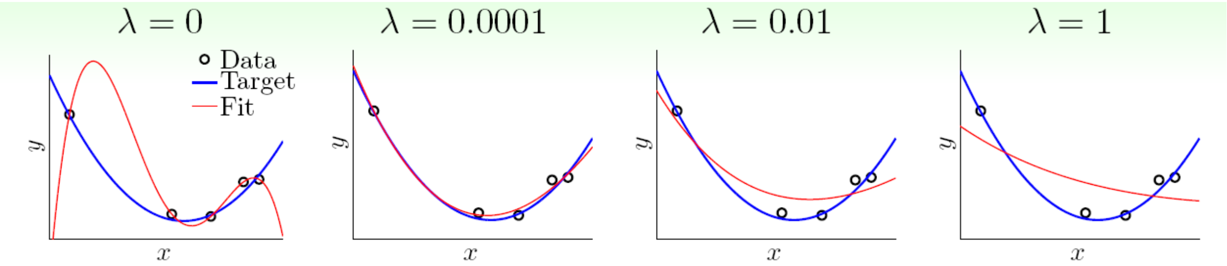
- optimal solution:

$$\mathbf{w}_{\text{REG}} \leftarrow (\mathrm{Z}^T\mathrm{Z} + \lambda\mathrm{I})^{-1}\mathrm{Z}^T\mathbf{y}$$

—called ridge regression in Statistics

minimizing unconstrained E_{aug} effectively
minimizes some C-constrained E_{in}

The Results



overfitting $\Longrightarrow$ $\Longrightarrow$ $\Longrightarrow$ underfitting

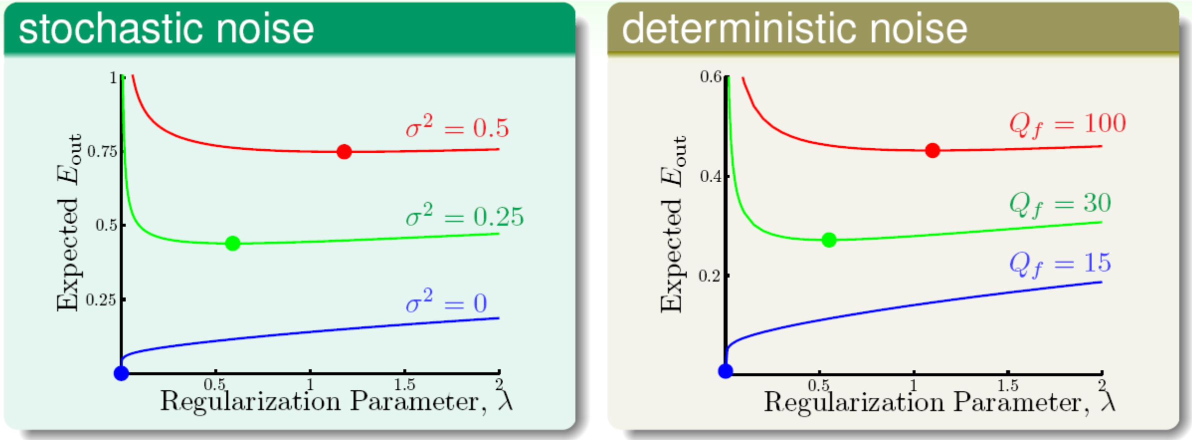
philosophy: *a little* ***regularization*** *goes a long way!*

call '$+\frac{\lambda}{N}\mathbf{w}^T\mathbf{w}$' **weight-decay** regularization:

larger λ
$\Longleftrightarrow$ prefer shorter $\mathbf{w}$
$\Longleftrightarrow$ effectively smaller C

—go with 'any' transform + linear model

The Optimal λ



- more noise $\Longleftrightarrow$ more regularization needed
  —more bumpy road $\Longleftrightarrow$ putting brakes more
- noise **unknown**—important to **make proper choices**

how to choose?
stay tuned for the next lecture! :-)

Regularization for Neural Network

basic choice:

old friend weight-decay (L2) regularizer $\Omega(\mathbf{w}) = \sum \left(w_{ij}^{(\ell)} \right)^2$

- 'shrink' weights:
  large weight $\rightarrow$ large shrink; small weight $\rightarrow$ small shrink
- want $w_{ij}^{(\ell)} = 0$ (sparse) to effectively **decrease** d_{VC}
  - L1 regularizer: $\sum \left| w_{ij}^{(\ell)} \right|$, but **not differentiable**
  - weight-elimination ('scaled' L2) regularizer:
    large weight $\rightarrow$ median shrink; small weight $\rightarrow$ median shrink

weight-elimination regularizer: $\sum \frac{\left(w_{ij}^{(\ell)} \right)^2}{1 + \left(w_{ij}^{(\ell)} \right)^2}$