

# Learning and Generalization

Toru Wakahara

## [1] 学習と汎化の問題

- 2乗和エラー関数最小化の意味

## [2] バイアスと分散

- trade-off 関係

## [3] 正則化と汎化能力

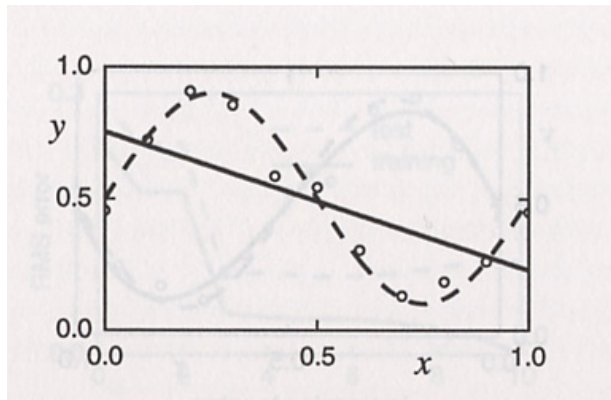
- Weight decay
- Early stopping

## [4] モデルの複雑さと汎化能力

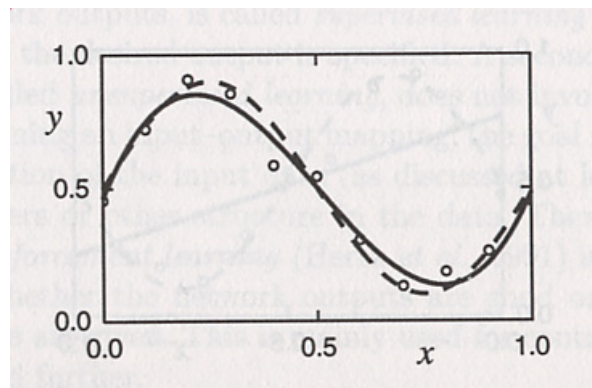
- Cross-validation
- Complexity criteria

# 学習と汎化の問題－回帰

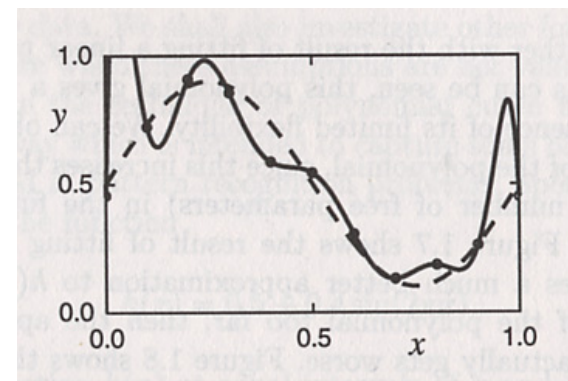
多項式による回帰:  $y(x) = w_0 + w_1x + \dots + w_Mx^M$



$M = 1$

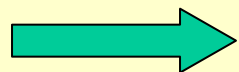


$M = 3$



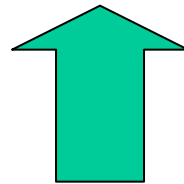
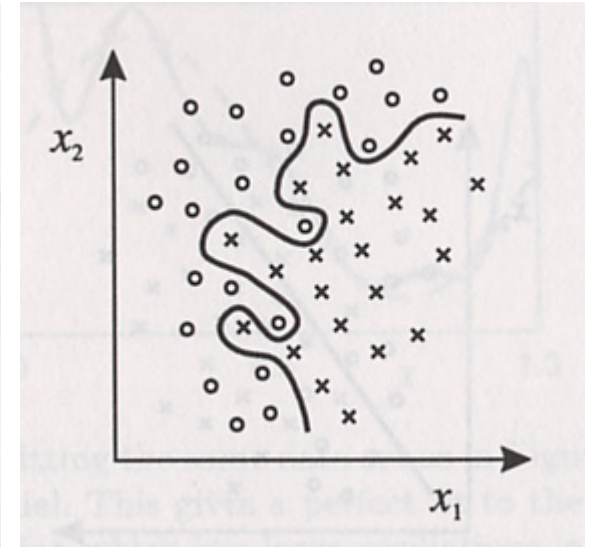
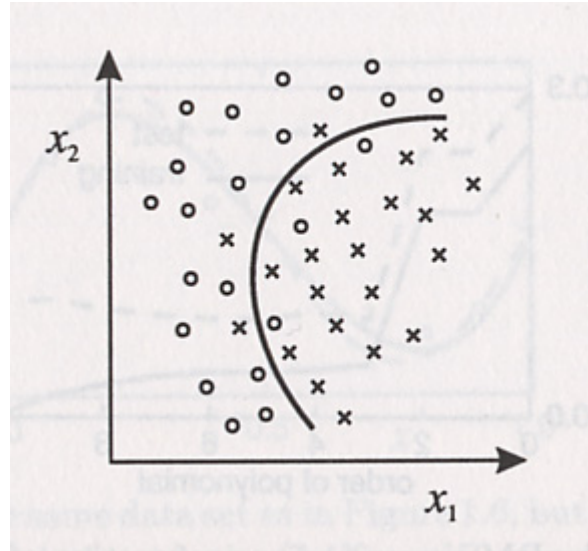
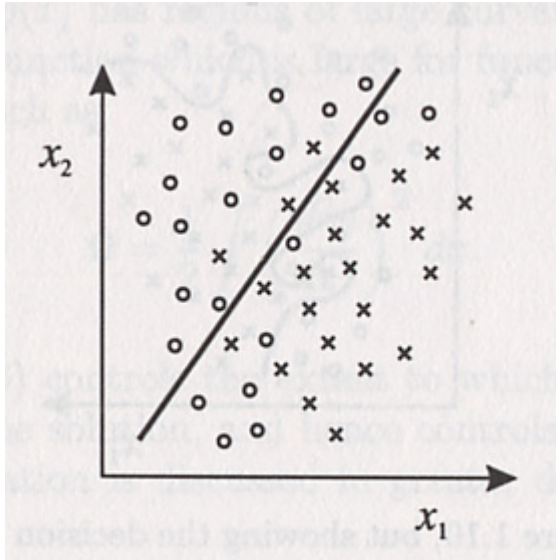
$M = 10$

汎化とは、学習データの正確な表現を求めることでなく、データの発生過程の統計的モデルを構築すること



モデルの複雑さはどう決定したらよいのか

# 学習と汎化の問題一分類

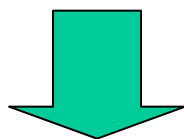


*Occam's razor* – William of Occam (1285–1349)

[データとの食い違い] + [モデルの複雑さ] → 最小

# 2乗和エラー関数の最小化の意味

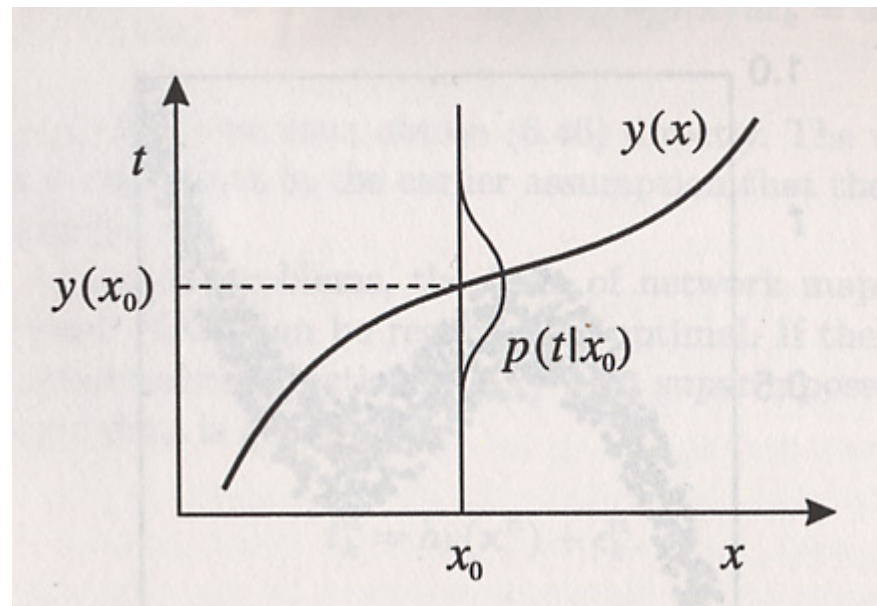
$$E = \frac{1}{2} \int \{ y(\mathbf{x}; \mathbf{w}) - \langle t | \mathbf{x} \rangle \}^2 p(\mathbf{x}) d\mathbf{x} \\ + \frac{1}{2} \int \{ \langle t^2 | \mathbf{x} \rangle - \langle t | \mathbf{x} \rangle^2 \} p(\mathbf{x}) d\mathbf{x}$$



$E$  の最小化:

$$y(\mathbf{x}; \mathbf{w}^*) = \langle t | \mathbf{x} \rangle$$

最適解

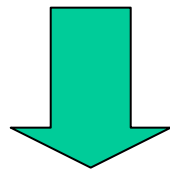


# バイアスと分散 (1)

$$E = \frac{1}{2} \int \{ y(\mathbf{x}; \mathbf{w}) - \langle t | \mathbf{x} \rangle \}^2 p(\mathbf{x}) d\mathbf{x} \quad : \mathbf{w} \text{ に依存し}$$
$$+ \frac{1}{2} \int \{ \langle t^2 | \mathbf{x} \rangle - \langle t | \mathbf{x} \rangle^2 \} p(\mathbf{x}) d\mathbf{x} \quad : \text{固有ノイズで}$$

最小化できる  
 $E$  の下限値

$\{ y(\mathbf{x}) - \langle t | \mathbf{x} \rangle \}^2$  : 特定の学習データ  $D$  に依存



$D$  に関する期待値 (集合平均)  
を評価してみよう

$$\mathcal{E}_D[\{ y(\mathbf{x}) - \langle t | \mathbf{x} \rangle \}^2]$$

## バイアスと分散 (2)

$$\begin{aligned}\{y(\mathbf{x}) - \langle t | \mathbf{x} \rangle\}^2 &= \{y(\mathbf{x}) - \mathcal{E}_D[y(\mathbf{x})] + \mathcal{E}_D[y(\mathbf{x})] - \langle t | \mathbf{x} \rangle\}^2 \\ &= \{y(\mathbf{x}) - \mathcal{E}_D[y(\mathbf{x})]\}^2 + \{\mathcal{E}_D[y(\mathbf{x})] - \langle t | \mathbf{x} \rangle\}^2 \\ &\quad + 2\{y(\mathbf{x}) - \mathcal{E}_D[y(\mathbf{x})]\}\{\mathcal{E}_D[y(\mathbf{x})] - \langle t | \mathbf{x} \rangle\}\end{aligned}$$

$D$  に関する期待値を取ると第3項は消えて次式を得る

$$\begin{aligned}\mathcal{E}_D[\{y(\mathbf{x}) - \langle t | \mathbf{x} \rangle\}^2] \\ &= \underbrace{\{\mathcal{E}_D[y(\mathbf{x})] - \langle t | \mathbf{x} \rangle\}^2}_{(\text{bias})^2} + \underbrace{\mathcal{E}_D[\{y(\mathbf{x}) - \mathcal{E}_D[y(\mathbf{x})]\}^2]}_{\text{variance}}\end{aligned}$$

# バイアスと分散 (3)

$$(\text{bias})^2 = \frac{1}{2} \int \{ \mathcal{E}_D[y(\mathbf{x})] - \langle t | \mathbf{x} \rangle \}^2 p(\mathbf{x}) d \mathbf{x}$$

$$\text{variance} = \frac{1}{2} \int \mathcal{E}_D[\{ y(\mathbf{x}) - \mathcal{E}_D[y(\mathbf{x})] \}^2] p(\mathbf{x}) d \mathbf{x}$$

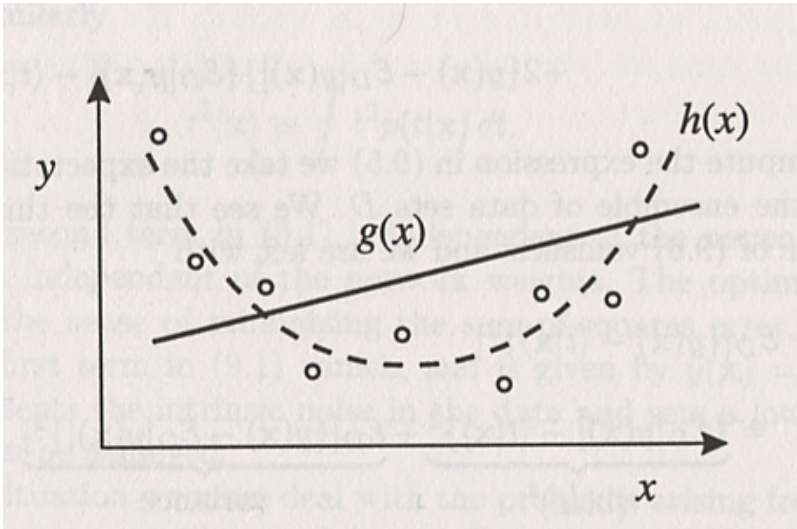
**bias**: 出力  $y(\mathbf{x})$  の集合平均が最適解  $\langle t | \mathbf{x} \rangle$  とどれだけ離れているかを評価したもの

**variance**: 出力  $y(\mathbf{x})$  がデータセット  $D$  の特定な選択にどれだけ敏感かを評価したもの

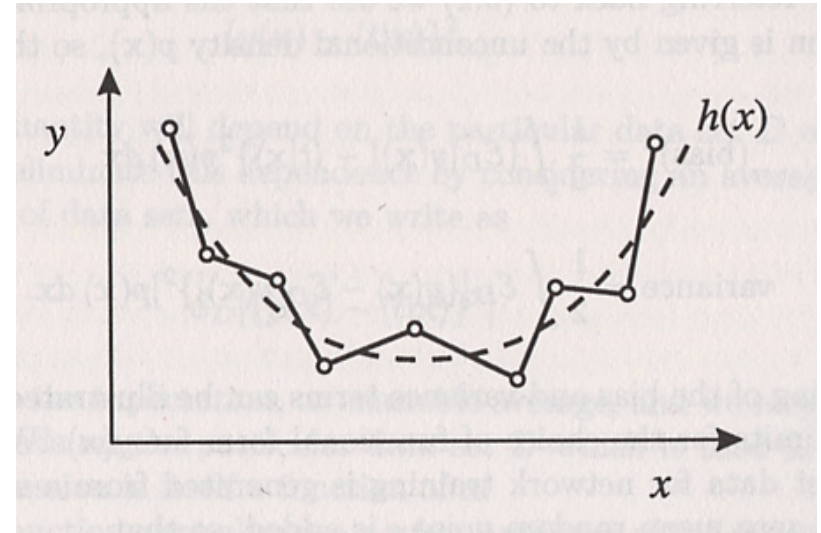
**bias と variance の間には trade-off 関係がある!**

# バイアスと分散 (4)

$$t^n = h(\mathbf{x}_n) + \varepsilon_n$$



variance = 0  
(bias)<sup>2</sup> 大



variance 大  
(bias)<sup>2</sup> 小

汎化能力から見て最適なバランスをどう決めるか

# 正則化と汎化能力

汎化能力: 学習データへの over-fitting でなく  
未知データに対する予測能力

## 正則化

$$\tilde{E} = E + \nu \Omega \quad \Omega: \text{正則化のペナルティ関数}$$

$\Omega$  の例:

$$\Omega = \frac{1}{2} \sum_i w_i^2$$

: *weight decay*

$$\Omega = \frac{1}{2} \sum_{n=1}^N \sum_{i=1}^d \sum_{k=1}^c \left( \frac{\partial^2 y_{n,k}}{\partial x_i^2} \right)^2$$

: *Tikhonov regularizer*

# 正則化 by weight decay (1)

エラー関数の微分:

$$\tilde{E} = E + \nu \Omega, \quad \Omega = \frac{1}{2} \sum_i w_i^2 \quad \rightarrow \quad \nabla \tilde{E} = \nabla E + \nu \mathbf{w}$$

$E$  を無視して  $\mathbf{w}$  についての更新を連続時間で考えると

$$\Delta \mathbf{w} = -\eta \nabla \tilde{E} = -\eta \nabla \nu \Omega = -\eta \nu \mathbf{w}$$

$$\therefore \frac{d \mathbf{w}}{dt} = -\frac{\eta}{\Delta} \nu \mathbf{w} \quad \rightarrow \quad \mathbf{w}(t) = \mathbf{w}(0) \exp\left(-\frac{\eta \nu}{\Delta} t\right)$$

$\mathbf{w}$  は指数的に減衰する

$\rightarrow$  *weight decay*

# 正則化 by weight decay (2)

エラー関数が2次の場合:  $b, H$  は定数で

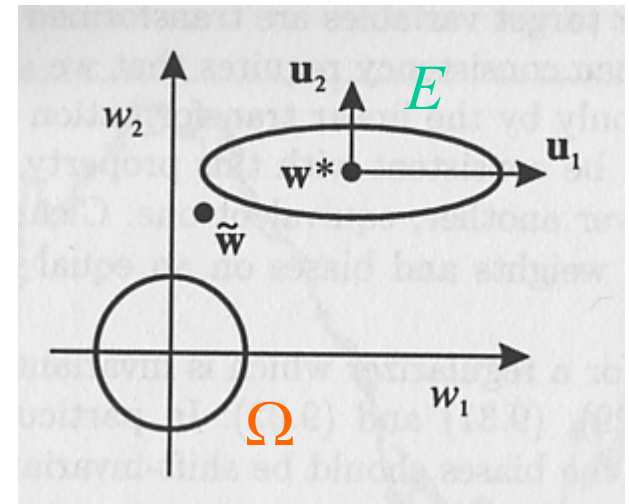
$$E(\mathbf{w}) = E_0 + \mathbf{b}^t \mathbf{w} + \frac{1}{2} \mathbf{w}^t \mathbf{H} \mathbf{w}$$

$E$  の最小点:  $\mathbf{b} + \mathbf{H} \mathbf{w}^* = 0$

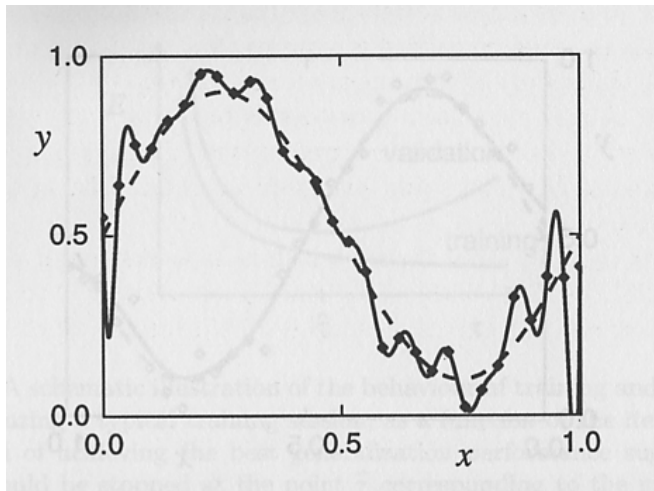
$\tilde{E}$  の最小点:  $\mathbf{b} + \mathbf{H} \tilde{\mathbf{w}} + \nu \tilde{\mathbf{w}} = 0$

$\mathbf{H} \mathbf{u}_j = \lambda_j \mathbf{u}_j$  の展開を用いると

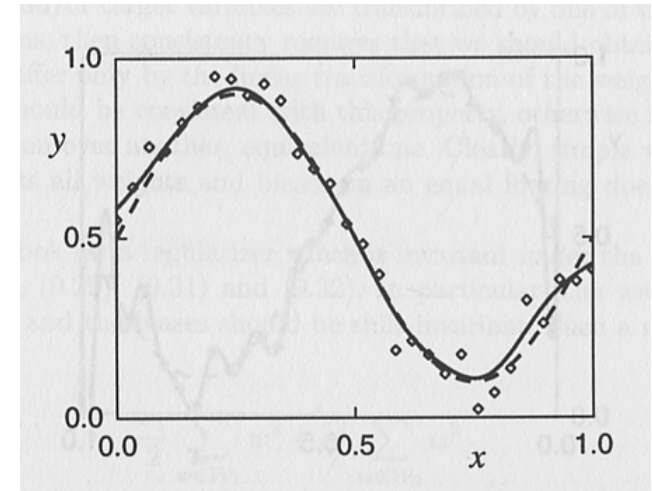
$$\mathbf{w}^* = \sum_j w_j^* \mathbf{u}_j, \quad \tilde{\mathbf{w}} = \sum_j \tilde{w}_j \mathbf{u}_j \quad \rightarrow \quad \tilde{w}_j = \frac{\lambda_j}{\lambda_j + \nu} w_j^*$$



# 正則化 by weight decay (3)



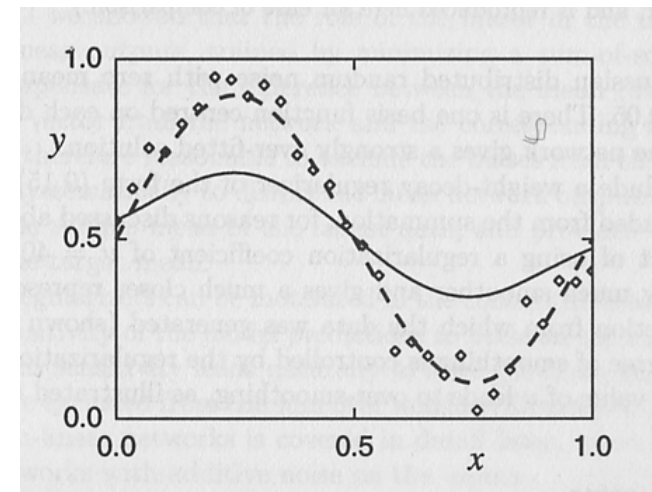
$$\tilde{E} = E + \nu \Omega$$



$$\nu = 40$$

雑音  $N(0.0, 0.05)$  を含む  
正弦波形:

$y = 0.5 + 0.4 \sin(2\pi x)$   
の30点の RBF による  
2乗和エラー  $E$  最小化  
で内挿した結果

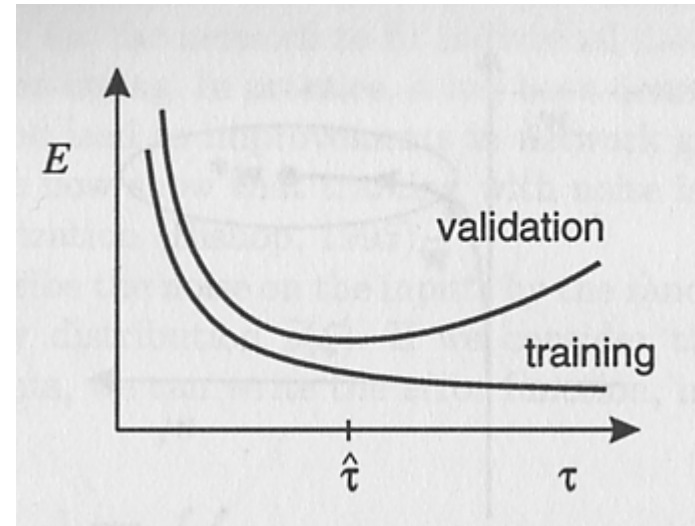


$$\nu = 1000$$

# Early stopping (1)

学習を途中で打ち切ることでモデルの複雑さを制御できる

→ over-fitting の抑制



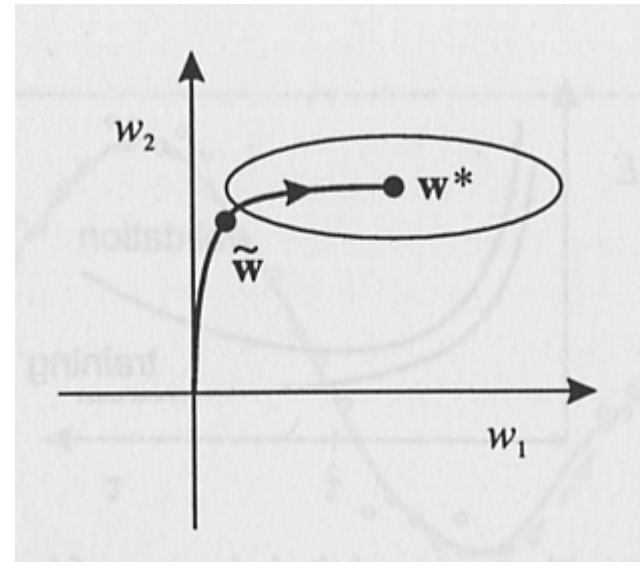
*training set* で学習を進めると  $E$  は減少し続ける



*validation set* での  $E$  が最小となった時点  $\hat{\tau}$  で学習を打ち切る

# Early stopping (2)

エラー関数が2次の場合,  
実は *early stopping* が  
正則化 by *weight decay* と  
似た効果を持つことを  
示すことができる！



$w$  が原点からスタートすると, 2次曲面の局所的な  
勾配方向  $-\nabla E$  に従い, 図のように  $w$  は移動する

途中でストップした点  $\tilde{w}$  は *weight decay* による  
正則化で得られる  $\tilde{w}$  と似ている

# 課題1

$E(\mathbf{w})$  が2次の場合の *early stopping* の振舞いを調べる。  
次式で The Hessian matrix  $H$  は正定値で定数とする。

$$E(\mathbf{w}) = E(\mathbf{w}^*) + \frac{1}{2}(\mathbf{w} - \mathbf{w}^*)^t H (\mathbf{w} - \mathbf{w}^*)$$

$\mathbf{w}^{(0)}$  を原点とし, 次式で  $\mathbf{w}^{(\tau)}$  を更新する。

$$\mathbf{w}^{(\tau)} = \mathbf{w}^{(\tau-1)} - \eta \nabla E$$

このとき,  $H$  の固有ベクトル, 固有値を用いて次の関係が  
成り立つことを証明しなさい。

$$w_j^{(\tau)} = \{1 - (1 - \eta \lambda_j)^\tau\} w_j^*$$

$$\text{where } w_j = \mathbf{w}^t \mathbf{u}_j, \quad H \mathbf{u}_j = \lambda_j \mathbf{u}_j$$

# 課題1 (続き)

続いて次の関係も示さない。

$$\mathbf{w}^{(\tau)} \rightarrow \mathbf{w}^* \quad \text{as } \tau \rightarrow \infty \quad \text{if } |1 - \eta\lambda_j| < 1$$

ここで、有限のステップ回数  $\tau$  でストップした場合、すなわち *early stopping* を行った場合、次の関係が成り立つことを証明しなさい。

$$w_j^{(\tau)} \cong w_j^* \quad \text{when } \lambda_j \gg (\eta\tau)^{-1} \quad (1)$$

$$|w_j^{(\tau)}| \ll |w_j^*| \quad \text{when } \lambda_j \ll (\eta\tau)^{-1} \quad (2)$$

# 課題1 (続き)

最後に、前頁で得られた式 (1), (2) と、*weight decay* による正則化で得られる式 (3) との対応関係について論じなさい。

*early stopping* の場合：

$$w_j^{(\tau)} \cong w_j^* \quad \text{when } \lambda_j \gg (\eta\tau)^{-1} \quad (1)$$

$$\left| w_j^{(\tau)} \right| \ll \left| w_j^* \right| \quad \text{when } \lambda_j \ll (\eta\tau)^{-1} \quad (2)$$

*weight decay regularizer* の場合：

$$\tilde{w}_j = \frac{\lambda_j}{\lambda_j + \nu} w_j^* \quad (3)$$

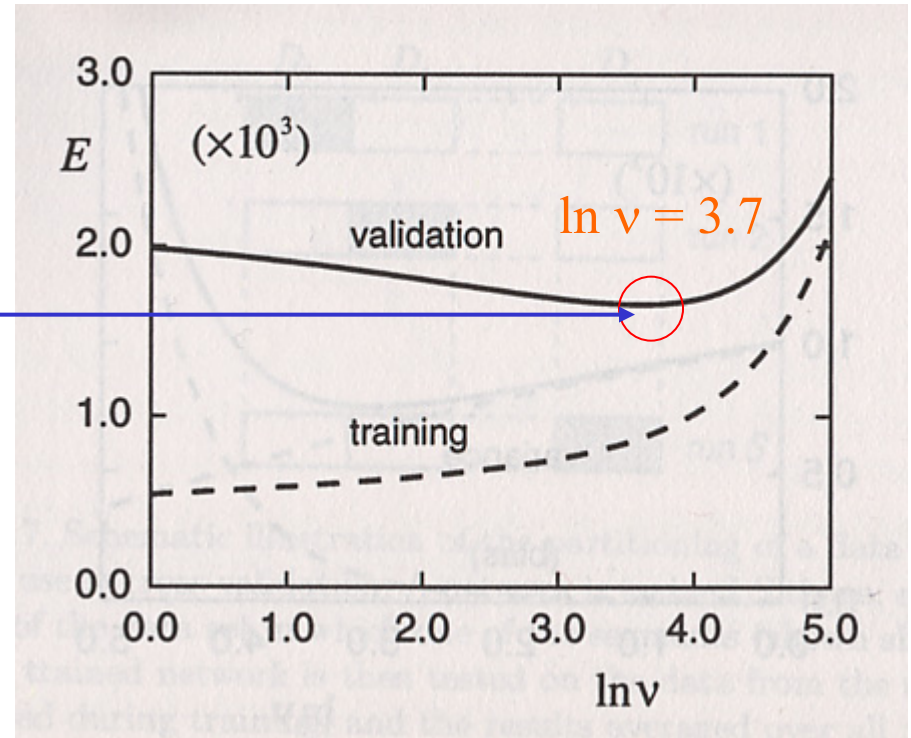
# 汎化能力の評価 (1)

正則化パラメータ  
 $\nu$  の選択:

$$\tilde{E} = E + \nu\Omega$$

[1] *hold-out method*  
data set を学習と評価  
用に2分割

- *training set*
- *validation set*



➡ 最後に独立なデータ  
*test set* で評価

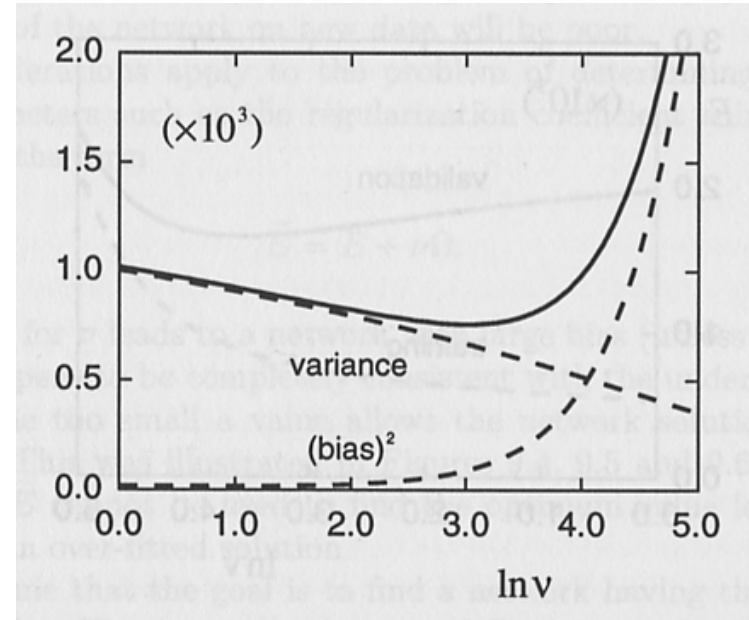
# 汎化能力の評価 (2)

RBF による30点の近似の例で,  
100データセットを発生して,  
bias-variance のtrade-offを調べる

$$\bar{y}(x) = \frac{1}{100} \sum_{i=1}^{100} y_i(x)$$

$$(\text{bias})^2 = \sum_n \{ \bar{y}(x_n) - h(x_n) \}^2$$

$$\text{variance} = \sum_n \frac{1}{100} \sum_{i=1}^{100} \{ y_i(x_n) - \bar{y}(x_n) \}^2$$



(bias)<sup>2</sup>+variance の和の  
最小に相当する値v\*は  
validation setでの最適値  
v\*とほぼ一致する！

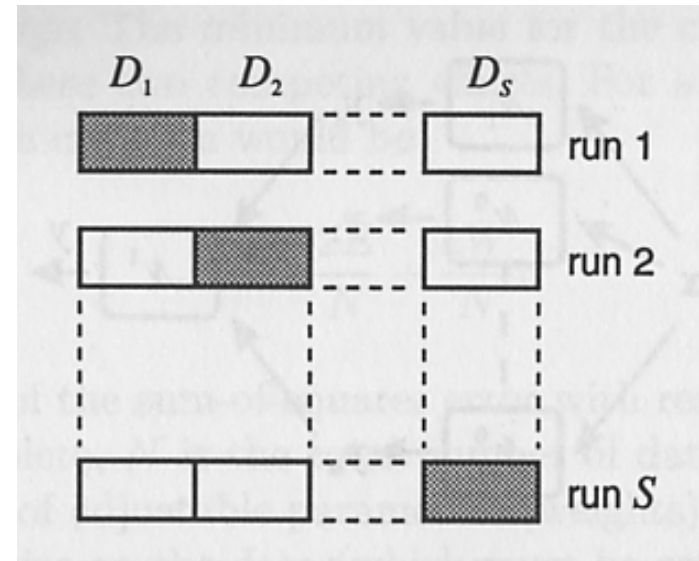
# 汎化能力の評価 (3)

## [2] *cross-validation*

data sets を  $S$  個に等分割して  $S - 1$  個の segments で学習して, 残り 1 個の segment で評価する.  
これを  $S$  回循環して平均化する.

## [3] *leave-one-out method*

[2] で  $S = N$  (データの総点数) とするもの.



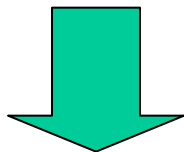
# モデルの複雑さの指標 (1)

validation set を用いずに汎化能力を評価する

→ モデルの複雑さそのものを評価

予測誤差:

PE = training error + complexity term



$$PE = \frac{2E}{N} + \frac{2W}{N} \sigma^2$$

$E$ : 2乗和エラー

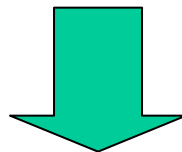
$N$ : 学習データ数

$W$ : モデルのパラメータ数

$\sigma^2$ : データノイズの分散

# モデルの複雑さの指標 (2)

予測誤差 (PE) の概念の一般化  
→ 正則化や非線形モデルも考慮



$$\text{GPE} = \frac{2E}{N} + \frac{2\gamma}{N} \sigma^2 \quad \text{where} \quad \gamma = \sum_{i=1}^W \frac{\lambda_i}{\lambda_i + \nu}$$

$\lambda_i \gg \nu$  となる固有値  $\lambda_i$  の個数が  $\gamma$  に寄与する

## 課題2

次に示すような線形ネットモデルと2乗和エラー関数を考える.

$$y_k = \sum_i w_{ki} x_i + w_{k0}$$

$$E = \frac{1}{2N} \sum_{n=1}^N \sum_k \{y_k(\mathbf{x}_n) - t_k^n\}^2$$

ここで, 入力  $x_i$  に対して, ランダム雑音  $\varepsilon_i$  が加えられるとする.

$$x_i \rightarrow x_i + \varepsilon_i$$

$$\langle \varepsilon_i \rangle = 0, \quad \langle \varepsilon_i \varepsilon_j \rangle = \delta_{ij} \nu$$

## 課題2(続き)

この雑音に乗った場合について、新しい2乗和エラー関数  $E$  を求めなさい。但し、雑音  $\varepsilon_i$  についての平均操作を用いること。

そして、得られた結果が、次式のような weight decay による正則化と等価であることを示しなさい。

$$\tilde{E} = E + \nu\Omega, \quad \Omega = \frac{1}{2} \sum_k \sum_{i \neq 0} w_{ki}^2$$

この結果より、雑音を加えた入力を用いた学習が、overfitting を抑制し、weight decay による正則化と同等な効果を持つことが分かる。