

CS599—Contents XI

- Learning and Generalization

- + bias-variance tradeoff

- + methods to improve generalization

- u regularization

- u training with noise

- u soft-weight sharing

- u growing and pruning algorithms

- u committee networks

- u cross validation

- Handout:

- u Class Notes

- Reading Assignment for Next Class

- u Bishop, Ch. 6 (browse the individual sections to obtain an idea about what the topics are)

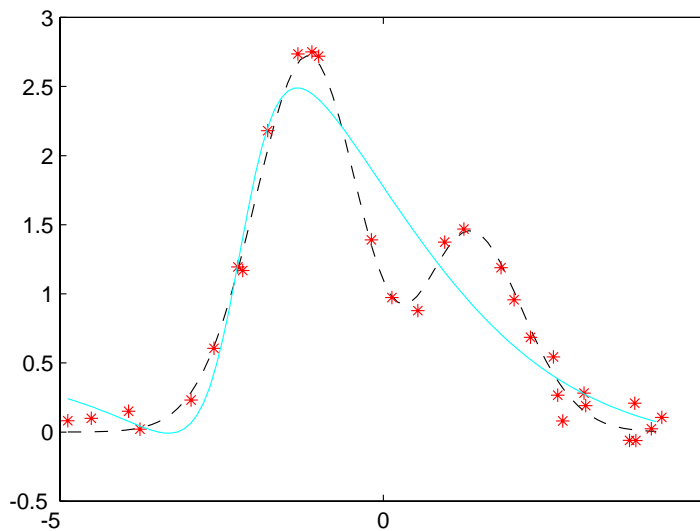
Bias-Variance Tradeoff

- Remember: The Means Squared Error can be decomposed into three terms

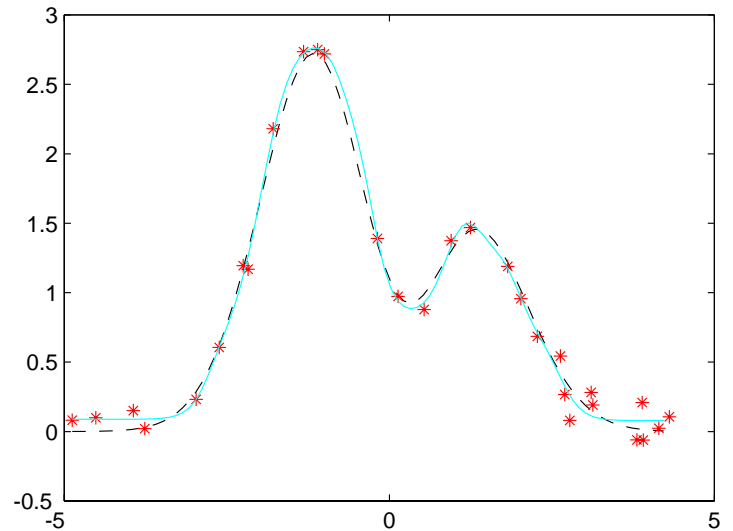
$$\begin{aligned} E\{J\} &= \sigma_{\varepsilon}^2 + (E\{y\} - f)^2 + E\{(y - E\{y\})^2\} = \\ &= \text{var}(\textit{noise}) + \textit{bias}^2 + \text{var}(\textit{estimate}) \end{aligned}$$

- Minimizing all these terms simultaneously can be ensured by:
 - + more data
 - + good prior information about how the learning problem
 - + special techniques that try to control bias & variance as a function of the available data

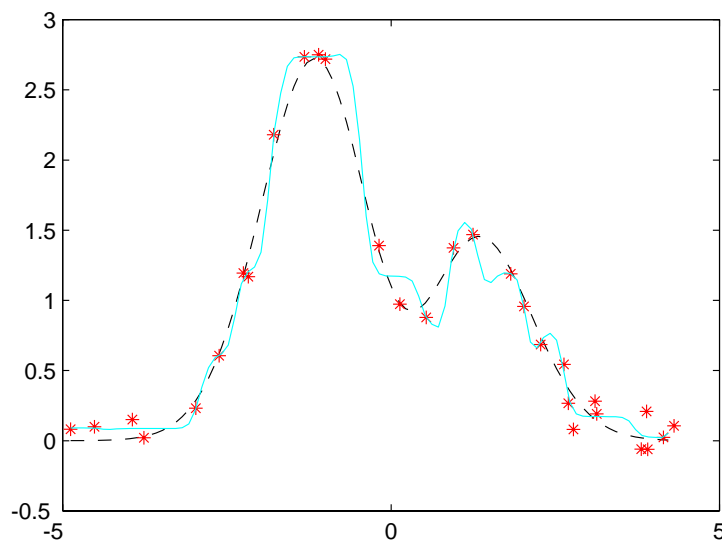
Underfitting/Overfitting



2 hidden units



10 hidden units



30 hidden units

Regularization

- Add a penalty to the cost function

$$\tilde{J} = J + \gamma P$$

$$\text{e.g., } P = \int \left(\frac{d^2 y}{dx^2} \right)^2 dx$$

- Methods of Regularization

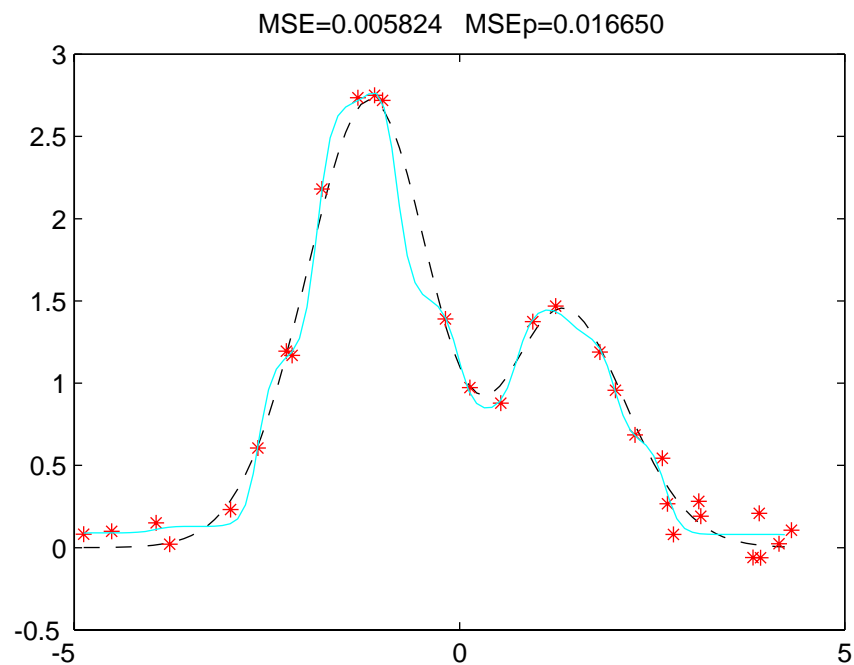
- + RBF networks are implicit regularizers
- + Weight decay
- + Early stopping
- + Curvature driven smoothing

Weight Decay

- Add Penalty Term

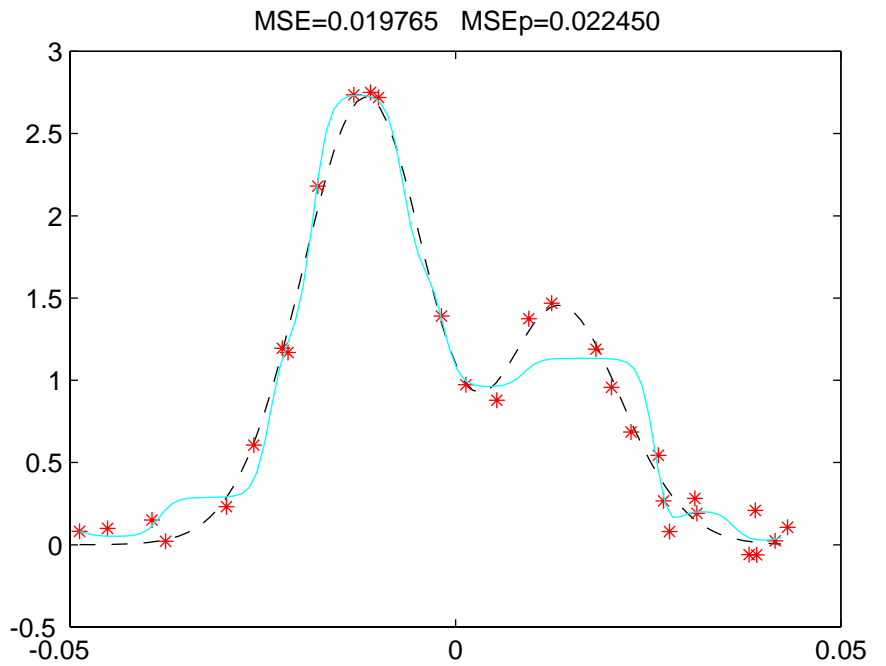
$$P = \frac{1}{2} \sum_i w_i^2$$

- Is weight decay consistent?
- Should the bias weights decay as well?

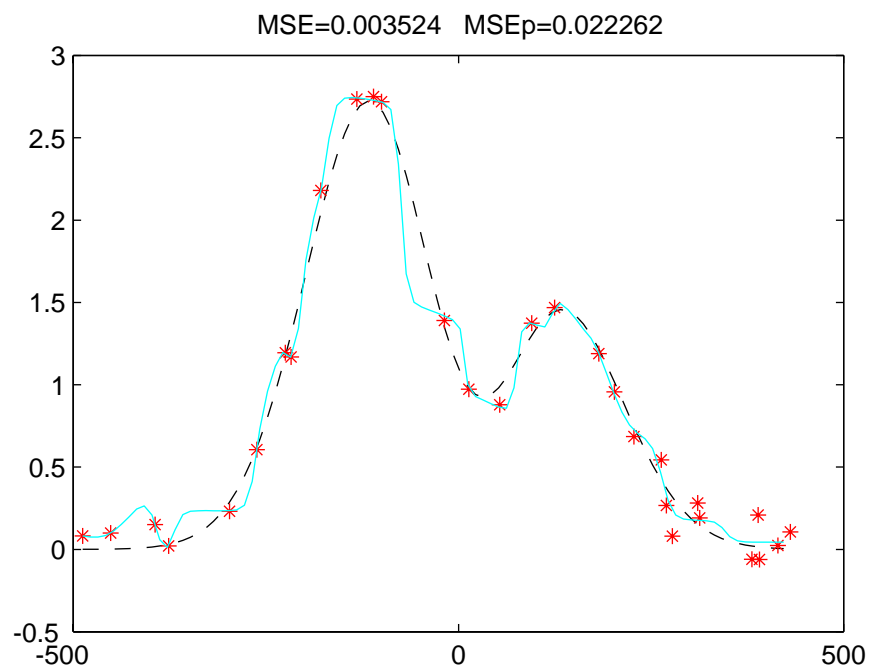


Inconsistency of Weight Decay

$x \rightarrow x * 0.01$

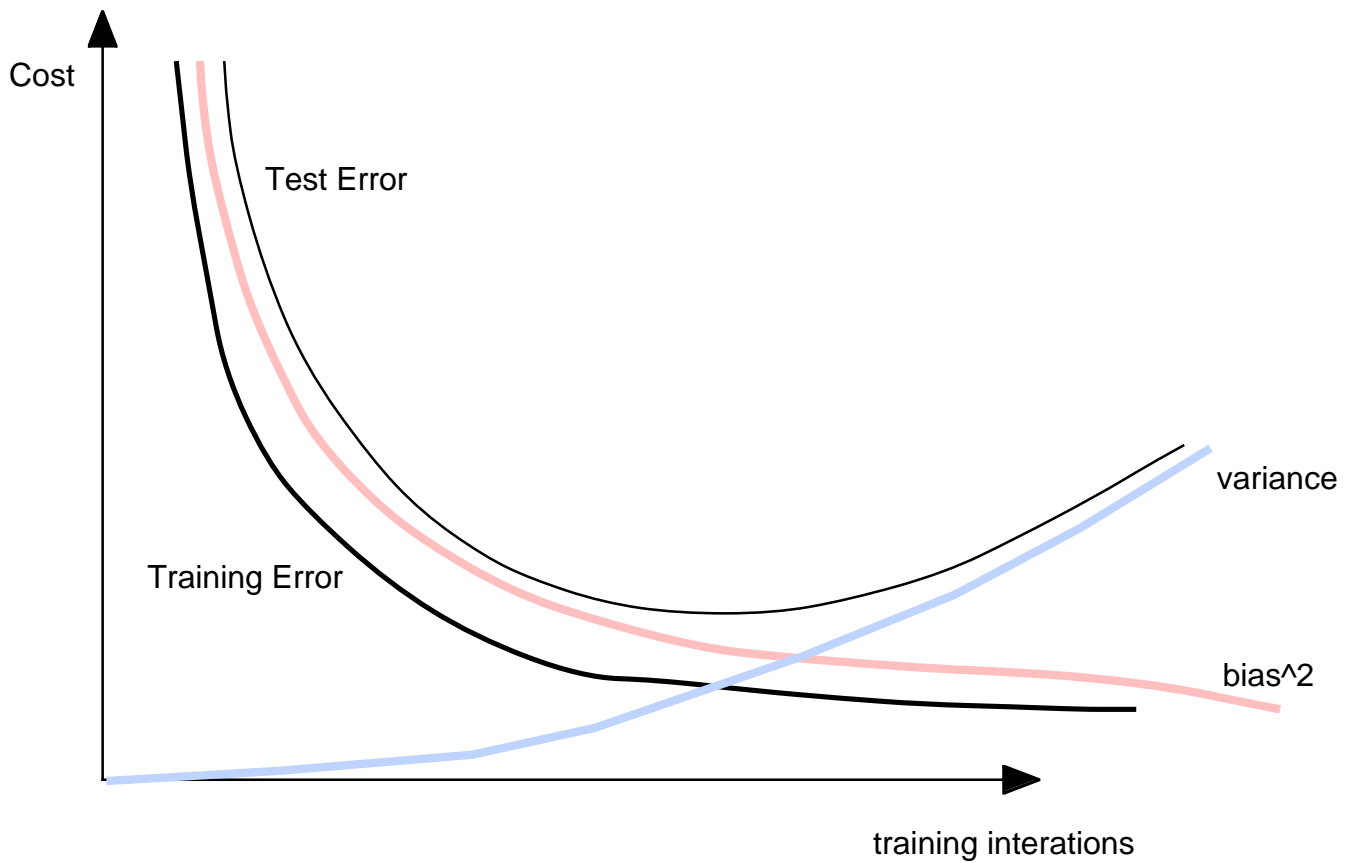


$x \rightarrow x * 100$



Early Stopping

- Use a validation set to notice when generalization error is the smallest
 - + however this point is not so easy to detect during training



Curvature Driven Smoothing

- Penalty based on Curvature

$$\tilde{J} = J + \gamma P$$

$$P = \frac{1}{2} \sum_{n=1}^N \sum_{i=1}^d \sum_{k=1}^c \frac{\partial^2 y_k^n}{\partial x_i^2}$$

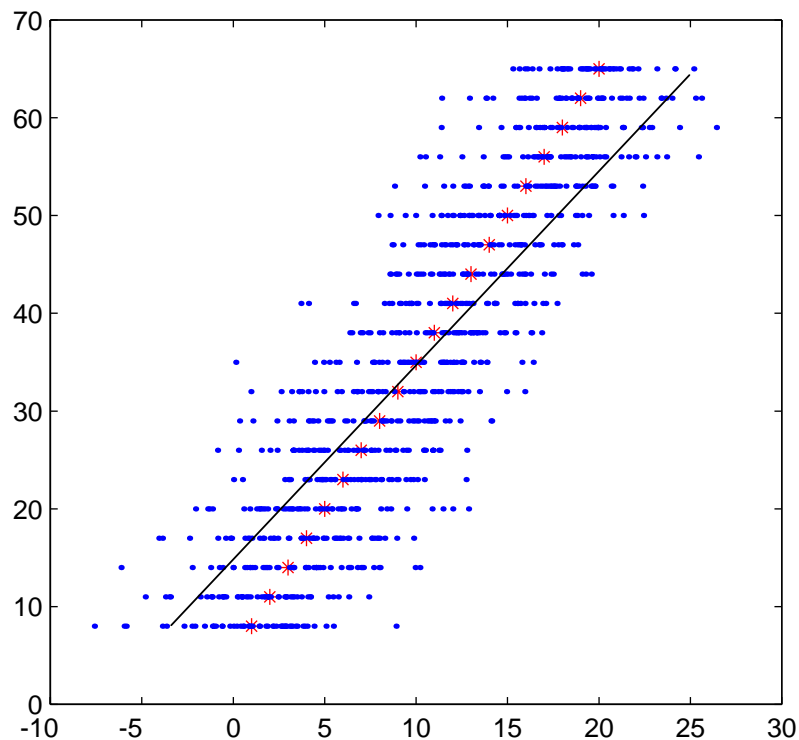
u need to calculate the second derivatives through network

Training with Noise

- Add a small amount of noise to the INPUT data
 - + this method can be shown to be equivalent to the following regularization function

$$\tilde{J} = J + \gamma P$$

$$P = \frac{1}{2} \sum_{n=1}^N \sum_{i=1}^d \frac{\partial y_k^n}{\partial x_i}$$



Soft Weight Sharing

- Weights are assumed to be generated from discrete “weight clusters”, i.e., a mixture model

+ for 1-D weights

$$p(w) = \sum_{k=1}^K \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{1}{2\sigma_k^2}(w - w_{0,k})^2\right)$$

- This results in a regularization penalty term

$$\tilde{J} = J + \gamma P$$

$$P = -\sum_{n=1}^N \ln\left(\sum_{k=1}^K \alpha_k p_k(w)\right)$$

u Derivatives of this cost function are straightforward (see Bishop Ch. 9.4)

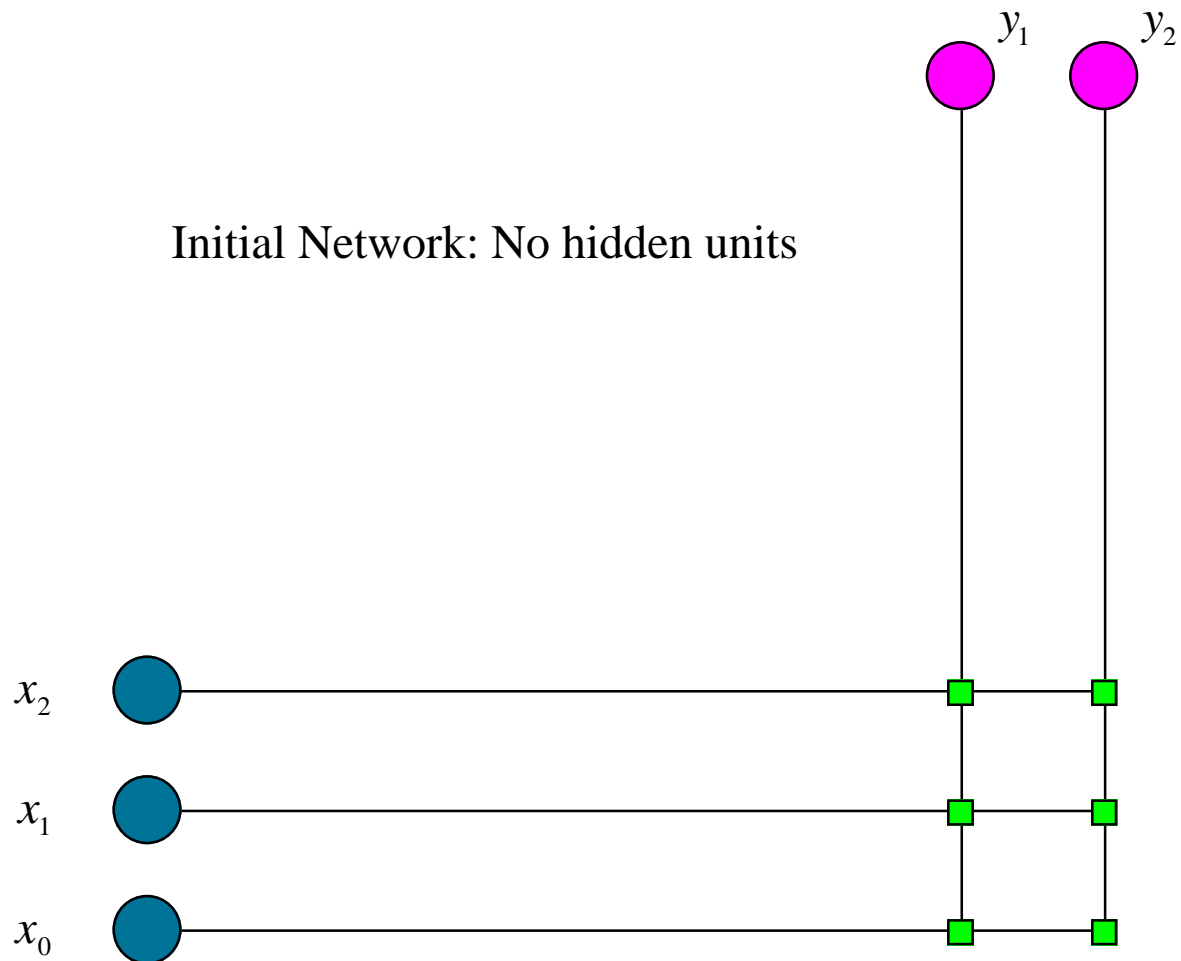
Constructive Algorithms

- Allocate the “right” number of units corresponding to the given data
 - + Growing Algorithms
 - + Pruning Algorithms
 - + Danger of Overfitting
 - + Danger of unprincipled heuristics

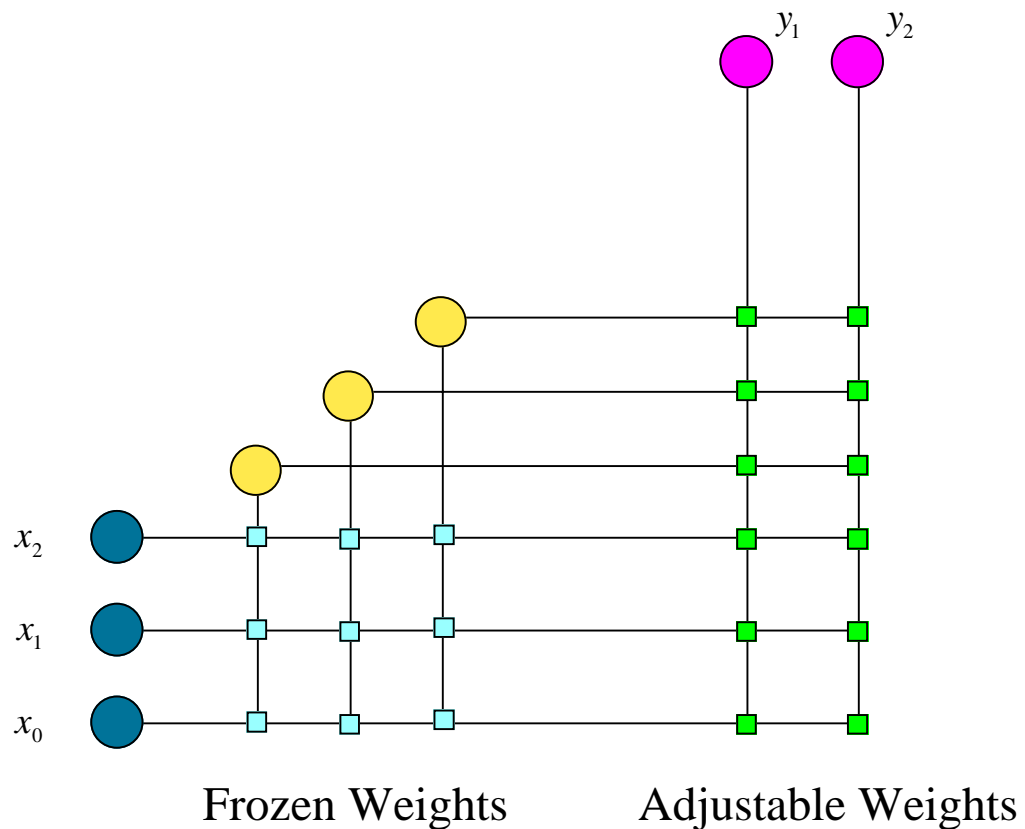
Example:

Cascade Correlation

- Add hidden units by creating an increasingly deeper-layered network



Cascade Correlation (cont'd)



How to add new unit?:

Train a “candidate pool” (possibly with different nonlinearities) and add the unit that maximizes:

$$S = \sum_{m=1}^M \left| \sum_{n=1}^n (z^n - \bar{z})(\epsilon_k^n - \bar{\epsilon}_k) \right|$$

Each candidate units adjusts its weights by gradient descent in S

Committee Networks

- Use average of several networks for prediction
+ this make efficient use of many trained networks

$$MSE_k = E\left\{(t(\mathbf{x}) - y_k(\mathbf{x}))^2\right\}$$

$$\text{Average } MSE = \frac{1}{K} \sum_{k=1}^K MSE_k$$

$$y_{com}(\mathbf{x}) = \frac{1}{K} \sum_{k=1}^K y_k(\mathbf{x})$$

$$\begin{aligned} \text{Committee } MSE &= E\left\{\left(t(\mathbf{x}) - \frac{1}{K} \sum_{k=1}^K y_k(\mathbf{x})\right)^2\right\} \\ &= E\left\{\left(\frac{1}{K} \sum_{k=1}^K e_k\right)^2\right\} \\ &= \frac{1}{K^2} \sum_{k=1}^K MSE_k \end{aligned}$$

Remarks:

- the error of the committee will always be smaller than the error of the WORST committee member
- a very bad committee member can make the committee perform very bad, even if all the other member are very good
- however, for reasonably chosen and trained members, the committee very often performs BETTER than the best member.