

Neuronale Netze (SS 2002), 27.5.

Architecture selection

- Theory says: one hidden layer with $|P|$ hidden neurons is enough for points P in $\mathbb{R}^n \times [0, 1]$, more hidden layer may save neurons.
- In practice: try several architectures and choose the best one. Thereby:
 - The best one has a small training error **and** generalizes to unseen data. The latter is tested computing the error on a set not used for training.
 - Algorithm:
divide P into training set P_1 and test set P_2 ,
train the architecture with $P_1 \rightarrow$ training error,
test the architecture on $P_2 \rightarrow$ test error,
architecture appropriate iff the test error is small.

This algorithm has a high variance depending on the separation.

- For scientific publications/validation of a new algorithm or architecture, . . . :
k-fold-cross-validation
divide P into P_1, \dots, P_k ,
train the architecture on $P \setminus P_i \rightarrow f_i$,
test f_i on $P_i \rightarrow E_i$,
estimation for the quality: $\sum E_i/k$

Training

Backprop is faced to severe problems:

- local instable maxima
- small derivatives in plateaus
- oscillation
- several of the above effects together in different dimensions

- sharp turns
- optima at the borders
- sgd'-factor

Various heuristics exist:

- **Online backpropagation:**

so-called stochastic gradient descent method

repeat:

 compute a permutation of the patterns

 for each pattern:

 compute the derivative and change the weights after each pattern

Pros and Cons: a little less stable, but avoids local instable optima, crosses small hills, there exist mathematical guarantees for the behavior

- **Backprop with momentum:** (only offline)

add in each step a small part of the direction of the previous change, i.e.

$$d_{ij} := -\eta \partial E / \partial w_{ij} + \alpha d_{ij}$$

Pros and Cons: larger steps in plateaus if the search goes into the same direction for several steps, damping of the step size if oscillation can be observed, possibility to cross small hills with the momentum, but the effect of reinforcing/supressing is limited