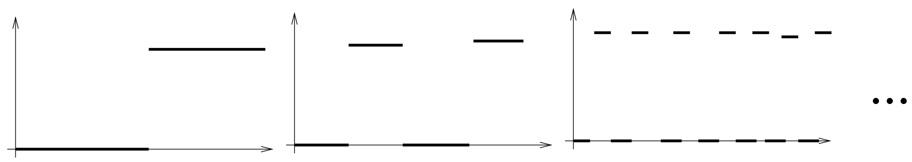
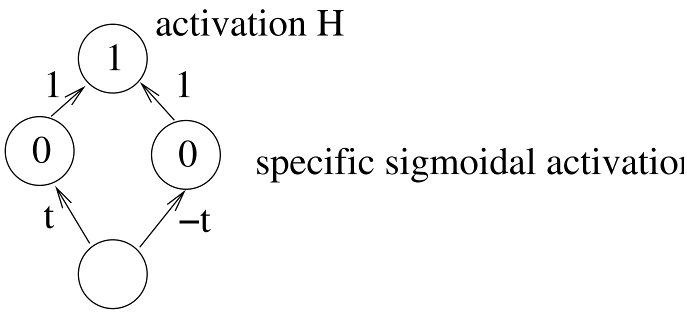
Neuronale Netze (SS 2002), 17.6.

Still some COLT theory

- Some examples:
  - Example: The class $\{f : [0,1] \to \{0,1\} \mid f$ equals one of the functions below $\}$ with the uniform distribution is not PAC learnable – the functions have pariwise distance $1/2$, i.e. e learning algorithm need sto precisely identify the underlying function. Since for every (general) set of points an infinite number of functions interpolates the points, this identification is not possible.



  - The function class $\{f : [0,1] \to \{0,1\} \mid f(x) = H(\cos(tx))$ for some $t \in \mathbb{R}\}$ is not PAC learnable – a function class with **only one parameter**!
  - Networks of the following form are not PAC-learnable:



    where the activation function is

$$\theta(x) = \frac{\arctan(x)}{\pi} + \frac{\cos(x)}{10(1+x^2)} + \frac{1}{2},$$

    a sigmoid-shaped function. The output can be computed as $H(cos(tx)/(5(1+x^2))) = H(cos(tx))$!!

- The **covering number** of a function class $\mathcal{F}$ is the smallest n such that n functions in $\mathcal{F}$ cover the entire space, i.e.

  $$N(\epsilon, \mathcal{F}, d_P) = \min\{n \mid \exists f_1, \ldots, f_n \in \mathcal{F}\, \forall f \in \mathcal{F}\, \exists f_i\, d_P(f, f_i) \leq \epsilon\}$$

  This might be infinite.

- $\mathcal{F}$ is PAC-learnable iff $N(\epsilon, \mathcal{F}, d_P)$ is finite for all ϵ. The minimum risk algorithm will do with $\sim \log N$ examples. Every algorithm needs $\sim \log N$ examples for valid generalization.

- $\mathcal{F}$ fulfills the **UCED-property** iff for all $\epsilon > 0$ the following holds

  $$P^m(\vec{x} \mid \sup_{f,g} |d_P(f,g) - \hat{d}_m(f,g,\vec{x})| > \epsilon) \to 0\ (m \to \infty)$$

  This means that the empirical error is representative for the generalization error for every two functions to be compared. In particular: every training algorithm with small training error will do!

- UCED $\Rightarrow$ PAC, but PAC is weaker than UCED (only one good algorithmn exists, not every algorithm with small error is ok for PAC).

- distribution independent PAC: ... the same holds with $\sup_P$

  distribution independent UCED: ... the same holds with $\sup_P$

  This means that one doesn't have to take care of the respective input distribution for the examples.

  The respective distribution independent notation is stronger than the distribution dependent one.

- $\mathcal{F}$ is distribution independent PAC $\iff$
  $\mathcal{F}$ is distribution independent UCED $\iff$
  the VC-dimension of $\mathcal{F}$ is finite !!!!!!!!!!!!!!!!!!

  $\sim$ VC-dimension examples are necessary and sufficient for training.

- The **VC-dimension** of $\mathcal{F}$ is the largest size of a set of points which can be shattered. A set $\{x_1, \ldots, x_m\}$ is **shattered** with $\mathcal{F}$ iff for all $d : \{x_1, \ldots, x_m\} \to \{0, 1\}$ a function $f \in \mathcal{F}$ exists with $d(x_i) = f(x_i)$.

  I.e. every function on $\{x_1, \ldots, x_m\}$ can be implemented within $\mathcal{F}$. No generalization can be expected on these points only!