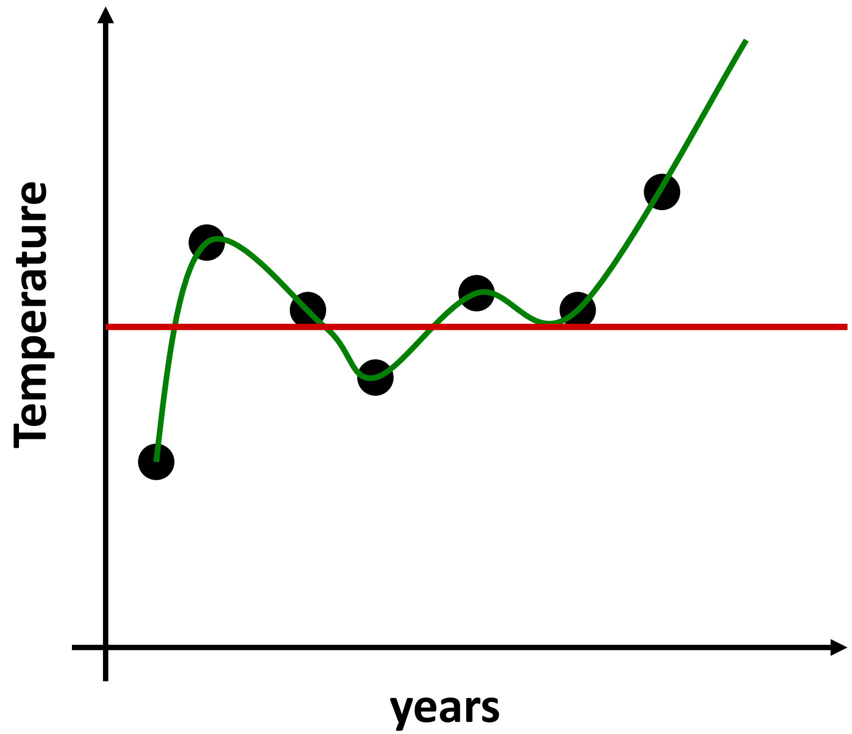
Stabilization matrix method (Ridge regression)

Morten Nielsen
Department of Health Technology,
DTU

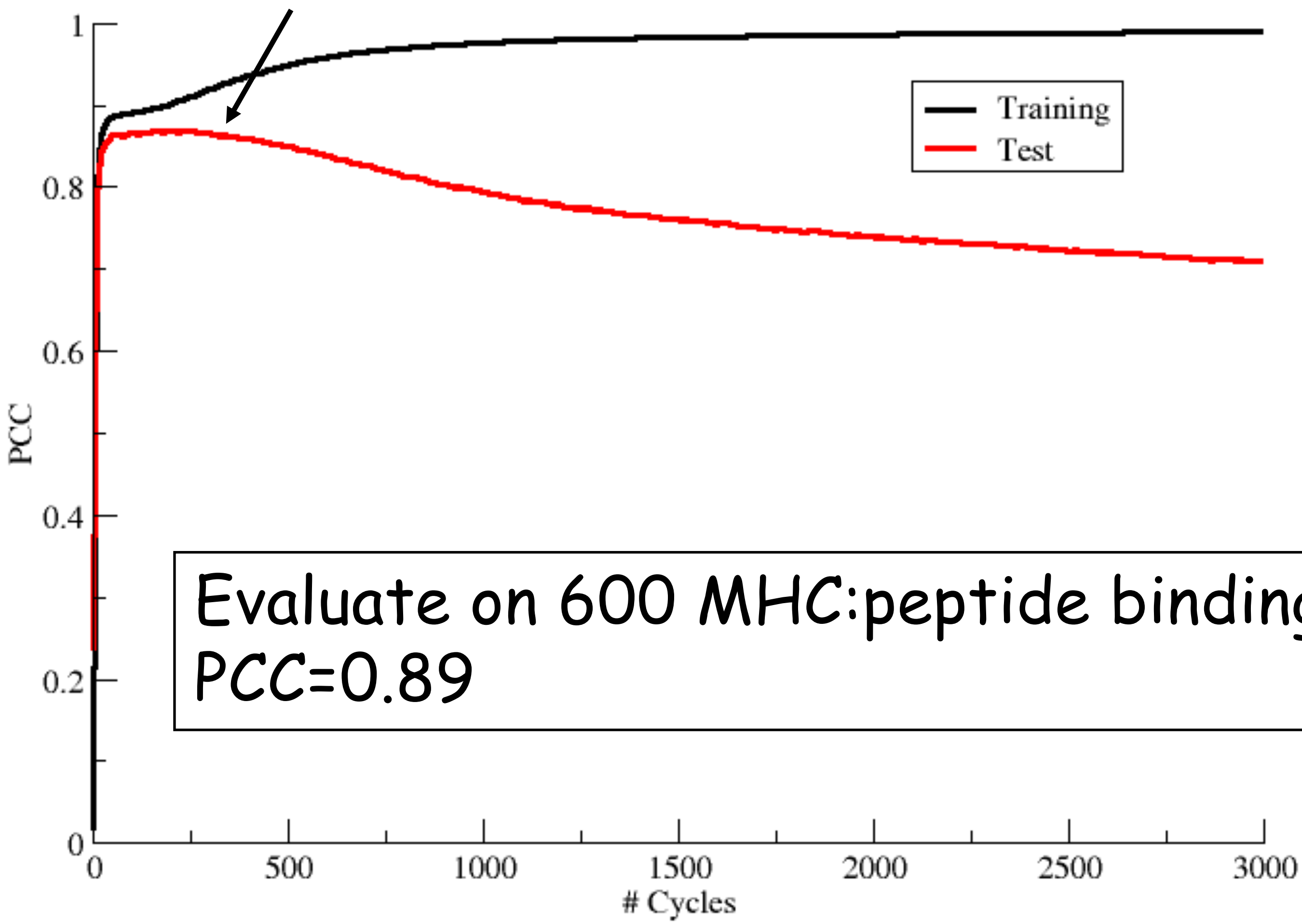
Data driven method training

- A prediction method contains a very large set of parameters
  - A matrix for predicting binding for 9meric peptides has 9x20=180 weights
- Over fitting is a problem



Model over-fitting (early stopping)

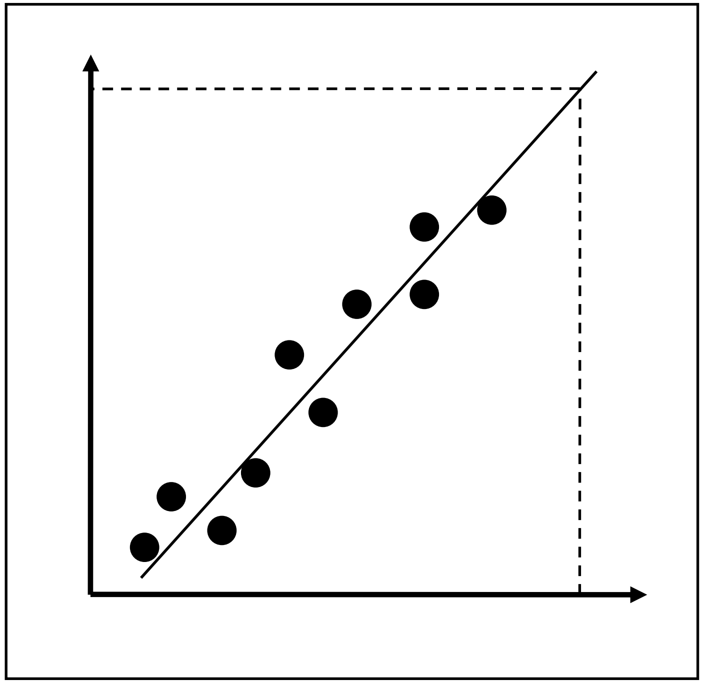
Stop training



Evaluate on 600 MHC:peptide binding data
PCC=0.89

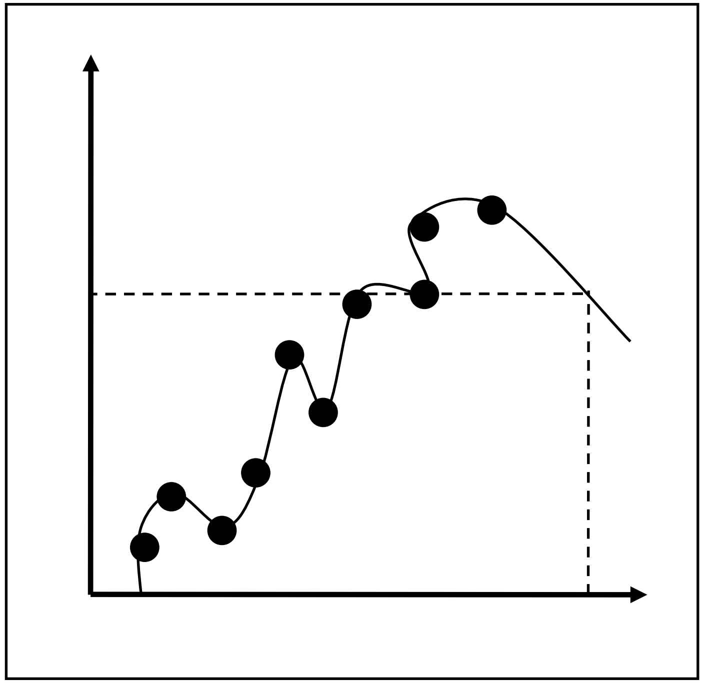
Stabilization matrix method
The mathematics

$$E = \frac{1}{2} \cdot \sum_i (O_i - t_i)^2$$



$y = ax + b$

2 parameter model
Good description, poor fit

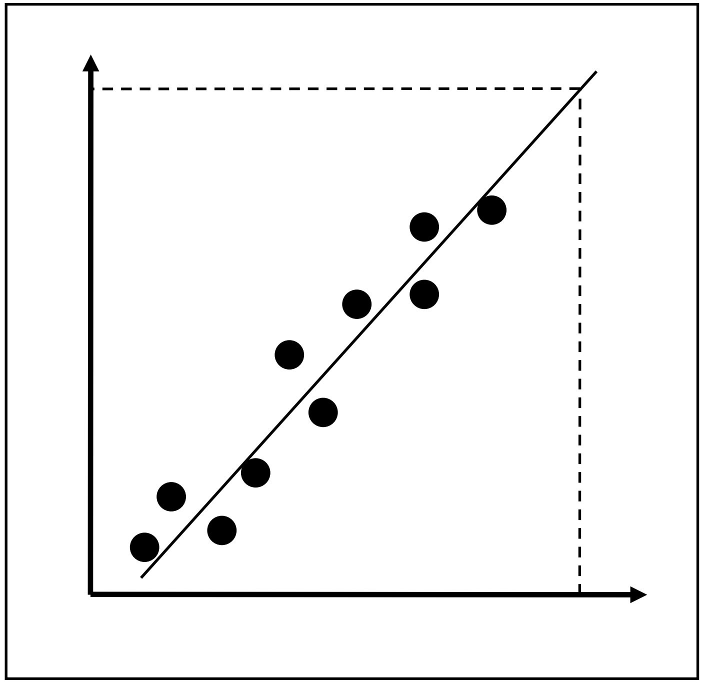


$y = ax^6+bx^5+cx^4+dx^3+ex^2+fx+g$

7 parameter model
Poor description, good fit

Stabilization matrix method
The mathematics

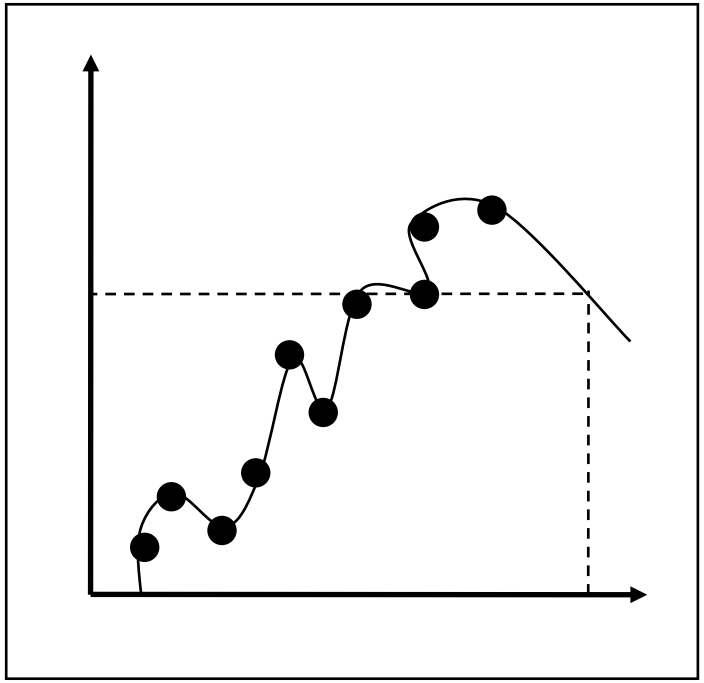
$$E = \frac{1}{2} \cdot \sum_i (O_i - t_i)^2 + \lambda \cdot \sum_l w_l^2$$



$y = ax + b$

2 parameter model

Good description, poor fit

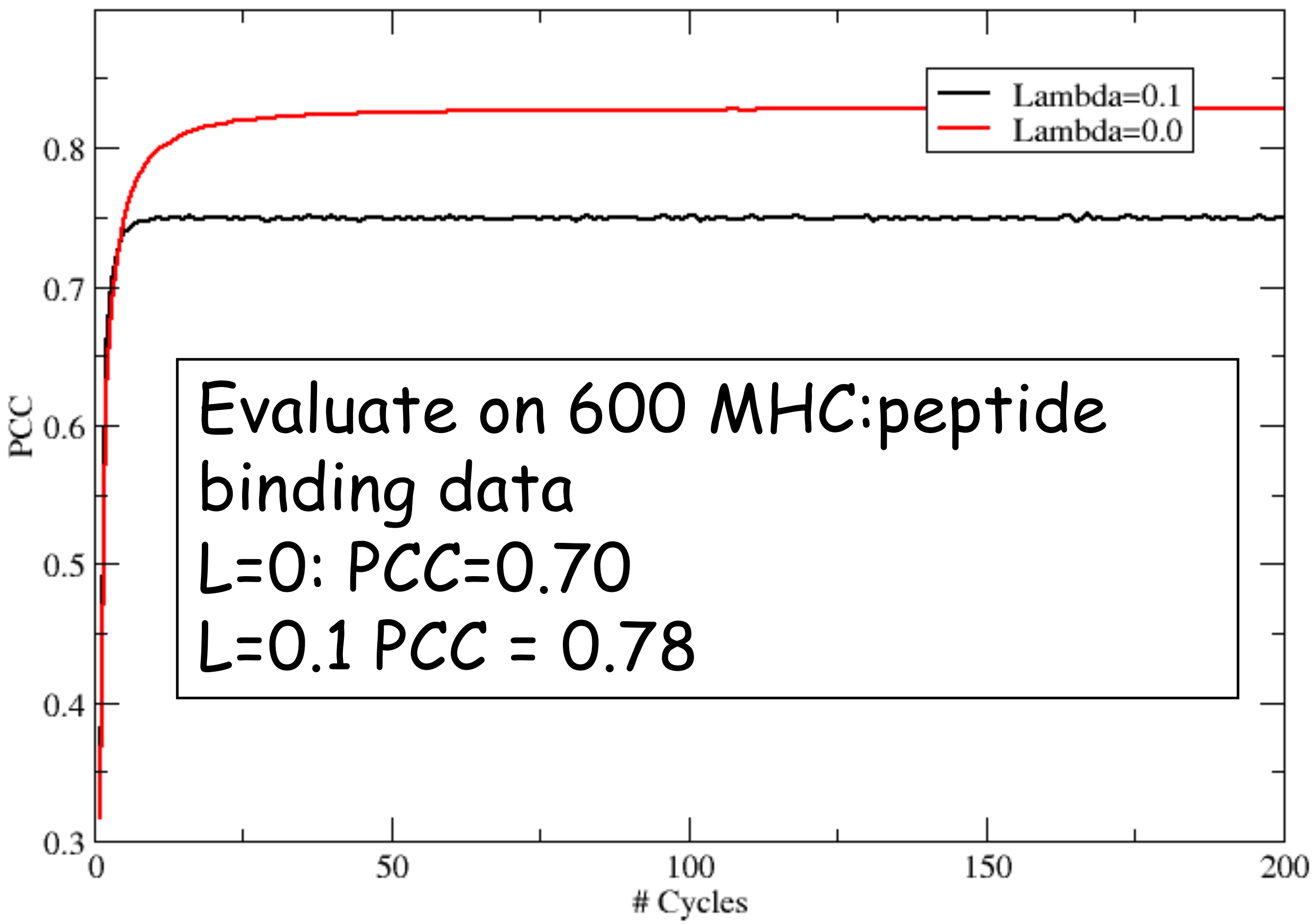


$y = ax^6+bx^5+cx^4+dx^3+ex^2+fx+g$

7 parameter model

Poor description, good fit

SMM training



Stabilization matrix method. The analytic solution

$$E = \frac{1}{2} \cdot \sum_i (O_i - t_i)^2 + \lambda \cdot \sum_l w_l^2$$

$$H \cdot w = p$$

$$\| H \cdot w - t \| + w^t \lambda w \rightarrow \min$$

$$w = (H^t H + \lambda)^{-1} H^t t$$

Each peptide is represented as 9*20 number (180)
H is a stack of such vectors of 180 values
t is the target value (the measured binding)
λ is a parameter introduced to suppress the effect of noise in the experimental data and lower the effect of overfitting

SMM - Stabilization matrix method - the numerical solution

$$E = \frac{1}{2} \cdot \sum_i (O_i - t_i)^2 + \lambda \cdot \sum_l w_l^2$$

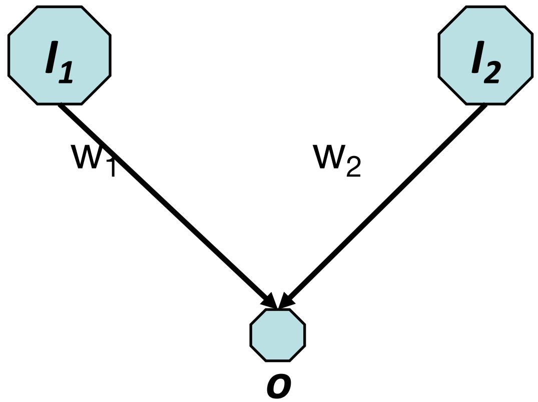
$$O = \sum_i I_i \cdot w_i$$

Sum over data points

Sum over weights

Linear function

$$O = I_1 \cdot w_1 + I_2 \cdot w_2$$



SMM - Stabilization matrix method

CENTERFORBIOLOGICALSEQUENCEANALYSIS CBS

Global error:

$$E = \tfrac{1}{2} \cdot \sum_i (O_i - t_i)^2 + \lambda \cdot \sum_l w_l^2$$

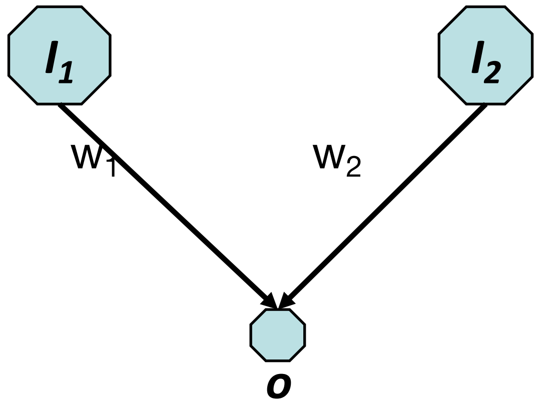
$$O = \sum_i I_i \cdot w_i$$

Sum over data points

Sum over weights

Linear function

$$O = I_1 \cdot w_1 + I_2 \cdot w_2$$



Per target error:

$$E = \sum_i E_i$$

$$E_i = \tfrac{1}{2} \cdot (O_i - t_i)^2 + \frac{\lambda}{N} \sum_l w_l^2$$

SMM - Stabilization matrix method
Do it yourself

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

$$E_{\text{per target}} = \tfrac{1}{2}\cdot(O-t)^2 + \boxed{\frac{\lambda}{N}}\sum_l w_l^2 = E_1 + E_2$$

λ per target

$$O = \sum_i I_i \cdot w_i$$

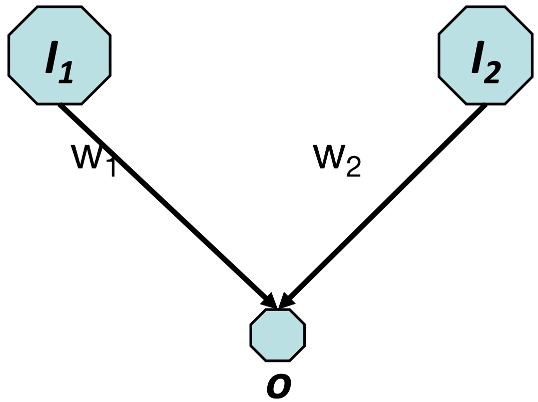
$$\frac{\partial E}{\partial w_i} = \frac{\partial E_1}{\partial w_i} + \frac{\partial E_2}{\partial w_i} = \frac{\partial E_1}{\partial O}\cdot\frac{\partial O}{\partial w_i} + \frac{\partial E_2}{\partial w_i}$$

$$\frac{\partial E_1}{\partial w_i} = ??$$

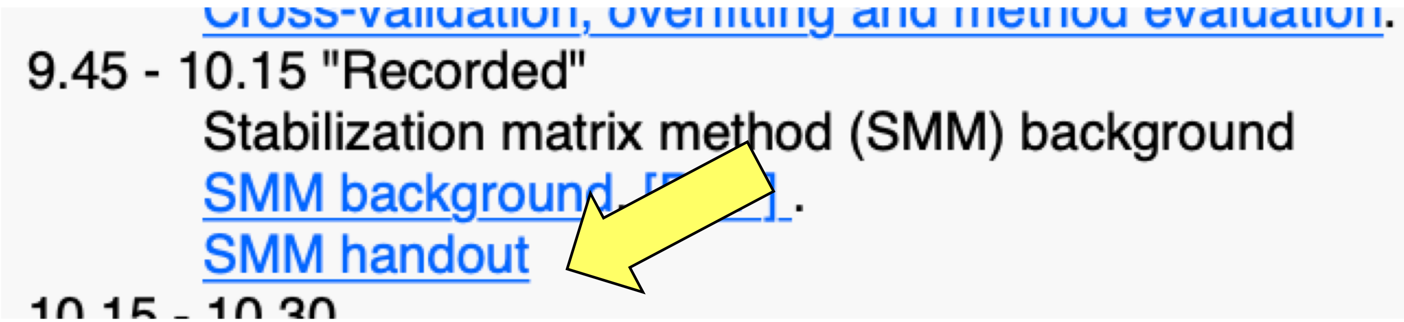
$$\frac{\partial E_2}{\partial w_i} = ??$$

Linear function

$$O = I_1 \cdot w_1 + I_2 \cdot w_2$$



And now you


Cross-validation, overfitting and method evaluation.
9.45 - 10.15 "Recorded"
Stabilization matrix method (SMM) background
SMM background [PDF] .
SMM handout
10.15 - 10.30

SMM - Stabilization matrix method

$$E_{\text{per target}} = \tfrac{1}{2} \cdot (O - t)^2 + \frac{\lambda}{N} \sum_l w_l^2$$

λ ***per target***

$$O = \sum_i I_i \cdot w_i$$

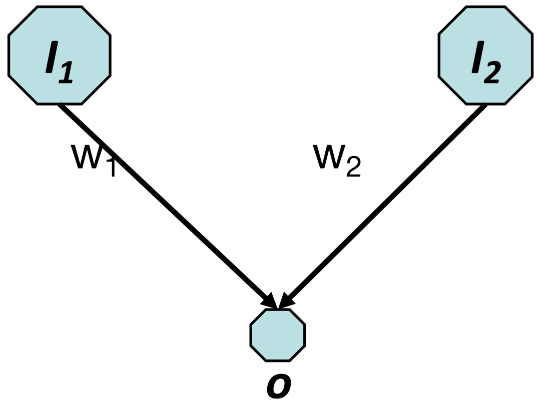
$$\frac{\partial E}{\partial w_i} = \frac{\partial E}{\partial O} \cdot \frac{\partial O}{\partial w_i}$$

$$\frac{\partial E}{\partial w_i} = (O - t) \cdot \frac{\partial O}{\partial w_i} + \frac{1}{\partial w_i} \left(\frac{\lambda}{N} \sum_l w_l^2\right)$$

$$= (O - t) \cdot I_i + \frac{2 \cdot \lambda}{N} \cdot w_i$$

Linear function

$$O = I_1 \cdot w_1 + I_2 \cdot w_2$$



SMM - Stabilization matrix method

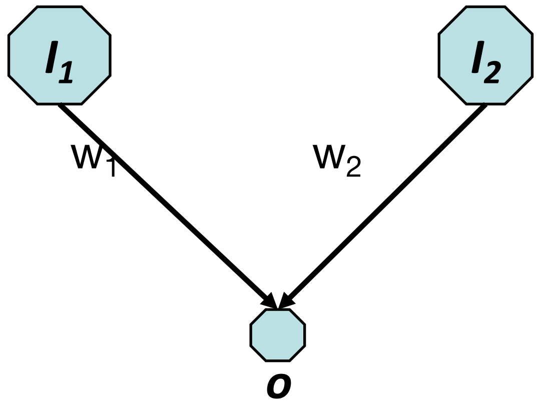
$$\frac{\partial E}{\partial w_i} = (O - t) \cdot I_i + \frac{2 \cdot \lambda}{N} \cdot w_i$$

$$w_i^{'} = w_i + \Delta w_i$$

$$\Delta w_i = -\varepsilon \cdot \frac{\partial E}{\partial w_i}$$

Linear function

$$O = I_1 \cdot w_1 + I_2 \cdot w_2$$



SMM - Stabilization matrix method
Monte Carlo

CENTERFORBIOLOGICALSEQUENCEANALYSIS CBS

Global:

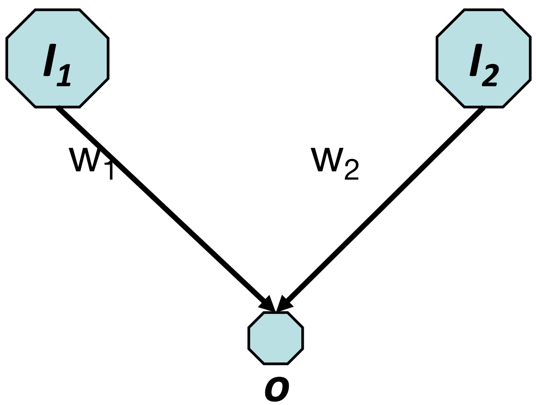
$$E = \frac{1}{2} \cdot \sum_i (O_i - t_i)^2 + \lambda \cdot \sum_l w_l^2$$

- Make random change to weights
- Calculate change in "global" error
- Update weights if MC move is accepted

Note difference between MC and GD in the use of "global" versus "per target" error

Linear function

$$O = I_1 \cdot w_1 + I_2 \cdot w_2$$



Training/evaluation procedure

- Define method
- Select data
- Deal with data redundancy
  - In method (sequence weighting)
  - In data (Hobohm)
- Deal with over-fitting either
  - in method (SMM regulation term) or
  - in training (stop fitting on test set performance)
- Evaluate method using cross-validation

A small doit tcsh script


```
#! /bin/tcsh -f
set DATADIR = /home/projects/mniel/ALGO/data/SMM/

foreach a ( A0101 A3002 )
mkdir -p $a
cd $a

# Here you can type the lambdas to test
foreach l ( 0 0.02 )

mkdir -p l.$l

cd l.$l

# Loop over the 5 cross validation configurations
foreach n ( 0 1 2 3 4 )

# Do training
smm -l $l ../f00$n > mat.$n

# Do evaluation
pep2score -mat mat.$n ../c00$n > c00$n.pred

end

# Do concatinated evaluation
echo $a $l `cat c00?.pred | grep -v "#" | gawk '{print $2,$3}' | xycorr` \
        `cat c00?.pred | grep -v "#" | gawk '{print $2,$3}' | gawk 'BEGIN{n+0; e=0.0}{n++; e += ($1-$2)*($1-$2)}END{print e/n}' `

cd ..

end

cd ..

end
```