The theory of on-line learning

Michael Biehl
University of Groningen
Institute for Mathematics and Computing Science
P.O. Box 407, NL–9700 AK Groningen

m.biehl@rug.nl
http://www.cs.rug.nl/~biehl

1 On–line Perceptron training

A guided tour through various learning scenarios

2 Learning by on–line gradient descent

Learning in continuous, multi–layered neural networks

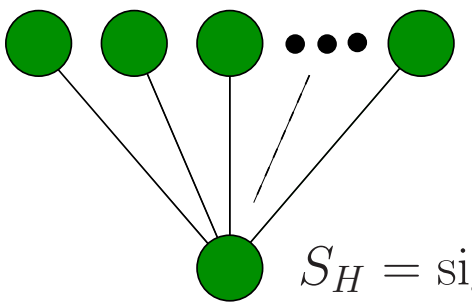
3 On–line learning from clustered data

Vector Quantization and
Learning Vector Quantization

1 On–line perceptron training

The perceptron

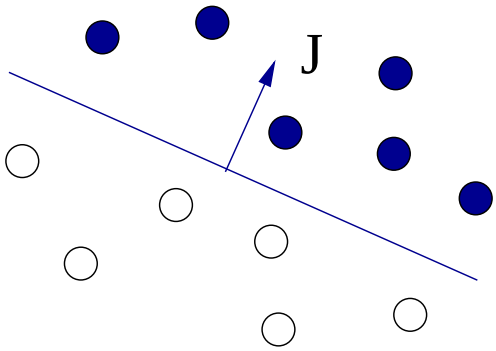
single layer feed-forward neural network:



$\boldsymbol{\xi}$ N-dim. inputs

$\mathbf{J}$ adaptive weights

$S_H = \mathrm{sign}(\mathbf{J} \cdot \boldsymbol{\xi})$ binary output (threshold unit)



implements a **linearly separable classification** of inputs

---

milestones:

Perceptron Convergence Theorem, Rosenblatt (1958)

Capacity, Winder (1963), Cover (1965), Schläfli (**1852**)

Perceptrons, Minsky and Papert (1969)

Statistical physics of perceptron weights, Gardner (1988)

Learning a linearly separable rule from reliable examples

- unknown **rule** $S_R(\boldsymbol{\xi}) = \mathrm{sign}(\mathbf{B}\cdot\boldsymbol{\xi}) = \pm 1$
  defines the correct classification, parametrized through a **teacher** perceptron with weights
  $\mathbf{B} \in I\!R^N, \qquad (\mathbf{B}^2 = 1)$

- only available information: **example data**

  $$I\!D = \{\, \boldsymbol{\xi}^\mu,\, S_R^\mu = S_R(\boldsymbol{\xi}^\mu)\,\}, \qquad \mu = 1, 2, \ldots, P$$

  (correct labels S_R^μ provided by the teacher)

- **training:** choice of student weights $\mathbf{J}$
  - $\mathbf{J}$ parametrizes a **hypothesis** $S_H(\boldsymbol{\xi}) = \mathrm{sign}(\mathbf{J}\cdot\boldsymbol{\xi})$
  - **supervised learning** based on the student performance with respect to the training data $I\!D$
  - here: **binary error measure**

$$\boxed{\varepsilon^\mu(\mathbf{J}) = \varepsilon\,[\,S_H(\boldsymbol{\xi}^\mu), S_R^\mu\,] = \begin{cases} 0 & \text{if } S_H(\boldsymbol{\xi}^\mu) = S_R^\mu \\ 1 & \text{if } S_H(\boldsymbol{\xi}^\mu) \neq S_R^\mu \end{cases}}$$

compares student output and rule classification

- **Off–line training:**

(Seung et al., 1992)
(Watkin et al., 1993)
(Opper and Kinzel, 1996)

  – guided by the minimization of a cost function $H(\mathbf{J})$, e.g. the **training error**

$$H(\mathbf{J}) \propto \varepsilon_t(\mathbf{J}) = \frac{1}{P} \sum_{\mu=1}^{P} \varepsilon^\mu$$

  (empirical mean of the error measure w.r.t. $\mathbb{D}$)

  – Equilibrium Statistical Mechanics treatment:
    – energy H of N degrees of freedom
    – ensemble of systems in thermal equilibrium at formal temperature $1/\beta$
    – disorder average over random training set (replicas),
      assume distribution $P(\boldsymbol{\xi})$ of inputs
    – macroscopic description, order parameters
    – typical properties of large systems, $P = \alpha N$

  – training strategies
    learnable rule: error free training $\varepsilon_t = 0$
    $\mathbf{J}$ inside *version space*
    – typical *Gibbs* student ($\beta \to \infty$)
    – maximum stability
    – Bayes optimal student, c.m. of version space, ...

On–line training

e.g. (Saad, 1998)

- **single presentation** of uncorrelated (*new*) $\{ \boldsymbol{\xi}^\mu, S_R^\mu \}$
- update of student weights:

$$\mathbf{J}(\mu) = \mathbf{J}(\mu\text{–}1) + \Delta\mathbf{J}\,[\,\mathbf{J}(\mu\text{–}1), \boldsymbol{\xi}^\mu, S_R^\mu)\,]$$

  disregards student performance on other examples
- learning dynamics in discrete time
- Statistical Physics approach:
  - consider sequence of independent, random $\boldsymbol{\xi}^\mu$
  - thermodynamic limit $N \to \infty$, $P = \alpha N$
  - recursion relation for order parameters
  - disorder average over latest example
    self–averaging properties
  - continuous time limit $\alpha = \mu/N$
    ordinary differential equations for evolution of order parameters

    ...
- $\Delta\mathbf{J}$ defines the actual **learning algorithm**
  - Hebbian learning
  - error correction, perceptron algorithm
  - *optimal* training prescription ?

    ...

Generalization

performance of the student (after training) with respect to arbitrary, *new* inputs $\boldsymbol{\xi} \notin \mathbb{D}$

- in practice: empirical mean of error measure over a set of test inputs $\{ \boldsymbol{\xi}^\tau \notin \mathbb{D} \}_{\tau=1,2,\ldots T}$

$$\varepsilon_{test} = \frac{1}{T} \sum_{\tau=1}^{T} \varepsilon \left[S_H(\boldsymbol{\xi}^\tau), S_R(\boldsymbol{\xi}^\tau) \right]$$

- in the theoretical analysis: average $\langle \ldots \rangle$ over the (assumed) probability density $P(\boldsymbol{\xi})$ of inputs,

generalization error $$\boxed{\varepsilon_g = \langle \varepsilon \left[S_H(\boldsymbol{\xi}), S_R(\boldsymbol{\xi}) \right] \rangle}$$

(here: $\varepsilon \in \{0, 1\}, \rightarrow \varepsilon_g$ is the error *probability*)

---

the simplest model distribution:

isotropic density $P(\boldsymbol{\xi})$, $\boldsymbol{\xi}$ uncorrelated with **B** and **J**

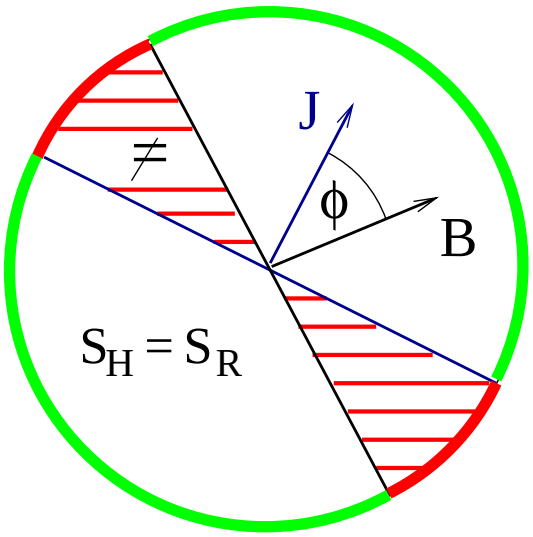
consider vectors of independent, identically distributed (*i.i.d.*) components ξ_j with

$$\boxed{\langle \xi_j \rangle = 0, \qquad \langle \xi_j \, \xi_k \rangle = \delta_{jk}}$$

e.g. independent Gaussian variables ξ_j , zero mean / unit variance

Geometric argument

projection of data into $(\mathbf{J}, \mathbf{B})$ –plane yields
isotropic (radially symmetric) density of inputs



probability for disagreement:

$$\varepsilon_g = \phi/\pi$$

$$= \frac{1}{\pi} \arccos\left(\frac{\mathbf{J}\cdot\mathbf{B}}{|\,\mathbf{J}\,|\,|\,\mathbf{B}\,|} \right)$$

for $|\,\mathbf{B}\,| = 1$:

$$\boxed{\varepsilon_g = \frac{1}{\pi} \arccos\left(\frac{R}{\sqrt{Q}} \right)}$$

overlap parameters

$$\boxed{R = \mathbf{J}\cdot\mathbf{B}, \quad Q = \mathbf{J}\cdot\mathbf{J}}$$

sufficient to quantify the success of training

$R = 0$, $\varepsilon_g = 1/2$ random guessing

$R = \sqrt{Q}$, $\varepsilon_g = 0$ perfect generalization

(actual value of student norm $Q \geq R^2$ is irrelevant)

Derivation for large N

given: $\mathbf{B}$, $\mathbf{J}$, uncorrelated random input $\boldsymbol{\xi}$ with

$$\langle \xi_j \rangle = 0, \ \langle \xi_j \, \xi_k \rangle = \delta_{jk}$$

consider student/teacher *fields*:

$$\boxed{x = \mathbf{J} \cdot \boldsymbol{\xi} = \sum_j J_j \, \xi_j, \quad y = \mathbf{B} \cdot \boldsymbol{\xi} = \sum_j B_j \, \xi_j}$$

sums of (many) independent random quantities

Central Limit Theorem (CLT):

joint density of (x, y) is, for $N \to \infty$, a two–dimensional Gaussian, fully specified by the first and second moments

$$\langle x \rangle = \sum_j J_j \langle \xi_j \rangle = 0, \qquad \langle y \rangle = \sum_j B_j \langle \xi_j \rangle = 0$$

$$\langle x^2 \rangle = \sum_{j,k} J_j J_k \langle \xi_j \xi_k \rangle = \sum_j J_j^2 = Q$$

$$\langle y^2 \rangle = \sum_{j,k} B_j B_k \langle \xi_j \xi_k \rangle = \sum_j B_j^2 = 1$$

$$\langle xy \rangle = \sum_{j,k} J_j B_k \langle \xi_j \xi_k \rangle = \sum_j J_j B_j = R$$

covariance matrix $\mathcal{C} = \begin{pmatrix} Q & R \\ R & 1 \end{pmatrix}$

$$P_{Q,R}(x, y) = \frac{1}{2\pi \sqrt{\mathrm{Det}\, \mathcal{C}}} \exp\left[-\frac{1}{2} (x, y) \, \mathcal{C}^{-1} \begin{pmatrix} x \\ y \end{pmatrix} \right]$$

$$P(x,y) \;=\; \frac{1}{2\pi\sqrt{Q-R^2}}\;\exp\left[-\frac{1}{2}\,\frac{x^2\;+Q\,y^2-2\,R\,x\,y}{Q-R^2}\right]$$

Note:

details of the input distribution are largely irrelevant

e.g. independent, binary $\xi_j \;=\; \pm 1$ with equal prob.

or independent, uniform $\xi_j \in [-a,a]$ with $\langle\, \xi_j^2 \,\rangle = 1$

---

generalization error:

$$\begin{aligned}
\varepsilon_g \;&=\; \left(\int_0^\infty dx \int_{-\infty}^0 dy \;+\; \int_{-\infty}^0 dx \int_0^\infty dy\right) P(x,y) \\
&=\; \ldots \qquad (exercise) \\
&=\; \frac{1}{\pi}\,\arccos\left(\frac{R}{\sqrt{Q}}\right)
\end{aligned}$$

In the following, the isotropic distribution is also assumed to describe the statistics of example data inputs, i.e. vectors $\boldsymbol{\xi}^\mu \in I\!\!D$ consist of i.i.d. components with

$$\langle\, \xi_j^\mu \,\rangle = \;0, \qquad \langle\, \xi_j^\mu\, \xi_k^\nu \,\rangle = \;\delta_{jk}\,\delta_{\nu\mu}$$

properties:

– no *spatial* correlations, $\langle \xi_j^\mu \, \xi_k^\nu \rangle = 0$ if $j \neq k$

– no distinguished directions in input space

– no *temporal* correlations, $\langle \xi_j^\mu \, \xi_j^\nu \rangle = 0$ if $\mu \neq \nu$

– no correlations with the rule, $P(\boldsymbol{\xi}^\mu, \mathbf{B}) \; = \; P(\boldsymbol{\xi}^\mu) \, P(\mathbf{B})$

– single presentation *without repetition*:
no correlation of $\boldsymbol{\xi}^\mu$ with $\mathbf{J}(\mu - 1)$, i.e. the student
after training with examples $\nu = 1, 2, \ldots, \mu$–1

consequences:

– average over data can be performed *step by step*

– actual choice of $\mathbf{B}$ is irrelevant,
it is not necessary to average over the teacher

Hebbian Learning revisited (Hebb,1949)

off–line interpretation (Vallet, 1989)

choice of student weights given $\mathbb{D} = \{ \boldsymbol{\xi}^\mu, S_R^\mu \}_{\mu=1}^P$

$$\mathbf{J}(P) = \frac{1}{N} \sum_{\mu=1}^{P} \boldsymbol{\xi}^\mu S_R^\mu$$

---

equivalent **on–line interpretation**

dynamics upon single presentation of examples:

$$\mathbf{J}(\mu) = \mathbf{J}(\mu\text{–}1) + \frac{1}{N} \boldsymbol{\xi}^\mu S_R^\mu$$

• from microscopics to macroscopics: **recursions** for

overlaps $\boxed{R(\mu) = \mathbf{J}(\mu) \cdot \mathbf{B}, \quad Q(\mu) = \mathbf{J}(\mu) \cdot \mathbf{J}(\mu)}$

$$R(\mu) = R(\mu\text{–}1) + \frac{1}{N} \underbrace{(\mathbf{B} \cdot \boldsymbol{\xi}^\mu)}_{y^\mu} \underbrace{S_R^\mu}_{\mathrm{sign}(y^\mu)}$$

$$Q(\mu) = Q(\mu\text{–}1) + \frac{2}{N} \underbrace{(\mathbf{J}(\mu\text{–}1) \cdot \boldsymbol{\xi}^\mu)}_{x^\mu} \underbrace{S_R^\mu}_{\mathrm{sign}(y^\mu)} + \frac{1}{N^2} \underbrace{(\boldsymbol{\xi}^\mu)^2}_{N}$$

•) average over latest example $\langle \ldots \rangle_\mu$

random input $\boldsymbol{\xi}^\mu$ enters only through the *fields*

$$\boxed{x^\mu \;=\; \mathbf{J}(\mu\text{–}1)\cdot\boldsymbol{\xi}^\mu, \qquad y^\mu \;=\; \mathbf{B}\cdot\boldsymbol{\xi}^\mu}$$

$\boldsymbol{\xi}^\mu$ and $\mathbf{J}(\mu\text{–}1)$, $\mathbf{B}$ are statistically independent! By the same argument as on page 8, the CLT applies and one obtains the joint density

$$P(x^\mu, y^\mu) \;=\; \frac{(2\pi)^{-1}}{\sqrt{Q-R^2}}\;\exp\left[-\frac{1}{2}\,\frac{x^{\mu 2}+Q\,y^{\mu 2}-2\,R\,x^\mu\,y^\mu}{Q-R^2}\right]$$

with $Q = Q(\mu\text{–}1)$ and $R = R(\mu\text{–}1)$ on the r.h.s.

---

here we need: (exercise)

$$\langle\,|\,y^\mu\,|\,\rangle_\mu \;=\; \sqrt{2/\pi}$$

$$\langle\,x^\mu\;\mathrm{sign}(y^\mu)\,\rangle_\mu \;=\; \sqrt{2/\pi}\;R(\mu\text{–}1)$$

$$\langle\,R(\mu)\,\rangle_\mu \;=\; R(\mu\text{–}1) \;+\; \frac{1}{N}\sqrt{\frac{2}{\pi}}$$

$$\langle\,Q(\mu)\,\rangle_\mu \;=\; Q(\mu\text{–}1) \;+\; \frac{1}{N}\left(2\sqrt{\frac{2}{\pi}}\,R(\mu\text{–}1) \;+\; 1\right)$$

- **average over all previous examples:** $[\,\ldots\,]_{\mu-1}$

$$\left\langle\, [\,R(\mu)\,]_{\mu-1} \,\right\rangle_\mu \;=\; [\,R(\mu)\,]_\mu \;=\; [\,R(\mu-1)\,]_{\mu-1} \;+\; \frac{1}{N}\sqrt{\frac{2}{\pi}}$$

define $R^\nu \;=\; [\,R(\nu)\,]\nu$ mean overlap after ν examples

$$R^\mu \;=\; R^{\mu-1} + \frac{1}{N}\sqrt{\frac{2}{\pi}} \qquad Q^\mu \;=\; Q^{\mu-1} + \frac{1}{N}\left(2\sqrt{\frac{2}{\pi}}R^{\mu-1} + 1\right)$$

(exercise: explicit summation)

- **continuous time limit,** $N \to \infty,\ \alpha = \mu/N,\ d\alpha = 1/N$

$$\frac{dR}{d\alpha} = \sqrt{\frac{2}{\pi}} \qquad \frac{dQ}{d\alpha} = 2\sqrt{\frac{2}{\pi}}R(\alpha) + 1$$

initial conditions (*tabular rasa*) $R(0) \;=\; Q(0) \;=\; 0$

mean values after training with $(\alpha\, N)$ examples

$$\boxed{R(\alpha) \;=\; \sqrt{\frac{2}{\pi}}\,\alpha \qquad\qquad Q(\alpha) \;=\; \frac{2}{\pi}\alpha^2 + \alpha}$$

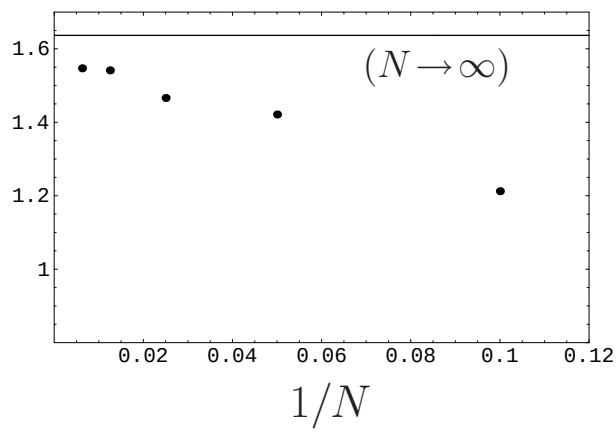
typical behavior? magnitude of fluctuations?

• **self–averaging properties** of order parameters

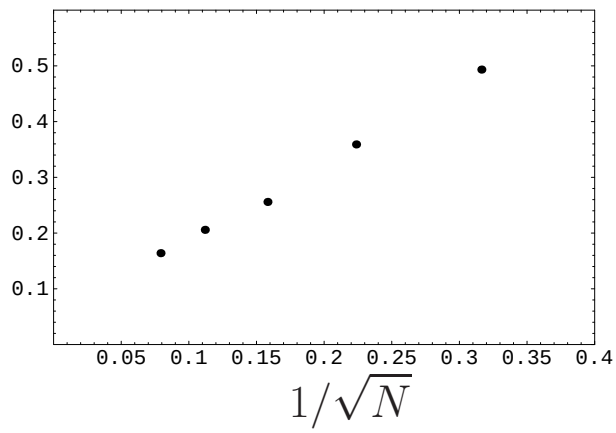
simulations: finite N, $\mathbf{J}(0) = 0$, $\alpha = 1.0$

50 independent sequences of random examples:

mean value of $Q(\alpha = 1)$



standard deviation



(for our purposes:)

A quantity $A(\mathbf{J})$ is called self–averaging if

– the width of its probability distribution vanishes

– the observation of a value A different from its mean occurs with vanishing probability

in the limit $\boldsymbol{N \to \infty}$

For a rigorous discussion see (Reents,Urbanczik, 1998)

Exercise

calculate $\left[R^2(\mu)\right]_\mu$ and show that $\left[R^2(\mu)\right]_\mu - \left[R(\mu)\right]_\mu^2 \propto 1/N$ for Hebbian learning. Hint: use the off–line interpretation and exploit the independence of examples.

• **deterministic differential equations**

describe the actual values of order parameters

explicit average over history is not necessary

• **learning curve**

α –dependence of order parameters (R, Q)

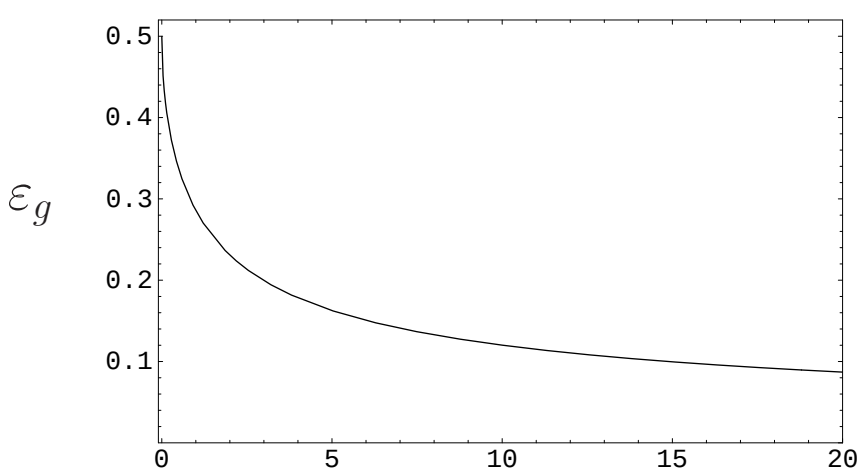
of (well–behaved) functions thereof, e.g.

normalized overlap $$\rho = \frac{\mathbf{J} \cdot \mathbf{B}}{|\,\mathbf{J}\,|\,|\,\mathbf{B}\,|} = \frac{R}{\sqrt{Q}}$$

$$\left[\frac{R(\mu)}{\sqrt{Q(\mu)}}\right]_\mu = \frac{[\,R(\mu)\,]_\mu}{\sqrt{[\,Q(\mu)\,]_\mu}} = \frac{R(\alpha)}{\sqrt{Q(\alpha)}} = \left(1 + \frac{\pi}{2\alpha}\right)^{-1/2}$$

generalization error

$$\varepsilon_g(\alpha) = \frac{1}{\pi} \arccos\left[\left(1 + \frac{\pi}{2\alpha}\right)^{-1/2}\right]$$



$\alpha = P/N$

small α –behavior ($\approx$ random guessing):

$$\rho(\alpha) = \frac{R(\alpha)}{\sqrt{Q(\alpha)}} \approx \sqrt{\frac{2}{\pi}}\sqrt{\alpha} \qquad \varepsilon_g \approx \frac{1}{2} - \frac{\sqrt{2}}{\pi^{3/2}}\sqrt{\alpha}$$

asymptotic behavior $(\boldsymbol{\alpha \to \infty})$:

$$\rho(\alpha) \approx 1 - \frac{\pi}{4\alpha} \qquad \varepsilon_g \approx \frac{1}{\sqrt{2\pi}}\,\alpha^{-1/2} \to 0$$

perfect generalization for $P = \alpha N,\ \alpha \to \infty$

Note:

$R, Q \to \infty$ with $\alpha \to \infty$, but $\rho = R/\sqrt{Q} \to 1$

only the **direction** of $\mathbf{B}$ is found!

Modified Hebbian Learning

$$\Delta\, \mathbf{J}\,(\,\mathbf{J}(\mu\text{–}1), \boldsymbol{\xi}^\mu, \mathrm{sign}[\mathbf{B}\cdot\boldsymbol{\xi}^\mu]\,) \;=\; \frac{1}{N}\, f\,(\,\ldots\,)\,\boldsymbol{\xi}^\mu\;\mathrm{sign}[\mathbf{B}\cdot\boldsymbol{\xi}^\mu]$$

modulation function f defines training algorithm
simplification: same $f_j \;=\; f$ in all components J_j

restriction: f may depend on **available quantities**

in principle: $f\,(\,\mathbf{J}(\mu\text{–}1), \boldsymbol{\xi}^\mu, \mathrm{sign}\, y^\mu\,)$

simplified : $f\,(\,Q(\mu\text{–}1), x^\mu, \mathrm{sign}\, y^\mu\,)$

$$\mathbf{J}(\mu) \;=\; \mathbf{J}(\mu\text{–}1) \;+\; \frac{1}{N}\,\boldsymbol{\xi}^\mu\;\mathrm{sign}[y^\mu]$$

$$\langle\, R(\mu)\,\rangle \;=\; R(\mu\text{–}1) \;+\; \frac{1}{N}\,\langle\, f\;|\,y^\mu\,|\,\rangle_\mu$$

$$\langle\, Q(\mu)\,\rangle \;=\; Q(\mu\text{–}1) \;+\; \frac{1}{N}\,\langle\, 2\,f\,x^\mu\;\mathrm{sign}(y^\mu) \;+\; f^2\,\rangle_\mu$$

single example averages on r.h.s. over $P(x^\mu, y^\mu)$
yield functions of $R(\mu\text{–}1), Q(\mu\text{–}1)$

assume and exploit **self–averaging** properties:

$$R^\mu = R^{\mu-1} + \frac{1}{N} F_R(R^{\mu-1}, Q^{\mu-1})$$
$$Q^\mu = Q^{\mu-1} + \frac{1}{N} F_Q(R^{\mu-1}, Q^{\mu-1})$$

continuous time limit:

$$\frac{dR}{d\alpha} = F_R(R,Q) = \langle f\, y\, S_R \rangle$$
$$\frac{dQ}{d\alpha} = F_Q(R,Q) = \langle 2\, f\, x\, S_R + f^2 \rangle$$

alternatively:

$$\frac{d\rho}{d\alpha} = F_\rho(\rho,Q) = \left\langle f\, S_R \left(\frac{y}{\sqrt{Q}} - \frac{\rho x}{Q} \right) - \frac{\rho}{2Q} f^2 \right\rangle$$

"Plug in and play"

- choose f to define the actual algorithm
- calculate r.h.s. of differential equations (perform Gaussian integrals over $P(x,y)$
- integrate (numerically) to obtain $R(\alpha), Q(\alpha), \varepsilon_g(\alpha)$
- investigate $(\alpha \to \infty)$ –asymptotics

Learning from mistakes

(Rosenblatt perceptron algorithm, 1959)

base training on the performance w.r.t. the new example:

$$\mathbf{J}(\mu) \;=\; \begin{cases} \mathbf{J}(\mu\text{–}1) & \text{if } \operatorname{sign}(\mathbf{J}(\mu\text{–}1)\cdot\boldsymbol{\xi}^\mu) \;=\; S_R^\mu \\ \mathbf{J}(\mu\text{–}1) + \frac{1}{N}\,\boldsymbol{\xi}^\mu S_R^\mu & \text{else} \end{cases}$$

modulation function

$$f(x^\mu, \operatorname{sign} y^\mu) = \Theta\,(\,-x^\mu\, S_R^\mu\,) = \Theta\,(\,-x^\mu\, y^\mu)\,)$$

corresponding on–line dynamics: (Biehl and Riegler, 1994)

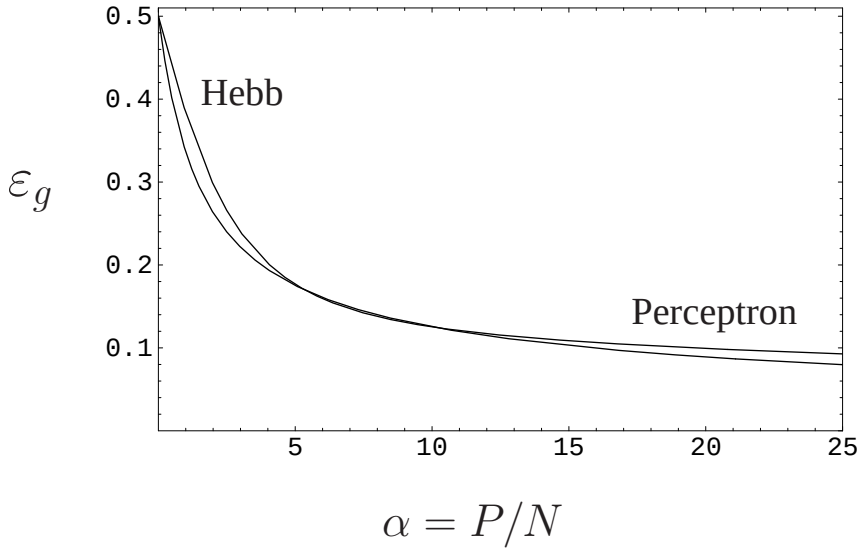
$$\frac{d\rho}{d\alpha} \;=\; \left\langle\, \Theta(-xy)\, S_R \left(\frac{y}{\sqrt{Q}} - \frac{\rho x}{Q} \right) - \frac{\rho}{2Q}\Theta(-xy)\,\right\rangle \;=\; \ldots$$

$$= \sqrt{\frac{1}{2\pi}} \left(\frac{1-\rho^2}{\sqrt{Q}} - \frac{\rho}{Q}\arccos(\rho) \right)$$

$$\frac{dQ}{d\alpha} \;=\; \langle\, 2\,\Theta(-xy)\, x\, S_R \;+\; \Theta(-xy)\,\rangle \;=\; \ldots$$

$$= \sqrt{\frac{2}{\pi}}\,(\rho-1)\,\sqrt{Q} \;+\; \frac{1}{\pi}\,\arccos(\rho)$$

Numerical integration $\rightarrow \rho(\alpha), Q(\alpha), \quad \varepsilon_g(\alpha)$



$\rho(0) = 0$

$Q(0) = 1$

Perceptron algorithm realizes perfect generalization:

$$\frac{d\rho}{d\alpha} = 0 \quad \text{for} \quad \rho = 1 \qquad \text{but} \quad Q, R \rightarrow \infty \ \text{ as } \ \alpha \rightarrow \infty$$

asymptotics:

$$\left.\begin{aligned} \varepsilon_g &= \frac{1}{\pi}\left(\frac{2}{3}\right)^{1/3} \alpha^{-1/3} \\ Q &= \frac{1}{\sqrt{2\pi}}\left(\frac{3}{2}\right)^{1/3} \alpha^{+1/3} \end{aligned}\right\} \qquad \text{for} \quad \alpha \rightarrow \infty$$

$\alpha^{-1/3} \gg \alpha^{-1/2}$!

why is simple minded Hebbian training *better* than sophisticated perceptron training ? (resolved later)

Necessary condition for perfect generalization

stationary, normalized overlap for $\rho = 1 \quad (x = \sqrt{Q}y)$

$$\left.\frac{d\rho}{d\alpha}\right|_{\rho=1} = \left\langle f\, S_R\left(\frac{y}{\sqrt{Q}} - \frac{\rho x}{Q}\right) - \frac{\rho}{2Q} f^2 \right\rangle\Big|_{\rho=1}$$

$$= \frac{-1}{2Q} \left\langle f^2(Q, x, \mathrm{sign}(x)) \right\rangle$$

$$= \frac{-1}{2Q} \int_{-\infty}^{\infty} Dt \;\; f^2(Q, x, \mathrm{sign}(x)) \;\stackrel{!}{=}\; 0$$

average over

$$\lim_{\rho\to 1} P(x,y) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}y^2\right] \delta(x - \sqrt{Q}y)$$

Gaussian measure $\quad Dt = \dfrac{dt\, e^{-t^2/2}}{\sqrt{2\pi}}$

Successfull *fine–tuning* for $\rho \to 1, \alpha \to \infty$, i.e. $\mathbf{J} \approx \mathbf{B}$, requires small <u>angular</u> changes. Two possibilities:

- $\lim_{\rho\to 1} Q \to \infty$, diverging student length
- $\lim_{\rho\to 1} f(Q, x, \mathrm{sign}(x)) = 0$
  - learning from mistakes only
  - explicit *annealing* $f = \eta(\alpha)\widetilde{f}$ with $\lim_{\alpha\to\infty} \eta(\alpha) = 0$

Optimal on–line training

Off–line perceptron (error–free training with $\varepsilon_t = 0$)

realizes $\varepsilon_g \propto \alpha^{-1}$ for large α

(e.g. <u>repeated</u> presentation of all examples)

- which is the best performance available within the limitations of modified Hebbian training?
- is it possible to achieve an asymptotic behavior $\varepsilon_g \propto \alpha^{-1}$ <u>on–line</u>?

Variational approach: (Kinouchi and Caticha, 1992)

choose modulation function f (within the framework of modified Hebbian learning)

such that the

$$\left\{ \begin{array}{ll} \text{decrease of} & \varepsilon_g \\ \text{increase of} & \rho \end{array} \right\} \text{per example is maximized}$$

(equivalent for the perceptron)

increase of ρ upon presentation of a single example:

$$\frac{\Delta\rho}{1/N} = f\, S_R \left(\frac{y}{\sqrt{Q}} - \frac{\rho\, x}{Q} \right) - \frac{\rho}{2\, Q} f^2 \qquad \text{(before averaging)}$$

necessary condition, (functional) derivative:

$$\frac{\delta}{\delta f}(N\Delta\rho) = S_R \left(\frac{y}{\sqrt{Q}} \frac{\rho\, x}{Q} \right) - \frac{\rho}{Q} f \stackrel{!}{=} 0$$

formally: $\Rightarrow \qquad f^* = S_R \left(\frac{\sqrt{Q}\, y}{\rho} - x \right)$

(Note: second (sufficient) condition for maximum is $\rho > 0$)

Problem (1):

f^* depends explicitly on teacher field y ,

only $S_R = \mathrm{sign}(y)$ is available from the training data

Solution:

average over unavailable information, i.e. over the

conditional density $\quad P(y|x, S_R) = \dfrac{P(x,y)\, \Theta(y\, S_R)}{\int_{-\infty}^{\infty} dy P(x,y)\, \Theta(y\, S_R)}$

One obtains:

$$f_{opt} \;=\; S_R \left(\frac{\sqrt{Q}}{\rho} \, \langle \, y \, \rangle_{y|(x,S_R)} \; - x \right)$$

(for general input distributions)

$$= \frac{\sqrt{Q}}{\sqrt{2\pi}} \frac{\sqrt{1-\rho^2}}{\rho} \qquad \frac{\exp\left[-\frac{1}{2} \frac{\rho^2}{1-\rho^2} \frac{x^2}{Q} \right]}{\Phi\left(\frac{\rho}{\sqrt{1-\rho^2}} \frac{x \, S_R}{\sqrt{Q}} \right)}$$

(for isotropic, independent data)

Problem (2):

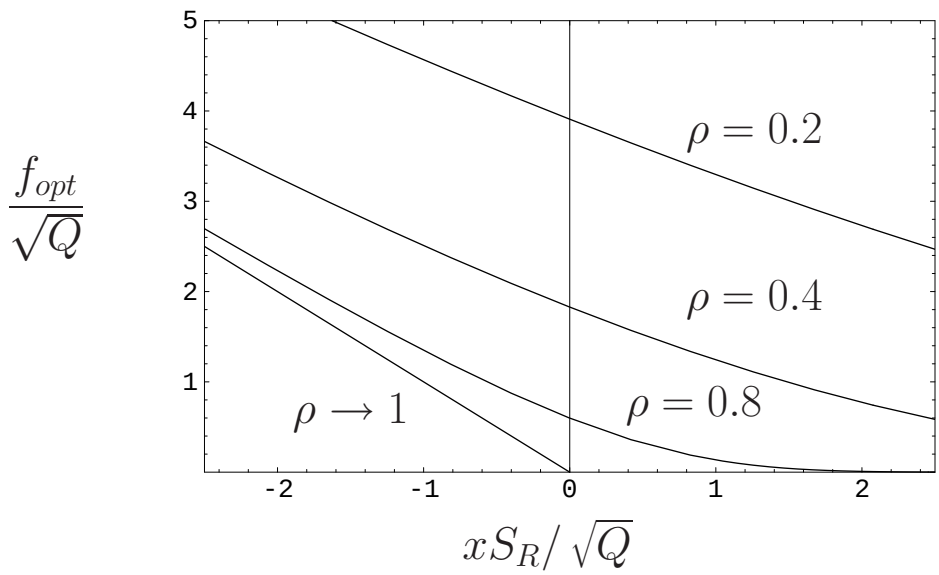
f_{opt} is of the modified Hebbian form,
but contains ρ as a parameter

f_{opt} requires information about the actual
performance (ρ, ε_g) of the student

Any realistic algorithm can at the very best *estimate* ρ ,
$\rightarrow$ analysis of *optimal on–line training* provides

- lower bounds for $\boldsymbol{\varepsilon_g(\alpha)}$
- insights in the general features of f_{opt} which allow for successful learning
- suggestions for how to construct practical algorithms

Features of f_{opt}



- **Learning from mistakes**

  most weight is given to examples with $xS_R < 0$

- **Adaptive embedding**

  f_{opt} increases with $-xS_R$

- **Performance dependent training**

  initially $\rho \approx 0$: $f_{opt} \approx const.$

  asymptotically $\rho \to 1$: $f_{opt} = -(xS_R)\Theta(-xS_R)$

  Hebbian learning $\to$ adaptive perceptron *(AdaTron)* training

Optimal on–line learning curve

Note: we can re–write differential equation for any f which does not depend on y explicitly:

$$\frac{d\rho}{d\alpha} = \frac{\rho}{Q}\left\langle f\, S_R \overbrace{\left(\frac{\sqrt{Q}}{\rho}\langle y\rangle_{y|(x,S_R)} - x\right)}^{f_{opt}} - \frac{1}{2} f^2 \right\rangle_{(x,S_R)}$$

$$= \frac{\rho}{Q}\left\langle f\, f_{opt} - \frac{1}{2} f^2 \right\rangle_{(x,S_R)}$$

$$\frac{d\rho}{d\alpha} = \frac{\rho}{Q}\left\langle \frac{1}{2} f_{opt}^2 \right\rangle$$

$$= \frac{1}{2\pi}\frac{(1-\rho^2)^{3/2}}{\rho}\int_{-\infty}^{\infty}\frac{dx}{\sqrt{2\pi}}\frac{\exp\left[-\frac{1}{2}(1+\rho^2)x^2\right]}{\Phi(\rho x)}$$

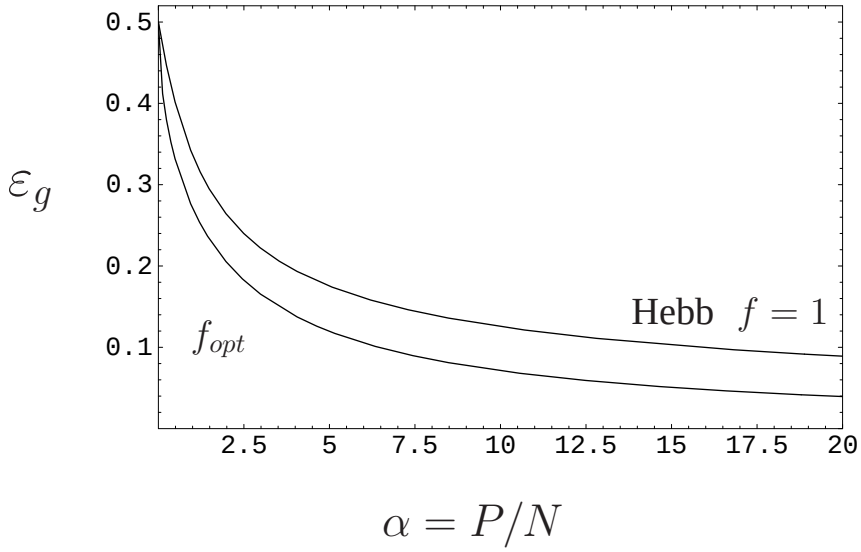
independent of Q –equation!

Exercise

obtain the optimal learning curves $\rho(\alpha)$ ($\varepsilon_g(\alpha)$)

by means of numerical integration, e.g. with *Mathematica*

Why is it difficult to use the initial condition $\rho(0) = 0$?



$\rho(0) \approx 0$

asymptotics

obviously $\left.\dfrac{d\rho}{d\alpha}\right|_{\rho=1} = 0$ (perfect generalization)

def. $z = \pi\varepsilon_g = \arccos(\rho)$, substitute $y = \rho x$

$$\frac{dz}{d\alpha} = -\frac{1}{2\pi}\tan^2 z \int_{-\infty}^{\infty} \frac{dy}{\sqrt{\pi}} \frac{\exp\left[-\frac{1}{2}\frac{1+\cos^2 z}{\cos^2 z}\right]}{\Phi(y)}$$

$$= -\frac{1}{2\pi} z^2 \int_{-\infty}^{\infty} \frac{dy}{\sqrt{2\pi}} \frac{e^{-y^2}}{\Phi(y)} + \mathcal{O}(z^4) \quad \text{for } z \approx 0$$

$$\varepsilon_g(\alpha \to \infty) = 2 \left(\int_{-\infty}^{+\infty} \frac{dy}{\sqrt{2\pi}} \frac{e^{-y^2}}{\Phi(y)} \right)^{-1} \boldsymbol{\alpha^{-1}} \approx \frac{\mathbf{0.88}}{\boldsymbol{\alpha}}$$

Optimal on–line asymptotics differs from optimal off–line result (*) by a factor 2 (exact)

(*) [Bayes optimal, c.m. in version space, opt. potential] requires repeated presentation of all examples

here: single presentation $(\varepsilon_t > 0)$, yet (almost) the same student performance as a function of α !

Practical algorithms with $\varepsilon_g \propto \alpha^{-1}$?

a) obtain a good estimate of $\rho(\varepsilon_g)$ while learning

Note that here (ideal situation):

$$\frac{1}{\sqrt{Q}}\frac{d\sqrt{Q}}{d\alpha} = \frac{1}{\rho}\frac{d\rho}{d\alpha}\,, \quad \Rightarrow \sqrt{Q(0)} = \rho(0) \text{ guarantees } \sqrt{Q(\alpha)} = \rho(\alpha)$$

available, perfect estimate $\sqrt{Q} = \rho$

b) use a simplified version of optimal modulation, e.g.

$$\lim_{\rho\to 1} f_{opt} = (-x\,S_R)\,\Theta(-x\,S_R) \qquad \text{(parameter free)}$$

(relaxation algorithm or AdaTron)

incorporates
- learning from mistakes
- adaptive embedding

presentation of $\boldsymbol{\xi}^\mu$:

$$\mathbf{J}(\mu) \;=\; \mathbf{J}(\mu\text{–}1) \;+\; \frac{1}{N}(-x^\mu\, S_R^\mu)\,\Theta(-x^\mu S_R^\mu)\;\boldsymbol{\xi}^\mu\, S_R^\mu$$

field before/after training

unchanged if $\quad x^\mu \;=\; \mathbf{J}(\mu\text{–}1)\cdot\boldsymbol{\xi}^\mu\, S_R > 0$ already,

else: $\quad \mathbf{J}(\mu)\cdot\boldsymbol{\xi}^\mu \;=\; x^\mu - \frac{1}{N}\, x^\mu\;\boldsymbol{\xi}^\mu\cdot\boldsymbol{\xi}^\mu \;=0$

misclassification is corrected by minimal change of $\mathbf{J}$

(identical with the zero temperature limit of *On–line Gibbs* learning (Kim and Sompolinksy, 1996))

Exercise

Show that the relaxation algorithm solves the following optimization problem:

minimize $\;|\,\mathbf{J}(\mu) - \mathbf{J}(\mu\text{–}1)\,|^2\;$ under the constraint $\;\mathbf{J}(\mu)\cdot\boldsymbol{\xi}^\mu \;\geq\; 0$

. . . diff. eqs. . . . asymptotics (Biehl and Riegler, 1994)

$$\varepsilon_g(\alpha\to\infty) \;=\; \frac{3}{2\,\alpha}$$

(coincides with the *worst student in version space* (Engel))

c) time–dep. learning rate (Barkai, Seung, Sompolinsky 1995)

normalized modified Hebbian learning: $(\mathbf{J}(\mu\text{–}1)^2 = 1)$

$$\mathbf{J}(\mu) = \frac{\mathbf{J}(\mu\text{–}1) + \frac{\eta}{N} f\, \boldsymbol{\xi}^\mu\, S_R^\mu}{\sqrt{1 + \frac{\eta}{N} f x^\mu S_R^\mu + \frac{\eta^2}{N} f^2}}$$

$$= \mathbf{J}(\mu\text{–}1)\left[1 - \frac{1}{N}\left(\eta f x^\mu S_R^\mu + \eta^2 f^2/2\right)\right]$$

$$+\frac{1}{N}\,\eta\, f\, \boldsymbol{\xi}^\mu S_R^\mu \quad + \quad \mathcal{O}(N^{-2})$$

$$\frac{d\rho}{d\alpha} = -\rho\left\langle \eta f x S_R + \eta^2 f^2/2 \right\rangle + \left\langle \eta\, f y S_R \right\rangle$$

For $f = \Theta(-xS_R)$ one obtains (exercise)

$$\frac{d\rho}{d\alpha} = \frac{\eta}{\sqrt{2\pi}}\left(1 - \rho^2\right) - \frac{\eta^2}{2\,\pi}\,\rho\,\arccos(\rho)$$

- constant η

$$\varepsilon_g(\alpha \to \infty) = \frac{\eta}{\sqrt{2\pi^3}} = const.$$

- (optimized) power law annealing: $\eta(\alpha) = 2\sqrt{2\pi}\big/\,\alpha$

$$\varepsilon_g(\alpha \to \infty) \approx \frac{1.27}{\alpha}$$

Learning a linearly separable rule from reliable examples

- exact description of on–line learning dynamics:
  - sequence of indep. random (isotropic) data
  - thermodynamic limit $N \to \infty$:
    typical behavior of large networks
    continuous time limit
- *Plug in and Play !*
  - Hebbian Learning $(f = 1)$, $\varepsilon_g(\alpha \to \infty) \propto \alpha^{-1/2}$
  - Rosenblatt perceptron $\Theta(-xS_R)$, $\varepsilon_g \propto \alpha^{-1/3}$
- variational approach $\to$ optimal modulation function
  - lower bound for typical $\varepsilon_g(\alpha)$
  - asymptotics: $\varepsilon_g(\alpha \to \infty) \approx 0.88 \,/\, \alpha$
- successful practical algorithms:
  - performance estimation
  - relaxation algorithm $\lim_{\rho \to 1} f_{opt}$
  - normalization, time dependent learning rate

Learning from noisy data (Biehl, Riegler, Stechert, 1995)

linearly separable rule $S_R = \mathrm{sign}(\mathbf{B} \cdot \boldsymbol{\xi})$

example data $\{ \boldsymbol{\xi}^\mu, S_T^\mu \}$

where the $S_T^\mu = \pm 1$ are *noisy* (stochastically corrupted) versions of $S_R^\mu = \mathrm{sign}\, y^\mu$

- modified Hebbian training:

$$\mathbf{J}(\mu) = \mathbf{J}(\mu - 1) + \frac{1}{N} f(Q(\mu\text{–}1), x^\mu, S_T^\mu) \; \boldsymbol{\xi}^\mu S_T^\mu$$

on–line dynamics:

$$\frac{dR}{d\alpha} = \langle f\, y\, S_T \rangle_{(x,y,S_T)} \qquad \frac{dQ}{d\alpha} = \langle 2\, f\, x\, S_T + f^2 \rangle_{(x,y,S_T)}$$

$$\frac{d\rho}{d\alpha} = \left\langle f\, S_R \left(\frac{y}{\sqrt{Q}} - \frac{\rho x}{Q} \right) - \frac{\rho}{2Q} f^2 \right\rangle_{(x,y,S_T)}$$

analogous to previous case (reliable data)

$\langle \ldots \rangle_{(x,y,S_T)}$ is the average over

(x, y) <u>and</u> the randomness in S_T (to be specified)

- two performance measures (inconsistent defs. in the literature!)

generalization error $\varepsilon_g = \langle \Theta(-x S_R) \rangle_{(x,y)} = \frac{1}{\pi} \arccos(\rho)$

compares student output with the true (learnable) rule

prediction error $\varepsilon_p = \langle \Theta(-x S_T) \rangle_{(x,y,S_T)}$

compares with corrupted label of a new example

$$\left.\begin{array}{l} \varepsilon_g = 0 \\ \varepsilon_p > 0 \end{array}\right\} \quad \text{for } \mathbf{J} \propto \mathbf{B} \qquad \text{(unpredictable noise in } S_T \text{)}$$

- **optimal modulation function**

$$f_{opt} = S_T \left(\frac{\sqrt{Q}}{\rho} \langle y \rangle_{y|(x,S_T)} - x \right)$$

average over $P(y|x, S_T)$

depends on the nature and strength of the considered noise which is assumed to be known within the optimization scheme

Two examples of stochastic corruption

(multiplicative) **output noise**

(additive) **weight noise** (input noise)

Output noise

$$S_T^\mu = \begin{cases} +S_R^\mu & \text{with prob.} \quad 1-\lambda \\ -S_R^\mu & \text{with prob.} \quad \lambda < 1/2 \quad \text{(inversion rate)} \end{cases}$$

random flip for each example, independent of input $\boldsymbol{\xi}^\mu$

single example averages

$$\langle\, g(x,y,S_T)\,\rangle_{(x,y,S_T)} = \lambda\,\langle\, g(x,y,-S_R)\,\rangle + (1-\lambda)\,\langle\, g(x,y,+S_R)\,\rangle$$

prediction error: $(0 \le \lambda \le 1/2,\ 0 \le \varepsilon_g \le 1/2)$

$$\begin{aligned} \varepsilon_p &= \lambda\,(1-\varepsilon_g) + (1-\lambda)\,\varepsilon_g \\ &= \lambda + (1-2\lambda)\varepsilon_g \qquad \geq \begin{cases} \lambda \\ \varepsilon_g \end{cases} \end{aligned}$$

Hebbian training $\quad (f = 1)$

$$\frac{dR}{d\alpha} = \langle\, y\, S_T \,\rangle_{(x,y,S_T)} = \langle\, |\, y\, |\, (\,(1-\lambda) S_R - \lambda S_R\,)\,\rangle_{(x,y)}$$

$$= \sqrt{\frac{2}{\pi}}\;(\,1 - 2\lambda\,)$$

$$\frac{dQ}{d\alpha} = 2\,\langle\, x\, S_T \,\rangle_{(x,y,S_T)} + 1 = 2\sqrt{\frac{2}{\pi}}\,(1-2\lambda)\,R + 1$$

analytically solvable as for $\lambda = 0$ ...

$$\varepsilon_g(\alpha) = \frac{1}{\pi}\,\arccos\left[\left(1 + \frac{\pi}{2(1-2\lambda)^2\alpha}\right)^{-1/2}\right]$$

$$= \left(\frac{1}{\sqrt{2\pi}\,(1-2\lambda)^2}\right)\alpha^{-1/2} \quad \text{for } \alpha \to \infty$$

perfect generalization still achieved

modified prefactor, diverges for $\lambda \to 1/2$ (random $S_T = \pm 1$)

Optimal modulation function

$$f_{opt} = \sqrt{Q}\,\frac{1-2\lambda}{\sqrt{2\pi}}\,\frac{\sqrt{1-\rho^2}}{\rho}\,\frac{\exp\left[-\frac{1}{2}\frac{\rho^2}{1-\rho^2}\frac{x^2}{Q}\right]}{(1-2\lambda)\,\Phi\left(\frac{\rho}{\sqrt{1-\rho^2}}\frac{xS_T}{\sqrt{Q}}\right)+\lambda}$$

makes use of unavailable/unknown

$\boldsymbol{\rho}$ performance estimation

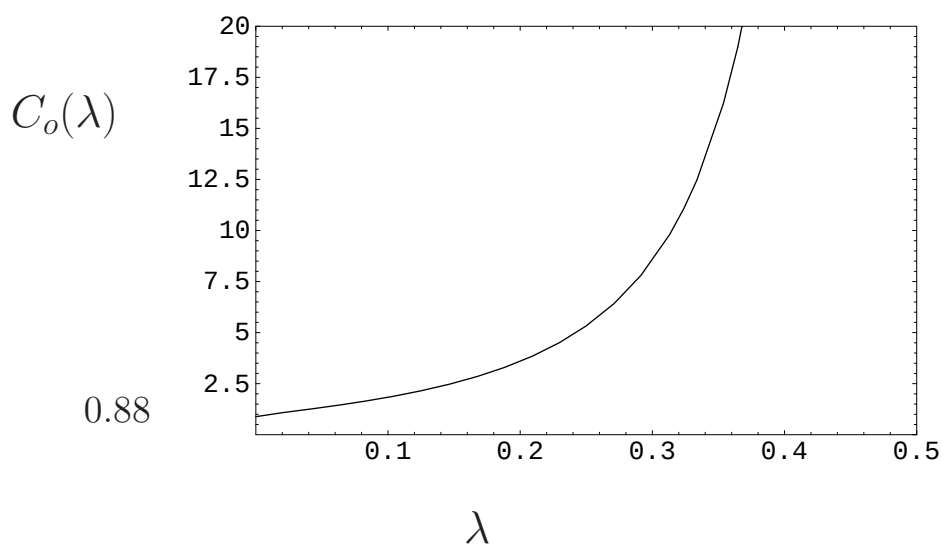
$\boldsymbol{\lambda}$ noise level estimation (see below)

---

asymptotics $\varepsilon_g(\alpha) = \dfrac{C_o(\lambda)}{\alpha}$

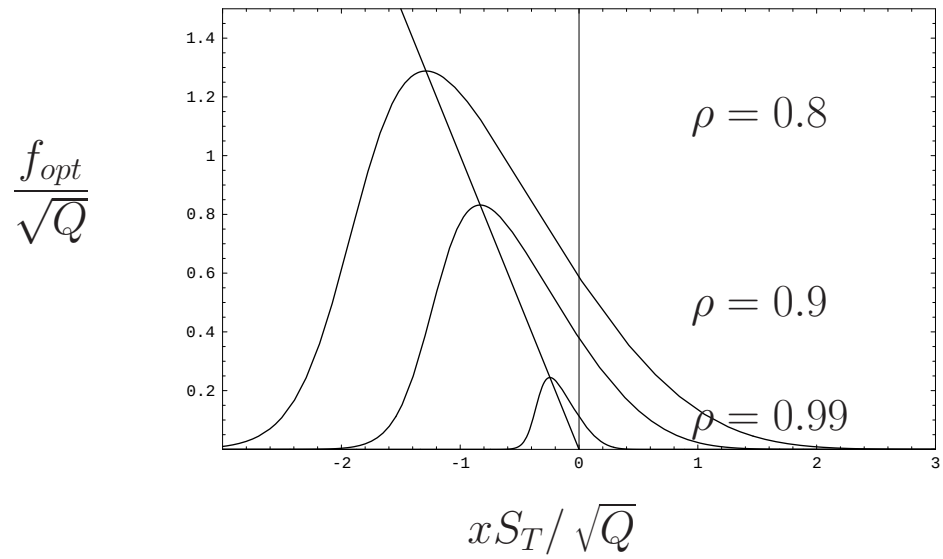
with $C_o(\lambda) = 2\,\dfrac{1}{(1-2\lambda)^2}\left[\int_{-\infty}^{+\infty} Dx\,\dfrac{1}{(1-2\lambda)\Phi(x)+\lambda}\right]^{-1}$

$2\times$ (optimal off–line result) !



optimal features

$\lambda = 0.01$



- **performance dependence**
- **learning from mistakes**
- **confidence**

  large $|x|$ indicate that a misclassification $\mathrm{sign}(x) = -S_T$ is most likely due to the presence of noise

Effective cut–off at negative $x \approx -c$,
roughly given by the width of the Gaussian factor in f_{opt}

$$c = \frac{\sqrt{1-\rho^2}}{\rho} = \tan(\pi \varepsilon_g) \quad \text{(performance dependent)}$$

$$\propto 1/\alpha \quad \text{(explicitly } \alpha \text{ –dependent)}$$

suggests a simplified algorithm: see (Kim,Sompolinsky, 1996)

$$f_{opt} = (-x\, S_T)\, \Theta(-x\, S_T)\, \Theta(x + c)$$

(Mis–) Estimation of the noise level (Copelli et al., 1997)

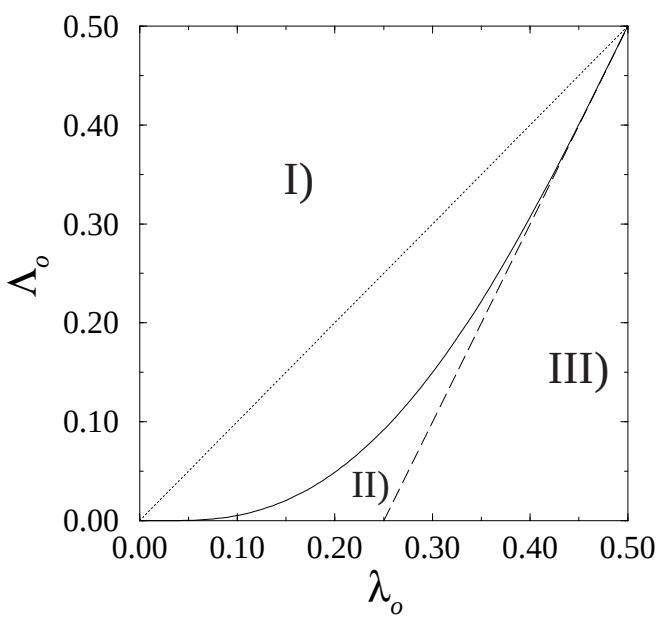
suboptimal modulation function f_Λ :
functional form of f_{opt} but true λ replaced with Λ

$$f_\Lambda = \sqrt{Q}\,\frac{1-2\Lambda}{\sqrt{2\pi}}\,\frac{\sqrt{1-\rho^2}}{\rho}\,\frac{\exp\left[-\frac{1}{2}\frac{\rho}{1-\rho^2}\frac{x^2}{Q}\right]}{(1-2\Lambda)\,\Phi\left(\frac{\rho}{\sqrt{1-\rho^2}}\frac{xS_T}{\sqrt{Q}}\right)+\Lambda}$$

over–estimation $\Lambda > \lambda$: $\varepsilon_g(\alpha\to\infty) = \dfrac{C(\lambda,\Lambda)}{\alpha}$
(perfect generalization !)

under–estimation $\Lambda < \lambda$:

Noise robustness diagram



I) robust region

$\varepsilon_g(\alpha\to\infty) = 0$

II) imperfect generalization

$0 < \varepsilon_g(\alpha\to\infty) < 1/2$

III) stable fixed point $\boldsymbol{\rho = 0}$

$\varepsilon_g(\alpha\to\infty) = 1/2$

(also valid for the *tree committee machine* with K h.u.)

On–line estimation of "hyper–parameter" Λ

$$\Lambda(\mu) = \Lambda(\mu\text{–}1) + \frac{1}{2N}\left[\,(1-\Lambda(\mu\text{–}1))\Theta(-x^\mu S_T^\mu) - \Lambda(\mu\text{–}1)\Theta(+x^\mu S_T^\mu)\,\right]$$

simple minded scheme:

– increases/decreases estimate if example is wrong/right

– obeys $0 \le \Lambda \le 1$

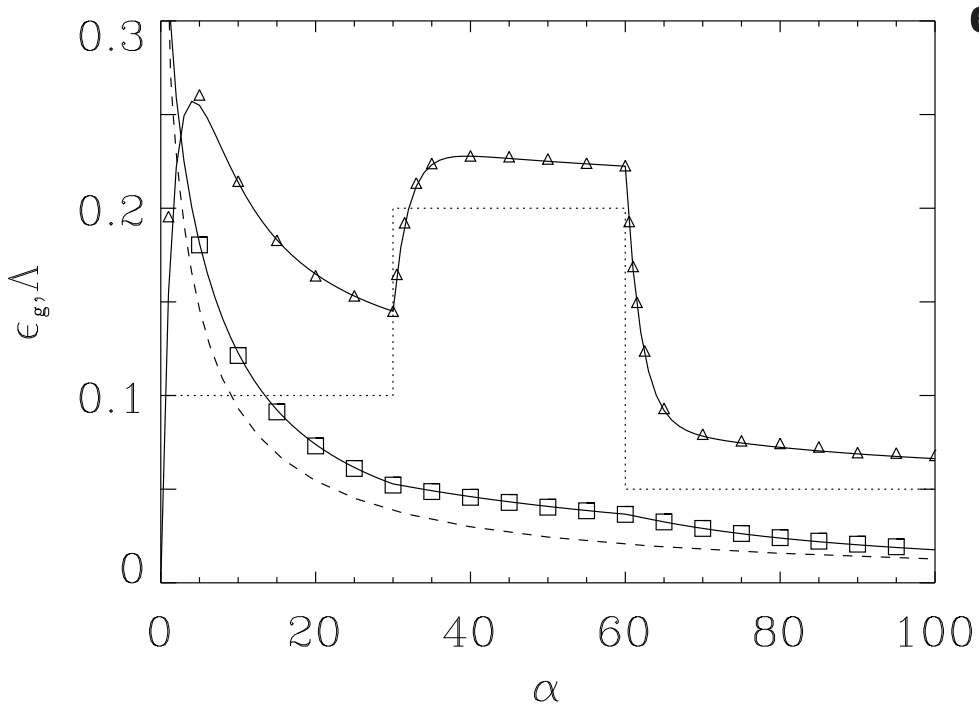
– disregards confidence in student output

average over (x, y, S_T) , continuous time limit $N \to \infty$,

self–averaging properties of $\Lambda(\alpha)$

system described by 3 coupled ODEs for ρ, Q and

$$\frac{d\Lambda}{d\alpha} = \frac{1}{2}\,(\,\varepsilon_p - \Lambda\,)) \qquad \text{(i.e. } \Lambda \to \varepsilon_p \to \lambda)$$



example:

noise level $\lambda(\alpha)$

estimate $\Lambda(\alpha)$

simulations

$(N = 1000)$

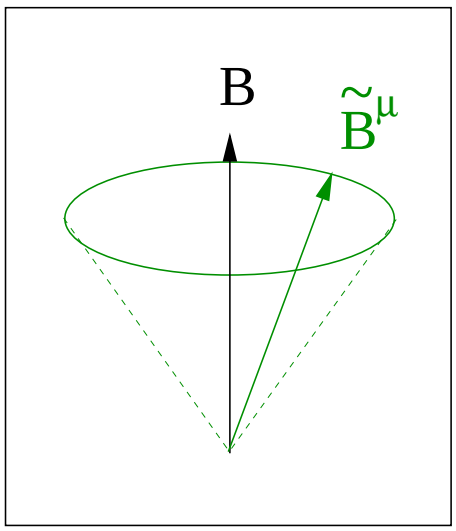
$\varepsilon_g(\alpha)$

Weight noise

S_T^μ determined according to a noisy version of the teacher

$$S_T^\mu = \text{sign}\left(\underbrace{\widetilde{\mathbf{B}}^\mu \cdot \boldsymbol{\xi}^\mu}_{\widetilde{y}^\mu}\right) \quad \text{where} \quad (\widetilde{\mathbf{B}}^\mu)^2 = 1 \quad \text{and} \quad \widetilde{\mathbf{B}}^\mu \cdot \mathbf{B} = \omega$$

(otherwise random and uncorrelated)

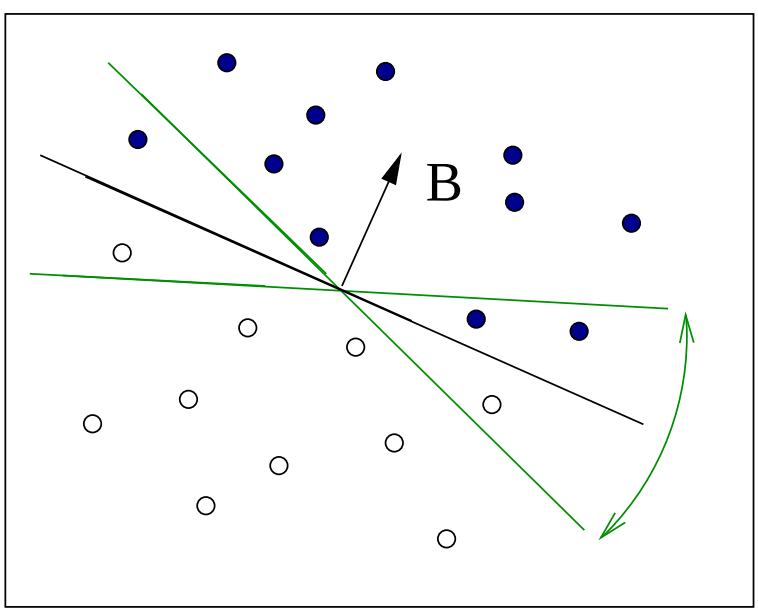


e.g. $\widetilde{\mathbf{B}}^\mu$ vector of i.i.d. components

with $\overline{\widetilde{B}_j^\mu} = \omega\, B_j$

$$\overline{(\widetilde{B}_j^\mu)^2} - \omega^2 B_j^2 = 1 - \omega^2$$

effect of the noise depends on teacher field $y^\mu = \mathbf{B} \cdot \boldsymbol{\xi}^\mu$



formally: (suppress $\overline{(\ldots)}$ for average over $\widetilde{\mathbf{B}}$)

$$\langle\, y \,\rangle = \langle\, \widetilde{y} \,\rangle = 0 \qquad \langle\, y^2 \,\rangle = \langle\, \widetilde{y}^2 \,\rangle = 1 \qquad \langle\, y\widetilde{y} \,\rangle = \omega$$

joint density:

$$P(\widetilde{y}, y) \;=\; \overbrace{\frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{1-\omega^2}} \exp\left[-\frac{1}{2} \frac{(\widetilde{y} - \omega y)^2}{1-\omega^2} \right]}^{P(\widetilde{y}|y)} \; \frac{1}{\sqrt{2\pi}} e^{-y^2/2}$$

inversion probability for $(S_T^\mu = -S_R^\mu)$

conditional

$$\langle\, \Theta(-\widetilde{y}y) \,\rangle_{(\widetilde{y}|y)} \;=\; \Phi\left(\frac{-\omega\,|\,y\,|}{\sqrt{1-\omega^2}} \right) \;=\; \begin{cases} 1/2 & \text{for} \quad |\,y\,| = 0 \\ 0 & \text{for} \quad |\,y\,| \to \infty \end{cases}$$

over–all rate $\qquad \langle\, \Theta(-\widetilde{y}y) \,\rangle_{(\widetilde{y},y)} \;=\; \dfrac{1}{\pi} \arccos(\omega)$

---

full description by means of joint Gaussian density $P(x, y, \widetilde{y})$

$$\langle\, x \,\rangle \;=\; \langle\, y \,\rangle \;=\; \langle\, \widetilde{y} \,\rangle \;= 0 \qquad \langle\, y^2 \,\rangle \;=\; \langle\, \widetilde{y}^2 \,\rangle \;= 1,$$

$$\langle\, x^2 \,\rangle = Q \qquad \langle\, xy \,\rangle \;=\; R \qquad \langle\, \widetilde{y}y \,\rangle \;= \omega$$

$$\langle\, x\widetilde{y} \,\rangle \;= \sum_{j,k} J_j\, \widetilde{B}_k \,\langle\, \xi_j \xi_k \,\rangle \;= \sum_j J_j\, \widetilde{B}_j \;=\; \omega\, R$$

Equivalent interpretation: input noise

noisy input $\widetilde{\boldsymbol{\xi}}^\mu$ with $(\widetilde{\boldsymbol{\xi}}^\mu)^2 = N, \quad \widetilde{\boldsymbol{\xi}}^\mu \cdot \boldsymbol{\xi}^\mu / N = \omega$

e.g. (i.i.d.) $P(\tilde{\xi}_j^\mu) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{1-\omega^2}} \exp\left[-\frac{1}{2} \frac{(\tilde{\xi}_j^\mu - \omega\, \xi_j^\mu)^2}{(1-\omega^2)} \right]$

- $\widetilde{\boldsymbol{\xi}}^\mu$ presented to the **teacher**

$$\mathbf{J}(\mu) = \mathbf{J}(\mu\text{–}1) + \frac{1}{N} f\, \boldsymbol{\xi}^\mu S_T^\mu \quad \text{with } S_T^\mu = \mathrm{sign}(\mathbf{B} \cdot \widetilde{\boldsymbol{\xi}}^\mu)$$

$\widetilde{y}^\mu = \mathbf{B} \cdot \widetilde{\boldsymbol{\xi}}^\mu$ same moments as above, e.g.

$$\overline{\langle\, \widetilde{y}^\mu y \,\rangle} = \sum_{jk} B_j B_k \left\langle\, \xi_j \overline{\tilde{\xi}_k} \,\right\rangle = \sum_{jk} B_j B_k \omega \left\langle\, \xi_j \xi_k \,\right\rangle = \omega$$

- $\widetilde{\boldsymbol{\xi}}^\mu$ presented to the **student**

$$\mathbf{J}(\mu) = \mathbf{J}(\mu\text{–}1) + \frac{1}{N} f\, \widetilde{\boldsymbol{\xi}}^\mu S_R^\mu \quad \text{with } S_R^\mu = \mathrm{sign}(\mathbf{B} \cdot \boldsymbol{\xi}^\mu)$$

fields $\widetilde{x} = \mathbf{J}(\mu\text{–}1) \cdot \widetilde{\boldsymbol{\xi}}^\mu$, $y = \mathbf{B} \cdot \boldsymbol{\xi}^\mu$, and $\widetilde{y} = \mathbf{B} \cdot \widetilde{\boldsymbol{\xi}}^\mu$
with the same statistics as $(x, \widetilde{y}, y)$ from above

---

- student weight noise $\mathbf{J}(\mu) \to \widetilde{\mathbf{J}}(\mu)$ is <u>not equivalent</u> as it changes the student/teacher overlap explicitly

Hebbian training $\quad f = 1$

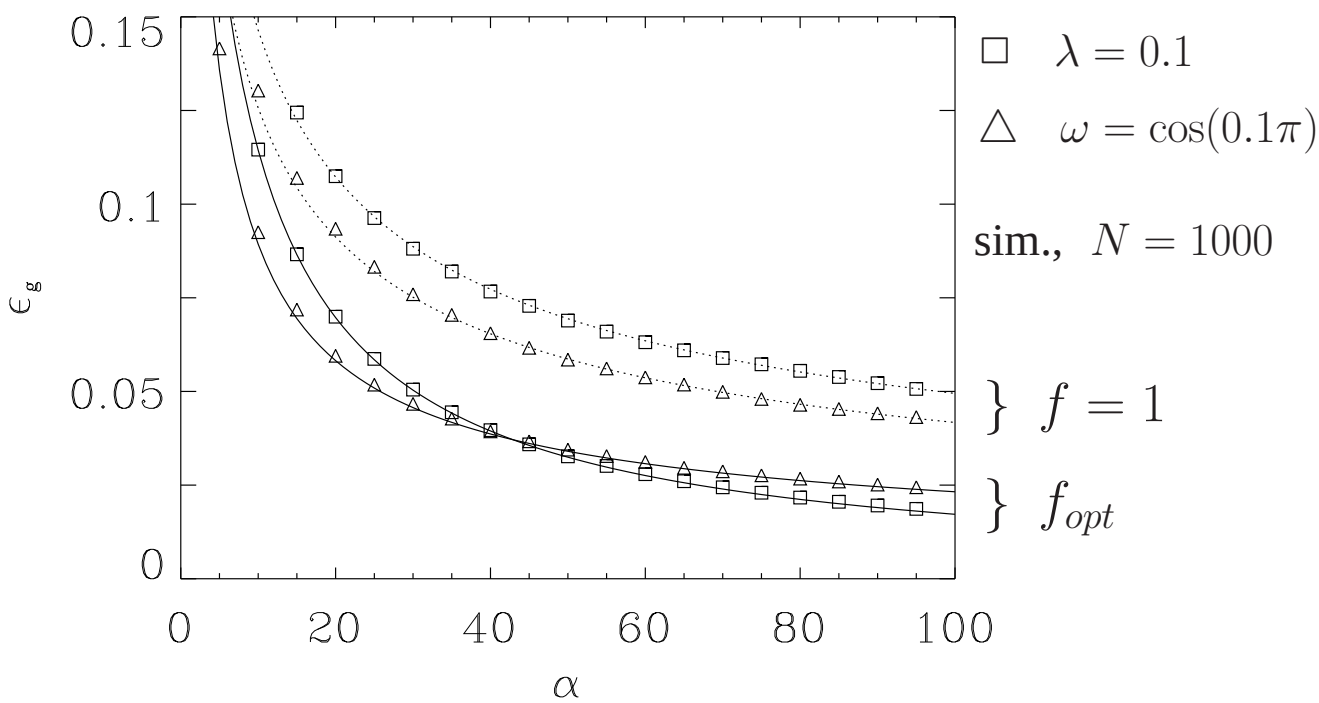
$$\frac{dR}{d\alpha} = \langle\, y \;\mathrm{sign}(\widetilde{y})\,\rangle_{(y,\widetilde{y})} = \sqrt{\frac{2}{\pi}}\,\omega$$

$$\frac{dQ}{d\alpha} = \langle\, x\,\mathrm{sign}(\widetilde{y})\,\rangle_{(x,\widetilde{y})} + 1 = \sqrt{\frac{2}{\pi}}\,\omega\, R + 1$$

formally identical with learning in the presence of output noise, replace $(1-2\lambda)$ with ω

$$\varepsilon_g(\alpha) = \frac{1}{\pi}\arccos\left[\left(1+\frac{\pi}{2\,\omega^2\,\alpha}\right)^{-1/2}\right]$$

comparison at equal inversion rates, i.e. for $\quad \lambda = \frac{1}{\pi}\arccos(\omega)$



generalization vs. prediction error :

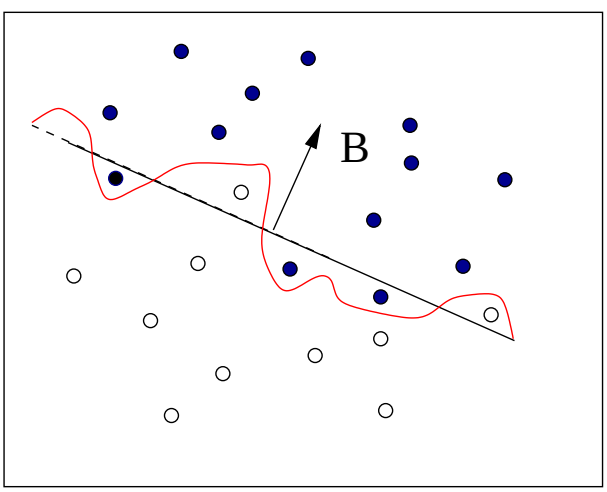
$$\varepsilon_p = \frac{1}{\pi} \arccos\left(\omega \cos(\pi\varepsilon_g)\right) = \frac{1}{\pi}\arccos(\omega\rho)$$

Asymptotical behavior for $\varepsilon_g \to 0$:

$$\varepsilon_p = \frac{1}{\pi}\arccos(\omega) + \frac{\omega\pi}{2\sqrt{1-\omega^2}}\,\varepsilon_g^2 = \varepsilon_p(\infty) + \mathcal{O}(\alpha^{-1})$$

Weight noise as a simple model of unlearnable rules

assume $S_R(\boldsymbol{\xi})$ is not linearly separable, but a more complicated function of $y = \mathbf{B}\cdot\boldsymbol{\xi}$, $S_R = \widetilde{g}(y)$



resembles the effect of weight noise

interpretation: <u>generalization error</u> ε_p

$$\left(\varepsilon_p(\alpha) - \varepsilon_p(\infty)\right) \propto \alpha^{-1}$$

Optimal modulation function

$$f_{opt} = \frac{\sqrt{Q}}{\sqrt{2\pi}} \, \frac{\omega\sqrt{1-\rho^2}}{\rho\sqrt{1-\omega^2\rho^2}} \, \frac{\exp\left[-\frac{1}{2}\frac{\omega^2\rho^2}{1-\omega^2\rho^2}\frac{x^2}{Q}\right]}{\Phi\left(\frac{\omega\rho}{\sqrt{1-\rho^2}}\frac{xS_T}{\sqrt{Q}}\right)}$$

asymptotics: in the limit $\alpha \to \infty$:

$$\varepsilon_g = \frac{(1-\omega^2)^{1/4}}{(\omega\pi)^{1/2}} \left[\int_{-\infty}^{+\infty} \frac{dt}{\sqrt{2\pi}} \frac{\exp\left(-\frac{1}{2}\frac{1+\omega^2}{\omega^2}t^2\right)}{\Phi(t)}\right]^{-1/2} \qquad \boldsymbol{\alpha^{-1/2}}$$

(coincides <u>exactly</u> with optimal off–line learning result!)

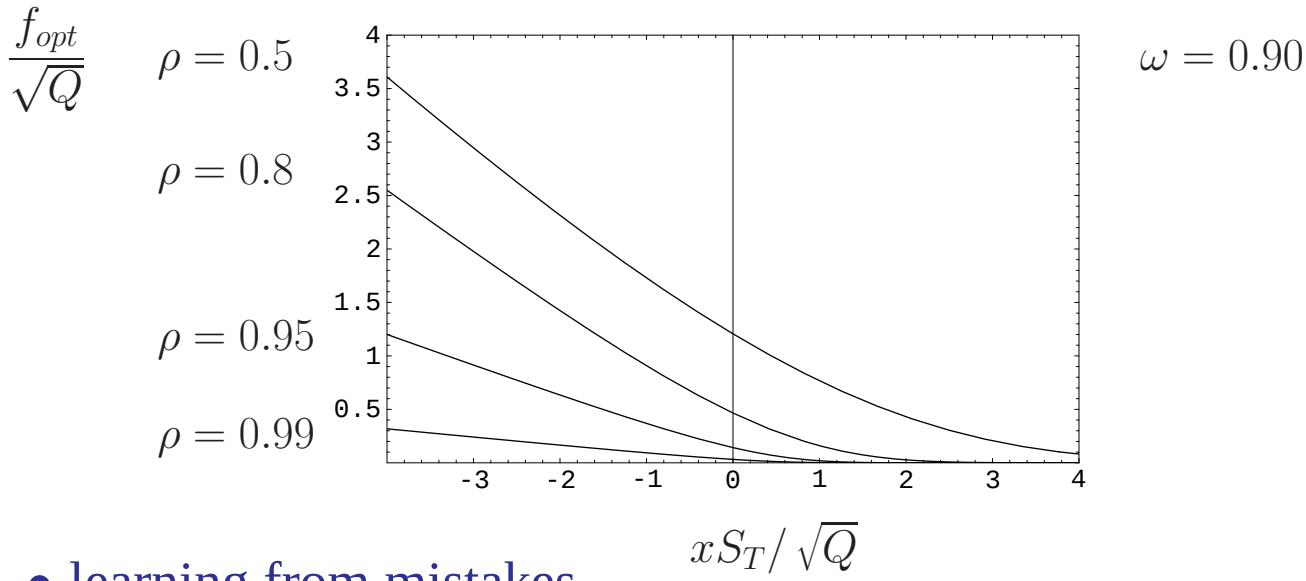
- small $\alpha, \ \rho \approx 0$

very weak effect of weight noise:
learning $\mathbf{B}$ or $\widetilde{\mathbf{B}}$ is not much different

- large $\alpha, \ \rho \approx 1$

very pronounced effect of weight noise:
fine tuning requires disagreement between student and teacher, i.e. $|\,y\,| \approx 0$ with very probable inversion

Features:



- learning from mistakes
- adaptive behavior for $x\,S_T < 0$
- <u>no cut–off</u>, but $f_{opt} \to 0$ with $\rho \to 1$
  student confidence is not helpful, because examples with large $|x|$ are corrupted by noise with very little probability

---

Note: the dependence of f_{opt} on the (unknown) noise level ω is only through ε_p :

$$f_{opt} = \frac{\sqrt{Q}}{\sqrt{2\pi}} \frac{\tan^2(\pi\varepsilon_g)}{\tan(\pi\varepsilon_p)} \frac{\exp\left[-\frac{1}{2}\frac{1}{\tan^2(\pi\varepsilon_p)}\frac{x^2}{Q}\right]}{\Phi\left(\frac{1}{\tan(\pi\varepsilon_p)}\frac{xS_T}{\sqrt{Q}}\right)}$$

$$\left(\tan(\pi\varepsilon_g) = \frac{\rho}{\sqrt{1-\rho^2}}, \qquad \tan(\pi\varepsilon_p) = \frac{\sqrt{1-\omega^2\rho^2}}{\omega\rho} \right)$$

Exercise: the limit of small prediction error $(\rho, \omega \to 1)$

$$\lim_{\varepsilon_p \to 0} f_{opt} = \frac{\tan^2(\pi\varepsilon_g)}{\tan^2(\pi\varepsilon_p)} \underbrace{\Theta(-x\, S_T)\,(-\,x\, S_T)}_{\text{(relaxation alg.)}}$$

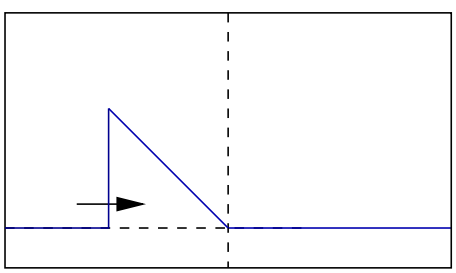
effective learning rate

$$\eta(\rho, \omega) = \frac{\tan^2[\arccos(\rho)]}{\tan^2[\arccos(\omega\rho)]} \qquad \text{(performance dependent)}$$

$$\eta(\alpha) = c(\omega)/\sqrt{\alpha} \qquad \text{(explicitly } \alpha \text{ –dep.)}$$

Summary

- **output noise** corrupts all examples equally efficient
  encourages **confidence** in the student if $|\,x\,|$ is large

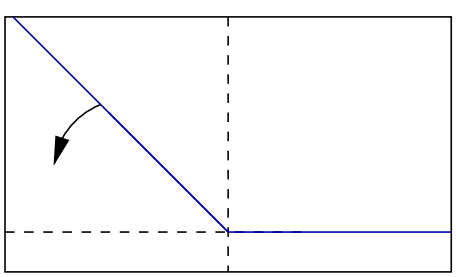


cut–off at $x = -c$

$c \propto \varepsilon_g \propto \alpha^{-1}$

On–line Gibbs algorithm

- **weight noise** affects mainly examples with $|\,y\,| \approx 0$
  hinders *fine–tuning* for $\mathbf{J} \approx \mathbf{B}$



decaying learning rate η

$\eta \propto \varepsilon_g \propto \alpha^{-1/2}$

Remarks and Questions

- In more realistic situations both **nature** and **strength** of the noise are unknown; requires on–line estimation of performance, and assignment of the *blame* to the different possible sources (according to $|x^\mu|$)
- In the presence of weight <u>and</u> output noise the more drastic effect of weight noise dominates, i.e.
  $\rightarrow \varepsilon_g \propto \alpha^{-1/2}$ asymptotically
- Why is Hebbian learning already optimal in the presence of weight noise? (apart from pre–factors)

General results w.r.t. "on–line vs. off–line learning":

Opper, 1995
Kinouchi and Caticha, 1995
van den Broeck, Reimann, 1996

optimal on–line is identical with optimal off–line asymptotics for *smooth* processes, e.g. weight noise

optimal on–line coincides with optimal off–line asymptotics apart from a prefactor (2) for *non–smooth* processes, e.g. output noise

Learning from clustered input data

(Meir (1995), Barkai et al (1993), Riegler et al (1996))

expect in realistic situations that the density of observed inputs reflects features which are relevant for the classification

examples: hand written digits “3” and “7”
medical data for diagnosis

a simple model:

$P(\boldsymbol{\xi})$ with a single symmetry breaking direction

$\mathbf{C} \in I\!R^N$ with $\mathbf{C}^2 = 1$, e.g. **all** $C_j = 1/\sqrt{N}$

$$P(\boldsymbol{\xi}) = 1/2 \sum_{\sigma=\pm 1} P(\boldsymbol{\xi}|\sigma)$$

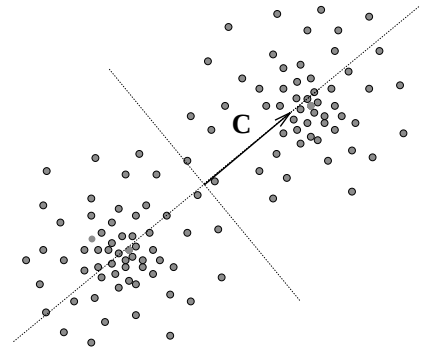
$$P(\boldsymbol{\xi}\,|\,\sigma) = = \prod_{j=1}^{N} \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}(\xi_j - m\sigma C_j)^2\right]$$

two **overlapping** clusters of data centered at $\pm m\mathbf{C}$

dummy variable σ denotes the cluster membership

conditional averages $\langle (\ldots) \rangle_\sigma$ over $P(\boldsymbol{\xi}|\sigma)$

$$\langle \xi_j \rangle_\sigma = m\,\sigma\,C_j = \mathcal{O}(1/\sqrt{N})$$
$$\langle \xi_j^2 \rangle_\sigma - m^2 C_j^2 = 1 \qquad (\boldsymbol{\xi}^2 = N)$$
$$\langle \xi_j \xi_k \rangle_\sigma - m^2 C_j C_k = 0 \text{ for } j \neq k$$



overlap $z = \mathbf{C} \cdot \boldsymbol{\xi}$:

$$\langle z \rangle_\sigma = \sum_j C_j \langle \xi_j \rangle_\sigma = m\,\sigma \qquad \left(\text{but: } \langle z \rangle = \frac{\langle z \rangle_{+1} + \langle z \rangle_{-1}}{2} = 0 \right)$$

$$\langle z^2 \rangle_\sigma = \sum_{j,k} C_j C_k \langle \xi_j \xi_k \rangle_\sigma$$

$$= \sum_j (1 + m^2 C_j^2) C_j^2 + \sum_{j,k(j \neq k)} C_j^2 C_k^2 m^2 = \ldots = 1 + m^2$$

i.e. $\langle z^2 \rangle_\sigma - \langle z \rangle_\sigma^2 = 1$, but $\langle z^2 \rangle - \langle z \rangle^2 = 1 + m^2$

---

vector $\mathbf{J}$ with $\mathbf{J} \cdot \mathbf{C} = D$, $\quad x = \mathbf{J} \cdot \boldsymbol{\xi}$

$$\langle x \rangle_\sigma = m\,D\,\sigma \qquad \langle x^2 \rangle_\sigma = Q + m^2 D^2$$

$$\langle x\,z \rangle_\sigma = D\,(1 + m^2)$$

$$P(x, y|\sigma) = \frac{(2\pi)^{-1}}{\sqrt{1 - D^2}} \exp\left[-\frac{1}{2} \frac{(x - Dy)^2}{1 - D^2} - \frac{1}{2}(y - m\sigma)^2 \right]$$

for $D = 0$ double Gaussian structure is not *visible*, and $P(x, y|\sigma)$ factorizes

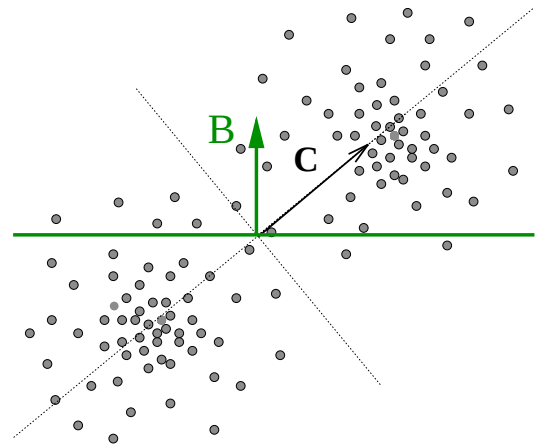
Linearly separable classification of clustered data

symmetry breaking dir. **C**

teacher vector **B**

$$S_R(\boldsymbol{\xi}) \;=\; \mathrm{sign}(\mathbf{B}\cdot\boldsymbol{\xi})$$

parameter $\quad \chi \;=\mathbf{B}\cdot\mathbf{C}$



χ quantifies the alignment of the rule with the underlying structure in input space, i.e. its relevance for the classification

expectation:

any $\chi > 0$ should be helpful in learning the concept **B**

$$x = \mathbf{J}\cdot\boldsymbol{\xi}, \quad y = \mathbf{B}\cdot\boldsymbol{\xi}, \quad z = \mathbf{C}\cdot\boldsymbol{\xi}$$

$$R \;=\; \mathbf{J}\cdot\mathbf{B}, \quad D \;=\mathbf{J}\cdot\mathbf{C}, \quad \chi \;=\; \mathbf{B}\cdot\mathbf{C}$$

$$P(x,y,z) \;=\; \frac{1}{2}\sum_{\sigma=\pm1} P(x,y,z|\sigma)$$ with conditional moments

$$\langle\, y\,\rangle_c \;=\; m\,\chi\,\sigma, \quad \left\langle\, y^2\,\right\rangle_c \;=\; 1 + m^2\,\chi^2$$

$$\langle\, xy\,\rangle_c \;=\; R + m^2\,\chi\,D, \quad \langle\, yz\,\rangle_c = \chi(1+m^2)$$

(others given above)

---

generalization error:

$$\varepsilon_g \;=\; \langle\, \Theta(-x S_R)\,\rangle_{(x,y)}$$

$$= \;\Phi(\chi\, m) \;-\; \Phi\left(\frac{mD}{\sqrt{Q}}\right) \;+\; 2 \int\limits_{-\infty}^{mD/\sqrt{Q}} Dt\, \Phi\left(\frac{tR/\sqrt{Q} \;-\; \chi\, m}{\sqrt{1 - R^2/Q}}\right)$$

---

differential equations for modified Hebbian learning:

(averages over full $P(x, y, z)$)

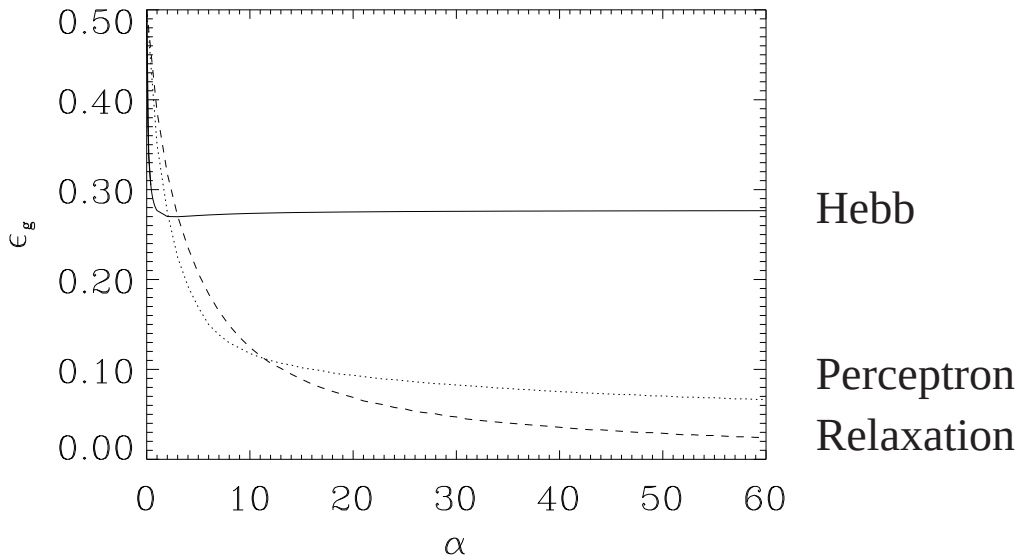
$$\frac{dR}{d\alpha} \;=\; \langle\, f\, y\, S_R\,\rangle, \quad \frac{dD}{d\alpha} \;=\; \langle\, f\, z\, S_R\,\rangle, \quad \frac{dQ}{d\alpha} \;=\; \left\langle\, 2 f\, x\, S_R + f^2\,\right\rangle$$

Hebbian Learning $(f = 1)$ (analytically solvable)

$$R(\alpha) \;=\; \left[\sqrt{\frac{2}{\pi}}\, e^{-\frac{1}{2}\chi^2 m^2} \;+ \chi\, m\; \mathrm{erf}\left(\frac{\chi m}{\sqrt{2}}\right)\right] \alpha$$

$$D(\alpha) \;=\; \left[\chi \sqrt{\frac{2}{\pi}}\, e^{-\frac{1}{2}\chi^2 m^2} \;+\; m\; \mathrm{erf}\left(\frac{\chi m}{\sqrt{2}}\right)\right] \alpha$$

$$Q(\alpha) \;=\; \left[1 + m\; \mathrm{erf}\left(\frac{\chi m}{2}\right) D + \frac{\sqrt{2}}{\pi}\, e^{-\frac{1}{2} m^2 \chi^2}\, R\right] \alpha \;\; + 1$$



Hebbian training **fails** to learn the l.s. rule (in general)

asymptotically: linear combination of teacher and **C**

$$\mathbf{J} \;\rightarrow\; \sqrt{\frac{2}{\pi}}\, e^{-\chi^2 m^2/2}\, \mathbf{B} \;+\; m \;\mathrm{erf}\left(\frac{\chi\, m}{\sqrt{2}}\right) \mathbf{C}$$

successful learning $\varepsilon_g \rightarrow 0$ is only possible if

$m = 0$ (irrelevant structure vector)

$\chi = 1$ (perfect alignment with the rule)

$\chi = 0$ (learning in the space orthogonal to **C**)

One can show more generally:

Hebb student can only generalize perfectly for $\alpha \rightarrow \infty$ when **B** is an eigenvector of the covariance matrix $\underline{M} = \left\langle \boldsymbol{\xi}\boldsymbol{\xi}^\top \right\rangle_{P(\boldsymbol{\xi})}$ with elements $M_{jk} = \langle \xi_j \xi_k \rangle$. Here: either $\mathbf{B} || \mathbf{C}$ or $\mathbf{B} \perp \mathbf{C}$

Hebbian perfect generalization with $\varepsilon_g \propto \alpha^{-1/2}$ is an artefact of assuming an isotropic data distribution!

Learning from mistakes

- Rosenblatt Perceptron $(f = \Theta(-xS_R)$

. . . asymptotics:

$$\varepsilon_g(\alpha \to \infty) = \frac{1}{\pi}\left(\frac{2}{3}e^{-\chi^2 m^2}\right)^{1/3} \boldsymbol{\alpha^{-1/3}}$$

The perceptron algorithm is fast for non–malicious input distributions (E.B. Baum, 1990)

(Exercise: suggest a candidate for a "malicious" input distribution)

- AdaTron, Relaxation $(f = (-xS_R)\,\Theta(-xS_R)$

. . . asymptotics:

$$\varepsilon_g(\alpha \to \infty) = \frac{\mathbf{3}}{\mathbf{2\alpha}}$$

independent of the structure (gain for small α)

2 Learning by on–line gradient descent

(Amari, 1967,1995)

Commonly used in practical applications:

multilayered neural networks with continuous activation functions, where output is a differentiable function of the adaptive parameters

learning based on the gradient of a continuous cost fct.

popular: **backpropagation of error**

see for instance (Chauvin and Rumelhart, 1995)

Exercise: The "linear perceptron"

student: $S_H = x = \mathbf{J} \cdot \boldsymbol{\xi} \in I\!R$

teacher: $S_R = y = \mathbf{B} \cdot \boldsymbol{\xi} \in I\!R$ with $\mathbf{B}^2 = 1$

training and performance evaluation are based on the quadratic error

$$\varepsilon(\boldsymbol{\xi}) = \frac{1}{2}(x-y)^2 \qquad \text{calculate} \quad \varepsilon_g = \langle\, \varepsilon(\boldsymbol{\xi}) \,\rangle$$

consider the training dynamics

$$\mathbf{J}(\mu) = \mathbf{J}(\mu\text{–}1) - \frac{\eta}{N} \nabla_J \varepsilon(\boldsymbol{\xi}^\mu)$$

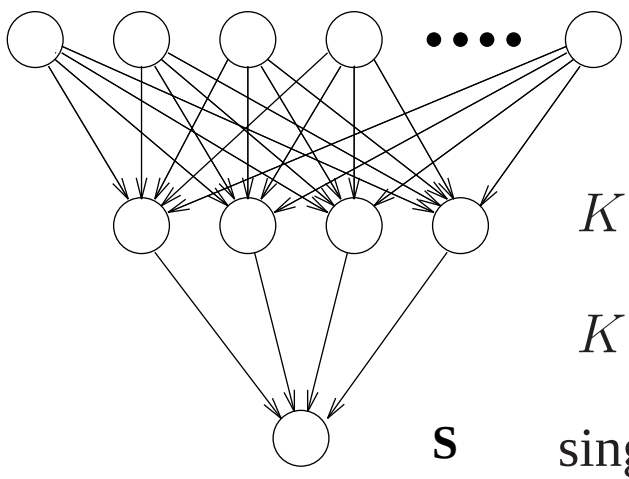
analyse on–line learning from random examples $\{\, \boldsymbol{\xi}^\mu, y^\mu \,\}$
in analogy to the simplest case considered for the sign–perceptron

Show that $\dfrac{d\varepsilon_g}{d\alpha} = (\eta^2 - 2\eta)\, \varepsilon_g$

and investigate the role of the learning rate η

Multilayered, feed–forward neural networks

example architecture: the "**soft–committee machine**"



ξ N –dim. inputs

K weight vectors $\mathbf{J}_k$

K *hidden units*

K weights w_j

single (linear) output unit

hidden unit **activation**:

$g(x_i)$ with argument $x_i = \mathbf{J}_i \cdot \boldsymbol{\xi}$

transfer function: **sigmoidal** $g(x)$, e.g. $g(x) = \tanh(x)$

here: (for analytical reasons)

$$g(x) = \mathrm{erf}\left(\frac{x}{\sqrt{2}}\right) = \frac{2}{\sqrt{\pi}} \int_0^{x/\sqrt{2}} e^{-t^2}\, dt, \qquad g'(x) = \frac{e^{-x^2/2}}{\sqrt{2\pi}}$$

total output:

$$S(\boldsymbol{\xi}) = \sum_{i=1}^{K} w_i\, g(x_i) = \sum_{i=1}^{K} w_i\, \mathrm{erf}\left(\frac{x_i}{\sqrt{2}}\right)$$

Note: (Cybenko, 1989)

a network with the slightly more complicated output

$$S(\boldsymbol{\xi}) = \sum_{i=1}^{K} w_i \, g(\mathbf{J}_i \cdot \boldsymbol{\xi} - \vartheta_i)$$

is a *universal approximator* if

– g continuous, sigmoidal

– $\{ \mathbf{J}_i, \vartheta_i, w_i \}$ adaptive parameters

– K unrestricted (might need very large K)

Student teacher scenario

assume the same architecture in student and teacher, but

K h.u. in the student vs. M h.u. in the teacher

$K = M$ perfectly learnable rules

$K > M$ *overlearnable* rules

$K < M$ unlearnable rules

We will focus on matching architecture $(K = M)$ in the following

The error measure

one (obvious) choice for continuous outputs is

$$\varepsilon(\boldsymbol{\xi}) \;=\; \frac{1}{2}\,(\,S_H(\boldsymbol{\xi}) - S_R(\boldsymbol{\xi})\,)^2$$

$$= \;\frac{1}{2}\left(\sum_{i=1}^{K} w_i g(x_i) \;-\; \sum_{n=1}^{M} v_n\, g(y_n)\right)^2$$

where $x_i \;=\; \mathbf{J}_i \cdot \boldsymbol{\xi}, \qquad y_i \;=\; \mathbf{B}_i \cdot \boldsymbol{\xi}$

popular off–line training energy:

$$H \;=\; \sum_{\mu=1}^{P} \varepsilon(\boldsymbol{\xi}^\mu) \;\; \text{for a given} \;\; \mathbb{D} \;=\; \{\,\boldsymbol{\xi}^\mu, S_R(\boldsymbol{\xi}^\mu)\,\}_{\mu=1,\ldots P}$$

On–line gradient descent

(Biehl, Schwarze, 1995)

sequence of examples in discrete time μ

$$\mathbf{J}_i(\mu) \;= \mathbf{J}_i(\mu\text{–}1) \;-\; \frac{\eta}{N}\,\nabla_{J_i}\;\varepsilon(\boldsymbol{\xi}^\mu)\Big|_{(\mu-1)}$$

$$w_i(\mu) \;=\; w_i(\mu\text{–}1)\;\frac{\eta}{N}\,\nabla_{w_i}\;\varepsilon(\boldsymbol{\xi}^\mu)\Big|_{(\mu-1)}$$

learning rate $\eta = \mathcal{O}(1) \quad for(\mathbf{J}_i^2 = \mathcal{O}(1))$

here: the same η everywhere in the network

isotropic, uncorrelated input data:

$$\langle\, \xi_j^\mu \,\rangle \;=\; 0, \qquad \langle\, \xi_j^\mu\, \xi_k^\nu \,\rangle \;=\; \delta_{\mu\nu}\, \delta_{jk}, \qquad (\boldsymbol{\xi}^\mu)^2 = N$$

CLT $\Rightarrow\; \{\, x_i \,\}_{i=1,2\ldots,K}\; and\; \{\, y_m \,\}_{m=1,2,\ldots,M}$ are
$K+M$ correlated Gaussians with zero mean and

$$\langle\, x_i\, x_k \,\rangle \;=\; Q_{ik} \;= \mathbf{J}_i \cdot \mathbf{J}_k \;= Q_{ki}$$

$$\langle\, x_i\, y_n \,\rangle \;=\; R_{in} \;= \mathbf{J}_i \cdot \mathbf{B}_n$$

$$\langle\, y_n\, y_m \,\rangle \;=\; T_{mn} \;= \mathbf{B}_m \cdot \mathbf{B}_n \;= T_{nm}$$

teacher–teacher overlaps T_{nm} and weights v_n
specify properties of the rule

order parameters and weights: $\{\, R_{in}, Q_{ik}, w_i \,\}$

are sufficient to describe large $(N \to \infty)$ students

microscopics: $K\,N \,+\, N$ degrees of freedom
macroscipics: $K(K-1)/2 + KM + K$ quantities

Generalization error: $\varepsilon_g = \langle \varepsilon(\boldsymbol{\xi}) \rangle_{P(\boldsymbol{\xi})}$

average over the Gaussian density $P(\{x_i, y_n\})$
can be performed analytically (for $g(x) = \mathrm{erf}(x/\sqrt{2})$)

(Saad and Solla, 1995)

$$\varepsilon_g = \frac{1}{\pi}\Bigg[\sum_{i,k=1}^{K} w_i\, w_k \arcsin\left(\frac{Q_{ik}}{\sqrt{1+Q_{ii}}\sqrt{1+Q_{kk}}}\right) + \sum_{n,m=1}^{M} v_n\, v_m \arcsin\left(\frac{T_{nm}}{\sqrt{1+T_{nn}}\sqrt{1+T_{mm}}}\right) - 2\sum_{i=1}^{K}\sum_{n=1}^{M} w_i\, v_n \arcsin\left(\frac{R_{in}}{\sqrt{1+Q_{ii}}\sqrt{1+T_{nn}}}\right)\Bigg]$$

The error measure and hence the generalization error reflect symmetries of the soft–committee machine

most important: **permutation symmetry**
invariance under permutation of *branches* $\{\mathbf{J}_i, w_i\}$

A simple case:

- hidden–to–output weights: $v_n = 1$
- *a priori* knowledge of v_n , fixed $w_i = v_i = 1$

---

weight vector update:

$$\mathbf{J}_i(\mu) = \mathbf{J}_i(\mu\text{–}1) - \frac{\eta}{N}\,[\,S_H(\boldsymbol{\xi}^\mu) - S_R(\boldsymbol{\xi}^\mu)\,]\,g'(x_i^\mu)\,\boldsymbol{\xi}^\mu$$

where $x_i^\mu = \mathbf{J}_i(\mu\text{–}1)\cdot\boldsymbol{\xi}^\mu$

Recursion relations e.g. for $R_{in} = \mathbf{J}_i\cdot\mathbf{B}_n$

$$R_{in}(\mu) = R_{in}(\mu\text{–}1) - \frac{\eta}{N}\,\delta_i^\mu\,y_n^\mu$$

with $\delta_i^\mu = g'(x_i^\mu)\,\Delta^\mu$

and $\Delta^\mu = S_H^\mu - S_R^\mu = \left[\sum_{i=1}^{K} g(x_i^\mu) - \sum_{n=1}^{M} g(y_i^\mu)\right]$

- average over latest example
- self–averaging assumption for all order parameters
- continuous time limite $\alpha = \mu/N$

$$\frac{dR_{in}}{d\alpha} = -\eta \, \langle \, \delta_i \, y_n \, \rangle$$

$$\frac{dQ_{ik}}{d\alpha} = -\eta \, \langle \, \delta_i \, x_i + \delta_j \, x_i \, \rangle + \eta^2 \, \langle \, \delta_i \, \delta_j \, \rangle$$

All averages on the r.h.s. can be performed analytically,
(for arbitrary K, M) (Saad,Solla,1995)

one obtains a system of first order ODE's
for the order parameters $\{ R_{in}(\alpha), Q_{ij}(\alpha) \}$

numerical integration yields **learning curve** $\boldsymbol{\varepsilon_g(\alpha)}$
specify: $K, M,$ matrix $T_{nm},$ learning rate η
<u>initial conditions</u>

Example: $\boldsymbol{K = M = 2, \quad T_{nm} = \delta_{nm}}$

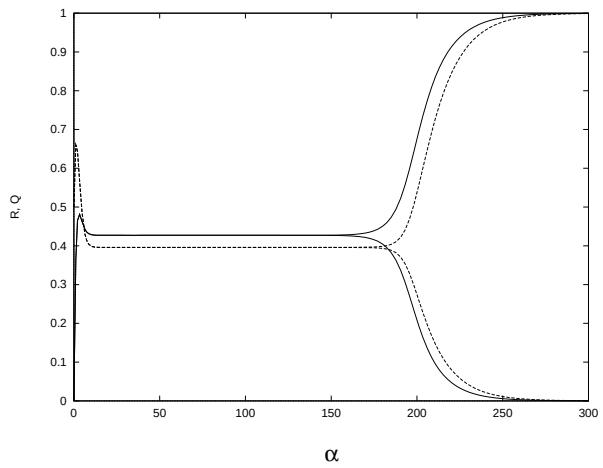
branch symmetric subspace:

$R_{ii} = R$ for $i = 1, 2$ $\quad R_{ij} = S$ for $i \neq j$

$Q_{ii} = Q$ for $i = 1, 2$ $\quad Q_{12} = C$

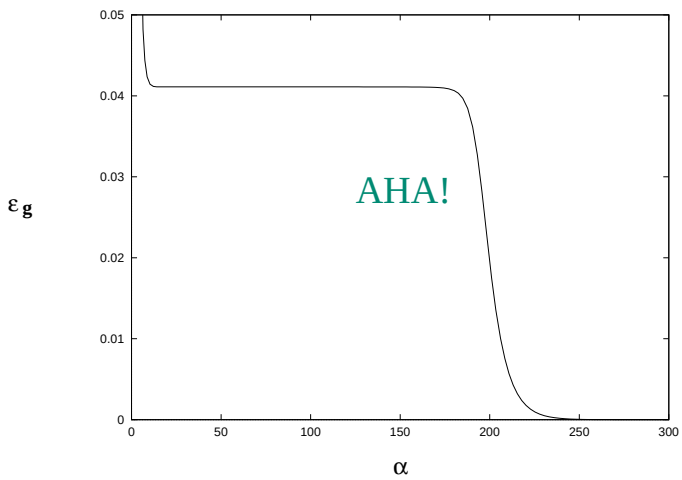
branch symmetry
is preserved under the dynamics if satisfied initially
is approached from quite general initial conditions

generic learning curves:



R_{ii}, Q_{ii}

R_{ij}, Q_{ij}



$\varepsilon_g \propto e^{-\lambda\alpha}$

Observations:

- perfect generalization is achieved
- exponential asymptotics: $\varepsilon_g \propto e^{-\lambda\alpha}$ for finite learning rates $\eta < \eta_c$
- learning process is dominated by quasi–stationary **plateau states** with $R_{ij}, Q_{ij}, \varepsilon_g \approx$ const.

Asymptotics

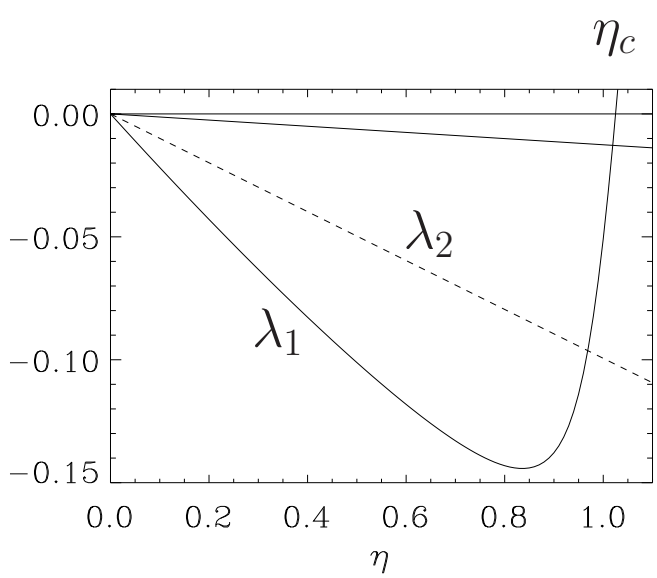
here: $R \to 1, S \to 0, Q \to 1, C \to 0$
def. $\mathbf{V}^{(4)} = (R-1, S, Q-1, C)^\top$

The dynamics equations can be linearized for $\mathbf{V}^{(4)} \to 0$, one obtains the form

$$\frac{d\mathbf{V}^{(4)}}{d\alpha} = \mathbf{m}^{(4)}\,\mathbf{V}^{(4)}$$

where the matrix $\mathbf{m}^{(4)}$ is given by derivatives of the eqs. of motion w.r.t. the order parameters (in the limit $\mathbf{V}^{(4)} \to 0$)
The decrease of the components of $\mathbf{V}^{(4)}$ is $\propto e^{-\lambda\alpha}$, where λ is the largest eigenvalue of $\mathbf{m}^{(4)}$

the two relevant eigenvalues:



successful learning only for subcritical learning rate $\eta \leq \eta_c$
(solid straight line: λ_2 for the system with adjustable w_i)

Plateau states

simplest (and most important) plateaus correspond to symmetric, **unspecialized** student configurations

$$\mathbf{J}_i \approx \mathbf{J} \propto \sum_{n=1}^{K} B_n \quad \text{for all} \quad i = 1, 2 \ldots, K$$

exact equality $\mathbf{J}_i = \mathbf{J}$ would be preserved by the learning algorithm, independent of the actual scenario!

As a consequence, a fixed point of the form

$$R_{in} = R = S, \; Q_{ik} = Q = C$$

always exists

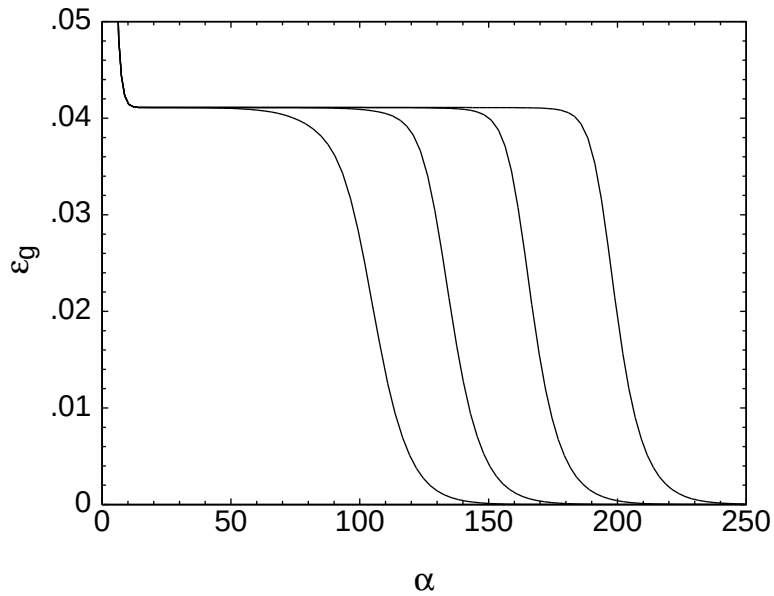
weak repulsion from the fixed point (small enough η) eventually allows for **student specialization**

(additional, less symmetric plateaus can be present!)

Plateau length

determined by
- properties of the fixed point, i.e. escape time $\tau_{esc} \propto \lambda_{esc}^{-1}$ (characteristic eigenvalue of linearization)
- initial deviation from the symmetry of the fixed point

$K = M = 2, \quad \eta = 1.5$



$$R_{11}(0) = R_{22}(0) = R_0 \approx 0$$
$$R_{12}(0) = R_0 + X_R$$
$$R_{21}(0) = R_0 - X_R$$

$$X_R = 10^{-6},\ 10^{-8},\ 10^{-10},\ 10^{-12}$$

deviation $X(\alpha) \sim |R_{12} - R_{11}| \propto e^{\tau_{esc}\alpha}$

plateau length for different values of X_R :

$$\left(\alpha_p(X_R^{(1)} - \alpha_p(X_R^{(2)}\right) = \tau_{esc} \ln \left(\frac{X_R^{(2)}}{X_R^{(1)}}\right)$$

What does this imply for realistic situations ?

(with no control of initial student/teacher overlaps)

Realistic initial conditions

e.g. random, uncorrelated students $\mathbf{J}_i(0)$

typical overlaps:

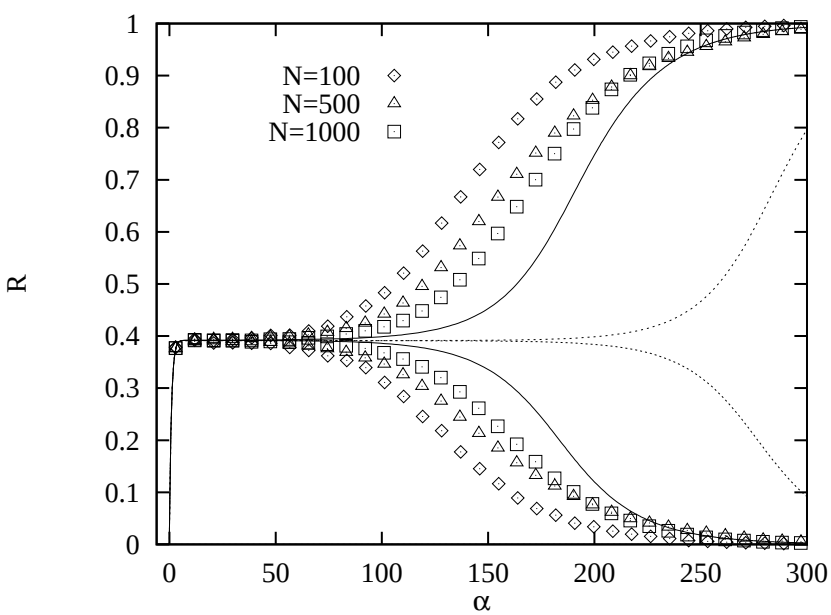
$$Q_{ii}(0) = \widetilde{Q} \qquad Q_{ik}(0) = \mathcal{O}\left(\frac{1}{\sqrt{N}}\right)$$

$$R_{in} == \mathcal{O}\left(\frac{1}{\sqrt{N}}\right)$$

expect also deviations $X_R \sim \mid R_{in} - R_{ii} \mid = \mathcal{O}\left(\frac{1}{\sqrt{N}}\right)$

plateau lengths $\alpha_p \propto \ln(N)$

diverges in the thermodynamic limit!



$K = 2, M = 1, \eta = 1$

— $\widetilde{Q} = 0.1$

$\cdots$ $\widetilde{Q} = 0.01$

perfect learning is a finite size effect

Further topics

- overlearnable/unlearnable scenarios
- adaptive second layer weights
- learning from noisy examples
- second order methods, *natural gradient descent*
- globally/locally optimal learning algorithms
- finite–size effects
- on–line learning from limited training sets
- relation on–line / off–line learning

“master references:”

D. Saad (ed.), *On–line learning in Neural Networks*, 1998

covers many of these topics and gives up to date refs.

References (very incomplete and not quite up to date)

S. Amari, *A theory of adaptive pattern classifiers*, IEEE Trans. **EC–16** (1967), 299

S. Amari, *Backpropagation and stochastic gradient descent method*, Neurocomputing **5**, 185

J.K.Anlauf and M.Biehl,*The AdaTron: an adaptive perceptron algorithm*, Europhys.Lett. **10** (1990) 687

N. Barkai, H.S. Seung, and H. Sompolinsky, *On–line learning of Dichotomies*, in: G. Tesauro et al. (eds.), *Advances in Neural Information Processing Systems 7*, Morgan Kaufmann (San Francisco) 1995

N. Barkai, H.S. Seung and H. Sompolinsky, *Scaling Laws in Learning of Classification Tasks*, Phys.Rev.Lett. **70** (1993) 3167

E.B. Baum, *The Perceptron Algorithm is Fast for Nonmalicious Distributions*, Neural. Comp. **2** (1990) 248

S. Becker, *Unsupervised Learning Procedures for Neural Networks*, Int.J. of Neural Systems, Vol. 2, Nos. 1&2 (1991), 17

M. Biehl, *An exactly solvable model of unsupervised learning*, Europhys.Lett. **25** (1994) 391

M. Biehl, A. Freking, G. Reents, *Dynamics of on–line competitive learning*, Europhys.Lett. **38** (1997) 73

M. Biehl and P. Riegler, *On–line learning with a perceptron*, Europhys.Lett. **25** (1994) 391

M. Biehl, P. Riegler, and C. Wöhler, *Transient dynamics of on–line learning in two–layered neural networks*, J.Phys. **A29** (1996) 4769

M. Biehl and E. Schlösser, *Dynamics of unsupervised principal component analysis*, J.Phys. **A31** (1998) L97

M. Biehl, P. Riegler, and M. Stechert, *Learning from noisy data: an exactly solvable model*, Phys.Rev. **E52** (1995) R4624

M. Biehl and H. Schwarze, *On–line learning of a time–dependent rule*, Europhys.Lett. **20** (1992) 733

M. Biehl and H. Schwarze, *Learning drifting concepts with neural networks*, J.Phys. **A26** (1993) 2651

M. Biehl and H. Schwarze, *Learning by online gradient descent*, J.Phys. **A28** (1995) 643

C. Bishop, *Neural networks for pattern recognition*, Oxford University Press (Oxford, UK) 1995

M. Copelli, R. Eichhorn, O. Kinouchi, M. Biehl, R. Simonetti, P. Riegler, and N. Caticha, *Noise Robustness in Multilayer Neural Networks*, Europhys.Lett. **37** (1997) 432

M. Copelli and N. Caticha,*On–line learning in the committee machine*, J.Phys. **A28** (1995) 1615

M. Copelli, O. Kinouchi, and N. Caticha, *Equivalence between learning in noisy perceptrons and tree committee machines*, Phys.Rev. **E53** (1996) 6341

Y. Chauvin and D.E. Rumelhart (eds.), *Backpropagation: Theory, Architectures, and Applications*, Lawrence Erlbaum (Hillsdale, NJ), 1995

G. Cybenko, *Approximation by Superpositions of a Sigmoidal Function*, Math. of Control, Signals, and Systems **2** (1989) 303

E. Gardner, *The Space of Interactions in Neural Network Models*, J.Phys. **A21** (1988) 257

E. Gardner and B. Derrida, *Optimal storage properties of neural network models*, J.Phys. **A21** (1988) 271

G. Györgyi and N. Tishby, *Statistical Theory of Learning a Rule*, in: *Neural Networks and Spin Glasses*, W.K. Theumann and R. Köberle (eds.), World Scientific (Singapore), 1990

G. Györgyi, *Inference of a Rule by a Neural Network with Thermal Noise*, Phys.Rev.Lett. **64** (1990) 2957

L.K. Hansen, R. Pathria and P. Salomon, *Stochastic Dynamics of Supervised Learning*, J.Phys. **A26** (1993) 63

D.O. Hebb: *The Organisation of Behavior*, Wiley (New York), 1949

J.A. Hertz, A. Krogh and R.G. Palmer, *Introduction to the Theory of Neural Computation*, Addison–Wesley (Redwood City, CA), 1991

T.M. Heskes and B. Kappen, *Learning Processes in neural networks*, Phys.Rev. **A44** (1991) 2718

T.M. Heskes and B. Kappen, *On–line learning processes in artificial neural networks*, in *Mathematical Approaches to Neural Networks*, ed. J.G. Taylor, North Holland (Amsterdam), 1993

O. Kinouchi and N. Caticha, *Optimal Generalization in Perceptrons*, J.Phys. **A25** (1992) 6243

O. Kinouchi and N. Caticha, *Lower bounds on generalization errors for drifting rules*, J.Phys. **A26** (1993) 6161

O. Kinouchi and N. Caticha, *On–line vs. off–line learning in the linear perceptron: a comparative study* Phys.Rev. **E52** (1995) 2878

J. Kim and H. Sompolinsky, *On–line Gibbs Learning*, Vorabdruck, Hebrew University (Jerusalem), 1995

W. Kinzel and P. Rujan, *Improving a Networks Generalization Ability by Selecting Examples* Europhys.Lett. **13** (1990) 472

A. Kuh, Th. Petsche and R.L. Rivest, *Incrementally Learning Time–varying Half–planes*, in: *Neural Information Processing Systems* **4**, ed. J.E. Moody, S.J. Hanson and R.P. Lippmann, Morgan Kaufmann (San Mateo, CA), 1992

R. Meir, *Learning Algorithms, Input Distributions and Generalization*, Neural Computation **7** (1995) 144

M.L. Minsky and S. Papert, *Perceptrons*, MIT Press (Cambridge, MA), 1969 and 1988.

B. Müller, J. Reinhardt and M.T. Strickland *Neural Networks, An Introduction*, 2. Auflage, Springer (Berlin), 1995

E. Oja, *A Simplified Neuron Model as a Principal Component Analyzer* J.Math.Biol. **15** (1982) 267

M. Opper, *Online vs. Offline Learning from Random Examples: General Results*, Phys.Rev.Lett. **77** (1996) 4671

M. Opper and D. Haussler, *Generalization Performance of Bayes Optimal Classification Algorithm for Learning a Perceptron*, Phys.Rev.Lett. **66** (1991) 2677

M. Opper and D. Haussler, *Calculation of the Learning Curve of Bayes Optimal Classification Algorithm for Learning a Perceptron With Noise* , in: *Proc. of the 4th Annual Workshop on Computational Learning Theory*, eds. M.K. Warmuth and L.G. Valliant, Morgan Kaufmann (San Mateo, CA), 1991

M. Opper and W. Kinzel, *Dynamics of Learning*, in: *Models of Neural Networks*, eds. E. Domany, J.L. van Hemmen and K. Schulten, Springer (Berlin), 1991

M. Opper and W. Kinzel, *Statistical Mechanics of Generalization*, in: *Physics of Neural Networks III*, eds. E. Domany, J.L. van Hemmen and K. Schulten, Springer (Berlin), in Druck

G. Radons, H. Schuster and D. Werner, in *Parallel Processing in Neural Systems and Computers*, eds. R. Eckmiller et al., Elsevier (North Holland), 1990

G. Reents and R. Urbanczik, *Self–averaging in on–line learning*, Phys.Rev.Lett. **80** (1998) 5448

P. Riegler and M. Biehl, *On–line backpropagation in two–layered networks*, J.Phys. **A28** (1995) L507

P. Riegler, M. Biehl, S.A. Solla, C. Marangi, *On–line learning from clustered input examples*, *Proc. of the 7th Italian Workshop on Neural Networks 1995*, World Scientific (Singapore) 1996

F. Rosenblatt, *Principles of Neurodynamics*, Spartan Books (New York, NY), 1962

D. Saad (ed.), *On–line learning in neural networks*, Cambridge University Press (Cambridge, UK) 1998

D. Saad and S.A. Solla, *Exact solution for on–line learning in multilayer neural networks*, Phys.Rev.Lett. **74** (1995) 4337, and
D. Saad and S.A. Solla, *On–line learning in soft committee machines*, Phys.Rev. **E52** (1995) 4225

H.S. Seung, H. Sompolinsky and N. Tishby, *Statistical Mechanics of Learning From Examples*, Phys.Rev. **A45** (1992) 6056

F. Vallet, *The Hebb Rule for Learning Separable Boolean Functions: Learning and Generalization*, Europhys.Lett. **8** (1989) 747

T.L.H. Watkin, A. Rau and M. Biehl, *The statistical mechanics of learning a rule*, Rev.Mod.Phys. **65** (1993) 499

C. van den Broeck and P. Reimann, *Unsupervised learning by examples: on–line vs. off–line*, Phys.Rev.Lett. **76** (1996) 2188

W. Wiegerinck and T. Heskes, *On-line learning with time–correlated patterns*, Europhys.Lett. **28** (1994) 451