1 **Rapid #: -668821**

IP: 128.193.162.52

CALL #:	**QA76.87 N4**
LOCATION:	**IQU :: Main Library :: sper**
TYPE:	Article
USER JOURNAL TITLE:	Neural Computation
OCLC JOURNAL TITLE:	Neural computation.
IQU CATALOG TITLE:	Neural computation
ARTICLE TITLE:	Does extra knowledge necessarily improve generalization?
ARTICLE AUTHOR:	Barber D, Saad D
VOLUME:	8
ISSUE:	1
YEAR:	1996
PAGES:	202-214
ISSN:	0899-7667
OCLC #:	
CROSS REFERENCE ID:	166435
VERIFIED:	

BORROWER:	**ORE :: Main Library**
PATRON:	**Bulatov,Yaroslav**
PATRON ID:	
PATRON ADDRESS:	
PATRON PHONE:	
PATRON FAX:	
PATRON E-MAIL:	
PATRON DEPT:	Engineering
PATRON STATUS:	Graduate
PATRON NOTES:	

System Date/Time: 11/30/2005 4:25:50 PM MST

Communicated by David Wolpert

Does Extra Knowledge Necessarily Improve Generalization?

David Barber
David Saad
Department of Physics, University of Edinburgh, Edinburgh EH9 3JZ, UK

The generalization error is a widely used performance measure employed in the analysis of adaptive learning systems. This measure is generally critically dependent on the knowledge that the system is given about the problem it is trying to learn. In this paper we examine to what extent it is necessarily the case that an increase in the knowledge that the system has about the problem will reduce the generalization error. Using the standard definition of the generalization error, we present simple cases for which the intuitive idea of "reducivity"—that more knowledge will improve generalization—does not hold. Under a simple approximation, however, we find conditions to satisfy "reducivity." Finally, we calculate the effect of a specific constraint on the generalization error of the linear perceptron, in which the signs of the weight components are fixed. This particular restriction results in a significant improvement in generalization performance.

1 Introduction

The employment of a priori knowledge in designing a learning machine is crucial to the success of the machine's ability to generalize well. Given that knowledge affects the generalization ability, our aim in this paper is to address the following question: does more knowledge necessarily improve generalization? Intuitively, the answer to this question would seem to be "yes," depending, of course, on the definitions of knowledge and generalization. Nevertheless, this question phrases a possible desiderata, which itself could affect the design of learning machines. We formulate the problem in the language of learning from examples (see, e.g., Hertz *et al.* 1991).

A training set of input/output pairs is generated by some teacher function, and the task is to find a student function whose outputs match closely the outputs of the teacher function on the training set. Constraints on the set of possible teacher functions that generate the training set are critical in narrowing down the search for a good student. Indeed, without any constraints it is an impossible task to find a student that generalizes to unseen examples (see, e.g., Wolpert 1992). A priori assumptions

are therefore made as to
imposed on the space of
assume that the spaces of
The learning problem is t
dent space, there is a stuc
teacher on all possible in
functions by $F(\Psi)$, and a
and $\theta \in \Psi$, where the ou
particular mapping that a
θ in the parameter space
noise-free set of training
each element of the traini
to find a student $f(x, \theta)$ t
set.[1] To measure the exte
an error measure $\epsilon(\theta, \theta^0$..
represented by the param
ment of minimizing the
set, and satisfying a prior
all the information that t
we review briefly the defi
lating the original questic
specific cases, beginning
version space. In Section
ear perceptron as the func
a summary of the main r
research.

2 General Theory

2.1 The Generalizatio

forms on the p training e
$\sum_{\sigma=1}^{p} \epsilon(\theta, \theta^0, x^\sigma)$. The stu
ror with respect to the p
a priori constraints. Thi
descent, resulting in a pc
$P^{\text{pri}}(\theta) \exp(-E_{\text{tr}}/T)$ where
of the stochastic algorithi
a priori constraint on th
bution of students becom
error and satisfy the a pri

[1] Extra regularization condit
considered here.

[2] We briefly note that the ass
known about the teacher functi
1994); in this paper, however, v

Improve Generalization?

ed performance measure em-
ning systems. This measure
knowledge that the system is
arn. In this paper we examine
hat an increase in the knowl-
em will reduce the generaliza-
of the generalization error, we
ive idea of "reducivity"—that
ation—does not hold. Under
ind conditions to satisfy "re-
of a specific constraint on the
ron, in which the signs of the
ticular restriction results in a
n performance.

designing a learning machine
bility to generalize well. Given
ability, our aim in this paper
s more knowledge necessarily
nswer to this question would
on the definitions of knowl-
is question phrases a possible
sign of learning machines. We
learning from examples (see,

is generated by some teacher
function whose outputs match
on the training set. Constraints
at generate the training set are
a good student. Indeed, with-
k to find a student that gener-
rt 1992). A priori assumptions

lassachusetts Institute of Technology

are therefore made as to the form of the teacher, that is, restrictions are imposed on the space of teacher functions. Throughout this paper we assume that the spaces of the teacher and student functions are the same. The learning problem is then realizable in the sense that among the student space, there is a student that will match perfectly the output of the teacher on all possible inputs. We denote the teacher/student space of functions by $F(\Psi)$, and a particular mapping as $y = f(x, \theta)$ for $f \in F(\Psi)$ and $\theta \in \Psi$, where the output is denoted by y, and the input by x. A particular mapping that a function performs is represented by the point θ in the parameter space Ψ. We assume that a teacher θ^0 generates the noise-free set of training data $\mathcal{L} = \left\{x^\sigma, f(x^\sigma, \theta^0)\right\}$, where $\sigma = 1..p$ indexes each element of the training set $\mathcal{L}$. In the learning problem, one attempts to find a student $f(x, \theta)$ that matches the teacher $f(x, \theta^0)$ on the training set.[1] To measure the extent to which the student has learned the teacher, an error measure $\epsilon(\theta, \theta^0, x)$ is defined. The set of admissible students, represented by the parameter space $\Theta \in \Psi$, is determined by the requirement of minimizing the error measure on all examples in the training set, and satisfying a priori constraints on the student. Hence Θ expresses all the information that the student has about the teacher.[2] In Section 2 we review briefly the definition of the generalization error, before formulating the original question more rigorously. We subsequently consider specific cases, beginning with the simplest possible—a one-dimensional version space. In Section 3, we analyze higher dimensions, using the linear perceptron as the function space $F(\Psi)$. In Section 4 we conclude with a summary of the main results of the paper and an outlook on further research.

2 General Theory

2.1 The Generalization Error. To measure how well the student performs on the p training examples, the training energy is formed, $E_{\text{tr}} \propto \sum_{\sigma=1}^{p} \epsilon(\theta, \theta^0, x^\sigma)$. The student is found by minimizing the training error with respect to the parameter θ, while also adhering to additional a priori constraints. This is typically achieved by stochastic gradient descent, resulting in a posttraining distribution of students, $P(\theta \mid \mathcal{L}) \propto P^{\text{pri}}(\theta) \exp(-E_{\text{tr}}/T)$ where the temperature, T, controls the randomness of the stochastic algorithm (see, e.g., Watkin *et al.* 1993). $P^{\text{pri}}(\theta)$ is the a priori constraint on the student. In the limit of zero T, the distribution of students becomes uniform over those that have zero training error and satisfy the a priori constraints; this space of student functions

[1] Extra regularization conditions on the student, such as weight decay, will not be considered here.

[2] We briefly note that the assumption that the set of admissible functions is all that is known about the teacher function is found also in the PAC approach (see, e.g., Haussler 1994); in this paper, however, we address somewhat different issues.

is known as the version space (Watkin *et al.* 1993), which we denote by Θ.[3] In Section 3.3, we present results for nonzero T, but for the rest of the paper, zero T is implied. To find the expected error that a student makes on a random example input, termed the generalization function, we average the error over the input distribution, $P(x)$, giving $\epsilon_f(\theta, \theta^0) = \int dx P(x) \epsilon(\theta, \theta^0, x)$.[4] Hence, given the teacher, $\epsilon_f(\theta, \theta^0)$ measures the expected error that a student θ makes, given that the teacher is θ^0 and that the student is θ. As the student does not know the teacher, we assume that Θ expresses all the information that the student has about the teacher. The generalization error is then defined as the expected performance of a randomly selected student from Θ, given a randomly selected teacher from Θ,

$$\epsilon_g(\Theta) = \left\langle \epsilon_f(\theta, \theta^0) \right\rangle_{\theta \subseteq \Theta, \theta^0 \subseteq \Theta} \tag{2.1}$$

where $\langle .. \rangle_{\theta^0 \subseteq \Theta}$ and $\langle .. \rangle_{\theta \subseteq \Theta}$ represent averages over the version space Θ.[5] We write $\epsilon_g(\Theta)$ to emphasize that the generalization error is a function of the version space. Intuitively, one expects that any further restrictions or a priori assumptions, resulting in a smaller version space, must necessarily reduce the generalization error. To test this intuition, we make the following definition.

Definition. $F(\Theta')$ is an "error reduced" function space of $F(\Theta)$ if $\epsilon_g(\Theta') < \epsilon_g(\Theta)$ for $\Theta' \subset \Theta$, and we say that "reducivity" holds.

In this paper we examine which subsets Θ' of Θ are error reducing, according to the preceding definition. We mention briefly that one can also consider the generalization error for a fixed teacher, $\epsilon_g(\theta^0, \Theta) = \langle \epsilon_f(\theta, \theta^0) \rangle_{\theta \subseteq \Theta}$, and check reducivity with the teacher assumed known. We show in a later section, however, that the main results of this paper also hold for $\epsilon_g(\theta^0, \Theta)$, and concentrate accordingly on $\epsilon_g(\Theta)$.

2.2 One-Dimensional Version Space. We begin with the simplest possible case of a one-dimensional version space, assuming that it can be parameterized by a connected interval on the real line, which we write, without loss of generality, as [0,a]. Furthermore, we assume that the generalization function can be written as, $\epsilon_f(\theta, \theta^0) = g(|\theta - \theta^0|)$, for some function $g(\cdot)$.[6] $\epsilon_g(\Theta)$ is then simply $\epsilon_g(a) = \int_0^a d\theta P(\theta) \int_0^a d\theta^0 P(\theta^0) g(|\theta - \theta^0|)$,

[3] The student distribution we consider is known also as exhaustive learning (see, e.g., Schwartz *et al.* 1990).

[4] An extension to the framework of this paper is to consider the off-training-set error (see, e.g., Wolpert 1992) in which the expected error of the student is calculated for test examples not included in the training set.

[5] In this joint average of $\epsilon_f(\theta, \theta^0)$ over the version space, we assume independence of the student and the teacher: As the training set is fixed, we write $P(\theta^0, \theta \mid \mathcal{L}) = P(\theta \mid \theta^0, \mathcal{L}) P(\theta^0 \mid \mathcal{L})$. With the assumption $P(\theta \mid \theta^0, \mathcal{L}) = P(\theta \mid \mathcal{L})$, we have that θ and θ^0 are independently distributed over Θ.

[6] In this assumption as to the form of the generalization function we have in mind a larger class of error measures than the square error measure, $\epsilon(\theta, \theta^0, x) =$

where $P(\cdot)$ is the para
tion, $P(\theta) = P(\theta^0) = 1/$

$$\epsilon_g(a) = \frac{2}{a^2} \int_0^a dy \int_{(}$$

for which the requirem

$$\int_0^a dx g(x) > \frac{2}{a} \int_0^a$$

This is equivalent to

$$a\langle g \rangle_a - \frac{2}{a} \int_0^a dx x \langle g$$

where $\langle g \rangle_x$ is the aver
monotonically increasi
holds for all monotoni

Unfortunately, for l
sible to separate the d
summation over the in
$\epsilon_f(\theta, \theta^0) \neq \sum_i^n g(|\theta_i - \theta_i^0|)$,
and more complicated

In the following se
beginning with an expl
that violates the error r

3 The Linear Perceptr

For the noise free linea
dimensional real vector
variable, $y \in \Re$ (see, e.
drawn independently a
matrix gaussian distribu
$\mathbf{w}^0 \cdot \mathbf{x} / \sqrt{n}$. Similarly, the
measure is taken to be
the teacher and student
gives, $\epsilon_f(\mathbf{w}, \mathbf{w}^0) = (\mathbf{w} -$
spherical constraint on
We proceed to analyze

3.1 A Two-Dimensi
dimensional linear perce
sional sphere of radius
represents the usual sph

$1/2 \left[f(x, \theta) - f(x, \theta^0) \right]^2$, for wh
for the linear function $f(x, \theta)$

[7] $w_1 = r \cos(\phi) \sin(\theta), w_2 =$
cal normalization condition.

et *al.* 1993), which we denote
ts for nonzero T, but for the
find the expected error that a
ut, termed the generalization
nput distribution, $P(x)$, giving
the teacher, $\epsilon_f(\theta, \theta^0)$ measures
given that the teacher is θ^0 and
not know the teacher, we as-
that the student has about the
efined as the expected perfor-
Θ, given a randomly selected

(2.1)

es over the version space Θ.[5]
eralization error is a function
ts that any further restrictions
aller version space, must nec-
o test this intuition, we make

nction space of $F(\Theta)$ if $\epsilon_g(\Theta') <$
ity" holds.

ets Θ' of Θ are error reduc-
We mention briefly that one
for a fixed teacher, $\epsilon_g(\theta^0, \Theta) =$
teacher assumed known. We
nain results of this paper also
gly on $\epsilon_g(\Theta)$.

We begin with the simplest
space, assuming that it can be
the real line, which we write,
ermore, we assume that the
$_f(\theta, \theta^0) = g(|\theta - \theta^0|)$, for some
$\int_0^a d\theta P(\theta) \int_0^a d\theta^0 P(\theta^0) g(|\theta - \theta^0|)$,

lso as exhaustive learning (see, e.g.,

o consider the off-training-set error
of the student is calculated for test

space, we assume independence of
fixed, we write $P(\theta^0, \theta \mid \mathcal{L}) = P(\theta \mid$
$P(\theta \mid \mathcal{L})$, we have that θ and θ^0 are

neralization function we have in
quare error measure, $\epsilon(\theta, \theta^0, x) =$

where $P(\cdot)$ is the parameter space distribution. For a uniform distribution, $P(\theta) = P(\theta^0) = 1/a$, and we can write

$$\epsilon_g(a) = \frac{2}{a^2} \int_0^a dy \int_0^y dx g(x)$$

for which the requirement of reducivity, i.e., $d\epsilon_g(a)/da > 0$ becomes

$$\int_0^a dx g(x) > \frac{2}{a} \int_0^a dy \int_0^y dx g(x)$$

This is equivalent to

$$a\langle g \rangle_a - \frac{2}{a} \int_0^a dx x \langle g \rangle_x > 0$$

where $\langle g \rangle_x$ is the average value of $g(\cdot)$ over the interval $[0, x]$. For a monotonically increasing function, $\langle g \rangle_a > \langle g \rangle_x$ $(a > x)$, and thus reducivity holds for all monotonic increasing functions defined on the real line.

Unfortunately, for higher dimensional cases, it is not generally possible to separate the dependence of the generalization function into a summation over the individual components of the parameter vector, i.e., $\epsilon_f(\theta, \theta^0) \neq \sum_i^n g(|\theta_i - \theta_i^0|)$, where n is the dimension of the parameterization, and more complicated effects can appear.

In the following sections we concentrate on the linear perceptron, beginning with an explicit example of a two-dimensional version space that violates the error reduction property.

3 The Linear Perceptron

For the noise free linear perceptron, the inputs are represented by n-dimensional real vectors, $\mathbf{x} \in \Re^n$, and the output is a single valued real variable, $y \in \Re$ (see, e.g., Hertz *et al.* 1991). The inputs $\mathbf{x}$ are assumed drawn independently and identically from a zero mean, unit covariance matrix gaussian distribution. The teacher outputs are given by $f(\mathbf{x}, \mathbf{w}^0) = \mathbf{w}^0 \cdot \mathbf{x}/\sqrt{n}$. Similarly, the student outputs are $f(\mathbf{x}, \mathbf{w}) = \mathbf{w} \cdot \mathbf{x}/\sqrt{n}$. The error measure is taken to be proportional to the squared difference between the teacher and student outputs, $\epsilon(\mathbf{w}, \mathbf{w}^0, \mathbf{x}) = (\mathbf{w} \cdot \mathbf{x} - \mathbf{w}^0 \cdot \mathbf{x})^2/2n$, which gives, $\epsilon_f(\mathbf{w}, \mathbf{w}^0) = (\mathbf{w} - \mathbf{w}^0)^2/2n$. We also impose the additional a priori spherical constraint on both the student and teacher, $\mathbf{w} \cdot \mathbf{w} = \mathbf{w}^0 \cdot \mathbf{w}^0 = n$. We proceed to analyze this model for a specific version space.

3.1 A Two-Dimensional Version Space. Let us consider the three-dimensional linear perceptron. A point on the surface of a three-dimensional sphere of radius $r = \sqrt{3}$ is given by the ordered pair (ϕ, θ), which represents the usual spherical polar coordinate parameterization.[7]

$1/2\,[f(x, \theta) - f(x, \theta^0)]^2$, for which the assumption $\epsilon_f(\theta, \theta^0) = g(|\theta - \theta^0|)$ would hold only for the linear function $f(x, \theta) = x\theta$ and $g(s) = s^2$.

[7] $w_1 = r\cos(\phi)\sin(\theta), w_2 = r\sin(\phi)\sin(\theta), w_3 = r\cos(\theta)$ where, $r = \sqrt{3}$ for the spherical normalization condition.

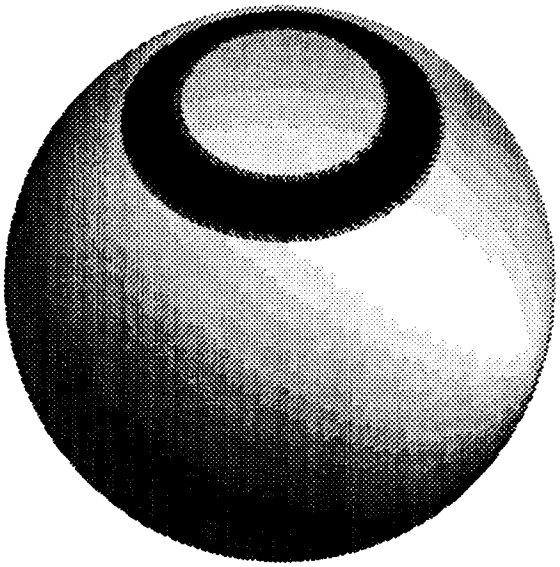


Figure 1: A sphere of radius $\sqrt{3}$. The shaded region represents the version space, $\Theta = \{\theta \in [0.4, 0.6], o \in [0, 2\pi]\}$. Making Θ smaller by pushing the inner boundary toward the outer boundary does not result in a reduction in generalization error.

The generalization function is $\epsilon_f(\mathbf{w}, \mathbf{w}^0) = 1 - \mathbf{w}\cdot\mathbf{w}^0/3$. We write this expression in spherical coordinates and average over the version space given by $\Theta = \{(o, \theta), o \in [a, b], \theta \in [c, d]\}$, to give

$$\epsilon_g(\Theta) = 1 - \frac{1}{(d-c)^2}\left\{\lambda\left[\cos(d) - \cos(c)\right]^2 + \left[\sin(d) - \sin(c)\right]^2\right\}$$

where $\lambda = 2\left[1 - \cos(b-a)\right]/(b-a)^2$. To violate reducivity we look for regions such that when we reduce the width of, for example, the interval $[c, d]$, the generalization error increases. Without loss of generality, we search for regions for which $\partial\epsilon_g(\Theta)/\partial c > 0$, and we plot one such region in Figure 1. To find such a region explicitly, we look for the boundary

at which $\partial\epsilon_g(\Theta)/\partial c = 0$, given by the equation

$$\Lambda = \frac{\sin c - \sin d}{\cos d - \cos c}\left\{\right.$$

In Figure 2a, we sho varies between 0 and 1, the value of λ; thus in re For $\lambda > \Lambda$, $\partial\epsilon_g(\Theta)/\partial c <$ (2) and (3) $\Lambda \notin [0, 1]$ an fixed, independent of [guaranteed. In region becomes increasingly n $\partial\epsilon_g(\Theta)/\partial c$ becomes less (3) is given by the solu the simplicity of the ex is nontrivial.

At this point, the re be guaranteed for conv surprisingly, we demor sufficient condition for

3.2 Euclidean Appr

we concentrate on versi be considered Euclidea corresponds to a region hypersphere appears " where $\mathbf{c}$ lies in the spac

$$\epsilon_g(\tilde{\Theta}) = \frac{1}{2n}\left\langle\left(\tilde{\mathbf{w}} -\right.\right.$$

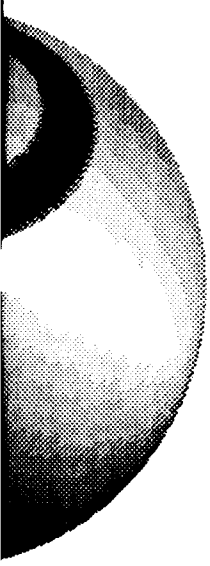
where $\tilde{\Theta}$ is the approxi uncorrelated, this can b

$$\epsilon_g(\tilde{\Theta}) = \frac{1}{n}\left(\left\langle\tilde{\mathbf{w}}^2\right\rangle_{\tilde{\mathbf{w}}}\right.$$

We now consider a For a uniform distribut can write, with a slight

$$\epsilon_g(\tilde{\Theta}') - \epsilon_g(\tilde{\Theta}) \approx -$$

[8] In general, a region is c within the region itself.



ed region represents the version
Θ smaller by pushing the inner
t result in a reduction in general-

$= 1 - \mathbf{w}\cdot\mathbf{w}^0$ 3. We write this
verage over the version space
give

$)]^2 + \{\sin(d) - \sin(c)\}^2\}$

violate reducivity we look for
h of, for example, the interval
Without loss of generality, we
and we plot one such region
ly, we look for the boundary

at which $\partial\epsilon_g(\Theta)/\partial c = 0$, and define $\Lambda(c,d) = \lambda[\partial\epsilon_g(\Theta)/\partial c = 0]$, which is given by the equation

$$\Lambda = \frac{\sin c - \sin d}{\cos d - \cos c}\left\{\frac{\sin c - \sin d + (d-c)\cos c}{\cos d - \cos c + (d-c)\sin c}\right\}$$

In Figure 2a, we show how this relates to reducivity. In region (1), Λ varies between 0 and 1, and $\partial\epsilon_g(\Theta)/\partial c$ can be of either sign, depending on the value of λ; thus in region (1), reducivity depends critically on $\delta = b-a$. For $\lambda > \Lambda$, $\partial\epsilon_g(\Theta)/\partial c < 0$, and for $\lambda < \Lambda$, $\partial\epsilon_g(\Theta)/\partial c > 0$. In both regions (2) and (3) $\Lambda \notin [0,1]$ and, as $\lambda \in [0,1]$ (Fig. 2b), the sign of $\partial\epsilon_g(\Theta)/\partial c$ is fixed, independent of $[a,b]$. In fact, in regions (2) and (3), reducivity is guaranteed. In region (2), as δ decreases (i.e., $[a,b]$ shrinks), $\partial\epsilon_g(\Theta)/\partial c$ becomes increasingly negative, whereas in region (3), for decreasing δ, $\partial\epsilon_g(\Theta)/\partial c$ becomes less negative. The boundary between regions (2) and (3) is given by the solution of $\cos d - \cos c - (d-c)\sin c = 0$. Despite the simplicity of the example, the behavior of reducivity on the sphere is nontrivial.

At this point, the reader may well conjecture that reducivity would be guaranteed for convex regions Θ and $\Theta' \subset \Theta$.[8] Perhaps somewhat surprisingly, we demonstrate in the next section that convexity is not a sufficient condition for reducivity.

3.2 Euclidean Approximation To The Version Space. For simplicity, we concentrate on version spaces small enough such that the region can be considered Euclidean. For the linear perceptron described above, this corresponds to a region small enough such that the curved surface of the hypersphere appears "flat." By writing $\mathbf{w} = \mathbf{c} + \tilde{\mathbf{w}}$, and $\mathbf{w}^0 = \mathbf{c} + \tilde{\mathbf{w}}^0$, where $\mathbf{c}$ lies in the space Θ, we have $\epsilon_t = \left(\tilde{\mathbf{w}} - \tilde{\mathbf{w}}^0\right)^2/2n$, and

$$\epsilon_g(\tilde{\Theta}) = \frac{1}{2n}\left\langle\left(\tilde{\mathbf{w}} - \tilde{\mathbf{w}}^0\right)^2\right\rangle_{\tilde{\mathbf{w}},\tilde{\mathbf{w}}^0\in\tilde{\Theta}}$$

where $\tilde{\Theta}$ is the approximately flat region on the sphere. As $\tilde{\mathbf{w}}$ and $\tilde{\mathbf{w}}^0$ are uncorrelated, this can be written in the form,

$$\epsilon_g(\tilde{\Theta}) = \frac{1}{n}\left(\left\langle\tilde{\mathbf{w}}^2\right\rangle_{\tilde{\mathbf{w}}\in\tilde{\Theta}} - \left\langle\tilde{\mathbf{w}}\right\rangle^2_{\tilde{\mathbf{w}}\in\tilde{\Theta}}\right)$$

We now consider an infinitesimal decrease in the space $\tilde{\Theta}' = \tilde{\Theta} - \Delta$. For a uniform distribution over the space, and ignoring terms in Δ^2, we can write, with a slight abuse of notation,

$$\epsilon_g(\tilde{\Theta}') - \epsilon_g(\tilde{\Theta}) \approx \frac{\Delta}{n\tilde{\Theta}}\left(\left\langle\tilde{\mathbf{w}}^2\right\rangle_{\tilde{\mathbf{w}}\in\tilde{\Theta}} - \left\langle\tilde{\mathbf{w}}^2\right\rangle_{\tilde{\mathbf{w}}\subset\Delta}\right) \tag{3.1}$$

[8] In general, a region is convex if the geodesic connecting any two points lies wholly within the region itself.

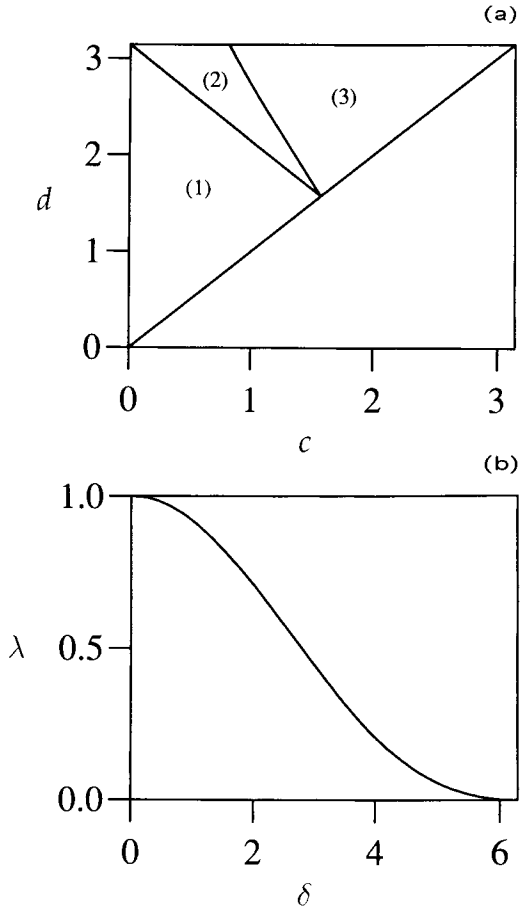


Figure 2: The version space is the region on the sphere given by $\Theta = \{(o, \theta), o \in [a, b], \theta \in [c, d]\}$. (a) In (1) reducivity depends on the region [a,b]. In (2) and (3) reducivity is guaranteed $[\partial \epsilon_g(\Theta)/\partial c < 0]$. In (2), as $[a, b]$ shrinks, $\partial \epsilon_g(\Theta)/\partial c$ becomes more negative, and vice versa in region (3). The region $c > d$ is unphysical. (b) The function λ versus $\delta = b - a$.

where Δ and $\tilde{\Theta}$ are the surface contents of Δ and $\tilde{\Theta}$, respectively. In equation 3.1, we have assumed, without loss of generality, that $\langle \tilde{\mathbf{w}} \rangle_{\tilde{\mathbf{w}} \in \tilde{\Theta}} = 0$, i.e., that the origin, **c**, is taken to be the centroid of $\tilde{\Theta}$. Reducivity holds then for the condition

$$\langle \tilde{\mathbf{w}}^2 \rangle_{\tilde{\mathbf{w}} \in \Delta} > \left\langle \tilde{\mathbf{w}}^2 \right\rangle_{\tilde{\mathbf{w}} \subset \tilde{\Theta}} \tag{3.2}$$

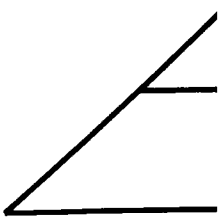


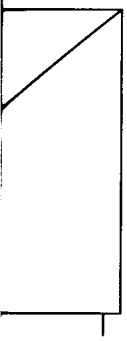
Figure 3: Counter exam
dition for reducivity. We
the position of the teacl
teacher.

Note that this is a ge
this, we can show th
not a sufficient condit
that equation 3.2 will
to the centroid, since
small. This observatio
example. Let the cor
Figure 3. By explicit c
angle $\circ = \pi/2$. We no
as shown, for which,
$\epsilon_g(\tilde{\Theta})$, demonstrating
reducivity.[9]

At this point we re
an example of a *fixed* t
edge results in an inc
example, consider a
the teacher to be pos
$\epsilon_g(\times, tri) = 1/6$. Takir
have $\epsilon_g(\times, trap) = 1/3$

[9]Note that the "distan
satisfy the triangle inequ
$n\epsilon_g(\text{trap}) = 0.32$, such tha

(a)



3

(b)

6

the sphere given by $\Theta = \{(\phi, \theta),$
depends on the region [a,b]. In
$\partial c < 0$]. In (2), as $[a, b]$ shrinks,
versa in region (3). The region
$\delta = b - a$.

of Δ and $\tilde{\Theta}$, respectively. In
ss of generality, that $\langle \tilde{\mathbf{w}} \rangle_{\tilde{\mathbf{w}} \subset \tilde{\Theta}} =$
entroid of $\tilde{\Theta}$. Reducivity holds

(3.2)

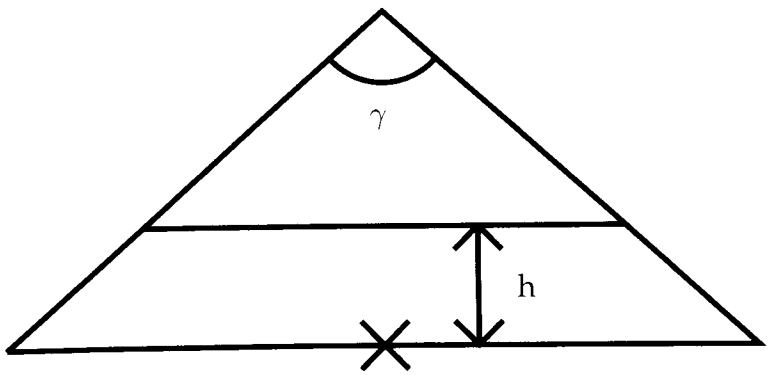


Figure 3: Counter example used to show that convexity is not a sufficient condition for reducivity. We take the hypotenuse to have length 2. The cross marks the position of the teacher for the example of reducivity violation for a given teacher.

Note that this is a general condition, holding for any dimension. Using this, we can show that convexity (for the linear perceptron at least) is not a sufficient condition for reducivity to hold. To do this, we observe that equation 3.2 will not be satisfied for regions, Δ, sufficiently close to the centroid, since then the left hand side of equation 3.2 will be small. This observation leads to the following two-dimensional counter example. Let the convex region $\tilde{\Theta}$ be the larger triangle as shown on Figure 3. By explicit calculation, one finds $n\epsilon_g(\text{tri}) = 2/9$ for the marked angle $\gamma = \pi/2$. We now take $\tilde{\Theta}'$, a convex subset of $\tilde{\Theta}$, to be the trapezium as shown, for which, in the limit $h \to 0$, $n\epsilon_g(\text{trap}) = 1/3$. Hence $\epsilon_g(\tilde{\Theta}') > \epsilon_g(\tilde{\Theta})$, demonstrating the insufficiency of convexity as a condition for reducivity.[9]

At this point we refer to Section 2.1 and note that we can readily find an example of a *fixed* teacher for which an increase in the student's knowledge results in an increase in $\epsilon_g(\theta^0, \Theta)$. In the above trapezium/triangle example, consider a very flat triangle, for which γ tends to π. We take the teacher to be positioned at the cross marked in Figure 3, for which, $\epsilon_g(\times, tri) = 1/6$. Taking again, $\tilde{\Theta}'$ to be the infinitely thin trapezium, we have $\epsilon_g(\times, trap) = 1/3$, which is larger than $\epsilon_g(\times, tri)$.

[9]Note that the "distance" measure, $\epsilon_t = (\tilde{\mathbf{w}} - \tilde{\mathbf{w}})^2/2n$ is not a metric (it does not satisfy the triangle inequality). For the metric, $\epsilon_t = |\tilde{\mathbf{w}} - \tilde{\mathbf{w}}^0|/2n$, $n\epsilon_g(\text{tri}) = 0.29$, and $n\epsilon_g(\text{trap}) = 0.32$, such that reducivity is still violated, though not as severely.

The geometry of the above situation may appear somewhat pathological. Such nonreducive situations can, however, be constructed for essentially any version space Θ. In passing, we mention another example to help clarify the situation.

For a two-dimensional ellipse with minor and major axes a and b, respectively, one readily finds $\langle \tilde{\mathbf{w}}^2 \rangle_{\text{ellipse}} = (a^2 + b^2)/4$. We see then that for a circle ($b = a$), all infinitesimal enlargements of the circle are "expansions" in the sense that they satisfy equation 3.2. For an ellipse ($b > a$) we can violate equation 3.2 by choosing the point on the perimeter about which we wish to expand to be close to the centroid ($\langle \tilde{\mathbf{w}}^2 \rangle_{\Delta} = a^2$) with $b > \sqrt{3}a$. We note that this violation of reducivity occurs for an eccentricity (b/a) that is not much larger than unity. In general, such nonexpansive enlargements can occur for the following reason: the centroid represents the best-guess student (within the euclidean approximation); adding space as close as possible to this student increases the weight on the distribution of weight space close to this best-guess, decreasing ϵ_g.

By examining equation 3.1, we note that the greatest decrease in generalization error is to be found for a region Δ farthest away from the centroid of the set. This is in line with the intuitive notion that we can improve generalization most by increasing our knowledge about the teacher in those regions that contribute most to the generalization error. One way to obtain this knowledge is to choose an input x such that the reply from the teacher yields information about the teacher in the desired region; this is the concept of query learning (see, e.g., Sollich 1994).

The previous arguments have been aimed at infinitesimal, local alterations to $\tilde{\Theta}$, and we consider briefly an example of global enlargement. We envisage situations in which the boundary of $\tilde{\Theta}$ can be expressed in a spherical coordinate system, $r = r(\phi, \theta, \ldots)$, which is the case for convex regions. The enlarged version space $\tilde{\Theta}'$ can then be defined by a new boundary, $r' = \lambda(\phi, \theta, \ldots) r(\phi, \theta, \ldots)$, for some $\lambda(\phi, \theta, \ldots) > 1$. Assuming we can bound λ by some extremum values, $\lambda_{\min} < \lambda(\phi, \theta, \ldots) < \lambda_{\max}$, it is then a simple matter to form an inequality such that the generalization error of the larger version space is greater than the generalization error of the smaller. For an enlargement $\lambda(\phi, \theta, \ldots)$ that preserves the origin as the centroid of both $\tilde{\Theta}$ and $\tilde{\Theta}'$, we require for reducivity in the two dimensional, $\lambda^2_{\min} > \lambda_{\max}$-sufficient, but by no means necessary.

3.3 Sign Constrained Weights. To now we have considered low-dimensional version spaces; here we calculate the generalization error of an infinitely large perceptron for a specific weight constraint. The sign of each weight is predetermined according to $\text{sgn}(w_i) = \rho_i$, where each ρ_i ($i = 1..n$) is either ± 1. This constraint has been studied previously in the context of pattern storage for the Hopfield network, for which it was found that the capacity was half that without the sign constraints (Amit *et al.* 1989).

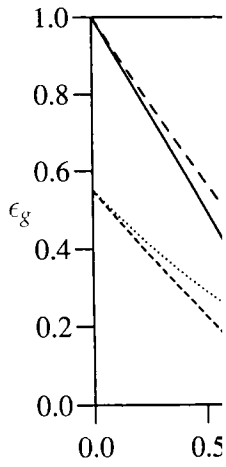


Figure 4: Comparison of the
and the spherical-sign constrai
spherical constraint; the spheri

By writing the output of
$|..|$ is the modulus, and trans
output can be written $y = \sum$
and hence symmetric, the a
that of constraining the we
the spherical constraint. Th
mechanics, following closel
This will enable us to obta
recourse to the euclidean a
required in statistical mech
dimension of the perceptron
proportional to the dimensi

A sketch of the calculatio
follows so closely that give
to that work, and point out
their analyses.

For the spherical constra
($T = 0$) reduces linearly wit
eralization error, $\epsilon_g = 1 - \alpha$,
ever, boundary effects resul
For $T = 0$ and $\alpha \geq 1$, the su
and $\epsilon_g = 0$. Nonzero T re

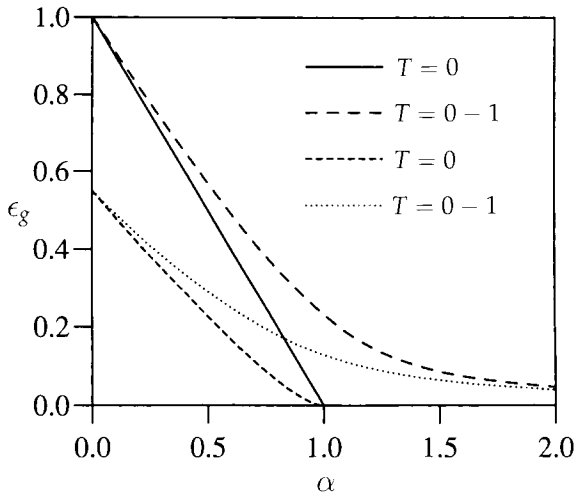


Figure 4: Comparison of the generalization error for the spherical constraint and the spherical-sign constraint. The curves beginning at 1 for $\alpha = 0$ are the spherical constraint; the spherical-sign curves begin at $1\text{-}\sqrt{2}/\pi$ for α=0.

By writing the output of the perceptron as $y = \sum x_i \mathrm{sgn}(w_i)|w_i|$, where $|..|$ is the modulus, and transforming the inputs according to, $x'_i = \rho_i x_i$, the output can be written $y = \sum x'_i|w_i|$. As the input distribution is gaussian and hence symmetric, the analysis of the sign constraint is equivalent to that of constraining the weights to be positive. In addition, we retain the spherical constraint. The method of calculation is that of statistical mechanics, following closely the exposition given in Seung *et al.* (1992). This will enable us to obtain results for any temperature, and without recourse to the euclidean approximation employed in Section 3.2. As is required in statistical mechanics calculations, we define the limit of the dimension of the perceptron such that the number of training patterns is proportional to the dimension of the perceptron, i.e., $p = \alpha n$.

A sketch of the calculation is given in the Appendix; as the calculation follows so closely that given by Seung *et al.* (1992), we refer the reader to that work, and point out only the major differences between our and their analyses.

For the spherical constraint alone, the dimension of the version space ($T = 0$) reduces linearly with α, resulting in a linear reduction of the generalization error, $\epsilon_g = 1 - \alpha$, $\alpha \leq 1$. For the spherical-sign constraint, however, boundary effects result in a small deviation from linearity (Fig. 4). For $T = 0$ and $\alpha \geq 1$, the subspace of solutions collapses to a single point and $\epsilon_g = 0$. Nonzero T results in an increase in generalization error,

affecting both the spherical and spherical-sign constraint similarly, such that for a given (α, T), $\epsilon_g^{\text{sign}} < \epsilon_g^{\text{sph}}$. For $\alpha = 0$, there is no information about the teacher other than that imposed by the a priori constraint, and we have $\epsilon_g^{\text{sph}} = 1$, and $\epsilon_g^{\text{sign}} = 1 - \sqrt{2}/\pi$.

4 Summary

We have examined the effect of constraints on the generalization error of simple learning systems, concentrating in particular on the linear perceptron. Assuming that both the student and teacher lie in the version space of constraints, we studied what effect increasing the constraint, by decreasing the version space, has on the generalization error. For a connected one-dimensional case, in which we assumed that the error function is simply a monotonically increasing function of the separation between the student and teacher, we showed that decreasing the version space necessarily decreases the generalization error. This, however, is not the case for higher dimensional version spaces, and we presented an explicit example. Furthermore, neither convexity of the version spaces, nor a metric generalization function is sufficient for the smaller version space to have lower generalization error. In general it is a nontrivial problem to predict whether reducing the version space will reduce the generalization error, and each case must be treated explicitly. We found that the generalization error of the spherical linear perceptron decreases under the additional weight component sign constraint.

Appendix

The sign constraint calculation follows closely that presented in Seung *et al.* (1992) and rather than entering into great detail, we refer the reader to Seung *et al.* (1992) and sketch here the main differences between the two calculations.

The free energy is separated into two terms, $F = G_0 - \alpha G_r$, where only the term G_0 is affected by the constraints upon the weights. We use the same notation for the order parameters as those in Seung *et al.* (1992), namely, that q is the normalized overlap between two replicas and R is the overlap between the student and the teacher. $\hat{q}$ and $\hat{R}$ are conjugate order parameters arising from the definition of the order parameters q and R. We write G_0 as,

$$G_0 = -\frac{1}{2}(1-q)\hat{q} - R\hat{R} + \frac{1}{n}\int_{-\infty}^{\infty} D\mathbf{z} \ln \int_{-\infty}^{\infty} d\mu(\mathbf{w}) \exp[\mathbf{w}\cdot(\mathbf{z}\sqrt{\hat{q}} + \mathbf{w}^0\hat{R})]$$

$D\mathbf{z}$ is the n-dimensional gaussian measure, $(2\pi)^{-n/2}\exp(-\mathbf{z}\cdot\mathbf{z}/2)d\mathbf{z}$. The

weight vector distribution

$$P(\mathbf{w}) = \frac{2^n}{V}\delta(\mathbf{w}\cdot\mathbf{w} - n)$$

and the corresponding me
face content of an n-spher
the integral representation
parameter λ) and perform
that $G_0^{\text{sign}} = G_0^{\text{sph}} + \sum_{i=1}^{n} J_i$
the free energy given by th
Seung *et al.* (1992), and J_i i

$$J_i = \sqrt{\frac{\lambda}{q\pi^2}} \sum_{i=1}^{n} \int_{-\infty}^{\infty} du \exp$$

There remains an explicit c
we average over teachers
give,

$$G_0^{\text{sign}} = G_0^{\text{sph}} + \sqrt{\frac{Q^2}{2\pi}} \int$$

where $Q = \sqrt{(\lambda+\hat{q})/(\hat{q}+$
sults necessary to find the

$$G_0^{\text{sph}} = \lambda - \frac{1}{2}(1-q)\hat{q}$$

$$G_r = \frac{1}{2}\ln[1+\hat{q}(1-q$$

The order parameters at
the free energy. The gener
$\epsilon_g = 1 - R$.

Acknowledgments

We thank Peter Sollich fo

References

Amit, D. J, Wong, K. Y. M.,
sign-constrained weight
Haussler, D. 1994. The prob
models. In *Foundations*
rowitz and S. Chipman,

weight vector distribution for the sign constraint is given by

$$P(\mathbf{w}) = \frac{2^n}{V}\delta(\mathbf{w}\cdot\mathbf{w} - n)\theta(\mathbf{w})d\mathbf{w}$$

and the corresponding measure is $d\mu(\mathbf{w}) = P(\mathbf{w})d\mathbf{w}$, where V is the surface content of an n-sphere, and $\theta(\cdot)$ is the theta function. Introducing the integral representation for the delta function (which gives rise to the parameter λ) and performing the saddle point approximation, we find that $G_0^{\text{sign}} = G_0^{\text{sph}} + \sum_{i=1}^{n} J_i\left(\lambda, q, R, \mathbf{w}^0\right)$, where G_0^{sph} is the contribution to the free energy given by the normal spherical constraint, as calculated in Seung *et al.* (1992), and J_i is

$$J_i = \sqrt{\frac{\lambda}{q\pi^2}}\int_{-\infty}^{\infty} du \exp\left\{-\frac{1}{2q}\left[4\lambda u^2 - 4\sqrt{\lambda}Rw_i^0 u + (w_i^0)^2\right]\right\}\ln\operatorname{erfc}(-u)$$

There remains an explicit dependence on the teacher weight $\mathbf{w}^0$ for which we average over teachers having the same measure as the students, to give,

$$G_0^{\text{sign}} = G_0^{\text{sph}} + \sqrt{\frac{\alpha^2}{2\pi}}\int_{-\infty}^{\infty} du \exp\left(-\alpha^2 u^2\right)\operatorname{erfc}\left(\frac{u\alpha R}{\sqrt{q}}\right)\ln\operatorname{erfc}(u)$$

where $\alpha = \sqrt{(\lambda + q)/(q + R^2)}$. For completeness, we state the further results necessary to find the free energy, namely

$$G_0^{\text{sph}} = \lambda - \frac{1}{2}(1 - q)q - R\hat{R}\frac{1}{2}q\hat{q} - R\hat{R} - \frac{1}{2}\ln(4\lambda) + \frac{\hat{R}^2 + \hat{q}}{4\lambda}$$

$$G_r = \frac{1}{2}\ln[1 + \beta(1 - q)] + \frac{\beta(q - 2R + 1)}{2[1 + \beta(1 - q)]}$$

The order parameters at $T = 1/\beta$ are found numerically by extremizing the free energy. The generalization error is then found from the relation, $\epsilon_g = 1 - R$.

Acknowledgments

We thank Peter Sollich for many stimulating discussions.

References

Amit, D. J, Wong, K. Y. M., and Campbell, C. 1989. Perceptron learning with sign-constrained weights. *J. Phys. A* **22**, 2039.

Haussler, D. 1994. The probably approximately correct (pac) and other learning models. In *Foundations of Knowledge Acquisition: Machine Learning*, A. Meyrowitz and S. Chipman, eds. Kluwer Academic Publishers, Boston.

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

Schwartz, D. B., Solla, S. A., and Samamlan, V. K. 1990. Exhaustive learning. *Neural Comp.* **2**(3), 374.

Seung, H. S., Sompolinsky, H., and Tishby, N. 1992. Statistical mechanics of learning from examples. *Phys. Rev. A* **45**, 6056–6091.

Sollich, P. 1994. Query construction, entropy and generalization in Neural Network models. *Phys. Rev. E* **49**, 4637–4651.

Watkin, T. L. H., Rau, A., and Biehl, M. 1993. The statistical mechanics of learning a rule. *Rev. Modern Phys.* **65**, 449–556.

Wolpert, D. H. 1992. On the connection between in-sample testing and generalisation error. *Complex Syst.* **6**, 47–94.

Received July 22, 1994; accepted May 31, 1995.