# TECHNICAL REPORTS

## Four Types of Learning Curves

Shun-ichi Amari, Naotake Fujita and Shigeru Shinomoto

METR 91 - 04 May 1991

MATHEMATICAL ENGINEERING SECTION

DEPARTMENT OF MATHEMATICAL ENGINEERING AND INSTRUMENTATION PHYSICS

FACULTY OF ENGINEERING, UNIVERSITY OF TOKYO

BUNKYO-KU, TOKYO, JAPAN

# Four Types of Learning Curves

**Shun-ichi Amari**
**Naotake Fujita**
Department of Mathematical Engineering and Information Physics
Univaersity of Tokyo, Tokyo 113, Japan
**Shigeru Shinomoto**
Department of Physics
Kyoto University, Kyoto 606, Japan

**Abstract**

In learning from examples, the generalization error $\epsilon(t)$ is the average probability that an incorrect decision is made by a machine trained by $t$ examples. The generalization error decreases as $t$ increases, and the curve $\epsilon(t)$ is called a learning curve. The present paper uses the Bayesian approach to show that, under the random phase approximation (annealed approximation), learning curves are classified into four asymptotic types depending on situations. When a machine is deterministic with noiseless teacher signals, 1) $\epsilon \sim at^{-1}$ when the correct machine parameter is unique in the parameter space, and 2) $\epsilon \sim at^{-2}$ when the set of the correct parameters has a finite measure. When teacher signals are noisy, 3) $\epsilon \sim at^{-1/2}$ for a deterministic machine, and 4) $\epsilon \sim c + at^{-1}$ for a stochastic machine.

## 1. Introduction

A number of approaches have so far been proposed for machine learning. A classical example is the perceptron algorithm proposed by Rosenblatt [1961], in which the convergence theorem was given. A general theory of parametric learning was proposed by Amari [1967], Rumelhart, Hinton and Williams [1986], White [1990], etc., based on the stochastic gradient descent algorithm. See, for example, Amari [1990] for this type of mathematical theories of neurocomputing.

A new framework of PAC learning was proposed by Valiant [1984], where he took both the computational complexity and stochastic evaluation of performance into account. The theory is successfully applied to neural networks by Baum and Haussler [1989], where the VC dimension of a dichotomy class plays an important role. However, the framework is too tight, and Haussler, Littlestone and Warmuth [1988] studied general convergence rate of a learning curve by removing the algorithmic complexity constraint, while Baum [1990] tried to remove the worst case constraint on the probability distribution.

A different approach is taken by Levin, Tishby and Solla [1990], where the statistical mechanical approach is unified with the Bayesian approach. See also

Schwartz, Samalkam, Solla and Denker [1990]. Here, a generalization error is defined by the probability, that a machine which has been trained by $t$ examples misclassify a novel example. The statistical average of the generalization error $\epsilon(t)$ is formulated under the Bayes formula. This theory can also be appreciated as a straightforward application of the predictive minimum description length method proposed by Rissanen [1986]. However, it is in general difficult to calculate the generalization error, and the "annealed approximation" is suggested (Levin, Tishby and Solla [1990]).

In the present paper, we discuss the generalization error along the line of the latter framework. It is not necessary, however, to use the statistical-mechanical framework nor to assume the Gibbs type probability distributions. Under the annealed or the random phase approximation, we obtained four types of scaling, showing how $\epsilon(t)$ decreases with $t$. The scaling does not depend on a specific structure of the target function, nor a specific architecture of the machine. They are universal in this sense. The scaling of a learning curve depends on underlying circumstances such as whether learning is noisy or noiseless, the behavior of a machine is deterministic or stochastic, and the correct machine parameter is unique or not.

## 2. Main Results

The problem is stated as follows. Let us consider a dichotomy of an $n$-dimensional Euclidean space $R^n$,

$$R^n = D_+ \cup D_-, \quad D_+ \cap D_- = \phi.$$

where $\mathbf{x} \in D_+$ is called a positive example and $\mathbf{x} \in D_-$ a negative example. A target signal $y$ accompanies each $\mathbf{x}$, where $y = 1$ for a positive example and $y = -1$ for a negative example. Given $t$ randomly chosen examples $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_t$ independently drawn from a probability distribution $p(\mathbf{x})$ together with corresponding target signals $y_1, \cdots, y_t$, a learning machine is required to guess the underlying dichotomy. A guessed dichotomy is evaluated by the generalization error probability $\epsilon(t)$ that the next example $\mathbf{x}_{t+1}$ produced by the same probability distribution is misclassified by the guessed dichotomy. We evaluate the average generalization error under the so-called annealed approximation, and give universal theorems on the convergence rate of $\epsilon(t)$ as $t$ tends to infinity.

More specifically, a machine considered here is specified by a set of continuous parameters $\mathbf{w} = (w_1, \cdots, w_m) \in R^m$ and it calculates a function $f(\mathbf{x}, \mathbf{w})$. The function $f(\mathbf{x}, \mathbf{w})$ specifies a dichotomy by

$$D_+ = \{\mathbf{x} \mid f(\mathbf{x}, \mathbf{w}) > 0\},$$

$$D_- = \{\mathbf{x} \mid f(\mathbf{x}, \mathbf{w}) < 0\},$$

when the output of a machine is uniquely determined by $f(\mathbf{x}, \mathbf{w})$. A machine is said to be deterministic in this case. When the output is not deterministic but its

probability is specified by a function of $f(\mathbf{x}, \mathbf{w})$, it is said to be stochastic. A neural network with modifiable synaptic weights gives a typical example of such a set of functions. For example, a layered feedforward neural network calculates a dichotomy function $f(\mathbf{x}, \mathbf{w})$, where $\mathbf{w}$ is a vector summarizing all the modifiable synaptic connection weights. The target teacher signal $y$ is said to be noiseless when $y$ is given by the sign of $f(\mathbf{x}, \mathbf{w}_0)$, while it is said to be noisy when $y$ is stochastically produced depending on the value $f(\mathbf{x}, \mathbf{w}_0)$, irrespective that the machine itself is deterministic or stochastic.

The following is the main results on the scaling of learning curves under the Bayesian framework and the annealed approximation.

Case 1. The average generalization error scales as

$$\epsilon(t) \sim \frac{m}{t},$$

when a machine is deterministic, the teacher signal is noiseless, and the machine giving correct classification is uniquely specified by the $m$-dimensional true parameter $\mathbf{w}_0$.

Case 2. The average generalization error scales as

$$\epsilon(t) \sim \frac{c}{t^2},$$

when a machine is deterministic, the teacher signal is noiseless, and the set of correct classifiers has a finite measure in the parameter space.

Case 3. The average generalization error scales as

$$\epsilon(t) \sim \frac{c}{\sqrt{t}},$$

when a machine is deterministic with a unique correct machine, but the teacher signal is noisy.

Case 4. The average generalization error scales as

$$\epsilon(t) \sim c_0 + \frac{c_1}{t},$$

when a machine is stochastic.

### 3. The average generalization error

We review here the Bayesian framework of learning along the line of Levin, Tishby and Solla [1990]. However, it is not necessary to use the statistical-mechanical framework nor to assume the Gibbs type distribution.

Let $p(y|\mathbf{x}, \mathbf{w})$ be the probability that a machine specified by $\mathbf{w}$ generates output $y$ when $\mathbf{x}$ is input. It is given by a monotone function

$$p(y=1|\mathbf{x}, \mathbf{w}) = h(f(\mathbf{x}, \mathbf{w})), \ 0 \leqq h(f) \leqq 1$$

of $f$ in the stochastic case. In the deterministic case,,

$$p(y|\mathbf{x}, \mathbf{w}) = \theta(yf(\mathbf{x}, \mathbf{w})),$$

where $\theta(z)=1$ when $z>0$ and 0 otherwise. Let $q(\mathbf{w})$ be a prior distribution of

parameter $\mathbf{w}$. Then, the joint probability density that the parameter $\mathbf{w}$ is chosen and $t$ examples of input-output pairs

$$\xi^{(t)}=[(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_t, y_t)]$$

are generated by the machine is

$$P(\mathbf{w}, \xi^{(t)})=q(\mathbf{w}) \prod_{i=1}^{t} p(y_i|\mathbf{x}_i, \mathbf{w})p(\mathbf{x}_i).$$

By using the Bayes formula, the posterior probability density of $\mathbf{w}$ is given by

$$Q(\mathbf{w}|\xi^{(t)})=\frac{P(\mathbf{w}, \xi^{(t)})}{Z(\xi^{(t)}) \prod p(\mathbf{x}_i)},$$

where

$$Z(\xi^{(t)})=\int q(\mathbf{w}) \prod_{i=1}^{t} p(y_i|\mathbf{x}_i, \mathbf{w})d\mathbf{w}$$

is the probability measure of $\mathbf{w}$'s generating $(y_1, \cdots, y_t)$ when inputs $(\mathbf{x}_1, \cdots, \mathbf{x}_t)$ are chosen.

In the deterministic case, the probability

$$Z(\xi^{(t)})=\int q(\mathbf{w}) \prod_{i=1}^{t} \theta[y_i f(\mathbf{x}_i, \mathbf{w})]d\mathbf{w}$$

is the measure of such $\mathbf{w}$ that are compatible with $t$ examples $\xi^{(t)}$, that is those $\mathbf{w}$ satisfying $yf(x_i, \mathbf{w})>0$, $i=1, \cdots, t$. Therefore, smaller this is, easier to guess the true $\mathbf{w}$. In the stochastic case, the probability $Z(\xi(t))$ can also be used as a measure of identifiability of the true $\mathbf{w}$.

The generalization error $\varepsilon_t^*$ based on $t$ examples $\xi^{(t)}$ is defined, in the deterministic case, by the probability that a machine with a randomly chosen $\mathbf{w}$ that classifies $t$ examples $\xi^{(t)}$ correctly fails to classify a new example $\mathbf{x}_{t+1}$. This is given by

$$\varepsilon_t^*=\text{Prob}\{y_{t+1}f(x_{t+1}, w)<0 \,|\, y_i f(x_i, w)>0, i=1, \cdots, t\}=1-\frac{Z_{t+1}}{Z_t}.$$

This quantity is also considered to show the generalization error in the stochastic case, because

$$\frac{Z_{t+1}}{Z_t}=\text{Prob}\{y_{t+1}|\xi^{(t)}, \mathbf{x}_{t+1}\}$$

is the predictive probability of $y_{t+1}$ given $\mathbf{x}_{t+1}$ under the condition that $t$ examples $\xi^{(t)}$ are observed.

The average generalization error $\varepsilon(t)$ is the average of $\varepsilon_t^*$ over all the possible examles $\xi^{(t)}$ and a new pair $(y_{t+1}, \mathbf{x}_{t+1})$

$$\varepsilon(t)=<\varepsilon_t^*>=1-<Z_{t+1}/Z_t>,$$

$<\ >$ denoting the expectation with respect to $\xi^{(t+1)}=(\xi(t)\ ,\ (y_{t+1}, \mathbf{x}_{t+1}))$. This quantity is closely related to the stochastic complexity $\varepsilon_t^c$ introduced by Rissanen [1986],

$$\varepsilon_t^c=-<[\ln(1-\varepsilon_t^*)]>=<\ln Z_t>-<\ln Z_{t+1}>$$

The actual evaluation of the quantity such as $<Z_{t+1}/Z_t>$ and $<\ln Z_t>$ is generally a very hard problem and has been obtained only for a few model systems (see for instance, Hansel and Sompolinsky [1990], Sompolinsky, Seung and Tishby [1990]. We show some examples later). In order to obtain a rough estimate of $\epsilon(t)$ or $\epsilon_t^c$, we introduce such approximations as,

$$<Z_{t+1}/Z_t> \sim <Z_{t+1}>/<Z_t>, \text{ and } <\ln Z_t> \sim \ln <Z_t>,$$

called the "annealed average" (Levin, Tishby and Solla [1990], see also Schwartz, Samalam, Solla and Denker [1990]). The approximations are valid if $Z_t$ does not depend sensitively on most probable $(\mathbf{x}_1, \cdots, \mathbf{x}_t)$. For this reason, we call it the random phase approximation. Validity of the approximation is still open. We will return this point in the final section. It is easy to show that the average generalization error $\epsilon(t)$ and the stochastic complexity $\epsilon_t^c$ are closely related in the asymptotic limit $t \to \infty$, and under the approximation,

$$\epsilon(t) \sim \epsilon_t^c$$

provided $\epsilon(t) \to 0$. Thus the remaining work is based on the evaluation of the average phase volume $<Z_t>$.

## 4. Case 1 : A unique correct deterministic machine with noiseless teacher

The expectation $<Z_t>$ is calculated for a deterministic machine as follows. Let $s(\mathbf{w})$ be the probability that a machine with $\mathbf{w}$ classifies a randomly chosen $\mathbf{x}$ as the true classifier with $\mathbf{w}_0$ does,

$$s(\mathbf{w}) = \text{Prob}\{f(\mathbf{x}, \mathbf{w}) \cdot f(\mathbf{x}, \mathbf{w}_0) > 0\}.$$

Then, since $y_i$ is the signum of $f(\mathbf{x}_i, \mathbf{w}_0)$,

$$<Z_t> = \text{Prob}\{y_i f(\mathbf{x}_i, \mathbf{w}) > 0,\ i = 1, \cdots, t\}$$

$$= \int q(\mathbf{w}) \text{Prob}\{y_i f(\mathbf{x}_i, \mathbf{w}) > 0,\ i = 1, \cdots, t \mid \mathbf{w}\} d\mathbf{w}$$

$$= \int q(\mathbf{w}) \{s(\mathbf{w})\}^t d\mathbf{w},$$

because $f(\mathbf{x}_i, \mathbf{w}) \cdot f(\mathbf{x}_i, \mathbf{w}_0) > 0$, $i = 1, \cdots, t$, are conditionally independent when $\mathbf{w}$ is fixed.

When $\mathbf{w}$ is slightly deviated from the true $\mathbf{w}_0$ in a unit direction $\mathbf{e}$, $|\mathbf{e}| = 1$,

$$\mathbf{w} = \mathbf{w}_0 + r\mathbf{e},$$

the regions $D_+(\mathbf{w})$ and $D_-(\mathbf{w})$ are slightly deviated from the true $D_+(\mathbf{w}_0)$ and $D_-(\mathbf{w}_0)$. The classifier with $\mathbf{w}$ misclassifies those examples that belong to $\triangle D$, which is the difference between $D_+(\mathbf{w})$ and $D_+(\mathbf{w}_0)$ or equivalently between $D_-(\mathbf{w})$ and $D_-(\mathbf{w}_0)$. Therefore, we have

$$s(\mathbf{w}) = 1 - \int_{\triangle D} p(\mathbf{x}) d\mathbf{x}.$$

We assume that the directional derivative

$$\alpha(\mathbf{e}) = \lim_{r \to 0} \frac{1}{r} \int_{\Delta D} p(\mathbf{x}) d\mathbf{x}$$

exists and is strictly positive for any direction e. This holds when the probability of x belonging to $\Delta D$ caused by a small change $\Delta$w in w is in proportion to $|\Delta \mathbf{w}|$.

We use a method similar to the saddle point approxiamtion to calculate $<Z_t>$,

$$<Z_t> = \int q(\mathbf{w}) \{s(\mathbf{w})\}^t d\mathbf{w}$$

$$= \int \exp\{t[\log s(\mathbf{w}) + \frac{1}{t} \log q(\mathbf{w})]\} d\mathbf{w}.$$

By expanding

$$\log s(\mathbf{w}) = -\alpha(\mathbf{e}) r + O(r^2).$$

and neglecting smaller order terms when $q(\mathbf{w})$ is regular, when $t$ is large,

$$<Z_t> = \int \exp\{-t\alpha(\mathbf{e}) r\} d\mathbf{w}.$$

Since the volume element $d\mathbf{w}$ is written as

$$d\mathbf{w} = r^{m-1} dr d\Omega,$$

where $d\Omega$ is angular volume element,

$$<Z_t> = \int \exp\{-t\alpha(\mathbf{e}) r\} r^{m-1} dr d\Omega$$

$$= \frac{C}{t^m},$$

where

$$C = (m-1)! \int \frac{1}{\{\alpha(\mathbf{e})\}^m} d\Omega$$

is a constant. From this, we have

$$\epsilon_t = 1 - <\frac{Z_{t+1}}{Z_t}> \doteq \frac{m}{t},$$

proving

**Theorem 1.** When $\mathbf{w}_0$ is unique, under a noiseless teacher and the annealed approximation, the generalization error rate of a deterministic machine decreases according to the universal formula

$$\epsilon(t) = \frac{m}{t},$$

where $m$ is the dimension number of w.

*Remark*: We have assumed in deriving the above result that the existence of non-zero directional drivative $\alpha(\mathbf{e})$ and a regular prior distribution $q(\mathbf{w})$ as regularity conditions. They hold in usual situations. However, it is possible to extend our result in more general cases.

When the set {w} of the correct classifiers forms an $k$-dimensional submanifold, we have

$$<Z_t> \propto t^{-(m-k)},$$

so that

$$\epsilon(t) \sim \frac{m-k}{t}.$$

In the case where the probability distribution $p(\mathbf{x})$ is densely concentrated or sparsely distributed in the neighborhood of the boundary of $D_+$ and $D_-$, we have the following expansion

$$s(\mathbf{w}) \sim 1 - a(\mathbf{e})r^{\alpha}, a > 0.$$

The result in this case is

$$\epsilon(t) \sim \frac{m}{at},$$

so that the $1/t$ law still holds.

## 5. Case 2 : Deterministic case with noiseless teacher, where a finite measure of correct classifiers exist.

In this case, $s(\mathbf{w}) = 1$ for $\mathbf{w} \in S_0$, where $S_0$ is the set of correct classifiers. We assume as a regularity condition that $S_0$ is a connected region having piecewise smooth boundary. Moreover, we assume that, when

$$\mathbf{w} = \mathbf{w}_{\omega} + r\mathbf{e}_{\omega},$$

where $\mathbf{w}_{\omega}$ is the value of $\mathbf{w}$ at position $\omega$ on $\partial S_0$ and $\mathbf{e}_{\omega}$ is the unit normal vector at $\omega$, $s(\mathbf{w})$ can be expanded as

$$s(\mathbf{w}) = \begin{cases} 1, & \mathbf{w} \in S_0 \\ 1 - a(\omega)r + O(r^2), & \mathbf{w} = \mathbf{w}_{\omega} + r\mathbf{e}_{\omega}. \end{cases}$$

The calculation of $<Z_t>$ proceeds in this case as

$$<Z_t> = \int q(\mathbf{w})\{s(\mathbf{w})\}^t d\mathbf{w}$$

$$= \int_{S_0} q(\mathbf{w}) d\mathbf{w} + \int \int q(\mathbf{w}) \exp\{-ta(\omega)r\} dr d\omega$$

$$= P_0 + \frac{C'}{t},$$

where $P_0$ is the measure of $S_0$ and

$$C' = \int_{\partial S_0} q(\mathbf{w}) \frac{1}{a(\omega)} d\omega.$$

From this follows

$$\epsilon(t) = 1 - (P_0 + \frac{C'}{t+1}) / (P_0 + \frac{C'}{t})$$

$$= \frac{B}{t^2},$$

where

$$B = \frac{C^*}{P_0}.$$

**Theorem 2.** When $S_0$ has a finite measure, $P_0 > 0$, the convergence rate $\epsilon(t)$ of a deterministic machine scales as

$$\epsilon(t) \sim \frac{B}{t^2},$$

where $B$ is a constant depending on $P_0$ and the function $f(\mathbf{x}, \mathbf{w})$.

Note that when $S_0$ tends to a point $\mathbf{w}_0$, $P_0$ tends to 0. This implies that $B$ tends to infinity, where the scaling changes as is shown in Theorem 1.

*Remark* : The above result is obtained from the annealed approximation of $<Z_{t+1} / Z_t>$. The above error probability $\epsilon(t)$ is, roughly speaking, based on the following learning scheme : Given $t$ examples $\xi^{(t)}$, choose one machine each time randomly which classifies the examples correctly. However, the scaling is exponential,

$$\epsilon(t) \sim \exp\{-ct\},$$

under the following scheme : Let a machine be randomly chosen such that it classifies $\xi^{(t)}$ correctly. Keep it when it classifies the $(t+1)$st example, otherwise choose one randomly such that it classifies the $t+1$ examples $\xi^{(t+1)}$ correctly.

## 6. Case 3 : A deterministic machine with noisy teacher

This section treats the case where the true classifier is unique and is deterministic machine with parameter $\mathbf{w}_0$ but teacher signals include stochastic error. The following is a typical example: The correct answer is 1 when $f(\mathbf{x}, \mathbf{w}_0) > 0$ and - 1 when $f(\mathbf{x}, \mathbf{w}_0) < 0$, but the teacher signal $y$ is 1 with probability $k(f(\mathbf{x}, \mathbf{w}_0))$ and is -1 with probabiality 1 - $k(f)$. A typical funcation $k$ is given by

$$k(u) = \frac{1}{1 + \exp\{-\beta u\}},$$

where $1/\beta$ is the so-called "temperature".

In this case, we cannot usually find any $\mathbf{w}$ consistent with $t$ examples $\xi^{(t)}$ when $t$ is large. We use instead a stochastic estimation $\hat{\mathbf{w}}_t$ from $t$ examples.

It is well known that the covariance matrix of the maximum likelihood estimator $\hat{\mathbf{w}}_t$ is asymptotically given by

$$E[(\mathbf{w}_t - \mathbf{w}_0)(\mathbf{w}_t - \mathbf{w}_0)'] = \frac{1}{t} G^{-1},$$

where $G$ is the Fisher information matrix. The Fisher information matrix is

explicitly given by

$$G=\beta^2\int k(1-k)\frac{\partial f}{\partial \mathbf{w}}(\frac{\partial f}{\partial \mathbf{w}})'p(\mathbf{x})d\mathbf{x},$$

where $k=k(f)$, $f=f(\mathbf{x}, \mathbf{w})$ (see Amari[1991]).

The expectation of the error probability is then given by

$$\epsilon(t)=1-<s(\mathbf{w}_t)>$$

$$=\frac{D}{\sqrt{t}},$$

where

$$D=\int a(\mathbf{e})(\mathbf{e}G^{-1}\mathbf{e}')d\Omega.$$

**Theorem 3.** When teacher signals include errors, the error rate $\epsilon(t)$ is asymptotically given by

$$\epsilon(t)\sim\frac{D}{\sqrt{t}}.$$

Here, the error probability is evaluated under the noiseless performance. When the temperature $\beta^{-1}$ tends to 0, the teacher becomes noiseless. It should be noted that the Fisher information $G$ tends to infinity in proportion to $\beta^2$. Hence, $D$ tends to 0 in this limit, and the scaling changes as is in Theorem 1.

## 7. Case 4 : Stochastic machine

In the case of a stochastic machine, teacher signals are also stochastic. The error probability $\epsilon(t)$ never tends to 0 in this case, but converges to $\epsilon_0>0$.

We have

$$<Z_t>=\int q(\mathbf{w})p(\xi^{(t)}|\mathbf{w})p(\xi^{(t)}|\mathbf{w}_0)\frac{1}{\prod p(\mathbf{x}_t)}d\mathbf{w}d\xi^{(t)}$$

$$=\int \exp\{t\log s(\mathbf{w})\}d\mathbf{w},$$

where

$$s(\mathbf{w})=\int p(y|\mathbf{x}, \mathbf{w})p(y|\mathbf{x}, \mathbf{w}_0)p(\mathbf{x})d\mathbf{x}dy.$$

Since $s(\mathbf{w})$ is smooth, we have the following expansion at its maximum $\mathbf{w}_0'$,

$$s(\mathbf{w})\sim c-(\mathbf{w}-\mathbf{w}_0')K(\mathbf{w}-\mathbf{w}_0')'$$

with a constant $c$ and a positive definite matrix $K$. Hence,

$$<Z_t>\sim c't^{-\frac{m}{2}},$$

so that

$$1-\frac{<Z_{t+1}>}{<Z_t>}=\varepsilon_0+\frac{a}{t}.$$

**Therem 4.** The generalization error scales as

$$\varepsilon(t)\sim\varepsilon_0+\frac{a}{t}$$

for a stochastic machine.

## 8. Discussions

We have thus obtained four typical asymptotic scaling laws of the generalizaiton error $\varepsilon(t)$ under the random phase approximation. In order to see the validity of the approximation, we calculate the exact $\varepsilon(t)$ of the following simple example : Predicting a half space of $R^2$, where signals $\mathbf{x}=(x_1, x_2)$ are normally distributed with mean 0 and the identity covariance matrix, $w$ is a scalar having a uniform prior $q(w)$ and

$$f(\mathbf{x}, w)=x_1\cos w+x_2\sin w.$$

It is not difficult to obtain the exact $\varepsilon(t)$, and it scales asymptotically as $\varepsilon(t)\sim 2/3t$, while the random phase approximation gives $\varepsilon(t)\sim 1/t$. This shows that the approximation gives the same scaling order but a different factor. It is interesting to see how the difference depends on the number $m$ of parameters $\mathbf{w}$.

Looking from the point of view of statistical inference, the deterministic case and stochastic case are quite different. The estimator $\hat{\mathbf{w}}_t$ from $t$ examples is usually subject to a normal distribution with a covariance matrix of order $1/t$ in the stochastic case. However, in the deterministic case, $\hat{\mathbf{w}}_t$ is usually not subject to a normal distribution. The squared error usually shows a stronger convergence. This is because the manifold of probability distributions has a Riemannian structure in the stochastic case (Amari [1985]), while it has a Finslerian structure in the deterministic case (Amari [1987]).

This suggests a difference of the validity of the random phase approximation in the two cases. We believe that the random phase approximation is valid in the stochastic case. This will be reported in another paper.

## References


Amari, S. [1967] : Theory of Adaptive Pattern Classifiers, *IEEE Trans.*, EC-16, No. 3, pp.299-307.

Amari, S. [1985] : *Differential-Geometrical Methods in Statistics*, Springer L. N. in Statistics, vol. 28, Springer.

Amari, S. [1987] : Dual connections on the Hilbert Bundles of statistical models. *Geometrization of Statistical Theory*, ed. C. T. J. Dodson, ULDM Lancaster UK., pp123-152.

Amari, S. [1990] : Mathematical Foundations of Neurocomputing. *Proceedings of the IEEE*, vol. 78, No. 9, pp.1443-1463.

Amari, S. [1991] : Dualistic Geometry of the Manifold of Higher-Order Neurons. *Neural Networks*, to appear.

Baum, E. B. [1990] : The perceptron algorithm is fast for nonmalicious distributions. *Neural Computation*, vol. 2, pp.248-260.

Baum, E. B. and Haussler, D. [1989] : What size net gives valid generalization. *Neural Computation*, vol. 1, pp.151-160.

Haussler, D., Littlestone, N. and Warmuth, K. [1988] : Predicting {0, 1} Functions on Randomly Drawn Points. *Proc. COLT'88* San Mateo, CA: Morgan Kaufmann, pp.280-295.

Hansel, D. and Sompolinsky, H. [1990] : Learning from examples in a single-layer neural network. *Europhys. Lett.*, 11, pp.687-692.

Levin, E., Tishby, N. and Solla, S. A. [1990] A Statistical Approach to Learning and Generalization in Layered Neural Networks. *Proceedings of the IEEE*, vol. 78, No. 10, pp.1568-1574.

Rissanen, J. [1986] : Stochastic complexity and modeling. *Ann. Statist.*, vol. 14, pp.1080-1100.

Rosenblatt, F. [1961] : *Principles of Neurodynamics*. Spartan.

Rumelhart, D., Hinton, G. E., and Williams, R. J. [1986] : L e a r n i n g internal representations by error propagation. *Parallel Distributed Processing : Explorations in the microstructure of cognition*, vol. 1: *Foundations*. MIT Press.

Schwartz, D. B., Samalam, V. K., Solla, S. A. and Denker, J. S. [1990] : Exhaustive learning. *Neural Computation*, 2, pp.375-385.

Sompolinsky, H., Seung, S. and Tishby N. : Learning from examples in large neural networks. to be published.

Valiant, L. G. [1984] : A theory of the learnable. *Comm. ACM*, v0l. 27, No. 11, pp.1134-1142.

White, H. [1989] : Learning in artificial neural networks : A statistical perspective. *Neural Computation*, vol. 1, pp.425-464.