Vladimir Semenovich Pugachev
Optimal learning systems

Kybernetika, Vol. 7 (1971), No. 5, (347)--376

Persistent URL: http://dml.cz/dmlcz/125145

KYBERNETIKA — VOLUME 7 (1971), NUMBER 5

Optimal Learning Systems

V. S. PUGACHEV

Optimal Bayes learning systems and the main properties of their algorithms are studied from the general standpoint of Learning System Theory. The general concepts of learning system, teacher, forms of learning, types of teachers, etc. are discussed. The general case of a real teacher which performs control processes with random errors is considered. The conventional case of an ideal teacher which has been the only case previously considered, is treated as a special case. A measure of closeness of the performance of a learning system to the optimal system with complete information is introduced. Studying of optimal learning systems enables the designer of learning systems to determine extreme possibilities of learning systems and to estimate the performance of various specific algorithms of learning.

1. INTRODUCTION

A *learning system* is called such a system which improves itself using the information contained in signals received by it.

The main feature of a learning system distinguishing it from conventional automatic systems is the dependence of its output not on the actual input, only, but also on some previously obtained signals. Let Z be the input to the system, $\hat{W}$ its output, Ξ' the set of previously obtained signals (teaching information) on which the output $\hat{W}$ depends. Then the input-output relation of the learning system has the general form

$$\hat{W} = A(\Xi')\, Z\,, \tag{1.1}$$

where $A(\Xi')$ is some operator, deterministic or random, depending on teaching information Ξ'. For usual automatic systems the operator A is completely independent of previously obtained signals.

It should be emphasized that the notion of learning system convers only systems improving themselves *while performing their functions.* All the information introduced into the system during its designing before it begins performing its functions can not be considered as teaching information. This information can only be treated as characterizing the initial organization of the system, its *capacities* to learn.

The period of time when the system receives the teaching signals Ξ' on the basis of which its operator is formed, is called *the period of learning*. The period of learning may, in particular, include all the time of system working, i.e. the system may be permanently learning during all its life-time.

If the system does not receive additional external information apart its usual inputs during the period of learning, it is called *self-learning system.*

If the system receives some additional information as well as its usual inputs while learning, we say that it is *taught by a teacher*. A source of the additional information, being a man-operator or another automatic system, is called *teacher.*

The teacher may teach a system in two different ways. The first way is to show the system how to perform its functions. This way we shall call *teaching by show.* The second way is to observe the actions of the system and introduce into it teacher's estimates of its performance. We shall call this mode of teaching *teaching by estimating system actions.*

These three main ways of teaching may be combined sequentially or in parallel. For example, the system may be taught at first by show, then by estimating its actions, and then may continue improving its performance by self-learning. The system with several outputs may be taught simultaneously by show at some of its outputs, by estimating its actions at other outputs and may be self-learning with respect to the remaining outputs.

The teacher may know exactly the desired output W corresponding to a given input Z and show this desired output to the system. We shall call such a teacher *the ideal teacher.* Only the ideal teacher has been considered in all previous works in the field of learning systems. But the teacher does not in general know the exact desired output W and can only elaborate more or less suitable estimates of W. For example, the instructor teaching a man to pilot an aeroplane can not pilot this plane with absolute accuracy. He will inevitably make random errors resulting in random deviations of the plane from the required state of motion. In other words, he will show to the pupil only his own output, but can not show the actual desired output.

We shall call a teacher whose output $\tilde{W}$ does not coincide with the desired output W *the real teacher.* The notion of the real teacher seems to be primarily introduced in the works of the author [1–3].

The algorithm of any real teacher is practically always stochastic, i.e. the output of the real teacher is random for any given input. So the algorithm of the real teacher can be defined only in the form of the conditional probability distribution of teacher output $\tilde{W}$ given the input Z. We shall call this conditional distribution *decision function* of the teacher.

The necessity of teaching a system arises only if the probability distribution of the input and desired output or some parameters of this distribution are unknown. E.g., the distribution of feedback signals is unknown, if the dynamical characteristics of the controlled system are unknown. But the feedback signals represent some of the

components of the input to the control system. So the control system must learn to master performing the functions for which it is designed in this case.

We shall call *algorithm of learning* the mode of treating teaching signals to improve system performance. Any algorithm of learning which provides the increase of system performance is admissible in principle.

Up to now only heuristic algorithms of learning based on designer's skill and intuition were practically employed in learning system design.

The most natural approach to design algorithms of learning seems to be statistical estimation of unknown probability distribution of signals involved. This approach was initiated by Špaček and developed by his pupils [4–18] who studied some general properties of learning processes. In particular, the methods of stochastic approximations were extended and first used to derive the algorithms of learning [6, 8, 10–13]. The use of the methods of stochastic approximations to find algorithms of self-adaptive systems was also first proposed [15–17] (see also [19]). Following Špaček ideas the author proposed in [20] one possible statistical algorithm of self-learning for systems based on optimal Bayes decisions. Some special statistical problems related to self-learning were solved in [21–25]. The algorithms of learning for pattern recognition systems based on stochastic approximation methods were first derived for the case of the ideal teacher in [26] (see also [27]).

We shall derive here the general optimal algorithms of learning using Bayes statistical decision approach.* The system using such an optimal algorithm of learning will be called *Bayes optimal learning system*. This system possesses the best possible performance among all the learning systems receiving the same teaching information.

The study of Bayes optimal learning systems enables us to establish the extreme performance of learning systems, i.e. *potential learnability* of learning systems. Comparing the performance of any projected learning system with the performance of the Bayes optimal learning system one can estimate the algorithm of learning employed in the designed system.

We use here the terminology of automatic control theory, since the problems of teaching automatic systems are of special interest for us. Alternately, using the terminology accepted in studies of biological aspects of learning, Z represents the stimulus, W the desired reaction, $\hat{W}$ the actual reaction of a learning system, $\hat{\bar{W}}$ the consequence caused by system reaction $\hat{W}$ to the stimulus Z.

2. BAYES OPTIMAL LEARNING SYSTEMS

We shall now give a precise formulation of the problem of finding a Bayes optimal learning system.

We suppose that all possible values z of the input Z represent elements of some

* This approach seems to be initiated by M. E. Shaikin [28] who solved one special problem of learning of pattern recognition systems.

set A, and all possible values w of the desired output W represent the elements of another set B. Let $\gamma(\Delta z; \theta \mid \lambda)$ be the family of probability measures on A, and $\varkappa(\Delta w; \theta \mid z, \lambda)$ the family of conditional probability measures on B given the value z of Z, λ being a parameter with values in a set L, and θ a parameter whose values are real numbers*. Suppose that the input Z and desired output W represent random variables in A and B respectively distributed in accordance with the probability measure $\gamma(\Delta z; \theta \mid \lambda)$ and conditional probability measure $\varkappa(\Delta w; \theta \mid z, \lambda)$ corresponding to some specific value of λ equal to the value $\lambda(\theta)$ of a random function $\Lambda(\theta)$ with values in L, the parameter θ taking different values in different cycles of the actions of the system (it may represent, in particular, the number of the cycle of system action).

If the true value of λ were known and introduced into the system at each cycle of its action, the system could make various statistical decisions, resulting in respective estimates of W. In particular, the system could evaluate the Bayes optimal estimate W_λ^* of W corresponding to any given loss function $l(W, \hat{W} \mid \lambda)$ which may, in general, depend on the parameter λ. We shall call the system whose output at each cycle represents the Bayes optimal estimate W_λ^* of the desired output W corresponding to the true value of λ *Bayes optimal system with complete information about* λ.

If the true value of λ is unknown, the necessary estimates can be made, in principle, on the basis of some observations. Thus the system must *learn* to elaborate estimates of the desired output by making some observations and processing the information thus obtained, before it will begin performing the functions for which it is designed. We shall suppose that the learning period during which the system makes observations consists of N cycles of system action corresponding to values $\theta_1, \ldots, \theta_N$ of the numerical parameter θ. The value of θ at the first cycle following the period of learning we shall leave without any indices.

If the system observes while learning only the inputs $Z_1, \ldots, Z_N$, it is *self-learning*. If it receives, apart from the inputs $Z_1, \ldots, Z_N$, some other signals, it is *learning with a teacher*. The teacher may show to the system estimates $\tilde{W}_1, \ldots, \tilde{W}_N$ of the desired outputs $W_1, \ldots, W_N$ corresponding to the inputs $Z_1, \ldots, Z_N$ or estimates of some functionals of the pairs $(Z_1, W_1), \ldots, (Z_N, W_N)$. This is the case of *teaching by show*. If the teacher observes the responses $\hat{W}_1, \ldots, \hat{W}_N$ of the system corresponding to the inputs $Z_1, \ldots, Z_N$ and shows to the system estimates $\tilde{W}_1, \ldots, \tilde{W}_N$ of its performance at each cycle, the system is *taught by estimating its actions*.

Thus the teacher output $\tilde{W}$ may have, in general, another nature than the desired and actual system outputs $W, \hat{W}$. Accordingly we shall suppose that all possible values $\tilde{w}$ of the teacher output $\tilde{W}$ represent elements of a set $\tilde{B}$ which may, certainly,

* Speaking about measures defined on a set, we mean, certainly, measures defined on properly determined σ-algebra of subsets of this set. This will be assumed throughout the paper without further mentioning. We denote random variables by capital letters and their possible values (samples) by respective small letters; the sets of elements $z, w, \ldots$ we denote $\Delta z, \Delta w, \ldots$, so Δz represents any subset of the set A, Δw any subset of the set B, etc.

coincide with B in the special case of teaching by show. The algorithm of the teacher is determined in general by the conditional probability measure $\delta_T(\Delta\tilde{w}; \theta \mid z, w, \hat{w}, \lambda)$ on $\tilde{B}$ given the values $z, w, \hat{w}$ of the input Z, desired output W, and actual system output $\hat{W}$, λ being equal to the same value $\lambda(\theta)$ of the random function $\Lambda(\theta)$ as before. In the special case of the ideal teacher δ_T is condensed in single point w and does not depend on $z, \hat{w}, \lambda$. In the case of the real teacher teaching the system by show δ_T is independent of $w, \hat{w}$. In the case of the real teacher teaching the system by estimating its actions δ_T is independent of w.

We shall denote for brevity by Ξ the set of all signals received by the learning system including all the signals received during the learning period and the input Z received at the first cycle after learning.

In the case of self-learning Ξ represents the set of teaching inputs $Z_1, \ldots, Z_N$ followed by the input Z to which the system must respond in an optimal way. The probability measure of Ξ in the Cartesian product $X = A^{N+1}$ is determined in this case by

$$\sigma(\Delta\xi \mid \lambda, \bar{\lambda}) = \gamma(\Delta z; \theta \mid \lambda) \prod_{i=1}^{N} \gamma(\Delta z_i; \theta_i \mid \lambda_i), \tag{2.1}$$

$\bar{\lambda}$ being the set of values $\lambda_1, \ldots, \lambda_N$ of the random function $\Lambda(\theta)$ corresponding to the values $\theta_1, \ldots, \theta_N$ of θ. In deriving (2.1) the assumption was made that $Z_1, \ldots, Z_N, Z$ are conditionally independent, i.e. independent for any given set of values of $\lambda_1, \ldots, \lambda_N, \lambda$. The measure $\sigma(\Delta\xi \mid \lambda, \bar{\lambda})$ represents the conditional probability measure of Ξ given the values $\lambda_1, \ldots, \lambda_N, \lambda$ of $\Lambda(\theta_1), \ldots, \Lambda(\theta_N), \Lambda(\theta)$.

In the case of a system taught by show Ξ represents the set of previously obtained inputs $Z_1, \ldots, Z_N$, corresponding teacher outputs $\tilde{W}_1, \ldots, \tilde{W}_N$, and the input Z to which the system must respond optimally. The conditional probability measure of Ξ in the Cartesian product $X = A^{N+1} \times \tilde{B}^N$ is given in this case by

$$\sigma(\Delta\xi \mid \lambda, \bar{\lambda}) = \gamma(\Delta z; \theta \mid \lambda) \prod_{i=1}^{N} \int_{\Delta z_i} \pi(\Delta\tilde{w}_i; \theta_i \mid z_i, \lambda) \, \mathrm{d}\gamma(z_i; \theta_i \mid \lambda_i), \tag{2.2}$$

where

$$\pi(\Delta\tilde{w}; \theta \mid z, \lambda) = \int_B \delta_T(\Delta\tilde{w}; \theta \mid z, w, \lambda) \, \mathrm{d}\varkappa(w; \theta \mid z, \lambda), \tag{2.3}$$

and triples $(Z_1, W_1, \tilde{W}_1), \ldots, (Z_N, W_N, \tilde{W}_N)$ and Z are assumed conditionally independent as before.

In the case of a system taught by estimating its actions Ξ represents the set of previously received inputs $Z_1, \ldots, Z_N$, corresponding system outputs $\hat{W}_1, \ldots, \hat{W}_N$, teacher outputs $\tilde{W}_1, \ldots, \tilde{W}_N$, and the input Z to which the system must respond optimally. The system outputs $\hat{W}_1, \ldots, \hat{W}_N$ represent random variables in B with conditional probability measures $\delta^1(\Delta\hat{w} \mid z), \ldots, \delta^N(\Delta\hat{w} \mid z)$ independent of λ (since the values $\lambda_1, \ldots, \lambda_N, \lambda$ of the random function $\Lambda(\theta)$ in the points $\theta_1, \ldots, \theta_N, \theta$ remain

unknown to the system; otherwise learning would be of no sense). Yet each of the measures $\delta^2, \ldots, \delta^N$ may depend on previously received teaching signals, namely $\delta^i(\Delta w \mid z)$ may depend also on the values $z_1, \ldots, z_{i-1}$, $\hat{w}_1, \ldots, \hat{w}_{i-1}$, $\tilde{w}_1, \ldots, \tilde{w}_{i-1}$ of $Z_1, \ldots, Z_{i-1}$, $\hat{W}_1, \ldots, \hat{W}_{i-1}$, $\tilde{W}_1, \ldots, \tilde{W}_{i-1}$ $(i = 2, \ldots, N)$. Hence the conditional probability measure of Ξ in the Cartesian product $X = A^{N+1} \times B^N \times \tilde{B}^N$ is given in this case by

$$\sigma(\Delta\xi \mid \lambda, \bar{\lambda}) = \gamma(\Delta z; \theta \mid \lambda) \prod_{i=1}^{N} \int_{\Delta z_i} \mathrm{d}\gamma(z_i; \theta_i \mid \lambda_i) \times \tag{2.4}$$

$$\times \int_{\Delta \tilde{w}_i} \pi(\Delta \tilde{w}_i; \theta_i \mid z_i, \hat{w}_i, \lambda_i)\, \mathrm{d}\delta^i(\hat{w}_i \mid z_i)\,,$$

where

$$\pi(\Delta\tilde{w}; \theta \mid z, \hat{w}, \lambda) = \int_B \delta_T(\Delta\tilde{w}; \theta \mid z, w, \hat{w}, \lambda)\, \mathrm{d}\varkappa(w; \theta \mid z, \lambda)\,, \tag{2.5}$$

and the quadruples $(Z_1, W_1, \hat{W}_1, \tilde{W}_1), \ldots, (Z_N, W_N, \hat{W}_N, \tilde{W}_N)$ and Z are assumed conditionally independent, as before.

The system is required to elaborate the optimal estimate W^* of the desired output W at the first cycle after learning using the set Ξ of all signals received.

The average loss (i.e. the expected value of the loss function) at the first cycle after learning is determined by

$$R(\delta) = \mathsf{E}l(W, \hat{W} \mid \Lambda) = \tag{2.6}$$

$$= \int_L \mathrm{d}_\lambda \int_{\bar{L}} \mathrm{d}_{\bar{\lambda}} A(\lambda, \bar{\lambda}) \int_X \mathrm{d}\sigma(\xi \mid \lambda, \bar{\lambda}) \int_B \mathrm{d}\delta(\hat{w} \mid \xi) \int_B l(w, \hat{w} \mid \lambda)\, \mathrm{d}\varkappa(w; \theta \mid z, \lambda)\,,$$

where $A(\Delta\lambda, \Delta\bar{\lambda})$ is the joint probability measure of the random variables $\Lambda(\theta)$, $\Lambda(\theta_1), \ldots, \Lambda(\theta_N)$ in the Cartesian product $L \times \bar{L} = \bar{L}^{N+1}$, and $\delta(\Delta\hat{w} \mid \Delta)$ is *the system decision function* representing the conditional probability measure of system output $\hat{W}$ at the first cycle after learning given the value ξ of Ξ.

The problem is then to find an optimal system decision function $\delta_{\mathrm{opt}}(\Delta\hat{w} \mid \xi)$ minimizing the average loss $R(\delta)$. The system using this optimal decision function will be the required *Bayes optimal learning system.*

To solve the problem we notice that for any sets $\Delta\xi$, $\Delta\lambda$, $\Delta\bar{\lambda}$ of values of ξ, λ, $\bar{\lambda}$ respectively

$$\int_{\Delta\lambda} \mathrm{d}_\lambda \int_{\Delta\bar{\lambda}} \mathrm{d}_{\bar{\lambda}} A(\lambda, \bar{\lambda}) \int_{\Delta\xi} \mathrm{d}\sigma(\xi \mid \lambda, \bar{\lambda}) = \int_{\Delta\xi} \mathrm{d}\beta(\xi) \int_{\Delta\lambda} \mathrm{d}_\lambda \int_{\Delta\bar{\lambda}} \mathrm{d}_{\bar{\lambda}} \eta(\lambda, \bar{\lambda} \mid \xi)\,, \tag{2.7}$$

where

$$\beta(\Delta\xi) = \int_L \mathrm{d}_\lambda \int_{\bar{L}} \sigma(\Delta\xi \mid \lambda, \bar{\lambda})\, \mathrm{d}_{\bar{\lambda}} A(\lambda, \bar{\lambda}) \tag{2.8}$$

is the unconditional probability measure of Ξ, and

$$\eta(\Delta\lambda, \Delta\bar{\lambda} \mid \xi) = \int_{\Delta\lambda} \mathrm{d}_{\lambda} \int_{\Delta\bar{\lambda}} \frac{\mathrm{d}\sigma(\xi \mid \lambda, \bar{\lambda})}{\mathrm{d}\beta(\xi)} \mathrm{d}_{\bar{\lambda}} A(\lambda, \bar{\lambda}) \tag{2.9}$$

the joint conditional probability measure of $\Lambda(\theta)$, $\Lambda(\theta_1)$, ..., $\Lambda(\theta_N)$ given the value ξ of Ξ. Using (2.7) and introducing the conditional probability measure of $\Lambda(\theta)$:

$$\Omega(\Delta\lambda \mid \xi) = \eta(\Delta\lambda, \bar{L} \mid \xi) = \int_{\Delta\lambda} \mathrm{d}_{\lambda} \int_{L} \frac{\mathrm{d}\sigma(\xi \mid \lambda, \bar{\lambda})}{\mathrm{d}\beta(\xi)} \mathrm{d}_{\bar{\lambda}} A(\lambda, \bar{\lambda}) , \tag{2.10}$$

(2.6) becomes

$$R(\delta) = \int_{X} \mathrm{d}\beta(\xi) \int_{B} \varrho(\xi, \hat{w}) \, \mathrm{d}\delta(\hat{w} \mid \xi) \tag{2.11}$$

with

$$\varrho(\xi, \hat{w}) = \int_{Z} \mathrm{d}\Omega(\lambda \mid \xi) \int_{B} l(w, \hat{w} \mid \lambda) \, \mathrm{d}\varkappa(w; \theta \mid z, \lambda) . \tag{2.12}$$

If for any ξ there exists a unique w^* satisfying

$$\varrho(\xi, w^*) \leqq \varrho(\xi, \hat{w}) \quad \text{for all} \quad \hat{w} , \tag{2.13}$$

then the optimal system decision function $\delta_{\mathrm{opt}}(\Delta\hat{w} \mid \xi)$ represents, obviously, the measure condensed in single point $w^* = Q\xi$ (Q being the deterministic operator establishing the correspondence between ξ and w^* satisfying (2.13)). Thus the Bayes optimal learning system is deterministic in this case and its output w^* corresponding to the value ξ of Ξ is determined as the unique value w^* of $\hat{w}$ giving the least value to $\varrho(\xi, \hat{w})$.

If for any ξ there exists a set C_ξ of values w^* satisfying (2.13), then the optimal system decision function $\delta_{\mathrm{opt}}(\Delta\hat{w} \mid \xi)$ represents an arbitrary probability measure on C_ξ. There exists then an infinite set of Bayes optimal learning systems, some of which are deterministic and others stochastic. Each of the latter includes a random mechanisme of choice of w^* from C_ξ for any given value ξ of Ξ.

As we see from (2.12) the algorithm of learning of the Bayes optimal learning system consists of elaborating conditional probability measure $\Omega(\Delta\lambda \mid \xi)$ of $\Lambda(\theta)$ at the first cycle of system action after receiving the teaching signals. $\Omega(\Delta\lambda \mid \xi)$ represents in fact the *posterior* probability measure of $\Lambda(\theta)$ after receiving the teaching signals, while $A(\Delta\lambda, \bar{L})$ is the *prior* probability measure of $\Lambda(\theta)$.

It should be emphasized that any Bayes optimal learning system is optimal for any value ξ of Ξ, i.e. for any given set of teaching signals and any input Z after receiving the teaching signals, the expected loss $\varrho(\xi, w^*)$ having the least possible value for any value of ξ.

Let us now estimate the effect of learning. To do this we compare the Bayes optimal learning system with the corresponding Bayes optimal system with complete information about the parameter λ. This Bayes optimal system minimizes the conditional loss

$$\varrho_\lambda(z, \hat{w}) = \int_B l(w, \hat{w} \mid \lambda)\, \mathrm{d}\varkappa(w; \theta \mid z, \lambda)\,. \tag{2.14}$$

If for any z there exists such w_λ^* for which

$$\varrho_\lambda(z, w_\lambda^*) \leqq \varrho_\lambda(z, \hat{w}) \quad \text{for all} \quad \hat{w}\,, \tag{2.15}$$

then the optimal decision function $\delta_\lambda(\Delta\hat{w} \mid z)$ is condensed in single point $\hat{w} = w_\lambda^* = A(\lambda)\, z$, $A(\lambda)$ being an operator depending on λ (and θ, certainly). If for any z there exists a set $C_{z\lambda}$ of w_λ^* satisfying (2.15), then the optimal decision function $\delta_\lambda(\Delta\hat{w} \mid z)$ represents an arbitrary probability measure on $C_{z\lambda}$.

In both cases (2.15) is satisfied for $\hat{w} = w^* = Q\xi$. Hence the relative amount of the loss due to the absence of complete information about the value of the parameter λ:

$$\varepsilon_{\lambda\xi}(\delta_{\mathrm{opt}}) = \frac{\varrho_\lambda(z, w^*) - \varrho_\lambda(z, w_\lambda^*)}{\varrho_\lambda(z, w_\lambda^*)} \tag{2.16}$$

can be taken as the measure of closeness of the Bayes optimal learning system to the Bayes optimal system with complete information.

To estimate the average performance of the learning system for all possible values of ξ given λ, we derive from (2.15)

$$\int_X \mathrm{d}\sigma(\xi \mid \lambda, \bar{\lambda}) \int_B \varrho_\lambda(z, \hat{w})\, \mathrm{d}\delta_\lambda(\hat{w} \mid z) \leqq \int_X \mathrm{d}\sigma(\xi \mid \lambda, \bar{\lambda}) \int_B \varrho_\lambda(z, \hat{w})\, \mathrm{d}\delta_{\mathrm{opt}}(\hat{w} \mid \xi)\,. \tag{2.17}$$

Averaging with respect to $\bar{\lambda}$ and denoting the conditional average loss corresponding to a given value of the parameter λ as

$$r_\lambda(\delta) = \int_L \mathrm{d}_{\bar{\lambda}} \frac{\mathrm{d}_\lambda A(\lambda, \bar{\lambda})}{\mathrm{d}_\lambda \int_L \mathrm{d}_{\bar{\gamma}} A(\lambda, \bar{\gamma})} \int_X \mathrm{d}\sigma(\xi \mid \lambda, \bar{\lambda}) \int_B \varrho_\lambda(z, \hat{w})\, \mathrm{d}\delta(\hat{w} \mid \xi)\,, \tag{2.18}$$

we obtain from (2.17) $r_\lambda(\delta_\lambda) \leqq r_\lambda(\delta_{\mathrm{opt}})$. So the relative amount of the conditional average loss

$$\varepsilon_\lambda(\delta_{\mathrm{opt}}) = \frac{r_\lambda(\delta_{\mathrm{opt}}) - r_\lambda(\delta_\lambda)}{r_\lambda(\delta_\lambda)} \tag{2.19}$$

can serve as a measure of performances of the Bayes optimal learning system for a given value of λ.

Finally, averaging (2.17) with respect to both $\bar{\lambda}$ and λ, we obtain $R(\delta_\lambda) \leqq R(\delta_{\text{opt}})$. So the relative amount of the average loss

$$\varepsilon(\delta_{\text{opt}}) = \frac{R(\delta_{\text{opt}}) - R(\delta_\lambda)}{R(\delta_\lambda)} \tag{2.20}$$

can be taken as a total measure of performance of the Bayes optimal learning system (the measure of its *potential learnability*) [2].*

Similarly the quantities $\varepsilon_{\lambda\xi}(\delta)$, $\varepsilon_\lambda(\delta)$ and $\varepsilon(\delta)$ for any $\delta(\Delta\hat{w} \mid \xi)$ can serve as respective measures of the result of learning for any algorithm of learning.

To estimate the result of learning it is sufficient to compare the values of $\varepsilon_{\lambda\xi}$, ε_λ or ε for the Bayes optimal learning system with the respective values of $\varepsilon_{\lambda\xi}$, ε_λ, ε for the Bayes optimal non-learning system which has no information about λ and does not receive teaching signals. To find the decision function $\delta'_{\text{opt}}(\Delta\hat{w} \mid z)$ of such a system it is sufficient to replace $\sigma(\Delta\xi \mid \lambda, \bar{\lambda})$ in all previous formulas by $\gamma(\Delta z; \theta \mid \lambda)$.

3. SOME GENERAL PROPERTIES OF OPTIMAL LEARNING PROCESSES

From (2.12), (2.10), (2.1), (2,2), (2.4) follow some general properties of optimal learning processes. Primarily, the Bayes optimal learning system represents a permanently learning system, namely it never ceases self-learning, as shown by (2.1), (2.2), (2.4) and (2.10).

It is clear that the learning of a system is possible only if the random variables $\Lambda(\theta_1), \ldots, \Lambda(\theta_N), \Lambda(\theta)$ are interdependent, i.e. if the unknown parameter λ varies sufficiently slowly. The learning is impossible, if $\Lambda(\theta_1), \ldots, \Lambda(\theta_N), \Lambda(\theta)$ are statistically independent, since $\Omega(\Delta\lambda \mid \xi)$ depends only on the value z of the input Z, and is quite independent of previously obtained teaching signals in such a case. Closer is the interdependence between $\Lambda(\theta_1), \ldots, \Lambda(\theta_N), \Lambda(\theta)$ to some determined functional relation, stronger is the effect of previously obtained teaching signals on $\Omega(\Delta\lambda \mid \xi)$, and therefore better are the results of learning. The most effective learning is attained when $\Lambda(\theta_1) = \ldots = \Lambda(\theta_N) = \Lambda(\theta)$, i.e. when the unknown parameter λ remains constant during learning and subsequent system action.

In the special case of the ideal teacher $\delta_T(\Delta\tilde{w}; \theta \mid z, w, \lambda)$ is unity for any set $\Delta\tilde{w}$ including w, and zero otherwise. So $\pi(\Delta\tilde{w}; \theta \mid z, \lambda) = \varkappa(\Delta\tilde{w}; \theta \mid z, \lambda)$ in this case, and (2.2) becomes

$$\sigma(\Delta\xi \mid \lambda, \bar{\lambda}) = \gamma(\Delta z; \theta \mid \lambda) \prod_{i=1}^{N} \int_{\Delta z_i} \varkappa(\Delta\tilde{w}_i; \theta_i \mid z_i, \lambda_i)\, \mathrm{d}\gamma(z_i; \theta_i \mid \lambda_i)\,. \tag{3.1}$$

* If $\varrho_\lambda(z, w^*_\lambda) = 0$ for some z, λ, then the loss $\varrho_\lambda(z, w^*)$ should be taken as a measure of performance of the learning system for these z, λ instead of $\varepsilon_{\lambda\xi}(\delta_{\text{opt}})$. Similarly, if $r_\lambda(\delta_\lambda) = 0$ for some λ, then the conditional average loss $r_\lambda(\delta_{\text{opt}})$ can serve as a measure of performance instead of $\varepsilon_\lambda(\delta_{\text{opt}})$. Finally, if $R(\delta_\lambda) = 0$, the average loss $R(\delta_{\text{opt}})$ would be a suitable measure of performance instead of $\varepsilon(\delta_{\text{opt}})$.

In the case of a real teacher δ_T in (2.3) and (2.5) is independent of w, and therefore $\pi = \delta_T$, and (2.2) and (2.4) take respectively the forms

$$\sigma(\Delta\xi \mid \lambda, \bar{\lambda}) = \gamma(\Delta z; \theta \mid \lambda) \prod_{i=1}^{N} \int_{\Delta z_i} \delta_T(\Delta\tilde{w}_i; \theta_i \mid z_i, \lambda_i)\, \mathrm{d}\gamma(z_i; \theta_i \mid \lambda_i)\,, \tag{3.2}$$

$$\sigma(\Delta\xi \mid \lambda, \bar{\lambda}) = \gamma(\Delta z; \theta \mid \lambda) \prod_{i=1}^{N} \int_{\Delta z_i} \mathrm{d}\gamma(z_i; \theta_i \mid \lambda_i) \times \tag{3.3}$$

$$\times \int_{\Delta\hat{w}_i} \delta_T(\Delta\tilde{w}_i; \theta_i \mid z_i, \hat{w}_i, \lambda_i)\, \mathrm{d}\delta^i(\hat{w}_i \mid z_i)\,.$$

The comparison of (3.1) with (3.2) shows that the real teacher teaching the system by show with $\delta_T(\Delta\tilde{w}; \theta \mid z, \lambda) = \varkappa(\Delta\tilde{w}; \theta \mid z, \lambda)$ is completely equivalent to the ideal teacher. In other words the real teacher whose decision function coincides with the conditional distribution of the desired output W for a given value z of the input Z, is completely equivalent to the ideal teacher from the point of view of its teaching capabilities.

If a real teacher teaching the system by show represents a deterministic system, then the measure δ_T is condensed in single point $\tilde{w} = A_T(\lambda)\, z$, $A_T(\lambda)$ being some deterministic operator depending on the corresponding value λ of $\Lambda(\theta)$ (and may be on θ also), then the posterior measure $\eta(\Delta\lambda, \Delta\bar{\lambda} \mid \xi)$ differs from zero only on the subset of $L \times \bar{L}$ determined by

$$\tilde{w}_i = A_T(\lambda_i)\, z_i \quad (i = 1, \ldots, N)\,. \tag{3.4}$$

If the random variables $\Lambda(\theta_1), \ldots, \Lambda(\theta_N), \Lambda(\theta)$ are strongly correlated, then (3.4) imposes rather strong restriction on the domain of possible values of $\Lambda(\theta)$ owing to which the aposteriori measure $\Omega(\Delta\lambda \mid \xi)$ will be condensed in a narrow domain near the unknown true value of λ. The effect of learning may be better in this case than in the case of the ideal teacher.

If, in particular $\Lambda(\theta_1) = \ldots = \Lambda(\theta_N) = \Lambda(\theta)$, i.e. the parameter λ is constant, then the posterior measure $\Omega(\Delta\lambda \mid \xi)$ differs from zero only on the subset of L determined by the equations

$$\tilde{w}_i = A_T(\lambda)\, z_i \quad (i = 1, \ldots, N)\,. \tag{3.5}$$

If there exists a finite number, say r, of such equations having a unique solution with respect to λ, then the unknown parameter λ is exactly determined after receiving r pairs $(z_1, \tilde{w}_1), \ldots, (z_r, \tilde{w}_r)$ of a teaching signals, and the optimal learning system becomes the Bayes optimal system with complete information about λ, for which $\varepsilon_{\lambda\xi}(\delta_{\mathrm{opt}}) = 0$.

Thus the deterministic real teacher whose operator $A_T(\lambda)$ admits a finite number of equations of the form (3.5) having a unique solution with respect to λ, is the best

possible teacher making the system completely learned after showing to it the respective finite number of teaching pairs of signals. The ideal teacher is thus by no means the best one in such a case. This result seems to be a paradox: the ideal teacher showing to the system *exact* values of the desired output corresponding to given values of the input is worse than the real teacher *committing errors* with probability one. The explanation of this paradox is very simple: since in the general case the input Z represents some signal distorted by noises, and the desired output W depends only on the signal, there exists no deterministic operator transforming Z into W. This is the reason why the relation between Z and W is much more difficult to determine from observations than the relation between Z and the output $\tilde{W}$ of a deterministic teacher.

It is also clear that in such cases stochastic real teachers with sufficiently small variances of $\tilde{W}$ can be better than the ideal teacher.

In special cases where the noises are absent or have small variances, the ideal teacher may certainly be better than any real teacher. The examples where the ideal teacher can make a system completely learned after showing to it a finite number of pairs input-output are numerous in pattern recognition theory (see, e.g., [29]).

Let us now consider an important special case where $\Lambda(\theta)$ represents a Markov random process. Let us partition the teaching cycles and the corresponding values $\theta_1, \ldots, \theta_N$ of θ into two groups $\theta_1, \ldots, \theta_K$ and $\theta_{K+1}, \ldots, \theta_N$ and let Ξ_1 be the set of teaching signals received in the first K cycles plus the $(K+1)$-th input Z_{K+1}, Ξ_2 the set of all remaining teaching signals plus the input Z at the first cycle after learning, λ_i^j the set $\lambda_i, \ldots, \lambda_j$, L_i^j the Cartesian product L^{j-i+1}, $\sigma_1(\Delta\xi_1 \mid \lambda_{K+1}, \lambda_1^K)$, $\beta_1(\Delta\xi_1)$, $\sigma_2(\Delta\xi_2 \mid \lambda, \lambda_{K+1}^N)$, $\beta_2(\Delta\xi_2)$ the respective probability measures σ and β. Then taking into account that in this case

$$A(\Delta\lambda, \Delta\bar{\lambda}) = \int_{\Delta\lambda_{K+1}} A_2(\Delta\lambda, \Delta\lambda_{K+2}^N \mid \lambda_{K+1})\, \mathrm{d}_{\lambda_{K+1}} A_1(\lambda_{K+1}, \Delta\lambda_1^K), \tag{3.6}$$

$A_1(\Delta\lambda_{K+1}, \Delta\lambda_1^K)$ being the probability measure of $\Lambda(\theta_1), \ldots, \Lambda(\theta_{K+1})$, and $A_2(\Delta\lambda, \Delta\lambda_{K+2}^N \mid \lambda_{K+1})$ the conditional probability measure of $\Lambda(\theta_{K+2}), \ldots, \Lambda(\theta_N)$, $\Lambda(\theta)$ given the value λ_{K+1} of $\Lambda(\theta_{K+1})$, we can rewrite (2.10) as

$$\Omega(\Delta\lambda \mid \xi) = \int_{\Delta\lambda} \mathrm{d}_\lambda \int_{L^N{}_{K+1}} \frac{\mathrm{d}\sigma_2(\xi_2 \mid \lambda, \lambda_{K+1}^N)}{\mathrm{d}\beta_2(\xi_2)}\, \mathrm{d}_{\lambda^N{}_{K+1}} A_3(\lambda, \lambda_{K+1}^N) \tag{3.7}$$

with

$$A_3(\Delta\lambda, \Delta\lambda_{K+1}^N) = \int_{\Delta\lambda_{K+1}} A_2(\Delta\lambda, \Delta\lambda_{K+2}^N \mid \lambda_{K+1})\, \mathrm{d}_{\lambda_{K+1}} \Omega_1(\lambda_{K+1} \mid \xi_1), \tag{3.8}$$

$$\Omega_1(\Delta\lambda_{K+1} \mid \xi_1) = \int_{\Delta\lambda_{K+1}} \mathrm{d}_{\lambda_{K+1}} \int_{L_1{}^K} \frac{\mathrm{d}\sigma_1(\xi_1 \mid \lambda_{K+1}, \lambda_1^K)}{\mathrm{d}\beta_1(\xi_1)}\, \mathrm{d}_{\lambda_1{}^K} A_1(\lambda_{K+1}, \lambda_1^K).$$

Formulas (3.7) and (3.8) show that the learning process can be realized recursively in this case, namely $\Omega(\Delta\lambda \mid \xi)$ can be calculated step by step after each teaching cycle using the posterior distribution previously obtained as the prior distribution. This fact permits to reduce essentially the capacity of the memory of computers.

Of course, this is also true for the special case of constant unknown parameter λ.

All the properties of the optimal learning processes studied above were previously established for the special case of constant unknown parameter λ in $[1-3]$.

We have restricted ourselves to studying learning processes for any finite number N of teaching cycles. The reason of this is that only a finite period of learning can be realized in practice. Yet the questions of convergence of learning processes when $N \to \infty$ are also of interest from the theoretical point of view. It is clear from the definitions of Section 2 that the optimal learning process is convergent for a given value of λ and a given sequence ξ of teaching signals, if $\varepsilon_{\lambda\xi}(\delta_{\text{opt}}) \to 0$ when $N \to \infty$. It is convergent for a given value of λ and almost all possible teaching sequences ξ, if $\varepsilon_{\lambda}(\delta_{\text{opt}}) \to 0$. Finally, it is convergent for almost all values of λ and almost possible sequences ξ, if $\varepsilon(\delta_{\text{opt}}) \to 0$.

4. CASE OF DISCRETE SYSTEMS

All the results obtained in previous sections are valid for the most general forms of inputs and outputs of systems. They may be elements of arbitrary sets. The cost of this generality is practical impossibility to use the theory for direct calculations. Yet the input and output of any real technical system represent some ordinary scalar or vector functions of time or other arguments. The large class of modern technical systems is the class of discrete or sampled-data systems. This class include, for instance, all digital computers. The inputs and outputs of such systems represent functions of discrete arguments taking always only finite number of values. Hence the sets of all the values of inputs and outputs may be considered as finite-dimensional vectors, and the sets A and B of the preceding theory represent ordinary finite-dimensional Euclidean spaces.

Thus in the case of discrete systems we may consider the input Z, desired and actual outputs of a system W, $\hat{W}$, and the teacher output $\tilde{W}$ as finite-dimensional random vectors and define their probability measures $\gamma(\Delta z; \theta \mid \lambda)$, $\varkappa(\Delta w; \theta \mid z, \lambda)$, $\delta_{\mathrm{T}}(\Delta\tilde{w}; \theta \mid z, w, \hat{w}, \lambda)$ by respective probability densities*:

$$\gamma(\Delta z; \theta \mid \lambda) = \int_{\Delta z} g(z; \theta \mid \lambda)\, \mathrm{d}z\,, \tag{4.1}$$

* If Z, W and $\tilde{W}$ represent discrete random variables, then the probability densities g, k and d_{T} represent linear combinations of Dirac delta-functions. If Z, W and $\tilde{W}$ have some possible values with non-zero probabilities besides continuous domains of possible values, then the respective probability densities g, k and d_{T} represent sums of functions integrable in Rieman sense and linear combinations of delta-functions.

$$\varkappa(\Delta w;\, \theta \mid z,\, \lambda) = \int_{\Delta w} k(w;\, \theta \mid z,\, \lambda)\, \mathrm{d}w\,, \tag{4.2}$$

$$\delta_{\mathrm{T}}(\Delta \tilde{w};\, \theta \mid z,\, w,\, \hat{w},\, \lambda) = \int_{\Delta \tilde{w}} d_{\mathrm{T}}(\tilde{w};\, \theta \mid z,\, w,\, \hat{w},\, \lambda)\, \mathrm{d}\tilde{w}\,. \tag{4.3}$$

To obtain results applicable for direct calculations in technical system design we suppose that the parameter λ represents also a finite-dimensional vector, and, consequently, L is an ordinary Euclidean space. Thus $\Lambda(\theta)$ of the preceding theory represents now a random vector function of θ, and the probability measures $A(\Delta\lambda, \Delta\bar{\lambda})$ and $\Omega(\hat{w}\lambda \mid \xi)$ can be defined by respective probability densities:

$$A(\Delta\lambda, \Delta\bar{\lambda}) = \int_{\Delta\lambda} \mathrm{d}\lambda \int_{\Delta\bar{\lambda}} \alpha(\lambda, \bar{\lambda})\, \mathrm{d}\bar{\lambda}\,, \tag{4.4}$$

$$\Omega(\Delta\lambda \mid \xi) = \int_{\Delta\lambda} \omega(\lambda \mid \xi)\, \mathrm{d}\lambda\,. \tag{4.5}$$

Using (4.1) – (4.5) we obtain from (2.10), (2.1), (2.2) and (2.4)

$$\omega(\lambda \mid \xi) = c(\xi)\, g(z;\, \theta \mid \lambda) \int_{\bar{L}} \alpha(\lambda, \bar{\lambda}) \prod_{i=1}^{N} g(z_i;\, \theta_i \mid \lambda_i) \times \tag{4.6}$$

$$\times\, p(\tilde{w}_i;\, \theta_i \mid z_i,\, \hat{w}_i,\, \lambda_i)\, \mathrm{d}\bar{\lambda}\,,$$

where integration extends over the domain of all possible values of the composed random vector $\bar{\Lambda}$ formed by all the components of the vectors $\Lambda(\theta_1), \ldots, \Lambda(\theta_N)$, the function $p(\tilde{w};\, \theta \mid z,\, \hat{w},\, \lambda)$ is unity in the case of self-learning, and is determined by

$$p(\tilde{w};\, \theta \mid z,\, \hat{w},\, \lambda) = \int_{B} d_{\mathrm{T}}(\tilde{w};\, \theta \mid z,\, w,\, \hat{w},\, \lambda)\, k(w;\, \theta \mid z,\, \lambda)\, \mathrm{d}w \tag{4.7}$$

in the case of learning with a teacher, and

$$c(\xi) = \left[\int_{L} g(z;\, \theta \mid \lambda)\, \mathrm{d}\lambda \int_{\bar{L}} \alpha(\lambda, \bar{\lambda}) \prod_{i=1}^{N} g(z_i;\, \theta_i \mid \lambda_i) \times \right. \tag{4.8}$$

$$\left. \times\, p(\tilde{w}_i;\, \theta_i \mid z_i,\, \hat{w}_i,\, \lambda_i)\, \mathrm{d}\bar{\lambda} \right]^{-1}$$

represents normalizing constant generally depending on ξ.

In the special case of the ideal teacher, d_{T} represents Dirac delta-function (certainly, multidimensional, in general), and $p = k$. In the case of any real teacher d_{T} is independent of w, and $p = d_{\mathrm{T}}$.

The value w^* of the output of the Bayes optimal learning system corresponding to a given value z of the input Z is determined in this case as the value of $\hat{w}$ minimizing

$$\varrho(\xi, \hat{w}) = \int_{L} \omega(\lambda \mid \xi)\, \mathrm{d}\lambda \int_{B} l(w, \hat{w} \mid \lambda)\, k(w;\, \theta \mid z,\, \lambda)\, \mathrm{d}w\,, \tag{4.9}$$

whereas the value w_λ^* of the output of the Bayes optimal system with complete information about λ corresponding to the same value z of the input Z is determined as the value of $\hat{w}$ minimizing

$$\varrho_\lambda(z, \hat{w}) = \int_B l(w, \hat{w} \mid \lambda)\, k(w; \theta \mid z, \lambda)\, \mathrm{d}w\,. \tag{4.10}$$

The formulas (2.18) and (2.6) determining the conditional average loss $r_\lambda(\delta)$ and average loss $R(\delta)$ become in this case respectively

$$r_\lambda(\delta) = \int_L \alpha_2(\bar{\lambda} \mid \lambda)\, \mathrm{d}\bar{\lambda} \int_X s(\xi \mid \lambda, \bar{\lambda})\, \mathrm{d}\xi \int_B \varrho_\lambda(z, \hat{w})\, d(\hat{w} \mid \xi)\, \mathrm{d}\hat{w}\,, \tag{4.11}$$

$$R(\delta) = \int_L r_\lambda(\delta)\, \alpha_1(\lambda)\, \mathrm{d}\lambda\,, \tag{4.12}$$

where

$$\alpha_1(\lambda) = \int_L \alpha(\lambda, \bar{\lambda})\, \mathrm{d}\bar{\lambda}\,, \quad \alpha_2(\bar{\lambda} \mid \lambda) = \frac{\alpha(\lambda, \bar{\lambda})}{\alpha_1(\lambda)}\,, \tag{4.13}$$

$$s(\xi \mid \lambda, \bar{\lambda}) = g(z; \theta \mid \lambda) \prod_{i=1}^{N} g(z_i; \theta_i \mid \lambda_i) \times d^i(\hat{w}_i \mid z_i)\, p(\bar{w}_i; \theta_i \mid z_i, \hat{w}_i, \lambda_i)\,, \tag{4.14}$$

and $d^i(\hat{w} \mid z), \ldots, d^N(\hat{w} \mid z), d(\hat{w} \mid \xi)$ represent the conditional probability densities of system output $\hat{W}$ at the respective cycles of its action determining its decision functions, of which $d(\hat{w} \mid \xi)$ is to be optimized.

Let us now consider the special case of learning pattern recognition systems designed to recognize which of the mutually exclusive patterns $A_1, \ldots, A_n$ is present in the input Z received.* The desired output W represents in this case the number of the pattern A_w, and therefore is a discrete random variable with n possible values $1, \ldots, n$. The actual output of the system $\hat{W}$ represents the number of the pattern determined by the system. If the refusal to decide is admissble, then $\hat{W}$ is a random variable with $n + 1$ possible values $0, 1, \ldots, n$, the value 0 being assigned to the refusal. As to the teacher output $\bar{W}$, it may represent either the number of the pattern determined by the teacher, or the number of the group of patterns, if the teacher shows to the system only to which of several groups into which the patterns are partitioned the input Z corresponds, or the estimate of the system response given by the teacher in the case of teaching by estimating system actions. Supposing in the latter case that the teacher estimate can assume only integer values $0, 1, \ldots, r$, we cover all the cases assuming that the teacher output $\bar{W}$ represents a random variable with integer possible values $0, 1, \ldots, r$ (r being equal to n in the first case).

* We call *pattern* any set of subjects having some common features in virtue of which they are considered as belonging to the same class. The pattern recognition problem consists in distinguishing between the subjects of different classes, i.e. of classifying objects into several classes.

It should be emphasized that this case covers also the problem of recognizing any finite sequence of patterns, since any such sequence may be considered as a single composed pattern.

Let $p_1, \ldots, p_n$ be the probabilities of the patterns $A_1, \ldots, A_n$ respectively, $f_w(z; \theta \mid \lambda)$ the conditional probability density of the input Z for a given pattern A_w. Then

$$g(z; \theta \mid \lambda) = \sum_{k=1}^{n} p_k f_k(z; \theta \mid \lambda) , \tag{4.15}$$

$$k(w; \theta \mid z, \lambda) = \sum_{k=1}^{n} P_k(z, \lambda, \theta)\, \delta(w - k) , \tag{4.16}$$

$$d_T(\tilde{w}; \theta \mid z, w, \hat{w}, \lambda) = \sum_{h=1}^{r} Q_h(z, w, \hat{w}, \lambda, \theta)\, \delta(\tilde{w} - h) , \tag{4.17}$$

and (4.7) becomes

$$p(\tilde{w}; \theta \mid z, \hat{w}, \lambda) = \sum_{h=0}^{r} \Psi_h(z, \hat{w}, \lambda, \theta)\, \delta(\tilde{w} - h) , \tag{4.18}$$

where

$$\Psi_h(z, \hat{w}, \lambda, \theta) = \sum_{k=1}^{n} P_k(z, \lambda, \theta)\, Q_h(z, k, \hat{w}, \lambda, \theta) . \tag{4.19}$$

Formula (4.6) takes the form

$$\omega(\lambda \mid \xi) = c(\xi)\, g(z; \theta \mid \lambda) \int_L \alpha(\lambda, \bar{\lambda}) \prod_{i=1}^{N} g(z_i; \theta_i \mid \lambda_i)\, \Psi_{\tilde{w}_i}(z_i, \hat{w}_i, \lambda_i, \theta_i)\, \mathrm{d}\bar{\lambda} , \tag{4.20}$$

where

$$c(\xi) = \left[\int_L g(z; \theta \mid \lambda)\, \mathrm{d}\lambda \int_L \alpha(\lambda, \bar{\lambda}) \prod_{i=1}^{N} g(z_i; \theta_i \mid \lambda_i)\, \Psi_{\tilde{w}_i}(z_i, \hat{w}_i, \lambda_i, \theta_i)\, \mathrm{d}\bar{\lambda} \right]^{-1} . \tag{4.21}$$

To obtain this formula from (4.6) strictly the function $\delta(x)$ should be replaced by some bounded function different from zero only in a narrow interval $(-\varepsilon, \varepsilon)$, $\varepsilon < 1/2$. Evidently this does not alter the problem. The this function will be cancelled in (4.20) and (4.21), and we obtain (4.20) in the limit when $\varepsilon \to 0$. Formula (4.20) can, of course, be obtained directly from the general formulas (2.10), (2.1), (2.2) and (2.4) without limiting processes.

Finally, (4.9) takes the form

$$\varrho(\xi, h) = \sum_{k=1}^{n} l_{kh} P_k^*(\xi, \lambda) \quad (h = \hat{w} = 0, 1, \ldots, n) , \tag{4.22}$$

where

$$P_k^*(\xi, \lambda) = \int_L P_k(z, \lambda, \theta)\, \omega(\lambda \mid \xi)\, \mathrm{d}\lambda \quad (k = 1, \ldots, n) . \tag{4.23}$$

The Bayes optimal learning system must choose the value h of $\hat{w}$ minimizing $\varrho(\xi, h)$, whereas the Bayes optimal system with complete information about λ must choose the value m minimizing

$$\varrho_\lambda(z, m) = \sum_{k=1}^{n} l_{km} P_k(z, \lambda, \theta) \quad (m = 0, 1, \ldots, n) . \tag{4.24}$$

Example. Let us consider a learning system designed to estimate the parameter U of the input

$$Z(t) = U\,\varphi(t) + X(t) , \tag{4.25}$$

where $\varphi(t)$ is a given function, U normally distributed random variable with unknown expected value λ and known variance D_u, $X(t)$ normally distributed random function independent of U with zero expected value and known covariance $K_x(t, t')$. The performance of the system is measured by its mean square error $\mathsf{E}(\hat{W} - W)^2$.

Considering the case of discrete systems we assume that the input Z excites the system at time instants $t = t_1, t_2, \ldots, t_n$, and the optimal estimate of U is required at the same time instants. The desired output W is in this case the parameter U. The Bayes optimal system with complete information about λ elaborates its output determined by

$$W_\lambda^*(t_\nu) = \frac{D_u H^{(\nu)} + \lambda}{b^{(\nu)} D_u + 1} \quad (\nu = 1, \ldots, n) , \tag{4.26}$$

where

$$H^{(\nu)} = \sum_{\mu=1}^{\nu} g_{\nu\mu}\, Z(t_\mu) , \quad b^{(\nu)} = \sum_{\mu=1}^{\nu} g_{\nu\mu}\, \varphi(t_\mu) , \tag{4.27}$$

the coefficients $g_{\nu\mu}$ being determined by the sets of linear algebraic equations

$$\sum_{\mu=1}^{\nu} g_{\nu\mu} K_x(t_\mu, t_\sigma) = \varphi(t_\sigma) \quad (\sigma = 1, \ldots, \nu;\ \nu = 1, \ldots, n) . \tag{4.28}$$

To determine the Bayes optimal learning system we suppose that the system receives at N cycles of the period of learning the values $z_1, \ldots, z_N$ of the input and the respective values $\tilde{w}_1, \ldots, \tilde{w}_N$ of teacher output, and then must elaborate the optimal estimate of the signal parameter U at every instant $t_1, \ldots, t_n$ of the first cycle after learning. As to the teacher we shall suppose that it elaborates at each cycle the Bayes optimal estimates $W_\lambda^*(t_\nu)$ $(\nu = 1, \ldots, n)$ of $W = U$ with random errors $\tilde{Y}(t_1), \ldots, \tilde{Y}(t_n)$ which represent the values of a normally distributed random function $\tilde{Y}(s)$ with zero expected value and known covariance $K_{\tilde{Y}}(s, s')$.

The probability density of the input Z and the decision function of the teacher at each of N teaching cycles are determined in this case by

$$g(z_i; \varepsilon_i \mid \lambda) = (2\pi)^{-n/2} \left|K_n\right|^{-1/2} \times \\ \times \exp\left\{-\tfrac{1}{2}\left[z_i - \zeta(\lambda)\right]' K_n^{-1}\left[z_i - \zeta(\lambda)\right]\right\} , \tag{4.29}$$

$$d_T(\tilde{w}_i; \varepsilon_i \mid z_i, \lambda) = (2\pi)^{-n/2} \left|K_T\right|^{-1/2} \times \\ \times \exp\left\{-\tfrac{1}{2}\left[\tilde{w}_i - v_i(\lambda)\right]' K_T^{-1}\left[\tilde{w}_i - v_i(\lambda)\right]\right\} , \tag{4.30}$$

where z_i represents $n \times 1$ matrix whose elements are the values $z_i(t_\nu)$ $(\nu = 1, \ldots, n)$ of the input $Z_i(t_\nu)$ at the i-th cycle, $\zeta(\lambda)$ the $n \times 1$ matrix with the elements $\zeta^{(\nu)}(\lambda) = \lambda\,\varphi(t_\nu)$ $(\nu = 1, \ldots, n)$, K_n the $n \times n$ covariance matrix of the input Z with the elements $D_u\,\varphi(t_\nu)\,\varphi(t_\mu) + K_x(t_\nu, t_\mu)$, $\tilde{w}_i$ the $n \times 1$ matrix whose elements are the values of teacher output

$$\tilde{W}(t_\nu) = \frac{D_u H^{(\nu)} + \lambda}{b^{(\nu)} D_u + 1} + \tilde{Y}(t_\nu) \quad (\nu = 1, \ldots, n) \tag{4.31}$$

at the i-th cycle, $v_i(\lambda)$ the $n \times 1$ matrix with the elements

$$v_i^{(\nu)}(\lambda) = \frac{D_u \eta_i^{(\nu)} + \lambda}{b^{(\nu)} D_u + 1}, \quad \eta_i^{(\nu)} = \sum_{\mu=1}^{\nu} g_{\nu\mu}\, z_i(t_\mu), \tag{4.32}$$

K_T the conditional covariance matrix of the random vector $\tilde{W}_i$ with the elements $K_{\tilde{y}}(t_\nu, t_\mu)$. The probability density of the input Z at each instant t_ν of the first cycle after learning is given by formula (4.29) with n replaced by ν $(\nu = 1, \ldots, n)$.

Assuming a normal prior distribution of the unknown parameter considered as a random variable Λ with the expected value m and variance D, (4.6) and (4.7) yield normal posterior distribution for Λ, the posterior expected value $\lambda^*(t_\nu)$ and variance $\Delta(t_\nu)$ of Λ being determined by

$$\lambda^*(t_\nu) = \Delta(t_\nu)\left\{\frac{\eta^{(\nu)}}{b^{(\nu)} D_u + 1} + \frac{1}{b^{(n)} D_u + 1} \sum_{i=1}^{N} \eta^{(n)} + \right. \\ \left. + \sum_{i=1}^{N} [\zeta_i - \Phi(\eta_i)] + \frac{m}{D}\right\}, \tag{4.33}$$

$$\Delta(t_\nu) = \\ = \frac{D(b^{(\nu)} D_u + 1)(b^{(n)} D_u + 1)}{(b^{(\nu)} D_u + 1)[(1 + N\varkappa D)(b^{(n)} D_u + 1) + N b^{(n)} D] + b^{(\nu)} D(b^{(n)} D_u + 1)}, \tag{4.34}$$

where, in addition to former notations,

$$\eta^{(\nu)} = \sum_{\mu=1}^{\nu} g_{\nu\mu}\, z(t_\mu), \quad \zeta_i = \sum_{\nu=1}^{n} h_\nu\, \tilde{w}_i(t_\nu), \tag{4.35}$$

$$\Phi(\eta) = \sum_{\nu=1}^{n} \frac{h_\nu \eta^{(\nu)}}{b^{(\nu)} D_u + 1}, \quad \varkappa = \sum_{\nu=1}^{n} \frac{h_\nu}{b^{(\nu)} D_u + 1}, \tag{4.36}$$

$z(t_1), \ldots, z(t_n)$ represent the values of the input $Z(t)$ at the first cycle after learning, and the coefficients $h_1, \ldots, h_n$ are determined by the set of linear algebraic equations

$$\sum_{\nu=1}^{n} h_\nu K_{\tilde{y}}(t_\nu, t_\sigma) = \frac{1}{b^{(\sigma)} D_u + 1} \quad (\sigma = 1, \ldots, n). \tag{4.37}$$

The output of the Bayes optimal learning system is given by

$$W^*(t_\nu) = \frac{D_u H^{(\nu)} + \lambda^*(t_\nu)}{b^{(\nu)} D_u + 1}. \tag{4.38}$$

Formula (2.20) gives in this case the following expression for the relative amount of the average loss at each step of the first cycle after learning:

$$\varepsilon(\delta_{\text{opt}}) = \frac{\Delta(t_\nu)}{D_u(b^{(\nu)}D_u + 1)} \quad (\nu = 1, \ldots, n). \tag{4.39}$$

For the Bayes optimal non-learning system $\lambda^*(t_\nu)$ and $\Delta(t_\nu)$ must be replaced by the respective prior values m and D. Formula (4.39) shows then that the performance of the learning system exceeds that of the non-learning system as D exceeds $\Delta(t_\nu)$.

As (4.34) shows $\Delta(t_\nu) \to 0$ when $N \to \infty$, $\lambda^*(t_\nu)$ tends to the unknown true value of λ, and $\varepsilon(\delta_{\text{opt}}) \to 0$. The optimal learning process is thus convergent in this case.

The case of self-learning may be obtained as a special case where the variance of the teacher output W is infinite. Equations (4.37) give in this case $h_1 = \ldots = h_n = 0$, and hence $\zeta_i = \Phi(\eta_i) = \varkappa = 0$.

The case of the ideal teacher, as we proved, is formally the same as the case of the real teacher whose decision function d_T coincides with the conditional probability density of the desired output W given the input Z. Thus to obtain the case of the ideal teacher, we must put

$$K_{\tilde{y}}(t_\nu, t_\mu) = \frac{D_u}{b^{(\sigma)}D_u + 1}, \quad \sigma = \max\{\nu, \mu\}. \tag{4.40}$$

Finally the case of the deterministic teacher which represents the Bayes optimal system with complete information about λ is obtained by putting $K_{\tilde{y}}(s, s') = 0$. Equations (4.37) give in this case $h_1 = \ldots = h_n = \infty$ after which (4.36) yeidl $\varkappa = \infty$, and formula (4.34) shows that $\Delta(t_\nu) = 0$ for any N and ν in this case. This means that the system is completely learned and becomes itself the Bayes optimal system with complete information about λ after receiving a single value (at any of the points $t_1, \ldots, t_n$) of the teacher output. This fact is quite clear, since $\tilde{Y}(s) = 0$ in this case, and (4.31) gives

$$\tilde{w}(t_\nu) = \frac{D_u\eta^{(\nu)} + \lambda}{b^{(\nu)}D_u + 1} \quad (\nu = 1, \ldots, n). \tag{4.41}$$

Any of these equations, say the first one, yields immediately the exact value of the unknown parameter λ. This example illustrates very well the fact established above that the ideal teacher showing to the system exact value of the desired output, i.e. the exact value of the parameter U to be estimated, is worse than the deterministic real teacher estimating U with random errors.

5. CASE OF CONTINUOUS SYSTEMS

In the case of systems with continuous input Z represents a random function of the argument t continuously varying in a certain domain T. The outputs W, $\hat{W}$ and $\tilde{W}$ may be discrete in this case, i.e. random vectors as in the preceding section. The random function $Z(t)$ is scalar in the case of systems with one input, and vector in the case of systems with several inputs. The argument t represents usually time, but may be any scalar or vector variable as well.

To obtain practically applicable results we shall suppose as before that λ is a finite-

dimensional vector. Then from (2.10), (2.1)–(2.5), (4.2)–(4.5) and (4.7) follows

$$\omega(\lambda \mid \xi) = \int_L \frac{\mathrm{d}\sigma(\xi \mid \lambda, \bar{\lambda})}{\mathrm{d}\beta(\xi)} \alpha(\lambda, \bar{\lambda})\, \mathrm{d}\bar{\lambda} \tag{5.1}$$

with

$$\frac{\mathrm{d}\sigma(\xi \mid \lambda, \bar{\lambda})}{\mathrm{d}\beta(\xi)} = \tag{5.2}$$

$$= \frac{\prod_{i=1}^{N} p(\hat{w}_i; \theta_i \mid z_i, \hat{w}_i, \lambda_i)}{\int_L \mathrm{d}\mu \int_L \frac{\mathrm{d}\gamma(z; \theta \mid \mu)}{\mathrm{d}\gamma(z; \theta \mid \lambda)} \prod_{i=1}^{N} \frac{\mathrm{d}\gamma(z_i; \theta_i \mid \mu_i)}{\mathrm{d}\gamma(z_i; \theta_i \mid \lambda_i)} p(\tilde{w}_i; \theta_i \mid z_i, \hat{w}_i, \mu_i)\, \alpha(\mu, \bar{\mu})\, \mathrm{d}\bar{\mu}} .$$

Thus it is sufficient to find Radon-Nikodym derivative $\mathrm{d}\gamma(z; \theta \mid \mu)/\mathrm{d}\gamma(z; \theta \mid \lambda)$.

We shall consider an important special case where $Z(t)$ statistically depends on a finite-dimensional random vector U and has normal conditional distribution for any possible values λ, u of Λ, U. In this case

$$\gamma(\hat{w}z; \theta \mid \lambda) = \int \gamma_1(\Delta z; \theta \mid \lambda, u) f(u; \theta \mid \lambda)\, \mathrm{d}u , \tag{5.3}$$

where $\gamma_1(\Delta z; \theta \mid \lambda, u)$ is Gaussian conditional probability measure of $Z(t)$ given the values λ, u of Λ, U, $f(u; \theta \mid \lambda)$ the conditional probability density of U (which may contain a linear combination of delta-functions) given the value λ of Λ, and integration extends over the region of all possible values of U.

From (5.3) we have

$$\frac{\mathrm{d}\gamma(z; \theta \mid \mu)}{\mathrm{d}\gamma(z; \theta \mid \lambda)} = \int \frac{f(v; \theta \mid \mu)\, \mathrm{d}v}{\int \frac{\mathrm{d}\gamma_1(z; \theta \mid \lambda, u)}{\mathrm{d}\gamma_1(z; \theta \mid \mu, v)} f(u; \theta \mid \lambda)\, \mathrm{d}u} , \tag{5.4}$$

and the problem is reduced to calculating Radon-Nikodym derivative of one Gaussian measure with respect to another Gaussian measure.

To calculate this derivative we suppose in addition that for any possible values λ, u of Λ, U the random function $Z(t)$ is representable by the series

$$Z(t) = \sum_{\nu=1}^{\infty} Z_\nu x_\nu(t) \quad (t \in T) , \tag{5.5}$$

where $\{x_\nu(t)\}$ is some set of functions independent of λ, u (but which may depend on the numerical parameter θ). As such set of functions the set of coordinate functions of a canonical expansion of the random function $Z(t)$ for some specific values λ_0, u_0 of λ, u may be used. Let $\{\Omega^{(\nu)}\}$ be the set of linear functionals satisfying the biortho-

 gonality condition

$$\Omega^{(\nu)} x_\mu = \delta_{\nu\mu} . \tag{5.6}$$

Such set may be determined, for instance, using the procedure indicated in [30], § 63. If $x_\nu(t)$ represent the coordinate functions of some canonical expansion, then the set $\{\Omega^{(\nu)}\}$ is automatically determined while finding this canonical expansion [30], §§ 60–62.

From (5.5) and (5.6) follows

$$Z_\nu = \Omega^{(\nu)} Z(t) . \tag{5.7}$$

Formulas (5.5) and (5.7) establish complete equivalence between the random function $Z(t)$, $t \in T$, and the set of random variables $\{Z_\nu\}$ in the sense that any set of values $\{z_\nu\}$ of $\{Z_\nu\}$ determine completely the corresponding sample $z(t)$, $t \in T$ of $Z(t)$ and vice versa.

Let us now find the conditional probability density $f_1(\lambda, u \mid z)$ of Λ, U given the value $z(t)$, $t \in T$ of $Z(t)$. We have in virtue of the above proved equivalence between $Z(t)$, $t \in T$, and $\{Z_\nu\}$

$$f_1(\lambda, u \mid z) = \lim_{n \to \infty} f_1^{(n)}(\lambda, u \mid z_1, \ldots, z_n) . \tag{5.8}$$

On the other hand

$$f_1(\lambda, u \mid z) = \frac{\alpha_1(\lambda) f(u; \theta \mid \lambda)}{\int_L \alpha_1(\mu) \, \mathrm{d}\mu \int \frac{\mathrm{d}\gamma_1(z; \theta \mid \mu, v)}{\mathrm{d}\gamma_1(z; \theta \mid \lambda, u)} f(v; \theta \mid \mu) \, \mathrm{d}v} . \tag{5.9}$$

Hence, once $f_1(\lambda, u \mid z)$ is found, the derivative $\mathrm{d}\gamma_1(z; \theta \mid \mu, v)/\mathrm{d}\gamma_1(z; \theta \mid \lambda, u)$ will be determined immediately. To determine $f_1^{(n)}(\lambda, u \mid z_1, \ldots, z_n)$ in (5.8) we notice that the joint conditional distribution of $Z_1, \ldots, Z_n$ is normal for any λ, u with the expected values

$$m_\nu(\lambda, u) = \Omega_t^{(\nu)} m_z(t \mid \lambda, u) \tag{5.10}$$

and covariance matrix $K_n(\lambda, u)$ whose elements are given by

$$k_{\nu\mu}(\lambda, u) = \Omega_t^{(\nu)} \Omega_\tau^{(\mu)} K_z(t, \tau \mid \lambda, u) , \tag{5.11}$$

where $m_z(t \mid \lambda, u)$, $K_z(t, t' \mid \lambda, u)$ are respectively the conditional expected value and conditional covariance of the random function $Z(t)$ given λ, u (they depend also on θ which is omitted for brevity). Therefore

$$f_1^{(n)}(\lambda, u \mid z_1, \ldots, z_n) = c_n \alpha_1(\lambda) f(u; \theta \mid \lambda) \times \tag{5.12}$$

$$\times \exp\{-\tfrac{1}{2}\sum_{\nu,\mu=1}^{n} k_{\nu\mu}^{-}(\lambda, u)\,\Omega_t^{(\nu)}\Omega_\tau^{(\mu)}[z(t) - m_z(t \mid \lambda, u)]\,[z(\tau) - m_z(\tau \mid \lambda, u)] - \tfrac{1}{2}\ln \Delta_n(\lambda, u)\},$$

where $k_{\nu\mu}^{-}(\lambda, u)$ represent the elements of the inverse matrix $K_n^{-1}(\lambda, u)$ (necessarily depending on n), c_n the normalizing constant independent of λ, u, and $\Delta_n(\lambda, u)$ the determinant of the $n \times n$ matrix $\Gamma_n(\lambda, u)$ with the elements

$$\gamma_{\mu\nu}(\lambda, u) = \frac{k_{\nu\mu}(\lambda, u)}{\sqrt{(D_\nu D_\mu)}}, \tag{5.13}$$

$\{D_\varkappa\}$ being an arbitrary sequence of positive numbers.

Now introducing the bilinear operator

$$Q_n(\lambda, u) = \sum_{\nu,\mu=1}^{n} [k_{\nu\mu}^{-}(\lambda, u) - D_\nu^{-1}\delta_{\nu\mu}]\,\Omega_t^{(\nu)}\Omega_\tau^{(\mu)} \tag{5.14}$$

and the linear operator

$$L_n(\lambda, u) = \sum_{\nu=1}^{n}[\sum_{\mu=1}^{n} k_{\nu\mu}^{-}(\lambda, u)\,\Omega_\tau^{(\mu)} m_z(\tau \mid \lambda, u)]\,\Omega_t^{(\nu)}, \tag{5.15}$$

we rewrite (5.12) in the form

$$f_1^{(n)}(\lambda, u \mid z_1, \ldots, z_n) = c_n\,\alpha_1(\lambda) f(u; \theta \mid \lambda) \exp\{-\tfrac{1}{2}Q_n(\lambda, u)\,z(t)\,z(\tau) + L_n(\lambda, u)\,z(t) - \tfrac{1}{2}\beta_n(\lambda, u)\}, \tag{5.16}$$

where

$$\beta_n(\lambda, u) = L_n(\lambda, u)_t\, m_z(t \mid \lambda, u) + \ln \Delta_n(\lambda, u). \tag{5.17}$$

Substituting (5.16) into (5.8), taking the limit with $n \to \infty$, and comparing the result with (5.9), we find

$$\frac{d\gamma_1(z; \theta \mid \mu, v)}{d\gamma_1(z; \theta \mid \lambda, u)} = \frac{\exp\{-\tfrac{1}{2}Q(\mu, v)\,z(t)\,z(\tau) + L(\mu, v)\,z(t) - \tfrac{1}{2}\beta(\mu, v)\}}{\exp\{-\tfrac{1}{2}Q(\lambda, u)\,z(t)\,z(\tau) + L(\lambda, u)\,z(t) - \tfrac{1}{2}\beta(\lambda, u)\}}, \tag{5.18}$$

where

$$Q(\lambda, u) = \lim_{n\to\infty} Q_n(\lambda, u), \quad L(\lambda, u) = \lim_{n\to\infty} L_n(\lambda, u), \tag{5.19}$$

and

$$\beta(\lambda, u) = \lim_{n\to\infty} \beta_n(\lambda, u) = L(\lambda, u)_t\, m_z(t \mid \lambda, u) + \ln \Delta(\lambda, u) \tag{5.20}$$

with $\Delta(\lambda, u) = \lim_{n\to\infty} \Delta_n(\lambda, u)$.

Now we introduce the random function

$$X(t) = \sum_{\nu=1}^{\infty} V_\nu \, x_\nu(t) \quad (t \in T), \tag{5.21}$$

where V_ν are uncorrelated random variables with zero expected values and variances equal to the respective numbers D_ν chosen so that the series

$$\sum_{\nu=1}^{\infty} D_\nu \, x_\nu^2(t) \tag{5.22}$$

be convergent for all $t \in T$. Then the random function $X(t)$ has bounded covariance

$$K_x(t, \tau) = \sum_{\nu=1}^{\infty} D_\nu \, x_\nu(t) \, x_\nu(\tau) \quad (t, \tau \in T). \tag{5.23}$$

Now it is easy to show that the operators Q and L satisfy respectively the linear equations

$$Q(\lambda, u)_{ts} \, K_z(t, \tau \mid \lambda, u) \, K_x(s, \sigma) = K_x(\tau, \sigma) - K_z(\tau, \sigma \mid \lambda, u) \quad (\tau, \sigma \in T), \tag{5.24}$$

$$L(\lambda, u)_t \, K_z(t, \tau \mid \lambda, u) = m_z(\tau \mid \lambda, u) \quad (\tau \in T). \tag{5.25}$$

Formula (5.18) determines the Radon-Nikodym derivative of two Gaussian measures, if they are absolutely continuous one with respect to other (i.e. equivalent), as is generally the case. But this formula is also valid, if these two measures are orthogonal, giving in such a case Radon-Nikodym derivative in the form of the delta-functional (which is the generalization of usual Dirac delta-function). From (5.18) follow as special cases some results of papers [31, 32].

Substituting (5.18) into (5.4) formulas (5.1) and (5.2) give again (4.6) with

$$g(z; \theta \mid \lambda) = \int f(u; \theta \mid \lambda) \exp \{ -\tfrac{1}{2} Q(\lambda, u) \, z(t) \, z(\tau) + L(\lambda, u) \, z(t) - \tfrac{1}{2} \beta(\lambda, u) \} \, \mathrm{d}u \,. \tag{5.26}$$

Thus in the case of systems with continuous input and discrete output the posterior probability density of $A(\theta)$ is determined by (4.6) and (4.7), $g(z; \theta \mid \lambda)$ being the functional of $z(t)$ given by (5.26) with operators Q and L satisfying the linear equations (5.24) and (5.25), and the function β determined by (5.20).

To solve equations (5.24) and (5.25) the method of canonical expansions of random functions may be applied in the general case. Representing the random function $Z(t)$, $t \in T$, for given λ, u by some canonical expansion, we find $P(\lambda, u, \sigma)_t = Q(\lambda, u)_{ts} \, K_x(s, \sigma)$ and $L(\lambda, u)$ as shown in [30], §§ 135, 136, after which using some canonical expansion of the form (5.21) of the random function $X(t)$, we obtain by the same techniques the operator $Q(\lambda, u)$. In various special cases other techniques can be used (see, for example, [30], §§ 128–133, and [33]).

The most difficult for practical calculations in evaluating $g(z; \theta \mid \lambda)$ is the finding of the infinite determinant $\Delta(\lambda, u)$. To derive a suitable expression for $\Delta(\lambda, u)$ we put

$$K_z(t, \tau \mid \lambda, u) = K_x(t, \tau) + R(t, \tau \mid \lambda, u) . \tag{5.27}$$

Then taking into account (5.6) and (5.23), formula (5.13) yields

$$\gamma_{\nu\mu}(\lambda, u) = \delta_{\nu\mu} + \frac{\Omega_t^{(\nu)}\Omega_\tau^{(\mu)}R(t, \tau \mid \lambda, u)}{\sqrt{(D_\nu D_\mu)}} . \tag{5.28}$$

Using the conventional expansion for $\Delta_n(\lambda, u)$ and passing to the limit with $n \to \infty$, we obtain in virtue of linearity of the functionals $\Omega^{(\nu)}$

$$\Delta(\lambda, u) = 1 + \sum_{p=1}^{\infty} \frac{1}{p!} \Theta_{t_1\tau_1}^{i_1 j_1} \dots \Theta_{t_p\tau_p}^{i_p j_p} \begin{vmatrix} R_{i_1 j_1}(t_1, \tau_1 \mid \lambda, u) & \dots & R_{i_1 j_p}(t_1, \tau_p \mid \lambda, u) \\ \dots\dots\dots\dots & \dots & \dots\dots\dots\dots \\ R_{i_p j_1}(t_p, \tau_1 \mid \lambda, u) & \dots & R_{i_p j_p}(t_p, \tau_p \mid \lambda, u) \end{vmatrix} , \tag{5.29}$$

where $R_{ij}(t, \tau \mid \lambda, u)$ represent the elements of the matrix $R(t, \tau \mid \lambda, u)$, and Θ is the bilinear operator

$$\Theta = \sum_{\nu=1}^{n} D_\nu^{-1} \Omega_t^{(\nu)} \Omega_\tau^{(\nu)} , \tag{5.30}$$

$\Theta_{t\tau}^{ij}$ acting on a matrix whose elements are numbered by i, j and represent functions of t, τ.

It is easy to show that the operator Θ satisfies the linear equation

$$\Theta_{ts} K_x(t, \tau) K_x(s, \sigma) = K_x(\tau, \sigma) \quad (\tau, \sigma \in T) . \tag{5.31}$$

Formula (5.29) is especially convenient, if $R(t, \tau \mid \lambda, u)$ is small as compared with $K_x(t, \tau)$. In this case only a few first members of the series in (5.29) will suffice. The numbers D_ν can be chosen optimally so that $K_x(t, \tau)$ coincide with the expected value of $K_z(t, \tau \mid \lambda, u)$ averaged with respect to λ, u.

Equations (5.24), (5.25) and (5.31) represent linear integral equations of the first kind, if $Z(t)$ is a scalar random function, and sets of simultaneous linear integral equations of the first kind, if $Z(t)$ is a vector random function. Finally, (5.24), (5.25) and (5.31) represent sets of linear algebraic equations, if T is a discrete set of values of the argument t (case of discrete systems).

The special case where $Z(t)$ represents the sum of two indepedent random functions, one of which has normal conditional distribution for any values λ, u of Λ, U, and other is independent of Λ, U normally distributed random functions, was first studied by L. P. Sysoev [34] who used canonical expansions in another way and obtained for this special case another expression for $\Delta(\lambda, u)$.

In the special case where $K_x(t, \tau \mid \lambda, u)$ is independent of λ, u, it can be taken as $K_x(t, \tau)$. Then (5.24) and (5.29) give $Q(\lambda, u) = 0$, $\Delta(\lambda, u) = 1$, and we obtain the results formerly derived in [35–37] (see also [30], § 144).

In the case of systems with continuous input and continuous output we obtain from (2.10), (2.2)–(2.5), (4.4) and (4.5) the expression (5.1) with

$$\frac{d\sigma(\xi \mid \lambda, \bar{\lambda})}{d\beta(\xi)} = \left[\int_L d\mu \int_L \frac{d\gamma(z; \theta \mid \mu)}{d\gamma(z; \theta \mid \lambda)} \times \right. \tag{5.32}$$

$$\left. \times \prod_{i=1}^{N} \frac{d\gamma(z_i; \theta_i \mid \mu_i)}{d\gamma(z_i; \theta_i \mid \lambda_i)} \frac{d\delta_T(\tilde{w}_i; \theta_i \mid z_i, \hat{w}_i, \mu_i)}{d\delta_T(\tilde{w}_i; \theta_i \mid z_i, \hat{w}_i, \lambda_i)} \alpha(\mu, \bar{\mu})\, d\bar{\mu} \right]^{-1}.$$

The Radon-Nikodym derivative

$$d\delta_T(\tilde{w}; \theta \mid z, \hat{w}, \mu) / d\delta_T(\tilde{w}; \theta \mid z, \hat{w}, \lambda)$$

can be calculated by the same techniques as $d\gamma(z; \theta \mid \mu)/d\gamma(z; \theta \mid \lambda)$ in the case where $\tilde{W}$ represents a random function $\tilde{W}(p)$, $p \in P$, depending on Λ and some other random vector A, and having normal conditional distribution for any $z(t)$, $t \in T$, $\hat{w}(s)$, $s \in S$, and any possible values λ, a of Λ, A. In the case of teaching by show δ_T is independent of $\hat{w}(s)$, and P coincides with S. In the case of the ideal teacher δ_T in (5.32) coincides with $\varkappa$.

Thus in the case of systems with continuous input and output formula (4.6) for the posterior probability density $\omega(\lambda \mid \xi)$ is valid with $g(z; \theta \mid \lambda)$ determined by (5.26) and

$$p(\tilde{w}; \theta \mid z, \hat{w}, \lambda) = d_T(\tilde{w}; \theta \mid z, \hat{w}, \lambda) = \tag{5.33}$$

$$= \int f_T(a; \theta \mid z, \hat{w}, \lambda) \exp\left\{-\tfrac{1}{2} Q_T(z, \hat{w}, \lambda, a)\, \tilde{w}(p)\, \tilde{w}(q) + \right.$$

$$\left. + L_T(z, \hat{w}, \lambda, a)\, \tilde{w}(p) - \tfrac{1}{2} \beta_T(z, \hat{w}, \lambda, a)\right\} da\,,$$

where $f_T(a; \theta \mid z, \hat{w}, \lambda)$ is the conditional probability density of the random vector A given $z(t)$, $t \in T$, $\hat{w}(s)$, $s \in S$, λ; $Q_T(z, \hat{w}, \lambda, a)$ the bilinear operator satisfying the linear equation

$$Q_T(z, \hat{w}, \lambda, a)_{pq} K_{\hat{w}}(p, \xi \mid z, \hat{w}, \lambda, a) K_y(q, \eta \mid z, \hat{w}) = \tag{5.34}$$

$$= K_y(\xi, \eta \mid z, \hat{w}) - K_{\hat{w}}(\xi, \eta \mid z, \hat{w}, \lambda, a) \quad (\xi, \eta \in P)\,,$$

$L_T(z, \hat{w}, \lambda, a)$ the linear operator satisfying the equation

$$L_T(z, \hat{w}, \lambda, a)_p K_{\hat{w}}(p, \xi \mid z, \hat{w}, \lambda, a) = m_{\hat{w}}(\xi \mid z, \hat{w}, \lambda, a) \quad (\xi \in P)\,, \tag{5.35}$$

and

$$\beta_T(z, \hat{w}, \lambda, a) = L_T(z, \hat{w}, \lambda, a)_p\, m_{\hat{w}}(p \mid z, \hat{w}, \lambda, a) + \ln \Delta_T(z, \hat{w}, \lambda, a)\,, \tag{5.36}$$

$\Delta_T(z, \hat{w}, \lambda, a)$ being the infinite determinant with the elements

$$\gamma^T_{\nu\mu}(z, \hat{w}, \lambda, a) = \frac{r_p^{(\nu)} r_q^{(\mu)} K_{\hat{w}}(p, q \mid z, \hat{w}, \lambda, a)}{\sqrt{(\Delta_\nu \Delta_\mu)}}, \tag{5.37}$$

$m_{\hat{w}}(p \mid z, \hat{w}, \lambda, a)$, $K_{\hat{w}}(p, q \mid z, \hat{w}, \lambda, a)$ the conditional expected value and covariance of the random function $\hat{W}(p)$, $r^{(\nu)}$ the linear functionals similar to $\Omega^{(\nu)}$, Δ_ν positive numbers similar to D_ν, and $K_y(p, q \mid z, \hat{w})$ the conditional covariance of the random function $Y(p)$ costructed in the same way as $X(t)$. The infinite determinant $\Delta_T(z, \hat{w}, \lambda, a)$ can be expressed by the formula similar to (5.29).

To obtain the expression for $\varrho(\xi, \hat{w})$ suitable to practical calculations, it is necessary to impose certain restrictions on the loss function.

At first we suppose that the loss function $l(w, \hat{w} \mid \lambda)$ is an ordinary function of the values of the functions $w(s)$ and $\hat{w}(s)$ at a finite set of points $s_1, \ldots, s_m$. In this case formula (4.9) is valid, $k(w; \theta \mid z, \lambda)$ being the joint conditional probability density of $W(s_1), \ldots, W(s_m)$.

Secondly we consider the loss function which is a functional of $w(s)$, $\hat{w}(s)$, of the form

$$l(w, \hat{w} \mid \lambda) = \Phi_s \sigma(s, w(s), \hat{w}(s) \mid \lambda), \tag{5.38}$$

where Φ is a linear functional in a space of ordinary functions of s, $s \in S$, and σ a function of s, $w(s)$, $\hat{w}(s)$ which may depend also on the derivatives $w'(s), w''(s), \ldots$ $\ldots, w^{(p)}(s), \hat{w}'(s), \hat{w}''(s), \ldots, \hat{w}^{(p)}(s)$, p being some positive integer. In this case

$$\varrho(\xi, \hat{w}) = \int_L \omega(\lambda \mid \xi)\, \mathrm{d}\lambda \Phi_s \int_B \sigma(s, w, \hat{w}(s) \mid \lambda)\, k(w; s, \theta \mid z, \lambda)\, \mathrm{d}w, \tag{5.39}$$

where $k(w; s, \theta \mid z, \lambda)$ is the joint conditional probability density of $W(s), W'(s), \ldots$ $\ldots, W^{(p)}(s)$ for a given value of the argument s.

To estimate the performance of the learned system only $\varepsilon_{\lambda\xi}(\delta)$ can be used in the general case. The quantities $\varepsilon_\lambda(\delta)$ and $\varepsilon(\delta)$ can be calculated practically only approximately. And only in some special cases one may succeed in obtaining exact values of $\varepsilon_\lambda(\delta)$ and $\varepsilon(\delta)$, as we shall see in the following example.

To obtain $\varepsilon_{\lambda\xi}(\delta)$, we use (4.9) and (4.10), or (5.39) and the corresponding expression for $\varrho_\lambda(z, \hat{w})$:

$$\varrho_\lambda(z, \hat{w}) = \Phi_s \int_B \sigma(s, w, \hat{w}(s) \mid \lambda)\, k(w; s, \theta \mid z, \lambda)\, \mathrm{d}w. \tag{5.40}$$

In the special case of recognizing system with continuous input (its output is always discrete) all the formulas of the preceding section are valid, if the input $Z(t)$ is a random function of the type considered, with $f_w(z; \theta \mid \lambda)$ given by the formula

similar to (5.26):

$$f_w(z;\theta \mid \lambda) = \int f(u;\theta \mid \lambda) \exp\{-\tfrac{1}{2}Q_w(\lambda,u)\, z(t)\, z(\tau) + \tag{5.41}$$

$$+ L_w(\lambda,u)\, z(t) - \tfrac{1}{2}\beta_w(\lambda,u)\}\, \mathrm{d}u\,,$$

operators Q_w and L_w being determined by the equations similar to (5.24) and (5.25), and the function β_w by the formula similar to (5.20).

Example. Let us consider the problem of the example of the preceding section, supposing that the input $Z(t)$ excites the system continuously in the interval of time $0 \leqq t \leqq T$ and the optimal estimate of the parameter U is required at any instant s of the interval $s_0 \leqq s \leqq T$. The output of the Bayes optimal system with complete information about λ is in this case determined by

$$W^*_\lambda(s) = \frac{D_u\, H(s) + \lambda}{b(s)\, D_u + 1} \quad (s_0 \leqq s \leqq T)\,, \tag{5.42}$$

where

$$H(s) = \int_0^s g(s,t)\, Z(t)\, \mathrm{d}t\,, \quad b(s) = \int_0^s g(s,t)\, \varphi(t)\, \mathrm{d}t\,, \tag{5.43}$$

the weighting function $g(s,t)$ being determined by the integral equation

$$\int_0^s g(s,t)\, K_x(t,\tau)\, \mathrm{d}t = \varphi(\tau) \quad (0 \leqq \tau \leqq T)\,. \tag{5.44}$$

To find the Bayes optimal learning system we suppose that the system receives the teacher output $\tilde{W}(s)$ continuously in the time interval $s_0 \leqq s \leqq T$ at each cycle of learning. We assume also that the teacher gives the Bayes optimal estimate $W^*_\lambda(s)$ of $W(s) = U$ with random error $\tilde{Y}(s)$ representing a normally distributed random function with zero expected value and the covariance $K_{\tilde{y}}(s,s')$. Using (5.20), (5.24)—(5.26) and (5.33)—(5.36) and taking into account that $A(\lambda,u) = A_T(z,w,\lambda,a) = 1$ in this case, we obtain

$$g(z;\theta \mid \lambda) = c_1 \exp\left\{-\frac{[b(T)\, D_u + 1]\,\lambda^2 - [D_u\, \eta(T) + \lambda]^2}{2D_u[b(T)\, D_u + 1]}\right\}, \tag{5.45}$$

$$d_T(\tilde{w};\theta \mid z,\lambda) = c_2 \exp\{[\zeta - \Phi(\eta)]\,\lambda - \tfrac{1}{2}\varkappa\lambda^2\}\,, \tag{5.46}$$

where c_1, c_2 are constants independent of λ, $\eta(s)$ is determined by (5.43) with $Z(t)$ replaced by $z(t)$, and

$$\zeta = \int_{s_0}^T h(s)\, \tilde{w}(s)\, \mathrm{d}s\,, \quad \Phi(\eta) = \int_{s_0}^T \frac{h(s)\, \eta(s)\, \mathrm{d}s}{b(s)\, D_u + 1}\,, \quad \varkappa = \int_{s_0}^T \frac{h(s)\, \mathrm{d}s}{b(s)\, D_u + 1}\,, \tag{5.47}$$

the weighting function $h(s)$ being determined by the integral equation

$$\int_{s_0}^T h(s)\, K_{\tilde{y}}(s,\sigma)\, \mathrm{d}s = \frac{1}{b(\sigma)\, D_u + 1} \quad (s_0 \leqq \sigma \leqq T)\,. \tag{5.48}$$

Then assuming as in the example of the preceding section the normal prior distribution for A with the expected value m and variance D, we obtain the normal posterior distribution $\omega(\lambda \mid \xi)$, the posterior expected value $\lambda^*(s)$ and variance $\Delta(s)$ of A being determined by

$$\lambda^*(s) = \Delta(s)\left\{\frac{\eta(s)}{b(s)\,D_u + 1} + \frac{1}{b(T)\,D_u + 1}\sum_{i=1}^{N}\eta_i(T) + \right. \tag{5.49}$$

$$\left. + \sum_{i=1}^{N}[\zeta_i - \Phi(\eta_i)] + \frac{m}{D}\right\},$$

$$\Delta(s) = \tag{5.50}$$

$$= \frac{D[b(s)\,D_u + 1]\,[b(T)\,D_u + 1]}{[b(s)\,D_u + 1]\,\{(1 + N\varkappa D)\,[b(T)\,D_u + 1] + Nb(T)\,D\} + b(s)\,[b(T)\,D_u + 1]}.$$

The output of the Bayes optimal learning system is determined by

$$W^*(s) = \frac{D_u H(s) + \lambda^*(s)}{b(s)\,D_u + 1}. \tag{5.51}$$

Finally, we shall estimate the results of learning. Using (4.9) and (4.10) with quadratic loss function $l(w, w \mid \lambda) = (w - w)^2$, we see that $\varepsilon_{\lambda\xi}(\delta_{\text{opt}})$ is independent of λ, ξ, and therefore

$$\varepsilon(\delta_{\text{opt}}) = \varepsilon_{\lambda}(\delta_{\text{opt}}) = \varepsilon_{\lambda\xi}(\delta_{\text{opt}}) = \frac{\Delta(s)}{D_u[b(s)\,D_u + 1]}. \tag{5.52}$$

Thus the fortunate circumstance – the independence of $\varrho(\xi, w^*)$ and $\varrho_{\lambda}(z, w_{\lambda}^*)$ of λ and ξ – gives in our case the possibility to find $\varepsilon_{\lambda}(\delta_{\text{opt}})$ and $\varepsilon(\delta_{\text{opt}})$ without tedious calculations. Such a case is, however, rarely encountered in problems of practice.

(Received August 22, 1967.)

REFERENCES

[1] Pugachev V. S.: Statistical Problems of Pattern Recognition Theory. The Third All-Union Conference on Automatic Control. Odessa, 1965.

[2] Pugachev V. S.: A Bayes Approach to the Theory of Learning Systems. The Third Congress of IFAC, London, 20–25 June, 1966.

[3] Pugachev V. S.: Optimal Algorithms of Learning of Automatic Systems in the Case of a Non-Ideal Teacher. Doklady Akademii Nauk SSSR *172* (1967), 5, 1039–1042 (in Russian).

[4] Fabian V., Špaček A.: Experience in Statistical Decision Problems. Czechoslovak Math. Journ. *6* (1956), 190–194.

[5] Špaček A.: An Elementary Experience Problem. Transactions of the Prague Conference on Information Theory, Statistical Decision Functions, Random Processes (1956). Prague 1957, 253–258.

[6] Driml M., Špaček A.: Continuous Random Decision Processes Controlled by Experience. Transactions of the First Prague Conference on Information Theory, Statistical Decision Functions, Random Processes (1956). Prague 1957, 43–60.

[7] Winkelbauer K.: Experience in Games of Strategy and in Statistical Decision. Transactions of the First Prague Conference on Information Theory, Statistical Decision Functions, Random Processes (1956). Prague 1957, 297—354.

[8] Hanš O.: Random Fixed Point Theorems. Transactions of the First Prague Conference on Information Theory, Statistical Decision Functions, Random Processes (1956). Prague 1957, 105—125.

[9] Prouza L.: Bemerkungen zur Lineare Prediktion mittels eines lernenden Filters. Transactions of the First Prague Conference on Information Theory, Statistical Decision Functions, Random Processes (1956). Prague 1957, 37—41.

[10] Driml M., Hanš O.: On Experience Theory Problems. Transactions of the Second Prague Conference on Information Theory, Statistical Decision Functions, Random Processes (1959). Prague 1960, 93—111.

[11] Driml M., Hanš O.: Continuous Stochastic Approximations. Transactions of the Second Prague Conference on Information Theory, Statistical Decision Functions, Random Processes (1959). Prague 1960, 113—122.

[12] Driml M., Nedoma J.: Stochastic Approximations for Continuous Random Processes. Transactions of the Second Prague Conference on Information Theory, Statistical Decision Functions, Random Processes (1959). Prague 1960, 145—158.

[13] Hanš O., Špaček A.: Random Fixed Point Approximation by Differentiable Trajectories. Transactions of the Second Prague Conference on Information Theory, Statistical Decision Functions, Random Processes (1959). Prague 1960, 203—213.

[14] Šefl O.: Filters and Predictors which Adapt Their Values to the Unknown Parameters of the Input Process. Transactions of the Second Prague Conference on Information Theory, Statistical Decision Functions, Random Processes (1959). Prague 1960, 597—608.

[15] Fabian V.: A Stochastic Approximation Method for Finding Optimal Conditions in Experimental Work and in Self-Adaptive Systems. Aplikace matematiky *6* (1961), 162—183 (in Czech).

[16] Fabian V.: A Block-scheme of an Automatic Optimizer. Aplikace matematiky 7 (1962), 426—440 (in Czech).

[17] Fabian V.: A New One-dimensional Stochastic Approximation Method for Finding a Local Minimum of a Function. Transactions of the Third Prague Conference on Information Theory, Statistical Decision Functions, Random Processes (1962). Prague 1964, 85—105.

[18] Šefl O.: Some Problems of Automatic Control Processes with Unknown Characteristics. Transactions of the Third Prague Conference on Information Theory, Statistical Decision Functions, Random Processes (1962). Prague 1964, 611—620 (in Russian).

[19] Gardner L. A.: Adaptive Predictors. Transactions of the Third Prague Conference on Information Theory, Statistical Decision Functions, Random Processes (1962). Prague 1964, 123—192.

[20] Pugachev V. S.: Method of Determining the Optimum System with a Non-linear Dependence of the Observed Function of the Parameters of Signal. Automatic and Remote Control. Proceedings of the First Congress of IFAC (1960). Butterworth, London 1961, 2, 702—705.

[21] Robbins H.: An Empirical Bayes Approach to Statistics. Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, 1 (1955), 157—164.

[22] Robbins H.: An Empirical Bayes Approach to Testing Statistical Hypothesis. Review of the International Statistical Institute *31* (1963), 2, 195—208.

[23] Robbins H.: An Empircial Bayes Approach to Problems of Statistical Decision Theory. Annals of Math. Statistics *35* (1964), 1, 1—20.

[24] Johns H. V.: Non-parametric Empircial Bayes Procedures. Annals of Math. Statistics *28* (1957), 649—669.

[25] Johns H. V.: An Empirical Bayes Approach to Non-parametric Two-way Classification.

Studies in Item Analysis and Prediction (edited by H. Solomon). Stanford University Press, 1961, 221—232.

[26] Cooper D. V.: Adaptive Pattern Recognition and Signal Detection Using Stochastic Approximation. IEEE Transactions on Electronic Computers *EC-13* (1964), 3, 306—307.

[27] Tsypkin Ya. Z.: Adaptation, Lerning and Self-learning in Automatic Systems. Avtomatika i Telemekhanika *27* (1966), 1, 23—61 (in Russian).

[28] Shaikin M. E.: A Bayes Approach to Self-adapting Filters Using a Finite Number of Teaching Samples. The Third All-Union Conference on Automatic Control. Odessa, 1965.

[29] Aizerman M. A., Braverman E. M., Rozonoer L. I.: Theoretical Foundations of Potential Functions Method in the Problem of Teaching Automata to Classify Input Situations. Avtomatika i Telemekhanika *25* (1964), 6, 917—936 (in Russian).

[30] Pugachev V. S.: Theory of Random Functions and its Application to Control Problems. Pergamon Press, 1965 (English translation of Russian book: В. С. Пугачев: Теория случайных функций и ее применение к задачам автоматического управления. Физматгиз, 1957, 1960, 1962).

[31] Shepp L. A.: Radon-Nikodym Derivatives of Gaussian Measures. Annals of Math. Statistics *37* (1966), 2, 321—354.

[32] Rozanov Yu. A.: On the Density of Gaussian Distributions and Wiener – Hopf Integral Equations. Teoria Veroyatnostei i ee Primeneniya *11* (1966), 1, 170—179 (in Russian).

[33] Pugachev V. S.: Methods for Solving Sets of Integral Equations Occuring in Prob'ems of Optimization of Multidimensional Systems. Proceedings of the Sixth All-Union Conference on Probability Theory and Mathematical Statistics (1960). Vilnius 1962, 233—237 (in Russian).

[34] Sysoev L. P.: Estimates of Parameters of Signals Modulated by Random Processes. Avtomatika i Telemekhanika *26* (1965), 10, 1255—1261 (in Russian).

[35] Pugachev V. S.: An Effective Method for Finding a Bayes Decision. Transactions of the Second Prague Conference on Information Theory, Statistical Decision Functions, Random Processes (1959). Prague 1960, 531—540 (in Russian).

[36] Pugachev V. S.: A method for Determining Optimum Systems Using General Bayes Criteria. IRE Transactions on Circuit Theory *CT-7* (1960), 4, 491—505.

[37] Pugachev V. S.: A Method for Determining an Optimal System Using a General Bayes Criterion. Izvestia Akademii Nauk SSSR, OTN, Energetika i Avtomatika (1960), 2, 83—97 (in Russian).

Optimální učící se systémy

V. S. PUGAČEV

V článku se probírají optimální Bayesovy učící se systémy a základní vlastnosti jejich algoritmů z hlediska obecné teorie učících se systémů. Diskutují se obecné koncepce učícího se systému, učitele, forem učení se, typy učitelů atd. Uvažuje se obecný případ reálného učitele řídícího procesy s náhodnými chybami. Případ ideálního učitele, který se dříve uvažoval výlučně, je pojímán jako zvláštní případ. Zavádí se míra odlišnosti učícího se systému od optimálního systému s úplnou informací.

V. S. Pugachev, Corresponding Member of the Academy of Sciences of the USSR, Institute of Control Problems, 81 Profsoyuznaya Street, Moscow V-485. USSR.