SAP
PAPER No. TA5-4

A MIXED STOCHASTIC OPTIMIZATION ALGORITHM AND ITS APPLICATIONS IN PATTERN RECOGNITION

L. P. Devroye
Osaka University, Japan

ABSTRACT

The problem of the minimization of a functional of several parameters is treated.It is supposed that the gradient of the functional is known but the stochastic approximation method cannot be applied since its convergence is not guaranteed due to the form of the functional.A mixed random search-stochastic approximation method insuring this convergence is developed.Several applications in the field of pattern recognition are highlighted.

INTRODUCTION

Consider the problem of the minimization of a functional $I(\underline{w})$ of n variables $w_1,\dots,w_n$ with $\underline{w} = \{w_1,\dots\dots,w_n\}$.It is supposed that $I(\underline{w})$ can be written as in (1) where $\underline{x}=\{x_1,\dots\dots,x_m\}$ is a vector of stationary random sequences or processes with distribution $p(\underline{x})$.

$$(1) \quad I(\underline{w})=\int_{\underline{x}} Q(\underline{x}/\underline{w})p(\underline{x})d\underline{x} = E_{\underline{x}}\{Q(\underline{x}/\underline{w})\}$$

If the function $Q(\underline{x}/\underline{w})$ is analytically given and differentiable,-denote the gradient by $\underline{\nabla}_{\underline{w}}Q(\underline{x}/\underline{w})$ - **the Robbins-Monro algorithm of stochastic** approximation [1] could be used for the minimization of (1).The convergence of the algorithm is the object of many papers [2],[4],etc. ,surveyed for instance in [10].Later,some authors modified the algorithm in order to speed up the rate of convergence while others dealt with the conditions of convergence if the roots of (2) are nonunique [3],[5-7] :

$$(2) \quad \underline{\nabla}_{\underline{w}}I(\underline{w}) = E_{\underline{x}}\{\underline{\nabla}_{\underline{w}}Q(\underline{x}/\underline{w})\} = \underline{0}$$

In most of the publications,it is required that $I(\underline{w})$ does not increase faster than a quadratic function with respect to $\underline{w}$ for $\underline{w}\to\infty$.In addition, there are conditions of convexity and unimodality which limit the class of functionals to be optimized by this stochastic approximation procedure.

The purpose of this paper is to present an algorithm for the extremization of (1) under the most general conditions on $I(\underline{w})$ (i.e.multimodal;not satisfying the conditions of convergence for the Robbins-Monro algorithm;...).Of course,for such a general purpose,the zero-gradient condition of optimality (2) won't be directly useful since (2) is also satisfied for local minima.We will present a mixed optimization scheme consisting of three basic parts : 1°)a gradient algorithm of the Robbins Monro type ;2°)a random search algorithm;3°)a probabilistic automaton which controls the choice of random search or gradient search at each moment.

Thereafter,some applications in the field of pattern recognition and feature extraction are highlighted without attempting,however,to improve known techniques in these areas.

RANDOM SEARCH

Let the functional $I(\underline{w})$ be defined in a closed bounded domain W of the Euclidean space R^n.A large group of direct optimization techniques is called random search.The most commonly used random search algorithms proceed as follows [9],[11-14] : let $\underline{w}(0)$ be chosen randomly in W.At the j-th iteration of the search,a test point $\underline{w}^*(j+1)$ is selected at random in accordance with a probability density function $f(\underline{w}^*/B_j)$ which is concentrated on W.B_j denotes the set of all the data that were used or known up to this j-th instant of time.Included in this information is the best estimate of the minimum,usually referred to as basepoint,$\underline{w}(j)$. Samples of size λ_j -a sample is the value of $Q(\underline{x}/\underline{w})$ for an independently and randomly generated $\underline{x}$ - are drawn from the distributions at $\underline{w}(j)$ and $\underline{w}^*(j+1)$ respectively.The sample means are denoted by $q(\underline{w}(j),\lambda_j)$ and $q(\underline{w}^*(j+1),\lambda_j)$ respectively. The basepoint is updated by the simple rule :

$$(3) \quad \underline{w}(j+1) = \begin{cases} \underline{w}^*(j+1) & \text{if } q(\underline{w}^*(j+1),\lambda_j) < q(\underline{w}(j),\lambda_j) - \varepsilon_j \\ \underline{w}(j) & \text{otherwise} \end{cases}$$

where $\varepsilon_j>0$ is a threshold.Usually,a solution region in W is defined by the minimum level I_M : I_M is such that for every j=1,2,... (4) holds.

$$(4) \quad \mathrm{Prob}\{I(\underline{w}^*(j+1))< I_M+\delta\} > 0$$

where δ is an arbitrary positive number.The so-called "solution region" contains all the points $\underline{w}$ for which $I(\underline{w})<I_M$.Condition (4) has two important consequences : first,$f(\underline{w}^*/B_j)$ should be positive for every $\underline{w}^*$ in W and for every j=1,2,... .Second, it will be impossible to detect minima that are defined on manifolds of dimension less than n since their measure in W is zero and,consequently,(4) is not influenced by such points.Thus,we may write that :

$$(5) \quad f(\underline{w}^*/B_j)> 0 \quad \text{for all } \underline{w}^*\in W \text{ and all } j=1,2,..$$

Another condition required by Gurin [13] concerns the variances of the random variables $q(\underline{w}(j),\lambda_j)$ and $q(\underline{w}^*(j+1),\lambda_j)$,denoted by $\sigma^2(\underline{w}(j),\lambda_j)$ and

$\sigma^2(\underline{w}^*(j+1),\lambda_j)$ respectively.Indeed,it is required that these variances tend to zero for $\lambda_j \to \infty$ .Of course,this condition of boundedness of the variance is satisfied in most of the problems that are dealt with.The most heavy requirements of [13] are well that

$$(6) \qquad \lim_{j\to\infty} \lambda_j = \infty$$

and that for every set of positive integers k_1,k_2,k_3 :

$$(7) \qquad \lim_{\frac{k_3}{k_2}\to 0} \frac{\Sigma_{k_3}}{\sum_{k=k_1+1,}^{k_1+k_2} \lambda_k} = 0$$

where the counter in (7) denotes any sum of k_3 terms of the sum in the denominator.If the average $\overline{\underline{w}}(j)$ is defined as :

$$(8) \qquad \overline{\underline{w}}(j) \triangleq \left(\sum_{k=1,}^{j} \lambda_k\right)^{-1} . \sum_{k=1,}^{j} \lambda_k . \underline{w}(k)$$

then,Gurin[13] proved that

$$(9) \qquad \lim_{j\to\infty} \mathrm{Prob}\{ |\overline{\underline{w}}(j)-\underline{w}_0|<4\delta\} = 1$$

where $\underline{w}_0$ is supposedly the unique global minimum of $I(\underline{w})$.If there is no uniqueness (i.e.the solution region defined by I_M consists of disjoint nonempty sets of points),then the convergence should be given preferably in terms of an "average performance" instead of an "average basepoint".The proof of convergence of an average performance for algorithm (3) is due to Saridis and Gilbert[14].In both publications [13-14],the threshold was constant during the search:$\varepsilon_j=\varepsilon_0$ for all j=1,2,... . Since ε_j appears in the limit for the average performance [14],[9] and since its value should be estimated sometimes without any good knowledge of the range of $I(\underline{w})$,it is better if ε_j is allowed to vary or to be updated during the search.This case is included in the most recent proof of convergence [9],and remarkable joint requirements upon the variations of the sequences$\{\lambda_j\}$ and $\{\varepsilon_j\}$ were obtained.At the same time,(5) could be replaced by the weaker condition :

$$(10) \qquad \sum_{j=1,}^{\infty} f(\underline{w}^*/B_j) = \infty \text{ for all } \underline{w}^* \in W$$

But,for the sake of simplicity,suppose here that the threshold is fixed $(\varepsilon_j=\varepsilon_0)$ and assume that the other conditions of Gurin are satisfied,then,in [9] it is proved that

$$(11) \qquad \lim_{j\to\infty} E\{\mathrm{Prob}\{I(\underline{w}(j))<I_M+\varepsilon_0\}\} = 1$$

where I_M satisfies (12) for $\underline{w}$ uniformly distributed in W and for every,however small, $\eta>0$:

$$(12) \qquad \mathrm{Prob}\ \{I(\underline{w})< I_M+\eta\} > 0$$

MIXED SEARCH

Suppose that $f(\underline{w}^*/B_j)$ is a mixture of the form :

$$(13) \qquad f(\underline{w}^*/B_j)=p(j).f_c(\underline{w}^*/B_j) +(1-p(j)).g(\underline{w}^*/B_j)$$

where $f_c(\underline{w}^*/B_j)$ and $g(\underline{w}^*/B_j)$ are probability density functions with domain not greater than W.Let $f_c(\underline{w}^*/B_j)$ satisfy (5).Then,of course,(13) also satisfies (5) if there exists a constant a such that :

$$(14) \qquad 1 \geqslant p(j) \geqslant a > 0 \quad \text{for all } j=1,2,...$$

But what is the meaning of (13) ? p(j) can be regarded as the probability of selection by an automaton of the strategy "generate $\underline{w}^*$ according to $f_c(\underline{w}^*/B_j)$".This probabilistic automaton has clearly only two actions -the second action being the decision to generate $\underline{w}^*(j+1)$ according to $g(\underline{w}^*/B_j)$ - and its only defining parameter,in casu p(j), is subjected to the constraint (14).Notice that (5) is satisfied for all possible density functions $g(\underline{w}^*/B_j)$,and in fact,$g(\underline{w}^*/B_j)$ need not be a probability density function in the classical sense of the word.Indeed,$g(\underline{w}^*/B_j)$ is only a symbolism for the event of the generation of $\underline{w}^*(j+1)$ by means of "another source of information" (which might even be a human intervention in the search process if there is a possible interaction between operator and computer).In the context of this paper,$g(\underline{w}^*/B_j)$ should of course incorporate the information contained in the given gradient $\underline{\nabla}_{\underline{w}} Q(\underline{x}/\underline{w})$. Since the comparison between $\underline{w}(j)$ and $\underline{w}^*(j+1)$(3) is based upon λ_j samples,it is intuitively felt that the effort of generating $\underline{w}^*(j+1)$ by the process symbolized by $g(\underline{w}^*/B_j)$ should be related in one way or another to this number λ_j.

Thus,we have here the special situation that the use of either gradient search alone or pure random search alone is bad since,in the first case,it is not sure that the convergence is towards the global minimum and in the second case,the valuable information,carried by the known gradients,is untouched.

For instance,let the operation symbolized by $g(\underline{w}^*/B_j)$ be defined as a gradient descent (15),starting from $\underline{w}(j)$ and broken off after μ_j iterations at $\underline{w}^*(j+1)$:

$$(15) \qquad \begin{aligned} &\underline{v}(0) = \underline{w}(j) \\ &\underline{v}(k) = \underline{v}(k-1)-\gamma(k).\underline{\nabla}_{\underline{w}} Q(\underline{x}(j,k)/\underline{v}(k-1)) \\ &\qquad\qquad\qquad \text{for } k=1,...,\mu_j \\ &\underline{w}^*(j+1) = \underline{v}(\mu_j) \end{aligned}$$

with the sequence $\{\gamma(k)\}$ satisfying :

$$(16) \qquad \gamma(k)>0 \ ; \sum_{k=1,}^{\infty} \gamma(k) = \infty \ ;\sum_{k=1,}^{\infty} \gamma^2(k)< \infty$$

It is well known that the sequence $\underline{v}(k)$ generated by (15) tends towards a stationary point of $I(\underline{w})$ for $k\to\infty$ and under certain assumptions on $Q(\underline{x}/\underline{w})$ [1-7]..Since λ_j tends to infinity for $j\to\infty$ (6), $\mu_j=K.\lambda_j$ (as we pointed out,the efforts should be related) also will become increasingly great.Thus, for large μ_j,the result $\underline{w}^*(j+1)$ of operation (15) will be close to the local minimum in which domain of attraction $\underline{w}(j)$ lies.Thus,besides the property that the sequence $I(\underline{w}(j))$ will converge for $j\to\infty$ in the sense of (11-12),our scheme will have the special additional feature that the basepoint $\underline{w}(j)$ will,for all large j,be close to one of the local minima.

The simple probabilistic automaton with parameter p(j) is allowed to change its structure,in casu p(j),during the search within the bounds (14),but it goes beyond the scope of this paper to propose rules for the continued learning of p(j) during the search.Let us now look at the density $f_c(\underline{w}^*/B_j)$ (13).The simpliest distribution satisfying (5) is of course the uniform distribution in W,which is often satisfactory for the purpose of multimodal search or rough localization of the global mini-

mum.If desired,more sophisticated densities [9], [11-12] could be used.As a rule,these densities are gaussian with fixed or adaptive parameters (mean,covariance matrix) or a mixture of gaussian and uniform densities with adaptive weights for each population in the mixture.In our problem with given gradients however,it is redundant to use such complex $f_c(\underline{w}^*/B_i)$.
A more general algorithm is obtained if the coefficients $\gamma(k)$ are also dependent upon the overall iteration counter j.Notice also that other powerful deterministic optimization techniques such as partan,... could be incorporated in the operation $g(\underline{w}^*/B_i)$,for instance in direct search problems,i.e. when the gradient is unknown.To the authors knowledge,this is the second approach of using all our know-how of deterministic optimization in stochastic problems.In the first approach,Kushner [6] proposed a combination of usual direct search techniques with one dimensional searches in given directions by a stochastic approximation algorithm of the type of Kiefer-Wolfowitz [8].Our approach is on the one hand applicable to a larger class of functionals,but,on the contrary,the cost of optimization increases due to the repeated comparisons (3).

FEATURE EXTRACTION

The feature extraction problem in pattern recognition may be approached from two points of view.Either an attempt is made to transform the sample space in such a way that the classes are more easily separable,or an attempt is made to reduce the dimensionality of the sample space.As stated in [15-16],the main difficulty in feature extraction is that the features must be evaluated in terms of the decision stage rather than on their own.Therefore it might be interesting if the feature extractor and the decision maker could be deigned simultaneously in order to extract only those features that are really relevant,i.e.that are useful in the decision process.
Let $\underline{z}$ be an input vector of measured or computed data $z_1,\dots,z_N$.The N-dimensional feature space is denoted by Z.It is assumed that there are a sufficient number of features in order to allow the use of a linear classifier.This can often be achieved by defining new features that are nonlinear functions of other features.Let us consider the 2 class problem with $y(\underline{z})\in\{-1,+1\}$ denoting the known desired classification of $\underline{z}$.The linear classifier is characterized by a N-dimensional weight vector $\underline{g}$ [20] such that $\underline{z}$ is classified into class $\mathrm{sgn}(\underline{z}^T.\underline{g})$ where sgn(.) denotes the sign function. The weight vector $\underline{g}$ is usually chosen in such a way that a given functional (17) is minimized :

$$(17)\qquad E_{\underline{z}}\{H(\underline{z}^T\underline{g};y(\underline{z}))\} = \int_Z H(\underline{z}^T\underline{g};y(\underline{z}))p(\underline{z}).d\underline{z}$$

$p(\underline{z})$ denotes the unknown probability density function of $\underline{z}$ in Z.In [20-21],several loss functions H(a,b) are surveyed. Most commonly used are $(a-b)^2$, $|a-b|$ and the "misclassification" $|\mathrm{sgn}\ a - \mathrm{sgn}\ b|\ /2$.
Let us introduce a new set of N variables $\underline{q} = \{q_1, q_2,..,q_N\}$,that will be used to decide which features will be used in the pattern recognizer and which won't be used.We further define the set of probabilities $\{q_i^*,i=1,..,N\}$ as :

$$(18)\qquad q_i^* \triangleq \mathrm{Min}\{1;|q_i|\}$$

During training, the feature extraction process is regarded as a random process : given the set $\underline{q}^*$,a "realization" of the feature extracting process is denoted by the random variable $\underline{r}=\{r_1,\dots,r_N\}$ with elements $r_i \in\{0,1\}$:the i-th feature is used in the decision stage ($r_i=1$) with probability $\alpha q_i^*+\beta$ and is omitted with the complementary probability.We will require thatα andβ be strictly positive and their sum should be strictly smaller than one.By $p(\underline{r}/\underline{q})$, we denote the probability density function of $\underline{r}$, given $\underline{q}$:

$$(19)\qquad p(\underline{r}/\underline{q})= \prod_{i=1,}^{N} (\delta(r_i)+(\alpha q_i^*+\beta)(\delta(r_i-1)-\delta(r_i)))$$

where $\delta(.)$ denotes the Dirac-function.Let further R denote the diagonal matrix with elements r_i.$E_{\underline{r}/\underline{q}}$ and $E_{R/\underline{q}}$ denote the expectations over all $\underline{r}$, resp. R for given $\underline{q}$.This randomization implies a redefinition of our goal-functional :

$$(20)\qquad E_{R/\underline{q}}\{E_{\underline{z}}\{H(\underline{z}^T R\underline{g};y(\underline{z}))\}\} = \int p(R/\underline{q}).[\int_Z H(\underline{z}^T R\underline{g};y(\underline{z})).p(\underline{z}).d\underline{z}].dR$$

Notice that

$$(21)\qquad \int\frac{\partial}{\partial q_i^*}.p(R/\underline{q})dR = E_{R/\underline{q}}\{\frac{\alpha(r_i-\alpha q_i^*-\beta)}{(1-\alpha q_i^*-\beta)(\alpha q_i^*+\beta)}\}$$

and,from (18) :

$$(22)\qquad \frac{\partial}{\partial q_i}p(R/\underline{q}) = \begin{matrix} \frac{\partial}{\partial q_i^*}p(R/\underline{q}).\mathrm{sgn}\ q_i & \text{if } |q_i|<1 \\ 0 & \text{if } |q_i|>1 \end{matrix}$$

Obviously,(20) is minimal for $q_i^*=1,i=1,.,N$ since each feature contributes -however small this contribution may be-to the minimization of (20) for usual choices of the loss function H.In order to select the most relevant features,the broader functional (23) is proposed :

$$(23)\qquad E_{R/\underline{q}}\{E_{\underline{z}}\{H(\underline{z}^T R\underline{g};y(\underline{z}))\}\} + A.\sum_{i=1,}^{N} d_i.|q_i|$$

where the number $d_i>0$ is proportional with the relative cost of the use of the i-th feature.A>0 is a constant,which value can be used to decrease or increase the number of features that are extracted.Indeed,the second term in (23) is proportional with the cost of "sampling" and the first term gives the cost of misclassification with the given choice of $\underline{g}$ and $\underline{q}$.
Although $p(R/\underline{q})$ is known (19),the computation of the expectation $E_{R/\underline{q}}$ is very difficult,especially if N is very high:each time,2^N products of N factors have to be computed.Therefore,$p(R/\underline{q})$ is, just as $p(\underline{z})$,treated as an unknown density.The gradient of (23) with respect to $\underline{g}$ is equal to the mean of (24) over $p(R/\underline{q})$ and $p(\underline{z})$ if H'(a,b) denotes the partial derivative of H(a,b) with respect to a:

$$(24)\qquad H'(\underline{z}^T R\underline{g};y(\underline{z})).R\underline{z}$$

Using (19-22),we also know that (25) defines a random variable with expected value over $p(R/\underline{q})$ and $p(\underline{z})$ equal to the gradient of functional (23) with respect to q_i.With the overall variable $\underline{w}$ being 2N-dimensional $\underline{w}=(\underline{g};\underline{q})$,with the random process $\underline{x}$ (1) being represented by the independent processes $\underline{z}$ and $\underline{r}$: $\underline{x}=(\underline{z};\underline{r})$ and with the performance index $\overline{I}(\underline{w})$ being defined as in (23),the gradient estimates (24-25) could be used for a gradient descent of the Robbins-Monro type (15),but since convergence in the global sense-even if H(a,b) is convex- is not guaranteed due to the addition of the new set of variables $\underline{q}$,a mixed search can be applied in this case.

$$(25)\quad \begin{cases} \operatorname{sgn} q_i.\{Ad_i + \dfrac{\alpha(r_i-\alpha q_i*-\beta)H(\underline{z}^T R\underline{g};y(\underline{z}))}{(1-\alpha q_i*-\beta)(\alpha q_i*+\beta)}\} & \text{if } |q_i|<1 \\ \operatorname{sgn} q_i.\ Ad_i & \text{if } |q_i|>1 \end{cases}$$

The main result is thus the simultaneous performance-directed design of a flexible feature extractor and a classifier.Since N is normally very high,this technique is only efficient when the computational algorithm is rather simple.The special field of application is when maximum profit is desired with the aid of the least possible number of features.From that point of view,there is a great flexibility in the design due to the addition of a cost-measuring part in (23).Moreover,sometimes it may be great advantage that the input vectors can be sequentially processed.

SUPERVISED LEARNING

It often happens in pattern recognition that the patterns $\underline{z}$ are not suitable for the construction of simple separating surfaces.The input patterns $\underline{z}$ are then transformed in a nonlinear way and a new pattern $\underline{\psi}(\underline{z})$ is generated which is defined in the "rectified space" Ψ[17] : $\underline{\psi}(\underline{z})=\{\psi_1(\underline{z}),\ldots,\psi_s(\underline{z})\}$. The space Z ($\underline{z}\in Z$) is N-dimensional while Ψ^s is s-dimensional.It is supposed that a trainer is available,i.e.the classification $y(\underline{z})$ of $\underline{z}$ is known.We will restrict ourselves again to the two-class problem with $y(\underline{z})\in\{0,1\}$.The results of this section can easily be extended towards more-class pattern-recognition,nonlinear function restoration and nonlinear estimation problems.The relationship between these problems is explained in [18]. Pattern $\underline{z}$ is classified into class $\operatorname{sgn}(\underline{g}^T\underline{\psi}(\underline{z}))$, where $\underline{g}$ is obtained from the minimization of a functional

$$(26)\quad E_{\underline{z}}\{H(\underline{g}^T\underline{\psi}(\underline{z});y(\underline{z}))\} = \int_Z H(\underline{g}^T\underline{\psi}(\underline{z});y(\underline{z}))p(\underline{z})d\underline{z}$$

where $p(\underline{z})$ and H(a,b) are defined as in the previous section.Stochastic approximation algorithms for the minimization of (26) have been extensively studied [20-21].Let us now broaden the class of functions $\psi_i(\underline{z})$ by defining a new set of parameters,say $\underline{u}_1,\ldots,\underline{u}_s$,where $\underline{u}_i$ is a set of characteristics influencing the transformation $\psi_i(\underline{z})$.The set $\{\underline{u}_1,..,\underline{u}_s\}$ will be denoted by U,and thus : $\underline{\psi}(\underline{z},U)=\{\psi_1(\underline{z},\underline{u}_1),\ldots,\psi_s(\underline{z},\underline{u}_s)\}$ .The gradient of (26) with respect to $\underline{g}$ and $\underline{u}_i$ are given in (27) and (28) respectively :

$$(27)\quad E_{\underline{z}}\{H'(\underline{g}^T\underline{\psi}(\underline{z},U);y(\underline{z})).\underline{\psi}(\underline{z},U)\}$$

$$(28)\quad E_{\underline{z}}\{H'(\underline{g}^T\underline{\psi}(\underline{z},U);y(\underline{z})).g_i.\nabla_{\underline{u}_i}\psi_i(\underline{z},\underline{u}_i)\}$$

Though the gradients are known,the algorithms of [1-7] might not be applicable especially if $\underline{\psi}(\underline{z},U)$ is a very nonlinear transformation.However,the gradients can be used in the mixed search scheme with $\underline{w}=(\underline{g},\underline{u}_1,\ldots,\underline{u}_s)$,$\underline{x}=\underline{z}$ and $I(\underline{w})$ defined by

$$(29)\quad E_{\underline{z}}\{H(\underline{g}^T\underline{\psi}(\underline{z},U);y(\underline{z}))\}$$

Let,for instance,$\underline{u}_i$(i=1,..,s) be points in the N-dimensional input space Z,and let $\psi_i(\underline{z},\underline{u}_i)$ (i=1,....,s) be potential functions,monotonically nonincreasing from 1 to 0 for $|\underline{z}-\underline{u}_i|$going from 0 to ∞. In [22-23],several functions are suggested and compared such as $(1+\xi^{-2}|\underline{z}-\underline{u}_i|^2)^{-1}$;$\exp(-\xi^{-2}|\underline{z}-\underline{u}_i|^2)$ and $(1-\xi^{-2}|\underline{z}-\underline{u}_i|^2).\omega(1-|\underline{z}-\underline{u}_i|^2\xi^{-2})$ where $\omega(.)$ is the step function.The role of the constant ξ is that of the radius of influence of $\underline{u}_i$ in Z.The component g_i can be considered as the "grade of membership" of the environment of $\underline{u}_i$ in the classes (+1) or (-1),respectively for positive or negative values of g_i.The gradient $\nabla_{\underline{u}_i}\psi_i(\underline{z},\underline{u}_i)$ is,for the given functions,of the form $\phi(|\underline{z}-\underline{u}_i|).\underline{z}-\underline{u}_i$ where $\phi(.)$ is again a nonincreasing function with $\phi(0)<\infty$ and $\phi(\infty)=0$.The product $\underline{g}^T\underline{\psi}(\underline{z},U)$,used in the decision,is mainly influenced by the higher values $\psi_i(\underline{z},\underline{u}_i)$ and thus,the classification of $\underline{z}$ is for the given choice of functions $\underline{\psi}(\underline{z},U)$ mainly determined by those weights g_i which correspond to points $\underline{u}_i$ that are close to $\underline{z}$ in Z.

Since the dimension n=(s+1)N is rather high,the field of application is clearly in high-precision pattern recognizers where the cost of training is not an important factor.A very interesting application is when $H(a,b)=(a-b)^2$ and

$$(30)\quad \psi_i(\underline{z},U) = \begin{cases} 1 \text{ if } |\underline{z}-\underline{u}_i|^2=\operatorname*{Min}_{k=1,..,s} |\underline{z}-\underline{u}_k|^2 \\ 0 \text{ otherwise} \end{cases}$$

Notice that all the gradients in (27) and (28) are zero,except the gradient with respect to g_i which is equal to $(g_i-y(\underline{z}))$ for one particular $\underline{z}$.The application of the gradient algorithm would thus influence only one weight for each processed $\underline{z}$. The classification rule is simple since it reduces to sgn g_i in case $\underline{z}$ is closest to $\underline{u}_i$ (30).Thus,we established a "nearest neighbour" rule [19] with s "mother points" $\underline{u}_1,\ldots,\underline{u}_s$ that do not belong to the set of training samples.With equal ease,(30) can be modified in a K-nearest neighbor transformation.Since the number of mother points in the nearest neighbor rule is usually very high,a method was developed in [24] for editing between the mother points and thus selecting a representative subset of points.By allowing the use of non-sample points as mother points,probably a reduction of this number is still possible for equal prestations,especially since these points $\underline{u}_1,\ldots,\underline{u}_s$ are determined by virtue of a search process minimizing functional (29).Since (28) is zero,the location of the $\underline{u}_i$ can only be modified by the direct search part in the mixed search.

The latter system (29-30) can also be used for the restoration of a multidimensional function of $\underline{z}$ by means of s flat levels.The domain Z is partitioned into s parts by $\underline{u}_1,\ldots,\underline{u}_s$ according to a nearest neighbor rule.(29) is a least square error criterion if $H(a,b) = (a-b)^2$ and if $y(\underline{z})$ represents the desired function value in $\underline{z}$.The optimal partition with respect to (29) and the optimal levels g_i,i=1,..,s (also with respect to (29)) are found simultaneously and by a sequential processing of the samples $\underline{z},y(\underline{z})$.

SELF-LEARNING

If $y(\underline{z})$ is not known,the problem turns into an unsupervised learning problem [26].Tsypkin [25] formulated the problem as follows :let the N-dimensional pattern space Z be partitioned in s disjoint parts Z_k $(k=1,..,s)$.For each region Z_k,a loss function $H_k(\underline{z},\underline{u}_1,...,\underline{u}_s)=H_k(\underline{z},U)$ is defined where $U=\{\underline{u}_1,...,\underline{u}_s\}$ are certain parameters of the regions and $\underline{z}$ denotes a pattern from the unknown distribution $p(\underline{z})$.The functions $H_k(\underline{z},U)$ evaluate the loss when the pattern $\underline{z}\in Z_k$.Let $\psi_i(\underline{z},U)$ (31) define the decision rule ($\underline{z}$ is classified into class or cluster i iff $\psi_i(\underline{z},\overline{U})=1$) :

$$(31)\quad \psi_i(\underline{z},U) = \begin{cases} 1 & \text{if } H_i(\underline{z},U)=\underset{k=1,..,s}{\mathrm{Min}} H_k(\underline{z},U) \\ 0 & \text{otherwise} \end{cases}$$

Then,the average risk of misclassification equals:

$$(32)\quad E_{\underline{z}}\{\Sigma H_k(\underline{z},U)\psi_k(\underline{z},U)\}=\Sigma\int_{Z_k} H_k(\underline{z},U).p(\underline{z})d\underline{z}$$

where Σ is for $k=1,..,s$.In general,(32) and its gradient with respect to $\underline{u}_i$ do not satisfy the usual Robbins-Monro conditions.As pointed out in [26],the convergence of the gradient algorithm proposed in [25] remains an open problem,since the risk function (32) may be non-convex or multimodal.The interesting algorithms of Tsypkin can be practically used if they are integrated in the mixed search scheme.

CONCLUSION

We focused our attention on the one-step design of pattern recognition systems,i.e.when unknown parameters in the data analyzer (i.e.variables in the feature extraction or feature transforming processes) and in the recognizer (such as the weights of a linear discriminant function) are simultaneously optimized with respect to a single performance index.Most of the variational approaches to one of these problems can be considered as special cases of our approach.The presented technique of optimization is a combination of random search and stochastic approximation.

BIBLIOGRAPHY

[1]H.ROBBINS,S.MONRO:"A stochastic approximation method",Ann.Math.Stat.,vol.22.No.1,1951

[2]J.A.BLUM:"Approximation methods which converge with probability one",Ann.Math.Stat.,vol.25,No.2, 1954

[3]I.P.DEVYATERIKOV,A.I.KAPLINSKII,Ya.Z.TSYPKIN: "Convergence of learning algorithms",Autom.and Remote Control,vol.30,No.10,1969

[4]E.M.BRAVERMAN,L.I.ROZONOER:"Convergence of random processes in learning machines theory",Autom. and Remote Control,vol.30,No.1,pp.57-77,1969

[5]B.M.LITVAKOV:"Convergence of recurrent algorithms for pattern recognition learning",Autom.and Remote Control,vol.29,No.1,pp.121-128,1968

[6]H.J.KUSHNER:"Stochastic approximation algorithms for the local optimization of functions with nonunique stationary points",IEEE Trans.on Automatic Control,vol.AC-17,No.5,pp.646-654,1972

[7]T.P.KRASULINA:"Robbins-Monro processes in the case of several roots",Autom.and Remote Control, vol.33,No.4,pp.580-585,1972

[8]E.KIEFER,J.WOLFOWITZ:"Stochastic estimation of the maximum of a regression function",Ann.Math. Stat.,vol.23,No.3,1952

[9]L.P.DEVROYE:"Multimodal stochastic optimization of arbitrary performance indices",submitted for publication in the IEEE Trans.on Systems,Man and Cybernetics.

[10]N.V.LOGINOV:"The methods of stochastic approximation",Autom.and Remote Control,vol.27,No.4, 1966

[11]J.MATYAS:"Random optimization",Autom.and Remote Control,vol.26,No.2,pp.244-251,1965

[12]L.D.COCKRELL,K.S.FU:"On search techniques in adaptive systems",Techn.Rept.TR-EE-70-1,Purdue University,Lafayette,Indiana,1970

[13]L.S.GURIN:"Random search in the presence of noise",Engineering Cybernetics,vol.4,No.3,pp.252-260,1966

[14]G.N.SARIDIS,H.D.GILBERT:"On the stochastic fuel-regulation problem",Techn.Rept.TR-EE-68-9, Purdue University,Lafayette,Indiana,1970

[15]G.NAGY:"State of art in pattern recognition", Proceedings of the IEEE,vol.57,No.5,pp.836-862, 1968

[16]M.D.LEVINE:"Feature extraction:a survey",Proceedings of the IEEE,vol.57,No.8,1969

[17]M.A.AIZERMAN,E.M.BRAVERMAN,L.I.ROZONOER:"Theoretical foundations for the method of potential functions in the problem of training of machines to separate functions into classes",Autom.and Remote Control,vol.25,No.6,pp.917-936,1964

[18]Ya.Z.TSYPKIN:"Adaptation,training and self-organization in automatic systems",Autom.and Remote Control,vol.27,No.1,pp.16-51,1966

[19]T.M.COVER,P.E.HART:"Nearest neighbour pattern classification",IEEE Trans.on Information Theory, vol.IT-13,No.1,pp.21-27,1967

[20]I.P.DEVYATERIKOV,A.I.PROPOI,Ya.Z.TSYPKIN:"Iterative learning algorithms for pattern recognition",Autom.and Remote Control,vol.28,No.1,pp.108-117,1967

[21]Y.C.HO,K.AGRAWALA:"On pattern classification algorithms.Introduction and survey",IEEE Trans.on Automatic Control,vol.AC-13,No.6,pp.676-690,1968

[22]O.A.BASHKIROV,E.M.BRAVERMAN,I.B.MUCHNIK:"Potential function algorithms for pattern recognition learning machines",Autom.and Remote Control, vol.25,No.5,pp.692-695,1964

[23]R.M.MALKINA,A.A.PERVOZVANSKIY:"Approximations in problems of constructions of functions and of pattern recognition",Engineering Cybernetics,vol. 7,No.1,1969

[24]D.L.WILSON:"Asymptotic properties of nearest neighbor rules using edited data",IEEE Trans.on Systems,Man and Cybernetics,vol.SMC-2,No.3,pp.408-421,1972

[25]Ya.Z.TSYPKIN:"Self-learning.What is it?",IEEE Trans.on Automatic Control,vol.AC-13,No.6,pp.608-612,1968

[26]A.A.DOREFOYUK:"Automatic classification algorithms",Autom.and Remote Control,vol.32,No.12, pp.1928-1958,1971