REMARKS ON THE OPTIMUM RATE OF CONVERGENCE OF THE ON-LINE IDENTIFICATION OF NON-STATIONARY SYSTEMS

By

I. BENCSIK

Department of Automation,
Technical University, Budapest

(Received March 18, 1972)
Presented by Prof. Dr. F. CSÁKI

Generally the optimum rate of convergence of recurrent identification algorithms for the adaptive control of noisy, sampled, *non-stationary*, *non-linear* systems is difficult to ensure due to the great number of contradictory circumstances. By modifying the rate of convergence based on the entropy variations of the samplongs, the optimum convergence factor in the stochastic sense can be numerically determined in a way that the informations arriving from the system could be included in the model at a maximum rate.

Introduction

It is a generally known fact that in some cases the recurrent algorithms are slowly convergent in the environment of the optimalization extreme value criterion, therefore attention is concentrated today to the acceleration of the convergence of these identification algorithms. Improvements of the algorithm suggested by various authors [1, 2, 3] are based on different principles, therefore this paper suggests a principle by which the efforts oriented towards the acceleration of the convergence might eventually be brought in harmony.

The on-line parameter-estimating algorithms, by interpreting the task of identification as formulated by ASTRÖM and EYKHOFF [4] take the informations stored in the preceding data with a lower weighting into consideration; this process is called also forgetting in the literature.

The aspects of determining the optimum forgetting and simultaneously the optimum rate of convergence are:

— The structure of the approximation may contain systematic errors due to the incomplete knowledge of the noise or the variables.

— In the case of parameters varying in time the rate of forgetting must be adapted to the rate of the parametric variation; further the presence of noise and the increase of noise dissipation, respectively, hamper the fast follow-up. Convergence can be ensured when the time constant of the parametric variations exceeds the noise correlation time.

— The need for following up the parametric variations due to the work point variation in non-linear systems can be determined by the average values of the input signals by the variation of their dispersion and the rate of convergence is to be adapted to the rate of the working point variation.

— The necessity of forgetting can be determined also by the deterioration of the closeness of approximation.

— The rate of convergence is optimum when the error probability of the approximation has its minimum in a minimum number of steps. Such a choice of the optimalization criterion results in the highest processing rate of new data arriving from the system.

In the following the amount of information of a sample, i.e. the entropy is determined, with consideration to the aspects listed above, then the relationship between the variation of the entropy and the factor of convergence is studied.

It is assumed that,

— the noise is not correlated with the signal, is zero-centered and finite, and its variance is known;

— the average values, the variations of the input and output signals are known, or they can be determined in the experimental phase.

Entropy

In this case, for its[1] definition that of FISHER [6] is suggested as against correct that given by SHANNON [5]:

$$I_H = \int_{-\infty}^{\infty} \left(\frac{\partial \ln f(y)}{\partial m} \right)^2 f(y)\, dy \tag{1}$$

where $f(y)$ — distribution density function

m — expected value of the distribution

y — output signal of the system, or of the model. For example, for the calculation of the entropy with normal distribution of y:

$$f(y) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[- \left(\frac{y-m}{\sqrt{2}\,\sigma} \right)^2 \right] \tag{2}$$

and

$$\ln f(y) = -\frac{1}{2} \ln 2\pi - \ln \sigma - \left(\frac{y-m}{\sqrt{2}\,\sigma} \right)^2 \tag{3}$$

[1] Indeed it is information.

hence

$$\frac{\partial \ln f(y)}{\partial m} = \frac{y-m}{\sigma^2} . \tag{4}$$

Therefore

$$I_H = \int_{-\infty}^{\infty} \left(\frac{y-m}{\sigma^2}\right) \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\left(\frac{y-m}{\sqrt{2}\,\sigma}\right)^2\right] dy = \frac{1}{\sigma^2} . \tag{5}$$

In the case of δy^2 constant, the entropy of measured data is seen to be identical between the samples. In the case of parameters varying with time, or of non-linear systems approximated as linear at the working point, — e.g. by a polynomial, — their effect on the entropy can be taken into consideration by weighting factors; so

— the deviation from the average rate of parametric variation is:

$$\alpha = \alpha(\ddot{u}, \ddot{y}) \tag{6}$$

where y — output signal

u — input signal vector

— the variation of the parameter average values, their fluctuation can be expressed as:

$$\beta = \beta(\dot{\bar{u}}, \dot{\bar{y}}, \dot{\bar{\sigma}}^2, \dot{\bar{\sigma}}_u^2) \tag{7}$$

Therefore in case of normal distribution:

$$I_H = \frac{1}{\alpha^T \beta\, \sigma_y^2 + \gamma\, \sigma_\xi^2} \tag{8}$$

α and β may be identified conceptually with the sensitivity coefficient, see e.g. in [7], and $\gamma = \gamma(\dot{\bar{\sigma}}_\xi^2)$. In α and β the sensitivity of identification for the individual parameters can be increased by the appropriate choice of the weighting factors, but for average values: $\alpha^T \beta = 1$. On the basis of the above said, α and β can be constructed to meet the requirements of the involved task.

The modification of the rate of forgetting on the basis of the numerical evaluation of the entropy:

— in case of $I_H(t)$ constant no forgetting is necessary; in the experimental phase forgetting of the ratio $1/t$ is necessary [2]; in the adaptive case forgetting has an average rate;

— in the case when $I_H(t)$ varies the rate of forgetting must be increased.

The necessity of forgetting follows also from the deterioration of the closeness of approximation and it can be determined by entropy variation of the information stored in the model. In the tested case, where the results supplied by the model made adequate to the system and the statistical characteristics of the output signal agree the entropy is:

$$I_m = \frac{1}{\sigma_m^2}\,. \tag{9}$$

When I_m decreases — with the deterioration of the approximation — the rate of forgetting must be increased and vice versa, so I_H and I_m have opposite effects on the convergence.

The optimum rate of convergence

The closeness of approximation is characterized by the quotient of both entropies: I_H/I_m, — which when considered as an argument of the F-test fundamental in variance analysis yields — the error probability.

At the optimum rate of convergence the new informations from the system get into the model at a minimum number of steps, if

$$\frac{1}{T}\int_{t-T}^{T} \frac{I_H}{I_m}\, dt = \text{extr.} \tag{10}$$

The value of T is given by the time constants of the process and it is advised to choose it higher by one order of magnitude than the smallest time constant of the system and the empirical average variances may be estimated on the basis of $N = 10\pi/T$, the number of samplings belonging to time T.

The quotient I_H/I_m has a known distribution, a known expected value and variance, as necessary for the evaluation of (10).

In the tested case of normally distributed the quotient is of second order beta-distribution (see e.g. [9]):

$$P\left(\frac{I_H}{I_m} < z\right) = \frac{\Gamma(N-1)}{\Gamma\left(\frac{N-1}{2}\right)} \int_0^z \frac{t^{\frac{N-3}{2}}}{(1+t)^{N-1}}\, dt \tag{11}$$

with Γ being the gamma function known from [10] and $z \geq 0$.

In the case of the model adequated to the system, for $N > 5$ the expected value and the variance are:

$$M\left\{\frac{I_H}{I_m}\right\} = \frac{N-1}{N-3}$$
$$D^2\left\{\frac{I_H}{I_m}\right\} = \frac{4\,(N-1)\,(N-2)}{(N-3)^2\,(N-5)}\,. \tag{12}$$

Respectively, the tables referring to the beta distribution are found in [8].

In the knowledge of the extreme value and variance of (10) of the factor of convergence is to be modified in each step without increasing (10) i.e.

$$\frac{\partial}{\partial r(t)}\,\frac{I_H(t)}{I_m(t)} \leq 0 \tag{13}$$

where $r(t)$ is the factor convergence (or some element of the matrix of convergence). To evaluate (13) the functional relationship between the factor of convergence and entropy is to be specified.

Be c the parametric vector, then the presently known linear recurrent algorithms offer a general solution to the problem, in the on-line case,

$$\frac{\partial c}{\partial t} = \mathbf{R}\cdot\left|\frac{\partial I}{\partial t}\right| = -\,\mathbf{R}\left(\frac{\partial I}{\partial t}\right)\cdot A\left(\frac{\partial I}{\partial c}\right) \tag{14}$$

when $\mathbf{R}$ is the matrix of convergence factor and operator A is the algorithm. In the discrete case:

$$\Delta C = \mathbf{R}\cdot|\Delta I| \tag{15}$$

The studied convergence factor versus entropy function can be chosen on the basis of (14), so in the on-line case the variation of the elements of the indefinite matrix $\mathbf{R}$ is proportional with $|\Delta I|$:

$$\Delta r[n] = r[n]\,|\Delta I[n-1]|$$

where

$$\Delta I[n] = I[n] - I[n-1] \tag{16}$$

for calculating the rate of forgetting the following empiric method is proposed:

$$r\,[n+1] = r\,[n]\,\frac{\dfrac{1}{N}\displaystyle\sum_{i=1}^{N}\frac{I_H\,[n-i]}{I_m\,[n-i]}}{\dfrac{1}{N-1}\displaystyle\sum_{i=2}^{N}\frac{I_H\,[n-i]}{I_m[n-i]}} \tag{17}$$

or taking (12) into consideration:

$$r[n+1] = r[n] \frac{\frac{(N-1)^2}{N-3}}{\sum_{i=2}^{N} \frac{I_H[n-i]}{Im[n-i]}} . \tag{18}$$

The initial value of $r(n)$ may be ≥ 1 in the adaptive phase. The optimum convergence factor on basis of (16) and (17):

$$r_{\text{opt}}[n+1] = r[n+1]\,[1 + |\Delta I_H(n+1)|] \tag{19}$$

Finally, the advantage of the established convergence rate optimization criterion is that it is independent of the functional relationship involved in the problem, its disadvantage is that the distribution function of the output signals must be known. The studied normal distribution may be of help for approximating other, — e.g. Poisson, binomial — distributions.

Further, the determination of the weighting factors in (8) may be difficult, but as the accuracy of the approximation is satisfactorily characterized by the quotient L_H/I_m even without the factors, so the latter choice is advisable on the assumption that the structure of the model has been chosen conveniently.

Conclusions

The use of the entropy definition (1) due to FISHER for the solution of on-line identification tasks permits the choice of the optimum convergence; because of its advantageous properties the convergence optimization criterion with the expected value according to (10) can be chosen as identification quality criterion, too therefore its application in practice is suggested; on the other hand the algorithms (14) and (15) serving for parameter estimation based on the variation of the entropy, open a new research trend, if the entropy is expressed as a function of the sampled values of output and input signals.

The author express herewith his thanks to Professor Dr. F. Csáki for his supporting the publication of this paper and to Dr. J. Kocsis senior research worker, for his valuable advices.

Summary

The use of the entropy definition by to FISHER for the solution of on-line identification tasks permits the choice of the optimum convergence. Due to its advantageous properties, the statistical convergence optimization criterion with the known expected values can be chosen as identification quality criterion, too, therefore its application in practice is suggested. The algorithms serving for parameter estimation based on the variation of the entropy, are a new trend in research, if the entropy is expressed as a function of sampled values of output and input signals.

References

1. Fu, K. S.—Chien, Y. T.—Nikolic, Z. I.—Wee, W. G.: On the stochastic approximation and related learning techniques. Purdue University, April 1966.
2. Cipkin, Y.: There does exist a theory of the synthesic of optimal adaptation system. Automatika i Telemekhanika **1**, (1968). 108.
3. Sakrison, D. J.: Application of stochastical approximation method to system optimization. MIT. Techn. Repont 391, July 10, 1962.
4. Aström, K. J.—Eykhoff, P.: System identification. A survey, IFAC Prague, **1970**.
5. Shannon, C. E.: A Mathematical Theory of Communication. Bell System Techn. J. 27, 1948.
6. Fisher, R. A.: Proc, Cambr. Phil. Soc. **22**, 700 (1925).
7. Kocsis, J.: Process optimization by adaptive programming. IFAC, Prague, 1970.
8. Pearson, K.: Tables of the Incomplete Beta Function. Cambr. Univ. Press. 1934.
9. Vince, I.: Mathematical Statistics. Műszaki Könyvkiadó, Budapest, 1968. (In Hungarian)
10. Bronstein, I. N.: Manual of Mathematics. Műszaki Könyvkiadó, 1963. Budapest. (In Hungarian)

István Bencsik 1052 Budapest, XI., Garami E. tér 3, Hungary