KYBERNETIKA — VOLUME 13 (1977), NUMBER 4

On the Acceleration of Adaptation Processes by Two-Step Principles

KLAUS FRITZSCH

The convergence behaviour of two-step adaptation algorithms is discussed and it is shown that there are cases where this class of algorithms is superior to the classic one-step algorithms. The convergence speeds of both classes of algorithms are compared by the criterion of mean square deviation from the optimum point.

1. INTRODUCTION

In this paper, an adaptation algorithm is a procedure which finds the minimum $\boldsymbol{w}^*$ of a performance functional

$$(1.1) \qquad J(\boldsymbol{w}) = E_{\boldsymbol{X}}\{Q(\boldsymbol{x}, \boldsymbol{w})\} ,$$

using only information about the random variables $Q(\boldsymbol{x}, \boldsymbol{w})$, $\boldsymbol{x} \in \boldsymbol{X}$ and/or their derivatives with respect to $\boldsymbol{w}$. A precise formulation of the problem will be given below. As known from the work of Tsypkin [1], Sacrison [2], and others the method of stochastic approximation (SA) has basic relevance for a large class of adaptation algorithms. In general, they take the form

$$(1.2) \qquad \boldsymbol{w}[n] = \boldsymbol{w}[n-1] - \gamma[n] \,.\, \boldsymbol{s}(\boldsymbol{x}[n] , \quad \boldsymbol{w}[n-1]) .$$

$\boldsymbol{s}(\boldsymbol{x}[n], \boldsymbol{w}[n-1]) = \boldsymbol{s}[n]$ is the so-called quasigradient which satisfies

$$(1.3) \qquad \nabla_{\boldsymbol{w}} J(\boldsymbol{w}[n-1])^{\mathrm{T}} E\{\boldsymbol{s}[n]\} \geqq 0 .$$

In the framework of SA the constraints for the sequence of the step widths $\gamma[n]$ are rather weak. Therefore, it is difficult to construct rapidly converging adaptation algorithms using only the apparatus of SA.

Albert and Gardner [3] were the first to give performance criteria for accelerated adaptation. The principle of optimality takes the following form: The step width has to be chosen in such a manner that

$$J_2 = E(w[n] - w^*)^2 \tag{1.4}$$

is minimum at each step. This leads to a local minimization giving suboptimal results. In most cases a global optimization of the sequence $\gamma[n]$ is unrealistic (see Łaski and Bzowy [4]). An algorithm optimized in the described manner remains to be of the one-step type, having two levels. The first level is given by eq. (1.2), the second one by the adjusting rule for $\gamma[n]$. In general, $\gamma[n]$ is a matrix.

The authors experience indicates that the acceleration obtained by applying two-level one-step algorithms of this kind is significant especially at the beginning of the adaptation. The main drawback is due to the fact that the computational expense is increasing rapidly when the dimensionality of the problem is becoming large. Therefore the question arises whether there are other principles which allow to accelerate convergence at a lower expense. Such a possibility is suggested by numerical analysis where multi-step algorithms are broadly used. To the authors knowledge Matyáš [5] was the first to use multi-step principles in adaptation. Some results obtained by the author in using two-level two-step algorithms for solving adaptation problems which are reported elsewhere [6] indicate that this approach may be useful. The aim of the present paper consists in decribing a two-level two-step algorithm and deriving some theorems relating to its convergence behaviour.

2. AN ADAPTIVE TWO-LEVEL TWO-STEP ALGORITHM

To be precise, in a two-step algorithm for $w[n]$ the right-hand side besides of $w[n - 1]$ contains also $w[n - 2]$ explicitely. In a two-level algorithm some parameters of the basic adjusting rule may be controlled by secondary rules.

In many practical cases where adaptation algorithms are employed one is confronted with the fact that the adaptation proceeds very slowly, keeping its direction in the mean. The vector differences $w[n] - w[n - 1]$ and $w[n - 1] - w[n - 2]$ are strongly correlated. A straightforward generalization of the one-step algorithm of eq. (1.2) leads to

$$w[n] = w[n - 1] + r_1[n]\,(w[n - 1] - w[n - 2]) - \gamma[n]\,s[n]\,. \tag{2.1}$$

The two-step term $w[n - 1] - w[n - 2]$ is multiplied by a factor $r_1[n]$ which is adjusted by a second-level rule

$$r_1[n] = (1 - \alpha_1)\,r_1[n - 1] + \alpha_2 \operatorname{sign}(\Delta Q[n]) + \alpha_3 \tag{2.2}$$

with

$$\alpha_1 > \alpha_2 + \alpha_3\,; \quad \alpha_1,\, \alpha_2\,,\, \alpha_3 > 0\,; \quad \alpha_1 < 1 \tag{2.3}$$

and

$$\Delta\, Q[n] = Q(x[n-1],\ w[n-2]) - Q(x[n],\ w[n-1])\,. \tag{2.4}$$

If the adaptation is proceeding toward better performance ($\Delta\, Q[n] > 0$), $r_1[n]$ is approximating $(\alpha_2 + \alpha_3)/\alpha_1$. If performance is becoming worse, $r_1[n]$ is decreasing, possibly taking negative values and stopping the undesired direction of adaptation.

The problem to be solved is whether two-level two-step algorithms offer a real advantage over the one-step algorithms. The comparison of the performance is made by means of eq. (1.4) and includes the following statements:

1) The two-level two-step algorithm according to eq. (2.1) and eq. (2.2) is stable;

2) it has the quasigradient property according to eq. (1.3);

3) its optimum convergence behaviour as estimated by eq. (1.4) is superior to that of the optimum one-step algorithm.

3. STABILITY AND QUASIGRADIENT PROPERTY

For adaptation algorithms with quasigradient property the following theorem holds.

Theorem 1 (Polyak and Tsypkin [7]). Let $\boldsymbol{H}$ be a bounded region of a finite dimensional space and let there be for $v,\ w \in \boldsymbol{H}$

$$J(w) \geqq J^* > -\infty\,; \tag{3.1}$$

$$|J(w) - J(v)| \leqq L|w - v|\,; \tag{3.2}$$

$$E\{s[n]^2\} \leqq \lambda[n] + K_1\, J(w[n-1]) + K_2\, \nabla J(w[n-1])^{\mathrm{T}}\, E\{s[n]\}\,; \tag{3.3}$$

$$\nabla J(w[n-1])^{\mathrm{T}}\, E\{s[n]\} \geqq \delta(\varepsilon) > 0 \quad \text{for} \quad J(w[n-1]) > J^* + \varepsilon\,, \quad \varepsilon > 0\,; \tag{3.4}$$

$$\gamma[n] \to 0\,,\ \sum_{n=0}^{\infty} \gamma[n] = \infty\,,\ \sum_{n=0}^{\infty} \gamma^2[n]\, \lambda[n] < \infty\,,\ \sum_{n=0}^{\infty} \gamma^2[n] < \infty\,. \tag{3.5}$$

Then with probability 1

$$\liminf_{n\to\infty} \nabla J(w[n-1])^{\mathrm{T}}\, E\{s[n]\} = 0\,; \tag{3.6}$$

$$\lim_{n\to\infty} w[n] = w^*\,; \tag{3.7}$$

$$\lim_{n\to\infty} J(w[n]) = J(w^*) = J^*\,: \tag{3.8}$$

$$\nabla J(w^*) = 0\,. \tag{3.9}$$

In order to extent Theorem 1 to the two-step case it is sufficient to show, that the assumptions (3.3) and (3.4) can be satisfied, using

$$(3.10)\qquad s'[n] = -\frac{r_1[n]}{\gamma[n]}(w[n-1] - w[n-2]) + s[n]$$

instead of $s[n]$. In this way, the two-step algorithm is formally taken as a one-step algorithm, and (3.3) and (3.4) are discussed for $s'[n]$, assuming that they are satisfied for $s[n]$.

Firstly, we note the following

Lemma. From the constraints of (2.3) it follows that

$$(3.11)\qquad |r_1[k]| \leqq \frac{\alpha_2 + \alpha_3}{\alpha_1} = \varrho_0 < 1 \quad \text{for any } k\,.$$

Using this and eq. (2.1), assuming

$$(3.12)\qquad \frac{1}{\gamma[l]} \sum_{k=0}^{l} \varrho_0^k \gamma[l-k-1] \leqq R_0 < \infty \quad \text{for} \quad l > 0$$

and employing the time shifting operator q, defined by

$$(3.13)\qquad qv[k] = v[k-1] \quad \text{for any } k\,,$$

one obtains

$$(3.14)\qquad \frac{|w[n-1] - w[n-2]|}{\gamma[n]} \leqq \frac{1}{\gamma[n]}(1 - r_1[n-1]\,q)^{-1}\,\gamma[n-1]\,s[n-1] \leqq$$
$$\leqq R_0 \,\Big|\max_{k=0,\dots,n-1} s[k]\Big| = R_1[n-1]\,.$$

This lemma states boundedness of the left-hand side of (3.14). The quasigradient property is stated in the following theorem.

Theorem 2. There exist constants α_1, α_2, α_3 which satisfy (2.3) so that if (3.4) holds for $s[n]$ it does also for $s'[n]$ with $r_1[n]$ according to eq. (2.2).

Proof. It is sufficient to show that

$$(3.15)\quad A_1 = \nabla J(w[n-1])^{\mathrm{T}}\, E\{r_1[n]\,(w[n-1] - w[n-2])\}/\gamma[n] \leqq 0\,.$$

According to the lemma this is equivalent to

$$(3.15\text{a})\quad A_1 = R_1[n-1]\,\nabla J(w[n-1])^{\mathrm{T}}\, E\{r_1[n]\}\,\operatorname{dir}(w[n-1] - w[n-2]) \leqq 0\,.$$

Employing eq. (2.2) and using $R_2[n-1] = R_1[n-1] \,.\, |\nabla J(w[n-1])|$ one obtains

$$A_1 \leqq (1-\alpha_1)\,\delta_1 + \alpha_3\delta_3\, R_2[n-1] + \alpha_2 S_2\, R_2[n-1]\,, \tag{3.16}$$

where δ_1 and δ_2 are constants and S_2 is the term

$$S_2 = \{\operatorname{dir} \nabla\, J(w[n-1)]\}^{\mathrm{T}}\, E\{\operatorname{sign} \Delta\, Q[n]\} \operatorname{dir}(w[n-1] - w[n-2])\,. \tag{3.17}$$

Using the following relations and abbreviations

$$\Delta\, Q_1[n] = Q(x[n-1]\,,\, w[n-2]) - Q(x[n]\,,\, w[n-2])\,, \tag{3.18a}$$

$$d[n-1] = J(w[n-1])^{\mathrm{T}} \operatorname{dir}(w[n-1] - w[n-2])\,, \tag{3.18b}$$

$$\nabla_w\, Q(x[n-1]\,,\, w[n-2]) = \nabla\, J(w[n-2]) + z[n]\,, \tag{3.18c}$$

$$z_1[n] = z[n]^{\mathrm{T}} \operatorname{dir}(w[n-1] - w[n-2]) \tag{3.18d}$$

and disregarding a term of second order, S_2 satisfies

$$S_2|\nabla\, J(w[n-1])| = E\{d[n-1] \operatorname{sign}(\Delta\, Q_1[n] + z_1[n] - d[n-1])\}\,. \tag{3.19}$$

It is easy to see that in assuming

$$P(\Delta\, Q_1[n] + z_1[n] > 0) = P(\Delta\, Q_1[n] + z_1[n] < 0) \tag{3.20}$$

and

$$P(0 < \Delta\, Q_1[n] + z_1[n] \leqq d[n-1]) > 0$$

(which is reasonable in most cases), one can always ensure

$$S_2 < -\,\delta_2 < 0\,. \tag{3.21}$$

Employing the above relations in (3.16) one obtains

$$A_1 \leqq [(1-\alpha_1)\,\delta_1 + \alpha_3\delta_3 - \alpha_2\delta_2]\, R_2[n-1]\,. \tag{3.22}$$

This means that $A_1 \leqq 0$ is always obtainable by a proper choice of the α's, q.e.d.

Under the assumptions of Theorem 1 we have the following theorem, relating to the boundedness of $s'[n]$.

Theorem 3. At the same time as (3.3) is satisfied for $s[n]$ it is also for $s'[n]$.

Proof. We have

$$\begin{aligned} E\{s'[n]^2\} &\leqq E\{r_1[n]^2\}\,(w[n-1] - w[n-2])^2/\gamma[n]^2 + \\ &+ E\{s[n]^2\} + 2[E\{(r_1[n])^2\}\,(w[n-1] - w[n-2])^2\, E\{s[n]^2\}]^{1/2}/\gamma[n]\,. \end{aligned} \tag{3.23}$$

Using the above lemma, one obtains

$$(3.24)\qquad E\{\mathbf{s}'[n]^2\} \leqq R_1[n-1]^2 + 2R_1[n-1]\,(1 + E\{\mathbf{s}[n]^2\}) + E\{\mathbf{s}[n]^2\} \leqq$$
$$\leqq \lambda'[n] + K_1\, J(\mathbf{w}[n-1]) + K_2\, \nabla\, J(\mathbf{w}[n-]1)^{\mathrm{T}}\, E\{\mathbf{s}[n]\}\,.$$

Employing Theorem 2 we get finally

$$(3.25)\qquad E\{\mathbf{s}'[n]^2\} \leqq \lambda[n] + K_1 J(\mathbf{w}[n-1]) + K_2\, \nabla\, J(\mathbf{w}[n-1])^{\mathrm{T}}\, E\{\mathbf{s}'[n]\}$$

and the proof is completed.

4. COMPARISON OF THE OPTIMAL CONVERGENCE SPEEDS

It is not stated by Theorem 2 that convergence can be really accelerated by two-step principles. For this purpose it is necessary to optimize the algorithms of both types separately and to compare the resulting convergence speeds. The comparison consists of three parts:

1) to obtain the optimal step width for both the one-step algorithm and the two-step one,

2) to determine the resulting convergence behaviour,

3) to discuss the cases where the two-step principle offers advantages.

The performance criterion of Albert and Gardner is used which gives only sub-optimal results, but has the advantage that the optimal step width can be obtained explicitely.

We define a regular problem by

$$(4.1)\qquad \nabla\, J(\mathbf{w}[n-1])^{\mathrm{T}}\,(\mathbf{w}[n-1] - \mathbf{w}^*) \geqq 0$$

and confine ourselves to such problems.

Using the following notations

$$E(a) = \bar{a} \quad \text{for any } a\,,$$
$$\mathbf{s}[n] = E(\mathbf{s}[n]) + \Delta\, \mathbf{s}[n] = \overline{\mathbf{s}[n]} + \Delta\, \mathbf{s}[n]\,,$$
$$\mathbf{v}[n] = \mathbf{w}[n] - \mathbf{w}^*\,,$$

we get for the one-step algorithm

$$(4.2)\qquad \gamma[n]_{\mathrm{opt}} = \frac{\overline{\mathbf{s}[n]}^{\mathrm{T}}\, \mathbf{v}[n-1]}{\overline{\mathbf{s}[n]}^2 + \overline{\Delta\, \mathbf{s}[n]^2}}\,.$$

Employing eq. (4.2), the optimal convergence behaviour in the sense of eq. (1.4) is given by

$$(4.3)\qquad \boldsymbol{v}[n]^2_{\text{opt}} = \boldsymbol{v}[n-1]^2 \frac{\overline{\Delta\, \boldsymbol{s}[n]^2}}{\overline{\boldsymbol{s}[n]^2} + \overline{\Delta\, \boldsymbol{s}[n]^2}}\,.$$

For the two-step algorithm a similar derivation can be made. Assuming

$$(4.4)\qquad E\{(\boldsymbol{w}[n-1] - \boldsymbol{w}^* - \gamma\, \boldsymbol{s}[n])^{\mathrm{T}} \Delta\, \boldsymbol{s}[n]\} = 0$$

and

$$(4.5)\qquad E\{r_1[n]\,(\boldsymbol{w}[n-1] - \boldsymbol{w}[n-2])^{\mathrm{T}} \Delta\, \boldsymbol{s}[n]\} = 0$$

we obtain

$$(4.6)\qquad \gamma[n]_{\text{opt}} = \frac{\boldsymbol{v}[n-1]^{\mathrm{T}}\, \overline{\boldsymbol{s}[n]} + \frac{1}{2}\overline{\boldsymbol{s}[n]}\, \overline{r_1[n]}\,(\boldsymbol{w}[n-1] - \boldsymbol{w}[n-2])}{\overline{\boldsymbol{s}[n]^2} + \overline{\Delta\, \boldsymbol{s}[n]^2}}\,.$$

Using this optimal step width for the two-step algorithm its optimal convergence behaviour is described by

$$(4.7)\qquad \boldsymbol{v}[n]^2_{\text{opt}} = \boldsymbol{v}[n-1]^2 \frac{\overline{\Delta\, \boldsymbol{s}[n]^2}}{\overline{\boldsymbol{s}[n]^2} + \overline{\Delta\, \boldsymbol{s}[n]^2}} + (\boldsymbol{w}[n-1] - \boldsymbol{w}[n-2])^2\,.$$
$$\cdot \left\{\overline{r_1[n]^2} - \tfrac{1}{4}\overline{r_1[n]}^2 \cdot \frac{\overline{\boldsymbol{s}[n]}^2}{\overline{\boldsymbol{s}[n]^2} + \overline{\Delta\, \boldsymbol{s}[n]^2}}\right\} +$$
$$+ \overline{r_1[n]}\, \boldsymbol{v}[n-1]^{\mathrm{T}}\,(\boldsymbol{w}[n-1] - \boldsymbol{w}[n-2]) \,.\, \frac{\overline{\Delta\, \boldsymbol{s}[n]^2}}{\overline{\boldsymbol{s}[n]^2} + \overline{\Delta\, \boldsymbol{s}[n]^2}}\,.$$

The first term of the right-hand side of eq. (4.7) corresponds to the right-hand side of eq. (4.2), the second term is of second order in $\boldsymbol{w}[n-1] - \boldsymbol{w}[n-2]$ and will be disregarded. The third term is a first-order one and will be discussed further. From Theorem 3 we have

$$(4.8)\qquad \overline{r_1[n]} \,.\, \nabla\, J(\boldsymbol{w}[n-1])^{\mathrm{T}}\,(\boldsymbol{w}[n-1] - \boldsymbol{w}[n-2]) \leqq 0\,.$$

The property of this term to be nonpositive is transferred to the third term of eq. (4.4) to such a degree as it is possible to identify the direction of the vector $\boldsymbol{v}[n-1]$ with that of the gradient $\nabla\, J(\boldsymbol{w}[n-1])$. Therefore we have

Theorem 4. Let the following assumptions be satisfied: (1.3), (3.1) to (3.5), (4.1), (4.4) and (4.5). Then there exist cases where the convergence behaviour in the sense of eq. (1.4) can be improved using two-step algorithms as given by (2.1) to (2.4) instead of the classic one-step algorithm as given by eq. (1.2) with optimal $\gamma[n]$.

This result indicates that the statements of Tsypkin and Polyak [8] concerning maximum performance of adaptation algorithms in the sense of eq. (1.4) have to be interpreted very carefully, taking into account the special form of the considered algorithms. According to Theorem 3 the effectivity of the two-step algorithms depends on the actual values of the parameters α_i.

In practical cases the existing problem knowledge has to be used in properly choosing these parameters. The two-step algorithms were tested in such cases as stochastic approximation of regression functions, generalized perceptron learning and random search. In all tests they proved to be superior to the one-step algorithms.

(Received May 10, 1976.)

REFERENCES

[1] Я. З. Цыпкин: Адаптация и обучение в автоматических системах. Изд. Наука, Москва 1968.

[2] D. J. Sacrison: Stochastic Approximation: a Recursive Method for Solving Regression Problems. In: Advances in Communication Systems (Ed. A. V. Balakrishnan). Academic Press, New York 1966.

[3] E. Albert, L. A. Gardner: Stochastic Approximation and Non-linear Regression. MIT-Press, Cambridge (Mass.) 1967.

[4] J. Łaski, A. Bzowy: On the Estimation of the Mean of a Stochastic Process. Automatica *11* (1975), 525—527.

[5] И. Матыаш: Случайная оптимизация. Автоматика и телемеханика *26* (1965), 246—253.

[6] K. Fritzsch: Über den Einsatz einer Klasse adaptiver Mehrschrittalgorithmen zur Lösung von Optimierungsproblemen. Elektronische Informationsverarbeitung und Kybernetik *13* (1977), 61—77.

[7] Я. З. Цыпкин, Б. Т. Поляк: Достижимая точность алгоритмов адаптации. Доклады АН СССР — Серия математика-физика *218* (1974), 523—535.

[8] Б. Т. Поляк, Я. З. Цыпкин: Псевдоградиентные алгоритмы адаптации и обучения. Автоматика и телемеханика (1973), 3, 45—68.

Dr. Klaus Fritzsch, Akademie der Wissenschaften der DDR, Zentralinstitut für Kybernetik und Informationsprozesse (Academy of Sciences of the G.D.R. — Central Institute of Cybernetics and Information Processes), DDR — 1199 Berlin, Rudower Chaussee. DDR.