KYBERNETIKA — VOLUME 30 (1994), NUMBER 3, PAGES 301 - 306

ON BATHER'S STOCHASTIC APPROXIMATION ALGORITHM

RAINER SCHWABE

Stochastic approximation procedures provide a useful technique for detecting the root of an unknown regression function. Based on the idea of averaging Bather (1989) proposed a new stochastic approximation algorithm at the Fourth Prague Symposium on Asymptotic Statistics. For this algorithm some results will be presented on the rate of convergence as well as on the behaviour for small to moderate sample sizes.

1. INTRODUCTION

Recursion formulae are frequently used for detecting characteristics of a function like roots or extrema which cannot be determined analytically. In particular, in the present note we will consider the problem of estimating the root θ of a real valued function f on $\mathbb{R}$, i.e. $f(\theta) = 0$, when only noisy observations of f are available.

For this situation Robbins and Monro [8] proposed the recursive scheme

$$X_{n+1} = X_n - a_n Y_n \tag{1}$$

where $Y_n = f(X_n) + U_n$ is an observation of f at X_n disturbed by some random noise U_n. Thus after n initial observations the next observation is to be taken at the setting equal to the estimate $\widehat{\theta}_n = X_{n+1}$ for θ. Typically steplengths $a_n = an^{-\alpha}$, $\alpha \leq 1$, will be considered.

Blum [2] proved almost sure (a.s.) consistency of the procedure, while results on the asymptotic normality were derived by Chung [3], Sacks [12], and Fabian [4]: $n^{-\alpha/2}(X_n - \theta)$ is asymptotically normal for $\alpha < 1$ and in case of harmonic steplengths ($\alpha = 1$) if additionally $2af'(\theta) > 1$. In particular, minimal variance $\sigma_0^2 = f'(\theta)^{-2}\sigma^2$ is achieved for $a_n = (nf'(\theta))^{-1}$.

In the following period methods taken over from the stability theory of ordinary differential equations (*o. d. e. methods*) have turned out to be of great use for proving the stability of stochastic approximation schemes (Ljung [6]). These methods also helped to establish that the recursive stochastic approximation procedures can essentially be represented by weighted averages of the error terms U_n under mild

regularity conditions (Kersting [5], Ruppert [9]). With this representation asymptotic results carry over directly from the weighted averages of the noise to the stochastic approximation schemes.

The performance of the Robbins-Monro procedure with harmonic steplengths, however, heavily depends on the unknown system parameter $f'(\theta)$, which may result in an unsatisfactory asymptotic behaviour, in particular, if $2af'(\theta) \leq 1$.

A first solution to overcome this problem has been proposed by Venter [15] who recommended to replace the unknown slope $f'(\theta)$ adaptively by an estimate $\widehat{f'(\theta)}$. In this procedure the outcome of the recursion formula $X_{n+1} = X_n - (n\widehat{f'(\theta)})^{-1}Y_n$ remains to be the estimate of θ, while the observations are to be taken at different design points $X_n \pm \delta_n$.

In a more recent approach Polyak [7] and Ruppert [10] independently suggested to improve the asymptotic behaviour by means of averaging. The settings X_n for the observations are determined according to the Robbins-Monro procedure (1), but θ is estimated by the average $\hat{\theta}_n = \overline{X_{n+1}} = (n+1)^{-1}\sum_{k=1}^{n+1} X_k$. This procedure proved to be asymptotically optimal, i.e. $\sqrt{n}(\hat{\theta}_n - \theta)$ is asymptotically normal with minimal variance σ_0^2, in case of larger steplengths ($\frac{1}{2} < \alpha < 1$).

For both procedures representations can be found in terms of weighted averages of the noise (cf. Schwabe [13, 14]). For further readings we refer to the recent survey articles by Ruppert [11] and Walk [16].

2. THE ALGORITHM

Alternatively Bather [1] skipped the original scheme (1) and formulated a recursion in terms of the averages

$$X_{n+1} = \overline{X_n} - na_n\overline{Y_n} \tag{2}$$

where $\overline{Y_n} = n^{-1}\sum_{k=1}^{n} Y_k = n^{-1}\sum_{k=1}^{n} f(X_k) + \overline{U_n}$ is the average of the observations. Similarly to the the averaging procedures due to Polyak [7] and Ruppert [10] the observations are to be taken at the settings X_n and the root θ is estimated by the averages $\hat{\theta}_n = \overline{X_{n+1}}$.

Taking into account that $(n+1)\overline{X_{n+1}} = n\overline{X_n} + X_{n+1}$ we obtain a recursion formula for the averages

$$\overline{X_{n+1}} = \overline{X_n} - \tfrac{n}{n+1}a_n\overline{Y_n} \tag{3}$$

which can be identified as an analogue to the Robbins-Monro procedure (1) with the design points replaced by their averages.

Similar considerations lead to

$$X_{n+1} = X_n - a_nY_n + n^{-1}c_n\left(X_n - \overline{X_{n-1}}\right) \tag{4}$$

where $c_n = a_{n-1}^{-1}(na_n - (n-1)a_{n-1})$, $n \geq 2$, $c_1 = 0$. In case of harmonic steplengths we obtain $c_n = 0$ and (2) coincides with the original Robbins-Monro procedure (1) (cf. Bather [1]) which produces a sub-optimal sequence of estimates $\hat{\theta}_n = \overline{X_{n+1}}$. For the more interesting case of larger steplengths $a_n = an^{-\alpha}$, $\alpha < 1$, c_n tends to $1 - \alpha$.

In any case $n^{-1}c_n = o(a_n)$ and the influence of the averages $\overline{X_{n-1}}$ on the choice of the design point X_{n+1} is dominated by the last observation Y_n.

In the next section we are going to illustrate the behaviour of Bather's [1] sequences $\hat{\theta}_n = \overline{X_{n+1}}$ and X_n in case of an underlying linear regression function f. In Section 4 we will consider a more general case. Details will be given elsewhere.

3. LINEAR REGRESSION FUNCTIONS

To illustrate the performance of the algorithm (2) we start with the situation of an underlying linear function, i.e. $f(x) = \lambda(x - \theta)$, with positive slope $\lambda > 0$ (cf. Schwabe [14]). In this case the recursion formula for the averages (3) can be written as $\overline{X_{n+1}} - \theta = (1 - \lambda\frac{n}{n+1}a_n)(\overline{X_n} - \theta) - \frac{n}{n+1}a_n\overline{U_n}$. For simplicity we assume that $\lambda a_n < 1$ for all n. Hence

$$\tilde{d}_n(\overline{X_{n+1}} - \theta) = \tilde{d}_{n-1}(\overline{X_n} - \theta) - \tfrac{n}{n+1}a_n\tilde{d}_n\overline{U_n} \ , \tag{5}$$

where $\tilde{d}_n$ is defined by $\tilde{d}_n = \prod_{k=1}^{n}(1 - \lambda\frac{k}{k+1}a_k)^{-1} = (1 - \lambda\frac{n}{n+1}a_n)^{-1}\tilde{d}_{n-1}$. Iterative evaluation of (5) yields

$$\overline{X_{n+1}} = \theta + \tilde{d}_n^{-1}(X_1 - \theta) - \tilde{Z}_n \ , \tag{6}$$

where $\tilde{Z}_n = \tilde{d}_n^{-1}\sum_{i=1}^{n}\frac{i}{i+1}a_i\tilde{d}_i\overline{U_i} = \tilde{d}_n^{-1}\sum_{k=1}^{n}U_k\sum_{i=k}^{n}\frac{1}{i+1}a_i\tilde{d}_i$ is a weighted average of the error terms U_n. In an analogous way we introduce the weighting sequence $d_n = \prod_{k=1}^{n}(1 - \lambda a_k)^{-1} = (1 - \lambda a_n)^{-1}d_{n-1}$ associated with the recursion formula (4) for the design points X_n. The sequences d_n and $\tilde{d}_n$ are asymptotically equivalent up to a multiplicative constant, i.e. $d_n/\tilde{d}_n \to \tilde{c} > 1$.

The asymptotic behaviour of the weighted averages $Z_{n+1} = d_n^{-1}\sum_{k=1}^{n}a_k d_k U_k$ plays an important role in stability considerations (see Ljung [6], Ruppert [10], and Walk [16], cf. also the Theorem in Section 4).

The sequence $\tilde{Z}_n$ is close to the arithmetic means $\overline{U_n}$ in the following sense:

Proposition. Let $\alpha < 1$. If $n^{\gamma}Z_n \to 0$ a.s., then

$$\tilde{Z}_n = \lambda^{-1}\overline{U_n} + o(n^{-\gamma-(1-\alpha)}) \quad \text{a.s.}$$

In many practical applications the requirement $n^{\gamma}Z_n \to 0$ a.s. is satisfied for every $\gamma < \frac{\alpha}{2}$. Hence, for $\alpha < 1$, the estimate $\hat{\theta}_n = \overline{X_{n+1}}$ can essentially be represented by the average of the error terms

$$\overline{X_{n+1}} = \tfrac{1}{\lambda n}\sum_{k=1}^{n}U_k + o(n^{-\frac{1}{2}-\varepsilon}) \quad \text{a.s.} \ , \tag{7}$$

for some $\varepsilon > 0$, and the asymptotic behaviour of $\lambda^{-1}\overline{U_n}$ carries over to $\hat{\theta}_n$. In particular, $\sqrt{n}(\hat{\theta}_n - \theta)$ is asymptotically normal with minimal variance σ_0^2, i.e. $\hat{\theta}_n = \overline{X_{n+1}}$ is asymptotically optimal.

For the design points X_n we obtain $X_{n+1} - \theta = (1 - \lambda n a_n)(\overline{X_n} - \theta) - n a_n \overline{U_n}$, which shows that X_{n+1} and $\overline{X_n}$ tend to lie on opposite sides of the root θ. This is a behaviour which is advantageous from the design point of view.

Furthermore, X_n can also be represented by a weighted average of the error terms

$$X_{n+1} = \theta - \lambda n a_n \tilde{d}_n \sum_{k=1}^{n} \left(\sum_{i=1}^{k-1} \tfrac{1}{i+1} a_i \tilde{d}_i \right) U_k + R_n \tag{8}$$

and the remainder term R_n is of order $o(n^{-(\alpha+\varepsilon)/2})$ a.s., for some $\varepsilon > 0$, under some additional regularity conditions on the error terms. From this formula the asymptotic normality of $n^{-\alpha/2}(X_n - \theta)$ can be derived which has been conjectured by Bather [1].

4. THE GENERAL CASE

In most applications the underlying regression function f will not be linear. In this case, however, some regularity conditions have to be satisfied: f is continuous, $f(x)(x - \theta) > 0$ for all $x \neq \theta$, f is linearly bounded, i.e. $|f(x)| \leq c_1 + c_2|x - \theta|$, and $\lambda = f'(\theta) > 0$. Additionally we assume that f is two times continuously differentiable in a neighbourhood of θ.

We can now present a result based on the behaviour of the weighted averages Z_n of the error terms defined in the previous section:

Theorem. Let $\frac{1}{2} < \alpha < 1$. If X_n is bounded and $n^{\alpha(1-\delta)/2} Z_n \to 0$ a.s., for some $\delta < \min(1 - \alpha, \alpha - \frac{1}{2})$, then

$$\begin{aligned} X_{n+1} &= \theta - Z_{n+1} + o(n^{-\frac{\alpha}{2}-\varepsilon}) \quad \text{a.s.}, \\ \overline{X_{n+1}} &= \theta - \tfrac{1}{n f'(\theta)} \sum_{k=1}^{n} U_n + o(n^{-\frac{1}{2}-\varepsilon}) \quad \text{a.s.}, \end{aligned}$$

for some $\varepsilon > 0$.

The proof is based on a refinement of the arguments given by Ljung [6] and Ruppert [9]. For the estimates $\hat{\theta}_n = \overline{X_{n+1}}$ and the design points X_n asymptotic results can directly be derived from the corresponding results of the averages of the error terms by means of the present Theorem. In particular, asymptotic normality and convergence rates of the iterated logarithm type can be obtained.

5. CONCLUSIONS

The algorithm (2) proposed by Bather [1] yields an asymptotically optimal sequence of estimates $\hat{\theta}_n = \overline{X_{n+1}}$ for the root θ of the unknown underlying regression function f. We note that for the general case the boundedness of X_n has still to be established which might be achieved by truncation arguments.

For even larger steplengths $a_n = an^{-\alpha}$, $\alpha \leq \frac{1}{2}$, a substantial bias may arise from the nonlinearity of f, which can be seen as follows: Starting from the recursion formula (2) a Taylor expansion of f gives

$$X_{n+1} = \theta + (1 - na_n f'(\theta))(\overline{X_n} - \theta) - na_n \overline{U_n} - \tfrac{1}{2} a_n \sum_{k=1}^{n} f''(\xi_k)(X_k - \theta)^2$$

for some ξ_k between X_k and θ. Even if X_k tends to θ as $k^{-\alpha/2}$ and, hence, $f''(\xi_k) \to f''(\theta)$, the remainder term $\frac{1}{2} a_n \sum_{k=1}^{n} f''(\xi_k)(X_k - \theta)^2$ does not vanish sufficiently fast for curved functions f with $f''(\theta) \neq 0$.

In Bather's [1] algorithm (2) the influence of inappropriate starting points decreases like d_n^{-1} (see (6)) and, hence, substantially faster than n^{-1}, which is the corresponding rate in the Robbins-Monro procedure with optimal steplengths and in the averaging procedures by Polyak [7] and Ruppert [10], at least in the linear case. In particular, for small to moderate sample sizes n this fact results in a better performance of Bather's [1] algorithm.

ACKNOWLEDGEMENT

The author wishes to express his thanks to Professor V. Fabian for helpful comments.

(Received March 3, 1994.)

REFERENCES

[1] J. A. Bather: Stochastic approximation: A generalisation of the Robbins-Monro procedure. In: Proc. Fourth Prague Symp. Asymptotic Statistics, Charles Univ. Prague, August 29–September 2, 1988 (P. Mandl and M. Hušková, eds.), Charles Univ., Prague 1989, pp. 13–27.

[2] J. R. Blum: Approximation methods which converge with probability one. Ann. Math. Statist. *25* (1954), 382–386.

[3] K. L. Chung: On a stochastic approximation method. Ann. Math. Statist. *25* (1954), 463–483.

[4] V. Fabian: On asymptotic normality in stochastic approximation. Ann. Math. Statist. *39* (1968), 1327–1332.

[5] G. Kersting: Almost sure approximation of the Robbins-Monro process by sums of independent random variables. Ann. Probab. *5* (1977), 954–965.

[6] L. Ljung: Strong convergence of a stochastic approximation algorithm. Ann. Statist. *6* (1978), 680–696.

[7] B. T. Polyak: New method of stochastic approximation type. Automat. Remote Control *51* (1990), 937–946.

[8] H. Robbins and S. Monro: A stochastic approximation method. Ann. Math. Statist. *22* (1951), 400–407.

[9] D. Ruppert: Almost sure approximations to the Robbins–Monro and Kiefer–Wolfowitz processes with dependent noise. Ann. Probab. *10* (1982), 178–187.

[10] D. Ruppert: Efficient Estimators from a Slowly Convergent Robbins-Monro Process. Technical Report No. 781, School of Operations Research and Industrial Engineering, Cornell Univ. Ithaca 1988.

[11] D. Ruppert: Stochastic approximation. In: Handbook of Sequential Analysis. (B. K. Ghosh and P. K. Sen, eds.), Marcel Dekker, New York 1991, pp. 503–529.
[12] J. Sacks: Asymptotic distribution of stochastic approximation procedures. Ann. Math. Statist. *29* (1958), 373–405.
[13] R. Schwabe: Strong representation of an adaptive stochastic approximation procedure. Stochastic Process. Appl. *23* (1986), 115–130.
[14] R. Schwabe: Stability results for smoothed stochastic approximation procedures. Z. Angew. Math. Mech. *73* (1993), 639-643.
[15] J. H. Venter: An extension of the Robbins–Monro procedure. Ann. Math. Statist. *38* (1967), 181–190.
[16] H. Walk: Foundations of stochastic approximation. In: Stochastic Approximation and Optimization of Random Systems, DMV Seminar Blaubeuren, May 28–June 4, 1989 (L. Ljung, G. Pflug and H. Walk, eds.), DMV Seminar, Vol. 17, Birkhäuser, Basel 1992, pp. 1–51.

Dr. Rainer Schwabe, Freie Universität Berlin, 1. Mathematisches Institut, Arnimallee 2–6, D-14 195 Berlin. Germany.