Stochastic Approximation and Stochastic Gradient methods–a overview and guide to the literature

Enrique Del Castillo
Applied Statistics Laboratory & Industrial Engineering Department
The Pennsylvania State University, University Park, PA 16802

November 2000

Abstract

A brief guide to the literature on Stochastic approximation/optimization methods is provided with application to process adjustment and process optimization problems in quality control, respectively. No intent is made of providing a literature review.

1 Newton's method and Stochastic approximation

Consider Newton's method for solving for the root of a function, i.e., find x such that $f(x) = 0$ is true. Here $f(x)$ is a deterministic function from $\Re$ to $\Re$, in other words, $f(x)$ can be measured without error. The recursive equation that is known to converge to the root is given by

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \tag{1}$$

where $f'(x) = df(x)/dx$ is known analytically. In this setting, convergence simply means that $x_n \to \theta$ as $n \to \infty$ (where θ is the closest root to x_1). The recursion is equivalent to steepest descent applied to $f(x)$.

The seminal work of Robbins and Monro (RM) [1] extended such approach to a stochastic function, that is, assume $y = y(x) = M(x) + \varepsilon$ where ε is a random variable with $E[\varepsilon] = 0$ so $M(x) = E[y|x]$ is the regression of y on x, a function which can be nonlinear in x. No parametric assumptions are made on the form of $M(x)$. Suppose we want to find a root of the equation $M(x) = 0$. Neither $M(x)$ nor $M'(x)$ are known analytically, and we cannot observe $M(x)$ directly, we can only observe $y(x)$, i.e., $y(x)$ is a measurable function. Suppose that instead of using (1) (which is not possible to use since the functions needed are not available to us) we use the recursion (sometimes called the "RM process"):

$$x_{n+1} = x_n - a_n y(x_n) \tag{2}$$

where the sequence $\{a_n\}_{n=1}^{\infty}$ needs to be determined and x_1 is selected arbitrarely. This recursion determines an experimental design on the factor space that will seek the root of the equation. RM asked a very simple question: under what conditions on $M(x)$, ε, and $\{a_n\}_{n=1}^{\infty}$ do we have convergence of the sequence of experimental factor levels $\{x_n\}_{n=1}^{\infty}$ to the root of $M(x) = 0$, call it θ? In their original paper, they showed that, if

a
$$a_n \to 0, \quad \sum_{n=1}^{\infty} a_n = \infty$$

b
$$\sum_{n=1}^{\infty} a_n^2 < \infty$$

c the regression function is such that $M(x) \leq 0$ if $x < \theta$ and $M(x) \geq 0$ if $x > \theta$,

d the distribution function of ε has finite tails

then using (2) we have
$$\lim_{n \to \infty} E[(x_n - \theta)^2] = 0$$
i.e., they show convergence in *mean square* of x_n to the root θ. A sequence that satisfies the first two conditions is the harmonic series $a_n = 1/n = \{1, 1/2, 1/3, ...\}$. Condition a) is needed because otherwise the search could stop before finding a root. Condition b) is needed to eventually eliminate the "noise" in the observations. Some of the assumptions under which mean square convergence is achieved were simplified and clarified by Dvoretzky [3]. In particular, instead of assumption d) it is only necessary that $\sigma_{\varepsilon}^2(x) < \infty$. Note that the variance can be non-homogeneous, a condition of considerable interest in response surface methods. A simplified proof of MS convergence for the RM process is provided by J.Semple [4]. The Dvoretzky conditions imply both MS convergence and convergence of x_n to θ with probability one.

2 Stochastic Gradient

Following a suggestion by RM, Kiefer and Wolfowitz [2] wrote a paper in which they propose to apply Robbins and Monro stochastic approximation approach to finding the root of $M'(x) = 0$, i.e., for finding the stationary point θ (say) of the regression function that is known to have a maximum (say). Since the slope of the regression function is not observable directly, they propose to use instead:

$$x_{n+1} = x_n + a_n \frac{y(x_n + c_n) - y(x_n - c_n)}{c_n} \qquad (3)$$

(if we wish to minimize we subtract the second term instead). Kiefer and Wolfowitz show that x_n as above converges in mean square to θ given that

$$a_n \to 0, \quad c_n \to 0, \quad \sum_{n=1}^{\infty} a_n = \infty, \qquad \sum_{n=1}^{\infty} a_n c_n < \infty, \qquad \sum_{n=1}^{\infty} a_n^2 c_n^2 < \infty$$

besides of some regularity conditions on $M(x)$. Blum [5] extended this procedure to the multivariate case, of particular interest in RSM. Kesten [6] provided an acceleration routine that should be useful for small samples.

3 Review papers

Given that the literature in these subjects is too large and the topic has been around for almost 50 years, it is very instructive to read reviews of these methods before consulting the original papers, which are quite technical in general. Of course, recent papers (post 1990) are not found in these reviews. The book by Wilde [7] contains a highly readable exposition of the first classic papers in stochastic approximation and stochastic optimization up to about 1965. A review on Stochastic Gradient up to 1970 is by Fabian [8]. A more recent review on Stochastic approximation methods is the article by Sampson [9]. An equally recent review on Stochastic Gradient (i.e., the KW method) is the article by Ruppert [10]. A very readable and interesting review of papers and results up to around 1989 is provided by Ruppert [11].

4 Emphasis on asymptotic results

It is quite evident, from reading the aforementioned reviews and papers, that there exist a large number of asymptotic results related to both the RM and the KW processes. Convergence in mean and with probability one has been proved in both cases under certain conditions. Also, the asymptotic normality of $\sqrt{n}(x_n - \theta)$ has been shown, etc. Most authors are interested in the long run behavior of x_n, although Frees and Ruppert [12] and Wu [13] discuss a MR process where they looked at MSE($\hat{\theta_n}$) for small n using simulation. To estimate the root θ, they propose using either the last x_n or the solution to $\widehat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ where the parameters are estimated using OLS from all the $x_n's$ and the $y's$ in the process. Robbins and Lai [14] and Wei [15] discuss an "Adaptive" RM procedure that considers the cost $\sum_{i=1}^{n}(x_n - \theta)^2$ (rather than $E[(x_n - \theta)^2]$), which they mention (correctly) that it should be more important for control. However, their interest is really in the asymptotic value of this sum, i.e., $lim_{n\to\infty} \sum_{i=1}^{n}(x_n - \theta)^2$. Interest in quality control of expensive parts should be on this sum for *finite and small n.*

In an early reference, Hodges and Lehmann [16] studied the behavior of $E[(x_n - \theta)^2]$ for small n, and considered the best selection of the constant c in the sequence $a_n = c/n$ studied by them. They concluded that using $c = 1/M'(\theta)$ was best. However, they didn't consider performance measures that are more relevant for control.

5 Applications and opportunities in process adjustment

The setup adjustment problem, where a machine is initially off-target by d units was studied by Grubbs [17]. He finds an adjustment scheme that is identical to using the RM estimate of d. Del Castillo and Pan [18] studied the connections between Grubbs' procedure, the RM process, and a Kalman filter approach. They studied the small sample Average Integrated Squared Deviation (AISD) provided by each rule, and concluded that Grubbs' rule was best in general. However, results from Frees and Ruppert [12] indicate that it is worth exploring the AISD performance of a RM-like rule where the weights are given by $C/(i\hat{\beta}_1 + r)$ where β_1 is the slope of $M(x)$ ($\beta_1 = 1$ is known for Grubbs procedure, although this would work also in case the gain or slope is not known) and C is some constant. These authors did not consider the AISD, but instead looked at $E[(x_n - \theta)^2]$ for small n, in an analysis closely related to the aforementioned one by Hodges and Lehmann . For this performance measure, Frees and Ruppert indicate that using $C = 1.5$ to 2 and $r = 1$ or 0 works best. It is of interest to derive AISD formulae for such weights and try to determine recommendations for C and r, as compared to simply using the harmonic rule $\{1/n\}$ as used by Grubbs.

A different area of application is for those setup adjustment problems in which overadjustment in one direction is much more expensive than under adjusting, i.e., the underlying loss function is not symmetric. A typical example is in a drilling process, where a hole too wide implies scrap whereas a hole too narrow can be reworked to save the part. RM-like approaches exist that guarantee approaching the root without overshooting it with some given probability [19, 20]. This can be used in such asymetric-loss problems.

Another area of considerable interest is applications of stochastic approximation (SA) to processes that exhibit dynamics, contrary to setup adjustment problems in which the process is assumed stable. Two very interesting, and practically forgotten papers by Comer [21, 22] provide several results of "adaptive proportional" controllers (he was actually referring to integral controllers). The idea is to use a RM-like process for determining the controllable factor. One of Comer's methods is comparable with another stochastic approximation algorithm recently developed by Patel and Jenkins [26], who proposed an adaptive EWMA controller with guaranteed stability. Their results are similar to a modification to the RM procedure proposed by Robbins and Sigmund [23]. No comparison of these methods for small or large samples are available.

An interesting recent paper is by Chen and Guo [24] who propose to use SA to find MSE-optimal EWMA weights in an EWMA controller when there is evidence of a shift, and use a constant EWMA weight otherwise. The idea of coupling Grubbs procedure with a control chart was originally proposed by Del Castillo [25]. It is not clear if always using SA (Grubbs) procedure rather than having a minimum weight value is better or not. How to detect the change point is very important, Chen and Guo used an EWMA control chart to trigger the RM-like adaptation of the weights, unaware that they were using Grubbs' procedure or SA.

6 Applications and opportunities in process optimization

The KW procedure provides an experimental optimization method for industrial processes, which, however, has not been used in practice where response surface methods are widely used. It is therefore of interest to investigate if the KW procedure can converge to a stationary point in fewer experiments than a traditional application of RSM. One of the difficulties of the KW method is that, given that the gradient of the response needs to be estimated, this results in slower convergence compared to the RM approach. Furthermore, the multivariate KW process studied by Blum requires $2k$ experiments at each iteration in order to estimate the gradient. A more recent approach by Spall [27], termed "Stochastic perturbation" requires only 2 experiments per iteration. Recent research in this area has increased thanks to the application of KW-like processes to Neural Network "learning" processes. Thinking in this type of application, Darken et al. [28] investigate a KW algorithm with weight sequence:

$$a_t = \eta_0 \frac{1 + \frac{ct}{\eta_0 \tau}}{1 + \frac{ct}{\eta_o \tau} + \tau \frac{t^2}{\tau^2}}$$

which equals to η_0 for times $t \ll \tau$. For $t \gg \tau$, this function behaves as c/t, the traditional RM/KW weights. The idea is to make the algorithm approach a "good" region rapidly, and only then start the convergence phase thanks to the RM/KW harmonic sequence, which, if used from the beginning, would make convergence too slow. Andradottir [29] provides another modification of the RM method for use in simulation optimization which appears to converge faster than the original RM process.

The convergence rate of KW-like processes can be improved if second-order information is used in the search. Ruppert [30] provides a stochastic version of Newton's method. The method's idea is to premultiply Y_n by some estimate of the inverse of the Hessian of $M(\theta)$. This approach is a RM, not a KW, method. It seems that only Fabian [8] has investigated incorporating second order information in KW processes but he did not provide any performance analysis. Nonlinear optimization algorithms such as the BFGS method provide a sequential method for approximating the Hessian of a deterministic function. It seems plausible that such a scheme can be put in a stochastic optimization setting, providing an algorithm that converges faster than traditional gradient-based KW methods. The emphasis of such investigation, however, should be small sample behavior, i.e., to optimize a process with the smallest number of experiments.

A different area of application of RM and KW processes is using them in certain parts of the traditional RSM framework. Once instance of this area is the use of the RM process to do the line searches needed in the steepest ascent phase of RSM. Investigation and comparison of such approach vs. stopping rules used in steepest ascent are of interest.

7 Relations with recursive estimation and adaptive control

One of the major areas of application of the RM process is for the solution of statistical estimation problems. Here, θ is a parameter that needs to be estimated. In particular, recursive estimation methods based on the RM approach have been investigated. An estimator θ_n is said to be recursive if it is a function only of θ_{n-1} and of y_n. The literature on recursive estimation is very large, but a good source of information is the book by Ljung and Soderstrom [31]. This reference emphasizes recursive estimation of parameters of models that define observations of time-dependent processes (i.e., time series). In general, based on their speed of convergence, this literature recommends a recursive version of the least squares algorithm rather than the use of stochastic gradient.

8 More papers

The literature of RM and KW methods has grown to a very considerable size over the years. Recent application to simulation optimization, process control, and Neural Network learning has generated a renewed interest in these methods. This paper just provided a guide to the most important and "classical" sources. A detailed literature search should be undertaken by anybody wishing to work in this areas. Since the literature is extremely technical, it is expected that this guide together with some of the references herein will provide a more gentle introduction to the topic, including some ideas for areas open to research.

References

[1] Robbins, H. and Monro, S. (1951). A Stochastic Approximation Method. Annals of Mathematical Statistics, 22, pp. 400-407. The "classsic" paper, and one of the most cited in the statistics area.

[2] Kiefer, J., and Wolfowtiz, J. (1952). Stochastic estimation of the maximum of a regression function. Ann. Math. Stat. 23, 452-466. The classic paper on stochastic optimization.

[3] Dvoretsky, A. (1955). On Stochastic Approximation. Proceedigns of the 3rd Berkeley Symposium on Mathematical Stats.

[4] Semple, J. (2000). Personal communication on a paper by Wolfowitz, PSU. Simplifies the proof of MS convergence using the idea of Cauchy convergence of series.

[5] Blum, J. (1954). Multivariate Stochastic Approximation Methods. Ann. Math. Stat. 25, 737-744. The first approach to extend the RM approach to multiple factors.

[6] Kesten, (1958). Accelerated Stochastic Approximation. Ann. MAth. Stat. 29, 41-59.

[7] Wilde, D. (1964). Optimum Seeking Methods. PRentice Hall, Englewood Cliffs, NJ. A very readable first place to look.

[8] Fabian, V. (1971). Stochastic Approximation. In Optimizing Methods in Statistics (J. Rustagi, ed.) 439-470, Academic Press, NY.

[9] Sampson. Stochastic approximation. Encyclopedia of Statistical Sciences, Kotz and Johnson, eds.

[10] Ruppert, D. Kiefer-Wofowitz procedure. Encyclopedia of Statistical Sciences, Kotz and Johnson, eds. A good and brief review.

[11] Ruppert, B. (1991). Stochastic Approximation. In Handbook of Sequential Analysis, B.K. Ghosh and P.K. Sen, eds. MArcel Dekker, NY. An excellent review of both RM and KW proceses up to around 1990.

[12] Frees, E.W., and Ruppert, D. (1991). Estimation following a Robbins-Monro designed experiment, JASA.

[13] Wu, C.J.F. (1985). Efficient sequential designs with binary data. JASA, 80, 974-984.

[14]

[15] Wei, C.Z. (1987). Multivariate adaptive stochastic approximation. Ann Stat. 15, 1115-1130.

[16] Hodges, J.L. jr., and Lehmann, E.L. Two approximations to the Robbins-Monro Process. Proceedings of the third Berkeley symposium on Mathematical Stats., pp. 95-104. Contains some of the most interesting results on small sample behavior of the RM process.

[17] Grubbs, F.E., (1954). "An Optimum Procedure for Setting Machines or Adjusting Processes," *Industrial Quality Control*, July, reprinted in *Journal of Quality Technology*, 1983, 15, 4, pp. 186-189.

[18] Del Castillo, E., and Pan, R. An Unifying View of Some Process Adjustment Methods. Submitted to JQT (2000).

[19] Gusev, S.V., and Krasulina, T.P. (1995). An algorithm for stochastic approximation with a preassigned probability of not exceeding a required threshold. Journal of Computer and Systems Sciences International. Sept. 35, 5.

[20] Krasulina, T.P. (1998). On the probability of not exceeding a desired threshold by a stochastic approximation. Automation and remote control, Oct. 59, 10.

[21] Comer, J.P. ,jr. (1965). Application of Stochastic Approximation to Process Control. Journal of the Royal Statistical Society (B), 27, 2, 321-331.

[22] Comer, J.P., jr. (1964). Some Stochastic approximation procedures for use in process control. Ann. math. Stat., 35, 3, 1136-1146.

[23] Robbins, H., and Siegmund, D. (1971). A convergence theorem for non-negative almost positive supermartingales and some applications. In Optimizing methods in statistics (J Rustagi, ed.) 237-257, Academic Press, NY.

[24] Guo, R.S., Chen, A., and Chen, J.J., (2000). "An Enhanced EWMA Controller for Processes Subject to Random Disturbances", in Moyne, J., Del Castillo, E., and Hurwitz, A., *Run to run control in semiconductor manufacturing*, CRC press.

[25] Del Castillo, E. (1998). "A Note on two Process Adjustment Models". *Quality and Reliability Engineering International*, 14, pp. 23-28.

[26] Patel, N.S., and Jenkins, S.T. (2000). Adaptive optimization of Run to Run Controllers: the EWMA example. IEEE Transactions on Semiconductor Manufacturing, 13,1, p. 97.

See also ch. 16 in Moyne et al., Run to Run Control in Semicondctor Manufaturing (2000).

[27] Spall, J.C. (1992). Multivariate Stochastic Approximation using a simultaneous perturbation gradient approximation. IEEE Transactions on Automatic Control, 37, 3. There are more recent references by Spall on this subject, including some limited results for small samples.

[28] Darken, C., Chang, J., Moody, J. (1992). Learning rate schedules for faster stochastic gradient search. Neural Networks for Signal Processing 2 –Proceedings of the 1992 IEEE Workshop, IEEE Press, NJ.

[29] Andradottir, S. (1995). A stochastic approximation algorithm with varying bounds. Operations Research, 43, 6, 1037–1048. Andradottir also has some other mid 90's papers in the subject.

[30] Ruppert, D. (1985). A Newton-Raphson version of the multivariate Robbins-Monro procedure. Ann. Math. Stat. 13, 236-245.

[31] Ljung, L. and T. Soderstrom (1987). *Theory and Practice of Recursive Identification*, Cambridge, Mass., The MIT Press.