Trakia Journal of Sciences, Vol. 3, No. 2, pp 17-21, 2005

Available online at:
http://www.uni-sz.bg

ISSN 1312-1723

Mini-review

STOCHASTIC APPROXIMATION FOR ESTIMATION OF BIOLOGICAL MODELS

Nikolay Iv. Petrov*

Trakia University – Stara Zagora, Yambol
Bulgaria

ABSTRACT

Stochastic approximation for estimation (SAE) is a class of optimisation algorithms, which computes, to an approximation, the gradient and/or the Hessian of the objective function by varying all the elements of the parameter vector simultaneously and, therefore, requires only a few objective function evaluations to obtain first or second-order information. Consequently, these algorithms are particularly well suited to problems involving a large number of design parameters. In this study, their potentials are assessed in the context of non-linear (NN) system identification. To pursue this objective, a challenging modelling application is considered, that is, dynamic modelling of batch animal cell cultures from sets of experimental data. The performance of the optimisation algorithms is discussed in terms of efficiency, accuracy and ease of use.

Key words: non-linear system identification; stochastic approximation; biotechnology

INTRODUCTION

The process of modelling requires an estimation of several unknown parameters from noisy measurement data. To achieve this aim, a least-squares or maximum–likelihood cost function (depending on the assumptions of the measurement noise) is usually minimised using a gradient-based optimisation method.

Several techniques for computing the gradient of the cost function are available, including finite difference approximation and analytic differentiation. This latter technique leads to back-propagation in neural networks or several sensitivity equations in the case of conventional first-principles models.

In the above-mentioned techniques, the computational expense required to estimate the current gradient direction, are directly proportional to the number of unknown model parameters, which become an issue for models involving a large number of parameters. This is typically the case in *non-linear* (NN) modelling, but can also occur when estimating parameters and initial conditions in first-principles models.

In contrast to standard finite differences, which approximate the gradient by varying the parameters one at a time, the *simultaneous perturbation* (SP) approximation of the gradient proposed by Spall [5] makes use of a very efficient technique based on a simultaneous (random) perturbation in all the parameters. Hence, one gradient evaluation requires only two evaluations of the cost function. This approach has first been applied to gradient estimation in a first-order *stochastic approximation* (SA) algorithm [5] and, more recently, to Hessian estimation in an accelerated second-order *stochastic approximation for estimation* (SAE) algorithm [6].

In previous works the authors applied the above–mentioned first- and second- order SA algorithms (1SAE and 2SAE) to weights and biases estimation in NNs and proposed several variations of the 1SAE algorithm [3, 7]. These simulation studies were limited to relatively simple examples but demonstrated the efficiency and modest computational costs of 1SAE. The objective of this paper is to extend these studies by evaluating:

- variants of 1SAE/2SAE algorithms, in which scaling of the gradient/Hessian estimates is introduced to avoid potential

*Correspondence to: *Nikolay Iv. Petrov, Trakia University – Stara Zagora, Yambol, Bulgaria, nikipetrov@lycos.com*

large variations in the course of the optimisation process;

- the performance of first- and second-order algorithms as applied to a challenging parameter estimation problem namely, identification of unknown parameters in a macroscopic model of batch animal cell cultures from experimental measurements of biomass, glucose, glutamine and lactate concentrations.

This paper is organised as follows: Section 2 introduces the basic principles of the first- and second-order SA algorithms used throughout this study. In section 3, the algorithms are applied to the maximum likelihood estimation of kinetic parameters and initial conditions of a bioprocess model from experimental measurements of several macroscopic component concentrations. Direct and cross-validation results demonstrate the good model agreement. Finally, section 4 is devoted to discussions and concluding remarks.

SAE ALGORITMS

Consider the problem of minimising a possibly noisy objective function $J(\theta)$ with respect to a vector θ of unknown parameters. 1 SAE is given by the following core recursion for the parameter vector θ [3; 7]. This is shown by equation:

(1) $\hat{\theta}_{k+1}(t) = \hat{\theta}_k(t) - a_k.\hat{g}_k\left(\hat{\theta}_k\right),$

in which a_k is a non-negative scalar gain coefficient, and $\hat{g}_k\left(\hat{\theta}_k\right)$ is an approximation of the criterion gradient obtained by varying all the elements of $\hat{\theta}_k(t)$ simultaneously, i.e.

(2)

$$\hat{g}_k(\theta_k) = \begin{bmatrix} \dfrac{J\left[\hat{\theta}_k(t) + c_k.\Delta_k\right] - J\left[\hat{\theta}_k(t) - c_k.\Delta_k\right]}{2c_k.\Delta_{k1}} \\ \cdots\cdots \\ \dfrac{J\left[\hat{\theta}_k(t) + c_k.\Delta_k\right] - J\left[\hat{\theta}_k(t) - c_k.\Delta_k\right]}{2c_k.\Delta_{kp}} \end{bmatrix},$$

where, c_k is a positive scalar and $\Delta_k = \left(\Delta_{k1}, \Delta_{k2}, \ldots, \Delta_{kp}\right)^T$ with symmetrically Bernouilli distributed random variables $\{\Delta_{ki}\}$.

In its original formulation, 1SAE makes use of decaying gain sequences $\{a_k\}$ and $\{c_k\}$ in the form

(3) $a_k = \dfrac{a}{(A+k+1)^{\alpha}}, \; c_k = \dfrac{c}{(k+1)^{\gamma}},$

which ensures asymptotic convergence results. However performance in finite samples can be different, and numerical experiments suggest that an adaptive gain sequence for parameter updating [1, 3, 5] can enhance convergence and stability (this is particularly true when solving a non convex parameter identification problem), i.e.

(4)

$a_k = \eta.a_{k-1}, \eta \geq 1, \quad if \quad J(\theta_k) < (1+\beta).J(\theta_{k-1})$

(5)

$a_k = \mu.a_{k-1}, \mu \geq 1, \quad if \quad J(\theta_k) < (1+\beta).J(\theta_{k-1})$

In addition to gain attenuation when the value of the criterion becomes worse, "locking" mechanisms [5, 6] are also applied, i.e. the current step is rejected and, starting from the previous parameter estimate, a new step is accomplished (with a new gradient evaluation and a reduced updating gain). The parameter β in equations (4) and (5) represents permissible increase in the criterion before step rejection and gain attenuation occur.

A constant gain sequence $c_k = c$ can be used for gradient approximation, the value c being selected so as to overcome the influence of (numerical or experimental) noise. In the optimum neighbourhood, however, a decaying sequence in the form (3) is required to evaluate the gradient with enough accuracy and avoid an amplification of the "slowing down" effect as an optimum is approached (note that this phenomenon is even more pronounced in the case of SP techniques since the gradient information is more delicate to "extract" in the – usually rather "flat" – neighbourhood of the optimum).

Finally, a *gradient smoothing* (GS) procedure is implemented, i.e., gradient approximations are averaged across iterations in the following way

(6)

$G_k = \rho_k.G_{k-1} + (1-\rho_k).\hat{g}_k\left(\hat{\theta}_k\right), \; 0 \leq \rho_k \leq 1, \; G_0 = 0,$

where, ρ_k is decreased in a way similar to equations (4) and (5) when step rejection occurs (i.e. $\rho_k = \mu.\rho_{k-1}$ with $\mu \leq 1$) and is reset to its initial value ρ_0 after a successful step.

The use of these numerical artifices, i.e. adaptive gain sequences, step rejection procedure and gradient smoothing,

significantly improves the effective practical performance of the algorithm (which, in the following, is denoted "adaptive1SP-GS") [1, 2, 4].

As relatively large excursions in the parameter space can be achieved, convergence can also be enhanced through scaling of the gradient estimate (2) at each iteration. This new feature is implemented here by normalising each direction of the gradient vector $\hat{g}_k\left(\theta_k\right)$ with respect to its largest component (infinity norm scaling)

(7) $\hat{\theta}_{k+1}\left(t\right)=\hat{\theta}_k\left(t\right)-a_k\dfrac{\hat{g}_k\left(\hat{\theta}_k\right)}{\left\|\hat{g}_k\left(\hat{\theta}_k\right)\right\|_\infty}$.

This latter version is denoted 1SP - GSS (Gradient Smoothing and Scaling).

Inequality constraints can also be taken into account by a projection algorithm introduced in [4], i.e. the current parameter estimate is projected onto a closed set included in the admissible region in such a way that no function evaluation is required outside this latter region. In this study, bound constraints (e.g., positivity constraints) are handled in this way.

The second-order algorithms 2SAE are based on the following two core recursions, one for the parameter vector θ, the second for the Hessian $H\left(\theta\right)$ of the criterion [6]:

(8) $\hat{\theta}_{k+1}\left(t\right)=\hat{\theta}_k\left(t\right)-a_k.\bar{\bar{H}}_k^{-1}.\hat{g}_k\left(\hat{\theta}_k\right)$,

(9) $\bar{\bar{H}}_k=f_k\left(\bar{H}_k\right)$,

(10) $\bar{H}_k=\dfrac{k}{k+1}\bar{H}_{k-1}+\dfrac{1}{k+1}\hat{H}_k$,

where: $\hat{H}_k$ is a per-iteration symmetric estimate of the Hessian matrix, which is computed from gradient approximations (or direct evaluations) using a simultaneous perturbation approach, $\bar{H}_k$ is a simple sample mean, and f_k is a mapping designed to cope with possible non-positive –definiteness of $\bar{H}_k$.

Again, the algorithm requires only a small number of function evaluations – at least four criterion evaluations to contract the gradient and Hessian estimates – independent of the number of unknown parameters.

Several variants of the mapping f_k have been considered in the literature:

- regularisation through addition of a diagonal perturbation matrix with small positive elements [5];
- a more elaborate regularisation technique recently proposed in which the eigenvalue matrix Λ_k of $\bar{H}_k$ is first "corrected", i.e. negative elements are replaced by a descending series of small positive eigenvalues, and a new matrix $\hat{\Lambda}_k$ is defined. Then, the orthogonal matrix P_k of eigenvectors is used to define the mapping $f_k\left(\bar{H}_k\right)=P_k.\hat{\Lambda}_k.P_k^T$;
- a simplified version of the preceding approach in which the "corrected" eigenvalue matrix $\hat{\Lambda}_k$ is replaced by a constant diagonal matrix defined by the geometric mean of all the eigenvalues.

Mapping (a) is easy to implement but relatively delicate to tune in practical situations (selection of the elements of the perturbation matrix). Mappings are potentially more efficient, but more complex to implement. In addition, some tuning is still required (to select the small positive eigenvalues that are substituted to the negative elements of Λ_k). In this study, a simple, tuning-free, Hessian estimate is considered. Following an idea originally introduced in [2], a diagonal approximation of the Hessian is built,

(11) $\hat{H}_k=diag\left\{\left[g_k\left(\hat{\theta}_k+c_k\Delta_k\right)-g_k\left(\hat{\theta}_k-c_k\Delta_k\right)\right]/\left(2c_k\Delta_k\right)\right\}$,

where the notation indicates a component-wise division of two vectors (in analogy with Matlab programming).

The gradients $g_k.\left(\hat{\theta}_k\pm c_k\Delta_k\right)$ are obtained by one-sided approximation (in order to limit the number of function evaluations)

(12)

$$g_k.\left(\hat{\theta}_k\pm c_k\Delta_k\right)=\begin{bmatrix}\dfrac{y\left(\hat{\theta}_k\pm c_k\Delta_k+\delta_k\Delta_k\right)-y\left(\hat{\theta}_k\pm c_k\Delta_k\right)}{\delta_k\Delta_{k1}}\\ \ldots\\ \dfrac{y\left(\hat{\theta}_k\pm c_k\Delta_k+\delta_k\Delta_k\right)-y\left(\hat{\theta}_k\pm c_k\Delta_k\right)}{\delta_k\Delta_{kp}}\end{bmatrix},$$

where, δ_k is a positive scalar (the sequence $\left\{\delta_k\right\}$ can be chosen in similar way as $\left\{c_k\right\}$, e.g. equation (3)) and $\Delta_k=\left(\Delta_{k1},\Delta_{k2},\ldots,\Delta_{kp}\right)^T$ with symmetrically Bernouilli distributed random variables $\left\{\Delta_{ki}\right\}$ (independent of $\left\{\Delta_{ki}\right\}$ in

equation (3)).

In the same spirit as Eq. (7), an infinity-norm scaling is introduced, i.e.

(13) $\bar{H}_k = \frac{k}{k+1}\bar{H}_{k-1} + \frac{1}{k+1}\hat{H}_k$,

(14) $\bar{\bar{H}}_k = \frac{abs\left(\bar{H}_k\right)}{\left\|\bar{H}_k\right\|_\infty}$,

where, $abs(\cdot)$ is a regularisation in which the absolute value of each of the (diagonal) elements of $\bar{H}_k$ is computed and $\left\|\bar{H}_k\right\|_\infty$ represents the largest of these elements.

This latter algorithm is denoted "adaptive 2SP-DHS" (2^{nd}- order Simultaneous Perturbation algorithm with Diagonal Hessian estimation and Scaling).

MODELLING OF ANIMAL CELL CULTURES

In [8] the authors propose a deterministic model of growth with constraints:

(15) $\frac{d\sigma}{dt} = k.\sigma$,

where, $d\sigma/dt$ is the intensity λ of the growth of the cellular number, and k is a coefficient with the sense of specific number growth.

The solution of this equation gives an exponential growth of the number of cells in the system:

(16) $\sigma(t) = \sigma_0.\exp(kt)$.

Modifications of the proposed equation are possible. The restricting factors on the process dynamics are usually accounted for by variation of the coefficient k. If with $\alpha > 0$ is denoted the degree of influence of a given restricting factor, this equation can be presented as:

(17) $\frac{d\sigma}{dt} = (k - \alpha.\sigma).\sigma$.

Consider batch animal cell cultures described by a simple macroscopic reaction scheme growth [6]:

(18) $\nu_{G\ln} G\ln \xrightarrow{\varphi_g} \wedge X$,

and scheme maintenance:

(19) $G + \nu_X.X \xrightarrow{\varphi_m} \nu_X.X + \nu_L.L$,

where, X, G, $G\ln$ and L represent biomass, glucose, glutamine and lactate, respectively, and $\nu_{G\ln}$, ν_X and ν_L are pseudo-stoichiometric coefficients. The symbol "$\rightarrow\wedge$" means that the growth reaction is auto-catalysed by X and the presence of "$\nu_X.X$" on both sides of the maintenance reaction means that X catalyses this latter reaction.

The growth rate φ_g and the maintenance rate φ_m are described by a general kinetic model structure proposed in [6]:

(20) $\varphi_g(X, G, G\ln) = \alpha_g.X^{\gamma_{g,X}}.G\ln^{\gamma_{g,G\ln}}.e^{-\beta_{g,G}G}$,

(21) $\varphi_m(X, G) = \alpha_m.X^{\gamma_{m,X}}G^{\gamma_{m,G}}.e^{-\beta_{m,X}.X}$.

Simple mass balances allow for the following dynamic model to be derived:

(22) $\frac{dX}{dt} = \varphi_g(X, G, G\ln)$, $X(0) = X_0$,

(23) $\frac{dG}{dt} = -\varphi_m(X, G)$, $G(0) = G_0$,

(24) $\frac{dG\ln}{dt} = -\nu_{G\ln}\varphi_g(X, G, G\ln)$, $G\ln(0) = G\ln_0$,

(25) $\frac{dL}{dt} = \nu_L.\varphi_m(X, G)$, $L(0) = L_0$,

where, $X(t), G(t), G\ln(t)$ and $L(t)$ denote the respective component concentrations.

Identification of bioprocess models is a delicate task and in [6], a systematic procedure is proposed, which allows the pseudo-stoichiometric coefficients to be estimated independently of the kinetic coefficients by minimising a maximum-likelihood criterion. This procedure also considers the estimation of the most likely initial conditions (since the concentration measurements are corrupted by noise at each sampling time, including the initial one).

In this study, it is assumed that the pseudo-stoichiometric coefficients have already been estimated following the above-mentioned procedure and that only the kinetic coefficients and the initial component concentrations have to be inferred from rare and asynchronous measurements of biomass, glucose, glutamine and lactate concentrations. The measurement equation is given by

(26) $y(t_i) = x(t_i) + \varepsilon(t_i)$, $i = 1, ..., N$,

where,

$x(t_i) = \left[X(t_i)\,G(t_i)\,G\ln(t_i)\,L(t_i)\right]^T$,

$y(t_i)$ and $\varepsilon(t_i)$ are the state, measurement

and noise vectors at time t_i, respectively. The measurement errors are assumed to be normally distributed, with noises having zero mean and variance matrix $Q(t_i)$.

Data are collected from seven batch experiments corresponding to different initial glucose and glutamine concentrations. Five of these experiments are used for parameter estimation, the two remaining ones being used for cross-validation tests.

The 28 unknown parameters (8 kinetic coefficients and 20 initial concentrations) are estimated by minimising a maximum likelihood cost function taking into account the measurement noises, i.e.

(27)

$$\min_{\theta} J_{m1}(\theta) = \min_{\theta} \frac{1}{2} \sum_{i=1}^{N} \left(y_i - \hat{x}_i(\theta)\right)^T . Q_i^{-1} . \left(y_i - \hat{x}_i(\theta)\right),$$

where, y_i, Q_i and $\hat{x}_i(\theta)$ are the measurement vector, the measurement error covariance matrix and the state estimate obtained by integration of the model equations (21-24) with the parameters θ at time t_i, respectively.

The tuning parameters of 1SP-GS are selected as follows: $c = 10^{-4}, \gamma = 0{,}15$ (a very slowly decaying sequence c_k is used for gradient evaluation), $a_0 = 10^{-6}; \eta = 1{,}01; \mu = 0{,}99; \beta = 0$ (no relative increase in the criterion is allowed), $\rho_0 = 0{,}99$. For 1SP-GSS, the same parameters are used, expert $a_0 = 10^{-3}$. Starting with the measured initial concentrations (which are effected by measurement errors) and in initial guess for the kinetic parameters corresponding to a criterion value $J_{m1} = 65761$, the minimisation problem (26) is repeated 10 times with both algorithms.

CONCLUSION

The following inferences could be drawn:

1. The simultaneous perturbation approach developed by Spall [3, 5] is a very powerful technique, which allows an approximation of the gradient of the objective function to be computed by effecting simultaneous random perturbations in all the parameters.
2. Therefore, this approach is particularly well suited to problems involving a relatively large number of design parameters.
3. In this study, variants of first- and second-order SP algorithms are considered and applied to the identification of the kinetic parameters and the initial conditions of a bioprocess model from experimental measurements of a few macroscopic components.

REFERENCES

[1] Renotte, C., Vande Wouwer, A., Remy, M., Neural Modeling and Control of a Heat Exchanger based on SPSA Techniques. *Proceedings of the Workshop,* ACC, 2000

[2]. Wouwer A. Vande, C. Renotte, M. Remy, On the Use of Simultaneous Perturbation Stochastic Approximation for Neural Network Training. *Proceedings of the Workshop*, ACC, 1999

[3]. Spall J., Multivariate Stochastic Approximation Using a Simultaneous Perturbation Gradient Approximation. *IEEE Trans. Automat. Contr,* 45, 2000

[4].Sadegh P., Constrained Optimization via Stochastic Approximation with a Simultaneous Perturbation Gradient Approximation. *Automatica,* 33, 1997

[5].Spall J., Adaptive Stochastic Approximation by the Simultaneous Perturbation Method. *IEEE Trans. Automat. Contr.* 45, 2000

[6].Bogaerts Ph., R. Hanus, Macroscopic Modeling of Bioprocesses with a View to Engineering Applications. In: Focus on Biotechnology, Vol. 4, Kluver, 2000

[7]. Petrov N,. Use Reliability of Risk Technical Systems. Tracian University, St. Zagora, Yambol, "Uchkov", ISBN 954-9978-26-5, 2002

[8].Kalchev I., S. Yordanov,. Biometric ecomonitoring: mathematical models, estimates and predictions, CPS'96, Volume 1, Bratislava, Slovak Republic, 14-15 May 1996, pp. 480-483