Stochastic approximation techniques applied to parameter estimation in a biological model

C. Renotte, A. Vande Wouwer*

Service d'Automatique, Faculté Polytechnique de Mons, 31 Boulevard Dolez, 7000 Mons, Belgium
(*tel: +32-(0)65-374141; fax: +32-(0)65-374136; email: Alain.VandeWouwer@fpms.ac.be)

Abstract: *Simultaneous perturbation stochastic approximation (SPSA) is a class of optimization algorithms which compute an approximation of the gradient and/or the Hessian of the objective function by varying all the elements of the parameter vector simultaneously and therefore, require 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 potentialities are assessed in the context of nonlinear system identification. To this end, a challenging modeling application is considered, i.e. dynamic modeling of batch animal cell cultures from sets of experimental data. The performance of the optimization algorithms are discussed in terms of efficiency, accuracy and ease of use.*

Keywords: stochastic approximation, optimization, nonlinear identification, biotechnology.

1 INTRODUCTION

Process modeling requires the estimation of several unknown parameters from noisy measurement data. To this end, a least-squares or maximum-likelihood cost function (depending on the assumptions on the measurement noise) is usually minimized using a gradient-based optimization method.

Several techniques for computing the gradient of the cost function are available, including finite difference approximations and analytic differentiation. This latter technique leads to backpropagation in neural networks or 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 is directly proportional to the number of unknown model parameters, which becomes an issue for models involving a large number of parameters. This is typically the case in NN modeling, 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 SPSA algorithm [6].

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

- variants of 1SPSA/2SPSA algorithms, in which scaling of the gradient/Hessian estimates is introduced to avoid potential large variations in the course of the optimization 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 organized as follows. Section 2 introduces the basic principles of the first- and second-order SPSA 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.

2 SPSA ALGORITHMS

Consider the problem of minimizing a, possibly noisy, objective function $J(\theta)$ with respect to a vector θ of unknown parameters

1SPSA is given by the following core recursion for the parameter vector θ [5].

$$\hat{\theta}_{k+1} = \hat{\theta}_k - a_k \, \hat{g}_k(\hat{\theta}_k) \tag{1}$$

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

$$\hat{g}(\theta_k) = \begin{bmatrix} \dfrac{J(\hat{\theta}_k + c_k\Delta_k) - J(\hat{\theta}_k - c_k\Delta_k)}{2c_k\Delta_{k1}} \\ \cdots \\ \dfrac{J(\hat{\theta}_k + c_k\Delta_k) - J(\hat{\theta}_k - c_k\Delta_k)}{2c_k\Delta_{kp}} \end{bmatrix} \tag{2}$$

where c_k is a positive scalar and $\Delta_k = (\Delta_{k1}, \Delta_{k_2}, ..., \Delta_{kp})^T$ with symmetrically Bernouilli distributed random variables $\{\Delta_{ki}\}$.

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

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

which ensure asymptotic convergence results. However, performance in finite samples can be different, and numerical experiments suggest that an adaptive gain sequence for parameter updating [3, 7] can enhance convergence and stability (this is particularly true when solving a non convex parameter identification problem), i.e.

$$\begin{aligned} a_k &= \eta \, a_{k-1}, \quad \eta \geq 1, \quad \text{if } J(\theta_k) < (1+\beta)J(\theta_{k-1}) \\ a_k &= \mu \, a_{k-1}, \quad \mu \leq 1, \quad \text{if } J(\theta_k) \geq (1+\beta)J(\theta_{k-1}) \end{aligned} \tag{4}$$

In addition to gain attenuation when the value of the criterion becomes worse, "blocking" mechanisms [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 (4) represents the 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 of c being selected so as to overcome the influence of (numerical or experimental) noise. In the optimum neighborhood, 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" - neighborhood of the optimum).

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

$$G_k = \rho_k G_{k-1} + (1 - \rho_k)\hat{g}_k(\hat{\theta}_k), \; 0 \leq \rho_k \leq 1, \; G_0 = 0 \tag{5}$$

where ρ_k is decreased in a way similar to (4) 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 "adaptive 1SP-GS") [3, 7].

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 normalizing each direction of the gradient vector $\hat{g}_k(\hat{\theta}_k)$ with respect to its largest component (infinity norm scaling)

$$\hat{\theta}_{k+1} = \hat{\theta}_k - a_k \frac{\hat{g}_k(\hat{\theta}_k)}{\left\| \hat{g}_k(\hat{\theta}_k) \right\|_\infty} \tag{6}$$

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 2SPSA are based on the following two core recursions, one for the parameter vector θ, the second for the Hessian $H(\theta)$ of the criterion [6]

$$\hat{\theta}_{k+1} = \hat{\theta}_k - a_k \, \overline{\overline{H}}_k^{-1} \, \hat{g}_k(\hat{\theta}_k) \;, \tag{7}$$

$$\overline{\overline{H}}_k = f_k(\overline{H}_k) \tag{8}$$

$$\overline{H}_k = \frac{k}{k+1}\overline{H}_{k-1} + \frac{1}{k+1}\hat{H}_k \tag{9}$$

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, $\overline{H}_k$ is a simple sample mean, and f_k is a mapping designed to cope with possible non-positive-definiteness of $\overline{H}_k$.

Again, the algorithm requires only a small number of function evaluations - at least four criterion evaluations to

construct 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:

a) regularization through addition of a diagonal perturbation matrix with small positive elements [6];
b) a more elaborate regularization technique recently proposed in [8], in which the eigenvalue matrix Λ_k of $\overline{H}_k$ is first "corrected", i.e. negative elements are replaced by a descending series of small positive eigenvalues, and a new $\hat{\Lambda}_k$ matrix is defined. Then, the orthogonal matrix P_k of eigenvectors is used to define the mapping $f_k(\overline{H}_k) = P_k \hat{\Lambda}_k P_k^T$;
c) 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 [8].

Mapping (a) is easy to implement, but relatively delicate to tune in practical situations (selection of the elements of the perturbation matrix). Mappings (b-c) 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 [7], a diagonal approximation of the Hessian is built,

$$\hat{H}_k = \mathrm{diag}\left((g_k(\hat{\theta}_k + c_k\Delta_k) - g_k(\hat{\theta}_k - c_k\Delta_k))./(2c_k\Delta_k) \right) \quad (10)$$

where the notation (./) indicates a componentwise division of two vectors (in analogy with Matlab programming).

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

$$g_k(\hat{\theta}_k \pm c_k\Delta_k) = \begin{bmatrix} \dfrac{y(\hat{\theta}_k \pm c_k\Delta_k + \tilde{c}_k\tilde{\Delta}_k) - y(\hat{\theta}_k \pm c_k\Delta_k)}{\tilde{c}_k\tilde{\Delta}_{k1}} \\ \cdots \\ \dfrac{y(\hat{\theta}_k \pm c_k\Delta_k + \tilde{c}_k\tilde{\Delta}_k) - y(\hat{\theta}_k \pm c_k\Delta_k)}{\tilde{c}_k\tilde{\Delta}_{kp}} \end{bmatrix} \quad (11)$$

where $\tilde{c}_k$ is a positive scalar (the sequence $\{\tilde{c}_k\}$ can be chosen in a similar way as $\{c_k\}$, e.g. Eq. (3)) and $\tilde{\Delta}_k = (\tilde{\Delta}_{k1}, \tilde{\Delta}_{k_2}, ..., \tilde{\Delta}_{kp})^T$ with symmetrically Bernouilli distributed random variables $\{\tilde{\Delta}_{ki}\}$ (independent of $\{\Delta_{ki}\}$ in (2)).

In the same spirit as Eq. (6), an infinity-norm scaling is introduced, i.e.

$$\overline{H}_k = \frac{k}{k+1}\overline{H}_{k-1} + \frac{1}{k+1}\hat{H}_k \quad (12)$$

$$\overline{\overline{H}}_k = \frac{\mathrm{abs}(\overline{H}_k)}{\left\|\overline{H}_k\right\|_\infty} \quad (13)$$

where abs(•) is a regularization in which the absolute value of each of the (diagonal) elements of $\overline{H}_k$ is computed and $\left\|\overline{H}_k\right\|_\infty$ 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).

3 MODELING OF ANIMAL CELL CULTURES

Consider batch animal cell cultures described by a simple macroscopic reaction scheme

growth: $\nu_{G\,ln}\, G\,ln \overset{\varphi_g}{\rightarrow}\!\!\hat{}\; X$ (14)

maintenance: $G + \nu_X\, X \overset{\varphi_m}{\rightarrow} \nu_X\, X + \nu_L\, L$ (15)

where X, G, Gln and L represent biomass, glucose, glutamine and lactate, respectively, and $\nu_{G\,ln}$, ν_X and ν_L are pseudo-stoechiometric coefficients. The symbol "$\rightarrow\!\!\hat{}$" means that the growth reaction is auto-catalyzed by X and the presence of "ν_X X" in both sides of the maintenance reaction means that X catalyzes this latter reaction.

The growth rate φ_g and the maintenance rate φ_m are described by a general kinetic model structure proposed in [2]

$$\varphi_g(X,G,G\,ln) = \alpha_g X^{\gamma_{g,X}} G\,ln^{\gamma_{g,G\,ln}}\, e^{-\beta_{g,G} G} \quad (16)$$

$$\varphi_m(X,G) = \alpha_m X^{\gamma_{m,X}} G^{\gamma_{m,G}}\, e^{-\beta_{m,X} X} \quad (17)$$

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

$$\frac{dX}{dt} = \varphi_g(X,G,G\,ln) \qquad X(0) = X_0 \quad (18)$$

$$\frac{dG}{dt} = -\varphi_m(X,G) \qquad G(0) = G_0 \quad (19)$$

$$\frac{dG\,ln}{dt} = -\nu_{G\,ln}\varphi_g(X,G,G\,ln) \qquad Gln(0) = Gln_0 \quad (20)$$

$$\frac{dL}{dt} = \nu_L \varphi_m(X,G) \qquad L(0) = L_0 \quad (21)$$

where X(t), G(t), Gln(t) and L(t) denote the respective

component concentrations.

Identification of bioprocess models is a delicate task and in [2], a systematic procedure is proposed, which allows the pseudo-stoechiometric coefficients to be estimated independently of the kinetic coefficients [1] by minimizing 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-stoechiometric 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

$$y(t_i) = x(t_i) + \varepsilon(t_i) \qquad i = 1,\ldots,N \qquad (22)$$

where $x(t_i) = [X(t_i)\ G(t_i)\ Gln(t_i)\ L(t_i)]^T$, $y(t_i)$ 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, white noises with 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 minimizing a maximum likelihood cost function taking into account the measurement noises, i.e.

$$\min_{\theta} J_{ml}(\theta) = \min_{\theta} \frac{1}{2} \sum_{i=1}^{N} (y_i - \hat{x}_i(\theta))^T Q_i^{-1} (y_i - \hat{x}_i(\theta)) \qquad (23)$$

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 (18-21) 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, except $a_0 = 10^{-3}$. Starting with the measured initial concentrations (which are affected by measurement errors) and an initial guess for the kinetic parameters corresponding to a criterion value $J_{ml} = 65761$, the minimization problem (23) is repeated 10 times with both algorithms. The evolution of the criterion value as a function of the number of iterations is represented in Fig. 1. Up to 50000 iterations are considered, which might appear quite large at first sight, but the computational cost is very modest as each iteration only requires two criterion evaluations (each of these evaluations involves 5 model simulations corresponding to the 5 experimental batches used in this identification phase). On the other hand, standard centered finite difference approximations would require 56 criterion evaluations per iteration !

From Fig. 1, it is apparent that the results obtained with 1SP-GSS are much less dispersed ($202 < J_{ml} < 221$) than those obtained with 1SP-GS ($233 < J_{ml} < 5304$)

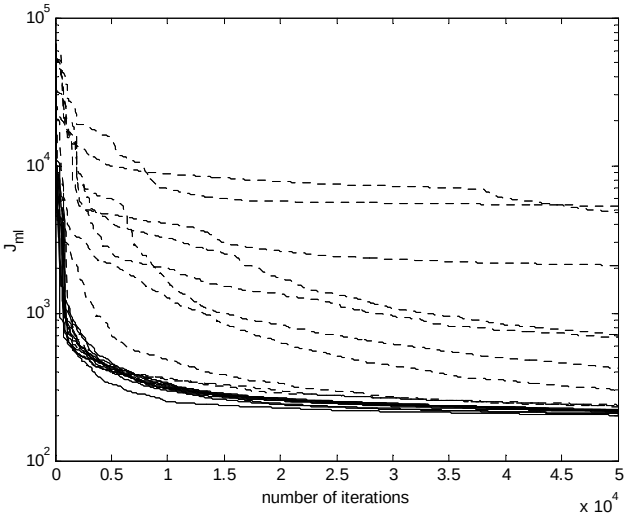

Fig. 1 - Evolution of the criterion as a function of the iteration number
(dashed lines: 1SP-GS; solid lines: 1SP-GSS)

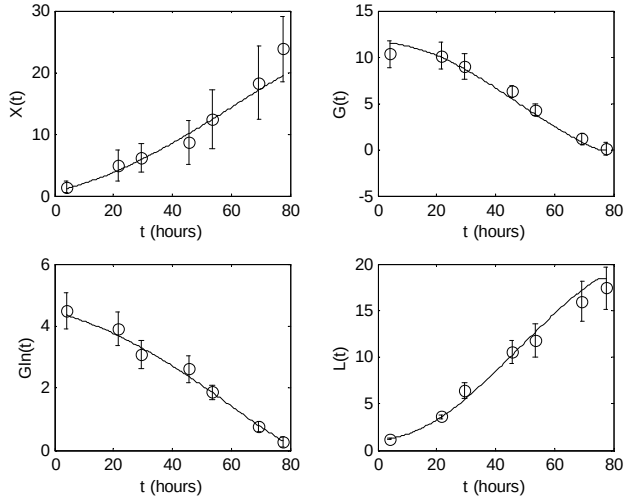

Fig. 2 - Direct validation (experiment 4)

The parameter estimates corresponding to the best run ($J_{ml} = 202$ with 1SP-GSS) are listed in Table 1. Fig. 2 compares the measurement data of one of the five experiments used in the parameter identification procedure with the model prediction (direct validation), whereas Fig. 3 shows the same kind of comparison with the measurement data of one of the remaining two experiments (cross-validation). In these graphs, the circled points are the measured data and the bars represent the 99% confidence intervals. The solid lines are the concentration trajectories predicted by the identified model. These figures demonstrate the excellent model

agreement.

Table 1. Parameter estimates

$\alpha_g = 0.0961$	$\alpha_m = 0.0321$
$\gamma_{g,X} = 0.4127$	$\gamma_{m,X} = 1.2278$
$\gamma_{g,G\ln} = 0.2094$	$\gamma_{m,G} = 0.0748$
$\beta_{g,G} = 0.0098$	$\beta_{m,X} = 0.1070$

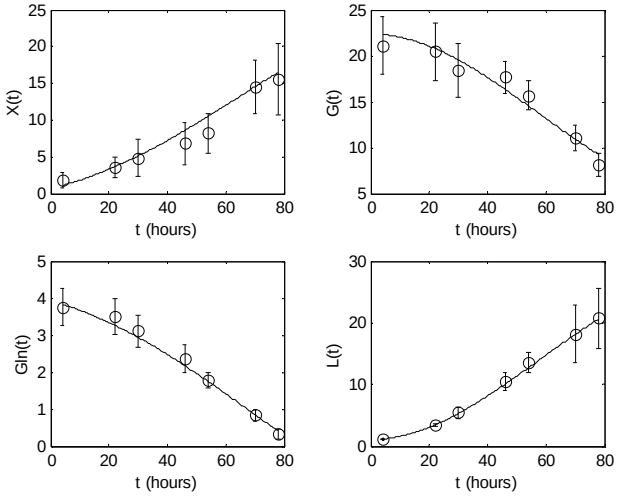

Fig. 3 - Cross-validation (experiment 1)

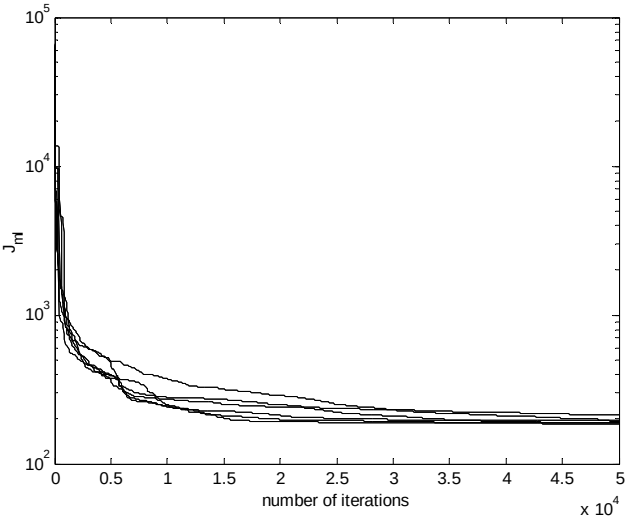

Fig. 4 – 2SP-DHS: evolution of the criterion as a function of the iteration number

For comparison purposes, cost function (23) is also minimized using 2SP-DHS, and the evolution of the criterion value as a function of the number of iterations is represented in Fig. 4, for 6 independent runs. The tuning parameters of 2SP-DHS are selected as follows: $c = 10^{-4}$, $\tilde{c} = 10^{-3}$, $\gamma = 0.15$, $a_0 = 10^{-3}$, $\eta = 1.01$, $\mu = 0.99$, $\beta = 0$, $\rho_0 = 0.99$.

The performance of 2SP-DHS is slightly better than for 1SP-GSS, both in terms of speed of convergence and accuracy. For instance, the final value of the cost function lies between 187 and 214 for the 6 independent runs considered in Fig. 4, whereas $202 < J_{ml} < 221$ when using 1SP-GSS (see Fig. 1). However, the benefits are small compared to the computational overhead (2 additional function evaluations/iteration are required to estimate the Hessian), so that we recommend the use of 1SP-GSS in most applications.

4 CONCLUSION

The simultaneous perturbation approach developed by Spall [5, 6, 8] 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. Therefore, this approach is particularly well-suited to problems involving a relatively large number of design parameters. 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.

ACKNOWLEDGEMENTS

The authors are very grateful to Prof. J. Wérenne (Université Libre de Bruxelles, Departement of Animal Cell Biotechnology) and Mr. M. Cherlet for providing the measurement data for the CHO-animal cell cultures.

REFERENCES

[1] G. Bastin, D. Dochain. *On-line estimation and adaptive control of bioreactors*, Elsevier, 1990.

[2] Ph. Bogaerts, R. Hanus. Macroscopic modelling of bioprocesses with a view to engineering applications. In*: Focus on Biotechnology, vol. IV (Engineering and Manufacturing for Biotechnology)* (Ph. Thonart, M. Hofman, Eds.), Kluwer, (2000).

[3] C. Renotte, A. Vande Wouwer, M. Remy. Neural Modeling and Control of a Heat Exchanger based on SPSA Techniques, *Proceedings ACC,* 3299-3303, (2000).

[4] P. Sadegh. Constrained Optimization via Stochastic Approximation with a Simultaneous Perturbation Gradient Approximation, *Automatica* **33**, 889-892, (1997).

[5] J.C. Spall. Multivariate Stochastic Approximation Using a Simultaneous Perturbation Gradient Approximation, *IEEE Trans. Automat. Contr.* **37**, 332-341 (1992).

[6] J.C Spall. Adaptive Stochastic Approximation by the Simultaneous Perturbation Method, *IEEE Trans. Automat. Contr.* **45,** 1839-1853, (2000).

[7] A. Vande Wouwer, C. Renotte, M. Remy. On the Use of Simultaneous Perturbation Stochastic Approximation for Neural Network Training, *Proceedings ACC,* 388-392, (1999).

[8] X. Zhu, J.C. Spall. A Modified Second Order SPSA Optimization Algorithm for Finite Samples, *Int. J. of Adaptive Control and Signal Processing* **16***,* 397-409 (2002).