Neural Processing Letters **7:** 81–89, 1998.

On-Line Learning Fokker-Planck Machine ⋆

J.A.K. SUYKENS, H. VERRELST and J. VANDEWALLE
Katholieke Universiteit Leuven, Department of Electrical Engineering, ESAT-SISTA, Kardinaal Mercierlaan 94, B-3001 Leuven (Heverlee), Belgium
E-mail: johan.suykens@esat.kuleuven.ac.be

Key words: RBF networks, Gaussian mixture distribution, global optimization, Fokker-Planck equation, constrained LMS, regularization

Abstract. In this letter we present an on-line learning version of the Fokker-Planck machine. The method makes use of a regularized constrained normalized LMS algorithm in order to estimate the time-derivative of the parameter vector of a radial basis function network. The RBF network parametrizes a transition density which satisfies a Fokker-Planck equation, associated to continuous simulated annealing. On-line learning using the constrained normalized LMS method is necessary in order to make the Fokker-Planck machine applicable to large scale nonlinear optimization problems.

1. Introduction

In (Suykens et al., 1996; Suykens & Vandewalle, 1995; Suykens & Vandewalle, 1996) the Fokker-Planck learning machine has been introduced as a new method for global optimization of differentiable cost functions. The method is derived from continuous simulated annealing (Gelfand & Mitter, 1991; Gelfand & Mitter, 1993; Kushner, 1987) (or recursive stochastic algorithms in a discrete time context) by considering the associated Fokker-Planck equation in the transition density. The step from the Fokker-Planck equation to the Fokker-Planck machine is made by parametrizing the density with a radial basis function network, corresponding to a Gaussian mixture distribution (Haykin, 1996; Amari, 1995; Streit & Luginbuhl, 1994) or Gaussian sum approximation (Alspach & Sorenson, 1972).

By sampling the search space and evaluating the Fokker-Planck equation in these points, a set of equations is obtained in the time-derivative of the parameter vector of the RBF network. Hence the Fokker-Planck machine is a population based method like genetic algorithms (Goldberg, 1989). However it is not driven by cost function values (survival of the fittest) but by the local geometry at the sampling

⋆ This research work was carried out at the ESAT laboratory and the Interdisciplinary Center of Neural Networks ICNN of the Katholieke Universiteit Leuven, in the following frameworks: the Belgian Programme on Interuniversity Poles of Attraction, initiated by the Belgian State, Prime Minister's Office for Science, Technology and Culture (IUAP P4–02 and IUAP P4–24), a Concerted Action Project MIPS (Modelbased Information Processing Systems) of the Flemish Community and the FWO (Fund for Scientific Research - Flanders) project G.0262.97 : Learning and Optimization: an Interdisciplinary Approach. The scientific responsibility rests with its authors.

points, characterized by the gradient and diagonal elements of the Hessian. The basic Fokker-Planck machine Suykens et al., 1996; Suykens & Vandewalle, 1995) has been extended with incorporation of local optimization steps and stochastic approximation smoothing of the cost function (Suykens & VandeWalle, 1996; Styblinski & Tang, 1990).

In (Suykens et al., 1996; Suykens & Vandewalle, 1995; Suykens & Vandewalle, 1996) an overdetermined set of equations (i.e. more points than the number of parameters in the RBF network) has been solved in order to track the evolution of the density. This has been done in batch mode. As a consequence the method is not directly applicable to high dimensional nonlinear optimization problems. In this letter we present a constrained normalized LMS (Least-Mean-Square) algorithm (Goodwin & Sin, 1984; Haykin, 1996; Widrow & Stearns, 1985) for solving the constrained set of equations. In this way the Fokker-Planck machine is on-line learning by updating the time-derivative of the RBF parameter vector, each time after sampling the search space at a certain point. In order to obtain convergence in the mean of the algorithm, it follows from simulation results that it is needed to apply regularization.

This letter is organized as follows. In Section 2 the basic principles of the Fokker-Planck machine are reviewed. In Section 3 the constrained normalized LMS algorithm is proposed. In Section 4 regularization of this algorithm is discussed. An example on regularization is presented in Section 5.

2. Fokker-Planck Machine

Consider the optimization problem

$$\min_{x \in \mathbb{R}^n} U(x) \tag{1}$$

where $U(.)$ is a twice continuously differentiable cost function defined on the n-dimensional search space $\mathbb{R}^n$. For global optimization of the cost function, recursive stochastic algorithms have been studied in (Gelfand & Mitter, 1991; Gelfand & Mitter, 1993; Kushner, 1987), associated with the following Langevin-type Markov diffusion

$$dx(t) = -\nabla U[x(t)]\, dt + \sigma(t)\, dw(t), \tag{2}$$

with state vector $x \in \mathbb{R}^n$, $w \in \mathbb{R}^n$ a Wiener process, noise intensity $\sigma(t)$ and cooling schedule $\sigma^2(t) = \sigma_0 / \log(t)$ (for t large) and σ_0 a fixed positive constant. This has been called continuous simulated annealing in (Gelfand & Mitter, 1991; Gelfand & Mitter, 1993).

In (Suykens et al., 1996; Suykens & Vandewalle, 1995; Suykens & Vandewalle, 1996) it has been interpreted as a special case of the general nonlinear stochastic differential equation

$$dx = f(x,t)\, dt + \sigma(x,t)\, dw \tag{3}$$

with state vector $x \in \mathbb{R}^n$, $w \in \mathbb{R}^m$ a Wiener process and $f : \mathbb{R}^n \times \mathbb{R} \mapsto \mathbb{R}^n$, $\sigma : \mathbb{R}^n \times \mathbb{R} \mapsto \mathbb{R}^{n \times m}$, the conditional transition density $p(x, t|x_0, t_0)$ satisfies the Fokker-Planck equation (Doob, 1953; Gihman & Skorohod, 1979; van Kampen, 1981; Wong, 1971)

$$\frac{\partial p}{\partial t} = -\sum_i \frac{\partial}{\partial x_i}(p f_i) + \frac{1}{2}\sum_i \sum_j \frac{\partial^2}{\partial x_i \partial x_j}(p\beta_{ij}) \tag{4}$$

where $\beta(x, t) = \sigma(x, t)\sigma(x, t)^T$ and $p(x, t|x_0, t_0)$ denotes the probability density of being in state x at time t, given the process is in state x_0 at time t_0. Continuous simulated annealing is a special case of (3) with $f(x, t) = -\nabla U(x)$, $m = n$ and $\sigma(x, t)$ the diagonal matrix $\sigma(t)I_n$ leading to the Fokker-Planck equation

$$\frac{\partial p}{\partial t} = \sum_i \frac{\partial U}{\partial x_i}\frac{\partial p}{\partial x_i} + \sum_i \frac{\partial^2 U}{\partial x_i^2}\, p + \frac{1}{2}\sigma^2(t)\, \sum_i \frac{\partial^2 p}{\partial x_i^2}. \tag{5}$$

Then the transition density has been parametrized by the RBF network, yielding the Gaussian mixture distribution

$$\hat{p}(x, t|x_0, t_0) = \sum_{i=1}^{n_h} w_i(t)\, N[x - s_i(t), R_i(t)], \quad \sum_{i=1}^{n_h} w_i(t) = 1, \quad w_i \geq 0 \tag{6}$$

with $N(s, R) = k|R|^{-1/2} \exp(-\frac{1}{2}s^T R^{-1} s)$, $k = (2\pi)^{-n/2}$ (Haykin, 1996; Streit & Luginbuhl, 1994; Amari, 1995; Alspach & Sorenson, 1972). n_h denotes the number of hidden neurons or centra, w_i the ith weight of the output layer, $s_i \in \mathbb{R}^n$ the center and $R_i \in \mathbb{R}^{n \times n}$ the covariance matrix related to the ith hidden neuron. The parametrization with $\sum_{i=1}^{n_h} w_i(t) = 1$ and $w_i \geq 0$ ensures that $\hat{p}$ is a density. The matrices R_i are assumed to be diagonal.

The Fokker-Planck equation in $\hat{p}$ is evaluated then at N points, which yields a constrained set of equations of the form

$$A(\theta)\,\dot{\theta} = b(\theta) \quad \text{such that} \quad c^T \dot{\theta} = 0 \tag{7}$$

where $\theta \in \mathbb{R}^q$ denotes the parameter vector of the RBF network and $A \in \mathbb{R}^{N \times q}$, $b \in \mathbb{R}^N$ and $c \in \mathbb{R}^q$ (Suykens et al., 1996; Suykens & Vandewalle, 1995; Suykens & Vandewalle, 1996). The constraint follows from the property $\sum_i w_i = 1$. It has been assumed that $N > q$. Hence the time-derivative of the RBF parameter vector is estimated from an overdetermined set of equations and is based on the knowledge of the gradient and diagonal elements of the Hessian at the sampling points.

This leads to the following basic algorithm for the Fokker-Planck machine:

1. *First generation: choose initial θ of RBF network.*
2. *Generate N points according to $\hat{p}$.*
3. *Calculate $\frac{\partial U}{\partial x_i}$, $\frac{\partial^2 U}{\partial x_i^2}$ at the N points.*
4. *Estimate $\dot{\theta}$ from the constrained linear least squares problem.*

5. *Next generation: compute* $\theta(t+dt) = \theta(t) + \dot{\theta}dt$.
6. *Remove centers with negative weights.*
7. *Update the noise intensity* σ*, according to the cooling scheme.*
8. *Go to 2, unless* n_g *is exceeded.*

Here σ_0, n_h, N, the number of generations n_g, the initial θ and the step size dt serve as input parameters for the algorithm. For a more sophisticated scheme with incorporation of local optimization steps and stochastic approximation smoothing of the cost function the reader is referred to (Suykens & Vandewalle, 1996).

3. On-line Learning using Constrained Normalized LMS

In (Suykens et al., 1996; Suykens & Vandewalle, 1995; Suykens & Vandewalle, 1996) a solution in least squares sense (Bolub & Van Loan, 1989) has been considered to the constrained set of equations (7):

$$\min_u \ \|Au - b\|_2^2 \quad \text{such that} \quad c^T u = 0 \tag{8}$$

where $\dot{\theta}$ is denoted by u. Taking the Lagrangian with Lagrange multiplier λ

$$\mathcal{L}(u, \lambda) = (Au - b)^T (Au - b) + \lambda\, c^T u \tag{9}$$

one obtains the following solution from the conditions for optimality $\frac{\partial \mathcal{L}}{\partial u} = 0$, $\frac{\partial \mathcal{L}}{\partial \lambda} = 0$:

$$\begin{cases} u \ = \ (A^T A)^{-1}(A^T b - \lambda c) \\ \lambda \ = \ \frac{c^T (A^T A)^{-1} A^T b}{c^T (A^T A)^{-1} c}. \end{cases} \tag{10}$$

Recursive (on-line) algorithms with systolic array implementations have been discussed e.g. in (Moonen & Vandewalle, 1991; Vanpoucke & Moonen, 1995). However, for high dimensional optimization problems it is not feasible to estimate u in this way due to a large A matrix and the matrix product $A^T A$. Therefore we will work with vector updates by employing LMS (Least-Mean-Square) type algorithms, which are well-known in adaptive filtering (Haykan, 1996; Goodwin & Sin, 1984; Widrow & Stearns, 1985). We apply a normalized LMS algorithm (Goodwin & Sin, 1984; Haykin, 1996), which shows faster convergence than LMS (Slock, 1993).

Now we derive a normalized LMS algorithm which takes into account the linear constraint. It corresponds to the minimizing solution to

$$\min_{u_{k+1}} \ \alpha^2 \|u_{k+1} - u_k\|_2^2 + (b_k - a_k^T u_{k+1})^2 \quad \text{s.t.} \quad c^T u_{k+1} = 0 \tag{11}$$

where a_k^T, b_k denote the kth row and kth element of A, b respectively. Each time the search space has been sampled, a_k and b_k are calculated, producing an estimate

u_{k+1} for this sample k. In this way the Fokker-Planck machine is learning on-line. We write (11) as

$$\min_{u_{k+1}} \| \begin{bmatrix} \alpha I \\ a_k^T \end{bmatrix} u_{k+1} - \begin{bmatrix} \alpha u_k \\ b_k \end{bmatrix} \|_2^2 \quad \text{s.t.} \quad c^T u_{k+1} = 0 \tag{12}$$

which brings the problem in the form (8). From the Matrix Inversion Lemma (Goodwin & Sin, 1984) $(\mathcal{A}+\mathcal{BC})^{-1} = \mathcal{A}^{-1} - \mathcal{A}^{-1}\mathcal{B}\,(\mathcal{I}+\mathcal{C}\,\mathcal{A}^{-1}\mathcal{B})^{-1}\mathcal{C}\mathcal{A}^{-1}$ with $\mathcal{A} = \alpha^2 I$, $\mathcal{B} = u_k$, $\mathcal{C} = u_k^T$ one derives the expressions

$$\begin{aligned} (A^T A)^{-1} &= \tfrac{1}{\alpha^2}(I - \tfrac{a_k a_k^T}{\alpha^2 + a_k^T a_k}) \\ (A^T A)^{-1} A^T b &= u_k + \tfrac{1}{\alpha^2 + a_k^T a_k} a_k (b_k - a_k^T u_k). \end{aligned} \tag{13}$$

This yields the constrained normalized LMS algorithm:

$$\begin{cases} u_{k+1} = u_k + \frac{1}{\alpha^2 + a_k^T a_k} a_k (b_k - a_k^T u_k) - \lambda (I - \frac{a_k a_k^T}{\alpha^2 + a_k^T a_k}) c \\ \lambda = \dfrac{c^T [u_k + \frac{1}{\alpha^2 + a_k^T a_k} a_k (b_k - a_k^T u_k)]}{c^T [I - \frac{a_k a_k^T}{\alpha^2 + a_k^T a_k}]\, c}. \end{cases} \tag{14}$$

The well-known normalized LMS algorithm is a special case for $\lambda = 0$. As starting point we take $u_0 = 0$ (or $\dot{\theta} = 0$) which means no update of the density.

In order to derive a condition for convergence in the mean of (14), let us write it as

$$u_{k+1} = (H_k - \frac{H_k c c^T H_k}{c^T H_k c}) u_k + (H_k - \frac{H_k c c^T H_k}{c^T H_k c}) \frac{1}{\alpha^2} a_k b_k \tag{15}$$

with

$$H_k = I - \frac{a_k a_k^T}{\alpha^2 + a_k^T a_k}.$$

Under certain assumptions (Haykin, 1996; Haykin, 1996) one can write

$$E\{u_{k+1}\} = F\, E\{u_k\} + F\, \frac{1}{\alpha^2} E\{a_k b_k\} \tag{16}$$

with

$$F = E\left\{ H_k - \frac{H_k c c^T H_k}{c^T H_k c} \right\}$$

where $E\{.\}$ denotes the expectation operator over the sample index k. $E\{a_k b_k\}$ is the cross-correlation matrix between the input vector a_k and the desired response

b_k. The linear system (16) with state vector $E\{u_k\}$ is stable (or in algorithmic sense convergent) if $\rho(F) < 1$, where $\rho(.)$ denotes the spectral radius of the matrix. Since F is symmetric this yields the condition

$$\max_i |\lambda_i(F)| < 1. \tag{17}$$

Note that for (unconstrained) LMS, convergence in the mean is determined by the covariance matrix $R = \{a_k a_k^T\}$ instead of by F. Fast convergence is obtained for $\rho(F)$ small. Furthermore, the matrix F does not depend on b_k. As a consequence it doesn't depend on the cost function, but only on the parameter vector of the RBF network itself. As will be demonstrated on an example in Section 5, convergence doesn't occur for RBFs with multiple centra, which is basically due to ill-conditioning of $A^T A$. In order to solve this problem we apply regularization.

4. Regularization

We consider the following regularization to the formulation (11):

$$\min_{u_{k+1}} \alpha^2 \|u_{k+1} - u_k\|_2^2 + (b_k - a_k^T u_{k+1})^2 + u_{k+1}^T \Lambda u_{k+1} \quad \text{s.t.} \quad c^T u_{k+1} = 0 \tag{18}$$

where Λ is a diagonal matrix with positive diagonal elements. A different weight can be given to the adaptations of the output weights, centra and variances in the RBF network, like has been done in (Suykens & Vandewalle, 1996). By formulating the Lagrangian and imposing conditions for optimality, the solution is given by

$$\begin{cases} u_{k+1} = S_k(\alpha^2 u_k + a_k b_k - \lambda c) \\ \lambda = \frac{c^T S_k(\alpha^2 u_k + a_k b_k)}{c^T S_k c} \end{cases} \tag{19}$$

where

$$S_k = \Delta^{-1} - \frac{\Delta^{-1} a_k a_k^T \Delta^{-1}}{1 + a_k^T \Delta^{-1} a_k}$$

with $\Delta = \alpha^2 I + \Lambda$. Like (14) this algorithm can be implemented with vector updates, avoiding the storage of matrices. Convergence in the mean occurs if

$$\rho\left(E\left\{ \alpha^2 \left(S_k - \frac{S_k c c^T S_k}{c^T S_k c} \right) \right\} \right) < 1. \tag{20}$$

The convergence can be influenced by the choice of Λ. Finally, the values of $E\{u_k\}$ will be used as an estimate for $\dot{\theta}$ at a certain generation in step 4 of the basic algorithm for the Fokker-Planck machine.

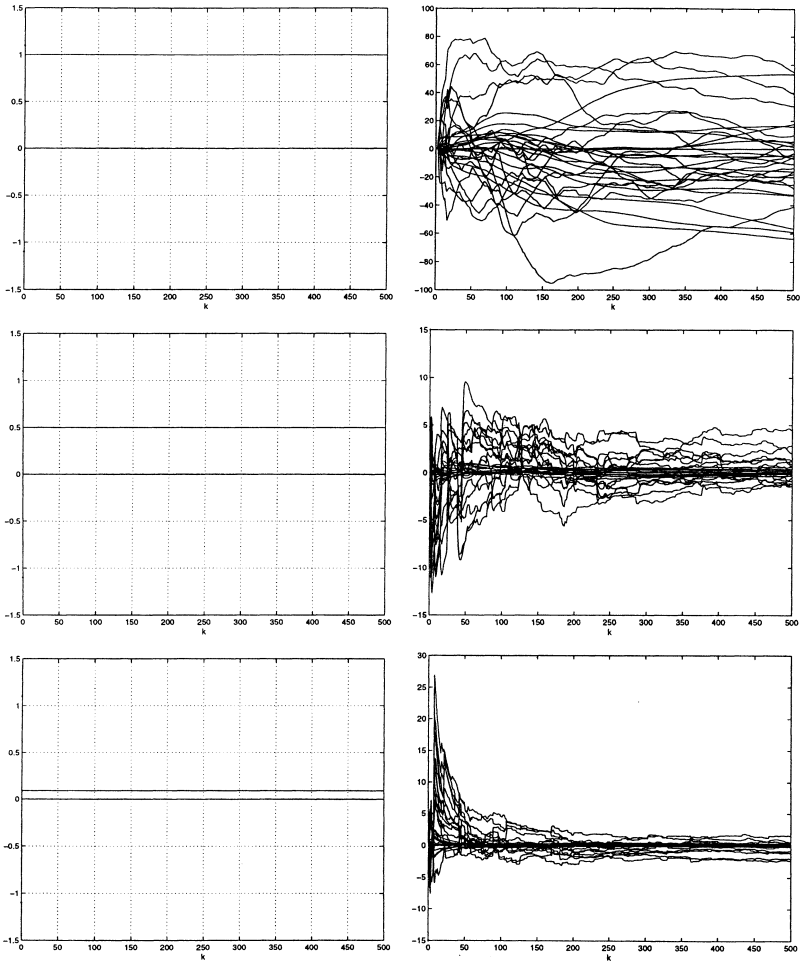

Figure 1. Illustration of the use of regularization in the constrained normalized LMS algorithm: (Left) eigenvalues of matrix F with respect to k; (Right) estimation of $E\{u_k\}$; (Top) no regularization; (Middle) $\Lambda = 0.01I$; (Bottom) $\Lambda = 0.1I$.

5. Example

We illustrate the constrained normalized LMS method on the cost function (Suykens & Vandewalle, 1996)

$$\begin{aligned} U(x) \;=\; & \tfrac{1}{2n}\textstyle\sum_{i=1}^{n} -4n\prod_{i=1}^{n}\cos(0.2x_i) - 4n\prod_{i=1}^{n}\cos(x_i) \\ & -4n\prod_{i=1}^{n}\cos(2x_i) - 4n\prod_{i=1}^{n}\cos(3x_i) \\ & -4n\prod_{i=1}^{n}\cos(4x_i) + 20n \end{aligned} \tag{21}$$

with $x \in \mathbb{R}^5$ ($n = 5$). When multiple centra are used in the RBF network, regularization is needed as is demonstrated on Figure 1 for an RBF with 3 centra. Convergence in the mean $E\{u_k\}$ is shown on the Figure. Without regularization $\rho(F)$ is equal to 1, resulting in a non-convergent algorithm. A nonzero Λ regularization matrix ($\Lambda = 0.01I$ and $\Lambda = 0.1$ are shown on the Figure) makes the algorithm convergent in the mean. $\alpha = 0.1$ has been chosen in the constrained LMS algorithm (19). Diagonal covariance matrices have been taken for the RBF, resulting in a 33 dimensional vector u_k. For the initial parameter vector of the RBF network, the centra have been randomly distributed in a hypercube $[-3, 3]^n$ and the standard deviations of the Gaussians have been taken equal to 3. The population consists of 500 points. Stochastic approximation smoothing of the cost function has been applied (Styblinski & Tang, 1990; Suykens & Vandewalle, 1996).

6. Conclusion

In this letter a constrained normalized LMS method has been discussed in order to track the evolution of an RBF parametrized density in the Fokker-Planck learning machine. In this way an on-line learning algorithm is obtained. The method works with vector updates which makes it suitable for solving high dimensional global optimization problems. This is not the case when the constrained linear least squares problem is solved in batch mode as has been done in previous work. Regularization is needed when multiple centra are used in the RBF network in order to obtain convergence in the mean and to avoid ill-conditioning.

Acknowledgement

We wish to thank Marc Moonen for stimulating discussions about the normalized LMS algorithm.

References

1. D.L. Alspach and H.W. Sorenson, “Nonlinear Bayesian estimation using Gaussian sum approximations”, IEEE Transactions on Automatic Control, Vol. 17, No. 4, 439–448, 1972.
2. S.-I. Amari, “Information geometry of the EM and em algorithms for neural networks”, Neural Networks, Vol. 8, No. 9, 1379–1408, 1995.
3. J.L. Doob, Stochastic processes, John Wiley & Sons, 1953.
4. S.B. Gelfand and S.K. Mitter, “Recursive stochastic algorithms for global optimization in $\mathbb{R}^d$”, SIAM Journal on Control and Optimization, Vol. 29, No. 5, 999–1018, 1991.
5. S.B. Gelfand and S.K. Mitter, “Metropolis-type annealing algorithms for global optimization in $\mathbb{R}^d$”, SIAM Journal on Control and Optimization, Vol. 31, No. 1, 111–131, 1993.
6. I.I. Gihman and A.V. Skorohod, The theory of stochastic processes I, II, III, Springer-Verlag: New York, 1979.
7. D.E. Goldberg, Genetic algorithms in search, optimization and machine learning, Addison-Wesley: Reading, MA, 1989.
8. G.H. Golub and C.F. Van Loan, Matrix Computations, Johns Hopkins University Press: Baltimore MD, 1989.
9. G.C. Goodwin and K.S. Sin, Adaptive filtering, Prediction and Control, Prentice-Hall: Englewood Cliffs, 1984.

10. S. Haykin, Adaptive Filter Theory, Prentice Hall, 1996.
11. S. Haykin, Neural Networks: a Comprehensive Foundation, Macmillan College Publishing Company: Englewood Cliffs, 1994.
12. H.J. Kushner, "Asymptotic global behavior for stochastic approximation and diffusions with slowly decreasing noise effects: global minimization via Monte Carlo", SIAM Journal on Applied Mathematics, 47, 169–185, 1987.
13. M. Moonen and J. Vandewalle, "A square root covariance algorithm for constrained recursive least squares estimation", Journal of VLSI Signal Processing, 3, 163–172, 1991.
14. T. Poggio and F. Girosi, "Networks for approximation and learning", Proceedings of the IEEE, Vol. 78, No. 9, 1481–1497, 1990.
15. D.T.M. Slock, "On the convergence behavior of the LMS and the normalized LMS algorithms", IEEE Transactions on Signal Processing, Vol. 41, No. 9, pp. 2811–2825, Sept. 1993.
16. R.L. Streit and T.E. Luginbuhl, "Maximum likelihood training of probabilistic neural networks", IEEE Transactions on Neural Networks, Vol. 5, No. 5, 764–783, 1994.
17. M.A. Styblinski and T.-S. Tang, "Experiments in nonconvex optimization: stochastic approximation with function smoothing and simulated annealing", Neural Networks, Vol. 3, 467–483, 1990.
18. J.A.K. Suykens, J.P.L. Vandewalle and B.L.R. De Moor, Artificial neural networks for modelling and control of non-linear systems, Kluwer Academic Publishers: Boston, 1996.
19. J.A.K. Suykens and J. Vandewalle, "Nonconvex optimization using a Fokker-Planck Learning Machine", 12th European Conference on Circuit Theory and Design (ECCTD 95), Istanbul Turkey, pp. 983-986, August 1995.
20. J.A.K. Suykens and J. Vandewalle, "A Fokker-Planck learning machine for global optimization", submitted for publication.
21. N.G. van Kampen, Stochastic processes in physics and chemistry, North-Holland, 1981.
22. F. Vanpoucke and M. Moonen, "Systolic robust adaptive beamforming with an adjustable constraint", IEEE Transactions on Aerospace and Electronic Systems, Vol. 31, No. 2, pp. 658–669, April 1995.
23. B. Widrow and S.D. Stearns, Adaptive signal processing, Prentice-Hall: Englewood Cliffs, 1985.
24. E. Wong, Stochastic processes in information and dynamical systems, McGraw-Hill, 1971.