Cooling Hybrid Learning for Neural Networks

Aristoklis D. Anastasiadis [a] *

[a]School of Computer Science and Information Systems, Birkbeck College, University of London, Malet Street, London WC1E 7HX, United Kingdom.

Inspired from our previous algorithm, the HLS, which was based on the theory of nonextensive statistical mechanics, as proposed by Tsallis, we develop a new evolving stochastic learning algorithm. The new scheme combines deterministic and stochastic search by employing a different adaptive stepsize for each weight, and a form of noise that is characterized by the nonextensive entropic index q that is regulated by a weight decay term. The behavior of the learning scheme can be more stochastic or deterministic depending on the trade off between T and q, which is done by introducing a formula between these two important learning parameters. Experimental study verifies that there are indeed improvements in learning speed, which constitutes the new evolving stochastic learning algorithm quicker than HLS.

Keywords: nonextensive statistics, neural systems, noise injection, pattern classification, HLS.

1. Introduction

Real neural cells' responses to identical stimuli have been found to be stochastic in nature, nevertheless the effect of noise on the operation of artificial neural networks has not been investigated in depth. Attempts to explore the benefits of introducing noise during learning have been based on the use of Gaussian distributions[1,4,9]. One of the most famous neural model operating with noise is the Boltzmann machine, [1,3], inspired by the Boltzman–Gibbs entropy $S_{BG} = -K \sum_i p_i ln p_i$ that provides exponential laws for describing stationary states and basic time–dependent phenomena, where $\{p_i\}$ are the probabilities of the microscopic configurations, and $K > 0$.

Simulated Annealing (SA) learning [5] has been explored for the Boltzmann machine. However, the problem with neural systems' error function E is not the well defined local minima but the nearly flat broad regions that do not allow the so-called Metropolis move of the classic SA to escape. Other attempts use hybrid schemes, such as:

$$w^{k+1} = w^k - \eta \nabla E(w^k) + \rho \cdot c \cdot 2^{-d \cdot k}, \qquad (1)$$

where k is the iteration, η is a common fixed stepsize for all weights, ρ is a constant controlling the initial intensity of the noise, $c \in (-0.5, +0.5)$ is a random number and d is the noise decay constant. This approach does not use the notion of the acceptance probability, such as the Metropolis algorithm in the classic SA, or the generalized acceptance probability in the generalised Simulated Annealing [12]. Instead, it implements a form of Langevin noise that has been proved quite effective for neural learning, and has motivated the development of other methods, such as the *Simulated Annealing Rprop*–SARprop and the SARprop with Restarts–ReSARprop [13].

In this paper, a new evolving stochastic learning scheme combining deterministic and stochastic search with a form of noise that is characterized by the nonextensive entropic index q, is discussed. This permits the modification of the error surface during training so that exploration of new regions of the error landscape is achieved. In addition, experiments were conducted to explore the influence of the entropic index q and temperature T on the convergence speed and stability of the proposed method. It is important to mention that the behavior of the learning scheme can be more stochastic or deterministic depending on the trade off between T and q. Finally, a formula between T and q is applied, and preliminary results are very promising.

2. The cooling hybrid learning scheme

This paper introduces a new cooling hybrid strategy for neural systems that builds on the theory of nonextensive statistical mechanics. In this approach noise is generated by a noise source that is characterized by the nonextensive entropic index q. In particular, the nonextensive entropy has been defined as [11]:

$$S_q \equiv K \, \frac{1 - \sum_{i=1}^{W} p_i^q}{q-1} \quad (q \in \mathbb{R}), \qquad (2)$$

where W is the total number of microscopic configurations, whose probabilities are $\{p_i\}$, and K is a conventional positive constant. When the entropic index $q = 1$, (2) recovers to Boltzmann–Gibbs entropy.

*Corresponding author, email: aris@dcs.bbk.ac.uk

The entropic index works like a biasing parameter: $q < 1$ privileges rare events (values of p close to 0 are benefited), while $q > 1$ privileges common events (values of p close to 1). The optimization of the entropic form (2) under appropriate constraints, [11], yields for the canonical ensemble

$$p_i \propto [1-(1-q)\beta E_i]^{\frac{1}{(1-q)}} \equiv e_q^{-\beta E_i}, \tag{3}$$

where β is a Lagrange parameter, $\{E_i\}$ is the energy spectrum, and the *q-exponential function*

$$e_q^x \equiv [1+(1-q)x]^{\frac{1}{(1-q)}} = \frac{1}{[1-(q-1)x]^{\frac{1}{(q-1)}}} \tag{4}$$

In our method, noise is generated according to a schedule:

$$Q(T,k) = e_q^{-T(\ln 2)\cdot k} = [1-(1-q)T(\ln 2)\cdot k]^{\frac{1}{1-q}}, \tag{5}$$

where T is the temperature; k indicates iterations. Noise is not applied proportionally to the size of each weight; instead a form of weight decay is used, which is considered beneficial for achieving a robust neural network that generalizes well. Thus, noise is introduced by formulating the *perturbed* energy function:

$$\tilde{E}(w^k) = E(w^k) + \mu \cdot \sum_{i=1}^{n} \frac{(w_i^k)^2}{[1+(w_i^k)^2]} \cdot Q(T,k), \tag{6}$$

where $E(w)$ is the error function, $\sum_i w_i^2/(1+w_i^2)$ is the weight decay bias term which can decay small weights more rapidly than large weights, and μ is a parameter that regulates the influence of the combined weight decay/noise effect. This form of weight decay modifies the energy landscape so that smaller weights are favoured at the beginning of the training but as learning progresses the magnitude of the weight decay is reduced to favor the growth of large weights. Thus, as the energy landscape is modified during training the search method is allowed to explore regions of the energy surface that were previously unavailable. Minimization of (6) requires calculating the gradient of the energy with respect to each weight

$$\tilde{g}_i(w^k) = g_i(w^k) + \acute{\mu} \cdot \frac{w_i^k}{[1+(w_i^k)^2]^2} \cdot Q(T,k), \tag{7}$$

where $\acute{\mu} > 0$ (in our experiments a fixed value of $\acute{\mu} = 0.01$ was used).

The proposed cooling hybrid scheme applies a sign–based weight adjustment, similar to HLS [2], on the perturbed energy function (6) using the gradient term of Equation (7). Also the learning rates are adapted by Rprop learning procedure [8]. Moreover, an additional condition is introduced in order to avoid using relatively small weight adjustments

$$\begin{aligned} &if \quad \left(\eta_i^{k-1} < \rho \cdot Q^2(T,k)\right) \quad then \\ &\eta_i^k = max\left(\eta_i^{k-1}\eta^- + 2c\rho \cdot Q^2(T,k), \Delta_{min}\right), \end{aligned} \tag{8}$$

where $0 < \rho < 1$ and $c \in (0,1)$ is a random number.

Lastly, inspired from previous work [12], we apply a cooling procedure. This is a formula between T and q, which gives good results in the Generalized Simulated Annealing method proposed by [12]. Using our method and applying the cooling formula, we manage to make the training algorithm, more deterministic. The CHLS is more stochastic, during the first epochs, and becomes deterministic with the passing of time. Thus, when we are close to the mimimum area, the algorithm converges fast. The cooling procedure is given by the next equation:

$$T = T_0 \cdot [\frac{2^{q-1}-1}{(1+epoch)^{q-1}-1}], q > 1 \tag{9}$$

where T_0 is the initial temperature, T is the current temperature, *epoch* is the number of iterations, and q is the Tsallis entropic index.

Below, a simple problem is used to visualize the behavior of the *Cooling Hybrid Learning Scheme*–(CHLS) and compare it with the *Hybrid Learning Scheme*–(HLS), and the Rprop algorithm. It is a single node with two weights and a logistic activation function. The energy landscape of Figure 1 has a global minimum and two local minima. Figure 1 shows that under the same initial conditions, HLS escapes a saddle point and a valley that leads to a local minimum, and converges to the global minimizer located at the center of the contour plot (Figure 1, left), while Rprop converges to the local minimizer (Figure 1, right).

3. Experimental study

We have evaluated the performance of the Cooling Hybrid Learning Scheme (CHLS) and compare it with the Rprop, and the HLS algorithms.

We have used well–studied problems from the UCI Repository of Machine Learning Databases of the University of California [6], as well as problems studied extensively by other researchers in an attempt to reduce as much as possible biases introduced by the size of the weights space. In all cases we have used networks with classic logistic activations. The guidelines of [8] and [2] were adopted for setting the learning parameters of Rprop and HLS respectively.

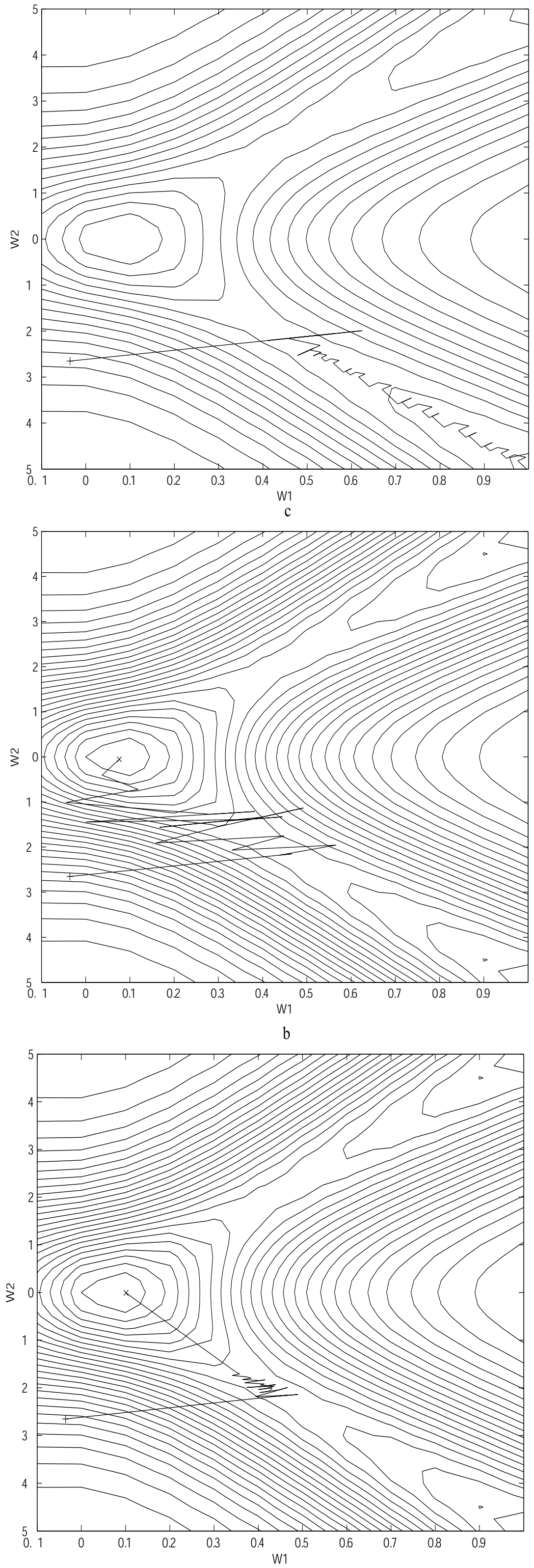

Figure 1. Weights trajectories of the Rprop (top), the Hybrid Learning Scheme HLS, and the Cooling Hybrid Learning Scheme CHLS (bottom).

Below, we report results from 100 independent trials for two UCI problems. These 100 random weight initializations are the same for the three learning algorithms, and the training and testing sets were created according to *Proben1* [7]. The statistical significance of the results has been analyzed using the Wilcoxon test [10]. This is a nonparametric method that is considered an alternative to the paired t–test. It assumes there is information in the magnitudes of the differences between paired observations, as well as the signs. All statements in the tables reported below refer to a significance level of 0.05.

3.1. Diabetes

The *diabetes1* benchmark is a real-world classification task which concerns deciding when a Pima Indian individual is diabetes positive or not [6,7]. There are 8 features representing personal data and results from a medical examination. The Proben1 collection suggests a 8–2–2–2 FNN (34 weights overall). The termination criterion is $E \leq 0.14$ within 2000 iterations.

In order to find the best value for the initial temperature and the tsallis entropic index q, we did 30 different runs. The graph 2 shows the best values of T and q for having better performance the CHLS algorithm.

Judging from the table 1 is obvious that the Rprop algorithm converges many times in local minima. The new stochastic learning algorithm overcomes this problem in the most cases. Its convergence success is 95% while HLS has 94% and Rprop 86%. Furthermore the CHLS is the fastest algorithm and improves significantly the Generalization success compared to Rprop. The cooling procedure seems to affect positive the learning speed of the new algorithm.

Table 1
Comparison of algorithms performance in the Diabetes problem for the converged runs

Diabetes			
Algorithm	Epochs	Generalization	Convergence
Rprop	413 (+)	75.6 (%) (+)	86 (%) (+)
HLS	210 (+)	75.8 (%) (−)	94 (%) (−)
CHLS	164	76.2 (%)	95 (%)

3.2. Cancer

The second benchmark is the *breast cancer diagnosis* problem which classifies a tumor as benign or malignant based on 9 features [6,7]. We have used

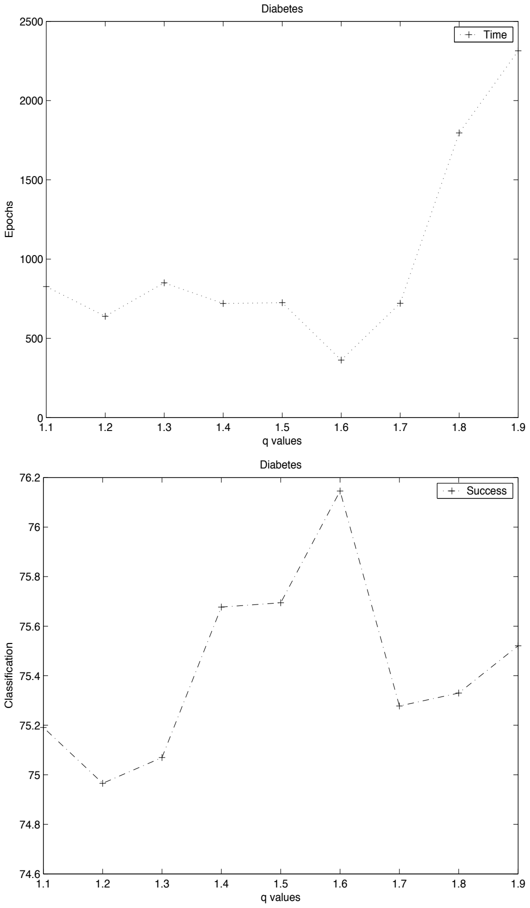

Figure 2. Optimal q based on time, and Success for the diabetes problem

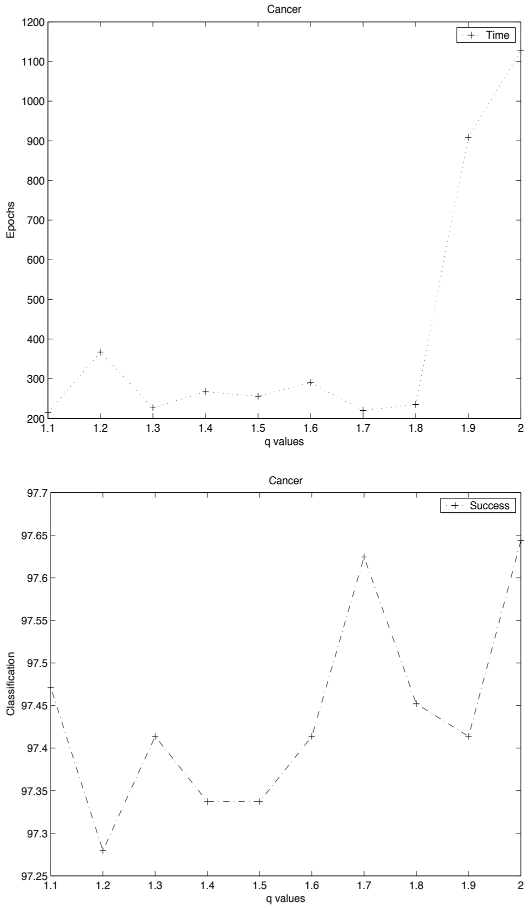

Figure 3. Optimal q based on time, and Success for the cancer problem

an FNN with 9–4–2–2 nodes (a total of 56 weights) as suggested in [7]. The termination criterion is $E \leq 0.02$ within 2000 iterations.

We did 30 different runs to find the best values for the T and q. Figure 3 shows the best values of these two important training parameters. As we can observe from this graph, the value of the q, which gives the best results, is the same for the learning speed and achieving the best classification success. The same phenomenon is occurred in the diabetes problem, as we can mention from the figure 2.

The comparative results are shown in Table 2. The new proposed scheme affects positive to meet fast the error goal. The HLS algorithm improves the convergence success compared to Rprop but the application of the nonextensive term with the cooling procedure in the learning phase, increases significantly the convergence speed and success of the CHLS.

4. Concluding remarks

In this paper we proposed a new evolving stochastic learning scheme that combines deterministic and stochastic search by employing a different adaptive stepsize for each weight, and a form of noise that is characterized by the nonextensive entropic index q. An adaptive formula that introduces a relationship between the T and q is applied. Preliminary experiments with the new scheme have been very encouraging: accelerated and reliable neural learning was achieved in most tested cases.

Further testing is of course necessary to fully ex-

Table 2
Comparison of algorithms performance in the Cancer problem for the converged runs

Cancer			
Algorithm	Epochs	Generalization	Convergence
Rprop	187 (+)	97.4(%) (−)	93(%) (+)
HLS	145 (+)	97.2(%) (−)	97(%) (−)
CHLS	127	97.3(%)	97(%)

plore the advantages and identify possible limitations of this cooling hybrid learning scheme. Moreover, exhaustive testing of the new method in other classes of problems will be done. We will also investigate the performance of CHLS in a restarting mode. Finally, we are going to explore further the properties of Tsallis entropy into Optimization methods in Artificial Intelligence applications.

5. Acknowledgements

I would like to thank Melanie Mitchell, Tom Carter and all the participants for engaging in extremely inspiring discussions in general for making CSSS05 a wonderful month! I would like to address special thanks to Constantino Tsallis for the very helpful discussions related to this work.

REFERENCES

1. D. Ackley. G. Hinton and T. Sejnowski, A learning algorithm for Boltzmann machines. *Cogn. Sci.*, 9, 147–169, 1985.
2. Anastasiadis A.D., Magoulas G.D., “Nonextensive statistical mechanics for hybrid learning of neural networks’, *Physica A: Statistical Mechanics and its Applications*, vol.344, pp. 372-382, 2004.
3. E. H. L. Arts and J. Korst, *Simulated Annealing and Boltzmann Machines*. New York: Wiley, 1989.
4. R. M. Burton and G. J. Mpitsos, Event dependent control of noise enhances learning in neural networks. *Neural Networks*, 5, 627-637, 1992.
5. S. Kirkpatrick, C.D. Gelatt Jr., and M.P. Vecchi, Optimization by simulated annealing. *Science*, 220, 671–680, 1983.
6. P.M. Murphy and D.W. Aha, UCI Repository of machine learning databases, Irvine, CA: University of California, Department of Information and Computer Science, 1994. http://www.ics.uci.edu/ mlearn/MLRepository.html.
7. L. Prechelt, PROBEN1–A set of benchmarks and benchmarking rules for neural network training algorithms, Technical report 21/94, Fakultt fr Informatik, Universitt Karlsruhe, 1994.
8. M. Riedmiller and H. Braun, A direct adaptive method for faster backpropagation learning: The Rprop algorithm. *Proc. Int. Conf. Neur. Net.*, San Francisco, CA, 586-591, 1993.
9. T. Rögnvaldsson, On Langevin updating in multilayer perceptrons. *Neural Computation*, 6, 916–926, 1994.
10. G. Snedecor and W. Cochran, *Statistical Methods*, Iowa State University Press, 8th edition, 1989.
11. C.Tsallis, Possible Generalization of Boltzmann-Gibbs Statistics. *J. Stat. Phys.*, 52, 479–487, 1988.
12. C. Tsallis and D. A. Stariolo, Generalized Simulated Annealing. *Physica A*, 233, 395–406, 1996.
13. N. K. Treadgold and T. D. Gedeon, Simulated Annealing and Weight Decay in Adaptive Learning: The SARPROP Algorithm. *IEEE Tr. Neural Networks*, 9, 4, 662–668, 1998.