Stochastics of on-line backpropagation[1]

Tom Heskes

Beckman Institute and Department of Physics,
University of Illinois at Urbana-Champaign,
405 North Mathews Avenue, Urbana, Illinois 61801, U.S.A.

Abstract

We study on-line backpropagation and show that the existing theoretical descriptions are strictly valid only on relatively short time scales or in the vicinity of (local) minima of the backpropagation error potential. *Qualitative* global features (e.g., why is it much easier to escape from local minima than from global minima) may also be explained by these local descriptions, but the current approaches cannot give accurate *quantitative* predictions of global properties (e.g., how long does it take to reach the global minimum starting from a local minimum).

1 Introduction

On-line backpropagation stands for backpropagation where at each learning step one of the training patterns is drawn at random from the training set and presented to the network. This is in contrast with batch-mode backpropagation where a weight change takes place on account of the whole training set. The random pattern presentation in on-line backpropagation leads to a special kind of noise, which helps to escape from local minima in the error function.

In the literature, several suggestions have been made to describe on-line backpropagation as a deterministic process with superimposed noise [1, 2, 3]. In this paper we will study the validity of this approach. We will discuss the usefulness of this description to explain global properties of on-line backpropagation, such as stationary solutions and mean first passage times.

2 Expansions of the master equation

At each learning step, a training pattern x^μ, with x^μ denoting the combination of input vector and desired output vector, is drawn at random from the total training set and presented to the network. The weight change follows

$$\Delta w \;=\; \eta\, f(w, x^\mu)\,, \tag{1}$$

with w the weight vector, which includes the strength of all synapses and thresholds, η the learning parameter, and $f(.,.)$ the backpropagation learning rule. In the following we will use one-dimensional notation for simplicity.

The learning process (1) can be described by the master equation [4, 1, 5]

$$\frac{\partial}{\partial t} P(w,t) \;+\; P(w,t) \;=\; \int dw'\, T(w|w')\, P(w',t)\,, \tag{2}$$

with the transition probability to go from an old state w' to a new one w,

$$T(w|w') \;=\; \frac{1}{p} \sum_{\mu=1}^{p} \delta(w - w' - \eta\, f(w', x^\mu))\,.$$

With a smart choice of the time-intervals between subsequent adaptations, the master equation (2) *exactly* describes the learning process (1) [5]. In general, this master equation cannot be

[1]Proceedings of the European Symposium on Artificial Neural Networks '94, pages 223–228

solved analytically. An option is to look for approximations valid for small learning parameters η.

The first step in most approximation schemes is to write the master equation in the form of its completely equivalent Kramers-Moyal expansion (see e.g. [6])

$$\frac{\partial}{\partial t}P(w,t) \;=\; \sum_{n=1}^{\infty} \frac{(-\eta)^n}{n!} \frac{\partial^n}{\partial w^n}[a_n(w)P(w,t)] \qquad \text{with} \qquad a_n(w) \;\equiv\; \frac{1}{p}\sum_{\mu=1}^{p} f^n(w,x^\mu)\,. \tag{3}$$

The Fokker-Planck equation uses only the drift $a_1(w)$ and the diffusion $a_2(w)$:

$$\frac{\partial}{\partial t}P(w,t) \;=\; -\eta\frac{\partial}{\partial w}[a_1(w)P(w,t)] \;+\; \frac{\eta^2}{2}\frac{\partial^2}{\partial w^2}[a_2(w)P(w,t)]\,. \tag{4}$$

It is often used to study on-line backpropagation [1, 2]. We will show that this approach is strictly valid only on relatively short time scales and/or to study local properties of on-line backpropagation.

A proper expansion is Van Kampen's "small-fluctuations expansion" [6]. It is based on the Ansatz that the evolution of w is given by a deterministic part $\phi(t)$ and superimposed noise with standard deviation of order $\sqrt{\eta}$:

$$w \;=\; \phi(t) \;+\; \sqrt{\eta}\,\xi\,. \tag{5}$$

Substitution of this Ansatz into the Kramers-Moyal expansion (3) and collecting all terms up to order η leads to a set of three differential equations [6, 5]:

$$\left\{ \begin{array}{rcl} \frac{1}{\eta}\frac{d}{dt}\phi(t) & = & a_1(\phi(t)) \\ \frac{1}{\eta}\frac{d}{dt}\left\langle \xi \right\rangle_t & = & a_1'(\phi(t))\left\langle \xi \right\rangle_t \\ \frac{1}{\eta}\frac{d}{dt}\left\langle \xi^2 \right\rangle_t & = & 2a_1'(\phi(t))\left\langle \xi^2 \right\rangle_t + a_2(\phi(t)) \end{array} \right. \tag{6}$$

where $\langle . \rangle_t$ stands for the ensemble average over $P(w,t)$ and $'$ denotes differentiation of a function with respect to its argument.

From the set of equations (6) we conclude that Van Kampen's Ansatz is valid if the derivative of the average learning rule $a_1'(\phi)$ is negative or on time scales $\leq \mathcal{O}(1/\eta)$. The generalization to higher dimensions is that the Hessian matrix $H(w)$, containing the second derivatives of the error potential $E(w)$, must be positive definite. Each of these so-called attraction regions with positive definite Hessian $H(\mathbf{w})$ contains one (local) minimum of the error potential $E(w)$. So, the small-fluctuations Ansatz (5) is valid in these attraction regions, but [on time scales $> \mathcal{O}(1/\eta)$] not outside of these attraction regions.

The small noise approximation (6) can also be obtained by substituting the Ansatz (5) into the Fokker-Planck equation (4), i.e., all terms $\geq \mathcal{O}(\eta^3)$ in the Kramers-Moyal expansion (3) vanish for small η. In this sense the Fokker-Planck equation (4) is equivalent to Van Kampen's equations (6). However, any (nonlinear) features that arise from using the Fokker-Planck equation beyond this small-noise approximation are spurious and cannot be taken seriously [6].

3 Qualitative explanation of global features

An important difference between on-line learning and simulated annealing or Langevin equations (see e.g. [7]), is that the noise in on-line learning processes is intrinsic and inhomogeneous, i.e., depends on the weight vector $\mathbf{w}$, whereas the noise in simulated annealing and Langevin equations is artificial and homogeneous, i.e., constant over the whole state space. If we define

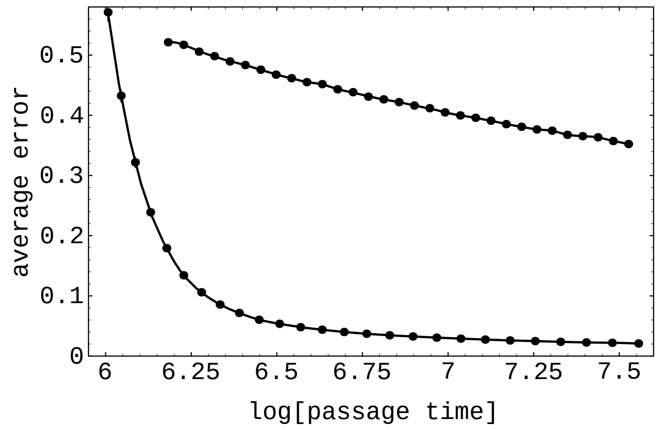

Figure 1: Average error (rescaled such that $E_{\mathrm{global}} \equiv 0$ and $E_{\mathrm{local}} \equiv 1$) versus logarithm of mean first passage time for on-line (lower line) and Langevin learning (upper line). On-line learning clearly yields a better performance.

temperature as the average increase in error potential due to the local fluctuations at a particular minimum $\mathbf{w}^*$, we obtain [5]

$$T(\mathbf{w}^*) \equiv \langle E(\mathbf{w}) - E(\mathbf{w}^*) \rangle_{I_{\mathbf{w}^*}} \approx \eta \, \mathrm{Tr}\left[a_2(\mathbf{w}^*) \right] ,$$

where the average is over the ensemble of all networks in the attraction region $I_{\mathbf{w}^*}$ of minimum $\mathbf{w}^*$. So, the local temperature is proportional to the learning parameter and to the local diffusion at the particular minimum.

As an example, let us consider the XOR problem with an additional pattern (see Appendix), which is known to have deep local minima [8]. At the global minima, the trace of the diffusion matrix is small, since all patterns are classified correctly. At the local minima, one of the patterns is misclassified, which leads to a much higher local temperature. Simulated annealing and Langevin equations have a "global temperature," i.e., the same local temperature at all minima. This difference suggests that the intrinsic noise of on-line learning makes it relatively more difficult to escape from lower lying minima and is therefore favorable.

To test the validity of this statement, we will compare on-line learning with Langevin learning, a discretized version of the Langevin equation [6], where Gaussian white noise $\boldsymbol{\xi}$ is added to the gradient of the *total* error:

$$\Delta \mathbf{w} = -\Delta t \, \boldsymbol{\nabla} E(\mathbf{w}) + \sqrt{2T \, \Delta t} \, \boldsymbol{\xi} \, . \tag{7}$$

All 500 learning networks start at a local minimum where four out of five patterns are classified correctly. For different values of the learning parameter η and temperature T (we keep $\Delta t = 1$ and do *not* take into account that Langevin learning is about $p = 5$ times slower), we calculate the mean first passage time τ into a region of weight space where all five patterns are classified correctly, i.e., where the output has the correct sign for all five patterns, and the average error for $10\tau \leq t \leq 20\tau$. For a faster escape out of the local minimum, one would like to choose a large learning parameter (high temperature), for a low asymptotic error a small learning parameter (low temperature). As can be seen from figure 1, on-line backpropagation is clearly better in dealing with this conflict: a lower (average) error can be reached in a shorter time.

4 Quantitative prediction of global features?

In the previous section we used local expansions of the master equation for a *qualitative* explanation of why on-line backpropagation might be a useful global minimization strategy if compared

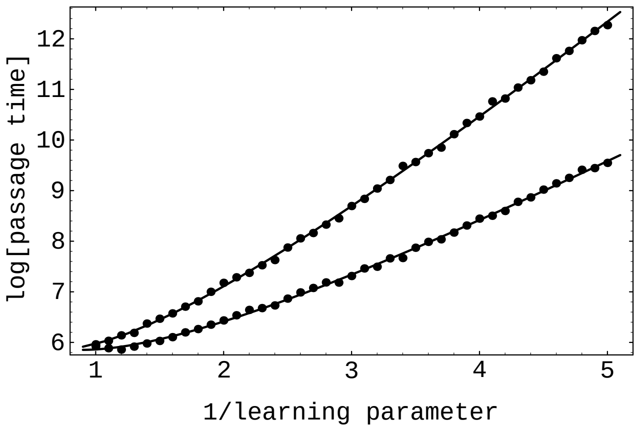

Figure 2: Logarithm of mean first passage times, starting from a local minimum into a region where all five patterns are classified correctly, versus one over the learning parameter. On-line (upper line) and Langevin learning (lower line). Theory based on only drift and diffusion cannot predict these curves.

with adding homogeneous noise. Now we would like to investigate whether we can apply any of these approaches to make *quantitative* statements about global properties. Therefore we suggest to again compare on-line learning with Langevin learning (7), but now with $\Delta t = \eta$ and *inhomogeneous* noise $\boldsymbol{\xi}$, chosen such that the drift vectors and diffusion matrices for both learning procedures are *exactly equal*. We start with an ensemble of 500 networks at the local minimum and calculate the mean first passage times into the region of weight space where all five patterns are classified correctly. Existing approaches (see e.g. [1, 5, 2, 3]) try to compute global properties of on-line learning using only the drift vector and the diffusion matrix, i.e., cannot make a difference between both learning procedures.

As can be seen from the results in figure 2, where the logarithm of the mean first passage times is plotted as a function of the reciprocal value of the learning parameter, the existing approaches are not sophisticated enough. The graphs of on-line and Langevin learning do not have the same slope (as suggested in [5]), nor do they converge in the limit of small learning parameters η (as suggested in [1, 2, 3]). So, although Fokker-Planck approaches, only based on drift and diffusion, can be used for a quantitative analysis of local properties of backpropagation (section 2) and possibly also for a qualitative explanation of global features (section 3), an application of these approaches to calculate global properties of on-line backpropagation is doomed to fail (section 4).

Appendix

The network, shown in figure 3(a), has $N = 9$ adaptive elements, combined in the weight vector $\mathbf{w} = (w_{10}, w_{11}, w_{12}, w_{20}, w_{21}, w_{22}, w_{30}, w_{31}, w_{32})^T$, two variable inputs, x_1 and x_2, and fixed inputs $x_0 = y_0 = -1$ to incorporate thresholds. Outputs are given by ($z_j = x_j$ for the hidden units and $z_j = y_j$ for the output unit)

$$y_i \;=\; \tanh\left[\sum_{j=0}^{2} w_{ij} z_j\right] .$$

To prevent the explosion of the weights, we add a so-called bias (with $\lambda = 0.01$ and $\alpha = 0.1$) to the squared backpropagation error:

$$E(\mathbf{w}) \;=\; \frac{1}{2p}\sum_{\mu=1}^{p}\left[y_3(\mathbf{w}, x_1^\mu, x_2^\mu) - x_3^\mu\right]^2 \;+\; \frac{\lambda}{4}\sum_{i=0}^{2}\sum_{j=0}^{2}\left[w_{ij}^2 - \alpha\right]^2 . \tag{8}$$

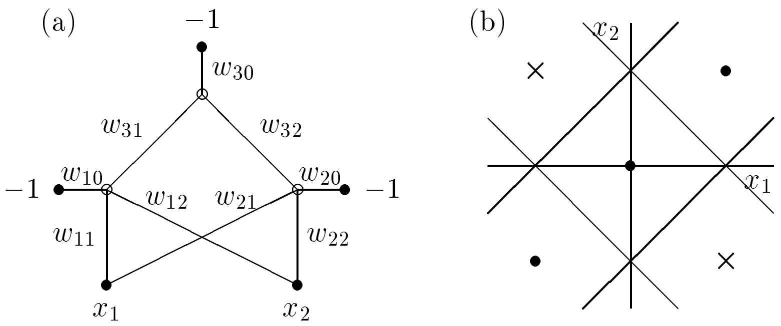

Figure 3: (a) Network structure. (b) XOR problem with one additional pattern.

After [8], we choose the set of $p = 5$ training patterns sketched in figure 3(b). Circles indicate negative desired output $x_3^\mu = -0.8$, crosses positive output $x_3^\mu = 0.8$. It is the usual XOR truth table with an additional pattern at the origin. Now the error potential (8) has not only global minima, but also deep local minima. The thick lines in figure 3(b) show the separation lines of the hidden units that lead to the optimal solution (all five patterns correctly classified), the thin lines those corresponding to the local minima (one pattern misclassified).

Acknowledgments

This work was supported by a grant from the National Institutes of Health (P41RR05969) to Klaus Schulten. I would like to thank Qing Sheng, Bert Kappen, and Klaus Schulten for valuable comments and discussions.

References

[1] G. Radons, H. Schuster, and D. Werner. Fokker-Planck description of learning in backpropagation networks. In *International Neural Network Conference 90 Paris*, pages 993–996, Dordrecht, 1990. Kluwer Academic.

[2] T. Leen and G. Orr. Weight-space probability densities and convergence times for stochastic learning. In *International Joint Conference on Neural Networks*. IEEE, 1992.

[3] L. Hansen, R. Pathria, and P. Salamon. Stochastic dynamics of supervised learning. *Journal of Physics A*, 26:63–71, 1993.

[4] H. Ritter and K. Schulten. Convergence properties of Kohonen's topology conserving maps: fluctuations, stability, and dimension selection. *Biological Cybernetics*, 60:59–71, 1988.

[5] T. Heskes and B. Kappen. On-line learning processes in artificial neural networks. In J. Taylor, editor, *Mathematical Foundations of Neural Networks*, pages 199–233. Elsevier, Amsterdam, 1993.

[6] N. van Kampen. *Stochastic Processes in Physics and Chemistry*. North-Holland, Amsterdam, 1992.

[7] T. Guillerm and N. Cotter. A diffusion process for global optimization in neural networks. In *IJCNN*, volume 1, pages 335–340, New York, 1991. IEEE.

[8] M. Gori and A. Tesi. On the problem of local minima in backpropagation. *IEEE Transactions on PAMI*, 14:76–86, 1992.