Error potentials for self-organization

Tom M. Heskes and Bert Kappen
Department of Medical Physics and Biophysics,
University of Nijmegen, Geert Grooteplein 21,
6525 EZ Nijmegen, The Netherlands,
e-mail: tom@mbfys.kun.nl

***Abstract*— We give an error potential for self-organizing learning rules. The gradient of this error potential leads to the well-known learning rule of Kohonen, except for the determination of the "winning" unit. The existence of an error potential facilitates a global description of the learning process. A one-dimensional topological map is treated as an example.**

I. Introduction

Sensory maps are a crucial first step in the information processing of the brain. The external information is represented in a orderly, topology-preserving manner, i.e., neighbouring units in the sensory map represent similar inputs. The formation of these maps is a process of self-organization for which several learning paradigms have been suggested [1, 2]. The proposal of Kohonen [3] does not aim at the modelling of all biological details, but tries to capture the most important features of self-organizing processes. Basically, this algorithm works as follows. Given a certain input vector from the environment, the unit with the smallest Euclidian distance to this vector is called the "winner." The weight vector of this unit and, to some extent, its neighbouring units, are moved towards the input vector. After this learning step, another input vector is drawn at random from the environment, etcetera. The properties of this learning procedure are studied in great detail [4, 5, 6]. Recently, a lot of effort is devoted to the search for an energy function that is minimized by the learning rule [7, 8, 9]. The existence of such an energy function or error potential facilitates a description of the global performance of the learning procedure [10]. The best possible state is the global minimum of the error potential, undesired (meta)stable configurations are simply local minima. In self-organizing learning rules possible local minima are topological defects like twists and kinks [11].

The definitions in Sec. II will be used in Sec. III to investigate whether it is possible to find an error potential for the original Kohonen learning rule. In Sec. IV we go the other way around. We start with a well-defined error potential and try to derive an on-line learning procedure. An example, kinks in a one-dimensional map, will be treated in Sec. V. Implications and further lines of research are discussed in Sec. VI.

II. The learning procedure

The network consists of n units labeled $1, \ldots, i, \ldots, n$. To each unit we ascribe an m-dimensional weight vector $\vec{w}_i$. The combination of all weight vectors is the N-dimensional state vector $\mathbf{W} = (\mathbf{w}_1, \ldots, \mathbf{w}_i, \ldots, \mathbf{w}_n)^T$, so $N = n \times m$. An on-line learning procedure, a repetition of the following three steps, takes care of the adaptation of this state vector.

1. *Pick an input vector from the environment.*

   This is why the learning procedure is called on-line: the network state is adjusted at each presentation of a training pattern. The environment Ω is a set of m-dimensional input vectors $\vec{x}$ with probability density function $\rho(\vec{x})$. The average of an arbitrary function $q(\vec{x})$ with respect to this environment Ω is written

   $$\langle q(\vec{x}) \rangle_\Omega \stackrel{\text{def}}{=} \int d^m x \, \rho(\vec{x}) \, q(\vec{x}) \, .$$

2. *Determine the winning unit.*

   We ascribe a "winning error" $g_i(\mathbf{W}, \vec{x})$ to each unit i. The unit with the smallest winning error is the "winner." For Kohonen's learning rule the winning error is the Euclidian distance between the input vector and the weight vector of the unit, i.e.,

   $$\textit{Kohonen:} \quad g_i(\mathbf{W}, \vec{x}) = \frac{1}{2} \, \|\vec{x} - \vec{w}_i\|^2 \, . \qquad (1)$$

This work was partly supported by the Dutch Foundation for Neural Networks and the Canon Foundation in Europe.

0-7803-0999-5/93/$03.00 ©1993 IEEE

For a nice mathematical description of this determination of the winner, we define, for any arbitrary vector $\mathbf{q}$, the following average with respect to the winning error (for notational convenience we will drop the arguments $\mathbf{W}$ and $\vec{x}$)

$$\langle\langle q\rangle\rangle_\beta \stackrel{\text{def}}{=} \frac{\sum_{i=1}^{n} q_i \exp[-\beta g_i]}{\sum_{i=1}^{n} \exp[-\beta g_i]} . \qquad (2)$$

In the limit $\beta \to \infty$ only the components with the smallest winning error survive, i.e.,

$$\langle\langle q\rangle\rangle_\infty \stackrel{\text{def}}{=} \lim_{\beta\to\infty} \langle\langle q\rangle\rangle_\beta = \frac{\sum_{i=1}^{n} q_i \delta_{g_i,g_{\min}}}{\sum_{i=1}^{n} \delta_{g_i,g_{\min}}}$$

$$\text{with} \quad g_{\min} \stackrel{\text{def}}{=} \min_i g_i . \qquad (3)$$

So, $\langle\langle\ldots\rangle\rangle_\infty$ stands for a "winner-take-all" mechanism with respect to the winning errors $g_i(\mathbf{W},\vec{x})$. This special average $\langle\langle q\rangle\rangle_\infty$ depends on $\mathbf{W}$ and $\vec{x}$ not only through $q_i(\mathbf{W},\vec{x})$, but also through the winning errors $g_i(\mathbf{W},\vec{x})$, which is important for differentation and integration. That part of the environment for which unit i is the winner is called the "receptive field" of unit i.

3. *Adapt the network state such that the local error of the winning unit becomes smaller.*

   Apart from the winning error, we also ascribe a "local error" $e_i(\mathbf{w},\vec{x})$ to each unit i. The learning rule is written formally

   $$\Delta\mathbf{W} = \eta\,\mathbf{F}(\mathbf{W},\vec{x}) = -\eta\,\langle\langle\nabla e(\mathbf{W},\vec{x})\rangle\rangle_\infty , \qquad (4)$$

   where η is the learning parameter and ∇ denotes the gradient with respect to the network state $\mathbf{W}$. This fairly cryptical and unusual definition of the learning rule becomes clearer if we write Kohonen's learning rule in these terms. We choose the local error

   $$\textit{Kohonen:} \quad e_k(\mathbf{W},\vec{x}) = \frac{1}{2}\sum_{i=1}^{n} h_{ki}\,\|\vec{x}-\vec{w}_i\|^2 . \qquad (5)$$

   The matrix with components h_{ij} defines the topological structure between the units. Usually, h_{ij} is a decreasing function of the distance between the units i and j in the topological map. The learning rule (4) now yields

   $$\textit{Kohonen:} \quad \Delta\vec{w}_i = \eta\,h_{\kappa(\mathbf{W},\vec{x})i}\,(\vec{x}-\vec{w}_i) ,$$

   with $\kappa(\mathbf{W},\vec{x})$ the winner, the unit with the smallest Euclidian distance between its weight vector and the input vector.

There are two closely related questions we would like to answer. The first question will be discussed in Sec. III, the second one in Sec. IV.

1. Is it possible to write the average learning rule (4) as the gradient of some global error potential, i.e., can we find a function $E(\mathbf{W})$ such that

   $$\langle\langle\langle\nabla e(\mathbf{W},\vec{x})\rangle\rangle_\infty\rangle_\Omega = \nabla E(\mathbf{W})\,?$$

2. Does the gradient of a global error potential of the form $E(\mathbf{W}) = \langle\langle\langle e(\mathbf{W},\vec{x})\rangle\rangle_\infty\rangle_\Omega$ lead to a self-organizing on-line learning rule, i.e., can we find learning rules $\mathbf{D}_i(\mathbf{W},\vec{x})$ such that

   $$\nabla\,\langle\langle\langle e(\mathbf{W},\vec{x})\rangle\rangle_\infty\rangle_\Omega = \langle\langle\langle\mathbf{D}(\mathbf{W},\vec{x})\rangle\rangle_\infty\rangle_\Omega\,?$$

III. Does there exist an error potential for Kohonen learning ?

To investigate whether there exists an error potential for Kohonen's learning rule or, more general, an error potential for learning rules of the form (4), we have to calculate the derivative of the average learning rule

$$\mathbf{F}(\mathbf{W}) \stackrel{\text{def}}{=} \langle\mathbf{F}(\mathbf{W},\vec{x})\rangle_\Omega .$$

The difficult part in calculating this derivative is the contribution of the winner-take-all mechanism. That is why we introduced the winner-take-all mechanism as a special limit of a weighted average with respect to the winning error. Using shorthand notation $\partial_\mu F_\nu \stackrel{\text{def}}{=} \partial F_\nu(\mathbf{W},\vec{x})/\partial W_\mu$, etcetera, we have

$$\partial_\mu F_\nu = -\partial_\mu \lim_{\beta\to\infty}\left\langle\langle\langle\partial_\nu e\rangle\rangle_\beta\right\rangle_\Omega = -\lim_{\beta\to\infty}\left\langle\partial_\mu\,\langle\langle\partial_\nu e\rangle\rangle_\beta\right\rangle_\Omega .$$

Here it is allowed to interchange the gradient and the limit, but *not* to interchange the gradient with respect to $\mathbf{W}$ and the average with respect to the winning errors $g_i(\mathbf{W},\vec{x})$, which is a function of $\mathbf{W}$. The gradient can be calculated using the definition (2). We obtain

$$\partial_\mu F_\nu = -\left\langle\left\langle\left\langle\partial^2_{\mu\nu} e\right\rangle\right\rangle_\infty\right\rangle_\Omega +$$

$$\lim_{\beta\to\infty}\beta\left\langle\left\langle\left\langle\left(\partial_\nu e - \langle\langle\partial_\nu e\rangle\rangle_\beta\right)\left(\partial_\mu g - \langle\langle\partial_\mu g\rangle\rangle_\beta\right)\right\rangle\right\rangle_\beta\right\rangle_\Omega . \qquad (6)$$

The average learning rule is the gradient of some error potential if and only if $\partial_\mu F_\nu = \partial_\nu F_\mu$. Obviously, the first term in (6) is symmetric. The second term is symmetric if and only if the local error is a monotonic function of the winning error, i.e., *if and only if the local error also determines the winning unit.* In fact, we might as well say that the local error must be equal to the winning error, since any monotonically increasing function of the winning error is totally equivalent to the winning error itself in the limit $\beta \to \infty$. So, only if the local and winning error are equal, the learning procedure (4) can be interpreted as a (stochastic) way to minimize a well-defined error criterion. The Kohonen choices (1) and (5) do not satisfy this requirement, except in the limit of no lateral interaction $h_{ij} = \delta_{ij}$.

IV. Does an error potential lead to an on-line learning rule ?

Here we do not ask whether it is possible to find an error potential corresponding to an existing learning rule, but we start with a well-defined error potential and try to derive a learning rule from it. An obvious choice for the error potential is

$$E(\mathbf{W}) \stackrel{\text{def}}{=} \langle\langle\langle e(\mathbf{W},\vec{x})\rangle\rangle_\infty\rangle_\Omega \,, \tag{7}$$

i.e., the local error of the winning unit, averaged over the whole environment. A similar error potential is suggested in [12]. It can be interpreted as a transmission error between neural layers [13]. The derivative of the error potential (7) obeys

$$\nabla E = \langle\langle\langle \nabla e\rangle\rangle_\infty\rangle_\Omega - \lim_{\beta\to\infty} \beta \left\langle \left\langle \left\langle \left(e - \langle\langle e\rangle\rangle_\beta\right)\left(\nabla g - \langle\langle \nabla g\rangle\rangle_\beta\right)\right\rangle\right\rangle_\beta \right\rangle_\Omega . \tag{8}$$

The first term is exactly the learning rule (4) averaged over the environment Ω. It is difficult (see [8]), to interpret the second term as the average of an on-line learning rule. Therefore we would like to get rid of it. For some combinations of $\mathbf{W}$ and $\vec{x}$ there is just one winner, say k. In this case

$$\lim_{\beta\to\infty} \left\langle \left\langle \left(e - \langle\langle e\rangle\rangle_\beta\right)\left(\nabla g - \langle\langle \nabla g\rangle\rangle_\beta\right)\right\rangle\right\rangle_\beta = (e_k - e_k)(\nabla g_k - \nabla g_k) = 0 .$$

However, there exist combinations of $\mathbf{W}$ and $\vec{x}$ for which there are two (or more) winning units, say k and l, with $g_k(\mathbf{W},\vec{x}) = g_l(\mathbf{W},\vec{x}) = g_{\min}(\mathbf{W},\vec{x})$, so, with $\vec{x}$ exactly on the boundary of the receptive fields of unit k and l. Then we have

$$\lim_{\beta\to\infty} \left\langle \left\langle \left(e - \langle\langle e\rangle\rangle_\beta\right)\left(\nabla g - \langle\langle \nabla g\rangle\rangle_\beta\right)\right\rangle\right\rangle_\beta = \frac{1}{4}(e_k - e_l)(\nabla g_k - \nabla g_l) .$$

This "boundary term" is zero if either $e_k = e_l$ or $\nabla g_k = \nabla g_l$. This second possibility is clearly not true for the Kohonen learning rule and is hard to satisfy in general. So, the only way to exclude boundary terms is to ensure that on these boundaries the local errors are equal. The conclusion is that the second term in (8) vanishes if and only if

$$g_i(\mathbf{W},\vec{x}) = g_j(\mathbf{W},\vec{x}) \;\Rightarrow\; e_i(\mathbf{W},\vec{x}) = e_j(\mathbf{W},\vec{x}) \quad \forall_{i,j,\mathbf{W},\vec{x}} .$$

We arrive at a similar conclusion as above: the gradient of an error potential of the form (7) leads to an on-line learning rule *if and only if the local error also determines the winning unit.*

In the meantime, we have proved that we may interchange taking the derivative and determining the winner, i.e., that

$$\langle\langle\langle \nabla e(\mathbf{W},\vec{x})\rangle\rangle_\infty\rangle_\Omega = \nabla \langle\langle\langle e(\mathbf{W},\vec{x})\rangle\rangle_\infty\rangle_\Omega ,$$

if the local error and the winning error are equal, i.e., if $e_i(\mathbf{W},\vec{x}) = g_i(\mathbf{W},\vec{x})$. With this particular choice, the on-line learning rule (4) performs stochastic gradient descent on the error potential (7). For the rest of the paper, we will choose the local error and the winning error equal to Kohonen's local error (5). The resulting learning rule is equal to Kohonen's learning rule except for the determination of the winning unit.

V. An example: kinks in one dimension

We consider a one-dimensional map consisting of three units. The weight vector is written $\mathbf{w} = (w_1, w_2, w_3)^T$. The input probability distribution obeys

$$\rho(x) = \theta(x)\,\theta(1-x) ,$$

i.e., x is drawn with equal probability from the interval $[0,1]$. The lateral interaction matrix h with components h_{ij} is defined

$$h = \frac{1}{1+\sigma}\begin{pmatrix} 1 & \sigma & 0 \\ \sigma & 1-\sigma & \sigma \\ 0 & \sigma & 1 \end{pmatrix} .$$

It is normalized such that $\sum_j h_{ij} = 1\ \forall_i$. σ gives the strength of the interaction between neighboring units in the map. $\sigma = 0$ means no lateral interaction.

Ordered configurations are called "lines." One of them, denoted by (123) since $w_1 < w_2 < w_3$, is drawn schematically in Fig. 1(a). The other one is (321), i.e., w_1 and w_3 are exchanged. There are four different disordered configurations called "kinks:" (132), (213), (231), and (312). The first one is sketched in Fig. 1(b).

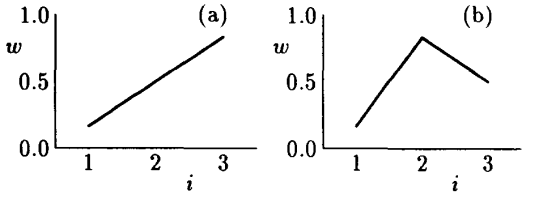

Figure 1: Configurations in a one-dimensional map. (a) Line. (b) Kink.

By numerical calculations, it can be proved that for $\sigma < \sigma^* \approx 0.082$ the error potential (7) has six minima: two lines are global minima, four kinks are local minima. At $\sigma = \sigma^*$ the local minima disappear and only two global minima remain.

We would like to picture how the error potential changes by going from the local to the global minimum. To get rid of two degrees of freedom, we bring in the following constraints. These are based on the fact that at a minimum always one of the weights is approximately equal to 1/6, the second to 1/2, and the third to 5/6. So, at the minima, the sum of the weights and the sum of the squared distances between the weights are approximately constant,

$$\begin{cases} w_1 + w_2 + w_3 \approx \frac{3}{2}\,, \\ (w_2 - w_1)^2 + (w_3 - w_2)^2 + (w_3 - w_1)^2 \approx \frac{2}{3}\,. \end{cases}$$

Combining these constraints ("$\approx$" is replaced with "="), $\mathbf{W}$ is totally parametrized by the angle ϕ,

$$\begin{cases} w_1(\phi) = \frac{1}{2} - \frac{2\sqrt{3}}{9}\cos\phi\,, \\ w_2(\phi) = \frac{1}{2} + \frac{\sqrt{3}}{9}\cos\phi - \frac{1}{3}\sin\phi\,, \\ w_3(\phi) = \frac{1}{2} + \frac{\sqrt{3}}{9}\cos\phi + \frac{1}{3}\sin\phi\,. \end{cases}$$

In terms of ϕ, the lines and the kinks are positioned as follows.

minimum	line	kink	kink	line	kink	kink
configuration	(123)	(213)	(231)	(321)	(312)	(132)
ϕ	$\pi/6$	$\pi/2$	$5\pi/6$	$7\pi/6$	$3\pi/2$	$11\pi/6$
phase	1/12	1/4	5/12	7/12	3/4	11/12

The error potential as a function of the "phase" $\phi/(2\pi)$ is plotted in Fig. 2 for $\sigma = 0$, 0.04, 0.08, and 0.12. At $\sigma = 0$ all minima are equally deep. For $\sigma > 0$, symmetry is broken: the kinks have a higher error potential than the lines. Eventually, the disordered local minima disappear.

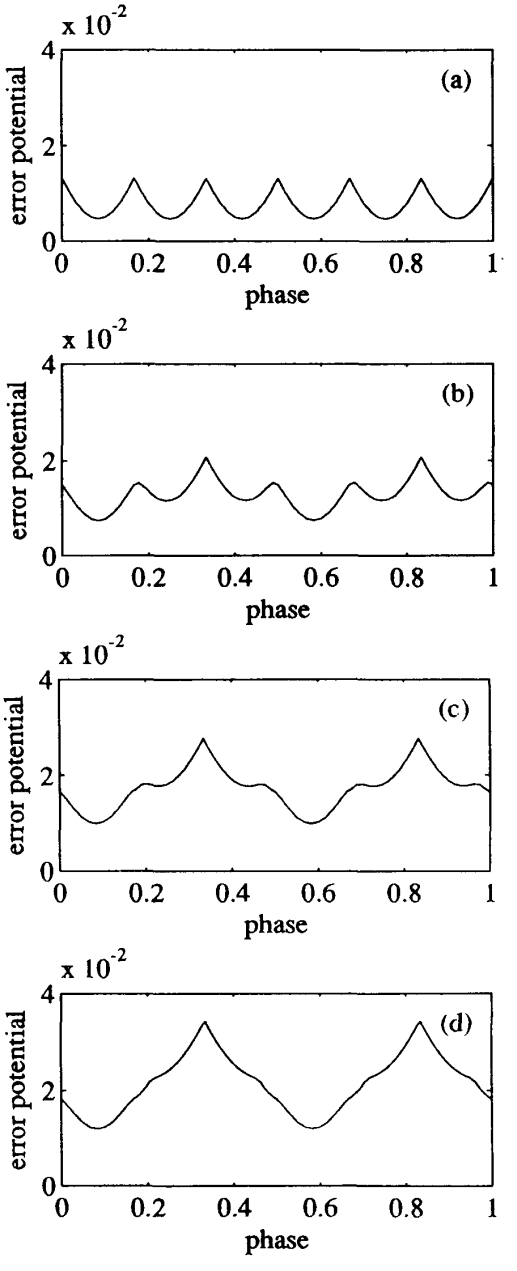

Figure 2: The error potential E as a function of the "phase" $\phi/(2\pi)$ for different values of the interaction strength σ. (a) $\sigma = 0$. (b) $\sigma = 0.04$. (c) $\sigma = 0.08$. (d) $\sigma = 0.12$.

VI. Discussion

In this paper an error potential and corresponding learning rule for the self-organization of topological maps are derived. The resulting learning rule is exactly equal to the one proposed by Kohonen, except for the determination of the winning unit. The disadvantage of our learning rule is that the determination of the winning unit is computationally more expensive. (The Euclidian distances, which require $n \times m$ multiplications, must be multiplied with the lateral interaction matrix h. This requires $n \times |h|$ extra multiplications, with $|h|$ the number of nonzero lateral connections for one unit.) The advantage is that we know exactly what error potential is minimized by the learning procedure. Furthermore, the existence of an error potential facilitates a global description of the learning process. The lower the error potential, the "better" the network state. Fixed points of the learning dynamics are minima of the error potential [14]. Local minima of the error potential correspond to topological defects, like kinks in one-dimensional maps or twists ("butterflies") in two-dimensional maps. Global minima are perfectly ordered configurations.

Results from a general study concerning learning in neural networks with local minima [10] can be applied to calculate transition times between different minima. In this context it means that we can calculate the transition times from topological defects to perfectly ordered configurations, i.e., the (average) time it takes to remove a kink in a one-dimensional map or to unfold a twist in a two-dimensional map [15]. Other research aims at the derivation of cooling schedules for the learning parameter that guarantee convergence to the global minimum of the error potential [16].

Acknowledgment

We would like to thank Dr. Andrzej Komoda for stimulating discussions.

References

[1] C. von der Malsburg. Self-organization of orientation sensitive cells in the striate cortex. *Kybernetik*, 14:85–100, 1973.

[2] A. Takeuchi and S. Amari. Formation of topographic maps and columnar microstructures. *Biological Cybernetics*, 35:63–72, 1979.

[3] T. Kohonen. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43:59–69, 1982.

[4] M. Cottrell and J. Fort. A stochastic model of retinotopy: a self-organizing process. *Biological Cybernetics*, 53:405–411, 1986.

[5] H. Ritter and K. Schulten. On the stationary state of Kohonen's self-organizing sensory mapping. *Biological Cybernetics*, 54:99–106, 1986.

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

[7] V. Tolat. An analysis of Kohonen's self-organizing maps using a system of energy functions. *Biological Cybernetics*, 64:155–164, 1990.

[8] T. Kohonen. Self-organizing maps: optimization approaches. In T. Kohonen, K. Mäkisara, O. Simula, and J. Kanga, editors, *Artificial Neural Networks*, pages 981–990, Amsterdam, 1991. North-Holland.

[9] E. Erwin, K. Obermayer, and K. Schulten. Self-organizing maps: ordering, convergence properties and energy functions. *Biological Cybernetics*, 67:47–55, 1992.

[10] T. Heskes, E. Slijpen, and B. Kappen. Learning in neural networks with local minima. *Physical Review A*, 46:5221–5231, 1992.

[11] T. Geszti. *Physical models of neural networks*. World Scientific, Singapore, 1990.

[12] S. Luttrell. Self-organisation: A derivation from first principles of a class of learning algorithms. In *International Joint Conference on Neural Networks*, volume 2, pages 495–498. IEEE Computer Society Press, 1989.

[13] H. Ritter, K. Obermayer, K. Schulten, and J. Rubner. Self-organizing maps and adaptive filters. In E. Domany, J. van Hemmen, and K. Schulten, editors, *Models of neural networks*, pages 281–306, Berlin, 1991. Springer.

[14] T. Heskes and B. Kappen. Learning processes in neural networks. *Physical Review A*, 44:2718–2726, 1991.

[15] T. Heskes. Transition times in self-organizing maps. *Submitted to Biological Cybernetics*, 1992.

[16] T. Heskes, E. Slijpen, and B. Kappen. Cooling schedules for learning in neural networks. *Submitted to Physical Review E*, 1992.