ESANN'2003 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 23-25 April 2003, d-side publi., ISBN 2-930307-03-X, pp. 93-98

Magnification Control in Winner Relaxing Neural Gas

Jens Christian Claussen[1] and Thomas Villmann[2]

[1]Christian-Albrechts-Universität zu Kiel, Institut für Theoretische Physik und Astrophysik, D–24098 Kiel, Leibniztr.15, Germany*
[2]Universität Leipzig, Klinik für Psychotherapie
D–04107 Leipzig, Karl–Tauchnitz–Str.25, Germany*

Abstract. We transfer the idea of winner relaxing learning from the self-organizing map to the neural gas to enable magnification control *independently* of the shape of the data distribution.

1. Introduction

Neural maps are a widely ranged type of neural vector quantizers which are commonly used e.g. in data visualization, feature extraction, principle component analysis, image processing, and classification tasks. A well studied approach is the Neural Gas Network (NG) [8]. An important advantage of the NG is the adaptation dynamics, which minimizes a potential, in contrast to the self-organizing map (SOM) [7] frequently used in vector quantization problems.

In the present paper we consider a new control scheme for the *magnification* of the map. Controlling the magnification factor is relevant for many applications in control theory or robotics, were (neural) vector quantizers are often used to determine the actual state of the system in a first step, which is an objective of the control task [9, 10].

The NG maps data vectors $\mathbf{v}$ from a (possibly high-dimensional) data manifold $\mathcal{D} \subseteq \mathbb{R}^d$ onto a set A of neurons i. This is formally written as $\Psi_{\mathcal{D}\to A} : \mathcal{D} \to A$. Each neuron i is associated with a pointer $\mathbf{w}_i \in \mathbb{R}^d$ all of which establish the set $\mathbf{W} = \{\mathbf{w}_i\}_{i\in A}$. The mapping description is a winner take all rule, i.e. a stimulus vector $\mathbf{v} \in \mathcal{D}$ is mapped onto the neuron $s \in A$ the pointer $\mathbf{w}_s$ of which is closest to the actually presented stimulus vector $\mathbf{v}$,

$$\Psi_{\mathcal{D}\to A} : \mathbf{v} \mapsto s(\mathbf{v}) = \underset{i\in A}{\operatorname{argmin}} \|\mathbf{v} - \mathbf{w}_i\| . \tag{1}$$

The neuron s is called *winner neuron*.

During the adaptation process a sequence of data points $\mathbf{v} \in \mathcal{D}$ is presented to the map with respect to the stimuli distribution $P(\mathcal{D})$. Each time the currently most proximate neuron s according to (1) is determined, and the pointer $\mathbf{w}_s$ as well as all pointers $\mathbf{w}_i$ of neurons in the neighborhood of s are shifted towards $\mathbf{v}$, according to

$$\triangle \mathbf{w}_i = \epsilon h_\lambda (i, \mathbf{v}, \mathbf{W}) (\mathbf{v} - \mathbf{w}_i) . \tag{2}$$

*email: claussen@theo-physik.uni-kiel.de and villmann@informatik.uni-leipzig.de

ESANN'2003 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 23-25 April 2003, d-side publi., ISBN 2-930307-03-X, pp. 93-98

The property of "being in the neighborhood of s" is represented by a neighborhood function $h_\lambda(i, \mathbf{v}, \mathbf{W})$. The neighborhood function is defined as

$$h_\lambda(i, \mathbf{v}, \mathbf{W}) = \exp\left(-\frac{k_i(\mathbf{v}, \mathbf{W})}{\lambda}\right), \tag{3}$$

where $k_i(\mathbf{v}, \mathbf{W})$ yields the number of pointers $\mathbf{w}_j$ for which the relation $\|\mathbf{v} - \mathbf{w}_j\| \leq \|\mathbf{v} - \mathbf{w}_i\|$ is valid [8]. In particular we have $h_\lambda(s, \mathbf{v}, \mathbf{W}) = 1.0$. We remark that in contrast to the SOM the neighborhood function is evaluated in the input space. Moreover, the adaptation rule for the weight vectors in average follows a potential dynamics [8].

The *magnification* of the trained map reflects the relation between the data density $P(\mathcal{D})$ and the density ρ of the weight vectors. For the NG the relation

$$P(\mathcal{D}) \propto \rho(\mathbf{w})^{\alpha_{\mathsf{NG}}} \tag{4}$$

with $\alpha_{\mathsf{NG}} = d/(d+2)$ has been derived [8]. The exponent α_{NG} is called *magnification factor* and depends on the intrinsic dimensionality of the data. However, the information transfer, in general, is not independent of the magnification of the map [11]. It is known that for a vector quantizer (or a neural map in our context) with optimal information transfer the relation $\alpha = 1$ holds. On the other hand, a vector quantizer which minimizes the mean distortion error $E_\gamma = \int_{\mathcal{D}} \|\mathbf{w}_s - \mathbf{v}\|^\gamma P(\mathbf{v})\, d\mathbf{v}$ need a magnification factor of $\alpha = d/(d+\gamma)$ with $\mathbf{v} \in \mathcal{D} \subseteq \mathbb{R}^d$ [11]. Hence, the NG minimizes the usual E_2 distortion error.

We now address the question how to extend the Neural Gas to achieve an *a priori* chosen optimization goal, i.e. an *a priori* chosen magnification factor.

2. Controlling the magnification

For the NG a solution of the magnification control problem can be realized by introducing an adaptive local learning step size $\epsilon_{s(\mathbf{v})}$ [10] according to a similar approach introduced for SOM [1]. The new (localized) learning rule reads as

$$\triangle \mathbf{w}_i = \epsilon_{s(\mathbf{v})} h_\lambda(i, \mathbf{v}, \mathbf{W})(\mathbf{v} - \mathbf{w}_i) \tag{5}$$

with the local learning parameters $\epsilon_i = \epsilon(\mathbf{w}_i)$ depending on the stimulus density P at the position of the weight vectors $\mathbf{w}_i$ via $\langle \epsilon_i \rangle = \epsilon_0 P(\mathbf{w}_i)^m$. The brackets $\langle \ldots \rangle$ denote the average in time, and $s(\mathbf{v})$ is the best–matching neuron with respect to (1). This approach finally leads to the new magnification law

$$\alpha' = \alpha_{NG} \cdot (m+1) \tag{6}$$

In real applications one has to estimate the generally unknown data distribution P. This may lead to numerical instabilities of the control mechanism [5]. Recently, an new approach for magnification control of the SOM was introduced [3] which is a generalization of a winner relaxing modification [6]. This approach provides a control scheme which is *independent* of the shape of the data distribution [3]. We transfer these ideas to the NG as follows:

ESANN'2003 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 23-25 April 2003, d-side publi., ISBN 2-930307-03-X, pp. 93-98

Analog to the generalized winner relaxing SOM [3], we introduce the *winner relaxing NG* by defining the learning rule as

$$\triangle \mathbf{w}_i = \epsilon h_\lambda \left(i, \mathbf{v}, \mathbf{W}\right) \left(\mathbf{v} - \mathbf{w}_i\right) + R\left(\xi, \kappa\right) \tag{7}$$

with the winner relaxing term

$$R\left(\xi, \kappa\right) = \left(\xi + \kappa\right) \left(\mathbf{v} - \mathbf{w}_i\right) \delta_{is} - \kappa \delta_{is} \sum_j h_\lambda \left(j, \mathbf{v}, \mathbf{W}\right) \left(\mathbf{v} - \mathbf{w}_j\right) \tag{8}$$

depending on weighting parameters ξ and κ. The original motivation in [6] was to derive a learning rule from an average reconstuction error including the effect of shifting voronoi borders, which results in an additional term for the winner, i. e. (7) with $\xi = 0, \kappa = \frac{1}{2}$. However, one can utilize the winner relaxing term to influence the magnification by global choice of κ as in [3].

We now derive a relation between the densities ρ and P in analogy to [8] for the winner relaxing learning (7). The procedure is very similar as in [8, 10]. We can write the average change $\langle \triangle \mathbf{w}_i \rangle$ for the winner relaxing NG (7) as

$$\langle \triangle \mathbf{w}_i \rangle = I_1 + I_2 + I_3 \tag{9}$$

$$\text{with} \qquad I_1 = \int P\left(\mathbf{v}\right) h_\lambda \left(i, \mathbf{v}, \mathbf{W}\right) \left(\mathbf{v} - \mathbf{w}_i\right) d\mathbf{v}, \tag{10}$$

$$I_2 = \int P\left(\mathbf{v}\right) \left(\xi + \kappa\right) \cdot \delta_{is} \cdot \left(\mathbf{v} - \mathbf{w}_i\right) d\mathbf{v} \tag{11}$$

$$\text{and} \qquad I_3 = -\int P\left(\mathbf{v}\right) \delta_{is} \kappa \sum_j h_\lambda \left(j, \mathbf{v}, \mathbf{W}\right) \left(\mathbf{v} - \mathbf{w}_j\right) d\mathbf{v} \tag{12}$$

The integral I_1 is the usual one according to the NG dynamics, yielding [8]

$$I_1 = \epsilon' \left(\partial_{\mathbf{r}} P - P \cdot \frac{d+2}{d} \cdot \frac{\partial_{\mathbf{r}} \rho}{\rho} \right) \tag{13}$$

$$\text{with} \qquad \epsilon' = \frac{\epsilon_0}{\left(\tau_d \cdot \rho\right)^{\frac{2+d}{d}}} \int_{\mathcal{D}} h_\lambda \left(\mathbf{x}\right) \cdot \left\|\mathbf{x}\right\|^2 \, d\mathbf{x}. \tag{14}$$

We further assume a continuum approach such that for an input $\mathbf{v}$ we have $\mathbf{w_s} = \mathbf{v}$ [9]. Then I_2 vanishes in the continuum limes because the integration over δ_{is} only contributes for $\mathbf{w}_s$, but in this case $(\mathbf{v} - \mathbf{w}_s) = 0$ holds. While I_2 may contribute in higher orders, no influence on the entropy was found for the choice $\kappa + \xi = 0$ instead of $\xi = 0$. We now pay attention to the I_3-integral:

The continuum assumption made above allows a turn over from sum $\sum_j h_\lambda \left(\mathbf{w}_j, \mathbf{v}, \mathbf{W}\right) \left(\mathbf{v} - \mathbf{w}_j\right)$ to the integral $\int h_\lambda \left(\mathbf{w}, \mathbf{v}, \mathbf{W}\right) \left(\mathbf{v} - \mathbf{w}\right) d\mathbf{w}$ in (12). A procedure completely analog to the derivation of the NG magnification gives

$$I_3 = \int P\left(\mathbf{v}\right) \delta_{is} \kappa \left[\int h_\lambda \left(\mathbf{x}\right) \cdot \mathbf{r}\left(\mathbf{x}\right) \cdot \mathbf{J}\left(\mathbf{x}\right) d\mathbf{x} \right] d\mathbf{v} \tag{15}$$

with the new integration variable $\mathbf{x}\left(\mathbf{r}\right) = \hat{\mathbf{r}} \cdot \mathbf{k}_i \left(\mathbf{r}\right)^{\frac{1}{d}}$ and the $d \times d$–Jacobian–matrix $\mathbf{J}\left(\mathbf{x}\right) = \det(\partial r_k / \partial x_l)$ with $\mathbf{r} = \mathbf{v} - \mathbf{w}_i$.

ESANN'2003 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 23-25 April 2003, d-side publi., ISBN 2-930307-03-X, pp. 93-98

I_3 only contributes to $\langle \triangle \mathbf{w}_i \rangle$ for the winning weight (realized by δ_{is}), i.e., for $\mathbf{w}_i = \mathbf{w}_s$ which is equal to $\mathbf{v}$ according to the continuum approach. Hence, the integration over $\mathbf{v}$ yields

$$I_3 = \kappa P(\mathbf{w}_i) \cdot \int h_\lambda(\mathbf{x}) \cdot \mathbf{r}(\mathbf{x}) \cdot \mathbf{J}(\mathbf{x}) \, d\mathbf{x} \tag{16}$$

If $h_\lambda(\mathbf{k}_i(\mathbf{r}))$ rapidly decreases to zero with increasing $\mathbf{r}$, we can replace the quantities $\mathbf{r}(\mathbf{x})$, $\mathbf{J}(\mathbf{x})$ by the first terms of their respective Taylor expansions around the point $\mathbf{x} = 0$ neglecting higher derivatives. We obtain

$$\mathbf{x}(\mathbf{r}) = \mathbf{r}\left(\tau_d \rho(\mathbf{w}_i)\right)^{\frac{1}{d}} \left(1 + \frac{\mathbf{r} \cdot \partial_{\mathbf{r}} \rho(\mathbf{w}_i)}{d \cdot \rho(\mathbf{w}_i)} + \mathcal{O}\left(\mathbf{r}^2\right)\right) \tag{17}$$

which corresponds to

$$\mathbf{r}(\mathbf{x}) = \mathbf{x}\left(\tau_d \rho(\mathbf{w}_i)\right)^{-\frac{1}{d}} \left(1 - \left(\tau_d \rho(\mathbf{w}_i)\right)^{-\frac{1}{d}} \cdot \frac{\mathbf{x} \cdot \partial_{\mathbf{r}} \rho(\mathbf{w}_i)}{d \cdot \rho(\mathbf{w}_i)} + \mathcal{O}\left(\mathbf{x}^2\right)\right) \tag{18}$$

with $\tau_d = \pi^{\frac{d}{2}} / \Gamma\left(\frac{d}{2} + 1\right)$ as the volume of a d–dimensional unit sphere [8]. Further,

$$\begin{aligned} \mathbf{J}(\mathbf{x}) &= \left(\mathbf{J}(0) + x_k \tfrac{\partial \mathbf{J}}{\partial x_k} + \ldots\right) \\ &= (\tau_d \cdot \rho)^{-1} \left(1 - (\tau_d \cdot \rho)^{-\frac{1}{d}} \left(1 + \tfrac{1}{d}\right) \cdot \mathbf{x} \cdot \tfrac{\partial_{\mathbf{r}} \rho}{\rho}\right) + \mathcal{O}\left(x^2\right) \end{aligned} \tag{19}$$

and, hence, $\frac{\partial \mathbf{J}}{\partial \mathbf{x}}\big|_{\mathbf{x}=0} = -(\tau_d \cdot \rho)^{-\left(1+\frac{1}{d}\right)} \frac{\partial_{\mathbf{r}} \rho}{\rho}$. Therefore, the integral in equation (16) can be rewritten as

$$\begin{aligned} I_3 &= \epsilon' \kappa P (\tau_d \cdot \rho)^{-\frac{1}{d}} \int_{\mathcal{D}} h_\lambda(\mathbf{x}) \cdot \mathbf{x} \cdot \\ &\quad \cdot \left((\tau_d \cdot \rho)^{-1} - \left(1 + \tfrac{1}{d}\right)(\tau_d \cdot \rho)^{-\left(1+\frac{1}{d}\right)} \cdot \mathbf{x} \cdot \tfrac{\partial_{\mathbf{r}} \rho}{\rho} + \ldots\right) \\ &\quad \cdot \left(1 - (\tau_d \cdot \rho)^{-\frac{1}{d}} \cdot \mathbf{x} \cdot \tfrac{\partial_{\mathbf{r}} \rho}{d \cdot \rho} + \ldots\right) \, d\mathbf{x} \end{aligned} \tag{20}$$

The integral terms in (20) of odd order in $\mathbf{x}$ vanish because of the rotational symmetry of $h_\lambda(\mathbf{x})$. Then (16) yields, neglecting terms in higher order in $\mathbf{x}$,

$$I_3 = \epsilon' \kappa P \frac{d+2}{d} \frac{\partial_{\mathbf{r}} \rho}{\rho} \tag{21}$$

Taking together (13) and (21), the stationary solution of (7) is given by

$$\langle \triangle \mathbf{w}_i \rangle = 0 = \partial_{\mathbf{r}} P - P \cdot \frac{d+2}{d} \cdot \frac{\partial_{\mathbf{r}} \rho}{\rho} + P \kappa \frac{d+2}{d} \frac{\partial_{\mathbf{r}} \rho}{\rho} \tag{22}$$

This differential equation has the same form as the one for the usual Neural Gas (13). The magnification exponent of the Winner Relaxing Neural Gas is now given by

$$\alpha_{\mathsf{WRNG}} = \frac{1}{1-\kappa} \frac{d}{d+2} = \frac{1}{1-\kappa} \alpha_{\mathsf{NG}}. \tag{23}$$

ESANN'2003 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 23-25 April 2003, d-side publi., ISBN 2-930307-03-X, pp. 93-98

Now two direct observations can be made: Firstly, the magnification exponent, same as in [3], appears to be independent of the additional diagonal term (controlled by ξ) for the winner. Therefore $\xi = 0$ again is the usual setting. Further, by adjusting κ appropriately, the magnification exponent can be adjusted, e.g. to the most interesting case of maximal mutual information (where $\hat{\alpha} = 1$)

$$\kappa_{\mathsf{opt}} = \frac{2}{d+2}. \tag{24}$$

If the same stability borders $|\kappa| = 1$ of the Generalized Winner-Relaxing SOM also apply here, one can expect to increase the NG exponent by positive values of κ, or to lower the NG exponent by a factor $1/2$ for $\kappa = -1$. In contrast to the Winner Enhancing SOM, where the relaxing term has to be inverted ($\kappa < 0$) to increase the magnification exponent, for the neural gas positive values of κ are required to increase the magnification exponent. While the exponent still remains dependent on the dimension of the data, once this dimension is known (for instance estimated according to [2] or [4]), the parameter κ can be set *a priori* to obtain a neural gas of maximal mutual information. In this approach it is not necessary to keep track of the local reconstruction errors and firing rate for each neuron to adjust a local learning rate.

However, one has to be cautious when transferring the $\lambda \to 0$ result obtained above (which would require to increase the number of neurons as well) to a realistic situation where a decrease of λ with time will be limited to a final finite value to avoid the stability problems found in [5]. If the neighborhood length in SOM is kept small but fixed for the limit of fine discretization, the neighborhood function of the second but one winner will again be of order 1 (as for the winner). For the NG however the neighborhood is defined by the rank list. As the winner is not present in the $I_2 + I_3$ integral, all terms share the factor $\mathrm{e}^{-\lambda}$ by $h_\lambda(k) = \mathrm{e}^{-\lambda} h_\lambda(k-1)$ which indicates that in the discretized algorithm κ has to be rescaled by $\mathrm{e}^{+\lambda}$ to agree with the continuum theory.

3. Numerical results

A numerical study shows that the winner-Relaxing sum can indeed be used to increase the mutual information of a map generated by the Neural Gas algorithm. Using a standard setup as in [5] of 50 Neurons and 10^7 training steps with a probability density $P(x_1 \dots x_d) = \prod_i \sin(\pi x_i)$, with fixed $\lambda = 1.5$ and ϵ decaying from 0.5 to 0.05, the entropy of the resulting map computed for an input dimension of 1, 2 and 3 is plotted in Fig. 1. The entropy shows a dimension-dependent maximum approximately at $\kappa = \frac{2}{d+2}\mathrm{e}^{\lambda}$. The scaling of the position of the entropy maximum with input dimension is in agreement with the continuum theory, as well as the prediction of the opposite sign of κ that has to be taken to increase mutual information. Our numerical investigation indicates that the prefactor in fact has to be taken in account for finite λ and a finite number of neurons.

To conclude, within a broad range around the optimal κ the entropy is close to the maximum $\sum_{i=1}^{N} P_i \log(P_i) = \log(N)$ given by information theory.

Acknowledgements: The authors want to thank Th. Martinetz for detailed comments and intensive discussion.

ESANN'2003 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 23-25 April 2003, d-side publi., ISBN 2-930307-03-X, pp. 93-98

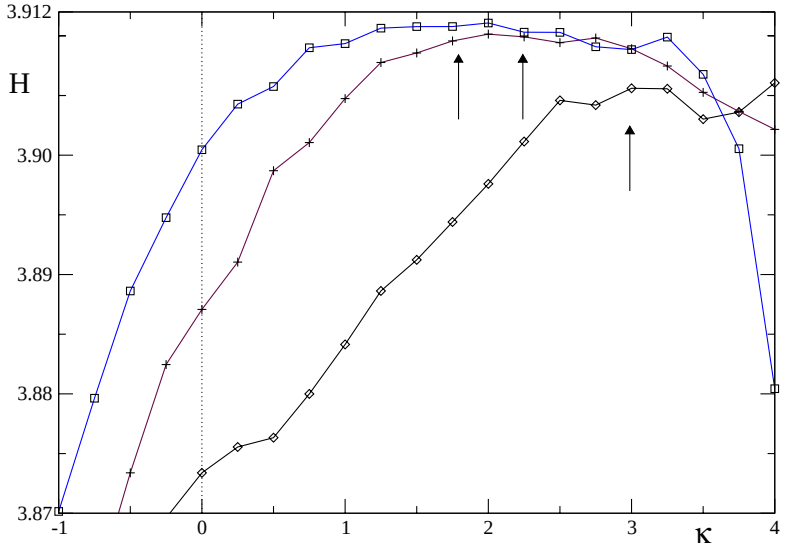

Figure 1: Plot of the entropy H curves for varying values of κ for one- ($\diamond$), two- ($+$), and three-dimensional ($\square$) data. The entropy has the maximum $\log(50) \simeq 3.912$ if the magnification equals unity [11]. The arrows indicate the rescaled κ_{opt}-values for the respective data dimensions.

References

[1] H.-U. Bauer, R. Der, and M. Herrmann. Controlling the magnification factor of self–organizing feature maps. *Neural Comp.*, 8(4):757–771, 1996.

[2] J. Bruske and G. Sommer. Dynamic cell structure learns perfectly topology preserving map. *Neural Computation*, 7(4):845–65, July 1995.

[3] J. C. Claussen. Generalized Winner-Relaxing Kohonen Feature Maps. *e-print cond-mat*, (http://arXiv.org/cond-mat/0208414), 2002.

[4] P. Grassberger and I. Procaccia. Measuring the strangeness of strange attractors. *Physica D*, 9:189–208, 1983.

[5] M. Herrmann and T. Villmann. Vector quantization by optimal neural gas. In W. Gerstner, A. Germond, M. Hasler, and J.-D. Nicoud, editors, *Artificial Neural Networks – Proc.ICANN'97, Lausanne*, pages 625–630. LNCS 1327, Springer Verlag Berlin Heidelberg, 1997.

[6] T. Kohonen. Self-Organizing Maps: Optimization approaches. In T. Kohonen, K. Mäkisara, O. Simula, and J. Kangas, eds., *Artificial Neural Networks*, volume II, pages 981–990, Amsterdam, Netherlands, 1991. North-Holland.

[7] T. Kohonen. *Self-Organizing Maps*, vol. 30 of *Springer Series in Information Sciences*. Springer, Berlin, Heidelberg, 1995. (2nd Ext. Ed. 1997).

[8] T. M. Martinetz, S. G. Berkovich, and K. J. Schulten. 'Neural-gas' network for vector quantization and its application to time-series prediction. *IEEE Trans. on Neural Networks*, 4(4):558–569, 1993.

[9] H. Ritter, T. Martinetz, and K. Schulten. *Neural Computation and Self-Organizing Maps: An Introduction*. Addison-Wesley, Reading, MA, 1992.

[10] T. Villmann. Controlling strategies for the magnification factor in the neural gas network. *Neural Network World*, 10(4):739–750, 2000.

[11] P. L. Zador. Asymptotic quantization error of continuous signals and the quantization dimension. *IEEE Trans. on Information Theory*, (28):149–159, 1982.