Asymptotic State of One-Dimensional SOM at Normal Point Density Input Passed Through Non-Linear Channel

IVO R. DRAGANOV, ANTOANETA A. POPOVA, NIKOLAY N. NESHOV
Radiocommunications and Videotechnologies Department
Technical University - Sofia
8 Kliment Ohridski Blvd., 1000 Sofia
BULGARIA
idraganov@tu-sofia.bg, antoaneta.popova@tu-sofia.bg, nicknesh@abv.bg

Abstract: In this paper an analysis is presented concerning the asymptotic state of the one-dimensional self-organizing map (SOM) with finite grid in the case of normal point distribution input passed through non-linear channel at first. The SOM distortion measure is analyzed with its value found approximately. The results obtained are considered useful enough in wide variety of practical cases where fine tuning of the SOM is needed.

Key Words: Self-Organizing Map, Normal Point Distribution, Non-Linear Channel, Distortion Measure

1 Introduction
It is well known fact that the area allocated for storing the most important feature set inside a self-organizing map (SOM) is proportional to the frequency of occurrence of that very same feature in the observations [1]. As the SOM structure tends to become very complex in the most of its real case applications often the magnification factor is used to describe heaping of feature vectors. It is simply the inverse of the point density around each neuron representing a cluster.

So far an investigation of the point density for the linear map is led in the presence of a very large number of codebook vectors over a finite area [2], [3]. It is revealed that the asymptotic point density is proportional to the probability of a certain feature vector occurring raised to some exponent depending of the number of neighbors including the winning neuron and some scalar factor.

In any case the initial neighbor function width may vary largely during the training process starting with huge values and ending with zero-order topology case – no neighbors except the winner are present. This boundary case is undesired since the learning process no longer maintains the order of the codebook vectors. The approximation accuracy of the probability of occurring for a feature and the minimum stability of ordering demanding more neighbor interactions are the two aspects to be balanced.

If we have no neighbors around the winner a simple scalar quantization case occurs. Then the power of the asymptotic function for the point density decreases, according to [3] below 1/3. Getting this power to higher values incrementally by trial and error approach seems a good solution but the following tendencies should be considered. If we try to approach the Bayesian classifier, i.e. to find the optimal classification border and the density functions of adjacent clusters are close to each other the latter could be replaced with any other pair of monotonic functions of densities. In such a case the practical SOM application is adaptable to simplification.

The other important property is that when feature dimensionality is increased in the order of hundreds of components per vector the power is close to 1 [3]. Similar research on the change of this power is done in [4] when the neighbor function is Gaussian kernel and its normalized second moment is independent variable. The resulting range for the power value in this case is from 1/3 to 2/3. Analogous results are presented in [5].

One recent research [6] investigates the influence of the normal point density input over the asymptotic state of a finite one-dimensional SOM and its distortion measure. In [7] a typical practical challenge is given concerning the location of the representative vectors of 256-QAM system operating in the presence of additive white Gaussian noise (AWGN) and third order nonlinearity which can be solved by applying the approach presented here. In part 2 such analysis is presented and in part 3 some computational results are given. In part 4 a conclusion is made.

2 SOM Analysis with Normal Point Density Input Passed Through Non-Linear Channel
Let one-dimensional feature space of \( x \) is considered. For our analysis to be correct the following assumptions should be granted: the number of points (feature vectors) must be large enough (e.g. by criteria given in [1]) and they must be stochastic variables so their differential probability for each cluster they fall into, i.e. the...
probability density \( p(x) \) could be defined. The codebook vectors \( m_i \) usually form regular optimal configuration and thus can not be stochastic. Their number is typically low in any cluster as well.

2.1 Asymptotic State of the One-Dimensional Finite-Grid SOM

Let suppose \( m_i \) and \( m_{i-1} \) are two neighboring points. A way of defining the point density is as \((m_{i-1} - m_i)^{-1}\) but it does not cover the samples around the boundaries of the clusters for which this density does not have meaning. So a better way of defining it is as the inverse of the width of the Voronoi set \((m_{i-1} - m_i)/2\). The case investigated here concerns input data with at least one neighbor at each side according to [1] is given by:

\[
m_i(t + 1) = m_i(t) + \varepsilon(t)(x(t) - m_i(t)), \text{ for } i \in \mathbb{N}_c,
\]

\[
m_i(t + 1) = m_i(t) \quad \text{for } i \notin \mathbb{N}_c,
\]

\[
c = \text{arg min}_i \|x(t) - m_i(t)\|,
\]

\[
N_c = \{\max(l, c-1), c, \min(k, c+1)\}
\]

where \( N_c \) is the neighbor set around node \( c \) and \( \varepsilon(l) \) is the learning-rate factor. The Voronoi set \( V_i \) around \( m_i \) is defined as:

\[
V_i = \left[\frac{m_{i-1} + m_i}{2}, \frac{m_i + m_{i+1}}{2}\right], \quad V_i = \left[0, \frac{m_i + m_{i-1}}{2}\right],
\]

\[
V_k = \left[\frac{m_{k-1} + m_k}{2}, 0\right], \quad \text{for } 1 < i < k,
\]

\[
U_j = V_{j-1} \cup V_j \cup V_{j+1}, U_i = V_i \cup V_{i+1},
\]

\[
U_k = V_{k-1} \cup V_k, \quad \text{for } 1 < i < k.
\]

Fig. 1. Non-linear transform of the input distribution

In this case \( U_i \) is the set of such \( x(t) \) which provoke changes in \( m_i(t) \) during one learning step. Following (1) and (2) we get to the well known stationary equilibrium for \( m_i \) coinciding for the general case [1]:

\[
m_i = E\{x \mid x \in U_i\}, \quad \forall i.
\]

In other words every \( m_i \) becomes centroid of the probability mass for each \( U_i \) and then for \( 2 < i < (k-1) \) the limits for \( U_i \) are:

\[
A_i = \frac{1}{2}(m_{i-2} + m_{i-1}), \quad B_i = \frac{1}{2}(m_{i+1} + m_{i+2}).
\]

For \( i = 1 \) and \( i = 2 \), \( A_i = 0 \), and for \( i = k - 1 \) and \( i = k \), \( B_i = 1 \).

The case investigated here concerns input data with the following distribution – typically one-dimensional vector \((\text{point(s) over } x)\) with AWGN present:

\[
p(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-x_0)^2}{2\sigma^2}} = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{x^2}{2\sigma^2}},
\]

where \( x_0 = 0 \) is set since \( 0 \leq x \leq 1 \). Suppose that this multitude of samples is passing through a channel with non-linearity approximated by a polynomial of third order of the type:

\[
y = a_3x^3 + a_2x^2 + a_1x + a_0.
\]

For most of the typical communication channels \( a_0 = 0, a_1 = 1, a_2 < 0 \) and \( a_3 > 0 \) – the latter two parameters are of the order of \( 10^8 \) and smaller with deviation of the same scale [7]. Now we need to find the statistical distribution of the samples toward \( y \) [8] that is the density \( q(y) \) (Fig. 1).

\[
q(y) = p(x) \frac{dx}{dy},
\]

If consider the inverse \( x = y^{-1}(y) \) then:

\[
\frac{dx}{dy} = \frac{d(y^{-1}(y))}{dy} = (y^{-1}(y))^'.
\]

Knowing that \( q(y) > 0 \) is always true and substituting (8) in (7) we finally find:
\[
q(y) = p \left[y^{-1}(y) \right] = \frac{d[y^{-1}(y)]}{dy}.
\]

If the general form of the inverse polynomial of (6) is:

\[
y^{-1}(y) = b_3 y^3 + b_2 y^2 + b_1 y,
\]

where \( b_i \) could easily be found from \( a_n \), for \( i = 1, 2 \) and 3, then according to (9):

\[
q(y) = \frac{3b_3 y^2 + 2b_2 y + b_1 + (b_3 y^3 + b_2 y^2 + b_1 y)^3}{\sqrt[2]{2 \pi \sigma}}.
\]

As (11) is too complex to be used in finding the centroids of the probability masses, Taylor series are used instead to the second order for simplicity:

\[
q(y) = \sum_{i=0}^{1} \frac{q^{(i)}(0)}{i!} y^i \approx q(0) + q'(0) y + \frac{q''(0)}{2} y^2,
\]

\[
q'(y) = \frac{1}{\sqrt[2]{2 \pi \sigma}} e^{- \frac{1}{2 \pi \sigma} \left[ 6b_3 y - 2b_2 - (b_3 y^3 + b_2 y^2 + b_1 y)^2 \right]}
\]

\[
q''(y) = \frac{1}{2 \pi \sigma} \left( (b_3 y^3 + b_2 y^2 + b_1 y)^3 \right) \left( (b_3 y^3 + b_2 y^2 + b_1 y)^2 \right).
\]

The solution of such a set could be done in the following way [1]. Let us have:

\[
\text{The stationary values of the centroids of the probability masses, Taylor series are used instead to the second order for simplicity:}
\]

\[
\text{As (11) is too complex to be used in finding the centroids of the probability masses, Taylor series are used instead to the second order for simplicity:}
\]

\[
\text{For improved accuracy more values of the } m_i \text{ are needed as we shall see in the next section.}
\]
From (21) and (22) the gradient of $E$ could be found as:

$$
\frac{\partial E}{\partial m_i} = \sum_{j \in \mathcal{N}_i} E_{i,j} + \sum_{j \in \mathcal{N}_i} E_{j,i} + \sum_{j \in \mathcal{N}_i} E_{i+1,j} + \sum_{j \in \mathcal{N}_i} E_{j+1,i},
$$

but here will not be given due to its long form and since the calculations are trivial.

To obtain minimal distortion for the SOM optimal values for $m_i$ should be found. A way of doing this is using the gradient-descent method according to:

$$
m_i(t+1) = m_i(t) - \lambda(t) \left( \frac{\partial E}{\partial m_i} \right),
$$

where the scalar factor $\lambda(t)$ is typically in the range from 0.001 to 0.01 but even larger values of the order of 10 are tolerable [1]. Here $E$ is of the fifth order, so at least one minimum can be found. The tendency is with the growth of $\lambda(t)$ the found minimum to become global not depending on the initial values of $m_i$.

### 3 Experimental Results

As a simulation environment we use Matlab® R2009B over MS® Windows® XP® Pro SP3.

First $\alpha$ from (20) is found for different number of grid points when $a_0 = 0$, $a_1 = 1$, $a_2 = -6.10^{-8}$ and $a_3 = 2.10^{12}$ – a typical non-linear communication channel from a cable television network. The more $m_i$ are used the more accurate are the results. For $i = 4$ and $j = k - 3$ assuring negligible border effects 10, 25, 50, and 100 grid points are used. The same experiment is done with normal probability functions in [7], so here a direct comparison can be made. The results are given in Table 1.

#### Table 1

The derived $\alpha$ for normal and distorted distribution of the input

<table>
<thead>
<tr>
<th>Grid points</th>
<th>Normal [7]</th>
<th>Distorted</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0.5850</td>
<td>0.5905</td>
</tr>
<tr>
<td>25</td>
<td>0.5980</td>
<td>0.5996</td>
</tr>
<tr>
<td>50</td>
<td>0.5980</td>
<td>0.6040</td>
</tr>
<tr>
<td>100</td>
<td>0.5992</td>
<td>0.6100</td>
</tr>
</tbody>
</table>

It is clearly seen from Table 1 the resulting $\alpha$ is greater for each case for the distorted distribution and the values are close to 0.6. Graphically the results are given on Fig. 2. The reason for the relative small differences between the alphas for both distributions is the small values for $a_3$ and $a_1$ typical for large number of practical cases making the probability density curves close one to another.

After computing the optimal $m_i$ using (19) and (20) for the same number of grid points – from 10 to 100, it is possible to estimate $\alpha$. Now when the minimal distortions have been assured this is done experimentally over real input data with distorted normal distribution around the winners. The results are given in Table 2.

#### Table 2

Experimentally estimated $\alpha$ for both distributions of the input at the minimal distortion measure

<table>
<thead>
<tr>
<th>Grid points</th>
<th>Normal [7]</th>
<th>Distorted</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0.2989</td>
<td>0.3010</td>
</tr>
<tr>
<td>25</td>
<td>0.3330</td>
<td>0.3332</td>
</tr>
<tr>
<td>50</td>
<td>0.3331</td>
<td>0.3334</td>
</tr>
<tr>
<td>100</td>
<td>0.3330</td>
<td>0.3332</td>
</tr>
</tbody>
</table>

Obviously the values obtained here again do not depend on the number of grids and again are even closer one to another despite the differences between the input distributions. The significant difference is with the results obtained by the theoretical derivations from (20). Now when we have the optimal $m_i$ found it is seen that the exponent of the approximated state of the SOM is virtually equal to 1/3. This is actually a case coinciding with the optimal vector quantization [1] unlike the case of $\alpha = 0.6$ and confirms the correctness of the optimal $m_i$ calculated. Graphically the results from Table 2 are given in Fig. 3.

In both cases, for $\alpha = 0.6$ and $\alpha = 0.3$, $m_i$ can be considered as forming an elastic network expanding into the input feature space following the order of appearance of the consecutive samples. The fact that the presented stochastic approximation (1) – (20) can not ensure the
optimal case when $\alpha = 0.3$ does not mean that it is useless – actually saving the iterations from (20) and the preliminary fixation of $m_i$ from (15)-(19) means considerable saving of computation time which is important in wide range of practical cases.

The results achieved prove the correctness of the suggested approach which is considered useful in a large number of practical cases where the input data with normal point density distribution is passed through non-linear channel.

**Acknowledgement**

This paper was supported by the National Fund for Scientific Research of the Bulgarian Ministry of Education and Science (Contract - BY-I-302/2007).

**References:**


