3 Point Density of the Model Vectors in the SOM

Teuvo Kohonen

3.1 Introduction

In the classical vector quantization (VQ) the objective is usually to approximate n-dimensional real signal vectors $\mathbf{x} \in \mathbb{R}^n$ using a finite number of quantized vectorial values $\mathbf{m}_i \in \mathbb{R}^n, i = 1, \ldots, N$ called the codebook vectors. One may want, e.g., to minimize the functional called the *distortion measure*:

$$E_{VQ} = \int \|\mathbf{x} - \mathbf{m}_c\|^r p(\mathbf{x}) d\mathbf{x} \ , \tag{23}$$

where r is some real-valued exponent, the integral is taken over the complete metric $\mathbf{x}$ space, $\mathbf{m}_c$ is the $\mathbf{m}_i$ closest to $\mathbf{x}$, i.e.,

$$c = \arg\min_i \{\|\mathbf{x} - \mathbf{m}_i\|\} \ , \tag{24}$$

the norm is usually assumed Euclidean, $p(\mathbf{x})$ is the probability density function of $\mathbf{x}$, and $d\mathbf{x}$ is a shorthand notation for the n-dimensional volume differential of the integration space. All the values of $\mathbf{x}$ that have the same $\mathbf{m}_c$ as their nearest neighbor are said to constitute the *Voronoi set* associated with $\mathbf{m}_c$. Under rather general conditions one can determine the point density $q(\mathbf{x})$ of the $\mathbf{m}_i$ as in the following expression [2, 8]:

$$q(\mathbf{x}) = \text{const.} \left[p(\mathbf{x})^{\frac{n}{n+r}} \right] \ . \tag{25}$$

A related problem occurs with the *self-organizing map (SOM)*, which resembles VQ, but in which the $\mathbf{m}_i$ are *ordered* in $\mathbb{R}^n$ according to their similarity. The SOM carries out a vector quantization, too, but the placement of the $\mathbf{m}_i$ in the signal space is restricted by the neighborhood relations.

A long-standing problem has been whether the SOM model vectors could be determined by the minimization of some objective function. For instance, Kohonen, 1991 [3] discussed the distortion measure

$$E = \int \sum_i h_{ci} \|\mathbf{x} - \mathbf{m}_i\|^2 p(\mathbf{x}) d\mathbf{x} = \sum_i \int_{\mathbf{x} \in V_i} \sum_j h_{ij} \|\mathbf{x} - \mathbf{m}_j\|^2 p(\mathbf{x}) d\mathbf{x} \ . \tag{26}$$

where V_i is the Voronoi set around $\mathbf{m}_i$. The gradient of E consists of two terms :

$$\frac{\partial E}{\partial \mathbf{m}_j} = G + H \ , \tag{27}$$

where G is obtained if the integration borders are kept fixed and the differentiation with respect to $\mathbf{m}_j$ is carried out in the integrand only, whereas in the computation of H, the integrand is held constant and the integration borders are let to vary when the $\mathbf{m}_j$ differential is taken.

In order to avoid the evaluation of the above integrals, one may try to resort to the classical method called the *stochastic approximation* [7]. If the inputs $\mathbf{x}$ are obtained as a sequence of samples $\{\mathbf{x}(t)\}$, one can compute at every time t the best tentative estimate of $\mathbf{m}_i$ so far, called $\mathbf{m}_i(t)$. The expression

$$E_1(t) = \sum_i h_{ci} \|\mathbf{x}(t) - \mathbf{m}_i(t)\|^2 \tag{28}$$

is taken as the sample of function E at time t. Following Robbins and Monro, at time t we approximate the gradient of E with respect to $\mathbf{m}_i$ by the gradient of $E_1(t)$ with respect to $\mathbf{m}_i(t)$. Then

$$\mathbf{m}_i(t+1) \;=\; \mathbf{m}_i(t) - \left(\frac{\varepsilon}{2}\right) \frac{\partial E_1(t)}{\partial \mathbf{m}_i(t)} \tag{29}$$

with ε a small number. However, it is not yet clear how good an approximation the Robbins-Monro process is in this case. We have now shown that the *point density derived from the SOM algorithm* and the *point density derived from the SOM distortion measure* are different already in the one-dimensional case.

3.2 Point Densities in a Simple One-Dimensional SOM

3.2.1 Asymptotic State of the One-Dimensional Finite-Grid SOM Algorithm

Consider a series of samples of the input $x(t) \in \mathbb{R}$, $t = 0, 1, 2, \ldots$ and a set of k model (codebook) values $m_i(t) \in \mathbb{R}$, $t = 0, 1, 2, \ldots$, whereupon i is the model index ($i = 1, \ldots, k$). For convenience assume $0 \le x(t) \le 1$.
The original one-dimensional self-organizing map (SOM) algorithm with at most one neighbor on each side of the best-matching m_i reads (Kohonen, 1997):

$$\begin{aligned}
m_i(t+1) &= m_i(t) + \varepsilon(t)[x(t) - m_i(t)] \;\text{ for }\; i \in N_c \;, \\
m_i(t+1) &= m_i(t) \;\text{ for }\; i \notin N_c \;, \\
c &= \arg\min_i\{|x(t) - m_i(t)|\} \;, \text{ and} \\
N_c &= \{\max(1, c-1), c, \min(k, c+1)\} \;,
\end{aligned} \tag{30}$$

where N_c is the neighborhood set around node c, and $\varepsilon(t)$ is a small scalar value called the learning-rate factor. In order to analyze the asymptotic values of the m_i, let us assume that the m_i are already ordered. The Voronoi set V_i around m_i is

$$\begin{aligned}
\text{for } 1 < i < k,\; V_i &= \left[\frac{m_{i-1} + m_i}{2}, \frac{m_i + m_{i+1}}{2}\right] \;, \\
V_1 &= \left[0, \frac{m_1 + m_2}{2}\right] ,\; V_k = \left[\frac{m_{k-1} + m_k}{2}, 1\right] \;, \;\text{ and denote} \\
\text{for } 1 < i < k,\; U_i &= V_{i-1} \cup V_i \cup V_{i+1} \;, U_1 = V_1 \cup V_2,\; U_k = V_{k-1} \cup V_k \;.
\end{aligned} \tag{31}$$

One can write the condition for stationary equilibrium of the m_i for a constant ε as:

$$\forall i, \;\; \mathrm{E}\,\{x - m_i | x \in U_i\} = 0 \;. \tag{32}$$

For $2 < i < k-1$ we have for the limits of the U_i:

$$A_i = \frac{1}{2}(m_{i-2} + m_{i-1}) \quad , B_i = \frac{1}{2}(m_{i+1} + m_{i+2}) \ . \tag{33}$$

For $i = 1$ and $i = 2$ we must take B_i as above, but $A_i = 0$; and for $i = k-1$ and $i = k$ we have A_i as above and $B_i = 1$.

Numerical example. Let $p(x) = 2x$ for $0 \leq x \leq 1$ and $p(x) = 0$ otherwise. The stationary values of the m_i are defined by the set of nonlinear equations

$$\forall i, \ m_i = \mathrm{E}\{x | x \in U_i\} = \frac{2(B_i^3 - A_i^3)}{3(B_i^2 - A_i^2)} \tag{34}$$

and the solution of (34) is sought by the so-called *contractive mapping*. Let us denote $\mathbf{z} = [m_1, m_2, \dots, m_k]^{\mathrm{T}}$. Then the equation to be solved is of the form $\mathbf{z} = f(\mathbf{z})$. Starting with the first approximation for $\mathbf{z}$ denoted $\mathbf{z}^{(0)}$, each improved approximation for the root is obtained recursively:

$$\mathbf{z}^{(s+1)} = f(\mathbf{z}^{(s)}) \ . \tag{35}$$

In the present case one may select for the first approximation of the m_i, e.g., equidistant values.
It may now be expedient to define the point density q_i around m_i as the inverse of the length of the Voronoi set, or $q_i = [(m_{i+1} - m_{i-1})/2]^{-1}$.
The problem expressed in a number of previous works, e.g., Ritter and Schulten (1986), Ritter (1991), and Dersch and Tavan (1995), is to find out whether q_i could be approximated by the functional form const.$[p(m_i)]^\alpha$. Previously this was only shown for the continuum limit, i.e. for an infinite number of grid points. The present numerical analysis allows us to derive results for finite-length grids, too. Assuming tentatively that the power law holds for the models m_i through m_j (leaving aside models near to the ends of the grid), we shall then have

$$\alpha = \frac{\log(m_{i+1} - m_{i-1}) - \log(m_{j+1} - m_{j-1})}{\log[p(m_j)] - \log[p(m_i)]} \ . \tag{36}$$

In Table 1, using $i = 4$ and $j = k-3$, between which the border effects may be assumed as negligible, the exponent α has been estimated for 10, 25, 50, and 100 grid points, respectively.

3.2.2 Optimum of the One-Dimensional SOM Distortion Measure with Finite Grid

In the previous example, (26) becomes

$$\begin{aligned} E &= 2\sum_i \sum_{j \in N_i} \int_{C_i}^{D_i} (x - m_j)^2 x dx \\ &= \sum_i \sum_{j \in N_i} m_j^2 (D_i^2 - C_i^2) - \frac{4}{3} m_j (D_i^3 - C_i^3) + \frac{1}{2}(D_i^4 - C_i^4) \end{aligned} \tag{37}$$

where the *neighborhood set of indices* N_i was defined in (30), and the borders C_i and D_i of the Voronoi set V_i are $C_1 = 0$, $D_k = 1$,

$$C_i = \frac{m_{i-1} + m_i}{2} \quad \text{for } 2 \le i \le k \text{ , and } D_i = \frac{m_i + m_{i+1}}{2} \quad \text{for } 1 \le i \le k-1 \text{ .} \tag{38}$$

The optimal values of the m_i are determined by the gradient method:

$$\forall i \,, \quad m_i(t+1) = m_i(t) - \lambda(t) \cdot \partial E / \partial m_i|_t \,, \tag{39}$$

where $\lambda(t)$ is a suitable small scalar factor. With $\lambda(t) > .01$ (even with $\lambda(t) = 10$) and starting with very different initial values for the m_i, the process has converged robustly to a unique global minimum. After computation of the optimal values $\{m_i\}$, the exponent α of the tentative power law was computed from (36) of the previous section and presented in Table 1 for different lengths of the grid. Clearly the cases discussed in Secs. 2.1 and 2.2 are qualitatively different.

Table 1: Exponent α derived from the SOM algorithm and the SOM distortion measure, respectively

Grid points	SOM algorithm	SOM distortion measure
10	0.5831	0.3281
25	0.5976	0.3331
50	0.5987	0.3333
100	0.5991	0.3331

3.3 Derivation of the VQ Point Density by the Calculus of Variations

The technique that will be used to approximate point densities for higher-dimensional SOMs will first be applied to the simpler VQ problem. If $p(\mathbf{x})$ is smooth and the placement of the $\mathbf{m}_i$ in the signal space is reasonably regular, one may try to approximate the Voronoi sets, which are polytopes in the n-dimensional space, by n-dimensional hyperspheres centered at the $\mathbf{m}_i$. This, of course, is a rough approximation, but it was in fact used already in the classical VQ papers [2, 8], and no better treatments exist for the time being.

Denoting the radius of the hypersphere by R, its hypervolume has the expression kR^n, where k is a numerical factor. If $p(\mathbf{x})$ is approximately constant over the polytope, the *elementary integral of the distortion* $\|\mathbf{x} - \mathbf{m}_i\|^r = \rho^r$ *over the hypersphere* is

$$D = nk \int_0^R p(\mathbf{x}) \cdot \rho^r \cdot \rho^{n-1} d\rho = \frac{nk}{n+r} \cdot p(\mathbf{x}) \cdot R^{n+r} \;; \tag{40}$$

notice that if $v(\rho)$ is the volume of the n-dimensional hypersphere with radius ρ, then $dv(\rho)/d\rho = nk\rho^{n-1}$ is the "hypersurface area" of the hypersphere.

The point density $q(\mathbf{x})$ is defined as $1/kR^n$. What we aim at first is the approximate "distortion density" that we denote by $I[\mathbf{x}, q(\mathbf{x})]$, where $q(\mathbf{x})$ is the point density of the $\mathbf{m}_i$ at the value $\mathbf{x}$:

$$I[\mathbf{x}, q(\mathbf{x})] = \frac{D}{kR^n} = \frac{n}{n+r} \cdot p(\mathbf{x}) \cdot R^r = \frac{np(\mathbf{x})}{n+r}[kq(\mathbf{x})]^{-\frac{r}{n}} \ . \tag{41}$$

In the continuum limit, the total distortion measure is the integral of the "distortion density" over the complete signal space:

$$\int I[\mathbf{x}, q(\mathbf{x})]d\mathbf{x} = \int \frac{np(\mathbf{x})}{n+r}[kq(\mathbf{x})]^{-\frac{r}{n}}d\mathbf{x} \ . \tag{42}$$

This integral is minimized under the restrictive condition that the sum of all quantization vectors shall always equal N; in the continuum limit the condition reads

$$\int q(\mathbf{x})d\mathbf{x} = N \ . \tag{43}$$

In the classical *calculus of variations* one often has to optimize a functional which in the one-dimensional case with one independent variable x and one dependent variable $y = y(x)$ reads

$$\int_a^b I(x, y, y_x)dx \ ; \tag{44}$$

here $y_x = dy/dx$, and a and b are fixed integration limits. If a restrictive condition

$$\int_a^b I_1(x, y, y_x)dx = \text{const.} \tag{45}$$

has to hold, the generally known Euler variational equation reads, using the Lagrange multiplier λ and denoting $K = I - \lambda I_1$,

$$\frac{\partial K}{\partial y} - \frac{d}{dx}\frac{\partial K}{\partial y_x} = 0 \ . \tag{46}$$

In the present case x is vectorial, denoted by $\mathbf{x}$, $y = q(\mathbf{x})$, and I and I_1 do not depend on $\partial q/\partial \mathbf{x}$. In order to introduce fixed, finite integration limits one may assume that $p(\mathbf{x}) = 0$ outside some finite support. Now we can write

$$I = \frac{nk^{-\frac{r}{n}}}{n+r} \cdot p(\mathbf{x}) \cdot [q(\mathbf{x})]^{-\frac{r}{n}} \ , \ I_1 = q(\mathbf{x}) \ , \ K = I - \lambda I_1 \ , \tag{47}$$

$$\frac{\partial K}{\partial q(\mathbf{x})} = -\frac{rk^{-\frac{r}{n}}}{n+1} \cdot p(\mathbf{x}) \cdot [q(\mathbf{x})]^{-\frac{n+r}{n}} - \lambda = 0 \, . \tag{48}$$

At every location $\mathbf{x}$ there then holds

$$q(\mathbf{x}) = C \cdot [p(\mathbf{x})]^{\frac{n}{n+r}} \ , \tag{49}$$

where the constant C can be solved by substitution of $q(\mathbf{x})$ into (43). Clearly (49) is identical with (25). We have now obtained the same result that earlier ensued from very intricate signal and error-theoretic probabilistic considerations.

3.4 The SOM Point Density Derived from the Distortion Measure for Equal Vector and Grid Dimensionalities

It is possible to carry out the following analysis with a rather general symmetric h_{ij}, but for simplicity, without much loss of generality, we may assume, like in the basic SOM theory, $h_{ij} = 1$ within a certain radius, relating to the distances measured along the grid from the node j; outside this radius $h_{ij} = 0$. This is called the *neighborhood* around grid point $\mathbf{m}_j$.

In the signal space this then means that if $p(\mathbf{x})$ and the point density of the $\mathbf{m}_i$ are changing slowly, in the first approximation we can take $h_{ij} = 1$ up to a distance aR from $\mathbf{m}_j$, where R is the radius of the hypersphere that approximates the Voronoi set V_j, and a is a numerical constant; in other words, the neighborhood shall contain a constant number of grid points everywhere over the SOM (except at the borders of the SOM).

For the elementary integral of the distortion *over the neighborhood* up to radius aR, with the exponent $r = 2$, we then obtain according to (40):

$$D = \frac{nk}{n+2} \cdot p(\mathbf{x}) \cdot (aR)^{n+2} \, , \tag{50}$$

and relating the "distortion density" to the "volume" of V_j,

$$I[\mathbf{x}, q(\mathbf{x})] = \frac{D}{kR^n} = \frac{na^{n+2}}{n+2} \cdot p(\mathbf{x}) \cdot [kq(\mathbf{x})]^{-\frac{2}{n}} \, . \tag{51}$$

We then directly obtain in analogy with equations (41) through (48) and taking $r = 2$ the result

$$q(\mathbf{x}) = C'[p(\mathbf{x})]^{\frac{n}{n+2}} \tag{52}$$

with another constant C' computed from the normalization condition.

Notice that (52), however, does not yet tell anything about the exponent if the *SOM algorithm* is used to determine the $\mathbf{m}_i$.

References

[1] Dersch, D.R., Tavan, P. (1995). Asymptotic level density in topological feature maps. *IEEE Trans. Neural Networks* 6:230-236.

[2] Gersho, A. (1979) Asymptotically optimal block quantization. *IEEE Trans. Inf. Theory* 25:373-380.

[3] Kohonen, T. (1991) Self-organizing maps: optimization approaches. In Kohonen, T., Mäkisara, K., Simula, O., Kangas, J. (Eds.), *Artificial Neural Networks* (vol. 2, pp. 981-990). Amsterdam: Elsevier.

[4] Kohonen, T. (1999) Comparison of SOM point densities based on different criteria. *Neural Computation*, in press.

[5] Ritter, H. (1991) Asymptotic level density for a class of vector quantization processes. *IEEE Trans. Neural Networks* 2:173-175.

[6] Ritter, H., Schulten, K. (1986) On the stationary state of Kohonen's self-organizing sensory mapping. *Biol. Cybern.* 54:99-106.

[7] Robbins, H., Monro, S. (1951) A stochastic approximation method. *Ann. Math. Statist.* 22:400-407.

[8] Zador, P.L. (1982) Asymptotic quantization error of continuous signals and the quantization dimension. *IEEE Trans. Inf. Theory* IT-28:139-149.