Concerning the Differentiability
of the Energy Function
in Vector Quantization Algorithms

Dominique Lepetz (1)

Max Némoz-Gaillard (2)

Michaël Aupetit (3)

(1) EMA-DM, 6 Av. de Clavières,30319 Alès cedex, FRANCE

dlepetz at wanadoo dot fr

(2) EMA-CMGD, 6 Av. de Clavières, 30319 Alès cedex, FRANCE

Max.Nemoz-Gaillard at ema dot fr

(3) Corresp. author - CEA, DASE, BP.12, 91680 Bruyères-le-Châtel, FRANCE

michael.aupetit at cea dot fr

June 29, 2018

1

Abstract

The adaptation rule for Vector Quantization algorithms, and consequently the convergence of the generated sequence, depends on the existence and properties of a function called the energy function, defined on a topological manifold. Our aim is to investigate the conditions of existence of such a function for a class of algorithms exemplified by the initial "K-means" (Mac-Queen, 1967) and Kohonen algorithms (Kohonen, 1982; Kohonen, 1988). The results presented here supplement previous studies, including (Tolat, 1990), (Erwin et al., 1992), (Cottrell et al., 1994),(Pagès, 1993) and (Cottrell et al., 1998). Our work shows that the energy function is not always a potential but at least the uniform limit of a series of potential functions which we call a pseudo-potential. It also shows that a large number of existing vector quantization algorithms developed by the Artificial Neural Networks community fall into this category. The framework we define opens the way to study the convergence of all the corresponding adaptation rules at once, and a theorem gives promising insights in that direction. We also demonstrate that the "K-means" energy function is a pseudo-potential but not a potential in general. Consequently, the energy function associated to the "Neural-Gas" is not a potential in general.

Keywords

Vector Quantization, K-means, Self-Organizing Maps, Neural-Gas, energy function, potential function, pseudo-potential

1   Introduction

In vector quantization theory (Gray and Neuhoff, 1998), a set of prototypes [1] $\underline{w} = (w_1, ..., w_n)$ is placed on a manifold $V \subset \mathbb{R}^d, d \geq 1$, in order to minimize

---
[1] also called "codebook vectors", "reference vectors", "units" or "neurons".

2

the following integral function, called the "energy function":

$$E_V(\underline{w}) = \int_V \frac{1}{2} \sum_{p=1}^n P(v)\psi_p(\underline{w}, v)(v - w_p)^2 dv = \int_V F(\underline{w}, v)dv \qquad (1)$$

where $P(v)$ indicates the probability density defined on V. We focus on the stochastic iterative approaches where at each time step, a datum v is drawn from the probability density function (pdf) P, and the prototypes $\underline{w}$ are adapted according to v using the adaptation rule:

$$\Delta w_p = \alpha \psi_p(\underline{w}, v)(v - w_p) \qquad (2)$$

where the adaptation step is tuned using the parameter α generally decreasing over the time (α is taken thereafter equal to 1 without restricting the general results), and ψ_p is a "neighborhood" function particular to each vector quantization algorithm. Here we focus on discontinuous ψ_p functions.

A main concern in the field of Vector Quantization, is to decide whether the adaptation rule (2) corresponds or not to a stochastic gradient descent along the energy function (1), *i.e.* whether this energy function is or is not a potential onto the entire manifold V. On one hand, if the energy function is a potential then the convergence of the prototypes obeying their adaptation rule toward a minimum of this energy function is well established, in particular in the stochastic optimization framework (Robbins and Monro, 1951; Albert and Gardner, 1967) with which this paper is concerned. For example, the energy function associated to the K-means algorithm (Mac-Queen, 1967; Ahalt et al., 1990), stochastic version of the LBG algorithm of Linde *et al.* (Linde et al., 1980), is a potential as long as the pdf P is continuous (Kohonen, 1991; Pagès, 1993; Cottrell et al., 1998).

On the other hand, if the energy function is not a potential, then very few is known about the convergence of the corresponding adaptation rule. For example, several results (Tolat, 1990; Erwin et al., 1992; Heskes and Kappen, 1993; Heskes, 1999) have already shown that for a continuous density P, the corresponding vector adaptation rule of the Kohonen Self-Organizing Map (SOM)

3

algorithm (Kohonen, 1982) (Kohonen, 1988) does not correspond to a stochastic gradient descent along a global energy function, and the convergence, although being observed in practice, turns out to be very difficult to prove, not to mention that most of the efforts have been carried out on the Kohonen rule (Cottrell et al., 1994; Cottrell et al., 1998; Benaïm et al., 1998).

All the vector quantization algorithms we study in this paper are variants of the K-means algorithm as we will see in section 5. We know these algorithms converge in practice toward acceptable value of their energy functions whenever they are proved to be associated or not to potentials. However, the theoretical study of their convergence is not available, so they remain largely heuristics. Among all these algorithms, the Neural-Gas (Martinetz and Schulten, 1994) deserves a particular attention. It has been claimed by its authors to be associated to a global potential in general, hence to a converging adaptation rule. We propose a counter-example with a discontinuous pdf P which demonstrates that this claim is not true. This shows that the study of the convergence of all these algorithms is still in its infancy and motivates the present work.

In this paper, we propose a framework which encompasses all these algorithms. We study this framework and we demonstrate that the energy function associated to these algorithms is not a potential in general. We also demonstrate that this energy function belongs to a broad class of functions which includes potential functions as a special case. The energy functions within this class are called "pseudo-potentials". The results we obtain do not depend on the continuity of the probability density function P, and give a first step toward an explanation why all the algorithms shown to belong to this framework succeed, in practice, in minimizing their associated energy function whether they are potentials or not. This framework should open up further avenues for a general study of the convergence properties of all the algorithms it contains at once.

In section 2, we present the framework of this study. In section 3, we define a "pseudo-potential" function, which can be approximated by a series

of potential functions: we define the concept of cellular manifold and this series of potentials. In section 4, we give the main theorem which states that an energy function of that framework is necessarily a pseudo-potential. We consider the K-means to show that pseudo-potentials are not always potentials. We discuss the consequence on the convergence of the corresponding adaptation rule. In section 5, we show that most of the common vector quantization algorithms belong to that framework. At last we conclude in section 6.

2   Framework

We consider $(\mathbb{R}^d, \| \, . \, \|)$ is the euclidean d-dimensional space associated to the euclidean norm. Let D be a non-empty bounded set in $\mathbb{R}^d$. Let δ be the diameter of D and V a topological manifold included in D. Let $\underline{w} = (w_1, ..., w_n)$ be a set of prototypes in D.

For p=1,...,n, the Voronoï cell associated to w_p is usually defined as (Okabe et al., 1992):

$$V_p = \{v \in V \mid \forall \, q = 1, ..., n \ \| \, v - w_p \, \| \leq \| \, v - w_q \, \|\} \tag{3}$$

The set of V_p, p=1,...,n provides a cellular decomposition of V.

For any l, the distance between w_l and v is denoted $d_l = \|w_l - v\|$.

We will show in section 5 that the neighborhood function ψ_p of various algorithms is constructed on the basis of the Heaviside step function of the distances d_l, denoted H such that $H(x < 0) = 0$ and $H(x \geq 0) = 1$. These step functions cause discontinuities of the corresponding energy functions or their derivatives, which appear at the Voronoï cells boundaries. This is the reason why we focused on the following class of neighborhood functions in the definition of our framework:

$$\psi_p(\underline{w}, v) = \phi_p(\{H(d_l^2 - d_m^2)\}_{lm}) \tag{4}$$

where ϕ_p is a bounded function.

We consider any probability density function P such that: $\int_V P(v)dv = 1$. In other words, all the results presented hereafter do not depend on the continuity of P.

3 Cellular manifolds and pseudo-potential

The discontinuities of the neighborhood functions $\psi_p(\underline{w}, v)$ occur onto the boundaries of the Voronoï cells. We shall consider a part of the manifold V called *cellular manifold* (and its complementary part called *tubular manifold*) which does not contain these boundaries to isolate them and to ease their study. This leads to the subsequent definition of pseudo-potential functions.

3.1 The family of cellular manifolds V^η

The cellular manifold is based on the Voronoï cells defined by the set of vectors w_p and which is arbitrarily close to the manifold V in the sense of the Lebesgue measure.

Let η be a number $\ll 1$; we denote $T_p^\eta(\underline{w}) = T_p^\eta$ the open tubular neighborhood, of thickness η, of the boundary of V_p, included in V.

This neighborhood is shown on figure 1 in $\mathbb{R}^2$.

Then for a given $\underline{w} = (w_1, ..., w_n)$ with $w_i \in V$, we define the cellular manifold $V^\eta(\underline{w})$ as the set of vectors of V which are not in the tubular neighborhood T_p^η for all p:

$$V^\eta(\underline{w}) = V \setminus (\bigcup_{p=1}^n T_p^\eta) \tag{5}$$

That means the smaller η, the closer to V the cellular manifold $V^\eta(\underline{w})$ which does not contain the boundaries of the Voronoï cells. In other words, $V \setminus V^\eta(\underline{w})$ "tends" towards the boundaries of the Voronoï cells while η tends towards 0. $V \setminus V^\eta(\underline{w})$ is called *tubular manifold*.

We can then state the following property:

$\mathbb{R}^d$ being provided with the product-measure of Lebesgue, $V^\eta(\underline{w})$ verifies:

$$meas(V \setminus V^\eta(\underline{w})) = O(\eta) \tag{6}$$

And we have in particular:

$$\lim_{\eta \to 0}(meas(V \setminus V^\eta(\underline{w}))) = 0$$

The proof of this property follows:

$$meas(V \setminus V^\eta(\underline{w})) = meas(\bigcup_{p=1}^n (T_p^\eta)) \leq \sum_{p=1}^n meas(T_p^\eta)$$

We have to consider two subcases according to the dimension d:

If d=1 : the boundaries of Voronoï cells are points thus their measure is null. One has in this case: $\sum_{p=1}^n meas(T_p^\eta) \leq (n-1)\eta$, whence the result.

If d>1 : we have $meas(T_p^\eta) \leq \eta \cdot meas_{\mathbb{R}^{d-1}}(boundary(V_p)) + O(\eta^2)$ where the residual term $O(\eta^2)$ is bounded by the following sum: each term of the sum is the product of the measures of the (d-k) cells ($k > 1$) of the polyhedral decomposition of the boundary of V_p by the volumes of the k-balls of radius η (i.e. $\frac{\pi^{\frac{k}{2}}\eta^k}{\Gamma(\frac{k}{2}+1)}$). However D is bounded, therefore all the measures of the boundaries of V_p are finished, whence the result.

3.2 Definition of a pseudo-potential

In general, a potential is defined as a differentiable function of its variables. We define a wider class of functions that we call *pseudo-potentials*, which contains potential functions as a special case. Pseudo-potentials do not verify in general the hypotheses of differentiability at every point but may be approached by a series of potential functions. Thus a potential is a pseudo-potential but the converse is false: a pseudo-potential is not necessarily differentiable everywhere and therefore is not necessarily a potential.

Definition: let Ω be a non empty and bounded set in $\mathbb{R}^d$ and $n \geq 1$ fixed. The function $E_\Omega \colon \Omega^n \to \mathbb{R}$ is called *pseudo-potential* if there exists a family of potential functions $E_\Omega^\eta \colon \Omega^n \to \mathbb{R}, \eta > 0$, such that

$$\lim_{\eta \to 0} \parallel E_\Omega - E_\Omega^\eta \parallel_\infty = 0$$

where $\parallel . \parallel_\infty$ denotes the norm of the uniform convergence.

In our case, we focus on the energy function $E_\Omega(\underline{w})$ defined by (1).

Introducing pseudo-potentials enables all these algorithms to be placed in the same framework (see section 5). In this framework, the neighborhood function belongs to the family defined in (4) and the associated energy function may not be differentiable on the boundaries of the Voronoï cells, hence is possibly not a potential on the whole manifold V.

Which leads us to the main result about the energy function E_V.

4 The energy function E_V is a pseudo-potential

We show that the energy function E_V defined in (1) under the hypotheses of the section 2, may be considered as the limit of a series of differentiable functions over the manifold V, without being itself differentiable over V, *i.e.* E_V is a pseudo-potential.

Theorem: The energy function E_V is a pseudo-potential with $\parallel E_V(\underline{w}) - E_V^\eta(\underline{w}) \parallel_\infty = O(\eta)$.

The first part of the theorem means E_V is not necessarily a potential over V, being not always differentiable on the boundaries of the Voronoï cells.

The second part means that the difference between the energy $E_V(\underline{w})$ and the energy $E_V^\eta(\underline{w})$ both defined on the whole V, is bounded by a value proportional to η, hence as small as wanted. In other words, even if $E_V(\underline{w})$ is not a potential, it is very close to be one.

4.1 Proof of the theorem

To prove that E_V is a pseudo-potential, we consider for $\eta > 0$, the functions E_V^η defined as:

$$\forall \underline{w} \in V^n, \ E_V^\eta : \underline{w} \longrightarrow \int_{V^\eta(\underline{w})} F(\underline{w}, v) dv$$

Then we first show that these functions which are defined on the whole manifold V, are differentiable on $V^\eta(\underline{w})$ (*i.e.* the domain where the integral is carried out). And second, we show that the difference $\| E_V(\underline{w}) - E_V^\eta(\underline{w}) \|_\infty$ equals $O(\eta)$, hence that $\lim_{\eta \to 0} \| E_\Omega - E_\Omega^\eta \|_\infty = 0$, fulfilling the conditions necessary for E_V to be a pseudo-potential.

The proof of the first part of the theorem rests on the behavior of the functions ψ_p. When the current v are far enough from the boundaries of the Voronoï cells, these functions behave like constants, while onto these boundaries they have discontinuities. We need to insure the differentiability according to $\underline{w}$ of the $F(\underline{w}, v)$ functions which depend on the ψ_p functions, and to control the integration domain $V^\eta(\underline{w})$ when $\underline{w}$ varies. This is the purpose of the two propositions which follow, to show that when the variation ζ of $\underline{w}$ remains lower than a given bound, the variations of $F(\underline{w}, v)$ (Proposition 1) and that of the integration domain V^η (Proposition 2) are negligible compared to the norm of ζ.

4.1.1 Invariance of the ψ_p functions

This proposition insures the invariance of the $\psi_p(\underline{w}, v)$ functions for v belonging to $V^\eta(\underline{w})$ and for sufficiently small variations ζ of the prototypes $\underline{w}$.

Proposition 1: For $\zeta = (\zeta_1, ..., \zeta_n) \in (\mathbb{R}^d)^n$ with $\underline{w} + \zeta = (w_1 + \zeta_1, ..., w_n + \zeta_n) \in V^n$, we denote $| \zeta | = \max_{p=1,n} \| \zeta_p \|$, and $d_r^\zeta = \| (w_r + \zeta_r) - v \|$. Thus,

9

we have:

$$\exists \, \nu > 0 \text{ such that } \forall r, p = 1, ..., n \text{ for } \mid \zeta \mid < \nu :$$
$$H((d_r^\zeta)^2 - (d_p^\zeta)^2) = H(d_r^2 - d_p^2), \forall v \in V^\eta(\underline{w}) \qquad (7)$$

Proof :

The proof is based on the existence of a bound denoted ν inside which the invariance of the Heaviside function according to ζ is insured. First, we consider the case where the Heaviside function takes the value 1 and then the case where it is 0:

(i) Considering $H(d_r^2 - d_p^2) = 1$, we must find a condition on ζ for which $(d_r^\zeta)^2 - (d_p^\zeta)^2 > 0$:

For $r \neq p$, we have $d_r^2 - d_p^2 > \eta^2$ and

$$(d_r^\zeta)^2 - (d_p^\zeta)^2 = d_r^2 - d_p^2 + 2\langle w_r - v \mid \zeta_r \rangle - 2\langle w_p - v \mid \zeta_p \rangle + \zeta_r^2 - \zeta_p^2 \qquad (8)$$

where $\langle . \mid . \rangle$ denotes the scalar product. However, for the scalar products, we have:

$$\langle w_r - v \mid \zeta_r \rangle \geq -2\delta \|\zeta_r\|$$

and

$$\langle v - w_p \mid \zeta_p \rangle \geq -2\delta \|\zeta_p\|$$

hence

$$(d_r^\zeta)^2 - (d_p^\zeta)^2 \geq \eta^2 - 4\delta(\|\zeta_r\| + \|\zeta_p\|) + \zeta_r^2 + \zeta_p^2$$

Finally, we have:

$$(d_r^\zeta)^2 - (d_p^\zeta)^2 \geq (\frac{1}{2}\eta^2 - \zeta_p^2 - 4\delta \parallel \zeta_p \parallel) + (\frac{1}{2}\eta^2 + \zeta_r^2 - 4\delta \parallel \zeta_r \parallel)$$
$$= a_1 + a_2$$

(ii) Considering $H(d_r^2 - d_p^2) = 0$, we must find a condition on ζ for which $(d_p^\zeta)^2 - (d_r^\zeta)^2 > 0$:

10

A similar calculation leads to:

$$(d_p^\zeta)^2 - (d_r^\zeta)^2 \geq (\frac{1}{2}\eta^2 + \zeta_p^2 - 4\delta \parallel \zeta_p \parallel) + (\frac{1}{2}\eta^2 - \zeta_r^2 - 4\delta \parallel \zeta_r \parallel)$$
$$= b_1 + b_2$$

The joint study of the polynomials in $\parallel \zeta_p \parallel$ defined by a_1 and b_1 shows that the conditions $a_1 \geq 0$ et $b_1 \geq 0$ are reached for:

$$\parallel \zeta_p \parallel \leq \mu = (4\delta^2 + \frac{\eta^2}{2})^{\frac{1}{2}} - 2\delta$$

Moreover, $a_2 \geq (\frac{1}{2}\eta^2 - 4\delta \parallel \zeta_r \parallel)$ and $b_2 \geq (\frac{1}{2}\eta^2 - 5\delta \parallel \zeta_r \parallel)$

The conditions $a_2 \geq 0$ et $b_2 \geq 0$ are met for $\parallel \zeta_r \parallel \leq \frac{\eta^2}{10\delta}$.

It is enough to take $\nu = \min\{\mu, \frac{\eta^2}{10\delta}\}$.

The proposition 1 means that considering a variation of the norm of $\underline{w}$ vectors lower than ν, ψ_p functions remain the same either within the energy function or within the adaptation rule. As a consequence, the function $F(\underline{w}, v)$ to integrate, which is a combination of ψ_p functions with continuous functions of $\underline{w}$, is continuous and differentiable over $V^\eta(\underline{w})$ according to $\underline{w}$. The nature of P as being continuous or not, does not affect this result because P does not depend on $\underline{w}$.

4.1.2 Variations of the integration domain

To study the variations of the energy function, it is necessary to study the variations of the integration domains.

This proposition insures that the variations of the integration domains $V^\eta(\underline{w})$ and $V \setminus V^\eta(\underline{w})$ remain small with small variations ζ of the prototypes.

Proposition 2: for $\mid \zeta \mid \ll 1$, we have :

(i) $\qquad \mid meas(V \setminus V^\eta(\underline{w} + \zeta)) - meas(V \setminus V^\eta(\underline{w})) \mid = O(\mid \zeta \mid^2);$

(ii) $\qquad \mid meas(V^\eta(\underline{w} + \zeta)) - meas(V^\eta(\underline{w})) \mid = O(\mid \zeta \mid^2).$

Proof : The proof of both equations is obtained by calculating the measure of the tubular neighborhood $T_p^\eta(\underline{w})$ of the Voronoï cells. The projections of these neighborhoods onto the coordinate axes verify :
$$| \, meas(T_p^\eta(\underline{w} + \zeta) - meas(T_p^\eta(\underline{w}) \, | = O(| \, \zeta \, |^2)$$
In just the same way as in property (6), we can write:

$$| \, meas(V \setminus V^\eta(\underline{w} + \zeta)) - meas(V \setminus V^\eta(\underline{w})) \, |$$
$$= \sum_{p=1}^{n} (meas(T_p^\eta(\underline{w} + \zeta)) - meas(T_p^\eta(\underline{w}))) = O(| \, \zeta \, |^2)$$

validating item (i) of the proposition.

Item (ii) is validated observing that:

$$meas(V^\eta(\underline{w})) + meas(V \setminus V^\eta(\underline{w})) = meas(V^\eta(\underline{w} + \zeta)) + meas(V \setminus V^\eta(\underline{w} + \zeta))$$

Hence, for small variations ζ of $\underline{w}$, the variations of the integration domains remain negligible compared to ζ.

4.1.3   Last step for the proof

We show that small variations ζ of $\underline{w}$ (*i.e.* less than the bound ν determined in Proposition 1) lead to a small variation of E_V^η which breaks down in a linear application plus other terms of higher order, hence that E_V^η is a potential for all $\underline{w} \in V$ and all $v \in V^\eta(\underline{w})$. Then we show that $E_V(\underline{w}) - E_V^\eta(\underline{w}) = O(\eta)$, $\forall \underline{w} \in V^n$ hence that $\lim_{\eta \to 0} \| \, E_V - E_V^\eta \, \|_\infty = 0$ demonstrating that E_V is a pseudo-potential, and at the same time that $\| \, E_V(\underline{w}) - E_V^\eta(\underline{w}) \, \|_\infty = O(\eta)$.

The difference $E_V^\eta(\underline{w} + \zeta) - E_V^\eta(\underline{w})$ may be written as:

$$E_V^\eta(\underline{w} + \zeta) - E_V^\eta(\underline{w})$$
$$= [E_V^\eta(\underline{w} + \zeta) - \int_{V^\eta(\underline{w})} F(\underline{w} + \zeta, v) dv] \; + \; [\int_{V^\eta(\underline{w})} F(\underline{w} + \zeta, v) dv - E_V^\eta(\underline{w})]$$
$$= \qquad\qquad [part \, 1] \qquad\qquad + \qquad\qquad [part \, 2]$$

The function $F(\underline{w}, v)$ being bounded on $V^n \times V$, Proposition 2 shows that [part 1] is $O(| \zeta |^2)$.

Proposition 1 leads to:

$$[part\,2] = - \langle \zeta_j \mid \int_{V^\eta(\underline{w})} P(v)\psi_j(\underline{w}, v)(v - w_j)dv \rangle$$
$$+ \frac{\| \zeta_j \|^2}{2} \int_{V^\eta(\underline{w})} P(v)\psi_j(\underline{w}, v)(v - w_j)^2 dv$$

The first term is of the form $L(\zeta)$, where L is a linear application and the second term is of higher order. Thus, we can write: $E_V^\eta(\underline{w} + \zeta) - E_V^\eta(\underline{w}) = L(\zeta) + O(|\zeta|^2)$, which means that $E_V^\eta(\underline{w})$ is differentiable for all $\underline{w} \in V$ and $v \in V^\eta(\underline{w})$.

Moreover, because the function $F(\underline{w}, v)$ is bounded on $V^n \times V$, we can write:

$$E_V(\underline{w}) - E_V^\eta(\underline{w}) = \int_{V \backslash V^\eta(\underline{w})} F(\underline{w}, v)dv$$
$$\leq \sup_{(\underline{w}, v) \in V^n \times V} F(\underline{w}, v).meas(V \backslash V^\eta)$$

whence, with the property (6) : $E_V(\underline{w}) - E_V^\eta(\underline{w}) = O(\eta)$, for all $\underline{w} \in V^n$.

The energy function E_V is then a pseudo-potential.

4.2   A pseudo-potential is not a potential in general

As far as the neighborhood functions ψ_p are of the form given in (4), the theorem ensures that the corresponding energy function is a pseudo-potential over the entire domain V, and at least a potential over $V^\eta(\underline{w})$. We also know that it exists energy functions in this framework (*i.e.* pseudo-potentials) which are potential over the entire domain V for continuous pdf P, *e.g.* the energy function of the K-means (see section 5.1) (Pagès, 1993). However, it remains to prove the existence of energy functions in this framework which are not potential over the entire domain V, *i.e.* the existence of pseudo-potentials which are not potentials.

Here we show that the energy function of the K-means does not correspond to a global potential for a particular discontinuous pdf P, hence is not a potential in general for all P.

In order to simplify the calculi, we consider only $n = 2$ prototypes $\underline{w} = (w_1, w_2)$ in a 1-dimensional space ($d = 1$). It is straightforward, though messy, to extend this result to higher dimensions and greater number of prototypes .

The neighborhood function of the K-means, associated to each prototype is defined as:

$$\psi_1(\underline{w}, v) = H(d_2^2 - d_1^2) = H(\| w_2 - v \|^2 - \| w_1 - v \|^2)$$
$$\psi_2(\underline{w}, v) = H(d_1^2 - d_2^2) = H(\| w_1 - v \|^2 - \| w_2 - v \|^2)$$

that we shorten $\psi_1(v)$ and $\psi_2(v)$ respectively. We have $\psi_i(v \in V_i) = 1$ and $\psi_i(v \notin V_i) = 0$.

These functions are part of the family given by equation (4), hence the corresponding energy function E_V is a pseudo-potential and E_V^η is a potential. Observing that $E_V = (E_V - E_V^\eta) + E_V^\eta$, we are going to show that E_V is not a potential by showing that $(E_V - E_V^\eta)$ is not a potential. The function $(E_V - E_V^\eta)$ is not a potential wrt $\underline{w}$ iff the variation of this function wrt some variation ζ of $\underline{w}$ cannot be written as $L(\zeta) + O(\zeta^2)$, $i.e.$ as a linear form of ζ.

Let $(w_1 + w_2)/2$ be the origin 0 of the directed line $(w_1 w_2)$. For a small positive variation $\zeta = \zeta_1$ of w_1 (see figure 2), with $0 < \zeta_1 < \eta$, we have:

$$
\begin{aligned}
\Delta(\zeta) = \ & (E_V - E_V^\eta)(\underline{w} + \zeta) - (E_V - E_V^\eta)(\underline{w}) \\
= \ & \int_{V \backslash V^\eta(\underline{w}+\zeta)} F(\underline{w} + \zeta, v) dv \ - \int_{V \backslash V^\eta(\underline{w})} F(\underline{w}, v) dv \\
= \ & \frac{1}{2} \int_{V \backslash V^\eta(\underline{w}+\zeta)} P(v) \left[\psi_1^{\zeta_1}(v)\delta_1^{\zeta_1}(v) + \psi_2^{\zeta_1}(v)\delta_2(v) \right] dv \\
& - \frac{1}{2} \int_{V \backslash V^\eta(\underline{w})} P(v) \left[\psi_1(v)\delta_1(v) + \psi_2(v)\delta_2(v) \right] dv
\end{aligned}
\tag{9}
$$

14

where

$$\delta_1(v) = d_1^2 = \|w_1 - v\|^2,$$
$$\delta_1^{\zeta_1}(v) = \|w_1 + \zeta_1 - v\|^2,$$
$$\delta_2(v) = d_2^2 = \|w_2 - v\|^2,$$

and

$$\psi_1^{\zeta_1}(v) = H(\delta_2(v) - \delta_1^{\zeta_1}(v))$$
$$\psi_2^{\zeta_1}(v) = H(\delta_1^{\zeta_1}(v) - \delta_2(v)).$$

The domains $V \setminus V^\eta(\underline{w} + \zeta)$ and $V \setminus V^\eta(\underline{w})$ are defined on the figure 2 and given below:

$$\begin{cases} V \setminus V^\eta(\underline{w} + \zeta) &= [P_2, 0] \cup [0, P_3] \cup [P_3, P_4] \cup [P_4, P_5] \\ V \setminus V^\eta(\underline{w}) &= [P_1, P_2] \cup [P_2, 0] \cup [0, P_3] \cup [P_3, P_4] \end{cases}$$

where

$$P_1 = -\eta/2$$
$$P_2 = -\eta/2 + \zeta_1/2$$
$$P_3 = \zeta_1/2$$
$$P_4 = \eta/2$$
$$P_5 = \eta/2 + \zeta_1/2$$

and $0 < \zeta_1 < \eta$ leads to $P_1 < P_2 < 0 < P_3 < P_4 < P_5$.

Let us consider a particular uniform density $P(v)$ defined as:

$$P(v) = \begin{cases} p = \frac{1}{\beta - \lambda} \text{ if } v \in [\lambda, \beta] \text{ with } \lambda \in [0, P_4] \text{ and } \beta \gg \eta \\ 0 \text{ else} \end{cases} \tag{10}$$

Notice that P is a discontinuous pdf at λ and β. We have also $0 < \zeta_1 < \eta << \beta$ hence $P_5 << \beta$. Then, for such a density and from (9) we get:

$$\Delta(\zeta) = \int_{P_2}^{P_5} F(\underline{w} + \zeta, v)dv - \int_{P_1}^{P_4} F(\underline{w}, v)dv$$

$$= \begin{cases} \dfrac{p}{2} \displaystyle\int_{\lambda}^{P_3} (\delta_1^{\zeta_1}(v) - \delta_2(v))dv + \dfrac{p}{2} \int_{P_4}^{P_5} \delta_2(v)dv \text{ if } \zeta_1 > 2\lambda \quad (i.e. \ \lambda < P_3) \\[20pt] \dfrac{p}{2} \displaystyle\int_{P_4}^{P_5} \delta_2(v)dv \text{ if } \zeta_1 \leq 2\lambda \quad (i.e. \ \lambda \geq P_3) \end{cases}$$

$$(11)$$

Developping equation (11) leads to:

$$\Delta(\zeta) = \begin{cases} \dfrac{p}{6} \left[(\lambda - w_2)^3 - (\lambda - w_1)^3 + w_2^3 - w_1^3 \right] \\[8pt] \qquad + \dfrac{p}{4} \left[2(w_1 - \lambda)^2 + (w_2 - \dfrac{\eta}{2})^2 - (w_1^2 + w_2^2) \right] \zeta_1 + o(\zeta_1^2) \text{ if } \zeta_1 > 2\lambda \\[16pt] \dfrac{p}{4}(w_2 - \dfrac{\eta}{2})^2 \zeta_1 + o(\zeta_1^2) \text{ if } \zeta_1 \leq 2\lambda \end{cases}$$

$$= \begin{cases} L_1(\zeta_1) + o(\zeta_1^2) \text{ if } \zeta_1 > 2\lambda \\[12pt] L_2(\zeta_1) + o(\zeta_1^2) \text{ if } \zeta_1 \leq 2\lambda \end{cases}$$

$$(12)$$

with $L_1 \neq L_2$. Therefore $\Delta(\zeta)$ is not a linear form of ζ which proves the non differentiability of $(E_V - E_V^\eta)$. Hence $(E_V - E_V^\eta)$ is not a potential and so, the energy function E_V is a pseudo-potential but not a potential in general.

4.3   What is important about this result

Consequence 1: The family of pseudo-potential functions includes potential

functions as a special case and it exists pseudo-potential functions which are not potentials.

Consequence 2: The previous example shows that a necessary condition for the energy function of the K-means to be a potential is that given any number n, position $\underline{w}$ and dimension d of the prototypes, the boundary of Voronoï cells never crosses any discontinuity of the pdf P whatever the value of the variation ζ. A sufficient condition for this to hold is P being continuous.

Consequence 3: The energy function of the K-means is not a potential at least for some discontinuous pdf P. This complements the result of Pagès (Pagès, 1993) stating this energy function is a potential for continuous P. Moreover, the algorithms presented in section 5, because they reduce to the K-means for specific values of their parameters, also share this property that prevent them from being potentials in general for all P and all setting of their parameters. In particular, this result holds for the Neural-Gas (Martinetz et al., 1993) despite the claim of its authors: the Neural-Gas is not a global potential at least for discontinuous P and width σ of the neighborhood function set to 0. This casts some doubt on the validity of their proof which do not specify any restriction on P and σ. As a consequence, the convergence of the associated adaptation rule in general still to be proved.

4.4   Consequence of the theorem concerning the convergence

The consequence of the theorem is promising concerning the eventual convergence of the adaptation rules associated to pseudo-potentials toward a local minimum. Indeed, from a mathematical point of view, talking about "derivatives" of the energy function $E_V(\underline{w})$ onto V according to some w_p does not make any sense because of the discontinuities of this function onto the Voronoï boundaries. The only possibility is to measure the variations of this function according to a small movement of the prototypes. We have already shown that the volume of the tubular neighborhood of the Voronoï

boundaries is in $O(\eta)$ (Equation (6)) so is bounded. Now this theorem shows that the variations of the energy function according to a bounded movement ζ of the prototypes, are also bounded.

Indeed, the theorem allows to write that $E_V(\underline{w}+\zeta)-E_V^\eta(\underline{w}+\zeta)=O(\eta)$ and $E_V(\underline{w})-E_V^\eta(\underline{w})=O(\eta)$ so $[E_V(\underline{w}+\zeta)-E_V(\underline{w})]-[E_V^\eta(\underline{w}+\zeta)-E_V^\eta(\underline{w})]=O(\eta)$. And E_V^η being a potential, then $E_V^\eta(\underline{w}+\zeta)-E_V^\eta(\underline{w})$ is bounded as a linear form of ζ which is bounded. Therefore $E_V(\underline{w}+\zeta)-E_V(\underline{w})=\Delta_V(\zeta)$ is also bounded although E_V is not a potential on V. As a consequence, the effects of the variation $\Delta_V(\zeta)$ of the energy function E_V according to ζ, on the dynamic of the prototypes remains negligible on average even for some data falling onto the Voronoï boundaries. In other words, the existence of a pseudo-potential for lack of a potential would be sufficient to ensure the convergence of the associated adaptation rule, although a rigorous proof is still to be carried out. The work of Bottou (Bottou, 1991) gives also insights in this direction but following a different way.

5   Consequence for existing rules

In this section, we show that the neighborhood function of a large number of algorithms can be written in the form of the equation (4), *i.e.* as a combination of Heaviside step functions of a difference of squared distances d_l. This demonstrates that the corresponding adaptation rule is associated to an energy function which is not necessarily a potential but at least a pseudo-potential.

5.1   K-means vector quantizer

The K-means vector quantizer (Mac-Queen, 1967) is the iterative version of the Linde-Buzzo-Gray batch learning technique for vector quantization (Linde et al., 1980). It consists in presenting one datum v at a time, then selecting the closest prototype w_{p^*} to it and moving it toward v. The corre-

sponding neighborhood function can be written as:

$$\psi_p^{[\text{K-means}]}(\underline{w}, v) = A_p(\underline{w}, v) = K_p(\underline{w}, v) \prod_{l=1}^{n} (K_l(\underline{w}, v)(H(l-p)-1)+1)$$

$$= \begin{cases} 1 \text{ if } v \in V_p \text{ and } p = \min(\{i \in (1, \ldots, n) | K_i(\underline{w}, v) = 1\}) \\ 0 \text{ else} \end{cases}$$

(13)

where the function K_p is an indicator function of the Voronoï cell V_p of w_p, defined as:

$$K_p(\underline{w}, v) = \prod_{k=1}^{n} H(d_k^2 - d_p^2)$$

$$= \begin{cases} 1 \text{ if } v \in V_p \\ 0 \text{ else} \end{cases}$$

(14)

The function A_p performs an additional sort over the index of the closest prototypes (the "winners") for which K_p is equal to 1, *i.e.* all the prototypes which are the closest to v. This is the algebraic writing of the algorithms which choose only one prototype among all the closest one in case of equality. Here, the choice is carried out according to the lowest index, it could be the highest one, or a random choice among the indices of all the winners. In case where all the winners are moved, then $\psi_p^{[\text{K-means}]}(\underline{w}, v) = K_p(\underline{w}, v)$ should be considered.

The K-means algorithm corresponds to a Hard Competitive Learning technique (Ahalt et al., 1990), where only the closest prototype to the datum is adapted at a time. To escape from local optima of the energy function, it has been improved by defining a neighborhood function which enables the winner to be adapted and also some of its neighbors. All the following algorithms belong to that class of Soft-Competitive Learning techniques (Ahalt et al., 1990), and each one defines its particular neighborhood function.

5.2 Self-Organizing Maps and other graph-based neighborhoods

The Self-Organizing Map (SOM) proposed by Kohonen (Kohonen, 1982) defines a set of connections between the prototypes, which corresponds to a graph G with a particular topology (*e.g.* a regular 2-dimensional grid). The winner being determined according to the datum v, the neighborhood function consists in weighting the adaptation step of the prototypes according to their closeness to the winner on the graph G.

The corresponding neighborhood function may be written as:

$$\psi_p^{[\text{SOM}]}(\underline{w}, v) = \sum_{q=1}^{n} A_q(\underline{w}, v) h_\sigma(D_{qp}(G)) \tag{15}$$

where h_σ is a non-increasing positive function with a tunable width σ (*e.g.* $h_\sigma(u) = e^{-\frac{u}{\sigma}}$) and $D_{ab}(G)$ is the distance between w_a and w_b in terms of the lowest number of edges separating them within the graph G.

Several other algorithms essentially differ from SOM by the fact they use a graph whose topology is not defined *a priori* but thanks to the data and the prototypes positions in the data space. This is the case in the Growing Neural-Gas (GNG) of (Fritzke, 1995b), where G is the Induced Delaunay Triangulation (IDT) (Martinetz and Schulten, 1994), in (Kangas et al., 1990) with Minimum Spanning Trees (MST), in (Mou and Yeung, 1994) with Gabriel Graphs, in the Growing Cell Structure (GCS) of (Fritzke, 1994) with a set of simplices with fixed dimension, and in the growing versions of SOM (GSOM) of (Fritzke, 1995a) and (Villmann and Bauer, 1997) with an adaptive grid structure. As far as n remains constant and the graph G remains the same, the neighborhood function of all these models is identical to the one of the SOM written above, and belongs to the framework we consider in this paper.

5.3 Neural-Gas

In the Neural-Gas (Martinetz et al., 1993), the prototypes are ranked in increasing order of their distance to the datum v. This rank is used to weight the adaptation rule of the prototypes. Martinetz et. al. give the corresponding neighborhood function :

$$\psi_p^{[\text{Neural-Gas}]}(\underline{w}, v) = h_\sigma(k_p(\underline{w}, v)) \text{ with } k_p(\underline{w}, v) = \sum_{q=1}^{n} \Upsilon(d_p^2 - d_q^2) \qquad (16)$$

where $\Upsilon(u) = 1 - H(-u)$, $\forall u$. The function k_p is the rank of the prototype w_p such that $k_p(\underline{w}, v) = j-1$ iff p is the jth closest vector to v (several prototypes may have the same rank). Note that the Neural-Gas could be included into the previous family of adaptive graph-based neighborhoods considering G as the graph which connects the n-nearest-neighbors of v among $\underline{w}$, in a chain where the ith nearest neighbor is connected to the $(i-1)^{th}$ ($\forall i > 1$) and the $(i+1)^{th}$ ($\forall i < n$).

5.4 Recruiting rules

One of us proposed the "Recruiting" Neural-Gas (Aupetit, 2000) as a way to cope with function approximation tasks using vector quantizers. A recruiting factor is added to the Neural-Gas adaptation rule. Such a factor is associated to each prototype and the winner imposes its own on the others. This tends to gather the prototypes around the one which has the highest recruiting factor. Then setting this factor proportional to the local output error approximating a function, enables more prototypes to be grouped together in areas of the input space where the corresponding output function is more difficult to approximate. This tends to decrease the global approximation error.

The corresponding neighborhood function may be written as:

$$\psi_p^{[\text{RecruitingNG}]}(\underline{w}, v) = h_\sigma(k_p(\underline{w}, v)) \sum_{q=1}^{n} A_q(\underline{w}, v)\epsilon_q \qquad (17)$$

21

where $\forall q, \epsilon_q \in [0, 1]$.Taking $\epsilon_q = \epsilon_p = 1$, $\forall p, q$ leads to the usual Neural-Gas.

Göppert and Rosenstiel (Göppert and Rosenstiel, 2000) proposed a similar approach with a SOM for which each prototype defines its own neighborhood's width σ_q tuned according to the local output approximation error. The corresponding neighborhood function may be written as:

$$\psi_p^{[\text{RecruitingSOM}]}(\underline{w}, v) = \sum_{q=1}^{n} A_q(\underline{w}, v) h_{\sigma_q}(D_{qp}(G)) \tag{18}$$

where $\forall q, \sigma_q \in [0, 1]$. Taking $\sigma_q = \sigma_p = \sigma$, $\forall p, q$ leads to the usual SOM.

In both approaches, as far as ϵ_q and σ_q remain independent of v and $\underline{w}$, the corresponding neighborhood function belongs to the framework we consider in this paper.

5.5 Concerning the algorithms with adaptive structures

We have shown that many vector quantization algorithms belong to our framework. However, considering dynamic approaches such as the algorithms which adapt either the number n of prototypes (GCS, GNG, GSOM), the graph of their neighborhood structure (GNG, GSOM), or the recruiting factor (RecruitingNG, RecruitingSOM), according to either the number of iterations, the position of the prototypes or the output approximation error, it is still difficult to define a framework taking into account these structural changes. That is why we considered these dynamic parameters to be fixed in such cases.

5.6 About some algorithms which do not belong to the present framework

We shall notice that the modified Self-Organizing Map proposed by Heskes and Kappen (Heskes and Kappen, 1993; Heskes, 1999) does not belong to

the present framework. Indeed the Heaviside step functions involved in the corresponding ψ_p neighborhood functions are not applied to a pair of square distances d_l directly, but to a sum over $\underline{w}$ of weighted square distances d_l. This prevents ψ_p from belonging to the family we consider in equation (4). However it seems possible to enlarge our framework in order to encompass the neighborhood function they proposed.

The γ-Observable Neighborhood has been proposed by one of us (Aupetit et al., 2002) as a neighborhood that decreases the number of iterations needed for the adaptation rule to converge toward an optimum of the energy function. The corresponding neighborhood function does not belong to the present framework. However, we have already defined an extension of this framework which encompasses this adaptation rule and thus which allows to demonstrate that the energy function associated to the γ-Observable Neighbors is also a pseudo-potential. This work has not been published yet.

6   Conclusion

In vector quantization, we propose a framework which ensures the existence of a family of potential functions (*i.e.* differentiable functions) which converges uniformly to the energy function that we call in such a case a "pseudo-potential". We demonstrate that a pseudo-potential is not necessarily differentiable everywhere, hence it is not always a potential. As a consequence, the corresponding adaptation rule does not necessarily perform a stochastic gradient descent along this energy function.

We also show how a large number of existing vector quantization algorithms belong to this framework, hence even if they are not associated to potentials, they are at least associated to pseudo-potentials. This framework allows to study at once the convergence of all these algorithms. At that point, although the pseudo-potentials are not necessarily potentials, a consequence of the theorem shows that the variations of the pseudo-potentials on the boundaries of the Voronoï cells remain bounded, so they have a negligi-

ble effect on the dynamic of the prototypes on average. This is a promising preliminary result about the convergence of the corresponding adaptation rules.

If the convergence of the adaptation rules associated to pseudo-potentials were demonstrated then the present framework would constitute an *a posteriori* justification of a large family of adaptation rules considered up to now as heuristic. Moreover, this framework makes possible the design of new adaptation rules respecting the hypotheses which ensure the existence of the corresponding pseudo-potential.

The results of this paper suggest two avenues for future research:

- investigating the convergence properties of the adaptation rules associated to pseudo-potentials in general.

- extending this framework to a wider class of neighborhood functions.

By introducing pseudo-potentials, we add a new concrete framework on the wasteland of non-potentials. Within this framework, the consequence of the theorem makes us hopeful to build new theorems which could insure at once the convergence with respect to a specific norm, of a large number of existing vector quantization algorithms which are not associated to potentials but at least to pseudo-potentials.

References

Ahalt, A., Krishnamurthy, A. K., Chen, P., and Melton, D. E. (1990). Competitive learning algorithms for vector quantization. *Neural Networks*, 3:277–290.

Albert, A. E. and Gardner, L. A. (1967). *Stochastic approximation and nonlinear regression*. MIT Press Cambridge, MA.

Aupetit, M. (2000). A "recruiting neural-gas" for function approximation. *Proc. of IEEE Int. Joint Conf. on Neural Networks (IJCNN'2000),Como, Italy*.

Aupetit, M., Couturier, P., and Massotte, P. (2002). γ-observable neighbours for vector quantization. *In Special Issue on Self-Organizing Maps, Neural Networks*, 15:1017–1027.

Benaïm, M., Fort, J. C., and Pagès, G. (1998). Convergence of the one-dimensional kohonen algorithm. *Adv. Appl. Prob.*, 30:850–869.

Bottou, L. (1991). Stochastic gradient learning in neural networks. *Proceedings of Neuro Nimes 91, Nimes, France*.

Cottrell, M., Fort, J. C., and Pagès, G. (1994). Two or three things that we know about the kohonen algorithm. *Proc. of European Symposium on Artificial Neural Networks (ESANN'1994), Brussels, Belgium*, pages 271–276.

Cottrell, M., Fort, J. C., and Pagès, G. (1998). Theoretical aspects of the som algorithm. *Neurocomputing*, 21:119–138.

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

Fritzke, B. (1994). Growing cell structures - a self-organizing network for unsupervised and supervised learning. *Neural Networks*, 7(9):1441–1460.

Fritzke, B. (1995a). Growing grid - a self-organizing network with constant neighborhood range and adaptation strength. *Neural Processing Letters*, 2(5):9–13.

Fritzke, B. (1995b). A growing neural-gas network learns topologies. *In G. Tesauro, D.S. Touretzky, T.K. Leen (Eds.), Advances in neural information processing systems*, 7.

Göppert, J. and Rosenstiel, W. (2000). Varying cooperation in som for improved function approximation. *Proc. of IEEE Int. Conf. on Neural Networks (ICNN'96), Piscataway, NJ*, 1:1–6.

Gray, R. M. and Neuhoff, D. L. (1998). Quantization. *IEEE Transactions on Information Theory*, 44:2325–2384.

Heskes, T. (1999). Energy functions for self-organizing maps. *In Kohonen Maps, E. Oja and S. Kaski (Eds.)*, pages 303–315.

Heskes, T. and Kappen, B. (1993). Error potentials for self-organization. *Proc. of IEEE Int. Conf. on Neural Networks (ICNN'93), New-York*, 3:1219–1223.

Kangas, J., Kohonen, T., and Laaksonen, J. (1990). Variants of self-organizing maps. *IEEE Trans. on Neural Networks*, 1(1):93–99.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics, Springer-Verlag*, 43:59–69.

Kohonen, T. (1988). *Self-Organization and Associative Memory Formation*. Springer-Verlag.

Kohonen, T. (1991). Self-organizing maps: optimization approaches. *Artificial Neural Networks, Elsevier Science Publishers*, 2:981–990.

Linde, Y., Buzzo, A., and Gray, R. M. (1980). An algorithm for vector quantizer design. *IEEE Trans. Commun.*, COM 28:84–95.

Mac-Queen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proc. of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Univ. of California Press, Berkeley*, 1:281–297.

Martinetz, T. M., Berkovitch, S. G., and Schulten, K. J. (1993). "neural-gas" network for vector quantization and its application to time-series prediction. *IEEE Trans. on Neural Networks*, 4(4):558–569.

Martinetz, T. M. and Schulten, K. J. (1994). Topology representing networks. *Neural Networks, Elsevier London*, 7:507–522.

Mou, K. L. and Yeung, D. Y. (1994). Gabriel networks: Self-organizing neural networks for adaptive vector quantization. *Proc. of the Int. Symp. on Speech, Image Processing and Neural Networks, Hong Kong*, 2:658–661.

Okabe, A., Boots, B., and Sugihara, K. (1992). *Spatial tessellations: concepts and applications of Voronoï diagrams*. John Wiley, Chichester.

Pagès, G. (1993). Voronoï tesselation, space quantization algorithms and numerical integration. *Proc. of European Symposium on Artificial Neural Networks (ESANN'1993), Brussels, Belgium*, pages 221–228.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *Ann. Math. Statist.*, 22:400–407.

Tolat, V. V. (1990). An analysis of kohonen's self-organizing maps using a system of energy functions. *Biol. Cybern.*, 64:155–164.

Villmann, T. and Bauer, H. U. (1997). The gsom-algorithm for growing hypercubical output spaces in self-organizing maps. *Proc. of the Workshop on Self-Organizing Maps (WSOM'97), Helsinky University of Technology, Finland*.
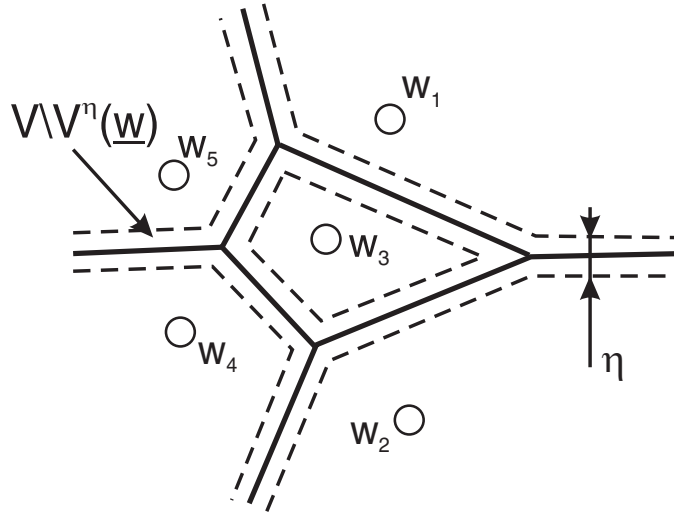
Figure 1: **Cellular and tubular manifolds**. The circles are the prototypes $\underline{w}$. We define tubular manifolds $V \setminus V^\eta(\underline{w})$ of thickness η (dotted lines) which contain the boundaries of the Voronoï cells (plain lines), and cellular manifolds denoted $V^\eta(\underline{w})$ complementary to the tubular manifolds.
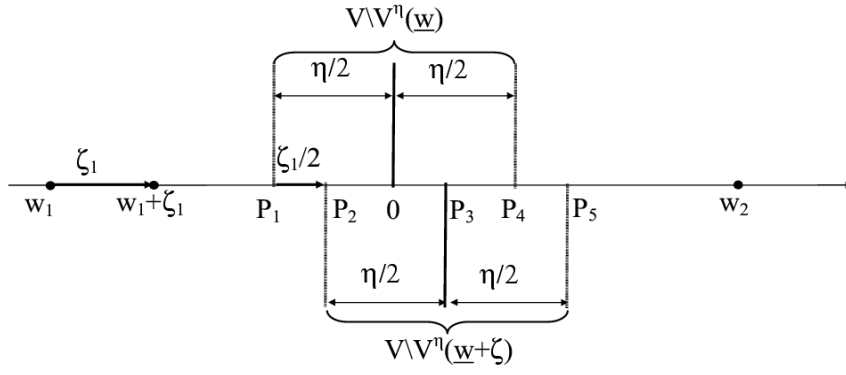
Figure 2: **Variation of the integration domain with** ζ_1 . The point O is the Voronoï boundary between w_1 and w_2. The point P_3 is the Voronoï boundary between $w_1 + \zeta_1$ and w_2.