ESANN'1996 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 24-25-26 April 1996, D-Facto public., ISBN 2-9600049-6-5, pp. 97-102

On the critical points of the 1-dimensional Competitive Learning Vector Quantization Algorithm

Damien LAMBERTON
Univ. de Marne-la-Vallée (France)

Gilles PAGÈS
Labo. de Proba. Paris 6, URA 224 & Univ. Paris 12 (France)

Abstract

In this contribution is established a specific property of the Competitive Learning Vector Quantization algorithm (also known as the Kohonen algorithm with 0 neighbor): in the 1-dimensional setting, that is when the examples ω^t to be coded are scalar with distribution μ, uniqueness of the equilibrium point is established under ln-concavity assumptions on the density f of the distribution μ. The proof relies on the celebrated (finite-dimensional) Mountain pass Lemma. A counter-example is exhibited when f does not satisfy this assumption.

Among all the adaptative algorithms for quantization the so-called *Competitive Learning Vector Quantization* algorithm (or simply Vector Quantization algorithm or Kohonen algorithm with 0 neighbor) can be seen as the ancestor of the myriad of adaptative quantifiers devised ever since (see [8]). The striking simplicity of its implementation, along with its quite satisfactory efficiency makes it one of the most used algorithm for quantization.

Among many interesting features, the VQ algorithm has the property to derive from a potential that is to be the stochastic gradient descent related to a differential function, namely

$$E_n^\mu(x_1, \cdots, x_n) := \int_{\mathbb{R}^d} \min_{1\leq i\leq n} \|x_i - u\|^2 \mu(du)$$

where $(x_1, \cdots, x_n) \in (\mathbb{R}^d)^n$ is a generic n-prototype, $\|\,.\,\|$ denotes the Euclidean norm and μ is the probability distribution of the $\mathbb{R}^d$-valued vector examples $\omega^1, \cdots, \omega^t, \cdots$ to be coded. The ω^t's are also assumed to be statistically independent. The distribution μ is supposed to have a second moment that is $\int_{\mathbb{R}^d} \|u\|^2 \mu(du)$ so that E_n^μ is finite. E_n^μ is called the *distortion*.

ESANN'1996 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 24-25-26 April 1996, D-Facto public., ISBN 2-9600049-6-5, pp. 97-102

The prototype vector(s) that best "sum(s) up" the distribution μ of the "examples" are those that achieve $\min_{(\mathbf{R}^d)^n} E^\mu$. The existence of such vectors was established in a quite general setting *e.g.* in [6] (or [5]). Furthermore such an "optimal" vector turns out to be an excellent n-tuple for numerical integration (see [6] or [5]). However the only way to be sure that the related VQ algorithm will converge toward such an optimal quantifier is to prove that all its possible limit points, that is the zeros of ∇E_n^μ, achieve the minimum value of E_n^μ. As a matter of fact the general theory of stochastic approximation shows that the VQ algorithm in its regular form cannot avoid getting "trapped" into some parasitic local minima (while metastable saddle points or local maximum are *almost surely* be avoided, see [4] or [2]).

Anyway as soon as the example dimension d is greater than 1 one cannot expect all critical points to be minimizers. Thus if a distribution μ has independent marginals μ^1 and μ^2 any n-grid vector $x^{*,(n_1)} \otimes x^{*,(n_2)}$ made up with a zero $x^{*,(n_1)}$ of $\nabla E_{n_1}^{\mu^1}$ and a zero $x^{*,(n_2)}$ of $\nabla E_{n_2}^{\mu^2}$ is a zero of ∇E_n^μ while any simulation shows that, usually, such grids are saddle points. Practitioners sum up this property by the heuristic rule "One d-dim vector quantifier works better than d parallel scalar quantifiers".

In one dimension, the true uniqueness of the zero of ∇E_n^μ is hopeless too since the distortion is a symmetrical function. However the question of *geometrical* uniqueness can reasonably be asked, for example by restricting our investigations to the n-tuples $(x_1, \cdots, x_n)$ with non decreasing components. It is shown in the paragraph below that a positive answer can be provided under some reasonable assumptions on the density f of the distribution μ. On the other hand, a counter-example to uniqueness is provided when f does not fulfill these requirements.

1 The main result

Theorem 1 *Assume that the ω^t-distribution μ has a density f and set $m_\mu := \inf\{u \,/\, f(u) \neq 0\} \in [-\infty, +\infty)$, $M_\mu := \sup\{u \,/\, f(u) \neq 0\} \in (-\infty, +\infty]$. If f satisfies*

$$(\mathcal{L}) \equiv \begin{cases} (i) & f \text{ is continuous and } f > 0 \text{ on } (m_\mu, M_\mu), \\ (ii) & \begin{cases} \bullet\ \ln(f) \text{ is strictly concave} \\ or \\ \bullet\ \ln(f) \text{ is concave and } f(m_\mu+) + f(M_\mu-) > 0, \end{cases} \end{cases} \tag{1}$$

then, $\nabla E_n^{2,\mu}$ admits a unique zero on $F_n^{\mu,+} := \{x \in (m_\mu, M_\mu)^n \,/\, x_1 < \cdots < x_i < \cdots < x_n\}$ which is a global minimum of E_n^μ.

A rather unexpected (and exciting) fact is that the proof of this theorem relies on a variant of the celebrated (finite-dimensional) Mountain pass Lemma that says

Theorem 2 *(see [7]) If $L : \mathbb{R}^d \longrightarrow \mathbb{R}_+$ (L for "landscape" when $d=2$) is a continuously differentiable function satisfying $\lim_{\|x\|\to+\infty} L(x)=+\infty$ and if two distinct zeros of ∇L are (strict) local minima then ∇L has a third zero which can be in no case a local minimum.*

Proof of theorem 1: Step 1: First E_n^μ is continuously differentiable on $F_n^{\mu,+}$ and $\nabla E_n^\mu := \left(\frac{\partial E_n^\mu}{\partial x_i}\right)_{1\le i\le n}$ admits a straightforward continuous extension on $\overline{F}_n^{\mu,+}$ given by

$$\forall i \in \{1,\cdots,n\},\ \frac{\partial E_n^\mu}{\partial x_i} = 2\int_{\tilde{x}_i}^{\tilde{x}_{i+1}} (x_i - u) f(u)du, \quad \text{where } \tilde{x}_i := \frac{x_i + x_{i-1}}{2},\ 2\le i\le n,$$

$\tilde{x}_1 := m_\mu$, $\tilde{x}_{n+1} := M_\mu$. The set $F_n^{\mu,+}$ being convex, one easily shows that, for every $\varepsilon \in (0,1/2)$, $Id - \varepsilon\nabla E_n^\mu$ leaves $F_n^{\mu,+}$ stable. Hence, its continuous extension, still denoted ∇E_n^μ, leaves the closure $\overline{F}_n^{\mu,+}$ stable too.

Furthermore, it has been shown in [3] (or [1]) that if the density f is compactly supported, the above ln-concavity assumption implies that the zero set of ∇E_n^μ satisfies $\{\nabla E_n^{2,\mu} = 0\} \subset F_n^{\mu,+}$ and is made up of (strict) local minima: this relies on the study of the eigenvalues of the Hessian $\nabla^2 E_n^\mu(x^*)$ when $\nabla E_n^\mu(x^*) = 0$. The extension to a general density functions defined on more general intervals of the real line is straightforward.

Step 2 (the density f is compactly supported): The obvious problem here is that E_n^μ cannot be differentiably defined on $\mathbb{R}^d$ and, anyway, does not go to $+\infty$ with $\|x\|$. However a careful reading of the proof shows that the Mountain pass Lemma admits the following extension:

The Mountain pass Lemma (compact case): Let $K \subset \mathbb{R}^n$ be the closure of a nonempty convex open set O (then $\overset{\circ}{K} = O$). If $L : K \longrightarrow \mathbb{R}$ is C^1 on $\overset{\circ}{K}$, ∇L admits a continuous extension on K satisfying $\{\nabla L = 0\} \subset \overset{\circ}{K}$ and if for every small enough $\varepsilon > 0$, $(Id - \varepsilon\nabla L)(\overset{\circ}{K}) \subset \overset{\circ}{K}$, then the conclusion of the regular theorem still holds.

Step 3 (the general case): The idea is to approximate the density function f by compactly supported densities that still fulfill the above assumption $(\mathcal{L})$. To this end, set for every $k \ge 1$, $f_k := \frac{f}{\int_k^k f(u)} \mathbf{1}_{[-k,k]}$ and $\mu_k(du) := f_k(u)du$. One readily shows that, at least for large enough k, f_k satisfies assumption $(\mathcal{L})$. Furthermore one readily checks that f_k converges *a.s.* to f and $\int u^2 |f_k(u) - f(u)| du \to 0$ as $k \to +\infty$. General theory on convergence of measures shows that $E_n^{\mu_k}$ converges to E_n^μ uniformly on compact sets of $F_n^{\mu,+}$.

Now assume that E_n^μ has two strict local minima in $F_n^{\mu,+}$. They lie in $F_n^{\mu_k,+}$ for large enough k. Subsequently, $E_n^{\mu_k}$ necessarily has itself two (strict)

ESANN'1996 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 24-25-26 April 1996, D-Facto public., ISBN 2-9600049-6-5, pp. 97-102

Theorem 2 *(see [7]) If $L : \mathbb{R}^d \longrightarrow \mathbb{R}_+$ (L for "landscape" when $d=2$) is a continuously differentiable function satisfying $\lim_{\|x\|\to+\infty} L(x)=+\infty$ and if two distinct zeros of ∇L are (strict) local minima then ∇L has a third zero which can be in no case a local minimum.*

Proof of theorem 1: Step 1: First E_n^μ is continuously differentiable on $F_n^{\mu,+}$ and $\nabla E_n^\mu := \left(\dfrac{\partial E_n^\mu}{\partial x_i}\right)_{1\le i\le n}$ admits a straightforward continuous extension on $\overline{F}_n^{\mu,+}$ given by

$$\forall\, i \in \{1,\cdots,n\},\ \frac{\partial E_n^\mu}{\partial x_i} = 2\int_{\tilde{x}_i}^{\tilde{x}_{i+1}} (x_i - u) f(u)du, \quad \text{where } \tilde{x}_i := \frac{x_i + x_{i-1}}{2},\ 2\le i\le n,$$

$\tilde{x}_1 := m_\mu$, $\tilde{x}_{n+1} := M_\mu$. The set $F_n^{\mu,+}$ being convex, one easily shows that, for every $\varepsilon \in (0,1/2)$, $Id - \varepsilon\nabla E_n^\mu$ leaves $F_n^{\mu,+}$ stable. Hence, its continuous extension, still denoted ∇E_n^μ, leaves the closure $\overline{F}_n^{\mu,+}$ stable too.

Furthermore, it has been shown in [3] (or [1]) that if the density f is compactly supported, the above ln-concavity assumption implies that the zero set of ∇E_n^μ satisfies $\{\nabla E_n^{2,\mu} = 0\} \subset F_n^{\mu,+}$ and is made up of (strict) local minima: this relies on the study of the eigenvalues of the Hessian $\nabla^2 E_n^\mu(x^*)$ when $\nabla E_n^\mu(x^*) = 0$. The extension to a general density functions defined on more general intervals of the real line is straightforward.

Step 2 (the density f is compactly supported): The obvious problem here is that E_n^μ cannot be differentiably defined on $\mathbb{R}^d$ and, anyway, does not go to $+\infty$ with $\|x\|$. However a careful reading of the proof shows that the Mountain pass Lemma admits the following extension:

The Mountain pass Lemma (compact case): Let $K \subset \mathbb{R}^n$ be the closure of a nonempty convex open set O (then $\overset{\circ}{K} = O$). If $L : K \longrightarrow \mathbb{R}$ is C^1 on $\overset{\circ}{K}$, ∇L admits a continuous extension on K satisfying $\{\nabla L = 0\} \subset \overset{\circ}{K}$ and if for every small enough $\varepsilon > 0$, $(Id - \varepsilon\nabla L)(\overset{\circ}{K}) \subset \overset{\circ}{K}$, then the conclusion of the regular theorem still holds.

Step 3 (the general case): The idea is to approximate the density function f by compactly supported densities that still fulfill the above assumption ($\mathcal{L}$). To this end, set for every $k \ge 1$, $f_k := \frac{f}{\int_k^k f(u)} \mathbf{1}_{[-k,k]}$ and $\mu_k(du) := f_k(u)du$. One readily shows that, at least for large enough k, f_k satisfies assumption ($\mathcal{L}$). Furthermore one readily checks that f_k converges *a.s.* to f and $\int u^2 |f_k(u) - f(u)| du \to 0$ as $k \to +\infty$. General theory on convergence of measures shows that $E_n^{\mu_k}$ converges to E_n^μ uniformly on compact sets of $F_n^{\mu,+}$.

Now assume that E_n^μ has two strict local minima in $F_n^{\mu,+}$. They lie in $F_n^{\mu_k,+}$ for large enough k. Subsequently, $E_n^{\mu_k}$ necessarily has itself two (strict)

ESANN'1996 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 24-25-26 April 1996, D-Facto public., ISBN 2-9600049-6-5, pp. 97-102

local minima (converging to those of E_n^μ when $k \uparrow +\infty$). Step 2 makes this impossible. $\diamond$

Remark: A geometrical consequence of the stability property $(Id-\varepsilon\nabla E_n^\mu)(\overline{F}_n^{\mu,+}) \subset \overline{F}_n^{\mu,+}$ is that, on the boundary $\partial F_n^{\mu,+}$ of $F_n^{\mu,+}$, the vector field $-\nabla E_n^\mu$ is oriented towards the interior of $F_n^{\mu,+}$ *i.e.*, for every $x_0 \in \partial F_n^{\mu,+}$ and every vector $\vec{u}$ satisfying $(\vec{u}|x - x_0) \geq 0$ for every $x \in F_n^{\mu,+}$ (there is some due to the convexity of $F_n^{\mu,+}$), one has $(-\nabla E_n^\mu(x_0)|\vec{u}) \geq 0$.

A counter-example to uniqueness when assumption $(\mathcal{L})$ fails (see also [1]): Let $\mu(du) := g(nu - [nu])du$ where $g \in \mathcal{C}([0,1])$ is a probability density function satisfying $g(u) = g(1-u)$ ($[\,.\,]$ denotes the integral part). $(\frac{2k-1}{2n})_{1\leq k\leq n}$ is an unstable equilibrium point of ∇E_n^μ, as soon as $g(0) \geq \frac{2n}{n-1}$. On the other hand, the existence of a minimum in $F_n^{\mu,+}$ is granted, so E_n^μ has at least two critical n-tuples.

However, note that this only stands as a counter-example to uniqueness of the equilibrium point in $F_n^{\mu,+}$: the set of local minima may still be reduced to a single n-tuple. It is then necessarily the absolute minimum on the whole $\overline{F}_n^{\mu,+}$ since the absolute minimum cannot lie on the boundary $\partial F_n^{\mu,+}$ (see [6]). Simulations plead in favor of such a situation in the above counter-example.

Examples: The above result embodies:

- the Normal distributions $m(du) := \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(u-m)^2}{2\sigma^2}} du$,
- the $\gamma(\alpha,\beta)$-distributions for $\alpha \geq 1$, $\beta > 0$ *i.e.* $\mu(du) := \frac{x^{\alpha-1}e^{-\frac{u}{\beta}}}{\beta^\alpha\Gamma(\alpha)} \mathbf{1}_{\{u\geq 0\}}du$,
hence all the Exponential distributions $\mu(du) := \lambda e^{-\lambda u}\mathbf{1}_{\{u\geq 0\}}du$, $\lambda > 0$, and the $\chi^2(k)$-distributions whenever $k \geq 2$,
- the $\beta(a,b)$-distributions whenever a, $b \geq 1$ *i.e.* $\mu(du) := \frac{u^{a-1}(1-u)^{b-1}}{B(a,b)} \mathbf{1}_{\{0\leq u\leq 1\}}du$ (and subsequently the $U([0,1])$-distribution for which $x^* := (\frac{2k-1}{2n})_{1\leq k\leq n}$),
- the *logistic* distribution given by $\mu(du) := \frac{du}{(1+e^u)(1+e^{-u})}$,
- any compactly supported distribution with strictly concave density.

One noticeable fact is that the $\chi^2(1)$-distribution is not included in that list. This is somewhat unexpected as this distribution is in some sense "surrounded" by distribution whose equilibrium is unique: if $\omega \sim \mathcal{N}(0;1)$, then $\omega^2 \sim \chi^2(1)$ and if ω_1, ω_2 are independent with $\chi^2(1)$ distribution then $\omega_1 + \omega_2 \sim \chi^2(2)$. Actually there is no real doubt that the $\chi^2(1)$ satisfies the uniqueness property for every $n \geq 1$.

Applying the usual Robbins-Monro approach for stochastic algorithm, one easily derives that the VQ algorithm converges toward its unique equilibrium, namely

Corollary 3 *Let μ be a probability measure on $\mathbb{R}$ whose density f satisfies assumption $(\mathcal{L})$. Let $(X^t)_{t\in\mathbb{N}}$ be the adaptative algorithm defined by $X^0 \in F_n^{\mu,+}$ and, for every $t \in \mathbb{N}$,*

$$X_i^{t+1} = X_i^t - \varepsilon_{t+1}\mathbf{1}_{]\tilde{X}_i,\tilde{X}_{i+1}]}(\omega^{t+1})(X_i^t - \omega^{t+1}), \quad 1 \le i \le n.$$

where $(\omega^t)_{t\in\mathbb{N}}$ is a sequence of independent identically μ-distributed stimuli and $(\varepsilon_t)_{t\geq 1}$ is a sequence of $(0,1/2)$-valued gain parameter such that $\sum_{t\geq 1}\varepsilon_t = +\infty$ and $\sum_{t\geq 1}\varepsilon_t^2 < +\infty$. Then there exists a unique n-tuple $x^* \in F_n^{\mu,+}$ s.t.

$$E_n^\mu(x^*) = \min_{\mathbf{R}^n} E_n^\mu \;\; \textit{and} \;\; X^t \stackrel{t\to+\infty}{\longrightarrow} x^*\text{-}a.s..$$

Remark: Following [4] and [2] a stochastic procedure *a.s.* cannot converge to a saddle point or a local maximum, so the above corollary will hold as soon as E_n^μ only has a single local minimum. But we do not know any natural assumption on the density f that would ensure such a property.

Acknowledgement: We are grateful to L. Jeanjean for a useful reference.

References

[1] C. Bouton, G. Pagès (1993). Self-organization and convergence of the one-dimensional Kohonen algorithm with non uniformly distributed stimuli, *Stochastic Processes and their Applications*, **47**, 249-274.

[2] O. Brandière, M. Duflo (1996). Les algorithmes stochastiques contournent-ils les pièges?, forthcoming in *Les Annales de l'I.H.P.*.

[3] J.-C. Fort, G. Pagès (1995). On the *a.s.* convergence of the Kohonen algorithm with a generalized neighborhood function, forthcoming in *Annals of Applied Probability*, **4**.

[4] V.A. Lazarev (1992). Convergence of stochastic approximation procedures in the case of a regression equation with several roots. Translated from *Problemy Pederachi informatsii*, vol. **28**, n°1.

[5] G. Pagès (1993). Voronoï Tessellation, space quantization algorithm and numerical integration, *Proceedings of the ESANN'93*, M. Verleysen éd., Editions D Facto (ISBN 2-9600049-0-6), Bruxelles, pp.221-228.

[6] G. Pagès (1996). Numerical integration by space quantization, Paris 12 Univ, technical report

[7] M. Struwe (1990). *Variational Methods (Application to non linear p.d.e & Hamiltonian Systems)*, Springer, 244p.

ESANN'1996 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 24-25-26 April 1996, D-Facto public., ISBN 2-9600049-6-5, pp. 97-102

[8] M. Verleysen (1995). Les principaux modèles de réseaux de neurones artificiels, in *Les Réseaux de Neurones en Finance : Conception et Applications*, E. de Bodt & E.F. Henrion eds., D Facto, (Bruxelles).

D. Lamberton, Univ. de Marne-La-Vallée, Equipe d'Analyse et de Mathématiques Appliquées, 2 rue de la Butte Verte, F-93166 Noisy-Le-Grand Cedex. dlamb@math.univ-mlv.fr

G. Pagès, Laboratoire de Probabilités, URA 224, Univ. P.&M. Curie, 4, Pl. Jussieu, F-75252 Paris Cedex 05. gpa@ccr.jussieu.fr