Convergence of Stochastic Vector Quantization and Learning Vector Quantization with Bregman Divergences ⋆

Christos N. Mavridis **John S. Baras**

Electrical and Computer Engineering Department and the Institute for Systems Research, University of Maryland, College Park, MD 20742 USA, (e-mail: {mavridis,baras}@umd.edu).

Abstract: Stochastic vector quantization methods have been extensively studied as supervised (classification) and unsupervised (clustering) learning algorithms, due to being online, data-driven, interpretable, robust, and fast to train and evaluate. As prototype-based methods, they depend on a dissimilarity measure. which, can be shown that, is both necessary and sufficient to belong to the family of Bregman divergences, if the mean value is used as the representative of the cluster. In this work, we investigate the convergence properties of stochastic vector quantization (VQ) and its supervised counterpart, Learning Vector Quantization (LVQ), using Bregman divergences. We employ the theory of stochastic approximation to study the conditions on the initialization and the Bregman divergence generating functions, under which, the algorithms converge to desired configurations. These results formally support the use of Bregman divergences, such as the Kullback-Leibler divergence, in vector quantization algorithms.

Keywords: learning algorithms, stochastic approximation, convergence proofs

1. INTRODUCTION

Vector quantization methods, originally proposed over 30 years ago for data compression (Gray, 1990), have been extensively studied and used as supervised (classification) and unsupervised (clustering) learning algorithms. Not only they constitute interpretable, robust, data-driven and topology-preserving algorithms (Uriarte and Martín, 2005), but they can be formulated as online, stochastic gradient descent algorithms, sparse in the sense of memory complexity, and fast to train and evaluate.

Despite not being able to compete with the accuracy of the state-of-the-art deep neural network architectures, they offer, in many cases, an appealing alternative because of their developed mathematical theory. As a result, they are still being studied in conjunction with current neural network architectures (Saralajew et al., 2018; Villmann et al., 2017a; Shah and Koltun, 2018), and still being used in standard classification problems (Villmann et al., 2017b), data clustering (Shah and Koltun, 2018), time series and speech analysis (Melchert et al., 2016; Wang et al., 2019), biomedical applications (Biehl, 2017), and topological data analysis (Zielinski et al., 2018). Recently, LVQ methods have shown impressive robustness against adversarial attacks, suggesting a valid reason to deploy them instead of neural network architectures in security critical applications (Saralajew et al., 2019).

As prototype-based learning methods, VQ and LVQ are usually based on metrics, such as the Euclidean distance. However, the utilization of non-standard metrics and general dissimilarity measures, has become a topic of increasing importance in data processing and pattern recognition, and in the case of prototype-based learning, the so called Bregman divergences, have recently been acknowledged to play an important role (Banerjee et al., 2005; Mwebaze et al., 2011; Villmann and Haase, 2011; Villmann et al., 2010). A key property of the family of Bregman Divergences is that their use as a distortion measure, is both sufficient and necessary for choosing the mean as a representative of a random set, when trying to minimize the expected value of the distortion. In addition, due to the correspondence between exponential families and Bregman divergences, the efficiency of soft-clustering algorithms using Expectation-Maximization (EM) methods, and Deterministic Annealing approaches (Rose, 1998), can be greatly improved (Banerjee et al., 2005).

Batch algorithms for Vector Quantization, which is a non-convex optimization problem, based on the generalized Linde-Buzo-Gray (LBG) algorithm (Gersho and Gray, 2012), have been shown to converge to a local minimum of the average distortion, if and only if a Bregman divergence is used as a distortion measure (Banerjee et al., 2005). On the other hand, convergence analysis of stochastic VQ and LVQ, which are, many times, preferred over batch learning algorithms due to their iterative nature, is more involved (Bottou, 1998; Bottou and Bengio, 1995; Baras and LaVigna, 1991; Baras and Dey, 1999). The reason is that they are stochastic, non-convex optimization problems, with not differentiable cost functions (although the latter can be dealt with by introducing differentiable approximations (Sato and Yamada, 1996; Hammer and Villmann, 2002; Nova and Estévez, 2016, 2014)). To our knowledge, convergence has only been studied under the assumption of using metrics.

In this work, we focus on the convergence properties of stochastic Vector Quantization (VQ) and Learning Vector Quantization (LVQ) using Bregman divergences as dissimilarity measures. We employ mathematical tools from stochastic approximation and control theory (Benveniste et al., 2012; Borkar, 2009; Bottou,

⋆ This work was partially supported by ONR grant N00014-17-1-2622.

1998) to investigate the conditions on the initialization and the Bregman divergence generating functions, under which, the algorithms converge. By making use of the o.d.e. method, we show convergence to desired configurations through standard Lyapunov stability arguments on their limiting ordinary differential equations. These results formally support the use of Bregman divergences, such as the Kullback-Leibler divergence, in vector quantization algorithms.

The rest of the paper is organized as follows: Section 2 defines the Bregman Divergences and introduces the stochastic approximation theory, and Sections 3 and 4 study the convergence properties of VQ and LVQ algorithms. In Section 5, initialization methods, LVQ variants, and practical applications are discussed, and, finally, Section 6 concludes the paper.

2. PRELIMINARIES

2.1 Bregman Divergences

A Bregman Divergence is a dissimilarity measure $d : H \times H \to [0,\infty)$, where H is a normed vector space, that generalizes the notion of a metric, and, in general, may not be symmetric or satisfy the triangle inequality. Formally it is defined in the following:

Definition 1 (Bregman Divergence)**.** *Let $\phi : H \to \mathbb{R}$, be a strictly convex function defined on a normed vector space $dom(\phi) = H$ such that ϕ is twice F-differentiable on H. The Bregman divergence $d_\phi : H \times H \to [0,\infty)$ is defined as:*

$$d_\phi(x,\mu) = \phi(x) - \phi(\mu) - \frac{\partial\phi}{\partial\mu}(\mu)(x-\mu),$$

where $x,\mu \in H$, and the continuous linear map $\frac{\partial\phi}{\partial\mu}(\mu) : H \to \mathbb{R}$ is the Fréchet derivative of ϕ at μ.

In this work, we will concentrate on nonempty, convex sets $S \subseteq H$, where H is a finite dimensional Hilbert space, and in particular, $H = \mathbb{R}^d$, where, in view of the Riesz-Fréchet theorem, and under the Euclidean inner product $\langle x,y\rangle = x^T y$, it is common to denote $\frac{\partial\phi}{\partial\mu}(\mu)\, s = \langle \nabla\phi(\mu), s\rangle$, $\forall s \in H$, so that the derivative of d_ϕ with respect to the second argument can be written as

$$\begin{aligned}\frac{\partial d_\phi}{\partial\mu}(x,\mu) &= \frac{\partial\phi(x)}{\partial\mu} - \frac{\partial\phi(\mu)}{\partial\mu} - \frac{\partial^2\phi(\mu)}{\partial\mu^2}(x-\mu) + \frac{\partial\phi(\mu)}{\partial\mu} \\ &= -\frac{\partial^2\phi(\mu)}{\partial\mu^2}(x-\mu) = -\left\langle \nabla^2\phi(\mu), (x-\mu)\right\rangle\end{aligned}$$

where $x,\mu \in S$, $\frac{\partial}{\partial\mu}$ represents differentiation with respect to the second argument of d_ϕ, and $\nabla^2\phi(\mu)$ represents the Hessian matrix of ϕ at μ.

Example 1. *As a first example, $\phi(x) = \langle x,x\rangle$, $x \in \mathbb{R}^d$, gives the squared Euclidean distance*

$$d_\phi(x,\mu) = \|x-\mu\|^2$$

for which $\frac{\partial d_\phi}{\partial\mu}(x,\mu) = -2(x-\mu)$.

Example 2. *Another interesting Bregman divergence, delineating the connection to information theory, is the generalized I-divergence which results from $\phi(x) = \langle x, \log x\rangle$, $x \in \mathbb{R}^d_{++}$ such that*

$$d_\phi(x,y) = \langle x, \log x - \log\mu\rangle - \langle \mathbb{1}, x-\mu\rangle$$

for which $\frac{\partial d_\phi}{\partial\mu}(x,\mu) = -diag^{-1}(\mu)(x-\mu)$, where $\mathbb{1} \in \mathbb{R}^d$ is the vector of ones, and $diag^{-1}(\mu) \in \mathbb{R}^{d\times d}_{++}$ is the diagonal matrix with diagonal elements the inverse elements of μ. It is easy to see that $\phi(x)$ reduces to the ubiquitous KL-divergence if $\langle \mathbb{1}, x\rangle = 1$.

An extensive overview of Bregman divergences in Vector Quantization applications can be found in (Banerjee et al., 2005). We summarize their key property in the following:

Theorem 1. *Let $X : \Omega \to S$ be a random variable defined in the probability space $(\Omega,\mathcal{F},\mathbb{P})$ such that $\mathbb{E}[X] \in ri(S)$, and let a distortion measure $d : S \times ri(S) \to [0,\infty)$, where $ri(S)$ denotes the relative interior of S. Then $\mu \triangleq \mathbb{E}[X]$ is the unique minimizer of $\mathbb{E}[d(X,s)]$ in $ri(S)$, if and only if d is a Bregman Divergence for any function ϕ that satisfies the definition.*

Proof. For necessity, identical arguments as in Appendix B of (Banerjee et al., 2005) are followed. For sufficiency, extending the work of Banerjee et al. (Banerjee et al., 2005), for any distribution of X:

$$\begin{aligned}&\mathbb{E}[d_\phi(X,s)] - \mathbb{E}[d_\phi(X,\mu)] = \\ &\quad= \phi(\mu) + \frac{\partial\phi}{\partial\mu}(\mu)(\mathbb{E}[X]-\mu) - \phi(s) - \frac{\partial\phi}{\partial s}(s)(\mathbb{E}[X]-s) \\ &\quad= \phi(\mu) - \phi(s) - \frac{\partial\phi}{\partial s}(s)(\mu-s) = d_\phi(\mu,s) \geq 0, \quad \forall s \in S\end{aligned}$$

with equality holding only when $s = \mu$ by the strict convexity of ϕ, which completes the proof. □

2.2 Stochastic Approximation

Theorem 2 ((Borkar, 2009))**.** *Almost surely, the sequence $\{x_n\} \in \mathbb{R}^d$ generated by the following stochastic approximation scheme:*

$$x_{n+1} = x_n + \alpha(n)[h(x_n) + M_{n+1}], n \geq 0 \tag{1}$$

with prescribed x_0, converges to a (possibly sample path dependent) compact, connected, internally chain transitive, invariant set of the o.d.e:

$$\dot{x}(t) = h(x(t)),\ t \geq 0, \tag{2}$$

where $x : \mathbb{R}_+ \to \mathbb{R}_d$ and $x(0) = x_0$, provided the following assumptions hold:

- *(A1) The map $h : \mathbb{R}^d \to \mathbb{R}^d$ is Lipschitz, i.e., $\exists L$ with $0 < L < \infty$ such that $\|h(x) - h(y)\| \leq L\|x-y\|$, $x,y \in \mathbb{R}^d$,*
- *(A2) The stepsizes $\{\alpha(n) \in \mathbb{R}_{++},\ n \geq 0\}$ satisfy $\sum_n \alpha(n) = \infty$, and $\sum_n \alpha^2(n) < \infty$,*
- *(A3) $\{M_n\}$ is a martingale difference sequence with respect to the increasing family of σ-fields $\mathcal{F}_n \triangleq \sigma(x_m, M_m,\ m \leq n)$, $n \geq 0$, i.e., $\mathbb{E}[M_{n+1}|\mathcal{F}_n] = 0$ a.s., for all $n \geq 0$, and, furthermore, $\{M_n\}$ are square-integrable with $\mathbb{E}\left[\|M_{n+1}\|^2|\mathcal{F}_n\right] \leq K\left(1 + \|x_n\|^2\right)$, a.s., where $n \geq 0$ for some $K > 0$,*
- *(A4) The iterates $\{x_n\}$ remain bounded a.s., i.e., $\sup_n \|x_n\| < \infty$ a.s.*

Given the conditions of Theorem 2, the following criteria for global and local convergence, respectively, hold:

Corollary 2.1 ((Borkar, 2009))**.** *Suppose there exists a radially unbounded Lyapunov function V, i.e., a continuously differentiable function $V : \mathbb{R}^d \to [0,\infty)$ such that $\lim_{\|x\|\to\infty} V(x) = \infty$, $H \triangleq \{x \in \mathbb{R}^d : V(x) = 0\} \neq \emptyset$, and $\langle h(x), \nabla V(x)\rangle \leq 0$ with the equality holding if and only if $x \in H$. Then, almost surely, $\{x_n\}$ converge to an internally chain transitive invariant set contained in H. If, the only internally chain transitive invariant sets for (2) are isolated equilibrium points, then, almost surely, $\{x_n\}$ converges to a possibly sample dependent equilibrium point.*

Corollary 2.2 ((Benveniste et al., 2012) Corollary 6, p.46)**.** *Assume x^* is a locally asymptotically stable equilibrium of (2) with domain of attraction D^*, and let Q be a compact subset of D^*. If $x_n \in Q$ for infinitely many n, then*

$$\lim_{n\to\infty} x_n = x^* \ a.s.$$

3. CONVERGENCE OF STOCHASTIC VECTOR QUANTIZATION

In this section, we focus on the unsupervised problem of prototype-based clustering. One can show based on Theorem 1, that the use of Bregman divergences in batch algorithms based on the generalized Lloyd algorithm, is both necessary and sufficient for local convergence (Banerjee et al., 2005). We extend this result to prove convergence of the stochastic Vector Quantization algorithm (Kohonen, 1995) based on Bregman divergences.

We begin with the definition of a Voronoi partition:

Definition 2 (Voronoi Partition)**.** *Let $S_h \subseteq S,\ h = 1,\dots,k$, such that $V \triangleq \{S_h\}_{h=1}^k$ forms a partition of S, i.e. $\bigcup_{h=1}^{k} S_h = S$, and $S_i \bigcap S_j = \emptyset,\ i \neq j \in \{1,\dots,k\}$. Then V is called a Voronoi partition with respect to $M \triangleq \{\mu_h\}_{h=1}^k \in S^k$, if*

$$S_h = \left\{ X \in S : h = \underset{\tau=1,\dots,k}{\arg\min}\ d(X,\mu_\tau) \right\},\ h = 1,\dots,k.$$

where $d : S \times S \to [0,\infty)$. If $d \equiv d_\phi$ is a Bregman divergence for an appropriately defined function ϕ, then S_h are convex, since the locus of equidistant points between two different points $\mu_1 \neq \mu_2 \in S$ is a hyperplane.

Then, the problem of divergence-based Vector Quantization can be stated as an optimization problem:

Problem 1. *Let $X : \Omega \to S$ be a random variable defined in the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and $d_\phi : S \times ri(S) \to [0,\infty)$ be a Bregman divergence with properly defined function ϕ. Let $V \triangleq \{S_h\}_{h=1}^k$ be a Voronoi partition of S with respect to d_ϕ and $M \triangleq \{\mu_h\}_{h=1}^k$, such that $\mu_h \in ri(S_h),\ h \in K,\ K \triangleq \{1,\dots,k\}$, and define the quantizer $Q : S \to S$ such that $Q(X) = \sum_{h=1}^k \mu_h \mathbb{1}_{[X\in S_h]}$.*

Then the problem is formulated as

$$\min_{M,V} J(Q) \triangleq \mathbb{E}_X \left[d_\phi \left(X, Q(X) \right) \right]$$

$$\Leftrightarrow \min_{\{\mu_h\}_{h=1}^k} J(Q) \triangleq \sum_{h=1}^{k} \mathbb{E}_X \left[d_\phi \left(X, \mu_h \right) \mathbb{1}_{[X\in S_h]} \right],$$

It is typically the case that the actual distribution of $X \in S$ is unknown, but a set of, independent, observations $\{X_i\}_{i=1}^n \in S$ that are identically distributed with X, are available. The stochastic vector quantization algorithm is used when the observed data are not available a priori but are being acquired online, or when the processing of the entire dataset in every iteration is computationally expensive, and is defined recursively for every $t \geq 0$ as:

$$\begin{aligned} \mu_h^{t+1} &= \mu_h^t + \alpha(t) \left(-\mathbb{1}_{\left[X_{t+1} \in S_h^{t+1}\right]} \right) \nabla_{\mu_h} d_\phi \left(X_{t+1}, \mu_h^t \right) \\ S_h^{t+1} &= \left\{ X \in S : h = \underset{\tau=1,\dots,k}{\arg\min}\ d_\phi(X, \mu_\tau^t) \right\},\ h \in K \end{aligned} \tag{3}$$

where μ_h^0 is given during initialization.

We employ the o.d.e. method introduced in Theorem 2 to show convergence of Algorithm (3) to a local minimum of $J(Q)$, as $n \to \infty$. In what follows we work in the same way for all $h \in K$. First, we define the functions $\Theta_h : S^k \times S \to H$ as

$$\Theta_h(\mu, X) = \left(-\mathbb{1}_{[X\in S_h]} \right) \nabla_{\mu_h} d_\phi \left(X, \mu_h \right) \tag{4}$$

and introduce, for $t \geq 0$, the increasing family of σ-fields $\mathcal{F}_t \triangleq \sigma\left(\mu_h^\tau, X_\tau,\ \tau \leq t\right)$, in order to define, for every $t \geq 0$, the differences

$$M_h^{t+1} \triangleq \Theta_h(\mu^t, X_{t+1}) - \mathbb{E}\left[\Theta_h(\mu^t, X_{t+1}) | \mathcal{F}_t\right]$$

which are martingale difference sequences, since, by definition, $\mathbb{E}\left[M_h^{t+1} | \mathcal{F}_t\right] = 0$ almost surely. Intuitively, we have expressed $\Theta_h(\mu^t, X_{t+1})$ as a perturbation of $\theta_h^t(\mu) \triangleq \mathbb{E}\left[\Theta_h(\mu^t, X_{t+1}) | \mathcal{F}_t\right]$, for all $t \geq 0$. Given the *iid* assumption on $\{X^t\}_{t=1}^n$, it is reasonable to assume the Markov property

$$\mathbb{P}[g(X_t, \mu_h^t) | \mathcal{F}_t] = \mathbb{P}[g(X_t, \mu_h^t) | \mu_h^t] \ a.s.$$

for every Borel measurable positive function g such that $\mathbb{E}\left[\left|g(X_t, \mu_h^t)\right|\right] < \infty$. Therefore, we can write

$$\theta_h^t(\mu) = \mathbb{E}\left[\Theta_h(\mu^t, X_{t+1}) | \mathcal{F}_t\right] = \mathbb{E}\left[\Theta_h(\mu^t, X_{t+1}) | \mu_h^t\right] \ a.s. \tag{5}$$

In other words, algorithm (3) is a stochastic approximation algorithm:

$$\mu^{t+1} = \mu^t + \alpha(t)\left[\theta^t(\mu) + M^{t+1}\right] \tag{6}$$

where $\mu^t = \left[\mu_1^t, \dots, \mu_k^t\right]^T$, $M^t = \left[M_1^t, \dots, M_k^t\right]^T$, and $\theta^t(\mu) = \left[\theta_1^t(\mu), \dots, \theta_k^t(\mu)\right]^T$. In order for (6) to satisfy the conditions of Theorem 2, we first select the stepsizes $\{\alpha(t)\}_{t\geq 0}$ to satisfy (A2), and define the functions

$$\begin{aligned} \theta_h(\mu) &= \lim_{t\to\infty} \mathbb{E}\left[\Theta_h(\mu^t, X_{t+1}) | \mu_h^t\right] \\ &= -\mathbb{E}_X\left[\mathbb{1}_{[X\in S_h]} \nabla_{\mu_h} d_\phi\left(X, \mu_h\right)\right] \end{aligned} \tag{7}$$

where, the expectation operator $\mathbb{E}_X[\cdot]$ is with respect to the random variable X, given the values of μ_h, S_h. In order to satisfy (A1), and (A3), we limit the choices of the Bregman divergence generating functions to those that satisfy the assumption:

Assumption 1. *The strictly convex functions $\phi : H \to \mathbb{R}$ are two times continuously F-differentiable on H, and $\frac{\partial^2 \phi(\mu)}{\partial \mu^2}$ is a Lipschitz bounded linear map for all $\mu \in S$, such that $\frac{\partial d_\phi}{\partial \mu}(x, \mu)$ is Lipschitz continuous in S, and $\left\|\nabla_\mu d_\phi(x,\mu)\right\|^2 \leq K_0(1 + \|\mu\|^2)$ for some $K_0 > 0$.*

We note that ϕ functions commonly used in defining Bregman divergences, such as the Euclidean, Mahalanobis and Itakura-Saito distance, as well as the generalized Kullback-Leibler (I) divergence, all satisfy Assumption 1. For example, as shown in Example 2, for the I divergence we get $\frac{\partial^2 \phi(\mu)}{\partial \mu^2} = diag^{-1}(\mu)$ which is a Lipschitz bounded linear map for all $\mu \in \mathbb{R}_{++}$. Now, each $\theta_h(\mu)$, and therefore $\theta(\mu) = \left[\theta^1(\mu), \dots, \theta^k(\mu)\right]^T$, is Lipschitz continuous. This is easy to see since

$$\begin{aligned} \theta_h(\mu) &= -\int_{S_h} \frac{\partial}{\partial \mu_h} d_\phi(x, \mu_h) dF(x) \\ &= \int_{S_h} \frac{\partial^2}{\partial \mu_h^2} \phi(\mu_h)(x - \mu) dF(x) \end{aligned}$$

where $F(x) = \mathbb{P}[x \leq X]$. Therefore, from Lebesgue theory, $\theta_h(\mu)$ is Lipschitz as long as $\frac{\partial^2}{\partial \mu_h^2}\phi(\mu_h)(x - \mu)$ is Lipschitz, which is a direct consequence of Assumption 1. Furthermore, given Algorithm (3), and the fact that $\mu^0 < \infty$, we can conclude that $\{\mu^t\}_{t=0}^n$ remains bounded almost surely. We have already shown that $\mathbb{E}\left[M_h^{t+1} | \mathcal{F}_t\right] = 0\ a.s.$, and, under Assumption 1:

$$\begin{aligned}\mathbb{E}\left[\left\|M_h^{t+1}\right\|^2|\mathcal{F}_t\right] &= \mathbb{E}_X\left[\left\|\Theta_h(\mu^t,X_{t+1})\right\|^2\right]-\left\|\theta_h^t(\mu)\right\|^2\\ &= \mathbb{E}_X\left[\left\|\mathbb{1}_{\left[X\in S_h^{t+1}\right]}\nabla_{\mu_h}d_\phi\left(X,\mu_h^t\right)\right\|^2\right]\\ &\quad-\left\|\mathbb{E}_X\left[\mathbb{1}_{[X\in S_h]}\nabla_{\mu_h}d_\phi\left(X,\mu_h\right)\right]\right\|^2\\ &\leq K_1\left(1+\left\|\mu_h^t\right\|^2\right)\end{aligned} \tag{8}$$

for some $K_1>0$. Therefore, by Theorem 2 and Corollary 2.2, μ^t converges to a locally asymptotically stable equilibrium μ^* of the o.d.e:

$$\dot{\mu}(t)=\theta\left(\mu(t)\right),\ t\geq 0, \tag{9}$$

where $\mu:\mathbb{R}_+\to S^k$, and $\mu(0)=\mu_0$, i.e., $\lim_{t\to\infty}\mu^t=\mu^*$almost surely, provided that μ^t visits a compact subset of the domain of attraction D^* of μ^* in (9), infinitely often. It should be mentioned that there is no general theory for the conditions under which μ visits D^* infinitely often, which, depends on both the initial conditions of (3) and the sample path $\{X^t\}_{t=1}^n$. Regarding the initial conditions μ^0, the convergence results above require that they are chosen close to a stable point μ^* of (9), i.e., within the domain of attraction D^*. We are interested in the asymptotically stable equilibria of (9). We recall that

$$\begin{aligned}\theta_h(\mu) &= -\mathbb{E}_X\left[\mathbb{1}_{[X\in S_h]}\nabla_{\mu_h}d_\phi\left(X,\mu_h\right)\right]\\ &= -\nabla_{\mu_h}\mathbb{E}_X\left[\mathbb{1}_{[X\in S_h]}d_\phi\left(X,\mu_h\right)\right]\end{aligned} \tag{10}$$

and define the functions $J_h(\mu)\triangleq\mathbb{E}_X\left[\mathbb{1}_{[X\in S_h]}d_\phi\left(X,\mu_h\right)\right]$ and $J(\mu)\triangleq\sum_{h=1}^k J_h(\mu)=\mathbb{E}_X\left[d_\phi\left(X,Q(X)\right)\right]$. Then

$$\theta(\mu)=-\nabla_\mu J(\mu) \tag{11}$$

where the cost function $J\geq 0$ can be treated as a potential function to be minimized, so that, by standard Lyapunov stability arguments, if $J(\mu^*)$ is a minimum of J, then μ^* is an asymptotically stable equilibrium point for (9) for some domain of attraction D^*. Therefore, we have shown the following:

Theorem 3. *The sequence $\{\mu^t\}$ generated by the stochastic vector quantization algorithm (3) converges almost surely to a local solution μ^* of Problem 1, as long as the function ϕ satisfies Assumption 1, the stepsizes satisfy $\sum_t\alpha(t)=\infty$, $\sum_t\alpha^2(t)<\infty$, and μ^t visits a compact subset of the domain of attraction D^* of μ^* infinitely often, $\mu^0\in D^*$.*

Furthermore, it is easy to see that as the number of clusters goes to infinity, i.e. as $k\to\infty$, then $J(Q)\to 0$, since $\mathbb{E}_X\left[d_\phi\left(X,\mu_h\right)\mathbb{1}_{[X\in S_h]}\right]\to 0$, for all $h\in K$.

4. CONVERGENCE OF LEARNING VECTOR QUANTIZATION

Learning vector quantization (LVQ) first introduced by Kohonen (Kohonen, 1995) is the supervised counterpart of the stochastic vector quantization algorithm, used for approximating the decision boundary of a pattern classification problem. It uses a set of training data for which the classes are known to divide the data space into a number of Voronoi cells represented by the corresponding Voronoi vectors and their associated class decisions.We are going to investigate the convergence properties of LVQ, based on Bregman divergences, in the case of binary classification, which can easily be generalized to any type of classification task (see, e.g. (Duda et al., 2012)). Let the following binary classification problem:

Problem 2. *Let $\{X,c\}\in S\times\{0,1\}$ defined in a probability space $(\Omega,\mathcal{F},\mathbb{P})$, $X:\Omega\to S$ be a random variable, and $c:S\to\{0,1\}$ its associated decision variable, such that c represents the actual class of X. Let $V\triangleq\{S_h\}_{h=1}^k$ be a Voronoi partition of S with respect to d_ϕ and $M\triangleq\{\mu_h\}_{h=1}^k$, $\mu_h\in ri(S_h)$, and define $C_\mu\triangleq\left\{c_{\mu_h}\right\}_{h=1}^k, c_{\mu_h}\in\{0,1\}, h\in K$, $K=\{1,\ldots,k\}$, such that c_{μ_h} represents the class of μ_h for all $h\in K$. Define the quantizer $Q:S\to\{0,1\}$ such that $Q(X)=\sum_{h=1}^k c_{\mu_h}\mathbb{1}_{[X\in S_h]}$.*

The minimum-error classification problem is then formulated as

$$\begin{aligned}\min_{\{\mu_h\}_{h=1}^k} J_B(Q) &\triangleq \pi_1\sum_{H_0}\mathbb{P}_1\left[X\in S_h\right]+\pi_0\sum_{H_1}\mathbb{P}_0\left[X\in S_h\right]\\ &= \pi_i+\sum_{H_i}\left(\pi_j\mathbb{P}_j\left[X\in S_h\right]-\pi_i\mathbb{P}_i\left[X\in S_h\right]\right)\end{aligned}$$

where $\pi_i=\mathbb{P}\left[c=i\right], \mathbb{P}_i\{\cdot\}=\mathbb{P}\{\cdot|c=i\}$, and H_i is defined as $H_i=\left\{h\in\{1,\ldots,k\}:c_{\mu_h}=i\right\}$, $i,j\in\{0,1\}, i\neq j$.

Remark 1. *We can generalize the definition of the minimum-error cost function J_B to a minimum-risk cost function*

$$J_R(Q)=\pi_1\sum_{H_0}\mathbb{E}_1\left[R(X)\mathbb{1}_{[X\in S_h]}\right]+\pi_0\sum_{H_1}\mathbb{E}_0\left[R(X)\mathbb{1}_{[X\in S_h]}\right]$$

where $\mathbb{E}_i$ denotes the expected value with respect to $\mathbb{P}_i$, $i,j\in\{0,1\}, i\neq j$, and $R:S\to\mathbb{R}_+$ is a risk function which assigns a miss-classification cost to each element in the domain of X.

Typically, the distribution of $\{X,c\}$ is not known, and, a sequence $\{X_i,c_i\}_{i=1}^n$ of independent random variables identically distributed with $\{X,c\}$ is being observed. The Learning Vector Quantization algorithm (LVQ) can be used when the observed data are not available a priori but are being acquired online, when the class indices of some observed data are not known apriori for training and need to be discovered, or when the processing of the entire dataset in every iteration is computationally expensive, and is defined recursively as follows

$$\begin{cases}\mu_h^{t+1} &= \mu_h^t-\alpha(t)\nabla_{\mu_h}d_\phi\left(X_{t+1},\mu_h^t\right),\ \text{if } c_{t+1}=c_{\mu_h}^t\\ \mu_h^{t+1} &= \mu_h^t+\alpha(t)\nabla_{\mu_h}d_\phi\left(X_{t+1},\mu_h^t\right),\ \text{if } c_{t+1}\neq c_{\mu_h}^t\end{cases} \tag{12}$$

where $h=\underset{\tau=1,\ldots,k}{\arg\min}\ d_\phi(X^{t+1},\mu_\tau^t)$, and μ_h^0 is given. We can write the LVQ algorithm as

$$\begin{aligned}\mu_h^{t+1} &= \mu_h^t+\alpha(t)\Theta_h(\mu^t,C_\mu^t,X_{t+1},c_{t+1})\nabla_{\mu_h}d_\phi\left(X_{t+1},\mu_h^t\right)\\ S_h^{t+1} &= \left\{X\in S:h=\underset{\tau=1,\ldots,k}{\arg\min}\ d_\phi\left(X,\mu_\tau^t\right)\right\},\ h=1,\ldots,k\end{aligned} \tag{13}$$

where, following the same methodology as in Section 3 for all $h\in K$, we have defined the functions

$$\Theta_h(\mu,C_\mu,X,c)=\left(-\mathbb{1}_{[X\in S_h]}\right)\left(\mathbb{1}_{\left[c=c_{\mu_h}\right]}-\mathbb{1}_{\left[c\neq c_{\mu_h}\right]}\right)\nabla_{\mu_h}d_\phi\left(X,\mu_h\right),$$

as well as the martingale difference sequences

$$M_h^{t+1}\triangleq\Theta_h(\mu^t,C_\mu^t,X_{t+1},c_{t+1})-\mathbb{E}\left[\Theta_h(\mu^t,C_\mu^t,X_{t+1},c_{t+1})|\mathcal{F}_t\right],$$

where $\mathcal{F}_t\triangleq\sigma\left(\mu_h^\tau,X_\tau,c_\tau,\ \tau\leq t\right)$, for $t\geq 0$, and, assuming similar independence and Markov properties as in Section 3, the functions $\theta_h^t(\mu)\triangleq\mathbb{E}\left[\Theta_h(\mu^t,C_\mu^t,X_{t+1},c_{t+1})|\mathcal{F}_t\right]=\mathbb{E}\left[\Theta_h(\mu^t,C_\mu^t,X_{t+1},c_{t+1})|\mu_h^t\right]$ a.s. Now (13) is a stochastic approximation algorithm in the form of (6) with stepsizes $\{\alpha(t)\}_{t\geq 0}$ satisfying (A2), while (A1), and (A3) are satisfied by Assumption 1, since $\theta(\mu)=\left[\theta^1(\mu),\ldots,\theta^k(\mu)\right]^T$ is Lipschitz continuous, with

$$\begin{aligned}\theta_h(\mu) &= \mathbb{E}_X\left[\Theta_h(\mu,C_\mu,X,c)\right]\\ &= \pi_0\mathbb{E}_0\left[\Theta_h(\mu,C_\mu,X,c)\right]+\pi_1\mathbb{E}_1\left[\Theta_h(\mu,C_\mu,X,c)\right]\\ &= -\delta_{\mu_h}\left(\pi_0\mathbb{E}_0\left[\mathbb{1}_{[X\in S_h]}\nabla_{\mu_h}d_\phi\left(X,\mu_h\right)\right]\right.\\ &\quad\left.-\pi_1\mathbb{E}_1\left[\mathbb{1}_{[X\in S_h]}\nabla_{\mu_h}d_\phi\left(X,\mu_h\right)\right]\right),\end{aligned} \tag{14}$$

where $\delta_{\mu_h} = \begin{cases} 1, & c_{\mu_h} = 0 \\ -1, & c_{\mu_h} = 1 \end{cases}$, and $\mathbb{E}\left[\left\|M_h^{t+1}\right\|^2 |\mathcal{F}_t\right] \leq K_1 (1 + \left\|\mu_h^t\right\|^2)$ for some $K_1 > 0$. However there is no guarantee that (A4) will be satisfied, i.e. $\sup_t \left\|\mu_h^t\right\| < \infty$ *a.s.*, and, in fact, in some cases the centroids $\mu_h,\ h \in K$ may diverge. Many variants of Algorithm (13) have been proposed to overcome this issue, as explained in Section 5, while, in an attempt to keep the same algorithm, Baras et. al in (Baras and LaVigna, 1991) proposed changing the decision policy of each centroid so that c_{μ_h} is updated in each iteration, according to the majority vote criterion, on the classes of the data in S_h. For now, we will assume that $\sup_t \left\|\mu_h^t\right\| < \infty$ *a.s*, and conclude that, according to Theorem 2 and Corollary 2.2, μ^t converges to a local asymptotically stable equilibrium μ^* of the o.d.e:

$$\dot{\mu}(t) = \theta(\mu(t)),\ t \geq 0, \tag{15}$$

where $\mu : \mathbb{R}_+ \to S^k$, and $\mu(0) = \mu^0$, provided that μ^t visits a compact subset of the domain of attraction D^* of μ^*, infinitely often. At this point, we seek potential asymptotically stable equilibrium points of (15). We note that

$$\begin{aligned} \theta_h(\mu) &= -\delta_{\mu_h}\left(\pi_0 \mathbb{E}_0\left[\mathbb{1}_{[X\in S_h]} \nabla_{\mu_h} d_\phi(X,\mu_h)\right] - \pi_1 \mathbb{E}_1\left[\mathbb{1}_{[X\in S_h]} \nabla_{\mu_h} d_\phi(X,\mu_h)\right]\right) \\ &= -\delta_{\mu_h} \nabla_{\mu_h}\left(\pi_0 \mathbb{E}_0\left[\mathbb{1}_{[X\in S_h]} d_\phi(X,\mu_h)\right] - \pi_1 \mathbb{E}_1\left[\mathbb{1}_{[X\in S_h]} d_\phi(X,\mu_h)\right]\right) \end{aligned} \tag{16}$$

and define the functions

$$J_{L_h}(\mu) \triangleq \delta_{\mu_h}\left(\pi_0 \mathbb{E}_0\left[\mathbb{1}_{[X\in S_h]} d_\phi(X,\mu_h)\right] - \pi_1 \mathbb{E}_1\left[\mathbb{1}_{[X\in S_h]} d_\phi(X,\mu_h)\right]\right)$$

and $J_L(\mu) \triangleq \sum_{h=1}^k J_h(\mu)$, where it is easy to show that

$$\begin{aligned} J_L &= \sum_{h=1}^k \delta_{\mu_h}\left(\pi_0 \mathbb{E}_0\left[\mathbb{1}_{[X\in S_h]} d_\phi(X,\mu_h)\right] - \pi_1 \mathbb{E}_1\left[\mathbb{1}_{[X\in S_h]} d_\phi(X,\mu_h)\right]\right) \\ &= J(Q) - 2J_{d_\phi}(Q) \end{aligned}$$

with $J(Q) = \sum_{h=1}^k \mathbb{E}\left[d_\phi(X,\mu_h)\,\mathbb{1}_{[X\in S_h]}\right]$ being the quantization error, and

$$J_{d_\phi}(Q) = \pi_1 \sum_{H_0} \mathbb{E}_1\left[d_\phi(X,\mu_h)\,\mathbb{1}_{[X\in S_h]}\right] + \pi_0 \sum_{H_1} \mathbb{E}_0\left[d_\phi(X,\mu_h)\,\mathbb{1}_{[X\in S_h]}\right]$$

being the minimum risk error associated with the risk function $R(X) \equiv d_\phi(X,\mu_h)$, for all $h \in K$. It is easy to see that, $-J(Q) \leq J_L \leq J(Q)$. However, if in each cluster $S_h,\ h = 1,\ldots,k$, with $c_{\mu_h} = i,\ i \in \{0,1\}$, it holds that

$$\pi_i \mathbb{E}_i\left[\mathbb{1}_{[X\in S_h]} d_\phi(X,\mu_h)\right] \geq \pi_j \mathbb{E}_j\left[\mathbb{1}_{[X\in S_h]} d_\phi(X,\mu_h)\right] \tag{17}$$

then it follows that $0 \leq J_L \leq J(Q)$, and the cost function J_L can be treated as a potential function, such that,

$$\theta(\mu) = -\nabla_\mu J_L(\mu) \tag{18}$$

The assumption that the class c_{μ_h} of μ_h corresponds to the class with the highest risk inside S_h, i.e. that (17) holds, is not so restrictive, as this holds true when the size of the Voronoi regions S_h gets smaller and the majority of the training samples inside S_h belong to class $c = i$, which is the case after some iterations of the algorithm. This can also be ensured by modifying the decision policy of the LVQ algorithm to incorporate an extension of the majority vote correction proposed in (Baras and LaVigna, 1991). Therefore, by standard Lyapunov stability arguments, μ^*, for which J_L is minimized, is an asymptotically stable equilibrium point for (15) for some domain of attraction D^*, such that, as the number of samples goes to infinity, then (13) moves $\mu \to \mu^*$, which, at least locally, minimizes J_L. Since $J_L = J(Q) - 2J_{d_\phi}(Q) \geq 0$, it follows that $J_{d_\phi}(Q) \leq \frac{1}{2}J(Q)$, that is, the misclassification risk is bounded above by a proportion of the clustering distortion. If, in addition, the number of clusters goes to infinity, i.e., if $k = k_t \to \infty$, it holds that $J(Q) \to 0$. Therefore, $J_{d_\phi}(Q) \to 0$ as well, and, because the size of the clusters $S_h,\ h = 1,\ldots,k$ goes to zero, following the arguments used in Ch.21, (Devroye et al., 2013), for convergence of Voronoi-type partitions, and assuming $\lim_{t\to\infty} k_t^2 \frac{\log t}{t} \to 0$, this implies that $J_B(Q)$ is minimized and that algorithm (13) converges to the Bayes classification error almost surely. Therefore, we have shown the following:

Theorem 4. *The sequence $\{\mu^t\}$ generated by the learning vector quantization algorithm (13) converges almost surely to a solution μ^* of Problem 2, as $k = k_t \to \infty$, provided that $\lim_{t\to\infty} k_t^2 \frac{\log t}{t} \to 0$, $\sum_t \alpha(t) = \infty$, $\sum_t \alpha^2(t) < \infty$, μ^t visits a compact subset of the domain of attraction D^* of μ^* infinitely often, $\mu^0 \in D^*$, $\sup_t \|\mu^t\| < \infty$ a.s., and the function ϕ satisfies Assumption 1.*

5. INITIALIZATION, VARIANTS, AND PRACTICAL IMPLICATIONS

It is apparent in the analysis of algorithms (3) and (13) that the initial configuration, as well as the number k of the clusters, play a key role in both the point of convergence, and the final minimum distortion achieved. This phenomenon is common in non-convex stochastic optimization problems, and annealing methods for avoiding local minima have been proposed, (Kirkpatrick et al., 1983; Rose, 1998). In particular, Deterministic Annealing (DA) (Rose, 1998; Miller et al., 1996) approaches, which make use of Gibbs distribution functions, become computationally simpler when based on Bregman divergences, due to their correspondence with exponential families (Banerjee et al., 2005), and can be used as a first step before applying vector quantization algorithms.

In order to guarantee satisfaction of Assumption (A4), Kohonen in (Kohonen, 1995) initially proposed LVQ2.1, and Sato et. al in (Sato and Yamada, 1996) extended it to the Generalized LVQ algorithm:

$$\begin{aligned} \mu_h^{t+1} &= \mu_h^t - \alpha(t) \nabla_{\mu_h} f\left(X_{t+1}, \mu_h^t, \mu_l^t\right) \\ \mu_l^{t+1} &= \mu_l^t - \alpha(t) \nabla_{\mu_l} f\left(X_{t+1}, \mu_h^t, \mu_l^t\right), \end{aligned} \tag{19}$$

where $h = \arg\min_{\tau: c_{\mu_\tau} = c_t} d_\phi(X^{t+1}, \mu_\tau^t)$, $l = \arg\min_{\tau: c_{\mu_\tau} \neq c_t} d_\phi(X^{t+1}, \mu_\tau^t)$, μ_h^0 is given, and the function $f : S \times S \times S \to \mathbb{R}$ is carefully selected (Sato and Yamada, 1996). Although out of the scope of this paper, the proposed methodology can be applied to show that, under the same assumptions, LVQ2.1 and GLVQ, minimize, at least locally and as $t, k \to \infty$, their error functions $J = \mathbb{E}\left[f\left(X_{t+1}, \mu_h^t, \mu_l^t\right)\right]$ with f depending on the algorithm. This leads to approximation of the Bayes decision boundary, as well.

The results presented in this work formally support the use of the family of Bregman divergences in vector quantization algorithms, which, because of their developed mathematical theory, can be used, in conjunction with current neural network architectures, in classification and clustering problems, time series analysis, biomedical applications, topological data analysis, and adversarial learning, where the robustness of LVQ methods against adversarial attacks has been promising. On the other hand, Bregman divergences, such as the Kullback-Leibler divergence, are mathematically related to various types of classification errors (via Stein's Theorem), which makes the associated learning algorithms more robust than algorithms based on Euclidean or other metrics for dissimilarity. For this reason, they consist a powerful tool when combined with VQ and LVQ learning algorithms. This aligns with the state of the art deep neural network architectures that are shifting from using the Euclidean measure (the simplest Bregman divergence

which is, at the same time, a metric) towards information-theoretic measures such as the unnormalized Kullback-Leibler divergence.

As a final note, the connection between vector quantization and stochastic approximation algorithms, suggests that further investigation may lead to interesting results on the convergence rates, as well as to the analysis of variants of these algorithms, such as Kohonen's Self Organizing Map.

6. CONCLUSION

In this work, we investigated the convergence of the unsupervised, stochastic vector quantization algorithm, and its supervised counterpart, learning vector quantization, based on Bregman divergences as dissimilarity measures. The convergence properties of the algorithms do not depend on the particular choice of the Bregman divergence, as long as its generating function satisfies the conditions mentioned, but, as expected, are shown to depend on conditions related to both the initialization of the weights and the given sample path. Our results formally support the use of Bregman divergences, such as the Kullback-Leibler divergence, in VQ and LVQ algorithms. The connection between vector quantization and stochastic approximation algorithms, shows that further investigation may lead to interesting results on the convergence rates, as well as to the analysis of variants of these algorithms.

REFERENCES

Banerjee, A., Merugu, S., Dhillon, I.S., and Ghosh, J. (2005). Clustering with bregman divergences. *Journal of machine learning research*, 6(Oct), 1705–1749.

Baras, J.S. and Dey, S. (1999). Combined compression and classification with learning vector quantization. *IEEE Transactions on Information Theory*, 45(6), 1911–1920.

Baras, J.S. and LaVigna, A. (1991). Convergence of a neural network classifier. In *Advances in Neural Information Processing Systems*, 839–845.

Benveniste, A., Métivier, M., and Priouret, P. (2012). *Adaptive algorithms and stochastic approximations*, volume 22. Springer Science & Business Media.

Biehl, M. (2017). Biomedical applications of prototype based classifiers and relevance learning. In *International Conference on Algorithms for Computational Biology*, 3–23. Springer.

Borkar, V.S. (2009). *Stochastic approximation: a dynamical systems viewpoint*, volume 48. Springer.

Bottou, L. (1998). Online learning and stochastic approximations. *On-line learning in neural networks*, 17(9), 142.

Bottou, L. and Bengio, Y. (1995). Convergence properties of the k-means algorithms. In *Advances in neural information processing systems*, 585–592.

Devroye, L., Györfi, L., and Lugosi, G. (2013). *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media.

Duda, R.O., Hart, P.E., and Stork, D.G. (2012). *Pattern classification*. John Wiley & Sons.

Gersho, A. and Gray, R.M. (2012). *Vector quantization and signal compression*, volume 159. Springer Science & Business Media.

Gray, R.M. (1990). Vector quantization. *Readings in speech recognition*, 1(2), 75–100.

Hammer, B. and Villmann, T. (2002). Generalized relevance learning vector quantization. *Neural Networks*, 15(8-9), 1059–1068.

Kirkpatrick, S., Gelatt, C.D., and Vecchi, M.P. (1983). Optimization by simulated annealing. *science*, 220(4598), 671–680.

Kohonen, T. (1995). *Learning Vector Quantization*, 175–189. Springer Berlin Heidelberg, Berlin, Heidelberg.

Melchert, F., Seiffert, U., Biehl, M., Hammer, B., Martinetz, T., and Villmann, T. (2016). Functional approximation for the classification of smooth time series. In *GCPR Workshop on New Challenges in Neural Computation 2016*, 24–31.

Miller, D., Rao, A.V., Rose, K., and Gersho, A. (1996). A global optimization technique for statistical classifier design. *IEEE Transactions on Signal Processing*, 44(12), 3108–3122.

Mwebaze, E., Schneider, P., Schleif, F.M., Aduwo, J.R., Quinn, J.A., Haase, S., Villmann, T., and Biehl, M. (2011). Divergence-based classification in learning vector quantization. *Neurocomputing*, 74(9), 1429–1435.

Nova, D. and Estévez, P.A. (2014). A review of learning vector quantization classifiers. *Neural Computing and Applications*, 25(3-4), 511–524.

Nova, D. and Estévez, P.A. (2016). A study on gmlvq convex and non-convex regularization. In *Advances in Self-Organizing Maps and Learning Vector Quantization*, 305–314. Springer.

Rose, K. (1998). Deterministic annealing for clustering, compression, classification, regression, and related optimization problems. *Proceedings of the IEEE*, 86(11), 2210–2239.

Saralajew, S., Holdijk, L., Rees, M., and Villmann, T. (2018). Prototype-based neural network layers: Incorporating vector quantization. *arXiv preprint arXiv:1812.01214*.

Saralajew, S., Holdijk, L., Rees, M., and Villmann, T. (2019). Robustness of generalized learning vector quantization models against adversarial attacks. *arXiv preprint arXiv:1902.00577*.

Sato, A. and Yamada, K. (1996). Generalized learning vector quantization. In *Advances in neural information processing systems*, 423–429.

Shah, S.A. and Koltun, V. (2018). Deep continuous clustering. *arXiv preprint arXiv:1803.01449*.

Uriarte, E.A. and Martín, F.D. (2005). Topology preservation in som. *International journal of applied mathematics and computer sciences*, 1(1), 19–22.

Villmann, T., Biehl, M., Villmann, A., and Saralajew, S. (2017a). Fusion of deep learning architectures, multilayer feedforward networks and learning vector quantizers for deep classification learning. In *2017 12th International Workshop on Self-Organizing Maps and Learning Vector Quantization, Clustering and Data Visualization (WSOM)*, 1–8. IEEE.

Villmann, T., Bohnsack, A., and Kaden, M. (2017b). Can learning vector quantization be an alternative to svm and deep learning?-recent trends and advanced variants of learning vector quantization for classification learning. *Journal of Artificial Intelligence and Soft Computing Research*, 7(1), 65–81.

Villmann, T. and Haase, S. (2011). Divergence-based vector quantization. *Neural Computation*, 23(5), 1343–1392.

Villmann, T., Haase, S., Schleif, F.M., Hammer, B., and Biehl, M. (2010). The mathematics of divergence based online learning in vector quantization. In *IAPR Workshop on Artificial Neural Networks in Pattern Recognition*, 108–119. Springer.

Wang, J., Wang, K.C., Law, M., Rudzicz, F., and Brudno, M. (2019). Centroid-based deep metric learning for speaker recognition. *arXiv preprint arXiv:1902.02375*.

Zielinski, B., Juda, M., and Zeppelzauer, M. (2018). Persistence codebooks for topological data analysis. *arXiv preprint arXiv:1802.04852*.