Journal of Machine Learning Research 12 (2011) 3431-3466 Submitted 12/10; Revised 11/11; Published 12/11

# Convergence of Distributed Asynchronous Learning Vector Quantization Algorithms

**Benoît Patra**[*] BENOIT.PATRA@UPMC.FR
*LSTA*
*Université Pierre et Marie Curie – Paris VI*
*Tour 15-25, 4 place Jussieu*
*75252 Paris cedex 05, France*

**Editor:** Gabor Lugosi

## Abstract

Motivated by the problem of effectively executing clustering algorithms on very large data sets, we address a model for large scale distributed clustering methods. To this end, we briefly recall some standards on the quantization problem and some results on the almost sure convergence of the competitive learning vector quantization (CLVQ) procedure. A general model for linear distributed asynchronous algorithms well adapted to several parallel computing architectures is also discussed. Our approach brings together this scalable model and the CLVQ algorithm, and we call the resulting technique the distributed asynchronous learning vector quantization algorithm (DALVQ). An in-depth analysis of the almost sure convergence of the DALVQ algorithm is performed. A striking result is that we prove that the multiple versions of the quantizers distributed among the processors in the parallel architecture asymptotically reach a consensus almost surely. Furthermore, we also show that these versions converge almost surely towards the same nearly optimal value for the quantization criterion.

**Keywords:** *k*-means, vector quantization, distributed, asynchronous, stochastic optimization, scalability, distributed consensus

## 1. Introduction

Distributed algorithms arise in a wide range of applications, including telecommunications, distributed information processing, scientific computing, real time process control and many others. Parallelization is one of the most promising ways to harness greater computing resources, whereas building faster serial computers is increasingly expensive and also faces some physical limits such as transmission speeds and miniaturization. One of the challenges proposed for machine learning is to build scalable applications that quickly process large amounts of data in sophisticated ways. Building such large scale algorithms attacks several problems in a distributed framework, such as communication delays in the network or numerous problems caused by the lack of shared memory.

Clustering algorithms are one of the primary tools of unsupervised learning. From a practical perspective, clustering plays an outstanding role in data mining applications such as text mining, web analysis, marketing, medical diagnostics, computational biology and many others. Clustering is a separation of data into groups of similar objects. As clustering represents the data with fewer clusters, there is a necessary loss of certain fine details, but simplification is achieved. The popular

[*] Also at LOKAD SAS, 70 rue Lemercier, 75017 Paris, France, email: benoit.patra@lokad.com.

competitive learning vector quantization (CLVQ) algorithm (see Gersho and Gray, 1992) provides a technique for building reliable clusters characterized by their prototypes. As pointed out by Bottou and Bengio (1995), the CLVQ algorithm can also be viewed as the on-line version of the widespread Lloyd's method (see Lloyd 2003, for the definition) which is referred to as batch *k*-means in Bottou and Bengio (1995). The CLVQ also belongs to the class of stochastic gradient descent algorithms (for more information on stochastic gradient descent procedures we refer the reader to Benveniste et al. 1990).

The analysis of parallel stochastic gradient procedures in a machine learning context has recently received a great deal of attention (see for instance Zinkevich et al. 2009 and McDonald et al. 2010). In the present paper, we go further by introducing a model that brings together the original CLVQ algorithm and the comprehensive theory of asynchronous parallel linear algorithms developed by Tsitsiklis (1984), Tsitsiklis et al. (1986) and Bertsekas and Tsitsiklis (1989). The resulting model will be called distributed asynchronous learning vector quantization (DALVQ for short). At a high level, the DALVQ algorithm parallelizes several executions of the CLVQ method concurrently on different processors while the results of these algorithms are broadcast through the distributed framework asynchronously and efficiently. Here, the term processor refers to any computing instance in a distributed architecture (see Bullo et al. 2009, chap. 1, for more details). Let us remark that there is a series of publications similar in spirit to this paper. Indeed in Frasca et al. (2009) and in Durham et al. (2009), a coverage control problem is formulated as an optimization problem where the functional cost to be minimized is the same of the quantization problem stated in this manuscript.

Let us provide a brief mathematical introduction to the CLVQ technique and DALVQ algorithms. The first technique computes quantization scheme for $d$ dimensional samples $\mathbf{z}_1, \mathbf{z}_2, \ldots$ using the following iterations on a $\left(\mathbb{R}^d\right)^\kappa$ vector,

$$w(t+1) = w(t) - \varepsilon_{t+1} H\left(\mathbf{z}_{t+1}, w(t)\right), \quad t \geq 0.$$

In the equation above, $w(0) \in \left(\mathbb{R}^d\right)^\kappa$ and the $\varepsilon_t$ are positive reals. The vector $H(\mathbf{z}, w)$ is the opposite of the difference between the sample $\mathbf{z}$ and its nearest component in $w$. Assume that there are $M$ computing entities, the data are split among the memory of these machines: $\mathbf{z}_1^i, \mathbf{z}_2^i, \ldots$, where $i \in \{1, \ldots, M\}$. Therefore, the DALVQ algorithms are defined by the $M$ iterations $\{w^i(t)\}_{t=0}^\infty$, called versions, satisfying (with slight simplifications)

$$w^i(t+1) = \sum_{j=1}^M a^{i,j}(t) w^j(\tau^{i,j}(t)) - \varepsilon_{t+1}^i H\left(\mathbf{z}_{t+1}^i, w^i(t)\right), \quad i \in \{1, \ldots, M\} \text{ and } t \geq 0.$$

The time instants $\tau^{i,j}(t) \geq 0$ are deterministic but unknown and the delays satisfy the inequality $t - \tau^{i,j}(t) \geq 0$. The families $\{a^{i,j}(t)\}_{j=1}^M$ define the weights of convex combinations.

As a striking result, we prove that multiple versions of the quantizers, distributed among the processors in a parallel architecture, asymptotically reach a consensus almost surely. Using the materials introduced above, it writes

$$w^i(t) - w^j(t) \xrightarrow[t\to\infty]{} 0, \quad (i,j) \in \{1, \ldots, M\}^2, \text{ almost surely (a.s.).}$$

Furthermore, we also show that these versions converge almost surely towards (the same) nearly optimal value for the quantization criterion. These convergence results are similar in spirit to the most satisfactory almost sure convergence theorem for the CLVQ algorithm obtained by Pagès (1997).

For a given time span, our parallel DALVQ algorithm is able to process much more data than a single processor execution of the CLVQ procedure. Moreover, DALVQ is also asynchronous. This means that local algorithms do not have to wait at preset points for messages to become available. This allows some processors to compute faster and execute more iterations than others, and it also allows communication delays to be substantial and unpredictable. The communication channels are also allowed to deliver messages out of order, that is, in a different order than the one in which they were transmitted. Asynchronism can provide two major advantages. First, a reduction of the synchronization penalty, which could bring a speed advantage over a synchronous execution. Second, for potential industrialization, asynchronism has greater implementation flexibility. Tolerance to system failures and uncertainty can also be increased. As in the case with any on-line algorithm, DALVQ also deals with variable data loads over time. In fact, on-line algorithms avoid tremendous and non scalable batch requests on all data sets. Moreover, with an on-line algorithm, new data may enter the system and be taken into account while the algorithm is already running.

The paper is organized as follows. In Section 2 we review some standard facts on the clustering problem. We extract the relevant material from Pagès (1997) without proof, thus making our exposition self-contained. In Section 3 we give a brief exposition of the mathematical framework for parallel asynchronous gradient methods introduced by Tsitsiklis (1984), Tsitsiklis et al. (1986) and Bertsekas and Tsitsiklis (1989). The results of Blondel et al. (2005) on the asymptotic consensus in asynchronous parallel averaging problems are also recalled. In Section 4, our main results are stated and proved.

## 2. Quantization and CLVQ Algorithm

In this section, we describe the mathematical quantization problem and the CLVQ algorithm. We also recall some convergence results for this technique found by Pagès (1997).

### 2.1 Overview

Let $\mu$ be a probability measure on $\mathbb{R}^d$ with finite second-order moment. The quantization problem consists in finding a "good approximation" of $\mu$ by a set of $\kappa$ vectors of $\mathbb{R}^d$ called quantizer. Throughout the document the $\kappa$ quantization points (or prototypes) will be seen as the components of a $\left(\mathbb{R}^d\right)^\kappa$-dimensional vector $w = (w_1, \ldots, w_\kappa)$. To measure the correctness of a quantization scheme given by $w$, one introduces a cost function called distortion, defined by

$$C_\mu(w) = \frac{1}{2} \int_{\mathbb{R}^d} \min_{1 \le \ell \le \kappa} \|\mathbf{z} - w_\ell\|^2 \, d\mu(\mathbf{z}).$$

Under some minimal assumptions, the existence of an optimal $\left(\mathbb{R}^d\right)^\kappa$-valued quantizer vector $w^\circ \in \operatorname{argmin}_{w \in \left(\mathbb{R}^d\right)^\kappa} C_\mu(w)$ has been established by Pollard (1981) (see also Sabin and Gray 1986, Appendix 2).

In a statistical context, the distribution $\mu$ is only known through $n$ independent random observations $\mathbf{z}_1, \ldots, \mathbf{z}_n$ drawn according to $\mu$. Denote by $\mu_n$ the empirical distribution based on $\mathbf{z}_1, \ldots, \mathbf{z}_n$, that is, for every Borel subset $A$ of $\mathbb{R}^d$

$$\mu_n(A) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{\mathbf{z}_i \in A\}}.$$

Much attention has been devoted to the convergence study of the quantization scheme provided by the empirical minimizers

$$w_n^\circ \in \operatorname*{argmin}_{w \in \left(\mathbb{R}^d\right)^\kappa} C_{\mu_n}(w).$$

The almost sure convergence of $C_\mu\left(w_n^\circ\right)$ towards $\min_{w\in\left(\mathbb{R}^d\right)^\kappa} C_\mu(w)$ was proved by Pollard (1981, 1982a) and Abaya and Wise (1984). Rates of convergence and nonasymptotic performance bounds have been considered by Pollard (1982b), Chou (1994), Linder et al. (1994), Bartlett et al. (1998), Linder (2001, 2000), Antos (2005) and Antos et al. (2005). Convergence results have been established by Biau et al. (2008) where $\mu$ is a measure on a Hilbert space. It turns out that the minimization of the empirical distortion is a computationally hard problem. As shown by Inaba et al. (1994), the computational complexity of this minimization problem is exponential in the number of quantizers $\kappa$ and the dimension of the data $d$. Therefore, exact computations are intractable for most of the practical applications.

Based on this, our goal in this document is to investigate effective methods that produce accurate quantizations with data samples. One of the most popular procedure is Lloyd's algorithm (see Lloyd, 2003) sometimes refereed to as batch $k$-means. A convergence theorem for this algorithm is provided by Sabin and Gray (1986). Another celebrated quantization algorithm is the competitive learning vector quantization (CLVQ), also called on-line $k$-means. The latter acronym outlines the fact that data arrive over time while the execution of the algorithm and their characteristics are unknown until their arrival times. The main difference between the CLVQ and the Lloyd's algorithm is that the latter run in batch training mode. This means that the whole training set is presented before performing an update, whereas the CLVQ algorithm uses each item of the training sequence at each update.

The CLVQ procedure can be seen as a stochastic gradient descent algorithm. In the more general context of gradient descent methods, one cannot hope for the convergence of the procedure towards global minimizers with a non convex objective function (see for instance Benveniste et al. 1990). In our quantization context, the distortion mapping $C_\mu$ is not convex (see for instance Graf and Luschgy 2000). Thus, just as in Lloyd's method, the iterations provided by the CLVQ algorithm converge towards local minima of $C_\mu$.

Assuming that the distribution $\mu$ has a compact support and a bounded density with respect to the Lebesgue measure, Pagès (1997) states a result regarding the almost sure consistency of the CLVQ algorithm towards critical points of the distortion $C_\mu$. The author shows that the set of critical points necessarily contains the global and local optimal quantizers. The main difficulties in the proof arise from the fact that the gradient of the distortion is singular on $\kappa$-tuples having equal components and the distortion function $C_\mu$ is not convex. This explains why standard theories for stochastic gradient algorithm do not apply in this context.

### 2.2 The Quantization Problem, Basic Properties

In the sequel, we denote by $\mathcal{G}$ the closed convex hull of $\operatorname{supp}(\mu)$, where $\operatorname{supp}(\mu)$ stands for the support of the distribution. Observe that, with this notation, the distortion mapping is the function

$C : \left(\mathbb{R}^d\right)^\kappa \longrightarrow [0,\infty)$ defined by

$$C(w) \triangleq \frac{1}{2}\int_{\mathcal{G}} \min_{1\leq \ell \leq \kappa} \|\mathbf{z} - w_\ell\|^2 \, d\mu(\mathbf{z}), \quad w = (w_1,\ldots,w_\kappa) \in \left(\mathbb{R}^d\right)^\kappa.$$

Throughout the document, with a slight abuse of notation, $\|.\|$ means both the Euclidean norm of $\mathbb{R}^d$ or $\left(\mathbb{R}^d\right)^\kappa$. In addition, the notation $\mathcal{D}_*^\kappa$ stands for the set of all vector of $\left(\mathbb{R}^d\right)^\kappa$ with pairwise distinct components, that is,

$$\mathcal{D}_*^\kappa \triangleq \left\{ w \in \left(\mathbb{R}^d\right)^\kappa \mid w_\ell \neq w_k \text{ if and only if } \ell \neq k \right\}.$$

Under some extra assumptions on $\mu$, the distortion function can be rewritten using space partition set called Voronoï tessellation.

**Definition 1** *Let $w \in \left(\mathbb{R}^d\right)^\kappa$, the Voronoï tessellation of $\mathcal{G}$ related to $w$ is the family of open sets $\{W_\ell(w)\}_{1\leq\ell\leq\kappa}$ defined as follows:*

- *If $w \in \mathcal{D}_*^\kappa$, for all $1 \leq \ell \leq \kappa$,*

$$W_\ell(w) = \left\{ v \in \mathcal{G} \;\middle|\; \|w_\ell - v\| < \min_{k\neq\ell} \|w_k - v\| \right\}.$$

- *If $w \in \left(\mathbb{R}^d\right)^\kappa \setminus \mathcal{D}_*^\kappa$, for all $1 \leq \ell \leq \kappa$,*
  - *if $\ell = \min\{k \mid w_k = w_\ell\}$,*

$$W_\ell(w) = \left\{ v \in \mathcal{G} \;\middle|\; \|w_\ell - v\| < \min_{w_k \neq w_\ell} \|w_k - v\| \right\}$$

  - *otherwise, $W_\ell(w) = \emptyset$.*

As an illustration, Figure 1 shows Voronoï tessellations associated to a vector $w$ lying in $([0,1] \times [0,1])^{50}$ whose components have been drawn independently and uniformly. This figure also highlights a remarkable property of the cell borders, which are portions of hyperplanes (see Graf and Luschgy, 2000).

Observe that if $\mu(H)$ is zero for any hyperplane $H$ of $\mathbb{R}^d$ (a property which is sometimes referred to as strong continuity) then using Definition 1, it is easy to see that the distortion takes the form:

$$C(w) = \frac{1}{2}\sum_{\ell=1}^{\kappa} \int_{W_\ell(w)} \|\mathbf{z} - w_\ell\|^2 \, d\mu(\mathbf{z}), \quad w \in \left(\mathbb{R}^d\right)^\kappa.$$

The following assumption will be needed throughout the paper. This assumption is similar to the peak power constraint (see Chou 1994 and Linder 2000). Note that most of the results of this subsection are still valid if $\mu$ satisfies the weaker strong continuity property.

**Assumption 1 (Compact supported density)** *The probability measure $\mu$ has a bounded density with respect to the Lebesgue measure on $\mathbb{R}^d$. Moreover, the support of $\mu$ is equal to its convex hull $\mathcal{G}$, which in turn, is compact.*

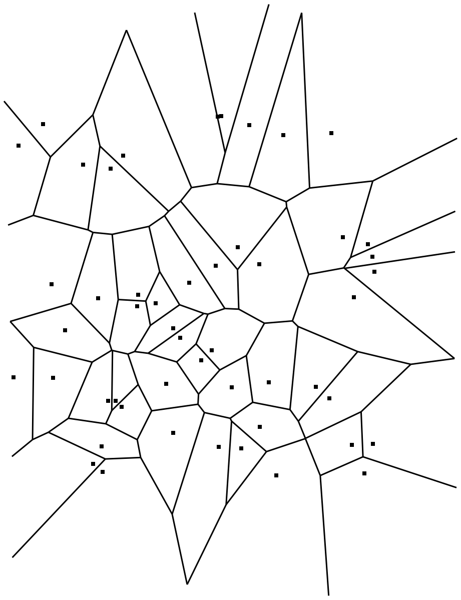

Figure 1: Voronoï tessellation of 50 points of $\mathbb{R}^2$ drawn uniformly in a square.

The next proposition states the differentiability of the distortion $C$, and provides an explicit formula for the gradient $\nabla C$ whenever the distortion is differentiable.

**Proposition 1 (Pagès 1997)** *Under Assumption 1, the distortion $C$ is continuously differentiable at every $w = (w_1, \dots, w_\kappa) \in \mathcal{D}_*^\kappa$. Furthermore, for all $1 \le \ell \le \kappa$,*

$$\nabla_\ell C(w) = \int_{W_\ell(w)} (w_\ell - \mathbf{z}) \, d\mu(\mathbf{z}).$$

Some necessary conditions on the location of the minimizers of $C$ can be derived from its differentiability properties. Therefore, Proposition 2 below states that the minimizers of $C$ have parted components and that they are contained in the support of the density. Thus, the gradient is well defined and these minimizers are necessarily some zeroes of $\nabla C$. For the sequel it is convenient to let $\overset{\circ}{A}$ be the interior of any subset $A$ of $\left(\mathbb{R}^d\right)^\kappa$.

**Proposition 2 (Pagès 1997)** *Under Assumption 1, we have*

$$\operatorname*{argmin}_{w \in \left(\mathbb{R}^d\right)^\kappa} C(w) \subset \operatorname*{argminloc}_{w \in \mathcal{G}^\kappa} C(w) \subset \overset{\circ}{\mathcal{G}^\kappa} \cap \{\nabla C = 0\} \cap \mathcal{D}_*^\kappa,$$

*where* $\mathrm{argminloc}_{w \in \mathcal{G}^\kappa} C(w)$ *stands for the set of local minimizers of $C$ over $\mathcal{G}^\kappa$.*

For any $\mathbf{z} \in \mathbb{R}^d$ and $w \in \left(\mathbb{R}^d\right)^\kappa$, let us define the following vector of $\left(\mathbb{R}^d\right)^\kappa$

$$H(\mathbf{z}, w) \triangleq \left( (w_\ell - \mathbf{z}) \, \mathbb{1}_{\{\mathbf{z} \in W_\ell(w)\}} \right)_{1 \le \ell \le \kappa}. \tag{1}$$

On $\mathcal{D}_*^\kappa$, the function $H$ may be interpreted as an observation of the gradient. With this notation, Proposition 1 states that

$$\nabla C(w) = \int_{\mathcal{G}} H(\mathbf{z}, w) d\mu(\mathbf{z}), \quad w \in \mathcal{D}_*^\kappa.$$

Let $\complement A$ stands for the complementary in $\left(\mathbb{R}^d\right)^\kappa$ of a subset $A \subset \left(\mathbb{R}^d\right)^\kappa$. Clearly, for all $w \in \complement \mathcal{D}_*^\kappa$, the mapping $H(., w)$ is integrable. Therefore, $\nabla C$ can be extended on $\left(\mathbb{R}^d\right)^\kappa$ *via* the formula

$$h(w) \triangleq \int_{\mathcal{G}} H(\mathbf{z}, w) d\mu(\mathbf{z}), \quad w \in \left(\mathbb{R}^d\right)^\kappa. \tag{2}$$

Note however that the function $h$, which is sometimes called the average function of the algorithm, is not continuous.

**Remark 1** *Under Assumption 1, a computation for all $w \in \mathcal{D}_*^\kappa$ of the Hessian matrix $\nabla^2 C(w)$ can be deduced from Theorem 4 of (Fort and Pagès, 1995). In fact, the formula established in this theorem is valid for cost functions which are more complex than $C$ (they are associated to Kohonen Self Organizing Maps, see Kohonen 1982 for more details). In Theorem 4, letting $\sigma(k) = \mathbb{1}_{\{k=0\}}$, provides the result for our distortion $C$. The resulting formula shows that $h$ is singular on $\complement \mathcal{D}_*^\kappa$ and, consequently, that this function cannot be Lipschitz on $\mathcal{G}^\kappa$.*

### 2.3 Convergence of the CLVQ Algorithm

The problem of finding a reliable clustering scheme for a data set is equivalent to find optimal (or at least nearly optimal) minimizers for the mapping $C$. A minimization procedure by a usual gradient descent method cannot be implemented as long as $\nabla C$ is unknown. Thus, the gradient is approximated by a single example extracted from the data. This leads to the following stochastic gradient descent procedure

$$w(t+1) = w(t) - \varepsilon_{t+1} H\left(\mathbf{z}_{t+1}, w(t)\right), \quad t \geq 0, \tag{3}$$

where $w(0) \in \overset{\circ}{\mathcal{G}^\kappa} \cap \mathcal{D}_*^\kappa$ and $\mathbf{z}_1, \mathbf{z}_2 \ldots$ are independent observations distributed according to the probability measure $\mu$.

The algorithm defined by the iterations (3) is known as the CLVQ algorithm in the data analysis community. It is also called the Kohonen Self Organizing Map algorithm with 0 neighbor (see for instance Kohonen 1982) or the on-line $k$-means procedure (see MacQueen 1967 and Bottou 1998) in various fields related to statistics. As outlined by Pagès in Pagès (1997), this algorithm belongs to the class of stochastic gradient descent methods. However, the almost sure convergence of this type of algorithm cannot be obtained by general tools such as Robbins-Monro method (see Robbins and Monro, 1951) or the Kushner-Clark's Theorem (see Kushner and Clark, 1978). Indeed, the main difficulty essentially arises from the non convexity of the function $C$, its non coercivity and the singularity of $h$ at $\complement \mathcal{D}_*^\kappa$ (we refer the reader to Pagès 1997, Section 6, for more details).

The following assumption set is standard in a gradient descent context. It basically upraises constraints on the decreasing speed of the sequence of steps $\{\varepsilon_t\}_{t=0}^\infty$.

**Assumption 2 (Decreasing steps)** *The $(0,1)$-valued sequence $\{\varepsilon_t\}_{t=0}^\infty$ satisfies the following two constraints:*

*1.* $\sum_{t=0}^{\infty} \varepsilon_t = \infty$.

*2.* $\sum_{t=0}^{\infty} \varepsilon_t^2 < \infty$.

An examination of identities (3) and (1) reveals that if $\mathbf{z}_{t+1} \in W_{\ell_0}(w(t))$, where the integer $\ell_0 \in \{1, \ldots, M\}$ then

$$w_{\ell_0}(t+1) = (1-\varepsilon_{t+1})\, w_{\ell_0}(t) + \varepsilon_{t+1}\mathbf{z}_{t+1}.$$

The component $w_{\ell_0}(t+1)$ can be viewed as the image of $w_{\ell_0}(t)$ by a $\mathbf{z}_{t+1}$-centered homothety with ratio $1-\varepsilon_{t+1}$ (Figure 2 provides an illustration of this fact). Thus, under Assumptions 1 and 2, the trajectories of $\{w(t)\}_{t=0}^{\infty}$ stay in $\overset{\circ}{\mathcal{G}}{}^{\kappa} \cap \mathcal{D}_*^{\kappa}$. More precisely, if

$$w(0) \in \overset{\circ}{\mathcal{G}}{}^{\kappa} \cap \mathcal{D}_*^{\kappa}$$

then

$$w(t) \in \overset{\circ}{\mathcal{G}}{}^{\kappa} \cap \mathcal{D}_*^{\kappa}, \quad t \geq 0, \text{ a.s.}$$

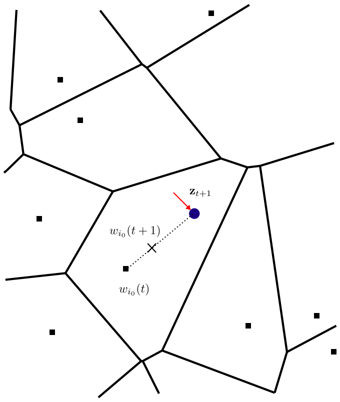

Figure 2: Drawing of a portion of a 2-dimensional Voronoï tessellation. For $t \geq 0$, if the vector $\mathbf{z}_{t+1} \in W_{\ell_0}(w(t))$ then $w_\ell(t+1) = w_\ell(t)$ for all $\ell \neq \ell_0$ and $w_{\ell_0}(t+1)$ lies in the segment $[w_{\ell_0}(t), \mathbf{z}_{t+1}]$. The update of the vector $w_{\ell_0}(t)$ can also be viewed as a $\mathbf{z}_{t+1}$-centered homothety with ratio $1-\varepsilon_{t+1}$.

Although $\nabla C$ is not continuous some regularity can be obtained. To this end, we need to introduce the following materials. For any $\delta > 0$ and any compact set $L \subset \mathbb{R}^d$, let the compact set $L_\delta^\kappa \subset \left(\mathbb{R}^d\right)^\kappa$ be defined as

$$L_\delta^\kappa \triangleq \left\{ w \in L^\kappa \mid \min_{k \neq \ell} \|w_\ell - w_k\| \geq \delta \right\}. \tag{4}$$

The next lemma that states on the regularity of $\nabla C$ will prove to be extremely useful in the proof of Theorem 4 and throughout Section 4.

**Lemma 2 (Pagès 1997)** *Assume that $\mu$ satisfies Assumption 1 and let $L$ be a compact set of $\mathbb{R}^d$. Then, there is some constant $P_\delta$ such that for all $w$ and $v$ in $L_\delta^\kappa$ with $[w,v] \subset \mathcal{D}_*^\kappa$,*

$$\|\nabla C(w) - \nabla C(v)\| \leq P_\delta \|w - v\|.$$

The following lemma, called G-lemma in Pagès (1997) is an easy-to-apply convergence results on stochastic algorithms. It is particularly adapted to the present situation of the CLVQ algorithm where the average function of the algorithm $h$ is singular.

**Theorem 3 (G-lemma, Fort and Pagès 1996)** *Assume that the iterations (3) of the CLVQ algorithm satisfy the following conditions:*

1. *$\sum_{t=1}^{\infty} \varepsilon_t = \infty$ and $\varepsilon_t \xrightarrow[t\to\infty]{} 0$.*
2. *The sequences $\{w(t)\}_{t=0}^{\infty}$ and $\{h(w(t))\}_{t=0}^{\infty}$ are bounded a.s.*
3. *The series $\sum_{t=0}^{\infty} \varepsilon_{t+1} (H(\mathbf{z}_{t+1}, w(t)) - h(w(t)))$ converge a.s. in $\left(\mathbb{R}^d\right)^\kappa$.*
4. *There exists a lower semi-continuous function $G : \left(\mathbb{R}^d\right)^\kappa \longrightarrow [0,\infty)$ such that*

$$\sum_{t=0}^{\infty} \varepsilon_{t+1} G(w(t)) < \infty, \quad \textit{a.s.}$$

*Then, there exists a random connected component $\Xi$ of $\{G = 0\}$ such that*

$$\operatorname{dist}(w(t), \Xi) \xrightarrow[t\to\infty]{} 0, \quad \textit{a.s.},$$

*where the symbol* dist *denotes the usual distance function between a vector and a subset of $\left(\mathbb{R}^d\right)^\kappa$. Note also that if the connected components of $\{G = 0\}$ are singletons then there exists $\xi \in \{G = 0\}$ such that $w(t) \xrightarrow[t\to\infty]{} \xi$ a.s.*

For a definition of the topological concept of connected component, we refer the reader to Choquet (1966). The interest of the G-lemma depends upon the choice of $G$. In our context, a suitable lower semi-continuous function is $\widehat{G}$ defined by

$$\widehat{G}(w) \triangleq \liminf_{v \in \mathcal{G}^\kappa \cap \mathcal{D}_*^\kappa,\, v \to w} \|\nabla C(v)\|^2, \qquad w \in \mathcal{G}^\kappa. \tag{5}$$

The next theorem is, as far as we know, the first almost sure convergence theorem for the stochastic algorithm CLVQ.

**Theorem 4 (Pagès 1997)** *Under Assumptions 1 and 2, conditioned on the event*

$$\left\{\liminf_{t\to\infty} \operatorname{dist}\left(w(t), \complement\mathcal{D}_*^\kappa\right) > 0\right\}, \quad \textit{one has}$$

$$\operatorname{dist}(w(t), \Xi_\infty) \xrightarrow[t\to\infty]{} 0, \quad \textit{a.s.},$$

*where $\Xi_\infty$ is some random connected component of $\{\nabla C = 0\}$.*

The proof is an application of the above G-lemma with the mapping $\widehat{G}$ defined by Equation (5). Theorem 4 states that the iterations of the CLVQ necessarily converge towards some critical points (zeroes of $\nabla C$). From Proposition 2 we deduce that the set of critical points necessarily contains optimal quantizers. Recall that without more assumption than $w(0) \in \overset{\circ}{\mathcal{G}}{}^{\kappa} \cap \mathcal{D}_*^{\kappa}$, we have already discussed the fact that the components of $w(t)$ are almost surely parted for all $t \geq 0$. Thus, it is easily seen that the two following events only differ on a set of zero probability

$$\left\{ \liminf_{t\to\infty} \operatorname{dist}\left(w(t), \complement\mathcal{D}_*^{\kappa}\right) > 0 \right\}$$

and

$$\left\{ \inf_{t\geq 0} \operatorname{dist}\left(w(t), \complement\mathcal{D}_*^{\kappa}\right) > 0 \right\}.$$

Some results are provided by Pagès (1997) for asymptotically stuck components but, as pointed out by the author, they are less satisfactory.

## 3. General Distributed Asynchronous Algorithm

We present in this section some materials and results of the asynchronous parallel linear algorithms theory.

### 3.1 Model Description

Let $s(t)$ be any $\left(\mathbb{R}^d\right)^{\kappa}$-valued vector and consider the following iterations

$$w(t+1) = w(t) + s(t), \quad t \geq 0. \tag{6}$$

Here, the model of discrete time described by iterations (6) can only be performed by a single computing entity. Therefore, if the computations of the vectors $s(t)$ are relatively time consuming then not many iterations can be achieved for a given time span. Consequently, a parallelization of this computing scheme should be investigated. The aim of this section is to discuss a precise mathematical description of a distributed asynchronous model for the iterations (6). This model for distributed computing was originally proposed by Tsitsiklis et al. (1986) and was revisited in Bertsekas and Tsitsiklis (1989, Section 7.7).

Assume that we dispose of a distributed architecture with $M$ computing entities called processors (or agents, see for instance Bullo et al. 2009). Each processor is labeled, for simplicity of notation, by a natural number $i \in \{1,\ldots,M\}$. Throughout the paper, we will add the superscript $i$ on the variables possessed by the processor $i$. In the model we have in mind, each processor has a buffer where its current version of the iterated vector is kept, that is, local memory. Thus, for agent $i$ such iterations are represented by the $\left(\mathbb{R}^d\right)^{\kappa}$-valued sequence $\left\{w^i(t)\right\}_{t=0}^{\infty}$.

Let $t \geq 0$ denote the current time. For any pair of processors $(i,j) \in \{1,\ldots,M\}^2$, the value kept by agent $j$ and available for agent $i$ at time $t$ is not necessarily the most recent one, $w^j(t)$, but more probably and outdated one, $w^j(\tau^{i,j}(t))$, where the deterministic time instant $\tau^{i,j}(t)$ satisfy $0 \leq \tau^{i,j}(t) \leq t$. Thus, the difference $t - \tau^{i,j}(t)$ can be seen as a communication delay. This is a modeling of some aspects of the network: latency and bandwidth finiteness.

We insist on the fact that there is a distinction between "global" and "local" time. The time variable we refer above to as $t$ corresponds to a global clock. Such a global clock is needed only for analysis purposes. The processors work without knowledge of this global clock. They have access to a local clock or to no clock at all.

The algorithm is initialized at $t = 0$, where each processor $i \in \{1,\ldots,M\}$ has an initial version $w^i(0) \in \left(\mathbb{R}^d\right)^\kappa$ in its buffer. We define the general distributed asynchronous algorithm by the following iterations

$$w^i(t+1) = \sum_{j=1}^{M} a^{i,j}(t) w^j(\tau^{i,j}(t)) + s^i(t), \quad i \in \{1,\ldots,M\} \text{ and } t \geq 0. \tag{7}$$

The model can be interpreted as follows: at time $t \geq 0$, processor $i$ receives messages from other processors containing $w^j(\tau^{i,j}(t))$. Processor $i$ incorporates these new vectors by forming a convex combination and incorporates the vector $s^i(t)$ resulting from its own "local" computations. The coefficients $a^{i,j}(t)$ are nonnegative numbers which satisfy the constraint

$$\sum_{j=1}^{M} a^{i,j}(t) = 1, \quad i \in \{1,\ldots,M\} \text{ and } t \geq 0. \tag{8}$$

As the combining coefficients $a^{i,j}(t)$ depend on $t$, the network communication topology is sometimes referred to as time-varying. The sequences $\left\{\tau^{i,j}(t)\right\}_{t=0}^{\infty}$ need not to be known in advance by any processor. In fact, their knowledge is not required to execute iterations defined by Equation (7). Thus, we do not necessary dispose of a shared global clock or synchronized local clocks at the processors.

As for now the descent terms $\left\{s^i(t)\right\}_{t=0}^{\infty}$ will be arbitrary $\left(\mathbb{R}^d\right)^\kappa$-valued sequences. In Section 4, when we define the distributed asynchronous learning vector quantization (DALVQ), the definition of the descent terms will be made more explicit.

### 3.2 The Agreement Algorithm

This subsection is devoted to a short survey of the results, found by Blondel et al. (2005), for a natural simplification of the general distributed asynchronous algorithm (7). This simplification is called agreement algorithm by Blondel et al. and is defined by

$$x^i(t+1) = \sum_{j=1}^{M} a^{i,j}(t) x^j(\tau^{i,j}(t)), \quad i \in \{1,\ldots,M\} \text{ and } t \geq 0. \tag{9}$$

where $x^i(0) \in \left(\mathbb{R}^d\right)^\kappa$. An observation of these equations reveals that they are similar to iterations (7), the only difference being that all descent terms equal 0.

In order to analyse the convergence of the agreement algorithm (9), Blondel et al. (2005) define two sets of assumptions that enforce some weak properties on the communication delays and the network topology. As shown in Blondel et al. (2005), if the assumptions contained in one of these two set hold, then the distributed versions of the agreement algorithm, namely the $x^i$, reach an asymptotical consensus. This latter statement means that there exists a vector $x^\star$ (independent of $i$) such that

$$x^i(t) \xrightarrow[t\to\infty]{} x^\star, \quad i \in \{1,\ldots,M\}.$$

The agreement algorithm (9) is essentially driven by the communication times $\tau^{i,j}(t)$ assumed to be deterministic but do not need to be known *a priori* by the processors. The following Assumption

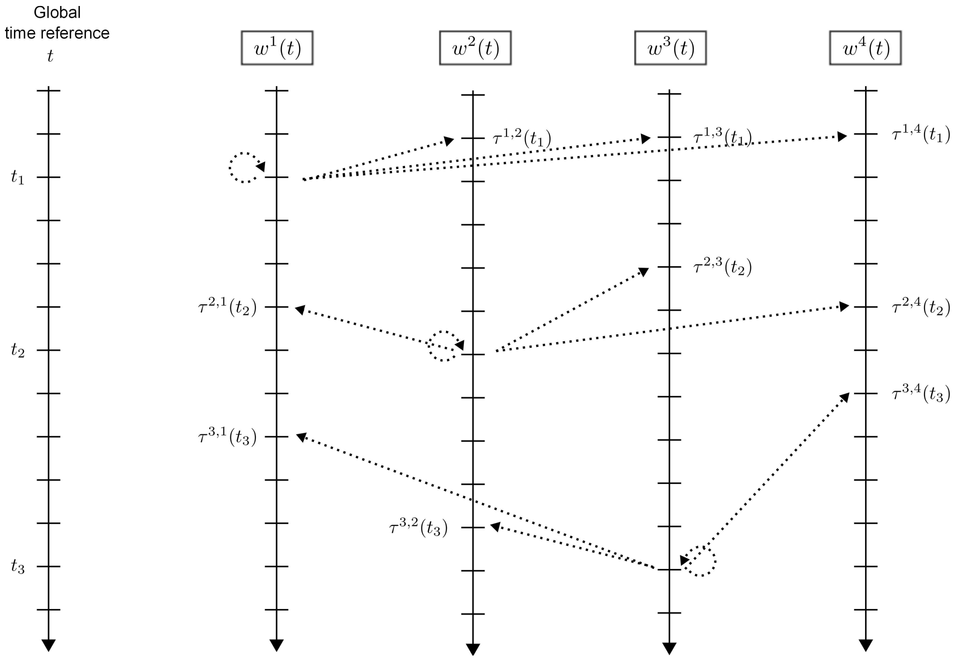

Figure 3: Illustration of the time delays introduced in the general distributed asynchronous algorithm. Here, there are $M = 4$ different processors with their own computations of the vectors $w^{(i)}$, $i \in \{1,2,3,4\}$. Three arbitrary values of the global time $t$ are represented ($t_1$, $t_2$ and $t_3$), with $\tau^{i,i}(t_k) = t_k$ for all $i \in \{1,2,3,4\}$ and $1 \leq k \leq 3$. The dashed arrows head towards the versions available at time $t_k$ for an agent $i \in \{1,2,3,4\}$ represented by the tail of the arrow.

3 essentially ensures, in its third statement, that the communication delays $t - \tau^{i,j}(t)$ are bounded. This assumption prevents some processor from taking into account some arbitrarily old values computed by others processors. Assumption 3 1. is just a convention: when $a^{i,j}(t) = 0$ the value $\tau^{i,j}(t)$ has no effect on the update. Assumption 3 2. is rather natural because processors have access to their own most recent value.

**Assumption 3 (Bounded communication delays)**

1. *If* $a^{i,j}(t) = 0$ *then one has* $\tau^{i,j}(t) = t, \quad (i,j) \in \{1,\ldots,M\}^2$ *and* $t \geq 0$,

2. $\tau^{i,i}(t) = t, \quad i \in \{1,\ldots,M\}$ *and* $t \geq 0$.

3. *There exists a positive integer* $B_1$ *such that*

$$t - B_1 < \tau^{i,j}(t) \leq t, \quad (i,j) \in \{1,\ldots,M\}^2 \textit{ and } t \geq 0.$$

The next Assumption 4 states that the value possessed by agent $i$ at time $t+1$, namely $x^i(t+1)$, is a weighted average of its own value and the values that it has just received from other agents.

**Assumption 4 (Convex combination and threshold)** *There exists a positive constant* $\alpha > 0$ *such that the following three properties hold:*

1. $a^{i,i}(t) \geq \alpha, \quad i \in \{1,\ldots,M\}$ *and* $t \geq 0$.
2. $a^{i,j}(t) \in \{0\} \cup [\alpha,1], \quad (i,j) \in \{1,\ldots,M\}^2$ *and* $t \geq 0$.
3. $\sum_{j=1}^{M} a^{i,j}(t) = 1, \quad i \in \{1,\ldots,M\}$ *and* $t \geq 0$.

Let us mention one particular relevant case for the choice of the combining coefficients $a^{i,j}(t)$. Let $i \in \{1,\ldots,M\}$ and $t \geq 0$, the set

$$N^i(t) \triangleq \left\{ j \in \{1,\ldots,M\} \in \{1,\ldots,M\} \mid a^{i,j}(t) \neq 0 \right\}$$

corresponds to the set of agents whose version is taken into account by processor $i$ at time $t$. For all $(i,j) \in \{1,\ldots,M\}^2$ and $t \geq 0$, the weights $a^{i,j}(t)$ are defined by

$$a^{i,j}(t) = \begin{cases} 1/\#N^i(t) & \text{if } j \in N^i(t); \\ 0 & \text{otherwise;} \end{cases}$$

where $\#A$ denotes the cardinal of any finite set $A$. The above definition on the combining coefficients appears to be relevant for practical implementations of the model DALVQ introduced in Section 4. For a discussion on others special interest cases regarding the choices of the coefficients $a^{i,j}(t)$ we refer the reader to Blondel et al. (2005).

The communication patterns, sometimes refereed to as the network communication topology, can be expressed in terms of directed graph. For a thorough introduction to graph theory, (see Jungnickel, 1999).

**Definition 5 (Communication graph)** *Let us fix* $t \geq 0$*, the communication graph at time* $t$*,* $(\mathcal{V}, E(t))$*, is defined by*

- *the set of vertices* $\mathcal{V}$ *is formed by the set of processors* $\mathcal{V} = \{1,\ldots,M\}$*,*
- *the set of edges* $E(t)$ *is defined via the relationship*

$$(j,i) \in E(t) \text{ if and only if } a^{i,j}(t) > 0.$$

Assumption 5 is a minimal condition required for a consensus among the processors. More precisely, it states that for any pair of agents $(i,j) \in \{1,\ldots,M\}^2$ there is a sequence of communications where the values computed by agent $i$ will influence (directly or indirectly) the future values kept by agent $j$.

**Assumption 5 (Graph connectivity)** *The graph* $(\mathcal{V}, \cup_{s \geq t} E(s))$ *is strongly connected for all* $t \geq 0$.

Finally, we define two supplementary assumptions. The combination of one of the two following assumptions with the three previous ones will ensure the convergence of the agreement algorithm. As mentioned above, if Assumption 5 holds then there is a communication path between any pair of agents. Assumption 6 below expresses the fact that there is a finite upper bound for the length of such paths.

**Assumption 6 (Bounded communication intervals)** *If $i$ communicates with $j$ an infinite number of times then there is a positive integer $B_2$ such that*

$$(i,j) \in E(t) \cup E(t+1) \cup \ldots \cup E(t+B_2-1), \quad t \geq 0.$$

Assumption 7 is a symmetry condition: if agent $i \in \{1,\ldots,M\}$ communicates with agent $j \in \{1,\ldots,M\}$ then $j$ has communicated or will communicate with $i$ during the time interval $(t-B_3, t+B_3)$ where $B_3 > 0$.

**Assumption 7 (Symmetry)** *There exists some $B_3 > 0$ such that whenever the pair $(i,j) \in E(t)$, there exists some $\tau$ that satisfies $|t-\tau| < B_3$ and $(j,i) \in E(\tau)$.*

To shorten the notation, we set

$$(\mathbf{AsY})_{\mathbf{1}} \equiv \begin{cases} \text{Assumption 3;} \\ \text{Assumption 4;} \\ \text{Assumption 5;} \\ \text{Assumption 6.} \end{cases} \qquad (\mathbf{AsY})_{\mathbf{2}} \equiv \begin{cases} \text{Assumption 3;} \\ \text{Assumption 4;} \\ \text{Assumption 5;} \\ \text{Assumption 7;} \end{cases}$$

We are now in a position to state the main result of this section. The Theorem 6 expresses the fact that, for the agreement algorithm, a consensus is asymptotically reached by the agents.

**Theorem 6 (Blondel et al. 2005)** *Under the set of Assumptions $(\mathbf{AsY})_{\mathbf{1}}$ or $(\mathbf{AsY})_{\mathbf{2}}$, there is a consensus vector $x^{\star} \in \left(\mathbb{R}^d\right)^{\kappa}$ (independent of i) such that*

$$\lim_{t\to\infty} \left\| x^i(t) - x^{\star} \right\| = 0, \quad i \in \{1,\ldots,M\}.$$

*Besides, there exist $\rho \in [0,1)$ and $L > 0$ such that*

$$\left\| x^i(t) - x^i(\tau) \right\| \leq L\rho^{t-\tau}, \quad i \in \{1,\ldots,M\} \text{ and } t \geq \tau \geq 0.$$

### 3.3 Asymptotic Consensus

This subsection is devoted to the analysis of the general distributed asynchronous algorithm (7). For this purpose, the study of the agreement algorithm defined by Equations (9) will be extremely fruitful. The following lemma states that the version possessed by agent $i \in \{1,\ldots,M\}$ at time $t \geq 0$, namely $w^i(t)$, depends linearly on the others initialization vectors $w^j(0)$ and the descent subsequences $\left\{s^j(\tau)\right\}_{\tau=-1}^{t-1}$, where $j \in \{1,\ldots,M\}$.

**Lemma 7 (Tsitsiklis 1984)** *For all $(i,j) \in \{1,\ldots,M\}^2$ and $t \geq 0$, there exists a real-valued sequence $\left\{\phi^{i,j}(t,\tau)\right\}_{\tau=-1}^{t-1}$ such that*

$$w^i(t) = \sum_{j=1}^{M} \phi^{i,j}(t,-1)\, w^j(0) + \sum_{\tau=0}^{t-1} \sum_{j=1}^{M} \phi^{i,j}(t,\tau)\, s^j(\tau).$$

For all $(i,j) \in \{1,\ldots,M\}^2$ and $t \geq 0$, the real-valued sequences $\left\{\phi^{i,j}(t,\tau)\right\}_{\tau=-1}^{t-1}$ do not depend on the value taken by the descent terms $s^i(t)$. The real numbers $\phi^{i,j}(t,\tau)$ are determined by the sequences $\left\{\tau^{i,j}(\tau)\right\}_{\tau=0}^{t}$ and $\left\{a^{i,j}(\tau)\right\}_{\tau=0}^{t}$ which do not depend on $w$. These last sequences are unknown in general, but some useful qualitative properties can be derived, as expressed in Lemma 8 below.

**Lemma 8 (Tsitsiklis 1984)** *For all* $(i,j) \in \{1,\ldots,M\}^2$*, let* $\left\{\phi^{i,j}(t,\tau)\right\}_{\tau=-1}^{t-1}$ *be the sequences defined in Lemma 7.*

1. *Under Assumption 4,*

$$0 \leq \phi^{i,j}(t,\tau) \leq 1, \quad (i,j) \in \{1,\ldots,M\}^2 \text{ and } t > \tau \geq -1.$$

2. *Under Assumptions* $(\mathbf{AsY})_\mathbf{1}$ *or* $(\mathbf{AsY})_\mathbf{2}$*, we have:*

   (a) *For all* $(i,j) \in \{1,\ldots,M\}^2$ *and* $\tau \geq -1$*, the limit of* $\phi^{i,j}(t,\tau)$ *as* $t$ *tends to infinity exists and is independent of* $j$*. It will be denoted* $\phi^i(\tau)$*.*

   (b) *There exists some* $\eta > 0$ *such that*

$$\phi^i(\tau) > \eta, \quad i \in \{1,\ldots,M\} \text{ and } \tau \geq -1.$$

   (c) *There exist a constant* $A > 0$ *and* $\rho \in (0,1)$ *such that*

$$\left|\phi^{i,j}(t,\tau) - \phi^i(\tau)\right| \leq A\rho^{t-\tau}, \quad (i,j) \in \{1,\ldots,M\}^2 \text{ and } t > \tau \geq -1.$$

Take $t' \geq 0$ and assume that the agents stop performing update after time $t'$, but keep communicating and merging the results. This means that $s^j(t) = 0$ for all $t \geq t'$. Then, Equations (7) write

$$w^i(t+1) = \sum_{j=1}^{M} a^{i,j}(t) w^j\left(\tau^{i,j}(t)\right), \quad i \in \{1,\ldots,M\} \text{ and } t \geq t'.$$

If Assumptions $(\mathbf{AsY})_\mathbf{1}$ or $(\mathbf{AsY})_\mathbf{2}$ are satisfied then Theorem 6 shows that there is a consensus vector, depending on the time instant $t'$. This vector will be equal to $w^\star(t')$ defined below (see Figure 4). Lemma 8 provides a good way to define the sequence $\{w^\star(t)\}_{t=0}^{\infty}$ as shown in Definition 9. Note that this definition does not involve any assumption on the descent terms.

**Definition 9 (Agreement vector)** *Assume that Assumptions* $(\mathbf{AsY})_\mathbf{1}$ *or* $(\mathbf{AsY})_\mathbf{2}$ *are satisfied. The agreement vector sequence* $\{w^\star(t)\}_{t=0}^{\infty}$ *is defined by*

$$w^\star(t) \triangleq \sum_{j=1}^{M} \phi^j(-1) w^j(0) + \sum_{\tau=0}^{t-1} \sum_{j=1}^{M} \phi^j(\tau) s^j(\tau), \quad t \geq 0.$$

It is noteworthy that the agreement vector sequence $w^\star$ satisfies the following recursion formula

$$w^\star(t+1) = w^\star(t) + \sum_{j=1}^{M} \phi^j(t) s^j(t), \quad t \geq 0. \tag{10}$$

## 4. Distributed Asynchronous Learning Vector Quantization

This section is devoted to the distributed asynchronous learning vector quantization techniques. We provide a definition and investigate the almost sure convergence properties of the techniques.

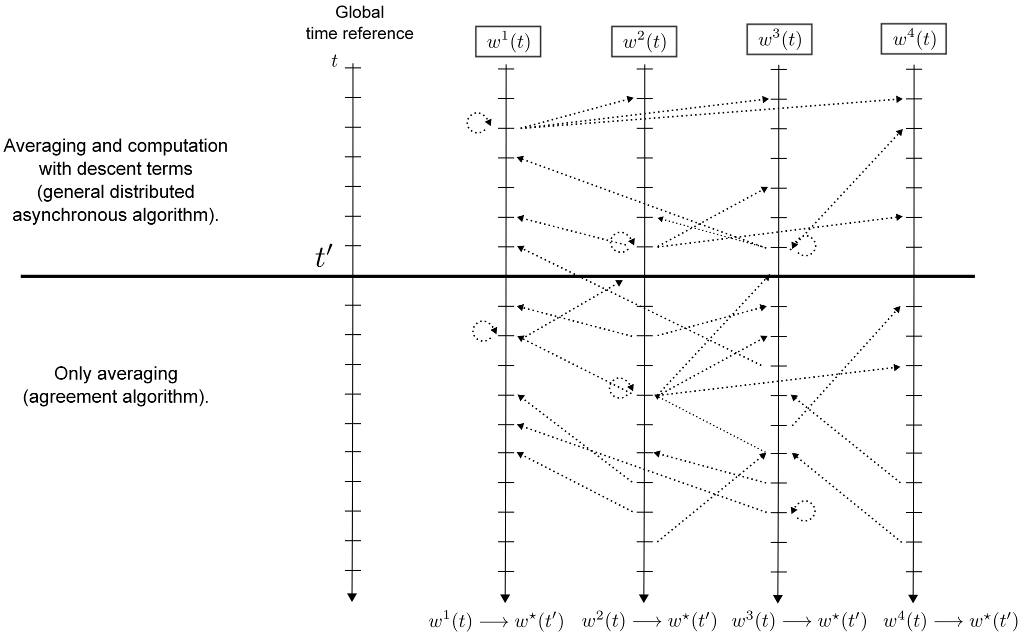
Global
time reference
$t$
$w^1(t)$
$w^2(t)$
$w^3(t)$
$w^4(t)$
Averaging and computation
with descent terms
(general distributed
asynchronous algorithm).
$t'$
Only averaging
(agreement algorithm).
$w^1(t) \longrightarrow w^\star(t')$
$w^2(t) \longrightarrow w^\star(t')$
$w^3(t) \longrightarrow w^\star(t')$
$w^4(t) \longrightarrow w^\star(t')$

Figure 4: The agreement vector at time $t'$, $w^\star(t')$ corresponds to the common value asymptotically achieved by all processors if computations integrating descent terms have stopped after $t'$, that is, $s^j(t) = 0$ for all $t \geq t'$.

### 4.1 Introduction, Model Presentation

From now on, and until the end of the paper, we assume that one of the two set of Assumptions $(\mathbf{AsY})_\mathbf{1}$ or $(\mathbf{AsY})_\mathbf{2}$ holds, as well as the compact-supported density Assumption 1. In addition, we will also assume that $0 \in \mathcal{G}$. For the sake of clarity, all the proofs of the main theorems as well as the lemmas needed for these proofs have been postponed at the end of the paper, in Annex.

Tsitsiklis (1984), Tsitsiklis et al. (1986) and Bertsekas and Tsitsiklis (1989) studied distributed asynchronous stochastic gradient optimization algorithms. In this series of publications, for the distributed minimization of a cost function $F : \left(\mathbb{R}^d\right)^\kappa \longrightarrow \mathbb{R}$, the authors considered the general distributed asynchronous algorithm defined by Equation (7) with specific choices for stochastic descent terms $s^i$. Using the notation of Section 3, the algorithm writes

$$w^i(t+1) = \sum_{j=1}^{M} a^{i,j}(t) w^j(\tau^{i,j}(t)) + s^i(t), \quad i \in \{1,\ldots,M\} \text{ and } t \geq 0,$$

with stochastic descent terms $s^i(t)$ satisfying

$$\mathbb{E}\left\{s^i(t) \mid s^j(\tau),\ j \in \{1,\ldots,M\} \text{ and } t > \tau \geq 0\right\} = -\varepsilon^i_{t+1} \nabla F\left(w^i(t)\right), \qquad i \in \{1,\ldots,M\} \text{ and } t \geq 0. \tag{11}$$

where $\left\{\varepsilon^i_t\right\}_{t=0}^\infty$ are decreasing steps sequences. The definition of the descent terms in Bertsekas and Tsitsiklis (1989) and Tsitsiklis et al. (1986) is more general than the one appearing in Equation

(11). We refer the reader to Assumption 3.2 and 3.3 in Tsitsiklis et al. (1986) and Assumption 8.2 in Bertsekas and Tsitsiklis (1989) for the precise definition of the descent terms. As discussed in Section 2, the CLVQ algorithm is also a stochastic gradient descent procedure. Unfortunately, the results from Tsitisklis et al. do not apply with our distortion function, $C$, since the authors assume that $F$ is continuously differentiable and $\nabla F$ is Lipschitz. Therefore, the aim of this section is to extend the results of Tsitsiklis et al. to the context of vector quantization and on-line clustering.

We first introduce the distributed asynchronous learning vector quantization (DALVQ) algorithm. To prove its almost sure consistency, we will need an asynchronous G-lemma, which is inspired from the G-lemma, Theorem 3, presented in Section 2. This theorem may be seen as an easy-to-apply tool for the almost sure consistency of a distributed asynchronous system where the average function is not necessary regular. Our approach sheds also some new light on the convergence of distributed asynchronous stochastic gradient descent algorithms. Precisely, Proposition 8.1 in Tsitsiklis et al. (1986) claims that the next asymptotic equality holds: $\liminf_{t\to\infty} \left\|\nabla F(w^i(t))\right\| = 0$, while our main Theorem 12 below states that $\lim_{t\to\infty} \left\|\nabla C(w^i(t))\right\| = 0$. However, there is a price to pay for this more precise result with the non Lipschitz gradient $\nabla C$. Similarly to Pagès (1997), who assumes that the trajectory of the CLVQ algorithm has almost surely asymptotically parted components (see Theorem 4 in Section 2), we will suppose that the agreement vector sequence has, almost surely, asymptotically parted component trajectories.

Recall that the goal of the DALVQ is to provide a well designed distributed algorithm that processes quickly (in term of wall clock time) very large data sets to produce accurate quantization. The data sets (or streams of data) are distributed among several queues sending data to the different processors of our distributed framework. Thus, in this context the sequence $\mathbf{z}_1^i, \mathbf{z}_2^i, \ldots$ stands for the data available for processor, where $i \in \{1, \ldots, M\}$. The random variables

$$\mathbf{z}_1^1, \mathbf{z}_2^1, \ldots, \mathbf{z}_1^2, \mathbf{z}_2^2, \ldots$$

are assumed to be independent and identically distributed according to $\mu$.

In the definition of the CLVQ procedure (3), the term $H\left(\mathbf{z}_{t+1}, w(t)\right)$ can be seen as an observation of the gradient $\nabla C\left(w(t)\right)$. Therefore, in our DALVQ algorithm, each processor $i \in \{1, \ldots, M\}$ is able to compute such observations using its own data $\mathbf{z}_1^i, \mathbf{z}_2^i, \ldots$. Thus, the DALVQ procedure is defined by Equation (7) with the following choice for the descent term $s^i$:

$$s^i(t) = \begin{cases} -\varepsilon_{t+1}^i H\left(\mathbf{z}_{t+1}^i, w^i(t)\right) & \text{if } t \in T^i; \\ 0 & \text{otherwise;} \end{cases} \tag{12}$$

where $\left\{\varepsilon_t^i\right\}_{t=0}^{\infty}$ are $(0,1)$-valued sequences. The sets $T^i$ contain the time instants where the version $w^i$, kept by processor $i$, is updated with the descent terms. This fine grain description of the algorithm allows some processors to be idle for computing descent terms (when $t \notin T^i$). This reflects the fact that the computing operations might not take the same time for all processors, which is precisely the core of asynchronous algorithms analysis. Similarly to time delays and combining coefficients, the sets $T^i$ are supposed to be deterministic but do not need to be known *a priori* for the execution of the algorithm.

In the DALVQ model, randomness arises from the data $\mathbf{z}$. Therefore, it is natural to let $\{\mathcal{F}_t\}_{t=0}^{\infty}$ be the filtration built on the $\sigma$-algebras

$$\mathcal{F}_t \triangleq \sigma\left(\mathbf{z}_s^i,\ i \in \{1, \ldots, M\} \text{ and } t \geq s \geq 0\right), \quad t \geq 0.$$

An easy verification shows that, for all $j \in \{1,\ldots,M\}$ and $t \geq 0$, $w^{\star}(t)$ and $w^{j}(t)$ are $\mathcal{F}_t$-measurable random variables.

For simplicity, the assumption on the decreasing speed of the sequences $\left\{\varepsilon_t^i\right\}_{t=0}^{\infty}$ is strengthened as follows. The notation $a \vee b$ stands for the maximum of two reals $a$ and $b$.

**Assumption 8** *There exist two real numbers $K_1 > 0$ and $K_2 \geq 1$ such that*

$$\frac{K_1}{t \vee 1} \leq \varepsilon_{t+1}^i \leq \frac{K_2}{t \vee 1}, \quad i \in \{1,\ldots,M\} \textit{ and } t \geq 0.$$

If Assumption 8 holds then the sequences $\left\{\varepsilon_t^i\right\}_{t=0}^{\infty}$ satisfy the standard Assumption 2 for stochastic optimization algorithms. Note that the choice of steps proportional to $1/t$ has been proved to be a satisfactory learning rate, theoretically speaking and also for practical implementations (see for instance Murata 1998 and Bottou and LeCun 2004).

For practical implementation, the sequences $\left\{\varepsilon_{t+1}^i\right\}_{t=0}^{\infty}$ satisfying Assumption 8 can be implemented without a global clock, that is, without assuming that the current value of $t$ is known by the agents. This assumption is satisfied, for example, by taking the current value of $\varepsilon_t^i$ proportional to $1/n_t^i$, where $n_t^i$ is the number of times that processor $i$ as performed an update, that is, the cardinal of the set $T^i \cap \{0,\ldots,t\}$. For a given processor, if the time span between consecutive updates is bounded from above and from below, a straightforward examination shows that the sequence of steps satisfy Assumption 8.

Finally, the next assumption is essentially technical in nature. It enables to avoid time instants where all processors are idle. It basically requires that, at any time $t \geq 0$, there is at least one processor $i \in \{1,\ldots,M\}$ satisfying $s^i(t) \neq 0$.

**Assumption 9** *One has $\sum_{j=1}^{M} \mathbb{1}_{\{t \in T^j\}} \geq 1$ for all $t \geq 0$.*

### 4.2 The Asynchronous G-lemma

The aim of this subsection is to state a useful theorem similar to Theorem 3, but adapted to our asynchronous distributed context. The precise Definition 9 of the agreement vector sequence should not cast aside the intuitive definition. The reader should keep in mind that the vector $w^{\star}(t)$ is also the asymptotical consensus if descent terms are zero after time $t$. Consequently, even if the agreement vector $\{w^{\star}(t)\}_{t=0}^{\infty}$ is adapted to the filtration $\{\mathcal{F}_t\}_{t=0}^{\infty}$, the vector $w^{\star}(t)$ cannot be accessible for a user at time $t$. Nevertheless, the agreement vector $w^{\star}(t)$ can be interpreted as a “probabilistic state” of the whole distributed quantization scheme at time $t$. This explains why the agreement vector is a such convenient tool for the analysis of the DALVQ convergence and will be central in our adaptation of G-lemma, Theorem 10.

Let us remark that Equation (10), writes for all $t \geq 0$,

$$\begin{aligned} w^{\star}(t+1) &= w^{\star}(t) + \sum_{j=1}^{M} \phi^j(t) s^j(t) \\ &= w^{\star}(t) - \sum_{j=1}^{M} \mathbb{1}_{\{t \in T^j\}} \phi^j(t) \varepsilon_{t+1}^j H\left(\mathbf{z}_{t+1}^j, w^j(t)\right). \end{aligned}$$

We recall the reader that the $[0,1]$-valued functions $\phi^j$’s are defined in Lemma 7.

Using the function $h$ defined by identity (2) and the fact that the random variables $w^\star(t)$ and $w^j(t)$ are $\mathcal{F}_t$-measurable then it holds

$$h(w^\star(t)) = \mathbb{E}\{H(\mathbf{z}, w^\star(t)) \mid \mathcal{F}_t\}, \quad t \geq 0.$$

and

$$h(w^j(t)) = \mathbb{E}\left\{H\left(\mathbf{z}, w^j(t)\right) \mid \mathcal{F}_t\right\}, \quad j \in \{1, \ldots, M\} \text{ and } t \geq 0.$$

where $\mathbf{z}$ is a random variable of law $\mu$ independent of $\mathcal{F}_t$.

For all $t \geq 0$, set

$$\varepsilon^\star_{t+1} \triangleq \sum_{j=1}^{M} \mathbb{1}_{\{t \in T^j\}} \phi^j(t) \varepsilon^j_{t+1}. \tag{13}$$

Clearly, the real numbers $\varepsilon^\star_t$ are nonnegative. Their strictly positiveness will be discussed in Proposition 3.

Set

$$\Delta M^{(1)}_t \triangleq \sum_{j=1}^{M} \mathbb{1}_{\{t \in T^j\}} \phi^j(t) \varepsilon^j_{t+1} \left(h(w^\star(t)) - h(w^j(t))\right), \quad t \geq 0, \tag{14}$$

and

$$\Delta M^{(2)}_t \triangleq \sum_{j=1}^{M} \mathbb{1}_{\{t \in T^j\}} \phi^j(t) \varepsilon^j_{t+1} \left(h(w^j(t)) - H\left(\mathbf{z}^j_{t+1}, w^j(t)\right)\right), \quad t \geq 0. \tag{15}$$

Note that $\mathbb{E}\left\{\Delta M^{(2)}_t\right\} = 0$ and, consequently, that the random variables $\Delta M^{(2)}_t$ can be seen as the increments of a martingale with respect to the filtration $\{\mathcal{F}_t\}_{t=0}^{\infty}$.

Finally, with this notation, equation (10) takes the form

$$w^\star(t+1) = w^\star(t) - \varepsilon^\star_{t+1} h(w^\star(t)) + \Delta M^{(1)}_t + \Delta M^{(2)}_t, \quad t \geq 0. \tag{16}$$

We are now in a position to state our most useful tool, which is similar in spirit to the G-lemma, but adapted to the context of distributed asynchronous stochastic gradient descent algorithm.

**Theorem 10 (Asynchronous G-lemma)** *Assume that* $(\mathbf{AsY})_\mathbf{1}$ *or* $(\mathbf{AsY})_\mathbf{2}$ *and Assumption 1 hold and that the following conditions are satisfied:*

1. $\sum_{t=0}^{\infty} \varepsilon^\star_t = \infty$ *and* $\varepsilon^\star_t \xrightarrow[t\to\infty]{} 0$.

2. *The sequences* $\{w^\star(t)\}_{t=0}^{\infty}$ *and* $\{h(w^\star(t))\}_{t=0}^{\infty}$ *are bounded a.s.*

3. *The series* $\sum_{t=0}^{\infty} \Delta M^{(1)}_t$ *and* $\sum_{t=0}^{\infty} \Delta M^{(2)}_t$ *converge a.s. in* $\left(\mathbb{R}^d\right)^\kappa$.

4. *There exists a lower semi-continuous function* $G : \left(\mathbb{R}^d\right)^\kappa \longrightarrow [0, \infty)$ *such that*

$$\sum_{t=0}^{\infty} \varepsilon^\star_{t+1} G(w^\star(t)) < \infty, \quad a.s.$$

*Then, there exists a random connected component* $\Xi$ *of* $\{G = 0\}$ *such that*

$$\operatorname{dist}(w^\star(t), \Xi) \xrightarrow[t\to\infty]{} 0, \quad a.s.$$

### 4.3 Trajectory Analysis

The Pagès's proof in Pagès (1997) on the almost sure convergence of the CLVQ procedure required a careful examination of the trajectories of the process $\{w(t)\}_{t=0}^{\infty}$. Thus, in this subsection we investigate similar properties and introduce the assumptions that will be needed to prove our main convergence result, Theorem 12.

The next Assumption 10 ensures that, for each processor, the quantizers stay in the support of the density.

**Assumption 10** *One has*

$$\mathbb{P}\left\{w^j(t) \in \mathcal{G}^{\kappa}\right\} = 1, \qquad j \in \{1,\dots,M\} \text{ and } t \geq 0.$$

Firstly, let us mention that since the set $\mathcal{G}^{\kappa}$ is convex, if Assumption 10 holds then

$$\mathbb{P}\{w^{\star}(t) \in \mathcal{G}^{\kappa}\} = 1, \quad t \geq 0.$$

Secondly, note that the Assumption 10 is not particularly restrictive. This assumption is satisfied under the condition: for each processor, no descent term is added while a combining computation is performed. This writes

$$a_{i,j}(t) = \delta_{i,j} \text{ and } \tau^{i,i}(t) = t, \qquad (i,j) \in \{1,\dots,M\}^2 \text{ and } t \in T^i.$$

This requirement makes sense for practical implementations.

Recall that if $t \notin T^i$, then $s^i(t) = 0$. Thus, Equation (7) takes the form

$$w^i(t+1) = \begin{cases} w^i(t+1) & = w^i(t) - \varepsilon^i_{t+1}\left(w^i(t) - \mathbf{z}^i_{t+1}\right) \\ & = \left(1-\varepsilon^i_{t+1}\right) w^i(t) + \varepsilon^i_{t+1}\mathbf{z}^i_{t+1} \end{cases} \text{if } t \in T^i; \\ \phantom{w^i(t+1) = } \begin{cases} w^i(t+1) & = \sum_{j=1}^{M} a^{i,j}(t) w^j(\tau^{i,j}(t)) \end{cases} \text{otherwise.}$$

Since $\mathcal{G}^{\kappa}$ is a convex set, it follows easily that if $w^j(0) \in \mathcal{G}^{\kappa}$, then $w^j(t) \in \mathcal{G}^{\kappa}$ for all $j \in \{1,\dots,M\}$ and $t \geq 0$ and, consequently, that Assumption 10 holds.

The next Lemma 11 provides a deterministic upper bound on the differences between the distributed versions $w^i$ and the agreement vector. For any subset $A$ of $\left(\mathbb{R}^d\right)^{\kappa}$, the notation $\mathrm{diam}(A)$ stands for the usual diameter defined by

$$\mathrm{diam}(A) = \sup_{x,y\in A}\left\{\|x-y\|\right\}.$$

**Lemma 11** *Assume* $(\mathbf{AsY})_1$ *or* $(\mathbf{AsY})_2$ *holds and that Assumptions 1, 8 and 10 are satisfied then*

$$\|w^{\star}(t) - w^i(t)\| \leq \sqrt{\kappa} M \,\mathrm{diam}(\mathcal{G}) A K_2 \theta_t, \quad i \in \{1,\dots,M\} \text{ and } t \geq 0, \text{ a.s.,}$$

*where* $\theta_t \triangleq \sum_{\tau=-1}^{t-1} \frac{1}{\tau \vee 1} \rho^{t-\tau}$, *$A$ and $\rho$ are the constants introduced in Lemma 8, $K_2$ is defined in Assumption 8.*

The sequence $\{\theta_t\}_{t=0}^{\infty}$ defined in Lemma 11 satisfies

$$\theta_t \xrightarrow[t\to\infty]{} 0 \text{ and } \sum_{t=0}^{\infty} \frac{\theta_t}{t} < \infty. \tag{17}$$

We give some calculations justifying the statements at the end of the Annex.

Thus, under Assumptions 8 and 10, it follows easily that

$$w^\star(t) - w^i(t) \xrightarrow[t\to\infty]{} 0, \quad i \in \{1,\dots,M\}, \text{ a.s.,}$$

and

$$w^i(t) - w^j(t) \xrightarrow[t\to\infty]{} 0, \quad (i,j) \in \{1,\dots,M\}^2, \text{ a.s.} \tag{18}$$

This shows that the trajectories of the distributed versions of the quantizers reach asymptotically a consensus with probability 1. In other words, if one of the sequences $\left\{w^i(t)\right\}_{t=0}^{\infty}$ converges then they all converge towards the same value. The rest of the paper is devoted to prove that this common value is in fact a zero of $\nabla C$, that is, a critical point.

To prove the result mentioned above, we will need the following assumption, which basically states that the components of $w^\star$ are parted, for every time $t$ but also asymptotically. This assumption is similar in spirit to the main requirement of Theorem 4.

**Assumption 11** *One has*

1. $\mathbb{P}\{w^\star(t) \in \mathcal{D}_*^\kappa\} = 1, \quad t \geq 0.$
2. $\mathbb{P}\left\{\liminf_{t\to\infty} \operatorname{dist}\left(w^\star(t), \complement\mathcal{D}_*^\kappa\right) > 0\right\} = 1, \quad t \geq 0.$

### 4.4 Consistency of the DALVQ

In this subsection we state our main theorem on the consistency of the DALVQ. Its proof is based on the asynchronous G-lemma, Theorem 10. The goal of the next proposition is to ensure that the first assumption of Theorem 10 holds.

**Proposition 3** *Assume* $(\mathbf{AsY})_\mathbf{1}$ *or* $(\mathbf{AsY})_\mathbf{2}$ *holds and that Assumptions 1, 8 and 9 are satisfied then* $\varepsilon_t^\star > 0$, $t \geq 0$, $\varepsilon_t^\star \xrightarrow[t\to\infty]{} 0$ *and* $\sum_{t=0}^{\infty} \varepsilon_t^\star = \infty$.

The second condition required in Theorem 10 deals with the convergence of the two series defined by Equations (14) and (15). The next Proposition 4 provides sufficient condition for the almost sure convergence of these series.

**Proposition 4** *Assume* $(\mathbf{AsY})_\mathbf{1}$ *or* $(\mathbf{AsY})_\mathbf{2}$ *holds and that Assumptions 1, 8, 10 and 11 are satisfied then the series* $\sum_{t=0}^{\infty} \Delta M_t^{(1)}$ *and* $\sum_{t=0}^{\infty} \Delta M_t^{(2)}$ *converge almost surely in* $\left(\mathbb{R}^d\right)^\kappa$.

This next proposition may be considered has the most important step in the proof of the convergence of the DALVQ. It establishes the convergence of a series of the form $\sum_{t=0}^{\infty} \varepsilon_{t+1} \left\|\nabla C\left(w(t)\right)\right\|^2$. The analysis of the convergence of this type of series is standard for the analysis of stochastic gradient method (see for instance Benveniste et al. 1990 and Bottou 1991). In our context, we pursue the fruitful use of the agreement vector sequence, $\{w^\star(t)\}_{t=0}^{\infty}$, and its related “steps”, $\{\varepsilon_t^\star\}_{t=0}^{\infty}$.

Note that under Assumption 11, we have $h\left(w^\star(t)\right) = \nabla C\left(w^\star(t)\right)$ for all $t \geq 0$, almost surely, therefore the sequence $\{\nabla C\left(w^\star(t)\right)\}_{t=0}^{\infty}$ below is well defined.

**Proposition 5** *Assume* $(\mathbf{AsY})_\mathbf{1}$ *or* $(\mathbf{AsY})_\mathbf{2}$ *holds and that Assumptions 1, 8, 10 and 11 are satisfied then*

1. *$C\left(w^{\star}(t)\right) \xrightarrow[t\to\infty]{} C_{\infty}$, a.s.,*
   *where $C_{\infty}$ is a $[0,\infty)$-valued random variable,*

2. $$\sum_{t=0}^{\infty} \varepsilon^{\star}_{t+1} \left\| \nabla C\left(w^{\star}(t)\right)\right\|^2 < \infty, \quad a.s. \tag{19}$$

Remark that from the convergence of the series given by Equation (19) one can only deduce that $\liminf_{t\to\infty} \left\|\nabla C\left(w^{\star}(t)\right)\right\| = 0$.

We are now in a position to state the main theorem of this paper, which expresses the convergence of the distributed version towards some zero of the gradient of the distortion. In addition, the convergence results (18) imply that if a version converges then all the versions converge towards this value.

**Theorem 12 (Asynchronous theorem)** *Assume $(\mathbf{AsY})_{\mathbf{1}}$ or $(\mathbf{AsY})_{\mathbf{2}}$ holds and that Assumptions 1, 8, 9, 10 and 11 are satisfied then*

1. *$w^{*}(t) - w^{i}(t) \xrightarrow[t\to\infty]{} 0, \quad i \in \{1,\ldots,M\}$, a.s.,*

2. *$w^{i}(t) - w^{j}(t) \xrightarrow[t\to\infty]{} 0, \quad (i,j) \in \{1,\ldots,M\}^2$, a.s.,*

3. *$\operatorname{dist}\left(w^{\star}(t), \Xi_{\infty}\right) \xrightarrow[t\to\infty]{} 0$, a.s.,*

4. *$\operatorname{dist}\left(w^{i}, \Xi_{\infty}\right) \xrightarrow[t\to\infty]{} 0, \quad i \in \{1,\ldots,M\}$, a.s.,*

*where $\Xi_{\infty}$ is some random connected component of the set $\{\nabla C = 0\} \cap \mathcal{G}^{\kappa}$.*

### 4.5 Annex

*Sketch of the proof of asynchronous G-lemma 10.* The proof is an adaptation of the one found by Fort and Pagès, Theorem 4 in Fort and Pagès (1996). The recursive equation (16) satisfied by the sequence $\{w^{\star}(t)\}_{t=0}^{\infty}$ is similar to the iterations (2) in Fort and Pagès (1996), with the notation of this paper:

$$X^{t+1} = X^{t} - \varepsilon_{t+1} h\left(X^{t}\right) + \varepsilon_{t+1}\left(\Delta M^{t+1} + \eta^{t+1}\right), \quad t \geq 0.$$

Thus, similarly, we define a family of continuous time stepwise function $\{u \mapsto \check{w}\,(t,u)\}_{t=1}^{\infty}$.

$$\check{w}^{\star}\,(0,u) \triangleq w^{\star}(s), \text{ if } u \in [\varepsilon^{\star}_1 + \ldots + \varepsilon^{\star}_s, \varepsilon^{\star}_1 + \ldots + \varepsilon^{\star}_{s+1}), \quad u \in [0,\infty).$$

and if $u < \varepsilon^{\star}_1$, $\check{w}^{\star}\,(0,u) = w^{\star}(0)$.

$$\check{w}^{\star}\,(t,u) \triangleq \check{w}^{\star}\left(0, \varepsilon^{\star}_1 + \ldots + \varepsilon^{\star}_t + u\right), \quad t \geq 1 \text{ and } u \in [0,\infty).$$

Hence, for every $t \in \mathbb{N}$,

$$\check{w}^{\star}(t,u) = \check{w}^{\star}(0,t) - \int_0^u h\left(\check{w}^{\star}(t,v)\right) dv + R_u(t), \quad u \in [0,\infty),$$

where, for every $t \geq 1$ and $u \in [\varepsilon_1^\star + \ldots + \varepsilon_{t+t'}^\star, \varepsilon_1^\star + \ldots + \varepsilon_{t+t'+1}^\star)$,

$$R_u(t) \triangleq \int_{\varepsilon_t^\star + \ldots + \varepsilon_{t+t'}^\star}^{\varepsilon_1^\star + \ldots + \varepsilon_t^\star + u} \breve{w}^\star(0,v)dv + \sum_{s=t+1}^{t+t'} \left(\Delta M_s^{(1)} + \Delta M_s^{(2)}\right).$$

The only difference between the families of continuous time functions $\{\breve{w}(t,u)\}_{t=1}^{\infty}$ and $\left\{X^{(t)}\right\}_{t=1}^{\infty}$ defined in Fort and Pagès (1996) is the remainder term $R_u(t)$. The convergence

$$\sup_{u \in [0,T]} \|R_u(t)\| \xrightarrow[t \to \infty]{} 0, \quad T > 0.$$

follows easily from the third assumption of Theorem 10. The rest of the proof follows similarly as in Fort and Pagès (1996, Theorem 4).

**Proof of Lemma 11** For all $i \in \{1, \ldots, M\}$, and all $t \geq 0$, and all $1 \leq \ell \leq \kappa$, we may write

$$\begin{aligned}
&\left\|w_\ell^i(t) - w_\ell^\star(t)\right\| \\
&\quad = \left\| \sum_{j=1}^{M} \left( \left(\phi^{i,j}(t,-1) - \phi^j(-1)\right) w_\ell^j(0) + \sum_{\tau=0}^{t-1} \left(\phi^{i,j}(t,\tau) - \phi^j(t)\right) s_\ell^j(\tau) \right) \right\| \\
&\qquad\quad (\text{by Definition 9 and Lemma 7}) \\
&\quad \leq \sum_{j=1}^{M} \left|\phi^{i,j}(t,-1) - \phi^j(-1)\right| \left\|w_\ell^j(0)\right\| + \sum_{\tau=0}^{t-1} \sum_{j=1}^{M} \left|\phi^{i,j}(t,\tau) - \phi^j(t)\right| \left\|s_\ell^j(\tau)\right\| \\
&\quad \leq A\rho^{t+1} \sum_{j=1}^{M} \left\|w_\ell^j(0)\right\| + A \sum_{\tau=0}^{t-1} \sum_{j=1}^{M} \rho^{t-\tau} \left\|s_\ell^j(\tau)\right\| \\
&\qquad\quad (\text{by Lemma 8}).
\end{aligned}$$

Thus,

$$\begin{aligned}
&\left\|w_\ell^i(t) - w_\ell^\star(t)\right\| \\
&\quad \leq A\rho^{t+1} \sum_{j=1}^{M} \left\|w_\ell^j(0)\right\| + A \sum_{\tau=0}^{t-1} \sum_{j=1}^{M} \rho^{t-\tau} \varepsilon_{\tau+1}^j \mathbb{1}_{\{\tau \in T^j\}} \left\|H(\mathbf{z}_{\tau+1}^j, w^j(\tau))_\ell\right\| \\
&\qquad\quad (\text{by Equation (12)}) \\
&\quad \leq A\rho^{t+1} \sum_{j=1}^{M} \left\|w_\ell^j(0)\right\| \\
&\qquad + A \sum_{\tau=0}^{t-1} \sum_{j=1}^{M} \rho^{t-\tau} \varepsilon_{\tau+1}^j \mathbb{1}_{\tau \in T^j} \mathbb{1}_{\left\{\mathbf{z}_{\tau+1}^j \in W_\ell(w^j(\tau))\right\}} \left\|w_\ell^j(\tau) - \mathbf{z}_{\tau+1}^j\right\|.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\left\|w_\ell^i(t)-w_\ell^\star(t)\right\| \\
&\quad\leq AM\operatorname{diam}(\mathcal{G})\rho^{t+1}+A\operatorname{diam}(\mathcal{G})K_2M\sum_{\tau=0}^{t-1}\frac{1}{\tau\vee 1}\rho^{t-\tau} \\
&\qquad\qquad\text{(because } 0\in\mathcal{G} \text{ and by Assumptions 8 and 10)} \\
&\quad\leq A\operatorname{diam}(\mathcal{G})K_2M\sum_{\tau=-1}^{t-1}\frac{1}{\tau\vee 1}\rho^{t-\tau}.
\end{aligned}$$

Consequently,

$$\begin{aligned}
&\left\|w^\star(t)-w^i(t)\right\| \\
&\quad=\sqrt{\sum_{\ell=1}^{\kappa}\left\|w_\ell^i(t)-w_\ell^\star(t)\right\|^2} \\
&\quad\leq\sqrt{\kappa}M\operatorname{diam}(\mathcal{G})AK_2\sum_{\tau=-1}^{t-1}\frac{1}{\tau\vee 1}\rho^{t-\tau}.
\end{aligned}$$

This proves the desired result. ■

Let us now introduce the following events: for any $\delta>0$ and $t\geq 0$,

$$A_\delta^t\triangleq\left\{w^\star(\tau)\in\mathcal{G}_\delta^\kappa\,,\ t\geq\tau\geq 0\right\}.$$

Recall that the $\mathcal{G}_\delta^\kappa$ is a compact subset of $\mathcal{G}^\kappa$ defined by Equality (4). The next lemma establishes a detailed analysis of security regions for the parted components of the sequences $\{w^\star(t)\}_{t=0}^\infty$ and $\left\{w^j(t)\right\}_{t=0}^\infty$.

**Lemma 13** *Let Assumptions 8 and 10 hold. Then,*

1. *there exists an integer* $t_\delta^1\geq 1$ *such that*

$$A_\delta^t\subset A_{\delta/2}^{t+1},\quad t\geq t_\delta^1.$$

*Moreover,*

$$w^\star(t)\in\mathcal{G}_\delta^\kappa\Rightarrow[w^\star(t),w^\star(t+1)]\subset\mathcal{G}_{\delta/2}^\kappa,\quad t\geq t_\delta^1.$$

2. *There exists an integer* $t_\delta^2\geq 1$ *such that*

$$w^\star(t)\in\mathcal{G}_\delta^\kappa\Rightarrow[w^\star(t),w^i(t)]\subset\mathcal{G}_{\delta/2}^\kappa,\quad i\in\{1,\ldots,M\}\ \textit{and}\ t\geq t_\delta^2.$$

**Proof of Lemma 13** *Proof of statement 1.* The proof starts with the observation that under Assumption 10 we have $w^j(t)\in\mathcal{G}^\kappa$, for all $i\in\{1,\ldots,M\}$ and $t\geq 0$. It follows that, for any $1\leq\ell\leq\kappa$,

$$\begin{aligned}
\left\|H\left(\mathbf{z}_{t+1}^j,w^j(t)\right)_\ell\right\| &\leq\left\|\mathbf{z}_{t+1}^j-w_\ell^j(t)\right\| \\
&\leq\operatorname{diam}(\mathcal{G}).
\end{aligned}$$

Let us now provide an upper bound on the norm of the differences between two consecutive values of the agreement vector sequence. We may write, for all $t \geq 0$ and all $1 \leq \ell \leq M$,

$$\begin{aligned}
&\|w_\ell^\star(t+1) - w_\ell^\star(t)\| \\
&\quad = \left\| \sum_{j=1}^{M} \phi^j(t) s_\ell^j(t) \right\| \\
&\quad \leq \sum_{j=1}^{M} \phi^j(t) \left\| s_\ell^j(t) \right\| \\
&\quad \leq \sum_{j=1}^{M} \varepsilon_{t+1}^j \mathbb{1}_{\{t \in T^j\}} \left\| H\left(\mathbf{z}_{t+1}^j, w^j(t)\right)_\ell \right\| \\
&\qquad\quad \big(\text{by Equation (12) and statement 1. of Lemma 8}\big) \\
&\quad \leq \frac{M \operatorname{diam}(\mathcal{G}) K_2}{t \vee 1} \qquad\qquad\qquad\qquad (20) \\
&\qquad\quad \big(\text{by Assumption 8}\big).
\end{aligned}$$

Take $t \geq \frac{4}{\delta} M \operatorname{diam}(\mathcal{G}) K_2$ and $1 \leq k \neq \ell \leq M$. Let $\alpha$ be a real number in the interval $[0,1]$. If $w^\star(t) \in \mathcal{G}_\delta^\kappa$ then

$$\begin{aligned}
&\|(1-\alpha)w_\ell^\star(t) + \alpha w_\ell^\star(t+1) - (1-\alpha)w_k^\star(t) - \alpha w_k^\star(t+1)\| \\
&\quad = \|w_\ell^\star(t) - w_k^\star(t) + \alpha\left(w_\ell^\star(t+1) - w_\ell^\star(t)\right) + \alpha\left(w_k^\star(t) - w_k^\star(t+1)\right)\| \\
&\quad \geq \|w_\ell^\star(t) - w_k^\star(t)\| - \|\alpha\left(w_\ell^\star(t+1) - w_\ell^\star(t)\right) + \alpha\left(w_k^\star(t) - w_k^\star(t+1)\right)\| \\
&\quad \geq \|w_\ell^\star(t) - w_k^\star(t)\| - \alpha\|w_\ell^\star(t+1) - w_\ell^\star(t)\| - \alpha\|w_k^\star(t) - w_k^\star(t+1)\| \\
&\quad \geq \delta - 2\alpha\frac{\delta}{4} \\
&\quad \geq \delta/2.
\end{aligned}$$

This proves that the whole segment $[w^\star(t), w^\star(t+1)]$ is contained in $\mathcal{G}_{\delta/2}^\kappa$.

*Proof of statement 2.* Take $t \geq 1$ and $1 \leq \ell \leq M$. If $w^\star(t) \in \mathcal{G}_\delta^\kappa$ then by Lemma 11, there exists $t_\delta^2$ such that

$$\left\|w_\ell^\star(t) - w_\ell^i(t)\right\| \leq \frac{\delta}{4}, \quad i \in \{1,\ldots,M\} \text{ and } t \geq t_\delta^2.$$

Let $k$ and $\ell$ two distinct integers between 1 and $M$. For any $t \geq t_\delta^2$,

$$\begin{aligned}
&\left\|\alpha w_k^i(t) + (1-\alpha)w_k^\star(t) - \alpha w_\ell^i(t) - (1-\alpha)w_\ell^\star(t)\right\| \\
&\quad = \left\|w_k^\star(t) - w_\ell^\star(t) + \alpha(w_k^i(t) - w_k^\star(t)) + \alpha(w_\ell^\star(t) - w_\ell^i(t))\right\| \\
&\quad \geq \|w_k^\star(t) - w_\ell^\star(t)\| - \alpha\left\|w_k^i(t) - w_k^\star(t)\right\| - \alpha\left\|w_\ell^\star(t) - w_\ell^i(t)\right\| \\
&\quad \geq \delta - 2\alpha\frac{\delta}{4} \\
&\quad \geq \delta/2.
\end{aligned}$$

This implies $[w^\star(t), w^i(t)] \subset \mathcal{G}_{\delta/2}^\kappa$, as desired. ■

**Proof of Proposition 3** By definition $\varepsilon^{\star}_{t+1}$ equals $\sum_{j=1}^{M}\mathbb{1}_{\{t\in T^j\}}\phi^j(t)\varepsilon^j_{t+1}$, for all $t\geq 0$ .

On the one hand, since the real number $\phi^j(t)$ belongs to the interval $[\eta,1]$ (by Lemma 8) $\varepsilon^{\star}_{t+1}$ is bounded from above by $\frac{MK_2}{t\vee 1}$ using the right-hand side inequality of Assumption 8.

On the other hand, $\varepsilon^{\star}_{t+1}$ is bounded from below by the nonnegative real number $\eta\frac{K_1}{t\vee 1}$ using the left-hand side inequality of Assumption 8. Note also that as Assumption 9 holds, this real number is a positive one. Therefore, it follows that

$$\varepsilon^{\star}_t \xrightarrow[t\to\infty]{} 0$$

and

$$\sum_{t=0}^{\infty}\varepsilon^{\star}_t=\infty.$$

■

**Proof of Proposition 4** *Consistency of* $\sum_{t=0}^{\infty}\Delta M_t^{(1)}$. Let $\delta$ be a positive real number and let $t\geq t^2_\delta$, where $t^2_\delta$ is given by Lemma 19. We may write

$$\begin{aligned}
&\mathbb{1}_{A^t_\delta}\sum_{j=1}^{M}\mathbb{1}_{\{t\in T^j\}}\phi^j(t)\varepsilon^j_{t+1}\left\|h\left(w^{\star}(t)\right)-h\left(w^j(t)\right)\right\|\\
&\leq \mathbb{1}_{\left\{[w^{\star}(t),w^j(t)]\subset\mathcal{G}^{\kappa}_{\delta/2}\right\}}\sum_{j=1}^{M}\phi^j(t)\varepsilon^j_{t+1}\left\|\nabla C\left(w^{\star}(t)\right)-\nabla C\left(w^j(t)\right)\right\|\\
&\qquad\text{(using statement 2. of Lemma 13 and the fact that } \nabla C=h \text{ on } \mathcal{D}^{\kappa}_{*})\\
&\leq \mathbb{1}_{\left\{[w^{\star}(t),w^j(t)]\subset\mathcal{G}^{\kappa}_{\delta/2}\right\}}P_{\delta/2}\sum_{j=1}^{M}\varepsilon^j_{t+1}\left\|w^{\star}(t)-w^j(t)\right\|\\
&\qquad\text{(by Lemma 2)}\\
&\leq\sqrt{\kappa}\,\mathrm{diam}(\mathcal{G})AK_2^2P_{\delta/2}M^2\frac{\theta_t}{t}\\
&\qquad\text{(by Lemma 11)}.
\end{aligned}$$

Thus, since $\sum_{t=0}^{\infty}\frac{\theta_t}{t}<\infty$, the series

$$\sum_{t=0}^{\infty}\mathbb{1}_{A^t_\delta}\sum_{j=1}^{M}\mathbb{1}_{\{t\in T^j\}}\phi^j(t)\varepsilon^j_{t+1}\left\|h\left(w^{\star}(t)\right)-h\left(w^j(t)\right)\right\|$$

is almost surely convergent. Under Assumption 11, we have

$$\mathbb{P}\left\{\bigcup_{\delta>0}\bigcap_{t\geq 0}A^t_\delta\right\}=1.$$

It follows that the series $\sum_{t=0}^{\infty}\Delta M_t^{(1)}$ converges almost surely in $\left(\mathbb{R}^d\right)^{\kappa}$.

*Consistency of* $\sum_{t=0}^{\infty}\Delta M_t^{(2)}$. The sequence of random variables $M_t^{(2)}$ defined, for all $t \geq 0$, by

$$\begin{aligned} M_t^{(2)} &\triangleq \sum_{\tau=0}^{t} \Delta M_\tau^{(2)} \\ &= \sum_{\tau=0}^{t}\sum_{j=1}^{M} \mathbb{1}_{\{\tau\in T^j\}}\varepsilon_{\tau+1}^{j}\phi^j(\tau)\left(h\left(w^j(\tau)\right)-H\left(\mathbf{z}_{\tau+1}^{j},w^j(\tau)\right)\right). \end{aligned}$$

is a vector valued martingale with respect to the filtration $\{\mathcal{F}_t\}_{t=0}^{\infty}$. It turns out that this martingale has square integrable increments. Precisely,

$$\sum_{t=0}^{\infty}\mathbb{E}\left\{\left\|M_{t+1}^{(2)}-M_t^{(2)}\right\|^2 \mid \mathcal{F}_t\right\}=\sum_{t=1}^{\infty}\mathbb{E}\left\{\left\|\Delta M_t^{(2)}\right\|^2 \mid \mathcal{F}_t\right\}<\infty.$$

Indeed, for all $j \in \{1,\ldots,M\}$ and $t \geq 1$,

$$\begin{aligned} &\sum_{\tau=1}^{t}\mathbb{E}\left\{\left\|\mathbb{1}_{\{\tau\in T^j\}}\varepsilon_{\tau+1}^{j}\left(h\left(w^j(\tau)\right)-H\left(\mathbf{z}_{\tau+1}^{j}(\tau),w^j(\tau)\right)\right)\right\|^2 \mid \mathcal{F}_\tau\right\} \\ &\quad\leq \sum_{\tau=1}^{t}\left(\varepsilon_{\tau+1}^{j}\right)^2\mathbb{E}\left\{\left\|h\left(w^j(\tau)\right)-H\left(\mathbf{z}_{\tau+1}^{j}(\tau),w^j(\tau)\right)\right\|^2 \mid \mathcal{F}_\tau\right\} \\ &\quad\leq 2\sum_{\tau=1}^{t}\left(\varepsilon_{\tau+1}^{j}\right)^2\mathbb{E}\left\{\left\|h\left(w^j(\tau)\right)\right\|^2+\left\|H\left(\mathbf{z}_{\tau+1}^{j}(\tau),w^j(\tau)\right)\right\|^2 \mid \mathcal{F}_\tau\right\} \\ &\quad\leq 4\kappa\,\mathrm{diam}(\mathcal{G})^2\sum_{\tau=1}^{t}\left(\varepsilon_{\tau+1}^{j}\right)^2 \\ &\qquad\qquad\text{(using Assumption 10)} \\ &\quad\leq 4\kappa\,\mathrm{diam}(\mathcal{G})^2K_2^2\sum_{\tau=1}^{t}\frac{1}{\tau^2}. \end{aligned}$$

We conclude that the series $\sum_{t\geq 1}\Delta M_t^{(2)}$ is almost surely convergent. ■

**Proof of Proposition 5** Denote by $\langle x,y\rangle$ the canonical inner product of two vectors $x,y\in\mathbb{R}^d$ and also, with a slight abuse of notation, the canonical inner product of two vectors $x,y\in\left(\mathbb{R}^d\right)^{\kappa}$. Let $\delta$ be a positive real number. Take any $t \geq \max\left\{t_\delta^1,t_\delta^2\right\}$, where $t_\delta^1$ and $t_\delta^2$ are defined as in Lemma 13. One has,

$$\begin{gathered} \mathbb{1}_{A_\delta^{t+1}}C\left(w^\star(t+1)\right)\leq \mathbb{1}_{A_\delta^{t}}C\left(w^\star(t+1)\right). \\ \text{(by definition } A_\delta^{t+1}\subset A_\delta^{t}\text{)} \end{gathered}$$

Consequently,

$$\begin{aligned}
&\mathbb{1}_{A_\delta^{t+1}} C\left(w^\star(t+1)\right) \\
&\quad \leq \mathbb{1}_{A_\delta^t} C\left(w^\star(t)\right) + \mathbb{1}_{A_\delta^t} \langle \nabla C(w^\star(t)), w^\star(t+1) - w^\star(t) \rangle \\
&\qquad + \mathbb{1}_{\left\{[w^\star(t), w^\star(t+1)] \subset \mathcal{G}^\kappa_{\delta/2}\right\}} \\
&\qquad\quad \times \left[ \sup_{z \in [w^\star(t), w^\star(t+1)]} \left\{ \left\| \nabla C(z) - \nabla C(w^\star(t)) \right\| \right\} \left\| w^\star(t+1) - w^\star(t) \right\| \right] \\
&\quad \leq \mathbb{1}_{A_\delta^t} C\left(w^\star(t)\right) + \mathbb{1}_{A_\delta^t} \langle \nabla C(w^\star(t)), w^\star(t+1) - w^\star(t) \rangle \\
&\qquad + P_{\delta/2} \left\| w^\star(t+1) - w^\star(t) \right\|^2 \\
&\qquad\quad \text{(using Lemma 2.)}
\end{aligned}$$

The first inequality above holds since the bounded increment formula above is valid by statement 1 of Lemma 13. Let us now bound separately the right hand side members of the second inequality.

Firstly, the next inequality holds by Inequality (20) provided in the proof of Lemma 13,

$$P_{\delta/2} \left\| w^\star(t+1) - w^\star(t) \right\|^2 \leq \kappa P_{\delta/2} \left( \frac{K_2 M \operatorname{diam}(\mathcal{G})}{t} \right)^2 .$$

Secondly,

$$\begin{aligned}
&\mathbb{1}_{A_\delta^t} \langle \nabla C(w^\star(t)), w^\star(t+1) - w^\star(t) \rangle \\
&\quad = \mathbb{1}_{A_\delta^t} \langle \nabla C(w^\star(t)), \sum_{j=1}^{M} \phi^j(t) s^j(t) \rangle \\
&\qquad\quad \text{(by Equation (10))} \\
&\quad = \mathbb{1}_{A_\delta^t} \sum_{j=1}^{M} \langle \nabla C(w^j(t)), \phi^j(t) s^j(t) \rangle \\
&\qquad + \mathbb{1}_{A_\delta^t} \sum_{j=1}^{M} \langle \nabla C(w^\star(t)) - \nabla C(w^j(t)), \phi^j(t) s^j(t) \rangle .
\end{aligned}$$

Thus,

$$\begin{aligned}
&\mathbb{1}_{A_\delta^t}\langle \nabla C(w^\star(t)), w^\star(t+1)-w^\star(t)\rangle \\
&\quad\le \mathbb{1}_{A_\delta^t}\sum_{j=1}^{M}\langle \nabla C(w^j(t)), \phi^j(t)s^j(t)\rangle \\
&\qquad + \mathbb{1}_{A_\delta^t}\sum_{j=1}^{M}\left|\langle \nabla C(w^\star(t))-\nabla C(w^j(t)), \phi^j(t)s^j(t)\rangle\right| \\
&\quad\le \mathbb{1}_{A_\delta^t}\sum_{j=1}^{M}\langle \nabla C(w^j(t)), \phi^j(t)s^j(t)\rangle \\
&\qquad + \sum_{j=1}^{M}\mathbb{1}_{A_\delta^t}\left\|\nabla C(w^\star(t))-\nabla C(w^j(t))\right\|\left\|\phi^j(t)s^j(t)\right\| \\
&\qquad\quad (\text{using Cauchy-Schwarz inequality}).
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\mathbb{1}_{A_\delta^t}\langle \nabla C(w^\star(t)), w^\star(t+1)-w^\star(t)\rangle \\
&\quad\le \mathbb{1}_{A_\delta^t}\sum_{j=1}^{M}\langle \nabla C(w^j(t)), \phi^j(t)s^j(t)\rangle \\
&\qquad + \sum_{j=1}^{M}\mathbb{1}_{\left\{[w^\star(t),w^j(t)]\subset \mathcal{G}_{\delta/2}^{\kappa}\right\}}\left\|\nabla C(w^\star(t))-\nabla C(w^j(t))\right\|\left\|\phi^j(t)s^j(t)\right\| \\
&\qquad (\text{by statement 2 of Lemma 13}) \\
&\quad\le \mathbb{1}_{A_\delta^t}\sum_{j=1}^{M}\langle \nabla C(w^j(t)), \phi^j(t)s^j(t)\rangle \\
&\qquad + P_{\delta/2}\sum_{j=1}^{M}\left\|w^\star(t)-w^j(t)\right\|\left\|\phi^j(t)s^j(t)\right\| \\
&\qquad (\text{using Lemma 2})
\end{aligned}$$

$$\begin{aligned}
&\mathbb{1}_{A_\delta^t}\langle \nabla C(w^\star(t)), w^\star(t+1)-w^\star(t)\rangle \\
&\quad\le \mathbb{1}_{A_\delta^t}\sum_{j=1}^{M}\langle \nabla C(w^j(t)), \phi^j(t)s^j(t)\rangle \\
&\qquad + P_{\delta/2}AK_2^2\kappa M^2\,\mathrm{diam}(\mathcal{G})^2\frac{\theta_t}{t} \\
&\qquad\quad (\text{using Lemma 11 and the upper bound (20)}).
\end{aligned}$$

Finally,

$$\begin{aligned}
&\mathbb{1}_{A_\delta^{t+1}} C\left(w^\star(t+1)\right) \\
&\quad \le \mathbb{1}_{A_\delta^t} C\left(w^\star(t)\right) + \mathbb{1}_{A_\delta^t} \sum_{j=1}^{M} \langle \nabla C(w^j(t)), \phi^j(t) s^j(t) \rangle \\
&\qquad + P_{\delta/2} A K_2^2 \kappa M^2 \operatorname{diam}(\mathcal{G})^2 \frac{\theta_t}{t} \\
&\qquad + \kappa P_{\delta/2} \left( \frac{K_2 M \operatorname{diam}(\mathcal{G})}{t} \right)^2 .
\end{aligned} \tag{21}$$

Set

$$\Omega_\delta^1 \triangleq P_{\delta/2} A K_2^2 \kappa M^2 \operatorname{diam}(\mathcal{G})^2$$

and

$$\Omega_\delta^2 \triangleq \kappa P_{\delta/2} \left( K_2 M \operatorname{diam}(\mathcal{G}) \right)^2 .$$

■

In the sequel, we shall need the following lemma.

**Lemma 14** *For all $t \ge \max\left\{t_\delta^1, t_\delta^2\right\}$, the quantity $W_t$ below is a nonnegative supermartingale with respect to the filtration $\{\mathcal{F}_t\}_{t=0}^{\infty}$:*

$$\begin{aligned}
W_t \triangleq \mathbb{1}_{A_\delta^t} C\left(w^\star(t)\right) + \eta K_1 \sum_{\tau=0}^{t-1} \mathbb{1}_{A_\delta^\tau} \frac{1}{\tau} \sum_{j=1}^{M} \mathbb{1}_{\{\tau \in T^j\}} \left\| \nabla C\left(w^j(\tau)\right) \right\|^2 \\
+ \Omega_\delta^1 \sum_{\tau=t}^{\infty} \frac{\theta(\tau)}{\tau} + \Omega_\delta^2 \sum_{\tau=t}^{\infty} \frac{1}{\tau^2}, \quad t \ge 1.
\end{aligned}$$

**Proof of Lemma 14** Indeed, using the upper bound provided by Equation (21),

$$\begin{aligned}
&\mathbb{E}\left\{\mathbb{1}_{A_\delta^{t+1}} C\left(w^\star(t+1)\right) \;\middle|\; \mathcal{F}_t\right\} \\
&\quad \leq \mathbb{1}_{A_\delta^t} C\left(w^\star(t)\right) + \mathbb{1}_{A_\delta^t} \sum_{j=1}^{M} \mathbb{E}\left\{\langle \nabla C(w^j(t)), \phi^j(t) s^j(t)\rangle \;\middle|\; \mathcal{F}_t\right\} \\
&\qquad + \Omega_\delta^1 \frac{1}{t}\theta_t + \Omega_\delta^2 \frac{1}{t^2} \\
&\quad = \mathbb{1}_{A_\delta^t} C\left(w^\star(t)\right) \\
&\qquad + \mathbb{1}_{A_\delta^t} \sum_{j=1}^{M} \left\langle \nabla C(w^j(t)), \mathbb{E}\left\{-\mathbb{1}_{\{t\in T^j\}} \phi^j(t) \varepsilon_{t+1}^j H(\mathbf{z}_{t+1}^j, w^j(t)))\rangle \;\middle|\; \mathcal{F}_t\right\}\right\rangle \\
&\qquad + \Omega_\delta^1 \frac{\theta_t}{t} + \Omega_\delta^2 \frac{1}{t^2} \\
&\quad = \mathbb{1}_{A_\delta^t} C\left(w^\star(t)\right) \\
&\qquad - \mathbb{1}_{A_\delta^t} \sum_{j=1}^{M} \mathbb{1}_{\{t\in T^j\}} \phi^j(t) \varepsilon_{t+1}^j \left\|\nabla C(w^j(t))\right\|^2 + \Omega_\delta^1 \frac{\theta_t}{t} + \Omega_\delta^2 \frac{1}{t^2} \\
&\quad \leq \mathbb{1}_{A_\delta^t} C\left(w^\star(t)\right) \\
&\qquad - \frac{\eta K_1}{t} \mathbb{1}_{A_\delta^t} \sum_{j=1}^{M} \mathbb{1}_{\{t\in T^j\}} \left\|\nabla C(w^j(t))\right\|^2 + \Omega_\delta^1 \frac{\theta_t}{t} + \Omega_\delta^2 \frac{1}{t^2}.
\end{aligned}$$

In the last inequality we used the fact that $\phi^j(t) \geq \eta$ (Lemma 8) and $\varepsilon_{t+1}^j \geq \frac{K_1}{t}$ (Assumption 8).

It is straightforward to verify that, we have $W_t - \mathbb{E}\{W_{t+1}|\mathcal{F}_t\} \geq 0$ which prove the desired result. ■

**Proof of Proposition 5 (continued)** Since $\{W_t\}_{t=1}^\infty$ is a nonnegative supermartingale (by Lemma 14), $W_t$ converges almost surely as $t \to \infty$ (see for instance Durrett 1990). Then, as $\sum_{\tau=t}^\infty \frac{\theta(\tau)}{\tau} \xrightarrow[t\to\infty]{} 0$ and $\sum_{\tau=t}^\infty \frac{1}{\tau^2} \xrightarrow[t\to\infty]{} 0$, we have

$$\mathbb{1}_{A_\delta^t} C(w^\star(t)) \xrightarrow[t\to\infty]{} C_\infty, \quad \text{a.s.,} \tag{22}$$

where $C_\infty \in [0, \infty)$ and, because the origin of the expression is increasing in $t$, the following series converges

$$\sum_{\tau=0}^\infty \mathbb{1}_{A_\delta^\tau} \frac{1}{\tau \vee 1} \sum_{j=1}^M \mathbb{1}_{\{\tau\in T^j\}} \left\|\nabla C\left(w^j(\tau)\right)\right\|^2 < \infty, \quad \text{a.s.} \tag{23}$$

*Proof of statement 1.* Assumption 11 means that

$$\mathbb{P}\left\{\bigcup_{\delta>0} \bigcap_{t\geq 0} A_\delta^t\right\} = 1.$$

Statement 1 follows easily from the convergence (22).

*Proof of statement 2.* The required convergence (19) is proven as follows. We have

$$\begin{aligned}
&\sum_{\tau=0}^{t} \varepsilon^{\star}_{\tau+1} \mathbb{1}_{A^{\tau}_{\delta}} \left\|\nabla C\left(w^{\star}(\tau)\right)\right\|^2 \\
&\quad \leq \sum_{\tau=0}^{t}\sum_{j=1}^{M} \phi^j(\tau) \mathbb{1}_{\{\tau\in T^j\}} \mathbb{1}_{A^{\tau}_{\delta}} \varepsilon^{j}_{\tau+1} \left\|\nabla C\left(w^{\star}(\tau)\right)\right\|^2 \\
&\qquad\qquad \text{(using Equality (13))} \\
&\quad \leq 2K_2 \sum_{\tau=0}^{t} \mathbb{1}_{A^{\tau}_{\delta}} \frac{1}{\tau\vee 1} \sum_{j=1}^{M} \mathbb{1}_{\{\tau\in T^j\}} \left\|\nabla C\left(w^{j}(\tau)\right)\right\|^2 \\
&\qquad\qquad \text{(using Assumption 9)} \\
&\qquad + 2K_2 \sum_{\tau=0}^{t} \mathbb{1}_{\left\{[w^{\star}(\tau),w^{j}(\tau)]\subset \mathcal{G}^{\kappa}_{\delta/2}\right\}} \frac{1}{\tau\vee 1} \sum_{j=1}^{M} \left\|\nabla C\left(w^{j}(\tau)\right) - \nabla C\left(w^{\star}(\tau)\right)\right\|^2 \\
&\qquad\qquad \text{(using Assumption 9 and statement 2 of Lemma 13.)}
\end{aligned}$$

Thus,

$$\begin{aligned}
&\sum_{\tau=0}^{t} \varepsilon^{\star}_{\tau+1} \mathbb{1}_{A^{\tau}_{\delta}} \left\|\nabla C\left(w^{\star}(\tau)\right)\right\|^2 \\
&\quad \leq 2K_2 \sum_{\tau=0}^{t} \mathbb{1}_{A^{\tau}_{\delta}} \frac{1}{\tau\vee 1} \sum_{j=1}^{M} \mathbb{1}_{\{\tau\in T^j\}} \left\|\nabla C\left(w^{j}(\tau)\right)\right\|^2 \\
&\qquad + 2K_2 P^2_{\delta/2} \sum_{\tau=0}^{t} \mathbb{1}_{\left\{[w^{\star}(\tau),w^{j}(\tau)]\subset \mathcal{G}^{\kappa}_{\delta/2}\right\}} \frac{1}{\tau\vee 1} \sum_{j=1}^{M} \left\|w^{j}(\tau) - w^{\star}(\tau)\right\|^2 \\
&\qquad\qquad \text{(by Lemma 2)}.
\end{aligned}$$

Thus,

$$\begin{aligned}
&\sum_{\tau=0}^{t} \varepsilon^{\star}_{\tau+1} \mathbb{1}_{A^{\tau}_{\delta}} \left\|\nabla C\left(w^{\star}(\tau)\right)\right\|^2 \\
&\quad \leq 2K_2 \sum_{\tau=0}^{t} \mathbb{1}_{A^{\tau}_{\delta}} \frac{1}{\tau\vee 1} \sum_{j=1}^{M} \mathbb{1}_{\{\tau\in T^j\}} \left\|\nabla C\left(w^{j}(\tau)\right)\right\|^2 \\
&\qquad + 2P^2_{\delta/2} K_2^3 \kappa M^3 A^2 \operatorname{diam}(\mathcal{G})^2 \sum_{\tau=1}^{t} \frac{1}{\tau\vee 1} \theta^2_{\tau} \\
&\qquad\qquad \text{(by Lemma 11)}.
\end{aligned}$$

Finally, using the convergence (23), one has

$$\sum_{\tau=0}^{\infty} \varepsilon^{\star}_{\tau+1} \mathbb{1}_{A^{\tau}_{\delta}} \left\|\nabla C\left(w^{\star}(\tau)\right)\right\|^2 < \infty, \quad \text{a.s.},$$

and the conclusion follows from the fact that Assumption 11 implies

$$\mathbb{P}\left\{\bigcup_{\delta>0}\bigcap_{t\geq 0} A^{t}_{\delta}\right\} = 1.$$

■

**Proof of Theorem 12** The proof consists in verifying the assumptions of Theorem 10 with the function $\widehat{G}$ defined by Equation (5).

It has been outlined that Assumption 10 implies that $w^\star(t)$ lie in the compact set $\mathcal{G}^\kappa$, almost surely, for all $t \geq 0$. Consequently, in the definition of $\widehat{G}(w^\star)$ the lim inf symbol can be omitted. For all $\mathbf{z} \in \mathcal{G}$ and all $t \geq 0$, we have $\|H(\mathbf{z}, w^\star(t))\| \leq \sqrt{\kappa}\,\mathrm{diam}\,(\mathcal{G})$, almost surely, whereas $\{h(w^\star(t))\}_{t=0}^{\infty}$ satisfies

$$h(w^\star(t)) = \mathbb{E}\left\{H\left(\mathbf{z}, w^\star(t)\right) \mid \mathcal{F}_t\right\}, \quad t \geq 0, \text{ a.s.}$$

Thus, the sequences $\{w^\star(t)\}_{t=0}^{\infty}$ and $\{h(w^\star(t))\}_{t=0}^{\infty}$ are bounded almost surely.

Proposition 3, respectively Proposition 4, respectively Proposition 5 show that the first assumption, respectively the third assumption, respectively the fourth assumption of Theorem 10 hold. This concludes the proof of the theorem. ■

*Justification of the statements (17).* Recall that the definition of $\theta$ is provided in Lemma 11. Let us remark that it is sufficient to analyse the behavior in $t$ of the quantity $\sum_{\tau=1}^{t-1} \rho^{t-\tau}/\tau$. Let $\varepsilon > 0$ then for all $t \geq \lfloor 1/\varepsilon \rfloor + 1$, we have

$$\begin{aligned}
&\sum_{\tau=1}^{t-1} \frac{\rho^{t-\tau}}{\tau} \\
&\quad = \sum_{\tau=1}^{\lfloor 1/\varepsilon \rfloor} \frac{\rho^{t-\tau}}{\tau} + \sum_{\tau=\lfloor 1/\varepsilon \rfloor+1}^{t-1} \frac{\rho^{t-\tau}}{\tau} \\
&\quad \leq \sum_{\tau=1}^{\lfloor 1/\varepsilon \rfloor} \rho^{t-\tau} + \varepsilon \sum_{\tau=\lfloor 1/\varepsilon \rfloor+1}^{t-1} \rho^{t-\tau} \\
&\quad \leq \frac{\rho^{t-\lfloor 1/\varepsilon \rfloor}}{1-\rho} + \frac{\varepsilon}{1-\rho} \\
&\qquad\qquad \text{(using the fact that } \rho \in (0,1)\text{)}.
\end{aligned}$$

Consequently, for $t$ sufficiently large we have

$$\sum_{\tau=1}^{t-1} \frac{\rho^{t-\tau}}{\tau} \leq \frac{2\varepsilon}{1-\rho}$$

which proves the first claim.
The second claim follows the same technique by letting "$\varepsilon = 1/\sqrt{t}$".
Thus, for $t \geq 1$ we have

$$\theta_t \leq \frac{\rho^{t-\lfloor \sqrt{t} \rfloor - 1}}{1-\rho} + \frac{1/\sqrt{t}}{1-\rho}.$$

Finally, for $T \geq 1$, it holds

$$\sum_{t=1}^{T} \sum_{\tau=1}^{t-1} \frac{\rho^{t-\tau}}{\tau} \leq \frac{1}{1-\rho} \left( \sum_{t=1}^{T} \rho^{n-\lfloor \sqrt{n} \rfloor - 1} + \sum_{t=1}^{T} \frac{1}{n^{3/2}} \right).$$

The two partial sums in the above parenthesis have finite limits which prove the second statement.

## References

E. A. Abaya and G. L. Wise. Convergence of vector quantizers with applications to optimal quantization. *SIAM Journal on Applied Mathematics*, 44(1):183–189, 1984.

A. Antos. Improved minimax bounds on the test and training distortion of empirically designed vector quantizers. *IEEE Transactions on Information Theory*, 51(11):4022–4032, 2005.

A. Antos, L. Györfi, and A. György. Individual convergence rates in empirical vector quantizer design. *IEEE Transactions on Information Theory*, 51(11):4013–4022, 2005.

P. L. Bartlett, T. Linder, and G. Lugosi. The minimax distortion redundancy in empirical quantizer design. *IEEE Transactions on Information Theory*, 44(5):1802–1813, 1998.

A. Benveniste, M. Métivier, and P. Priouret. *Adaptive Algorithms and Stochastic Approximations*. Springer-Verlag, 1990.

D. P. Bertsekas and J. N. Tsitsiklis. *Parallel and Distributed Computation: Numerical Methods*. Prentice-Hall, Inc., 1989.

G. Biau, L. Devroye, and G. Lugosi. On the performance of clustering in hilbert spaces. *IEEE Transactions on Information Theory*, 54(2):781–790, 2008.

V. D. Blondel, J. M. Hendrickx, A. Olshevsky, and J. N. Tsitsiklis. Convergence in multiagent coordination, consensus, and flocking. In *Decision and Control, 2005 and 2005 European Control Conference.*, 2005.

L. Bottou. Stochastic gradient learning in neural networks. In *Proceedings of Neuro-Nîmes 91*, 1991.

L. Bottou. Online algorithms and stochastic approximations. In *Online Learning and Neural Networks*. Cambridge University Press, 1998.

L. Bottou and Y. Bengio. Convergence properties of the $k$-means algorithm. In *Advances in Neural Information Processing Systems*. MIT Press, 1995.

L. Bottou and Y. LeCun. Large scale online learning. In *Advances in Neural Information Processing Systems 16*. MIT Press, 2004.

F. Bullo, J. Cortés, and S. Martínez. *Distributed Control of Robotic Networks*. Princeton University Press, 2009.

G. Choquet. *Topology*. Academic Press, 1966.

P. A. Chou. The distortion of vector quantizers trained on $n$ vectors decreases to the optimum at $o_p(1/n)$. In *In Proceedings of IEEE International Symposium on Information Theory*, 1994.

J. W. Durham, R. Carli, P. Frasca, and F. Bullo. Discrete partitioning and coverage control with gossip communication. In *ASME Conference Proceedings*. ASME, 2009.

R. Durrett. *Probability: Theory and Examples*. Duxbury Press, 1990.

J. C. Fort and G. Pagès. On the a.s. convergence of the kohonen algorithm with a general neighborhood function. *The Annals of Applied Probability*, 5(4):1177–1216, 1995.

J. C. Fort and G. Pagès. Convergence of stochastic algorithms: from the kushner-clark theorem to the lyapounov functional method. *Advances in Applied Probability*, 28(4):1072–1094, 1996.

P. Frasca, R. Carli, and F. Bullo. Multiagent coverage algorithms with gossip communication: Control systems on the space of partitions. In *American Control Conference*, 2009.

A. Gersho and R. M. Gray. *Vector Quantization and Signal Compression*. Kluwer, 1992.

S. Graf and H. Luschgy. *Foundations of Quantization for Probability Distributions*. Springer-Verlag, 2000.

M. Inaba, N. Katoh, and H. Imai. Applications of weighted voronoi diagrams and randomization to variance-based *k*-clustering: (extended abstract). In *SCG '94: Proceedings of the Tenth Annual Symposium on Computational Geometry*, 1994.

D. Jungnickel. *Graphs, Networks and Algorithms*. Springer-Verlag, 1999.

T. Kohonen. Analysis of a simple self-organizing process. *Biological Cybernetics*, 44(2):135–140, 1982.

H. J. Kushner and D. S. Clark. *Stochastic Approximation for Constrained and Unconstrained Systems*. Springer-Verlag, 1978.

T. Linder. On the training distortion of vector quantizers. *IEEE Transactions on Information Theory*, 46(4):1617–1623, 2000.

T. Linder. Learning-theoretic methods in vector quantization. In *Lecture Notes for the Advanced School on the Principles of Nonparametric Learning*, 2001.

T. Linder, G. Lugosi, and K. Zeger. Rates of convergence in the source coding theorem, in empirical quantizer design, and in universal lossy source coding. *IEEE Transactions on Information Theory*, 40(6):1728–1740, 1994.

S. Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2):129–137, 2003.

J. B. MacQueen. Some methods of classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1967.

R. McDonald, K. Hall, and G. Mann. Distributed training strategies for the structured perceptron. In *HLT 2010: Human Language Technologies*, 2010.

N. Murata. A statistical study of on-line learning. In *Online Learning and Neural Networks*, 1998.

G. Pagès. A space vector quantization for numerical integration. *Journal of Applied and Computational Mathematics*, 89(1):1–38, 1997.

D. Pollard. Stong consistency of *k*-means clustering. *The Annals of Statistics*, 9(1):135–140, 1981.

D. Pollard. Quantization and the method of *k*-means. *IEEE Transactions on Information Theory*, 28(2):199–205, 1982a.

D. Pollard. A central limit theorem for *k*-means clustering. *The Annals of Probability*, 28(4):199–205, 1982b.

H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.

M. J. Sabin and R. M. Gray. Global convergence and empirical consistency of the generalized lloyd algorithm. *IEEE Transactions on Information Theory*, 32(2):148–155, 1986.

J. Tsitsiklis. *Problems in Decentralized Decision Making and Computation*. PhD thesis, Department of EECS, MIT, Cambridge, USA, 1984.

J. Tsitsiklis, D. Bertsekas, and M. Athans. Distributed asynchronous deterministic and stochastic gradient optimization algorithms. *IEEE Transactions on Automatic Control*, 31(9):803–812, 1986.

M. Zinkevich, A. Smola, and J. Langford. Slow learners are fast. In *Advances in Neural Information Processing Systems 22*. Curran Associates, 2009.