



**HAL**  
open science

## (anti-Omega<sup>x</sup>Σ<sub>z</sub>) based k-set agreement algorithms

Zohir Bouzid, Corentin Travers

► **To cite this version:**

Zohir Bouzid, Corentin Travers. (anti-Omega<sup>x</sup>Σ<sub>z</sub>) based k-set agreement algorithms. OPODIS, Dec 2010, Tozeur, Tunisia. pp.189-204, 10.1007/978-3-642-17653-1\_16 . hal-00992778

**HAL Id: hal-00992778**

**<https://inria.hal.science/hal-00992778>**

Submitted on 19 May 2014

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# (anti- $\Omega^x \times \Sigma_z$ )-based $k$ -set Agreement Algorithms

Zohir Bouzid<sup>1</sup> and Corentin Travers<sup>2\*</sup>

<sup>1</sup> University Pierre et Marie Curie - Paris 6, LIP6-CNRS 7606, France.  
zohir.bouzid@lip6.fr

<sup>2</sup> LaBRI University Bordeaux 1 travers@labri.fr

**Abstract.** This paper considers the  $k$ -set agreement problem in a crash-prone asynchronous message passing system enriched with failure detectors. Two classes of failure detectors have been previously identified as necessary to solve asynchronous  $k$ -set agreement: the class anti-leader anti- $\Omega^k$  and the weak-quorum class  $\Sigma_k$ . The paper investigates the families of failure detector (anti- $\Omega^x$ )<sub>1 ≤ x ≤ n</sub> and ( $\Sigma_z$ )<sub>1 ≤ z ≤ n</sub>. It characterizes in an  $n$  processes system equipped with failure detectors anti- $\Omega^x$  and  $\Sigma_z$  for which values of  $k, x$  and  $z$   $k$ -set-agreement can be solved. While doing so, the paper (1) disproves previous conjunctures about the weakest failure detector to solve  $k$ -set-agreement in the asynchronous message passing model and, (2) introduces the first indulgent algorithm that tolerates a majority of processes failures.

**Keywords:** Set-agreement, asynchrony, failure detectors, indulgent algorithms.

## 1 Introduction

The  $k$ -set-agreement problem  $k$ -set-agreement [9] is one of the fundamental problem in fault tolerant distributed computing. In this problem,  $n$  processes starting each with an initial private value are required to agree on at most  $k$  values chosen among their initial values. The problem generalizes the *consensus* problem, which corresponds to the case where  $k = 1$ . In an *asynchronous* system, it is well known that 1-set-agreement is impossible as soon as at least one process may fail by crashing [16], whereas the case  $k = n$  does not require any coordination at all. For intermediate values of  $k$  ( $1 < k < n$ ), asynchronous  $k$ -set agreement tolerating  $t$  crash failures is possible if and only if  $k > t$  [6, 24, 29].

*Failure detectors* A *failure detector* is a distributed oracle that provides processes with possibly unreliable information on failures [8]. According to the quality of the information, several classes of failure detectors can be defined. Starting with [26, 30], the failure detector approach has been investigated to alleviate the  $k$ -set-agreement impossibility in asynchronous systems. An algorithm that tolerates

---

\* On leave from University Pierre et Marie Curie - Paris 6. Supported by the ANR project SPREADS and by the INRIA project REGAL.

unreliable failure detection is said to be *indulgent* towards its failure detector [18, 20]. Informally, an indulgent algorithm is always *safe*: it never violates the safety part of the problem it is supposed to solve, even when the underlying failure detector gives false information about failures.

*The quest for the weakest failure detector for  $k$ -set-agreement* Given a distributed problem  $P$ , a natural question is to determine the *weakest failure detector* for  $P$ , that is a failure detector  $D$  which is both *sufficient* to solve the problem – there is an asynchronous algorithm based on  $D$  that solves  $P$  – and *necessary*, in the sense that any failure detector  $D'$  that allows solving  $P$  can be used to emulate  $D$ .

The question of the weakest failure detector class for  $k$ -set agreement ( $1 < k < n$ ) has been first stated in [28]. This line of research [10, 11, 19, 23] culminated with the work of Zieliński who established that the failure detector class  $\text{anti-}\Omega^{n-1}$  is the weakest to solve  $(n-1)$ -set-agreement in the wait-free shared memory model [31]. This has later been generalized to any  $k$ ,  $1 \leq k < n$  by three independent groups [2, 14, 17]. Informally, a failure detector  $\text{anti-}\Omega^k$  outputs sets of  $n-k$  process ids such that some non faulty process id eventually never appear in the outputs.

The situation is different in the message passing model where the answer is known only for the two boundaries cases, i.e.,  $k=1$  (consensus) and  $k=n-1$  [13]. For consensus ( $k=1$ ), it has been shown that the class of *eventual leader* failure detector  $\Omega = \text{anti-}\Omega^1$  is the weakest failure detector in the asynchronous message passing model in which a majority of processes are non-faulty ( $t < \frac{n}{2}$ ) [7]. This result is generalized to the *wait-free* environment in [12] where it is shown that  $\Omega \times \Sigma$  is the weakest failure detector class for consensus when  $t < n$ . Intuitively, failure detector  $\Sigma$  provides a reliable quorum system: when queried, a failure detector of the class  $\Sigma$  returns a sets of processes ids, such that (1) any two sets intersect and (2) eventually, every set contains only ids of correct processes. Actually,  $\Sigma$  is the weakest failure detector to implement a register in the message passing model [5, 12].

Recently, the failure detector family  $(\Sigma_k \times \Omega^k)_{1 \leq k < n}$  has been conjectured to be the weakest failure detector classes for  $k$ -set-agreement [4]. Failure detector  $\Sigma_k$  and  $\Omega^k$  generalizes the classes  $\Sigma$  and  $\Omega$  respectively. Intuitively, a failure detector  $\Sigma_k$  allows up to  $k$  partitions: any collection of  $k+1$  sets outputs by the failure detector contain at least two intersecting sets.  $\Omega^k$ , which has been introduced by Neiger [27], outputs sets of  $k$  ids that eventually converge to a set including the id of a non-faulty process. It is shown in [4] that  $\Sigma_{n-1} \times \Omega^{n-1}$  is equivalent to the loneliness failure detector  $\mathcal{L}$  which is the weakest failure detector class for  $(n-1)$ -set-agreement [13]. Before this paper, nothing specific was known about the power of  $\Sigma_x \times \Omega^x$  to solve  $k$ -set-agreement, for  $1 < x < n-1$ .

*Content of the paper* The paper investigates in the message passing model the computational power of the failure detector families  $(\Sigma_x)_{1 \leq x \leq n}$  and  $(\text{anti-}\Omega^z)_{1 \leq z \leq n}$  as far as  $k$ -set-agreement is concerned. Its main contributions are the following:

1. It has been shown that  $\Sigma_k$  is necessary to solve  $k$ -set-agreement, for each  $k, 1 \leq k \leq n - 1$  [4]. Moreover, for  $k = 1$ ,  $\Sigma_1 = \Sigma$  alone is not powerful enough to solve consensus whereas  $\Sigma_{n-1}$  is sufficient to solve  $(n - 1)$ -set-agreement [4, 13]. We give necessary and sufficient conditions on the values of  $k$ ,  $x$  and  $n$  in order to  $k$ -set-agreement to be solvable in an  $n$  processes message passing system enriched with  $\Sigma_x$  (Theorem 1, section 3). Roughly speaking, we show that  $\Sigma_x$  allows to eliminate at most  $\lfloor \frac{n}{x+1} \rfloor$  initial values, thereby generalizing prior results for the cases  $k = 1$  [11] and  $k = n - 1$  [13].
2. The paper then investigates the combined power of  $\Sigma_x$  and anti- $\Omega^z$ . For  $k \geq xz$ , we present a  $k$ -set-agreement algorithm that tolerates any number of failures (Section 5).

To ensure safety, namely that no more than  $x$  values are decided, we design a non-trivial generalization of the **alpha** abstraction which is at the core of indulgent consensus [21]. Our abstraction (called **alpha** <sub>$x$</sub> , section 4) can be seen as an obstruction-free object that allows processes to store and retrieve at most  $x$  distinct values. Its implementation relies solely on a failure detector of the class  $\Sigma_x$ . Of note, as  $\Sigma_x$  can be simulated in an asynchronous message passing system when  $t < \frac{xn}{x+1}$ , we obtain a  $xz$ -set-agreement algorithm which is indulgent (towards the underlying failure detector of the class anti- $\Omega^z$ ) and tolerates  $t < \frac{xn}{x+1}$  failures. To our knowledge, every prior indulgent algorithm assumes a majority of correct processes ( $t < n/2$ ) or relies on a strong failure detector (e.g.,  $\Sigma$ ) that cannot be implemented in the asynchronous message passing model when a majority of processes may fail ( $t \geq n/2$ ).

3. Finally, we show that for large enough values of  $n$ , there is no  $k$ -set-agreement algorithm based on  $\Sigma_x \times \Omega^z$  if  $k < xz$  (Theorem 2, section 5). This last result has two noteworthy corollaries. First, as anti- $\Omega^z$  can easily be simulated using the output of  $\Omega^z$ , it implies that the previous algorithm is optimal. Second, it rules out  $\Pi_k = \Sigma_k \times \Omega^k$  as a weakest failure candidate for  $k$ -set-agreement, thus disproving Bonnet and Raynal’s conjuncture [4].

*Roadmap* The paper is made up of 6 sections. Section 2 describes the computing model and the families of failure detector we are interested in. Section 3 investigates the power of  $\Sigma_x$  with respect to the solvability of  $k$ -set agreement. The **alpha** <sub>$k$</sub>  abstraction is introduced in section 4, which presents also an  $\Sigma_k$ -based implementation. Section 5 then describes an indulgent  $k$ -set agreement algorithm that relies on the previous abstraction and a failure detector of the class anti- $\Omega^x$ . A matching impossibility result is also presented. Finally, section 6 provides some concluding remarks. Due to space limitations, some proofs are presented in a companion technical report [?].

## 2 System Model and Failures Detectors

*Asynchronous message passing system with process crash failures* The system consists in a set of  $n$  processes denoted  $\Pi = \{p_1, \dots, p_n\}$ . Processes are asyn-

chronous and may fail by crashing. Processes communicate via sending and receiving messages over an asynchronous network. Each pair of processes is connected by a bi-directional channel. The channels are asynchronous but reliable. Reliable means that there is no creation, alteration or loss of messages whereas asynchronous means that message transfer delays are finite but unbounded.

Processes may fail by *crashing*, i.e., prematurely stop executing their code. A process is *correct* in an execution if it never crashes in this execution; otherwise it is *faulty*.  $t(1 \leq t < n)$  denotes an upper bound on the number of processes that can crash in a run. Given an execution, *Correct* denotes the set of correct processes.

*Notation* As in [25],  $\mathcal{MP}_{n,t}$  denotes the asynchronous distributed system made of  $n$  processes, among which at most  $t$  may crash in any run.  $\mathcal{MP}_{n,t}[X]$  denotes a system enriched with a failure detector of a class  $X$ .

*The  $k$ -set agreement problem* In the  $k$ -set agreement problem, each process proposes a value and has to decide a value such that the following properties are satisfied: (*Validity*) A decided value is a proposed value; (*Termination*) Every correct process eventually decides a value; (*Agreement*) The number of distinct decided values is at most  $k$ .

*Families of failure detector classes* For process  $p_i$ ,  $FD_i^\tau$  is the value output by the failure detector at time  $\tau$ .

- *The eventual leader family*  $(\Omega^k)_{1 \leq k \leq n}$ . This family has been introduced in [27] to generalize the class of failure detectors  $\Omega$  defined in [7], with  $\Omega^1 = \Omega$ . A failure detector of the class  $\Omega^k$  maintains at each process  $p_i$  a set of processes of size at most  $k$  (denoted  $LEADER_i$ ) that satisfies the following property:
  - (Eventual multiple leadership). There is a time after which the sets  $LEADER_i$  contains forever the same set of processes and at least one process of this set is correct.
- *The quorum family*  $(\Sigma_k)_{1 \leq k \leq n}$  [4]. A failure detector of the class  $\Sigma_k$  maintains at each process  $p_i$  a variable  $TRUSTED_i$  that contains a set of processes. The family generalizes the “quorum” failure detector  $\Sigma = \Sigma_1$  introduced in [12]. The sets output by a failure detector of the class  $\Sigma_z$  satisfy:
  - (Completeness) There is a time after which every set  $TRUSTED_i$  contains only correct processes.
  - (Intersection) For every set  $\mathcal{Q} = \{Q_1, \dots, Q_{k+1}\}$  of  $k+1$  sets output by the failure detector, there exists  $Q_i, Q_j \in \mathcal{Q}, i \neq j$  such that  $Q_i \cap Q_j \neq \emptyset$ .
 Of note, a failure detector  $\Sigma_k$  can be implemented in  $\mathcal{MP}_{n,t}$  provided that  $\frac{kn}{k+1} > t$ . To simulate a failure detector query, a process sends a REQUEST message to all processes and waits for matching RESPONSES. The set  $X$  made of the ids of the senders of the first  $n-t$  responses received defines the result of the query. It is easy to see that completeness is ensured: eventually, only correct processes send responses. The intersection property follows from

the fact that each simulated query returns a set of  $n-t \geq \lfloor \frac{n}{k+1} \rfloor + 1$  identities. Hence, any collection of  $k+1$  such sets contains at least two intersecting sets.

- *The anti- $\Omega$  family* (anti- $\Omega^k$ ) $_{1 \leq k \leq n}$  [31]. A failure detector of the class anti- $\Omega^k$  outputs at each process  $p_i$  a set ANTI-LEADER $_i$  of  $n-k$  processes ids. anti- $\Omega^1$  is equivalent to  $\Omega$ . In every run, there is a correct process such that eventually each set output by the failure detector does not contain the identity of this process.

- (Anti-leadership)  $\exists p_c \in \text{Correct}, \exists \tau$  such that  $\forall \tau' \geq \tau, \forall p_i \in \Pi, c \notin \text{ANTI-LEADER}_i^{\tau'}$ .

### 3 $\Sigma_z$ and $k$ -set-agreement

Among other results, [11] shows that there is a  $k$ -set-agreement algorithm based on  $\Sigma_1$  if  $k > n/2$ . On the other side ( $k = n-1$ ), in [13] a  $(n-1)$ -set agreement message passing algorithm is presented. The algorithm relies on a failure detector called  $\mathcal{L}$ , which has been proved in [4] to be equivalent to  $\Sigma_{n-1}$ . Actually, it is also shown in [13] that failure detector  $\mathcal{L}$  is the weakest failure detector for  $(n-1)$ -set-agreement in the wait-free message passing model ( $t = n-1$ ). We generalize these boundary results to the entire family  $(\Sigma_z)_{1 \leq z \leq n}$ . Specifically, we present a  $k$ -set-agreement algorithm based on  $\Sigma_z$ , provided that  $k \geq n - \lfloor \frac{n}{z+1} \rfloor$ . A simple matching impossibility result is also presented.

**Theorem 1.** *The  $k$ -set-agreement problem can be solved in  $\mathcal{MP}_{n,n-1}[\Sigma_z]$  if and only if  $k \geq n - \lfloor \frac{n}{z+1} \rfloor$*

*Solving  $k$ -set-agreement with  $\Sigma_z$*  The algorithm combines ideas borrowed from the  $(n-1)$ -set-agreement protocol based on failure detector  $\mathcal{L}$  presented in [13] and a  $k$ -set-agreement protocol based on  $\sigma_{2k}$  [11]. In short, a failure detector of the class  $\sigma_{2k}$  provides the properties of the class  $\Sigma$  only to a subset of size  $2k$  of the system. The algorithm is described in Figure 1.

Let  $A_1, \dots, A_{z+1}$  be a partition of the set of processes such that  $\forall i, 1 \leq i \leq z, |A_i| = \lfloor \frac{n}{z+1} \rfloor$  and  $|A_{z+1}| = \lfloor \frac{n}{z+1} \rfloor + (n \bmod (z+1))$ . Each process in set  $A_i$  tries to decide the proposal of some process that belongs to some partition  $A_j, j < i$ . To that end, each process  $p \in A_i$  first sends its proposal to all processes in “higher” partitions, i.e., the processes that belong to the sets  $A_{i+1}, \dots, A_{z+1}$  (line 1). When a process receives a value  $w$  from a “lower” partition, it decides that value after broadcasting a *DEC* message carrying that value (line 5). A process that has not yet decided also decides  $w$  when it receives such a message *DEC*( $w$ ) (Task T3). Note that the initial values of the processes in the “highest” partition ( $A_{z+1}$ ) cannot be decided using this mechanism. Hence at most  $n - |A_{z+1}| = z \lfloor \frac{n}{z+1} \rfloor$  are decided in that way.

The mechanism sketched above allows every correct process to eventually decide as soon as at least two partitions contain correct processes. However, it may happen that all correct processes are contained in a single partition  $A_i$ . We notice that in that case, the failure detector output at each process

is eventually contained in  $A_i$  (by the completeness property of the class  $\Sigma_z$ ). Henceforth, to prevent processes from waiting for values forever, each process  $p_i$  periodically checks its failure detector output; If the current set of trusted processes is contained in  $p_i$ 's partition,  $p_i$  is allowed to decide its initial value (task T2, lines 6-8). The proof shows (Lemma 1) that the total number of decided values is at most  $k = z \lfloor \frac{n}{z+1} \rfloor + (n \bmod (z+1))$ .

```

init  $A_1, \dots, A_{z+1}$  sets of processes such that  $\forall i, j, i \neq j, A_i \cap A_j = \emptyset; \bigcup A_i = \Pi;$ 
 $\forall i \in [1..z] |A_i| = \lfloor \frac{n}{z+1} \rfloor; |A_{z+1}| = \lfloor \frac{n}{z+1} \rfloor + n \bmod (z+1)$ 

propose( $v$ ) % code for process  $p \in A_i$ 
(1) foreach  $q \in \bigcup_{j>i} A_j$  do send VAL( $v$ ) to  $q$  endfor
(2) start tasks T1, T2, T3

(3) Task T1: when VAL( $w$ ) is received do
(4) foreach  $q \in \Pi$  do send DEC( $w$ ) to  $q$  enddo
(5) decide  $w$ ; return

(6) Task T2: repeat  $X \leftarrow \Sigma_z\text{-QUERY}()$  until  $X \subseteq A_i$ 
(7) foreach  $q \in \Pi$  do send DEC( $v$ ) to  $q$  enddo
(8) decide  $v$ ; return

(9) Task T3: when DEC( $w$ ) is received
(10) foreach  $q \in \Pi$  do send DEC( $w$ ) to  $q$  enddo
(11) decide( $w$ ); return

```

**Fig. 1.**  $k$ -set agreement algorithm in  $\mathcal{MP}_{n,n-1}[\Sigma_z]$ ,  $k = z \lfloor \frac{n}{z+1} \rfloor + (n \bmod (z+1))$

**Lemma 1.** *The protocol described in the figure 1 solves  $k$ -set agreement in  $\mathcal{MP}_{n,n-1}[\Sigma_z]$  for  $k \geq n - \lfloor \frac{n}{z+1} \rfloor$*

*An impossibility result* Together with Lemma 1, the following lemma completes the proof of Theorem 1.

**Lemma 2.**  *$\forall n, k, z$  such that  $k < n - \lfloor \frac{n}{z+1} \rfloor$ , there is no  $k$ -set-agreement algorithm in  $\mathcal{MP}_{n,n-1}[\Sigma_z]$*

## 4 The $\mathbf{Alpha}_k$ abstraction

This section presents the  $\mathbf{Alpha}_k$  abstraction that generalizes the  $\mathbf{Alpha}$  abstraction introduced by Guerraoui and Raynal in [21] to capture the safety part of indulgent consensus<sup>3</sup>. In the very same way, the abstraction  $\mathbf{Alpha}_k$  captures the safety part of eventual failure detector based  $k$ -set-agreement algorithms. In short, the  $\mathbf{Alpha}_k$  abstraction can be viewed as a shared object intended to store at most  $k$  values. A process accesses the object via the operation  $\mathbf{propose}(\cdot)$  with as parameter a value it is willing to store and gets back one of the values

<sup>3</sup> Another generalization has been introduced in [28]. The implementation presented there relies on atomic registers which are not available in our settings.

actually stored in the object. However, in case of concurrent accesses,  $\text{propose}(\cdot)$  operations may not store any value and return the special value  $\perp$ , which is the object initial value.

More precisely, an  $\text{alpha}_k$  object exports one operation  $\text{propose}(v, r)$  with input parameters a value  $v$  and a round number  $r$ . As in [21], distinct processes must input distinct round numbers and each process must use strictly increasing round number. The  $\text{Alpha}_k$  abstraction is specified by the following properties, where  $\perp$  is a special value that cannot be proposed:

- *Termination.* Every invocation of  $\text{propose}(\cdot)$  by any non-faulty process returns.
- *Validity.* If the invocation  $\text{propose}(v, r)$  returns  $v' \neq \perp$ , then  $\text{propose}(v', r')$  with  $r' \leq r$  has been invoked.
- *k-Quasi-Agreement.* Let  $V$  be the set of non- $\perp$  values that are returned by  $\text{propose}(\cdot)$  invocations.  $|V| \leq k$ .
- *Conditional non- $\perp$  convergence.* Let  $I = \text{propose}(\cdot, r)$  be a terminating invocation. If for every invocation  $I' = \text{propose}(\cdot, r')$  that starts before  $I$  returns, we have  $r' < r$ ,  $I$  returns a non- $\perp$  value.

#### 4.1 Implementing $\text{Alpha}_k$ with $\Sigma_k$

The algorithm implementing  $\text{Alpha}_k$  in an asynchronous message passing system is described in Figure 2. The algorithm relies on an underlying failure detector of the class  $\Sigma_k$ . It tolerates any number of failures.

*Algorithm principles* At any time, each process  $p_i$  has a value  $v$  (initially  $\perp$ ) stored in the local variable  $val_i$  and a pair of integers  $\langle r, \rho \rangle$  stored in the variables  $\langle lre_i, pos_i \rangle$ . The pair  $\langle r, \rho \rangle$  can be seen as the *priority* of value  $v$  from  $p_i$ 's point view. As in [21], *lre* stands for *last round entered*.  $r$  is the highest round number passed as a parameter of a  $\text{propose}(\cdot)$  operation so far, as far as  $p_i$  knows. Furthermore, each round  $r$  is associated with a sequence of *positions* numbered from 1 to  $2^r$ . When  $\langle lre_i, pos_i, val_i \rangle = \langle r, \rho, v \rangle$ , we say that *value  $v$  has reached position  $\rho$  in round  $r$* . Also, based on its position  $\rho$  at round  $r$ , value  $v$  *logically occupies* a position  $\rho'$  at round  $r + \delta$ , for each  $\delta > 0$ .  $\rho'$  is defined by the following function  $g$ :

$$g(\rho, \delta) = 2^\delta(\rho - 1) + 1$$

Any pair of triplets  $\langle r, \rho, v \rangle, \langle r', \rho', v' \rangle$ ,  $r \leq r'$  can be compared via the function  $g$ :  $\langle r, \rho, v \rangle \prec \langle r', \rho', v' \rangle$ , i.e.,  $v$  has a priority lower than  $v'$  iff  $g(\rho, r' - r) < \rho'^4$ .

An operation  $\text{propose}(v, r)$  returns a value  $v' \neq \perp$  (possibly  $v' \neq v$ ) only if  $v'$  has obtained a priority high enough so that no more than  $k - 1$  values  $\neq v'$  can be awarded higher priority. Operationally, a process  $p_i$  that invokes  $\text{propose}(v, r)$  proceeds as follows:

<sup>4</sup> When  $g(\rho, r' - r) = \rho'$ ,  $v$  has a lower priority if  $v < v'$ . One can check that the  $\prec$  relation is transitive, so  $g$  induces a total order on triplets  $\langle r, \rho, v \rangle$ .

- In the first phase (lines 1-7), process  $p_i$  broadcasts the message  $\text{REQ\_R}(r)$  in order (1) to inform other processes that it has entered round  $r$  and (2) to collect triplets  $\langle \text{round}, \text{position}, \text{value} \rangle$  held by other processes.

When a process  $p_j$  receives a message  $\text{REQ\_R}(r)$ , it first updates its round and the position of its value (using the function  $g$ ) if  $r > \text{pre}_j$ . It then sends back the current value of its variables  $\langle \text{pre}_i, \text{pos}_i, \text{val}_i \rangle$  in a response message  $\text{RSP\_R}$  (lines 17-18).

$p_i$  is done collecting  $\langle \text{round}, \text{position}, \text{value} \rangle$  triplets when it has received such values from each process  $p_j$  in a *quorum*, that is a set of processes returned by a query to the underlying failure detector  $\Sigma_k$ . If  $p$  discovers that another  $\text{propose}(\cdot)$  operation with input  $r' > r$  has already started, it returns  $\perp$  (line 5). Note that this does not violate the conditional convergence property. Otherwise,  $p_i$  selects among the values received the triplet with the highest priority, and updates its  $\langle \text{pre}_i, \text{pos}_i, \text{val}_i \rangle$  accordingly (lines 6). In the case no triplets contain a value  $\neq \perp$ ,  $p_i$  selects its own value with position 0 (line 7).

- The second phase (lines 8-16) consists in a repeat loop. In each iteration of the loop,  $p_i$  tries to increment the position of the value currently stored in  $\text{val}_i$ . To that end, it first broadcasts a request message  $\text{REQ\_W}$  that carries  $p_i$ 's current value together with its position and the current round  $r$  (lines 9).

Process  $p_j$  that has learned that a round  $> r$  has been started ignores the content of the messages  $\text{REQ\_W}(\langle r, \rho, v \rangle)$  it receives. Otherwise,  $p_j$  updates its round number and the position of its value. In addition, it adopts the received value if it has higher priority (lines 19-24). Finally,  $p_j$  answers with a message  $\text{RSP\_W}$  that carries the updated values of its variables  $\langle \text{pre}_i, \text{pos}_i, \text{val}_i \rangle$  (lines 25).

As in the first phase,  $p_i$  stops collecting responses matching its request when a response message  $\text{RSP\_W}(\cdot)$  has been received from each process  $p_j$  in a quorum  $Q$ . Similarly, if one of the response carries a round number  $> r$ ,  $p_i$  returns  $\perp$ . If this not the case,  $p_i$  adopts among the values received the triplet with the highest priority, and updates its  $\langle \text{pre}_i, \text{pos}_i, \text{val}_i \rangle$  variables accordingly (lines 14). Since  $p_i$  always receives a response from itself, the value of  $\text{pos}_i$  at the end of the iteration is greater than the value of this variable at the end of the previous iteration. Finally, if the current value  $v$  of  $p_i$  reaches the last position associated with round  $r$ ,  $v$  is returned (lines 15-16).

*k-Quasi agreement* The main difficulty is to guarantee that  $\text{propose}(\cdot)$  invocations return collectively no more than  $k$  non- $\perp$  values. Value  $v_1$  is returned at round  $r_1$  if it reaches position  $\rho_1 = 2^{r_1}$  and it has been adopted by a quorum  $Q_1$ . This means that for each process  $q \in Q_1$ , there is a point in time  $\tau_q$  at which we have  $\langle \text{pre}_q, \text{pos}_q, \text{val}_q \rangle = \langle r_1, \rho_1, v_1 \rangle$ . However, because quorums may not intersect, another value  $v' \neq v_1$  may reach an arbitrary high position and consequently replaces the value  $v_1$  at each process  $q \in Q_1$ . For example, this

```

init  $lre_i \leftarrow 0$ ;  $val_i \leftarrow \perp$ ;  $pos_i \leftarrow 0$ ;

function propose( $r, v$ )
(1) for each  $j \in \Pi$  send REQ_R( $r$ ) end for;
(2) repeat  $Q \leftarrow \Sigma_k\text{-QUERY}()$ 
(3) until ( $\forall p_j \in Q \cup \{p_i\} : \text{RSP\_R}(r, \langle lre_j, pos_j, val_j \rangle)$  has been received from  $p_j$ )
(4) let  $RCV = \{ \langle lre_j, pos_j, val_j \rangle : \text{RSP\_R}(r, \langle lre_j, pos_j, val_j \rangle) \text{ has been received} \}$ ;
(5) if ( $\exists lre : \langle lre, -, - \rangle \in RCV : lre > lre_i$ ) then return ( $\perp$ ) endif
(6) let  $pos_M = \max\{pos : \langle r, pos, v \rangle \in RCV\}$ ;
 $val_i \leftarrow \max\{v : \langle r, pos_M, v \rangle \in RCV\}$ ;  $pos_i \leftarrow pos_M$ ;
(7) if  $val_i = \perp$  then  $val_i \leftarrow v$  endif
(8) repeat  $pos_i \leftarrow pos_i + 1$ ;
(9) for each  $p_j \in \Pi$  send REQ_W( $\langle r, pos_i, val_i \rangle$ ) to  $p_j$  end for
(10) repeat  $Q \leftarrow \Sigma_k\text{-QUERY}()$ 
(11) until ( $\forall p_j \in Q \cup \{p_i\} : \text{RSP\_W}(r, pos_i, \langle lre_j, pos_j, val_j \rangle)$  has been received from  $p_j$ )
(12) let  $RCV = \{ \langle lre_j, pos_j, val_j \rangle : \text{RSP\_W}(r, pos_i, \langle lre_j, pos_j, val_j \rangle) \text{ has been received} \}$ ;
(13) if ( $\exists lre_j, \langle lre_j, -, - \rangle \in RCV : lre_j > r$ ) then return ( $\perp$ ) end if
(14) let  $pos_M = \max\{pos : \langle r, pos, v \rangle \in RCV\}$ ;
 $val_i \leftarrow \max\{v : \langle r, pos_M, v \rangle \in RCV\}$ ;  $pos_i \leftarrow pos_M$ ;
(15) until ( $pos_i = 2^r$ )
(16) return ( $val_i$ )

when REQ_R( $rd$ ) is received from  $p_j$ 
(17) if  $rd > lre_i$  then  $pos_i \leftarrow g(pos_i, rd - lre_i)$ ;  $lre_i \leftarrow rd$  end if
(18) send RESP_R( $rd, \langle lre_i, pos_i, val_i \rangle$ ) to  $p_j$ 

when REQ_W( $\langle rd, pos_j, val_j \rangle$ ) is received from  $p_j$ 
(19) if ( $rd \geq lre_i$ ) then  $pos_i \leftarrow g(pos_i, rd - lre_i)$ ;  $lre_i \leftarrow rd$ 
(20) case  $pos_j > pos_i$  then  $val_i \leftarrow val_j$ ;  $pos_i \leftarrow pos_j$ 
(21)  $pos_i = pos_j$  then  $val_i \leftarrow \max(v_i, v_j)$ 
(22)  $pos_j < pos_i$  then nop
(23) end case
(24) end if
(25) send RSP_W( $rd, pos_j, \langle lre_i, pos_i, val_i \rangle$ ) to  $p_j$ 

```

**Fig. 2.** Implementing  $\text{Alpha}_k$  with  $\Sigma_k$  (code for  $p_i$ )

might happen if the quorums output by the failure detector during  $\text{propose}(r', -)$  invocations with  $r' > r_1$  do not intersect with  $Q_1$ . In these invocations,  $v'$  may be selected at the end of the first phase and its position can be increased in the second phase. In that case,  $v'$  has an higher priority than  $v_1$ , i.e., a process  $q \in Q_1$  that receives  $\langle r', \rho', v' \rangle$  will adopts  $v'$ .

The key idea of the algorithm is as follows. Fix some round  $r' > r_1$ . In order to value  $v'$  to “overtake” value  $v_1$  in round  $r'$ ,  $v'$  has to be adopted by a quorum  $Q'$  that does *not intersect* with  $Q_1$ . Consider the positions associated with round  $r'$ . At the beginning of round  $r'$ , an odd position  $x$  might be logically occupied by a value  $v$ . This is the case if for some process  $p$  and some round  $r < r'$ , we have  $\langle lre_p, pos_p, val_p \rangle = \langle r, \rho, v \rangle$  and  $g(\rho, r' - r) = x$ . Differently, by definition of  $g$ , each even position is initially free. Let  $x'$  and  $x_1 = g(\rho_1, r' - r_1)$  be the positions logically occupied by values  $v'$  and  $v_1$  respectively at the beginning of round  $r'$ . Observe that positions are increased by step of 1 and  $x' + 2 \leq x_1$ . So, to reach position  $x_1$  value  $v'$  must first successfully go through position  $x_1 - 1$ . This can only happen if there is quorum  $Q'$  that adopts  $\langle r', x_1 - 1, v' \rangle$ . For each process  $q \in Q_1$ , the value  $v_1$  held by  $q$  has an higher priority, since it logically occupies position  $x_1$ . So  $q$  cannot adopt  $\langle r', x_1 - 1, v' \rangle$ , hence  $Q' \cap Q_1 \neq \emptyset$ .

The rationale above can be extended to a chain of values  $v_1, \dots, v_\ell$  that each reaches higher and higher priorities to imply the existence of  $\ell$  pairwise disjoint quorums. As any collection of  $k + 1$  quorums contains at least two intersecting quorums, the length of such a chain is at most  $k$ . In particular, this implies that at most  $k$  distinct values are returned – see the second part of the proof for more details (Lemmas 6–10).

*Remark* The algorithm is generic in the sense that the parameter  $k$  is never explicitly used in the code. In order to implement an  $\text{Alpha}_{k'}$  abstraction, it is sufficient to replace the underlying failure detector by a failure detector in the class  $\Sigma_{k'}$ . On the other hand, the algorithm uses  $2^r$  positions per round. We have also developed along the same principles an algorithm that uses  $O(r^{k-1})$  positions per round. However, determining which is the round  $r'$  position corresponding to a round  $r < r'$  position, i.e., defining the equivalent of the  $g$  function, is more involved. As a result, the correctness proof is more intricate.

## 4.2 Proof

Consider a *well-formed* execution, in which processes execute the algorithm described in Figure 2 when  $\text{propose}(\cdot)$  is invoked. An execution is well-formed if the following conditions are fulfilled: (1) Only round number  $r > 0$  are used as input parameters; (2) For any invocations  $\text{propose}(\cdot, r)$  and  $\text{propose}(\cdot, r')$  performed by processes  $p$  and  $p'$  respectively, if  $p \neq p'$  then  $r \neq r'$  and, if  $p = p'$  and  $\text{propose}(\cdot, r)$  is invoked before  $\text{propose}(\cdot, r')$  then  $r < r'$ .

**Lemma 3 (Termination).** *Every invocation of  $\text{propose}(\cdot)$  by a correct process terminates.*

**Lemma 4 (Validity).** *Suppose that the invocation  $\text{propose}(r, v)$  returns  $v' \neq \perp$ . Then  $\text{propose}(r', v')$  with  $r' \leq r$  has been invoked by some process.*

**Lemma 5 (Conditional non- $\perp$  convergence).** *Let  $I = \text{propose}(r, \cdot)$  be a terminating invocation. If for every invocation  $I' = \text{propose}(r', \cdot)$  that starts before  $I$  returns we have  $r' < r$ ,  $I$  returns a non- $\perp$  value.*

*k-quasi agreement* The next lemma is central in the proof of the  $k$ -quasi agreement property. In the following, a quorum is a set of processes returned by a query to the underlying  $\Sigma_k$  failure detector.

**Lemma 6.** *Let  $V$  be the set non- $\perp$  values that are returned by the  $\text{propose}(\cdot)$  invocations.  $|V| = x \Rightarrow \exists x$  quorums  $Q_1, \dots, Q_x, \forall 1 \leq i < j \leq x, Q_i \cap Q_j = \emptyset$ .*

The  $k$ -quasi agreement property then follows easily from Lemma 6.

**Lemma 7 ( $k$ -quasi agreement).** *Suppose that the protocol described in Figure 2 is instantiated with a failure detector of the class  $\Sigma_k$ . The total number of non- $\perp$  values that are returned by the  $\text{propose}(\cdot)$  invocations is at most  $k$ .*

*Proof.* Assume for contradiction that  $x > k$  non- $\perp$  values are returned. It then follows from Lemma 6 that at least  $k + 1$  disjoint quorums are output by the underlying failure detector  $\Sigma_k$ . This contradicts the intersection property of the class  $\Sigma_k$ .  $\square$

In order to prove Lemma 6, we define a sequence  $S = s_1, \dots, s_i = \langle r_i, \rho_i, v_i \rangle, \dots$  where for each  $i$ ,  $r_i$  is a round number,  $\rho_i$  a position associated to round  $r_i$ , and  $v_i$  a value. The sequence  $S$  is defined inductively as follows:

- $r_1$  is the smallest round  $r$  such that the invocation  $\text{propose}(\_, r)$  returns a non- $\perp$  value, if any.  $\rho_1 = 2^{r_1}$  and  $v_1$  is the value returned by that invocation.
- Suppose that  $s_1, \dots, s_{i-1}$  have been defined.  $r_i$  is the first round  $r > r_{i-1}$  during which a value  $v \neq \{v_1, \dots, v_{i-1}\}$  reaches a position  $\geq g(\rho_{i-1}, r - r_{i-1})$  (if such a round exists), i.e.,  $r_i = \min \{r : r > r_{i-1}, \exists p_x, \exists v \notin \{v_1, \dots, v_{i-1}\}, \langle \text{lire}_x, \text{pos}_x, \text{val}_x \rangle = \langle r, g(\rho_{i-1}, r - r_{i-1}), v \rangle\}$ .  $v_i$  is then this value, and we define  $\rho_i = g(\rho_{i-1}, r_i - r_{i-1}) - 1$ .

In the next lemma we give a formula for computing values  $\rho_i$ .

**Lemma 8.** *Suppose that  $|S| \geq \ell$ .  $\forall i, 2 \leq i \leq \ell$ ,  $\rho_i = 2^{r_i} (1 - \frac{1}{2^{r_1}} - \dots - \frac{1}{2^{r_{i-1}}})$*

Suppose that value  $v$  reaches position  $\rho$  in round  $r$ , i.e., there exists a process  $p_i$  for which we have  $\langle \text{lire}_i, \text{pos}_i, \text{val}_i \rangle = \langle r, \rho, v \rangle$  at some time. For every round  $r' \geq r$ , value  $v$  then logically occupies round  $r$  position  $g(\rho, r' - r)$ . Indeed, if process  $p_i$  later receives a read or write request carrying round  $r' \geq r$ ,  $\text{pos}_i$  is updated to the value  $g(\rho, r' - r)$  (at line 17 or line 19). Given a round  $r$ , we can then define the highest position logically occupied by value  $v$  as follows:

**Definition 1.** *Given a value  $v$  and a round number  $r$ , let  $\text{mpos}(v, r)$  denotes the maximal position logically occupied by value  $v$  at the beginning of round  $r$ . Formally,  $\text{mpos}(v, r) = \max\{g(\rho', r - r') : \exists p_j, r' < r \text{ and a time at which } \langle \text{lire}_j, \text{pos}_j, \text{val}_j \rangle = \langle r', \rho', v \rangle\}$ ; if no invocation  $\text{propose}(r', v)$  with  $r' < r$  occurs,  $\text{mpos}(v, r) = 0$ .*

**Lemma 9.** *Suppose that  $|S| \geq \ell$ . Let  $\langle r, \rho, v \rangle$  be the value of process  $p_i$  variables  $\langle \text{lire}_i, \text{pos}_i, \text{val}_i \rangle$  at some time. If  $r \leq r_\ell$  and  $v \notin \{v_1, \dots, v_\ell\}$ ,  $g(\rho, r_\ell - r) < \rho_\ell$ .*

**Lemma 10.** *Let  $V$  be the set non- $\perp$  values that are returned by the  $\text{propose}(\cdot)$  invocations. If  $|V| = x$ ,  $s_1, \dots, s_x$  are well defined.*

We are now ready to prove Lemma 6. To do so we associate to each  $s_i \in S$  a quorum  $Q_i$ . Intuitively, the processes in  $Q_i$  are those processes that allow value  $v_i$  to reach position  $\rho_i$  during round  $r_i$ . Each process  $q \in Q_i$  hence holds the triplet  $\langle r_i, \rho_i, v_i \rangle$  at some time. Note that, after that time, the round  $r$  and position  $\rho$  are always such that  $r \geq r_i$  and  $\rho \geq g(\rho_i, r - r_i)$ . The crucial observation is that  $q$  cannot allow any value  $v_j \neq v_i$  to reach position  $\rho_j$ , essentially because either  $r_i > r_j$  (in the case  $i > j$ ) or  $g(\rho_i, r_j - r_i) > \rho_j$  (if  $j > i$ ).

*Proof of Lemma 6.* Suppose that  $|V| = x$ . Let  $\ell, 1 \leq \ell \leq x$ . We first bound  $\text{mpos}(v_\ell, r_\ell)$ . Suppose that  $\langle r, \rho, v_\ell \rangle$  are stored by some process  $p$ , with  $r \leq r_\ell$ . There are two cases:

- $1 \leq r \leq r_{\ell-1}$ . Since  $v_\ell \notin \{v_1, \dots, v_{\ell-1}\}$ , it follows from Lemma 9 that  $g(\rho, r_{\ell-1} - r) < \rho_{\ell-1}$ . Hence,  $g(g(\rho, r_{\ell-1} - r), r_\ell - r_{\ell-1}) < g(\rho_{\ell-1}, r_\ell - r_{\ell-1})$  from which we have  $g(\rho, r_\ell - r) < g(\rho_{\ell-1}, r_\ell - r_{\ell-1})$ .
- $r_{\ell-1} < r < r_\ell$ . By definition of  $s_\ell$ , we have  $\rho < g(\rho_{\ell-1}, r - r_{\ell-1})$ . Therefore  $g(\rho, r_\ell - r) < g(g(\rho_{\ell-1}, r - r_{\ell-1}), r_\ell - r)$  which implies  $g(\rho, r_\ell - r) < g(\rho_{\ell-1}, r_\ell - r_{\ell-1})$ .

We conclude that  $mpos(v_\ell, r_\ell) < g(\rho_{\ell-1}, r_\ell - r_{\ell-1}) = \rho_\ell + 1$ . By definition of  $g(\cdot)$ ,  $\rho + 1$  is odd. Similarly, as there exists  $r' < r_\ell, \rho'$  such that  $mpos(v_\ell, r_\ell) = g(\rho', r_\ell - r')$ ,  $mpos(v_\ell, r_\ell)$  is odd. Consequently  $mpos(v_\ell, r_\ell) < \rho_\ell$ .

We now define a quorum  $Q_\ell$  associated with the triplet  $\langle r_\ell, \rho_\ell, v_\ell \rangle$ . By definition of  $s_\ell$ ,  $r_\ell$  is the first round during which value  $v_\ell$  reaches a position  $\geq \rho_\ell + 1$ . There is a (unique) process  $p_\ell$  that invokes **propose**( $\cdot$ ) with input parameter  $r_\ell$ . Otherwise, round  $r_\ell$  is never entered and value  $v_\ell$  cannot reach position  $g(\rho_{\ell-1}, r_\ell - r_{\ell-1}) = \rho_\ell + 1$  in round  $r_\ell$ .

Note that (1) value  $v_\ell$  reaches a position  $\geq \rho_\ell + 1$  in round  $r_\ell$ , (2) the highest position logically occupied by  $v_\ell$  at the beginning of round  $r_\ell$  is  $< \rho_\ell$ . Moreover, (3) only process  $p_\ell$  increases positions in round  $r_\ell$ , and (4)  $p_\ell$  tries to move at most one value from position  $\phi$  to position  $\phi + 1$ , for every position  $\phi$ . It then follows that  $p_\ell$  successfully moves value  $v_\ell$  from position  $\rho_\ell - 1$  to position  $\rho_\ell + 1$ . In more details, this means that the variable  $pos_\ell$  successively contains the values  $\rho_\ell - 1, \rho_\ell, \rho_\ell + 1$  while the variables  $\langle lre_\ell, val_\ell \rangle$  keep the values  $\langle r_\ell, v_\ell \rangle$ .

In particular, let us consider the iteration of the repeat loop (lines 8-15) in which  $pos_\ell = \rho_\ell$ . Let  $Q_\ell$  be the quorum that allows the inner repeat loop to terminate (lines 10-11). Observe that  $Q_\ell$  is a set of process returned by a query to failure detector  $\Sigma_k$ . For each  $q \in Q_\ell$ , the message **RSP.W** received from  $q$  must carry the triplet  $\langle r_\ell, \rho_\ell, v_\ell \rangle$ . If not,  $p_\ell$  either picks another pair  $\langle \rho, v \rangle$  with  $v \neq v_\ell$  and  $\rho \geq \rho_\ell$  or returns  $\perp$ . In both case,  $p$  stops moving value  $v_\ell$ . It cannot move  $v_\ell$  later in the same round, as the highest position occupied by  $v_\ell$  is  $\rho_\ell$ , and in subsequent iterations, only values located at position  $> \rho_\ell$  can be moved.

Consequently, it follows that  $\forall p_i \in Q_\ell$  there exists a time  $\tau_i^\ell$  at which we have  $\langle lre_i, pos_i, val_i \rangle = \langle r_\ell, \rho_\ell, v_\ell \rangle$ .

Finally, we establish that  $\forall i, j, 1 \leq i < j \leq \ell$ ,  $Q_i \cap Q_j$ . Observe that if  $\langle r_1, \rho_1, v_1 \rangle$  and  $\langle r_2, \rho_2, v_2 \rangle$  are the values of the same process variables  $\langle lre, pos, val \rangle$  at times  $\tau_1 < \tau_2$  respectively,  $(r_1 = r_2 \wedge \rho_1 \leq \rho_2) \vee (r_1 < r_2 \wedge g(\rho_1, r_2 - r_1) \leq \rho_2)$  ( $\star\star\star$ ).

Assume for contradiction that  $\exists \ell, m, 1 \leq \ell < m \leq x$  such that  $Q_\ell \cap Q_m \neq \emptyset$ . Let  $p_i \in Q_\ell \cap Q_m$ . There are two cases:

- $\tau_i^\ell < \tau_i^m$ . In that case,  $p_\ell$  sends first a message **RSP.W** carrying  $\langle r_\ell, \rho_\ell, v_\ell \rangle$  and later a message **RSP.W** carrying  $\langle r_m, \rho_m, v_m \rangle$ . Note that the two triplets are the values at times  $\tau_i^\ell$  and  $\tau_i^m$  respectively of the variables  $\langle lre_i, pos_i, val_i \rangle$ . We have:

$$g(\rho_\ell, r_m - r_\ell) = 2^{r_m} \left(1 - \sum_{j=1}^{\ell} \frac{1}{2^{r_j}}\right) + 1 \quad \text{and} \quad \rho_m = 2^{r_m} \left(1 - \sum_{j=1}^{m-1} \frac{1}{2^{r_j}}\right)$$

from which we obtain  $\rho_m < g(\rho_\ell, r_m - r_\ell)$ , contradicting observation  $(\star\star\star)$ .  
 –  $\tau_i^\ell > \tau_i^m$ . This implies that  $lre_i$  first contains  $r_m$  and later  $r_\ell < r_m$ , which is impossible according to observation  $(\star\star\star)$ .  $\square$

## 5 A $k$ -set agreement algorithm

This section presents an  $(\text{anti-}\Omega^x \times \Sigma_z)$ -based  $k$ -set-agreement protocol, and a matching impossibility result on solving  $k$ -set agreement in the family of systems  $(\mathcal{MP}_{n,n-1}[\text{anti-}\Omega^x, \Sigma_z])_{1 \leq x, z \leq n}$ . The main results of this section are summarized by the following theorem:

**Theorem 2.** *The  $k$ -set-agreement problem can be solved in  $\mathcal{MP}_{n,n-1}[\text{anti-}\Omega^x, \Sigma_z]$  if  $k \geq xz$ . Moreover, if  $2xz \leq n$ , the  $k$ -set-agreement problem cannot be solved in  $\mathcal{MP}_{n,n-1}[\Omega^x, \Sigma_z]$  if  $k < xz$ .*

### 5.1 Solving $k$ -set agreement with $\text{anti-}\Omega^x$ and $\Sigma_z$

For the system  $\mathcal{MP}_{n,n-1}[\text{anti-}\Omega^x, \Sigma_z]$ , we describe a  $k$ -set agreement algorithm that requires  $k \geq xz$ . From a computability point of view, our algorithm is optimal if  $n$  is large enough: we later establish that if  $k < xz$  and  $n \geq 2xz$  there is no  $k$ -set agreement algorithm in  $\mathcal{MP}_{n,n-1}[\text{anti-}\Omega^x, \Sigma_z]$  (Corollary 1).

*Using  $\Omega$  and  $\Sigma_z$  to solve  $k$ -set agreement for  $k \geq z$*  The algorithm is a simple adaptation of the generic  $\Omega$ -based consensus algorithm presented in [21], in which an  $\text{Alpha}_k$  object is used in place of an  $\text{Alpha}$  object. For completeness, the algorithm is described in Figure 3. The fact that any decided value has been returned by an invocation of  $\text{Alpha}_k.\text{propose}(\cdot)$  guarantees validity and agreement. Because eventually a unique correct process considers itself the leader, there is a time after which only this process invokes  $\text{Alpha}_k.\text{propose}(\cdot)$ . Hence, by the conditional convergence property of the object, there is an invocation that returns a non- $\perp$  value. This value is then broadcast, allowing every non-faulty process to decide, therefore ensuring termination.

```

SA_propose( $v$ )
(1)  $dec_i \leftarrow \perp$ ;  $r_i \leftarrow i$ ;
(2) while ( $dec_i = \perp$ ) do
(3)   if  $\Omega\text{-QUERY}() = i$  then  $dec_i \leftarrow \text{Alpha}_k.\text{propose}(r_i, v)$ 
(4)    $r_i \leftarrow r_i + n$  end if end do
(5) for each  $p_j \in \Pi$  do send  $\text{DECIDE}(dec_i)$  to  $p_j$  end do

when  $\text{DECIDE}(w)$  is received do
(6) for each  $p_j \in \Pi$  do send  $\text{DEC}(w)$  to  $p_j$  end do
(7) decide  $w$ ; return

```

**Fig. 3.**  $k$ -set agreement algorithm in  $\mathcal{MP}_{n,n-1}[\Omega, \Sigma_k]$ , code for  $p_i$

Using anti- $\Omega^x$  and  $\Sigma_z$  to solve  $k$ -set agreement for  $k \geq xz$  Our algorithm is based on a failure detector vector- $\Omega^x$  [31]. A failure detector of the class vector- $\Omega^x$  is a vector of  $x$  sub-detectors,  $\Omega_1, \dots, \Omega_x$ , such that at least one  $\Omega_i$  is a failure detector of the class  $\Omega$ . When  $k = n - 1$ , the vector- $\Omega$  failure detector proposed in [31] is obtained. It was shown there how vector- $\Omega$  can be implemented from anti- $\Omega^{n-1}$  in the wait-free asynchronous shared memory model, and how it can be used to solve  $(n - 1)$ -set agreement. The failure detector vector- $\Omega^x$  was also presented in [31]. It is claimed there that the algorithm to transform anti- $\Omega^{n-1}$  into vector- $\Omega^{n-1}$  (Figure 1 in [31], see also [3]), can be generalized to transform anti- $\Omega^x$  into vector- $\Omega^x$ . A close look at the transformation algorithm reveals that it can be easily adapted to the message passing case if a reliable broadcast primitive is available. As reliable broadcast can be implemented in an asynchronous message passing system in which any number of processes may fail [22], vector- $\Omega^x$  can be implemented in  $\mathcal{MP}_{n,n-1}[\text{anti-}\Omega^x]$ .

To solve  $k$ -set agreement in  $\mathcal{MP}_{n,n-1}[\text{anti-}\Omega^x, \Sigma_z]$ , processes simulate outputs of a failure detector vector- $\Omega^x$ . We associate to each sub-detector  $\Omega_i, 1 \leq i \leq x$  an instance of the  $(\Omega, \Sigma_z)$ -based  $z$ -set agreement algorithm described in Figure 3. Each processes participates simultaneously in each of the  $x$  instances, and terminates as soon as it decides in one instance.

It follows from the fact that at least one sub detector  $\Omega_i$  is a failure detector of the class  $\Omega$  that at least one instance terminates. Moreover, since at most  $z$  values are decided in each instance, the total number of decided value is upper bounded by  $xz$ . Therefore,

**Lemma 11.** *Let  $1 \leq k, x, z \leq n$ . There is a  $k$ -set agreement algorithm in  $\mathcal{MP}_{n,n-1}[\text{anti-}\Omega^x, \Sigma_z]$  if  $k \geq xz$ .*

## 5.2 An impossibility result

This section investigates  $k$ -set-agreement solvability when the system is enriched with failure detectors of both classes  $\Omega^y$  and  $\Sigma_z$ . The main result is Lemma 13 which establishes that there is no  $k$ -set agreement algorithm in the wait-free environment ( $t = n - 1$ ) where failure detectors  $\Omega^y$  and  $\Sigma_z$  are provided if  $k < yz$ .

**Lemma 12.** *Let  $k, 1 \leq k \leq n$  and  $x, 1 \leq 2x \leq n$ . If  $k < x$ , there is no  $k$ -set agreement algorithm in  $\mathcal{MP}_{n,n-1}[\Omega^x, \Sigma]$ .*

**Lemma 13.** *Let  $k, 1 \leq k \leq n$  and  $x, z, 1 \leq 2xz \leq n$ . If  $k < xz$ , there is no  $k$ -set agreement algorithm in  $\mathcal{MP}_{n,n-1}[\Omega^x, \Sigma_z]$ .*

Given a failure detector  $\Omega^x$ , it is easy to simulate a anti- $\Omega^x$  failure detector by outputting the complement of the sets leader output by  $\Omega^x$ . Therefore,

**Corollary 1.** *Let  $k, 1 \leq k \leq n$  and  $x, z, 1 \leq 2xz \leq n$ . If  $k < xz$ , there is no  $k$ -set agreement algorithm in  $\mathcal{MP}_{n,n-1}[\text{anti-}\Omega^x, \Sigma_z]$ .*

Bonnet and Raynal introduce in [4] the failure detector class  $\Pi_k$  as a weakest failure detector candidate for message passing  $k$ -set-agreement. Next corollary disproves this conjuncture.

**Corollary 2.** *Let  $k, n : 1 < k < n - 1$  and  $2k^2 \leq n$ . There is no  $k$ -set agreement algorithm in  $\mathcal{MP}_{n,n-1}[\Pi_k]$ .*

*Proof.* [4] proves that  $\Pi_k$  is equivalent to  $\Sigma_k \times \Omega^k$ . The corollary then directly follows from Lemma 13 □

## 6 Concluding remarks

The paper has investigated the computational power of the failure detector classes  $\Sigma_x$  and anti- $\Omega^z$  as far as  $k$ -set-agreement is concerned in the  $n$ -processes message passing asynchronous model. The main result is that for large enough values of  $n$ , namely  $n > 2kz$ ,  $k$ -set agreement is possible if and only if  $k \geq xz$ .

The main open question is the weakest failure detector for message passing  $k$ -set-agreement, for  $1 < k < n - 1$ . Our  $xz$ -set agreement algorithm may help to demonstrate the sufficiency of weakest failure detector candidate. Another interesting avenue for future research is the complexity of  $k$ -set-agreement tolerating  $t > n/2$  failures. When a majority of processes does not fail, it has been shown that the price of indulgence is constant [1, 15]. Is it still true when a majority of processes failures has to be tolerated?

## References

1. D. Alistarh, S. Gilbert, R. Guerraoui, and C. Travers. Of choices, failures and asynchrony: The many faces of set agreement. In *ISAAC 2009*, LNCS # 5878, pp. 943–953.
2. A. F. Anta, S. Rajsbaum, and C. Travers. Brief announcement: weakest failure detectors via an egg-laying simulation. In *PODC 2009*, ACM Press, pp. 290–291.
3. A. F. Anta, S. Rajsbaum, and C. Travers. Weakest failure detectors via an egg-laying simulation (preliminary version). Technical report, Universidad Rey Juan Carlos. Reports on Systems and Communications, vol IX, no 2, Jan 2009. <http://gsyc.es/tr-docs/RoSAC-2009-2.pdf>.
4. F. Bonnet and M. Raynal. Looking for the weakest failure detector for  $k$ -set agreement in message-passing systems: Is  $\pi_k$  the end of the road? In *SSS 2009*, LNCS # 5873, pp. 149–164.
5. F. Bonnet and M. Raynal. A simple proof of the necessity of the failure detector sigma to implement an atomic register in asynchronous message-passing systems. *Information Processing Letters*, 110(4):153–157, 2010.
6. E. Borowsky and E. Gafni. Generalized FLP impossibility result for  $t$ -resilient asynchronous computations. In *STOC 1993*, ACM Press, pp. 91–100.
7. T. D. Chandra, V. Hadzilacos, and S. Toueg. The weakest failure detector for solving consensus. *J. ACM*, 43(4):685–722, 1996.
8. T. D. Chandra and S. Toueg. Unreliable failure detectors for reliable distributed systems. *J. ACM*, 43(2):225–267, 1996.

9. S. Chaudhuri. More choices allow more faults: Set consensus problems in totally asynchronous systems. *Information and Computation*, 105(1):132–158, 1993.
10. W. Chen, J. Zhang, Y. Chen, and X. Liu. Weakening failure detectors for  $k$ -set agreement via the partition approach. In *DISC 2007*, LNCS # 4731, pp. 123–138.
11. C. Delporte-Gallet, H. Fauconnier, and R. Guerraoui. Sharing is harder than agreeing. In *PODC 2008*, ACM press, pp. 85–94.
12. C. Delporte-Gallet, H. Fauconnier, R. Guerraoui, V. Hadzilacos, P. Kouznetsov, and S. Toueg. The weakest failure detectors to solve certain fundamental problems in distributed computing. In *PODC 2004*, ACM press, pp. 338–346.
13. C. Delporte-Gallet, H. Fauconnier, R. Guerraoui, and A. Tielmann. The weakest failure detector for message passing set-agreement. In *DISC 2008*, LNCS # 5218, pp. 109–120.
14. C. Delporte-Gallet, H. Fauconnier, R. Guerraoui, and A. Tielmann. The disagreement power of an adversary: extended abstract. In *PODC 2009*, ACM Press, pp. 288–289.
15. P. Dutta and R. Guerraoui. The inherent price of indulgence. *Distributed Computing*, 18(1):85–98, 2005.
16. M. J. Fischer, N. A. Lynch, and M. Paterson. Impossibility of distributed consensus with one faulty process. *J. ACM*, 32(2):374–382, 1985.
17. E. Gafni and P. Kuznetsov. The weakest failure detector for solving  $k$ -set agreement. In *PODC 2009*, ACM Press, pp. 83–91.
18. R. Guerraoui. Indulgent algorithms (preliminary version). In *PODC 2000*, ACM Press, pp. 289–297.
19. R. Guerraoui, M. Herlihy, P. Kuznetsov, N. A. Lynch, and C. C. Newport. On the weakest failure detector ever. *Distributed Computing*, 21(5):353–366, 2009.
20. R. Guerraoui and N. A. Lynch. A general characterization of indulgence. *Transactions on Autonomous and Adaptive Systems*, 3(4), 2008.
21. R. Guerraoui and M. Raynal. The alpha of indulgent consensus. *The Computer Journal*, 50(1):53–67, 2007.
22. V. Hadzilacos and S. Toueg. Reliable broadcast and related problems. *Distributed Systems*, pp. 97–145, 1993.
23. M. Herlihy and L. D. Penso. Tight bounds for  $k$ -set agreement with limited-scope failure detectors. *Distributed Computing*, 18(2):157–166, 2005.
24. M. Herlihy and N. Shavit. The topological structure of asynchronous computability. *J. ACM*, 46(6):858–923, 1999.
25. A. Mostéfaoui, S. Rajsbaum, M. Raynal, and C. Travers. On the computability power and the robustness of set agreement-oriented failure detector classes. *Distributed Computing*, 21(3):201–222, 2008.
26. A. Mostéfaoui and M. Raynal.  $k$ -set agreement with limited accuracy failure detectors. In *PODC 2000*, ACM Press, pp. 143–152.
27. G. Neiger. Failure detectors and the wait-free hierarchy. In *PODC 1995*, ACM Press, pp. 100–109.
28. M. Raynal and C. Travers. In search of the holy grail: Looking for the weakest failure detector for wait-free set agreement. In *OPDIS 2006*, LNCS # 4305, pp. 3–19.
29. M. E. Saks and F. Zaharoglou. Wait-free  $k$ -set agreement is impossible: The topology of public knowledge. *SIAM J. Comput.*, 29(5):1449–1483, 2000.
30. J. Yang, G. Neiger, and E. Gafni. Structured derivations of consensus algorithms for failure detectors. In *PODC 1998*, ACM Press, pp. 297–306.
31. P. Zieliński. Anti-omega: the weakest failure detector for set agreement. In *PODC 2008*, ACM Press, pp. 55–64.