Approximate Byzantine Fault-Tolerance in Distributed Optimization

Shuo Liu[⋆] Nirupam Gupta[†] Nitin H. Vaidya[†]

{[⋆]sl1539, [†]*firstname.lastname*}@georgetown.edu
Department of Computer Science
Georgetown University
Washington DC, USA

Abstract

This paper studies the problem of Byzantine fault-tolerance in distributed multi-agent optimization. In this problem, each agent has a local cost function, and in the fault-free case, the objective is to design a distributed algorithm that allows all the agents to find a minimum point of all the agents' aggregate cost function. We consider a scenario where certain number of agents might be Byzantine faulty, i.e., these agents may not follow a prescribed algorithm and may share arbitrary information regarding their local cost functions. In the presence of such faulty agents, it is generally impossible to find a minimum point of all the agents' aggregate cost function. A more reasonable goal, however, is to design an algorithm that allows all the non-faulty agents to compute, either *exactly* or *approximately*, the minimum point of only the non-faulty agents' aggregate cost function.

From prior work we know that in the presence of up to f (out of n) Byzantine faulty agents, a deterministic algorithm can compute a minimum point of the non-faulty agents' aggregate cost *exactly* if and only if the non-faulty agents' cost functions satisfy a certain redundancy property named $2f$*-redundancy* [20]. However, the $2f$-redundancy property can only be guaranteed in ideal systems free from noises (or uncertainties), and therefore, the objective of exact fault-tolerance is unsuitable for many practical settings that inevitably suffer from noises. In this paper, we consider the problem of *approximate* fault-tolerance - a generalization of exact fault-tolerance where the goal is to only compute an approximation of a minimum point of the non-faulty agents' aggregate cost function. Upon defining approximate fault-tolerance later as (f, ϵ)-resilience where ϵ is the approximation error, we show that it can be achieved under a weaker redundancy condition than $2f$-redundancy. We present necessary and sufficient conditions for achieving (f, ϵ)-resilience in a synchronous distributed system with server-based architecture. Then, we consider a special case when the agents' cost functions are differentiable. Here, we analyse the approximate fault-tolerance of the distributed gradient-descent method, which is a prominent distributed optimization algorithm in this particular case, when equipped with a *gradient-filter* or *robust gradient aggregation*; such as *comparative gradient elimination* (CGE) or *coordinate-wise trimmed mean* (CWTM).

1 Introduction

The problem of distributed optimization in multi-agent systems has gained significant attention in recent years [7, 25, 16]. In this problem, each agent has a *local cost function* and, when the agents are fault-free, the goal is to design algorithms that allow the agents to collectively minimize the aggregate of their cost functions. To be precise, suppose that there are n agents in the system and let $Q_i(x)$ denote the local cost function of agent i, where x is a d-dimensional vector of real values, i.e., $x \in \mathbb{R}^d$. A traditional distributed optimization algorithm outputs a *global minimum* x^* such that

$$x^* \in \arg\min_{x \in \mathbb{R}^d} \sum_{i=1}^{n} Q_i(x). \tag{1}$$

As a simple example, $Q_i(x)$ may denote the cost for an agent i (which may be a robot or a person) to travel to location x from their current location, and x^* is a location that minimizes the total cost of meeting for all the agents. Such multi-agent optimization is of interest in many practical applications, including distributed machine learning [7], swarm robotics [29], and distributed sensing [28].

We consider the distributed optimization problem in the presence of up to f Byzantine faulty agents, originally introduced by Su and Vaidya [32]. The Byzantine faulty agents may behave arbitrarily [22]. In particular, the non-faulty agents may share arbitrary incorrect and inconsistent information in order to bias the output of a distributed optimization algorithm. For example, consider an application of multi-agent optimization in the case of distributed sensing where the agents (or *sensors*) observe a common *object* in order to collectively identify the object. However, the faulty agents may send arbitrary observations concocted to prevent the non-faulty agents from making the correct identification [11, 13, 26, 33]. Similarly, in the case of distributed learning, which is another application of distributed optimization, the faulty agents may send incorrect information based on *mislabelled* or arbitrary concocted data points to prevent the non-faulty agents from learning a *good* classifier [1, 2, 4, 9, 10, 12, 19, 35].

1.1 Background: Exact Fault-Tolerance

In the *exact fault-tolerance* problem, the goal is to design a distributed algorithm that allows all the non-faulty agents to compute a minimum point of the aggregate cost of only the non-faulty agents [20]. Specifically, suppose that in a given execution, set $\mathcal{B}$ with $|\mathcal{B}| \leq f$ is the set of Byzantine agents, where notation $|\cdot|$ denotes the set cardinality, and $\mathcal{H} = \{1, \ldots, n\} \setminus \mathcal{B}$ denotes the set of non-faulty (i.e., honest) agents. Then, a distributed optimization algorithm has exact fault-tolerance if it outputs a point $x^*_{\mathcal{H}}$ such that

$$x^*_{\mathcal{H}} \in \arg\min_{x \in \mathbb{R}^d} \sum_{i \in \mathcal{H}} Q_i(x). \tag{2}$$

However, since the identity of the Byzantine agents is a priori unknown, in general, exact fault-tolerance is unachievable [32]. Specifically, it is shown in [20, 21] that exact fault-tolerance can be achieved *if and only if* the agents' cost functions satisfy the $2f$-*redundancy* property defined below. Let $|\cdot|$ denote set cardinality.

Definition 1 ($2f$**-redundancy).** *The agents' cost functions are said to have* $2f$-redundancy *property if and only if for for every pair of subsets* S, $\widehat{S} \subseteq \{1, \ldots, n\}$ *with* $|S| = n - f$, $\left|\widehat{S}\right| \geq n - 2f$ *and* $\widehat{S} \subseteq S$,

$$\arg\min_{x \in \mathbb{R}^d} \sum_{i \in \widehat{S}} Q_i(x) = \arg\min_{x \in \mathbb{R}^d} \sum_{i \in S} Q_i(x).$$

In principle, the $2f$-redundancy property can be realized by design for many applications of multi-agent distributed optimization including distributed sensing and distributed learning (see [18, 20]). However, practical realization of $2f$-redundancy can be difficult due to the presence of *noise* in the real-world systems. This motivates us to consider a generalization of exact fault-tolerance, namely the problem of *approximate fault-tolerance*, which is described and defined as follows.

1.2 Approximate Fault-Tolerance

Unlike exact fault-tolerance, in approximate fault-tolerance it is acceptable for an algorithm to output an approximate minimum point of the non-faulty aggregate cost function. As stated below, we formally define approximate fault-tolerance by (f, ϵ)-*resilience* where $\epsilon \in \mathbb{R}_{\geq 0}$ is the measure of approximation. Recall that the Euclidean distance between a point x and a non-empty set X in the space $\mathbb{R}^d$, denoted by $\text{dist}(x, X)$, is defined to be

$$\text{dist}(x, X) = \inf_{y \in X} \|x - y\| \tag{3}$$

where $\|\cdot\|$ denotes the Euclidean norm.

Definition 2 ((f, ϵ)**-resilience).** *A distributed optimization algorithm is said to be* (f, ϵ)*-resilient if it outputs a point* $\widehat{x} \in \mathbb{R}^d$ *such that for every subset* S *of non-faulty agents with* $|S| = n - f$,

$$\text{dist}\left(\widehat{x}, \arg\min_{x \in \mathbb{R}^d} \sum_{i \in S} Q_i(x)\right) \leq \epsilon,$$

despite the presence of up to f *Byzantine agents.*

Alternately, a distributed optimization algorithm is (f, ϵ)-*resilient* if it outputs a point within ϵ distance from a minimum point of the aggregate cost function of at least $n - f$ non-faulty agents. As there can be at most f Byzantine faulty agents whose identity remains unknown the two scenarios where; (1) there are exactly f Byzantine agents, and (2) there are less than f Byzantine agents are indistinguishable. Thus, estimation for the minimum point of the aggregate cost functions of $n - f$ non-faulty agents is indeed a reasonable goal [32].

In this paper, we consider *deterministic* algorithms which, given a fixed set of inputs from the server and the agents, always output the same point in $\mathbb{R}^d$. Thus, a deterministic $(f,\, \epsilon)$-resilient algorithm produces a unique output point in all of its executions with identical inputs from the server and all the agents (including the faulty agents). Note that in the deterministic framework, exact fault-tolerance is equivalent to $(f,\, 0)$-resilience, i.e., a deterministic $(f,\, 0)$-resilient algorithm achieves exact fault-tolerance, and vice-versa. Proof of this claim is given in Appendix A.

As a key result, we show that $(f,\, \epsilon)$-*resilience* requires a *weaker redundancy* condition, in comparison to $2f$-redundancy, named $(2f,\, \epsilon)$-*redundancy* which is defined as follows. Recall that the Euclidean Hausdorff distance between two sets X and Y in $\mathbb{R}^d$, which we denote by $\mathrm{dist}\,(X,\, Y)$, is defined to be [24]

$$\mathrm{dist}\,(X,\, Y) \triangleq \max \left\{ \sup_{x \in X} \mathrm{dist}\,(x,\, Y)\,, \;\; \sup_{y \in Y} \mathrm{dist}\,(y,\, X) \right\}. \tag{4}$$

Definition 3 (**$(2f,\, \epsilon)$-redundancy**)**.** *The agents' cost functions are said to have* $(2f,\, \epsilon)$-redundancy *property if and only if for every pair of subsets* $S,\, \widehat{S} \subseteq \{1, \ldots,\, n\}$ *with* $|S| = n - f,\; \left|\widehat{S}\right| \geq n - 2f$ *and* $\widehat{S} \subseteq S$*,*

$$\mathrm{dist}\left(\arg\min_{x \in \mathbb{R}^d} \sum_{i \in S} Q_i(x),\, \arg\min_{x \in \mathbb{R}^d} \sum_{i \in \widehat{S}} Q_i(x) \right) \leq \epsilon. \tag{5}$$

Upon comparison between the Definitions 1 and 3, it is easy to see that $2f$-redundancy is equivalent to $(2f,\, 0)$-redundancy. It is also obvious that $2f$-redundancy implies $(2f,\, \epsilon)$-redundancy for all $\epsilon \geq 0$. However, the converse need not be true. Thus, the $(2f,\, \epsilon)$-redundancy property with $\epsilon > 0$ is weaker than $2f$-redundancy.

1.3 Summary of Our Contributions

In the **first part** of the paper, i.e., Section 3, we show two key results:

- $(f,\, \epsilon)$-resilience is feasible only if $(2f,\, \epsilon)$-redundancy property holds true.
- If $(2f,\, \epsilon)$-redundancy property holds true then $(f,\, 2\epsilon)$-resilience is achievable.

As $(f,\, 0)$-resilience is equivalent to exact fault-tolerance (see Section 1.2), and $(2f,\, 0)$-redundancy is equivalent to $2f$-redundancy, the above results generalize the known result that exact fault-tolerance is feasible if and only if the $2f$-redundancy condition holds true (see [20, 21]).

In the **second part**, i.e., Sections 4 and 5, we consider the case when agents' cost functions are assumed differentiable, such as in machine learning [6] and state estimation or regression [18, 31, 33]. Here, we study the problem of approximate fault-tolerance in the distributed gradient-descent (DGD) method - an iterative distributed optimization algorithm commonly used in this particular case.

- In Section 4, we present a generic convergence result of the DGD method when equipped with a *gradient-filter* (or *robust gradient aggregation*), which is a common fault-tolerance mechanism [4, 12, 20].

- Then, upon assuming the $(2f,\, \epsilon)$-redundancy property, we present approximate fault-tolerance guarantees of two specific gradient-filters -
  - the comparative gradient elimination (CGE) gradient-filter [17], and
  - the coordinate-wise trimmed mean (CWTM) gradient-filter [36].

  The above two filters are computationally cheapest of all the existing gradient-filters, and have wide applicability, for eg., see [17, 18, 31, 36].
- In Section 5, we present empirical comparisons between the approximate fault-tolerance achieved by the two aforementioned gradient-filters.

System architecture: The results in this paper apply to two different system architectures illustrated in Figure 1. The system is assumed *synchronous.* In the server-based architecture, the server is assumed to be trustworthy, but up to f agents may be Byzantine faulty. The trusted server helps solve the distributed optimization problem in coordination with the agents. In the peer-to-peer architecture, the agents are connected to each other by a complete network, and up to f of these agents may be Byzantine faulty. Provided that $f < \frac{n}{3}$, an algorithm for the server-based architecture can be simulated in the peer-to-peer system using the well-known *Byzantine broadcast* primitive [23]. For simplicity of presentation, the rest of this paper considers the server-based architecture.

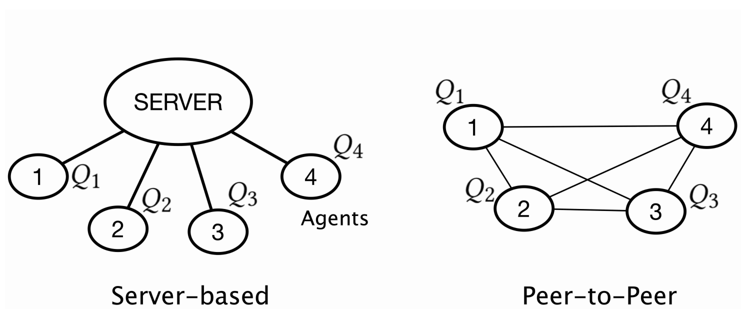


Figure 1: System architecture.

Before we present our results in detail, we review below other related work on approximate fault-tolerance and gradient-filters.

2 Other Related Work

In the past, different definitions of approximate fault-tolerance, besides $(f,\, \epsilon)$-resilience that we use, have been formulated to analyze Byzantine fault-tolerance of different distribute optimization algorithms [15, 32]. As we discuss below in Section 2.1, the difference between these other definitions and our definition of $(f,\, \epsilon)$-resilience arises mainly due to the applicability of the distributed optimization problem. Later, in Section 2.2, we discuss prior work on gradient-filters for conferring Byzantine fault-tolerance to the distributed gradient-descent method - an iterative distributed optimization algorithm used when the agents' cost functions are differentiable [6].

2.1 Alternate Definitions of Approximate Fault-Tolerance

As proposed by Su and Vaidya, 2016 [32], an alternate way of measuring approximation (or accuracy) in fault-tolerance is by use of scalar coefficients. Specifically, instead of a minimum point of the *uniformly weighted* aggregate of non-faulty agents' cost functions, a distributed optimization algorithm may output a minimum point of a *non-uniformly weighted* aggregate of non-faulty costs, i.e., $\sum_{i \in \mathcal{H}} \alpha_i \, Q_i(x)$, where $\mathcal{H}$ denotes the set of at least $n - f$ non-faulty agents, and $\alpha_i \geq 0$ for all $i \in \mathcal{H}$. As is suggested in [32], upon re-scaling the coefficients such that $\sum_{i \in \mathcal{H}} \alpha_i = 1$, we can measure approximation in fault-tolerance using two metrics; 1) the number of coefficients in $\{\alpha_i, \, i \in \mathcal{H}\}$ that are positive, and 2) the minimum positive value amongst the coefficients: $\min \{\alpha_i; \, \alpha_i > 0, \, i \in \mathcal{H}\}$. Results on the achievability of such approximation in the case of scalar optimization problems, i.e., when $d = 1$, can be found in [32, 34]. However, we are not aware of any such results for the case of higher-dimensional optimization problem, i.e., when $d > 1$.

Another alternate way of measuring approximation is by the value of the non-faulty aggregate cost function, or its gradient. For instance, as discussed in [15], for the case of differentiable cost functions a resilient distributed optimization algorithm may aim to output a point $x_\Pi \in \mathbb{R}^d$, such that each element of the aggregate non-faulty gradient $\sum_{i \in \mathcal{H}} \nabla Q_i(x_\Pi)$ has an absolute value upper bounded by ϵ. As yet another alternative, a resilient algorithm Π may aim to output a point x_Π such that the non-faulty aggregate cost $\sum_{i \in \mathcal{H}} Q_i(x_\Pi)$ is within ϵ of the true minimum cost $\min_x \sum_{i \in \mathcal{H}} Q_i(x)$. However, these definitions of approximate resilience are sensitive to scaling of the cost functions. In particular, if the elements of $\sum_{i \in \mathcal{H}} \nabla Q_i(x_\Pi)$ are bounded by ϵ then the elements of $\sum_{i \in \mathcal{H}} \alpha \nabla Q_i(x_\Pi)$ are bounded by $\alpha \epsilon$, where α is a positive scalar value. On the other hand, both $\sum_{i \in \mathcal{H}} Q_i(x)$ and $\sum_{i \in \mathcal{H}} \alpha Q_i(x_\Pi)$ have identical minimum point regardless of the value of α. Therefore, when the objective is to approximate a minimum point of the non-faulty aggregate cost $\arg\min_x \sum_{i \in \mathcal{H}} Q_i(x)$, which is the indeed the case in this paper, use of function (or gradient) values to measure approximation is not a suitable choice.

2.2 Gradient-Filters

In the past, several gradient-filters have been proposed to *robustify* the distributed gradient-descent (DGD) method against Byzantine faulty agents, for example, see [1, 4, 14, 15, 17, 27, 32, 36] and references therein. The DGD method is an iterative distributed optimization algorithm wherein the server maintain an estimate of the solution (a valid minimum point), and updates it iteratively using gradients computed by the agents of their respective cost functions. In the traditional DGD method, the server updates its estimates using the average (or sum) of agents' gradients. However, gradient averaging is rendered ineffective when Byzantine faulty agents send arbitrary incorrect gradients [32].

A gradient-filter refers to the technique of *robust aggregation of agents' gradients*, which mitigates the detrimental impact of incorrect gradients. To name a few gradient-filters, that are provably effective against Byzantine faulty agents, we have comparative gradient elimination (CGE) [18, 17], coordinate-wise trimmed mean

(CWTM) [32, 36], geometric median-of-means (GMoM) [12], KRUM [4], and the spectral gradient-filters [15]. However, different gradient-filters guarantee Byzantine fault-tolerance under different assumptions on non-faulty agents' cost functions.

In this paper, we first propose a generic result (in Theorem 3) to establish convergence of the DGD method when coupled with a gradient-filter. The result holds true regardless of the gradient-filter used. Then, we use this convergence result to derive the approximate fault-tolerance guarantees (as per Definition 2) of two specific gradient-filters; CGE and CWTM, under the $(2f,\, \epsilon)$-redundancy property. The reason for choosing these two filters is that they are the computationally cheapest of all, with $\mathcal{O}(n(\log n + d))$ per iteration time complexity [17, 31].

It is worth noting that both CGE and CWTM gradient-filters can guarantee exact fault-tolerance under certain assumptions, besides $2f$-redundancy, as show [18, 17, 31]. However, unlike CGE, the fault-tolerance property of CWTM is only known for the special distributed optimization problems of distributed linear regression [31], and distributed machine learning [36]. In Section 4.2, we present approximate fault-tolerance properties of CGE and CWTM gradient-filters that are applicable to the more general distributed optimization problem, and also encapsulates the special case of exact fault-tolerance.

3 Necessary and Sufficient Conditions for $(f,\, \epsilon)$-Resilience

In this section, we present formal details on necessary and sufficient conditions for $(f,\, \epsilon)$-resilience. Throughout this paper we assume, as stated below, that the non-faulty agents' cost functions and their aggregates have well-defined minimum points. Otherwise, the problem of optimization is rendered vacuous.

Assumption 1. *We assume that for any non-empty set S of non-faulty agents, the set $\arg\min_{x \in \mathbb{R}^d} \sum_{i \in S} Q_i(x)$ is closed and non-empty.*

Moreover, we also assume that $f < n/2$. Lemma 1, stated below, shows that $(f,\, \epsilon)$-resilience is impossible in general when $f \geq n/2$.

Lemma 1. *If $f \geq n/2$ then there cannot exist a deterministic $(f,\, \epsilon)$-resilient algorithm for any $\epsilon \geq 0$.*

Proof. We prove the lemma by contradiction. We consider a case when $n = 2$, $d = 1$, i.e., $Q_i : \mathbb{R} \to \mathbb{R}$ for all i, and all the cost functions have unique minimum points. Suppose that $f = 1$, and that there exists a deterministic $(f,\, \epsilon)$-resilient algorithm Π for some $\epsilon \geq 0$. Without loss of generality, we suppose that agent 2 is Byzantine faulty. We denote $x_1 = \arg\min_{x \in \mathbb{R}} Q_1(x)$.

The Byzantine agent 2 can choose to behave as a non-faulty agent with cost function $\widetilde{Q}_2(x) = Q_1(x - x_1 - 2\epsilon - \delta)$ where δ is some positive real value. Now, note that the minimum point of $\widetilde{Q}_2(x)$, which we denoted by x_2, is unique and equal to $x_1 + 2\epsilon + \delta$. Therefore, $|x_1 - x_2| = 2\epsilon + \delta > 2\epsilon$. As the identity of Byzantine agent is a priori unknown, the server cannot distinguish between scenarios; (i) agent 1 is non-faulty, and (ii) agent 2 is non-faulty. Now, being deterministic algorithm,

Π should produce the same output in both the scenarios. In scenario (i), as Π is assumed $(f,\, \epsilon)$-resilient, its output must lie in the interval $[x_1 - \epsilon,\, x_1 + \epsilon]$. Similarly, in scenario (ii), the output of Π must lie in the interval $[x_2 - \epsilon,\, x_2 + \epsilon]$. However, as $|x_1 - x_2| > 2\epsilon$, the two intervals $[x_1 - \epsilon,\, x_1 + \epsilon]$ and $[x_2 - \epsilon,\, x_2 + \epsilon]$ do not overlap. Therefore, Π cannot be $(f,\, \epsilon)$-resilient in both the scenarios simultaneously, which is a contradiction to the assumption that Π is $(f,\, \epsilon)$-resilient.

The above argument extends easily for the case when $n > 2$, and $f > n/2$. □

We show below the **necessity of** $(2f,\, \epsilon)$**-redundancy for** $(f,\, \epsilon)$**-resilience**.

Theorem 1 (Necessity)**.** *Suppose that Assumption 1 holds true. There exists a deterministic* $(f,\, \epsilon)$*-resilient distributed optimization algorithm where* $\epsilon \geq 0$ only if *the agents' cost functions satisfy the* $(2f, \epsilon)$*-redundancy property.*

Proof. To prove the theorem we present a scenario when the agents' cost functions (if non-faulty) are scalar functions, i.e., $d = 1$ and for all i, $Q_i : \mathbb{R} \to \mathbb{R}$, and the minimum point of an aggregate of one or more agents' cost functions is uniquely defined. Obviously, if a condition is necessary in the aforementioned case then it is so in the more general case of vector functions with non-unique minimum points.

To prove the necessity condition, we also assume that the server has full knowledge of all the agents' cost functions. This may not hold true in practice, where instead the server may only have partial information about the agents' cost functions. Indeed, this forces the Byzantine faulty agents to a priori fix their cost functions. However, in reality the Byzantine agents may send arbitrary information over time to the server that need not be consistent with a fixed cost function. Thus, necessity of $(2f, \epsilon)$-redundancy under this assumption implies its necessity in general.

The proof is by contradiction. Specifically, we show the following:

> *If the non-faulty cost functions do not satisfy the* $(2f, \epsilon)$*-redundancy property then there cannot exist a* deterministic (f, ϵ)*-resilient distributed optimization algorithm.*

Recall that we have assumed that for a non-empty set of agents T the aggregate cost function $\sum_{i \in T} Q_i(x)$ has a unique minimum point. To be precise, for each non-empty subset of agents T, we define

$$x_T = \arg\min_x \sum_{i \in T} Q_i(x).$$

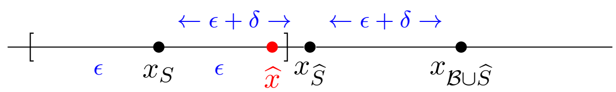
Suppose that the agents' cost functions **do not** satisfy the $(2f, \epsilon)$-redundancy property stated in Definition 3. Then, there exists a real number $\delta > 0$ and a pair of subsets S, $\widehat{S}$ with $\widehat{S} \subset S$, $|S| = n - f$, and $n - 2f \leq \left|\widehat{S}\right| < n - f$ such that

$$\left\| x_{\widehat{S}} - x_S \right\| \geq \epsilon + \delta. \tag{6}$$

Now, suppose that $n - f - \left|\widehat{S}\right|$ agents in the remainder set $\{1, \ldots,\, n\} \backslash S$ are Byzantine faulty. Let us denote the set of faulty agents by $\mathcal{B}$. Note that $\mathcal{B}$ is non-empty with

$|\mathcal{B}| = n - f - \left|\widehat{S}\right| \leq f$. Similar to the non-faulty agents, the faulty agents send to the server cost functions that are scalar, and the aggregate of one of more agents' cost functions in the set $S \cup \mathcal{B}$ is unique. However,

> the aggregate cost function of the agents in the set $\mathcal{B} \cup \widehat{S}$ minimizes at a unique point $x_{\mathcal{B}\cup\widehat{S}}$ which is $\left\|x_{\widehat{S}} - x_S\right\|$ distance away from $x_{\widehat{S}}$, similar to x_S, but lies on the other side of $x_{\widehat{S}}$ as shown in the figure below. Note that it is always possible to pick such functions for the faulty agents.



Note that the distance between the two points x_S and $x_{\mathcal{B}\cup\widehat{S}}$ is $2\epsilon + 2\delta$, i.e.,

$$\left\|x_S - x_{\mathcal{B}\cup\widehat{S}}\right\| = 2\epsilon + 2\delta. \tag{7}$$

We now show below, by contradiction, that there cannot exist a deterministic $(f,\, \epsilon)$-resilient distributed optimization algorithm.

Suppose, toward a contradiction, that there exists an $(f,\, \epsilon)$-resilient deterministic optimization algorithm named Π. As the identity of Byzantine faulty agents is a priori unknown to the server, and the cost functions sent by the Byzantine faulty agents have similar properties as the non-faulty agents, the server cannot distinguish between the following two possible scenarios; i) S is the set of non-faulty agents, and ii) $\mathcal{B} \cup \widehat{S}$ is the set of non-faulty agents. Note that both the sets S and $\mathcal{B} \cup \widehat{S}$ contain $n - f$ agents.

As the cost functions received by the server are identical in both of the above scenarios, being a deterministic algorithm, Π should have identical output in both the cases. We let $\widehat{x}$ denote the output of Π. In scenario (i) when the set of honest agents is given by S with $|S| = n - f$, as Π is assumed (f, ϵ)-resilient, by Definition 2 the output

$$\widehat{x} \in [x_S - \epsilon,\, x_S + \epsilon] \tag{8}$$

as shown in the figure above. By the same logic, in scenario (ii) when the set of honest agents is $\mathcal{B} \cup \widehat{S}$ with $\left|\mathcal{B} \cup \widehat{S}\right| = n - f$ the output

$$\widehat{x} \in [x_{\mathcal{B}\cup\widehat{S}} - \epsilon,\, x_{\mathcal{B}\cup\widehat{S}} + \epsilon]. \tag{9}$$

However, (7) implies that $\widehat{x}$ cannot satisfy both (8) and (9) simultaneously. Therefore, if algorithm Π is (f, ϵ)-resilient in scenario (i) then it cannot be so in scenario (ii), and vice-versa. This is a contradiction to the assumption that Π is $(f,\, \epsilon)$-resilient, and therefore proves the impossibility of $(f,\, \epsilon)$-resilience when the $(2f,\, \epsilon)$-redundancy property is violated.

$\square$

Next, we show that $(2f,\, \epsilon)$**-redundancy suffices for** $(f,\, 2\epsilon)$**-resilience**.

Theorem 2 (Sufficiency)**.** *Suppose that Assumption 1 holds true. For a real value $\epsilon \geq 0$, if the agents' cost functions satisfy the $(2f, \epsilon)$-redundancy property then $(f, 2\epsilon)$-resilience is achievable.*

Proof. The proof is constructive where we assume that all the agents send their individual cost functions to the server. We assume that $f > 0$ to avoid the trivial case of $f = 0$. Throughout the proof we write the notation $\arg\min_{x \in \mathbb{R}^d}$ simply as $\arg\min$, unless otherwise stated. Consider the algorithm presented below comprising three steps.

Step 1: Each agent sends their cost function to the server. An honest agent sends its actual cost function, while a faulty agent may send an arbitrary function.

Step 2: For each set T of received functions, $|T| = n - f$, the server computes a point

$$x_T \in \arg\min \sum_{i \in T} Q_i(x).$$

For each subset $\widehat{T} \subset T$, $\left|\widehat{T}\right| = n - 2f$, the server computes

$$r_{T\widehat{T}} \triangleq \text{dist}\left(x_T, \, \arg\min \sum_{i \in \widehat{T}} Q_i(x)\right), \tag{10}$$

and

$$r_T = \max_{\substack{\widehat{T} \subset T, \\ \left|\widehat{T}\right| = n-2f}} r_{T\widehat{T}}. \tag{11}$$

Step 3: The server outputs x_S such that

$$S = \arg\min_{\substack{T \subset \{1, \ldots, n\}, \\ |T| = n-f}} r_T. \tag{12}$$

We show below that above algorithm is $(f, \, 2\epsilon)$-resilient under $(2f, \, \epsilon)$-redundancy. For a non-empty set of agents T, we denote

$$X_T = \arg\min \sum_{i \in T} Q_i(x).$$

Consider an arbitrary set of non-faulty agents G with $|G| = n - f$. Such a set is guaranteed to exist as there are at most f faulty agents, and therefore, at least $n - f$ non-faulty agents in the system. Consider an arbitrary set $\widehat{T}$ such that $\widehat{T} \subset G$ and $\left|\widehat{T}\right| = n - 2f$. By Definition 3 of $(2f, \, \epsilon)$-redundancy,

$$\text{dist}\left(X_G, \, X_{\widehat{T}}\right) \leq \epsilon. \tag{13}$$

Recall from (10) that $r_{G\widehat{T}} = \text{dist}\left(x_G, \, X_{\widehat{T}}\right)$. As $x_G \in X_G$, by Definition (4) of Hausdorff set distance, $\text{dist}\left(x_G, \, X_{\widehat{T}}\right) \leq \text{dist}\left(X_G, \, X_{\widehat{T}}\right)$. Therefore,

$$r_{G\widehat{T}} \leq \text{dist}\left(X_G, \, X_{\widehat{T}}\right).$$

Substituting from (13) in the above inequality implies that

$$r_{G\widehat{T}} \leq \epsilon. \tag{14}$$

Now, recall from (11) that

$$r_G = \max_{\substack{\widehat{T} \subset G, \\ \left|\widehat{T}\right| = n-2f}} r_{G\widehat{T}}.$$

As $\widehat{T}$ in (14) is an arbitrary subset of G with $\left|\widehat{T}\right| = n - 2f$,

$$r_G = \max_{\substack{\widehat{T} \subset G, \\ \left|\widehat{T}\right| = n-2f}} r_{G\widehat{T}} \;\leq\; \epsilon. \tag{15}$$

From (12) and (15) we obtain that

$$r_S \leq r_G \leq \epsilon. \tag{16}$$

As $|G| = n - f$, for every set of agents T with $|T| = n - f$, $|T \cap G| \geq n - 2f$. Therefore, for the set S defined in (12), there exists a subset $\widehat{G}$ of G such that $\widehat{G} \subset S$ and $\left|\widehat{G}\right| = n - 2f$. For such a set $\widehat{G}$, by definition of r_S in (11), we obtain that

$$r_{S\widehat{G}} \triangleq \text{dist}\left(x_S, \, X_{\widehat{G}}\right) \leq r_S.$$

Substituting from (16) above, we obtain that

$$\text{dist}\left(x_S, \, X_{\widehat{G}}\right) \leq \epsilon. \tag{17}$$

As $\widehat{G}$ is a subset of G, all the agents in $\widehat{G}$ are non-faulty. Therefore, by Assumption 1, $X_{\widehat{G}}$ is a closed set. Recall that $\text{dist}\left(x_S, \, X_{\widehat{G}}\right) = \inf_{x \in X_{\widehat{G}}} \|x_S - x\|$. The closedness of $X_{\widehat{G}}$ implies that there exists a point $z \in X_{\widehat{G}}$ such that

$$\|x_S - z\| = \inf_{x \in X_{\widehat{G}}} \|x_S - x\| = \text{dist}\left(x_S, \, X_{\widehat{G}}\right).$$

The above, in conjunction with (17), implies that

$$\|x_S - z\| \leq \epsilon. \tag{18}$$

Moreover, as $z \in X_{\widehat{G}}$ where $\widehat{G} \subset G$ with $\left|\widehat{G}\right| = n - 2f$ and $|G| = n - f$, the $(2f, \, \epsilon)$-redundancy condition stated in Definition 3 implies that

$$\text{dist}\left(z, \, X_G\right) \leq \epsilon.$$

Similar to an argument made above, under Assumption 1, X_G is a closed set, and therefore, there exists $x^* \in X_G$ such that

$$\|z - x^*\| = \text{dist}\left(z, \, X_G\right) \leq \epsilon. \tag{19}$$

By triangle inequality, (18) and (19) implies that

$$\|x_S - x^*\| \leq \|x_S - z\| + \|z - x^*\| \leq 2\epsilon. \tag{20}$$

Finally, recall that set G in the above inequality is an arbitrary set of $n-f$ non-faulty agents. Therefore, (20) proves the theorem. $\square$

In the next part of the paper, i.e., Sections 4 and 5, we consider the case when the (non-faulty) agents' cost functions are assumed differentiable. Specifically, we present and study a generic fault-tolerance method of *gradient-filtering* for conferring approximate fault-tolerance to a commonly used distributed optimization algorithm - the distributed gradient-descent method.

4 Distributed Gradient-Descent Method

In this section, we consider a setting wherein the non-faulty agents' cost functions are differentiable. For this particular setting, we study the approximate fault-tolerance of the distributed gradient-descent method when coupled with a *gradient-filter* described below. We consider the server-based system architecture, shown in Fig. 1, assuming a synchronous system.

The distributed gradient-descent method is an iterative algorithm wherein the server maintains an *estimate* of a minimum point, and updates it iteratively using gradients sent by the agents. Specifically, in each iteration $t \in \{0, 1, \ldots\}$, the server starts with an estimate x^t and broadcasts to all the agents. Each non-faulty agent i sends back to the sever the gradient of its cost function at x^t, i.e., $\nabla Q_i(x^t)$. However, Byzantine faulty agents may send arbitrary incorrect vectors as their gradients to the server. The initial estimate, named x^0, is chosen arbitrarily by the server.

A *gradient-filter* is a vector function, denoted by $\mathsf{GradFilter}$, that maps the n gradients received by the server from all the n agents to a d-dimensional vector, i.e., $\mathsf{GradFilter} : \mathbb{R}^{d \times n} \to \mathbb{R}^d$. For example, an average of all the gradients as in the case of the traditional distributed gradient-descent method is technically a gradient-filter. However, averaging is not quite robust against Byzantine faulty agents [4, 32]. The real purpose of a gradient-filter is to mitigate the detrimental impact of incorrect gradients sent by the Byzantine faulty agents. In other words, a gradient-filter *robustifies* the traditional gradient-descent method against Byzantine faults. We show that if a gradient-filter satisfies a certain property then it can confer fault-tolerance to the distributed gradient-descent method.

We first formally describe below the steps in each iteration of the distributed gradient-descent method implemented on a synchronous server-based system. Note that we constrain the estimates computed by the server to a compact convex set $\mathcal{W} \subset \mathbb{R}^d$. The set $\mathcal{W}$ can be arbitrarily large. For a vector $x \in \mathbb{R}^d$, its projection onto $\mathcal{W}$, denoted by $[x]_{\mathcal{W}}$, is defined to be

$$[x]_{\mathcal{W}} = \arg\min_{y \in \mathcal{W}} \left\| x - y \right\|. \tag{21}$$

As $\mathcal{W}$ is compact and convex set, $[x]_{\mathcal{W}}$ is unique for each x (see [8]).

4.1 Steps in t-th iteration

In each iteration $t \in \{0, 1, \ldots\}$ the server updates its current estimate x^t to x^{t+1} using Steps S1 and S2 described as follows.

S1: The server requests from each agent the gradient of its local cost function at the current estimate x^t. Each non-faulty agent i will then send to the server the gradient $\nabla Q_i(x^t)$, whereas a faulty agent may send an incorrect arbitrary value for the gradient.

The gradient received by the server from agent i is denoted as g_i^t. If no gradient is received from some agent i, agent i must be faulty (because the system is assumed to be synchronous) – in this case, the server eliminates the agent i from the system, updates the values of n, f, and re-assigns the agents indices from 1 to n.

S2: **[Gradient-filtering]** The server applies a gradient-filter $\mathsf{GradFilter}$ to the n received gradients and computes $\mathsf{GradFilter}\left(g_1^t, \ldots, g_n^t\right) \in \mathbb{R}^d$. Then, the server updates its estimate to

$$x^{t+1} = \left[x^t - \eta_t \, \mathsf{GradFilter}\left(g_1^t, \ldots, g_n^t\right)\right]_{\mathcal{W}} \tag{22}$$

where η_t is the step-size of positive value for iteration t.

We present below, in Theorem 3, a generic convergence result for the above algorithm. The proof of Theorem 3 is deferred to Appendix C.

Theorem 3. *Consider the update law* (22) *in the above iterative algorithm, with diminishing step-sizes* $\{\eta_t, \, t = 0, \, 1, \ldots\}$ *satisfying* $\sum_{t=0}^{\infty} \eta_t = \infty$ *and* $\sum_{t=0}^{\infty} \eta_t^2 < \infty$. *Suppose that* $\left\|\mathsf{GradFilter}\left(g_1^t, \ldots, g_n^t\right)\right\| < \infty$ *for all* t. *For some point* $x^* \in \mathcal{W}$, *if there exists real-valued constants* $\mathsf{D}^* \in [0, \max_{x \in \mathcal{W}} \|x - x^*\|)$ *and* $\xi > 0$ *such that for each iteration* t,

$$\phi_t = \left\langle x^t - x^*, \, \mathsf{GradFilter}\left(g_1^t, \ldots, g_n^t\right)\right\rangle \geq \xi \text{ when } \left\|x^t - x^*\right\| \geq \mathsf{D}^*, \tag{23}$$

then $\lim_{t \to \infty} \left\|x^t - x^*\right\| \leq \mathsf{D}^*$.

Note that the values D^* and ξ in the statement of Theorem 3 need not be independent of each other. As shown in the subsequent section, the generic convergence result shown in Theorem 3 helps us obtain approximate fault-tolerance properties of different gradient-filters, under $(2f, \epsilon)$-redundancy and certain standard assumptions. We consider two particular gradient-filters, namely *Comparative Gradient Elimination* (CGE) and *Coordinate-Wise Trimmed Mean* (CWTM).

4.2 Gradient-Filters and their Fault-Tolerance Properties

In this subsection, we present precise approximate fault-tolerance guarantees of two specific gradient-filters; the Comparative Gradient Elimination (CGE) [17, 18], and the Coordinate-Wise Trimmed Mean (CWTM) [31, 36]. Note that differentiability of non-faulty agents' cost functions, which is already assumed for the gradient-descent method, implies Assumption 1 (see [8]). We additionally make Assumptions 2, 3 and 4 about the non-faulty agents' cost functions. Similar assumptions are made in prior work on fault-free distributed optimization [3, 7, 25].

Assumption 2 (Lipschitz smoothness)**.** *For each non-faulty agent* i, *we assume that the gradient of its cost function* $\nabla Q_i(x)$ *is Lipschitz continuous, i.e., there exists a finite real value* $\mu > 0$ *such that*

$$\left\|\nabla Q_i(x) - \nabla Q_i(x')\right\| \leq \mu \left\|x - x'\right\|, \quad \forall x, x' \in \mathcal{W}.$$

Assumption 3 (Strong convexity). *For a non-empty set of non-faulty agents $\mathcal{H}$, let $Q_{\mathcal{H}}(x)$ denote the average cost function of the agents in $\mathcal{H}$, i.e.,*

$$Q_{\mathcal{H}}(x) = \frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} Q_i(x).$$

For each such set $\mathcal{H}$ with $|\mathcal{H}| = n - f$, we assume that $Q_{\mathcal{H}}(x)$ is strongly convex, i.e., there exists a finite real value $\gamma > 0$ such that

$$\left\langle \nabla Q(x) - \nabla Q(x'),\, x - y \right\rangle \geq \gamma \left\| x - x' \right\|^2, \quad \forall x, x' \in \mathcal{W}.$$

Note that, under Assumptions 2 and 3, $\gamma \leq \mu$. This inequality is proved in Appendix B. Now, recall that the iterative estimates of the algorithm in Section 4.1 are constrained to a compact convex set $\mathcal{W} \subset \mathbb{R}^d$.

Assumption 4 (Existence). *For each set of non-faulty agents $\mathcal{H}$ with $|\mathcal{H}| = n - f$, we assume that there exists a point $x_{\mathcal{H}} \in \arg\min_{x \in \mathbb{R}^d} \sum_{i \in \mathcal{H}} Q_i(x)$ such that $x_{\mathcal{H}} \in \mathcal{W}$.*

We now present the fault-tolerance properties of the CGE and CWTM gradient-filters in Sections 4.2.1 and 4.2.2 below, respectively.

4.2.1 CGE Gradient-Filter

To apply the CGE gradient-filter in Step S2, the server sorts the n gradients received from the n agents at the completion of Step S1 as per their Euclidean norms (ties broken arbitrarily):

$$\left\| g^t_{i_1} \right\| \leq \ldots \leq \left\| g^t_{i_{n-f}} \right\| \leq \left\| g^t_{i_{n-f+1}} \right\| \leq \ldots \leq \left\| g^t_{i_n} \right\|.$$

That is, the gradient with the smallest norm, $g^t_{i_1}$, is received from agent i_1, and the gradient with the largest norm, $g^t_{i_n}$, is received from agent i_n. Then, the output of the CGE gradient-filter is the vector sum of the $n - f$ gradients with smallest $n - f$ Euclidean norms. Specifically,

$$\mathsf{GradFilter}\left(g^t_1, \ldots, g^t_n\right) = \sum_{j=1}^{n-f} g^t_{i_j}. \tag{24}$$

We show below, in Theorem 4, that when the fraction of Byzantine faulty agents f/n is bounded then the algorithm in Section 4.1 with the CGE gradient-filter in Step S2 is $(f, \mathcal{O}(\epsilon))$-resilient, under $(2f, \epsilon)$-redundancy and the aforementioned assumptions. To present the formal result we define below some parameters.

- We define a *fault-tolerance margin*

$$\alpha = 1 - \frac{f}{n}\left(1 + \frac{2\mu}{\gamma}\right) \tag{25}$$

  that determines the maximum fraction of Byzantine faulty agents that can be tolerated in an execution of the algorithm.

- We define a coefficient

$$\mathsf{D} = \frac{4\mu f}{\alpha\,\gamma} = \frac{4\mu f}{\gamma - \frac{f}{n}\,(\gamma + 2\mu)} \tag{26}$$

that measures the resilience of the algorithm.

Both α and D depend upon f, the most number of Byzantine faulty agents in any given execution of the algorithm. Note that, under Assumptions 3 and 4, for each non-empty set of non-faulty agents $\mathcal{H}$ with $|\mathcal{H}| = n - f$, the aggregate cost function $\sum_{i \in \mathcal{H}} Q_i(x)$ has a unique minimum point, denoted by $x_{\mathcal{H}}$, in the set $\mathcal{W}$. Specifically,

$$x_{\mathcal{H}} = \arg\min_{x \in \mathbb{R}^d} \sum_{i \in \mathcal{H}} Q_i(x) \cap \mathcal{W}. \tag{27}$$

Theorem 4. *Suppose that the non-faulty agents' cost functions satisfy the* $(2f,\, \epsilon)$*-redundancy property, and Assumptions 2, 3 and 4. Consider the iterative algorithm in Section 4.1 with the CGE gradient-filter defined in* (24)*, and diminishing step-sizes* $\{\eta_0, \eta_1, \ldots\}$ *in* (22) *satisfying:* $\sum_{t=1}^{\infty} \eta_t = \infty$ *and* $\sum_{t=1}^{\infty} \eta_t^2 < \infty$*. If* $\alpha > 0$ *then for each set of* $n - f$ *non-faulty agents* $\mathcal{H}$*,*

$$\lim_{t \to \infty} \left\| x^t - x_{\mathcal{H}} \right\| \leq \mathsf{D}\,\epsilon.$$

Thus, by Definition 2, the algorithm is asymptotically $(f,\, \mathsf{D}\epsilon)$-resilient*.*

The proof of Theorem 4 relies on Theorem 3, and is deferred to Appendix D. Essentially, to prove Theorem 4 we show that the CGE gradient-filter defined in (24) satisfies the conditions specified in Theorem 3 for $\mathsf{D}^* = \mathsf{D}\,\epsilon$, and $x^* = x_{\mathcal{H}}$ for all set of non-faulty agents $\mathcal{H}$ with $|\mathcal{H}| = n - f$.

According to Theorem 4, if $\alpha > 0$, or equivalently, the fraction of Byzantine faulty agents f/n is below the threshold value of $1/\left(1 + 2(\mu/\gamma)\right)$ then, under $(2f,\, \epsilon)$-redundancy property and other standard assumptions, the distributed gradient-descent method with CGE gradient-filter is $(f,\, \mathsf{D}\epsilon)$-resilient. As $\gamma \leq \mu$ under Assumptions 2 and 3 (see Appendix B), the fault-tolerance property of CGE gradient-filter stated in Theorem 4 requires $f/n < 1/3$ or $f < n/3$.

Note that, as the number of Byzantine faulty agents f decreases the value of D decreases, i.e., the resilience of the algorithm improves. Also, note that the value of D is equal to 0 when $f = 0$, and therefore, the algorithm converges to the actual minimum point of all the agents' aggregate cost function in the fault-free case.

4.2.2 CWTM Gradient-Filter

To apply the CWTM gradient-filter in Step S2, the server sorts the n gradients received from the n agents at the completion of Step S1 as per their individual elements. For a vector $v \in \mathbb{R}^d$, we let $v[k]$ denote its k-th element. Specifically, for each $k \in \{1, \ldots,\, d\}$, the server sorts the k-th elements of the gradients (ties broken arbitrarily):

$$g^t_{i_1[k]}[k] \leq \ldots \leq g^t_{i_{f+1}[k]}[k] \leq \ldots \leq g^t_{i_{n-f}[k]}[k] \leq \ldots \leq g^t_{i_n[k]}[k].$$

That is, the gradient with the smallest k-th element, $g^t_{i_1[k]}$, is received from agent $i_1[k]$, and the gradient with the largest norm, $g^t_{i_n[k]}$, is received from agent $i_n[k]$. For each k, the server eliminates the largest f and the smallest f elements of the gradients received. Then, the output of the CWTM gradient-filter is a vector whose k-th element is equal to the average of the remaining $n-2f$ gradients' k-th elements. That is, for each $k \in \{1, \ldots, d\}$,

$$\mathsf{GradFilter}\left(g^t_1, \ldots, g^t_n\right)[k] = \frac{1}{n-2f} \sum_{j=f+1}^{n-f} g^t_{i_j[k]}[k]. \tag{28}$$

We show below in Theorem 5 that when the separation between the gradients of non-faulty agents' cost functions is small enough then the CWTM gradient-filter can guarantee approximate fault-tolerance under $(2f, \epsilon)$-redundancy. To formally present our result we make the following additional assumption about the distance between the gradients of two non-faulty agents' cost functions.

Assumption 5. *For two non-faulty agents i and j, we assume that there exists $\lambda > 0$ such that for all $x \in \mathcal{W}$,*

$$\left\| \nabla Q_i(x) - \nabla Q_j(x) \right\| \leq \lambda \max \left\{ \left\| Q_i(x) \right\|, \left\| Q_j(x) \right\| \right\}.$$

Obviously, owing to the triangle triangle inequality, Assumption 5 is always true for $\lambda = 2$. However, as shown below in Theorem 5, we can presently guarantee fault-tolerance of CWTM gradient-filter when $\lambda < \gamma/(\mu\sqrt{d})$ where μ and γ are the Lipschitz smoothness and strong convexity coefficients, respectively defined in Assumption 2 and 3. Recall the definition of $x_{\mathcal{H}}$ from (27).

Theorem 5. *Suppose that the non-faulty agents' cost functions satisfy the $(2f, \epsilon)$-redundancy property, and Assumptions 2, 3, 4 and 5. Consider the iterative algorithm in Section 4.1 with the CWTM gradient-filter defined in (28), and diminishing step-sizes $\{\eta_0, \eta_1, \ldots\}$ in (22) satisfying: $\sum_{t=1}^{\infty} \eta_t = \infty$ and $\sum_{t=1}^{\infty} \eta_t^2 < \infty$. If $\lambda < \gamma/(\mu\sqrt{d})$ then for each set of $n-f$ non-faulty agents $\mathcal{H}$,*

$$\lim_{t \to \infty} \left\| x^t - x_{\mathcal{H}} \right\| \leq \left(\frac{2n\lambda}{(\gamma/\mu\sqrt{d}) - \lambda} \right) \epsilon.$$

Proof of Theorem 5 is similar to Theorem 4, and is deferred to Appendix E.

By Definition 2 of approximate fault-tolerance, Theorem 5 implies the algorithm with CWTM gradient-filter is asymptotically (f, σ)-*resilient* where

$$\sigma = \left(\frac{2n\lambda}{(\gamma/\mu\sqrt{d}) - \lambda} \right) \epsilon.$$

Note that smaller is the value of λ, smaller is the value of σ and therefore, better is the resilience guarantee of the CWTM gradient-filter.

Unlike CGE gradient-filter, the obtain resilience of CWTM in Theorem 5 is independent of f, as long as the separation between the gradients of non-faulty agents' cost functions is sufficiently small, i.e., $\lambda < \gamma/(\mu\sqrt{d})$. However, unlike the sufficient condition for resilience of CGE gradient-filter, the condition on λ to guarantee the resilience of CWTM gradient-filter depends upon the dimension d of the optimization problem. Larger dimension result in a tighter bound on λ.

5 Numerical experiments

In this section, we present simulation results to empirically compare the approximate fault-tolerance achieved by the aforementioned gradient-filters; CGE and CWTM. For the simulation, we consider the problem of distributed linear regression, which is a special distributed optimization problem with quadratic cost functions [18].

5.1 Problem description

We consider a synchronous server-based system, as shown in Figure 1, wherein $n = 6$, $d = 2$, and $f = 1$. Each agent $i \in \{1, \ldots, n\}$ has a data point represented by a triplet (A_i, B_i, N_i) where A_i is a d-dimensional row vector, $B_i \in \mathbb{R}$ as the response, and a noise value $N_i \in \mathbb{R}$. Specifically, for all $i \in \{1, \ldots, n\}$,

$$B_i = A_i x^* + N_i \quad \text{where} \quad x^* = \begin{pmatrix} 1 \\ 1 \end{pmatrix}. \tag{29}$$

The collective data is represented by a triplet of matrices (A, B, N) where the i-th row of A, B, and N are equal to A_i, B_i and N_i, respectively. The specific values are as follows.

$$A = \begin{pmatrix} 1 & 0 \\ 0.8 & 0.5 \\ 0.5 & 0.8 \\ 0 & 1 \\ -0.5 & 0.8 \\ -0.8 & 0.5 \end{pmatrix}, \; B = \begin{pmatrix} 0.9108 \\ 1.3349 \\ 1.3376 \\ 1.0033 \\ 0.2142 \\ -0.3615 \end{pmatrix}, \; \text{and } N = \begin{pmatrix} -0.0892 \\ 0.0349 \\ 0.0376 \\ 0.0033 \\ -0.0858 \\ -0.0615 \end{pmatrix}. \tag{30}$$

It should be noted that

$$B = Ax^* + N. \tag{31}$$

We let A_S, B_S and N_S represent matrices of dimensions $|S| \times 2$, $|S| \times 1$ and $|S| \times 1$ obtained by stacking the rows $\{A_i, \, i \in S\}$, $\{B_i, \, i \in S\}$ and $\{N_i, \, i \in S\}$, respectively, in the increasing order. From (31), observe that for every non-empty set S,

$$B_S = A_S x^* + N_S. \tag{32}$$

Recall from basic linear algebra that if A_S is full-column rank, i.e., $\text{rank}\,(A_S) = d = 2$ then x^* is the unique solution of the set of equations in (32). Note that for every set S with $|S| \geq n - 2f = 6 - 2 = 4$, the matrix A_S is full rank. Specifically,

$$\text{rank}\,(A_S) = d = 2, \quad \forall S \subseteq \{1, \ldots, 6\}, \; |S| \geq 4. \tag{33}$$

In this particular distributed optimization problem, each agent i has a quadratic cost function defined to be

$$Q_i(x) = (B_i - A_i x)^2, \quad \forall x \in \mathbb{R}^2.$$

For an arbitrary non-empty set of agents S, we define

$$Q_S(x) = \sum_{i \in S} Q_i(x) = \sum_{i \in S} (B_i - A_i x)^2, \quad \forall x \in \mathbb{R}^2.$$

Therefore,

$$Q_S(x) = \sum_{i \in S} (B_i - A_i x)^2 = \|B_S - A_S x\|^2 . \tag{34}$$

As matrix A_S is full rank for every S with $|S| \geq 4$,

$$\arg\min_{x \in \mathbb{R}^2} Q_S(x) = \arg\min_{x \in \mathbb{R}^2} \|B_S - A_S x\|^2 = (x \mid A_S x = B_S) . \tag{35}$$

Therefore, $Q_S(x)$ has a unique minimum point when $|S| \geq 4$. Henceforth, we write notation $\arg\min_{x \in \mathbb{R}^2}$ simply as $\arg\min$, unless otherwise stated.

5.2 Properties of agents' cost functions

Due to the rank condition (33), the agents' cost functions satisfy the $(2f,\, \epsilon)$-*redundancy* property, stated in Definition 3, with $\epsilon = 0.0890$. The steps for computing ϵ are described below.

1. For each set $S \subset \{1, \ldots, 6\}$ with $|S| = n - f = 5$, compute $x_S \in \mathbb{R}^2$ such that $B_S = A_S x_S$. Note that, due to (35), $x_S = \arg\min Q_S(x)$.
2. For each set $S \subset \{1, \ldots, 6\}$ with $|S| = n - f = 5$ do the following:
   (a) For each set $\widehat{S} \subseteq S$ with $\left|\widehat{S}\right| \geq n - 2f = 4$, compute $x_{\widehat{S}}$ such that $B_{\widehat{S}} = A_{\widehat{S}} x_{\widehat{S}}$. Note that, due to (35), $x_{\widehat{S}} = \arg\min Q_{\widehat{S}}(x)$.
   (b) Compute
   $$\epsilon_S = \max_{\widehat{S} \subseteq S,\, \left|\widehat{S}\right| \geq 4} \left\| x_S - x_{\widehat{S}} \right\| .$$
   In this particular case, both the sets of minimum points $\arg\min Q_S(x)$ and $\arg\min Q_{\widehat{S}}(x)$ are singleton with points x_S and $x_{\widehat{S}}$, respectively. Therefore,
   $$\left\| x_S - x_{\widehat{S}} \right\| = \mathrm{dist}\left(\arg\min Q_S(x),\, \arg\min Q_{\widehat{S}}(x)\right) .$$
3. In the final step, we compute
   $$\epsilon = \max_{|S| = n - f} \epsilon_S .$$

For each agent i, its cost function $Q_i(x)$ has Lipschitz continuous gradients, i.e., satisfy Assumption 2, with Lipschitz coefficient

$$\mu = \overline{v}_i \tag{36}$$

where $\overline{v}_i$ denotes the largest eigenvalue of $A_i^T A_i$. Also, for every set of agents S with $|S| = n - f = 5$, their average cost function $(1/|S|)Q_S(x)$ is strongly convex, i.e., satisfy Assumption 3, with the strong convexity coefficient

$$\gamma = \frac{1}{|S|} \underline{v}_S \tag{37}$$

where $\underline{v}_S$ is the smallest eigenvalue of $A_S^T A_S$. Derivations of (36) and (37) can found in [18, Section 10].

5.3 Simulation

In our experiments, we simulate the following fault behaviors for the faulty agent.

- *gradient-reverse*: the faulty agent *reverses* its true gradient. Suppose the correct gradient of a faulty agent i at step t is s_i^t, the agent i will send the incorrect gradient $g_i^t = -s_i^t$ to the server.
- *random*: the faulty agent sends a randomly chosen vector in $\mathbb{R}^d$. In our experiments, the faulty agent in each step chooses i.i.d. Gaussian random vector with mean 0 and a isotropic covariance matrix with standard deviation of 200.

We simulate the distributed gradient-descent algorithm described in Section 4.1 by assuming agent 1 to be Byzantine faulty. It should be noted that the identity of the faulty agent is not used in any way during the simulations. Here, the set of non-faulty agents is $\mathcal{H} = \{2, \ldots, 6\}$ and $|\mathcal{H}| = n - f = 5$. Therefore, in this particular case $\mathcal{H}$ is the only set of $n - f$ non-faulty agents. From (35), we obtain that the minimum point of the aggregate cost function $\sum_{i \in \mathcal{H}} Q_i(x)$, denoted by $x_{\mathcal{H}}$, is equal to the solution of the following set of linear equations:

$$B_{\mathcal{H}} = A_{\mathcal{H}} x_{\mathcal{H}}.$$

Specifically, $x_{\mathcal{H}} = \begin{pmatrix} 1.0780 \\ 0.9825 \end{pmatrix}$. Also, note from our earlier deductions in (36) and (37) that in this particular case, the non-faulty agents' cost functions satisfy Assumptions 2 and 3 with $\mu = 1$ and $\gamma = 0.356$, respectively.

Parameters: We use the following parameters for implementing the algorithm. In the update rule (22), we use step size $\eta_t = 1.5/(t+1)$ for iteration $t = 0, 1, \ldots$. Note that this particular step-size is diminishing and satisfies the conditions: $\sum_{t=0}^{\infty} \eta_t = \infty$ and $\sum_{t=0}^{\infty} \eta_t^2 = 3\pi^2/8 < \infty$ (see [30]). We assume the convex compact $\mathcal{W} \subset \mathbb{R}^d$ to be a 2-dimensional hypercube $[-1000, 1000]^2$. Note that $x_{\mathcal{H}} \in \mathcal{W}$, i.e., Assumption 4 holds true. In all the simulation results presented below, the initial estimate $x^0 = (-0.0085, -0.5643)$.

In every execution, we observe that the iterative estimates produced by the algorithm practically converge after 400 iterations. Thus, to measure the approximate fault-tolerance achieved by the different gradient-filter, i.e., CGE and CWTM, we define the output of the algorithm to be $x_{\text{out}} = x^{500}$. The outputs for the two gradient-filters, under different faulty behaviors, are shown in Table 1. Note that $\text{dist}\,(x_{\mathcal{H}}, x_{\text{out}}) = \|x_{\mathcal{H}} - x_{\text{out}}\|$. The results for the case when the faulty agent sends *random* faulty gradients are only shown for a randomly chosen execution.

Conclusion: As shown in Table 1, in all executions, the distances between $x_{\mathcal{H}}$ the output of the algorithm x_{out} in case of both CGE and CWTM gradient-filters are smaller than $\epsilon = 0.0890$. For the said executions, we plot in Figure 2 the values of the aggregate cost function $\sum_{i \in \mathcal{H}} Q_i(x^t)$ (referred as *loss*) and the approximation error $\left\|x^t - x_{\mathcal{H}}\right\|$ (referred as *distance*) for iteration t ranging from 0 to 500. We also

Table 1: *Algorithm's outputs with gradient-filters CGE and CWTM, and the approximation errors, corresponging to executions when the faulty agent 1 exhibits two different types of Byzantine faults; gradient-reverse and random. Recall that* $x_{\text{out}} = x^{500}$ *and* $\text{dist}(x_{\mathcal{H}}, x_{\text{out}}) = \|x_{\mathcal{H}} - x_{\text{out}}\|$.

	gradient-reverse		**random**	
	x_{out}	$\text{dist}(x_{\mathcal{H}}, x_{\text{out}})$	x_{out}	$\text{dist}(x_{\mathcal{H}}, x_{\text{out}})$
CGE	$\begin{pmatrix}1.0541\\0.9826\end{pmatrix}$	0.0239	$\begin{pmatrix}1.0779\\0.9826\end{pmatrix}$	4.72×10^{-5}
CWTM	$\begin{pmatrix}1.0645\\0.9924\end{pmatrix}$	0.0167	$\begin{pmatrix}1.0775\\0.9840\end{pmatrix}$	1.51×10^{-3}

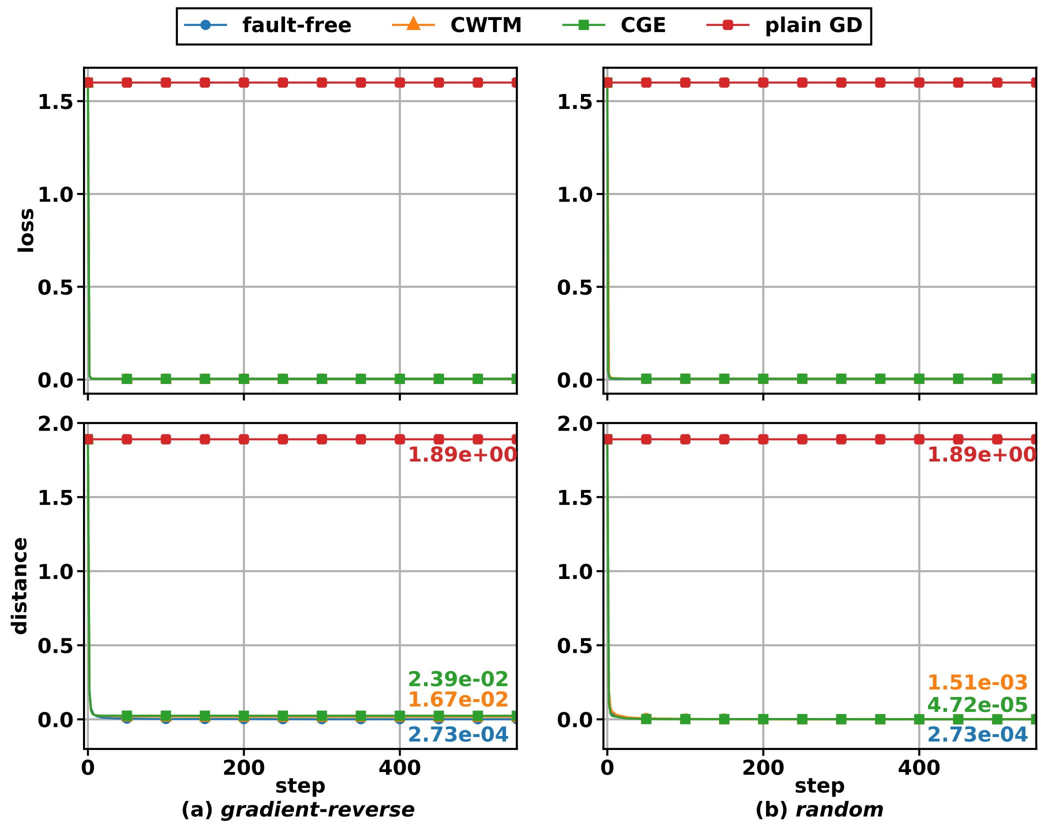


Figure 2: *The losses, i.e.,* $\sum_{i\in\mathcal{H}} Q_i(x^t)$, *and distances, i.e.,* $\left\|x^t - x_{\mathcal{H}}\right\|$, *versus the number of iterations in the algorithm. The final approximation errors, i.e.,* $\left\|x^{500} - x_{\mathcal{H}}\right\|$, *are annotated in the same colors of their corresponding plots. For the executions shown, agent* 1 *is assumed to be Byzantine faulty. The different columns show the results when the faulty agent exhibits the different types of faults:* (a) *gradient-reverse, and* (b) *random. Apart from the plots with CGE (in* green*) and CWTM (in* yellow*) gradient-filters, we also plot the fault-free distributed gradient-descent (DGD) method when all agents are free from faults (in* blue*), and the DGD method without any gradient-filters when agent* 1 *is Byzantine faulty (in* red*).*

show the plots of the fault-free distributed gradient-descent (DGD) method when all agents are free from faults, and the DGD method without any gradient-filter when agent 1 is Byzantine faulty. The details for iteration t ranging from 0 to 80 are also highlighted in Figure 3.

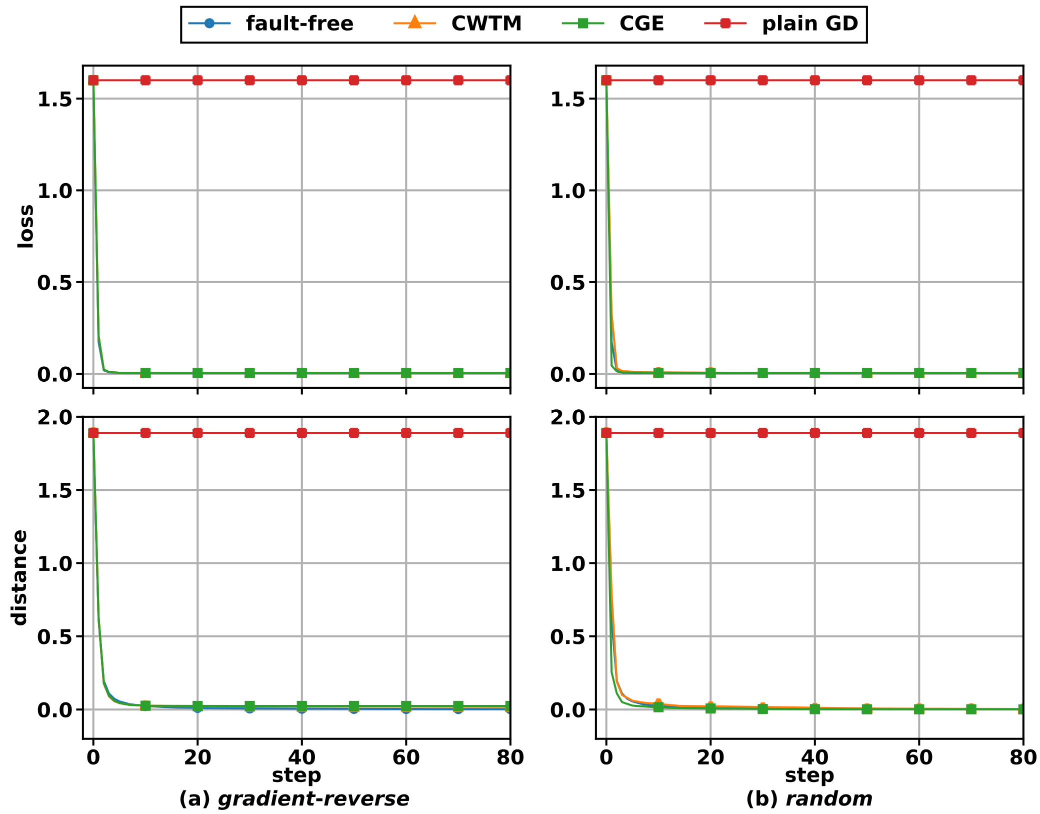


Figure 3: *The losses, i.e., $\sum_{i \in \mathcal{H}} Q_i(x^t)$, and distances, i.e., $\left\| x^t - x_{\mathcal{H}} \right\|$, versus the number of iterations in the algorithm, magnified for the initial 80 iterations in the training process. The meaning of plots are the same as in Figure 2.*

6 Summary

In this paper, we have studied the problem of *approximate* Byzantine fault-tolerance, which is a generalization of the *exact* fault-tolerance problem studied in prior work [20]. Unlike the exact fault-tolerance, the goal in approximate fault-tolerance is to design a distributed optimization problem that produces only an approximation of a minimum point of the aggregate cost function of at least $n - f$ non-faulty agents, in the presence of up to f (out of n) Byzantine faulty agents.

We have defined approximate fault-tolerance formally as (f, ϵ)-resilience where $\epsilon \in \mathbb{R}_{\geq 0}$ denotes the approximation error. In the first part of the paper, i.e, Section 3, we have obtained necessary and sufficient conditions for the achievability of (f, ϵ)-resilience. These results generalize the prior result which states that exact fault-tolerance is achievable if and only if $2f$-redundancy property is satisfied [20, 21]. In the second part of the paper, i.e., Sections 4 and 5, we have considered the case when agents' cost functions are differentiable. For this particular case, first we have derived a generic approximate fault-tolerance property of the distributed gradient-descent method when equipped with Byzantine robust gradient aggregation or *gradient-filter*. Then, we have obtained specific approximate fault-tolerance guarantees for two well-known gradient-filters; comparative gradient elimination (CGE) and coordinate-wise trimmed mean (CWTM). Finally, in Section 5, we have presented empirical results comparing the approximate fault-tolerance achieved by the two aforementioned gradient-filters.

7 Acknowledgements

Research reported in this paper was sponsored in part by the Army Research Laboratory under Cooperative Agreement W911NF- 17-2-0196, and by the National Science Foundation award 1842198. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory, National Science Foundation or the U.S. Government.

References

[1] Dan Alistarh, Zeyuan Allen-Zhu, and Jerry Li. Byzantine stochastic gradient descent. In *Advances in Neural Information Processing Systems*, 2018.

[2] Jeremy Bernstein, Jiawei Zhao, Kamyar Azizzadenesheli, and Anima Anandkumar. signsgd with majority vote is communication efficient and Byzantine fault tolerant. *arXiv preprint arXiv:1810.05291*, 2018.

[3] Dimitri P Bertsekas and John N Tsitsiklis. *Parallel and distributed computation: numerical methods*, volume 23. Prentice hall Englewood Cliffs, NJ, 1989.

[4] Peva Blanchard, Rachid Guerraoui, Julien Stainer, et al. Machine learning with adversaries: Byzantine tolerant gradient descent. In *Advances in Neural Information Processing Systems*, pages 119–129, 2017.

[5] Léon Bottou. Online learning and stochastic approximations. *On-line learning in neural networks*, 17(9):142, 1998.

[6] Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311, 2018.

[7] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, Jonathan Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1), 2011.

[8] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[9] Xinyang Cao and Lifeng Lai. Distributed gradient descent algorithm robust to an arbitrary number of byzantine attackers. *IEEE Transactions on Signal Processing*, 67(22):5850–5864, 2019.

[10] Moses Charikar, Jacob Steinhardt, and Gregory Valiant. Learning from untrusted data. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 47–60, 2017.

[11] Yuan Chen, Soummya Kar, and Jose MF Moura. Resilient distributed estimation through adversary detection. *IEEE Transactions on Signal Processing*, 66(9), 2018.

[12] Yudong Chen, Lili Su, and Jiaming Xu. Distributed statistical machine learning in adversarial settings: Byzantine gradient descent. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 1(2):44, 2017.

[13] Michelle S Chong, Masashi Wakaiki, and Joao P Hespanha. Observability of linear systems under adversarial attacks. In *American Control Conference*, pages 2439–2444. IEEE, 2015.

[14] Georgios Damaskinos, Rachid Guerraoui, Rhicheek Patra, Mahsa Taziki, et al. Asynchronous Byzantine machine learning (the case of sgd). In *International Conference on Machine Learning*, pages 1153–1162, 2018.

[15] Ilias Diakonikolas, Gautam Kamath, Daniel M Kane, Jerry Li, Jacob Steinhardt, and Alistair Stewart. Sever: A robust meta-algorithm for stochastic optimization. *arXiv preprint arXiv:1803.02815*, 2018.

[16] John C Duchi, Alekh Agarwal, and Martin J Wainwright. Dual averaging for distributed optimization: Convergence analysis and network scaling. *IEEE Transactions on Automatic control*, 57(3):592–606, 2011.

[17] Nirupam Gupta, Shuo Liu, and Nitin H Vaidya. Byzantine fault-tolerant distributed machine learning using stochastic gradient descent (sgd) and norm-based comparative gradient elimination (cge). *arXiv preprint arXiv:2008.04699*, 2020.

[18] Nirupam Gupta and Nitin H Vaidya. Byzantine fault tolerant distributed linear regression. *arXiv preprint arXiv:1903.08752*, 2019.

[19] Nirupam Gupta and Nitin H Vaidya. Byzantine fault-tolerant parallelized stochastic gradient descent for linear regression. In *2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 415–420. IEEE, 2019.

[20] Nirupam Gupta and Nitin H Vaidya. Fault-tolerance in distributed optimization: The case of redundancy. In *Proceedings of the 39th Symposium on Principles of Distributed Computing*, pages 365–374, 2020.

[21] Nirupam Gupta and Nitin H Vaidya. Resilience in collaborative optimization: redundant and independent cost functions. *arXiv preprint arXiv:2003.09675*, 2020.

[22] Leslie Lamport, Robert Shostak, and Marshall Pease. The Byzantine generals problem. *ACM Transactions on Programming Languages and Systems*, 4(3), 1982.

[23] Nancy A Lynch. *Distributed algorithms*. Elsevier, 1996.

[24] James R Munkres. Topology, 2000.

[25] Angelia Nedic and Asuman Ozdaglar. Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, 54(1), 2009.

[26] Miroslav Pajic, James Weimer, Nicola Bezzo, Paulo Tabuada, Oleg Sokolsky, Insup Lee, and George J Pappas. Robustness of attack-resilient state estimators. In *ICCPS'14: ACM/IEEE 5th International Conference on Cyber-Physical Systems (with CPS Week 2014)*, pages 163–174. IEEE Computer Society, 2014.

[27] Adarsh Prasad, Arun Sai Suggala, Sivaraman Balakrishnan, and Pradeep Ravikumar. Robust estimation via robust gradient estimation. *arXiv preprint arXiv:1802.06485*, 2018.

[28] Michael Rabbat and Robert Nowak. Distributed optimization in sensor networks. In *Proceedings of the 3rd international symposium on Information processing in sensor networks*, pages 20–27, 2004.

[29] Robin L Raffard, Claire J Tomlin, and Stephen P Boyd. Distributed optimization for cooperative agents: Application to formation flight. In *2004 43rd IEEE Conference on Decision and Control (CDC)*, volume 3, pages 2453–2459. IEEE, 2004.

[30] Walter Rudin. *Principles of mathematical analysis*, volume 3. McGraw-hill New York, 1964.

[31] Lili Su and Shahin Shahrampour. Finite-time guarantees for Byzantine-resilient distributed state estimation with noisy measurements. *arXiv preprint arXiv:1810.10086*, 2018.

[32] Lili Su and Nitin H Vaidya. Fault-tolerant multi-agent optimization: optimal iterative distributed algorithms. In *Proceedings of the 2016 ACM symposium on principles of distributed computing*, pages 425–434, 2016.

[33] Lili Su and Nitin H Vaidya. Non-bayesian learning in the presence of byzantine agents. In *International symposium on distributed computing*. Springer, 2016.

[34] Lili Su and Nitin H Vaidya. Byzantine-resilient multi-agent optimization. *IEEE Transactions on Automatic Control*, 2020.

[35] Cong Xie, Oluwasanmi Koyejo, and Indranil Gupta. Generalized Byzantine-tolerant sgd. *arXiv preprint arXiv:1802.10116*, 2018.

[36] Dong Yin, Yudong Chen, Kannan Ramchandran, and Peter Bartlett. Byzantine-robust distributed learning: Towards optimal statistical rates. In *International Conference on Machine Learning*, pages 5636–5645, 2018.

A Appendix: The Special Case of $(f,\, 0)$-Resilience

We show that $(f,\, 0)$-*resilience*, stated in Definition 2, and *exact fault-tolerance*, defined in Section 1.1, are equivalent in the deterministic framework. Specifically, we show that a deterministic $(f,\, 0)$-resilient algorithm achieves exact fault-tolerance, and a deterministic exact fault-tolerant algorithm is $(f,\, 0)$-resilient. We consider the server-based system architecture. Notation $\arg\min_{x \in \mathbb{R}^d}$ is simply written as $\arg\min$, unless otherwise stated.

First, we show that $(f,\, 0)$-resilience implies exact fault-tolerance. Suppose that there exists a deterministic $(f,\, 0)$-resilient algorithm Π. Consider an arbitrary execution $E_{\mathcal{H}}$ of Π wherein $\mathcal{H} \subseteq \{1, \ldots, n\}$ denotes the set of all the non-faulty agents, and let $\widehat{x}$ denote the output. Recall that, as there are at most f faulty agents, $|\mathcal{H}| \geq n - f$. To prove that Π has exact fault-tolerance, it suffices to show that, in execution $E_{\mathcal{H}}$, $\widehat{x}$ is a minimum point of the aggregate cost function of all non-faulty agents $\sum_{i \in \mathcal{H}} Q_i(x)$.

By Definition 2 of $(f,\, 0)$-resilience, for every set $S \subseteq \mathcal{H}$ with $|S| = n - f$,

$$\widehat{x} \in \arg\min \sum_{i \in S} Q_i(x).$$

Therefore, for every set S with $S \subseteq \mathcal{H}$ and $|S| = n - f$,

$$\sum_{i \in S} Q_i(\widehat{x}) \leq \sum_{i \in S} Q_i(x), \quad \forall x \in \mathbb{R}^d. \tag{38}$$

Now, note that there are $\binom{|\mathcal{H}|}{n-f}$ subsets in $\mathcal{H}$ of size $n - f$, and each agent $i \in \mathcal{H}$ is contained in $\binom{|\mathcal{H}|-1}{n-f-1}$ of those subsets. Therefore,

$$\sum_{\substack{S \subseteq \mathcal{H}, \\ |S| = n-f}} \sum_{i \in S} Q_i(x) = \binom{|\mathcal{H}| - 1}{n - f - 1} \sum_{i \in \mathcal{H}} Q_i(x). \tag{39}$$

Substituting from (38) in (39) we obtain that

$$\sum_{i \in \mathcal{H}} Q_i(\widehat{x}) \leq \sum_{i \in \mathcal{H}} Q_i(x), \quad \forall x \in \mathbb{R}^d$$

The above implies that

$$\widehat{x} \in \arg\min \sum_{i \in \mathcal{H}} Q_i(x).$$

The above proves that Π has exact fault-tolerance in execution $E_{\mathcal{H}}$.

Now, we show that exact fault-tolerance implies $(f,\, 0)$-resilience. Suppose that Π is a deterministic algorithm with exact fault-tolerance. Similar to above, consider an arbitrary execution $E_{\mathcal{H}}$ of Π wherein set $\mathcal{H}$ comprises all the non-faulty agents, and $\widehat{x}$ is its output. Therefore,

$$\widehat{x} \in \arg\min \sum_{i \in \mathcal{H}} Q_i(x).$$

To prove that Π is $(f, 0)$-resilient, it suffices to show that in execution $E_{\mathcal{H}}$ for every set $S \subseteq \mathcal{H}$ with $|S| = n - f$, $\widehat{x}$ is a minimum point of the aggregate cost function $\sum_{i \in S} Q_i(x)$. This is trivially true when $|\mathcal{H}| = n - f$. We assume below that $|\mathcal{H}| > n - f$.

Consider an arbitrary subset S of $\mathcal{H}$ with $|S| = n - f$. Consider an execution E_S wherein S is the set of all non-faulty agents, with the remaining agents in $\{1, \ldots, n\} \setminus S$ being Byzantine faulty. Suppose that the inputs from all the agents to the server in E_S are identical to their inputs in $E_{\mathcal{H}}$. Therefore, as Π is a deterministic algorithm, its output in execution E_S is same as that in execution $E_{\mathcal{H}}$, i.e., $\widehat{x}$. Moreover, as Π is assumed to have exact fault-tolerance,

$$\widehat{x} \in \arg\min \sum_{i \in S} Q_i(x).$$

As S is an arbitrary subset S of $\mathcal{H}$ with $|S| = n - f$, the above proves that Π is $(f, 0)$-resilient in execution $E_{\mathcal{H}}$.

B Appendix: Proof of $\gamma \leq \mu$

We show below that if Assumptions 2 and 3 hold true simultaneously then $\gamma \leq \mu$.

Consider an arbitrary set of $n - f$ non-faulty agents $\mathcal{H}$, and two arbitrary non-identical points $x, y \in \mathbb{R}^d$, i.e., $x \neq y$. If Assumption 2 holds true then

$$\left\| \nabla Q_i(x) - \nabla Q_i(y) \right\| \leq \mu \left\| x - y \right\|, \quad \forall i \in \mathcal{H}.$$

Therefore, owing to the Cauchy-Schwartz inequality, for all $i \in \mathcal{H}$,

$$\langle x - y, \, \nabla Q_i(x) - \nabla Q_i(y) \rangle \leq \left\| x - y \right\| \, \left\| \nabla Q_i(x) - \nabla Q_i(y) \right\| \leq \mu \left\| x - y \right\|^2. \tag{40}$$

From (40) we obtain that

$$\sum_{i \in \mathcal{H}} \langle x - y, \, \nabla Q_i(x) - \nabla Q_i(y) \rangle \leq \mu \left| \mathcal{H} \right| \, \left\| x - y \right\|^2. \tag{41}$$

If Assumption 3 holds true then

$$\sum_{i \in \mathcal{H}} \langle x - y, \, \nabla Q_i(x) - \nabla Q_i(y) \rangle \geq \gamma \left| \mathcal{H} \right| \, \left\| x - y \right\|^2. \tag{42}$$

As x, y are arbitrary non-identical points, (41) and (42) together imply that $\gamma \leq \mu$.

C Appendix: Proof of Theorem 3

The proof of Theorem 3 relies on the following sufficient criterion for the convergence of non-negative sequences.

Lemma 2 (Bottou, 1998 [5]). *Consider a sequence of real values $\{u_t,\, t = 0,\, 1, \ldots\}$. If $u_t \geq 0$, $\forall t$ then*

$$\sum_{t=0}^{\infty}(u_{t+1} - u_t)_+ < \infty \implies \begin{cases} u_t \underset{t\to\infty}{\longrightarrow} u_\infty < \infty \\ \\ \sum_{t=0}^{\infty}(u_{t+1} - u_t)_- > -\infty \end{cases} \tag{43}$$

where the operators $(\cdot)_+$ and $(\cdot)_-$ are defined as follows for a real scalar x,

$$(x)_+ = \begin{cases} x & , \quad x > 0 \\ 0 & , \quad otherwise \end{cases}, \; and \; (x)_- = \begin{cases} 0 & , \quad x > 0 \\ x & , \quad otherwise \end{cases}$$

Recall from the statement of Theorem 3 that $x^* \in \mathcal{W}$ where $\mathcal{W}$ is a compact convex set. We define, for all $t \in \{0, 1, \ldots\}$,

$$e_t = \left\| x^t - x^* \right\|. \tag{44}$$

Next, we define a univariate real-valued function $\psi : \mathbb{R} \to \mathbb{R}$:

$$\psi(x) = \begin{cases} 0, & x < (\mathsf{D}^*)^2, \\ \left(x - (\mathsf{D}^*)^2\right)^2, & x \geq (\mathsf{D}^*)^2. \end{cases} \tag{45}$$

Let $\psi'(x)$ denote the derivative of ψ at x. Specifically,

$$\psi'(x) = \max\left\{0, 2\left(x - (\mathsf{D}^*)^2\right)\right\}. \tag{46}$$

We show below that $\psi'(x)$ is a Lipschitz continuous function with Lipschitz coefficient of 2. From (46), we obtain that

$$\left|\psi'(x) - \psi'(y)\right| = \begin{cases} 2\left|x - y\right| & , \quad \text{both } x,\, y \geq (\mathsf{D}^*)^2 \\ 2\left|x - (D^*)^2\right| & , \quad x \geq (\mathsf{D}^*)^2,\, y < (\mathsf{D}^*)^2 \\ 0 & , \quad \text{both } x,\, y < (\mathsf{D}^*)^2 \end{cases} \tag{47}$$

Note from (47) that for the case when $x \geq (\mathsf{D}^*)^2$, $y < (\mathsf{D}^*)^2$,

$$\left|\psi'(x) - \psi'(y)\right| = 2\left|x - (D^*)^2\right| < 2\left|x - y\right|.$$

Similarly, due to symmetry, when $x < (\mathsf{D}^*)^2$, $y \geq (\mathsf{D}^*)^2$ then $\left|\psi'(x) - \psi'(y)\right| = 22\left|y - (D^*)^2\right| < 2\left|x - y\right|$. Therefore, from (46) we obtain that

$$\left|\psi'(x) - \psi'(y)\right| \leq 2\left|x - y\right|, \quad \forall x,\, y \in \mathbb{R}. \tag{48}$$

The Lipschitz continuity of $\psi'(x)$, shown in (48), implies that [6, Section 4.1]

$$\psi(y) - \psi(x) \leq (y - x)\psi'(x) + (y - x)^2, \quad \forall x,\, y \in \mathbb{R}. \tag{49}$$

Now, for each $t \in \{0, 1, \ldots\}$, we define

$$h_t = \psi(e_t^2). \tag{50}$$

From (49) and (50), for all t, we obtain that

$$h_{t+1} - h_t = \psi\left(e_{t+1}^2\right) - \psi\left(e_t^2\right) \leq \left(e_{t+1}^2 - e_t^2\right) \cdot \psi'\left(e_t^2\right) + \left(e_{t+1}^2 - e_t^2\right)^2.$$

From now on we use ψ'_t as a shorthand for $\psi'(e_t^2)$. From above we have

$$h_{t+1} - h_t \leq \left(e_{t+1}^2 - e_t^2\right) \psi'_t + \left(e_{t+1}^2 - e_t^2\right)^2. \tag{51}$$

Now, recall from (22) that for all $t \in \{0, 1, \ldots\}$,

$$x^{t+1} = \left[x^t - \eta_t \, \mathsf{GradFilter}\left(g_1^t, \ldots, g_n^t\right)\right]_{\mathcal{W}} \tag{52}$$

By the non-expansion property of Euclidean projection onto a closed convex set,

$$\left\|x^{t+1} - x^*\right\| \leq \left\|x^t - x^* - \eta_t \, \mathsf{GradFilter}\left(g_1^t, \ldots, g_n^t\right)\right\|.$$

Recall from (44) that e_t denotes $\left\|x^t - x^*\right\|$ for all t. Upon squaring the both sides in the above inequality, we obtain that

$$e_{t+1}^2 \leq e_t^2 - 2\eta_t \left\langle x_t - x^*, \, \mathsf{GradFilter}\left(g_1^t, \ldots, g_n^t\right)\right\rangle + \eta_t^2 \left\|\mathsf{GradFilter}\left(g_1^t, \ldots, g_n^t\right)\right\|^2. \tag{53}$$

Recall, from (23) in the statement of Theorem 3, that

$$\phi_t = \left\langle x_t - x^*, \, \mathsf{GradFilter}\left(g_1^t, \ldots, g_n^t\right)\right\rangle, \quad \forall t.$$

Substituting from the above in (53), we obtain that

$$e_{t+1}^2 \leq e_t^2 - 2\eta_t \phi_t + \eta_t^2 \left\|\mathsf{GradFilter}\left(g_1^t, \ldots, g_n^t\right)\right\|^2. \tag{54}$$

As $\psi'_t \geq 0$, $\forall t$, substituting from (54) in (51) we get

$$h_{t+1} - h_t \leq \left(-2\eta_t \phi_t + \eta_t^2 \left\|\mathsf{GradFilter}\left(g_1^t, \ldots, g_n^t\right)\right\|^2\right) \psi'_t + \left(e_{t+1}^2 - e_t^2\right)^2. \tag{55}$$

Note that for an arbitrary t,

$$\left|e_{t+1}^2 - e_t^2\right| = \left(e_{t+1} + e_t\right) \left|e_{t+1} - e_t\right|. \tag{56}$$

As $\mathcal{W}$ is assumed compact, there exists $\Gamma = \max_{x \in \mathcal{W}} \left\|x - x^*\right\| < \infty$. We assume $\Gamma > 0$, otherwise $\mathcal{W} = \{x^*\}$ and the theorem is trivial. Recall from the update rule (22), which is re-stated above in (52), that $x^t \in \mathcal{W}$ for all t, and that $x^* \in \mathcal{W}$. Therefore,

$$e_t = \left\|x^t - x^*\right\| \leq \max_{x \in \mathcal{W}} \left\|x - x^*\right\| = \Gamma, \quad \forall t. \tag{57}$$

From (57), for all t, we obtain that

$$e_{t+1} + e_t \leq 2\Gamma.$$

Substituting from above in (56) implies that

$$\left|e_{t+1}^2 - e_t^2\right| \leq 2\Gamma \left|e_{t+1} - e_t\right|, \quad \forall t. \tag{58}$$

From triangle inequality, we get

$$\left|e_{t+1} - e_t\right| = \left|\left\|x^{t+1} - x^*\right\| - \left\|x^t - x^*\right\|\right| \leq \left\|x^{t+1} - x^t\right\|.$$

Substituting from above in (58) we obtain that

$$\left|e_{t+1}^2 - e_t^2\right| \leq 2\Gamma \left\|x^{t+1} - x^t\right\|. \tag{59}$$

Due to the non-expansion property of Euclidean projection onto a closed convex set, from (52) we obtain that

$$\left\|x^{t+1} - x^t\right\| = \left\|\left[x^t - \eta_t \, \mathsf{GradFilter}\left(g_1^t, \ldots, g_n^t\right)\right]_{\mathcal{W}} - x^t\right\| \leq \eta_t \left\|\mathsf{GradFilter}\left(g_1^t, \ldots, g_n^t\right)\right\|.$$

Substituting from above in (59) we obtain that

$$\left|e_{t+1}^2 - e_t^2\right| \leq 2\eta_t \Gamma \left\|\mathsf{GradFilter}\left(g_1^t, \ldots, g_n^t\right)\right\|.$$

Thus,

$$\left(e_{t+1}^2 - e_t^2\right)^2 \leq 4\eta_t^2 \, \Gamma^2 \, \left\|\mathsf{GradFilter}\left(g_1^t, \ldots, g_n^t\right)\right\|^2. \tag{60}$$

Substituting from (60) in (55) we obtain that, for all t,

$$\begin{aligned} h_{t+1} - h_t &\leq \left(-2\eta_t \phi_t + \eta_t^2 \left\|\mathsf{GradFilter}\left(g_1^t, \ldots, g_n^t\right)\right\|^2\right) \psi_t' \\ &\quad + 4\eta_t^2 \Gamma^2 \left\|\mathsf{GradFilter}\left(g_1^t, \ldots, g_n^t\right)\right\|^2 \\ &= -2\eta_t \phi_t \psi_t' + \eta_t^2 \left(\psi_t' + 4\Gamma^2\right) \left\|\mathsf{GradFilter}\left(g_1^t, \ldots, g_n^t\right)\right\|^2. \end{aligned} \tag{61}$$

Recall from (57) that $e_t \leq \Gamma$. Also, by assumption, $\mathsf{D}^* < \max_{x \in \mathcal{W}} \|x - x^*\| = \Gamma$. Recall that ψ_t' is short for $\psi'(e_t^2)$. Therefore, from (46) we obtain that

$$0 \leq \psi_t' \leq 2\left(\Gamma^2 - (\mathsf{D}^*)^2\right) \leq 2\Gamma^2, \quad \forall t. \tag{62}$$

As the statement of Theorem 3 assumes that $\left\|\mathsf{GradFilter}\left(g_1^t, \ldots, g_n^t\right)\right\| < \infty$ for all t, there exists a real value $\mathsf{M} < \infty$ such that

$$\left\|\mathsf{GradFilter}\left(g_1^t, \ldots, g_n^t\right)\right\| \leq \mathsf{M}, \quad \forall t. \tag{63}$$

Substituting from (62) and (63) in (61) we obtain that

$$h_{t+1} - h_t \leq -2\eta_t \phi_t \psi_t' + 6\eta_t^2 \Gamma^2 \mathsf{M}^2. \tag{64}$$

We now use Lemma 2 to prove that $h_\infty = 0$ as follows.

For an iteration t, we consider below two possible cases: (i) $e_t < \mathsf{D}^*$, and (ii) $e_t = \mathsf{D}^* + \delta$ for some $\delta \geq 0$.

Case i) In this particular case, $\psi_t' = 0$. Therefore, due to Cauchy-Schwartz inequality,

$$|\phi_t| = \left|\left\langle x^t - x^*, \, \mathsf{GradFilter}\left(g_1^t, \ldots, g_n^t\right)\right\rangle\right| \leq e_t \left\|\mathsf{GradFilter}\left(g_1^t, \ldots, g_n^t\right)\right\|.$$

Substituting from (63) above, we obtain that $|\phi_t| \leq \Gamma\mathsf{M} < \infty$. Therefore,

$$\phi_t \psi_t' = 0. \tag{65}$$

Case ii) In this particular case, from (46), we obtain that

$$\psi_t' = 2\left((\mathsf{D}^* + \delta)^2 - (\mathsf{D}^*)^2\right) = 2\delta(2\mathsf{D}^* + \delta).$$

Now, by assumption, $\phi_t \geq \xi$ when $e_t \geq \mathsf{D}^*$ where $\xi > 0$. Therefore,

$$\phi_t \psi_t' \geq 2\xi\delta(2\mathsf{D}^* + \delta). \tag{66}$$

From (65) and (66) above, we obtain that

$$\phi_t \psi_t' \geq 0, \quad \forall t. \tag{67}$$

Substituting the above in (64) implies that

$$h_{t+1} - h_t \leq 6\eta_t^2 \Gamma^2 \mathsf{M}^2, \quad \forall t.$$

Recall that notation $(\cdot)_+$ from Lemma 2. The above inequality implies that

$$(h_{t+1} - h_t)_+ \leq 6\eta_t^2 \Gamma^2 \mathsf{M}^2.$$

As $\eta_t^2 < \infty$, $\forall t$, and constants L, $\mathsf{M} < \infty$, the above implies that

$$\sum_{t=0}^{\infty} (h_{t+1} - h_t)_+ \leq 6\Gamma^2 \mathsf{M}^2 \sum_{t=0}^{\infty} \eta_t^2 < \infty.$$

As $h_t \geq 0$ for all t, the above in conjunction with Lemma 2 implies that

$$\begin{aligned} &h_t \xrightarrow{t \to \infty} h_\infty < \infty, \text{ and} \\ &\sum_{t=0}^{\infty} (h_{t+1} - h_t)_- > -\infty. \end{aligned} \tag{68}$$

Note that $h_\infty - h_0 = \sum_{t=0}^{\infty} (h_{t+1} - h_t)$. Thus, from (64) we obtain that

$$h_\infty - h_0 \leq -2 \sum_{t=0}^{\infty} \eta_t \phi_t \psi_t' + 6\Gamma^2 \mathsf{M}^2 \sum_{t=0}^{\infty} \eta_t^2. \tag{69}$$

By Definition (50), $h_t \geq 0$ for all t. Therefore, from (69) above we obtain that

$$2\left|\sum_{t=0}^{\infty} \eta_t \phi_t \psi_t'\right| \leq h_0 + h_\infty + 6\Gamma^2 \mathsf{M}^2 \sum_{t=0}^{\infty} \eta_t^2. \tag{70}$$

By assumption, $\sum_{t=0}^{\infty} \eta_t^2 < \infty$. From (68), $h_\infty < \infty$. Substituting from (57) that $e_t < \infty$, $\forall t$ in Definition of h_t (50), we obtain that $h_0 = \psi(e_0^2) < \infty$. Therefore, (70) implies that

$$2\left|\sum_{t=0}^{\infty} \eta_t \phi_t \psi_t'\right| < \infty.$$

Recall from (67) that $\phi_t \psi_t' \geq 0$ for all t. Thus, from above we obtain that

$$\sum_{t=0}^{\infty} \eta_t \phi_t \psi_t' < \infty. \tag{71}$$

Finally, we reason below by contradiction that $h_\infty = 0$.

Note that for any $\zeta > 0$, there exists a unique positive value β such that $\zeta = 2\beta\left(2\mathsf{D}^* + \sqrt{\beta}\right)^2$. Suppose that $h_\infty = 2\beta(2\mathsf{D}^* + \sqrt{\beta})^2$ for some positive value β. As the sequence $\{h_t\}_{t=0}^{\infty}$ converges to h_∞ (see (68)), there exists some finite $\tau \in \mathbb{Z}_{\geq 0}$ such that for all $t \geq \tau$,

$$|h_t - h_\infty| \leq \beta\left(2\mathsf{D}^* + \sqrt{\beta}\right)^2 \implies h_t \geq h_\infty - \beta\left(2\mathsf{D}^* + \sqrt{\beta}\right)^2.$$

As $h_\infty = 2\beta(2\mathsf{D}^* + \sqrt{\beta})^2$, the above implies that

$$h_t \geq \beta\left(2\mathsf{D}^* + \sqrt{\beta}\right)^2, \quad \forall t \geq \tau. \tag{72}$$

Therefore (cf. (45) and (50)), for all $t \geq \tau$,

$$\left(e_t^2 - (\mathsf{D}^*)^2\right)^2 \geq \beta\left(2\mathsf{D}^* + \sqrt{\beta}\right)^2, \text{ or}$$
$$\left|e_t^2 - (\mathsf{D}^*)^2\right| \geq \sqrt{\beta}\left(2\mathsf{D}^* + \sqrt{\beta}\right).$$

Thus, for each $t \geq \tau$, either

$$e_t^2 \geq (\mathsf{D}^*)^2 + \sqrt{\beta}\left(2\mathsf{D}^* + \sqrt{\beta}\right) = \left(\mathsf{D}^* + \sqrt{\beta}\right)^2, \tag{73}$$

or

$$e_t^2 \leq (\mathsf{D}^*)^2 - \sqrt{\beta}\left(2\mathsf{D}^* + \sqrt{\beta}\right) < (\mathsf{D}^*)^2. \tag{74}$$

If the latter, i.e. (74), holds true for some $t' \geq \tau$ then

$$h_{t'} = \psi\left(e_{t'}^2\right) = 0,$$

which contradicts (72). Therefore, (72) implies (73) for all $t \geq \tau$.

From above we obtain that if $h_\infty = 2\beta(2\mathsf{D}^* + \sqrt{\beta})^2$ then there exists $\tau < \infty$ such that for all $t \geq \tau$,

$$e_t \geq \mathsf{D}^* + \sqrt{\beta}.$$

Thus, from (66) we obtain that

$$\phi_t \psi_t' \geq 2\xi\sqrt{\beta}(2\mathsf{D}^* + \sqrt{\beta}), \quad \forall t \geq \tau.$$

Therefore,

$$\sum_{t=\tau}^{\infty} \eta_t \phi_t \psi_t' \geq 2\xi\sqrt{\beta}(2\mathsf{D}^* + \sqrt{\beta}) \sum_{t=\tau}^{\infty} \eta_t = \infty.$$

This is a contradiction to (71). Therefore, $h_\infty = 0$, and by definition of h_t in (50),

$$h_\infty = \lim_{t\to\infty} \psi(e_t^2) = 0.$$

Hence, by definition of $\psi(\cdot)$ in (45),

$$\lim_{t\to\infty} \left\|x^t - x^*\right\| \leq \mathsf{D}^*.$$

D Appendix: Proof of Theorem 4

In this section we present the proof of Theorem 4. Throughout this section we assume $f > 0$ to ignore the trivial case of $f = 0$.

Consider an arbitrary set $\mathcal{H}$ of non-faulty agents with $|\mathcal{H}| = n - f$. Recall that under Assumptions 3 and 4, the aggregate cost function $\sum_{i \in \mathcal{H}} Q_i(x)$ has a unique minimum point in set $\mathcal{W}$, which we denote by $x_{\mathcal{H}}$. Specifically,

$$x_{\mathcal{H}} = \arg\min_{x \in \mathbb{R}^d} \sum_{i \in \mathcal{H}} Q_i(x) \cap \mathcal{W}. \tag{75}$$

To prove the theorem we make use of the result stated in Theorem 3. Specifically, we show is that the CGE gradient-filter satisfies the conditions of Theorem 3 for $\mathsf{D}^* = \mathsf{D}\,\epsilon$, and $x^* = x_{\mathcal{H}}$. The rest follows easily from the convergence result stated in Theorem 3. Recall from (24) that for CGE gradient-filter, in update rule (22),

$$\mathsf{GradFilter}\left(g_1^t, \ldots, g_n^t\right) = \sum_{j=1}^{n-f} g_{i_j}^t, \quad \forall t. \tag{76}$$

First, we show that $\left\|\sum_{j=1}^{n-f} g_{i_j}^t\right\|$ is finite for all t.

Consider a subset $S_1 \subset \mathcal{H}$ with $|S_1| = n - 2f$. Triangle inequality implies that

$$\left\|\sum_{j \in S_1} \nabla Q_j(x) - \sum_{j \in S_1} \nabla Q_j(x_{\mathcal{H}})\right\| \leq \sum_{j \in S_1} \left\|\nabla Q_j(x) - \nabla Q_j(x_{\mathcal{H}})\right\|, \quad \forall x \in \mathbb{R}^d.$$

Under Assumption 2, i.e., Lipschitz continuity of non-faulty gradients, for each non-faulty agent j, $\left\|\nabla Q_j(x) - \nabla Q_j(x_{\mathcal{H}})\right\| \leq \mu \left\|x - x_{\mathcal{H}}\right\|$. Substituting this above implies that

$$\left\|\sum_{j \in S_1} \nabla Q_j(x) - \sum_{j \in S_1} \nabla Q_j(x_{\mathcal{H}})\right\| \leq |S_1|\, \mu \, \left\|x - x_{\mathcal{H}}\right\|. \tag{77}$$

As $|S_1| = n - 2f$, the $(2f, \epsilon)$-redundancy property defined in Definition 3 implies that for all $x_1 \in \arg\min_x \sum_{j \in S_1} Q_j(x)$,

$$\left\|x_1 - x_{\mathcal{H}}\right\| \leq \epsilon.$$

Substituting from above in (77) implies that, for all $x_1 \in \arg\min_x \sum_{j \in S_1} Q_j(x)$,

$$\left\|\sum_{j \in S_1} \nabla Q_j(x_1) - \sum_{j \in S_1} \nabla Q_j(x_{\mathcal{H}})\right\| \leq |S_1|\, \mu \, \left\|x_1 - x_{\mathcal{H}}\right\| \leq |S_1|\, \mu\epsilon. \tag{78}$$

For all $x_1 \in \arg\min_x \sum_{j \in S_1} Q_j(x)$, $\nabla Q_j(x_1) = 0$. Thus, (78) implies that

$$\left\|\sum_{j \in S_1} \nabla Q_j(x_{\mathcal{H}})\right\| \leq |S_1|\, \mu\epsilon. \tag{79}$$

Now, consider an arbitrary non-faulty agent $i \in \mathcal{H} \setminus S_1$. Let $S_2 = S_1 \cup \{i\}$. Using similar arguments as above we obtain that under the $(2f, \epsilon)$-redundancy property and Assumption 2, for all $x_2 \in \arg\min_x \sum_{j \in S_2} Q_j(x)$,

$$\left\| \sum_{j \in S_2} \nabla Q_j(x_{\mathcal{H}}) \right\| = \left\| \sum_{j \in S_2} \nabla Q_j(x_2) - \sum_{j \in S_2} \nabla Q_j(x_{\mathcal{H}}) \right\| \leq |S_2| \, \mu\epsilon. \tag{80}$$

Note that $\sum_{j \in S_2} \nabla Q_j(x) = \sum_{j \in S_1} \nabla Q_j(x) + \nabla Q_i(x)$. From triangle inequality,

$$\|\nabla Q_i(x_{\mathcal{H}})\| - \left\| \sum_{j \in S_1} \nabla Q_j(x_{\mathcal{H}}) \right\| \leq \left\| \sum_{j \in S_1} \nabla Q_j(x_{\mathcal{H}}) + \nabla Q_i(x_{\mathcal{H}}) \right\|. \tag{81}$$

Therefore, for each non-faulty agent $i \in \mathcal{H}$,

$$\begin{aligned} \|\nabla Q_i(x_{\mathcal{H}})\| &\leq \left\| \sum_{j \in S_1} \nabla Q_j(x_{\mathcal{H}}) + \nabla Q_i(x_{\mathcal{H}}) \right\| + \left\| \sum_{j \in S_1} \nabla Q_j(x_{\mathcal{H}}) \right\| \leq |S_2| \, \mu\epsilon + |S_1| \, \mu\epsilon \\ &= (n - 2f + 1)\mu\epsilon + (n - 2f)\mu\epsilon = (2n - 4f + 1)\mu\epsilon. \end{aligned} \tag{82}$$

Now, for all x and $i \in \mathcal{H}$, by Assumption 2,

$$\|\nabla Q_i(x) - \nabla Q_i(x_{\mathcal{H}})\| \leq \mu \|x - x_{\mathcal{H}}\|.$$

By triangle inequality,

$$\|\nabla Q_i(x)\| \leq \|\nabla Q_i(x_{\mathcal{H}})\| + \mu \|x - x_{\mathcal{H}}\|.$$

Substituting from (82) above we obtain that

$$\|\nabla Q_i(x)\| \leq (2n - 4f + 1)\mu\epsilon + \mu \|x - x_{\mathcal{H}}\| \leq 2n\mu\epsilon + \mu \|x - x_{\mathcal{H}}\|. \tag{83}$$

We use the above inequality (83) to show below that $\left\| \sum_{j=1}^{n-f} g^t_{i_j} \right\|$ is bounded for all t. Recall that for each iteration t,

$$\left\|g^t_{i_1}\right\| \leq ... \leq \left\|g^t_{i_{n-f}}\right\| \leq \left\|g^t_{i_{n-f+1}}\right\| \leq ... \leq \left\|g^t_{i_n}\right\|.$$

As there are at most f Byzantine agents, for each t there exists $\sigma_t \in \mathcal{H}$ such that

$$\left\|g^t_{i_{n-f}}\right\| \leq \left\|g^t_{i_{\sigma_t}}\right\|. \tag{84}$$

As $g^t_j = \nabla Q_j(x^t)$ for all $j \in \mathcal{H}$, from (84) we obtain that

$$\left\|g^t_{i_j}\right\| \leq \left\|\nabla Q_{\sigma_t}(x^t)\right\|, \quad \forall j \in \{1, \ldots, n-f\}, \; t.$$

Substituting from (83) above we obtain that for every $j \in \{1, \ldots, n-f\}$,

$$\left\|g^t_{i_j}\right\| \leq \left\|g^t_{i_{n-f}}\right\| \leq 2n\mu\epsilon + \mu \left\|x^t - x_{\mathcal{H}}\right\|.$$

Therefore, from triangle inequality,

$$\left\|\sum_{j=1}^{n-f} g_{i_j}^t\right\| \leq \sum_{j=1}^{n-f} \left\|g_{i_j}^t\right\| \leq (n-f)\left(2n\mu\epsilon + \mu\left\|x^t - x_{\mathcal{H}}\right\|\right). \tag{85}$$

Recall from (75) that $x_{\mathcal{H}} \in \mathcal{W}$. Let $\Gamma = \max_{x \in \mathcal{W}} \|x - x_{\mathcal{H}}\|$. As $\mathcal{W}$ is a compact set, $\Gamma < \infty$. Recall from the update rule (22) that $x^t \in \mathcal{W}$ for all t. Thus, $\left\|x^t - x_{\mathcal{H}}\right\| \leq \max_{x \in \mathcal{W}} \|x - x_{\mathcal{H}}\| = \Gamma < \infty$. Substituting this in (85) implies that

$$\left\|\sum_{j=1}^{n-f} g_{i_j}^t\right\| \leq (n-f)\left(2n\mu\epsilon + \mu\Gamma\right) < \infty. \tag{86}$$

Recall that in this particular case, $\sum_{j=1}^{n-f} g_{i_j}^t = \mathsf{GradFilter}\left(g_1^t, \ldots, g_n^t\right)$ (see (76)). Therefore, from above we obtain that

$$\left\|\mathsf{GradFilter}\left(g_1^t, \ldots, g_n^t\right)\right\| < \infty, \quad \forall t. \tag{87}$$

Next, we show that for an arbitrary $\delta > 0$ there exists $\xi > 0$ such that

$$\phi_t \triangleq \left\langle x^t - x_{\mathcal{H}}, \sum_{j=1}^{n-f} g_{i_j}^t \right\rangle \geq \xi \;\text{ when }\; \left\|x^t - x_{\mathcal{H}}\right\| \geq \mathsf{D}\,\epsilon + \delta.$$

Consider an arbitrary iteration t. Note that, as $|\mathcal{H}| = n - f$, there are at least $n - 2f$ agents that are common to both sets $\mathcal{H}$ and $\{i_1, ..., i_{n-f}\}$. We let $\mathcal{H}^t = \{i_1, ..., i_{n-f}\} \cap \mathcal{H}$. The remaining set of agents $\mathcal{B}^t = \{i_1, ..., i_{n-f}\} \setminus \mathcal{H}^t$ comprises of only faulty agents. Note that $\left|\mathcal{H}^t\right| \geq n - 2f$ and $\left|\mathcal{B}^t\right| \leq f$. Therefore,

$$\phi_t = \left\langle x^t - x_{\mathcal{H}}, \sum_{j \in \mathcal{H}^t} g_j^t \right\rangle + \left\langle x^t - x_{\mathcal{H}}, \sum_{k \in \mathcal{B}^t} g_k^t \right\rangle. \tag{88}$$

Consider the first term in the right-hand side of (88). Note that

$$\begin{aligned}\left\langle x^t - x_{\mathcal{H}}, \sum_{j \in \mathcal{H}^t} g_j^t \right\rangle &= \left\langle x^t - x_{\mathcal{H}}, \sum_{j \in \mathcal{H}^t} g_j^t + \sum_{j \in \mathcal{H} \setminus \mathcal{H}^t} g_j^t - \sum_{j \in \mathcal{H} \setminus \mathcal{H}^t} g_j^t \right\rangle \\ &= \left\langle x^t - x_{\mathcal{H}}, \sum_{j \in \mathcal{H}} g_j^t \right\rangle - \left\langle x^t - x_{\mathcal{H}}, \sum_{j \in \mathcal{H} \setminus \mathcal{H}^t} g_j^t \right\rangle.\end{aligned}$$

Recall that $g_j^t = \nabla Q_j(x^t)$, $\forall j \in \mathcal{H}$. Therefore,

$$\left\langle x^t - x_{\mathcal{H}}, \sum_{j \in \mathcal{H}^t} g_j^t \right\rangle = \left\langle x^t - x_{\mathcal{H}}, \sum_{j \in \mathcal{H}} \nabla Q_j(x^t) \right\rangle - \left\langle x^t - x_{\mathcal{H}}, \sum_{j \in \mathcal{H} \setminus \mathcal{H}^t} \nabla Q_j(x^t) \right\rangle. \tag{89}$$

Due to the strong convexity assumption (i.e., Assumption 3), for all $x,\, y \in \mathbb{R}^d$,

$$\left\langle x - y,\, \nabla \sum_{j \in \mathcal{H}} Q_j(x) - \nabla \sum_{j \in \mathcal{H}} Q_j(y) \right\rangle \geq \left| \mathcal{H} \right| \gamma \left\| x - y \right\|^2 .$$

As $x_{\mathcal{H}}$ is minimum point of $\sum_{j \in \mathcal{H}} Q_j(x)$, $\nabla \sum_{j \in \mathcal{H}} Q_j(x_{\mathcal{H}}) = 0$. Thus,

$$\begin{aligned} \left\langle x^t - x_{\mathcal{H}},\, \sum_{j \in \mathcal{H}} \nabla Q_j(x^t) \right\rangle &= \left\langle x^t - x_{\mathcal{H}},\, \nabla \sum_{j \in \mathcal{H}} Q_j(x^t) - \nabla \sum_{j \in \mathcal{H}} Q_j(x_{\mathcal{H}}) \right\rangle \\ &\geq \left| \mathcal{H} \right| \gamma \left\| x^t - x_{\mathcal{H}} \right\|^2 . \end{aligned} \tag{90}$$

Now, due to the Cauchy-Schwartz inequality,

$$\begin{aligned} \left\langle x^t - x_{\mathcal{H}},\, \sum_{j \in \mathcal{H} \backslash \mathcal{H}^t} \nabla Q_j(x^t) \right\rangle &= \sum_{j \in \mathcal{H} \backslash \mathcal{H}^t} \left\langle x^t - x_{\mathcal{H}},\, \nabla Q_j(x^t) \right\rangle \\ &\leq \sum_{j \in \mathcal{H} \backslash \mathcal{H}^t} \left\| x^t - x_{\mathcal{H}} \right\| \left\| \nabla Q_j(x^t) \right\| . \end{aligned} \tag{91}$$

Substituting from (90) and (91) in (89) we obtain that

$$\left\langle x^t - x_{\mathcal{H}},\, \sum_{j \in \mathcal{H}^t} g_j^t \right\rangle \geq \gamma \left| \mathcal{H} \right| \left\| x^t - x_{\mathcal{H}} \right\|^2 - \sum_{j \in \mathcal{H} \backslash \mathcal{H}^t} \left\| x^t - x_{\mathcal{H}} \right\| \left\| \nabla Q_j(x^t) \right\| . \tag{92}$$

Next, we consider the second term in the right-hand side of (88). From the Cauchy-Schwartz inequality,

$$\left\langle x^t - x_{\mathcal{H}},\, g_k^t \right\rangle \geq - \left\| x^t - x_{\mathcal{H}} \right\| \left\| g_k^t \right\| .$$

Substituting from (92) and above in (88) we obtain that

$$\phi_t \geq \gamma \left| \mathcal{H} \right| \left\| x^t - x_{\mathcal{H}} \right\|^2 - \sum_{j \in \mathcal{H} \backslash \mathcal{H}^t} \left\| x^t - x_{\mathcal{H}} \right\| \left\| \nabla Q_j(x^t) \right\| - \sum_{k \in \mathcal{B}^t} \left\| x^t - x_{\mathcal{H}} \right\| \left\| g_k^t \right\| . \tag{93}$$

Recall that, due to the sorting of the gradients, for an arbitrary $k \in \mathcal{B}^t$ and an arbitrary $j \in \mathcal{H} \backslash \mathcal{H}^t$,

$$\left\| g_k^t \right\| \leq \left\| g_j^t \right\| = \left\| \nabla Q_j(x^t) \right\| . \tag{94}$$

Recall that $\mathcal{B}^t = \{ i_1, \ldots, i_{n-f} \} \backslash \mathcal{H}^t$. Thus, $\left| \mathcal{B}^t \right| = n - f - \left| \mathcal{H}^t \right|$. Also, as $\left| \mathcal{H} \right| = n - f$, $\left| \mathcal{H} \backslash \mathcal{H}^t \right| = n - f - \left| \mathcal{H}^t \right|$. That is, $\left| \mathcal{B}^t \right| = \left| \mathcal{H} \backslash \mathcal{H}^t \right|$. Therefore, (94) implies that

$$\sum_{k \in \mathcal{B}^t} \left\| g_k^t \right\| \leq \sum_{j \in \mathcal{H} \backslash \mathcal{H}^t} \left\| \nabla Q_j(x^t) \right\| .$$

Substituting from above in (93), we obtain that

$$\phi_t \geq \gamma \left| \mathcal{H} \right| \left\| x^t - x_{\mathcal{H}} \right\|^2 - 2 \sum_{j \in \mathcal{H} \backslash \mathcal{H}^t} \left\| x^t - x_{\mathcal{H}} \right\| \left\| \nabla Q_j(x^t) \right\| .$$

Substituting from (83), i.e., $\|\nabla Q_i(x)\| \leq 2n\mu\epsilon + \mu \|x - x_{\mathcal{H}}\|$, above we obtain that

$$\begin{aligned}
\phi_t &\geq \gamma \left|\mathcal{H}\right| \left\|x^t - x_{\mathcal{H}}\right\|^2 - 2\left|\mathcal{H}\backslash\mathcal{H}^t\right| \left\|x^t - x_{\mathcal{H}}\right\| \left(2n\mu\epsilon + \mu \left\|x^t - x_{\mathcal{H}}\right\|\right) \\
&\geq \left(\gamma\left|\mathcal{H}\right| - 2\mu\left|\mathcal{H}\backslash\mathcal{H}^t\right|\right) \left\|x^t - x_{\mathcal{H}}\right\|^2 - 4n\mu\epsilon\left|\mathcal{H}\backslash\mathcal{H}^t\right| \left\|x^t - x_{\mathcal{H}}\right\|.
\end{aligned}$$

As $|\mathcal{H}| = n - f$ and $\left|\mathcal{H}\backslash\mathcal{H}^t\right| \leq f$, the above implies that

$$\begin{aligned}
\phi_t &\geq (\gamma(n-f) - 2\mu f) \left\|x^t - x_{\mathcal{H}}\right\|^2 - 4n\mu\epsilon f \left\|x^t - x_{\mathcal{H}}\right\| \\
&= (\gamma(n-f) - 2\mu f) \left\|x^t - x_{\mathcal{H}}\right\| \left(\left\|x^t - x_{\mathcal{H}}\right\| - \frac{4n\mu\epsilon f}{\gamma(n-f) - 2\mu f}\right) \\
&= n\gamma\left(1 - \frac{f}{n}\left(1 + \frac{2\mu}{\gamma}\right)\right) \left\|x^t - x_{\mathcal{H}}\right\| \left(\left\|x^t - x_{\mathcal{H}}\right\| - \frac{4\mu f\,\epsilon}{\gamma\left(1 - \frac{f}{n}\left(1 + \frac{2\mu}{\gamma}\right)\right)}\right).
\end{aligned} \tag{95}$$

Recall from (25) and (26), respectively, that

$$\alpha = 1 - \frac{f}{n}\left(1 + \frac{2\mu}{\gamma}\right) \quad \text{and} \quad \mathsf{D} = \frac{4\mu f}{\alpha\gamma}.$$

Substituting from above in (95) we obtain that

$$\phi_t \geq \alpha\, n\gamma \left\|x^t - x_{\mathcal{H}}\right\| \left(\left\|x^t - x_{\mathcal{H}}\right\| - \mathsf{D}\,\epsilon\right). \tag{96}$$

As it is assumed that $\alpha > 0$, (96) implies that for an arbitrary $\delta > 0$,

$$\phi_t \geq \alpha\, n\gamma\, \delta(\mathsf{D}\epsilon + \delta) \quad \text{when} \quad \left\|x^t - x_{\mathcal{H}}\right\| \geq \mathsf{D}\epsilon + \delta.$$

The above satisfies condition (23) in Theorem 3 with $\mathsf{D}^* = \mathsf{D}\epsilon + \delta$ and $\xi = \alpha\, n\gamma\, \delta(\mathsf{D}\epsilon + \delta)$. Recall from (87) that $\left\|\mathsf{GradFilter}\left(g_1^t, \ldots,\, g_n^t\right)\right\| < \infty,\ \forall t$. Therefore, upon using Theorem 3 we obtain that

$$\lim_{t\to\infty} \left\|x^t - x_{\mathcal{H}}\right\| \leq \mathsf{D}\,\epsilon + \delta.$$

Note that the above inequality holds true for an arbitrary $\delta > 0$. Therefore,

$$\lim_{t\to\infty} \left\|x^t - x_{\mathcal{H}}\right\| \leq \mathsf{D}\,\epsilon + \delta, \quad \forall \delta > 0. \tag{97}$$

Reasoning by contraction, (97) implies that $\lim_{t\to\infty} \left\|x^t - x_{\mathcal{H}}\right\| \leq \mathsf{D}\,\epsilon$.

E Appendix: Proof of Theorem 5

In this section we present the proof of Theorem 5. Throughout this section we assume $f > 0$ to ignore the trivial case of $f = 0$. The proof closely follows that of Theorem 4, and we may borrow some notation and results directly from Appendix D.

Consider an arbitrary set $\mathcal{H}$ of non-faulty agents with $|\mathcal{H}| = n - f$. Recall that under Assumption 3 the minimum point of the aggregate cost function $\sum_{i\in\mathcal{H}} Q_i(x)$, denoted by $x_{\mathcal{H}}$, is unique. Specifically,

$$x_{\mathcal{H}} = \arg\min_{x\in\mathbb{R}^d} \sum_{i\in\mathcal{H}} Q_i(x).$$

To prove the theorem we make use of the result stated in Theorem 3. Specifically, we show that the CWTM gradient-filter satisfies the conditions of Theorem 3 for $\mathsf{D}^* = 4\lambda\mu f\,\epsilon$, and $x^* = x_{\mathcal{H}}$. The rest follows easily from the convergence result stated in Theorem 3. Recall from (28) that for CWTM gradient-filter, for all $l \in \{1, \ldots, d\}$,

$$\mathsf{GradFilter}\left(g_1^t, \ldots, g_n^t\right)[l] = \frac{1}{n-2f} \sum_{j=f+1}^{n-f} g^t_{i_j[l]}[l], \quad \forall t. \tag{98}$$

First, we show that $\sum_{j=f+1}^{n-f} g^t_{i_j[l]}[l]$ is finite for all l and t.

From (83) in Appendix D, we know that under $(2f,\, \epsilon)$-redundancy property and Assumption 2, for each non-faulty agent $i \in \mathcal{H}$,

$$\left\|\nabla Q_i(x)\right\| \leq 2n\mu\epsilon + \mu \left\|x - x_{\mathcal{H}}\right\|. \tag{99}$$

The above implies that for all $i \in \mathcal{H}$, $l \in \{1, \ldots, d\}$ and x,

$$\left|\nabla Q_i(x)[l]\right| \leq 2n\mu\epsilon + \mu \left\|x - x_{\mathcal{H}}\right\|. \tag{100}$$

Recall that for all l and t,

$$g^t_{i_1[l]}[l] \leq \ldots \leq g^t_{i_{f+1}[l]}[l] \leq \ldots \leq g^t_{i_{n-f}[l]}[l] \leq \ldots \leq g^t_{i_n[l]}[l].$$

As there are at most f Byzantine agents and $|\mathcal{H}| = n - f$, for all l and t there exists a pair of non-faulty agents $\sigma_t^1[l]$, $\sigma_t^2[l] \in \mathcal{H}$ such that

$$g^t_{i_{n-f}[l]}[l] \leq g^t_{i_{\sigma_t^1[l]}}[l], \text{ and } g^t_{i_{f+1}[l]}[l] \geq g^t_{i_{\sigma_t^2[l]}}[l]. \tag{101}$$

As $g_j^t = \nabla Q_j(x^t)$ for all $j \in \mathcal{H}$, from (101) we obtain that for all $j \in \{f+1, \ldots, n-f\}$, l and t,

$$\left|g^t_{i_j[l]}[l]\right| \leq \max\left\{\left|\nabla Q_{\sigma_t^1[l]}(x^t)[l]\right|, \left|\nabla Q_{\sigma_t^1[l]}(x^t)[l]\right|\right\}.$$

Substituting from (100) above we obtain that for all $j \in \{f+1, \ldots, n-f\}$, l and t,

$$\left|g^t_{i_j[l]}[l]\right| \leq 2n\mu\epsilon + \mu \left\|x^t - x_{\mathcal{H}}\right\|.$$

Therefore, owing to the triangle inequality,

$$\left|\sum_{j=f+1}^{n-f} g^t_{i_j[l]}[l]\right| \leq \sum_{j=f+1}^{n-2f} \left|g^t_{i_j[l]}[l]\right| \leq (n-f)\left(2n\mu\epsilon + \mu\left\|x^t - x_{\mathcal{H}}\right\|\right). \tag{102}$$

Let $\Gamma = \max_{x \in \mathcal{W}} \left\|x - x_{\mathcal{H}}\right\|$. As $\mathcal{W}$ is a compact set, $\Gamma < \infty$. Recall from the update rule (22) that $x^t \in \mathcal{W}$ for all t. Thus, $\left\|x^t - x_{\mathcal{H}}\right\| \leq \max_{x \in \mathcal{W}} \left\|x - x_{\mathcal{H}}\right\| = \Gamma < \infty$. Substituting this in (102) implies that for all $l \in \{1, \ldots, d\}$,

$$\left|\sum_{j=f+1}^{n-f} g^t_{i_j[l]}[l]\right| \leq (n-2f)\left(2n\mu\epsilon + \mu\Gamma\right) < \infty.$$

Substituting from above in (98) we obtain that

$$\left\|\mathsf{GradFilter}\left(g_1^t, \ldots, g_n^t\right)\right\| < \infty, \quad \forall t. \tag{103}$$

Now, consider an arbitrary iteration t and $l \in \{1, \ldots, d\}$. From prior works on CWTM gradient-filter for the scalar case [32], i.e., when $d = 1$, we know that trimmed mean of the l-th elements of the gradients lies in the convex hull of l-th elements of the non-faulty agents' gradients in set $\mathcal{H}$. Specifically,

$$\min_{i \in \mathcal{H}} g_i^t[l] \leq \mathsf{GradFilter}\left(g_1^t, \ldots, g_n^t\right)[l] \leq \max_{i \in \mathcal{H}} g_i^t[l]. \tag{104}$$

Obviously,

$$\min_{i \in \mathcal{H}} g_i^t[l] \leq \frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} g_i^t[l] \leq \max_{i \in \mathcal{H}} g_i^t[l]. \tag{105}$$

Therefore, from (104) and (105) we obtain that

$$\left|\mathsf{GradFilter}\left(g_1^t, \ldots, g_n^t\right)[l] - \frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} g_i^t[l]\right| \leq \max_{i \in \mathcal{H}} g_i^t[l] - \min_{i \in \mathcal{H}} g_i^t[l].$$

As $\max_{i \in \mathcal{H}} g_i^t[l] - \min_{i \in \mathcal{H}} g_i^t[l] = \max_{i,\, j \in \mathcal{H}} \left|g_i^t[l] - g_j^t[l]\right|$, and $g_i^t = \nabla Q_i(x^t)$ for all $i \in \mathcal{H}$, the above can be re-written as follows.

$$\left|\mathsf{GradFilter}\left(g_1^t, \ldots, g_n^t\right)[l] - \frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} \nabla Q_i(x^t)[l]\right| \leq \max_{i,\, j \in \mathcal{H}} \left|\nabla Q_i(x^t)[l] - \nabla Q_j(x^t)[l]\right|. \tag{106}$$

Note that for any two $i,\, j \in \mathcal{H}$,

$$\left|\nabla Q_i(x^t)[l] - \nabla Q_j(x^t)[l]\right| \leq \left\|\nabla Q_i(x^t) - \nabla Q_j(x^t)\right\|. \tag{107}$$

Substituting from Assumption 5, $\|\nabla Q_i(x) - \nabla Q_j(x)\| \leq \lambda \max\left\{\|\nabla Q_i(x)\|, \|\nabla Q_j(x)\|\right\}$, in (107) we obtain that

$$\left|\nabla Q_i(x^t)[l] - \nabla Q_j(x^t)[l]\right| \leq \lambda \max\left\{\left\|\nabla Q_i(x^t)\right\|, \left\|\nabla Q_j(x^t)\right\|\right\}. \tag{108}$$

Substituting from (99) above we obtain that

$$\left|\nabla Q_i(x^t)[l] - \nabla Q_j(x^t)[l]\right| \leq \lambda \left(2n\mu\, \epsilon + \mu \left\|x^t - x_{\mathcal{H}}\right\|\right). \tag{109}$$

Finally, substituting from (109) in (106) we obtain that, for all l,

$$\left|\mathsf{GradFilter}\left(g_1^t, \ldots, g_n^t\right)[l] - \frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} \nabla Q_i(x^t)[l]\right| \leq \lambda \left(2n\mu\, \epsilon + \mu \left\|x^t - x_{\mathcal{H}}\right\|\right).$$

As $\|x\| = \sqrt{\sum_{l=1}^{d} |x[l]|^2}$ for $x \in \mathbb{R}^d$, the above implies that

$$\left\|\mathsf{GradFilter}\left(g_1^t, \ldots, g_n^t\right) - \frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} \nabla Q_i(x^t)\right\| \leq \sqrt{d}\lambda \left(2n\mu\, \epsilon + \mu \left\|x^t - x_{\mathcal{H}}\right\|\right). \tag{110}$$

Now, note that

$$\mathsf{GradFilter}\left(g_1^t, \ldots, g_n^t\right) = \frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} \nabla Q_i(x^t) + \left(\mathsf{GradFilter}\left(g_1^t, \ldots, g_n^t\right) - \frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} \nabla Q_i(x^t)\right). \tag{111}$$

Recall from Theorem 3 that ϕ_t, for each t, is defined to be

$$\phi_t = \left\langle x^t - x_{\mathcal{H}}, \, \mathsf{GradFilter}\left(g_1^t, \ldots, g_n^t\right)\right\rangle.$$

Substituting from (111) above we obtain that

$$\phi_t = \left\langle x^t - x_{\mathcal{H}}, \, \frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} \nabla Q_i(x^t)\right\rangle + \left\langle x^t - x_{\mathcal{H}}, \, \mathsf{GradFilter}\left(g_1^t, \ldots, g_n^t\right) - \frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} \nabla Q_i(x^t)\right\rangle. \tag{112}$$

Recall from Assumption 3 that $Q_{\mathcal{H}}(x) = (1/|\mathcal{H}|) \sum_{i \in \mathcal{H}} Q_i(x)$. Thus, the first term on the right-hand side of (112),

$$\left\langle x^t - x_{\mathcal{H}}, \, \frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} \nabla Q_i(x^t)\right\rangle = \left\langle x^t - x_{\mathcal{H}}, \, \nabla Q_{\mathcal{H}}(x^t)\right\rangle.$$

Substituting from the Assumption 3 above, and recalling that $\nabla Q_{\mathcal{H}}(x_{\mathcal{H}}) = 0$, we obtain that

$$\left\langle x^t - x_{\mathcal{H}}, \, \frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} \nabla Q_i(x^t)\right\rangle \geq \gamma \left\|x^t - x_{\mathcal{H}}\right\|^2. \tag{113}$$

Next, we consider the second term on the right-hand side of (112). From Cauchy-Schwartz inequality,

$$\left\langle x^t - x_{\mathcal{H}}, \, \mathsf{GradFilter}\left(g_1^t, \ldots, g_n^t\right) - \frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} \nabla Q_i(x^t)\right\rangle \geq - \left\|x^t - x_{\mathcal{H}}\right\| \left\|\mathsf{GradFilter}\left(g_1^t, \ldots, g_n^t\right) - \frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} \nabla Q_i(x^t)\right\|.$$

Substituting from (110) above we obtain that

$$\left\langle x^t - x_{\mathcal{H}}, \, \mathsf{GradFilter}\left(g_1^t, \ldots, g_n^t\right) - \frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} \nabla Q_i(x^t)\right\rangle \geq -\sqrt{d}\lambda \left\|x^t - x_{\mathcal{H}}\right\| \left(2n\mu \, \epsilon + \mu \left\|x^t - x_{\mathcal{H}}\right\|\right). \tag{114}$$

Substituting from (113) and (114) in (112) we obtain that

$$\begin{aligned} \phi_t &\geq \gamma \left\|x^t - x_{\mathcal{H}}\right\|^2 - \sqrt{d}\lambda \left\|x^t - x_{\mathcal{H}}\right\| \left(2n\mu \, \epsilon + \mu \left\|x^t - x_{\mathcal{H}}\right\|\right) \\ &= \left(\gamma - \sqrt{d}\lambda\mu\right) \left\|x^t - x_{\mathcal{H}}\right\| \left(\left\|x^t - x_{\mathcal{H}}\right\| - \frac{2\sqrt{d}n\mu\lambda}{(\gamma - \sqrt{d}\mu\lambda)} \, \epsilon\right). \end{aligned} \tag{115}$$

The above inequality is similar to (96) in the proof of Theorem 4 in Appendix D where by assumption $\gamma - \sqrt{d}\lambda\mu > 0$. Also, recall from (103) that in this particular case of CWTM gradient-filter $\left\|\mathsf{GradFilter}\left(g_1^t, \ldots, g_n^t\right)\right\| < \infty, \ \forall t$. Therefore, using similar arguments as in Appendix D, we obtain that

$$\lim_{t \to \infty} \left\|x^t - x_{\mathcal{H}}\right\| \leq \frac{2\sqrt{d}n\mu\lambda}{(\gamma - \sqrt{d}\mu\lambda)} \, \epsilon = \left(\frac{2n\lambda}{(\gamma/\mu\sqrt{d}) - \lambda}\right) \epsilon.$$