

Fast and Robust Distributed Learning in High Dimension

El-Mahdi El-Mhamdi
EPFL
elmahdi.elmhamdi@epfl.ch

Rachid Guerraoui
EPFL
rachid.guerraoui@epfl.ch

Sébastien Rouault
EPFL
sebastien.rouault@epfl.ch

Abstract—Could a gradient aggregation rule (GAR) for distributed machine learning be both robust and fast? This paper answers by the affirmative through MULTI-BULYAN. Given n workers, f of which are arbitrary malicious (Byzantine) and $m = n - f$ are not, we prove that MULTI-BULYAN can ensure a strong form of Byzantine resilience, as well as an $\frac{m}{n}$ slowdown, compared to averaging, the fastest (but non Byzantine resilient) rule for distributed machine learning. When $m \approx n$ (almost all workers are correct), MULTI-BULYAN reaches the speed of averaging. We also prove that MULTI-BULYAN’s cost in local computation is $O(d)$ (like averaging), an important feature for ML where d commonly reaches 10^9 , while robust alternatives have at least quadratic cost in d .

Our theoretical findings are complemented with an experimental evaluation which, in addition to supporting the linear $O(d)$ complexity argument, conveys the fact that MULTI-BULYAN’s parallelisability further adds to its efficiency.

Index Terms—Distributed Systems, Byzantine Resilience, Machine Learning, Stochastic Gradient Descent, High Dimension, Non-Convex Optimization

I. INTRODUCTION

The ongoing data deluge has been both a blessing and a burden for machine learning system designers. It was a blessing since machine learning provably performs better with more training data [21], and a burden since the quantity of available data is huge. The set of parameters in machine learning is now in the order of a gigabyte [10], while training data is several orders of magnitude beyond that [10] and rarely available in the same location. In short, distributed machine learning is not an option anymore; it is the only way to deliver results in a reasonable time for the user.

At the core of the recent success in machine learning lies Gradient Descent (GD). GD is an algorithm that consists in the very simple idea of updating a parameter in the opposite direction of the gradient of a cost function. The parameter tunes the algorithm being trained, and the cost function typically represents how bad the actual parameter performs. Stochastic Gradient Descent (SGD), a lightweight version of GD, is the workhorse of today’s machine learning. When all involved machines are reliable, SGD can be easily distributed. The general recipe is that several machines (called *workers*) carry the gradient computation, and a *server* aggregates the resulting gradients in order to update the parameter. This

setting is now widely called the *parameter-server* setting [15] and is the dominant standard in distributed machine learning.

In the mainstream declination of this parameter-server setting [1], [10], [15], [17], the *gradient aggregation rule* (GAR) consists mostly in averaging the received gradients. When none of the workers misbehaves (no Byzantine workers) and the system is synchronous, it can be easily proven that averaging the gradients is optimal [7]. Indeed, averaging (or problem-specific variants of it [32]) requires less steps to train the model by optimizing the use of multiple workers. However, as was recently shown [6], [8], [27], averaging is brittle to the most simple form of adversarial behavior, requiring only one single machine to behave in a malicious manner.

Among the proposed alternatives to averaging, Krum [6], [11] attracted significant attention and has been used as a benchmark in most of the body of work on Byzantine-resilient gradient descent (e.g. [3], [4], [8], [18], [19], [23], [27]–[30]). Unlike averaging, which incorporates the sum of all proposed gradients (including the Byzantine ones), Krum selects the gradient that is the closest to its $\approx n - f$ neighbors¹. This ensures that the model update is performed using a gradient from a worker that behaves correctly.

A key feature of Krum is that, unlike other tools from traditional robust statistics [20], it does not suffer neither from computability nor from complexity issues. In particular, Krum can be computed in a linear time in d , the dimension² of the model being trained. However, Krum suffers from two important limitations. (1) By basing its selection on a distance measurement, Krum suffers from what is known in high-dimensional machine learning as the *curse of dimensionality*. Basically, distances become almost meaningless when the dimension of the vectors is too large. In particular, because models are high dimensional, and the landscape where they are being optimized can be highly non convex, a strong adversary

¹More specifically, it is $n - f - 2$, and the “ -2 ” comes from the fact that each correct worker knows that in its neighbors, there are only $n - f - 1$ other workers guaranteed to be correct, besides itself. For safety reasons, the worker exclude 1 worker from its $n - f - 1$ neighbourhood, so that the distance to the remaining $n - f - 2$ that are selected is guaranteed to be upper bounded by another *correct* worker.

²This parameter can attain values as high as 10^9 or 10^{11} , thus making any solution from the usual security and fraud detection toolbox unpractical. The latter tends to have a quadratic or even super-quadratic cost as they mostly rely on Principal Component Analysis (PCA) as a key ingredient [24].

can make a Byzantine resilient GAR converge (as it is proven to do), but to a highly non desirable local minima. Specifically, the adversary benefits from a margin of \sqrt{d} to attack distance-based Byzantine resilient GARs, and make them converge to a bad local minima. (2) By selecting only the best behaving gradient (the closest one to its neighbors), Krum ends up using only one gradient for the update, and does not benefit from the variance reduction that comes from leveraging many workers as in averaging. Potentially, Krum can have a slowdown of up to $\frac{1}{n}$ compared to using all gradients, when none are Byzantine. (This slowdown is computed as the number of steps it takes averaging to converge, divided by the number of steps it takes Krum to converge.) Given that f workers are Byzantine, it is understandable that any Byzantine-resilient GAR would have some slowdown compared to averaging, as it keeps supposing that not all n workers are reliable. However, $\frac{1}{n}$ is still too far from a desirable $\frac{n-f}{n}$ (using all $n - f$ non Byzantine workers).

We ask the question, whether we could devise a gradient aggregation rule for distributed SGD that is both robust and fast? We answer by the affirmative through MULTI-BULYAN, which we prove circumvents the two aforementioned limitations.

MULTI-BULYAN builds on previous work. The intuition behind MULTI-BULYAN (combining MULTI-KRUM and BULYAN) was sketched in [9] without however any precise description or analysis. BULYAN [12] was proposed on top of Krum, and was proved to divide the attackers leeway by the required $\frac{1}{\sqrt{d}}$ that stems from the so called curse of dimensionality and circumvents the first limitation. For the second limitation, an experimental variant of Krum, coined MULTI-KRUM was proposed. This does not only take the (single) closest gradient to its $\approx n - f$ neighbors, like Krum does. MULTI-KRUM goes further and selects the (multiple) $\approx n - f$ closest gradients to their $\approx n - f$ closest neighbors. However, it was not clear whether (a) MULTI-KRUM inherits the Byzantine resilience merits of Krum while having a better slowdown of $\approx \frac{n-f}{n}$ and whether (b) BULYAN can also be used on top of MULTI-KRUM to produce the same leeway reduction.

In this paper, we first define two notions of Byzantine resilience, *weak* and *strong*, in order to take into account the specificity of high dimensional non-convex optimization. We also explain some of the non-intuitive aspects of the latter, such as why "mild" adversarial behavior (noise) can sometimes accelerate learning. We then prove that MULTI-KRUM is indeed Byzantine resilient in the same sense as Krum. MULTI-KRUM guarantees convergence despite f Byzantine workers. Hence, BULYAN can be used on top of MULTI-KRUM to guarantee not only convergence, but convergence without the high-dimensional vulnerability of non-convex landscapes. (We call the obtained combination MULTI-BULYAN.)

We prove (a) that MULTI-BULYAN ensures strong Byzantine resilience and (b) that it is $\approx \frac{n-2f}{n}$ times as fast as the optimal algorithm (averaging) in the absence of Byzantine workers.

In particular, when $f \ll n$, as is the case in many practical situations [2], this slowdown tends to 1 and MULTI-BULYAN is as fast as the optimal (but non Byzantine resilient) averaging.

Paper organization. In section II, we present the algorithms being studied, as well as a toolbox of formal definitions: weak, strong, and (α, f) -Byzantine resilience. We also present a necessary context on non-convex optimization, as well as its interplay with the high dimensionality of machine learning together with the \sqrt{d} leeway it provides to strong attackers. We then introduce some background on non-convex optimization and why it changes the requirement for Byzantine resilient gradient descent. We also formalize our definitions of weak and strong Byzantine resilience (the latter was only informally sketched before). In Section III, we give our proofs of Byzantine resilience and slow down of MULTI-KRUM and in Section IV our proofs of Byzantine resilience and slow down of MULTI-BULYAN. In Section V, we evaluate MULTI-KRUM and MULTI-BULYAN on two practically important metrics, namely (1) the aggregation time achieved on actual hardware and (2) the maximum top-1 cross-accuracy reached using a given GAR. Section VI discusses how our work bridges the gap between the Byzantine faults threat model of distributed computing and the data-poisoning threat model of machine learning while proposing new directions.

Our experimental claims could be reproduced using this repository: <https://github.com/anonconfsubmit/submit-8618.git> with the password 8AS5lds_lzeczmi95ash.

II. BACKGROUND

A. Stochastic Gradient Descent

For illustration, but without loss of generality, consider a learning task that consists in making accurate predictions for the labels of each data instance ξ_i using a high dimensional model (for example, a neural network); we denote the d parameters of that model by the vector \mathbf{x} . Each data instance has a set of *features* (image pixels), and a set of labels (e.g., {cat, person}). We refer to the j^{th} coordinate of a vector \mathbf{x}_i with $x_{i,j}$. The model is trained with the popular backpropagation algorithm based on SGD. Specifically, SGD addresses the following optimization problem.

$$\min_{\mathbf{x} \in \mathcal{R}^d} Q(\mathbf{x}) \triangleq \mathbb{E}_{\xi} F(\mathbf{x}; \xi) \quad (1)$$

where ξ is a random variable representing a total of B data instances and $F(\mathbf{x}; \xi)$ is the cost function. The function $Q(\mathbf{x})$ is smooth but not convex.

SGD computes the gradient ($\mathbf{G}(\mathbf{x}, \xi) \triangleq \nabla F(\mathbf{x}; \xi)$) and then updates the model parameters (\mathbf{x}) in a direction opposite to that of the gradient (descent). The vanilla SGD update rule given a sequence of learning rates $\{\gamma_k\}$ at any given step³ is the following:

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \gamma_k \cdot \mathbf{G}(\mathbf{x}^{(k)}, \xi) \quad (2)$$

³A step denotes an update in the model parameters.

The popularity of SGD stems from its ability to employ noisy approximations of the actual gradient. In a distributed setup, SGD employs a *mini-batch* of $b < B$ training instances for the gradient computation:

$$\mathbf{G}(\mathbf{x}, \xi) = \sum_{i=1}^b \mathbf{G}(\mathbf{x}, \xi_i) \quad (3)$$

The size of the mini-batch (b) induces a trade-off between the robustness of a given update (noise in the gradient approximation) and the time required to compute this update. The mini-batch also affects the amount of parallelism (Equation 3) that modern computing clusters (multi-GPU etc.) largely benefit from. Scaling the mini-batch size to exploit additional parallelism requires however a non-trivial selection of the sequence of learning rates [13]. A very important assumption for the convergence properties of SGD is that each gradient is an unbiased estimation of the actual gradient, which is typically ensured through uniform random sampling, i.e., gradients that are on expectation equal to the actual gradient (Figure II-A).

B. Algorithms

MULTI-BULYAN relies on two algorithmic components: MULTI-KRUM [6] and BULYAN [12]. The former rule requires that $n \geq 2f + 3$ and the second requires that $n \geq 4f + 3$. MULTI-BULYAN (along with MULTI-KRUM and BULYAN used as functions) is presented in Algorithm 1.

Intuitively, the goal of MULTI-KRUM is to select the gradients that deviate less from the “majority” based on their relative distances. Given gradients $\mathbf{G}_1 \dots \mathbf{G}_n$ proposed by workers 1 to n respectively, MULTI-KRUM selects the m gradients with the smallest sum of scores (i.e., ℓ_2 norm from the other gradients) as follows:

$$(m) \arg \min_{i \in \{1, \dots, n\}} \sum_{i \rightarrow j} \|\mathbf{G}_i - \mathbf{G}_j\|^2 \quad (4)$$

where given a function $X(i)$, $(m) \arg \min(X(i))$ denotes the indexes i with the m smallest $X(i)$ values, and $i \rightarrow j$ means that \mathbf{G}_j is among the $n - f - 2$ closest gradients to \mathbf{G}_i . BULYAN in turn takes the aforementioned m vectors, computes their coordinate-wise median and produces a gradient which coordinates are the average of the $m - 2f$ closest values to the median.

C. Byzantine Resilience

Intuitively, Byzantine resilience requires a *GAR* to guarantee convergence despite the presence of f Byzantine workers. It can be formally stated as follows.

Definition 1 (Weak Byzantine resilience). *We say that a GAR ensures weak f -Byzantine resilience if the sequence $\mathbf{x}^{(k)}$ (Equation 2) converges almost surely to some \mathbf{x}^* where $\nabla Q(\mathbf{x}^*) = 0$, despite the presence of f Byzantine workers.*

Algorithm 1 The Gradient Aggregation Rule (GAR) of MULTI-BULYAN.

Require: \mathcal{I} : Set of all items; \mathcal{U} : Set of all users.
Require: n : number of workers, m : minimum MULTI-KRUM size.
Parameter Server:
1: **function** MULTI-KRUM ($f, [G_1, \dots, G_k]$):
2: $S = Dict()$ \triangleright Dictionary: (key=gradient, value=score).
3: $m = k - f - 2$
4: **for** $i \in [1, \dots, k]$ **do**:
5: neighbors = $k - f - 2$ closest (ℓ_2 norm) vectors to G_i
6: $S[G_i] = \sum_{G \in neighbors} \|G_i - G\|^2$
7: $G^{winner} = S.getSmallestValues(1).getKeys()$
8: $G^{output} = Average(S.getSmallestValues(m).getKeys())$
9: **return** G^{winner}, G^{output}
10: **end function**
11:
12: **function** MULTI-BULYAN ($f, [G_1, \dots, G_n]$):
13: $\theta = n - 2f - 2$
14: $\beta = \theta - 2f$
15: $G^{ext} = Array[\theta][d]$ \triangleright extracted gradients.
16: $G^{agr} = Array[\theta][d]$ \triangleright aggregated gradients.
17: $M = Array[d]$ \triangleright coordinate-wise medians.
18: $C = Array[\beta][d]$
19: **for** $i \in [0, \dots, \theta]$ **do**:
20: $G^{ext}[i], G^{agr}[i] = MULTI-KRUM(f, [G_1, \dots, G_n] \setminus G^{ext})$
21: $M = Median(G^{ext})$
22: **for** $j \in [0, \dots, d - 1]$ **do**:
23: $C[:,j] = Argpartition(\|G^{agr}[:,j] - M[j]\|, \beta) \triangleright \beta$ closest coordinates to the median.
return $Average(G^{agr}[C])$
24: **end function**
25:
26: **function** GAR($f, [G_1, \dots, G_n]$):
return MULTI-BULYAN ($f, [G_1, \dots, G_n]$)
27: **end function**

On the other hand, strong Byzantine resilience requires that this convergence does not lead to “bad” optimums, and is related to more intricate problem of non-convex optimization, which, in the presence of Byzantine workers, is highly aggravated by the dimension of the problem as explained in what follows.

a) Specificity of non-convex optimization.

Non-convex optimization is one of the earliest established NP-hard problems [14]. In fact, many interesting but hard questions in machine learning have their source in the non-convexity of the cost function.

In distributed machine learning, the non-convexity of the cost function creates two non-intuitive behaviours that are important to highlight.

(1) A “mild” Byzantine worker can make the system converge faster. For instance, it has been reported several times in the literature that noise accelerates learning [7], [14]. This can be understood from the “S” (stochasticity) of SGD: as (correct) workers cannot have a full picture of the surrounding landscape of the cost function, they can only draw a sample at random and estimate the best direction based on that sample, which can be, and is probably biased compared to the true gradient. Moreover, due to non-convexity,

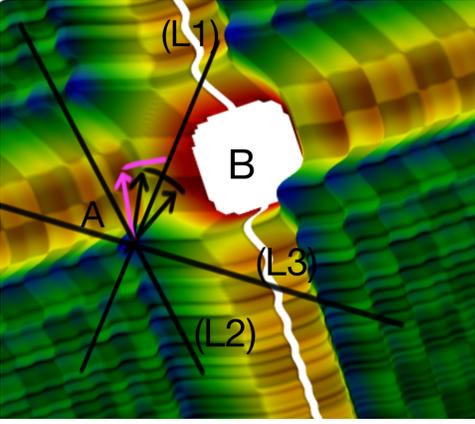


Fig. 1: In a non-convex situation, two correct vectors (black arrows) are pointing towards the deep optimum located in area B, both vectors belong to the plane formed by lines L1 and L2. A Byzantine worker (magenta) is taking benefit from the third dimension, and the non-convex landscape, to place a vector that is heading towards one of the bad local optimums of area A. This Byzantine vector is located in the plane (L1,L3). Due to the variance of the correct workers on the plane (L1,L2), the Byzantine one has a budget of about $\sqrt{3}$ times the disagreement of the correct workers, to put as a deviation towards A, on the line (L3), while still being selected by a weak Byzantine resilient *GAR*, since its projection on the plane (L1,L2) lies exactly on the line (L1), unlike that of the correct workers. In very high dimensions, the situation is amplified by \sqrt{d} .

even the true gradient might be leading to the local minima where the parameter server is. By providing a wrong direction (i.e. not the true gradient, or a correct stochastic estimation), a Byzantine worker whose resources cannot face the high-dimensional landscape of the cost function, might end up providing a direction to get out of that local minima.

(2) Combined with high dimensional issues, non-convexity explains the need for strong Byzantine resilience. Unlike the "mild" Byzantine worker, a strong adversary with more resources than the workers and the server, can see a larger picture and provide an attack that requires a stronger requirement. Namely, a requirement that would cut the \sqrt{d} leeway offered to an attacker in each dimension. Figure 1 provides an illustration.

This motivates the following formalization of strong Byzantine resilience.

Definition 2 (Strong Byzantine resilience). *We say that a GAR ensures strong f -Byzantine resilient if it ensures weak Byzantine resilience and if for every $i \in [1, d]$, there exists a correct gradient G (i.e., computed by a non-Byzantine worker) s.t. $\mathbb{E}|GAR_i - G_i| = O(\frac{1}{\sqrt{d}})$. The expectation is taken over the random samples (ξ in Equation 1) and v_i denotes the i^{th} coordinate of a vector v .*

b) *Weak vs. strong Byzantine resilience.*

To attack non-Byzantine resilient *GARs* such as averaging, it only takes the computation of an estimate of the gradient, which can be done in $O(n.d)$ operations per round by a Byzantine worker. This attack is reasonably cheap: within the usual cost of the workload of other workers, $O(d)$, and the server, $O(n.d)$.

To attack weakly Byzantine-resilient *GARs* however, one needs to find the 'most legitimate but harmful vector possible', i.e one that will (1) be selected by a weakly Byzantine-resilient *GAR*, and (2) be misleading convergence (red arrow in Figure 1). To find this vector, an attacker has to first collect every correct worker's vector (before they reach the server), and solve an optimization problem (by linear regression) to approximate this harmful but legitimate vector [12]. If the desired quality of the approximation is ϵ , the Byzantine worker would need at least $\Omega(\frac{n.d}{\epsilon})$ operation to reach it with regression. This is a tight lower bound for a regression problem in d dimensions with n vectors [14]. In practice, if the required precision is of order 10^{-9} , with 100 workers and a neural network model of dimension 10^9 , the cost of the attack becomes quickly prohibitive ($\approx 10^{20}$ operations to be done in each step by the attacker).

To summarize, weak Byzantine resilience can be enough as a practical solution against attackers whose resources are comparable to the server's. However, strong Byzantine resilience remains the only provable solution against attackers with significant resources.

For the sake of our theoretical analysis, we recall the definition of (α, f) -Byzantine resilience [6] (Definition 3). This definition is a sufficient condition (as proved in [6] based on [7]) for weak Byzantine resilience. Even-though the property of (α, f) -Byzantine resilience is a sufficient, but not a necessary condition for (weak) Byzantine resilience, it has been so far used as the defacto standard [6], [27] to guarantee (weak) Byzantine resilience for SGD. We will therefore follow this standard and require (α, f) -Byzantine resilience from any *GAR* that is plugged into MULTI-BULYAN, in particular, we will require it from MULTI-KRUM. The theoretical analysis done in [12] guarantees that BULYAN inherits it.

Intuitively, Definition 3 states that the gradient aggregation rule *GAR* produces an output vector that lives, on average (over random samples used by SGD), in the cone of angle α around the true gradient. We simply call this the "correct cone".

Definition 3 ((α, f) -Byzantine resilience (as in [6])). *Let $0 \leq \alpha < \pi/2$ be any angular value, and any integer $0 \leq f \leq n$. Let V_1, \dots, V_n be any independent identically distributed random vectors in \mathbb{R}^d , $V_i \sim G$, with $\mathbb{E}G = g$. Let B_1, \dots, B_f be any random vectors in \mathbb{R}^d , possibly dependent*

on the V_i 's. An aggregation rule GAR is said to be (α, f) -Byzantine resilient if, for any $1 \leq j_1 < \dots < j_f \leq n$, vector

$$GAR = GAR(V_1, \dots, \underbrace{B_1}_{j_1}, \dots, \underbrace{B_f}_{j_f}, \dots, V_n)$$

satisfies (i) $\langle \mathbb{E}GAR, g \rangle \geq (1 - \sin \alpha) \cdot \|g\|^2 > 0$ ⁴ and (ii) for $r = 2, 3, 4$, $\mathbb{E} \|GAR\|^r$ is bounded above by a linear combination of terms $\mathbb{E} \|G\|^{r_1} \dots \mathbb{E} \|G\|^{r_{n-1}}$ with $r_1 + \dots + r_{n-1} = r$.

c) *Choice of f .*

The properties of the existing Byzantine-resilient SGD algorithms all depend on one important parameter, i.e., the number of *potentially* Byzantine nodes f . It is important to notice that f denotes a contract between the designer of the fault-tolerant solution and the user of the solution (who implements a service on top of the solution and deploys it in a specific setting). As long as the number of Byzantine workers is less than f , the solution is safe. Fixing an optimal value for f is an orthogonal problem. For example, if daily failures in a data center are about 1%, $f = 0.01.n$ would be a suggested choice to tune the algorithm, and suffer from only a 99% slowdown.

The performance (convergence time) of certain existing Byzantine-resilient SGD algorithms in a non-Byzantine environment is independent of the choice of f . These algorithms do not exploit the full potential of the choice of f . Modern large-scale systems are versatile and often undergo important structural changes while providing online services (e.g., addition or maintenance of certain worker nodes). Intuitively, there should be a fine granularity between the level of pessimism (i.e., value of f) and the performance of the SGD algorithm in the setting with no Byzantine failures.

III. MULTI-KRUM: WEAK BYZANTINE RESILIENCE AND SLOWDOWN

Let n be any integer greater than 2, f any integer s.t $f \leq \frac{n-2}{2}$ and m an integer s.t $m \leq n - f - 2$. Let $\tilde{m} = n - f - 2$.

Before proving the strong Byzantine resilience of MULTI-BULYAN, we first prove the (α, f) -Byzantine resilience of MULTI-KRUM (Lemma 1), then prove its almost sure convergence (Lemma 2) based on that, which proves its weak Byzantine resilience (Theorem 1).

In all what follows, expectations are taken over random samples used by correct workers to estimate the gradient, i.e the "S" (stochasticity) that is inherent to SGD. It is worth noting that this analysis in expectation is not an average case analysis from the point of view of Byzantine fault tolerance. For instance, the Byzantine worker is always assumed to

⁴Having a scalar product that is lower bounded by this value guarantees that the GAR of MULTI-KRUM lives in the aforementioned cone.

follow arbitrarily bad policies and the analysis is a worst-case one.

The Byzantine resilience proof (Lemma 1) relies on the following observation: given $m \leq n - f - 2$, and in particular $m = n - f - 2$ ⁵, m -Krum averages m gradients that are all in the "correct cone", and a cone is a convex set, thus stable by averaging. The resulting vectors therefore also live in that cone. The angle of the cone will depend on a variable $\eta(n, f)$ as in [6], the value of $\eta(n, f)$ itself depends on m . This is what enables us to use m vectors as the basis of our MULTI-KRUM, unlike [6] where a restriction is made on $m = 1$. In short, the convexity of a cone allows us to prove lemma 1 almost as in [6], the same derivation applies this time to the averaged m vectors instead of the single vector chosen by Krum.

The proof of Lemma 2 (deferred to the appendix) is the same as the one in [6] which itself draws on the rather classic analysis of SGD made by L.Bottou [7]. The key concepts are (1) a global confinement of the sequence of parameter vectors and (2) a bound on the statistical moments of the random sequence of estimators built by the GAR of MULTI-KRUM. As in [6], [7], reasonable assumptions are made on the cost function Q , those assumption are not restrictive and are common in practical machine learning.

Theorem 1 (Byzantine resilience and slowdown of MULTI-KRUM). *Let m be any integer s.t. $m \leq n - f - 2$. (i) MULTI-KRUM has weak Byzantine resilience against f failures. (ii) In the absence of Byzantine workers, MULTI-KRUM has a slowdown (expressed in ratio with averaging) of $\frac{\tilde{m}}{n}$.*

Proof. Proof of (i). To prove (i), we will require Lemma 1 and Lemma 2, then conclude by construction of MULTI-KRUM as an m -Krum algorithm with $m = n - f - 2$.

Lemma 1. *Let V_1, \dots, V_n be any independent and identically distributed random d -dimensional vectors s.t $V_i \sim G$, with $\mathbb{E}G = g$ and $\mathbb{E} \|G - g\|^2 = d\sigma^2$. Let B_1, \dots, B_f be any f random vectors, possibly dependent on the V_i 's. If $2f + 2 < n$ and $\eta(n, f)\sqrt{d} \cdot \sigma < \|g\|$, where*

$$\eta(n, f) \stackrel{\text{def}}{=} \sqrt{2 \left(n - f + \frac{f \cdot m + f^2 \cdot (m + 1)}{m} \right)},$$

then the GAR function of MULTI-KRUM is (α, f) -Byzantine resilient where $0 \leq \alpha < \pi/2$ is defined by

$$\sin \alpha = \frac{\eta(n, f) \cdot \sqrt{d} \cdot \sigma}{\|g\|}.$$

Proof. Without loss of generality, we assume that the Byzantine vectors B_1, \dots, B_f occupy the last f positions in the list of arguments of MULTI-KRUM, i.e., MULTI-KRUM = MULTI-KRUM($V_1, \dots, V_{n-f}, B_1, \dots, B_f$).

⁵The slowdown question is an incentive to take the highest value of m among those that satisfy Byzantine resilience, in this case \tilde{m} .

An index is *correct* if it refers to a vector among V_1, \dots, V_{n-f} . An index is *Byzantine* if it refers to a vector among B_1, \dots, B_f . For each index (correct or Byzantine) i , we denote by $\delta_c(i)$ (resp. $\delta_b(i)$) the number of correct (resp. Byzantine) indices j such that $i \rightarrow j$ (the notation we introduced in Section 3 when defining MULTI-KRUM), i.e the number of workers, among the m neighbors of i that are correct (resp. Byzantine).

We have $\delta_c(i) + \delta_b(i) = m$, $n - 2f - 2 \leq \delta_c(i) \leq m$ and $\delta_b(i) \leq f$.

We focus first on the condition (i) of (α, f) -Byzantine resilience. We determine an upper bound on the squared distance $\|\mathbb{E}\text{MULTI-KRUM} - g\|^2$. Note that, for any correct j , $\mathbb{E}V_j = g$.

We denote by i_* the index of the worst scoring among the m vectors chosen by the MULTI-KRUM function, i.e one that ranks with the m^{th} smallest score in Equation 4.

Though we follow the same derivation of [6], one should keep in mind that the manipulated object now is not the winner, minimizing Equation 4, but the "worst (m^{th}) possible vector to choose" and prove that it also lies in the correct cone. This allows us to average it with the $m - 1$ vectors with smaller scores, and by convexity of a cone, prove that the resulting MULTI-KRUM vector is also in that cone.

$$\begin{aligned} \|\mathbb{E}\text{MULTI-KRUM} - g\|^2 &\leq \|\mathbb{E}(\text{MULTI-KRUM} \\ &\quad - \frac{1}{\delta_c(i_*)} \sum_{i_* \rightarrow \text{correct } j} V_j)\|^2 \\ \text{(Jensen inequality)} &\leq \mathbb{E} \|\text{MULTI-KRUM} \\ &\quad - \frac{1}{\delta_c(i_*)} \sum_{i_* \rightarrow \text{correct } j} V_j\|^2 \\ &\leq \sum_{\text{correct } i} \mathbb{E} \|V_i \\ &\quad - \frac{1}{\delta_c(i)} \sum_{i \rightarrow \text{correct } j} V_j\|^2 \mathbb{I}(i_* = i) \\ &\quad + \sum_{\text{byz } k} \mathbb{E} \|B_k \\ &\quad - \frac{1}{\delta_c(k)} \sum_{k \rightarrow \text{correct } j} V_j\|^2 \mathbb{I}(i_* = k) \end{aligned}$$

where \mathbb{I} denotes the indicator function⁶. We examine the case $i_* = i$ for some correct index i .

$$\begin{aligned} \|V_i - \frac{1}{\delta_c(i)} \sum_{i \rightarrow \text{correct } j} V_j\|^2 &= \|\frac{1}{\delta_c(i)} \sum_{i \rightarrow \text{correct } j} V_i - V_j\|^2 \\ \text{(Jensen inequality)} &\leq \frac{1}{\delta_c(i)} \sum_{i \rightarrow \text{correct } j} \|V_i - V_j\|^2 \\ \mathbb{E} \|V_i - \frac{1}{\delta_c(i)} \sum_{i \rightarrow \text{correct } j} V_j\|^2 &\leq \frac{1}{\delta_c(i)} \sum_{i \rightarrow \text{correct } j} \mathbb{E} \|V_i - V_j\|^2 \end{aligned}$$

⁶ $\mathbb{I}(P)$ equals 1 if the predicate P is true, and 0 otherwise.

$$\leq 2d\sigma^2.$$

We now examine the case $i_* = k$ for some Byzantine index k . The fact that k minimizes the score implies that for all correct indices i

$$\begin{aligned} \sum_{k \rightarrow \text{correct } j} \|B_k - V_j\|^2 + \sum_{k \rightarrow \text{byz } l} \|B_k - B_l\|^2 &\leq \\ \sum_{i \rightarrow \text{correct } j} \|V_i - V_j\|^2 + \sum_{i \rightarrow \text{byz } l} \|V_i - B_l\|^2. \end{aligned}$$

Then, for all correct indices i

$$\begin{aligned} \|B_k - \frac{1}{\delta_c(k)} \sum_{k \rightarrow \text{correct } j} V_j\|^2 &\leq \frac{1}{\delta_c(k)} \sum_{k \rightarrow \text{correct } j} \|B_k - V_j\|^2 \\ &\leq \frac{1}{\delta_c(k)} \sum_{i \rightarrow \text{correct } j} \|V_i - V_j\|^2 \\ &\quad + \frac{1}{\delta_c(k)} \underbrace{\sum_{i \rightarrow \text{byz } l} \|V_i - B_l\|^2}_{D^2(i)}. \end{aligned}$$

We focus on the term $D^2(i)$. Each correct process i has m neighbors, and $f+1$ non-neighbors. Thus there exists a correct worker $\zeta(i)$ which is farther from i than any of the neighbors of i . In particular, for each Byzantine index l such that $i \rightarrow l$, $\|V_i - B_l\|^2 \leq \|V_i - V_{\zeta(i)}\|^2$. Whence

$$\begin{aligned} \|B_k - \frac{1}{\delta_c(k)} \sum_{k \rightarrow \text{correct } j} V_j\|^2 &\leq \frac{1}{\delta_c(k)} \sum_{i \rightarrow \text{correct } j} \|V_i - V_j\|^2 \\ &\quad + \frac{\delta_b(i)}{\delta_c(k)} \|V_i - V_{\zeta(i)}\|^2 \\ \text{taking expectation on both sides:} & \\ \mathbb{E} \|B_k - \frac{1}{\delta_c(k)} \sum_{k \rightarrow \text{correct } j} V_j\|^2 &\leq \\ \frac{1}{\delta_c(k)} \cdot (\delta_c(i) \cdot 2d\sigma^2 + \delta_b(i) \cdot \sum_{\text{correct } j \neq i} \mathbb{I}(\zeta(i) = j) \mathbb{E} \|V_i - V_j\|^2) & \\ \leq (\frac{\delta_c(i)}{\delta_c(k)} \cdot \frac{\delta_b(i)}{\delta_c(k)} (m+1)) 2d\sigma^2 & \\ \leq (\frac{m}{n-2f-2} & \\ + \frac{f}{n-2f-2} \cdot (m+1)) 2d\sigma^2. & \end{aligned}$$

Combining everything we obtain

$$\begin{aligned} \|\mathbb{E}\text{MULTI-KRUM} - g\|^2 &\leq \\ (n-f)2d\sigma^2 + f \cdot (\frac{m}{n-2f-2} + \frac{f}{n-2f-2} \cdot (m+1)) 2d\sigma^2 & \\ \leq 2(n-f + \underbrace{\frac{f \cdot m + f^2 \cdot (m+1)}{n-2f-2}}_{\eta^2(n,f)}) d\sigma^2. & \end{aligned}$$

By assumption, $\eta(n, f)\sqrt{d}\sigma < \|g\|$, i.e., $\mathbb{E}\text{MULTI-KRUM}$ belongs to a ball centered at g with radius $\eta(n, f) \cdot \sqrt{d} \cdot \sigma$. This implies

$$\begin{aligned} \langle \mathbb{E}\text{MULTI-KRUM}, g \rangle &\geq \\ (\|g\| - \eta(n, f) \cdot \sqrt{d} \cdot \sigma) \cdot \|g\| &= (1 - \sin \alpha) \cdot \|g\|^2. \end{aligned}$$

To sum up, condition (i) of the (α, f) -Byzantine resilience property holds. We now focus on condition (ii).

$$\begin{aligned} \mathbb{E}\|\text{MULTI-KRUM}\|^r &= \\ \sum_{\text{correct } i} \mathbb{E} \|V_i\|^r \mathbb{I}(i_* = i) + \sum_{\text{byz } k} \mathbb{E} \|B_k\|^r \mathbb{I}(i_* = k) & \\ \leq (n - f) \mathbb{E} \|G\|^r + \sum_{\text{byz } k} \mathbb{E} \|B_k\|^r \mathbb{I}(i_* = k). & \end{aligned}$$

Denoting by C a generic constant, when $i_* = k$, we have for all correct indices i

$$\begin{aligned} \|B_k - \frac{1}{\delta_c(k)} \sum_{k \rightarrow \text{correct } j} V_j\| &\leq \\ \sqrt{\frac{1}{\delta_c(k)} \sum_{i \rightarrow \text{correct } j} \|V_i - V_j\|^2 + \frac{\delta_b(i)}{\delta_c(k)} \|V_i - V_{\zeta(i)}\|^2} & \\ \leq C \cdot \left(\sqrt{\frac{1}{\delta_c(k)}} \cdot \sum_{i \rightarrow \text{correct } j} \|V_i - V_j\| + \sqrt{\frac{\delta_b(i)}{\delta_c(k)}} \cdot \|V_i - V_{\zeta(i)}\| \right) & \\ \leq C \cdot \sum_{\text{correct } j} \|V_j\| \quad (\text{triangular inequality}). & \end{aligned}$$

The second inequality comes from the equivalence of norms in finite dimension. Now

$$\begin{aligned} \|B_k\| &\leq \\ \|B_k - \frac{1}{\delta_c(k)} \sum_{k \rightarrow \text{correct } j} V_j\| + \left\| \frac{1}{\delta_c(k)} \sum_{k \rightarrow \text{correct } j} V_j \right\| & \\ \leq C \cdot \sum_{\text{correct } j} \|V_j\| & \\ \|B_k\|^r &\leq C \cdot \sum_{r_1 + \dots + r_{n-f} = r} \|V_1\|^{r_1} \dots \|V_{n-f}\|^{r_{n-f}}. \end{aligned}$$

Since the V_i 's are independent, we finally obtain that $\mathbb{E} \|\text{MULTI-KRUM}\|^r$ is bounded above by a linear combination of terms of the form $\mathbb{E} \|V_1\|^{r_1} \dots \mathbb{E} \|V_{n-f}\|^{r_{n-f}} = \mathbb{E} \|G\|^{r_1} \dots \mathbb{E} \|G\|^{r_{n-f}}$ with $r_1 + \dots + r_{n-f} = r$. This completes the proof of condition (ii). \square

Lemma 2. Assume that (i) the cost function Q is three times differentiable with continuous derivatives, and is non-negative, $Q(x) \geq 0$; (ii) the learning rates satisfy $\sum_t \gamma_t = \infty$ and $\sum_t \gamma_t^2 < \infty$; (iii) the gradient estimator satisfies $\mathbb{E}G(x, \xi) = \nabla Q(x)$ and $\forall r \in \{2, \dots, 4\}$, $\mathbb{E}\|G(x, \xi)\|^r \leq A_r + B_r \|x\|^r$ for some constants A_r, B_r ; (iv) there exists a constant $0 \leq \alpha < \pi/2$ such that for all x

$$\eta(n, f) \cdot \sqrt{d} \cdot \sigma(x) \leq \|\nabla Q(x)\| \cdot \sin \alpha;$$

(v) finally, beyond a certain horizon, $\|x\|^2 \geq D$, there exist $\epsilon > 0$ and $0 \leq \beta < \pi/2 - \alpha$ such that

$$\begin{aligned} \|\nabla Q(x)\| &\geq \epsilon > 0 \\ \frac{\langle x, \nabla Q(x) \rangle}{\|x\| \cdot \|\nabla Q(x)\|} &\geq \cos \beta. \end{aligned}$$

Then the sequence of gradients $\nabla Q(x_t)$ converges almost surely to zero.

The proof of Lemma 2 is exactly as in [6] and is deferred to the appendix.

We conclude the proof of (i) by recalling the definition of MULTI-KRUM , as the instance of m -Krum with $m = n - f - 2$.

a) *Proof of (ii).*

(ii) is a consequence of the fact that m -Krum is the average of m estimators of the gradient (line 8 in Algorithm 1). In the absence of Byzantine workers, all those estimators will not only be from the "correct cone", but from correct workers (Byzantine workers can also be in the correct cone, but in this case there are none). As SGD converges in $O(\frac{1}{m})$, where m is the number of used estimators of the gradient, the slowdown result follows. \square

IV. MULTI-BULYAN: STRONG BYZANTINE RESILIENCE AND SLOWDOWN

Let n be any integer greater than 2, f any integer s.t $f \leq \frac{n-3}{4}$ and m an integer s.t $m \leq n - 2f - 2$. Let $\tilde{m} = n - 2f - 2$.

Theorem 2 (Byzantine resilience and slowdown of MULTI-BULYAN). (i) MULTI-BULYAN provides strong Byzantine resilience against f failures. (ii) MULTI-BULYAN requires $O(d)$ local computation. (iii) When no worker is Byzantine, MULTI-BULYAN has a $\frac{\tilde{m}}{n}$ slowdown relative to averaging.

Proof. (i) If the number of iterations over MULTI-KRUM is $n - 2f$, then the leeway, defined by the coordinate-wise distance between the output of BULYAN and a correct gradient is upper bounded by $\mathcal{O}(\frac{1}{\sqrt{d}})$. This is due to the fact that BULYAN relies on a component-wise median, that, as proven in [12] guarantees this bound. The proof is then a direct consequence of Theorem 1 and the properties of BULYAN [12]. (ii) The linear cost in d is the consequence of running through the coordinates only in a single loop in BULYAN (lines 23-25 in Algorithm 1) and a computation of euclidean distances in MULTI-KRUM (line 6 in Algorithm 1). Finally, (iii) is a consequence of averaging \tilde{m} gradients in MULTI-BULYAN (returned value in line 24 of Algorithm 1). \square

V. EXPERIMENTS

We report on the performance MULTI-BULYAN (and its component MULTI-KRUM) over two metrics: (1) the aggregation time of our implementations of MULTI-KRUM and MULTI-BULYAN, compared to the implementation of MEDIAN in PyTorch, and (2) the maximum top-1 cross-accuracy reached on a commonly used classification task in the ML literature, compared to mere averaging and MEDIAN.

A. Setup

We run our experiments on the following hardware: (CPU) Intel® Core™ i7-8700K @ 3.70GHz, (GPU) Nvidia GeForce GTX 1080 Ti, and (RAM) 64 GB.

We report on the aggregation time, i.e. the time needed by a GAR to aggregate its input gradients and provide the output gradient. This metric is arguably the empirical counterpart of the asymptotic complexity, respectively $\mathcal{O}(n^2d)$, $\mathcal{O}(n^2d)$ and $\mathcal{O}(nd)$ for MULTI-KRUM, MULTI-BULYAN and MEDIAN. To study the empirical behaviors of MULTI-KRUM and MULTI-BULYAN compared to MEDIAN, we then vary both n and d over a realistic range of values. Namely we set $(n, d) \in \{7, 9, 11, \dots, 35, 37, 39\} \times \{10^5, 10^6, 10^7\}$ and $f = \lfloor \frac{n-3}{4} \rfloor$.

The protocol for one run is the following. n gradients are independently sampled in $\mathcal{U}(0, 1)^d$. These gradients are moved over to the GPU main memory. The command queue is then flushed on the GPU with `torch.cuda.synchronize()`, ensuring no kernel is pending on the CUDA stream. The timer is then started. The GAR is called on the GPU with the n input gradients. The command queue is then flushed again, waiting for the GAR’s execution to fully complete. The timer is finally stopped. There are 7 runs per values of (n, d) , from which we remove the 2 furthest execution times from the median of the execution times, and we report on the average and standard deviation of the 5 remaining measurements in Figure 2.

We report on the maximum top-1 cross-accuracy reached by a distributed training process using either MULTI-KRUM, MULTI-BULYAN, MEDIAN or mere averaging for aggregation. We set $n = 11$ workers and $f = 2$. There is no attack thought: this experiment highlights the benefits of averaging more gradients per aggregation step, as MULTI-KRUM and MULTI-BULYAN do, over aggregation rules that keep (the equivalent of) only one gradient, e.g. MEDIAN.

The classification task we consider is Fashion-MNIST [25] (60000 training points and 10000 testing points). The model that we train is a convolutional network, composed of two 2D-convolutional layers followed by two fully-connected layers. The first convolutional layer has 20 channels (kernel-size 5, stride 1, no padding) and the second 50 channels (same kernel-size, stride and padding). Each convolutional layer uses the *ReLU* activation function followed by a 2D-maxpool of size 2×2 . The first fully-connected layer has 500 hidden units, employing *ReLU*, and the second has 10 output units.

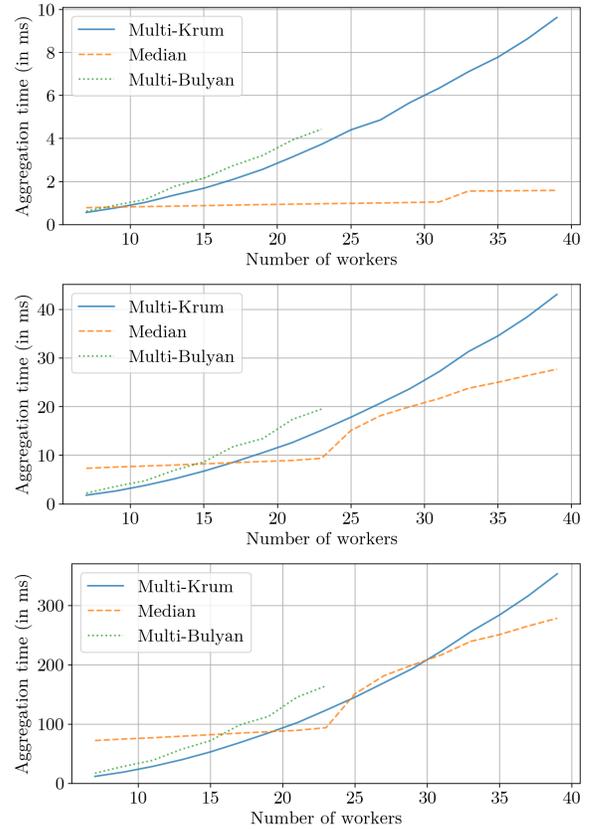


Fig. 2: Aggregation time function of the number of aggregated gradients. From top to bottom: $d = 10^5, 10^6, 10^7$. Each experiment is repeated 7 times, and we report on the average and standard deviation of the 5 aggregation times closest to the median. The standard deviation in our measurements is very small, barely visible on the graphs.

We train the model using a cross-entropy loss (*log-softmax* normalization + *negative log likelihood* loss) over 3000 steps, with a fixed learning rate of 0.1 and momentum 0.9. To compute their gradients, each worker employs minibatches of size $b \in \{5, 10, 15, \dots, 45, 50\}$. Every 100 steps we measure the top-1 cross-accuracy of the model over the whole testing set, and we keep the highest accuracy achieved over the whole training. For reproducibility purpose we seed each training, repeated 5 times with seeds 1 to 5. We report on the average and standard deviation of the highest accuracy achieved using each GAR and batch size in Figure 3.

B. Experimental Results

In Figure 2, the first observation that we can make is that the computational cost of both MULTI-KRUM and MULTI-BULYAN indeed appears quadratic in n , the number of workers. The number of workers n is kept below 24 for MULTI-BULYAN due to a limited amount of available on-die *shared memory* on the GPU we used. Regarding MEDIAN, for which we expect a linear increase with n , the tendency is not clear.

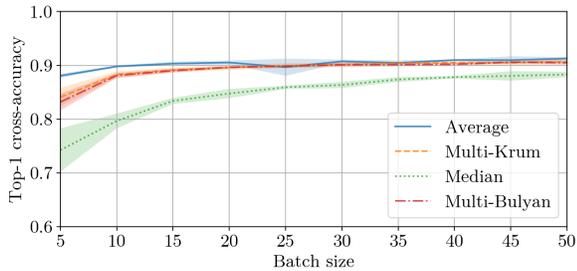


Fig. 3: Maximum top-1 cross-accuracy reached by the model with a given GAR and gradient batch size. Each experiment is repeated 5 times, with seeds 1 to 5 for reproducibility purpose, and we report on the average and standard deviation of the measured maximum accuracies.

The MEDIAN that we used for comparison is provided by the state-of-the-art machine learning framework PyTorch.

In Figure 2, and despite a higher asymptotic complexity, MULTI-KRUM and MULTI-BULYAN achieve lower aggregation times than MEDIAN for respectively $n \leq 7, n \leq 9$ ($d = 10^5$), $n \leq 15, n \leq 13$ ($d = 10^6$) and $n \leq 17, n \leq 15$ ($d = 10^7$). Essentially, the higher the dimension of the model, the higher the number of workers up to which MULTI-BULYAN is more competitive than the MEDIAN.

For reference, ResNet-50 contains $d \approx 24M$ parameters. For such neural network sizes, major DNN frameworks already show scaling issues when employing only 8 workers [16]. This inherent limitation the practitioner has to apply on the number of workers not to saturate the standard parameter server (even when using high-throughput 56 Gbps *IP-over-InfiniBand* networks [16]) would actually make MULTI-KRUM and MULTI-BULYAN faster than MEDIAN in reasonable deployments (where n is typically smaller than 20). The steady performance of MULTI-KRUM is mostly explained by the fact that its most computationally intensive part, the gradients’ pairwise distances computation, is also naturally parallelizable on GPU: it consists in many additions and multiplications executed in parallel. The remaining computations for MULTI-KRUM merely consists in ordering *scalar* values. The same applies for MULTI-BULYAN: our implementation does the costly pairwise distance computation only once, and since $f \approx \frac{n}{4}$ the median of MULTI-BULYAN is computed over a substantially reduced set of pre-aggregated gradients.

The empirical “slowdown” effect of each GAR is captured in Figure 3. Each of the studied GAR *throw away* gradients that are, in these experiments, all correct. Compared to mere averaging the $n = 11$ gradients, aggregating less gradients per step has a tangible impact on the model performance: either more training steps, or higher batch sizes per worker, is needed to compensate. By averaging only (the equivalent of) one gradient per step, MEDIAN shows in this Byzantine-free settings a tangible loss in top-1 cross-accuracy compared to MULTI-BULYAN and MULTI-KRUM, which both achieve

almost the same performance as averaging. As an additional note, the convolutional model in Figure 3 has $d = 431080$ parameters, for $n = 11$ workers (and $f = 2$). For these settings, both MULTI-KRUM and MULTI-BULYAN also have smaller aggregation times than MEDIAN.

VI. CONCLUDING REMARKS

From poisoning to Byzantine faults. We have proven the Byzantine resilience guarantees of MULTI-BULYAN and its component MULTI-KRUM, as well as their slowdown with respect to the fastest (but non Byzantine resilient) gradient aggregation rule: averaging. We also introduced two notions of Byzantine resilience (weak and strong), which we believe are practically interesting in their own right. The first to guarantee convergence and the second to protect against high-dimensional vulnerabilities.

Our notion of strong Byzantine resilience is robust to all kinds of worker failures: software bugs, hardware faults, corrupt data and malicious attacks. In particular, it also encompasses poisoning attacks [5], an important topic in *adversarial machine learning*. Mediated cases of poisoning include social platforms being perturbed by few, acute outlying data-points. While averaging approaches (and even weakly Byzantine-resilient approaches) are vulnerable to such attacks, MULTI-BULYAN tolerates them (unless they originate from a majority of users) as reported, e.g. in some of the recent work on backdooring federated learning [26].

We also argue that, even when no obvious network of machines exist, the distributed point of view presented in this paper remains relevant. What is machine learning if not an attempt to *aggregate* knowledge from distributed sources? As a concrete instance, an account on a social network posting its own content and interacting (e.g. like, comment) with content from other accounts can be seen as a *worker*. Indeed: each of these account generates data-points that, in turn, produce gradients that can be used to update e.g. a recommendation model. Byzantine-resilient aggregations would then be able to filter out the gradients malicious accounts would be producing. Through the distributed computing lense, SGD is an algorithm that eventually reaches *agreement* between data sources.

The no-free lunch variance requirement. Since 2017, many alternatives to averaging have been proposed ([4], [6], [8], [8], [12], [18], [19], [22], [23], [28]–[31] to list a few). One common aspect underlying all these methods, including ours, is their reliance on some “quality gradient” from the non Byzantine workers. This requirement is expressed in terms of how low should the variance of the gradients be. It is important to note that this requirement is not new in machine learning (it is independent of Byzantine resilience requirements), as an unbounded variance provably prevents convergence [7].

Recently, attention has been brought [3] to the hypothesis on bounded variance made in the works that are based on Krum, BULYAN, trimmed mean and variants. This hypothesis

was used also in the present work. It is important to note that the limitations that are pointed do not contradict what has been proven in this paper. Precisely, what we prove is that, as long as the variance of the stochastic gradient is controlled by the norm of the real gradient ($\eta(n, f)\sqrt{d} \cdot \sigma < \|g\|$), MULTI-BULYAN keeps making progress (i.e. keeps improving the accuracy of the model). What has been showed in [3] is that, when the models have converged (and no accuracy gain is made anymore), it is possible to inject erroneous gradients and make them accepted by MULTI-BULYAN. This is not a surprise, as the assumption does not hold when convergence has happened (the norm of the gradients becomes close to zero). In practice, this situation is already prevented by what is called *early stopping* [14]. Because the behavior of SGD can lead to bad models after an excessive number of rounds (due to, among other reasons, the fact that the hypothesis on bounded variance does not hold anymore close to convergence [7]), practitioners tend to have a test-set (different than the training-set) on which the model is tested, when the accuracy on the test-set starts degrading close to convergence, the training is stopped and the previous value of the model is kept. The attack of [3] would therefore only have been effective if it was preventing SGD from progress in the *early steps*, not later and close to convergence.

So far, we have proven that MULTI-BULYAN is (1) faster than Krum and the Median (the two leading Byzantine resilient GARs), as it relies on MULTI-KRUM, and (2) more robust than Krum and the Median, as it can make use of BULYAN. An interesting open question is whether any progress can be made on the optimality of the control ratio $\eta(n, f)$ of MULTI-KRUM and thus of MULTI-BULYAN? In other words, could we make $\eta(n, f)$ smaller, and still prove the convergence of MULTI-KRUM and MULTI-BULYAN, while tolerating even smaller values of the variance? If the answer is negative, are MULTI-KRUM and MULTI-BULYAN pluggable in the available variance-reducing methods [14] for SGD?

REFERENCES

- [1] ABADI, M., BARHAM, P., CHEN, J., CHEN, Z., DAVIS, A., DEAN, J., DEVIN, M., GHEMAWAT, S., IRVING, G., ISARD, M., ET AL. Tensorflow: A system for large-scale machine learning. In *OSDI* (2016).
- [2] BARROSO, L. A., CLIDARAS, J., AND HÖLZLE, U. The datacenter as a computer: An introduction to the design of warehouse-scale machines. *Synthesis lectures on computer architecture* 8, 3 (2013), 1–154.
- [3] BARUCH, M., BARUCH, G., AND GOLDBERG, Y. A little is enough: Circumventing defenses for distributed learning. *Advances in Neural Information Processing Systems* (2019).
- [4] BERNSTEIN, J., ZHAO, J., AZIZZADENESHELI, K., AND ANANDKUMAR, A. signSGD with majority vote is communication efficient and fault tolerant. In *International Conference on Learning Representations* (2019).
- [5] BIGGIO, B., AND LASKOV, P. Poisoning attacks against support vector machines. In *ICML* (2012).
- [6] BLANCHARD, P., EL MHAMDI, E. M., GUERRAOUI, R., AND STAINER, J. Machine learning with adversaries: Byzantine tolerant gradient descent. In *NIPS* (2017), pp. 118–128.
- [7] BOTTOU, L. Online learning and stochastic approximations. *Online learning in neural networks* 17, 9 (1998), 142.
- [8] CHEN, L., CHARLES, Z., PAPAILIOPOULOS, D., ET AL. Draco: Robust distributed training via redundant gradients. *arXiv preprint arXiv:1803.09877* (2018).

- [9] DAMASKINOS, G., EL-MHAMDI, E.-M., GUERRAOUI, R., GUIRGUIS, A., AND ROUAULT, S. Aggregator: Byzantine machine learning via robust gradient aggregation. In *Proceedings of the 2nd Conference on Systems and Machine Learning (SysML)* (2019).
- [10] DEAN, J., CORRADO, G., MONGA, R., CHEN, K., DEVIN, M., MAO, M., SENIOR, A., TUCKER, P., YANG, K., LE, Q. V., ET AL. Large scale distributed deep networks. In *NIPS* (2012), pp. 1223–1231.
- [11] EL MHAMDI, E. M. *Robust Distributed Learning*. PhD thesis, 2020.
- [12] EL MHAMDI, E. M., GUERRAOUI, R., AND ROUAULT, S. The hidden vulnerability of distributed learning in Byzantium. In *Proceedings of the 35th International Conference on Machine Learning (Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018)*, vol. 80 of *Proceedings of Machine Learning Research*, PMLR, pp. 3521–3530.
- [13] GOYAL, P., DOLLÁR, P., GIRSHICK, R., NOORDHUIS, P., WESOLOWSKI, L., KYROLA, A., TULLOCH, A., JIA, Y., AND HE, K. Accurate, large minibatch sgd: training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677* (2017).
- [14] HAYKIN, S. S. *Neural networks and learning machines*, vol. 3. Pearson Upper Saddle River, NJ, USA, 2009.
- [15] LI, M., ANDERSEN, D. G., PARK, J. W., SMOLA, A. J., AHMED, A., JOSIFOVSKI, V., LONG, J., SHEKITA, E. J., AND SU, B.-Y. Scaling distributed machine learning with the parameter server. In *OSDI* (2014), pp. 583–598.
- [16] LUO, L., NELSON, J., CEZE, L., PHANISHAYEE, A., AND KRISHNAMURTHY, A. Parameter box: High performance parameter servers for efficient distributed deep neural network training. *CoRR abs/1801.09805* (2018).
- [17] MENG, X., BRADLEY, J., YAVUZ, B., SPARKS, E., VENKATARAMAN, S., LIU, D., FREEMAN, J., TSAI, D., AMDE, M., OWEN, S., ET AL. MLlib: Machine learning in apache spark. *JMLR* 17, 1 (2016), 1235–1241.
- [18] MUÑOZ-GONZÁLEZ, L., CO, K. T., AND LUPU, E. C. Byzantine-robust federated machine learning through adaptive model averaging. *arXiv preprint arXiv:1909.05125* (2019).
- [19] RAJPUT, S., WANG, H., CHARLES, Z., AND PAPAILIOPOULOS, D. Detox: A redundancy-based framework for faster and more robust gradient aggregation. *Neural Information Processing Systems* (2019).
- [20] ROUSSEEUW, P. J. Multivariate estimation with high breakdown point. *Mathematical statistics and applications* 8 (1985), 283–297.
- [21] SHALEV-SHWARTZ, S., AND BEN-DAVID, S. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.
- [22] SU, L., AND XU, J. Securing distributed machine learning in high dimensions. *arXiv preprint arXiv:1804.10140* (2018).
- [23] TIANXIANG, W., ZHENG, Z., CHANGBING, T., AND HAO, P. Aggregation rules based on stochastic gradient descent in byzantine consensus. In *2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC)* (2019), IEEE, pp. 317–324.
- [24] WOLD, S., ESBENSEN, K., AND GELADI, P. Principal component analysis. *Chemometrics and intelligent laboratory systems* 2, 1-3 (1987), 37–52.
- [25] XIAO, H., RASUL, K., AND VOLLGRAF, R. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *CoRR abs/1708.07747* (2017).
- [26] XIE, C., HUANG, K., CHEN, P.-Y., AND LI, B. {DBA}: Distributed backdoor attacks against federated learning. In *International Conference on Learning Representations* (2020).
- [27] XIE, C., KOYEJO, O., AND GUPTA, I. Generalized Byzantine-tolerant sgd. *arXiv preprint arXiv:1802.10116* (2018).
- [28] YANG, Z., AND BAJWA, W. U. Bridge: Byzantine-resilient decentralized gradient descent. *arXiv preprint arXiv:1908.08098* (2019).
- [29] YANG, Z., AND BAJWA, W. U. Byrdie: Byzantine-resilient distributed coordinate descent for decentralized learning. *IEEE Transactions on Signal and Information Processing over Networks* (2019).
- [30] YANG, Z., GANG, A., AND BAJWA, W. U. Adversary-resilient inference and machine learning: From distributed to decentralized. *arXiv preprint arXiv:1908.08649* (2019).
- [31] YIN, D., CHEN, Y., RAMCHANDRAN, K., AND BARTLETT, P. Byzantine-robust distributed learning: Towards optimal statistical rates. *arXiv preprint arXiv:1803.01498* (2018).
- [32] ZHANG, S., CHOROMANSKA, A. E., AND LECUN, Y. Deep learning with elastic averaging sgd. In *NIPS* (2015), pp. 685–693.