Supplementary proofs : AGGREGATHOR Byzantine-resilience and Convergence Speed.

Abstract

In [1], Krum, the first provably Byzantine resilient algorithm for SGD, was introduced. Krum only uses one worker per step, which hampers its speed of convergence, especially in best case conditions when none of the workers is actually Byzantine. The idea behind MULTI-KRUM, of using $m > 1$ different workers per step was mentioned in [1], without however any proof neither on its Byzantine resilience nor on its slowdown. The present technical report closes this open problem and provides proofs of (weak) Byzantine resilience, convergence, and $\sqrt{\frac{m}{n}}$ slowdown of MULTI-KRUM compared to the optimal averaging in the absence of Byzantine workers. Based on that, and on the theoretical work of [3], we prove the similar $\sqrt{\frac{m}{n}}$ slowdown of AGGREGATHOR and its (strong) Byzantine resilience. We deduce that AGGREGATHOR ensures strong Byzantine resilience and the very fact that it is $\sqrt{\frac{m}{n}}$ times as fast as the optimal algorithm (averaging) in the absence of Byzantine workers.

AGGREGATHOR is the composition of MULTI-KRUM and BULYAN, which can be viewed as generalization (also using m different workers per step to leverage the fact that f, possibly less than a minority can be faulty) of *Bulyan*, the defense mechanism of [3]. Before presenting in Section 2, our proofs of convergence and slow down of MULTI-KRUM and in Section 3 our proofs of convergence and slow down of BULYAN and hence AGGREGATHOR, we introduce in Section 1 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.

1 Theoretical Context

Intuitively, weak 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 $\boldsymbol{x}^{(k)}$ (Equation 2 in the main paper) converges almost surely to some $\boldsymbol{x}^*$ where $\boldsymbol{\nabla} Q(\boldsymbol{x}^*) = 0$, despite the presence of f Byzantine workers.*

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

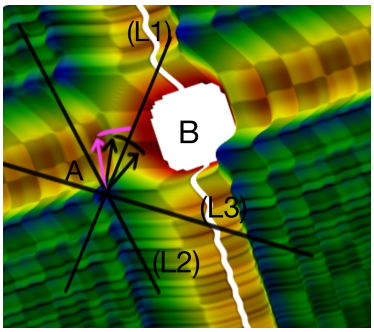

Figure 1: In a non-convex situation, two correct vectors (black arrows) are pointing towards the deep optimum located in area B, both vectors belong to the plane formed by lines L1 and L2. A Byzantine worker (magenta) is taking benefit from the third dimension, and the non-convex landscape, to place 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 $\sqrt{d}$.

optimization, which, in the presence of Byzantine workers, is highly aggravated by the dimension of the problem as explained in what follows.

Specificity of non-convex optimization. Non-convex optimization is one of the earliest established NP-hard problems [4]. In fact, many, if not all of the interesting but hard questions in machine learning boil down to one answer: "because the cost function is not convex".

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 [2, 4]. This can be understood from the "S" (stochasticity) of SGD: as (correct) workers cannot have a full picture of the surrounding landscape of the loss, they can only draw a sample at random and estimate the best direction based on that sample, which can be - and is probably - different the true gradient. But on expectation (over samples) this gradient estimate is equal to the true gradient. Moreover, due to non-convexity, 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 might end up providing a direction to get out of that local minima ! Unless of course when the computational resources of that Byzantine worker can face the high-dimensional landscape of the loss and find a truly misleading

update vector.

(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 for every $i \in [1, d]$, there exists a correct gradient $\boldsymbol{G}$ (i.e., computed by a non-Byzantine worker) s.t. $\mathbb{E}|\boldsymbol{GAR}_i - \boldsymbol{G}_i| = O(\frac{1}{\sqrt{d}})$. The the expectation is taken over the random samples (ξ in Equation 4 of the main paper)and $\boldsymbol{v}_i$ denotes the i^{th} coordinate of a vector $\boldsymbol{v}$.*

For the sake of our theoretical analysis, we also introduce the definition of (α, f)–Byzantine resilience (Definition 3). This definition is a sufficient condition (as proved in [1] based on [2]) for weak Byzantine resilience that we introduce and require from GARs in our main paper (Section 2, Definition 1). Eventhough 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 [1, 6] 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 AGGREGATHOR, in particular, we will require it from MULTI-KRUM. The theoretical analysis done in [3] 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)**.** *Let $0 \leq \alpha < \pi/2$ be any angular value, and any integer $0 \leq f \leq n$. Let $V_1, \ldots, V_n$ be any independent identically distributed random vectors in $\mathbb{R}^d$, $V_i \sim G$, with $\mathbb{E}G = g$. Let $B_1, \ldots, B_f$ 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 $1 \leq j_1 < \cdots < j_f \leq n$, vector*

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

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

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 the weak Byzantine resilience of MULTI-KRUM (Theorem 1).

[1] Having a scalar product that is lower bounded by this value guarantees that the GAR of MULTI-KRUM lives in the aformentioned cone. For a visualisation of this requirement, see the ball and inner triangle of Figure 2

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 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 $m = n-f-2$ [2], 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 [1], the value of $\eta(n.f)$ itself depends on m. This is what enables us to use multi-Krum as the basis of our MULTI-KRUM, unlike [1] where a restriction is made on $m = 1$.

The proof of Lemma 2 is the same as the one in [1] which itself draws on the rather classic analysis of SGD made by L.Bottou [2]. 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 [1,2], reasonable assumptions are made on the cost function Q, those assumption are not restrictive and are common in practical machine learning.

2 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 $\tilde{m} = n-f-2$.

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 $\Omega(\sqrt{\frac{\tilde{m}}{n}})$.*

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

Lemma 1. *Let $V_1, \ldots, V_n$ be any independent and identically distributed random d-dimensional vectors s.t $V_i \sim G$, with $\mathbb{E}G = g$ and $\mathbb{E}\left\|G-g\right\|^2 = d\sigma^2$. Let $B_1, \ldots, B_f$ 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*

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

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

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 $B_1, \ldots, B_f$ occupy the last f positions in the list of arguments of MULTI-KRUM, i.e., MULTI-KRUM $=$ MULTI-KRUM$(V_1, \ldots, V_{n-f}, B_1, \ldots, B_f)$. An index is *correct* if it refers to a vector among $V_1, \ldots, V_{n-f}$. An index is *Byzantine* if it refers to a vector among $B_1, \ldots, B_f$. 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 $i \to j$ (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

$$\begin{aligned} \delta_c(i) + \delta_b(i) &= m \\ n - 2f - 2 \leq \delta_c(i) &\leq m \\ \delta_b(i) &\leq f. \end{aligned}$$

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 5 of the main paper (Section 3).

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

where $\mathbb{I}$ denotes the indicator function[3]. We examine the case $i_* = i$ for some correct index i.

$$\left\| V_i - \frac{1}{\delta_c(i)} \sum_{i \to \text{ correct } j} V_j \right\|^2 = \left\| \frac{1}{\delta_c(i)} \sum_{i \to \text{ correct } j} V_i - V_j \right\|^2$$

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

$$\leq \frac{1}{\delta_c(i)} \sum_{i\to \text{ correct } j} \|V_i - V_j\|^2 \quad \text{(Jensen inequality)}$$

$$\mathbb{E}\left\|V_i - \frac{1}{\delta_c(i)} \sum_{i\to \text{ correct } j} V_j\right\|^2 \leq \frac{1}{\delta_c(i)} \sum_{i\to \text{ correct } j} \mathbb{E}\,\|V_i - V_j\|^2$$
$$\leq 2d\sigma^2.$$

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

$$\sum_{k\to \text{ correct } j} \|B_k - V_j\|^2 + \sum_{k\to \text{ byz } l} \|B_k - B_l\|^2 \leq \sum_{i\to \text{ correct } j} \|V_i - V_j\|^2 + \sum_{i\to \text{ byz } l} \|V_i - B_l\|^2 .$$

Then, for all correct indices i

$$\left\|B_k - \frac{1}{\delta_c(k)} \sum_{k\to \text{ correct } j} V_j\right\|^2 \leq \frac{1}{\delta_c(k)} \sum_{k\to \text{ correct } j} \|B_k - V_j\|^2$$
$$\leq \frac{1}{\delta_c(k)} \sum_{i\to \text{ correct } j} \|V_i - V_j\|^2 + \frac{1}{\delta_c(k)} \underbrace{\sum_{i\to \text{ byz } l} \|V_i - B_l\|^2}_{D^2(i)} .$$

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 $i \to l$, $\|V_i - B_l\|^2 \leq \left\|V_i - V_{\zeta(i)}\right\|^2$. Whence

$$\left\|B_k - \frac{1}{\delta_c(k)} \sum_{k\to \text{ correct } j} V_j\right\|^2 \leq \frac{1}{\delta_c(k)} \sum_{i\to \text{ correct } j} \|V_i - V_j\|^2 + \frac{\delta_b(i)}{\delta_c(k)} \left\|V_i - V_{\zeta(i)}\right\|^2$$

$$\mathbb{E}\left\|B_k - \frac{1}{\delta_c(k)} \sum_{k\to \text{ correct } j} V_j\right\|^2 \leq \frac{\delta_c(i)}{\delta_c(k)} \cdot 2d\sigma^2 + \frac{\delta_b(i)}{\delta_c(k)} \sum_{\text{correct } j\neq i} \mathbb{E}\,\|V_i - V_j\|^2 \mathbb{I}(\zeta(i) = j)$$
$$\leq \left(\frac{\delta_c(i)}{\delta_c(k)} \cdot + \frac{\delta_b(i)}{\delta_c(k)}(m+1)\right) 2d\sigma^2$$
$$\leq \left(\frac{m}{n-2f-2} + \frac{f}{n-2f-2} \cdot (m+1)\right) 2d\sigma^2.$$

Putting everything back together, we obtain

$$\|\mathbb{E}\textsc{Multi-Krum} - g\|^2 \leq (n-f)2d\sigma^2 + f \cdot \left(\frac{m}{n-2f-2} + \frac{f}{n-2f-2} \cdot (m+1)\right) 2d\sigma^2$$
$$\leq \underbrace{2\left(n - f + \frac{f\cdot m + f^2 \cdot (m+1)}{n-2f-2}\right)}_{\eta^2(n,f)} d\sigma^2.$$

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

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

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

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

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

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

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 $\mathbb{E}\,\|V_1\|^{r_1}\cdots\mathbb{E}\,\|V_{n-f}\|^{r_{n-f}} = \mathbb{E}\,\|G\|^{r_1}\cdots\mathbb{E}\,\|G\|^{r_{n-f}}$ with $r_1+\cdots+r_{n-f}=r$. This completes the proof of condition *(ii)*. □

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 $\forall r \in \{2, \ldots, 4\},\ \mathbb{E}\|G(x,\xi)\|^r \leq A_r + B_r\|x\|^r$ *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*

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

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

Proof. For the sake of simplicity, we write $\textsc{Multi-Krum}_t = \textsc{Multi-Krum}(V_1^t, \ldots, V_n^t)$. Before proving the main claim of the proposition, we first show that the sequence x_t is almost surely globally confined within the region $\|x\|^2 \leq D$.

(Global confinement). Let $u_t = \phi(\|x_t\|^2)$ where

$$\phi(a) = \begin{cases} 0 & \text{if } a < D \\ (a-D)^2 & \text{otherwise} \end{cases}$$

Note that

$$\phi(b) - \phi(a) \leq (b-a)\phi'(a) + (b-a)^2. \tag{1}$$

This becomes an equality when $a, b \geq D$. Applying this inequality to $u_{t+1} - u_t$ yields

$$\begin{aligned} u_{t+1} - u_t \leq\ & \left(-2\gamma_t\langle x_t, \textsc{Multi-Krum}_t\rangle + \gamma_t^2\|\textsc{Multi-Krum}_t\|^2\right) \cdot \phi'(\|x_t\|^2) \\ & + 4\gamma_t^2\langle x_t, \textsc{Multi-Krum}_t\rangle^2 - 4\gamma_t^3\langle x_t, \textsc{Multi-Krum}_t\rangle\|\textsc{Multi-Krum}_t\|^2 + \gamma_t^4\|\textsc{Multi-Krum}_t\|^4 \\ & \leq -2\gamma_t\langle x_t, \textsc{Multi-Krum}_t\rangle\phi'(\|x_t\|^2) + \gamma_t^2\|\textsc{Multi-Krum}_t\|^2\phi'(\|x_t\|^2) \\ & + 4\gamma_t^2\|x_t\|^2\|\textsc{Multi-Krum}_t\|^2 + 4\gamma_t^3\|x_t\|\|\textsc{Multi-Krum}_t\|^3 + \gamma_t^4\|\textsc{Multi-Krum}_t\|^4. \end{aligned}$$

Let $\mathcal{P}_t$ denote the σ-algebra encoding all the information up to round t. Taking the conditional expectation with respect to $\mathcal{P}_t$ yields

$$\begin{aligned} \mathbb{E}\left(u_{t+1} - u_t | \mathcal{P}_t\right) \leq\ & -2\gamma_t\langle x_t, \mathbb{E}\textsc{Multi-Krum}_t\rangle + \gamma_t^2\mathbb{E}\left(\|\textsc{Multi-Krum}_t\|^2\right)\phi'(\|x_t\|^2) \\ & + 4\gamma_t^2\|x_t\|^2\mathbb{E}\left(\|\textsc{Multi-Krum}_t\|^2\right) + 4\gamma_t^3\|x_t\|\mathbb{E}\left(\|\textsc{Multi-Krum}_t\|^3\right) + \gamma_t^4\mathbb{E}\left(\|\textsc{Multi-Krum}_t\|^4\right). \end{aligned}$$

Thanks to condition *(ii)* of (α, f)-Byzantine resilience, and the assumption on the first four moments of G, there exist positive constants A_0, B_0 such that

$$\mathbb{E}\left(u_{t+1} - u_t | \mathcal{P}_t\right) \leq -2\gamma_t\langle x_t, \mathbb{E}\textsc{Multi-Krum}_t\rangle\phi'(\|x_t\|^2) + \gamma_t^2\left(A_0 + B_0\|x_t\|^4\right).$$

Thus, there exist positive constant A, B such that

$$\mathbb{E}\left(u_{t+1} - u_t | \mathcal{P}_t\right) \leq -2\gamma_t\langle x_t, \mathbb{E}\textsc{Multi-Krum}_t\rangle\phi'(\|x_t\|^2) + \gamma_t^2\left(A + B \cdot u_t\right).$$

When $\|x_t\|^2 < D$, the first term of the right hand side is null because $\phi'(\|x_t\|^2) = 0$. When $\|x_t\|^2 \geq D$, this first term is negative because (see Figure 2)

$$\langle x_t, \mathbb{E}\text{MULTI-KRUM}_t \rangle \geq \|x_t\| \cdot \|\mathbb{E}\text{MULTI-KRUM}_t\| \cdot \cos(\alpha + \beta) > 0.$$

Hence

$$\mathbb{E}\left(u_{t+1} - u_t | \mathcal{P}_t\right) \leq \gamma_t^2 \left(A + B \cdot u_t\right).$$

We define two auxiliary sequences

$$\mu_t = \prod_{i=1}^{t} \frac{1}{1 - \gamma_i^2 B} \xrightarrow[t\to\infty]{} \mu_\infty$$
$$u'_t = \mu_t u_t.$$

Note that the sequence μ_t converges because $\sum_t \gamma_t^2 < \infty$. Then

$$\mathbb{E}\left(u'_{t+1} - u'_t | \mathcal{P}_t\right) \leq \gamma_t^2 \mu_t A.$$

Consider the indicator of the positive variations of the left-hand side

$$\chi_t = \begin{cases} 1 & \text{if } \mathbb{E}\left(u'_{t+1} - u'_t | \mathcal{P}_t\right) > 0 \\ 0 & \text{otherwise} \end{cases}$$

Then

$$\mathbb{E}\left(\chi_t \cdot (u'_{t+1} - u'_t)\right) \leq \mathbb{E}\left(\chi_t \cdot \mathbb{E}\left(u'_{t+1} - u'_t | \mathcal{P}_t\right)\right) \leq \gamma_t^2 \mu_t A.$$

The right-hand side of the previous inequality is the summand of a convergent series. By the quasi-martingale convergence theorem [5], this shows that the sequence u'_t converges almost surely, which in turn shows that the sequence u_t converges almost surely, $u_t \to u_\infty \geq 0$.

Let us assume that $u_\infty > 0$. When t is large enough, this implies that $\|x_t\|^2$ and $\|x_{t+1}\|^2$ are greater than D. Inequality 1 becomes an equality, which implies that the following infinite sum converges almost surely

$$\sum_{t=1}^{\infty} \gamma_t \langle x_t, \mathbb{E}\text{MULTI-KRUM}_t \rangle \phi'(\|x_t\|^2) < \infty.$$

Note that the sequence $\phi'(\|x_t\|^2)$ converges to a positive value. In the region $\|x_t\|^2 > D$, we have

$$\begin{aligned}
\langle x_t, \mathbb{E}\text{MULTI-KRUM}_t \rangle &\geq \sqrt{D} \cdot \|\mathbb{E}\text{MULTI-KRUM}_t\| \cdot \cos(\alpha + \beta) \\
&\geq \sqrt{D} \cdot \left(\|\nabla Q(x_t)\| - \eta(n, f) \cdot \sqrt{d} \cdot \sigma(x_t)\right) \cdot \cos(\alpha + \beta) \\
&\geq \sqrt{D} \cdot \epsilon \cdot (1 - \sin\alpha) \cdot \cos(\alpha + \beta) > 0.
\end{aligned}$$

This contradicts the fact that $\sum_{t=1}^{\infty} \gamma_t = \infty$. Therefore, the sequence u_t converges to zero. This convergence implies that the sequence $\|x_t\|^2$ is bounded, i.e., the vector x_t is confined in a bounded region containing the origin. As a consequence, any continuous function of x_t is also bounded, such as, e.g., $\|x_t\|^2$, $\mathbb{E}\|G(x_t, \xi)\|^2$ and all the derivatives of the cost function $Q(x_t)$. In the sequel, positive constants K_1, K_2, etc... are introduced whenever such a bound is used.

(Convergence). We proceed to show that the gradient $\nabla Q(x_t)$ converges almost surely to zero. We define

$$h_t = Q(x_t).$$

Using a first-order Taylor expansion and bounding the second derivative with K_1, we obtain

$$|h_{t+1} - h_t + 2\gamma_t \langle \textsc{Multi-Krum}_t, \nabla Q(x_t)\rangle| \leq \gamma_t^2 \|\textsc{Multi-Krum}_t\|^2 K_1 \text{ a.s.}$$

Therefore

$$\mathbb{E}\left(h_{t+1} - h_t | \mathcal{P}_t\right) \leq -2\gamma_t \langle \mathbb{E}\textsc{Multi-Krum}_t, \nabla Q(x_t)\rangle + \gamma_t^2 \mathbb{E}\left(\|\textsc{Multi-Krum}_t\|^2 | \mathcal{P}_t\right) K_1. \tag{2}$$

By the properties of (α, f)-Byzantine resiliency, this implies

$$\mathbb{E}\left(h_{t+1} - h_t | \mathcal{P}_t\right) \leq \gamma_t^2 K_2 K_1,$$

which in turn implies that the positive variations of h_t are also bounded

$$\mathbb{E}\left(\chi_t \cdot (h_{t+1} - h_t)\right) \leq \gamma_t^2 K_2 K_1.$$

The right-hand side is the summand of a convergent infinite sum. By the quasi-martingale convergence theorem, the sequence h_t converges almost surely, $Q(x_t) \to Q_\infty$.

Taking the expectation of Inequality 2, and summing on $t = 1, \ldots, \infty$, the convergence of $Q(x_t)$ implies that

$$\sum_{t=1}^{\infty} \gamma_t \langle \mathbb{E}\textsc{Multi-Krum}_t, \nabla Q(x_t)\rangle < \infty \text{ a.s.}$$

We now define

$$\rho_t = \|\nabla Q(x_t)\|^2.$$

Using a Taylor expansion, as demonstrated for the variations of h_t, we obtain

$$\rho_{t+1} - \rho_t \leq -2\gamma_t \langle \textsc{Multi-Krum}_t, \left(\nabla^2 Q(x_t)\right) \cdot \nabla Q(x_t)\rangle + \gamma_t^2 \|\textsc{Multi-Krum}_t\|^2 K_3 \text{ a.s.}$$

Taking the conditional expectation, and bounding the second derivatives by K_4,

$$\mathbb{E}\left(\rho_{t+1} - \rho_t | \mathcal{P}_t\right) \leq 2\gamma_t \langle \mathbb{E}\textsc{Multi-Krum}_t, \nabla Q(x_t)\rangle K_4 + \gamma_t^2 K_2 K_3.$$

The positive expected variations of ρ_t are bounded

$$\mathbb{E}\left(\chi_t \cdot (\rho_{t+1} - \rho_t)\right) \leq 2\gamma_t \mathbb{E}\langle \mathbb{E}\textsc{Multi-Krum}_t, \nabla Q(x_t)\rangle K_4 + \gamma_t^2 K_2 K_3.$$

The two terms on the right-hand side are the summands of convergent infinite series. By the quasi-martingale convergence theorem, this shows that ρ_t converges almost surely.

We have

$$\langle \mathbb{E}\textsc{Multi-Krum}_t, \nabla Q(x_t)\rangle \geq \left(\|\nabla Q(x_t)\| - \eta(n, f) \cdot \sqrt{d} \cdot \sigma(x_t)\right) \cdot \|\nabla Q(x_t)\|$$

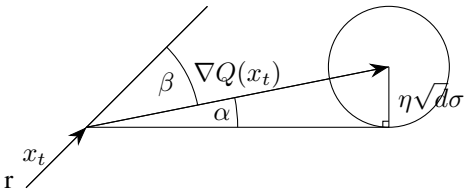

Figure 2: Condition on the angles between x_t, $\nabla Q(x_t)$ and the the GAR of MULTI-KRUM vector $\mathbb{E}$MULTI-KRUM$_t$, in the region $\|x_t\|^2 > D$.

$$\geq \underbrace{(1 - \sin\alpha)}_{>0} \cdot \rho_t.$$

This implies that the following infinite series converge almost surely

$$\sum_{t=1}^{\infty} \gamma_t \cdot \rho_t < \infty.$$

Since ρ_t converges almost surely, and the series $\sum_{t=1}^{\infty} \gamma_t = \infty$ diverges, we conclude that the sequence $\|\nabla Q(x_t)\|$ converges almost surely to zero. □

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

Proof of (ii). **(ii)** is a consequence of the fact that m-Krum is the average of m estimators of the gradient. 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 $O(\frac{1}{\sqrt{m}})$, where m is the number of used estimators of the gradient, the slowdown result follows. □

3 AGGREGATHOR: 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 AGGREGATHOR)**.** *(i)* AGGREGATHOR *provides strong Byzantine resilience against f failures. (ii) In the absence of Byzantine workers,* AGGREGATHOR *has a slowdown (expressed in ratio with averaging) of* $\Omega(\sqrt{\frac{\tilde{m}}{n}})$.

Proof. 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 [3] guarantees this bound. The proof is then a direct consequence of Theorem 1 and the properties of Bulyan [3] □

References

[1] BLANCHARD, P., EL MHAMDI, E. M., GUERRAOUI, R., AND STAINER, J. Machine learning with adversaries: Byzantine tolerant gradient descent. In *Neural Information Processing Systems* (2017), pp. 118–128.

[2] BOTTOU, L. Online learning and stochastic approximations. *Online learning in neural networks 17*, 9 (1998), 142.

[3] 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), J. Dy and A. Krause, Eds., vol. 80 of *Proceedings of Machine Learning Research*, PMLR, pp. 3521–3530.

[4] HAYKIN, S. S. *Neural networks and learning machines*, vol. 3. Pearson Upper Saddle River, NJ, USA:, 2009.

[5] MÉTIVIER, M. *Semi-Martingales*. Walter de Gruyter, 1983.

[6] XIE, C., KOYEJO, O., AND GUPTA, I. Generalized Byzantine-tolerant sgd. *arXiv preprint arXiv:1802.10116* (2018).