The Hidden Vulnerability of Distributed Learning in Byzantium — Supplementary Material

El Mahdi El Mhamdi [1] Rachid Guerraoui [1] Sébastien Rouault [1]

A. Brute's (α, f)–Byzantine–resilience proof

A.1. Background

Definition 1 ((α, f)–Byzantine–resilience)**.**
Let $(\alpha, f) \in [0, \pi/2] \times [0 \,..\, n]$ be any angle and any integer. Let $n \in \mathbb{N}$ with $n > f$.
Let $(V_1 \dots V_{n-f}) \in \left(\mathbb{R}^d\right)^{n-f}$ be independent, identically distributed random vectors, with $V_i \sim \mathcal{G}$ and $\mathbb{E}[\mathcal{G}] = G$.
Let $(B_1 \dots B_f) \in \left(\mathbb{R}^d\right)^f$ be random vectors, possibly dependent between them and the vectors $(V_1 \dots V_{n-f})$.
Then, an aggregation rule $\mathcal{F}$ is said to be (α, f)-Byzantine-resilient if, for any $1 \leq j_1 < \dots < j_f \leq n$, the vector:

$$F = \mathcal{F}\left(V_1, \dots, \underbrace{B_1}_{j_1}, \dots, \underbrace{B_f}_{j_f}, \dots, V_n\right)$$

satisfies:

1. $\langle \mathbb{E}[F], G \rangle \geq (1 - \sin\alpha) \cdot \|G\|^2 > 0$
2. $\forall r \in \{2, 3, 4\}$, $\mathbb{E}\|F\|^r$ *is bounded above by a linear combination of the terms* $\mathbb{E}\|\mathcal{G}\|^{r_1} \cdot \dots \cdot \mathbb{E}\|\mathcal{G}\|^{r_{n-1}}$, *with* $r_1 + \dots + r_{n-1} = r$.

A.2. Definition

Let $(n, f) \in \mathbb{N}^2$ with $n \geq 2f + 1$.
Let $(V_1 \dots V_{n-f}) \in \left(\mathbb{R}^d\right)^{n-f}$ be independent, identically distributed random vectors, with $V_i \sim \mathcal{G}$ and $\mathbb{E}[\mathcal{G}] = G$.
Let $(B_1 \dots B_f) \in \left(\mathbb{R}^d\right)^f$ be random vectors, possibly dependent between them and the vectors $(V_1 \dots V_{n-f})$.
Let $\|\cdot\|_p$ be the ℓ_p–norm, with $p \in \mathbb{N}^* \cup \{+\infty\}$.

Let $\mathcal{Q} = \{V_1 \dots V_n\}$ be the set of submitted gradients.
Let $\mathcal{R} = \{\mathcal{X} \mid \mathcal{X} \subset \mathcal{Q}, |\mathcal{X}| = n - f\}$ be the set of all the subsets of $\mathcal{Q}$ with a cardinality of $n - f$.
Let $\mathcal{S} = \underset{\mathcal{X} \in \mathcal{R}}{\arg\min}\left(\max_{(V_i, V_j) \in \mathcal{X}^2}\left(\|V_i - V_j\|_p\right)\right)$.
Then, the aggregated gradient $F = \frac{1}{n-f} \sum_{V \in \mathcal{S}} V$.

A.3. Proof

Let $\forall (i, j) \in [1 \,..\, n - f]^2$, $i \neq j$ be $\bar{\sigma} \triangleq \mathbb{E}\|V_i - V_j\|_p$. Under the assumption that $2 f \bar{\sigma} < (n - f)\|G\|_p$, we will prove that this rule is (α, f)–Byzantine–resilient.

Trivial case: $\forall i \in [1 \,..\, f]$, $B_i \notin \mathcal{S}$.
As the aggregated gradient F is the arithmetic mean of unbiased vectors V_j, we have $\mathbb{E}[F] = G$, and points *1.* and *2.* of definition 1 are trivially satisfied.

Otherwise, without loss of generality, let $b \in [1 \,..\, f]$ and $\mathcal{S} = \{V_1 \dots V_{n-f-b}, B_1 \dots B_b\}$, $\bar{\mathcal{R}} = \mathcal{R} \setminus \mathcal{S}$. It holds:

$$\begin{aligned}
&\forall \bar{\mathcal{S}} \in \bar{\mathcal{R}},\ \exists X_i \in \bar{\mathcal{S}} \setminus \mathcal{S},\ \exists X_j \in \bar{\mathcal{S}} \setminus \{X_i\}, \\
&\forall X_k \in \mathcal{S},\ \forall X_l \in \mathcal{S} \setminus \{X_k\}, \\
&\|X_k - X_l\|_p < \|X_i - X_j\|_p
\end{aligned}$$

We can also notice that: $\exists \mathcal{V} \in \bar{\mathcal{R}}$, $\forall i \in [1 \,..\, f]$, $B_i \notin \mathcal{V}$. Then, by combining this observation with the previous one:

$$\begin{aligned}
&\forall a \in [1 \,..\, b], B_a \in \mathcal{S} \\
&\qquad \Rightarrow\ \exists (x_a, y_a) \in [1 \,..\, n - f]^2,\ x_a \neq y_a, \\
&\qquad\quad \forall k \in [1 \,..\, n - f - b], \\
&\qquad\quad \|B_a - V_k\|_p < \|V_{x_a} - V_{y_a}\|_p
\end{aligned}$$

This last observation will be reused in the following.

We can compute the aggregated gradient:

$$F = \frac{1}{n-f}\left(\sum_{i=1}^{n-f-b} V_i + \sum_{i=1}^{b} B_i\right)$$

and compare it with the average of the non–Byzantine ones:

$$\begin{aligned}
\widehat{G} &= \frac{1}{n-f} \sum_{i=1}^{n-f} V_i \\
F - \widehat{G} &= \frac{1}{n-f}\left(\sum_{i=1}^{b} B_i - \sum_{i=n-f-b+1}^{n-f} V_i\right) \\
&= \frac{1}{n-f} \sum_{i=1}^{b} B_i - V_{i+n-f-b}
\end{aligned}$$

$$\begin{aligned}\left\|F - \widehat{G}\right\|_p &\leq \frac{1}{n-f} \sum_{i=1}^{b} \left\|B_i - V_{i+n-f-b}\right\|_p \\ &\leq \frac{1}{n-f} \sum_{i=1}^{b} \Big(\left\|B_i - V_1\right\|_p \\ &\qquad + \left\|V_1 - V_{i+n-f-b}\right\|_p\Big) \\ &\leq \frac{1}{n-f} \sum_{i=1}^{b} \Big(\left\|V_{x_i} - V_{y_i}\right\|_p \\ &\qquad + \left\|V_1 - V_{i+n-f-b}\right\|_p\Big)\end{aligned}$$

We can then compute the expected value of this distance, and with $\mathbb{E}\Big[\widehat{G}\Big] \triangleq G$ and the Jensen's inequality:

$$\begin{aligned}\left\|\mathbb{E}[F] - G\right\|_p &\leq \mathbb{E}\left\|F - \widehat{G}\right\|_p \\ &\leq \frac{1}{n-f} \sum_{i=1}^{b} \bar{\sigma} + \bar{\sigma} \\ &\leq \frac{2\,b\,\bar{\sigma}}{n-f} \leq \frac{2\,f\,\bar{\sigma}}{n-f}\end{aligned}$$

So, under the assumption that $2\,f\,\bar{\sigma} < (n-f)\left\|G\right\|_p$, we verify that $\left\|\mathbb{E}[F] - G\right\|_p < \left\|G\right\|_p$, and so: $\langle\mathbb{E}[F]\,, G\rangle > 0$.

Point *2.* can also be verified formally, $\forall r \in \{\,2, 3, 4\,\}$:

$$\mathbb{E}\left\|F\right\|_p^r \leq \frac{n-f-b}{n-f}\mathbb{E}\left\|\mathcal{G}\right\|_p^r + \frac{1}{n-f}\sum_{i=1}^{b}\mathbb{E}\left\|B_i\right\|_p^r$$

Then, by using the binomial theorem twice:

$$\begin{aligned}\left\|B_i\right\|_p^r &\leq \sum_{r_1+r_2=r}\binom{r}{r_1}\left\|B_i - V_k\right\|_p^{r_1}\left\|V_k\right\|_p^{r_2} \\ &\quad \text{with } k \in [1\mathinner{..}n-f-d] \\ \left\|B_i - V_k\right\|_p^{r_1} &\leq \left\|V_x - V_y\right\|_p^{r_1} \\ &\leq \sum_{r_3+r_4=r_1}\binom{r_1}{r_3}\left\|V_x\right\|_p^{r_3}\left\|V_y\right\|_p^{r_4}\end{aligned}$$

Finally, as $(V_1 \ldots V_{n-f})$ are *independent, identically distributed* random variables following the same distribution $\mathcal{G}$, we have that $\forall\,(i,j) \in [1\mathinner{..}n-f]^2\,,\, i \neq j$, $\mathbb{E}\Big[\left\|V_i\right\|_p^{r_1}\left\|V_j\right\|_p^{r_2}\Big] = \mathbb{E}\left\|\mathcal{G}\right\|_p^{r_1}\cdot\mathbb{E}\left\|\mathcal{G}\right\|_p^{r_2}$, and so $\mathbb{E}\left\|B_i\right\|_p^r$ is bounded as described in point *2.* of definition 1.

B. Approximation of α_m, with $p \in \mathbb{N}^*$

B.1. Prior conventions and assumptions

Let remind: $\forall i \in [1\mathinner{..}n-f]\,,\, V_i = \left(v_1^{(i)} \ldots v_d^{(i)}\right) \sim \mathcal{G}$. We model each coordinate as a *normal distribution*:

$$\begin{aligned}&\forall j \in [1\mathinner{..}d]\,,\, \exists\,(\mu_j, \sigma_j) \in \mathbb{R}^2, \\ &\forall i \in [1\mathinner{..}n-f]\,,\, v_j^{(i)} \sim \mathcal{N}\left(\mu_j, {\sigma_j}^2\right)\end{aligned}$$

We assume $d \gg 1$, and we will write $\bar{\delta}$ for:

$$\begin{aligned}\forall\,(i,j) \in [1\mathinner{..}n-f]^2\,,\, i \neq j,\ \bar{\delta} &= \frac{1}{d}\sum_{k=1}^{d}\mathbb{E}\left|v_k^{(i)} - v_k^{(j)}\right| \\ &= \frac{2}{d\sqrt{\pi}}\sum_{k=1}^{d}\sigma_k\end{aligned}$$

and note that: $\displaystyle \frac{1}{d}\sum_{k=1}^{d}\mathbb{E}\left|v_k^{(i)} - \mu_k\right| = \frac{\sqrt{2}}{d\sqrt{\pi}}\sum_{k=1}^{d}\sigma_k = \frac{\bar{\delta}}{\sqrt{2}}$

Then, $\forall\,(i,j) \in [1\mathinner{..}n-f]^2\,,\, i \neq j$, we can approximate:

$$\begin{aligned}\left\|V_i - V_j\right\|_p &= \left(\sum_{k=1}^{d}\left|v_k^{(i)} - v_k^{(j)}\right|^p\right)^{\frac{1}{p}} \\ &\approx \left(d\,\bar{\delta}^p\right)^{\frac{1}{p}}\end{aligned}$$

Let $E = (0 \ldots 0, 1, 0 \ldots 0) \in \mathbb{R}^d$ the attacked coordinate. Then, with $\alpha_m > 0$, $B = \overline{V} + \alpha_m\,E$, we can approximate:

$$\begin{aligned}\left\|B - V_i\right\|_p &= \left(\left(\sum_{k=1}^{d}\left|v_k^{(i)} - \bar{v_k}\right|^p\right)\right. \\ &\qquad \left. - \left|v_e^{(i)} - \bar{v_e}\right|^p + \left|v_e^{(i)} - \bar{v_e} + \alpha_m\right|^p\right)^{\frac{1}{p}} \\ &\approx \left({\alpha_m}^p + \sum_{k=1}^{d}\left|v_k^{(i)} - \mu_k\right|^p\right)^{\frac{1}{p}} \\ &\approx \left({\alpha_m}^p + d\left(\frac{\bar{\delta}}{\sqrt{2}}\right)^p\right)^{\frac{1}{p}}\end{aligned}$$

B.2. Attack against Brute

We only study the *worst case* scenario, where $n = 2f + 1$, maximizing the proportion of Byzantine workers.

Assuming B is selected by Brute:

$$\begin{aligned}&B \in \mathcal{S} \\ \Rightarrow\ &\exists\,(x,y) \in [1\mathinner{..}n-f]^2\,,\, x \neq y,\end{aligned}$$

$$\forall k \in [1\,..\,n-f-b]\,,\ \|B - V_k\|_p < \|V_x - V_y\|_p$$
$$\rightsquigarrow\ \left(\alpha_m{}^p + d\left(\frac{\bar{\delta}}{\sqrt{2}}\right)^p\right)^{\frac{1}{p}} < \left(d\,\bar{\delta}^p\right)^{\frac{1}{p}}$$
$$\rightsquigarrow\ \alpha_m < \left(\left(1 - \frac{1}{\sqrt{2}^p}\right) d\right)^{\frac{1}{p}} \bar{\delta}$$

This is a *necessary*, approximated condition. It is only to give broad insights on the relation between some hyper–parameters and α_m: with p, q constants, $\alpha_m = \mathcal{O}\left(\bar{\delta}\ \sqrt[p]{d}\right)$.

B.3. Attack against Krum/GeoMed

We only study the *worst case* scenario, where $n = 2f + 3$, maximizing the proportion of Byzantine workers.
Let $q \in \{\,1, 2\,\}$, $q = 1$ for GeoMed and $q = 2$ for Krum.

First, we approximate the Byzantine submission's score:

$$s\left(B\right) \approx 2\,\|B - V_i\|_p^q$$
$$\approx 2\left(\alpha_m{}^p + d\left(\frac{\bar{\delta}}{\sqrt{2}}\right)^p\right)^{\frac{q}{p}}$$

$\forall i \in [1\,..\,n-f]$, let $b \in [0\,..\,f]$ be how many B belongs to the $n - f - 2$ closest vectors to V_i. Then the score of V_i is:

$$s\left(V_i\right) \approx b\,\|B - V_i\|_p^q + \left(f + 1 - b\right) \|V_j - V_i\|_p^q$$
$$\approx b\left(\alpha_m{}^p + d\left(\frac{\bar{\delta}}{\sqrt{2}}\right)^p\right)^{\frac{q}{p}} + \left(f + 1 - b\right) \left(d\,\bar{\delta}^p\right)^{\frac{q}{p}}$$

Finally, B is selected $\Rightarrow\ \forall i \in [1\,..\,n-f]\,,\ s\left(B\right) \lessapprox s\left(V_i\right)$

$$\underset{\forall i}{\overset{}{\underset{\uparrow}{\Rightarrow}}}\ \left(2 - b\right)\left(\alpha_m{}^p + d\left(\frac{\bar{\delta}}{\sqrt{2}}\right)^p\right)^{\frac{q}{p}} \lessapprox \left(f + 1 - b\right) \left(d\,\bar{\delta}^p\right)^{\frac{q}{p}}$$

$$\underset{\exists i}{\underset{\uparrow}{\Rightarrow}}\ \alpha_m \lessapprox \left(\left(\frac{f + 1 - b}{2 - b}\right)^{\frac{p}{q}} - \frac{1}{\sqrt{2}^p}\right)^{\frac{1}{p}} d^{\frac{1}{p}}\,\bar{\delta}$$

This last implication is always true: there *must* be at least one non–Byzantine vector V_j for which $b \in \{\,0, 1\,\}$; else α_m could increase unbounded, which would be absurd.

In conclusion, with p, q constants: $\alpha_m = \mathcal{O}\left(\bar{\delta}\ \sqrt[q]{f}\ \sqrt[p]{d}\right)$.

C. Supplementary experiments

C.1. Attack on Brute, Krum and GeoMed

On MNIST, here we use $\eta_0 = 1$, $r_\eta = 10000$, a batch size of 83 images (256 for Brute), and for the workers:

Krum/GeoMed	30 non–Byzantines + 27 Byzantines
Brute	6 non–Byzantines + 5 Byzantines
Average	30 non–Byzantines + 0 Byzantines

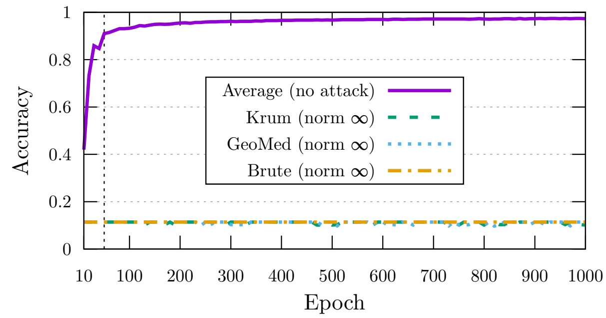

Figure 1. MNIST: accuracy on the testing set up to epoch 1000, comparing the presented aggregation rules under our attack. The attack was maintained only up to epoch 50 (dotted line). The *average* is the reference: it is the accuracy the model would have shown if only non–Byzantine gradients had been selected.

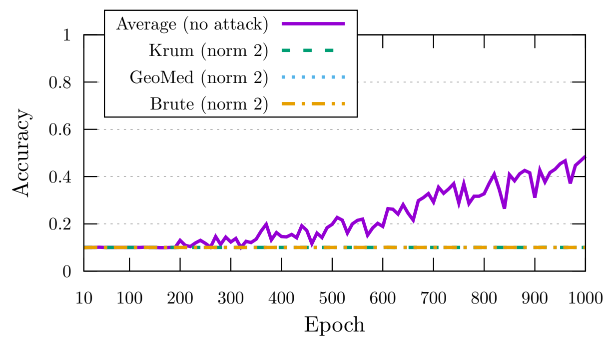

Figure 2. CIFAR–10: accuracy on the testing set up to epoch 1000, comparing the presented aggregation rules under our attack. The *average* is the reference: it is the accuracy the model would have shown if only non–Byzantine gradients had been selected.

On CIFAR–10, we use $\eta_0 = 0.5$, $r_\eta = 2000$, a batch size of 128 images (256 for Brute), and for the worker counts:

Krum/GeoMed	21 non–Byzantines + 18 Byzantines
Brute	6 non–Byzantines + 5 Byzantines
Average	21 non–Byzantines + 0 Byzantines

In Figure 1, the attack is maintained only up to 50 epochs. The attack variant for ℓ_∞ norm–based gradient aggregation rules exhibited a very strong impact. None of the presented gradient aggregation rules prevented the stochastic gradient descent from being *pushed* and remaining in a sub–space of *ineffective* models, and for at least 1000 epochs.

In Figure 2, the attack is never stopped. Again, none of the presented gradient aggregation rules prevented the stochastic gradient descent from being *pushed* and remaining in a sub–space of *ineffective* models, for at least 1000 epochs.