Variance Reduction is an Antidote to Byzantines: Better Rates, Weaker Assumptions and Communication Compression as a Cherry on the Top

Eduard Gorbunov
MBZUAI

Samuel Horváth
MBZUAI

Peter Richtárik
KAUST

Gauthier Gidel
Mila & UdeM
Canada CIFAR AI Chair

Federated Learning One-World Seminar

February 7, 2024

I am on the job market for Assistant Professor position!

- Postdoc at MBZUAI (Abu Dhabi, UAE) hosted by Samuel Horváth and Martin Takáč (from September 2022)
- Previous positions: - junior researcher at MIPT (2020-2022)
  - remote postdoc at Mila (2022), hosted by Gauthier Gidel
- PhD in Computer Science, MIPT (2020-2021), Supervisors: Alexander Gasnikov and Peter Richtárik
- Research interests: Stochastic Optimization, Distributed Optimization, Variational Inequalities, Derivative-Free Optimization
- Selected awards: Ilya Segalovich Award 2019 (highly selective), best reviewer award (ICLR 2021, ICML 2021-2022, NeurIPS 2020-2022)
- **See more about me on my website: eduardgorbunov.github.io**

E. Gorbunov, S. Horváth, P. Richtárik, G. Gidel. *Variance Reduction is an Antidote to Byzantines: Better Rates, Weaker Assumptions and Communication Compression as a Cherry on the Top* (ICLR 2023)

Samuel Horváth
Assistant professor at MBZUAI

Peter Richtárik
Professor at KAUST

Gauthier Gidel
Assistant professor at Mila and UdeM

Outline

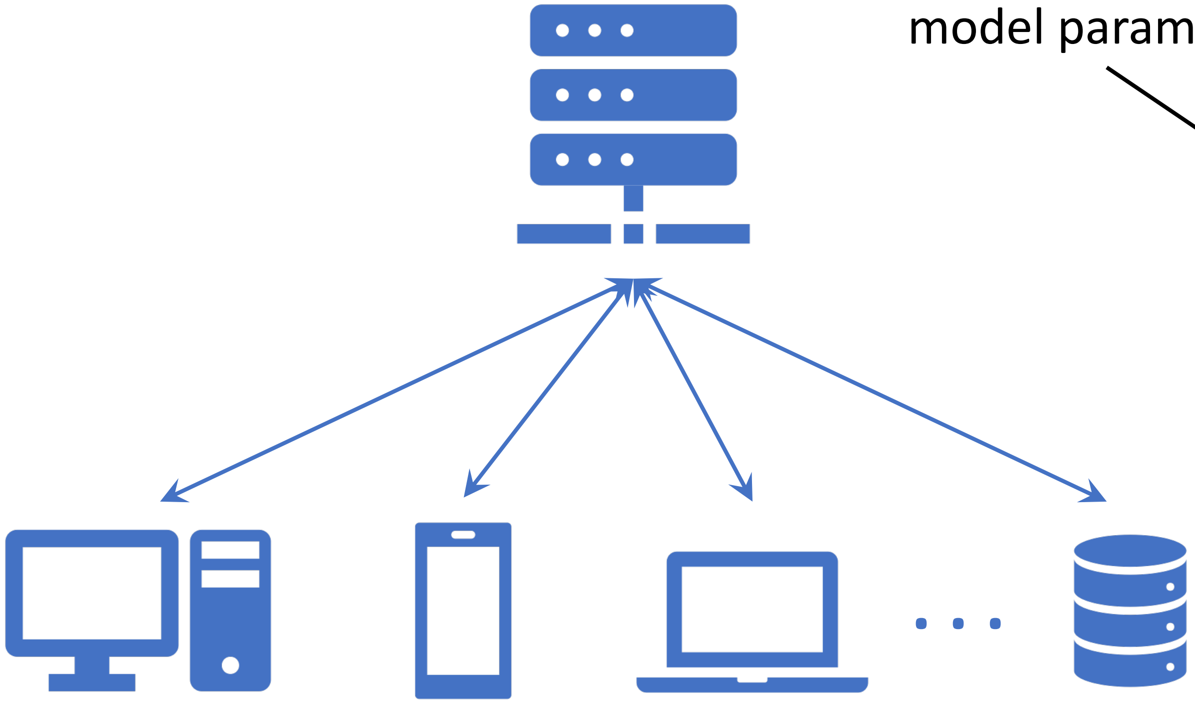
1. Byzantine-robust training

2. Robust aggregation

3. Variance reduction and Byzantine-robustness

Byzantine-Robust Training

The Problem

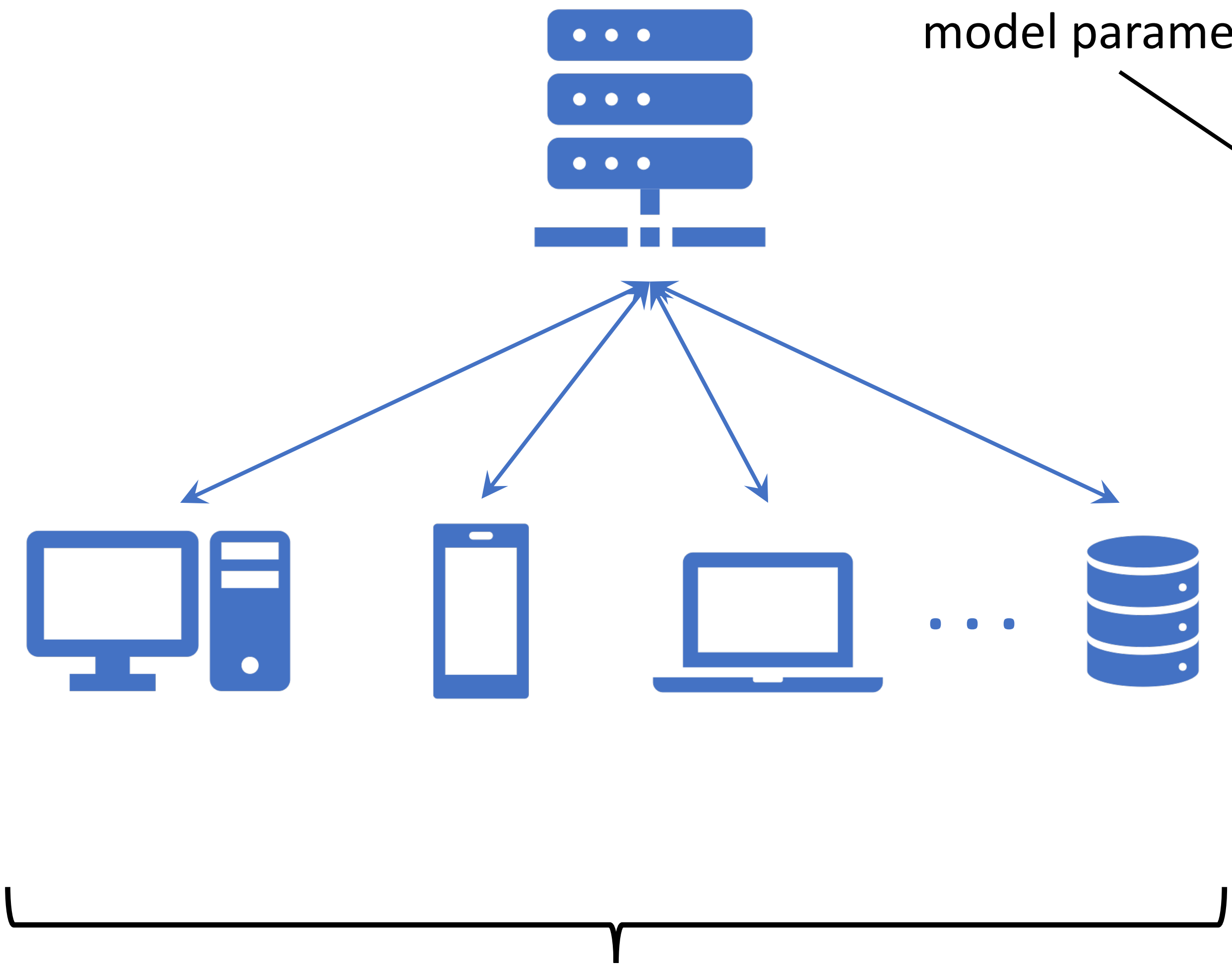


model parameters

$$\min_{x \in \mathbb{R}^d} \left\{ f(x) := \frac{1}{n} \sum_{i=1}^{n} f_i(x) \right\}$$

# of parameters

The Problem



$$\min_{x \in \mathbb{R}^d} \left\{ f(x) := \frac{1}{n} \sum_{i=1}^{n} f_i(x) \right\}$$

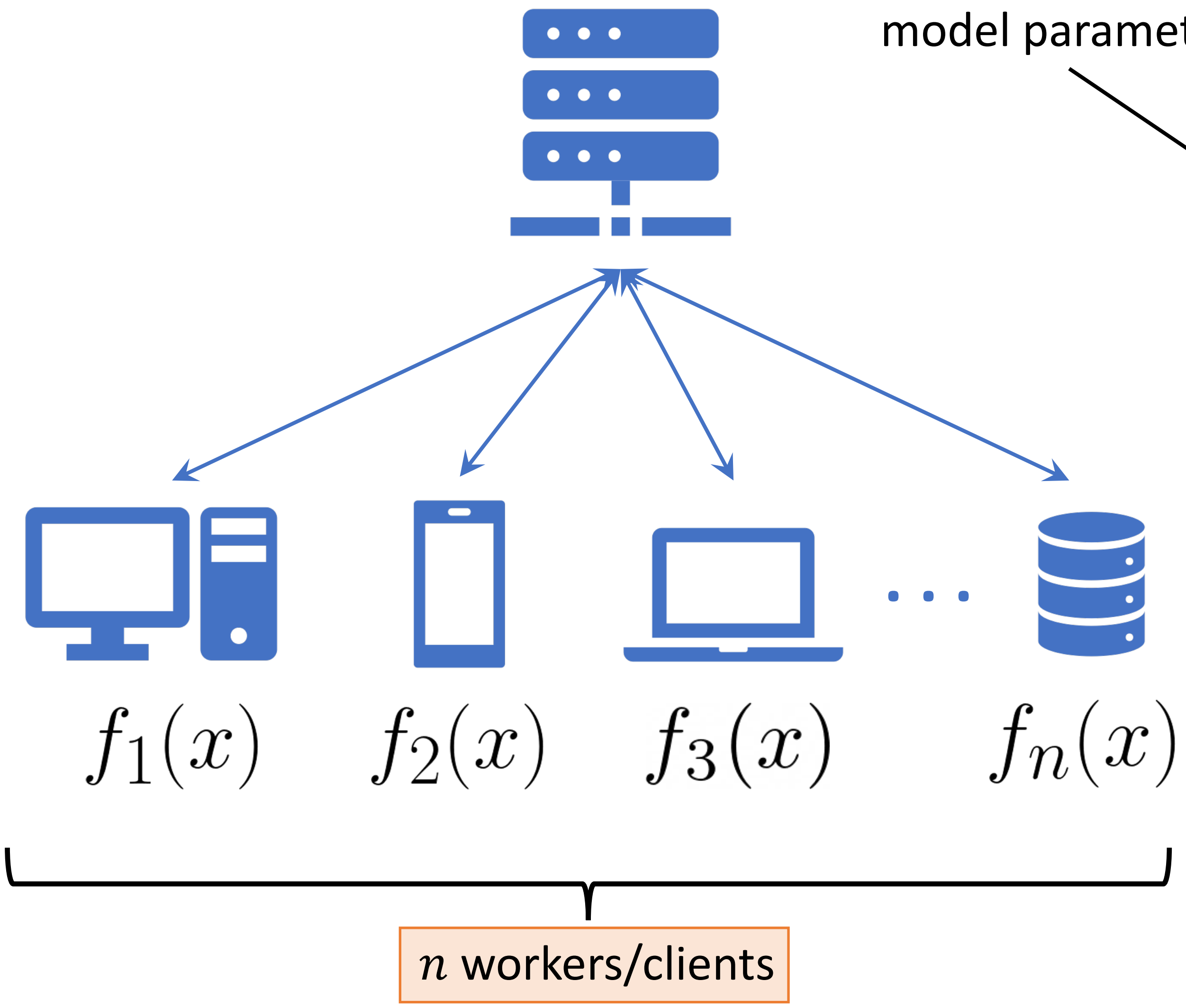
model parameters

# of parameters

# of workers/clients

n workers/clients

The Problem



$$\min_{x \in \mathbb{R}^d} \left\{ f(x) := \frac{1}{n} \sum_{i=1}^{n} f_i(x) \right\}$$

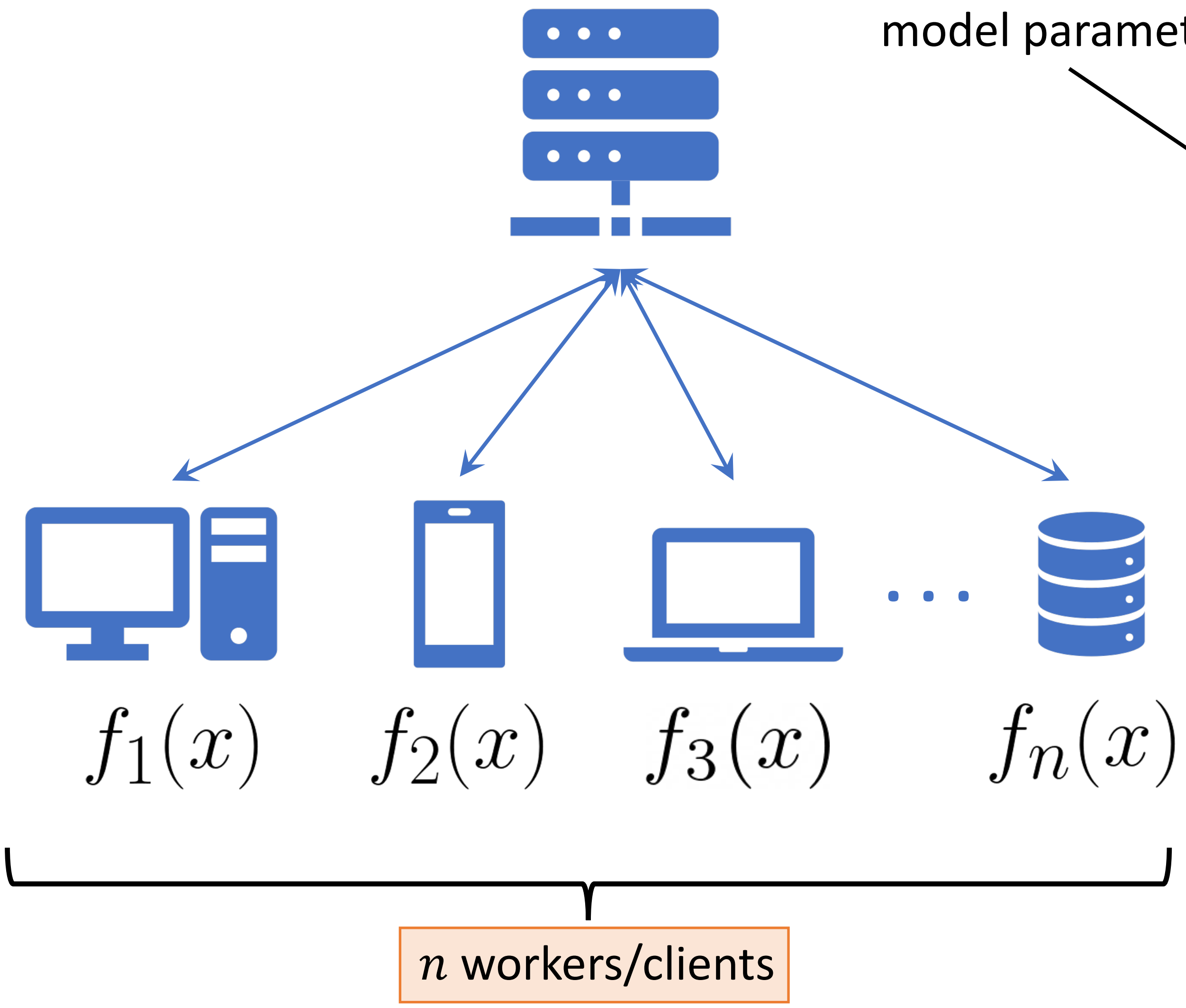
model parameters

# of parameters

# of workers/clients

loss on the data accessible on worker i

The Problem



$$\min_{x \in \mathbb{R}^d} \left\{ f(x) := \frac{1}{n} \sum_{i=1}^{n} f_i(x) \right\}$$

model parameters

\# of parameters

\# of workers/clients

loss on the data accessible on worker i

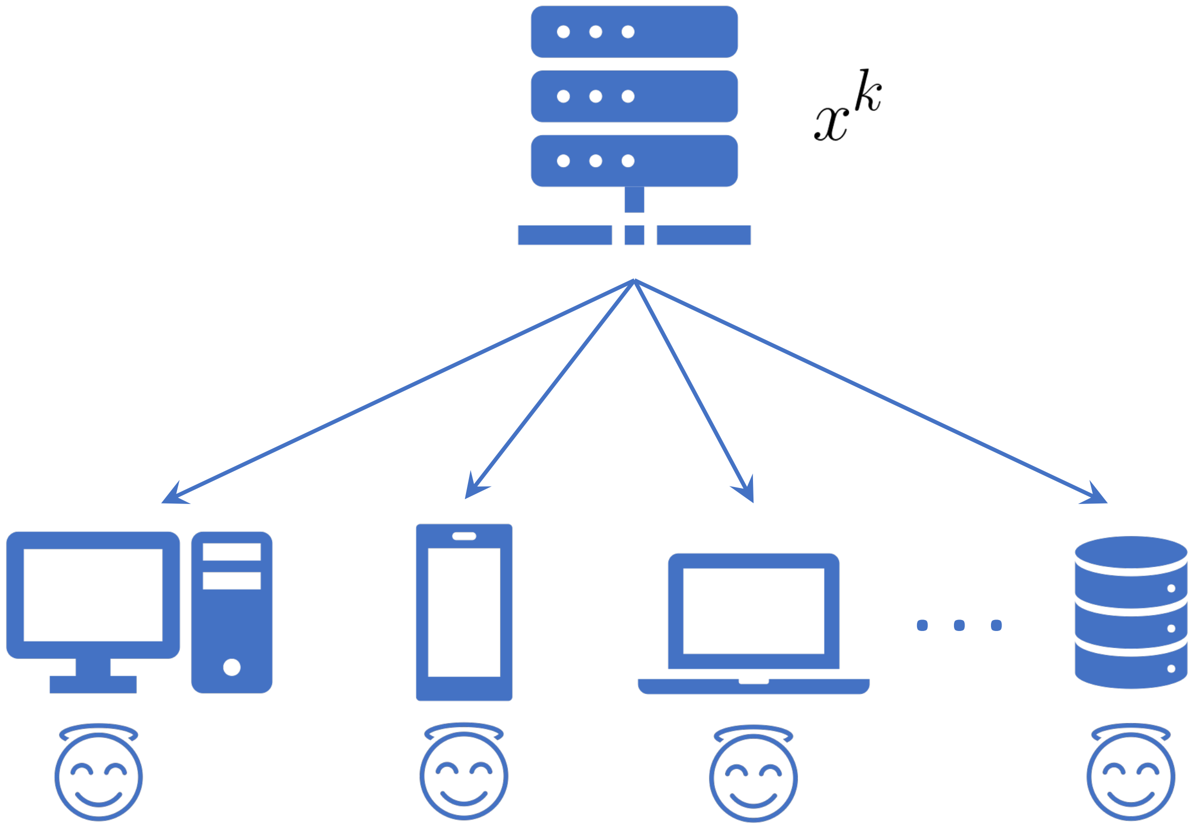
Key features:

- The problem is hard to solve for one client
- Clients do not know each other

Parallel SGD

Iteration k:

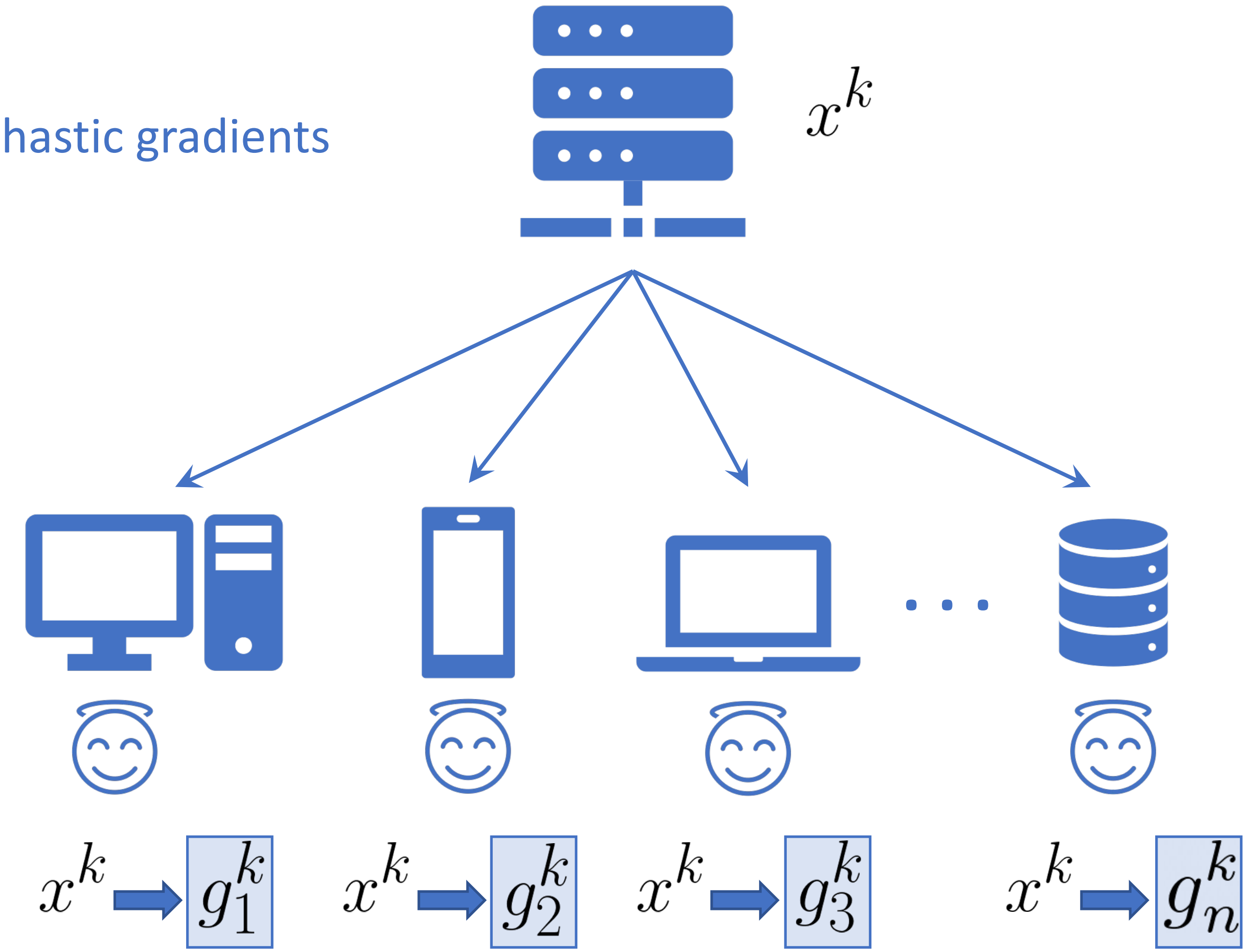
1. Server broadcasts x^k



Parallel SGD

Iteration k:

1. Server broadcasts x^k
2. Workers compute stochastic gradients

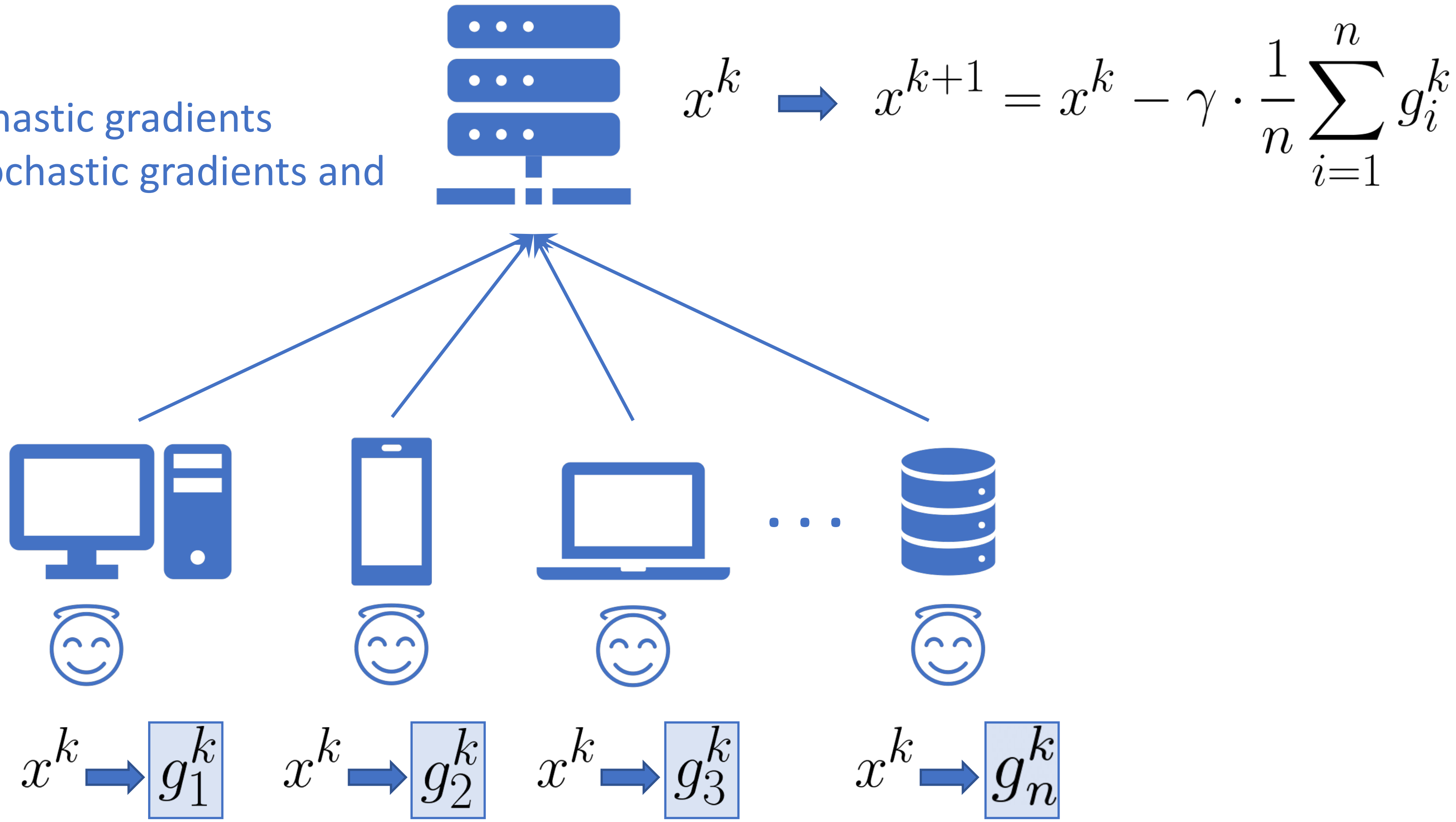


$$\mathbb{E}_k[g_i^k] = \nabla f_i(x^k)$$

Parallel SGD

Iteration k:

1. Server broadcasts x^k
2. Workers compute stochastic gradients
3. Server averages the stochastic gradients and makes an SGD step

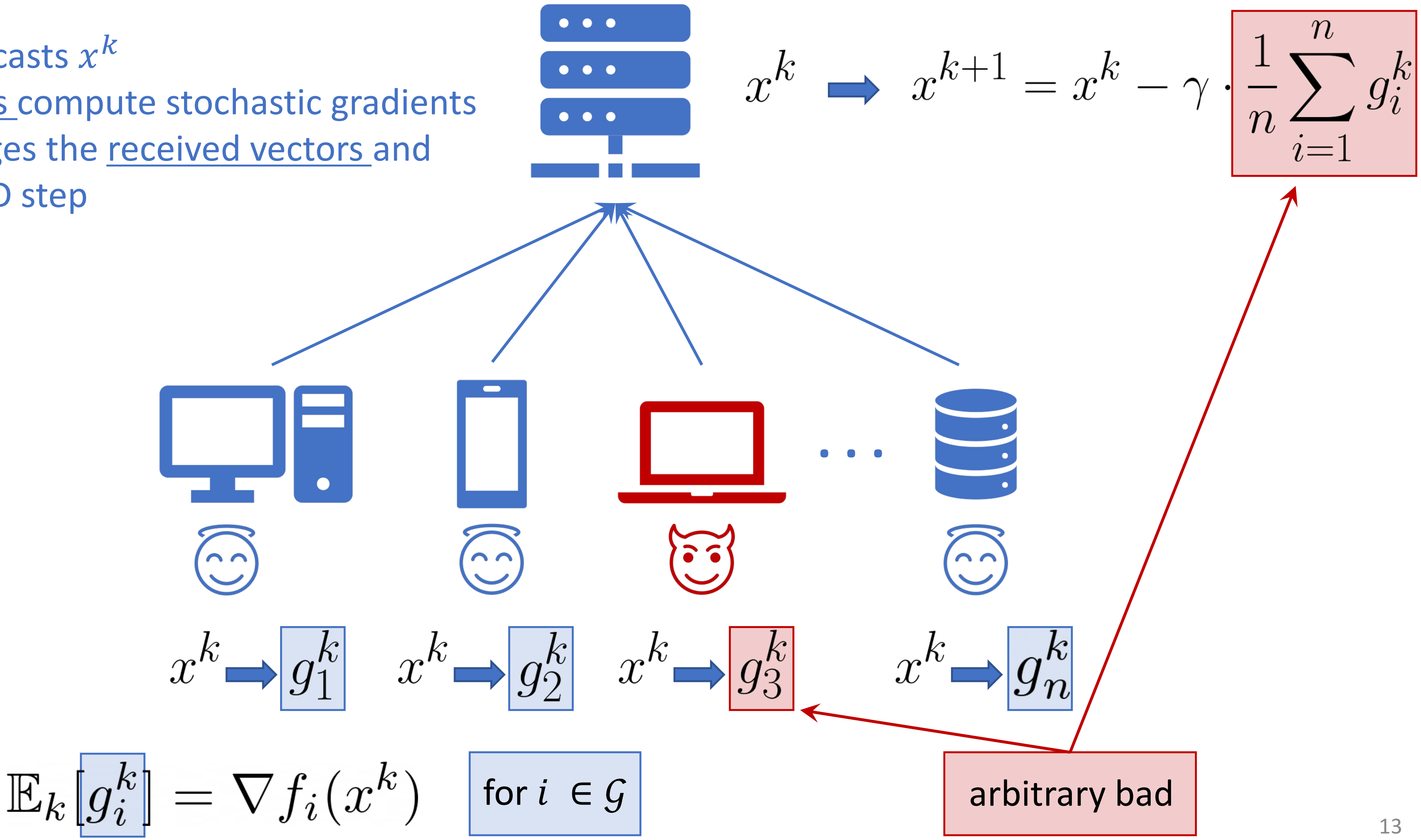


$$\mathbb{E}_k[g_i^k] = \nabla f_i(x^k)$$

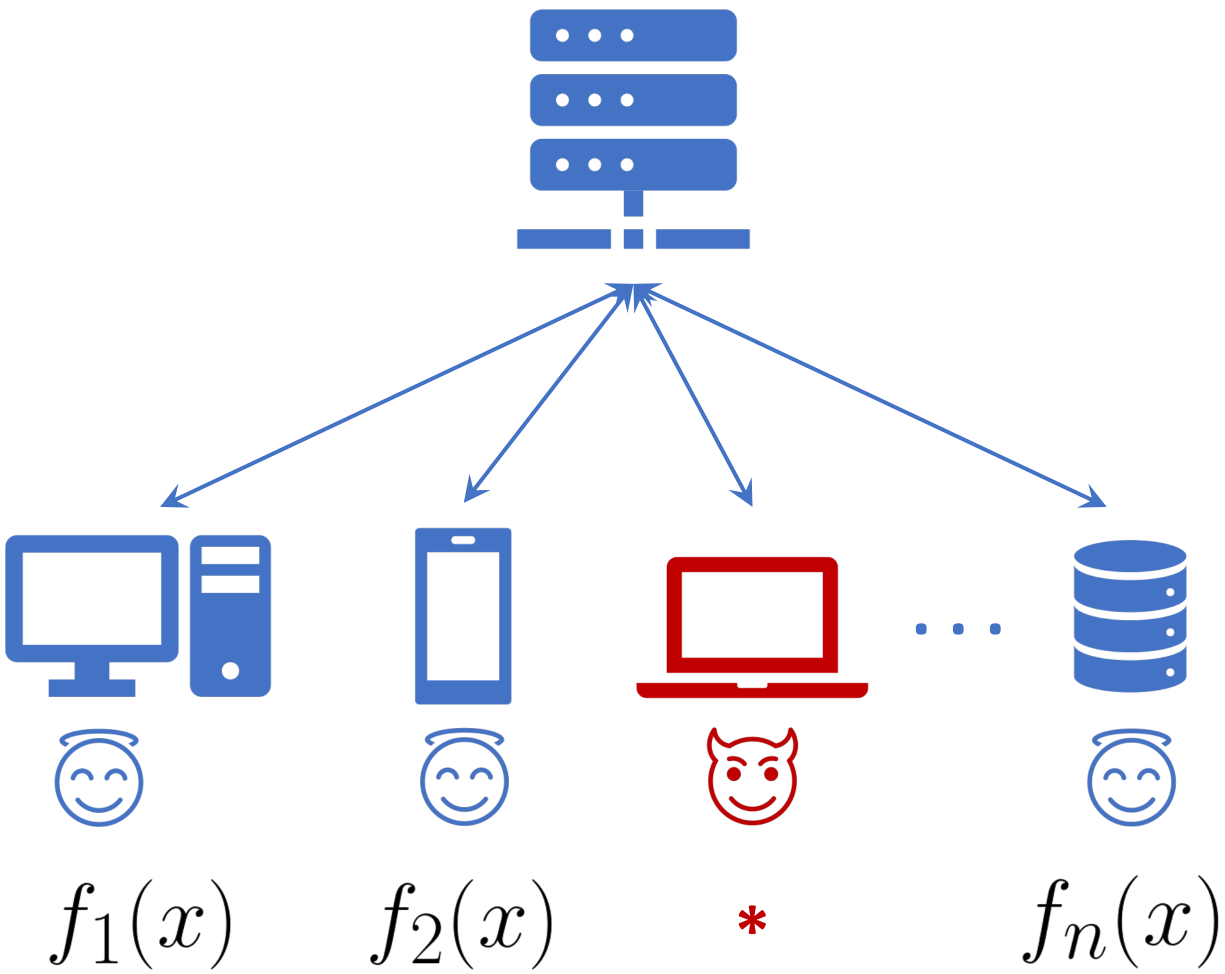
Parallel SGD Is Fragile

Iteration k:

1. Server broadcasts x^k
2. Good workers compute stochastic gradients
3. Server averages the received vectors and makes an SGD step



The Refined Problem Formulation

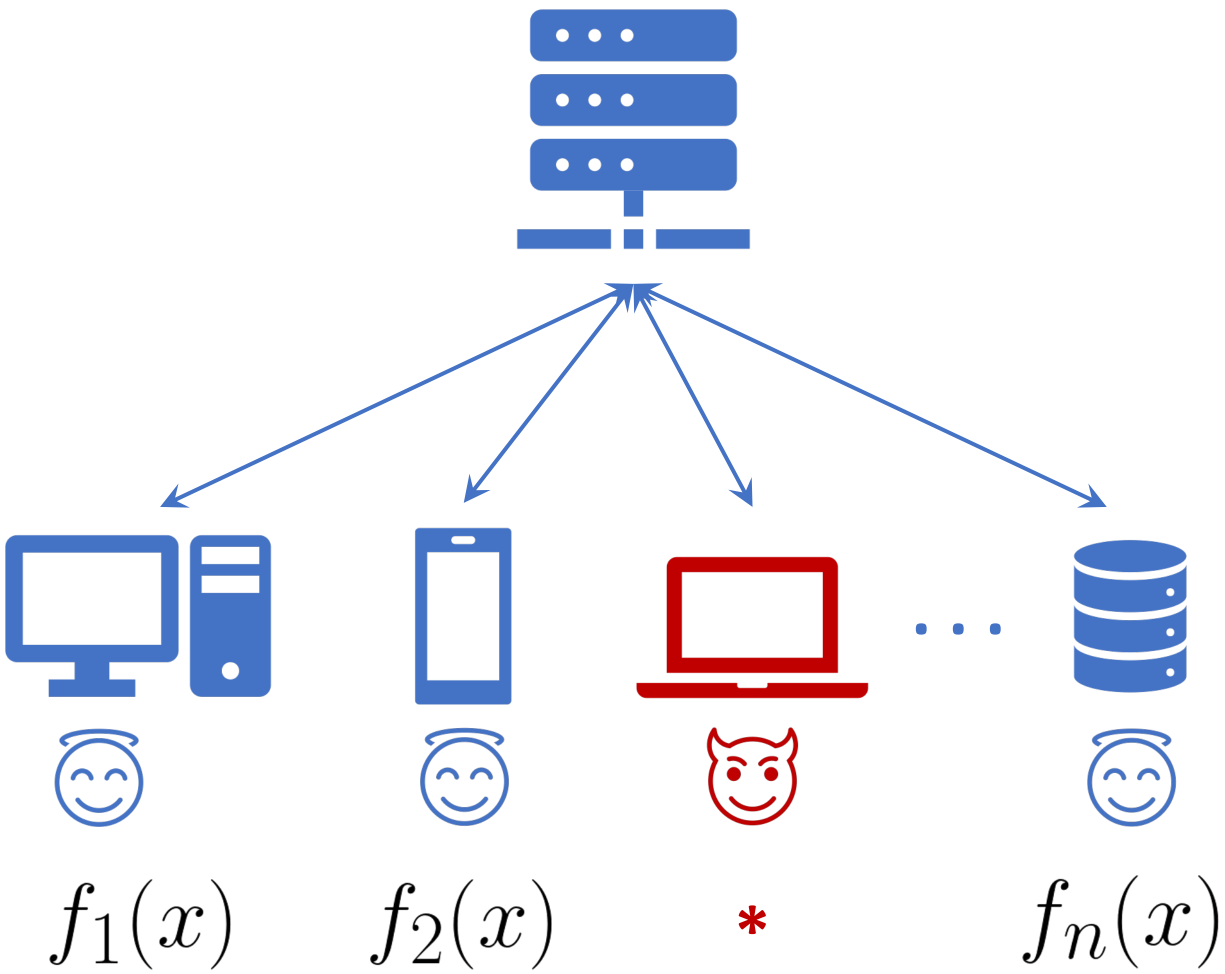


$$\min_{x \in \mathbb{R}^d} \left\{ f(x) := \frac{1}{G} \sum_{i \in \mathcal{G}} f_i(x) \right\}$$

Good workers form the majority:

- $\mathcal{G}$ – good workers
- $\mathcal{B}$ – Byzantines (see the page "Byzantine fault" in Wikipedia)
- $\mathcal{G} \sqcup \mathcal{B} = [n], \quad |\mathcal{G}| = G, \quad |\mathcal{B}| = B$
- $B \leq \delta n, \quad \delta < {}^1\!/_2$
- Byzantines are omniscient

The Refined Problem Formulation



$$\min_{x \in \mathbb{R}^d} \left\{ f(x) := \frac{1}{G} \sum_{i \in \mathcal{G}} f_i(x) \right\}$$

Good workers form the majority:

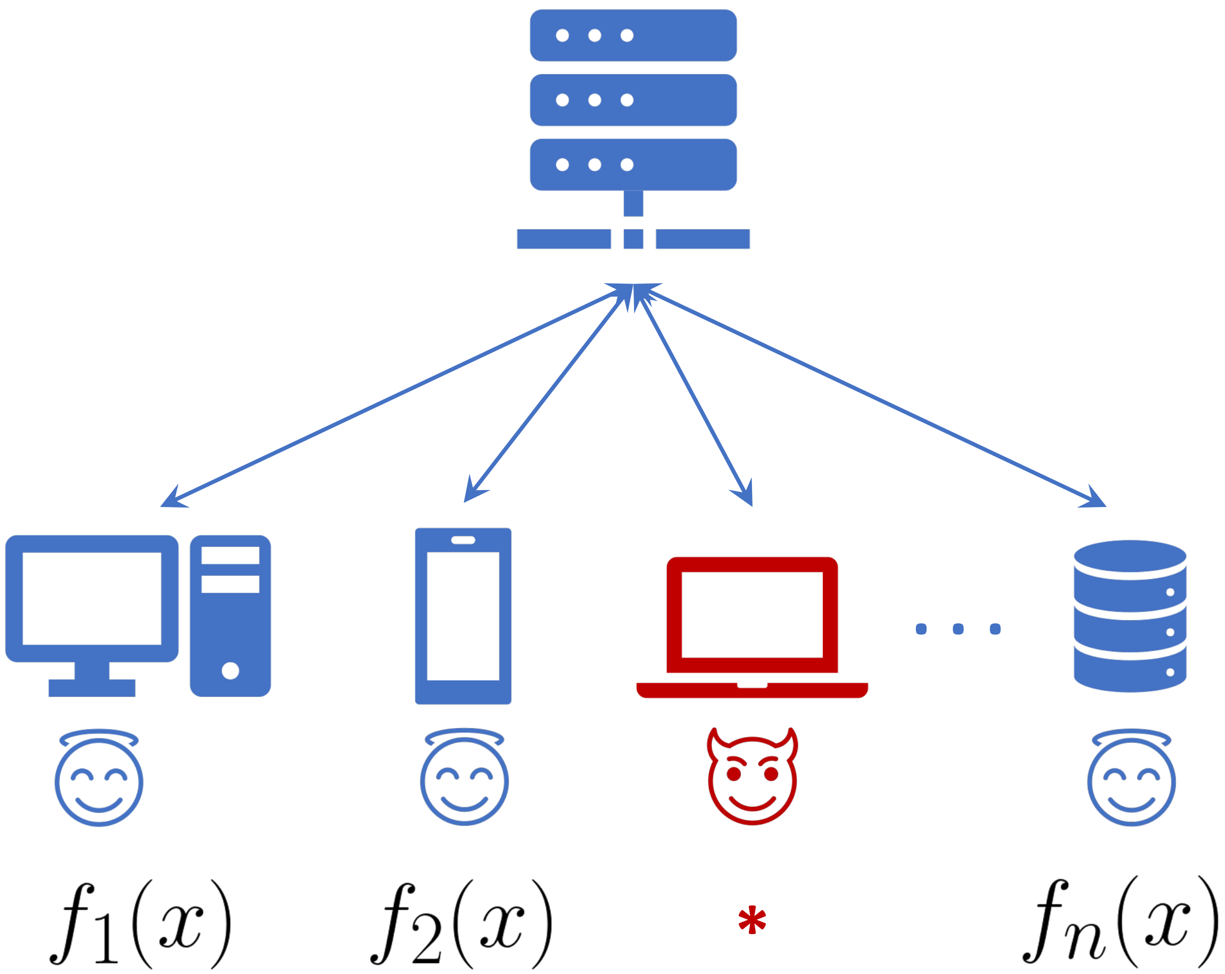
- $\mathcal{G}$ – good workers
- $\mathcal{B}$ – Byzantines (see the page “Byzantine fault” in Wikipedia)
- $\mathcal{G} \sqcup \mathcal{B} = [n], \quad |\mathcal{G}| = G, \quad |\mathcal{B}| = B$
- $B \leq \delta n, \quad \delta < 1/2$
- Byzantines are omniscient

On the heterogeneity:

- Loss functions on good peers cannot be arbitrary heterogeneous
- In this talk, we will assume that

$$\forall\, i \in \mathcal{G} \;\rightarrow\; f_i = f$$

The Refined Problem Formulation



Question: how to solve such problems?

$$\min_{x\in\mathbb{R}^d}\left\{f(x) := \frac{1}{G}\sum_{i\in\mathcal{G}} f_i(x)\right\}$$

Good workers form the majority:

- $\mathcal{G}$ – good workers
- $\mathcal{B}$ – Byzantines (see the page "Byzantine fault" in Wikipedia)
- $\mathcal{G} \sqcup \mathcal{B} = [n], \quad |\mathcal{G}| = G, \quad |\mathcal{B}| = B$
- $B \leq \delta n, \quad \delta < {}^1/_2$
- Byzantines are omniscient

On the heterogeneity:

- Loss functions on good peers cannot be arbitrary heterogeneous
- In this talk, we will assume that

$$\forall\, i \in \mathcal{G} \;\rightarrow\; f_i = f$$

Robust Aggregation

“Middle-Seekers” Aggregators

Natural idea: replace the averaging with more robust aggregation rule!

$$x^{k+1} = x^k - \gamma g^k \quad \Rightarrow \quad x^{k+1} = x^k - \gamma \widehat{g}^k$$

$$g^k = \frac{1}{n}\sum_{i=1}^{n} g_i^k \quad \Rightarrow \quad \widehat{g}^k = \mathtt{RAgg}\left(g_1^k, g_2^k, \ldots, g_n^k\right)$$

Question: how to choose aggregator?

"Middle-Seekers" Aggregators

- Geometric median (RFA):

Pillutla, K., Kakade, S. M., & Harchaoui, Z. (2019). Robust aggregation for federated learning. arXiv preprint arXiv:1912.13445.

$$\widehat{g}^k = \arg\min_{g \in \mathbb{R}^d} \sum_{i=1}^{n} \|g - g_i^k\|_2$$

"Middle-Seekers" Aggregators

- Geometric median (RFA):

Pillutla, K., Kakade, S. M., & Harchaoui, Z. (2019). Robust aggregation for federated learning. arXiv preprint arXiv:1912.13445.

$$\widehat{g}^k = \arg\min_{g \in \mathbb{R}^d} \sum_{i=1}^{n} \|g - g_i^k\|_2$$

- Coordinate-wise median (CM):

Yin, D., Chen, Y., Kannan, R., & Bartlett, P. (2018, July). Byzantine-robust distributed learning: Towards optimal statistical rates. *In International Conference on Machine Learning* (pp. 5650-5659). PMLR.

$$\widehat{g}^k = \arg\min_{g \in \mathbb{R}^d} \sum_{i=1}^{n} \|g - g_i^k\|_1$$

“Middle-Seekers” Aggregators

- Geometric median (RFA):

Pillutla, K., Kakade, S. M., & Harchaoui, Z. (2019). Robust aggregation for federated learning. arXiv preprint arXiv:1912.13445.

$$\widehat{g}^k = \arg\min_{g \in \mathbb{R}^d} \sum_{i=1}^{n} \|g - g_i^k\|_2$$

- Coordinate-wise median (CM):

Yin, D., Chen, Y., Kannan, R., & Bartlett, P. (2018, July). Byzantine-robust distributed learning: Towards optimal statistical rates. *In International Conference on Machine Learning* (pp. 5650-5659). PMLR.

$$\widehat{g}^k = \arg\min_{g \in \mathbb{R}^d} \sum_{i=1}^{n} \|g - g_i^k\|_1$$

- Krum estimator:

Blanchard, P., El Mhamdi, E. M., Guerraoui, R., & Stainer, J. (2017, December). Machine learning with adversaries: Byzantine tolerant gradient descent. *In Proceedings of the 31st International Conference on Neural Information Processing Systems* (pp. 118-128).

$$\widehat{g}^k = \arg\min_{g \in \{g_1^k, \ldots, g_n^k\}} \sum_{i \in \mathcal{N}_{n-B-2}(g)} \|g - g_i^k\|_2^2$$

indices of the closest $n - B - 2$ workers to g

Simple Example When "Middle-Seekers" Are Good

Let $d = 1, \mathcal{G} = \{1, 2, 3, 4\}, \mathcal{B} = \{5, 6\}, g_1^k = 1.5, g_2^k = 2, g_3^k = 2.5, g_4^k = 3$, and Byzantines are trying to shift the estimator via sending $g_5^k = g_6^k = 1000$. In this case,

Simple Example When "Middle-Seekers" Are Good

Let $d = 1, \mathcal{G} = \{1,2,3,4\}, \mathcal{B} = \{5,6\}, g_1^k = 1.5, g_2^k = 2, g_3^k = 2.5, g_4^k = 3$, and Byzantines are trying to shift the estimator via sending $g_5^k = g_6^k = 1000$. In this case,

- Average of the good workers: $\bar{g}^k = \frac{1}{4}\sum_{i=1}^{4} g_4^k = 2.25$
- Average estimator: $g^k = \frac{1}{6}\sum_{i=1}^{6} g_i^k = 335$
- Median: $\hat{g}^k$ – any number from $[2.5, 3]$
- Krum estimator: $\hat{g}^k = 2$ or 2.5

Simple Example When "Middle-Seekers" Are Good

Let $d = 1, \mathcal{G} = \{1,2,3,4\}, \mathcal{B} = \{5,6\}, g_1^k = 1.5, g_2^k = 2, g_3^k = 2.5, g_4^k = 3$, and Byzantines are trying to shift the estimator via sending $g_5^k = g_6^k = 1000$. In this case,

- Average of the good workers: $\bar{g}^k = \frac{1}{4}\sum_{i=1}^4 g_4^k = 2.25$
- Average estimator: $g^k = \frac{1}{6}\sum_{i=1}^6 g_i^k = 335$
- Median: $\hat{g}^k$ – any number from $[2.5, 3]$
- Krum estimator: $\hat{g}^k = 2$ or 2.5

"Middle-seekers" can be good for reducing the effect of outliers

When “Middle-Seekers” Can Be Bad

Karimireddy, S. P., He, L., & Jaggi, M. (2021, July). Learning from history for byzantine robust optimization. *In International Conference on Machine Learning* (pp. 5311-5319). PMLR.

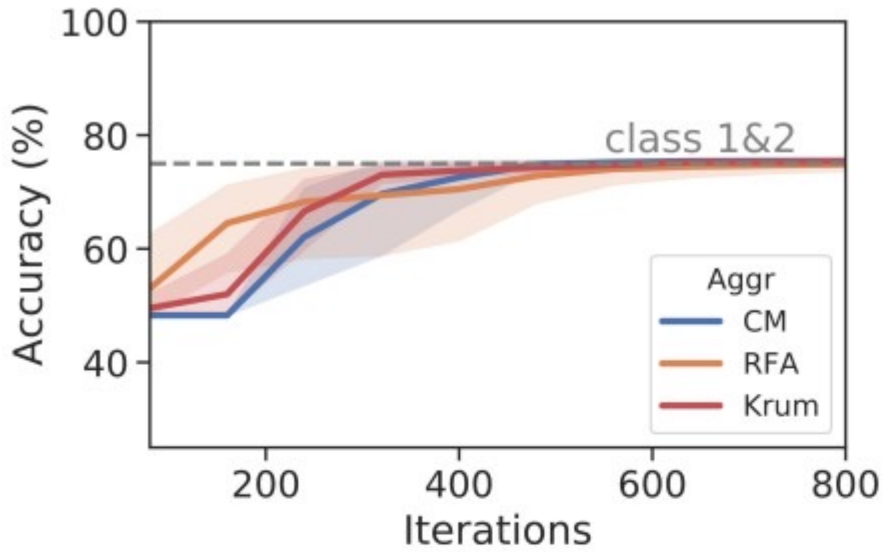


Figure 1: Failure of existing methods on imbalanced MNIST dataset. Only the head classes (class 1 and 2 here) are learnt, and the rest 8 classes are ignored. See Sec. 7.1.

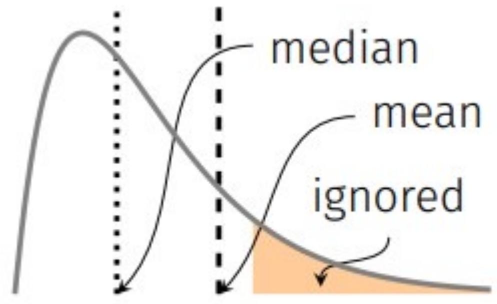
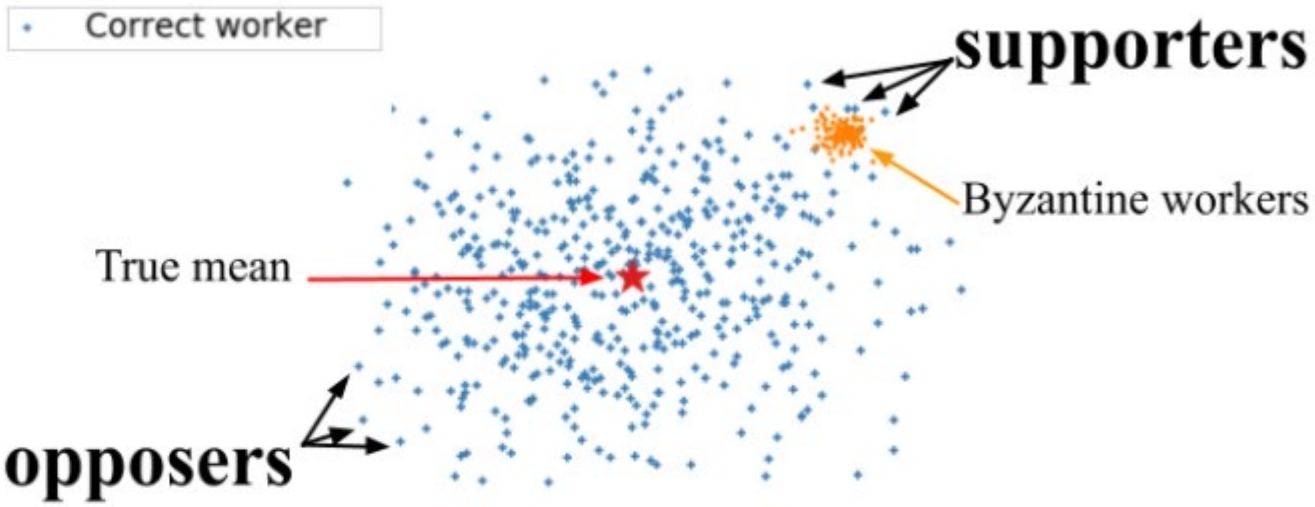


Figure 2: For fat-tailed distributions, median based aggregators ignore the tail. This bias remains even if we have infinite samples.

A Little Is Enough (ALIE) Attack

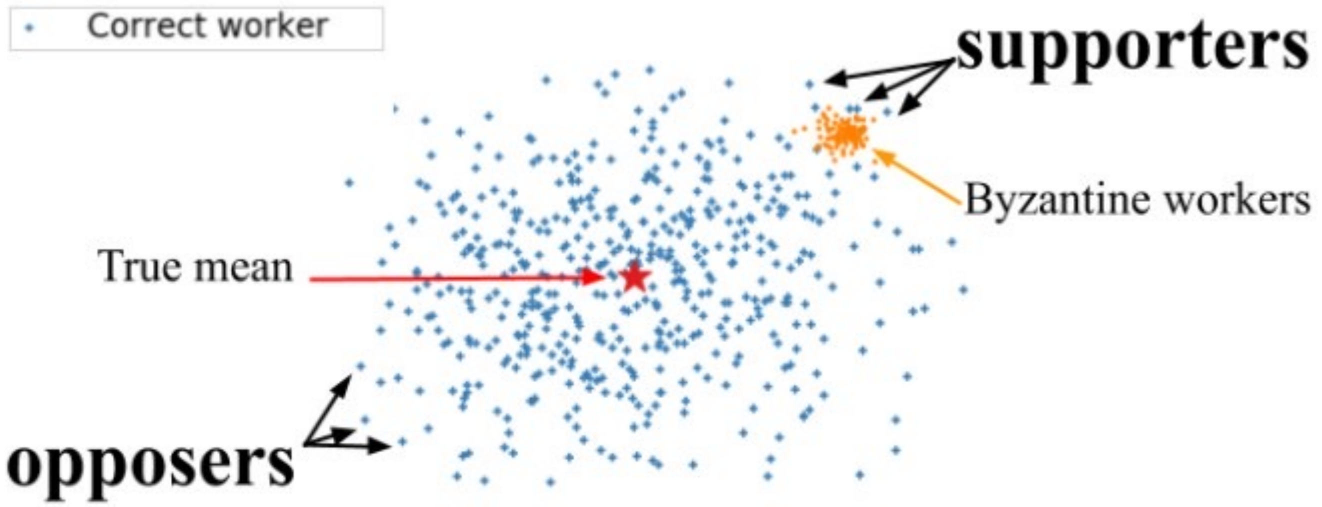
Baruch, G., Baruch, M., & Goldberg, Y. (2019). A little is enough: Circumventing defenses for distributed learning. *Advances in Neural Information Processing Systems*, 32.



Byzantines send the following vectors: $g_i^k = \mu_{\mathcal{G}} - z\sigma_{\mathcal{G}}$

A Little Is Enough (ALIE) Attack

Baruch, G., Baruch, M., & Goldberg, Y. (2019). A little is enough: Circumventing defenses for distributed learning. *Advances in Neural Information Processing Systems*, 32.



Byzantines send the following vectors: $g_i^k = \mu_{\mathcal{G}} - z\sigma_{\mathcal{G}}$

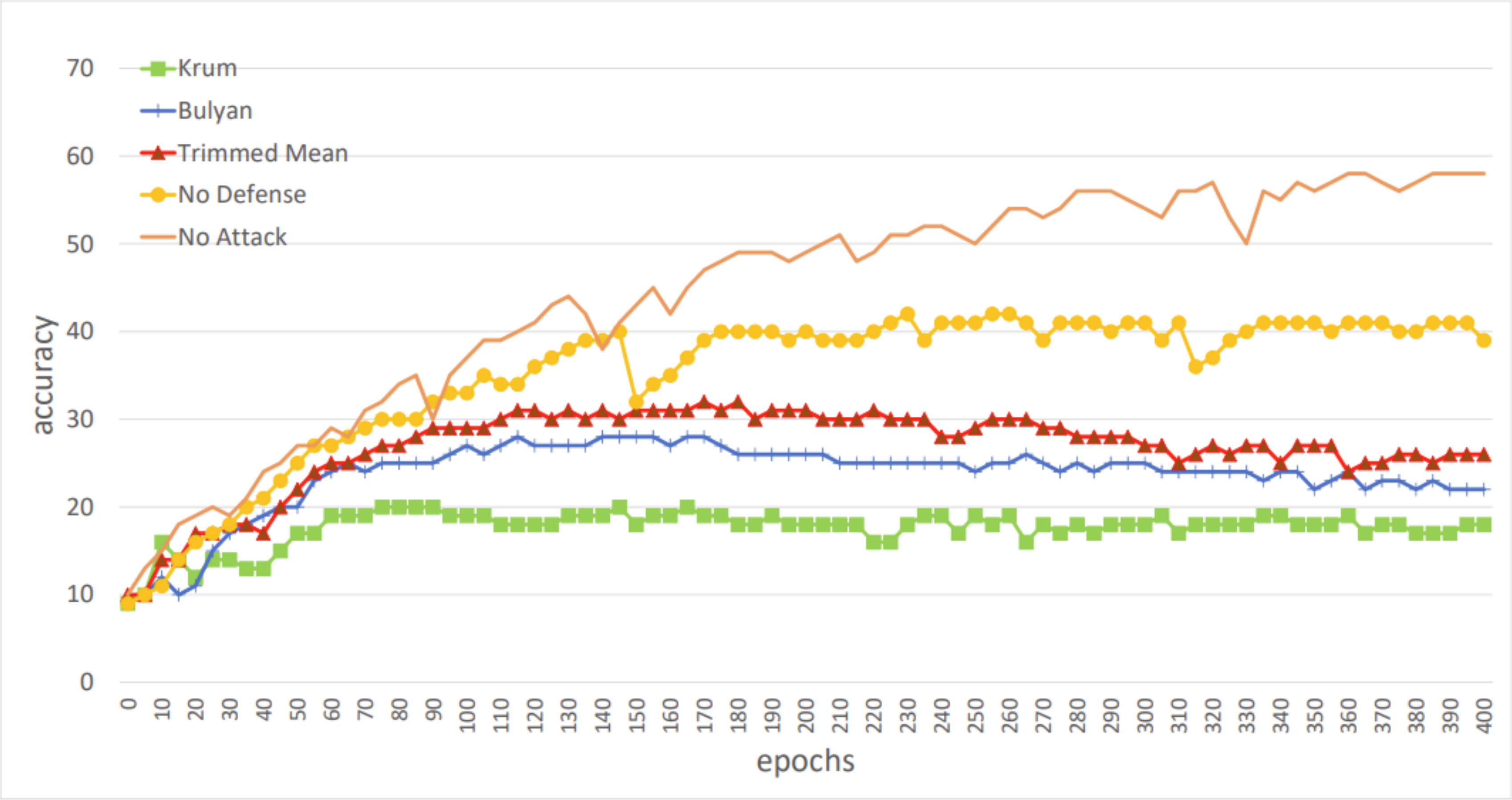
mean of the good workers

coordinate-wise standard deviation of good workers

- Byzantines choose z such that they are close to the “boundary of the cloud”
- Since Byzantines are closer to the mean, “middle-seekers” will treat opposers as outliers

The Result of ALIE Attack on the Training @ CIFAR10

Baruch, G., Baruch, M., & Goldberg, Y. (2019). A little is enough: Circumventing defenses for distributed learning. *Advances in Neural Information Processing Systems*, 32.



“No defense” strategy is more robust! Formal definition of robust aggregation is required!

Robust Aggregation Formalism

Karimireddy, S. P., He, L., & Jaggi, M. (2021, July). Learning from history for byzantine robust optimization. *In International Conference on Machine Learning* (pp. 5311-5319). PMLR.

Definition of (δ, c)–robust aggregator

Let $g_1 \dots, g_n$ be random variables such that there exist a good subset $\mathcal{G} \subseteq [n]$ of size $G \geq (1-\delta)n > {}^{n}/_{2}$ such that $\{g_i\}_{\{i \in \mathcal{G}\}}$ are independent and for all fixed pairs of good workers $i, j \in \mathcal{G}$ we have

$$\mathbb{E}\left[\|g_i - g_j\|^2\right] \leq \sigma^2.$$

Let $\bar{g} = \frac{1}{G}\sum_{i \in \mathcal{G}} g_i$. Then $\hat{g} = \mathrm{RAgg}(g_1, \dots, g_n)$ is called (δ, c)–robust aggregator if for some $c > 0$

$$\mathbb{E}\left[\|\hat{g} - \bar{g}\|^2\right] \leq c\delta\sigma^2$$

Robust Aggregation Formalism

Karimireddy, S. P., He, L., & Jaggi, M. (2021, July). Learning from history for byzantine robust optimization. *In International Conference on Machine Learning* (pp. 5311-5319). PMLR.

Definition of (δ, c)–robust aggregator

Let $g_1 \dots, g_n$ be random variables such that there exist a good subset $\mathcal{G} \subseteq [n]$ of size $G \geq (1-\delta)n > {}^{n}/_{2}$ such that $\{g_i\}_{\{i \in \mathcal{G}\}}$ are independent and for all fixed pairs of good workers $i, j \in \mathcal{G}$ we have

$$\mathbb{E}\left[\|g_i - g_j\|^2\right] \leq \sigma^2.$$

Let $\bar{g} = \frac{1}{G}\sum_{i \in \mathcal{G}} g_i$. Then $\hat{g} = \mathrm{RAgg}(g_1, \dots, g_n)$ is called (δ, c)–robust aggregator if for some $c > 0$

$$\mathbb{E}\left[\|\hat{g} - \bar{g}\|^2\right] \leq c\delta\sigma^2$$

- Medians and Krum estimators do not satisfy this definition
- **Question:** do such aggregators exist?

Bucketing Fixes “Middle-Seekers”

Karimireddy, S. P., He, L., & Jaggi, M. (2022). Byzantine-Robust Learning on Heterogeneous Datasets via Bucketing. *In International Conference on Learning Representations*.

Bucketing takes $\{g_1, \ldots, g_n\}$, positive integer s, and aggregator Aggr as an input and returns

$$\widehat{g} = \mathtt{Aggr}(y_1, \ldots, y_{\lceil n/s \rceil})$$

Bucketing Fixes “Middle-Seekers”

Karimireddy, S. P., He, L., & Jaggi, M. (2022). Byzantine-Robust Learning on Heterogeneous Datasets via Bucketing. *In International Conference on Learning Representations*.

Bucketing takes $\{g_1, \ldots, g_n\}$, positive integer s, and aggregator Aggr as an input and returns

$$\widehat{g} = \mathtt{Aggr}(y_1, \ldots, y_{\lceil n/s \rceil})$$

where $y_i = \frac{1}{s} \sum_{k=s(i-1)+1}^{\min\{si,n\}} x_{\pi(k)}$ and $\pi = (\pi(1), \ldots, \pi(n))$ is a random permutation of $[n]$

Bucketing Fixes "Middle-Seekers"

Karimireddy, S. P., He, L., & Jaggi, M. (2022). Byzantine-Robust Learning on Heterogeneous Datasets via Bucketing. *In International Conference on Learning Representations*.

Bucketing takes $\{g_1, \ldots, g_n\}$, positive integer s, and aggregator Aggr as an input and returns

$$\widehat{g} = \texttt{Aggr}(y_1, \ldots, y_{\lceil n/s \rceil})$$

where $y_i = \frac{1}{s} \sum_{k=s(i-1)+1}^{\min\{si,n\}} x_{\pi(k)}$ and $\pi = (\pi(1), \ldots, \pi(n))$ is a random permutation of $[n]$

For any $\delta \leq \delta_{\max}$ and $s = \lfloor \delta_{\max}/\delta \rfloor$

- Krum $\circ$ Bucketing is (δ, c)–robust aggregator with $c = \mathcal{O}(1)$ and $\delta_{\max} < 1/4$

Bucketing Fixes "Middle-Seekers"

Karimireddy, S. P., He, L., & Jaggi, M. (2022). Byzantine-Robust Learning on Heterogeneous Datasets via Bucketing. *In International Conference on Learning Representations*.

Bucketing takes $\{g_1, \ldots, g_n\}$, positive integer s, and aggregator Aggr as an input and returns

$$\widehat{g} = \texttt{Aggr}(y_1, \ldots, y_{\lceil n/s \rceil})$$

where $y_i = \frac{1}{s} \sum_{k=s(i-1)+1}^{\min\{si,n\}} x_{\pi(k)}$ and $\pi = (\pi(1), \ldots, \pi(n))$ is a random permutation of $[n]$

For any $\delta \leq \delta_{\max}$ and $s = \lfloor {}^{\delta_{\max}}/_{\delta} \rfloor$

- Krum ∘ Bucketing is (δ, c)–robust aggregator with $c = \mathcal{O}(1)$ and $\delta_{\max} < {}^1/_4$
- RFA ∘ Bucketing is (δ, c)–robust aggregator with $c = \mathcal{O}(1)$ and $\delta_{\max} < {}^1/_2$

Bucketing Fixes "Middle-Seekers"

Karimireddy, S. P., He, L., & Jaggi, M. (2022). Byzantine-Robust Learning on Heterogeneous Datasets via Bucketing. *In International Conference on Learning Representations*.

Bucketing takes $\{g_1, \dots, g_n\}$, positive integer s, and aggregator Aggr as an input and returns

$$\widehat{g} = \mathtt{Aggr}(y_1, \dots, y_{\lceil n/s \rceil})$$

where $y_i = \frac{1}{s} \sum_{k=s(i-1)+1}^{\min\{si,n\}} x_{\pi(k)}$ and $\pi = (\pi(1), \dots, \pi(n))$ is a random permutation of $[n]$

For any $\delta \leq \delta_{\max}$ and $s = \lfloor {}^{\delta_{\max}}/_{\delta} \rfloor$

- Krum ∘ Bucketing is (δ, c)–robust aggregator with $c = \mathcal{O}(1)$ and $\delta_{\max} < {}^1/_4$
- RFA ∘ Bucketing is (δ, c)–robust aggregator with $c = \mathcal{O}(1)$ and $\delta_{\max} < {}^1/_2$
- CM ∘ Bucketing is (δ, c)–robust aggregator with $c = \mathcal{O}(d)$ and $\delta_{\max} < {}^1/_2$

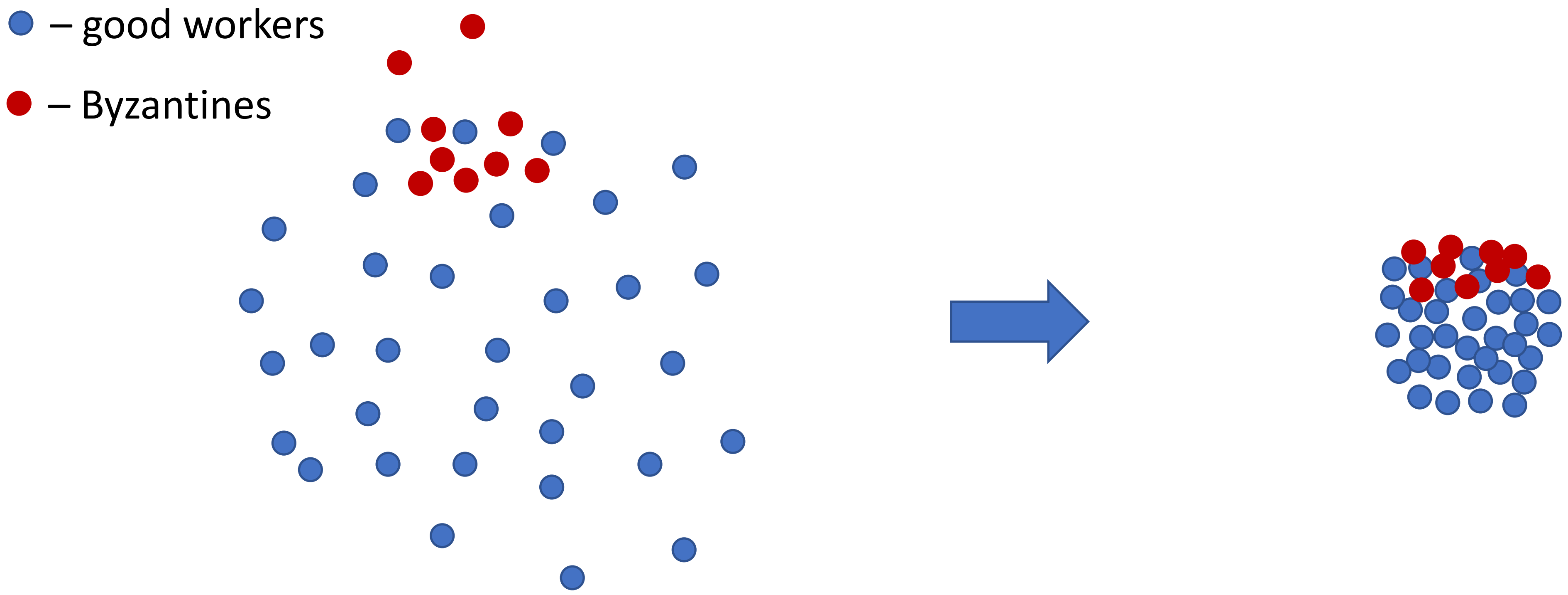
Moreover, these estimators are agnostic to σ^2!

Variance Reduction and Byzantine-Robustness

Why Variance Reduction?

Wu, Z., Ling, Q., Chen, T., & Giannakis, G. B. (2020). Federated variance-reduced stochastic gradient descent with robustness to byzantine attacks. IEEE Transactions on Signal Processing, 68, 4583-4596.

Natural idea: if the variance of good vectors gets smaller, it becomes progressively harder for Byzantines to shift the result of the aggregation from the true mean



- **Large variance** allows Byzantines to hide in noise and still create large bias
- Hard to detect outliers

- **Small variance** does not allow Byzantines to create large bias easily
- Easy to detect outliers

Byrd-SAGA: Byzantine-Robust SAGA

Wu, Z., Ling, Q., Chen, T., & Giannakis, G. B. (2020). Federated variance-reduced stochastic gradient descent with robustness to byzantine attacks. IEEE Transactions on Signal Processing, 68, 4583-4596.

Finite-sum optimization:

$$\min_{x \in \mathbb{R}^d} \left\{ f(x) := \frac{1}{m} \sum_{j=1}^{m} f_j(x) \right\}$$

# of samples in the dataset

loss on j-th sample

Byrd-SAGA: Byzantine-Robust SAGA

Wu, Z., Ling, Q., Chen, T., & Giannakis, G. B. (2020). Federated variance-reduced stochastic gradient descent with robustness to byzantine attacks. IEEE Transactions on Signal Processing, 68, 4583-4596.

Finite-sum optimization:

$$\min_{x\in\mathbb{R}^d}\left\{f(x):=\frac{1}{m}\sum_{j=1}^{m}f_j(x)\right\}$$

m: # of samples in the dataset

$f_j(x)$: loss on j-th sample

Byrd-SAGA:

- Good workers compute SAGA-estimators
- Server uses geometric median aggregator

$$x^{k+1}=x^k-\gamma\widehat{g}^k$$

$$\widehat{g}^k=\mathtt{RFA}(g_1^k,\ldots,g_n^k)$$

$$g_i^k=\begin{cases}\nabla f_{j_{i_k}}(x^k)-\nabla f_{j_{i_k}}(\phi^k_{i,j_{i_k}})+\frac{1}{m}\sum_{j=1}^{m}\nabla f_j(\phi^k_{i,j}), & \text{if } i\in\mathcal{G},\\ *, & \text{if } i\in\mathcal{B}\end{cases}$$

$$\phi^{k+1}_{i,j}=\begin{cases}\phi^k_{i,j}, & \text{if } j\neq j_{i_k},\\ x^k, & \text{if } j=j_{i_k}\end{cases}\quad\forall i\in\mathcal{G}$$

Complexity of Byrd-SAGA

Wu, Z., Ling, Q., Chen, T., & Giannakis, G. B. (2020). Federated variance-reduced stochastic gradient descent with robustness to byzantine attacks. IEEE Transactions on Signal Processing, 68, 4583-4596.

Assumptions:

- μ–strong convexity of f: $$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|y - x\|^2 \quad \forall x, y \in \mathbb{R}^d$$
- L–smoothness of $f_1, \ldots, f_m$: $$\|\nabla f_j(y) - \nabla f_j(x)\| \leq L\|y - x\| \quad \forall x, y \in \mathbb{R}^d, j \in [m]$$

Complexity of Byrd-SAGA

Wu, Z., Ling, Q., Chen, T., & Giannakis, G. B. (2020). Federated variance-reduced stochastic gradient descent with robustness to byzantine attacks. IEEE Transactions on Signal Processing, 68, 4583-4596.

Assumptions:

- μ–strong convexity of f: $$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|y - x\|^2 \quad \forall x, y \in \mathbb{R}^d$$
- L–smoothness of $f_1, \ldots, f_m$: $$\|\nabla f_j(y) - \nabla f_j(x)\| \leq L\|y - x\| \quad \forall x, y \in \mathbb{R}^d, j \in [m]$$

Theorem:
Let $\delta < {}^1/_2$ and the above assumptions hold. Then, there exists a choice of the stepsize γ such that the mini-batched version of Byrd-SAGA (with batchsize b) produces x^k satisfying $\mathbb{E}\left[\|x^k - x^*\|^2\right] \leq \varepsilon$ after

$$\mathcal{O}\left(\frac{m^2 L^2}{b^2(1 - 2\delta)\mu^2} \log \frac{1}{\varepsilon}\right)$$ iterations

Reflecting on the Complexities

- Complexity of Byrd-SAGA ($b = 1,\ \delta > 0$): $\mathcal{O}\left(\frac{m^2L^2}{(1-2\delta)\mu^2}\log\frac{1}{\varepsilon}\right)$
- Complexity of Byrd-SAGA ($b = 1,\ \delta = 0$): $\mathcal{O}\left(\frac{m^2L^2}{\mu^2}\log\frac{1}{\varepsilon}\right)$
- Complexity of SAGA ($b = 1,\ \delta = 0$): $\mathcal{O}\left(\left(m + \frac{L}{\mu}\right)\log\frac{1}{\varepsilon}\right)$

Reflecting on the Complexities

- Complexity of Byrd-SAGA ($b = 1,\ \delta > 0$): $\mathcal{O}\left(\frac{m^2 L^2}{(1-2\delta)\mu^2} \log \frac{1}{\varepsilon}\right)$
- Complexity of Byrd-SAGA ($b = 1,\ \delta = 0$): $\mathcal{O}\left(\frac{m^2 L^2}{\mu^2} \log \frac{1}{\varepsilon}\right)$
- Complexity of SAGA ($b = 1,\ \delta = 0$): $\mathcal{O}\left(\left(m + \frac{L}{\mu}\right) \log \frac{1}{\varepsilon}\right)$

The reason for such a dramatic deterioration in the complexity of Byrd-SAGA in comparison to SAGA:

$$\mathbb{E}_k[\widehat{g}^k] \neq \nabla f(x^k)$$

Analysis of SAGA/SVRG-based methods is very sensitive to unbiasedness!

Biased VR: You Cannot "Break" What Is Already "Broken"!

SARAH/Geom-SARAH/PAGE (1 node case):

$$x^{k+1} = x^k - \gamma g^k$$

Nguyen, L. M., Liu, J., Scheinberg, K., & Takáč, M. (2017, July). SARAH: A novel method for machine learning problems using stochastic recursive gradient. In International Conference on Machine Learning (pp. 2613-2621). PMLR.

Horváth, S., Lei, L., Richtárik, P., & Jordan, M. I. (2022). Adaptivity of stochastic gradient methods for nonconvex optimization. SIAM Journal on Mathematics of Data Science, 4(2), 634-648.

Li, Z., Bao, H., Zhang, X., & Richtárik, P. (2021, July). PAGE: A simple and optimal probabilistic gradient estimator for nonconvex optimization. In International Conference on Machine Learning (pp. 6286-6295). PMLR.

Biased VR: You Cannot "Break" What Is Already "Broken"!

SARAH/Geom-SARAH/PAGE (1 node case):

$$x^{k+1} = x^k - \gamma g^k$$

$$g^k = \begin{cases} \nabla f(x^k), & \text{with prob. } p \\ g^{k-1} + \frac{1}{b} \sum_{j \in J_k} \left(\nabla f_j(x^k) - \nabla f_j(x^{k-1}) \right), & \text{with prob. } 1-p \end{cases}$$

Nguyen, L. M., Liu, J., Scheinberg, K., & Takáč, M. (2017, July). SARAH: A novel method for machine learning problems using stochastic recursive gradient. In International Conference on Machine Learning (pp. 2613-2621). PMLR.

Horváth, S., Lei, L., Richtárik, P., & Jordan, M. I. (2022). Adaptivity of stochastic gradient methods for nonconvex optimization. SIAM Journal on Mathematics of Data Science, 4(2), 634-648.

Li, Z., Bao, H., Zhang, X., & Richtárik, P. (2021, July). PAGE: A simple and optimal probabilistic gradient estimator for nonconvex optimization. In International Conference on Machine Learning (pp. 6286-6295). PMLR.

Biased VR: You Cannot "Break" What Is Already "Broken"!

SARAH/Geom-SARAH/PAGE (1 node case):

$$x^{k+1} = x^k - \gamma g^k$$

$$g^k = \begin{cases} \nabla f(x^k), & \text{with prob. } p \\ g^{k-1} + \frac{1}{b} \sum_{j \in J_k} \left(\nabla f_j(x^k) - \nabla f_j(x^{k-1}) \right), & \text{with prob. } 1-p \end{cases}$$

J_k– indices in the mini-batch, $|J_k| = b$

Nguyen, L. M., Liu, J., Scheinberg, K., & Takáč, M. (2017, July). SARAH: A novel method for machine learning problems using stochastic recursive gradient. In International Conference on Machine Learning (pp. 2613-2621). PMLR.

Horváth, S., Lei, L., Richtárik, P., & Jordan, M. I. (2022). Adaptivity of stochastic gradient methods for nonconvex optimization. SIAM Journal on Mathematics of Data Science, 4(2), 634-648.

Li, Z., Bao, H., Zhang, X., & Richtárik, P. (2021, July). PAGE: A simple and optimal probabilistic gradient estimator for nonconvex optimization. In International Conference on Machine Learning (pp. 6286-6295). PMLR.

Biased VR: You Cannot "Break" What Is Already "Broken"!

SARAH/Geom-SARAH/PAGE (1 node case):

$$x^{k+1} = x^k - \gamma g^k$$

$p \sim {}^{b}/_{m}$ – probability of computing the full gradient

$$g^k = \begin{cases} \nabla f(x^k), & \text{with prob. } p \\ g^{k-1} + \frac{1}{b} \sum_{j \in J_k} \left(\nabla f_j(x^k) - \nabla f_j(x^{k-1}) \right), & \text{with prob. } 1 - p \end{cases}$$

J_k – indices in the mini-batch, $|J_k| = b$

Nguyen, L. M., Liu, J., Scheinberg, K., & Takáč, M. (2017, July). SARAH: A novel method for machine learning problems using stochastic recursive gradient. In International Conference on Machine Learning (pp. 2613-2621). PMLR.

Horváth, S., Lei, L., Richtárik, P., & Jordan, M. I. (2022). Adaptivity of stochastic gradient methods for nonconvex optimization. SIAM Journal on Mathematics of Data Science, 4(2), 634-648.

Li, Z., Bao, H., Zhang, X., & Richtárik, P. (2021, July). PAGE: A simple and optimal probabilistic gradient estimator for nonconvex optimization. In International Conference on Machine Learning (pp. 6286-6295). PMLR.

Biased VR: You Cannot "Break" What Is Already "Broken"!

SARAH/Geom-SARAH/PAGE (1 node case):

$$x^{k+1} = x^k - \gamma g^k$$

$p \sim {}^{b}/_{m}$ – probability of computing the full gradient

$$g^k = \begin{cases} \nabla f(x^k), & \text{with prob. } p \\ g^{k-1} + \frac{1}{b} \sum_{j \in J_k} \left(\nabla f_j(x^k) - \nabla f_j(x^{k-1}) \right), & \text{with prob. } 1 - p \end{cases}$$

J_k – indices in the mini-batch, $|J_k| = b$

$$\mathbb{E}_k[g^k] \neq \nabla f(x^k)$$

Estimator is biased from the beginning!

Nguyen, L. M., Liu, J., Scheinberg, K., & Takáč, M. (2017, July). SARAH: A novel method for machine learning problems using stochastic recursive gradient. In International Conference on Machine Learning (pp. 2613-2621). PMLR.

Horváth, S., Lei, L., Richtárik, P., & Jordan, M. I. (2022). Adaptivity of stochastic gradient methods for nonconvex optimization. SIAM Journal on Mathematics of Data Science, 4(2), 634-648.

Li, Z., Bao, H., Zhang, X., & Richtárik, P. (2021, July). PAGE: A simple and optimal probabilistic gradient estimator for nonconvex optimization. In International Conference on Machine Learning (pp. 6286-6295). PMLR.

New Method: Byz-PAGE

$$x^{k+1} = x^k - \gamma \widehat{g}^k \qquad \widehat{g}^k = \texttt{ARAggr}(g_1^k, \ldots, g_n^k)$$

New Method: Byz-PAGE

(δ, c)–robust aggregator agnostic to the variance, e.g., Krum/RFA/CM ∘ Bucketing

$$x^{k+1} = x^k - \gamma \widehat{g}^k \qquad \widehat{g}^k = \boxed{\texttt{ARAggr}(g_1^k, \ldots, g_n^k)}$$

New Method: Byz-PAGE

(δ, c)–robust aggregator agnostic to the variance, e.g., Krum/RFA/CM ∘ Bucketing

$$x^{k+1} = x^k - \gamma \widehat{g}^k \qquad \widehat{g}^k = \texttt{ARAggr}(g_1^k, \ldots, g_n^k)$$

$$g_i^k = \begin{cases} \nabla f(x^k), & \text{with prob. } p \\ g^{k-1} + \frac{1}{b} \sum\limits_{j \in J_k} \left(\nabla f_j(x^k) - \nabla f_j(x^{k-1})\right), & \text{with prob. } 1-p \end{cases} \quad \forall i \in \mathcal{G}$$

Geom-SARAH/PAGE–estimator

Complexity of Byz-PAGE (Simplified)

Assumptions:

- f is lower-bounded: $$f_* = \inf_{x \in \mathbb{R}^d} f(x) > -\infty$$
- L–smoothness of $f_1, \ldots, f_m$: $$\|\nabla f_j(y) - \nabla f_j(x)\| \leq L\|y - x\| \quad \forall x, y \in \mathbb{R}^d, j \in [m]$$

Complexity of Byz-PAGE (Simplified)

Assumptions:

- f is lower-bounded: $$f_* = \inf_{x \in \mathbb{R}^d} f(x) > -\infty$$
- L–smoothness of $f_1, \ldots, f_m$: $$\|\nabla f_j(y) - \nabla f_j(x)\| \leq L\|y - x\| \quad \forall x, y \in \mathbb{R}^d, j \in [m]$$

Theorem 1:

Let the above assumptions hold and ARAggr be (δ, c)–robust aggregator. Then, there exists a choice of the stepsize γ such that Byz-PAGE produces $\hat{x}^k$ satisfying $\mathbb{E}\left[\left\|\nabla f(\hat{x}^k)\right\|^2\right] \leq \varepsilon^2$ after

Complexity of Byz-PAGE (Simplified)

Assumptions:

- f is lower-bounded: $f_* = \inf_{x \in \mathbb{R}^d} f(x) > -\infty$
- L–smoothness of $f_1, \dots, f_m$: $\|\nabla f_j(y) - \nabla f_j(x)\| \leq L\|y - x\| \quad \forall x, y \in \mathbb{R}^d, j \in [m]$

Theorem 1:

Let the above assumptions hold and ARAggr be (δ, c)–robust aggregator. Then, there exists a choice of the stepsize γ such that Byz-PAGE produces $\hat{x}^k$ satisfying $\mathbb{E}\left[\|\nabla f(\hat{x}^k)\|^2\right] \leq \varepsilon^2$ after

$$\mathcal{O}\left(\frac{\left(1 + \sqrt{\frac{c\delta m^2}{b^3} + \frac{m}{b^2 n}}\right) L\left(f(x^0) - f_*\right)}{\varepsilon^2}\right) \quad \text{iterations}$$

Complexity of Byz-PAGE: PŁ Case (Simplified)

Assumptions:

- f has a minimizer: $$x^* = \arg\min_{x \in \mathbb{R}^d} f(x)$$
- L–smoothness of $f_1, \ldots, f_m$: $$\|\nabla f_j(y) - \nabla f_j(x)\| \leq L\|y - x\| \quad \forall x, y \in \mathbb{R}^d, j \in [m]$$
- f is μ–PŁ function: $$\|\nabla f(x)\|^2 \geq 2\mu\left(f(x) - f\left(x^*\right)\right)$$

Complexity of Byz-PAGE: PŁ Case (Simplified)

Assumptions:

- f has a minimizer: $$x^* = \arg\min_{x \in \mathbb{R}^d} f(x)$$
- L–smoothness of $f_1, \ldots, f_m$: $$\|\nabla f_j(y) - \nabla f_j(x)\| \leq L\|y - x\| \quad \forall x, y \in \mathbb{R}^d, j \in [m]$$
- f is μ–PŁ function: $$\|\nabla f(x)\|^2 \geq 2\mu \left(f(x) - f\left(x^*\right)\right)$$

Theorem 2:
Let the above assumptions hold and ARAggr be (δ, c)–robust aggregator. Then, there exists a choice of the stepsize γ such that Byz-PAGE produces x^k satisfying $\mathbb{E}\left[f\left(x^k\right) - f(x^*)\right] \leq \varepsilon$ after

Complexity of Byz-PAGE: PŁ Case (Simplified)

Assumptions:

- f has a minimizer: $$x^* = \arg\min_{x \in \mathbb{R}^d} f(x)$$
- L–smoothness of $f_1, \ldots, f_m$: $$\|\nabla f_j(y) - \nabla f_j(x)\| \le L\|y - x\| \quad \forall x, y \in \mathbb{R}^d, j \in [m]$$
- f is μ–PŁ function: $$\|\nabla f(x)\|^2 \ge 2\mu\left(f(x) - f\left(x^*\right)\right)$$

Theorem 2:
Let the above assumptions hold and ARAggr be (δ, c)–robust aggregator. Then, there exists a choice of the stepsize γ such that Byz-PAGE produces x^k satisfying $\mathbb{E}\left[f\left(x^k\right) - f\left(x^*\right)\right] \le \varepsilon$ after

$$\mathcal{O}\left(\left(\frac{m}{b} + \frac{\left(1 + \sqrt{\frac{c\delta m^2}{b^3} + \frac{m}{b^2 n}}\right)L}{\mu}\right)\log\frac{1}{\varepsilon}\right)$$ iterations

Comparison with SOTA Results

Method	Assumptions	Complexity (NC)	Complexity (PŁ)
BR-SGDm [Karimireddy et al., 2021, 2022]	UBV	$\frac{1}{\varepsilon^2} + \frac{\sigma^2(c\delta+1/n)}{b\varepsilon^4}$	✗
BR-MVR [Karimireddy et al., 2021]	UBV	$\frac{1}{\varepsilon^2} + \frac{\sigma\sqrt{c\delta+1/n}}{\sqrt{b}\varepsilon^3}$	✗
BTARD-SGD [Gorbunov et al., 2021a]	UBV(1)	$\frac{1}{\varepsilon^2} + \frac{n^2\delta\sigma^2}{Cb\varepsilon^2} + \frac{\sigma^2}{nb\varepsilon^4}$	$\frac{1}{\mu} + \frac{\sigma^2}{nb\mu\varepsilon} + \frac{n^2\delta\sigma}{C\sqrt{b\mu\varepsilon}}$
Byrd-SAGA (2) [Wu et al., 2020]	Smooth $f_{i,j}$	✗	$\frac{m^2}{b^2(1-2\delta)\mu^2}$
Byz-VR-MARINA Cor. E.1 & Cor. E.5	As. 2.4	$\frac{1+\sqrt{\frac{c\delta m^2}{b^3}+\frac{m}{b^2 n}}}{\varepsilon^2}$	$\frac{1+\sqrt{\frac{c\delta m^2}{b^3}+\frac{m}{b^2 n}}}{\mu} + \frac{m}{b}$

- Byz-VR-MARINA = version of Byz-PAGE with communication compression
- NC = general non-convex functions
- PŁ = Polyak-Łojasiewicz-functions (BTARD-SGD and Byrd-SAGA are analyzed under strong convexity)
- UBV = uniformly bounded variance assumption: $\mathbb{E}\left[\left\|\nabla f_j(x) - \nabla f(x)\right\|^2\right] \leq \sigma^2$
- As. 2.4 = generalization of smoothness and data-similarity that incorporates non-uniform sampling of stochastic gradients

Remarks on the Results and One Extension

Remarks on the results:

- We achieve **new SOTA theoretical results** for Byzantine-robust learning
- When $\delta = 0$ (no Byzantines), the derived complexity bounds recover the known ones for Geom-SARAH/PAGE
- Therefore, the terms that are not affected by δ are **unimprovable**
- **Open question:** are the derived upper bounds optimal?

Remarks on the Results and One Extension

Remarks on the results:

- We achieve **new SOTA theoretical results** for Byzantine-robust learning
- When $\delta = 0$ (no Byzantines), the derived complexity bounds recover the known ones for Geom-SARAH/PAGE
- Therefore, the terms that are not affected by δ are **unimprovable**
- **Open question:** are the derived upper bounds optimal?

The extension to the compressed communication case:

- Byz-PAGE:

$$g_i^k = \begin{cases} \nabla f\left(x^k\right), & \text{with prob. } p \\ g^{k-1} + \frac{1}{b} \sum\limits_{j \in J_k} \left(\nabla f_j\left(x^k\right) - \nabla f_j\left(x^{k-1}\right)\right), & \text{with prob. } 1-p \end{cases}$$

Remarks on the Results and One Extension

Remarks on the results:

- We achieve **new SOTA theoretical results** for Byzantine-robust learning
- When $\delta = 0$ (no Byzantines), the derived complexity bounds recover the known ones for Geom-SARAH/PAGE
- Therefore, the terms that are not affected by δ are **unimprovable**
- **Open question:** are the derived upper bounds optimal?

The extension to the compressed communication case:

- Byz-PAGE:

$$g_i^k = \begin{cases} \nabla f\left(x^k\right), & \text{with prob. } p \\ g^{k-1} + \frac{1}{b}\sum\limits_{j\in J_k}\left(\nabla f_j\left(x^k\right) - \nabla f_j\left(x^{k-1}\right)\right), & \text{with prob. } 1-p \end{cases}$$

- Byz-VR-MARINA:

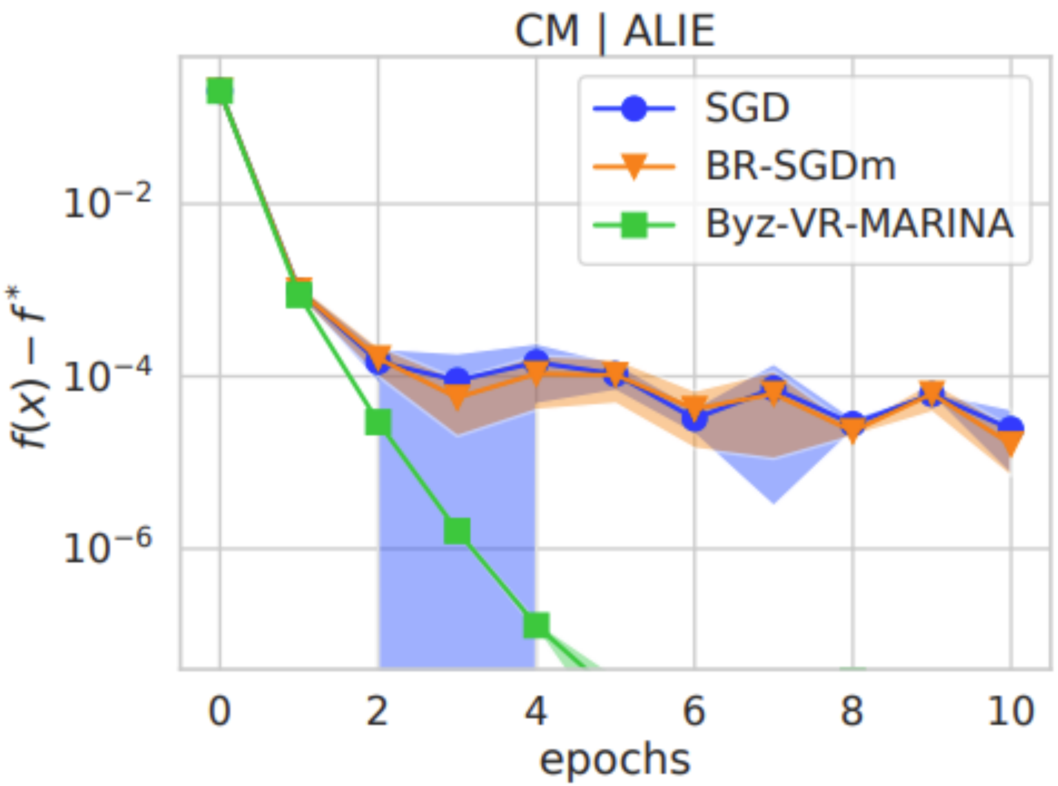
$$g_i^k = \begin{cases} \nabla f\left(x^k\right), & \text{with prob. } p \\ g^{k-1} + \mathcal{Q}\left(\frac{1}{b}\sum\limits_{j\in J_k}\left(\nabla f_j\left(x^k\right) - \nabla f_j\left(x^{k-1}\right)\right)\right), & \text{with prob. } 1-p \end{cases}$$

unbiased compression operator

Gorbunov, E., Burlachenko, K. P., Li, Z., & Richtárik, P. (2021, July). MARINA: Faster non-convex distributed learning with compression. In International Conference on Machine Learning (pp. 3788-3798). PMLR.

Numerical Results

- We tested the proposed method on the logistic regression tasks
- In this experiment, we have 4 good workers and 1 Byzantine
- As predicted by the derived results, the proposed method has linear convergence
- Competitors struggle to achieve better loss
- The results are consistent for all tested attacks



Concluding Remarks

In the Paper We Also Have

- Analysis of the version with compression (Byz-VR-MARINA)
- Analysis under bounded heterogeneity
- Non-uniform sampling of stochastic gradients
- Analysis takes into account data-similarity
- Additional experiments

Recent Follow Up Works

Ahmad Rammal, Kaja Gruntkowska, Nikita Fedin, Eduard Gorbunov, Peter Richtárik. *Communication Compression for Byzantine Robust Learning: New Efficient Algorithms and Improved Rates* (AISTATS 2024)

Workers send only compressed vectors
Better complexities when compression is used
Support of biased compression operators and error feedback

Grigory Malinovsky, Peter Richtárik, Samuel Horváth, Eduard Gorbunov. *Byzantine Robustness and Partial Participation Can Be Achieved Simultaneously: Just Clip Gradient Differences* (arXiv:2311.14127)

🤯 Provable convergence even if <u>Byzantine workers can form majority</u> during some rounds!

Thank you!