Byzantine Resilient Machine Learning

Algorithms to cure poisoned SGD

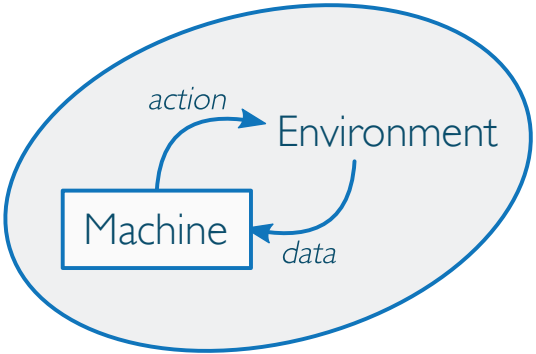
Sébastien Rouault
sebastien.rouault@epfl.ch

Distributed Computing Laboratory, EPFL

May 6, 2019

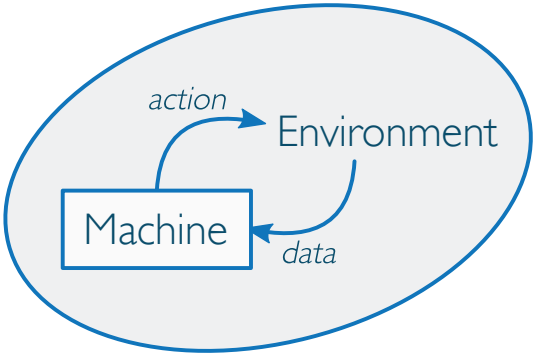
Introduction

Software security



Introduction

Software security



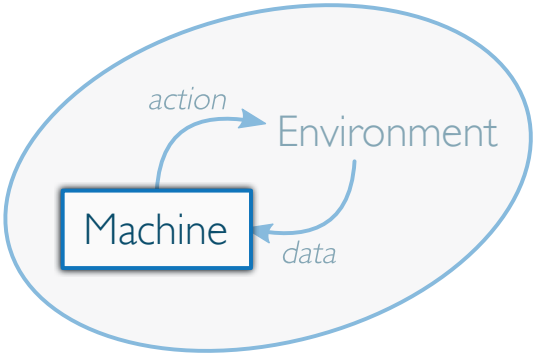
Allow intended actions

&

Prevent harmful actions

Introduction

Software security



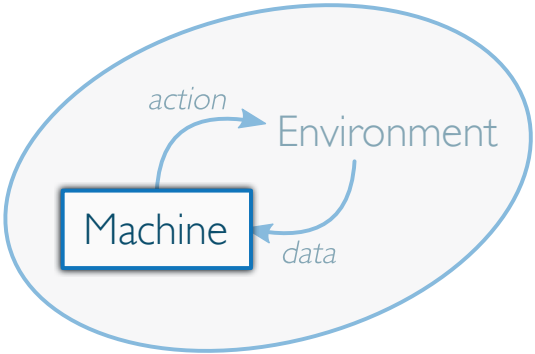
data + state

↓

action + state

Introduction

Software security



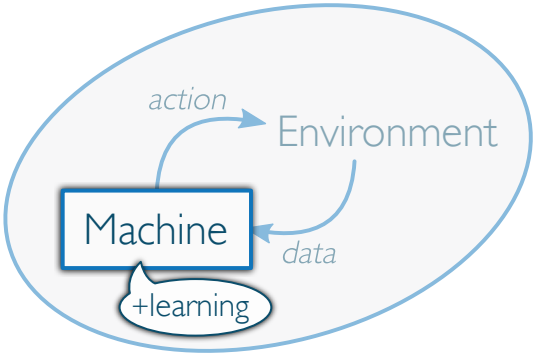
data + state

↓

action + state

Introduction

Software security, for ML



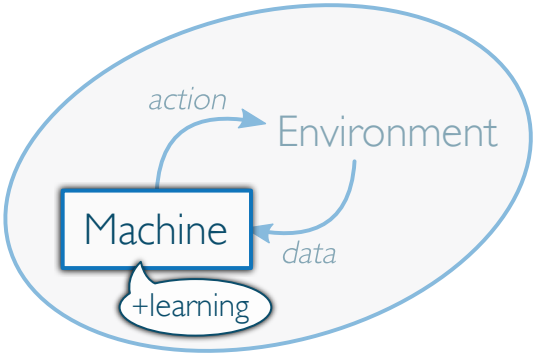
data + state

↓ +

action + state

Introduction

Software security, for ML



data is code

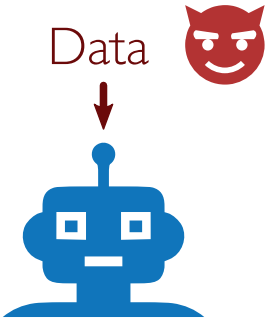
(should be treated as such)

Attacker side

Taxonomy of attacks

3 families of attacks:

- Evasion
- Exploratory
- Poisoning

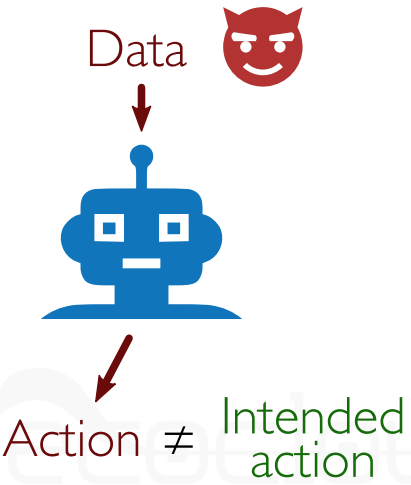


Attacker side

Taxonomy of attacks

3 families of attacks:

- Evasion
- Exploratory
- Poisoning

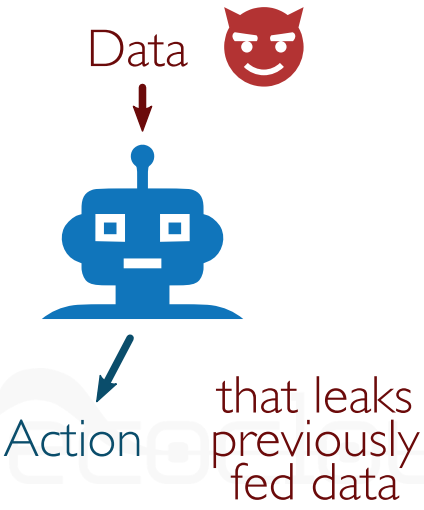


Attacker side

Taxonomy of attacks

3 families of attacks:

- Evasion
- Exploratory
- Poisoning

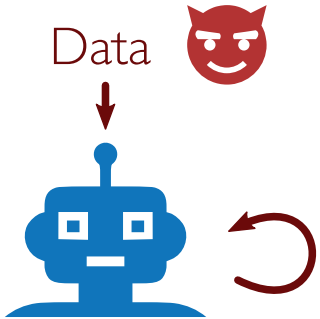


Attacker side

Taxonomy of attacks

3 families of attacks:

- Evasion
- Exploratory
- Poisoning

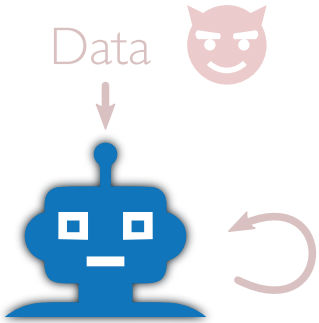


Attacker side

Taxonomy of attacks

3 families of attacks:

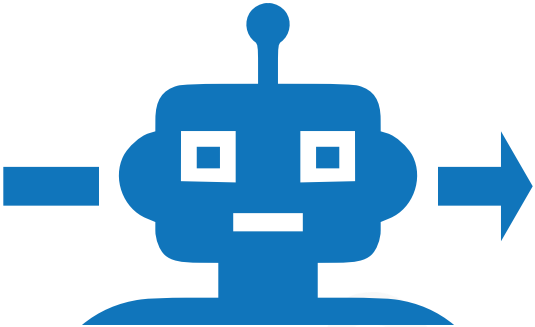
- Evasion
- Exploratory
- Poisoning



Attacker side

Training a classifier/predictor

…



"Boat"

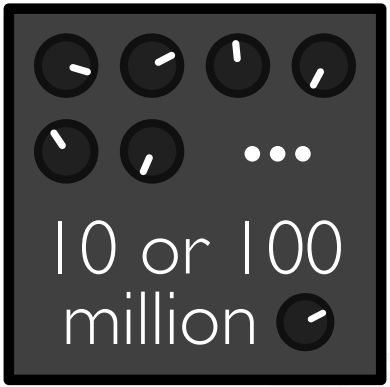
"Goat"

…

Attacker side

Training a classifier/predictor

...

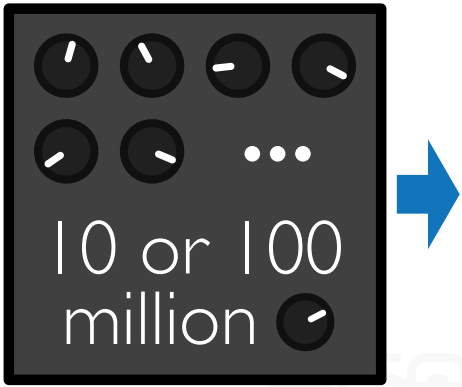


"random"
data

Attacker side

Training a classifier/predictor

…



"Boat"

"Goat"

…

Attacker side

Training: stochastic gradient descent (SGD)

Attacker side

Training: stochastic gradient descent (SGD)

Training loop:

1. Estimate gradient
2. Turn potentiometers following the gradient
3. Loop back to step 1.

Attacker side

Training: stochastic gradient descent (SGD)

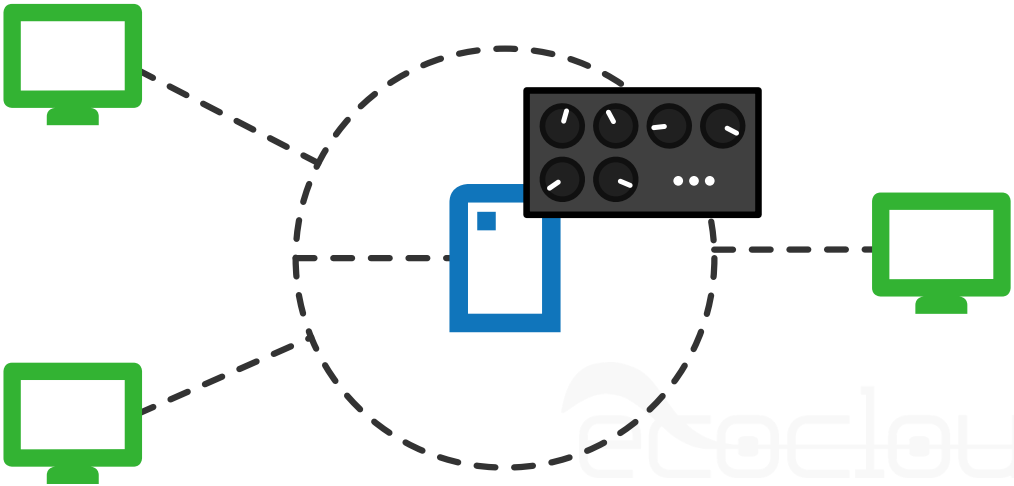
Computationally costly

&

Fully parallelizable

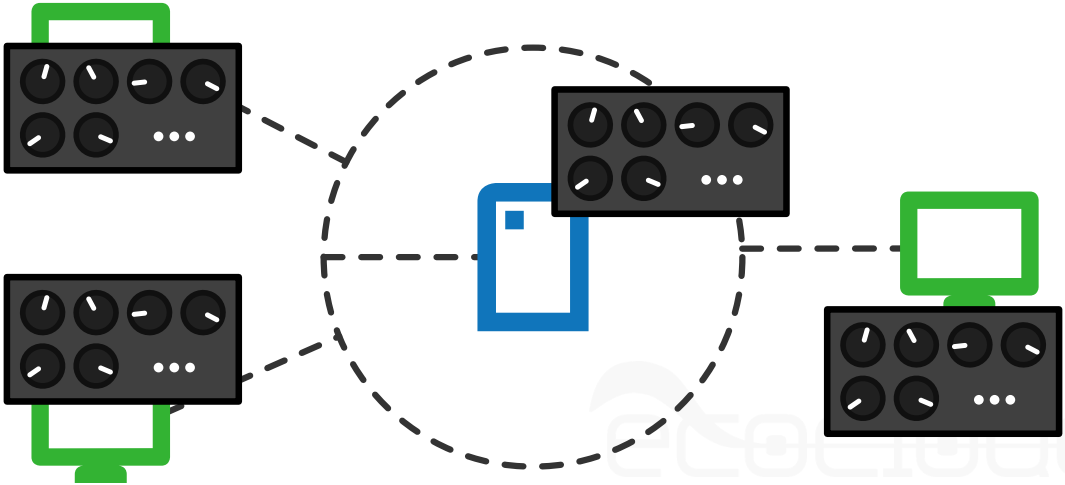
Attacker side

Training: distributed SGD



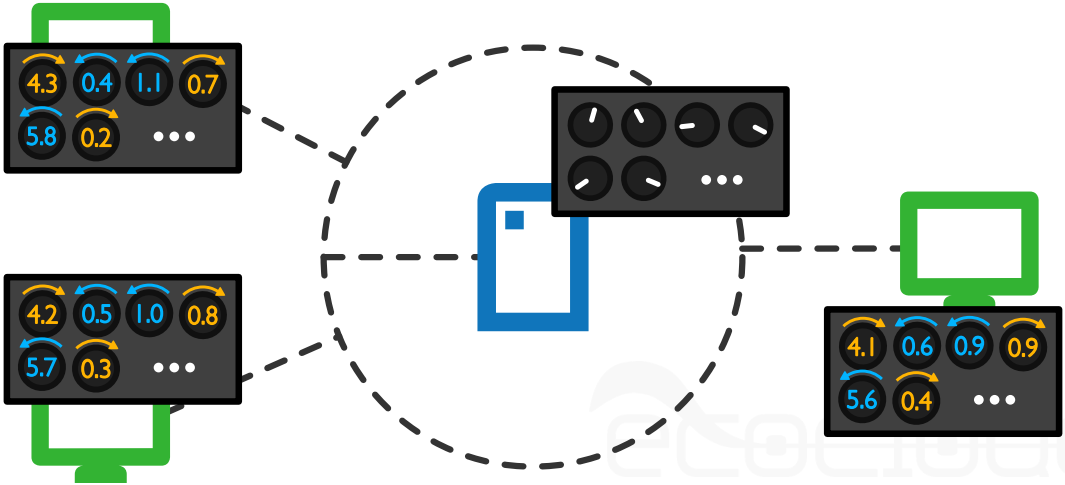
Attacker side

Training: distributed SGD



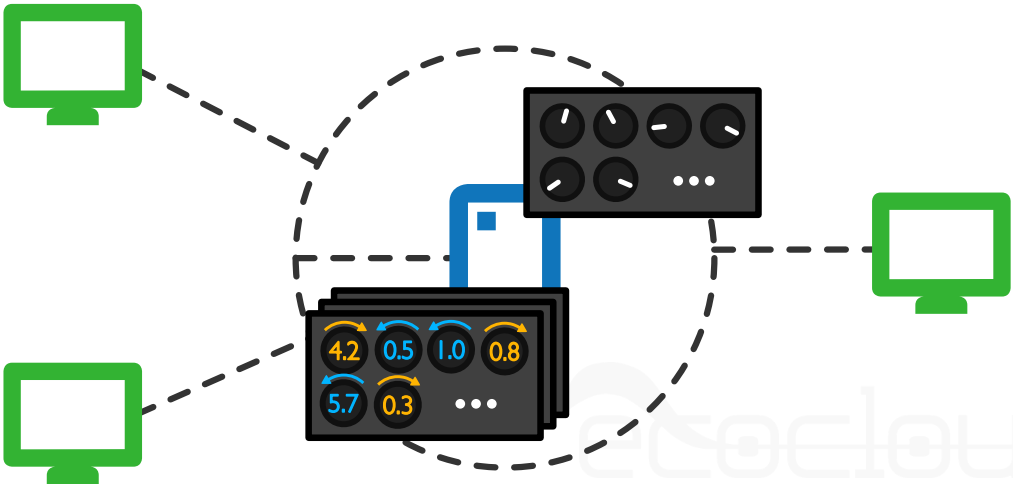
Attacker side

Training: distributed SGD



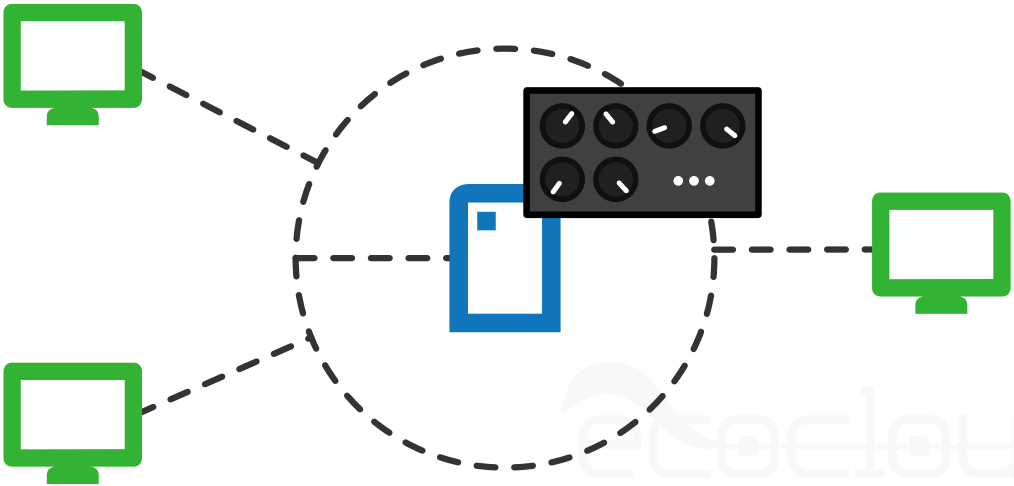
Attacker side

Training: distributed SGD



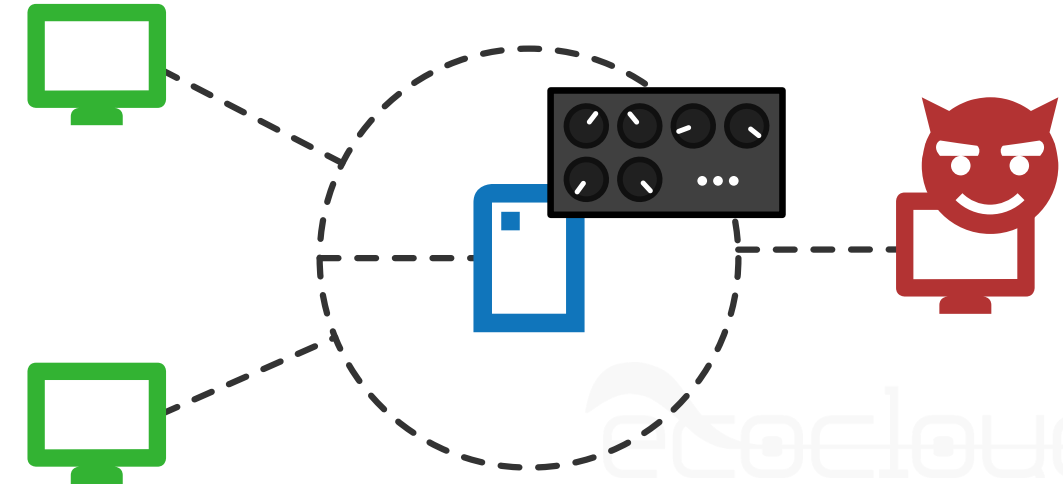
Attacker side

Training: distributed SGD



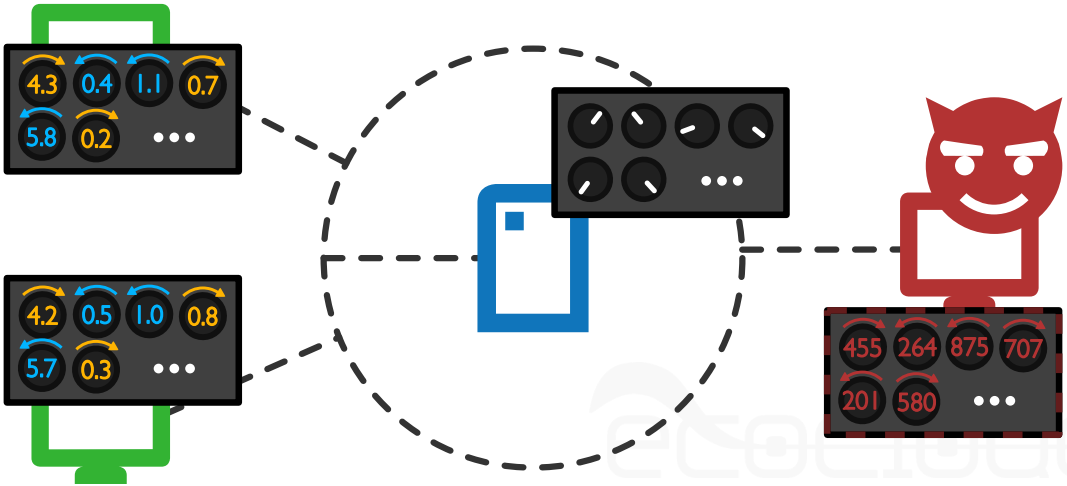
Attacker side

Poisoning: distributed SGD



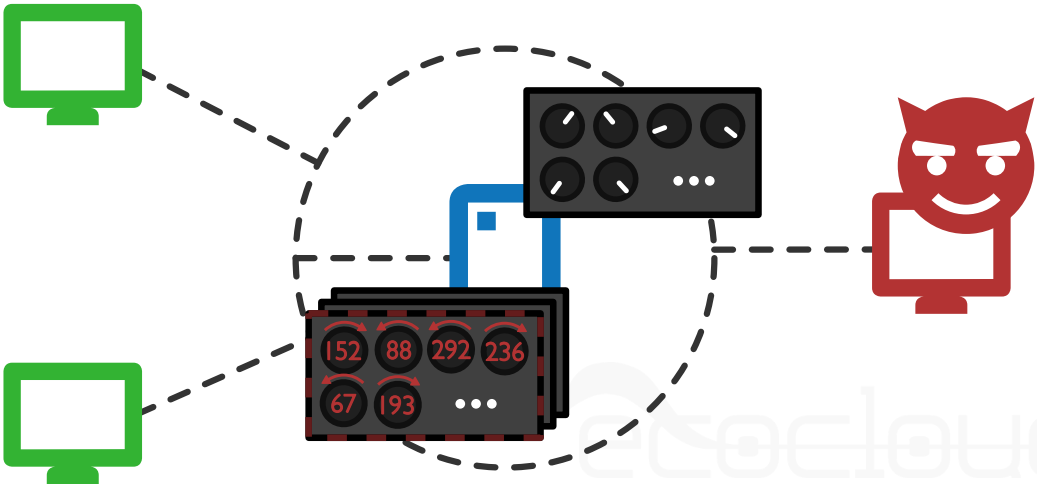
Attacker side

Poisoning: distributed SGD



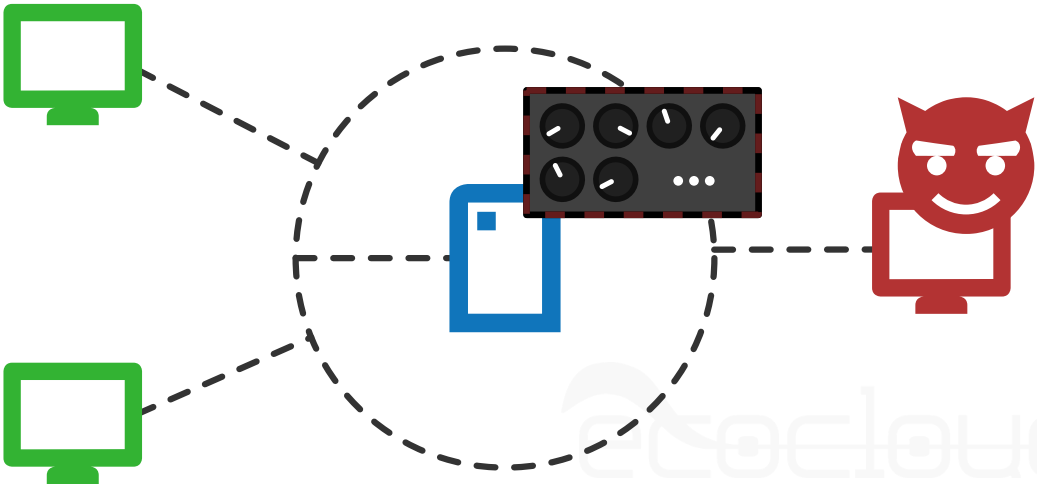
Attacker side

Poisoning: distributed SGD



Attacker side

Poisoning: distributed SGD



Defender side

A cure for poisoned SGD

Minority of 

are malicious 

1. Filter out or 2. limit the impact of 

Defender side

A cure for poisoned SGD

ecocloud
an EPFL research center

1. Gradient redundancy

- Draco [CWCP18]

2. Statistical robustness

- Krum [BEMGS17]
- Bulyan [EMGR18]
- Kardam [DEMG+18]
- Trimmed-mean [YCRB18]

Defender side

A system to cure poisoned SGD

Defender side

A system to cure poisoned SGD

ecocloud
an EPFL research center

TensorFlow

+


krum.cc


bulyan.cc


krum.cu


bulyan.cu

...

Defender side

A system to cure poisoned SGD

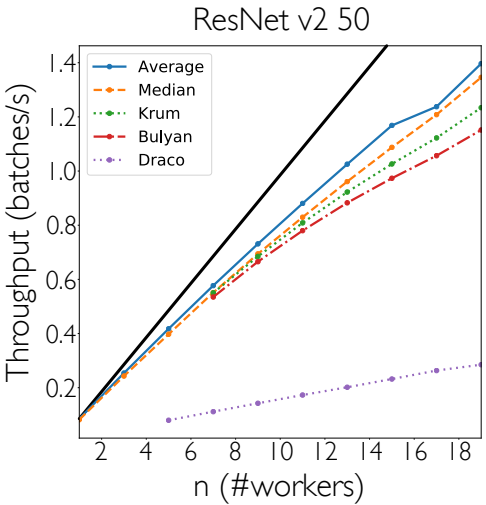
+

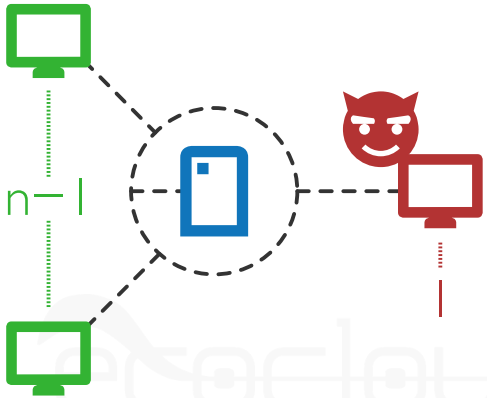
...

+ Design patch

Defender side

Main performance result





Defender side

More results and code

ecocloud
an EPFL research center

AGGREGATHOR: Byzantine Machine Learning via Robust Gradient Aggregation

- Experiments with UDP under congestion
- https://github.com/LPD-EPFL/AggregaThor

Defender side

Current and future evolutions

PyTorch

Conclusion

ecocloud
an EPFL research center

data is code

Statistical robustness

Off–the–shelf implementations

Conclusion

ecocloud
an EPFL research center

data is code

Statistical robustness

Off–the–shelf implementations

Conclusion

ecocloud
an EPFL research center

data is code

Statistical robustness

Off–the–shelf implementations

References I

- Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer, Machine learning with adversaries: Byzantine tolerant gradient descent, Neural Information Processing Systems, 2017, pp. 118–128.
- Lingjiao Chen, Hongyi Wang, Zachary Charles, and Dimitris Papailiopoulos, Draco: Byzantine-resilient distributed training via redundant gradients, International Conference on Machine Learning, 2018, pp. 902–911.
- Georgios Damaskinos, El Mahdi El Mhamdi, Rachid Guerraoui, Rhicheek Patra, Mahsa Taziki, et al., Asynchronous byzantine machine learning (the case of sgd), ICML, 2018, pp. 1153–1162.
- El Mahdi El Mhamdi, Rachid Guerraoui, and Sébastien Rouault, The hidden vulnerability of distributed learning in Byzantium, Proceedings of the 35th International Conference on Machine Learning (Stockholmsmässan, Stockholm Sweden) (Jennifer Dy and Andreas Krause, eds.), Proceedings of Machine Learning Research, vol. 80, PMLR, 10–15 Jul 2018, pp. 3521–3530.
- Dong Yin, Yudong Chen, Kannan Ramchandran, and Peter Bartlett, Byzantine-robust distributed learning: Towards optimal statistical rates, arXiv preprint arXiv:1803.01498 (2018).