Attack-Resistant Federated Learning with Residual-based Reweighting

Shuhao Fu[1], Chulin Xie[2], Bo Li[3], Qifeng Chen[1]

[1]HKUST,

[2]Zhejiang University

[3]UIUC

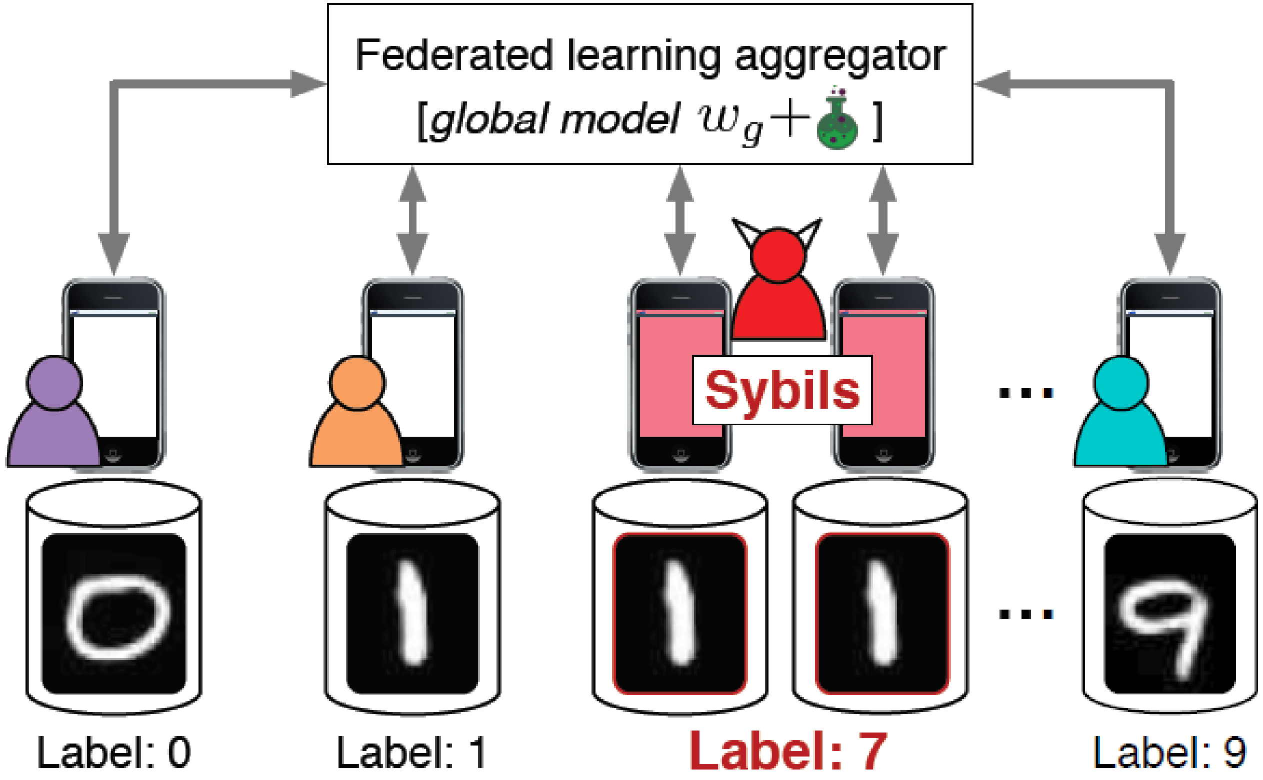
Contents

/01
Motivation

Challenges

Targeted Model Poisoning

- Derived from data poisoning
- Label flipping attack
    - Change the label of data so the model will misclassify test samples

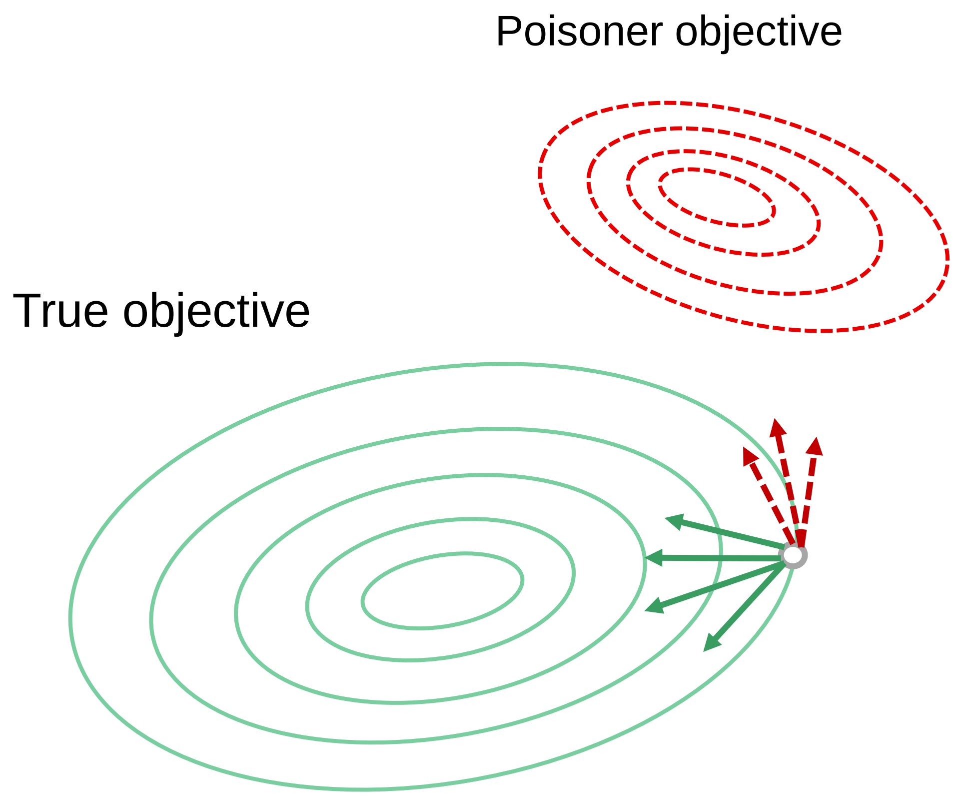


In the MNIST example, the adversary changes the label of digit 1 to 7 and uploads the poisoned model.

Observations

Attackers have a **different objective** than honest users.

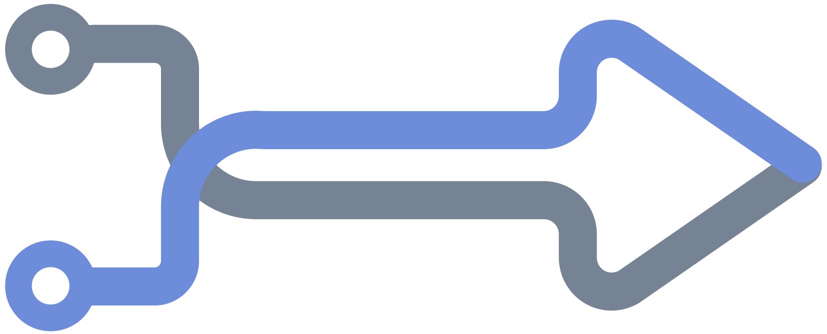
The malicious objective is more and more **obvious** as training **converges**



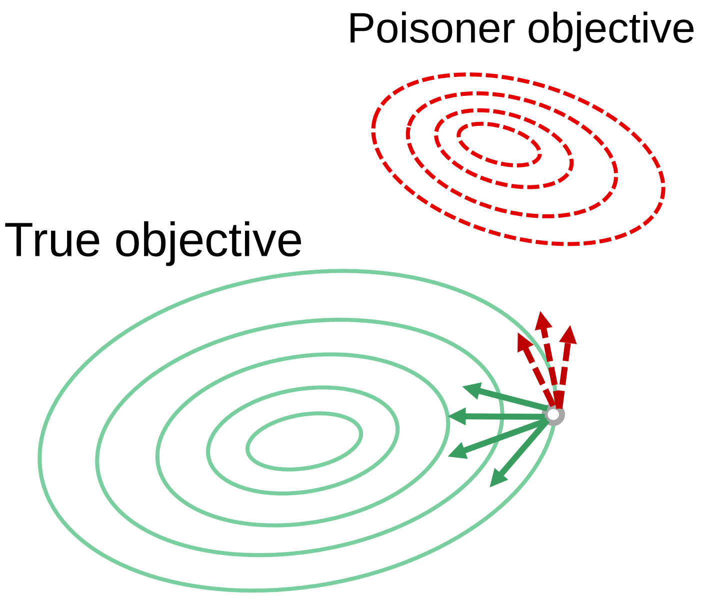
The objectives of models

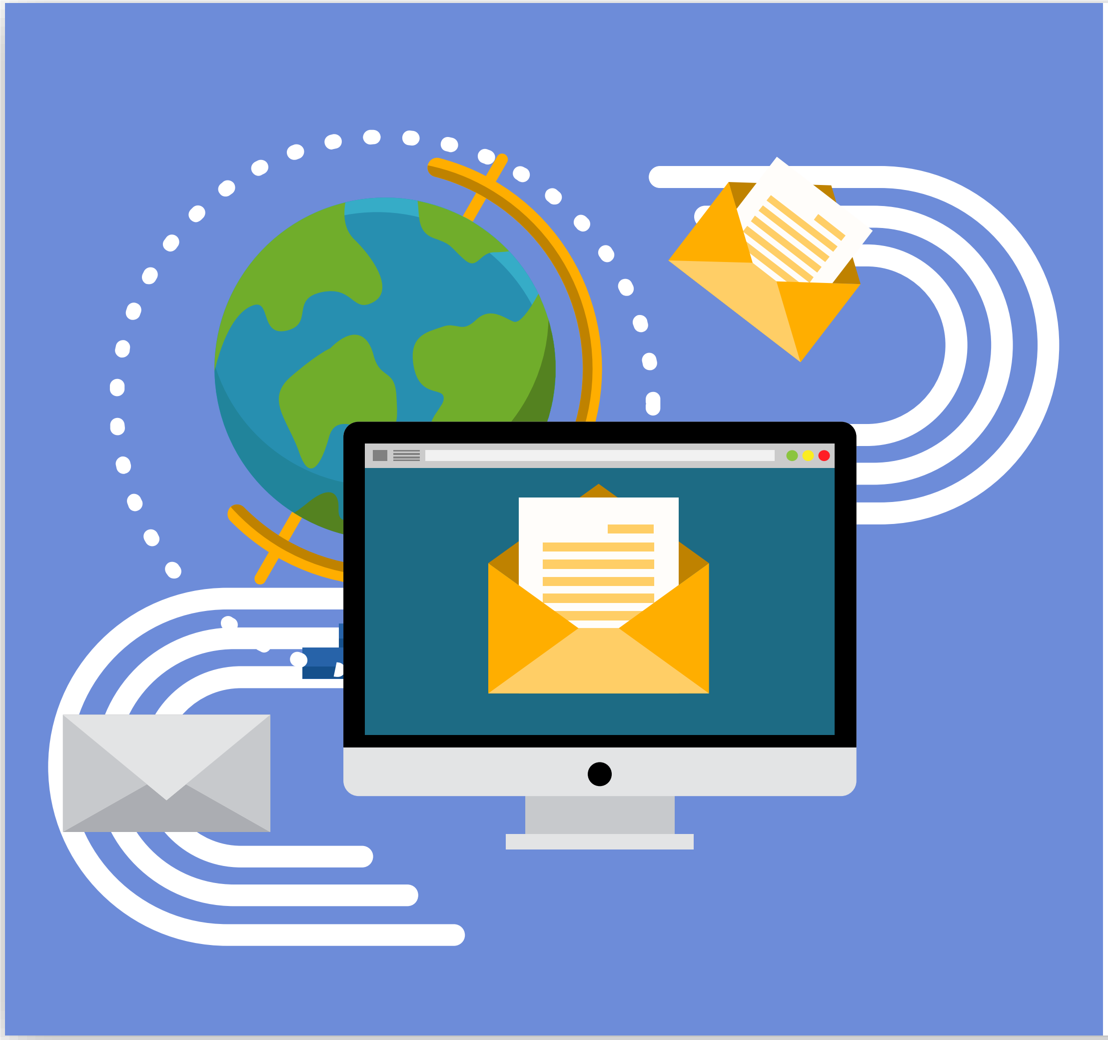
Intuition

- Attackers have a **different objective** than honest users.
- The malicious objective is more and more **obvious** as training **converges**.



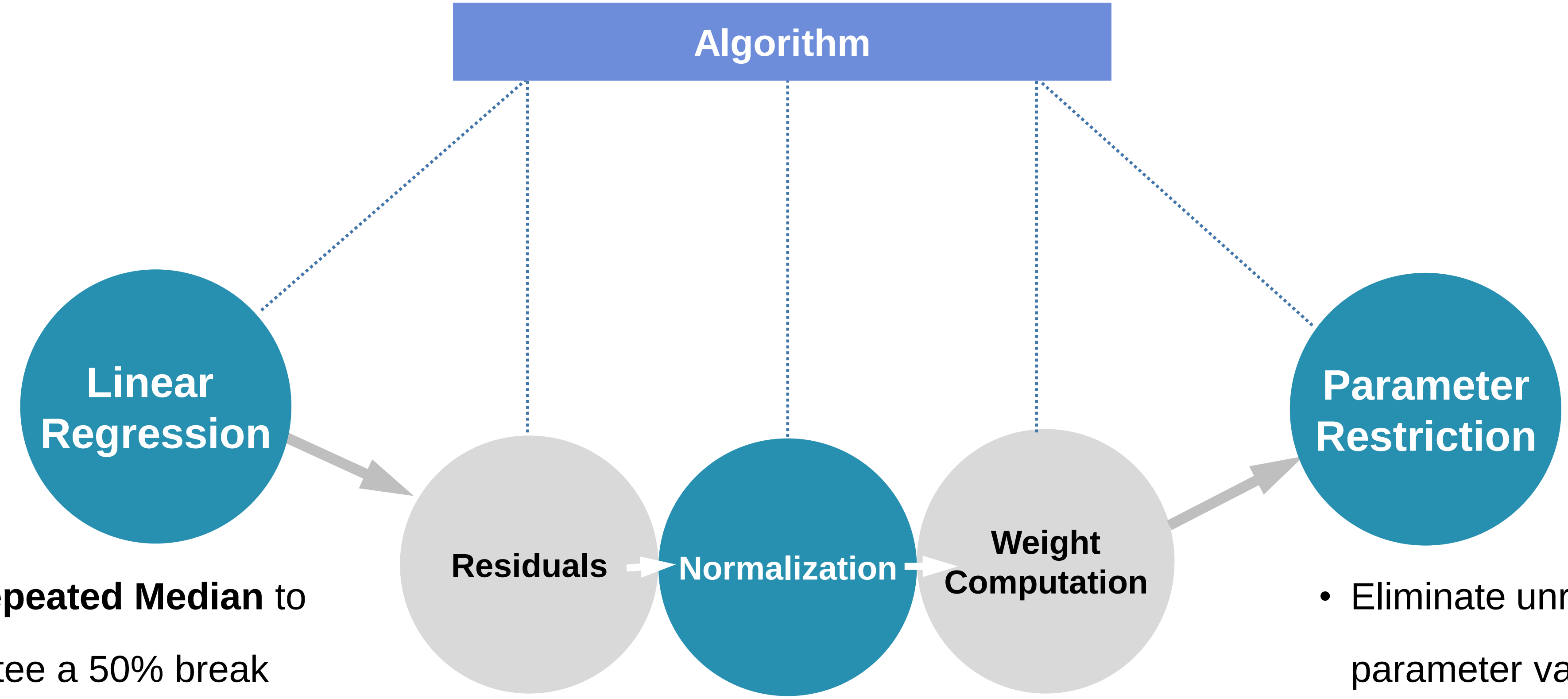
Can we design an algorithm to **detect malicious objective**, especially when the model **converges**?





/02
Method

Our Algorithm



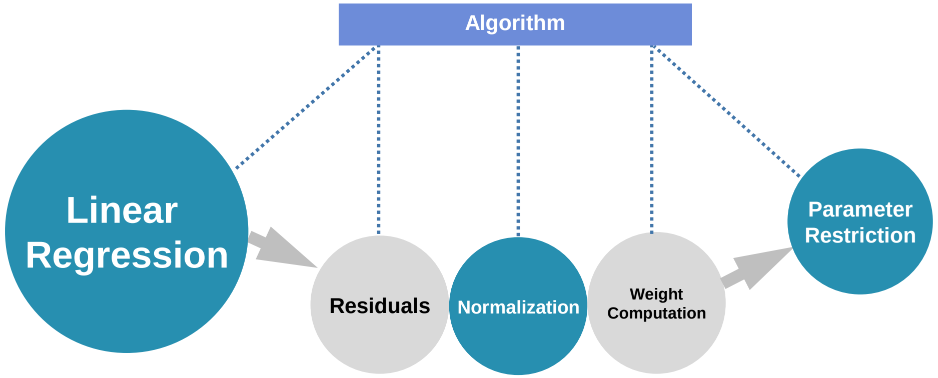
- Use **Repeated Median** to guarantee a 50% break down point

- Assign weights to models based on the **coordinate-wise residual**

- Eliminate unreasonably large parameter values

Our Algorithm

Notations: For each user k in [K], where [K] = 1, 2, …, K. We use $M^{(k)}$ to denote its model and $y_n^{(k)}$ to denote its n-th parameter. We collect each $y_n^{(k)}$ to form $\boldsymbol{y_n} = [y_n^1, y_n^2, \ldots, y_n^K]$.



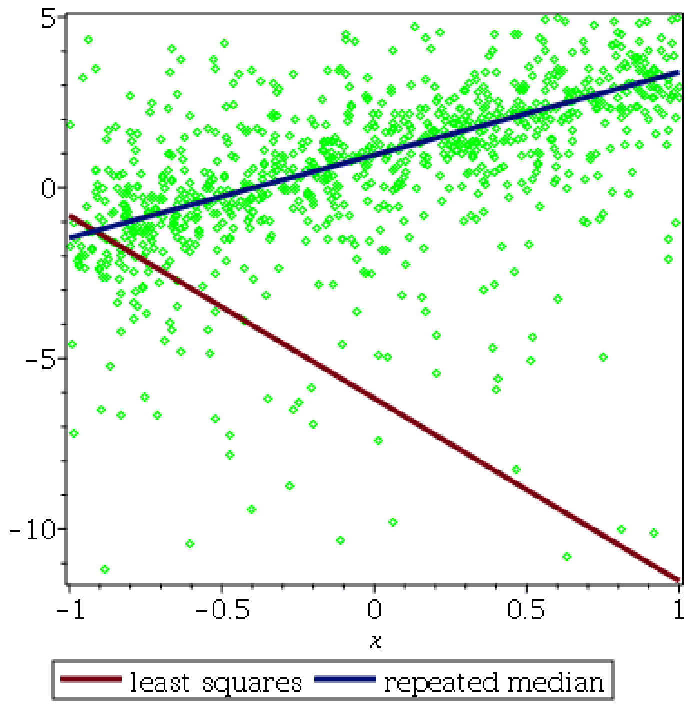
Step 1: Linear Regression

- **Repeated Median Estimation** to estimate a robust distribution of coordinate-wise parameter n.

$$\boldsymbol{y_n} = \beta_{n0} + \beta_{n1}\boldsymbol{x_n}$$

$$\beta_{n1} = \operatorname*{median}_{i} \operatorname*{median}_{i \neq j} \frac{y_n^{(j)} - y_n^{(i)}}{x_n^{(j)} - x_n^{(i)}}$$

$$\beta_{n0} = \operatorname*{median}_{i} \operatorname*{median}_{i \neq j} \frac{x_n^{(j)} y_n^{(i)} - x_n^{(i)} y_n^{(j)}}{x_n^{(j)} - x_n^{(i)}}$$



Our Algorithm

Step 2: Weight Computation

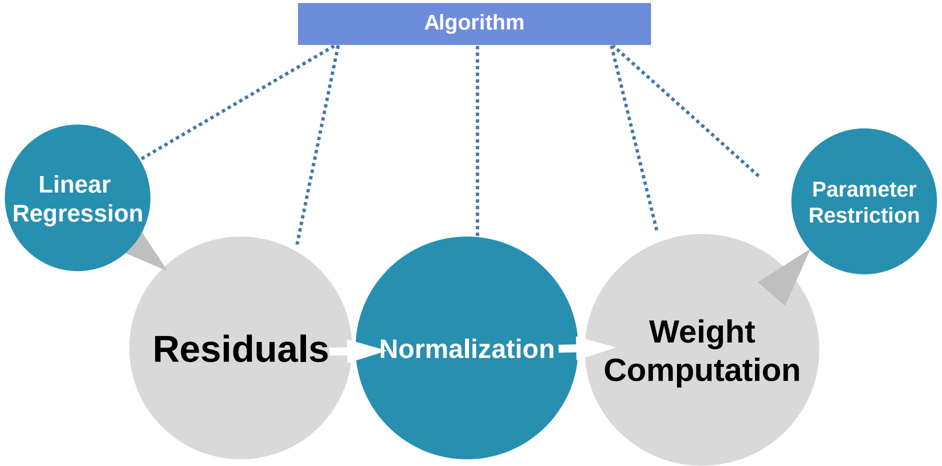
- Compute the **residuals (r)** between parameters and the estimated line.

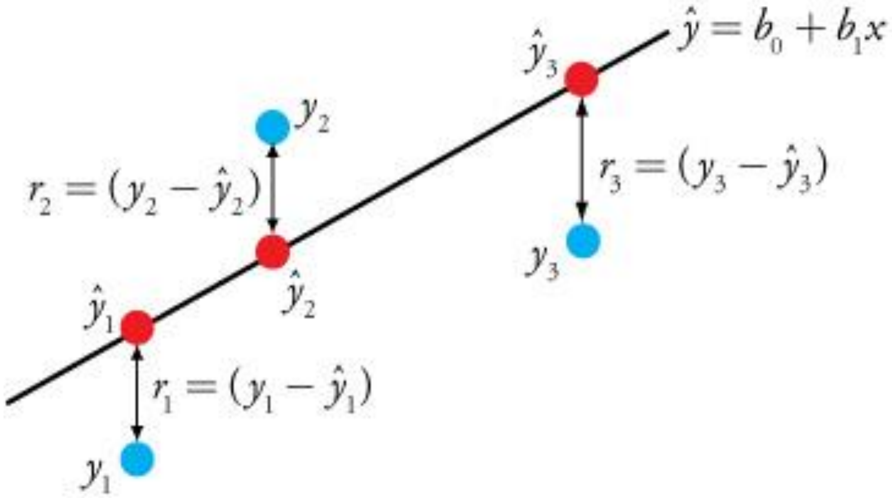
$$\boldsymbol{r_n} = \boldsymbol{y_n} - \beta_{n0} - \beta_{n1}\boldsymbol{x_n}$$

- **Normalize** residuals

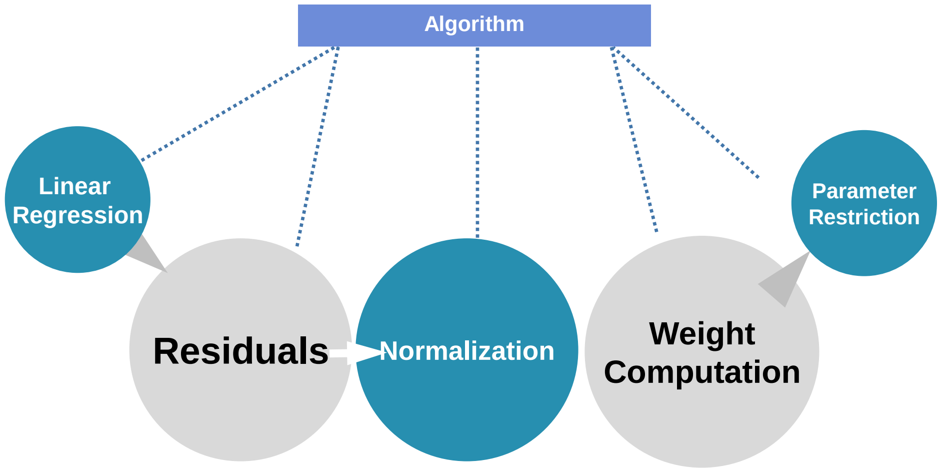
$$e_n^{(k)} = \frac{r_n^{(k)}}{\tau_n}$$, where $$\tau_n = \gamma\widetilde{|r_n|}(1 + \frac{5}{K-1})$$

and $$\widetilde{|r_n|} = \mathrm{median}(|\boldsymbol{r_n}|)$$





Our Algorithm



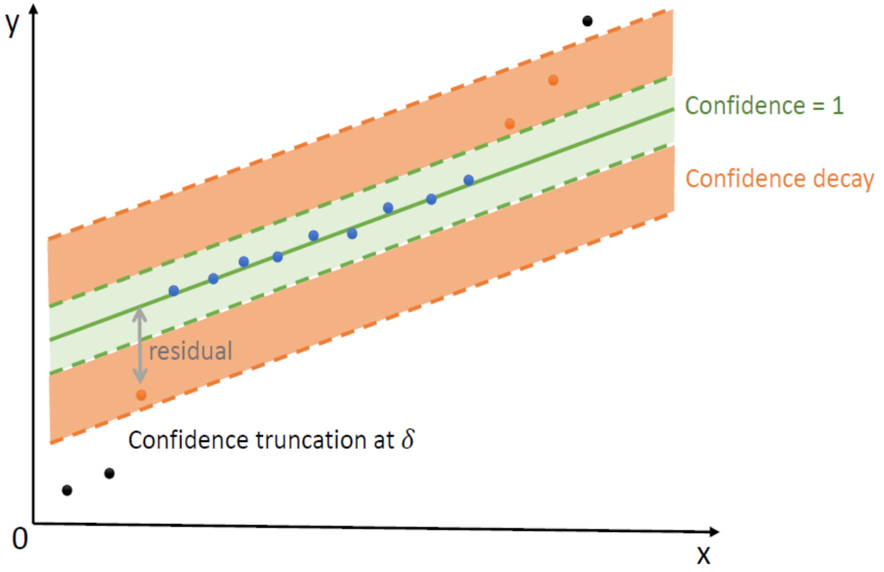
Step 2: Weight Computation (Cont.)

- Assign **weights** according to **normalized residuals (e)**.

$$w_n^{(k)} = \frac{\sqrt{1-h_{kk}}}{e_n^{(k)}} \Psi\left(\frac{e_n^{(k)}}{\sqrt{1-h_{kk}}}\right)$$

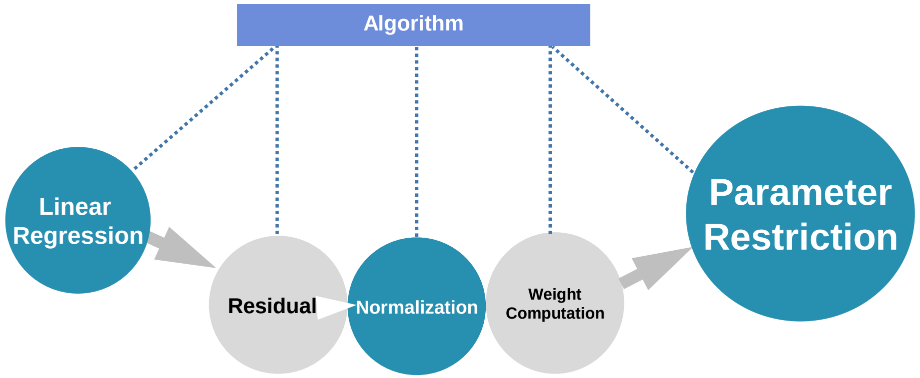
, where $\Psi(x) = max[-Z, min(Z, x)]$ with $Z = \lambda\sqrt{2/K}$
and h_{kk} is the k-th diagonal of matrix $H_n = x_n(x_n^T x_n)^{-1} x_n^T$

- Reweight weights by $w_n \leftarrow w_n \sigma(w_n)$
- Weights with larger variations will receive larger weights in the final model.



Our Algorithm

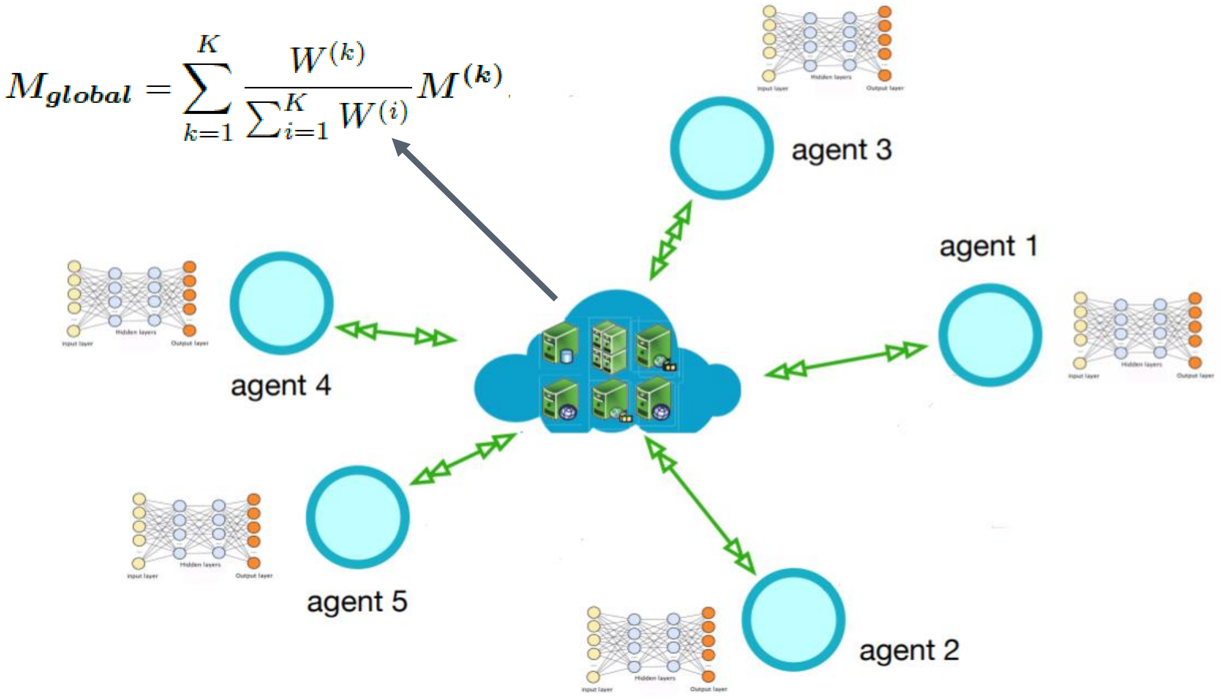
Step 3: Extreme value correction

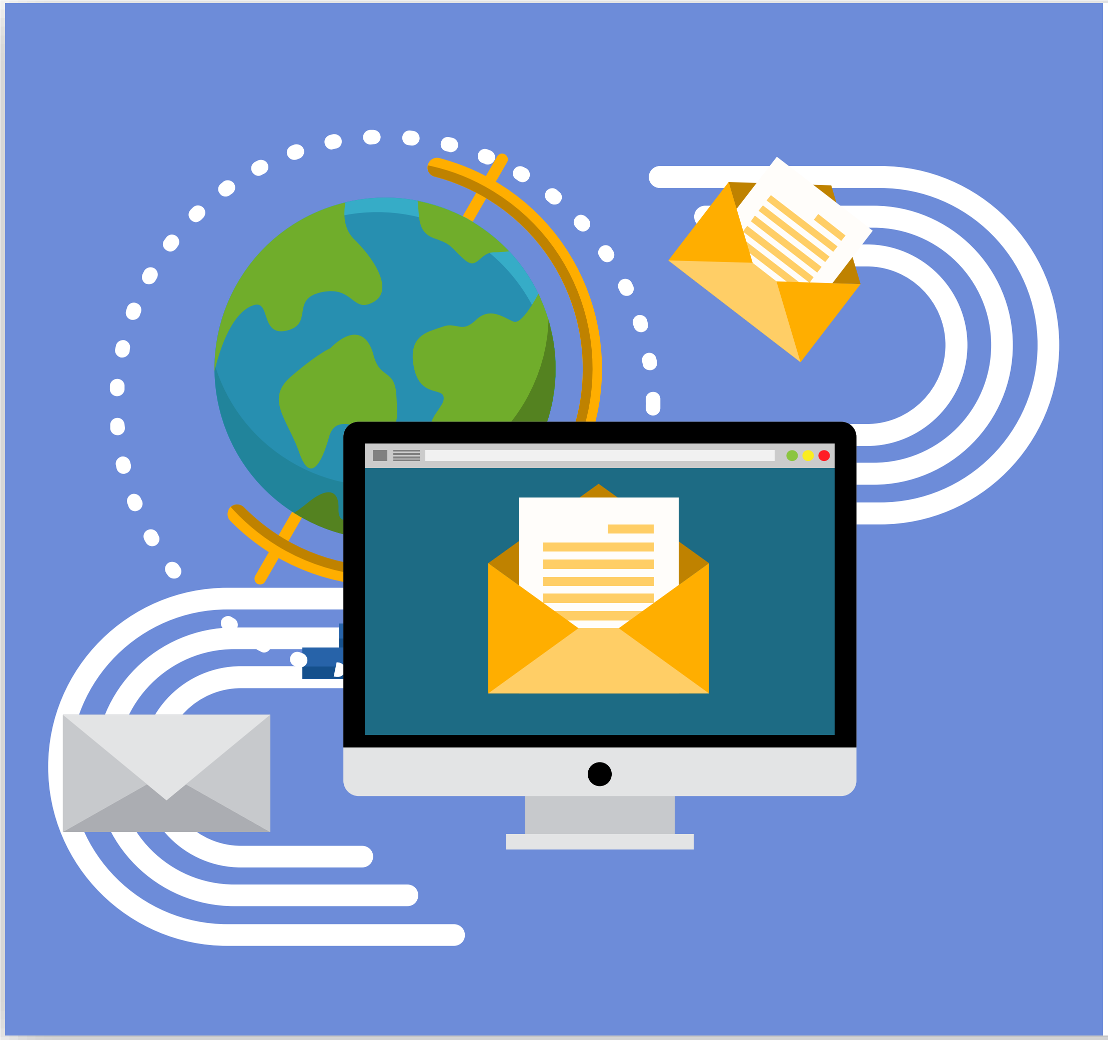


- For parameter with $w_n^{(k)}$ less than a **threshold δ**, we change its value to the corresponding value on the estimated line.
- This step removes the extreme values.

Final Step: Reweighted Aggregation

$$M_{global} = \sum_{k=1}^{K} \frac{W^{(k)}}{\sum_{i=1}^{K} W^{(i)}} M^{(k)}$$





/03

Experiments

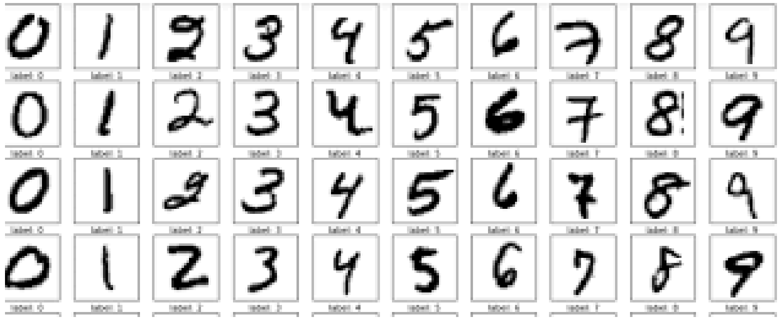
Experiments

Two Scenarios

Datasets

Label-flipping Attacks

- MNIST & CIFAR-10



Backdoor Attacks

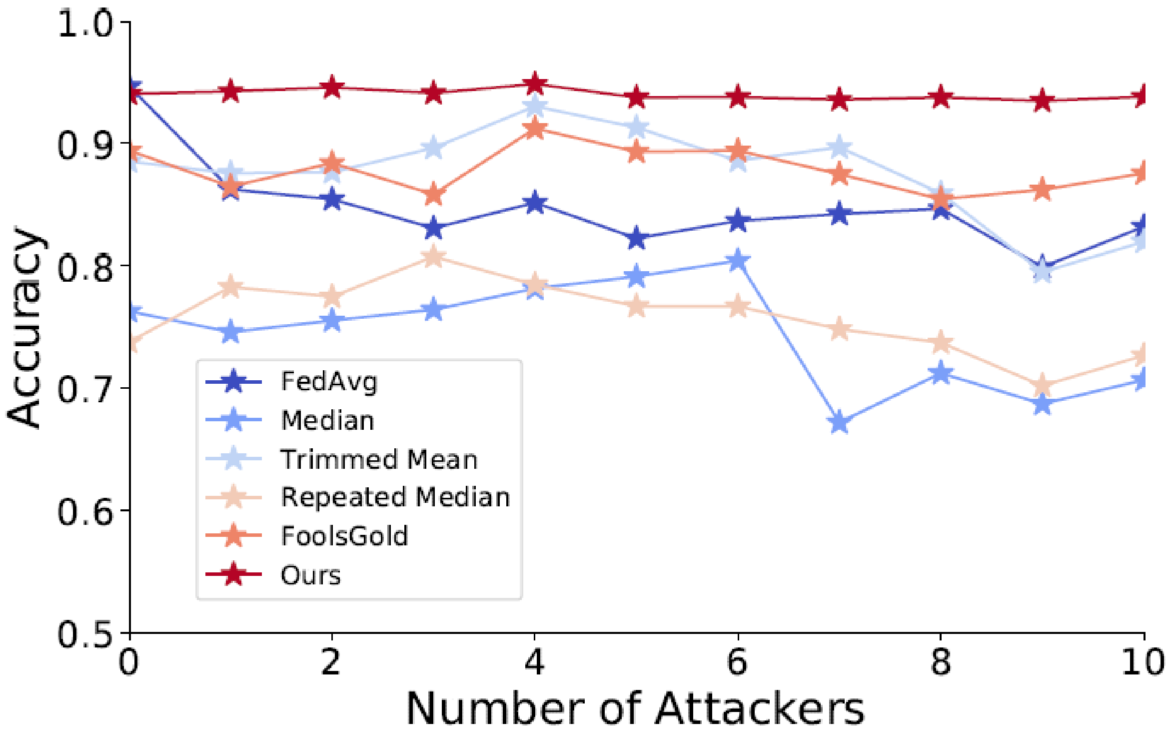

Clean image


Backdoored image

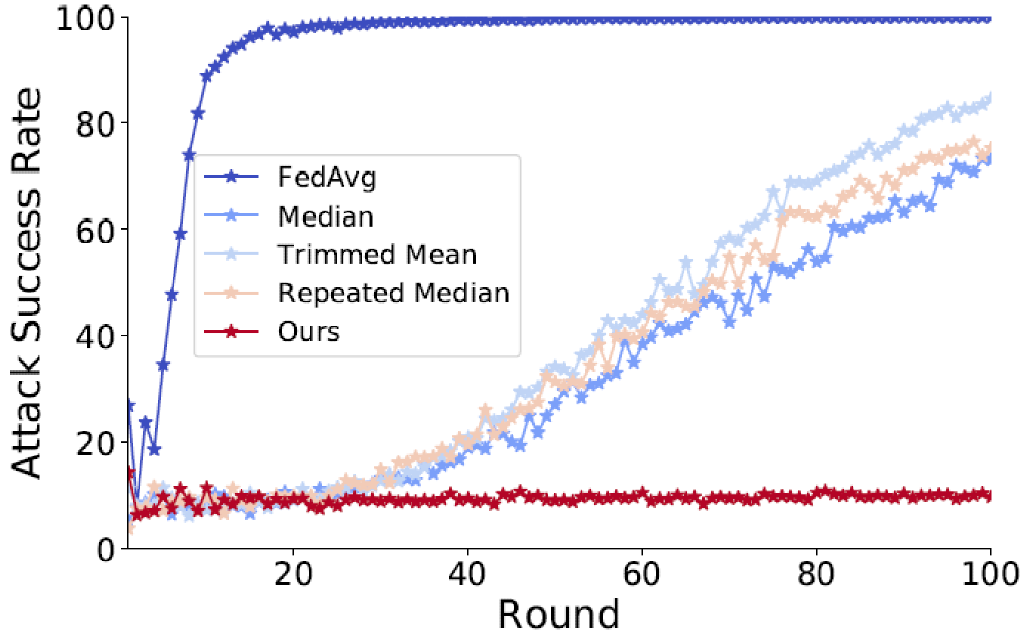
- Amazon Reviews Dataset

```
{
  "reviewerID": "A2SUAM1J3GNN3B",
  "asin": "0000013714",
  "reviewerName": "J. McDonald",
  "helpful": [2, 3],
  "reviewText": "I bought this for my husband who plays the piano.
He is having a wonderful time playing these old hymns.  The music  is
at times hard to read because we think the book was published for
singing from more than playing from.  Great purchase though!",
  "overall": 5.0,
  "summary": "Heavenly Highway Hymns",
  "unixReviewTime": 1252800000,
  "reviewTime": "09 13, 2009"
}
```

MNIST Dataset



Label-flipping Attack



Backdoor Attacker

Model Poisoning Attacks

- CIFAR-10 Dataset

# of Attackers	0	1	2	3	4
FedAvg	88.96%	85.74%	82.49%	82.35%	82.11%
Median	88.11%	87.69%	87.15%	85.85%	82.01%
Trimmed Mean	88.70%	88.52%	**87.44%**	85.36%	82.35%
Repeated Median	88.60%	87.76%	86.97%	85.77%	81.82%
FoolsGold	9.70%	9.57%	10.72%	11.42%	9.98%
Ours	**89.17%**	**88.60%**	86.66%	**86.09%**	**85.81%**

- Amazon Review Dataset

# of Attackers	0	1	2	3	4
FedAvg	**91.81%**	86.91%	24.97%	12.52%	9.78%
Median	91.73%	**91.87%**	**91.79%**	91.43%	91.17%
Trimmed Mean	**91.81%**	91.82%	91.82%	91.49%	91.26%
Repeated Median	91.55%	88.41%	23.22%	11.70%	9.62%
FoolsGold	50.79%	49.45%	47.44%	49.71%	49.95%
Ours	91.71%	91.79%	91.76%	**91.67%**	**91.38%**

Ablation Study

		Original		Median Estimator		Theil-Sen Estimator		Gaussian Weighting	
		Number of attackers		Number of attackers		Number of Attackers		Number of Attackers	
λ (or σ in Gaussian)	Delta	0	9	0	9	0	9	0	9
1	0.01	94.41%	94.54%	94.46%	**95.19%**	93.76%	92.87%	84.32%	92.22%
1	0.05	93.36%	91.15%	93.37%	93.79%	94.43%	92.70%	87.33%	90.65%
1	0.1	86.93%	89.39%	83.77%	90.93%	92.77%	94.31%	88.31%	89.23%
1	0.2	84.77%	91.40%	70.84%	80.79%	93.28%	93.63%	83.22%	90.28%
2	0.01	**94.95%**	94.86%	93.34%	94.36%	94.38%	49.28%	91.07%	92.70%
2	0.05	91.45%	93.14%	93.41%	94.86%	**95.62%**	91.65%	90.85%	93.00%
2	0.1	93.08%	91.84%	94.02%	93.48%	92.29%	93.07%	88.61%	93.15%
2	0.2	86.09%	91.43%	88.84%	92.68%	92.21%	91.70%	90.54%	90.80%
3	0.01	93.83%	94.89%	94.67%	94.68%	94.45%	75.83%	92.46%	93.18%
3	0.05	93.76%	**95.86%**	93.67%	94.52%	94.86%	**94.72%**	93.30%	**94.25%**
3	0.1	94.74%	94.13%	93.11%	91.30%	92.32%	94.70%	92.09%	93.65%
3	0.2	89.11%	93.25%	93.67%	93.76%	94.00%	93.20%	90.88%	93.26%
5	0.01	92.62%	93.77%	93.68%	84.26%	94.69%	93.27%	**94.10%**	93.58%
5	0.05	94.53%	95.28%	94.23%	94.72%	93.67%	79.91%	92.78%	93.69%
5	0.1	94.23%	94.47%	**94.88%**	94.69%	94.60%	92.85%	92.81%	93.83%
5	0.2	92.60%	94.23%	92.90%	93.87%	93.51%	91.41%	91.72%	92.93%

Table 4: The results of the controlled experiments by replacing the linear estimator or the weighting scheme with alternative methods. All the experiments are performed on the MNIST dataset with label-flipping attacks.

Thanks!