ByMI: Byzantine Machine Identification with False Discovery Rate Control

Chengde Qian [1] Mengyuan Wang[1] Haojie Ren [2] Changliang Zou [1]

[1]School of Statistics and Data Sciences, LPMC, KLMDASR and LEBPS, Nankai University, Tianjin, China
[2]School of Mathematical Sciences, Shanghai Jiao Tong University

ICML International Conference On Machine Learning

Introduction

- **Distributed system**: Samples $\{\mathbf{s}_i\}_{i=1}^N \overset{i.i.d.}{\sim} P_0$ are evenly stored in m machines $\mathcal{M}_1, \dots \mathcal{M}_m$ with $N = mn$.
- **Empirical risk minimization task**: $\min_{\boldsymbol{\theta}} \sum_{j=1}^m \sum_{i \in \mathcal{M}_j} \ell(\mathbf{s}_i; \boldsymbol{\theta})$.
- **Local gradient averages**: $\mathbf{g}_j(\boldsymbol{\theta}) = n^{-1} \sum_{i \in \mathcal{M}_j} \nabla_{\boldsymbol{\theta}} \ell(\mathbf{s}_i; \boldsymbol{\theta}) \in \mathbb{R}^d$.
- **Distributed optimization procedure**: $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \alpha m^{-1} \sum_{j=1}^m \mathbf{g}_j(\boldsymbol{\theta}_t)$.

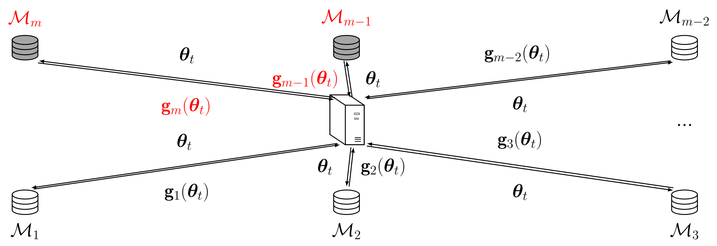


Figure 1. Byzantine attacks in distributed learning system.

- **Byzantine machines**: $\lfloor \varrho m \rfloor$ local machines on which samples are poisoned

$$\begin{cases} \mathbf{s}_i \sim P_0 & \text{if } \mathbf{s}_i \in \mathcal{M}_j \in \mathcal{G} \text{ (Good)} \\ \mathbf{s}_i \nsim P_0 & \text{if } \mathbf{s}_i \in \mathcal{M}_j \in \mathcal{B} \text{ (Byzantine)}. \end{cases}$$

- Byzantine robust learning: aggregating $\{\mathbf{g}_j\}$ via robust mean estimators (Yin et al., 2018; Zhu et al., 2023)
- A complementary task: the Byzantine machine identification

Formulate the Identification as Multiple Testing

- Translate the identification into statistical multiple testing:

$$\mathbb{H}_{0j} : \mathcal{M}_j \in \mathcal{G} \quad \text{v.s.} \quad \mathbb{H}_{1j} : \mathcal{M}_j \in \mathcal{B} \quad j \in [m].$$

- A mean test framework:

$$\mathbb{H}_{0j} : \mathbb{E}[\mathbf{g}_j(\boldsymbol{\theta}_0)] = \boldsymbol{\mu}^* \quad \text{v.s.} \quad \mathbb{H}_{1j} : \mathbb{E}[\mathbf{g}_j(\boldsymbol{\theta}_0)] \neq \boldsymbol{\mu}^* \quad j \in [m],$$

with some given $\boldsymbol{\theta}_0$. The true center $\boldsymbol{\mu}^*$ can be robustly estimated

- **Criterion for multiple testing**: false discovery proportion and true positive proportion,

$$\text{FDP}(\widehat{\mathcal{B}}) = \frac{|\widehat{\mathcal{B}} \cap \mathcal{G}|}{|\widehat{\mathcal{B}}| \vee 1}, \quad \text{TPP}(\widehat{\mathcal{B}}) = \frac{|\widehat{\mathcal{B}} \cap \mathcal{B}|}{|\mathcal{B}|}.$$

Goal: $\text{FDR} = \mathbb{E}(\text{FDP}) \leq \alpha$.

Failure of traditional Outlier detection via p-values

- **Outlier detection methods** (Filzmoser et al., 2008; Ro et al., 2015):
  1. Obtain robust center $\hat{\boldsymbol{\mu}}$
  2. Build some mean test statistics and obtain the p-values based on the asymptotic theory
  3. Apply the BH method (Benjamini and Hochberg, 1995)
- Problem of the strategy: the asymptotic distributions of the statistics are heavily dimension-dependent.

Hotelling's T^2 for fixed/small d v.s. Gaussian for large d.

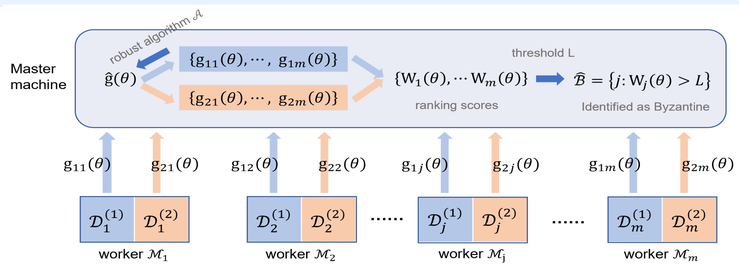
- Our solution: A sample-splitting-based, p-value-free and dimension-agnostic procedure, ByMI.

ByMI Procedure

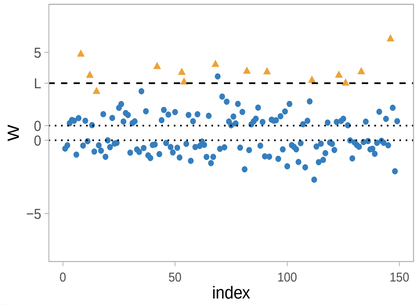
1. Randomly split the samples on each $\mathcal{M}_j$ into $\{\mathcal{D}_j^{(k)}\}_{k=1,2}$ with equal size.
2. Obtain a robust mean estimator $\widehat{\mathbf{g}}(\boldsymbol{\theta})$ based on $\mathcal{D}_j^{(1)}$
3. Compute score $W_j = \{\mathbf{g}_{1j}(\boldsymbol{\theta}) - \widehat{\mathbf{g}}(\boldsymbol{\theta})\}^\top \boldsymbol{\Omega} \{\mathbf{g}_{2j}(\boldsymbol{\theta}) - \widehat{\mathbf{g}}(\boldsymbol{\theta})\}$, $j \in [m]$.
4. Choose the threshold $L > 0$ as

$$L = \inf\left\{ \ell > 0 : \frac{1 + \#\{j : W_j \leq -\ell\}}{\#\{j : W_j \geq \ell\} \vee 1} \leq \alpha \right\},$$

FDR level $\alpha > 0$; Detected Byzantine machines $\widehat{\mathcal{B}} = \{\mathcal{M}_j : W_j \geq L\}$.



- **ByMI is p-value free and dimension-insensitive**. Conditional on $\{\mathcal{D}_j^{(1)}\}$, W_j is an univariate projection of $\mathbf{g}_{2j}(\boldsymbol{\theta}) - \widehat{\mathbf{g}}(\boldsymbol{\theta})$ and enjoys the symmetric property.



$$\begin{aligned} &\text{FDP}(L) \\ =&\frac{\#\{j : W_j \leq -L\}}{\#\{j : W_j \geq L\} \vee 1} \\ \times&\frac{\#\{j : W_j \geq L, j \in \mathcal{G}\}}{\#\{j : W_j \leq -L\}} \\ \leq&\alpha \times \frac{\#\{j : W_j \geq L, j \in \mathcal{G}\}}{\#\{j : W_j \leq -L, j \in \mathcal{G}\}} \\ \approx&\alpha \text{ (by symmetricity of } W_j). \end{aligned}$$

- Choices of $\boldsymbol{\Omega}$.
  Scale matrix: $\boldsymbol{\Omega} = \text{diag}\{\hat{\sigma}_1^{-2}, \dots, \hat{\sigma}_d^{-2}\}$
  Projection matrix: $\boldsymbol{\Omega} = \mathbf{v}_1 \mathbf{v}_1^\top$, where $\mathbf{v}_1$ is the first eigenvector of $\text{Cov}(\{\mathbf{g}_{1j}(\boldsymbol{\theta})\})$.

Theoretical Guarantees

- **Conditions**: $\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}^*\| = O(\delta_\mu)$; L_q-L_2-norm equivalence of the gradients;
- **Finite-sample FDR control.**

Denote $\omega_n = n^{-(1-2\eta^2)\kappa/2}$ with some $\eta \in (0, 1/\sqrt{2})$. With probability at least $1 - O(mn^{-\eta^2\kappa/2})$ ($\kappa = \min(1, q-2)$),

$$\text{FDR}(\hat{\mathcal{B}}) \leq \alpha + O\left(\sqrt{\omega_n + n^{\frac{1}{2}+\eta^2\kappa}\delta_\mu}\right).$$

- **Finite-sample FDP control under signal condition.**

Further assume that $\|\boldsymbol{\mu}_j^* - \boldsymbol{\mu}^*\| \geq C(\sqrt{\frac{\log n}{n}} + \delta_\mu + d^{\frac{1}{2}} n^{-\frac{1}{2}+\frac{\kappa_2}{q}})$. With probability at least $1 - O(mn^{-\frac{\eta^2\kappa}{2}} + \psi_m^{-(1-\delta)})$,

$$\text{FDP}(\hat{\mathcal{B}}) \leq \alpha\left[1 + O\left(s_{nm} + n^{\frac{1}{2}}\delta_\mu\right)\right],$$

where $s_{nm} = n^{-\frac{(1-\eta^2)\kappa}{2}}(\log n)^{\frac{1}{2}} + mn^{-\frac{\eta^2\kappa}{2}} + (\alpha\psi_m)^{-\delta/3}$, $\eta \in (0, 1)$.

Real data analysis

Identification result.

Attack	Method	MNIST FDR	MNIST TPR	MNIST P_a	F-MNIST FDR	F-MNIST TPR	F-MNIST P_a	CIFAR10 FDR	CIFAR10 TPR	CIFAR10 P_a
OOD	ByMI-Filter	6.8	98.1	94.2	7.1	92.8	83.4	5.5	81.7	73.6
	ByMI-Filter+	6.6	97.2	95.8	6.5	85.0	79.0	5.7	77.1	73.8
	ByMI-GEOM	10.2	97.4	93.4	9.1	92.1	80.8	7.4	76.4	67.2
	RMDP-BH	88.3	99.9	99.8	68.9	98.6	96.8	54.1	84.6	75.0
	Krum	27.1	93.8	81.6	36.4	81.7	55.4	52.5	59.4	34.8
	FABA	25.2	96.2	89.6	32.2	87.1	69.8	50.6	61.8	45.4
	Zeno	26.2	94.8	84.8	35.6	82.8	56.2	54.8	56.5	32.4
IPM	ByMI-Filter	6.3	99.6	99.6	6.5	99.6	99.6	5.4	100.0	100.0
	ByMI-Filter+	6.8	99.2	99.2	6.0	97.0	97.0	7.7	100.0	100.0
	ByMI-GEOM	10.2	96.2	96.2	10.4	97.0	97.0	10.0	99.6	99.6
	RMDP-BH	89.2	96.2	96.2	75.3	78.6	78.6	79.6	7.6	7.6
	Krum	75.2	31.9	22.6	59.1	52.6	42.2	78.9	26.4	17.2
	FABA	88.5	14.8	8.2	72.7	35.1	26.8	96.1	4.9	2.6
	Zeno	88.3	15.0	7.8	71.2	37.0	25.8	95.2	6.0	2.2

Table 1. FDR(%), TPR(%) and P_a(%) of the OOD and IPM attacks when $\varrho = 0.1$. ($a = 0.2$ in the IPM attack.)

Apply to robust learning task.

- We apply ByMI to detect Byzantine machines and use the simple mean aggregation of the left local gradients to train the model.

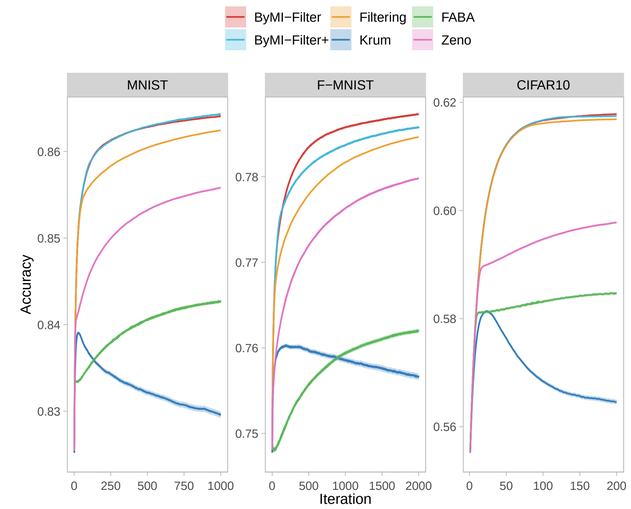


Figure 2. Test accuracy under the IPM attack with the contamination level $\varrho = 0.3$ when $a = 2$.

References

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995.

Peter Filzmoser, Ricardo Maronna, and Mark Werner. Outlier identification in high dimensions. *Computational Statistics & Data Analysis*, 52(3):1694–1711, 2008.

Kwangil Ro, Changliang Zou, Zhaojun Wang, and Guosheng Yin. Outlier detection for high-dimensional data. *Biometrika*, 102(3):589–599, 2015.

Dong Yin, Yudong Chen, Ramchandran Kannan, and Peter Bartlett. Byzantine-robust distributed learning: Towards optimal statistical rates. In *International Conference on Machine Learning*, pages 5650–5659. PMLR, 2018.

Banghua Zhu, Lun Wang, Qi Pang, Shuai Wang, Jiantao Jiao, Dawn Song, and Michael I. Jordan. Byzantine-Robust Federated Learning with Optimal Statistical Rates. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent, editors, *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 3151–3178. PMLR, 2023.