ByMI: Byzantine Machine Identification with False Discovery Rate Control

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

Abstract

Various robust estimation methods or algorithms have been proposed to hedge against Byzantine failures in distributed learning. However, there is a lack of systematic approaches to provide theoretical guarantees of significance in detecting those Byzantine machines. In this paper, we develop a general detection procedure, ByMI, via error rate control to address this issue, which is applicable to many robust learning problems. The key idea is to apply the sample-splitting strategy on each worker machine to construct a score statistic integrated with a general robust estimation and then to utilize the symmetry property of those scores to derive a data-driven threshold. The proposed method is dimension insensitive and p-value free with the help of the symmetry property and can achieve false discovery rate control under mild conditions. Numerical experiments on both synthetic and real data validate the theoretical results and demonstrate the effectiveness of our proposed method on Byzantine machine identification.

1. Introduction

With the rapid growth of the dataset size and the decentralization of data sources, distributed and federated learning, where the worker machines locally preserve the data and only communicate summarized information with the master machine, have received substantial attention (Kairouz et al., 2021). In such a distributed learning system, there is usually a small fraction of worker machines that send any arbitrary information due to malicious attacks on worker machines and communication channels, or the variation and contamination in the data sources (Blanchard et al., 2017). Those abnormal machines are called *Byzantine machines*. Such breaks in the distributed system are modeled as the *Byzantine failures* (Lamport et al., 1982) and have serious adverse effects on learning performance.

1.1. Motivation & Related Works

Byzantine robust learning. Recently, there is a great deal of effort to develop distributed-learning algorithms with the properties of *Byzantine resilience* and *Byzantine robustness*, that are provably robust against Byzantine failures. Typically, the studies of Byzantine-robust distributed learning focus on aggregating those workers' messages, e.g. the averages of gradients, via some robust estimation algorithms in the master machine. This has an intimate connection with robust estimation in statistics literature (Shi et al., 2022). A family of aggregation methods is replacing the simple sample mean with robust location estimations. One popular choice is to take the median instead of the sample mean, such as the coordinate-wise median and trimmed mean (Yin et al., 2019; 2018), the geometric median (Chen et al., 2017), and the coordinate-wise composite quantile (Tu et al., 2021). However, those median-type algorithms suffer a bias dependent on the dimension due to Byzantine failures. Some recent works adopt the computational-efficient high-dimensional mean estimations to correct the bias term to a dimension-agnostic one. For instance, Yin et al. (2019) applies the filtering algorithm (Lai et al., 2016; Diakonikolas et al., 2017) as the aggregation rule and Zhu et al. (2023) considers both the filtering algorithm and the first-order approach (Cheng et al., 2020; Zhu et al., 2022).

Another direction is to detect and delete Byzantine machines and further make estimations based on some *reputation scores* which measure the trustworthiness of worker machines, such as Krum (Blanchard et al., 2017), FABA (Xia et al., 2019) and Zeno (Xie et al., 2019). As commented by Cheng et al. (2019), all of these methods suffer similar dimension-dependent bias as the median-type algorithms. That's partly because those methods take the number of Byzantine machines as predetermined and are unable to give a significant guarantee against underestimation or overestimation, which may hamper their applicability.

Outlier detection methods. As discussed above, the identification of Byzantine machines plays an important role in the resilience task but has received less attention. This is rel-

[*]Equal contribution [1]School of Statistics and Data Sciences, LPMC, KLMDASR and LEBPS, Nankai University, Tianjin, China [2]School of Mathematical Sciences, Shanghai Jiao Tong University. Correspondence to: Haojie Ren <haojieren@sjtu.edu.cn>, Changliang Zou <zoucl@nankai.edu.cn>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

evant to outlier detection in the statistics regime since each Byzantine machine performs like one outlier. Traditionally, one outlier detection procedure is generally to obtain some robust center estimates and then to compute p-values based on some efficient tests to evaluate whether it is one outlier (Filzmoser et al., 2008; Ro et al., 2015; Zimek et al., 2012). It has been revealed that the performance of this kind of methods heavily depends on the approximation accuracy of p-values (Efron, 2004; Liu & Shao, 2014), which are obtained from the asymptotic distribution when the sample size or dimension goes to infinity with some specific rate. However, it is often unrealistic, especially for one complex distributed learning system where the approximation distribution is hard to estimate. In addition, as many machine learning algorithms have to face the situation that the parameter dimension is much larger than the sample size in each machine, it makes the traditional p-value dependent methods largely ineffective (Bottou et al., 2018). Hence, it is important to design a detection procedure that is insensitive or free of dimensions.

1.2. Our Contributions

In this paper, we suggest a *p-value free* and *dimension insensitive* detection procedure, named as *Byzantine Machines Identification* (ByMI). To avoid falsely identifying too many normal machines, we consider controlling the false discovery rate (FDR), the expectation of the proportion of the false discoveries among all the discoveries. The FDR control has been fully explored in the literature of multiple testing and is a particularly useful tool to maintain the ability to detect true alternatives without excessive false positive ones (Benjamini & Hochberg, 1995; Benjamini & Yekutieli, 2001; Du et al., 2021; Zou et al., 2020). The proposed method integrates the classical Byzantine robust estimation with the generic idea of the sample-splitting strategy to construct a series of score statistics with the symmetry property, which plays an important role in distinguishing Byzantine and normal machines. Then the ByMI entails choosing a data-driven threshold by exploiting the empirical distributions between the negative and positive statistics.

To our best knowledge, this is the first effort to systematically identify those Byzantine machines with error rate control. ByMI's main contributions/advantages include:

- Under a unified framework, ByMI addresses how to detect Byzantine machines in the regime of gradient functions. It is dimension insensitive since the ranking scores are some univariate projection of gradients and applicable to many Byzantine problems, such as the mean estimation and the communication-efficient distributed learning procedure.

- ByMI is p-value free and can achieve the generic finite-sample upper bound of FDR without strong model/distribution assumptions. Under mild conditions, we show that the proposed ByMI method yields valid FDR control and sure-detection property.

- ByMI can be easily coupled with robust estimations. Extensive numerical experiments indicate that ByMI is able to yield accurate FDR control, while significantly detecting most Byzantine machines compared to existing outlier detection algorithms.

2. Byzantine Machines Identification Procedure

2.1. Problem Formulation

In the distributed system, assume N independent samples $\{\mathbf{s}_i\}_{i=1}^N$ are evenly stored in $m+1$ machines $\mathcal{M}_0, \mathcal{M}_1 \dots \mathcal{M}_m$, each of which contains n observations and $N = (m+1)n$. Here, $\mathbf{s}$ could either be a p-variate random vector $\mathbf{x} \in \mathbb{R}^p$ or $(y, \mathbf{x})$ with y and $\mathbf{x}$ being respectively the response variable and p-variate covariates. Note that $\mathcal{M}_0$ is the master machine that is in charge of integrating information from worker machines $\mathcal{M}_1, \dots, \mathcal{M}_m$ and cannot be corrupted. Considering Byzantine failures in the system, there exist $\lfloor \varrho m \rfloor$ Byzantine machines on which samples are poisoned, where $\varrho \in [0, 1]$ is the proportion of Byzantine machines. Denote the Byzantine machines set and the good/normal machines set as $\mathcal{B}$ and $\mathcal{G}$, respectively, with $\mathcal{B} \cup \mathcal{G} = \{\mathcal{M}_1, \dots, \mathcal{M}_m\}$. Assume the normal data are i.i.d drawn from P_0. We consider the behavior of the Byzantine machines by the Huber contamination model (Huber, 1964),

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

Our goal is to propose a procedure to identify the Byzantine machines set $\mathcal{B}$. From the perspective of multiple testing, the null hypothesis of the j-th machine is that it is normal, and the alternative asserts it is a Byzantine machine. Namely, the Byzantine machines detection problem is translated to the multiple testing problem:

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

If one detection procedure yields the Byzantine machines set estimation $\widehat{\mathcal{B}}$, the false discovery proportion (FDP) and true positive proportion (TPP) with $\widehat{\mathcal{B}}$ are

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

The false discovery rate (FDR) and true positive rate (TPR) are defined as the expectation of the FDP($\widehat{\mathcal{B}}$) and TPP($\widehat{\mathcal{B}}$)

respectively. One reliable detection procedure is to control FDR at a target level and identify as many Byzantine machines as well.

2.2. Byzantine Machines Identification Procedure

We consider a generic distributed risk minimization framework. Let $\ell(\mathbf{s}, \boldsymbol{\theta})$ be the loss function with parameter $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^d$. Denote its gradient function as $\mathbf{g}(\mathbf{s}, \boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \ell(\mathbf{s}, \boldsymbol{\theta}) \in \mathbb{R}^d$. In a distributed learning problem, each worker machine $\mathcal{M}_j$ computes an empirical gradient based on its local samples by $\mathbf{g}_j(\boldsymbol{\theta}_0) = \frac{1}{n} \sum_{i \in \mathcal{M}_j} \mathbf{g}(\mathbf{s}_i, \boldsymbol{\theta}_0)$ with some given parameter $\boldsymbol{\theta}_0$. Then these gradients are transmitted to the master machine $\mathcal{M}_0$ for further updating or estimating parameters. Note that $\mathbf{g}_j(\boldsymbol{\theta}_0)$ reflects the state of the worker machine, since we usually have $\mathbb{E}[\mathbf{g}_j(\boldsymbol{\theta}_0)] \neq \mathbb{E}[\mathbf{g}_k(\boldsymbol{\theta}_0)]$ where $\mathcal{M}_j \in \mathcal{B}$ is a Byzantine machine but $\mathcal{M}_k \in \mathcal{G}$ is a normal one. With this regime, the original problem (1) is reframed into a multiple testing problem on detecting the difference in mean of these gradients $\{\mathbf{g}_j\}_{j=1}^m$. Denote $\boldsymbol{\mu}^*$ as the mean of $\mathbf{g}_j(\boldsymbol{\theta}_0)$ when $\mathcal{M}_j \in \mathcal{G}$ is one normal machine. Then the alternative hypotheses can be written as $\mathbb{H}_{1j} : \mathbb{E}[\mathbf{g}_j(\boldsymbol{\theta}_0)] \neq \boldsymbol{\mu}^*$ for $j \in [m]$.

The standard procedure for multiple testing is to build some mean test statistics for each worker machine and their asymptotic distribution under null, such as Hotelling's T^2 for fixed d or other modified tests in high-dimension regime (Bai & Saranadasa, 1996; Chen & Qin, 2010), and then apply the Benjamini-Hochberg (BH) method to the approximated p-values of those test statistics (Benjamini & Hochberg, 1995). However, the performance of the BH method heavily depends on the accuracy of p-values from the asymptotic distribution, which usually involves some unknown quantity related to the gradient populations and may be different based on the diverging rate of d relative to n. It implies that traditional mean tests become ineffective or practically infeasible in modern machine learning models, as it's challenging to estimate these quantities for complex gradients, and it's hard to determine the asymptotic distribution to approximate p-values.

This promotes the development of our p-value free and dimension-insensitive detection procedure, named as *Byzantine Machines Identification* (ByMI). We construct one new test statistic with the sample-splitting strategy and employ the empirical distribution in place of the asymptotic distribution to achieve FDR control.

Step 1. The first step of our procedure is to randomly split the samples on each worker machine $\mathcal{M}_j$ into two sets $\mathcal{D}_j^{(1)}$ and $\mathcal{D}_j^{(2)}$ of equal size $\frac{n}{2}$, $j \in [m]$. Write $\mathbf{g}_{1j}(\boldsymbol{\theta})$ and $\mathbf{g}_{2j}(\boldsymbol{\theta})$ as the empirical gradient functions based on $\mathcal{D}_j^{(1)}$ and $\mathcal{D}_j^{(2)}$ with some given $\boldsymbol{\theta}$, respectively.

Step 2. Based on $\{\mathbf{g}_{1j}(\boldsymbol{\theta})\}_{j=0}^m$, we employ a particular algorithm $\mathcal{A}$ to obtain the robust mean estimator of $\boldsymbol{\mu}^*$, denoted as $\widehat{\mathbf{g}}(\boldsymbol{\theta})$. Many methods can be chosen as $\mathcal{A}$, such as those median-type algorithms (Su & Xu, 2019; Tu et al., 2021; Yin et al., 2018) and the dimension-agnostic algorithms (Cheng et al., 2020; Diakonikolas et al., 2017; Lai et al., 2016; Zhu et al., 2022). In general, one more robust and precise estimator leads to more reliable detection results (Ro et al., 2015).

Step 3. Then, we construct the ranking score which provides evidence that $\mathcal{M}_j$ may be one Byzantine machine. For $j \in [m]$, let

$$W_j = \{\mathbf{g}_{1j}(\boldsymbol{\theta}) - \widehat{\mathbf{g}}(\boldsymbol{\theta})\}^\top \boldsymbol{\Omega} \{\mathbf{g}_{2j}(\boldsymbol{\theta}) - \widehat{\mathbf{g}}(\boldsymbol{\theta})\}. \quad (2)$$

Here $\boldsymbol{\Omega}$ can serve as a rough scale estimator for standardization or can play as a projection matrix for the projection-based detection methods (Ren et al., 2017), which will be further discussed in Section 2.4. Notice that W_j's play an important role in distinguishing Byzantine and normal machines. Intuitively, a large positive W_j indicates that $\mathcal{M}_j$ is likely to be the Byzantine machine. For $\mathcal{M}_j \in \mathcal{G}$, W_j is (asymptotic) symmetric with mean zero due to the central limit theorem and independence between $\mathcal{D}_j^{(1)}$ and $\mathcal{D}_j^{(2)}$.

Step 4. That further inspires us to choose the threshold $L > 0$ as

$$L = \inf \left\{ \ell > 0 : \frac{1 + \#\{j : W_j \leq -\ell\}}{\#\{j : W_j \geq \ell\} \vee 1} \leq \alpha \right\}, \quad (3)$$

for the target FDR level $\alpha > 0$. Finally, the identified Byzantine machine set is $\widehat{\mathcal{B}} = \{\mathcal{M}_j : W_j \geq L\}$. If the set is empty, we simply set $L = +\infty$. Intuitively, $\#\{j : W_j \leq -\ell\}$ is an overestimation of $\#\{j : W_j \leq -\ell, \mathcal{M}_j \in \mathcal{G}\}$, which further is a good approximation to $\#\{j : W_j \geq \ell, \mathcal{M}_j \in \mathcal{G}\}$, the number of false discoveries, due to W_j's symmetry property for those normal machines. Thus, it implies that the fraction in (3) is an overestimation of FDP.

The test statistic W_j in (2) indeed has a similar form to the traditional mean test, i.e., $W_j' = \{\mathbf{g}_j(\boldsymbol{\theta}) - \widehat{\mathbf{g}}(\boldsymbol{\theta})\}^\top \boldsymbol{\Omega} \{\mathbf{g}_j(\boldsymbol{\theta}) - \widehat{\mathbf{g}}(\boldsymbol{\theta})\}$ or its variants when $\boldsymbol{\Omega}$ estimates the precision matrix of $\mathbf{g}_j$ (Chen & Qin, 2010), but they are distinguished in that our ByMI procedure does not rely on the p-values from the asymptotic distribution. This is especially important since the asymptotic distribution heavily depends on the dimension d relative to n. The asymptotic distribution of W_j' can be a chi-square for one fixed or small d and a normal distribution for a large d, making practical determination of which asymptotic behavior challenging. In contrast, conditional on $\mathcal{D}_j^{(1)}$, the proposed W_j in (2) can be regarded as a univariate projection of $\mathbf{g}_{2j}(\boldsymbol{\theta}) - \widehat{\mathbf{g}}(\boldsymbol{\theta})$ and enjoys the symmetric property regardless of the gradient dimension d. Benefiting from the joint use of the proposed

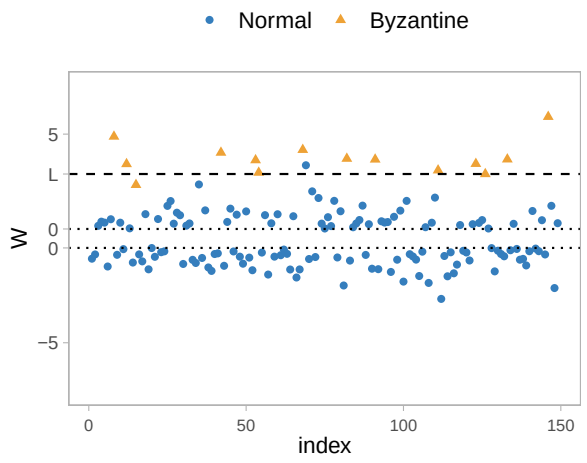

Figure 1. Scatter plot of W_j. The W_j's of normal machines are symmetric with 0.

W_j and the threshold L, the empirical distribution of the negative statistics can be used to approximate that of the positive ones in place of asymptotic distribution for calibration, giving the proposed ByMI method an edge over existing methods in terms of the accuracy of FDR control. A similar idea has been discussed in (Chen et al., 2023) where similar symmetric statistics based on the sample-splitting strategy were proposed for change point detection.

The idea of ByMI is illustrated in Figure 1, where we display the scatter plot of score statistics W_j when there are $m+1 = 150$ machines containing one master and 14 Byzantines. The detailed settings are shown in Section 4.2. We can see that W_j's of normal machines are symmetric with 0, but Byzantine machines own large W_j's. Only a few false discoveries are beyond the threshold L in (3).

Benefiting from only passing two gradient vectors $\mathbf{g}_{1j}(\boldsymbol{\theta})$ and $\mathbf{g}_{2j}(\boldsymbol{\theta})$ from each single worker machine to the master one in **Step 1**, it is worth emphasizing that the ByMI method is communication-efficient (Jordan et al., 2019) and can freely conjugate with the learning procedure. The proposed procedure is also computationally efficient. The running time of **Step 2** is similar to the commonly used robust learning algorithms. Meanwhile, the computation of W_j and the implementation of the detection process in **Steps 3-4** incur a small run-time overhead due to only computing a d-dimension vector and searching a m size set, respectively. The process of the ByMI method is displayed in Figure 2.

2.3. Some Examples

ByMI detects the Byzantine machines in a general gradient form. It covers a wide range of applications. We list several introductory examples.

Mean estimation. Let $\{\mathbf{s}_i\}_{i=1}^N$ be i.i.d. observations from the distribution of $\mathbf{s} \in \mathbb{R}^p$ with $\mathbb{E}[\mathbf{s}] = \boldsymbol{\theta}$. Consider the loss function $\ell(\mathbf{s}, \boldsymbol{\theta}) = \|\mathbf{s} - \boldsymbol{\theta}\|_2^2/2$ and its gradient $\mathbf{g}(\mathbf{s}, \boldsymbol{\theta}) = \mathbf{s} - \boldsymbol{\theta}$. Here, there is no need to give $\boldsymbol{\theta}$ for Byzantine machine identification. On each machine $\mathcal{M}_j$, the empirical gradient function $\mathbf{g}_j(\boldsymbol{\theta})$ can be replaced by $\bar{\mathbf{s}}_j = \sum_{i \in \mathcal{M}_j} \mathbf{s}_i$. In our detection procedure, $\{\bar{\mathbf{s}}_{1j}\}_{j=0}^m$ is aggregated to be a robust mean estimation $\hat{\boldsymbol{\theta}}$ and then we construct ranking scores with $\hat{\boldsymbol{\theta}}$ in (2) to detect Byzantine machines.

Linear regression model. Let $\mathbf{s} = (y, \mathbf{x})$ and $y = \mathbf{x}^\top \boldsymbol{\theta} + \varepsilon$, where $\mathbf{x} \in \mathbb{R}^p$ is the covariates and ε is the noise. We choose the square loss function $\ell(\mathbf{s}, \boldsymbol{\theta}) = \|y - \mathbf{x}^\top \boldsymbol{\theta}\|_2^2/2$. And the gradient becomes $\mathbf{g}(\mathbf{s}, \boldsymbol{\theta}) = \mathbf{x}(y - \mathbf{x}^\top \boldsymbol{\theta}) \in \mathbb{R}^p$. When set $\boldsymbol{\theta} = \mathbf{0}$, it reduces to a more concise expression $\mathbf{g}(\mathbf{x}, \mathbf{0}) = \mathbf{x}y$ which measures the covariance of $\mathbf{x}$ and y.

Multicategory classification with cross-entropy loss. We consider a K-category classification task here. The sample $\mathbf{s}$ contains a pair of covariates $\mathbf{x} \in \mathbb{R}^p$ and a response variable $y \in [K]$ and it can be transformed to the one-hot representation $\mathbf{y} = (y_1, \ldots, y_K)$ where the y-th entry is one and other entries are zero. We can employ the cross-entropy loss $\ell(\mathbf{s}, \boldsymbol{\theta}) = -\sum_{k=1}^K y_k \log\{\exp(\boldsymbol{\theta}_k^\top \mathbf{x}_k)/\sum_{j=1}^K \exp(\boldsymbol{\theta}_j^\top \mathbf{x}_j)\}$ and consider the corresponding gradient function $\mathbf{g}(\mathbf{s}, \boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \ell(\mathbf{s}, \boldsymbol{\theta})$ with $\boldsymbol{\theta} = \left(\boldsymbol{\theta}_1^\top, \ldots, \boldsymbol{\theta}_K^\top\right)^\top \in \mathbb{R}^d$ and $d = pK$.

2.4. Choice of $\boldsymbol{\Omega}$

There are different choices of $\boldsymbol{\Omega}$ when we construct the ranking scores W_j.

Scale Matrix. The performance of our procedure is not sensitive to its choice when $\boldsymbol{\Omega}$ only serves to standardize the components of $\mathbf{g}_j(\boldsymbol{\theta})$ so that they are aggregated fairly. We suggest adopting a diagonal estimator $\boldsymbol{\Omega} = \mathrm{diag}\{\hat{\sigma}_1^{-2}, \ldots, \hat{\sigma}_d^{-2}\}$, where $\hat{\sigma}_k^2$ is variance estimator of the k-th component of $\mathbf{g}_j(\boldsymbol{\theta})$ and can be obtained on $\bigcup_{j=1}^m \mathcal{D}_j^{(1)}$ by some robust estimation. More details are discussed in Supplementary Material.

Projection Matrix. Instead of taking all d dimensions for W_j, we can also adopt the projection matrix $\boldsymbol{\Omega} = \mathbf{v}_1 \mathbf{v}_1^\top$ where $\mathbf{v}_1$ is the first eigenvector of the covariance matrix of $\{\mathbf{g}_{1j}(\boldsymbol{\theta})\}$. Such a choice is inspired by the dimension-agnostic robust mean estimation approach (Diakonikolas et al., 2017) which proved that the outliers that significantly affect the mean estimation should lie in the direction $\mathbf{v}_1$.

3. Statistical Performance Guarantees

This section provides statistical guarantees of the ByMI procedure. We begin with a general finite sample result about the FDR control, in the sense that it requires no model

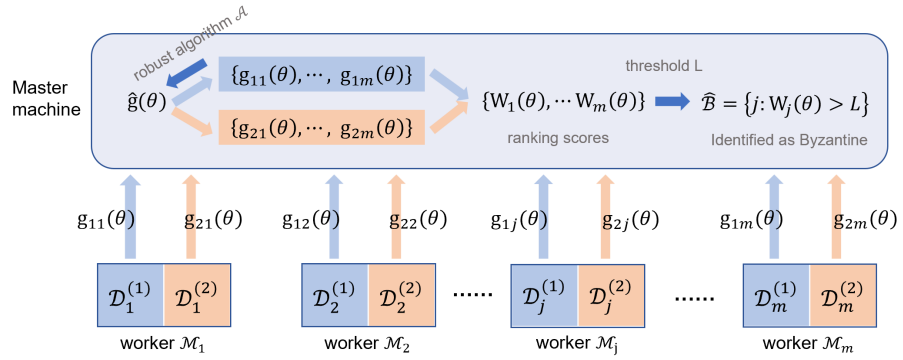

Figure 2. The implementation of ByMI procedure.

or distribution assumptions. Denote $\mathcal{D}_k = \bigcup_{j=1}^m \mathcal{D}_j^{(k)}$. For simplicity of notations, let's set $\mathbf{z}_{ij} := \mathbf{g}(\mathbf{s}_i, \theta)$ for $i \in \mathcal{M}_j$ and $\overline{\mathbf{z}}_{kj} := \mathbf{g}_{kj}(\boldsymbol{\theta})$ for $k = 1, 2$ and $j \in [m]$.

Lemma 3.1. *For $j \in [m]$, denote $\Delta_j = |\mathbb{P}(W_j > 0 \mid \mathcal{D}_1, |W_j|) - 1/2|$. For any $\alpha \in (0,1)$, the ByMI procedure satisfies*

$$\mathrm{FDR}(\hat{\mathcal{B}}) \triangleq \mathbb{E}[\mathrm{FDP}(\hat{\mathcal{B}})] \leq \min_{\epsilon \geq 0}\Big\{\alpha(1+5\epsilon) + \mathbb{E}\big[\mathbb{P}(\max_{j\in\mathcal{G}} \Delta_j > \epsilon \mid \mathcal{D}_1)\big]\Big\}.$$

Here Δ_j measures the asymmetricity of ranking scores W_j for those normal machines $\mathcal{G}$, which implies that a tight control of Δ_j's leads to effective FDR control of the proposed ByMI procedure. Consider the ideal case where W_j's in $\mathcal{G}$ are all symmetrically distributed around zero. Then, we have $\Delta_j = 0$ for all $j \in \mathcal{G}$, and therefore the FDR is exactly controlled by setting $\epsilon = 0$, i.e, $\mathrm{FDR}(\hat{\mathcal{B}}) \leq \alpha$.

Next, we turn to a stringent finite-sample result of FDR control under some mild conditions on the samples.

Assumption 3.2 (Moments)**.** Samples $\{\mathbf{z}_{ij}\}_{i\in[n],j\in[m]}$ are with bounded q-th centered moments ($q > 2$) and satisfies L_q-L_2 norm equivalence condition with parameter γ_q, i.e.

$$\max_{\mathbf{v}\in\mathcal{S}^{d-1}} \frac{(\mathbb{E}|\mathbf{v}^\top(\mathbf{z}_{ij} - \boldsymbol{\mu}^*)|^q)^{\frac{1}{q}}}{(\mathbb{E}|\mathbf{v}^\top(\mathbf{z}_{ij} - \boldsymbol{\mu}^*)|^2)^{\frac{1}{2}}} \leq \gamma_q.$$

Assumption 3.3 (Robust estimation)**.** Assume the robust estimation satisfies that $\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}^*\| = O(\delta_\mu)$, where δ_μ is some positive sequence that converges to zero.

The moment condition in Assumption 3.2 is commonly used in the literature (Mendelson & Zhivotovskiy, 2020). Assumption 3.3 sets theoretical minimal requirements for the accuracy of the robust estimator $\hat{\boldsymbol{\mu}}$ obtained by $\mathcal{D}_1$. As discussed in Lemma 3.1, we cannot expect that the (asymptotic) symmetric property of W_j can be achieved for one arbitrary robust estimation. This can be satisfied by a large category of robust estimations. For example, for the median-type methods (Chen et al., 2017; Yin et al., 2018), $\delta_\mu = O(\sqrt{\frac{d}{nm}} + \sqrt{\frac{r}{nm}} + \varrho\sqrt{\frac{d}{n}})$ with probability at least $1 - e^{-r}$, while for those algorithms whose bias is free of the dimension (Cheng et al., 2020; Zhu et al., 2022), $\delta_\mu = O(\sqrt{\frac{d}{nm}} + \sqrt{\frac{r}{nm}} + \sqrt{\frac{\varrho}{n}})$.

Theorem 3.4. *(FDR control) Suppose Assumptions 3.2- 3.3 hold. Denote $\kappa = \min(1, q-2)$ and $\omega_n = n^{-\frac{(1-2\eta^2)\kappa}{2}}$ with some $\eta \in (0, \frac{1}{\sqrt{2}})$. With probability at least $1 - O(mn^{-\frac{\eta^2\kappa}{2}})$,*

$$\mathrm{FDR}(\hat{\mathcal{B}}) \leq \alpha + O\Big(\sqrt{\omega_n + n^{\frac{1}{2}+\eta^2\kappa}\delta_\mu}\Big). \tag{4}$$

With the assumption of signals, we further obtain a finite-sample result of FDP control.

Assumption 3.5 (Signals)**.** Denote $\mathcal{B}_\mu$ the set of identifiable Byzantine machines and $\boldsymbol{\mu}_j^*$ the mean of the j-th machine. There is a constant $C > 0$ such that, for $j \in \mathcal{B}_\mu$, $\|\boldsymbol{\mu}_j^* - \boldsymbol{\mu}^*\| \geq C(\sqrt{\frac{\log n}{n}} + \delta_\mu + d^{\frac{1}{2}}n^{-\frac{1}{2}+\frac{\kappa_2}{q}})$ where $0 < \kappa_2 < \frac{q}{2}$. Assume that $\psi_m = |\mathcal{B}_\mu|$ is sufficiently large.

Assumption 3.5 refers to the minimum signal magnitudes to distinguish a Byzantine machine from the well-behaved one. Note that the last term is essentially from the upper bound, $\sup_{j\in[m]}\|\overline{\mathbf{z}}_{1j} - \boldsymbol{\mu}_j^*\| = O(d^{\frac{1}{2}}n^{-\frac{1}{2}+\frac{\kappa_2}{q}})$ with probability at least $1 - mn^{-\kappa_2}$ for any $0 < \kappa_2 < \frac{q}{2}$ owing to Assumption 3.2. It can be further improved by assuming more light tails like the sub-Gaussian condition.

Theorem 3.6. *(FDP control) Suppose Assumptions 3.2, 3.3 and 3.5 hold. Denote $\kappa = \min(1, q-2)$ and $s_{nm} = n^{-\frac{(1-\eta^2)\kappa}{2}}(\log n)^{\frac{1}{2}} + mn^{-\frac{\eta^2\kappa}{2}} + (\alpha\psi_m)^{-\delta/3}$ with some $\eta \in (0,1)$. With probability at least $1 - O(mn^{-\frac{\eta^2\kappa}{2}} + \psi_m^{-(1-\delta)})$,*

$$\mathrm{FDP}(\hat{\mathcal{B}}) \leq \alpha\Big[1 + O\Big(s_{nm} + n^{\frac{1}{2}}\delta_\mu\Big)\Big]. \tag{5}$$

Theorem 3.4 and Theorem 3.6 imply that the effect of dimension d in FDR or FDP control is due to the accuracy of robust mean estimation in Assumption 3.3. Indeed, the ByMI is dimension-agnostic if building W_j with $\boldsymbol{\mu}^*$ instead of its robust estimation $\hat{\boldsymbol{\mu}}$. Besides FDR control, the next result shows that the ByMI method is also capable of maintaining the "sure-detection property". This property ensures that all identifiable Byzantine machines $\mathcal{B}_\mu$ can be detected.

Corollary 3.7. *Under the conditions in Theorem 3.6, with probability at least* $1 - O(mn^{-\frac{\eta^2\kappa}{2}} + \psi_m^{-(1-\delta)})$*, we have* $\mathcal{B}_\mu \subseteq \hat{\mathcal{B}}$.

As a byproduct, in Corollary 3.7, we have the sure-detection property of the ByMI procedure, which says that all the identifiable Byzantine machines $\mathcal{B}_\mu$ can be detected by the ByMI procedure with probability tending to one. As a byproduct, Corollary 3.7 establishes the sure-detection property of the ByMI procedure. It states that all identifiable Byzantine machines in $\mathcal{B}_\mu$ can be detected by the ByMI procedure with high probability.

4. Experiments and Evaluation

We illustrate the breadth of applicability of the ByMI procedure through experiments on synthetic data and real-data applications. We implement the proposed ByMI method in conjunction with two algorithms $\mathcal{A}$ for robust mean estimators. One is the geometric median (Minsker, 2015) of all empirical gradients, and the other one is the filtering estimator proposed in (Diakonikolas et al., 2017; Lai et al., 2016; Zhu et al., 2022). These two adopt the scale estimator for $\boldsymbol{\Omega}$ and are denoted as ByMI-GEOM and ByMI-Filter, respectively. Also, we choose $\boldsymbol{\Omega}$ as the projection matrix described in 2.4 in conjunction with the filtering mean estimator, named as ByMI-Filter+. Other algorithms $\mathcal{A}$ are studied in the Supplementary. The target FDR level is fixed as $\alpha = 0.1$.

Benchmarks. We compare the ByMI procedure with three benchmarks. The first one is to implement the well-known outlier detection algorithms in high-dimension, RMDP (Ro et al., 2015) with the empirical gradients $\mathbf{g}_j(\boldsymbol{\theta})$ of each machine. This method builds a minimum diagonal product estimator based on modified Mahalanobis distance and identifies outliers with its asymptotic distribution. To make a fair comparison, we embed the RMDP with the classical BH procedure (Benjamini & Hochberg, 1995) to achieve the FDR control, referred to as RMDP-BH. The other three competitors are Krum (Blanchard et al., 2017), FABA (Xia et al., 2019), and Zeno (Xie et al., 2019) from machine learning literature. Those methods employ some distance-based scores and roughly detect a given number ϱm of those machines with large scores as Byzantine. Specifically, the Krum and FABA adopt some Euclidean distances and Zeno proposes stochastic descendant scores. More details of the benchmarks can be found in Supplementary Material.

Performance Measures. The empirical FDR and TPR are evaluated using the average of FDP and TPP from 500 replications, respectively. The proportion that all identifiable Byzantine machines are detected, that is, $P_a = \Pr(\mathcal{B} \subseteq \widehat{\mathcal{B}}(L))$ is computed to evaluate the sure-detection property.

4.1. Results on Synthetic Data

- Scenario A (Mean Estimation): The data on normal machines are i.i.d from $\mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} = (0.2^{|i-j|})_{p\times p}$; meanwhile the data on Byzantine machines are i.i.d from $\mathcal{N}_p(b\log(p)\cdot \boldsymbol{v_p}, 0.5\mathbf{I}_p)$, where b is the shift size and $\boldsymbol{v_p} \in \mathbb{R}^p$ is a normalized vector with p independent random variables from $\mathcal{U}(0,1)$.

- Scenario B (Regression Model): We consider the linear model $y = \mathbf{x}^\top\boldsymbol{\theta} + \varepsilon$ where $\mathbf{x} \sim \mathcal{N}_p(\mathbf{0}, \mathbf{I}_p)$, $\varepsilon \sim \mathcal{N}(0,1)$ and $\boldsymbol{\theta} = (\mathbf{1}_s, 0, \cdots, 0)^\top$ with $s = \lfloor 0.1p \rfloor$. Two kinds of Byzantine machines are investigated: (a) the model is corrupted where the parameter on Byzantine machines is $\boldsymbol{\theta_c} = (\mathbf{1}_{s_c}, 0, \cdots, 0)^\top$ with $s_c = \lfloor 0.07p \rfloor$; (b) the data is contaminated where $\mathbf{x}_i$'s on Byzantine machines are replaced by $\tilde{\mathbf{x}}_i = 0.8\mathbf{x}_i + 3\boldsymbol{v}_p$ where $\boldsymbol{v}_p \in \mathbb{R}^p$ is same as Scenario A, and Y_i's are added with a constant bias $c = 1$. For simplicity, we compute the gradients at $\boldsymbol{\theta} = \mathbf{0}$.

We fix the number of worker machines as $m = 1,000$ and the local sample size as $n = 200$ so that the entire sample size is $N = 200 \times 1001$ including one master. We randomly choose $\lfloor \varrho m \rfloor$ worker machines as Byzantine ones.

Results. Figure 3 reports the FDR, TPR and P_a curves against the shift size b with the contamination ratio $\varrho = 0.05$ under Scenario A. We see that the FDR levels of ByMI-based (ByMI-Filter, ByMI-GEOM, ByMI-Filter+) are close to the nominal level. All methods also achieve satisfactory TPR and P_a under all the scenarios. In contrast, RMDP-BH yields slightly inflated FDRs under $p = 50$ but a little conservative one under $p = 100$. Also, RMDP-BH leads to lower TPR and P_a compared to the proposed ByMI. This can be understood that the p-values of RMDP-BH are from one asymptotic distribution, which may be sensitive to the dimension or other model settings in finite sample cases. The Krum and FABA result in overly inflated FDR levels, and accordingly, they do not perform well in terms of P_a. It implies that both distance-based methods detect a fixed number as Byzantine machines and some true Byzantines would be missed.

Figure 4 presents the boxplots of empirical FDP and TPP under Scenario B. The FDPs of ByMI-Filter and ByMI-Filter+ vary in an acceptable range of the target level while

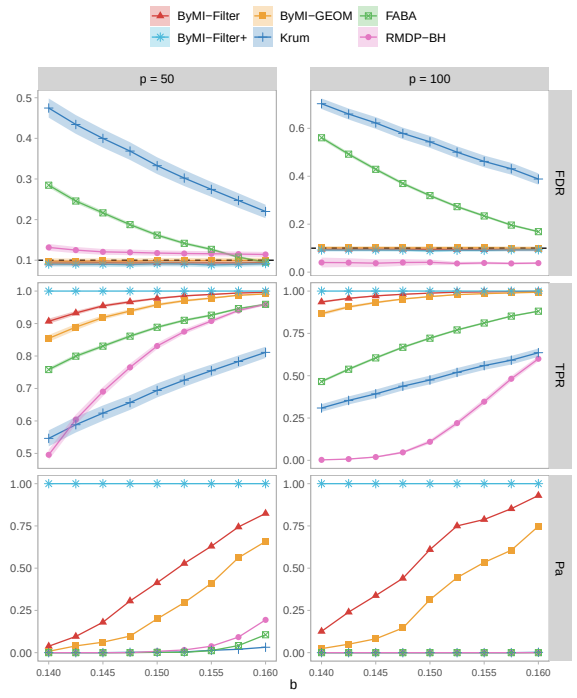

Figure 3. FDR, TPR and P_a over shift size b when $\varrho = 0.05$ and $p = 50, 100$ under Scenario A. The black dashed lines indicate the FDR level $\alpha = 0.1$.

the FDPs of ByMI-GEOM slightly exceed the nominal level and the TPPs of ByMI-Filter and ByMI-Filter+ are higher than ByMI-GEOM. This is from the fact that the robust estimation based on filtering has dimension-agnostic superiority against the bias made by the adversarial attacks compared to the geometric median, which may be conducive to improving the signal-to-noise. Meanwhile, other benchmarks deliver overly inflated FDPs under all the settings. This further demonstrates the effectiveness of the proposed ByMI method: it is data-driven and p-value free which allows FDR control and achieves reliable TPR.

4.2. Results on Real Data

Datasets. The MNIST (LeCun et al., 1998), Fashion MNIST (Xiao et al., 2017) and CIFAR10 (Krizhevsky, 2009) datasets are used to verify the performance of our ByMI.

- MNIST includes $60,000$ training images and $10,000$ testing images with size 28×28 pixels and their corresponding labels from 0 to 9.
- Fashion MNIST (F-MNIST) contains $60,000$ training images and 10,000 testing images with size 28×28 pixels, each belonging to 10 fashion items.
- CIFAR10 consists of $50,000$ training images and $10,000$ testing images, each of size 32×32 pixels, belonging to 10 different classes.

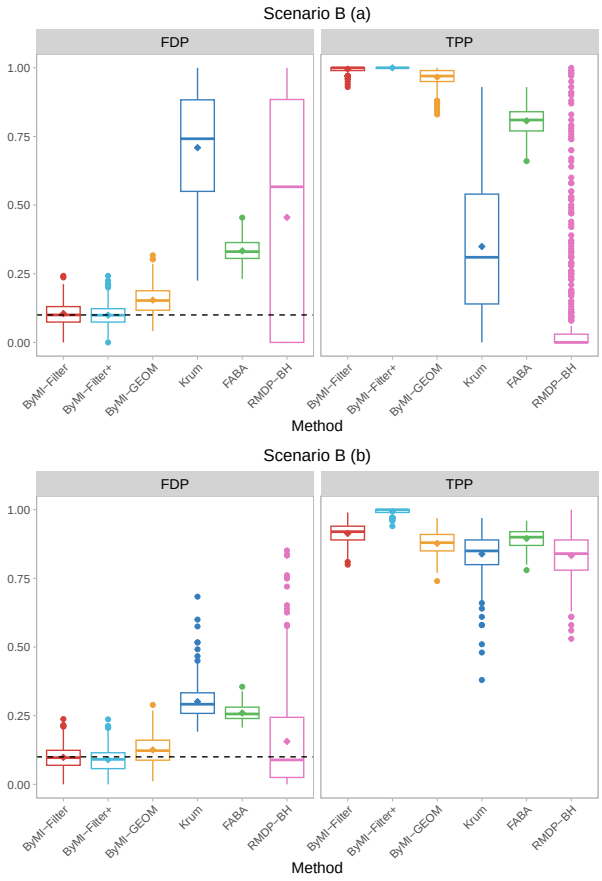

Figure 4. Boxplots of FDP and TPP when the dimension $p = 100$ and the contamination ratio $\varrho = 0.1$ under Scenario B. The black dashed lines indicate the FDR level $\alpha = 0.1$.

As the main focus is the detection task, we adopt the pre-trained Resnet18 model in Pytorch and take the output of the last layer with the dimension 512. We further use the PCA on the features to get decorrelated covariates $\{\mathbf{x}_i\}$ with dimension $p = 20$. A multinomial logistic regression model is adopted where the total parameter dimension is $d = 200$. For MNIST and F-MNIST, all of the samples in the training set are randomly divided into $m + 1 = 150$ machines (including one master machine) with an equal sample size $n = 400$. For CIFAR10, we fix $m + 1 = 125$ and $n = 400$.

Attacks. For Byzantine machines, we conduct both the out-of-distribution (OOD) attack (Fort et al., 2021) and the IPM attack (Xie et al., 2020). Specifically, for the OOD attack we replace the covariates $\mathbf{x}_i$'s on Byzantine machines by $\tilde{\mathbf{x}}_i = 0.7\mathbf{x}_i + \boldsymbol{\varepsilon}_p$ where $\boldsymbol{\varepsilon}_p$ is from $\mathcal{N}_p(\boldsymbol{\nu}_p, \sigma^2 \boldsymbol{I}_p)$ with $\boldsymbol{\nu}_p \in \mathbb{R}^p$ randomly sampled from the standard multivariate normal distribution and $\sigma = 0.2$. For the IPM attack, the Byzantine gradients are assigned as $-a\bar{\mathbf{g}}$, where $\bar{\mathbf{g}} = \frac{1}{|\mathcal{G}|}\sum_{j\in\mathcal{G}} \mathbf{g}_j(\boldsymbol{\theta})$

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

		MNIST			F-MNIST			CIFAR10		
Attack	Method	FDR	TPR	P_a	FDR	TPR	P_a	FDR	TPR	P_a
OOD Attack	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 Attack	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

and $a > 0$.

We train the multinomial logistic regression model using distributed gradient descent. At the beginning, we obtain an initial parameter $\boldsymbol{\theta}_0$ which is trained by the master machine only, and then we deliver $\boldsymbol{\theta}_0$ to each worker machine to compute the local gradients. In each iteration, those local gradients on worker machines are sent to the master for parameter aggregation and updating. Our goal is to make use of the local gradients to detect Byzantine machines.

Results of Byzantine machine detection. Table 1 reports the experiment results of the OOD attack and the IPM attack. For simplicity, we compare the performance of different detected methods in the first iteration under the contamination ratio $\varrho = 0.1$. Three ByMI-based methods perform reasonably well. The FDRs are controlled under or close to the nominal level $\alpha = 0.1$ across all the settings. Meanwhile, the ByMI-based methods yield quite high TPRs and P_a, which clearly demonstrates the efficiency of our proposed method. In addition, RMDP-BH also has satisfactory TPRs, but it yields an overly inflated FDR level since the p-values for the BH procedure are approximated by an asymptotic distribution. In contrast, those distance-based methods (Krum, FABA and Zeno) select a given number, i.e, $\lfloor 1.2\varrho m \rfloor$ as Byzantine machines, but fewer true Byzantine machines can be correctly detected concerning the large FDR as well as the small P_a. It implies that these three are hard to guarantee the sure-detection property with simple distance ranking.

Application to robust distributed learning. Besides the measurements of FDR, TPR and P_a, we further study the performance of the ByMI method in the distributed learning tasks. We apply ByMI to detect Byzantine machines and use the simple mean aggregation of the left local gradients to train the model. Figure 5 shows the test accuracy among the learning processes. Filtering refers to aggregating local gradients by the filtering algorithm in each round. The ByMI-based methods deliver higher classification accuracy compared to other benchmarks. In contrast, Krum, FABA and Zeno perform not well and are susceptible to Byzantine failures. It's worth mentioning that ByMI-Filter and ByMI-Filter+ attain higher accuracy and achieve the FDR control in the identification task as well compared to the original Filtering method. In summary, ByMI not only

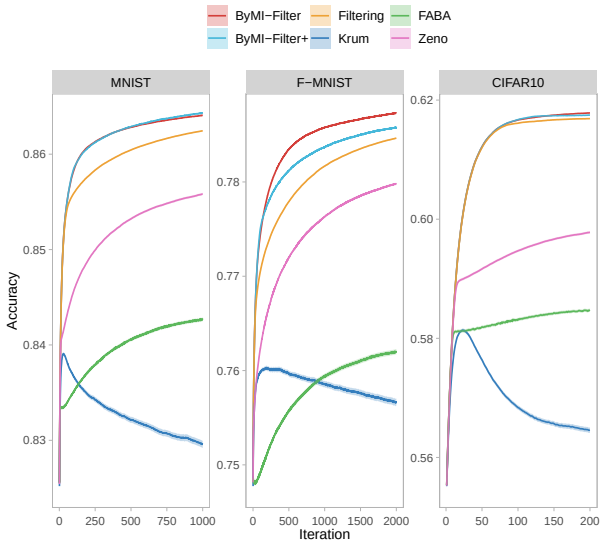

Figure 5. Test accuracy of different methods under the IPM attack with the contamination level $\varrho = 0.3$ when $a = 2$.

works well for identifying Byzantine machines with FDR control but also improves the robustness and interpretability in the distributed learning tasks with Byzantine corruptions.

5. Concluding Remarks

Identification of Byzantine machines is very important but obtains less attention to conduct robust learning algorithms in a distributed system. This paper proposes a data-driven detection procedure, ByMI, to address this issue via FDR control under a unified framework. The ByMI method is easy to implement and communication-efficient because only the local gradients are transmitted. It is shown that the proposed method can control FDR with reliable robust estimators while retaining all the identifiable Byzantine machines under mild conditions. Thus, it could serve as a useful tool for further robust inference or system diagnostics.

We conclude this paper with two remarks. First, we achieve the FDR control by sample-splitting strategy. In practice, one may prefer to use the whole data to find the Byzantine machines without accuracy sacrifice. It is of interest to further improve ByMI or investigate what ByMI could contribute. Secondly, we mainly consider the behavior of the Byzantine machines by the Huber contamination model. How to adapt ByMI to other aggressive behaviors, such as attacks on the transmission paths, deserves further study.

Acknowledgements

The authors would like to express their sincere appreciation to the anonymous reviewers for their valuable comments and constructive feedback. This research was supported by the National Key R&D Program of China (Grant Nos. 2022YFA1003703, 2022YFA1003800), the National Natural Science Foundation of China (Grant Nos. 11925106, 12101398, 12231011, 11931001, 12226007, 12326325) and Shanghai Sailing Program.

Impact Statement

This paper presents work whose goal is to improve the robustness and trustworthiness of distributed and federated machine learning systems by introducing advanced statistical testing tools. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Bai, Z. and Saranadasa, H. Effect of high dimension: by an example of a two sample problem. Statistica Sinica, pp. 311–329, 1996.

Barber, R. F., Candès, E. J., and Samworth, R. J. Robust inference with knockoffs. The Annals of Statistics, 48 (3):1409 – 1431, 2020.

Benjamini, Y. and Hochberg, Y. 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.

Benjamini, Y. and Yekutieli, D. The control of the false discovery rate in multiple testing under dependency. The Annals of Statistics, 29(4):1165–1188, 2001.

Blanchard, P., El Mhamdi, E. M., Guerraoui, R., and Stainer, J. Machine learning with adversaries: Byzantine tolerant gradient descent. In Advances in Neural Information Processing Systems, volume 30, pp. 1–11. Curran Associates, Inc., 2017.

Bottou, L., Curtis, F. E., and Nocedal, J. Optimization methods for large-scale machine learning. SIAM review, 60(2):223–311, 2018.

Chen, H., Ren, H., Yao, F., and Zou, C. Data-driven selection of the number of change-points via error rate control. Journal of the American Statistical Association, 118(542):1415–1428, 2023.

Chen, S. X. and Qin, Y.-L. A two-sample test for high-dimensional data with applications to gene-set testing. The Annals of Statistics, 38(2):808–835, 2010.

Chen, Y., Su, L., and Xu, J. Distributed statistical machine learning in adversarial settings: Byzantine gradient descent. Proceedings of the ACM on Measurement and Analysis of Computing Systems, 1(2):1–25, 2017.

Cheng, Y., Diakonikolas, I., and Ge, R. High-dimensional robust mean estimation in nearly-linear time. In Proceedings of the 30th annual ACM-SIAM symposium on discrete algorithms, pp. 2755–2771. SIAM, 2019.

Cheng, Y., Diakonikolas, I., Ge, R., and Soltanolkotabi, M. High-dimensional Robust Mean Estimation via Gradient Descent. In Proceedings of the 37th International Conference on Machine Learning, pp. 1768–1778. PMLR, 2020.

Diakonikolas, I., Kamath, G., Kane, D. M., Li, J., Moitra, A., and Stewart, A. Being robust (in high dimensions) can be practical. In Precup, D. and Teh, Y. W. (eds.), Proceedings of the 34th International Conference on Machine Learning, volume 70 of Proceedings of Machine Learning Research, pp. 999–1008. PMLR, 2017.

Du, L., Guo, X., Sun, W., and Zou, C. False discovery rate control under general dependence by symmetrized

data aggregation. Journal of the American Statistical Association, pp. 1–15, 2021.

Efron, B. Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. Journal of the American Statistical Association, 99(465):96–104, 2004.

Filzmoser, P., Maronna, R., and Werner, M. Outlier identification in high dimensions. Computational Statistics & Data Analysis, 52(3):1694–1711, 2008.

Fort, S., Ren, J., and Lakshminarayanan, B. Exploring the limits of out-of-distribution detection. Advances in Neural Information Processing Systems, 34:7068–7081, 2021.

Huber, P. J. Robust Estimation of a Location Parameter. The Annals of Mathematical Statistics, 35(1):73–101, 1964.

Jordan, M. I., Lee, J. D., and Yang, Y. Communication-Efficient Distributed Statistical Inference. Journal of the American Statistical Association, 114(526):668–681, 2019.

Kairouz, P., McMahan, H. B., Avent, B., and Bellet. et. al, A. Advances and Open Problems in Federated Learning. Foundations and Trends® in Machine Learning, 14(1–2): 1–210, 2021.

Kamp, M., Boley, M., Missura, O., and Gärtner, T. Effective parallelisation for machine learning. Advances in Neural Information Processing Systems, 30, 2017.

Krizhevsky, A. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Lai, K. A., Rao, A. B., and Vempala, S. Agnostic estimation of mean and covariance. In 2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS), pp. 665–674. IEEE, 2016.

Lamport, L., Shostak, R., and Pease, M. The byzantine generals problem. ACM Transactions on Programming Languages and Systems, 4(3):382–401, 1982.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11):2278–2324, 1998.

Liu, W. and Shao, Q.-M. Phase transition and regularized bootstrap in large-scale t-tests with false discovery rate control. The Annals of Statistics, 42(5):2003–2025, 2014.

Mendelson, S. and Zhivotovskiy, N. Robust covariance estimation under $L_{4}-L_{2}$ norm equivalence. The Annals of Statistics, 48(3):1648 – 1664, 2020.

Minsker, S. Geometric median and robust estimation in banach spaces. Bernoulli, 21(4):2308–2335, 2015.

Petrov, V. On probabilities of moderate deviations. Journal of Mathematical Sciences, 109(6):2189–2191, 2002.

Ren, H., Chen, N., and Zou, C. Projection-based outlier detection in functional data. Biometrika, 104(2):411–423, 2017.

Ro, K., Zou, C., Wang, Z., and Yin, G. Outlier detection for high-dimensional data. Biometrika, 102(3):589–599, 2015.

Shi, J., Wan, W., Hu, S., Lu, J., and Zhang, L. Y. Challenges and approaches for mitigating byzantine attacks in federated learning. In 2022 IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), pp. 139–146. IEEE, 2022.

Su, L. and Xu, J. Securing distributed gradient descent in high dimensional statistical learning. Proceedings of the ACM on Measurement and Analysis of Computing Systems, 3(1):1–41, 2019.

Tu, J., Liu, W., Mao, X., and Chen, X. Variance reduced median-of-means estimator for byzantine-robust distributed inference. Journal of Machine Learning Research, 22(1):3780–3846, 2021.

Xia, Q., Tao, Z., Hao, Z., and Li, Q. FABA: An Algorithm for Fast Aggregation against Byzantine Attacks in Distributed Neural Networks. In Proceedings of the 28th International Joint Conference on Artificial Intelligence, pp. 4824–4830, Macao, China, 2019. International Joint Conferences on Artificial Intelligence Organization.

Xiao, H., Rasul, K., and Vollgraf, R. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. arXiv preprint arXiv:1708.07747, 2017.

Xie, C., Koyejo, S., and Gupta, I. Zeno: Distributed stochastic gradient descent with suspicion-based fault-tolerance. In Proceedings of the 36th International Conference on Machine Learning, volume 97 of Proceedings of Machine Learning Research, pp. 6893–6901. PMLR, 2019.

Xie, C., Koyejo, O., and Gupta, I. Fall of empires: Breaking byzantine-tolerant sgd by inner product manipulation. In Uncertainty in Artificial Intelligence, pp. 261–270. PMLR, 2020.

Yin, D., Chen, Y., Kannan, R., and Bartlett, P. Byzantine-robust distributed learning: Towards optimal statistical rates. In Proceedings of the 35th International Conference on Machine Learning, volume 80 of

Proceedings of Machine Learning Research, pp. 5650–5659. PMLR, 2018.

Yin, D., Chen, Y., Kannan, R., and Bartlett, P. Defending Against Saddle Point Attack in Byzantine-Robust Distributed Learning. In Proceedings of the 36th International Conference on Machine Learning, pp. 7074–7084. PMLR, 2019.

Zhu, B., Jiao, J., and Steinhardt, J. Robust estimation via generalized quasi-gradients. Information and Inference: A Journal of the IMA, 11(2):581–636, 2022.

Zhu, B., Wang, L., Pang, Q., Wang, S., Jiao, J., Song, D., and Jordan, M. I. Byzantine-robust federated learning with optimal statistical rates. In Proceedings of The 26th International Conference on Artificial Intelligence and Statistics, volume 206 of Proceedings of Machine Learning Research, pp. 3151–3178. PMLR, 2023.

Zimek, A., Schubert, E., and Kriegel, H.-P. A survey on unsupervised outlier detection in high-dimensional numerical data. Statistical Analysis and Data Mining: The ASA Data Science Journal, 5(5):363–387, 2012.

Zou, C., Ren, H., Guo, X., and Li, R. A new procedure for controlling false discovery rate in large-scale t-tests. arXiv preprint arXiv:2002.12548, 2020.

A. More Details on the Related Works

We offer an expanded introduction to the literature associated with Byzantine machine detection, serving as a complement to Section 1.

Byzantine detection approach in machine learning literature. Given the number ϱm of Byzantine machines, for the j-th machine, Krum (Blanchard et al., 2017) proposes using the average Euclidean distance of the $(1-\varrho)m$ neighbors of $\mathbf{g}_j$ to detect the Byzantine machines. The larger the distance the higher the possibility that the j-th machine is a Byzantine one. FABA (Xia et al., 2019) simply adopts the Euclidean distance of $\mathbf{g}_j$ to the mean of $\{\mathbf{g}_j\}$, deletes the $\mathbf{g}_j$ with the largest distance, and then updates the mean vector. The above process will stop until ϱm gradient vectors are deleted and the final mean vector will be returned. Zeno (Xie et al., 2019) suggests the *stochastic descendant score* $S_j = \sum_{i\in\mathcal{M}_0} \ell(\mathbf{s}_i, \boldsymbol{\theta}) - \ell(\mathbf{s}_i, \boldsymbol{\theta} - \gamma\mathbf{g}_j) - \rho\|\mathbf{g}_j\|^2$, the descendant value of the loss function with the data in the master machine on the current parameter $\boldsymbol{\theta}$, learning rate γ and a regularized parameter ρ. All these algorithms require a pre-specified proportion of Byzantine machines, which is set to 1.2 times the true proportion ϱ during experiments.

The RMDP-BH procedure. The RMDP approach (Ro et al., 2015) measures the departure of machine $\mathcal{M}_j$ from the center by the modified Mahalanobis distance $(\mathbf{g}_j - \boldsymbol{\mu}^*)^\top D^{-1}(\mathbf{g}_j - \boldsymbol{\mu}^*)$ where D is a diagnoal matrix with the marginal variance of $\mathbf{g}_j$ in the normal machines on the diagnoal. As $n \to \infty$ and $d \to \infty$, the modified Mahalanobis distance is asymptotical normal in the normal machines, i.e. $\frac{(\mathbf{g}_j-\boldsymbol{\mu}^*)^\top D^{-1}(\mathbf{g}_j-\boldsymbol{\mu}^*)-d}{[2\,\mathrm{tr}(\mathbf{R})]^{1/2}} \xrightarrow{\mathrm{d}} \mathcal{N}(0,1)$. The RMDP-BH approach replaces the true center $\boldsymbol{\mu}^*$ by the minimum diagonal product estimator and the p-value p_j of each machine $\mathcal{M}_j$ can be computed based on the asymptotical normality. Finally provided the p-values set $\{p_j\}_{j=1}^m$, the BH procedure (Benjamini & Hochberg, 1995) is adopted.

B. The Finite Sample Control of False Discovery Proportion

B.1. Notations and Prelimits

Let $[m] = \{1, 2, \ldots, m\}$, $\mathcal{G} = \{j \in [m] : \text{the } j\text{-th machine is a normal one}\}$ and $m_0 = |\mathcal{G}|$. For a vector $\mathbf{z}$, denote $\|\mathbf{z}\|_q$ the ℓ_q-norm and $\|\mathbf{z}\| = \|\mathbf{z}\|_2$ the Euclidean norm.

Let Σ be the covariance of $\{\mathbf{z}_{ij}\}$ on the clean machines. To ease the notation, we set $\mathbf{t}_{kj} = \sqrt{n}(\overline{\mathbf{z}}_{kj} - \hat{\boldsymbol{\mu}})$, $\overline{\mathbf{t}}_{kj} = \sqrt{n}(\overline{\mathbf{z}}_{kj} - \boldsymbol{\mu}_j^*)$. Without loss of generality we choose the identity matrix as $\boldsymbol{\Omega}$ and $W_j = \frac{1}{n}\mathbf{t}_{1j}^\top\mathbf{t}_{2j}$ in the theory. Otherwise we can replace the samples $\mathbf{z}_{ij}$ by $\boldsymbol{\Omega}^{1/2}\mathbf{z}_{ij}$ and the theory is still valid. Note that the ByMI procedure is scale-invariant, i.e. the identification set $\hat{\mathcal{B}}$ remains the same when $\{W_j\}$ are multiplied by a positive scalar, we cancel the factor $\frac{1}{n}$ and set $W_j = \mathbf{t}_{1j}^\top\mathbf{t}_{2j}$ thereafter.

B.2. Useful lemmas

Lemma B.1 (Berry-Esseen Inequality (Petrov, 2002))**.** *Suppose that $X_1, \ldots, X_n$ are independent random variables with mean zero, satisfying $\mathbb{E}[|X_j|^{2+q}] < \infty$, for some $q > 0$. Denote $\kappa = \min(1, q)$. Let $B_n = \sum_{i=1}^n \mathbb{E}X_i^2$ and $L_n = B_n^{-1-\frac{\kappa}{2}} \sum_{i=1}^n \mathbb{E}|X_i|^{2+\kappa}$. There exists a universal constant $A > 0$, such that*

$$\max_{-\infty<x<\infty} |F_n(x) - \Phi(x)| \le AL_n, \tag{6}$$

where $\Phi(\cdot)$ is the distribution function of the standard Gaussian distribution and $F_n(x)$ is the distribution function of the normalized summation, i.e. $F_n(x) \triangleq \mathbb{P}[B_n^{-\frac{1}{2}} \sum_{i=1}^n X_i \le x]$. When $X_1, \ldots, X_n$ are identically distributed with $\mathbb{E}X_1^2 = \sigma^2$ and $\mathbb{E}|X_1|^{2+\kappa} = \gamma^{2+\kappa}$, we have $L_n = \frac{\gamma^{2+\kappa}}{\sigma^{2+\kappa}n^{\kappa/2}}$.

Lemma B.2 (Moderate deviations for finite-moment random variables (Petrov, 2002))**.** *Under the same conditions in Lemma B.1, for any constant $0 < \eta < 1$ and $0 \le x \le \eta(2\log\frac{1}{L_n})^{\frac{1}{2}}$,*

$$\left|\frac{1-F_n(x)}{1-\Phi(x)} - 1\right| \le CL_n^{1-\eta^2}(\log\frac{1}{L_n})^{\frac{1}{2}}, \tag{7}$$

and

$$\left|\frac{F_n(-x)}{\Phi(-x)} - 1\right| \le CL_n^{1-\eta^2}(\log\frac{1}{L_n})^{\frac{1}{2}}, \tag{8}$$

where $C = 4\sqrt{\pi}\eta A$.

Lemma B.3. *Assume that the random vector* $\mathbf{x} \in \mathbb{R}^d$ *satisfies the* L_q-L_2 *norm equivalence condition with mean* $\boldsymbol{\mu}^*$ *and covariance* Σ. *We have for any fixed vector* $\boldsymbol{\mu} \in \mathbb{R}^d$,

$$\mathbb{P}(\|\mathbf{x} - \boldsymbol{\mu}\| > T + \|\boldsymbol{\mu} - \boldsymbol{\mu}^*\|) \leq \gamma_q^q \|\Sigma\|^{\frac{q}{2}} d^{\frac{q}{2}} T^{-q}. \tag{9}$$

Proof. Since $\|\mathbf{x} - \boldsymbol{\mu}^*\|^q \leq d^{\frac{q}{2}-1}\|\mathbf{x} - \boldsymbol{\mu}^*\|_q^q$, we have

$$\begin{aligned}&\mathbb{P}(\|\mathbf{x} - \boldsymbol{\mu}\| > T + \|\boldsymbol{\mu} - \boldsymbol{\mu}^*\|) \leq \mathbb{P}(\|\mathbf{x} - \boldsymbol{\mu}^*\| > T)\\ \leq&\mathbb{P}(\|\mathbf{x} - \boldsymbol{\mu}^*\|_q^q > d^{-\frac{q}{2}+1} T^q) \leq \mathbb{E}[\|\mathbf{x} - \boldsymbol{\mu}\|_q^q] d^{\frac{q}{2}-1} T^{-q}.\end{aligned}$$

By the L_q-L_2 equivalence,

$$\mathbb{E}[\|\mathbf{x} - \boldsymbol{\mu}\|_q^q] \leq d\|\Sigma\|^{\frac{q}{2}} \gamma_q^q.$$

Hence we obtain

$$\mathbb{P}(\|\mathbf{x} - \boldsymbol{\mu}\| > T + \|\boldsymbol{\mu} - \boldsymbol{\mu}^*\|) \leq \gamma_q^q \|\Sigma\|^{\frac{q}{2}} d^{\frac{q}{2}} T^{-q}.$$

□

B.3. Proof of Lemma 3.1

Denote $\Delta_j = |\mathbb{P}(W_j > 0 \mid \mathcal{D}_1, |W_j|) - \frac{1}{2}|$. Fix $\epsilon > 0$ and for any threshold $t > 0$, denote

$$R_\epsilon(t) = \frac{\sum_{j \in \mathcal{G}} \mathbf{1}(W_j \geq t, \Delta_j \leq \epsilon)}{1 + \sum_{j \in \mathcal{G}} \mathbf{1}(W_j \leq -t)}. \tag{10}$$

Assume that the event $\mathcal{A} = \{\Delta \triangleq \max_{j \in \mathcal{G}} \Delta_j \leq \epsilon\}$ holds. Then by the definition of L,

$$\begin{aligned}&\frac{\sum_{j \in \mathcal{G}} \mathbf{1}(W_j \geq L, \Delta_j \leq \epsilon)}{1 \vee \sum_{j \in [m]} \mathbf{1}(W_j \geq L)} = \frac{1 + \sum_{j \in [m]} \mathbf{1}(W_j \leq -L)}{1 \vee \sum_{j \in [m]} \mathbf{1}(W_j \geq L)} \times \frac{\sum_{j \in \mathcal{G}} \mathbf{1}(W_j \geq L, \Delta_j \leq \epsilon)}{1 + \sum_{j \in [m]} \mathbf{1}(W_j \leq -L)}\\ \leq&\alpha \times R_\epsilon(L).\end{aligned}$$

Denote L_j the critical value like L while replacing W_j by $|W_j|$. Let $\mathbf{W}_{-j} = \{W_k\}_{k \neq j}$. The following equations are all conditional on $\mathcal{D}_1$.

$$\begin{aligned}&\mathbb{E}[R_\epsilon(L)] = \sum_{j \in \mathcal{G}} \mathbb{E}\Big[\frac{\mathbf{1}(W_j \geq L, \Delta_j \leq \epsilon)}{1 + \sum_{j \in \mathcal{G}} \mathbf{1}(W_j \leq -L)}\Big]\\ =&\sum_{j \in \mathcal{G}} \mathbb{E}\Big[\frac{\mathbf{1}(W_j \geq L_j, \Delta_j \leq \epsilon)}{1 + \sum_{k \in \mathcal{G}, k \neq j} \mathbf{1}(W_k \leq -L_j)}\Big]\\ =&\sum_{j \in \mathcal{G}} \mathbb{E}\Big[\mathbb{E}\Big\{\frac{\mathbf{1}(W_j \geq L_j, \Delta_j \leq \epsilon)}{1 + \sum_{k \in \mathcal{G}, k \neq j} \mathbf{1}(W_k \leq -L_j)} \;\Big|\; |W_j|, \mathbf{W}_{-j}\Big\}\Big]\\ =&\sum_{j \in \mathcal{G}} \mathbb{E}\Big[\frac{\mathbb{P}[W_j > 0 \mid |W_j|, \mathcal{D}_1]\mathbf{1}(|W_j| \geq L_j, \Delta_j \leq \epsilon)}{1 + \sum_{k \in \mathcal{G}, k \neq j} \mathbf{1}(W_k \leq -L_j)}\Big]\\ \leq&\sum_{j \in \mathcal{G}} \mathbb{E}\Big[\frac{(\frac{1}{2} + \Delta_j)\mathbf{1}(|W_j| \geq L, \Delta_j \leq \epsilon)}{1 + \sum_{k \in \mathcal{G}, k \neq j} \mathbf{1}(W_k \leq -L_j)}\Big]\\ \leq&(\frac{1}{2} + \epsilon)\Big[\sum_{j \in \mathcal{G}} \mathbb{E}\Big\{\frac{\mathbf{1}(W_j \geq L_j, \Delta_j \leq \epsilon)}{1 + \sum_{k \in \mathcal{G}, k \neq j} \mathbf{1}(W_k \leq -L_j)}\Big\} + \sum_{j \in \mathcal{G}} \mathbb{E}\Big\{\frac{\mathbf{1}(W_j \leq -L_j)}{1 + \sum_{k \in \mathcal{G}, k \neq j} \mathbf{1}(W_k \leq -L_j)}\Big\}\Big]\\ \leq&(\frac{1}{2} + \epsilon)\Big[\mathbb{E}\{R_\epsilon(L)\} + \sum_{j \in \mathcal{G}} \mathbb{E}\Big\{\frac{\mathbf{1}(W_j \leq -L_j)}{1 + \sum_{k \in \mathcal{G}, k \neq j} \mathbf{1}(W_k \leq -L_j)}\Big\}\Big].\end{aligned}$$

Recall that the above result is conditional on $\mathcal{A}$. The second term in the last equation is equal to 0 if for all $j \in \mathcal{G}$, $W_j > -L_j$ and otherwise $\sum_{j\in\mathcal{G}} \mathbb{E}\{\frac{\mathbf{1}(W_j \le -L_j)}{1+\sum_{k\in\mathcal{G},k\neq j}\mathbf{1}(W_k\le -L_j)}\} = \sum_{j\in\mathcal{G}} \mathbb{E}\{\frac{\mathbf{1}(W_j \le -L_j)}{1+\sum_{k\in\mathcal{G},k\neq j}\mathbf{1}(W_k\le -L_k)}\} = 1$, see Barber et al. (2020). In summary we have that conditional on $\mathcal{A}$, $\sum_{j\in\mathcal{G}} \mathbb{E}\{\frac{\mathbf{1}(W_j \le -L_j)}{1+\sum_{k\in\mathcal{G},k\neq j}\mathbf{1}(W_k\le -L_j)}\} \le 1$. Finally, we obtain

$$\begin{aligned}\mathrm{FDR}(\hat{\mathcal{B}}) =& \mathbb{E}\left[\frac{|j\in\mathcal{G}: W_j \ge L|}{|j\in[m]: W_j \ge L|}\right] \le \min_{\epsilon\ge 0}\left[\alpha\mathbb{E}[R_\epsilon(L)] + \mathbb{E}\Big\{\mathbb{P}(\max_{j\in\mathcal{G}}\Delta_j > \epsilon \mid \mathcal{D}_1)\Big\}\right] \\ \le& \min_{\epsilon\ge 0}\left[\alpha(1+5\epsilon) + \mathbb{E}\Big\{\mathbb{P}(\max_{j\in\mathcal{G}}\Delta_j > \epsilon \mid \mathcal{D}_1)\Big\}\right],\end{aligned}$$

where $\Delta_j = |\mathbb{P}(W_j > 0 \mid \mathcal{D}_1, |W_j|) - \frac{1}{2}|$.

B.4. Proof of Theorem 3.4

The above analysis provides a general finite-sample guarantee for FDR control, in the sense that it requires no model or distribution assumptions. The quantity Δ_j at the end measures the asymmetricity of ranking scores W_j for those normal machines $\mathcal{G}$. The lemma implies that a tight control of Δ_j's leads to effective FDR control of the proposed ByMI procedure. To prove Theorem 3.4, it is sufficient to control the probability $\mathbb{P}(\max_{j\in\mathcal{G}}\Delta_j > \epsilon \mid \mathcal{D}_1)$.

When conditioning on $\mathcal{D}_1$, we use the notation $\mathbf{u}_j = \mathbf{t}_{1j}$ to emphasize that $\mathbf{t}_{1j}$ is a fixed vector without randomness. In other words, the randomness of $\mathbf{t}_{1j}$ comes from $\mathcal{D}_1$ at all. Let $s_j = (\mathbf{u}_j^\top \Sigma \mathbf{u}_j)^{\frac{1}{2}}$ and $t_j^* = \eta s_j (2\log\frac{1}{L_n})^{\frac{1}{2}} - |\sqrt{n}\mathbf{u}_j^\top(\boldsymbol{\mu}^* - \hat{\boldsymbol{\mu}})|$. We choose the scale parameter s_j because W_j/s_j is approximately standard Gaussian for $j \in \mathcal{G}$. Let $F_j(\cdot)$ and $f_j(\cdot)$ be the distribution function and density function of W_j/s_j conditional on $\mathcal{D}_1$, respectively. Here by Lemma B.1, $F_j(\cdot)$ should be closed to $\Phi(\cdot)$, the distribution function of standard Gaussian variable.

Define the event $\mathcal{C} = \bigcap_{j\in\mathcal{G}}\{|W_j| \le t_j^*\}$. For any $j \in \mathcal{G}$, By Lemma B.1,

$$\mathbb{P}(|W_j| \ge t_j^* \mid \mathcal{D}_1) \le 2\mathbb{P}\Big(\mathbf{u}_j^\top \bar{\mathbf{t}}_{2j} \ge \eta s_j (2\log\frac{1}{L_n})^{\frac{1}{2}} \;\Big|\; \mathcal{D}_1\Big) \lesssim L_n^{\eta^2} = O(n^{-\frac{\eta^2\kappa}{2}}).$$

Thus, we obtain

$$\begin{aligned}\mathbb{P}(\max_{j\in\mathcal{G}}\Delta_j > \epsilon \mid \mathcal{D}_1) =& \mathbb{P}(\max_{j\in\mathcal{G}}\Delta_j > \epsilon \mid \mathcal{C}, \mathcal{D}_1) + \mathbb{P}(\mathcal{C}^\complement \mid \mathcal{D}_1) \\ =& \mathbb{P}(\max_{j\in\mathcal{G}}\Delta_j > \epsilon \mid \mathcal{C}, \mathcal{D}_1) + \sum_{j\in\mathcal{G}}\mathbb{P}(|W_j| \ge t_j^* \mid \mathcal{D}_1) \\ \le& \mathbb{P}(\max_{j\in\mathcal{G}}\Delta_j > \epsilon \mid \mathcal{C}, \mathcal{D}_1) + O(|\mathcal{G}|n^{-\frac{\eta^2\kappa}{2}}) \\ \le& \mathbb{P}(\max_{j\in\mathcal{G}}\Delta_j > \epsilon \mid \mathcal{C}, \mathcal{D}_1) + O(mn^{-\frac{\eta^2\kappa}{2}})\end{aligned}$$

Given $\mathcal{C}$, we have

$$\begin{aligned}\max_{j\in\mathcal{G}}\Delta_j \le& \max_{j\in\mathcal{G}}\big|\mathbb{P}(W_j > 0 \mid \mathcal{D}_1, |\overline{W}_j|) - 1/2\big| \\ \le& \max_{j\in\mathcal{G}}\sup_{0\le t\le t_j^*/s_j}\left|\frac{f_j(t)}{f_j(-t)} - 1\right|.\end{aligned}$$

By Lemma B.1, we have $|F_j(t) - \Phi(t)| \le AL_n + \sqrt{n}\delta_\mu$. Let $\alpha_n = AL_n = O(n^{-\frac{\kappa}{2}})$ as in Lemma B.1 and $h =$

$\sqrt{\alpha_n + n^{\frac{1}{2}}\delta_\mu}$. By the Taylor's expansion, $F_j(t+h) = F_j(t) + hf_j(t) + O(h^2)$. Therefore for any $0 \le t \le t_j^*/s_j$,

$$\begin{aligned} f_j(t) =& \frac{F_j(t+h) - F_j(t)}{h} + \xi_h \\ =& \frac{\big(F_j(t+h) - \Phi(t+h)\big) - \big(F_j(t) - \Phi(t)\big) + \Phi(t+h) - \Phi(t)}{h} + \xi_h \\ =& O\Big(\frac{\alpha_n + n^{\frac{1}{2}}\delta_\mu}{h}\Big) + \phi(t) + \xi_h \\ =& \phi(t) + \xi_h, \end{aligned}$$

where $\xi_h = O(h)$. Similarly, we have $f_j(-t) = \phi(-t) + \xi_h^-$ with $\xi_h^- = O(h)$. Therefore,

$$\left|\frac{f_j(t)}{f_j(-t)} - 1\right| = \left|\frac{\phi(t) + \xi_h}{\phi(-t) + \xi_h^-} - 1\right| = \frac{O(h)}{\left|\phi(-t) + \xi_h^-\right|}.$$

In fact, $\phi(t) = \phi(-t) \ge \phi(t_j^*/s_j) \ge \phi(\eta(2\log\frac{1}{L_n})^{\frac{1}{2}}) = Cn^{-\frac{\eta^2\kappa}{2}}$ for some constant $C > 0$. Hence we can choose a constant $0 < \eta < \frac{1}{\sqrt{2}}$ such that $n^{-\frac{\eta^2\kappa}{2}} \gtrsim h$ and it holds that

$$\left|\frac{f_j(t)}{f_j(-t)} - 1\right| = \frac{O(h)}{|\phi(-t) + \xi_h^-|} = O\big(hn^{\frac{\eta^2\kappa}{2}}\big) = O\Big(\sqrt{n^{-\frac{(1-2\eta^2)\kappa}{2}} + n^{\frac{1}{2}+\eta^2\kappa}\delta_\mu}\Big).$$

Therefore,

$$\mathbb{P}\left(\max_{j\in\mathcal{G}} \Delta_j > O\Big(\sqrt{n^{-\frac{(1-2\eta^2)\kappa}{2}} + n^{\frac{1}{2}+\eta^2\kappa}\delta_\mu}\Big) \;\Big|\; \mathcal{D}_1\right) \le O(mn^{-\frac{\eta^2\kappa}{2}}).$$

By applying Lemma 3.1 with $\epsilon = \sqrt{n^{-\frac{(1-2\eta^2)\kappa}{2}} + n^{\frac{1}{2}+\eta^2\kappa}\delta_\mu}$, we obtain that with probability at least $1 - O(mn^{-\frac{\eta^2\kappa}{2}})$,

$$\mathrm{FDR}(\hat{\mathcal{B}}) \le \alpha + O\Big(\sqrt{n^{-\frac{(1-2\eta^2)\kappa}{2}} + n^{\frac{1}{2}+\eta^2\kappa}\delta_\mu}\Big).$$

B.5. Proof of Theorem 3.6

By the definition of the threshold L, one obtains

$$\mathrm{FDP} = \frac{\sum_{j\in\mathcal{G}} \mathbf{1}(W_j \ge L)}{1 \vee \sum_{j\in[m]} \mathbf{1}(W_j \ge L)} = \frac{\sum_{j\in[m]} \mathbf{1}(W_j \le -L)}{1 \vee \sum_{j\in[m]} \mathbf{1}(W_j \ge L)} \times \frac{\sum_{j\in\mathcal{G}} \mathbf{1}(W_j \ge L)}{\sum_{j\in[m]} \mathbf{1}(W_j \le -L)} \le \alpha R(L), \tag{11}$$

where $R(L) = \frac{\sum_{j\in\mathcal{G}} \mathbf{1}(W_j \ge L)}{\sum_{j\in\mathcal{G}} \mathbf{1}(W_j \le -L)}$. The result follows if $R(L) - 1 = o(1)$.

Denote $G(t) = \frac{1}{|\mathcal{G}|}\sum_{j\in\mathcal{G}} \mathbb{P}(W_j \ge t \mid \mathcal{D}_1)$ and $G_-(t) = \frac{1}{|\mathcal{G}|}\sum_{j\in\mathcal{G}} \mathbb{P}(W_j \le -t \mid \mathcal{D}_1)$. We will first provide two finite sample results that uniformly control the ratio-type processes like $\frac{G(t)}{G_-(t)}$.

Lemma B.4. *Let* $r_n = L_n^{1-\eta^2}(\log\frac{1}{L_n})^{\frac{1}{2}} = O(n^{-\frac{(1-\eta^2)\kappa}{2}}(\log n)^{\frac{1}{2}})$ *where* $0 < \eta < 1$. *Assume that conditional on* $\mathcal{D}_1$, $\sqrt{n}\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}^*\| = O(1)$. *Uniformly for* $0 \le t \le G_-^{-1}(1/m)$,

$$\left|\frac{G(t)}{G_-(t)} - 1\right| \le O\Big(r_n + mn^{-\frac{\eta^2\kappa}{2}} + n^{\frac{1}{2}}\|\boldsymbol{\mu}^* - \hat{\boldsymbol{\mu}}\|\Big).$$

Lemma B.5. *For any* $1 < v < m$ *be sufficiently large and* $0 < \delta < 1$, *we have with probability* $1 - O(v^{-(1-\delta)})$,

$$\sup_{0\le t\le G^{-1}(v/m)} \left|[m_0 G(t)]^{-1} \sum_{j\in\mathcal{G}} \mathbf{1}(W_j \ge t) - 1\right| \lesssim v^{-\delta/3}, \tag{12}$$

$$\sup_{0\le t\le G_-^{-1}(v/m)} \left|[m_0 G_-(t)]^{-1} \sum_{j\in\mathcal{G}} \mathbf{1}(W_j \le -t) - 1\right| \lesssim v^{-\delta/3}. \tag{13}$$

By the definition of the threshold rule, we have

$$L = \inf\Big\{t \geq 0 : 1 + \sum_j \mathbf{1}(W_j \leq -t) \leq \alpha \max\Big(\sum_j \mathbf{1}(W_j \geq t), 1\Big)\Big\}. \tag{14}$$

We will first derive an upper bound of L so that Lemma B.4 and Lemma B.5 can be applied.

For any $j \in \mathcal{G}$, fixed $\mathbf{t}_{1j} = \mathbf{u}_j$. For $t_* = \sup_{j\in\mathcal{G}}(2\log\frac{1}{L_n}\mathbf{u}_j^\top\Sigma\mathbf{u}_j)^{\frac{1}{2}} + \sqrt{n}|\mathbf{u}_j^\top(\boldsymbol{\mu}^* - \hat{\boldsymbol{\mu}})| = \sup_{j\in\mathcal{G}}\|\mathbf{u}_j\| \times \Theta(\sqrt{n}\delta_\mu + \sqrt{\log n})$, By Lemma B.1,

$$\mathbb{P}(\sum_{j\in\mathcal{G}} \mathbf{1}(W_j \geq t_*) > 0 \mid \mathcal{D}_1) \leq \sum_{j\in\mathcal{G}} \mathbb{P}(W_j \geq t_* \mid \mathcal{D}_1) \lesssim m_0 n^{-\frac{\kappa}{2}} \leq mn^{-\frac{\kappa}{2}},$$

$$\mathbb{P}(\sum_{j\in\mathcal{G}} \mathbf{1}(W_j \leq -t_*) > 0 \mid \mathcal{D}_1) \leq \sum_{j\in\mathcal{G}} \mathbb{P}(W_j \leq -t_* \mid \mathcal{D}_1) \lesssim m_0 n^{-\frac{\kappa}{2}} \leq mn^{-\frac{\kappa}{2}}.$$

Therefore

$$\mathbb{P}(\sum_{j\in\mathcal{G}} \mathbf{1}(W_j \leq -t_*) = 0 \mid \mathcal{D}_1) = 1 - O(mn^{-\frac{\kappa}{2}}), \tag{15}$$

For the other side where $j \in \mathcal{B}_\mu$, denote $\mathbf{v}_j = \sqrt{n}(\overline{\mathbf{x}}_{1j} - \hat{\boldsymbol{\mu}})$. Observe that for $j \in \mathcal{B}_\mu$,

$$\begin{aligned}
&\mathbb{P}(W_j < t_* \mid \mathcal{D}_1) = \mathbb{P}(n(\overline{\mathbf{x}}_{1j} - \hat{\boldsymbol{\mu}})^\top(\overline{\mathbf{x}}_{2j} - \hat{\boldsymbol{\mu}}) < t_* \mid \mathcal{D}_1)\\
=&\mathbb{P}(\mathbf{v}_j^\top[\sqrt{n}(\overline{\mathbf{x}}_{2j} - \boldsymbol{\mu}_j^*)] < t_* - n(\overline{\mathbf{x}}_{1j} - \hat{\boldsymbol{\mu}})^\top(\boldsymbol{\mu}_j^* - \hat{\boldsymbol{\mu}}) \mid \mathcal{D}_1)\\
=&\mathbb{P}\left(\frac{\mathbf{v}_j^\top(\sqrt{n}(\overline{\mathbf{x}}_{2j} - \boldsymbol{\mu}_j^*))}{\sqrt{\mathbf{v}_j\Sigma_j\mathbf{v}_j}} < \frac{t_* - n(\overline{\mathbf{x}}_{1j} - \hat{\boldsymbol{\mu}})^\top(\boldsymbol{\mu}_j^* - \hat{\boldsymbol{\mu}})}{\sqrt{\mathbf{v}_j\Sigma_j\mathbf{v}_j}} \,\middle|\, \mathcal{D}_1\right)
\end{aligned}$$

By the condition on signals and the choice of t_*, the Berry-Esseen bound Lemma B.1 ensures that

$$\mathbb{P}(W_j < t_* | \mathcal{D}_1) \lesssim n^{-\frac{\kappa}{2}}.$$

By taking the union bound, we obtain

$$\mathbb{P}(\sum_{j\in[m]} \mathbf{1}(W_j \geq t_*) \geq \psi_m \mid \mathcal{D}_1) \geq \mathbb{P}(\sum_{j\in\mathcal{B}_\mu} \mathbf{1}(W_j \geq t_*) \geq \psi_m \mid \mathcal{D}_1) = 1 - O(\psi_m n^{-\frac{\kappa}{2}}). \tag{16}$$

Let $\tilde{t} = G_-^{-1}(\frac{\alpha\psi_m}{m})$. By Lemma B.5, with probability at least $1 - O(\psi_m^{-(1-\delta)})$, $\frac{\alpha\psi_m}{m} = G_-(\tilde{t}) = \frac{1}{m_0}\sum_{j\in\mathcal{G}}\mathbf{1}(W_j \leq -\tilde{t})[1 + O(\psi_m^{-\delta/3})]$. Hence combining with Equation 15, one obtains $\tilde{t} \leq t^*$. It also implies that

$$\mathbb{P}(\sum_{j\in[m]} \mathbf{1}(W_j \geq \tilde{t}) \geq \psi_m \mid \mathcal{D}_1) \geq \mathbb{P}(\sum_{j\in\mathcal{B}_\mu} \mathbf{1}(W_j \geq \tilde{t}) \geq \psi_m \mid \mathcal{D}_1) = 1 - O(\psi_m n^{-\frac{\kappa}{2}}). \tag{17}$$

Therefore with probability at least $1 - O(mn^{-\frac{\eta^2\kappa}{2}} + \psi_m^{-(1-\delta)} + \psi_m n^{-\frac{\kappa}{2}}) = 1 - O(mn^{-\frac{\eta^2\kappa}{2}} + \psi_m^{-(1-\delta)})$,

$$1 + \sum_{j\in[m]} \mathbf{1}(W_j \leq -\tilde{t}) = 1 + \sum_{j\in\mathcal{G}} \mathbf{1}(W_j \leq -\tilde{t}) = \frac{\alpha\psi_m m_0}{m[1 + O(\psi_m^{-\delta/3})]} \leq \alpha\sum_j \mathbf{1}(W_j \geq \tilde{t}).$$

Thus $L \leq \tilde{t} = G_-^{-1}(\frac{\alpha\psi_m}{m})$. Under the above event, one can apply Lemma B.4 and Lemma B.5 so that

$$\frac{\sum_{j\in\mathcal{G}}\mathbf{1}(W_j \geq L)}{\sum_{j\in\mathcal{G}}\mathbf{1}(W_j \leq -L)} = 1 + O\Big(n^{-\frac{(1-\eta^2)\kappa}{2}}(\log n)^{\frac{1}{2}} + mn^{-\frac{\eta^2\kappa}{2}} + n^{\frac{1}{2}}\|\boldsymbol{\mu}^* - \hat{\boldsymbol{\mu}}\| + (\alpha\psi_m)^{-\delta/3}\Big).$$

Accordingly,

$$\mathrm{FDP} \leq \alpha\left[1 + O\Big(n^{-\frac{(1-\eta^2)\kappa}{2}}(\log n)^{\frac{1}{2}} + mn^{-\frac{\eta^2\kappa}{2}} + n^{\frac{1}{2}}\|\boldsymbol{\mu}^* - \hat{\boldsymbol{\mu}}\| + (\alpha\psi_m)^{-\delta/3}\Big)\right]. \tag{18}$$

B.6. Proof of Corollary 3.7

Corollary 3.7 follows from Equation (17).

B.7. Proof of Lemma B.4 & Lemma B.5

Proof of Lemma B.4. For any $j \in \mathcal{G}$, fixed $\mathbf{t}_{1j} = \mathbf{u}_j$. By Lemma B.2, for $0 \le t \le \eta(2\log\frac{1}{L_n}\mathbf{u}_j^\top\Sigma\mathbf{u}_j)^{\frac{1}{2}} + \sqrt{n}\mathbf{u}_j^\top(\boldsymbol{\mu}^* - \hat{\boldsymbol{\mu}})$, $\mathbb{P}(W_j \ge t \mid \mathcal{D}_1) = \mathbb{P}(\mathbf{u}_j^\top\overline{\mathbf{t}}_{2j} \ge t - \sqrt{n}\mathbf{u}_j^\top(\boldsymbol{\mu}^* - \hat{\boldsymbol{\mu}}) \mid \mathcal{D}_1) = \overline{\Phi}(\frac{t-\sqrt{n}\mathbf{u}_j^\top(\boldsymbol{\mu}^*-\hat{\boldsymbol{\mu}})}{(\mathbf{u}_j^\top\Sigma\mathbf{u}_j)^{\frac{1}{2}}})[1+O(r_n)] = \overline{\Phi}(\frac{t}{(\mathbf{u}_j^\top\Sigma\mathbf{u}_j)^{\frac{1}{2}}})[1+O(r_n + n^{\frac{1}{2}}\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}^*\|)]$. Similarly, for $0 \le t \le \eta(2\log\frac{1}{L_n}\mathbf{u}_j^\top\Sigma\mathbf{u}_j)^{\frac{1}{2}} - \sqrt{n}\mathbf{u}_j^\top(\boldsymbol{\mu}^* - \hat{\boldsymbol{\mu}})$,

$$\mathbb{P}(W_j \le -t \mid \mathcal{D}_1) = \Phi\Big(-\frac{t}{(\mathbf{u}_j^\top\Sigma\mathbf{u}_j)^{\frac{1}{2}}}\Big)\Big[1 + O(r_n + n^{\frac{1}{2}}\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}^*\|)\Big].$$

Else if $\eta(2\log\frac{1}{L_n}\mathbf{u}_j^\top\Sigma\mathbf{u}_j)^{\frac{1}{2}} - |\sqrt{n}\mathbf{u}_j^\top(\boldsymbol{\mu}^* - \hat{\boldsymbol{\mu}})| \le t \le G_-^{-1}(1/m)$, by the Berry-Esseen Inequality Lemma B.1,

$$\mathbb{P}(W_j \ge t \mid \mathcal{D}_1) \le \mathbb{P}\Big(\mathbf{u}_j^\top\overline{\mathbf{t}}_{2j} \ge \eta(2\log\frac{1}{L_n}\mathbf{u}_j^\top\Sigma\mathbf{u}_j)^{\frac{1}{2}} \;\Big|\; \mathcal{D}_1\Big) \lesssim L_n^{\eta^2} = O(n^{-\frac{\eta^2\kappa}{2}}).$$

The same result holds for $\mathbb{P}(W_j \le -t \mid \mathcal{D}_1)$. Now for a fixed $t > 0$, we can divide $j \in [m]$ into two sets by the above conditions on $\mathbf{u}_j$. Let $\mathcal{I}_{0,1} = \{j \in \mathcal{G} : t \le \eta(2\log\frac{1}{L_n}\mathbf{u}_j^\top\Sigma\mathbf{u}_j)^{\frac{1}{2}} - \sqrt{n}|\mathbf{u}_j^\top(\boldsymbol{\mu}^* - \hat{\boldsymbol{\mu}})|\}$ and $\mathcal{I}_{0,2} = \mathcal{G} \setminus \mathcal{I}_{0,1}$. One can obtain

$$\begin{aligned}
&\Big|\frac{G(t)}{G_-(t)} - 1\Big| = \frac{|\sum_{j\in\mathcal{G}}\mathbb{P}(W_j \ge t \mid \mathcal{D}_1) - \mathbb{P}(W_j \le -t \mid \mathcal{D}_1)|}{\sum_{j\in\mathcal{G}}\mathbb{P}(W_j \le -t \mid \mathcal{D}_1)} \\
\le&\frac{|\sum_{j\in\mathcal{I}_{0,1}}\mathbb{P}(W_j \ge t \mid \mathcal{D}_1) - \mathbb{P}(W_j \le -t \mid \mathcal{D}_1)|}{\sum_{j\in\mathcal{G}}\mathbb{P}(W_j \le -t \mid \mathcal{D}_1)} + \frac{|\sum_{j\in\mathcal{I}_{0,2}}\mathbb{P}(W_j \ge t \mid \mathcal{D}_1) - \mathbb{P}(W_j \le -t \mid \mathcal{D}_1)|}{\sum_{j\in\mathcal{G}}\mathbb{P}(W_j \le -t \mid \mathcal{D}_1)} \\
\le&\frac{|\sum_{j\in\mathcal{I}_{0,1}}\mathbb{P}(W_j \ge t \mid \mathcal{D}_1) - \mathbb{P}(W_j \le -t \mid \mathcal{D}_1)|}{\sum_{j\in\mathcal{I}_{0,1}}\mathbb{P}(W_j \le -t \mid \mathcal{D}_1)} \\
+&\frac{\max(\sum_{j\in\mathcal{I}_{0,2}}\mathbb{P}(W_j \ge t \mid \mathcal{D}_1), \sum_{j\in\mathcal{I}_{0,2}}\mathbb{P}(W_j \le -t \mid \mathcal{D}_1))}{m_0/m} \\
\lesssim& r_n + n^{\frac{1}{2}}\|\boldsymbol{\mu}^* - \hat{\boldsymbol{\mu}}\| + mL_n^{\eta^2} = O\Big(n^{-\frac{(1-\eta^2)\kappa}{2}}(\log n)^{\frac{1}{2}} + mn^{-\frac{\eta^2\kappa}{2}} + n^{\frac{1}{2}}\|\boldsymbol{\mu}^* - \hat{\boldsymbol{\mu}}\|\Big).
\end{aligned}$$

□

Proof of Lemma B.5. Let $z_0 < z_1 < \cdots < z_s \le 1$ and $t_i = G^{-1}(z_i)$ where $z_0 = \frac{v}{m}$ and $z_i = \frac{v}{m}(1+\xi)^i$ with $s = \lfloor\frac{\log(v/m)}{\log(1+\xi)}\rfloor$ with some sufficiently small constant $\xi > 0$. Note that $G(t_i)/G(t_{i+1}) = z_i/z_{i+1} = 1/(1+\xi) = 1 - O(\xi)$ uniformly in i. It is therefore enough to derive the convergence rate of the supremum,

$$D = \sup_{0\le i\le s}\left|\frac{\sum_{j\in\mathcal{G}}\{\mathbf{1}(W_j \ge t_i) - \mathbb{P}(W_j \ge t_i)\}}{m_0 G(t_i)}\right|.$$

Note that the process of variance satisfies,

$$\begin{aligned}
D(t) &= \mathbb{E}\Big[\Big\{\sum_{j\in\mathcal{G}}\mathbf{1}(W_j \ge t) - \mathbb{P}(W_j \ge t \mid \mathcal{D}_1)\Big\}^2 \;\Big|\; \mathcal{D}_1\Big] \\
&= \sum_{j\in\mathcal{G}}\mathbb{E}\Big[\Big\{\mathbf{1}(W_j \ge t) - \mathbb{P}(W_j \ge t \mid \mathcal{D}_1)\Big\}^2 \;\Big|\; \mathcal{D}_1\Big] \le m_0 G(t).
\end{aligned}$$

We obtain

$$\begin{aligned}
\mathbb{P}(D \ge a) &\le \sum_{0\le i\le s}\mathbb{P}\left(\left|\frac{\sum_{j\in\mathcal{G}}\{\mathbf{1}(W_j \ge t_i) - \mathbb{P}(W_j \ge t_i)\}}{m_0 G(t_i)}\right| \ge a\right) \\
&\le \frac{1}{a^2}\sum_{0\le i\le s}\frac{1}{m_0 G(t_i)} \lesssim \frac{m}{a^2 m_0 v\xi} \lesssim \frac{1}{a^2\xi v}.
\end{aligned}$$

Hence with probability at least $1 - O(\frac{1}{a^2\xi v})$,

$$\sup_{0\le t\le G^{-1}(v/m)} \Big| [m_0 G(t)]^{-1} \sum_{j\in\mathcal{G}} \mathbf{1}(W_j \ge t) - 1 \Big| \lesssim a + \xi,$$

Finally, we simplify the above result. Consider the case that $\xi = a = v^{-\delta/3}$, we have with probability at least $1-O(v^{-(1-\delta)})$,

$$\sup_{0\le t\le G^{-1}(v/m)} \Big| [m_0 G(t)]^{-1} \sum_{j\in\mathcal{G}} \mathbf{1}(W_j \ge t) - 1 \Big| \lesssim v^{-\delta/3}.$$

□

C. Pseudocode of ByMI

The following algorithm outlines the steps of the ByMI method.

Algorithm 1 The ByMI Procedure

Input: Machines $\{\mathcal{M}_j\}_{j=0}^m$, $\boldsymbol{\theta}$

1: Randomly split the samples on $\mathcal{M}_j$ into two sets $\mathcal{D}_j^{(1)}$ and $\mathcal{D}_j^{(2)}$ of equal size $\frac{n}{2}$, for $j \in [m]$.
2: Compute the empirical gradients $\mathbf{g}_{1j}(\boldsymbol{\theta})$ and $\mathbf{g}_{2j}(\boldsymbol{\theta})$ based on $\mathcal{D}_j^{(1)}$ and $\mathcal{D}_j^{(2)}$ with $\boldsymbol{\theta}$, for $j \in [m]$.
3: Adopt the robust mean algorithm $\mathcal{A}$ to obtain the robust mean estimator $\widehat{\mathbf{g}}(\boldsymbol{\theta})$ based on $\{\mathbf{g}_{1j}(\boldsymbol{\theta})\}_{j=0}^m$.
4: Calculate the ranking scores $\{W_j\}_{j=1}^m$ according to (2).
5: Compute the threshold L in (3).

Output: $\widehat{\mathcal{B}} = \{\mathcal{M}_j : W_j \ge L\}$.

D. Additional Numerical Results

D.1. Details of the Implementation

All the experiments are conducted on an Ubuntu 20.04 LTS server with 64 Intel(R) Xeon(R) Gold 5218 CPUs @ 2.30GHz, 128G RAM and the R platform with version 4.0.2. The code is available at https://github.com/mywang99/ByMI.

D.2. Choice of Scale Matrix

Here we compare three choices of the scale matrix $\boldsymbol{\Omega} = \mathrm{diag}\{\hat{\sigma}_1^{-2}, \ldots, \hat{\sigma}_d^{-2}\}$.

Choice 1. Take $\hat{\sigma}_k = \mathrm{MAD}\left(g_1^{(k)}(\boldsymbol{\theta}), \cdots, g_m^{(k)}(\boldsymbol{\theta})\right)$, $k = 1, \cdots, d$, where $g_j^{(k)}(\boldsymbol{\theta})$ is the k-th component of $\mathbf{g}_j(\boldsymbol{\theta})$.

Choice 2. Adopt $\boldsymbol{\Omega}^{-1} = \mathrm{diag}(\widehat{\Sigma}_0)$ where $\widehat{\Sigma}_0$ is the sample covariance of the samples on the master machine $\mathcal{M}_0$.

Choice 3. Let $(\hat{\sigma}_1^2, \ldots, \hat{\sigma}_d^2)^\top = \widehat{\mathbf{g}}_2(\boldsymbol{\theta}) - \widehat{\mathbf{g}}(\boldsymbol{\theta})$, where $\widehat{\mathbf{g}}_2(\boldsymbol{\theta})$ and $\widehat{\mathbf{g}}(\boldsymbol{\theta})$ are the robust mean estimators of $\mathbb{E}[\mathbf{g}(\boldsymbol{\theta})^2]$ and $\mathbb{E}[\mathbf{g}(\boldsymbol{\theta})]$ respectively.

In the comparison, we choose the filtering estimator as the robust aggregator. Table 2 presents the results of ByMI-Filter when $p = 50$ and 100 under Scenario A in Section 4.1. The result shows that all the choices can achieve FDR control and there is little difference in TPRs and P_a, which implies that ByMI is not sensitive to the scale estimator. Note that Choice 1 is adopted across all the experiments.

D.3. Results of Other Robust Estimators Combined with ByMI

In this section, we examine the performance of different ByMI-based methods and the scale matrix $\boldsymbol{\Omega}^{-1} = \mathrm{diag}\{\hat{\sigma}_1^2, \ldots, \hat{\sigma}_d^2\}$ is adopted here. In the simulation studies, we consider five robust mean estimators which are listed as follows.

Table 2. FDR(%), TPR(%) and P_a(%) adopting different scale estimator choices when $b = 0.15$ under Scenario A.

	$p = 50$			$p = 100$		
Method	FDR	TPR	P_a	FDR	TPR	P_a
Choice 1	9.2	97.5	38.6	8.9	98.7	57.0
Choice 2	9.4	97.8	40.6	9.1	98.8	60.4
Choice 3	9.4	98.1	47.2	9.1	99.0	64.0

- MOM, the median of means estimator (Yin et al., 2018),
- VRMOM, the variance reduced median-of-means estimator (Tu et al., 2021),
- GEOM, the geometric median (Minsker, 2015),
- GD, the first-order method (Cheng et al., 2020),
- Filtering, the filtering algorithm (Diakonikolas et al., 2017; Lai et al., 2016; Zhu et al., 2022).

Figure 6 reports the FDR and TPR curves over the shift size b with the contamination level $\varrho = 0.05$ under Scenario A in Section 4.1. As can be seen, the FDR levels of almost all ByMI-based methods are close to the nominal level $\alpha = 0.1$. In the high-dimensional case, ByMI combined with dimension-dependent robust estimators, such as GEOM, MOM and VRMOM, fail to achieve high TPRs and give larger FDRs. With the superiority of dimension-agnostic property, ByMI-GD and ByMI-Filter control FDRs better and their TPRs are also higher.

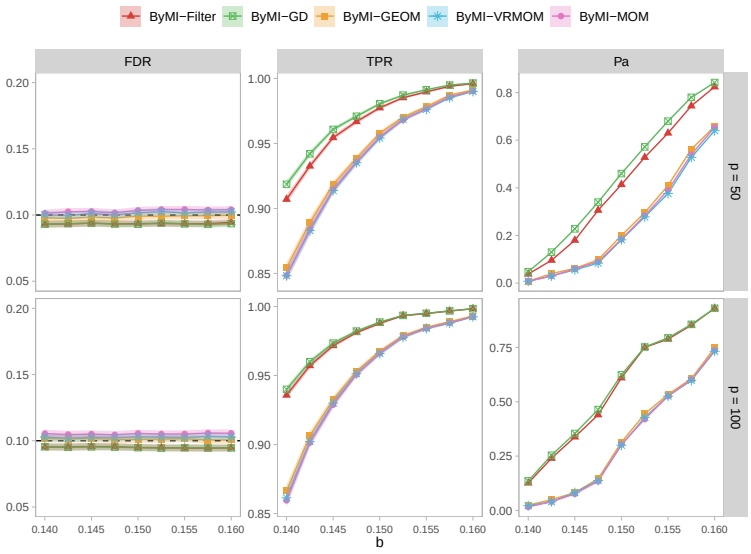

Figure 6. FDR, TPR and P_a of ByMI-type methods over shift size b when $\varrho = 0.05$ and $p = 50, 100$ under Scenario A. The black dashed lines indicate the FDR level $\alpha = 0.1$.

We also consider the combination of ByMI and the Randon point method (Kamp et al., 2017). Table 3 reports the empirical results of FDR, TPR and P_a under Scenario A when $b = 0.15$ based on 500 replications, demonstrating that the Radon point method shows comparable performance to the geometric median.

D.4. Results on Real Data

In this section, we conduct more experiments on the MNIST, F-MNIST and CIFAR10 datasets in various settings. The results of other ByMI-based methods are also reported.

Table 3. FDR(%), TPR(%) and P_a (%) of ByMI-type methods when $b = 0.15$ under Scenario A.

	$p = 50$			$p = 100$		
Method	FDR	TPR	P_a	FDR	TPR	P_a
ByMI-Filter	9.5	97.7	39.4	9.6	98.7	57.0
ByMI-Filter+	9.0	100.0	100.0	9.3	100.0	100.0
ByMI-GEOM	10.0	95.7	20.8	10.3	96.6	27.6
ByMI-Radon	10.3	95.6	18.8	10.6	96.8	28.6

D.4.1. Results for the deep learning task

We try the convolution neural network used in Zhu et al. (2023) with the IPM attack and the MNIST dataset. The results of FDR, TPR and P_a in the first iteration are summarized in Table 4 based on 50 repeats. Both ByMI-Filter and ByMI-Filter+ can control the FDR error and achieve high TPRs.

In theory, the problem is that the dimension d of the gradients here will be enormous. In this case, the projection-based alternatives like ByMI-Filter+ (in which we project the d-variate score to one dimension) can still be valid. For the ByMI method where $\mathbf{\Omega}$ is the identity matrix or with full-rank, the mean estimation error δ_μ can be large and it brings some difficulties to obtain the FDR control theory. However, the results in Table 4 indicate that ByMI remains effective in numerical experiments. This could be attributed to the fact that the matrix comprising all sample gradients (scores) of the deep neural network has a low-rank structure, which implies a relatively small effective dimension.

Table 4. FDR(%), TPR(%) and P_a (%) of the IPM attack for the MNIST dataset in the deep learning task.

	$\varrho = 0.1$			$\varrho = 0.2$		
Method	FDR	TPR	P_a	FDR	TPR	P_a
ByMI-Filter	0.7	70.0	70.0	5.0	80.0	80.0
ByMI-Filter+	0.7	60.0	60.0	3.6	80.0	80.0
ByMI-GEOM	1.3	70.0	70.0	9.4	70.0	70.0
Krum	48.3	66.4	60.0	44.3	69.1	65.0
FABA	48.3	66.4	50.0	65.6	42.8	40.0
Zeno	45.6	70.0	50.0	45.7	67.4	40.0

D.4.2. Extra Results on Byzantine Machine Identification

In the following experiments, we additionally compute the *estimation error*. To be specific, we use the average of the local gradients on the normal machines as the oracle $\mathbf{g}^*$ and measure the estimation error by the ℓ_2-norm loss $\|\hat{\mathbf{g}} - \mathbf{g}^*\|$ of the aggregated gradient estimators.

Table 5 shows the performance of the proposed ByMI method in conjunction with various robust estimators and other detection methods under the OOD attack when the contamination level $\varrho = 0.15$ and 0.2. Similar to the experiments in Section 4.2, we summarize the results in the first iteration. For the ByMI-based methods, ByMI combined with dimension-agnostic robust estimators, ByMI-Filter, ByMI-Filter+ and ByMI-GD, perform well reasonably. Meanwhile, ByMI combined with other less robust estimators, such as MOM, VRMOM and GEOM, are unable to control the FDR at the desired level. Due to the failure of p-value in the non-asymptotic environment, RMDP-BH detects almost all the machines as Byzantine machines. Compared to ByMI-based methods, distance-based methods (Krum, FABA and Zeno) give larger FDRs and lower TPRs. A similar phenomenon can be observed in Table 6, which summarizes results against different contamination ratios under the IPM attack based on 500 replications.

D.4.3. Results for the Random Gradient Attack

Besides the OOD attack and IPM attack, we conduct experiments on the random gradient attack. In detail, we replace the gradient $\mathbf{g}_i$ of the sample $(y_i, \mathbf{x}_i)$ on Byzantine machines by a Gaussian vector $\mathcal{N}_d(\bar{\mathbf{g}} + 0.8\mathbf{u}_d, 0.7\,\mathrm{diag}(\mathbf{s}_g))$ where $\bar{\mathbf{g}}$ is the sample mean vector, $\mathbf{s}_g$ corresponds to the sample variances of all the gradients $\{\mathbf{g}_i\}$ and $\mathbf{u}_d$ is uniformly sampled

Table 5. FDR(%), TPR(%), P_a(%), $|\hat{\mathcal{B}}|$ and Error (‰) of the OOD attack.

		$\varrho = 0.15$					$\varrho = 0.2$				
Dataset	Method	FDR	TPR	P_a	$\lvert\hat{\mathcal{B}}\rvert$	Error	FDR	TPR	P_a	$\lvert\hat{\mathcal{B}}\rvert$	Error
MNIST	ByMI-Filter	7.9	99.0	94.8	23.8	4.1	8.9	99.4	96.2	31.9	5.4
	ByMI-Filter+	7.5	99.8	99.8	23.9	3.3	8.9	100.0	100.0	32.0	4.7
	ByMI-GD	8.2	98.9	94.6	23.9	4.2	9.1	99.4	96.2	31.9	5.5
	ByMI-GEOM	16.3	97.6	90.8	26.4	7.1	27.9	98.1	90.6	41.4	12.3
	ByMI-VRMOM	16.6	98.0	92.2	26.6	7.0	24.5	98.6	93.2	39.3	11.0
	ByMI-MOM	16.9	98.1	92.4	26.7	7.0	24.9	98.7	93.6	39.6	11.1
	RMDP-BH	82.6	100.0	100.0	132.9	170.9	78.0	100.0	100.0	135.7	184.5
	Krum	21.6	96.2	81.4	27.0	9.4	22.1	96.7	84.8	36.0	11.9
	FABA	20.5	97.6	90.6	27.0	8.6	21.0	98.0	92.2	36.0	10.6
	Zeno	23.0	94.5	75.2	27.0	10.1	24.6	93.6	69.4	36.0	14.0
F-MNIST	ByMI-Filter	8.3	94.1	84.0	22.9	7.3	10.1	97.1	83.4	31.6	8.5
	ByMI-Filter+	8.3	96.3	93.6	23.4	6.0	9.4	99.5	99.0	32.1	6.4
	ByMI-GD	9.0	93.4	82.4	22.9	7.9	10.9	96.4	81.2	31.7	9.2
	ByMI-GEOM	14.9	91.0	76.0	24.4	11.2	24.4	92.2	70.0	37.2	17.1
	ByMI-VRMOM	14.8	92.5	79.6	24.7	10.4	22.1	94.9	75.0	36.5	14.4
	ByMI-MOM	15.1	92.6	81.2	24.8	10.5	22.3	95.1	76.0	36.7	14.4
	RMDP-BH	62.2	98.6	96.0	74.7	47.7	55.2	98.7	96.6	77.6	42.0
	Krum	30.2	85.6	57.2	27.0	18.0	28.5	88.7	60.2	36.0	21.4
	FABA	27.8	88.6	71.0	27.0	15.4	28.6	88.6	72.0	36.0	21.0
	Zeno	34.0	81.0	43.6	27.0	20.4	37.1	78.1	34.6	36.0	30.8
CIFAR10	ByMI-Filter	7.3	87.0	74.0	17.2	11.8	6.7	84.2	69.4	22.0	16.5
	ByMI-Filter+	8.7	91.9	88.2	18.4	9.4	8.4	95.1	92.8	25.2	10.1
	ByMI-GD	7.2	87.2	75.0	17.2	11.6	6.7	85.7	70.0	22.4	15.5
	ByMI-GEOM	13.6	79.1	64.4	17.5	19.3	21.7	74.4	57.2	25.8	32.0
	ByMI-VRMOM	13.0	82.7	67.0	18.0	17.0	17.6	77.6	61.6	24.5	27.5
	ByMI-MOM	14.4	83.6	68.0	18.6	17.1	19.5	78.6	62.4	25.5	27.7
	RMDP-BH	49.8	84.2	76.8	35.3	32.7	49.1	76.0	69.6	37.8	42.3
	Krum	51.7	61.7	36.6	23.0	39.2	53.4	58.2	37.6	30.0	56.1
	FABA	51.8	61.6	44.6	23.0	38.3	54.8	56.6	45.0	30.0	56.8
	Zeno	57.8	54.0	26.6	23.0	43.9	61.1	48.6	22.2	30.0	65.4

from the unit sphere. Table 7 reports the empirical results of FDR, TPR and P_a under the random gradient attack with different contamination levels based on 500 replications. Similar to the OOD attack and IPM attack, ByMI combined with dimension-agnostic robust estimators, ByMI-Filter, ByMI-FIlter+ and ByMI-GD, perform well reasonably. In contrast, RMDP-BH and distance-based methods (Krum, FABA and Zeno) do little to identify the Byzantine machines.

D.4.4. Results for larger contamination levels

We conduct more experiments on the three real datasets under the three attacks with $\varrho = 0.3, 0.35, 0.4, 0.45$. All experiments are repeated 500 times. From Table 9, we can see that the ByMI-type methods obtain high TPRs. ByMI-Filter controls the FDR well when ϱ increases to 0.4. ByMI-Filter+ and ByMI-GD can control the FDR even when $\varrho = 0.45$. Similar phenomenons can be observed in Table 8 and Table 10.

Table 6. FDR(%), TPR(%), P_a(%), $|\hat{\mathcal{B}}|$ and Error (‰) of the IPM attack when the attack parameter $a = 0.2$.

		$\varrho = 0.15$					$\varrho = 0.2$				
Dataset	Method	FDR	TPR	P_a	$\lvert\hat{\mathcal{B}}\rvert$	Error	FDR	TPR	P_a	$\lvert\hat{\mathcal{B}}\rvert$	Error
MNIST	ByMI-Filter	7.5	100.0	100.0	23.9	3.6	8.3	100.0	100.0	31.8	4.9
	ByMI-Filter+	7.9	100.0	100.0	24.1	3.4	8.7	100.0	100.0	32.0	4.6
	ByMI-GD	8.0	100.0	100.0	24.1	3.7	8.7	99.8	99.8	31.9	5.2
	ByMI-GEOM	21.6	96.2	96.2	28.2	9.5	37.3	93.4	93.4	46.6	18.9
	ByMI-VRMOM	31.2	97.4	97.4	33.1	11.5	48.7	97.4	97.4	59.2	21.6
	ByMI-MOM	38.5	98.2	98.2	37.8	13.4	60.5	99.0	99.0	76.0	28.3
	RMDP-BH	86.3	82.2	82.2	130.4	201.4	85.4	71.0	71.0	130.2	252.0
	Krum	68.1	39.1	27.2	27.0	47.4	71.6	35.2	25.6	36.0	72.2
	FABA	96.5	4.3	2.8	27.0	74.2	98.9	1.4	1.0	36.0	108.4
	Zeno	95.4	5.7	2.6	27.0	72.2	97.9	2.6	1.2	36.0	105.3
F-MNIST	ByMI-Filter	7.7	100.0	100.0	24.0	4.4	8.6	99.8	99.8	31.9	6.4
	ByMI-Filter+	8.1	100.0	100.0	24.1	4.2	9.7	100.0	100.0	32.3	6.2
	ByMI-GD	8.5	100.0	100.0	24.2	4.7	9.1	100.0	100.0	32.1	6.4
	ByMI-GEOM	20.7	95.8	95.8	27.8	11.3	36.0	92.8	92.8	45.4	22.7
	ByMI-VRMOM	26.2	98.0	98.0	30.7	11.8	41.6	95.8	95.8	51.4	23.2
	ByMI-MOM	35.0	98.6	98.6	35.8	14.4	57.1	98.8	98.8	70.8	31.2
	RMDP-BH	77.7	62.6	62.6	71.5	85.1	85.4	38.6	38.6	74.2	141.3
	Krum	51.7	59.3	47.8	27.0	41.3	53.5	57.7	48.6	36.0	61.0
	FABA	81.7	22.4	15.6	27.0	74.8	94.5	6.9	5.4	36.0	128.3
	Zeno	76.5	28.9	16.2	27.0	67.5	88.2	14.6	5.6	36.0	114.4
CIFAR10	ByMI-Filter	7.4	100.0	100.0	19.6	5.4	6.5	100.0	100.0	25.9	6.3
	ByMI-Filter+	9.1	100.0	100.0	20.0	6.0	8.1	100.0	100.0	26.3	7.1
	ByMI-GD	7.5	100.0	100.0	19.6	5.4	6.6	100.0	100.0	25.9	6.5
	ByMI-GEOM	22.4	96.8	96.8	23.6	15.2	40.4	96.4	96.4	42.3	30.6
	ByMI-VRMOM	23.4	98.0	98.0	24.2	14.8	36.8	98.4	98.4	39.9	25.9
	ByMI-MOM	41.3	99.4	99.4	33.4	22.8	64.1	99.8	99.8	70.2	53.3
	RMDP-BH	86.0	2.0	2.0	18.8	109.4	89.2	0.2	0.2	19.9	146.9
	Krum	74.5	32.5	22.8	23.0	79.2	88.4	14.5	11.6	30.0	144.3
	FABA	98.3	2.1	1.8	23.0	112.0	100.0	0.0	0.0	30.0	162.0
	Zeno	97.1	3.7	1.8	23.0	108.7	99.9	0.1	0.0	30.0	159.7

Table 7. FDR(%), TPR(%) and P_a(%) of the random gradient attack with different contamination levels.

		MNIST			F-MNIST			CIFAR10		
	Method	FDR	TPR	P_a	FDR	TPR	P_a	FDR	TPR	P_a
$\varrho = 0.1$	ByMI-Filter	6.0	100.0	100.0	6.9	99.4	99.2	5.6	100.0	100.0
	ByMI-Filter+	6.6	98.0	98.0	6.4	82.6	82.4	7.1	100.0	100.0
	ByMI-GD	6.5	99.8	99.8	7.2	99.0	99.0	6.0	100.0	100.0
	ByMI-GEOM	9.2	99.0	98.8	10.0	98.2	98.0	9.9	100.0	100.0
	ByMI-VRMOM	11.9	99.1	99.0	12.6	98.4	98.4	10.7	100.0	100.0
	ByMI-MOM	13.5	99.2	99.2	14.7	98.4	98.4	15.5	100.0	100.0
	RMDP-BH	90.3	94.6	94.6	90.2	45.4	44.8	82.5	10.5	9.8
	Krum	87.8	15.7	5.8	89.7	13.3	4.2	82.5	21.9	8.6
	FABA	100.0	0.1	0.0	99.3	0.9	0.0	99.9	0.1	0.0
	Zeno	99.8	0.3	0.0	98.9	1.4	0.0	99.6	0.5	0.0
$\varrho = 0.15$	ByMI-Filter	6.8	100.0	100.0	7.1	100.0	100.0	7.3	100.0	100.0
	ByMI-Filter+	8.6	100.0	100.0	8.9	99.6	99.6	10.3	100.0	100.0
	ByMI-GD	7.3	100.0	100.0	7.9	100.0	100.0	7.5	100.0	100.0
	ByMI-GEOM	18.6	98.4	98.4	17.8	95.0	94.8	20.5	99.4	99.2
	ByMI-VRMOM	24.7	99.8	99.8	21.7	98.0	98.0	21.2	99.4	99.4
	ByMI-MOM	29.8	100.0	100.0	29.4	98.2	98.2	36.5	99.8	99.8
	RMDP-BH	87.2	83.2	82.8	91.5	21.2	20.6	85.7	0.8	0.6
	Krum	81.0	23.4	4.0	84.3	19.3	3.0	73.2	34.3	19.0
	FABA	100.0	0.0	0.0	99.9	0.2	0.0	100.0	0.0	0.0
	Zeno	100.0	0.0	0.0	99.6	0.5	0.0	99.8	0.2	0.0
$\varrho = 0.2$	ByMI-Filter	7.4	100.0	100.0	7.8	99.8	99.8	6.7	100.0	100.0
	ByMI-Filter+	9.7	100.0	100.0	10.3	100.0	100.0	8.7	100.0	100.0
	ByMI-GD	7.9	100.0	100.0	8.5	99.6	99.6	7.0	100.0	100.0
	ByMI-GEOM	34.2	94.2	94.2	31.7	88.8	88.4	38.6	99.2	99.0
	ByMI-VRMOM	42.3	98.5	98.2	35.0	96.1	95.8	34.1	99.8	99.8
	ByMI-MOM	52.3	99.2	99.2	50.3	98.6	98.6	59.5	100.0	100.0
	RMDP-BH	86.0	68.2	67.4	95.2	5.8	5.6	89.2	0.0	0.0
	Krum	88.4	14.4	7.6	87.5	15.5	5.8	94.9	6.4	4.0
	FABA	100.0	0.0	0.0	100.0	0.0	0.0	100.0	0.0	0.0
	Zeno	100.0	0.0	0.0	100.0	0.0	0.0	100.0	0.0	0.0

Table 8. FDR(%), TPR(%) and $P_a(\%)$ of the OOD attack with larger contamination levels.

		MNIST			F-MNIST			CIFAR10		
	Method	FDR	TPR	P_a	FDR	TPR	P_a	FDR	TPR	P_a
$\varrho = 0.3$	ByMI-Filter	8.6	98.4	91.4	9.0	92.5	76.8	8.2	84.2	66.6
	ByMI-Filter+	8.9	100.0	100.0	9.7	99.6	99.2	9.9	98.8	97.4
	ByMI-GD	9.0	98.5	92.0	9.4	92.8	75.6	8.3	84.9	67.0
	ByMI-GEOM	51.9	95.5	79.6	43.4	84.0	56.8	40.3	69.9	41.4
	ByMI-VRMOM	45.0	96.4	83.2	36.4	86.7	62.4	32.8	73.8	46.8
	ByMI-MOM	45.2	96.5	83.4	37.0	87.3	63.6	34.7	75.7	49.8
	RMDP-BH	67.1	99.8	98.6	51.3	93.1	83.6	62.8	56.9	38.8
	Krum	24.9	92.1	82.6	39.7	74.0	59.8	76.8	28.3	25.4
	FABA	24.8	92.3	89.2	42.2	70.9	63.2	73.3	32.5	29.6
	Zeno	31.4	84.2	40.8	47.7	64.2	12.0	74.8	30.6	5.0
$\varrho = 0.35$	ByMI-Filter	10.0	98.0	89.6	10.3	91.8	71.8	9.2	81.8	57.4
	ByMI-Filter+	9.5	100.0	99.8	9.6	99.8	99.2	9.3	99.1	98.0
	ByMI-GD	9.4	98.3	91.4	10.1	92.8	72.2	8.7	84.5	63.0
	ByMI-GEOM	56.0	94.6	74.2	49.7	83.6	51.0	44.6	67.4	37.4
	ByMI-VRMOM	51.2	95.6	77.6	44.0	85.7	54.4	38.9	70.3	41.6
	ByMI-MOM	51.4	95.7	77.8	44.8	86.6	56.4	41.7	72.9	43.4
	RMDP-BH	62.2	99.4	96.0	50.5	88.4	69.6	71.6	47.1	27.8
	Krum	36.5	77.0	75.2	61.7	46.4	45.6	88.6	14.0	14.0
	FABA	31.6	82.9	80.8	58.1	50.7	47.8	81.7	22.5	22.0
	Zeno	38.2	74.9	15.0	55.9	53.4	4.4	80.6	23.9	2.2
$\varrho = 0.4$	ByMI-Filter	12.4	98.6	87.2	12.8	88.6	64.0	11.4	78.7	53.0
	ByMI-Filter+	9.3	100.0	100.0	10.2	99.9	99.8	9.8	99.5	98.0
	ByMI-GD	9.4	99.3	92.2	10.4	91.2	72.0	9.1	83.5	62.6
	ByMI-GEOM	56.7	96.0	73.0	52.6	82.4	48.6	47.9	69.7	38.0
	ByMI-VRMOM	54.0	96.6	77.6	48.5	84.1	53.0	44.1	71.7	42.0
	ByMI-MOM	54.1	96.7	78.0	49.2	84.5	53.6	45.1	73.5	44.4
	RMDP-BH	57.8	99.5	96.0	54.8	83.2	55.0	75.3	44.4	20.4
	Krum	57.7	51.6	51.6	87.7	15.0	15.0	98.2	2.2	2.2
	FABA	42.6	70.1	69.2	76.2	29.1	27.6	91.7	10.2	10.0
	Zeno	44.5	67.7	5.6	64.4	43.4	0.4	85.8	17.4	0.0
$\varrho = 0.45$	ByMI-Filter	23.0	95.7	76.0	23.3	86.5	49.4	16.1	72.3	42.0
	ByMI-Filter+	9.7	100.0	100.0	10.2	99.6	99.4	12.2	97.5	96.0
	ByMI-GD	14.3	98.5	88.8	12.8	93.6	69.8	10.1	82.5	61.2
	ByMI-GEOM	53.7	93.7	51.4	51.7	83.0	26.4	49.5	68.4	25.4
	ByMI-VRMOM	52.3	95.0	66.6	49.8	85.8	40.4	46.0	71.2	36.8
	ByMI-MOM	52.4	95.2	66.8	50.3	86.5	40.6	48.3	73.0	38.2
	RMDP-BH	53.8	98.9	90.4	61.7	76.6	42.2	81.6	35.6	13.2
	Krum	99.8	0.2	0.2	100.0	0.0	0.0	100.0	0.0	0.0
	FABA	68.6	38.0	37.2	95.3	5.7	5.2	97.3	3.3	3.2
	Zeno	59.5	48.9	0.0	73.7	31.7	0.0	89.0	13.6	0.0

Table 9. FDR(%), TPR(%) and P_a(%) of the IPM attack with larger contamination levels.

		MNIST			F-MNIST			CIFAR10		
	Method	FDR	TPR	P_a	FDR	TPR	P_a	FDR	TPR	P_a
$\varrho = 0.3$	ByMI-Filter	9.0	99.8	99.8	9.1	100.0	100.0	8.1	100.0	100.0
	ByMI-Filter+	8.9	100.0	100.0	9.4	100.0	100.0	9.4	100.0	100.0
	ByMI-GD	8.7	100.0	100.0	9.0	100.0	100.0	8.0	99.8	99.8
	ByMI-GEOM	66.6	99.0	99.0	65.9	98.4	98.4	67.8	99.6	99.6
	ByMI-VRMOM	65.5	99.6	99.6	62.0	98.0	98.0	57.3	97.6	97.6
	ByMI-MOM	68.7	100.0	100.0	68.3	100.0	100.0	68.9	100.0	100.0
	RMDP-BH	95.4	15.4	15.4	96.8	7.8	7.8	97.2	0.0	0.0
	Krum	100.0	0.0	0.0	94.3	7.0	7.0	100.0	0.0	0.0
	FABA	100.0	0.0	0.0	99.5	0.6	0.6	100.0	0.0	0.0
	Zeno	100.0	0.0	0.0	98.3	2.1	0.6	100.0	0.0	0.0
$\varrho = 0.35$	ByMI-Filter	9.9	99.4	99.4	10.1	99.0	99.0	9.4	100.0	100.0
	ByMI-Filter+	9.2	100.0	100.0	9.9	100.0	100.0	9.9	100.0	100.0
	ByMI-GD	8.7	100.0	100.0	9.3	99.6	99.6	8.3	100.0	100.0
	ByMI-GEOM	64.2	100.0	100.0	63.9	100.0	100.0	64.8	100.0	100.0
	ByMI-VRMOM	62.8	99.8	99.8	60.9	99.4	99.4	58.3	99.0	99.0
	ByMI-MOM	64.3	100.0	100.0	64.2	100.0	100.0	64.8	100.0	100.0
	RMDP-BH	99.2	2.4	2.4	98.7	2.4	2.4	99.2	0.0	0.0
	Krum	100.0	0.0	0.0	99.7	0.4	0.4	100.0	0.0	0.0
	FABA	100.0	0.0	0.0	100.0	0.0	0.0	100.0	0.0	0.0
	Zeno	100.0	0.0	0.0	99.6	0.4	0.0	100.0	0.0	0.0
$\varrho = 0.4$	ByMI-Filter	11.9	98.6	98.6	11.5	98.4	98.4	11.0	99.8	99.8
	ByMI-Filter+	9.7	100.0	100.0	9.8	100.0	100.0	9.6	100.0	100.0
	ByMI-GD	9.7	100.0	100.0	9.3	99.6	99.6	8.3	100.0	100.0
	ByMI-GEOM	60.3	100.0	100.0	60.2	100.0	100.0	60.4	100.0	100.0
	ByMI-VRMOM	58.7	100.0	100.0	57.4	99.6	99.6	55.8	99.6	99.6
	ByMI-MOM	60.0	100.0	100.0	59.9	100.0	100.0	60.2	100.0	100.0
	RMDP-BH	99.7	0.8	0.8	100.0	0.0	0.0	100.0	0.0	0.0
	Krum	100.0	0.0	0.0	100.0	0.0	0.0	100.0	0.0	0.0
	FABA	100.0	0.0	0.0	100.0	0.0	0.0	100.0	0.0	0.0
	Zeno	100.0	0.0	0.0	100.0	0.0	0.0	100.0	0.0	0.0
$\varrho = 0.45$	ByMI-Filter	18.7	88.4	88.4	18.1	89.0	89.0	14.4	98.2	98.2
	ByMI-Filter+	10.0	100.0	100.0	9.8	100.0	100.0	10.1	100.0	100.0
	ByMI-GD	9.6	100.0	100.0	9.9	99.8	99.8	8.8	99.8	99.8
	ByMI-GEOM	54.9	100.0	100.0	54.8	100.0	100.0	55.5	100.0	100.0
	ByMI-VRMOM	53.6	100.0	100.0	52.6	99.6	99.6	51.5	99.4	99.4
	ByMI-MOM	54.8	100.0	100.0	54.7	100.0	100.0	55.5	100.0	100.0
	RMDP-BH	99.6	0.8	0.8	100.0	0.0	0.0	100.0	0.0	0.0
	Krum	100.0	0.0	0.0	100.0	0.0	0.0	100.0	0.0	0.0
	FABA	100.0	0.0	0.0	100.0	0.0	0.0	100.0	0.0	0.0
	Zeno	100.0	0.0	0.0	100.0	0.0	0.0	100.0	0.0	0.0

Table 10. FDR(%), TPR(%) and P_a(%) of the random gradient attack with larger contamination levels.

		MNIST			F-MNIST			CIFAR10		
	Method	FDR	TPR	P_a	FDR	TPR	P_a	FDR	TPR	P_a
$\varrho = 0.3$	ByMI-Filter	8.0	100.0	100.0	8.1	100.0	100.0	7.8	100.0	100.0
	ByMI-Filter+	9.9	100.0	100.0	9.8	100.0	100.0	10.1	100.0	100.0
	ByMI-GD	8.5	100.0	100.0	8.6	99.8	99.8	8.0	100.0	100.0
	ByMI-GEOM	66.3	98.8	98.8	67.0	93.6	93.6	67.4	100.0	100.0
	ByMI-VRMOM	64.2	98.8	98.8	58.5	91.0	90.8	55.8	99.4	99.0
	ByMI-MOM	67.7	100.0	100.0	67.3	99.6	99.6	68.6	100.0	100.0
	RMDP-BH	95.4	15.6	14.6	97.9	0.2	0.2	97.8	0.0	0.0
	Krum	100.0	0.0	0.0	100.0	0.0	0.0	100.0	0.0	0.0
	FABA	100.0	0.0	0.0	100.0	0.0	0.0	100.0	0.0	0.0
	Zeno	100.0	0.0	0.0	100.0	0.0	0.0	100.0	0.0	0.0
$\varrho = 0.35$	ByMI-Filter	8.8	100.0	100.0	8.9	99.6	99.6	7.4	100.0	100.0
	ByMI-Filter+	10.0	100.0	100.0	10.0	100.0	100.0	9.9	100.0	100.0
	ByMI-GD	8.5	100.0	100.0	8.7	99.8	99.8	7.2	100.0	100.0
	ByMI-GEOM	64.1	100.0	100.0	63.9	100.0	100.0	64.7	100.0	100.0
	ByMI-VRMOM	62.6	99.2	99.2	60.2	96.6	96.0	57.2	99.0	99.0
	ByMI-MOM	63.9	100.0	100.0	63.7	100.0	100.0	64.6	100.0	100.0
	RMDP-BH	98.2	5.2	4.6	99.6	0.0	0.0	99.6	0.0	0.0
	Krum	100.0	0.0	0.0	100.0	0.0	0.0	100.0	0.0	0.0
	FABA	100.0	0.0	0.0	100.0	0.0	0.0	100.0	0.0	0.0
	Zeno	100.0	0.0	0.0	100.0	0.0	0.0	100.0	0.0	0.0
$\varrho = 0.4$	ByMI-Filter	9.9	99.2	99.0	10.4	96.8	96.6	9.4	99.8	99.8
	ByMI-Filter+	10.1	100.0	100.0	10.0	100.0	100.0	10.1	100.0	100.0
	ByMI-GD	8.4	100.0	100.0	9.0	100.0	100.0	7.9	100.0	100.0
	ByMI-GEOM	60.1	100.0	100.0	60.0	100.0	100.0	60.3	100.0	100.0
	ByMI-VRMOM	58.5	100.0	100.0	57.4	98.2	98.2	55.5	99.4	99.2
	ByMI-MOM	59.7	100.0	100.0	59.6	100.0	100.0	60.1	100.0	100.0
	RMDP-BH	99.5	1.2	1.2	99.8	0.0	0.0	100.0	0.0	0.0
	Krum	100.0	0.0	0.0	100.0	0.0	0.0	100.0	0.0	0.0
	FABA	100.0	0.0	0.0	100.0	0.0	0.0	100.0	0.0	0.0
	Zeno	100.0	0.0	0.0	100.0	0.0	0.0	100.0	0.0	0.0
$\varrho = 0.45$	ByMI-Filter	11.6	97.6	97.4	12.8	93.2	92.8	10.9	99.2	99.2
	ByMI-Filter+	10.1	100.0	100.0	10.0	100.0	100.0	10.0	100.0	100.0
	ByMI-GD	9.0	100.0	100.0	9.4	99.8	99.8	8.3	100.0	100.0
	ByMI-GEOM	54.7	100.0	100.0	54.6	100.0	100.0	55.5	100.0	100.0
	ByMI-VRMOM	53.6	100.0	100.0	52.6	99.4	99.4	52.3	99.1	98.8
	ByMI-MOM	54.6	100.0	100.0	54.5	100.0	100.0	55.4	100.0	100.0
	RMDP-BH	99.8	0.4	0.4	100.0	0.0	0.0	100.0	0.0	0.0
	Krum	100.0	0.0	0.0	100.0	0.0	0.0	100.0	0.0	0.0
	FABA	100.0	0.0	0.0	100.0	0.0	0.0	100.0	0.0	0.0
	Zeno	100.0	0.0	0.0	100.0	0.0	0.0	100.0	0.0	0.0