Faster Convergence and More Accurate Models: How Anomaly Detection and Exclusion Enhances Federated Learning

Anonymous Author(s)
Affiliation
Address
email

Abstract

This paper investigates how principled anomaly detection in Federated Learning (FL) can make the learning more efficient and effective, while also addressing the traditional trade-off between system robustness and performance. We prove theoretically and verify on practical cases that detecting and removing anomalous models, whether due to adversarial behavior or data corruption, benefits learning efficiency by boosting convergence and producing more accurate models. Specifically, we conduct a theoretical investigation of FL convergence with and without defenses that detect anomalous models and exclude corresponding clients, which proves that removing clients supplying anomalous updates in conventional FL algorithms results in faster convergence. While state-of-the-art anomaly detection mechanisms typically introduce an additional quadratic computational overhead, we reduce the anomaly detection computational complexity by introducing a novel Proportional-Integral–Derivative-inspired Model Anomaly Detection and Exclusion (PID-MADE) algorithm for not only detecting anomalous clients but also excluding them from the training process. This approach complements standard aggregation strategies, such as FedAvg, preserving the original linear time complexity. Empirical evaluation on several benchmark datasets confirms that our method, combined with standard FL aggregation, not only improves security by effectively identifying and removing anomalous clients but also enhances learning efficiency compared to state-of-the-art approaches. The results emphasize that anomaly detection measures in FL, which improve security and privacy protection, can coexist with, and even enhance learning efficiency, providing a more effective framework for federated model training.

1 Introduction

Federated Learning (FL) is a decentralized Machine Learning (ML) paradigm that enables multiple clients to collaboratively train a shared model without exposing their private data (1). FL is becoming indispensable in modern ML privacy-sensitive applications in various domains, such as healthcare and finance, where data sharing is restricted due to ethical and legal constraints. However, FL systems are prone to learning efficiency degradation due to anomalous client updates, which can occur because of malicious actions or data corruption.(2; 3). Introducing anomaly detection and exclusion mechanisms imposes computational and communication overhead, potentially slowing down the model's convergence. For developers, balancing the trade-off between exclusion effectiveness, learning efficiency and fairness is particularly challenging when transitioning to production environments, where both system's security and integrity and learning efficiency are critical.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

Algorithm	Computation	#MC Estimation?	Temporal Adaptivity?
PID-MADE (Ours)	$\boldsymbol{O(nd)}$	**No**	**Yes**
Krum	$O(n^2d)$	Yes	No
Multi-Krum	$O(n^2d)$	Yes	No
Bulyan	$O(n^2d)$	Yes	No
RFA	$O(n^2d)$	No	No

Table 1: Comparison of our PID-inspired approach with previous work in terms of the computational burden and requirement on the prior knowledge of the number of malicious clients (#MC) in the system

Several defense mechanisms have been proposed to detect and exclude anomalous client updates, typically by analyzing gradients or model updates using distance-based metrics, but these methods often incur significant overhead, risk excluding benign clients, and rely on prior estimates of the number of anomalies, which are rarely accurate in practice (4; 5; 6).

To address these challenges, we propose a novel approach inspired by the Proportional-Integral-Derivative (PID) control concept, widely used in engineering, to enhance the security and performance of FL systems. As the approach is targeted at Model Anomaly Detection and Exclusion (MADE), we name it PID-MADE. PID-MADE introduces a scoring mechanism that incorporates both current and historical client behavior, using a distance metric to compute PID-based scores for client updates. Clients whose scores exceed a threshold are flagged as anomalous and excluded from the aggregation process, which results in reducing computational costs, training time, and improving the learning effectiveness and efficiency of the FL system.

In this paper, we present the following contributions. **(1)** We provide a formal theoretical analysis demonstrating that any FL algorithm incorporating MADE provably converges. **(2)** We further provide theoretical and empirical evidence demonstrating that FL with MADE converges faster than undefended FL in the presence of anomalies. **(3)** We introduce PID-MADE, a novel FL anomalies detection mechanism designed to reduce computational burdens while incorporating temporal information about clients. A key advantage of PID-MADE is that it does not require the estimated number of malicious clients, unlike methods such as Krum and its derivatives. We provide a formal analysis demonstrating PID-MADE's linear computational complexity and provide theoretically-grounded recommendations on setting anomaly detection thresholds for FL. Experimental results demonstrate that PID-MADE performs on par with, and in some cases outperforms, state-of-the-art defenses in terms of learning efficiency while growing linear in time. **(4)** To facilitate broader adoption and further research, we implement our approach as a software framework, which we made available to the public[1] to benefit the academic and professional communities. Together, these contributions demonstrate that enhancing FL with MADE systems can simultaneously improve learning efficiency and effectiveness, providing an improved framework for federated model training.

2 Related Work

Anomalies in FL generally can fall into two categories: malicious, such as (7; 8; 9; 10; 11; 12; 13; 14; 15), and technological occurring due to inherent noise in data. A variety of defense and anomaly detection mechanisms have been introduced. Similarity-based defenses, such as (4), focus on identifying robust aggregation strategies to reduce the influence of outlier updates. Privacy-preserving methods such as differential privacy (DP) (16; 17) and homomorphic encryption (18) provide guarantees for protecting client data but introduce computational and communication overhead. Secure aggregation and secure multi-party computation (SMPC) (19; 20) enable encrypted communication and ensure the integrity of the aggregated updates. ELSA (21), RoFL (22), or Prio (23) aim to improve FL security holistically by integrating secure aggregation mechanisms, cryptographic techniques, protocol-level modifications, and using multiple-server architectures. While effective, these approaches typically require substantial changes to the FL workflow and incur significant computational costs, thus not considered in this work.

[1] https://drive.google.com/file/d/1VSTeE6ynMPQcnGUu_nIZO0_mkQdni8DH/view?usp=drive_link (anonymized)

Symbol	Description
i	Client index
t	Communication round
w_t^i	Set of weights sent by client i to the server at round t
μ_t	Model parameters centroid computed by the server at round t
d_t^i	Euclidean distance of the i-th client model from the centroid at round t
$\mathcal{A}$	Set of all clients participating in the learning process
$\mathcal{G} \subset \mathcal{A}$	Subset of "good" clients not excluded during learning
$w_t^{\mathcal{A}}$	Set of weights for all clients in $\mathcal{A}$ at round t: $\mathcal{A} = \{w_t^{i_1}, w_t^{i_2}, \ldots, w_t^{i_{\lvert\mathcal{A}\rvert}}\}$
$w_t^{\mathcal{G}}$	Set of weights for all clients in $\mathcal{G}$ at round t: $\mathcal{G} = \{w_t^{i_1}, w_t^{i_2}, \ldots, w_t^{i_{\lvert\mathcal{G}\rvert}}\}$
w^*	Optimal model parameters that minimize the loss function
$\mu_t^{\mathcal{G}}$	Centroid of the "good" models at round t
$\mu_t^{\mathcal{A}}$	Centroid of all models at round t

Figure 1: Notation used in the paper

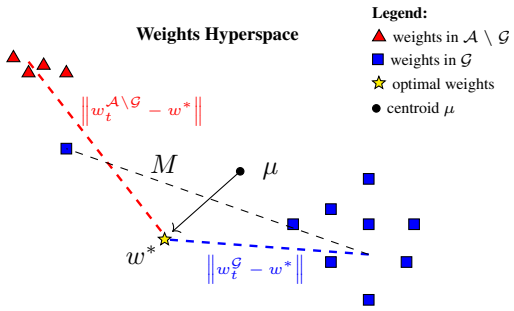


Figure 2: Visualization of definition (1) in the hyperspace of model weights. M is the margin that allows us to include some of the weights that are further away from the majority of good weights

We compare our method against state-of-the-art FL defenses, such as **Krum** and **Multi-Krum** (5), **Bulyan** (4), and **RFA** (6). Krum and Multi-Krum use geometric distances to select updates that are closest to others in the parameter space, making them robust against malicious clients. However, they require prior knowledge of the maximum number of malicious clients, which is impractical in real-world scenarios, and both have a computational complexity of $O(n^2 d)$, where n is the number of clients and d is the gradient space's dimension. Bulyan extends these methods with additional filtering steps for improved robustness but inherits the same computational complexity. RFA employs the geometric median instead of the mean to aggregate updates. Unlike Krum-based methods, RFA does not require an estimation of malicious clients. Computing the geometric median, which involves iterative algorithms like Weiszfeld's with a complexity of $O(n^2 d)$, is computationally intensive, making RFA less scalable in real-world deployments.

Our PID-MADE approach addresses these limitations by improving the computational complexity and eliminating the need for prior knowledge of the number of malicious clients. It incorporates an integral component accounting for the historical behavior of client updates for more reliable evaluations over time. Unlike previous methods that assess clients in isolation per round, PID-MADE leverages temporal adaptivity to detect persistent anomalies across training rounds, thereby improving robustness to the impacts of these anomalies. Furthermore, our approach enhances security while improving learning efficiency, accelerating model training compared to the baseline methods. We formally analyze the computational complexity and convergence of our approach in Sec. 3. In Table 1 we compare our approach with Krum, Multi-Krum, Bulyan, and RFA in terms of computational complexity, and the need for prior knowledge used as an input to the system.

3 FL Defense Formal Analysis

In this section, we present a theoretical analysis to evaluate the feasibility of enhancing FL learning efficiency by excluding anomalous clients from the aggregation process. Our approach involves formulating lemmas and theorems, and analyzing the convergence of the FL process with and without clients sending anomalous updates. First, we establish that model anomaly detection and exclusion accelerates convergence, and then we quantify the acceleration. Figure 1 summarizes the notation and terminology used throughout the paper. Due to space limitations we present all the proofs in the Appendix.

We propose FL defense that aims to identify and separate anomalous clients ("bad") from benign ones ("good") to prevent biased model updates that hinder convergence. This is achieved by analyzing the distribution of model updates and removing outliers based on their distance from the centroid position. This can be described by the following definition of anomalous model weights.

Definition 1 (Anomalous Model Weights): We say that weights submitted by a client are anomalous if they satisfy the following *separation condition*: Assume some training round t. The minimal

distance between the aggregated anomalous client model updates ($w_t^{\mathcal{A}\setminus\mathcal{G}}$) and the optimal model (w^*) must be greater than the maximum distance between the aggregated "good" client model updates ($w_t^{\mathcal{G}}$) and the optimal model, plus the margin M: $\min_{\mathcal{A}\setminus\mathcal{G}} \left\| w_t^{\mathcal{A}\setminus\mathcal{G}} - w^* \right\| > \max_{\mathcal{G}} \left\| w_t^{\mathcal{G}} - w^* \right\| + M$, where M is a sensitivity margin for outliers. Figure 2 illustrates **Definition 1**. As M increases, more outlier clients may join $\mathcal{G}$. The centroid μ_t is used as a proxy for the optimal model, with malicious clients (red) further from it than benign ones (blue). The defense works as long as $|\mathcal{G}| > |\mathcal{A} \setminus \mathcal{G}|$, i.e. the honest majority persists.

Criterion 1 (Anomaly Signature in FL): In real-world FL deployments, some client updates may deviate drastically from the benign population because of an attacker's poisoned data or simply corrupted measurements. We treat any such persistently "outlying" update as an anomaly. Formally, we say an anomaly in FL satisfies: $\forall \varepsilon > 0, \nexists N \in \mathbb{N}$ s.t. $\forall t \geq N, \left\| w_t^{\mathcal{A}} - w^* \right\| < \varepsilon$. In other words, no matter how small a tolerance ε we choose, there is no round after which some client i's updates remain within that tolerance of w^*. This criterion follows from **Definition 1.**

3.1 Convergence

Lemma 1 (Variance Reduction through Outlier Removal): Let $\{a_i\}$ be a set of points on a number line with scalar values where $a_i \in \mathbb{R}, i \in \mathbb{N}$ and $a_1 < a_2 < \ldots < a_N$, $N \geq 2$. We consider one of those points, a_N, an outlier point a_o, meaning that a_o satisfies $(a_o - \mu)^2 = \max_{1 \leq i \leq N}(a_i - \mu)^2$. Let us form a new set of points by simply removing a_o from the original set. Then, if σ^2 is the variance of the original set and σ'^2 is the variance of the new set, we have that $\sigma'^2 \leq \sigma^2$.

Theorem 1 (Convergence Preservation under Anomalous Model Exclusion): Consider global models $m^{\mathcal{A}}$ composed by the aggregation of all local models $w_t^{\mathcal{A}}$ and $m^{\mathcal{G}}$ composed by the aggregation of models after exclusion $w_t^{\mathcal{G}}$ through FedAvg. If $\forall \varepsilon > 0, \exists N_1 \in \mathbb{N}$ s.t. $\forall t \geq N_1, \left\| w_t^{\mathcal{A}} - w^* \right\| < \varepsilon$, then $\exists N_2 \in \mathbb{N}$ s.t. $\forall t \geq N_2, \left\| w_t^{\mathcal{G}} - w^* \right\| < \varepsilon$. That is, assuming the original learning algorithm converges, an algorithm augmented with anomaly detection and exclusion also converges.

Commentary: removing anomalous clients from the FL aggregation does not violate the convergence of the original algorithm if it still converges even under the attacks or anomalies. If the original model $m^{\mathcal{A}}$ does converge, this implies that the attack is not strong enough, which in practice can occur due to various reasons, such as a low proportion of malicious clients or the attack goal was to make it converge to the wrong model (10). With the convergence of $m^{\mathcal{A}}$, we can guarantee that if the anomaly detection and exclusion is applied, $m^{\mathcal{G}}$ will always converge to the correct model and faster than $m^{\mathcal{A}}$, which is shown in the next part of the theorem. Furthermore, we make a stronger assumption that even if $m^{\mathcal{A}}$ does not converge, $m^{\mathcal{G}}$ will still converge. While we do not have a theoretical guarantee, this is suggested by the empirical evidence, which we present in Sec. 5.

Theorem 2 (Accelerated Convergence under Anomaly Exclusion): If N_1 is the round, on which the conventional FL with all clients (no clients removed) converges on $w_t^{\mathcal{A}}$, that is $\forall t \geq N_1, \left\| w_t^{\mathcal{A}} - w^* \right\| < \varepsilon$, and N_2 is the round, on which FL with good clients only (some bad clients are removed) converges on $w_t^{\mathcal{G}}$, that is $\forall t \geq N_2, \left\| w_t^{\mathcal{G}} - w^* \right\| < \varepsilon$, then $N_2 \leq N_1$.

Commentary: the implication of this theorem is that, when using only the updates from clients without outlier updates (as in $w_k^{\mathcal{G}}$), the convergence towards the optimal model will be faster than when aggregating updates from all clients, including those with outlier updates (as in $w_k^{\mathcal{A}}$). This is because the outlier updates, which may significantly deviate from the optimal model, distort the global model, causing it to remain far from an optimal solution for a longer period. In contrast, when outlier updates are excluded, the model converges faster to the optimal model, as the updates are more consistent and less prone to distortion. The result shows that the anomaly detection mechanism that excludes outlier updates from the aggregation improves FL convergence, thus accelerating the learning process in FL. In Sec. 5, we demonstrate our verification of the convergence on practical use cases. While the preceding theorems establish improved convergence, the next result shows a general upper-bound on this accelerated rate (**Theorem 3**).

Theorem 3 (Enhanced Convergence Rate under Anomaly Exclusion): The norm of the distances between good models' and optimal model weights is bounded by the distances between all models'

and optimal model weights, that is $\exists N \in \mathbb{N}$ s.t. $\forall t \geq N$, $\left\|w_t^{\mathcal{G}} - w^*\right\| \leq C\left\|w_t^{\mathcal{A}} - w^*\right\|$, where C is a constant if the number of malicious clients does not change during learning and $C = \sqrt{\frac{|\mathcal{G}|}{|\mathcal{A}|}} \leq 1$.

Proof Sketch: by removing clients sending anomalous updates to the server, we remove the outliers in the weights dimension, which also reduces the variance, as we show in **Lemma 1**. Comparing the variance of benign model weights around the centroid $\mu_t^{\mathcal{G}}$ and the variance of all model weights around $\mu_t^{\mathcal{A}}$, and further rewriting in vector notation using the Euclidean norm, the bound follows by taking the square root.

Commentary: the practical significance of this relationship is that the model with weights ${w_t}^{\mathcal{G}}$ at some round $t \geq N$ will converge quicker than $w_t^{\mathcal{A}}$, and the weights $w_t^{\mathcal{G}}$ will be $\sqrt{\frac{|\mathcal{A}|}{|\mathcal{G}|}}$ times closer to the optimal model than the weights $w_t^{\mathcal{A}}$. Moreover, as long as $|\mathcal{G}| > |\mathcal{A} \setminus \mathcal{G}|$, with the number of anomalous clients $|\mathcal{A} \setminus \mathcal{G}|$ growing, the model protected by our defense will converge quicker at a rate which depends on $|\mathcal{A} \setminus \mathcal{G}|$ if the number of anomalous clients increases during learning. Specifically, if we take the derivative of $\sqrt{\frac{|\mathcal{G}|}{|\mathcal{A}\setminus\mathcal{G}|+|\mathcal{G}|}}$ with respect to $\frac{|\mathcal{A}\setminus\mathcal{G}|}{|\mathcal{G}|}$, the convergence increases with a rate proportional to $\frac{1}{\sqrt{(\frac{|\mathcal{A}\setminus\mathcal{G}|}{|\mathcal{G}|}+1)^3}}$. This also means that as the number of anomalous clients grows, the gain in the convergence rate diminishes.

4 PID-MADE Approach

We develop a novel PID control-inspired algorithm to detect and exclude anomalous updates from FL aggregation. PID provides for a feedback mechanism with three components – proportional, integral, and derivative – widely used in automated control systems since its formalization by (24). The goal of the PID controller is to minimize the error value $e(t)$ over time by adjusting the control variable $c(t)$. The error is calculated as the difference between the setpoint and the control variable. The control function $c(t)$ is given by $c(t) = K_p e(t) + K_I \int_0^t e(\varphi) d\varphi + K_d \frac{de(t)}{dt}$, where $e(t)$ is the error value at time t, and the coefficients K_p, K_I, and K_d determine the weights of the proportional, integral, and derivative components.

In our approach, the proportional term reacts to instantaneous deviations, the integral term identifies persistent drifts by accounting for historical trends, and the derivative term anticipates future changes. These components together enable the effective detection of an abnormal client behavior. Our algorithm measures the error as the distance between a client's updates and the *centroid*. It aims at flagging and excluding significant deviations from the centroid. In the following, we describe how we adopt the PID principle for detecting anomalous clients in FL. The error value is calculated as the Euclidean distance of client i's model from the centroid μ_t of all submitted models: $D_t^{(i)}(w_t^{(i)}, \mu_t) = \|w_t^{(i)} - \mu_t\|$. Although we defined the centroid μ_t as both the mean $\frac{1}{N}\sum_{i=0}^{N} w_t^{(i)}$ and the geometric median: $\arg\min_y \sum_{i=0}^{N} \|w_t^{(i)} - y\|$ depending on the setting, we illustrate our findings in the following sections with the mean estimate. The PID score for each client i is calculated as:

$$u_t^{(i)} = \underbrace{D_t^{(i)}(w_t^{(i)}, \mu_t)}_{proportional} + \underbrace{K_I \sum_{x=0}^{t-1} D_x^{(i)}(w_x^{(i)}, \mu_x)}_{integral} + \underbrace{K_d(D_t^{(i)}(w_t^{(i)}, \mu_t) - D_{t-1}^{(i)}(w_{t-1}^{(i)}, \mu_{t-1}))}_{derivative}. \quad (1)$$

For each training round: (1) the server distributes the global model to the clients, (2) the clients train the model locally for a number of epochs and send it back to the server, (3) the server computes the centroid μ_t and our PID score as detailed in Formula 1 and excludes any clients above the threshold τ, derivation of which we describe in Sec. 4.1. The full mechanism integrated with FedAvg is summarized in Algorithm 2.

4.1 PID-MADE: Anomaly Detection and Exclusion Mechanism

We detect and exclude anomalies by calculating PID scores for each client based on the distance from μ_t and comparing them against the threshold τ, which is derived from the upper bound of PID scores for non-anomalous clients. To implement this mechanism, we address two major challenges:

(1) how to estimate the unknown optimal model and (2) how to estimate the threshold τ. Since the optimal model is unknown, we estimate it with the centroid $\mu_t = \frac{1}{N}\sum_{i=0}^{N} w_t^{(i)}$ at each round. This yields a biased estimate in highly non-IID settings, in which case the geometric median can be used instead for more robust estimation. To derive the threshold τ, we analyze the PID metric in Formula 1 and first provide a permissive upper bound which is free from assumptions, but leads to a high false negative rate. To improve this bound, we introduce specific assumptions which allow us to provide a tighter estimate of τ.

Theorem 4 (Permissive Upper Bound for Benign PID Scores): The permissive upper bound of the PID score for the good client is given by $t \cdot \left(\Delta_{max} + O(\frac{f}{N})\right)$, where Δ_{max} is the maximal deviation from the centroid, f is the number of anomalies, N is the number of all clients, and t is the number of training rounds. This overly permissive bound provides a zero false-positive rate, but may yield a high false-negative rate. Although impractical as a detection threshold, it serves as a useful baseline from which we derive tighter, more effective bound estimates.

In **Theorems 5** and **6**, we provide tighter and more practical upper bounds on PID scores of non-anomalous clients in a cross-device case. Before we introduce **Theorems 5** and **6**, we present **Lemma 2** and **Assumption 1** which are necessary for us to prove the theorems.

Lemma 2 (Bounded Centroid Shift): The centroid shift is bounded by $O\left(\frac{f}{N}\right)$ (see Appendix, proof of **Theorem 4**).

Let $\Delta_t = \|w_t^{(i)} - \mu_t\|$ be the deviation of client i's update from the centroid at any round t.

Assumption 1 (Uncorrelated Deviations): The sequence of random variables $\{\Delta_t\}_{t=0}^{T}$ satisfies $\forall 0 \leq t \neq y \leq T : \mathrm{Cov}(\Delta_t, \Delta_y) = 0$.

Although **Assumption 1** does not strictly hold in realistic federated settings, nonzero covariances $\mathrm{Cov}(\Delta_t, \Delta_y)$ can only increase the true variance of the PID score – meaning that the threshold we derive will be more conservative. Empirically, we observe that applying the threshold τ derived under **Assumption 1** sufficiently separates benign from anomalous clients. A fully rigorous threshold would account for each pairwise covariance term, however, estimating all $\mathrm{Cov}(\Delta_t, \Delta_y)$ online would impose significant overhead, and in practice the independence-based approximation already provides a tight, computationally efficient bound, which we derive in **Theorems 5** and **6**.

Theorem 5 (Chebyshev Threshold): Let us introduce the random variable U_t representing PID scores. Under **Assumption 1** and using **Lemma 2**, without knowing the distribution of PID scores, with probability at most α, the PID scores of good clients will be within $z\sigma_t$ of the sample average of PID scores $\bar{u}_t$, where σ_t is the standard deviation of PID scores at round t. Formally, $\Pr[U_t - \bar{u}_t \geq z\sigma_t] \leq \alpha$, where $\alpha = \frac{1}{\sqrt{z}}$ represents the desired alarm rate (i.e. false positive rate). With a probability of at least $1 - \alpha$ the benign clients will be under the threshold $\tau = \bar{u}_t + z\sigma_t$. Equivalently, no more than α-fraction of benign clients exceed τ. The threshold derivation follows Chebyshev's inequality (see the Appendix).

Theorem 6 (Gaussian Threshold): If we assume $\Delta_t \sim \mathcal{N}(\mu_\Delta, \sigma_\Delta^2)$, then the PID scores become also Gaussian, $U_t \sim \mathcal{N}(\bar{u}_t, \sigma_t^2)$. Then, the exact Gaussian threshold $\tau_{Gauss} = \bar{u}_t + z_{1-\alpha}\sigma_t$ ensures a false positive rate of α, where $z_{1-\alpha}$ is the z-score corresponding to desired α.

Theorems 5 and **6** give us an opportunity to efficiently select the threshold value based on the detection statistics we want to achieve in practice and to satisfy specification requirements set up in an application. To transfer this theoretical foundation into practice and filter out anomalous clients we compute the expected PID score as sample average $\bar{u}_t = \frac{1}{N}\sum_{i=0}^{N} u_i$. Any $u_t^{(i)}$ greater than $\bar{u}_t + \alpha\sigma_t$ is flagged as an anomaly and excluded from aggregation. As we show in Sec. 5, this empirical threshold estimation is effective even when Δ_t are not Gaussian, which is often the case in practice. The full algorithm is presented in Algorithm 2, where the input is the set of client models $\mathcal{A}$ and desired alarm rate α, and the output is the non-anomalous client set Q and the aggregated model. Unlike previous methods (5; 4; 6), our PID-based approach is adaptive and does not require prior knowledge of the number of malicious clients. The integral term accumulates historical deviations, making persistent attackers identifiable over time. Additionally, our method has a linear time complexity of $O(nd)$ which we prove in **Lemma 3**.

Algorithm 2 FedAvg with PID-MAD

Input: $\mathcal{A}$, set of clients with private local data, alarm rate α
Output: Q, aggregated global model
Input: $\mathcal{A}$, $|\mathcal{A}|$ clients with private local data
Output: Q, aggregated global model
Clients Execute
  receive global model from the server
  for each local epoch **do**
    execute training algorithm (e.g. SGD)
  end for
  push the local model w_t^i to the aggregation server
Server Executes
  $Q \leftarrow \mathcal{A}$
  for each round $t = 1, 2, \ldots$ **do**
    receive w_t^i from local clients
    compute $\mu_t, u_t^{(i)}, \bar{u}_t, \sigma_t$
    for each client $i \in Q$ **do**
      $Q \leftarrow Q \setminus \{w_t^{(i)} : u^{(i)} \leq \tau = \bar{u}(t) + \alpha\sigma_t\}$
    end for
    Perform aggregation of weights in Q based on FedAvg.
    Distribute aggregated global model back to the clients.
  end for

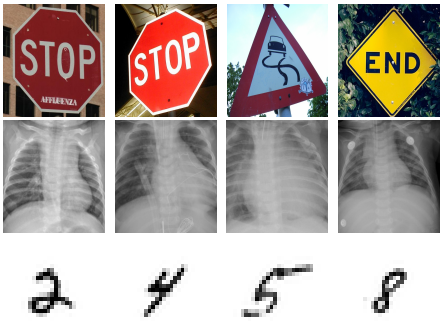


Figure 3: Example images from our study. Top: ITS; Middle: PneumoniaMNIST; Bottom: FEMNIST.

Lemma 3 (Computational Complexity): Algorithm 2 runs in $O(nd)$ time, where n refers to the number of clients and d is the dimension of the model parameter space. *Proof*: the computation of the centroid μ_t and of $u_t^{(i)}, \bar{u}_t, \sigma_t$ are linear $O(nd)$, keeping total complexity linear $O(nd)$.

5 Empirical Evaluation

Anomaly Model: based on **Criterion 1** of model anomalies, introduced in Sec. 3, we implement and evaluate untargeted data poisoning attacks, focusing on a practical case likely in real-world scenarios (10). Importantly, data poisoning serves as a proxy for a broader class of anomalies, capturing not only malicious behaviors but also inadvertent deviations arising from corrupted, mislabeled, or non-representative client data. Thus, our evaluation encompasses both adversarial and non-adversarial sources of model anomalies.

Emprical Study Setup: we evaluate our PID-based approach on three image datasets: Intelligent Transportation Systems (ITS), FEMNIST (25), and PneumoniaMNIST (26). From FEMNIST, we employ a numerical labels subset, and we use the entire PneumoniaMNIST. The ITS dataset consists of traffic sign images, with around 600 stop sign images and over 3,000 traffic sign images from the Open Images V6 dataset[2]. We poison data by flipping labels for certain clients. Each client uses 90% of their data for training and 10% for validation. Example images are shown in Figure 3. As a classifier, we employ a convolutional model with a sequence of convolutional, max pooling, and fully-connected layers, followed by dropout and ReLU activation functions. The final layer is a softmax with cross-entropy loss. As a FL framework, we employ Flower (27). The experiments run on a single node with varying numbers of clients. Our PID-MADE aggregation algorithm is implemented using Flower's API and we are making it available to the public (see footnote on page 2). In our experiments, we use the mean centroid computation for PID-MADE, where μ_t is the average of submitted local model updates, and set the alarm rate $\alpha = 2$. Our ablation study (see the Appendix) on selecting and fine-tuning the PID-MADE coefficients demonstrated that $K_I = 0.8$ and $K_d = 0.2$, selected heuristically, make anomalous clients more distinguishable from the "good" ones. The experiments were conducted on a system equipped with an AMD Ryzen 5 7600 CPU, 32 GB of RAM, and an NVIDIA RTX 4060TI GPU with 16 GB of dedicated memory, running the Ubuntu 22.04 OS.

5.1 Experimental Results

Figures 4(a), 4(b), and 4(c) present the loss function performance of various FL defense mechanisms across three datasets: FEMNIST (100 rounds, 20 clients, 2 with flipped labels acting as anomalies), PneumoniaMNIST (50 rounds, 8 clients, 2 anomalous), and ITS (50 rounds, 8 clients, 2 anomalous). On FEMNIST, while all methods, including the undefended FedAvg baseline, show rapid initial loss reduction, FedAvg plateaus and even slightly increases towards the end of training, indicating

[2]https://storage.googleapis.com/openimages/web/download.html

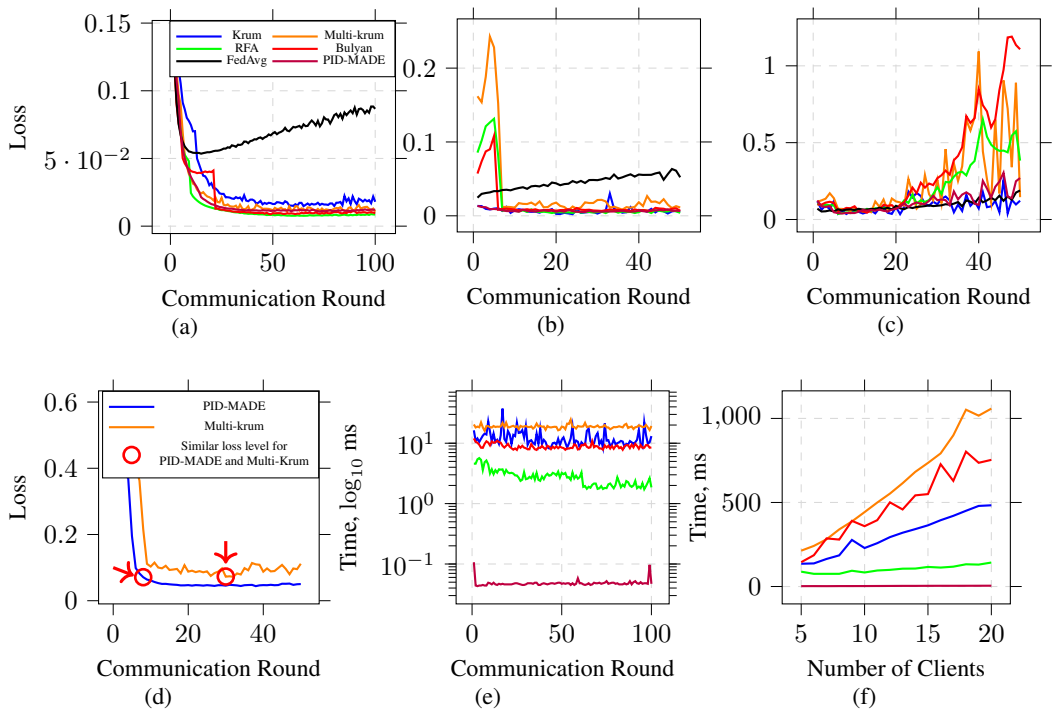


Figure 4: Empirical evaluation results: loss demonstrated by various aggregation strategies with and without defense mechanisms over (a) – FEMNIST (b) – PneumoniaMNIST, and (c) – ITS; (d) – comparison of learning efficiency and convergence demonstrated by PID-MADE and Multi-Krum on FEMNIST; (e) – changes in score calculation time taken by each defense method in the training process over FEMNIST; (f) – changes in score calculation with increasing number of clients for tested defense mechanisms over FEMNIST

convergence difficulties. In contrast, all defenses (Krum, Multi-Krum, Bulyan, RFA, and PID-MADE) achieve consistently lower loss and more stable convergence than FedAvg after the initial phase, demonstrating improved learning efficiency. On PneumoniaMNIST, a similar trend is observed: FedAvg plateaus at a higher loss compared to the defenses. Though some initial volatility is present, particularly with Multi-Krum, the defenses ultimately converge to significantly lower loss values, again showing improved learning efficiency. Highlighting the real-world complexities of the ITS dataset, PID-MADE demonstrates learning efficiency competitive to Krum and FedAvg, and better than other methods. ITS dataset includes a variety of traffic sign images, notably both traffic signs and stop signs that are visually very similar. Even with some slight deviation in its loss curve, PID-MADE maintains its robustness. This suggests PID-MADE's potential as an efficient anomaly detection and exclusion mechanism even in the cases with complex real-world data.

Figure 4(d) presents a comparison of the convergence behavior of the proposed PID-MADE mechanism against Multi-Krum. The experiment was conducted with 18 benign and 2 anomalous clients. The plot illustrates the values of the loss function over communication rounds for both methods. Initially, both PID-MADE and Multi-Krum show a rapid decrease in loss, indicating effective initial learning. However, a key difference emerges in their subsequent convergence behavior. PID-MADE achieves a significantly lower loss earlier in the training process compared to Multi-Krum. Specifically, PID-MADE reaches a stable, low loss within approximately 10 communication rounds, whereas Multi-Krum takes considerably longer to reach a comparable level of performance, requiring more than 20 rounds. The faster convergence of PID-MADE demonstrates a clear advantage in terms of learning efficiency. This results in reduced communication overhead and a faster overall training time, which is especially beneficial in resource-constrained FL settings.

In Figures 4(e) and 4(f) we demonstrate results on time which various defense algorithms require to assess the updates on the aggregation server. Here we employed FEMNIST data with 100 communication rounds and 20 clients, 2 of which were anomalous. Figure 4(e) illustrates the change in score calculation time for each defense mechanism across the 100 communication rounds. Our

Method	Avg FP	Avg FN	ΔFP	ΔFN	Recall	Precision	F1-score
MK f=2	0.14	2.14	-3.21	2.11	0.47	0.93	0.62
MK f=4	1.98	1.98	-1.37	1.95	0.51	0.51	0.51
MK f=7	4.63	1.63	1.28	1.60	0.59	0.34	0.43
MK f=9	6.45	1.45	3.10	1.42	0.64	0.28	0.39
PID-MADE	3.35	0.03	-	-	**0.72**	**0.56**	**0.72**

Table 2: Comparison of Detection Metrics Across Methods. 4 clients acting as anomalies.

PID-MADE consistently exhibits the lowest score calculation time throughout all rounds, remaining significantly lower (approximately 10^{-1} ms on the log scale, or roughly 0.1 ms in absolute terms) compared to all other defenses. Figure 4(f) presents the scaling behavior of the absolute score calculation time with an increasing number of clients. In comparison to all other evaluated defenses, our PID-MADE demonstrates a linear increase in calculation time as the number of clients grows. This suggests that all other defenses tested become more computationally expensive as the number of clients increases, making them less suitable for large-scale FL scenarios.

Table 2 highlights the classic precision–recall trade-off across methods on FEMNIST dataset and shows that our PID-based detector achieves the best overall balance. The Multi-Krum variants with small rejection budgets (e.g. $f = 2$) incur very few false positives (Avg FP=0.14) and yield high precision (0.93) but miss many anomalies (Avg FN=2.14, Recall=0.47), resulting in a moderate F1 of 0.62. As f increases, Multi-Krum flags more anomalies (Recall rises from 0.51 to 0.64) but at the cost of dramatically more false alarms (Avg FP up to 6.45) and sharply reduced precision (down to 0.28), dragging F1 scores below 0.50. By contrast, PID-MADE attains near-perfect detection (Avg FN ≈ 0, Recall = 0.72) with a moderate false-positive rate (Avg FP = 3.35), yielding both higher precision (0.56) and the highest F1-score (0.72). PID-MADE outperforms all Multi-Krum settings in achieving a superior trade-off between sensitivity and specificity.

6 Limitations

The choice of PID coefficients is critical and should be informed by application-specific considerations. We offer the following guidelines: if anomalous clients are expected to exhibit consistent deviations over time, increasing the coefficient of the integral term can help accumulate and amplify these persistent shifts. Conversely, if immediate discrepancies in submitted updates are more indicative of anomalies, assigning a higher weight to the proportional term enhances sensitivity to such deviations. We assume that benign updates share roughly similar deviation patterns. In highly non-IID environments, where legitimate clients' data distributions vary dramatically, the PID threshold can misclassify rare-but-valid updates as anomalies. In this work, we only theoretically analyzed the PID metric's behavior in a cross-device setting, but verified its practicality in the cross-silo setting. A deeper theoretical analysis with respect to tailoring the threshold to account for distribution heterogeneity and cross-silo setting is left for future work.

7 Conclusion

We demonstrated that augmenting FL with anomaly detection and exclusion improves learning efficiency. Our theoretical analysis provided a foundation for understanding how FL anomaly exclusion mechanisms contribute to faster convergence of the global model. We have shown theoretically and verified empirically that FL with defenses converges faster than conventional FL. As another key contribution, we introduced PID-MADE, a novel FL detection mechanism offering several key advantages over existing approaches. Notably, PID-MADE operates without requiring the estimate of expected anomalies, unlike other methods such as Krum and its derivatives, freeing users from specifying this potentially difficult-to-determine parameter in practice. Furthermore, PID-MADE's theoretical analysis demonstrated, and empirical validation confirmed, linear computational complexity while maintaining similar or even better learning efficiency, a critical factor for scalability in large-scale FL deployments. Finally, we also provided theoretically justified recommendations for threshold selection, which were verified empirically and demonstrated PID-MADE's superior performance against state-of-the-art methods.

References

[1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*, pp. 1273–1282, PMLR, 2017.

[2] X. Zhang, Y. Kang, K. Chen, L. Fan, and Q. Yang, "Trading off privacy, utility, and efficiency in federated learning," *ACM Transactions on Intelligent Systems and Technology*, vol. 14, no. 6, pp. 1–32, 2023.

[3] Z. Yan, D. Li, Z. Zhang, and J. He, "Accuracy–security tradeoff with balanced aggregation and artificial noise for wireless federated learning," *IEEE Internet of Things Journal*, vol. 10, no. 20, pp. 18154–18167, 2023.

[4] E. M. E. Mhamdi, R. Guerraoui, and S. Rouault, "The Hidden Vulnerability of Distributed Learning in Byzantium," July 2018. Issue: arXiv:1802.07927 arXiv: 1802.07927 [cs, stat].

[5] P. Blanchard, E. M. El Mhamdi, R. Guerraoui, and J. Stainer, "Machine learning with adversaries: Byzantine tolerant gradient descent," *Advances in neural information processing systems*, vol. 30, 2017.

[6] K. Pillutla, S. M. Kakade, and Z. Harchaoui, "Robust Aggregation for Federated Learning," *IEEE Transactions on Signal Processing*, vol. 70, pp. 1142–1154, 2022. arXiv: 1912.13445 [cs, stat].

[7] N. Rodríguez-Barroso, D. Jiménez-López, M. V. Luzón, F. Herrera, and E. Martínez-Cámara, "Survey on federated learning threats: Concepts, taxonomy on attacks and defences, experimental study and challenges," *Information Fusion*, vol. 90, pp. 148–173, 2023.

[8] V. Tolpegin, S. Truex, M. E. Gursoy, and L. Liu, "Data Poisoning Attacks Against Federated Learning Systems," Aug. 2020. arXiv:2007.08432 [cs].

[9] H. Wang, K. Sreenivasan, S. Rajput, H. Vishwakarma, S. Agarwal, J.-y. Sohn, K. Lee, and D. Papailiopoulos, "Attack of the tails: Yes, you really can backdoor federated learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 16070–16084, 2020.

[10] V. Shejwalkar, A. Houmansadr, P. Kairouz, and D. Ramage, "Back to the Drawing Board: A Critical Evaluation of Poisoning Attacks on Production Federated Learning," in *2022 IEEE Symposium on Security and Privacy (SP)*, pp. 1354–1371, IEEE, May 2022. Place: San Francisco, CA, USA.

[11] G. Baruch, M. Baruch, and Y. Goldberg, "A little is enough: Circumventing defenses for distributed learning," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[12] A. N. Bhagoji, S. Chakraborty, P. Mittal, and S. Calo, "Analyzing federated learning through an adversarial lens," in *International conference on machine learning*, pp. 634–643, PMLR, 2019.

[13] V. Shejwalkar and A. Houmansadr, "Manipulating the Byzantine: Optimizing Model Poisoning Attacks and Defenses for Federated Learning," in *Proceedings 2021 Network and Distributed System Security Symposium*, (Virtual), Internet Society, 2021.

[14] M. Fang, X. Cao, J. Jia, and N. Z. Gong, "Local Model Poisoning Attacks to Byzantine-Robust Federated Learning," Nov. 2021. Issue: arXiv:1911.11815 arXiv: 1911.11815 [cs].

[15] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, and V. Shmatikov, "How To Backdoor Federated Learning," Aug. 2019. Issue: arXiv:1807.00459 arXiv: 1807.00459.

[16] H. B. McMahan, D. Ramage, K. Talwar, and L. Zhang, "Learning differentially private recurrent language models," *arXiv preprint arXiv:1710.06963*, 2017.

[17] Z. Bu, J. Dong, Q. Long, and W. J. Su, "Deep learning with gaussian differential privacy," *Harvard data science review*, vol. 2020, no. 23, 2020.

[18] J. Ma, S.-A. Naas, S. Sigg, and X. Lyu, "Privacy-preserving federated learning based on multi-key homomorphic encryption," *International Journal of Intelligent Systems*, vol. 37, no. 9, pp. 5880–5901, 2022.

[19] P. Kairouz, Z. Liu, and T. Steinke, "The distributed discrete gaussian mechanism for federated learning with secure aggregation," in *International Conference on Machine Learning*, pp. 5201–5212, PMLR, 2021.

[20] Y. Li, T.-H. Chang, and C.-Y. Chi, "Secure federated averaging algorithm with differential privacy," in *2020 IEEE 30th international workshop on machine learning for signal processing (MLSP)*, pp. 1–6, IEEE, 2020.

[21] M. Rathee, C. Shen, S. Wagh, and R. A. Popa, "Elsa: Secure aggregation for federated learning with malicious actors," in *2023 IEEE Symposium on Security and Privacy (SP)*, pp. 1961–1979, IEEE, 2023.

[22] L. Burkhalter, H. Lycklama, A. Viand, N. Küchler, and A. Hithnawi, "Rofl: Attestable robustness for secure federated learning," *arXiv preprint arXiv:2107.03311*, vol. 21, 2021.

[23] H. Corrigan-Gibbs and D. Boneh, "Prio: Private, robust, and scalable computation of aggregate statistics," in *14th USENIX symposium on networked systems design and implementation (NSDI 17)*, pp. 259–282, 2017.

[24] N. Minorsky, "Directional stability of automatically steered bodies," *Journal of the American Society for Naval Engineers*, vol. 34, no. 2, pp. 280–309, 1922.

[25] S. Caldas, S. M. K. Duddu, P. Wu, T. Li, J. Konečnỳ, H. B. McMahan, V. Smith, and A. Talwalkar, "Leaf: A benchmark for federated settings," *arXiv preprint arXiv:1812.01097*, 2018.

[26] J. Yang, R. Shi, D. Wei, Z. Liu, L. Zhao, B. Ke, H. Pfister, and B. Ni, "Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification," *Scientific Data*, vol. 10, no. 1, p. 41, 2023.

[27] D. J. Beutel, T. Topal, A. Mathur, X. Qiu, J. Fernandez-Marques, Y. Gao, L. Sani, K. H. Li, T. Parcollet, P. P. B. de Gusmão, *et al.*, "Flower: A friendly federated learning research framework," *arXiv preprint arXiv:2007.14390*, 2020.

Faster Convergence and More Accurate Models: How Anomaly Detection and Exclusion Enhances Federated Learning

(Technical Appendices and Supplementary Materials)

Anonymous Author(s)
Affiliation
Address
email

1 Proofs

1.1 Proof of Criterion 1

Assume that $\forall \varepsilon > 0, \exists N \in \mathbb{N}$ s.t. $\forall t \geq N$, $\|w_t^{\mathcal{A}} - w^*\| < \varepsilon$. This means that by **Definition 1** $\min_{\mathcal{A}\setminus\mathcal{G}} \|w_t^{\mathcal{A}\setminus\mathcal{G}} - w^*\| < \varepsilon$, which can only happen when there are no anomalies. Hence, we have reached a contradiction with **Definition 1**, and thus for every $\varepsilon > 0$ there is no such N for which $\|w_t^{\mathcal{A}} - w^*\| < \varepsilon$ is satisfied.

1.2 Proof of Lemma 1

We will show that removing outliers reduces the variance for a set of points on a number line with scalar values. Let $\{a_i\}$ be a set where $a_i \in \mathbb{R}, i \in \mathbb{N}$ and $a_1 < a_2 < \ldots < a_N$. We consider one of those points, a_N, an outlier point a_o, meaning that a_o significantly deviates from the rest of the points. The mean $\bar{a}$ of $\{a_i\}$ is given as

$$\bar{a} = \frac{1}{N} \sum_{i=1}^{N} a_i. \tag{1}$$

If we exclude a_o, the new mean $\bar{a}'$ is

$$\bar{a}' = \frac{1}{N-1} \sum_{i=1}^{N-1} a_i. \tag{2}$$

But (1) can be rewritten as

$$\bar{a} = \frac{1}{N} \left(\sum_{i=1}^{N-1} a_i + a_o \right) \tag{3}$$

$$\bar{a} = \frac{N-1}{N} \bar{a}' + \frac{a_o}{N} \tag{4}$$

Equivalently,

$$\bar{a} - \bar{a}' = \frac{a_o - \bar{a}'}{N} \tag{5}$$

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

Variance σ^2 of the set without outlier removal:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(a_i - \bar{a})^2 \tag{6}$$

$$\sigma^2 = \frac{1}{N}\left(\sum_{i=1}^{N-1}(a_i - \bar{a})^2 + (a_o - \bar{a})^2\right) \tag{7}$$

Variance $(\sigma')^2$ of the set with a_o removed:

$$(\sigma')^2 = \frac{1}{N-1}\sum_{i=1}^{N-1}(a_i - \bar{a}')^2 \tag{8}$$

The deviation of each term a_i around the mean $\bar{a}$ is

$$a_i - \bar{a} = a_i - \bar{a}' - (\bar{a} - \bar{a}') \tag{9}$$

Using (5):

$$a_i - \bar{a} = a_i - \bar{a}' - \frac{a_o - \bar{a}'}{N} \tag{10}$$

$$(a_i - \bar{a})^2 = \left(a_i - \bar{a}' - \frac{a_o - \bar{a}'}{N}\right)^2 \tag{11}$$

$$= (a_i - \bar{a}')^2 - 2(a_i - \bar{a}')\left(\frac{a_o - \bar{a}'}{N}\right) + \left(\frac{a_o - \bar{a}'}{N}\right)^2 \tag{12}$$

$$\begin{aligned}&\sum_{i=1}^{N-1}(a_i - \bar{a})^2 = \\ &\sum_{i=1}^{N-1}(a_i - \bar{a}')^2 - 2\left(\frac{a_o - \bar{a}'}{N}\right)\sum_{i=1}^{N-1}(a_i - \bar{a}') + \\ &(N-1)\left(\frac{a_o - \bar{a}'}{N}\right)^2\end{aligned} \tag{13}$$

$\sum_{i=1}^{N-1}(a_i - \bar{a}') = 0$ due to sum of deviations around the mean being zero. Then (13) reduces to

$$\begin{aligned}&\sum_{i=1}^{N-1}(a_i - \bar{a})^2 = \\ &\sum_{i=1}^{N-1}(a_i - \bar{a}')^2 + (N-1)\left(\frac{a_o - \bar{a}'}{N}\right)^2\end{aligned} \tag{14}$$

Plugging (14) into (7) we get

$$\sigma^2 = \frac{1}{N}\left[\sum_{i=1}^{N-1}(a_i - \bar{a}')^2 + (N-1)\left(\frac{a_o - \bar{a}'}{N}\right)^2 + (a_o - \bar{a})^2\right] \tag{15}$$

Using (8):

$$\sigma^2 = \frac{N-1}{N}\sigma'^2 + \frac{(N-1)}{N}\left(\frac{a_o - \bar{a}'}{N}\right)^2 + \frac{(a_o - \bar{a})^2}{N} \tag{16}$$

Given that a_o is sufficiently large, from (16) it follows that $\sigma^2 > \sigma'^2$.

1.3 Proof of Theorems 1 and 3

According to *lemma 1* (the inequality here is not strict because we might not remove any model weights at all):

$$\frac{1}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\left(w_t^i-\mu_t^{\mathcal{G}}\right)^2 \leq \frac{1}{|\mathcal{A}|}\sum_{j\in\mathcal{A}}\left(w_t^j-\mu_t^{\mathcal{A}}\right)^2 \tag{17}$$

Multiplying by $|\mathcal{G}|$ both sides and additionally multiplying the right side by $\frac{|\mathcal{A}|}{|\mathcal{A}|}$ yields:

$$\sum_{i\in\mathcal{G}}\left(w_t^i-\mu_t^{\mathcal{G}}\right)^2 \leq \frac{|\mathcal{G}|}{|\mathcal{A}|}\sum_{j\in\mathcal{A}}\left(w_t^j-\mu_t^{\mathcal{A}}\right)^2 \tag{18}$$

In vector notation using the Euclidean norm:

$$\left\|w_t^{\mathcal{G}}-\mu_t^{\mathcal{G}}\right\|^2 \leq \frac{|\mathcal{G}|}{|\mathcal{A}|}\left\|w_t^{\mathcal{A}}-\mu_t^{\mathcal{A}}\right\|^2 \tag{19}$$

Because centroid $\mu_t^{\mathcal{A}}$ minimizes $\left\|w_t^{\mathcal{A}}-\mu_t^{\mathcal{A}}\right\|^2$:

$$\left\|w_t^{\mathcal{G}}-\mu_t^{\mathcal{G}}\right\|^2 \leq \frac{|\mathcal{G}|}{|\mathcal{A}|}\left\|w_t^{\mathcal{A}}-\mu_t^{\mathcal{A}}\right\|^2 \leq \frac{|\mathcal{G}|}{|\mathcal{A}|}\left\|w_t^{\mathcal{A}}-\mu_t^{\mathcal{G}}\right\|^2 \tag{20}$$

$$\lim_{t\to\infty}\left\|\mu_t^{\mathcal{G}}-w^*\right\| = 0$$

For $t \geq N$:

$$\left\|w_t^{\mathcal{G}}-w^*\right\|^2 \leq \frac{|\mathcal{G}|}{|\mathcal{A}|}\left\|w_t^{\mathcal{A}}-w^*\right\|^2 \tag{21}$$

Finally,

$$\left\|w_t^{\mathcal{G}}-w^*\right\| \leq \sqrt{\frac{|\mathcal{G}|}{|\mathcal{A}|}}\left\|w_t^{\mathcal{A}}-w^*\right\| \quad \square \tag{22}$$

In the Appendix, we also provide additional commentary to theorems 1 and 3.

1.4 Proof of Theorem 2

Let assume that possibly $N_2 > N_1 : \exists N_1 < k < N_2$, such that for $w_k^{\mathcal{A}}$ and $w_k^{\mathcal{G}}$, there exists round k such that:

$$\left\|w_k^{\mathcal{A}}-w^*\right\| < \varepsilon \quad \text{and} \quad \left\|w_k^{\mathcal{G}}-w^*\right\| > \varepsilon, \tag{23}$$

meaning that in round k

$$\left\|w_k^{\mathcal{A}}-w^*\right\| \leq \left\|w_k^{\mathcal{G}}-w^*\right\| \tag{24}$$

that contradicts to the definition given in (1). In (24), k is sufficiently large such that the outlier updates significantly affect the global model $w_k^{\mathcal{A}}$, causing it to deviate beyond ε-distance from w^*.

Additional commentary to Theorems 1 and 3

If we further split $w_t^{\mathcal{A}}$ into "good" $w_t^{\mathcal{G}}$ and "bad" $w_t^{\mathcal{B}}$ clients ($\mathcal{B} = \{w_t^{i_1}, w_t^{i_2}, \ldots, w_t^{i_{|\mathcal{B}|}}\}$), we can derive the following, more detailed bound for the relation $\frac{\|w_t^{\mathcal{A}}-w^*\|}{\|w_t^{\mathcal{G}}-w^*\|}$:

$$w_t^{\mathcal{A}} = \frac{|\mathcal{G}|}{|\mathcal{B}|+|\mathcal{G}|}w_t^{\mathcal{G}} + \frac{|\mathcal{B}|}{|\mathcal{B}|+|\mathcal{G}|}w_t^{\mathcal{B}}$$

$$\left\|w_t^{\mathcal{A}}-w^*\right\| = \frac{|\mathcal{G}|}{|\mathcal{B}|+|\mathcal{G}|}\left\|w_t^{\mathcal{G}}-w^*\right\| + \frac{|\mathcal{B}|}{|\mathcal{B}|+|\mathcal{G}|}\left\|w_t^{\mathcal{B}}-w^*\right\|$$

Dividing both sides by $\|w_t^{\mathcal{G}}-w^*\|$ yields

$$\frac{\left\|w_t^{\mathcal{A}}-w^*\right\|}{\left\|w_t^{\mathcal{G}}-w^*\right\|} = \frac{|\mathcal{G}|}{|\mathcal{B}|+|\mathcal{G}|} + \frac{|\mathcal{B}|}{|\mathcal{B}|+|\mathcal{G}|}\frac{\left\|w_t^{\mathcal{B}}-w^*\right\|}{\left\|w_t^{\mathcal{G}}-w^*\right\|}$$

In comparison to Theorem 1.2, here we provide an equality, i.e. we can quantify the relation $\frac{\|w_t^{\mathcal{A}}-w^*\|}{\|w_t^{\mathcal{G}}-w^*\|}$. However, since w^* in practice is unknown our approximation can only be based on μ_t. This would further increase the term $\frac{\|w_t^{\mathcal{B}}-w^*\|}{\|w_t^{\mathcal{G}}-w^*\|}$ to $\frac{\|w_t^{\mathcal{B}}-\mu_t\|}{\|w_t^{\mathcal{G}}-\mu_t\|}$.

1.5 Proof of Theorem 4

Proof: PID score:

$$u_t^{(i)} = D_t^{(i)}(w_t^{(i)}, \mu_t) + K_I \sum_{x=0}^{t-1} D_x^{(i)}(w_x^{(i)}, \mu_x) + K_d(D_t^{(i)}(w_t^{(i)}, \mu_t) - D_{t-1}^{(i)}(w_{t-1}^{(i)}, \mu_{t-1})) \quad (25)$$

where

$$D_t^{(i)}(w_t^{(i)}, \mu_t) = \left\| w_t^{(i)} - \mu_t \right\|.$$

Substitute into (25):

$$u_t^{(i)} = \left\| w_t^{(i)} - \mu_t \right\| + K_I \sum_{x=0}^{t-1} \left\| w_x^{(i)} - \mu_x \right\| + (1 - K_I)\Big\{ \left\| w_t^{(i)} - \mu_t \right\| - \left\| w_{t-1}^{(i)} - \mu_{t-1} \right\| \Big\} \quad (26)$$

Assume minority of clients are anomalous, i.e. $f < \frac{N}{2}$, where f is the number of anomalies. Let's first show that the centroid $\mu_t = \frac{1}{N}\sum_i w_t^{(i)}$ will not be shifted significantly.

$$\mu_t = \frac{1}{N}\sum_i w_t^{(i)} = \frac{1}{N}\left(\sum_{i \in \mathcal{G}} w_t^{(i)} + \sum_{j \in \mathcal{B}} w_t^{(j)} \right). \quad (27)$$

Consider the purely good centroid $\mu_t^{\mathcal{G}}$:

$$\mu_t^{\mathcal{G}} = \frac{1}{|\mathcal{G}|}\sum_{i \in \mathcal{G}} w_t^{(i)} = \frac{1}{N-f}\sum_{i \in \mathcal{G}} w_t^{(i)}. \quad (28)$$

Let's subtract $\mu_t^{\mathcal{G}}$ from both sides of (4):

$$\mu_t - \mu_t^{\mathcal{G}} = \frac{1}{N}\sum_{i \in \mathcal{G}} w_t^{(i)} - \frac{1}{N-f}\sum_{i \in \mathcal{G}} w_t^{(i)} + \frac{1}{N}\sum_{j \in \mathcal{B}} w_t^{(j)}. \quad (29)$$

$$\mu_t - \mu_t^{\mathcal{G}} = \left(\frac{1}{N} - \frac{1}{N-f} \right) \sum_{i \in \mathcal{G}} w_t^{(i)} + \frac{1}{N}\sum_{j \in \mathcal{B}} w_t^{(j)}.$$

Take the Euclidean norm on both sides and apply triangle inequality:

$$\left\| \mu_t - \mu_t^{\mathcal{G}} \right\| \leq \left| \frac{1}{N} - \frac{1}{N-f} \right| \sum_{i \in \mathcal{G}} \left\| w_t^{(i)} \right\| + \frac{1}{N}\sum_{j \in \mathcal{B}} \left\| w_t^{(j)} \right\|. \quad (30)$$

Note that

$$\left| \frac{1}{N} - \frac{1}{N-f} \right| = \left| \frac{N - f - N}{N(N-f)} \right| = \frac{f}{N(N-f)}. \quad (31)$$

Now, **assume** both anomalous and benign norms are bound with some constant ζ. This means that $\sum_{i \in \mathcal{G}} \left\| w_t^{(i)} \right\| \leq (N-f)\zeta$ and $\sum_{j \in \mathcal{B}} \left\| w_t^{(i)} \right\| \leq f\zeta$. Considering this and (31), we rewrite (30) as:

$$\left\| \mu_t - \mu_t^{\mathcal{G}} \right\| \leq \frac{f(N-f)}{N(N-f)}\zeta + \frac{f}{N}\zeta \leq \frac{2f}{N}\zeta \leq O\Big(\frac{f}{N}\Big). \quad (32)$$

(32) Provides an upper bound on the centroid shift, which refer to as **Bounded Centroid Shift** in **Lemma 2** of the main paper. This lemma allows us to bound the PID score for benign clients. Let's analyze each term in (8) by rewriting it with the benign centroid $\mu_t^{\mathcal{G}}$:

$$u_t^{(i)\in\mathcal{G}} = \underbrace{\left\|w_t^{(i)\in\mathcal{G}} - \mu_t\right\|}_{Proportional} + K_I \underbrace{\sum_{x=0}^{t-1}\left\|w_x^{(i)\in\mathcal{G}} - \mu_x\right\|}_{Integral} + (1-K_I)\underbrace{\left\{\left\|w_t^{(i)\in\mathcal{G}} - \mu_t\right\| - \left\|w_{t-1}^{(i)\in\mathcal{G}} - \mu_{t-1}\right\|\right\}}_{Derivative}. \tag{33}$$

First, consider the proportional part. Add and subtract $\mu_t^{\mathcal{G}}$:

$$\left\|w_t^{(i)\in\mathcal{G}} - \mu_t + \mu_t^{\mathcal{G}} - \mu_t^{\mathcal{G}}\right\| \le \left\|w_t^{(i)\in\mathcal{G}} - \mu_t^{\mathcal{G}}\right\| + \left\|\mu_t - \mu_t^{\mathcal{G}}\right\| \le \left\|w_t^{(i)\in\mathcal{G}} - \mu_t^{\mathcal{G}}\right\| + O\Big(\frac{f}{N}\Big). \tag{34}$$

Assuming bounded heterogeneity between good clients, the upper bound on $\left\|w_t^{(i)\in\mathcal{G}} - \mu_t^{\mathcal{G}}\right\|$ is some Δ_{max}. Then (34) is bounded above by $\Delta_{max} + O(\frac{f}{N})$.

Second, look at the integral part. Same manipulation:

$$\sum_{x=0}^{t-1}\left\|w_x^{(i)\in\mathcal{G}} - \mu_x + \mu_t^{\mathcal{G}} - \mu_t^{\mathcal{G}}\right\| \le \sum_{x=0}^{t-1}\left\|w_x^{(i)\in\mathcal{G}} - \mu_t^{\mathcal{G}}\right\| + \sum_{x=0}^{t-1}\left\|\mu_x - \mu_t^{\mathcal{G}}\right\| = \sum_{x=0}^{t-1}\left\|w_x^{(i)\in\mathcal{G}} - \mu_t^{\mathcal{G}}\right\| + t\cdot O\Big(\frac{f}{N}\Big) = t\Big(\Delta_{max} + O(\frac{f}{N})\Big) \tag{35}$$

Third, we do the same analysis on the derivative part:

$$\left\|w_t^{(i)\in\mathcal{G}} - \mu_t\right\| - \left\|w_{t-1}^{(i)\in\mathcal{G}} - \mu_{t-1}\right\| \le \left\|w_t^{(i)\in\mathcal{G}} - \mu_t^{\mathcal{G}}\right\| + O\Big(\frac{f}{N}\Big) - \left\|w_{t-1}^{(i)\in\mathcal{G}} - \mu_{t-1}^{\mathcal{G}}\right\| + O\Big(\frac{f}{N}\Big) \le O\Big(\frac{f}{N}\Big). \tag{36}$$

Combining (34), (35), (36), we get that for a good client, the upper bound on PID value $u_t^{(i)\in\mathcal{G}}$ is $\Delta_{max} + O(\frac{f}{N}) + t\Big(\Delta_{max} + O(\frac{f}{N})\Big) + O(\frac{f}{N})$, which can be simplified to $t\Big(\Delta_{max} + O(\frac{f}{N})\Big)$.

The permissive upper bound of a threshold for the PID score of good clients $\{i : i \in \mathcal{G}\}$ is $t\Big(\Delta_{max} + O(\frac{f}{N})\Big)$. This upper bound ensures zero false positive rate, however, the false negative rate can be expected to be high. This bound is not usable, however, it provides a starting point for us to derive a more tight and practical threshold.

1.6 Proof of Theorem 5

Proof:

$$\mathbb{E}[U_t] = \mathbb{E}\Big[\|w_t^{(i)} - \mu_t\|\Big] + K_I \sum_{x=0}^{t-1}\mathbb{E}\Big[\|w_x^{(i)} - \mu_x\|\Big] + K_D\mathbb{E}\Big[\|w_t^{(i)} - \mu_t\| - \|w_{t-1}^{(i)} - \mu_{t-1}\|\Big] \tag{37}$$

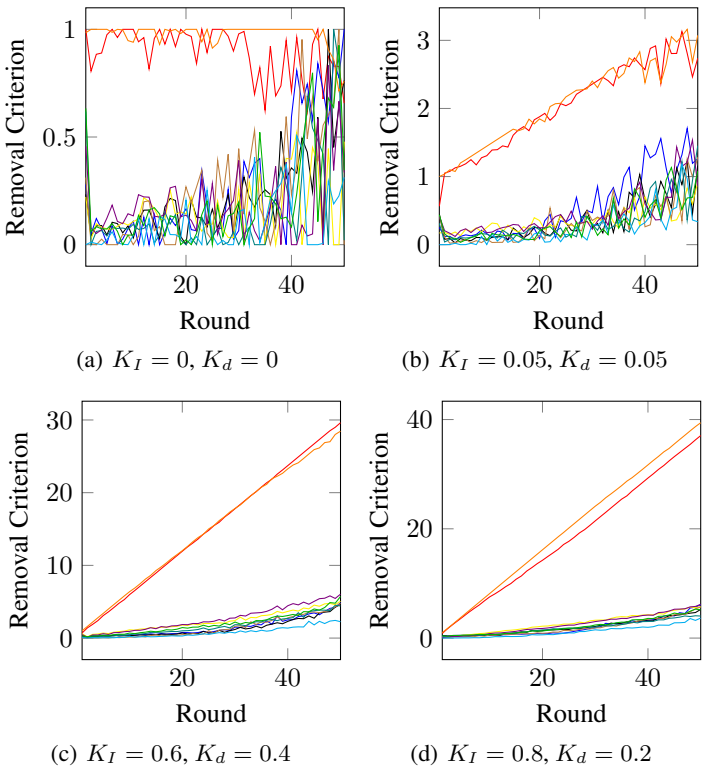
Using $\mu_\Delta = \mathbb{E}[\Delta_t]$

$$\mathbb{E}[U_t] = \mu_\Delta + K_I t\mu_\Delta + 2K_D\mu_\Delta \approx \mu_\Delta(1 + K_I t) \tag{38}$$

Next, derive the variance of our PID Score. Due to Bienaymé identity we would have additional covariance terms, but those can be neglected due to assumption (1). The variance of U_t then becomes:

$$\sigma_t^2 = \text{Var}[U_t] = \text{Var}[\Delta_t] + K_I^2\sum_{x=0}^{t-1}\text{Var}[\Delta_x] + 2K_D\text{Var}[\Delta_t] = \sigma_\Delta^2 + K_I^2\sigma_\Delta^2 + 2K_D^2\sigma_\Delta^2 \tag{39}$$

Finally, using Chebyshev's inequality we can state that with a probability of at least $1-\alpha$ the benign clients will be under the threshold $\tau = \bar{u}_t + z\sigma_t$. Equivalently, no more than α-fraction of benign clients exceed τ.



(a) $K_I = 0$, $K_d = 0$ (b) $K_I = 0.05$, $K_d = 0.05$

(c) $K_I = 0.6$, $K_d = 0.4$ (d) $K_I = 0.8$, $K_d = 0.2$

Figure 1: The effect of varying coefficients for the integral and derivative parts. 8 benign and 2 anomalous datasets were generated based on the FEMNIST dataset.

1.7 Proof of Theorem 6

Proof: the Gaussian threshold follows directly from **Theorem 5** under the Gaussian assumption $\Delta_t \sim \mathcal{N}(\mu_\Delta, \sigma_\Delta^2)$. If α is the desired alarm rate, than using the standard normal distribution and the z-score corresponding to $1-\alpha$ gives us $\Pr[U_t > \tau_{Gauss}] = 1 - \Phi(z_{1-\alpha}) = \alpha$.

2 Ablation Study: PID Coefficients

We begin our empirical study by investigating the impact of varying coefficient values of the proportional, integral, and derivative component of equation 25. First, consider Figure 1(a), where $K_p = 1$, $K_I = 0$ and $K_d = 0$ as well. In this case we observe a metric convergence pattern similar to the one in Krum and Multi-Krum due to the fact that PID value degrades to the P component's value only, i.e. we are simply comparing the distances between each client's submitted model weights and the centroid. As we increase K_I and K_d slightly, to 0.05, the distinction between benign and anomalous clients becomes more prominent (Figure 1(b), anomalous clients are represented by the orange and red curves at the top), however, the convergence of anomalous and benign clients still persists. As we increase K_I even more, the compounding effect of PID takes over, and the anomalous clients become clearly identifiable throughout the entire learning process (Figures 1(c), 1(d)). Even though PID acts based on the distance values, because of the compounding of the integral term we can avoid the mix-up between the anomlaous and benign clients. For the remaining parts of our empirical evaluation we heuristically set the values $K_I = 0.8$ and $K_d = 0.2$.

3 Performance

Table 1: Last round accuracy and loss for each method across datasets. This table illustrates the results in Fig. 4 of the main paper.

Metric	Method	FEMNIST	ITS	Pneumonia	Std(FEMNIST)	Std(ITS)	Std(Pneumonia)
Accuracy	Krum	0.9434	0.9057	0.9868	0.1719	0.0528	0.0094
	Multi-Krum	0.9537	0.9306	0.9860	0.1352	0.0463	0.0762
	RFA	0.9730	0.9405	0.9895	0.1108	0.0452	0.0796
	Bulyan	0.9732	0.9057	0.9925	0.1027	0.0448	0.0440
	PID	0.9715	0.9306	0.9895	0.1311	0.0450	0.0092
Loss	Krum	0.0180	0.1223	0.0056	0.0275	0.0477	0.0043
	Multi-Krum	0.0107	0.1452	0.0112	0.0235	0.2551	0.0583
	RFA	0.0086	0.3831	0.0041	0.0221	0.1776	0.0343
	Bulyan	0.0095	1.1076	0.0068	0.0222	0.3492	0.0234
	PID	0.0121	0.2699	0.0066	0.0226	0.0666	0.0017

3.1 PID Computation Time

Total clients	Time (ms)
5	2.225
6	2.394
7	2.160
8	2.575
9	2.602
10	2.757
11	3.051
12	3.348
13	3.771
14	3.752
15	4.126
16	4.260
17	4.196
18	4.329
19	4.559
20	4.767

Table 2: This graph illustrates Fig. 4(f) of the main paper. Time required for computing the PID score as the number of participating clients increases from 5 to 20 clients.

Code and Dataset Artifacts

The experiments were conducted on a system equipped with an AMD Ryzen 5 7600 CPU, 32 GB of RAM, and an NVIDIA RTX 4060TI GPU with 16 GB of dedicated memory, running the Ubuntu 22.04 OS. Our code may be used for the reproduction and further reconfiguration of our experimental setup. Additionally, it provides the ability to collect and save metrics necessary for the further analysis. We also provide the datasets that we used to facilitate the reproduction of our empirical study experiments. All the shared materials can be found by this anonymized link that does not disclose the authors' identities: https://drive.google.com/file/d/1VSTeE6ynMPQcnGUu_nIZO0_mkQdni8DH/view?usp=drive_link.

Datasets used in the experiments are initially downloaded from AWS by the execution script and later can be found in the *datasets/* folder of the archive.

Guidelines for the experiment setup configuration and execution are included in the README with the code artifacts found in the link above.