HISTORY-GUIDED DIRECTIONAL AGGREGATION FOR BYZANTINE-ROBUST FEDERATED LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Defending federated learning against untargeted poisoning requires suppressing malicious influence while preserving useful information from clients with non-IID data. We propose History-Guided Directional Aggregation (HGDA), which uses a small trusted server dataset to assess clients over time and weight the distinct update directions they currently submit. At each communication round, the server trains a copy of the global model on this dataset to obtain reference updates at two checkpoints. HGDA separately accumulates positive and negative alignments between client and reference updates within a sliding window to score each client. Each current direction receives the largest score among the clients submitting it, so repeated submissions share an aggregation weight while nearby directions remain separate. The server uses the aggregated directions to update the global model with its own optimizer, independently of client and reference update magnitudes. HGDA requires no estimate of the Byzantine population. We analyze convergence with server Adam, showing how residual adversarial influence and local-training error affect progress on the benign clients' objective. Experiments on five datasets demonstrate broad Byzantine robustness across diverse poisoning attacks and Byzantine fractions under full and partial participation, with strong accuracy where representative baselines severely degrade or collapse. HGDA remains effective under poisoning when the trusted dataset is not representative of client data and has incomplete class coverage. It also demonstrates resilience to scheduled and adaptive on-off attacks, in which clients alternate between benign training and poisoning.

1 INTRODUCTION

Federated learning trains a shared model by aggregating locally computed updates without centralizing client data (McMahan et al., 2017). The server receives updates without directly observing the local data or computation that produced them, leaving training exposed to malicious submissions (Fang et al., 2020). Strong client-admission and device protections can keep malicious participation low enough for simple aggregation defenses to be effective (Shejwalkar et al., 2022). However, verifying that a device is genuine does not ensure that its training contribution is benign (Bonawitz et al., 2019). Attackers can also inject software-emulated clients rather than compromise genuine devices, reducing the cost of malicious participation (Mozaffari et al., 2024). Recent attacks show that injected clients can substantially degrade model accuracy despite existing robust aggregation rules (Xie et al., 2025). These failures motivate stronger aggregation rules that preserve model utility across a broad range of adversarial conditions. We therefore study robustness across different attacks and levels of malicious participation, abstracting away how attackers compromise devices or inject clients.

Defending against these attacks is especially difficult when benign clients have non-IID data. Their different local objectives produce diverse updates, while coordinated attackers can craft harmful updates that resemble benign ones (Baruch et al., 2019; Shejwalkar & Houmansadr, 2021). Differences among uploads can therefore reflect useful data diversity as well as adversarial manipulation. The challenge is to suppress malicious influence while preserving useful information from heterogeneous benign clients.

We consider untargeted poisoning attacks that aim to reduce the global model's overall predictive accuracy. The server holds a small trusted dataset for the learning task (Xie et al., 2020b; Bao et al.,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

2024). The service provider can collect and label this dataset without pooling clients' private data (Cao et al., 2021). In each communication round, the server trains a copy of the current global model on this dataset to obtain a reference update. Like a benign client with non-IID data, the server can provide a useful training signal for the shared task even when its data distribution differs from those of the clients. Because this signal comes from a trusted source, the server can use it to assess client contributions rather than relying solely on comparisons among their updates.

The trusted reference and a benign client can reflect different local data, so their updates need not align closely in every round. We therefore keep a *history* for each client, recording how its updates align with or oppose trusted training over recent rounds. Both types of evidence inform the assessment, allowing opposing updates to offset accumulated support.

The server uses these assessments to weight current client updates. Repeated submissions of the same direction can amplify its influence without adding new directional information, as illustrated by the Mimic attack (Karimireddy et al., 2022). Yet merging nearby but different directions can remove useful variation among benign clients (Bao et al., 2024; Fung et al., 2020). We therefore use histories to assess clients and assign aggregation weights to the distinct directions they currently submit.

We propose *History-Guided Directional Aggregation* (HGDA, Figure 1). HGDA accumulates positive and negative directional evidence separately within a sliding window. It assesses each client against early and later checkpoints of one trusted-data training path, with the weaker assessment determining the client's score. Among clients with positive scores, those submitting the same current direction share the largest score rather than adding their scores together.

Client updates contribute through their directions rather than their magnitudes, while the trusted references guide assessment without setting the global step size. The server uses the aggregated directions to update the global model with its own optimizer. HGDA requires no estimate of the Byzantine population.

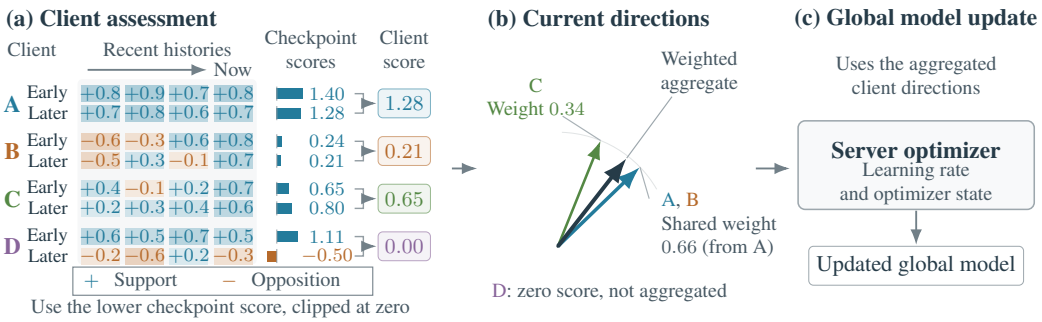


Figure 1: HGDA overview with illustrative histories. (a) The lower of two checkpoint scores is clipped at zero to give the client score. (b) Clients submitting the same unit direction share their largest score, while nearby directions remain distinct. Zero-score clients are excluded. (c) The server optimizer updates the global model from the weighted aggregate. Figure 2 details (a).

Our contributions are as follows.

- We introduce HGDA, which uses trusted data to assess clients over recent rounds and weights their distinct current directions. It keeps nearby directions separate, controls repeated influence and leaves the global update scale to the server optimizer, independently of client and reference magnitudes.
- We establish convergence bounds for HGDA with server Adam. The analysis connects signed client histories and direction-level weights to progress on the benign clients' objective, quantifying the roles of residual adversarial influence and local-training error.
- Experiments on five datasets demonstrate broad Byzantine robustness across poisoning attacks and Byzantine fractions of 1–80%, under full and partial participation. HGDA maintains strong accuracy where representative baselines suffer severe degradation or collapse.

- Ablations and sensitivity studies clarify the roles of history, reference choice and direction-level weighting. HGDA learns effectively from heterogeneous clients under poisoning despite bias and incomplete class coverage in the trusted dataset. It also shows resilience when clients repeatedly switch between benign and malicious behavior, including when attackers adapt their timing to client scores.

2 Related Work

Aggregation from client updates. Robust aggregation rules can limit malicious influence using distances or coordinate statistics among client updates, without requiring server data. Krum selects an update using nearest-neighbor distances, while Bulyan combines candidate selection with coordinate-wise filtering (Blanchard et al., 2017; El-Mhamdi et al., 2018). Coordinate-wise median and trimmed mean operate on individual coordinates (Yin et al., 2018), whereas RFA uses a geometric median (Pillutla et al., 2022). Under non-IID data, these comparisons must also accommodate variation among benign clients. Bucketing and nearest neighbor mixing address this variation by averaging random groups (Karimireddy et al., 2022) or local neighborhoods (Allouah et al., 2023) before robust aggregation. GAS instead scores lower-dimensional parts of the updates before selecting which full updates to aggregate (Liu et al., 2023).

Trusted server data. Other defenses use a small trusted dataset at the server to evaluate client contributions, rather than relying only on comparisons among client updates. Zeno scores updates by estimated loss reduction and a norm penalty (Xie et al., 2019), while Sageflow filters and weights client models using prediction entropy and loss on public server data (Park et al., 2021). Other methods use server-computed gradients as references. Zeno++ filters asynchronous gradients against a validation gradient (Xie et al., 2020b), and FLTrust weights client directions by their positive current-round cosine similarity to a server-trained update, then rescales the aggregate by that reference's norm (Cao et al., 2021). Under label skew, BOBA uses class-specific server gradients to estimate client label proportions and filter updates (Bao et al., 2024). It uses a Byzantine-count bound and assumes shared class-conditional distributions.

Client histories. Robust aggregation can also draw on earlier rounds, either by retaining past updates or by tracking client behavior. Learning from History applies robust aggregation to worker momentum, which averages gradients across rounds (Karimireddy et al., 2021). DeMoA accommodates partial participation by combining fresh client momentum with cached momentum from clients absent in the current round (Otsuka et al., 2026). For client assessment, FLDetector accumulates discrepancies between predicted and received updates to detect malicious clients (Zhang et al., 2022). FoolsGold downweights clients with similar accumulated update vectors (Fung et al., 2020).

Directions and repeated submissions. In weighted aggregation, an update's influence depends on its magnitude as well as its weight. Fed-NGA forms a weighted average of normalized gradients, removing update magnitude from each client's contribution (Zuo et al., 2025). Copies of an otherwise benign update can also bias the aggregate, as demonstrated by Mimic under heterogeneous data (Karimireddy et al., 2022).

HGDA assesses each client through its signed alignment histories with two trusted references. These assessments determine the weights of distinct current directions, using the largest client score for each direction and keeping nearby directions separate. The server optimizer determines the global update scale independently of client and reference magnitudes.

3 Method

HGDA uses trusted references to assess clients over time and weight their distinct current directions. Figure 2 illustrates client scoring, and Algorithm 1 in Appendix A.1 gives the full procedure.

3.1 Setting and server references

Let N be the number of clients and $\mathcal{C}_t$ the set participating in round t, with $n_t = |\mathcal{C}_t|$. Benign clients train from the global model $\theta_t \in \mathbb{R}^d$ on local, potentially non-IID data and submit updates

Δ_i^t. Byzantine clients may submit arbitrary updates, including coordinated copies, to reduce overall predictive accuracy. The server holds a small clean dataset D_s from the learning task, separate from clients' training data, and keeps both this dataset and its reference updates private. HGDA requires no Byzantine-count estimate.

Each round, the server trains a copy of θ_t on D_s, with $\phi_{t,e}$ denoting its parameters after e SGD epochs and $\phi_{t,0} = \theta_t$. The displacement from θ_t can change direction during training, so a client update may align with an early reference but oppose a later one (Figure 2(a)). HGDA retains both views using checkpoints at $E_s < E_l$. For $h \in \{s, l\}$, normalizing the increments gives the reference direction q_h^t, client direction u_i^t and their cosine similarity $r_{i,h}^t$:

$$q_h^t = \frac{\phi_{t,E_h} - \theta_t}{\|\phi_{t,E_h} - \theta_t\|_2}, \qquad u_i^t = \frac{\Delta_i^t}{\|\Delta_i^t\|_2}, \qquad r_{i,h}^t = \langle u_i^t, q_h^t \rangle. \tag{1}$$

The sign of $r_{i,h}^t$ indicates alignment or opposition to reference h. Reference increments are nonzero. Zero client increments receive zero alignment and are excluded from aggregation.

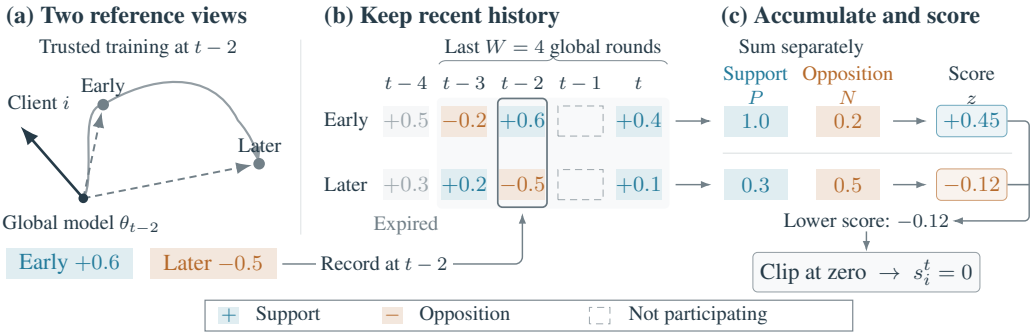


Figure 2: Client assessment in HGDA ($W = 4$, $\lambda = 1$ in this example). (a) Client i's update aligns with the early reference but opposes the later one. (b) Both alignments enter column $t-2$. Faded cells have expired and dashed cells indicate nonparticipation. (c) Support and opposition accumulate separately. The lower score, clipped at zero, gives $s_i^t = 0$ despite positive current-round alignments.

3.2 Scoring Recent Client Updates

HGDA records these signed alignments separately for each reference over the latest W global rounds (Figure 2(b)). For client i, the observed rounds are

$$\mathcal{W}_{i,t}^{(W)} = \{\tau : \max(0, t - W + 1) \le \tau \le t,\ i \in \mathcal{C}_\tau\}. \tag{2}$$

Unobserved rounds contribute no evidence.

For each reference, positive alignments and the magnitudes of negative alignments accumulate separately as support and opposition (Figure 2(c)). With $[x]_+ = \max(x, 0)$,

$$P_{i,h}^{t,W} = \sum_{\tau \in \mathcal{W}_{i,t}^{(W)}} [r_{i,h}^\tau]_+, \qquad N_{i,h}^{t,W} = \sum_{\tau \in \mathcal{W}_{i,t}^{(W)}} [-r_{i,h}^\tau]_+. \tag{3}$$

We form signed scores $z_{i,h}^{t,W}$ with a regularization parameter $\lambda > 0$ that moderates their magnitudes when evidence is sparse. We use $\lambda = 1$ by default and evaluate its sensitivity in Section 4.4. The smaller reference score determines the client score s_i^t:

$$z_{i,h}^{t,W} = \frac{P_{i,h}^{t,W} - N_{i,h}^{t,W}}{\sqrt{P_{i,h}^{t,W} + N_{i,h}^{t,W} + 2\lambda}}, \qquad s_i^t = \left[\min_{h \in \{s,l\}} z_{i,h}^{t,W}\right]_+, \tag{4}$$

Support across rounds can offset occasional disagreement, while opposing updates reduce the score. The square-root denominator moderates score growth, so equal additions of support yield

progressively smaller increases. Taking the lower checkpoint score requires net support under both references, so a favorable assessment under one cannot mask opposition under the other.

As the window advances, new observations replace expired evidence, allowing the assessment to adapt to changing client behavior. Section 4.5 evaluates resilience to on-off attacks.

Figure 3 shows how reference choice and history jointly affect scoring. The later reference gives fewer Byzantine updates positive scores under Min-Max, whereas the early reference does so under sign-flip. Accumulating evidence from only one reference can reinforce a favorable assessment of Byzantine updates: under Min-Max, early-only history raises their positive-score rate from 56.6% to 78.5%.

With both references, a 100-round history raises the benign positive-score rate from 57.8% to 83.2% under Min-Max and from 54.5% to 78.2% under sign-flip, while lowering the Byzantine rates below 1% in both attacks. Accumulated support thus helps retain benign contributions, while opposition identified through either reference can exclude Byzantine updates.

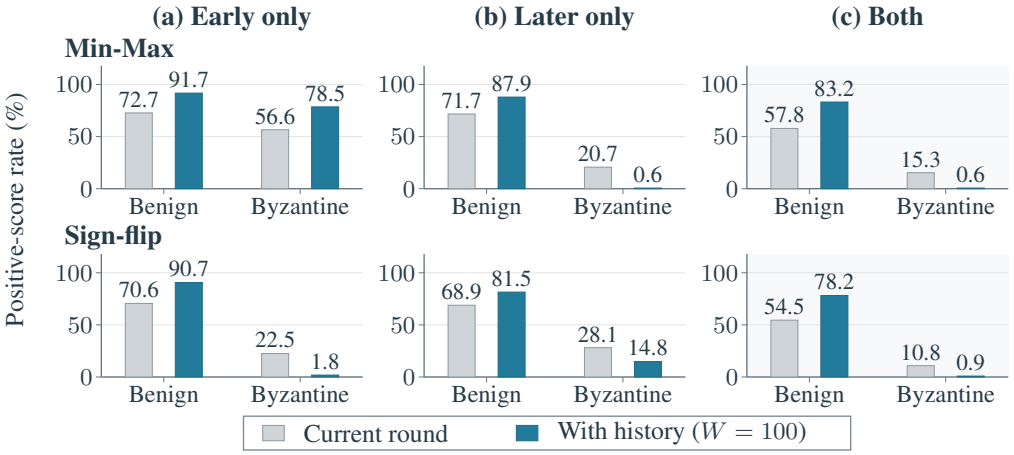


Figure 3: Reference choice and history in client scoring on CIFAR-10 (50% Byzantine). Bars show the percentage of observed updates with positive scores using the current round alone or a 100-round history. Appendix D.1 gives the protocol and further comparisons.

3.3 Direction-level aggregation and server optimization

HGDA groups the current nonzero updates from positive-score clients by their unit directions. Let $\mathcal{G}_t$ denote these groups and v_G^t the direction of group G. Each group receives one score, normalized across groups to form the aggregate:

$$b_G^t = \max_{i \in G} s_i^t, \qquad \pi_G^t = \frac{b_G^t}{\sum_{H \in \mathcal{G}_t} b_H^t}, \qquad \bar{u}_t = \sum_{G \in \mathcal{G}_t} \pi_G^t v_G^t. \tag{5}$$

Each group inherits its largest client score, preserving its strongest assessment without summing over copies or allowing low-scoring copies to dilute it. Nearby directions remain separate to preserve useful benign variation. The convex aggregate satisfies $\|\bar{u}_t\|_2 \leq 1$ and is not normalized again. Appendix G gives aggregation invariance and score concentration results.

The server updates the global model from $\bar{u}_t$ using bias-corrected Adam, whose learning rate and state control the parameter step (Reddi et al., 2021). Neither client nor reference update magnitudes

set the update scale. If no nonzero update has a positive score, the trainable parameters and optimizer state remain unchanged. Appendix A.2 gives the optimizer, buffer and computational details.

3.4 Convergence analysis

We use signed histories and groupwise maxima to bound the weight on directions supplied only by Byzantine clients (Lemma 1). This bound links HGDA's assessment to optimization of the benign population objective with server Adam.

For a smooth, lower-bounded objective, Proposition 5 bounds the smallest gradient norm attained over T applied server updates. If the residual adversarial weight leaves a uniformly positive descent margin under Adam's coordinate scaling and the local-training error is uniformly bounded, a learning rate proportional to $T^{-1/2}$ yields an $O(T^{-1/2})$ optimization term. The remaining term is proportional to local-training error, which includes data heterogeneity, stochastic-gradient and clipping errors and local drift. Appendix H gives the assumptions, proofs and additional learning-rate results.

4 Experiments

We evaluate HGDA across attacks, Byzantine fractions and participation conditions, then study trusted-set coverage, design choices and on-off resilience.

4.1 Experimental setup

Tasks and participation. FEMNIST (Caldas et al., 2018), Purchase-100 (Shokri et al., 2017) and Tiny ImageNet (Le & Yang, 2015) use 500 rounds, sampling 100 of 3400, 50 of 1000 and 10 of 100 clients per round, respectively. MNIST and CIFAR-10 use full participation with 100 clients for 200 and 1500 rounds. FEMNIST retains natural writer partitions. Other tasks use Dirichlet label skew with concentration 0.5. Byzantine identities and their population fraction m are fixed within each run. Under partial participation, sampling varies the round-level Byzantine fraction (Allouah et al., 2024) and produces sparser client histories. We test 1–60% on FEMNIST, 10–60% on Purchase, 10–40% on Tiny ImageNet and 20–80% under full participation.

Attacks and comparisons. All tasks use sign-flip, label-flip, Mimic and IPM, covering reversed updates, corrupted labels, copied updates and opposition to the benign mean. MNIST and CIFAR-10 add Gaussian, SameValue, ALIE, Min-Sum and Min-Max, including coordinated shifts constrained by benign-update statistics (Baruch et al., 2019; Shejwalkar & Houmansadr, 2021). We compare against Multi-Krum, Bulyan, coordinate median, trimmed mean, RFA, Fed-NGA, FLTrust-Early, FLTrust-Late and BOBA, with FedAvg and FLDetector+Median where evaluated. FedAvg (benign only), which aggregates only benign participants' updates, provides an empirical reference. Count-based defenses receive the true round-level Byzantine count, a favorable setting that eliminates count-estimation error. Under partial participation, we resample infeasible cohorts for Multi-Krum, Bulyan and trimmed mean. Appendices A.4 and B detail the comparisons and attacks.

Training and metrics. We use five local epochs, $E_s = 1$, $\lambda = 1$ and $W = 100$. The later reference approximately matches the average client's SGD step count. Full-participation studies also include $W = 10$. Trusted sets contain 100 examples, or 620 on FEMNIST. Evaluation excludes the reserved server pool. Shared maximum round budgets give slower methods time to learn under the same client-side protocol. We report best test accuracy within this budget. Appendices A.3 and C give the settings, budget rationale and complete results.

4.2 Robustness to Byzantine updates

Robustness across conditions. HGDA remains effective across the attack and Byzantine-fraction grids under partial participation (Figure 4). Its curves degrade gradually rather than showing the sharp collapses seen in several baselines. On FEMNIST and Tiny ImageNet, accuracy also stays close to FedAvg (benign only), indicating that aggregation preserves useful learning from benign clients.

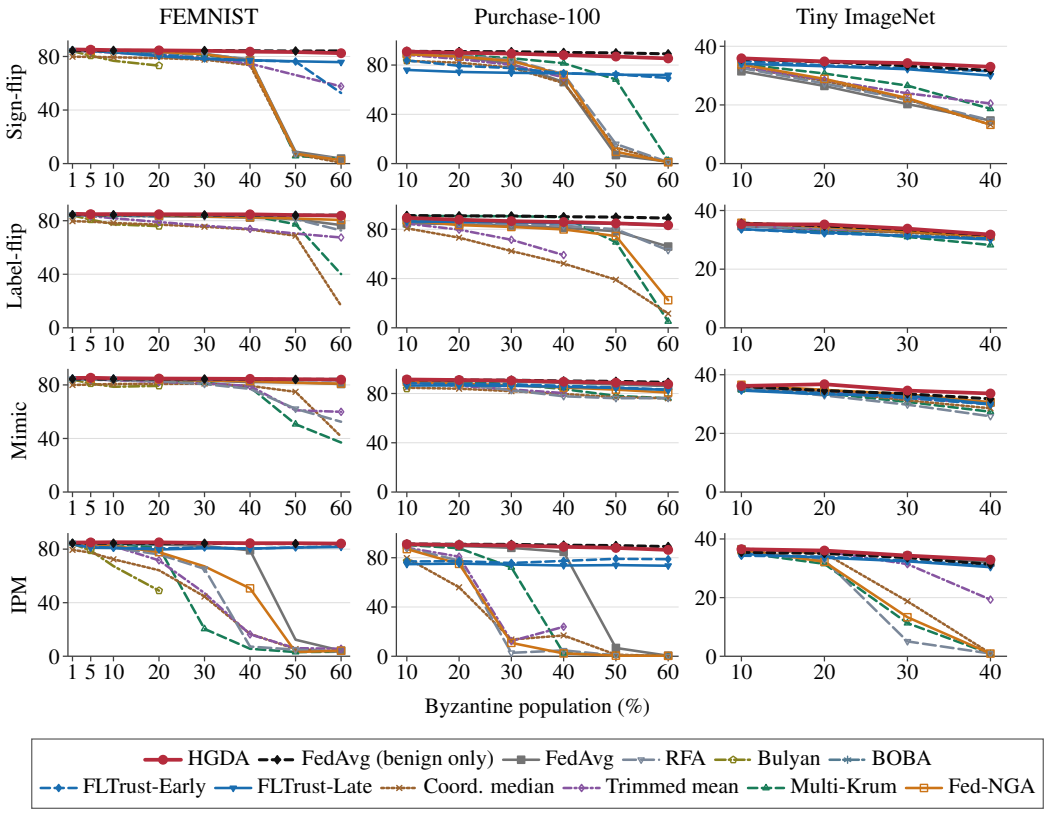


Figure 4: Robustness to Byzantine updates under partial participation. Best test accuracy (%) over three seeds.

Attack-dependent degradation. Several baselines are effective against one attack but fail under another. On Purchase, Multi-Krum leads the label-flip comparison at 10–40% Byzantine clients, yet drops to 2.47% under IPM at 40%, versus 88.85% for HGDA.

Full participation and broader attacks. The full-participation grids show the same consistency across a broader attack set (Figure 5). On CIFAR-10, HGDA with $W = 100$ reaches 80.10–81.99% accuracy across all nine attacks at $m = 40\%$. On MNIST, it reaches at least 97.50% across all 43 attacked conditions within 100 rounds (Appendix C.7).

4.3 Learning beyond the trusted set

The trusted sets contain 620 examples on FEMNIST and 100 on Purchase and Tiny ImageNet, at most 0.1% of the clients' combined training-set size. They cover 57–59 of 62, 60–65 of 100 and 72–84 of 200 classes, respectively. On Purchase, we compare macro accuracy on present and absent classes at each run's best overall-accuracy checkpoint.

Accuracy on absent classes stays within three percentage points of that on present classes across all attacks and fractions in Table 1. Even at 60% Byzantine clients, it remains above 80% across all four attacks (Appendix E). HGDA can therefore use a limited trusted set for assessment while still learning effectively from the broader client population.

4.4 Client assessment and direction-level weighting

We now relate the score behavior in Figure 3 to training accuracy and isolate the role of direction-level weighting (Table 2).

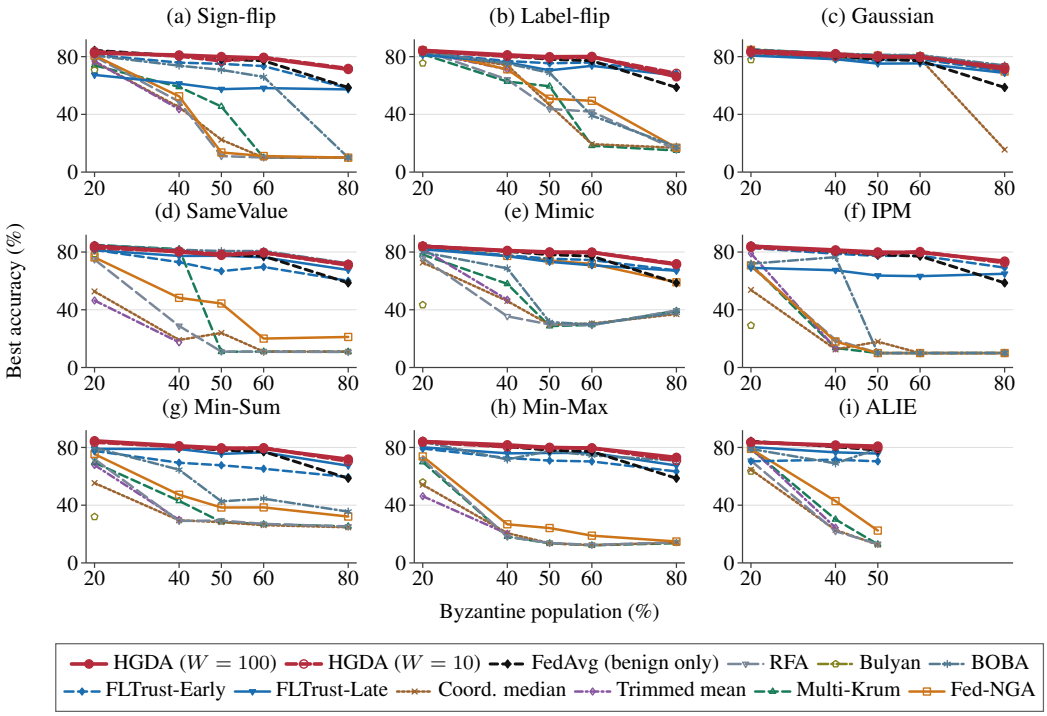


Figure 5: Robustness to Byzantine updates on CIFAR-10 under full participation. Best test accuracy (%).

Table 1: Macro accuracy (%) on Purchase-100 classes present in or absent from the trusted set (over three seeds).

	$m = 10\%$		$m = 40\%$	
Attack	Present	Absent	Present	Absent
Sign-flip	91.29 ± 0.89	89.78 ± 0.91	88.23 ± 0.88	85.80 ± 0.79
Label-flip	89.58 ± 1.07	87.78 ± 1.15	86.19 ± 0.76	83.34 ± 1.03
Mimic	91.99 ± 0.84	89.62 ± 0.90	89.93 ± 0.53	87.52 ± 0.61
IPM	91.49 ± 1.27	88.98 ± 1.06	89.14 ± 0.88	86.83 ± 0.92
Clean: Present 91.88 ± 0.48, Absent 90.66 ± 0.10.				

Reference choice. The early and later references complement each other across attacks (Figure 3). Using both improves CIFAR-10 label-flip accuracy by 3.41 and 8.13 percentage points over early-only and later-only scoring, while remaining close to early-only under sign-flip (Table 2(a)). The Purchase checkpoint sweeps show similar accuracy with single- and dual-reference configurations (Appendix E.4). Together, these results support using both references for broad robustness: the joint assessment improves performance where a single reference is less reliable and maintains accuracy where either reference works well.

History and sparse observations. On CIFAR-10 at $m = 50\%$, a 100-round history improves sign-flip accuracy by 10.73 points over using the current round alone (Table 2(b)), consistent with the better score separation in Figure 3. Label-flip accuracy remains near 80% across windows. The FEMNIST sweeps show low parameter sensitivity with sparse histories: under 10% sign-flip, varying W over 10–300 or λ over 0.1–10 changes accuracy by less than 0.3 points in either sweep (Table 2(d,e)).

Table 2: Design studies and on-off resilience. Best test accuracy (%). CIFAR-10 uses $m = 80\%$, $W = 10$ in (a,c) and $m = 50\%$ in (b,f). Panel (f) uses $W = 100$ and lists adaptive switch-off thresholds. FEMNIST panels (d,e) use 10% sign-flip.

(a) Reference choice

Reference	Sign-flip	Label-flip
Early	72.42	64.98
Later	68.78	60.26
Both	71.76	68.39

(b) CIFAR-10 W

W	Sign-flip	Label-flip
1	69.28	79.97
10	77.04	79.08
100	80.01	79.92

(c) Aggregation level

Attack	Client-level	HGDA (direction-level)
Mimic	70.08	71.83
Adaptive Mimic	64.14	71.83

(d) FEMNIST W

W	Best
10	84.94
50	84.74
100	84.77
300	84.89

(e) FEMNIST λ

λ	Best
0.1	84.86
1	84.77
5	84.77
10	85.03

(f) CIFAR-10 on-off

	Scheduled (on fraction)				Adaptive		Continuous
Attack	33%	50%	67%	80%	0.1	0.4	
Sign-flip	83.82	80.88	79.67	79.38	82.10	84.53	80.01
Label-flip	85.29	82.39	82.68	80.39	84.56	84.80	79.92

Direction-level weighting. A copied benign update can receive a high score yet distort aggregation if each copy gets its own weight. A client-level variant weights clients separately while retaining HGDA's scoring rule and server optimizer. HGDA's advantage grows from 1.75 points under Mimic to 7.69 under adaptive Mimic, which copies the highest-scoring benign update (Table 2(c)). Direction-level weighting controls this repeated influence while retaining the useful direction itself. Appendix D gives the training trajectories.

4.5 Resilience to changing client behavior

On-off attacks test whether temporal assessment remains useful when the same clients switch between benign training and poisoning. Scheduled attacks are active in 33–80% of post-warmup rounds, while adaptive attackers use previous-round client scores to time their switches. On CIFAR-10 at $m = 50\%$ and $W = 100$, the trajectories show repeated returns to high accuracy after attack-induced drops, including under score-adaptive timing (Appendix F). Across these schedules, HGDA reaches 79.38–85.29% best accuracy (Table 2(f)).

5 Limitations

HGDA groups updates with the same direction to control repeated influence without discarding directional information. Extending this grouping to nearby updates would introduce a clustering decision: replacing distinct directions with one representative could suppress useful contributions from benign clients with related but different data distributions. A natural extension is to limit the collective weight of coordinated near-copies while preserving benign directional diversity, rather than simply relaxing the matching rule.

HGDA uses a fixed window measured in global rounds, so the number of retained observations varies with client participation. A longer window provides more evidence for assessment, while a shorter one removes older observations sooner. Adapting this retention period to clients' participation patterns and behavioral changes is a direction for future work.

6 Conclusion

HGDA uses trusted data to assess clients over recent rounds and weights their distinct current directions, leaving the global update scale to the server optimizer. Our convergence analysis relates progress on the benign clients' objective to residual adversarial influence and local-training error. Experiments on five datasets demonstrate broad robustness across poisoning attacks, Byzantine fractions and participation conditions. HGDA also learns effectively from heterogeneous clients under poisoning despite limited trusted-set coverage and shows resilience to scheduled and adaptive on-off attacks.

AI USE STATEMENT

The authors formulated the research question, developed HGDA and designed the experiments. Under their direction, generative AI tools assisted with manuscript drafting and editing, literature retrieval and the derivation and checking of convergence proofs. The tools also supported running author-specified experiments and interpreting and presenting their results. The authors made the final research and editorial decisions and take responsibility for the paper's content.

REFERENCES

Youssef Allouah, Sadegh Farhadkhani, Rachid Guerraoui, Nirupam Gupta, Rafael Pinot, and John Stephan. Fixing by mixing: A recipe for optimal byzantine ML under heterogeneity. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent (eds.), *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 1232–1300. PMLR, 25–27 Apr 2023.

Youssef Allouah, Sadegh Farhadkhani, Rachid Guerraoui, Nirupam Gupta, Rafael Pinot, Geovani Rizk, and Sasha Voitovych. Byzantine-robust federated learning: Impact of client subsampling and local updates. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 1078–1114. PMLR, 21–27 Jul 2024.

Wenxuan Bao, Jun Wu, and Jingrui He. BOBA: byzantine-robust federated learning with label skewness. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li (eds.), *International Conference on Artificial Intelligence and Statistics, 2-4 May 2024, Palau de Congressos, Valencia, Spain*, Proceedings of Machine Learning Research, pp. 892–900. PMLR, 2024.

Gilad Baruch, Moran Baruch, and Yoav Goldberg. A little is enough: Circumventing defenses for distributed learning. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 8632–8642, 2019.

Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. Machine learning with adversaries: Byzantine tolerant gradient descent. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 119–129, 2017.

Keith Bonawitz, Hubert Eichner, Wolfgang Grieskamp, Dzmitry Huba, Alex Ingerman, Vladimir Ivanov, Chloé Kiddon, Jakub Konečný, Stefano Mazzocchi, Brendan McMahan, Timon Van Overveldt, David Petrou, Daniel Ramage, and Jason Roselander. Towards federated learning at scale: System design. In A. Talwalkar, V. Smith, and M. Zaharia (eds.), *Proceedings of Machine Learning and Systems*, volume 1, pp. 374–388, 2019.

Sebastian Caldas, Peter Wu, Tian Li, Jakub Konečný, H. Brendan McMahan, Virginia Smith, and Ameet Talwalkar. LEAF: A benchmark for federated settings. *arXiv preprint arXiv:1812.01097*, 2018.

Xiaoyu Cao, Minghong Fang, Jia Liu, and Neil Zhenqiang Gong. FLTrust: Byzantine-robust federated learning via trust bootstrapping. In *28th Annual Network and Distributed System Security Symposium, NDSS 2021, virtually, February 21-25, 2021*. The Internet Society, 2021.

El-Mahdi El-Mhamdi, Rachid Guerraoui, and Sébastien Rouault. The hidden vulnerability of distributed learning in Byzantium. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3521–3530. PMLR, 10–15 Jul 2018.

Minghong Fang, Xiaoyu Cao, Jinyuan Jia, and Neil Zhenqiang Gong. Local model poisoning attacks to byzantine-robust federated learning. In Srdjan Capkun and Franziska Roesner (eds.),

29th USENIX Security Symposium, USENIX Security 2020, August 12-14, 2020, pp. 1605–1622. USENIX Association, 2020.

Clement Fung, Chris J. M. Yoon, and Ivan Beschastnikh. The limitations of federated learning in sybil settings. In Manuel Egele and Leyla Bilge (eds.), *23rd International Symposium on Research in Attacks, Intrusions and Defenses, RAID 2020, San Sebastian, Spain, October 14-15, 2020*, pp. 301–316. USENIX Association, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 770–778. IEEE Computer Society, 2016.

Sai Praneeth Karimireddy, Lie He, and Martin Jaggi. Learning from history for byzantine robust optimization. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, Proceedings of Machine Learning Research, pp. 5311–5319. PMLR, 2021.

Sai Praneeth Karimireddy, Lie He, and Martin Jaggi. Byzantine-robust learning on heterogeneous datasets via bucketing. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.

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

Ya Le and Xuan Yang. Tiny ImageNet visual recognition challenge. CS231n course project report, Stanford University, 2015.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proc. IEEE*, 86(11):2278–2324, 1998.

Cody Lewis, Vijay Varadharajan, and Nasimul Noman. Attacks against federated learning defense systems and their mitigation. *Journal of Machine Learning Research*, 24(30):1–50, 2023.

Liping Li, Wei Xu, Tianyi Chen, Georgios B. Giannakis, and Qing Ling. RSA: byzantine-robust stochastic aggregation methods for distributed learning from heterogeneous datasets. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pp. 1544–1551. AAAI Press, 2019.

Yuchen Liu, Chen Chen, Lingjuan Lyu, Fangzhao Wu, Sai Wu, and Gang Chen. Byzantine-robust learning on heterogeneous data via gradient splitting. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, Proceedings of Machine Learning Research, pp. 21404–21425. PMLR, 2023.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In Aarti Singh and Xiaojin (Jerry) Zhu (eds.), *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017, 20-22 April 2017, Fort Lauderdale, FL, USA*, Proceedings of Machine Learning Research, pp. 1273–1282. PMLR, 2017.

Hamid Mozaffari, Sunav Choudhary, and Amir Houmansadr. Fake or compromised? making sense of malicious clients in federated learning. In Joaquin Garcia-Alfaro, Rafał Kozik, Michał Choraś, and Sokratis Katsikas (eds.), *Computer Security – ESORICS 2024*, pp. 187–207, Cham, 2024. Springer Nature Switzerland. ISBN 978-3-031-70879-4.

Kaoru Otsuka, Yuki Takezawa, and Makoto Yamada. Delayed momentum aggregation: Communication-efficient byzantine-robust federated learning with partial participation. In *Forty-third International Conference on Machine Learning*, 2026.

Jungwuk Park, Dong-Jun Han, Minseok Choi, and Jaekyun Moon. Sageflow: Robust federated learning against both stragglers and adversaries. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 840–851. Curran Associates, Inc., 2021.

Krishna Pillutla, Sham M. Kakade, and Zaïd Harchaoui. Robust aggregation for federated learning. *IEEE Trans. Signal Process.*, 70:1142–1154, 2022.

Sashank J. Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečný, Sanjiv Kumar, and Hugh Brendan McMahan. Adaptive federated optimization. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

Virat Shejwalkar and Amir Houmansadr. Manipulating the byzantine: Optimizing model poisoning attacks and defenses for federated learning. In *28th Annual Network and Distributed System Security Symposium, NDSS 2021, virtually, February 21-25, 2021*. The Internet Society, 2021.

Virat Shejwalkar, Amir Houmansadr, Peter Kairouz, and Daniel 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, 2022.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017*, pp. 3–18. IEEE Computer Society, 2017.

Cong Xie, Sanmi Koyejo, and Indranil Gupta. Zeno: Distributed stochastic gradient descent with suspicion-based fault-tolerance. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, Proceedings of Machine Learning Research, pp. 6893–6901. PMLR, 2019.

Cong Xie, Oluwasanmi Koyejo, and Indranil Gupta. Fall of empires: Breaking byzantine-tolerant SGD by inner product manipulation. In Ryan P. Adams and Vibhav Gogate (eds.), *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, volume 115 of *Proceedings of Machine Learning Research*, pp. 261–270. PMLR, 2020a.

Cong Xie, Sanmi Koyejo, and Indranil Gupta. Zeno++: Robust fully asynchronous SGD. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, Proceedings of Machine Learning Research, pp. 10495–10503. PMLR, 2020b.

Yueqi Xie, Minghong Fang, and Neil Zhenqiang Gong. Model poisoning attacks to federated learning via multi-round consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15454–15463, June 2025.

Dong Yin, Yudong Chen, Kannan Ramchandran, and Peter Bartlett. Byzantine-robust distributed learning: Towards optimal statistical rates. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 5650–5659. PMLR, 10–15 Jul 2018.

Zaixi Zhang, Xiaoyu Cao, Jinyuan Jia, and Neil Zhenqiang Gong. FLDetector: Defending federated learning against model poisoning attacks via detecting malicious clients. In Aidong Zhang and Huzefa Rangwala (eds.), *KDD '22: The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, August 14 - 18, 2022*, pp. 2545–2555. ACM, 2022.

Shiyuan Zuo, Xingrun Yan, Rongfei Fan, Li Shen, Puning Zhao, Jie Xu, and Han Hu. Efficient federated learning against byzantine attacks and data heterogeneity via aggregating normalized gradients. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 28998–29035. Curran Associates, Inc., 2025.

A Algorithm, implementation and experimental setup

A.1 Complete server-side pseudocode

Algorithm 1 combines reference construction, client assessment and direction-level aggregation into one communication round, following Section 3. The histories $\mathcal{H} = \{\mathcal{H}_{i,h}\}$ store observed cosines indexed by global round and are initialized empty before round 0. The optimizer state o_t contains Adam's moment estimates and update counter.

Algorithm 1 HGDA: one server round

Require: Parameters θ_t, optimizer state o_t, uploads $\{\Delta_i^t : i \in \mathcal{C}_t\}$
Require: Trusted data D_s, checkpoints $E_s < E_l$, window W, regularizer $\lambda > 0$, histories $\mathcal{H}$
Ensure: Updated parameters θ_{t+1}, optimizer state o_{t+1} and histories $\mathcal{H}$

$\ell_t \leftarrow \max(0, t - W + 1)$
Server references
Starting from θ_t, run one SGD trajectory on D_s through epoch E_l
Retain ϕ_{t,E_s} and ϕ_{t,E_l} from this same trajectory
$q_h^t \leftarrow (\phi_{t,E_h} - \theta_t)/\|\phi_{t,E_h} - \theta_t\|_2$ for $h \in \{s, l\}$
Recent alignment scores
for each $i \in \mathcal{C}_t$ **do**
  $u_i^t \leftarrow \Delta_i^t/\|\Delta_i^t\|_2$ if $\|\Delta_i^t\|_2 > 0$, otherwise $u_i^t \leftarrow 0$
  for each $h \in \{s, l\}$ **do**
    $r_{i,h}^t \leftarrow \langle u_i^t, q_h^t \rangle$ and $\mathcal{H}_{i,h}[t] \leftarrow r_{i,h}^t$
    $\mathcal{W}_{i,t}^{(W)} \leftarrow \{\tau : \ell_t \le \tau \le t,\ \mathcal{H}_{i,h}[\tau] \text{ is recorded}\}$
    $P_{i,h}^{t,W} \leftarrow \sum_{\tau \in \mathcal{W}_{i,t}^{(W)}} [\mathcal{H}_{i,h}[\tau]]_+$
    $N_{i,h}^{t,W} \leftarrow \sum_{\tau \in \mathcal{W}_{i,t}^{(W)}} [-\mathcal{H}_{i,h}[\tau]]_+$
    $z_{i,h}^{t,W} \leftarrow (P_{i,h}^{t,W} - N_{i,h}^{t,W})/\sqrt{P_{i,h}^{t,W} + N_{i,h}^{t,W} + 2\lambda}$
  end for
  $s_i^t \leftarrow [\min\{z_{i,s}^{t,W}, z_{i,l}^{t,W}\}]_+$
end for
Current-direction aggregation
$\mathcal{S}_t \leftarrow \{i \in \mathcal{C}_t : \|\Delta_i^t\|_2 > 0,\ s_i^t > 0\}$
if $\mathcal{S}_t = \varnothing$ **then**
  $(\theta_{t+1}, o_{t+1}) \leftarrow (\theta_t, o_t)$
else
  Partition $\mathcal{S}_t$ into $\mathcal{G}_t$ by direction equality $u_i^t = u_j^t$
  for each $G \in \mathcal{G}_t$ **do**
    $v_G^t \leftarrow u_i^t$ for any $i \in G$, and $b_G^t \leftarrow \max_{i \in G} s_i^t$
  end for
  $\pi_G^t \leftarrow b_G^t / \sum_{H \in \mathcal{G}_t} b_H^t$ for each $G \in \mathcal{G}_t$
  $\bar{u}_t \leftarrow \sum_{G \in \mathcal{G}_t} \pi_G^t v_G^t$
  $(\theta_{t+1}, o_{t+1}) \leftarrow \textsc{ServerStep}(\theta_t, o_t, \bar{u}_t)$
end if
return $\theta_{t+1}, o_{t+1}, \mathcal{H}$

Here $[x]_+ = \max(x, 0)$ and ServerStep applies Adam to the aggregated client directions without renormalizing $\bar{u}_t$. Each participating client records its current alignments before aggregation, so the histories advance even when $\mathcal{S}_t$ is empty. Nonparticipating clients add no observations, and entries older than ℓ_t do not enter the sums. Appendix A.2 details the server update and the separate treatment of model buffers.

A.2 Implementation details

From client scores to direction weights. The implementation first normalizes the scores of clients with nonzero current updates. Let $\widetilde{s}_i^t = s_i^t$ for such an update and zero otherwise. When the total eligible score is positive, define

$$a_i^t = \frac{\widetilde{s}_i^t}{\sum_{j \in \mathcal{C}_t} \widetilde{s}_j^t}.$$

Thus, a_i^t is client i's share of the total score before direction grouping. Each direction then receives the largest a_i^t among the clients submitting it. Normalizing these group scores cancels the common client-level denominator and gives Equation (5). The adaptive on-off protocol in Appendix B.2 uses a_i^t to track changes in the attacking clients' scores.

Server optimizer. The server uses Adam to turn $\bar{u}_t$ into a global parameter update. Let k count the server updates applied so far, and let $\mathbf{m}_k$ and $\mathbf{v}_k$ denote Adam's first- and second-moment estimates, initialized at zero. For the next update, increment k and compute

$$\begin{aligned} \mathbf{m}_k &= \beta_1 \mathbf{m}_{k-1} + (1-\beta_1)\bar{u}_t, \quad \mathbf{v}_k = \beta_2 \mathbf{v}_{k-1} + (1-\beta_2)\bar{u}_t^{\odot 2}, \\ \widehat{\mathbf{m}}_k &= \frac{\mathbf{m}_k}{1-\beta_1^k}, \qquad\qquad \widehat{\mathbf{v}}_k = \frac{\mathbf{v}_k}{1-\beta_2^k}, \\ \theta_{t+1} &= \theta_t + \eta_t \frac{\widehat{\mathbf{m}}_k}{\sqrt{\widehat{\mathbf{v}}_k} + \varepsilon}. \end{aligned}$$

Squaring, square roots and division of vectors are elementwise. The step is added to θ_t because the client updates are parameter increments from that model. The server learning rate η_t follows the task-specific schedule in Appendix A.3. If no update has positive eligible score, the parameters, moments and update counter remain unchanged.

Model buffers. Model buffers, such as running batch-normalization statistics, are separate from the trainable parameters. We average them using client sample-count weights and publish them in every round. Buffer values are converted to floating point for averaging and then cast back to their original data types.

Histories and computation. Both references come from one server-training trajectory. Normalizing client updates and comparing them with the references costs $O(n_t d)$ per round, in addition to reference training. The implementation stores the two alignment histories in scalar arrays of size $N \times T$, where T is the allocated training horizon, using $O(NT)$ space. To score participating clients, it reads only their entries from the most recent W global rounds, at a cost of $O(n_t W)$. Direction grouping compares the current normalized vectors for equality.

Numerical handling. A client increment with norm at most 10^{-12} is treated as zero and records zero cosine for that round. An eligible-score sum at most 10^{-12} leaves the parameter and optimizer states unchanged. The implementation requires nonzero server references and a nonzero directional aggregate when there is positive support.

A.3 Data and training settings

Table A.1: Datasets, models and participation settings.

Dataset	Model	Participants	Rounds	Seeds
FEMNIST	CNN	100/3400	500	3
Purchase-100	MLP	50/1000	500	3
Tiny ImageNet	ResNet-50	10/100	500	3
MNIST	LeNet	100/100	200	3
CIFAR-10	ResNet-20	100/100	1500	1

Models and client partitions. Table A.1 summarizes the five tasks. Its Participants column gives the number selected per round followed by the total client population. MNIST uses a LeNet-style model (LeCun et al., 1998), CIFAR-10 uses ResNet-20 (Krizhevsky, 2009; He et al., 2016), FEMNIST uses a two-convolution CNN, Purchase-100 uses an MLP and Tiny ImageNet uses ResNet-50. All models are trained from scratch. FEMNIST retains the natural writer partition, with one client per writer. The other tasks use Dirichlet label skew with concentration 0.5.

Table A.2: Model architectures.

Task	Architecture
MNIST	LeNet-style convolutional classifier
CIFAR-10	ResNet-20
FEMNIST	Two convolution layers, widths 16 and 32, then FC 512 and 62
Purchase-100	MLP: $600 \to 1024 \to 100$
Tiny ImageNet	ResNet-50

Trusted data and evaluation. Before training, we reserve a server pool P_0 from the dataset's test split $\mathcal{T}$ and select the trusted dataset $D_s \subset P_0$ from that pool. Clients use the training split, and test accuracy is evaluated on the remaining examples $\mathcal{T} \setminus P_0$. MNIST and CIFAR-10 use a balanced 1000-example server pool and 100 trusted examples, with ten trusted examples per class. On the other three tasks, HGDA's trusted examples are sampled without class quotas. Table A.3 gives the pool and trusted-set sizes, and Appendix E examines the resulting class coverage and learning on classes absent from D_s.

Table A.3: Learning rates and trusted-set sizes.

Task	$\lvert P_0\rvert$	$\lvert D_s\rvert$	Local LR	Server Adam LR	W
MNIST	1000	100	0.01	0.002	10, 100
CIFAR-10	1000	100	0.01	0.002	10, 100
FEMNIST	6200	620	$0.3 \to 0$	0.007	100
Purchase-100	1000	100	$0.3 \to 0$	$0.015 \to 0.001$	100
Tiny ImageNet	2000	100	0.4	$0.015 \to 0.001$	100

Local training and server optimization. All main protocols use five local SGD epochs, batch size 64 and gradient-norm clipping at 10, without local momentum or weight decay. Arrows in Table A.3 denote cosine learning-rate schedules. FEMNIST and Purchase use a local schedule from 0.3 to 0 with a cosine horizon of 501 rounds for their 500-round training budgets. Server-reference SGD follows the same learning-rate schedule as client training.

HGDA's server Adam uses $(\beta_1, \beta_2) = (0.8, 0.9)$ and $\varepsilon = 10^{-4}$. Its learning rate is constant on MNIST, CIFAR-10 and FEMNIST. On Purchase and Tiny ImageNet, it follows a cosine schedule from 0.015 to 0.001 over 500 rounds. The score regularizer is $\lambda = 1$, and the history windows are listed in Table A.3.

Communication-round budgets. We compare aggregation rules under matched attacks and Byzantine fractions, using a common maximum round budget within each task. Learning speed depends on the client data distribution, local training and how each rule selects, weights and scales updates. A shared horizon gives slower methods time to learn under the same client-side protocol. The 1500-round CIFAR-10 horizon is also used in previous Byzantine-robust FL evaluations (Cao et al., 2021; Allouah et al., 2024).

We measure each run's best test accuracy within the budget in Table A.1. MNIST uses rounds 0–199, CIFAR-10 uses rounds 0–1499 and the other main grids use rounds 0–499. HGDA begins scoring from the first round, using the available observations as its history window fills. Appendix C.7 reports learning progress at shorter cutoffs on MNIST and CIFAR-10.

Reference checkpoint selection. The early reference is obtained after one epoch on the trusted dataset, so $E_s = 1$. We choose the later checkpoint E_l to approximately match the number of SGD steps taken by an average client during local training. Matching step counts rather than epoch counts accounts for differences between server and client dataset sizes. For batch size B, local epochs E and total client training size n_{train}, we set

$$E_l = \max\left(1, \text{round}\left(\frac{E\lceil n_{\text{train}}/(NB)\rceil}{\lceil |D_s|/B\rceil}\right)\right).$$

The numerator approximates the number of local SGD steps for an average client, and the denominator is the number of SGD steps in one epoch on D_s. The reference pair is fixed across attacks within each task, with $E_l > E_s$. Single-reference ablations use one checkpoint. Appendix E.4 evaluates single- and dual-reference assessment and varies checkpoint placement.

Server learning-rate selection. We selected the Purchase server schedule by comparing six candidates on clean training with seed 2026, followed by three-seed comparisons of four candidates on clean training and on the four attacks at $m = 40\%$. The selected cosine schedule, $0.015 \to 0.001$ over 500 rounds, is used for clean training, the attack comparisons, the class-coverage analysis and the reference-checkpoint study.

A.4 Compared methods and feasibility

FedAvg and the benign-only reference. FedAvg averages client models with weights proportional to their local sample counts (McMahan et al., 2017). FedAvg (benign only) applies this rule to the benign participants in each round, using the known Byzantine identities to exclude malicious uploads and renormalizing the weights over the remaining clients. Its accuracy serves as an empirical indicator of learning from the available benign data under the corresponding data partition and participation setting. As the Byzantine fraction increases, this reference helps interpret how well each defense sustains learning from a smaller benign population. We include it for context and exclude it from the method rankings.

Aggregation from client updates. Our Multi-Krum implementation averages candidates with the smallest nearest-neighbor distance scores (Blanchard et al., 2017). Bulyan combines Krum candidate selection with coordinate filtering (El-Mhamdi et al., 2018). Coordinate-wise median takes the median in each parameter coordinate, while trimmed mean removes f_t values from each tail before averaging (Yin et al., 2018). Here f_t is the actual Byzantine count among round-t participants. RFA uses a smoothed Weiszfeld iteration to approximate the geometric median (Pillutla et al., 2022). Fed-NGA combines whole-model normalized increments using client sample-count weights (Zuo et al., 2025). We use the schedule $\eta_t = \eta_0/\sqrt{0.002t + 1}$, with $\eta_0 = 1$ except for Tiny ImageNet, where $\eta_0 = 7$.

Methods using trusted data. FLTrust scores client updates against a reference trained on trusted server data (Cao et al., 2021). Our FLTrust-Early and FLTrust-Late comparisons use the references at E_s and E_l, respectively, for scoring. Both rescale client updates to the norm of the later server reference before aggregation. BOBA requires server examples from every class and enough eligible clients for its class-conditioned subspace (Bao et al., 2024). The FEMNIST BOBA comparison uses ten server examples per class to meet this requirement. BOBA is not part of the Purchase and Tiny ImageNet main grids, whose cohort sizes are smaller than their numbers of classes.

History-based detection. FLDetector+Median uses history-based detection followed by median aggregation and can restart training after removing detected clients. We evaluate it on the full-participation tasks, following its protocol of predicting each client's update from consecutive rounds (Zhang et al., 2022).

Byzantine counts and client sampling. For Multi-Krum, Bulyan, trimmed mean and BOBA, we supply the true Byzantine count f_t among the clients participating in each round. This oracle information is favorable to these baselines because it eliminates Byzantine-count estimation error. Under partial participation, feasibility checks resample cohorts for Bulyan, trimmed mean and Multi-Krum when their aggregation rules cannot be evaluated. Appendix C gives the corresponding participant-count requirements and explains every N/A, NR and DNF condition by dataset.

B ATTACK PROTOCOLS

Byzantine clients alter local training or submitted parameter increments to reduce the global model's overall predictive accuracy. Their identities are fixed within each run, while the participating subset can change under partial participation. Client sampling follows Appendix C. All update increments below are measured relative to the current global model θ_t.

B.1 STATIC ATTACKS

Sign-flip and label-flip. Sign-flip first performs local training, then replaces the resulting increment Δ_i^t by $-\Delta_i^t$. Label-flip instead poisons local training data by transforming each label as $y \mapsto C-1-y$, where C is the number of classes, before local SGD (Fang et al., 2020; Li et al., 2019).

Gaussian and SameValue. Gaussian replaces each malicious increment with independent zero-mean Gaussian noise of variance 200 in every coordinate. SameValue replaces it with $100\mathbf{1}_d$, where $\mathbf{1}_d$ is the all-ones vector.

Attacks constructed from benign updates. ALIE, IPM, Min-Sum, Min-Max and Mimic use the benign updates from the current round to construct malicious submissions. Let $\mathcal{B}_t$ denote the benign participants, and let μ_t and σ_t be the coordinate-wise mean and sample standard deviation of their increments $\{\Delta_j^t : j \in \mathcal{B}_t\}$.

ALIE and IPM. ALIE makes the Byzantine clients submit $\mu_t - z_t\sigma_t$, shifting the benign mean along coordinates with greater variation (Baruch et al., 2019). The shift parameter is chosen from a normal quantile using the current participant counts. With n_t participants, f_t Byzantine clients and $b_t = n_t - f_t$ benign clients, we use

$$h_t = \lfloor n_t/2 \rfloor + 1 - f_t, \qquad z_t = \Phi^{-1}\left(\frac{b_t - h_t}{b_t}\right).$$

Here h_t counts the benign clients needed alongside the Byzantine clients for a strict majority, and Φ is the standard normal cumulative distribution function. IPM instead submits $-\mu_t$, opposing the mean benign update (Xie et al., 2020a).

Min-Max and Min-Sum. Both attacks construct a common malicious increment $x_t(\gamma) = \mu_t - \gamma\sigma_t$ and use binary search to increase the shift while satisfying a distance constraint (Shejwalkar & Houmansadr, 2021). Min-Max bounds the distance from the malicious increment to every benign increment by the largest distance between benign increments:

$$\max_{j\in\mathcal{B}_t} \|x_t(\gamma) - \Delta_j^t\|_2^2 \le \max_{i,j\in\mathcal{B}_t} \|\Delta_i^t - \Delta_j^t\|_2^2.$$

Min-Sum bounds the total squared distance to the benign increments by the smallest such total attained by a benign increment:

$$\sum_{j\in\mathcal{B}_t} \|x_t(\gamma) - \Delta_j^t\|_2^2 \le \min_{i\in\mathcal{B}_t} \sum_{j\in\mathcal{B}_t} \|\Delta_i^t - \Delta_j^t\|_2^2.$$

The resulting submissions are displaced from the benign mean while remaining within these bounds on benign-update distances.

Mimic. All Byzantine clients copy the same benign client's increment in the current round (Karimireddy et al., 2022). The copied update can be useful on its own, but repeated submissions can increase its relative influence under client-level aggregation. This attack therefore tests how aggregation handles repeated contributions under heterogeneous client data.

Attack coverage. All five datasets are evaluated under sign-flip, label-flip, Mimic and IPM. MNIST and CIFAR-10 also use Gaussian, SameValue, ALIE, Min-Sum and Min-Max. On these two datasets, ALIE uses $m = 20\%, 40\%, 50\%$, and the other static attacks use $m = 20\%, 40\%, 50\%, 60\%, 80\%$. The partial-participation grids are given in Appendices C.3–C.5.

B.2 Adaptive Mimic and on-off schedules

Adaptive Mimic. Adaptive Mimic uses HGDA's current client scores to choose which benign update to copy. All Byzantine clients submit the update of the highest-scoring benign client, concentrating their copies on a direction with strong score support. For FLTrust, the copied benign update is selected using its corresponding trust score. Section 4.4 and Figure A.3 compare the resulting influence under client-level and direction-level aggregation.

Scheduled on-off attacks. On-off attacks alternate between benign training and poisoning (Lewis et al., 2023). During off rounds, Byzantine clients train on clean labels and submit ordinary updates. During on rounds, they use sign-flip or label-flip. The scheduled variants start with 20 benign warm-up rounds, then repeat cycles with (on, off) lengths $(1, 2)$, $(1, 1)$, $(2, 1)$ or $(4, 1)$. These schedules attack in 33%, 50%, 67% and 80% of post-warmup rounds, respectively. Continuous sign-flip and label-flip provide the always-on comparisons.

Score-adaptive on-off attacks. The adaptive variants choose when to switch using the preceding round's client scores rather than a fixed cycle. Let $\mathcal{M}_t$ be the Byzantine participants and a_i^t the normalized client scores defined in Appendix A.2. Their relative score share is

$$\rho_t = \frac{\sum_{i \in \mathcal{M}_t} a_i^t}{|\mathcal{M}_t|/|\mathcal{C}_t|},$$

The numerator is the Byzantine clients' share of the total score before direction grouping, and the denominator is their share of the participating clients. Thus, $\rho_t = 1$ corresponds to the score share they would receive under uniform client weighting. The Byzantine clients switch off together when this relative share falls below τ_{off}, then resume attacking when it rises above τ_{on}. The decision for round t uses ρ_{t-1}, and the current phase is retained between the thresholds. We use $\tau_{\text{on}} = 0.8$ and $\tau_{\text{off}} \in \{0.1, 0.4\}$. Appendix F reports the training trajectories and phase-specific accuracies and explains how entering and expiring observations change the scores during these cycles.

C Complete robustness matrices

The following matrices give the complete results for the training settings in Appendix A.3 and the attacks in Appendix B. Tables report best test accuracy, averaged over repeated runs, with standard deviations for MNIST and the partial-participation tasks. Bold and underline indicate the highest and second-highest means, excluding FedAvg (benign only).

Participant counts and table markers. The Byzantine fraction m in a table heading refers to the full client population. In round t, n_t clients participate and f_t of them are Byzantine. Under full participation, these counts remain fixed unless a defense removes clients. Partial participation allows f_t to vary across rounds even with a fixed population. We call mn_t the nominal Byzantine count for a cohort of size n_t. The tables distinguish three reasons for a missing accuracy:

- **N/A**: the condition was excluded from the evaluation grid because its nominal cohort violates the method's participant-count requirement.
- **NR**: the condition was not run despite satisfying the nominal participant-count requirement. This marker occurs only for Purchase-100 Bulyan at $m = 20\%$, as explained in Appendix C.4.
- **DNF**: training was launched, but at least one run stopped before completing the evaluation budget. The dataset-specific notes below identify the operation that caused each interruption.

Requirements of the aggregation rules. Three methods account for the participant-count exclusions in these tables.

- **Bulyan** first selects $n_t - 2f_t$ candidate updates, then averages $n_t - 4f_t$ values near the median in each coordinate. Its two-stage procedure requires $n_t \geq 4f_t + 3$ (El-Mhamdi et al., 2018).

- **Trimmed mean** removes the f_t largest and f_t smallest values in each coordinate (Yin et al., 2018). Averaging the remaining values requires $n_t - 2f_t > 0$.
- **BOBA** fits a $(C-1)$-dimensional affine subspace to $n_t - f_t$ selected updates, where C is the number of classes (Bao et al., 2024). At least C updates are needed to support this affine dimension, giving the requirement $n_t - f_t \geq C$. This is separate from its requirement for trusted examples from every class.

Sampling under partial participation. Each candidate cohort is sampled without replacement from the fixed client population. For evaluated Bulyan and trimmed-mean conditions, a cohort that violates the corresponding requirement above is redrawn. Multi-Krum similarly requires $n_t - f_t - 2 > 0$ to compute its nearest-neighbor scores. The sampler also checks that IPM and Mimic have a benign update from which to construct their attacks. Consequently, the accepted cohorts for these methods follow the sampling distribution conditioned on their respective requirements. The notes below specify where this conditioning permits results at population fractions above a method's per-round threshold, and where conditions were left outside the evaluation grid.

C.1 MNIST

All 100 clients participate in each of the 200 training rounds.

Bulyan and trimmed mean: N/A conditions. With 100 participants, Bulyan requires $f_t \leq 24$, and trimmed mean requires $f_t \leq 49$. For sign-flip, label-flip, Gaussian, SameValue, Mimic, Min-Sum, Min-Max and IPM, Bulyan is therefore N/A at $m = 40\%, 50\%, 60\%, 80\%$, and trimmed mean is N/A at $m = 50\%, 60\%, 80\%$. ALIE is evaluated at $m = 20\%, 40\%, 50\%$, so its Bulyan entries are N/A at 40% and 50%, and its trimmed-mean entry is N/A at 50%. These exclusions follow directly from the fixed Byzantine count under full participation.

FLDetector+Median: DNF conditions. FLDetector predicts client updates from their recent history, detects clients whose received updates differ from those predictions and restarts training after removing the detected clients. The five DNF entries arise at two different stages of this procedure:

- **SameValue at** $m = 60\%$ **and** 80% (Table A.8). A majority of clients submit the same constant update, so coordinate-wise median aggregation repeatedly returns that update. In these runs, the compact linear system used for the detector's L-BFGS prediction is singular. Training stops when prediction begins.
- **ALIE at** $m = 40\%$ **and** 50% (Table A.9). The detector removes benign clients while retaining the Byzantine clients, leaving a Byzantine majority for the restarted training phase. ALIE's quantile-based shift depends on the benign and Byzantine counts and is no longer valid for this remaining population. The restarted run therefore stops when constructing the attack.
- **Min-Max at** $m = 80\%$ (Table A.12). In one run, the detector removes all benign clients, leaving 80 Byzantine clients for the restart. Min-Max constructs its perturbation using distances among benign updates. With no benign updates remaining, that construction cannot proceed and the run terminates.

Table A.4: MNIST without attack. Best test accuracy (%).

Method	0%
FedAvg (benign only)	97.56 ± 0.07
Multi-Krum	97.30 ± 0.08
Bulyan	97.30 ± 0.08
Coord. median	96.30 ± 0.12
Trimmed mean	97.30 ± 0.08
RFA	97.43 ± 0.07
Fed-NGA	$\underline{98.44} \pm 0.07$
FLTrust-Early	97.11 ± 0.45
FLTrust-Late	96.83 ± 0.68
BOBA	97.30 ± 0.08
FLDetector+Median	96.30 ± 0.12
HGDA ($W = 10$)	98.31 ± 0.31
HGDA ($W = 100$)	$\mathbf{98.76} \pm 0.07$

Table A.5: MNIST under sign-flip. Best test accuracy (%).

Method	20%	40%	50%	60%	80%
FedAvg (benign only)	97.57 ± 0.05	97.55 ± 0.09	97.44 ± 0.13	97.48 ± 0.11	96.93 ± 0.17
Multi-Krum	95.31 ± 0.55	82.10 ± 15.01	68.14 ± 8.34	14.27 ± 4.48	11.66 ± 2.75
Bulyan	93.95 ± 0.49	N/A	N/A	N/A	N/A
Coord. median	94.66 ± 0.25	88.09 ± 1.20	32.20 ± 3.10	11.16 ± 1.46	7.96 ± 1.83
Trimmed mean	94.97 ± 0.14	88.00 ± 1.19	N/A	N/A	N/A
RFA	96.04 ± 0.27	90.85 ± 0.92	38.69 ± 26.92	12.64 ± 3.99	7.30 ± 2.57
Fed-NGA	$\underline{97.90} \pm 0.05$	96.54 ± 0.22	48.64 ± 34.73	12.53 ± 3.81	5.49 ± 2.67
FLTrust-Early	95.86 ± 1.89	95.09 ± 1.53	94.38 ± 2.04	91.74 ± 2.52	88.85 ± 0.60
FLTrust-Late	94.34 ± 2.97	92.55 ± 2.22	91.84 ± 2.39	89.77 ± 0.43	88.84 ± 0.64
BOBA	95.01 ± 0.33	78.60 ± 21.56	69.17 ± 13.40	76.19 ± 21.29	65.60 ± 15.99
FLDetector+Median	94.66 ± 0.25	88.09 ± 1.20	31.99 ± 3.46	12.10 ± 1.56	7.96 ± 1.83
HGDA ($W = 10$)	97.89 ± 0.42	$\underline{97.85} \pm 0.29$	$\underline{97.73} \pm 0.34$	$\underline{97.60} \pm 0.19$	$\underline{96.96} \pm 0.46$
HGDA ($W = 100$)	$\mathbf{98.68} \pm 0.04$	$\mathbf{98.56} \pm 0.08$	$\mathbf{98.33} \pm 0.23$	$\mathbf{98.12} \pm 0.28$	$\mathbf{97.59} \pm 0.46$

Table A.6: MNIST under label-flip. Best test accuracy (%).

Method	20%	40%	50%	60%	80%
FedAvg (benign only)	97.57 ± 0.05	97.55 ± 0.09	97.44 ± 0.13	97.48 ± 0.11	96.93 ± 0.17
Multi-Krum	97.27 ± 0.10	97.18 ± 0.14	96.97 ± 0.18	22.66 ± 6.07	12.88 ± 4.28
Bulyan	95.70 ± 0.30	N/A	N/A	N/A	N/A
Coord. median	96.22 ± 0.20	95.91 ± 0.25	64.53 ± 21.47	15.75 ± 2.53	11.26 ± 2.34
Trimmed mean	96.68 ± 0.08	95.91 ± 0.30	N/A	N/A	N/A
RFA	97.53 ± 0.12	97.03 ± 0.03	48.67 ± 32.36	15.94 ± 4.79	10.01 ± 0.91
Fed-NGA	97.84 ± 0.13	97.80 ± 0.32	85.62 ± 15.31	96.61 ± 0.64	18.37 ± 8.97
FLTrust-Early	96.91 ± 0.39	96.74 ± 0.30	96.54 ± 0.16	96.56 ± 0.31	95.82 ± 0.42
FLTrust-Late	96.53 ± 0.47	96.50 ± 0.57	96.43 ± 0.33	96.56 ± 0.20	95.99 ± 0.34
BOBA	97.11 ± 0.21	96.11 ± 0.87	95.01 ± 2.15	92.41 ± 1.87	54.24 ± 34.01
FLDetector+Median	96.22 ± 0.20	95.91 ± 0.25	64.53 ± 21.47	15.75 ± 2.53	11.26 ± 2.34
HGDA ($W = 10$)	$\underline{98.10} \pm 0.45$	$\underline{98.26} \pm 0.27$	$\mathbf{98.52} \pm 0.06$	$\underline{98.30} \pm 0.27$	$\underline{97.69} \pm 0.22$
HGDA ($W = 100$)	$\mathbf{98.67} \pm 0.11$	$\mathbf{98.65} \pm 0.06$	$\underline{98.47} \pm 0.24$	$\mathbf{98.36} \pm 0.07$	$\mathbf{97.83} \pm 0.34$

Table A.7: MNIST under Gaussian. Best test accuracy (%).

Method	20%	40%	50%	60%	80%
FedAvg (benign only)	97.57 ± 0.05	97.55 ± 0.09	97.44 ± 0.13	97.48 ± 0.11	96.93 ± 0.17
Multi-Krum	97.27 ± 0.09	97.23 ± 0.06	97.16 ± 0.24	97.18 ± 0.06	96.63 ± 0.24
Bulyan	95.79 ± 0.17	N/A	N/A	N/A	N/A
Coord. median	96.39 ± 0.16	96.43 ± 0.14	96.30 ± 0.19	96.42 ± 0.13	24.16 ± 4.49
Trimmed mean	96.88 ± 0.08	96.54 ± 0.16	N/A	N/A	N/A
RFA	97.45 ± 0.06	97.41 ± 0.11	97.34 ± 0.15	97.35 ± 0.06	96.84 ± 0.20
Fed-NGA	$\underline{98.26} \pm 0.07$	98.10 ± 0.02	97.97 ± 0.16	97.67 ± 0.29	96.83 ± 0.26
FLTrust-Early	97.09 ± 0.45	97.01 ± 0.45	97.00 ± 0.33	96.99 ± 0.52	96.61 ± 0.24
FLTrust-Late	96.74 ± 0.69	96.63 ± 0.79	96.73 ± 0.42	96.79 ± 0.43	96.12 ± 0.32
BOBA	97.27 ± 0.09	97.23 ± 0.06	97.16 ± 0.24	97.18 ± 0.06	96.63 ± 0.24
FLDetector+Median	95.04 ± 0.18	95.03 ± 0.17	94.84 ± 0.21	95.01 ± 0.33	94.36 ± 0.78
HGDA ($W = 10$)	98.07 ± 0.35	$\underline{98.27} \pm 0.30$	$\underline{98.03} \pm 0.19$	$\underline{97.74} \pm 0.33$	$\underline{97.39} \pm 0.28$
HGDA ($W = 100$)	$\mathbf{98.69} \pm 0.05$	$\mathbf{98.64} \pm 0.09$	$\mathbf{98.49} \pm 0.17$	$\mathbf{98.32} \pm 0.08$	$\mathbf{97.88} \pm 0.08$

Table A.8: MNIST under SameValue. Best test accuracy (%).

Method	20%	40%	50%	60%	80%
FedAvg (benign only)	97.57 ± 0.05	97.55 ± 0.09	97.44 ± 0.13	97.48 ± 0.11	96.93 ± 0.17
Multi-Krum	97.27 ± 0.09	97.23 ± 0.06	19.88 ± 0.24	12.74 ± 6.09	19.78 ± 0.06
Bulyan	95.76 ± 0.14	N/A	N/A	N/A	N/A
Coord. median	96.03 ± 0.16	88.55 ± 1.58	65.99 ± 2.37	12.74 ± 6.09	12.74 ± 6.09
Trimmed mean	95.36 ± 0.15	85.99 ± 1.56	N/A	N/A	N/A
RFA	97.10 ± 0.33	93.47 ± 0.46	19.90 ± 0.23	13.09 ± 5.95	13.54 ± 5.46
Fed-NGA	92.91 ± 2.95	89.54 ± 0.59	90.22 ± 1.86	84.74 ± 2.90	24.16 ± 6.92
FLTrust-Early	97.06 ± 0.53	96.94 ± 0.56	96.69 ± 0.78	96.60 ± 1.04	95.97 ± 1.07
FLTrust-Late	96.61 ± 0.93	96.51 ± 1.05	96.59 ± 0.76	96.57 ± 0.97	95.97 ± 0.97
BOBA	97.27 ± 0.09	97.23 ± 0.06	97.16 ± 0.24	97.18 ± 0.06	96.63 ± 0.24
FLDetector+Median	96.03 ± 0.17	86.44 ± 2.39	57.01 ± 1.12	DNF	DNF
HGDA ($W = 10$)	$\underline{98.07} \pm 0.38$	$\underline{98.25} \pm 0.28$	$\underline{98.04} \pm 0.19$	$\underline{97.78} \pm 0.27$	$\underline{97.44} \pm 0.20$
HGDA ($W = 100$)	$\mathbf{98.71} \pm 0.07$	$\mathbf{98.65} \pm 0.08$	$\mathbf{98.49} \pm 0.17$	$\mathbf{98.33} \pm 0.06$	$\mathbf{97.87} \pm 0.05$

Table A.9: MNIST under ALIE. Best test accuracy (%).

Method	20%	40%	50%
FedAvg (benign only)	97.57 ± 0.05	97.55 ± 0.09	97.44 ± 0.13
Multi-Krum	96.86 ± 0.23	92.19 ± 0.62	24.47 ± 2.54
Bulyan	96.30 ± 0.40	N/A	N/A
Coord. median	96.30 ± 0.26	88.84 ± 0.59	24.47 ± 2.54
Trimmed mean	96.91 ± 0.11	90.16 ± 0.55	N/A
RFA	96.66 ± 0.40	89.78 ± 1.17	24.48 ± 2.54
Fed-NGA	$\underline{98.43} \pm 0.02$	94.89 ± 0.42	87.74 ± 1.66
FLTrust-Early	97.25 ± 0.35	96.13 ± 0.63	95.28 ± 1.37
FLTrust-Late	96.98 ± 0.64	96.07 ± 0.64	96.10 ± 0.55
BOBA	96.92 ± 0.20	96.00 ± 0.29	97.04 ± 0.25
FLDetector+Median	96.30 ± 0.26	DNF	DNF
HGDA ($W = 10$)	98.19 ± 0.32	$\underline{98.25} \pm 0.30$	$\underline{98.10} \pm 0.16$
HGDA ($W = 100$)	$\mathbf{98.69} \pm 0.01$	$\mathbf{98.63} \pm 0.06$	$\mathbf{98.46} \pm 0.17$

Table A.10: MNIST under Mimic. Best test accuracy (%).

Method	20%	40%	50%	60%	80%
FedAvg (benign only)	97.57 ± 0.05	97.55 ± 0.09	97.44 ± 0.13	97.48 ± 0.11	96.93 ± 0.17
Multi-Krum	96.49 ± 0.18	92.50 ± 1.31	67.56 ± 11.21	67.56 ± 11.21	54.89 ± 13.06
Bulyan	93.51 ± 1.51	N/A	N/A	N/A	N/A
Coord. median	96.12 ± 0.25	92.32 ± 1.05	67.56 ± 11.21	67.56 ± 11.21	54.89 ± 13.06
Trimmed mean	96.51 ± 0.23	92.80 ± 1.08	N/A	N/A	N/A
RFA	97.09 ± 0.08	87.23 ± 3.85	67.57 ± 11.22	67.56 ± 11.21	54.89 ± 13.06
Fed-NGA	$\underline{98.28} \pm 0.22$	98.11 ± 0.17	97.50 ± 0.22	97.09 ± 0.32	93.50 ± 1.93
FLTrust-Early	96.89 ± 0.35	96.71 ± 0.60	97.01 ± 0.32	97.07 ± 0.31	95.31 ± 0.23
FLTrust-Late	96.84 ± 0.65	96.70 ± 0.72	97.04 ± 0.11	97.09 ± 0.20	95.46 ± 0.21
BOBA	96.57 ± 0.14	93.65 ± 1.08	67.66 ± 11.24	67.55 ± 11.26	54.90 ± 13.14
FLDetector+Median	95.08 ± 0.86	89.37 ± 4.20	66.37 ± 12.08	66.44 ± 12.09	52.61 ± 14.36
HGDA ($W = 10$)	98.10 ± 0.34	$\underline{98.27} \pm 0.28$	$\underline{98.00} \pm 0.13$	$\underline{97.77} \pm 0.29$	$\underline{97.47} \pm 0.17$
HGDA ($W = 100$)	$\mathbf{98.69} \pm 0.08$	$\mathbf{98.63} \pm 0.07$	$\mathbf{98.48} \pm 0.14$	$\mathbf{98.31} \pm 0.07$	$\mathbf{97.82} \pm 0.08$

Table A.11: MNIST under Min-Sum. Best test accuracy (%).

Method	20%	40%	50%	60%	80%
FedAvg (benign only)	97.57 ± 0.05	97.55 ± 0.09	97.44 ± 0.13	97.48 ± 0.11	96.93 ± 0.17
Multi-Krum	96.88 ± 0.18	95.58 ± 0.37	94.76 ± 1.11	94.29 ± 0.87	94.83 ± 1.27
Bulyan	95.41 ± 0.77	N/A	N/A	N/A	N/A
Coord. median	95.69 ± 0.37	94.92 ± 0.94	94.76 ± 1.11	94.29 ± 0.87	94.82 ± 1.27
Trimmed mean	96.81 ± 0.25	95.00 ± 1.02	N/A	N/A	N/A
RFA	96.70 ± 0.37	94.90 ± 0.89	94.76 ± 1.11	94.29 ± 0.87	94.83 ± 1.26
Fed-NGA	$\underline{98.28} \pm 0.06$	97.10 ± 0.07	95.32 ± 0.48	94.10 ± 0.98	91.07 ± 0.65
FLTrust-Early	96.78 ± 0.52	95.53 ± 2.58	94.32 ± 2.98	94.11 ± 3.51	92.22 ± 1.37
FLTrust-Late	96.71 ± 0.51	95.33 ± 2.08	95.27 ± 2.49	95.01 ± 2.90	93.26 ± 1.34
BOBA	96.70 ± 0.21	95.76 ± 0.19	95.21 ± 0.61	95.00 ± 0.50	95.06 ± 0.95
FLDetector+Median	95.69 ± 0.37	94.74 ± 1.11	94.11 ± 0.67	93.27 ± 1.63	92.88 ± 0.23
HGDA ($W = 10$)	98.09 ± 0.41	$\underline{98.17} \pm 0.34$	$\underline{98.09} \pm 0.32$	$\underline{97.79} \pm 0.34$	$\underline{97.36} \pm 0.31$
HGDA ($W = 100$)	$\mathbf{98.72} \pm 0.01$	$\mathbf{98.64} \pm 0.03$	$\mathbf{98.48} \pm 0.20$	$\mathbf{98.36} \pm 0.10$	$\mathbf{97.86} \pm 0.07$

Table A.12: MNIST under Min-Max. Best test accuracy (%).

Method	20%	40%	50%	60%	80%
FedAvg (benign only)	97.57 ± 0.05	97.55 ± 0.09	97.44 ± 0.13	97.48 ± 0.11	96.93 ± 0.17
Multi-Krum	97.22 ± 0.07	79.51 ± 5.24	23.95 ± 2.49	24.83 ± 3.29	43.52 ± 24.10
Bulyan	95.36 ± 0.30	N/A	N/A	N/A	N/A
Coord. median	95.60 ± 0.24	87.10 ± 0.52	23.95 ± 2.49	24.83 ± 3.29	43.68 ± 24.39
Trimmed mean	95.43 ± 0.21	86.11 ± 1.13	N/A	N/A	N/A
RFA	96.81 ± 0.30	90.15 ± 1.16	23.95 ± 2.49	24.83 ± 3.29	43.49 ± 24.05
Fed-NGA	97.13 ± 0.36	89.14 ± 1.11	87.56 ± 1.01	87.12 ± 1.84	73.34 ± 5.01
FLTrust-Early	96.96 ± 0.45	95.86 ± 0.93	95.27 ± 1.25	94.29 ± 1.29	92.80 ± 0.75
FLTrust-Late	96.66 ± 0.54	96.11 ± 0.50	96.10 ± 0.54	95.12 ± 0.69	93.13 ± 1.40
BOBA	97.19 ± 0.14	97.13 ± 0.04	97.06 ± 0.25	96.94 ± 0.17	96.40 ± 0.57
FLDetector+Median	94.32 ± 1.08	86.35 ± 1.73	23.95 ± 2.49	26.38 ± 1.87	DNF
HGDA ($W = 10$)	$\underline{98.10} \pm 0.33$	$\underline{98.28} \pm 0.25$	$\underline{98.08} \pm 0.26$	$\underline{97.84} \pm 0.40$	$\underline{97.35} \pm 0.19$
HGDA ($W = 100$)	$\mathbf{98.69} \pm 0.02$	$\mathbf{98.64} \pm 0.03$	$\mathbf{98.48} \pm 0.18$	$\mathbf{98.33} \pm 0.01$	$\mathbf{97.84} \pm 0.06$

Table A.13: MNIST under IPM. Best test accuracy (%).

Method	20%	40%	50%	60%	80%
FedAvg (benign only)	97.57 ± 0.05	97.55 ± 0.09	97.44 ± 0.13	97.48 ± 0.11	96.93 ± 0.17
Multi-Krum	92.54 ± 0.78	12.19 ± 2.13	7.03 ± 2.35	8.04 ± 4.05	8.05 ± 2.13
Bulyan	87.69 ± 1.42	N/A	N/A	N/A	N/A
Coord. median	91.63 ± 0.22	62.41 ± 7.46	7.38 ± 2.18	8.04 ± 4.05	8.05 ± 2.13
Trimmed mean	92.61 ± 0.15	60.00 ± 9.54	N/A	N/A	N/A
RFA	91.54 ± 0.63	62.01 ± 4.13	7.03 ± 2.35	8.04 ± 4.05	8.05 ± 2.13
Fed-NGA	97.57 ± 0.16	80.67 ± 5.80	6.74 ± 3.57	5.65 ± 3.45	6.61 ± 4.60
FLTrust-Early	96.08 ± 1.82	96.39 ± 1.52	95.86 ± 2.27	96.11 ± 2.06	94.86 ± 2.93
FLTrust-Late	93.83 ± 2.45	95.04 ± 1.82	94.75 ± 3.01	95.29 ± 2.37	94.98 ± 1.94
BOBA	93.78 ± 1.05	84.01 ± 12.37	11.50 ± 2.29	11.95 ± 1.60	25.44 ± 24.57
FLDetector+Median	91.28 ± 0.41	62.41 ± 7.46	9.78 ± 2.91	10.61 ± 3.74	11.48 ± 1.67
HGDA ($W = 10$)	$\underline{97.90} \pm 0.44$	$\underline{98.12} \pm 0.17$	$\underline{97.94} \pm 0.16$	$\underline{97.53} \pm 0.09$	$\underline{97.46} \pm 0.28$
HGDA ($W = 100$)	$\mathbf{98.70} \pm 0.05$	$\mathbf{98.64} \pm 0.09$	$\mathbf{98.45} \pm 0.16$	$\mathbf{98.31} \pm 0.08$	$\mathbf{97.87} \pm 0.05$

C.2 CIFAR-10

All 100 clients participate in each of the 1500 training rounds.

Computational budget and repeated runs. A 1500-round CIFAR-10 run with full participation and five local epochs per round takes approximately 9–10 hours on a single NVIDIA GeForce RTX 3090. To cover the full range of attacks and Byzantine fractions within the computational budget, we use one seed (2026) for the complete CIFAR-10 grid and three seeds (2025, 2026 and 2027) to compare all methods under sign-flip with 20% Byzantine clients. The other four datasets have complete three-seed evaluations across their respective grids. Appendix C.6 examines run-to-run consistency using these results and the repeated CIFAR-10 experiments.

Bulyan and trimmed mean: N/A conditions. The 100-client cohort gives the same count requirements as MNIST: $f_t \leq 24$ for Bulyan and $f_t \leq 49$ for trimmed mean. Across sign-flip, label-flip, Gaussian, SameValue, Mimic, Min-Sum, Min-Max and IPM, Bulyan is N/A at $m = 40\%, 50\%, 60\%, 80\%$, while trimmed mean is N/A at $m = 50\%, 60\%, 80\%$. For the smaller ALIE grid, Bulyan is N/A at $m = 40\%, 50\%$, and trimmed mean is N/A at $m = 50\%$. Full participation fixes the Byzantine count in each of these conditions, so the required candidate set or untrimmed coordinate set cannot be formed.

Table A.14: CIFAR-10 without attack. Best test accuracy (%).

Method	0%
FedAvg (benign only)	85.53
Multi-Krum	85.41
Bulyan	<u>85.99</u>
Coord. median	84.62
Trimmed mean	85.92
RFA	85.63
Fed-NGA	**86.04**
FLTrust-Early	84.79
FLTrust-Late	81.72
BOBA	85.63
HGDA ($W = 10$)	84.76
HGDA ($W = 100$)	84.86

Table A.15: CIFAR-10 under sign-flip. Best test accuracy (%).

Method	20%	40%	50%	60%	80%
FedAvg (benign only)	84.33	80.49	78.17	77.14	58.63
Multi-Krum	73.93	59.01	45.41	10.00	10.00
Bulyan	70.69	N/A	N/A	N/A	N/A
Coord. median	76.42	44.98	22.41	10.00	10.00
Trimmed mean	76.86	43.57	N/A	N/A	N/A
RFA	80.03	48.90	11.09	10.09	10.00
Fed-NGA	80.76	52.51	13.60	11.13	10.00
FLTrust-Early	81.23	75.86	74.98	73.43	58.10
FLTrust-Late	67.33	61.30	57.46	58.21	57.28
BOBA	80.76	73.73	70.83	65.96	10.03
HGDA ($W = 10$)	**83.44**	<u>80.38</u>	<u>77.04</u>	**79.49**	**71.76**
HGDA ($W = 100$)	<u>82.41</u>	**81.11**	**80.01**	<u>79.01</u>	<u>71.17</u>

Table A.16: CIFAR-10 under label-flip. Best test accuracy (%).

Method	20%	40%	50%	60%	80%
FedAvg (benign only)	84.33	80.49	78.17	77.14	58.63
Multi-Krum	81.69	62.71	59.48	18.24	14.94
Bulyan	75.48	N/A	N/A	N/A	N/A
Coord. median	82.47	74.09	46.38	19.37	16.83
Trimmed mean	83.32	71.12	N/A	N/A	N/A
RFA	**84.40**	64.18	43.78	41.84	15.47
Fed-NGA	84.33	71.19	50.83	49.40	16.92
FLTrust-Early	82.67	76.94	75.41	75.92	68.38
FLTrust-Late	81.13	76.03	70.51	73.61	66.98
BOBA	83.51	75.26	68.73	39.11	17.64
HGDA ($W = 10$)	83.91	80.24	79.08	79.66	**68.39**
HGDA ($W = 100$)	84.36	**81.23**	**79.92**	**80.09**	66.04

Table A.17: CIFAR-10 under Gaussian. Best test accuracy (%).

Method	20%	40%	50%	60%	80%
FedAvg (benign only)	84.33	80.49	78.17	77.14	58.63
Multi-Krum	**85.09**	81.94	80.61	80.18	73.00
Bulyan	77.79	N/A	N/A	N/A	N/A
Coord. median	82.66	78.61	78.76	78.64	15.57
Trimmed mean	84.47	80.00	N/A	N/A	N/A
RFA	84.74	81.09	80.84	80.97	72.16
Fed-NGA	84.67	81.18	80.69	80.13	69.67
FLTrust-Early	82.71	78.30	77.17	78.26	69.88
FLTrust-Late	80.72	78.20	75.17	75.32	68.69
BOBA	84.74	**82.16**	**81.30**	**81.10**	**73.89**
HGDA ($W = 10$)	82.93	80.39	80.47	79.49	72.57
HGDA ($W = 100$)	83.86	81.99	79.74	80.01	71.32

Table A.18: CIFAR-10 under SameValue. Best test accuracy (%).

Method	20%	40%	50%	60%	80%
FedAvg (benign only)	84.33	80.49	78.17	77.14	58.63
Multi-Krum	**84.92**	**82.06**	10.90	11.04	10.94
Bulyan	78.77	N/A	N/A	N/A	N/A
Coord. median	52.69	19.04	23.96	10.98	10.98
Trimmed mean	46.43	17.46	N/A	N/A	N/A
RFA	74.81	28.80	11.02	10.89	10.74
Fed-NGA	76.29	48.26	44.36	20.04	21.17
FLTrust-Early	81.51	72.97	66.77	69.67	60.06
FLTrust-Late	81.32	77.29	77.39	76.37	67.58
BOBA	84.81	81.61	**80.68**	**80.72**	**72.18**
HGDA ($W = 10$)	82.90	80.52	78.46	79.53	70.92
HGDA ($W = 100$)	84.17	80.10	77.73	79.64	71.44

Table A.19: CIFAR-10 under ALIE. Best test accuracy (%).

Method	20%	40%	50%
FedAvg (benign only)	84.33	80.49	78.17
Multi-Krum	78.93	30.16	12.90
Bulyan	63.47	N/A	N/A
Coord. median	64.57	22.82	12.59
Trimmed mean	79.34	24.39	N/A
RFA	70.77	21.96	13.09
Fed-NGA	79.01	42.83	22.41
FLTrust-Early	70.41	71.53	70.38
FLTrust-Late	80.27	76.61	75.78
BOBA	78.96	68.97	79.02
HGDA ($W = 10$)	**83.70**	81.04	79.37
HGDA ($W = 100$)	83.57	**81.67**	**80.91**

Table A.20: CIFAR-10 under Mimic. Best test accuracy (%).

Method	20%	40%	50%	60%	80%
FedAvg (benign only)	84.33	80.49	78.17	77.14	58.63
Multi-Krum	78.21	57.94	28.69	29.51	39.27
Bulyan	43.37	N/A	N/A	N/A	N/A
Coord. median	72.70	46.07	29.63	30.48	36.94
Trimmed mean	82.12	46.93	N/A	N/A	N/A
RFA	75.72	35.47	29.82	29.32	39.56
Fed-NGA	83.86	77.32	74.20	71.77	59.04
FLTrust-Early	82.89	77.74	75.40	74.44	67.27
FLTrust-Late	81.94	77.10	73.19	70.80	67.04
BOBA	79.92	68.68	31.47	29.87	38.02
HGDA ($W = 10$)	**84.03**	80.74	79.59	**79.88**	**71.83**
HGDA ($W = 100$)	83.99	**81.07**	**80.04**	79.72	71.32

Table A.21: CIFAR-10 under Min-Sum. Best test accuracy (%).

Method	20%	40%	50%	60%	80%
FedAvg (benign only)	84.33	80.49	78.17	77.14	58.63
Multi-Krum	69.49	42.98	28.39	26.80	25.31
Bulyan	31.98	N/A	N/A	N/A	N/A
Coord. median	55.36	29.54	28.24	26.16	24.68
Trimmed mean	67.80	29.83	N/A	N/A	N/A
RFA	71.92	29.10	29.33	27.04	25.21
Fed-NGA	75.39	47.23	38.39	38.47	32.12
FLTrust-Early	77.60	69.40	67.64	65.24	59.47
FLTrust-Late	79.06	78.89	75.51	76.64	67.33
BOBA	79.79	64.58	42.66	44.51	35.53
HGDA ($W = 10$)	83.64	80.53	78.94	**79.78**	70.46
HGDA ($W = 100$)	**84.59**	**81.12**	**79.68**	79.19	**71.87**

Table A.22: CIFAR-10 under Min-Max. Best test accuracy (%).

Method	20%	40%	50%	60%	80%
FedAvg (benign only)	84.33	80.49	78.17	77.14	58.63
Multi-Krum	69.91	18.23	13.60	12.20	13.66
Bulyan	55.89	N/A	N/A	N/A	N/A
Coord. median	54.13	20.63	13.62	12.32	14.04
Trimmed mean	46.22	20.19	N/A	N/A	N/A
RFA	71.69	18.16	13.78	12.49	14.31
Fed-NGA	73.79	26.76	24.17	18.89	14.84
FLTrust-Early	79.23	72.69	70.89	70.29	63.34
FLTrust-Late	80.12	75.92	76.18	76.23	67.50
BOBA	<u>83.93</u>	71.64	77.63	74.46	70.91
HGDA ($W = 10$)	83.68	<u>80.44</u>	<u>79.80</u>	<u>79.57</u>	<u>71.01</u>
HGDA ($W = 100$)	**84.14**	**81.80**	**80.00**	**79.60**	**73.09**

Table A.23: CIFAR-10 under IPM. Best test accuracy (%).

Method	20%	40%	50%	60%	80%
FedAvg (benign only)	84.33	80.49	78.17	77.14	58.63
Multi-Krum	71.51	13.52	10.00	10.00	10.00
Bulyan	29.11	N/A	N/A	N/A	N/A
Coord. median	53.76	12.51	17.89	10.00	10.00
Trimmed mean	79.03	13.30	N/A	N/A	N/A
RFA	70.17	19.14	10.00	10.00	10.00
Fed-NGA	70.74	17.97	10.00	10.00	10.00
FLTrust-Early	82.87	78.80	77.31	77.47	69.19
FLTrust-Late	69.07	67.39	63.70	63.28	64.94
BOBA	71.92	76.48	10.00	10.00	10.10
HGDA ($W = 10$)	<u>83.29</u>	<u>80.71</u>	<u>78.70</u>	**80.28**	<u>72.60</u>
HGDA ($W = 100$)	**84.08**	**81.43**	**80.00**	<u>79.70</u>	**73.66**

C.3 FEMNIST

Each round samples 100 of the 3400 clients, with a training budget of 500 rounds. The following conditions apply to all four attacks: sign-flip, label-flip, Mimic and IPM.

Bulyan: N/A at $m = 30\%, 40\%, 50\%, 60\%$. Bulyan requires at most 24 Byzantine clients in a 100-client cohort. The nominal counts at these four fractions are 30, 40, 50 and 60, respectively, so they were excluded from the evaluation grid. The reported conditions at $m = 1\%, 5\%, 10\%, 20\%$ use cohort resampling to enforce $f_t \leq 24$ in every training round.

BOBA: DNF at $m = 30\%$. FEMNIST has 62 classes, so BOBA needs at least 62 selected updates to fit its 61-dimensional affine subspace. With $n_t = 100$, this requires $f_t \leq 38$. The nominal count at $m = 30\%$ satisfies this requirement, and these runs were launched with ordinary client sampling. As participation varies, however, a sampled cohort can contain more than 38 Byzantine clients. The runs terminate when this occurs because BOBA's subspace-fitting step no longer has enough updates. For example, $f_t = 39$ leaves only $100 - 39 = 61$ updates for the fit.

BOBA: N/A at $m = 40\%, 50\%, 60\%$. The nominal retained counts are 60, 50 and 40, respectively, all below the 62 updates required for the subspace fit. These conditions were excluded before training. BOBA's trusted data contain ten examples from each class, so class coverage is supplied separately from the participant-count requirement.

Trimmed mean at high population fractions. The trimmed-mean results at $m = 50\%$ and 60% use cohort resampling until $f_t \leq 49$. Thus, although these columns specify 50% and 60% Byzantine clients in the population, each accepted 100-client cohort contains fewer than 50 Byzantine clients and leaves a nonempty set of coordinate values after trimming.

Table A.24: FEMNIST without attack. Best test accuracy (%).

Method	0%
FedAvg (benign only)	84.42 ± 0.10
FedAvg	84.42 ± 0.10
Multi-Krum	84.51 ± 0.05
Bulyan	84.52 ± 0.02
Coord. median	79.90 ± 0.17
Trimmed mean	84.53 ± 0.09
RFA	84.07 ± 0.08
Fed-NGA	84.73 ± 0.05
FLTrust-Early	$\underline{84.78} \pm 0.21$
FLTrust-Late	84.75 ± 0.19
BOBA	84.58 ± 0.02
HGDA ($W = 100$)	$\mathbf{85.20} \pm 0.14$

Table A.25: FEMNIST under sign-flip. Best test accuracy (%).

(a) 1–20% Byzantine clients

Method	1%	5%	10%	20%
FedAvg (benign only)	84.43 ± 0.11	84.35 ± 0.08	84.52 ± 0.03	84.43 ± 0.06
FedAvg	84.50 ± 0.11	84.28 ± 0.08	84.11 ± 0.03	83.41 ± 0.14
Multi-Krum	84.49 ± 0.05	84.36 ± 0.14	83.93 ± 0.05	82.87 ± 0.16
Bulyan	84.04 ± 0.06	80.62 ± 0.20	76.79 ± 0.39	73.16 ± 0.22
Coord. median	79.81 ± 0.16	79.73 ± 0.19	79.38 ± 0.35	78.78 ± 0.25
Trimmed mean	84.43 ± 0.05	83.73 ± 0.10	82.82 ± 0.12	81.25 ± 0.15
RFA	84.07 ± 0.09	83.95 ± 0.21	83.80 ± 0.21	83.03 ± 0.20
Fed-NGA	84.61 ± 0.10	84.40 ± 0.15	$\underline{84.22} \pm 0.18$	$\underline{83.42} \pm 0.11$
FLTrust-Early	$\underline{84.71} \pm 0.30$	$\underline{84.50} \pm 0.29$	83.95 ± 0.26	81.24 ± 0.70
FLTrust-Late	84.52 ± 0.34	84.18 ± 0.15	83.11 ± 0.10	80.27 ± 0.61
BOBA	84.42 ± 0.10	84.32 ± 0.07	83.97 ± 0.04	83.06 ± 0.08
HGDA ($W = 100$)	$\mathbf{85.25} \pm 0.02$	$\mathbf{85.08} \pm 0.05$	$\mathbf{84.77} \pm 0.32$	$\mathbf{84.56} \pm 0.16$

(b) 30–60% Byzantine clients

Method	30%	40%	50%	60%
FedAvg (benign only)	84.40 ± 0.16	84.25 ± 0.07	84.08 ± 0.08	84.17 ± 0.14
FedAvg	$\underline{82.01} \pm 0.29$	77.32 ± 0.50	8.86 ± 4.14	3.76 ± 2.22
Multi-Krum	81.12 ± 0.52	76.80 ± 1.09	5.67 ± 0.14	3.59 ± 2.07
Bulyan	N/A	N/A	N/A	N/A
Coord. median	77.63 ± 0.11	73.56 ± 0.97	7.13 ± 2.03	0.66 ± 0.46
Trimmed mean	78.98 ± 0.15	74.64 ± 0.75	66.14 ± 0.97	57.64 ± 3.49
RFA	81.63 ± 0.25	$\underline{77.93} \pm 0.43$	7.29 ± 0.32	3.72 ± 2.18
Fed-NGA	81.72 ± 0.17	76.77 ± 0.56	7.43 ± 2.24	2.50 ± 1.14
FLTrust-Early	78.56 ± 1.27	77.05 ± 0.65	76.23 ± 0.72	52.74 ± 39.31
FLTrust-Late	78.14 ± 0.83	77.26 ± 0.65	$\underline{76.41} \pm 0.39$	$\underline{75.78} \pm 0.44$
BOBA	DNF	N/A	N/A	N/A
HGDA ($W = 100$)	$\mathbf{84.27} \pm 0.26$	$\mathbf{83.61} \pm 0.37$	$\mathbf{83.30} \pm 0.55$	$\mathbf{82.45} \pm 0.60$

Table A.26: FEMNIST under label-flip. Best test accuracy (%).

(a) 1–20% Byzantine clients

Method	1%	5%	10%	20%
FedAvg (benign only)	84.43 ± 0.11	84.44 ± 0.06	84.42 ± 0.12	84.47 ± 0.08
FedAvg	84.27 ± 0.12	83.96 ± 0.06	83.72 ± 0.16	83.39 ± 0.20
Multi-Krum	84.49 ± 0.06	84.41 ± 0.09	84.29 ± 0.11	84.24 ± 0.09
Bulyan	83.98 ± 0.05	80.83 ± 0.17	77.29 ± 0.23	75.94 ± 0.32
Coord. median	79.64 ± 0.35	79.26 ± 0.22	78.28 ± 0.32	77.08 ± 0.33
Trimmed mean	84.38 ± 0.07	83.43 ± 0.18	81.84 ± 0.08	79.01 ± 0.12
RFA	84.02 ± 0.03	83.84 ± 0.09	83.79 ± 0.12	83.62 ± 0.14
Fed-NGA	84.58 ± 0.08	84.36 ± 0.01	84.01 ± 0.22	83.93 ± 0.16
FLTrust-Early	$\underline{84.68} \pm 0.31$	$\underline{84.51} \pm 0.31$	$\underline{84.49} \pm 0.25$	84.32 ± 0.31
FLTrust-Late	84.58 ± 0.36	84.45 ± 0.27	84.46 ± 0.17	$\underline{84.46} \pm 0.28$
BOBA	84.44 ± 0.08	84.16 ± 0.05	83.76 ± 0.14	83.24 ± 0.06
HGDA ($W = 100$)	$\mathbf{85.19} \pm 0.10$	$\mathbf{84.89} \pm 0.21$	$\mathbf{84.94} \pm 0.34$	$\mathbf{84.82} \pm 0.28$

(b) 30–60% Byzantine clients

Method	30%	40%	50%	60%
FedAvg (benign only)	84.32 ± 0.08	84.32 ± 0.18	84.04 ± 0.10	84.17 ± 0.19
FedAvg	82.97 ± 0.16	82.30 ± 0.27	81.06 ± 0.32	76.73 ± 1.47
Multi-Krum	84.03 ± 0.21	83.86 ± 0.14	77.37 ± 10.02	40.09 ± 1.80
Bulyan	N/A	N/A	N/A	N/A
Coord. median	75.49 ± 0.30	73.44 ± 0.65	68.76 ± 1.16	16.46 ± 2.65
Trimmed mean	76.28 ± 0.65	73.89 ± 0.40	70.40 ± 0.80	67.45 ± 0.91
RFA	83.20 ± 0.22	82.67 ± 0.37	81.21 ± 0.30	72.71 ± 2.00
Fed-NGA	83.15 ± 0.11	82.57 ± 0.45	81.66 ± 0.38	80.56 ± 0.40
FLTrust-Early	$\underline{84.27} \pm 0.33$	$\underline{84.21} \pm 0.20$	$\underline{84.19} \pm 0.18$	$\underline{83.95} \pm 0.08$
FLTrust-Late	84.23 ± 0.29	84.18 ± 0.15	84.15 ± 0.35	$\mathbf{83.97} \pm 0.18$
BOBA	DNF	N/A	N/A	N/A
HGDA ($W = 100$)	$\mathbf{84.80} \pm 0.69$	$\mathbf{84.76} \pm 0.10$	$\mathbf{84.23} \pm 0.12$	83.75 ± 0.28

Table A.27: FEMNIST under Mimic. Best test accuracy (%).

(a) 1–20% Byzantine clients

Method	1%	5%	10%	20%
FedAvg (benign only)	84.43 ± 0.11	84.36 ± 0.05	84.42 ± 0.12	84.44 ± 0.08
FedAvg	84.42 ± 0.10	84.27 ± 0.12	84.07 ± 0.24	83.74 ± 0.14
Multi-Krum	84.53 ± 0.11	84.54 ± 0.07	84.27 ± 0.15	83.40 ± 0.15
Bulyan	84.02 ± 0.04	81.06 ± 0.04	78.62 ± 0.47	79.30 ± 0.22
Coord. median	79.94 ± 0.20	80.24 ± 0.16	80.68 ± 0.42	80.75 ± 0.13
Trimmed mean	84.41 ± 0.05	84.03 ± 0.10	83.52 ± 0.16	82.18 ± 0.24
RFA	84.12 ± 0.12	83.88 ± 0.03	83.65 ± 0.08	82.79 ± 0.14
Fed-NGA	84.71 ± 0.27	84.61 ± 0.17	84.54 ± 0.15	84.07 ± 0.22
FLTrust-Early	$\underline{84.74} \pm 0.24$	$\underline{84.71} \pm 0.32$	$\underline{84.72} \pm 0.27$	84.43 ± 0.22
FLTrust-Late	84.63 ± 0.33	84.58 ± 0.28	84.62 ± 0.16	$\underline{84.47} \pm 0.21$
BOBA	84.47 ± 0.09	84.47 ± 0.12	84.33 ± 0.23	83.87 ± 0.16
HGDA ($W = 100$)	$\mathbf{85.19} \pm 0.05$	$\mathbf{85.36} \pm 0.36$	$\mathbf{85.01} \pm 0.10$	$\mathbf{84.78} \pm 0.31$

(b) 30–60% Byzantine clients

Method	30%	40%	50%	60%
FedAvg (benign only)	84.42 ± 0.14	84.25 ± 0.05	84.11 ± 0.14	84.21 ± 0.16
FedAvg	83.20 ± 0.26	82.09 ± 1.17	81.86 ± 0.58	80.66 ± 1.53
Multi-Krum	81.68 ± 0.34	78.33 ± 0.88	50.67 ± 39.17	36.98 ± 24.33
Bulyan	N/A	N/A	N/A	N/A
Coord. median	81.00 ± 0.19	79.29 ± 0.53	74.84 ± 0.35	41.40 ± 34.08
Trimmed mean	81.17 ± 0.44	79.18 ± 0.60	61.05 ± 27.71	59.92 ± 24.47
RFA	80.62 ± 0.16	77.04 ± 0.42	62.13 ± 18.64	52.49 ± 33.38
Fed-NGA	83.25 ± 0.26	82.88 ± 0.23	81.53 ± 0.44	80.96 ± 0.32
FLTrust-Early	84.20 ± 0.25	$\underline{84.02} \pm 0.06$	83.85 ± 0.33	83.49 ± 0.25
FLTrust-Late	$\underline{84.28} \pm 0.27$	83.97 ± 0.14	$\underline{83.92} \pm 0.33$	$\underline{83.61} \pm 0.34$
BOBA	DNF	N/A	N/A	N/A
HGDA ($W = 100$)	$\mathbf{84.68} \pm 0.18$	$\mathbf{84.53} \pm 0.33$	$\mathbf{84.29} \pm 0.26$	$\mathbf{83.93} \pm 0.24$

Table A.28: FEMNIST under IPM. Best test accuracy (%).

(a) 1–20% Byzantine clients

Method	1%	5%	10%	20%
FedAvg (benign only)	84.43 ± 0.11	84.42 ± 0.09	84.42 ± 0.12	84.33 ± 0.10
FedAvg	84.49 ± 0.12	$\underline{84.40} \pm 0.02$	$\underline{84.08} \pm 0.21$	$\underline{83.52} \pm 0.16$
Multi-Krum	84.50 ± 0.08	84.39 ± 0.06	83.93 ± 0.09	81.16 ± 0.48
Bulyan	83.99 ± 0.08	78.93 ± 0.12	67.34 ± 1.24	48.99 ± 3.01
Coord. median	79.44 ± 0.11	77.40 ± 0.24	72.42 ± 0.36	64.24 ± 0.46
Trimmed mean	84.37 ± 0.06	83.62 ± 0.07	82.27 ± 0.21	71.64 ± 0.52
RFA	83.92 ± 0.12	83.02 ± 0.27	81.16 ± 0.10	76.05 ± 0.34
Fed-NGA	$\underline{84.58} \pm 0.12$	83.85 ± 0.15	82.53 ± 0.02	77.76 ± 0.35
FLTrust-Early	84.32 ± 0.17	81.87 ± 0.52	80.98 ± 0.62	80.50 ± 0.92
FLTrust-Late	83.49 ± 0.71	81.09 ± 1.01	80.91 ± 0.69	79.71 ± 1.03
BOBA	$\underline{84.58} \pm 0.06$	84.33 ± 0.14	84.06 ± 0.15	83.23 ± 0.21
HGDA ($W = 100$)	$\mathbf{84.95} \pm 0.26$	$\mathbf{85.02} \pm 0.41$	$\mathbf{85.18} \pm 0.26$	$\mathbf{85.07} \pm 0.33$

(b) 30–60% Byzantine clients

Method	30%	40%	50%	60%
FedAvg (benign only)	84.37 ± 0.09	84.17 ± 0.10	84.12 ± 0.14	84.20 ± 0.11
FedAvg	$\underline{82.34} \pm 0.21$	79.11 ± 0.11	12.33 ± 11.18	4.13 ± 2.54
Multi-Krum	20.43 ± 18.69	5.57 ± 0.05	3.12 ± 1.84	3.37 ± 2.21
Bulyan	N/A	N/A	N/A	N/A
Coord. median	44.59 ± 2.25	16.70 ± 12.98	5.35 ± 0.21	3.37 ± 2.21
Trimmed mean	47.17 ± 1.06	16.35 ± 9.02	6.02 ± 0.62	5.59 ± 0.03
RFA	65.30 ± 0.12	7.22 ± 3.13	4.91 ± 1.19	3.37 ± 2.21
Fed-NGA	67.12 ± 0.60	50.70 ± 0.98	3.12 ± 1.84	4.13 ± 2.54
FLTrust-Early	81.27 ± 0.44	$\underline{80.43} \pm 1.06$	$\underline{81.29} \pm 0.42$	$\underline{82.21} \pm 0.23$
FLTrust-Late	80.66 ± 0.46	80.40 ± 1.07	81.15 ± 0.18	81.59 ± 0.72
BOBA	DNF	N/A	N/A	N/A
HGDA ($W = 100$)	$\mathbf{84.66} \pm 0.05$	$\mathbf{84.44} \pm 0.29$	$\mathbf{84.43} \pm 0.27$	$\mathbf{84.17} \pm 0.10$

C.4 PURCHASE-100

Each round samples 50 of the 1000 clients, with a training budget of 500 rounds. The following conditions apply to sign-flip, label-flip, Mimic and IPM throughout the $m = 10\%$–60% grid.

Bulyan: NR at $m = 20\%$. For 50 participants, Bulyan requires $f_t \leq 11$. At $m = 20\%$, the nominal count is ten, but approximately 28.5% of uniformly sampled cohorts contain at least 12 Byzantine clients. We did not run these four attack conditions because of the frequent participant-count violations, giving the NR entries. The reported $m = 10\%$ results use cohort resampling to keep $f_t \leq 11$.

Bulyan: N/A at $m = 30\%, 40\%, 50\%, 60\%$. The nominal Byzantine counts are 15, 20, 25 and 30, respectively. All exceed the limit of 11, so these four conditions were excluded from the evaluation grid for each attack.

Trimmed mean: N/A at $m = 50\%, 60\%$. With 50 participants, trimming both tails requires $f_t \leq 24$. The nominal counts of 25 and 30 leave no coordinate values to average after removing f_t values from each tail, so these conditions were excluded from the grid. The reported $m = 10\%, 20\%, 30\%, 40\%$ conditions use resampling to meet the per-round requirement.

BOBA and the number of classes. Purchase-100 has 100 classes but only 50 participants per round. Even a clean cohort provides fewer than the 100 updates needed for BOBA's 99-dimensional affine fit. BOBA is therefore absent from both the clean and attacked Purchase-100 tables.

Table A.29: Purchase-100 without attack. Best test accuracy (%).

Method	0%
FedAvg (benign only)	91.33 ± 0.05
FedAvg	$\underline{91.33} \pm 0.05$
Multi-Krum	91.29 ± 0.07
Bulyan	91.29 ± 0.12
Coord. median	84.61 ± 0.26
Trimmed mean	91.29 ± 0.06
RFA	91.24 ± 0.08
Fed-NGA	90.10 ± 0.08
FLTrust-Early	88.84 ± 0.45
FLTrust-Late	86.74 ± 0.33
HGDA ($W = 100$)	$\mathbf{91.79} \pm 0.45$

Table A.30: Purchase-100 under sign-flip. Best test accuracy (%).

(a) 10–30% Byzantine clients

Method	10%	20%	30%
FedAvg (benign only)	91.13 ± 0.07	90.93 ± 0.11	90.83 ± 0.04
FedAvg	89.43 ± 0.21	86.79 ± 0.24	81.02 ± 0.61
Multi-Krum	$\underline{89.94} \pm 0.11$	$\underline{88.19} \pm 0.22$	$\underline{85.57} \pm 0.12$
Bulyan	83.47 ± 0.30	NR	N/A
Coord. median	83.40 ± 0.17	81.89 ± 0.15	77.92 ± 1.05
Trimmed mean	88.34 ± 0.29	84.85 ± 0.25	80.20 ± 0.65
RFA	89.87 ± 0.10	87.92 ± 0.13	84.21 ± 0.33
Fed-NGA	88.78 ± 0.14	86.90 ± 0.05	83.29 ± 0.24
FLTrust-Early	83.73 ± 2.25	79.40 ± 5.06	77.53 ± 4.01
FLTrust-Late	76.04 ± 2.38	74.40 ± 1.65	73.69 ± 1.16
HGDA ($W = 100$)	$\mathbf{90.95} \pm 0.85$	$\mathbf{89.92} \pm 0.92$	$\mathbf{89.31} \pm 0.70$

(b) 40–60% Byzantine clients

Method	40%	50%	60%
FedAvg (benign only)	90.20 ± 0.07	89.89 ± 0.28	89.07 ± 0.24
FedAvg	66.15 ± 3.72	6.90 ± 1.89	1.58 ± 0.36
Multi-Krum	$\underline{81.50} \pm 1.06$	68.33 ± 3.89	2.47 ± 1.34
Bulyan	N/A	N/A	N/A
Coord. median	65.34 ± 1.97	12.99 ± 0.94	0.83 ± 0.48
Trimmed mean	70.01 ± 0.94	N/A	N/A
RFA	72.64 ± 1.68	15.92 ± 6.36	1.19 ± 0.41
Fed-NGA	72.06 ± 1.15	9.61 ± 5.89	1.07 ± 0.18
FLTrust-Early	73.35 ± 5.04	$\underline{72.43} \pm 3.38$	69.49 ± 2.91
FLTrust-Late	73.27 ± 0.53	72.15 ± 1.40	$\underline{71.93} \pm 1.97$
HGDA ($W = 100$)	$\mathbf{88.03} \pm 1.04$	$\mathbf{86.90} \pm 0.87$	$\mathbf{85.40} \pm 1.77$

Table A.31: Purchase-100 under label-flip. Best test accuracy (%).

(a) 10–30% Byzantine clients

Method	10%	20%	30%
FedAvg (benign only)	91.14 ± 0.08	90.91 ± 0.12	90.81 ± 0.03
FedAvg	84.81 ± 0.22	84.17 ± 0.28	83.03 ± 0.40
Multi-Krum	$\mathbf{90.96} \pm 0.04$	$\mathbf{90.70} \pm 0.08$	$\mathbf{90.62} \pm 0.16$
Bulyan	83.99 ± 0.27	NR	N/A
Coord. median	80.79 ± 0.08	73.20 ± 0.11	62.34 ± 0.56
Trimmed mean	84.59 ± 0.07	79.71 ± 0.29	71.47 ± 0.82
RFA	88.32 ± 0.09	85.14 ± 0.24	84.35 ± 0.20
Fed-NGA	87.40 ± 0.07	83.39 ± 0.19	81.87 ± 0.07
FLTrust-Early	88.16 ± 0.62	87.31 ± 0.69	86.56 ± 0.85
FLTrust-Late	86.24 ± 0.35	86.10 ± 0.26	85.80 ± 0.39
HGDA ($W = 100$)	$\underline{89.28} \pm 1.10$	$\underline{87.76} \pm 1.44$	$\underline{86.58} \pm 1.35$

(b) 40–60% Byzantine clients

Method	40%	50%	60%
FedAvg (benign only)	90.18 ± 0.08	89.89 ± 0.27	89.06 ± 0.23
FedAvg	81.14 ± 0.32	78.40 ± 1.78	65.99 ± 2.61
Multi-Krum	$\mathbf{89.84} \pm 0.07$	69.49 ± 34.56	5.21 ± 1.68
Bulyan	N/A	N/A	N/A
Coord. median	52.19 ± 0.98	39.04 ± 0.56	11.52 ± 4.68
Trimmed mean	59.06 ± 0.92	N/A	N/A
RFA	82.86 ± 0.07	80.06 ± 0.92	62.97 ± 4.34
Fed-NGA	79.86 ± 0.27	74.53 ± 1.33	22.41 ± 5.64
FLTrust-Early	85.69 ± 0.86	84.60 ± 1.35	$\underline{83.32} \pm 1.63$
FLTrust-Late	84.83 ± 0.46	$\mathbf{84.85} \pm 0.58$	$\mathbf{83.71} \pm 0.81$
HGDA ($W = 100$)	$\underline{85.89} \pm 0.80$	$\underline{84.84} \pm 1.33$	83.27 ± 0.98

Table A.32: Purchase-100 under Mimic. Best test accuracy (%).

(a) 10–30% Byzantine clients

Method	10%	20%	30%
FedAvg (benign only)	91.14 ± 0.08	90.92 ± 0.13	90.84 ± 0.04
FedAvg	$\underline{90.91} \pm 0.06$	$\underline{90.44} \pm 0.17$	$\underline{89.79} \pm 0.09$
Multi-Krum	90.70 ± 0.09	89.56 ± 0.24	87.61 ± 0.18
Bulyan	83.89 ± 0.16	NR	N/A
Coord. median	84.67 ± 0.26	83.90 ± 0.21	81.91 ± 0.74
Trimmed mean	88.97 ± 0.21	86.23 ± 0.27	83.27 ± 0.73
RFA	90.41 ± 0.07	87.77 ± 0.12	82.67 ± 0.71
Fed-NGA	89.44 ± 0.15	88.15 ± 0.14	86.64 ± 0.30
FLTrust-Early	88.27 ± 0.63	87.93 ± 0.59	87.42 ± 0.91
FLTrust-Late	86.48 ± 0.57	86.52 ± 0.53	86.25 ± 0.67
HGDA ($W = 100$)	$\mathbf{91.40} \pm 0.99$	$\mathbf{90.97} \pm 0.25$	$\mathbf{90.46} \pm 0.51$

(b) 40–60% Byzantine clients

Method	40%	50%	60%
FedAvg (benign only)	90.18 ± 0.09	89.87 ± 0.29	89.04 ± 0.27
FedAvg	$\underline{88.65} \pm 0.26$	$\underline{87.36} \pm 0.14$	$\underline{86.01} \pm 0.32$
Multi-Krum	83.56 ± 0.37	78.10 ± 0.31	76.10 ± 0.80
Bulyan	N/A	N/A	N/A
Coord. median	79.62 ± 0.32	77.08 ± 0.60	76.12 ± 0.57
Trimmed mean	80.38 ± 0.53	N/A	N/A
RFA	77.73 ± 0.61	76.16 ± 0.52	76.23 ± 0.76
Fed-NGA	84.84 ± 0.34	83.14 ± 0.18	80.77 ± 0.90
FLTrust-Early	86.26 ± 0.82	84.89 ± 0.66	83.15 ± 0.40
FLTrust-Late	85.37 ± 0.75	84.54 ± 0.70	82.87 ± 0.46
HGDA ($W = 100$)	$\mathbf{89.54} \pm 0.61$	$\mathbf{88.79} \pm 0.59$	$\mathbf{87.58} \pm 0.56$

Table A.33: Purchase-100 under IPM. Best test accuracy (%).

(a) 10–30% Byzantine clients

Method	10%	20%	30%
FedAvg (benign only)	91.14 ± 0.08	90.92 ± 0.11	90.83 ± 0.04
FedAvg	$\underline{90.50} \pm 0.07$	$\underline{89.53} \pm 0.00$	$\underline{88.02} \pm 0.06$
Multi-Krum	90.33 ± 0.16	87.73 ± 0.22	72.19 ± 1.87
Bulyan	77.24 ± 0.89	NR	N/A
Coord. median	79.70 ± 0.24	55.97 ± 1.08	13.60 ± 0.64
Trimmed mean	88.10 ± 0.08	80.87 ± 0.13	12.35 ± 2.21
RFA	87.63 ± 0.12	77.74 ± 0.75	2.80 ± 0.72
Fed-NGA	87.03 ± 0.21	75.01 ± 0.95	10.87 ± 0.37
FLTrust-Early	76.96 ± 4.67	77.24 ± 4.06	75.73 ± 3.13
FLTrust-Late	74.77 ± 1.83	75.20 ± 1.00	74.45 ± 0.94
HGDA ($W = 100$)	$\mathbf{90.97} \pm 1.14$	$\mathbf{90.37} \pm 0.67$	$\mathbf{89.96} \pm 0.21$

(b) 40–60% Byzantine clients

Method	40%	50%	60%
FedAvg (benign only)	90.19 ± 0.08	89.88 ± 0.31	89.08 ± 0.25
FedAvg	$\underline{84.74} \pm 0.28$	6.75 ± 2.46	0.44 ± 0.17
Multi-Krum	2.47 ± 0.21	0.81 ± 0.18	0.23 ± 0.02
Bulyan	N/A	N/A	N/A
Coord. median	16.90 ± 1.00	1.27 ± 0.44	0.23 ± 0.02
Trimmed mean	23.95 ± 1.31	N/A	N/A
RFA	4.77 ± 3.04	0.70 ± 0.05	0.23 ± 0.02
Fed-NGA	2.30 ± 0.31	0.51 ± 0.12	0.36 ± 0.07
FLTrust-Early	77.43 ± 4.02	$\underline{79.09} \pm 3.54$	$\underline{78.81} \pm 3.60$
FLTrust-Late	73.70 ± 1.21	74.01 ± 2.15	73.49 ± 2.75
HGDA ($W = 100$)	$\mathbf{88.85} \pm 0.93$	$\mathbf{87.96} \pm 0.79$	$\mathbf{86.35} \pm 0.54$

C.5 TINY IMAGENET

Each round samples 10 of the 100 clients, with a training budget of 500 rounds. The following conditions apply to all four attacks: sign-flip, label-flip, Mimic and IPM.

Bulyan: N/A at $m = 20\%, 30\%, 40\%$. With ten participants, Bulyan's requirement $10 \geq 4f_t + 3$ allows at most one Byzantine client. The nominal counts at these fractions are two, three and four, respectively, so the corresponding conditions were excluded from the grid. At the reported $m = 10\%$ condition, the sampler redraws any cohort with two or more Byzantine clients, ensuring $f_t \leq 1$.

Trimmed mean and the sampled Byzantine count. Trimmed mean requires $f_t \leq 4$ for a ten-client cohort. It is evaluated at all four population fractions, $m = 10\%, 20\%, 30\%, 40\%$, with resampling whenever a cohort contains five or more Byzantine clients.

BOBA and the number of classes. Tiny ImageNet has 200 classes. A ten-client cohort cannot provide the 200 updates required for BOBA's 199-dimensional affine fit, even without Byzantine clients. BOBA is therefore absent from both the clean and attacked Tiny ImageNet tables.

Table A.34: Tiny ImageNet without attack. Best test accuracy (%).

Method	0%
FedAvg (benign only)	36.86 ± 0.47
FedAvg	36.86 ± 0.47
Multi-Krum	37.32 ± 0.95
Bulyan	36.59 ± 0.48
Coord. median	37.83 ± 0.15
Trimmed mean	36.75 ± 0.31
RFA	37.13 ± 0.08
Fed-NGA	$\underline{38.20} \pm 0.21$
FLTrust-Early	35.27 ± 0.59
FLTrust-Late	35.70 ± 0.38
HGDA ($W = 100$)	$\mathbf{38.26} \pm 0.46$

Table A.35: Tiny ImageNet under sign-flip. Best test accuracy (%).

Method	10%	20%	30%	40%
FedAvg (benign only)	35.66 ± 0.06	34.65 ± 0.46	33.36 ± 0.37	31.75 ± 0.91
FedAvg	31.48 ± 0.62	26.48 ± 1.26	20.30 ± 0.25	14.69 ± 1.69
Multi-Krum	33.93 ± 0.72	30.71 ± 1.64	26.59 ± 0.94	18.69 ± 1.34
Bulyan	34.57 ± 0.31	N/A	N/A	N/A
Coord. median	32.86 ± 0.44	28.14 ± 0.40	21.82 ± 0.44	13.29 ± 1.17
Trimmed mean	32.90 ± 0.27	28.52 ± 0.87	23.97 ± 0.39	20.50 ± 0.78
RFA	32.58 ± 0.50	27.31 ± 0.98	21.67 ± 0.57	14.78 ± 0.20
Fed-NGA	33.78 ± 0.17	28.89 ± 0.89	22.33 ± 1.12	13.22 ± 0.55
FLTrust-Early	$\underline{34.96} \pm 0.46$	$\underline{33.49} \pm 0.24$	$\underline{32.52} \pm 0.51$	$\underline{31.48} \pm 1.11$
FLTrust-Late	34.04 ± 0.46	33.19 ± 1.02	32.17 ± 0.37	30.01 ± 0.26
HGDA ($W = 100$)	$\mathbf{35.90} \pm 0.62$	$\mathbf{34.84} \pm 0.12$	$\mathbf{34.27} \pm 0.58$	$\mathbf{33.01} \pm 0.59$

Table A.36: Tiny ImageNet under label-flip. Best test accuracy (%).

Method	10%	20%	30%	40%
FedAvg (benign only)	35.67 ± 0.07	34.75 ± 0.16	33.53 ± 0.43	31.52 ± 1.12
FedAvg	34.70 ± 0.49	33.40 ± 0.09	32.57 ± 0.20	31.44 ± 0.60
Multi-Krum	35.07 ± 0.13	33.05 ± 0.43	31.00 ± 0.95	28.23 ± 0.99
Bulyan	$\underline{35.40} \pm 0.72$	N/A	N/A	N/A
Coord. median	34.85 ± 0.49	33.64 ± 0.45	32.44 ± 0.33	30.71 ± 0.54
Trimmed mean	34.78 ± 0.32	33.56 ± 0.19	$\underline{33.22} \pm 0.05$	31.54 ± 1.20
RFA	34.73 ± 0.48	33.88 ± 0.43	33.07 ± 0.39	$\underline{31.72} \pm 0.76$
Fed-NGA	$\mathbf{35.81} \pm 0.15$	$\underline{34.38} \pm 0.54$	32.90 ± 0.08	31.21 ± 0.30
FLTrust-Early	33.59 ± 0.34	32.22 ± 0.91	31.44 ± 1.00	30.43 ± 0.87
FLTrust-Late	33.64 ± 0.16	32.58 ± 0.73	31.27 ± 0.93	30.11 ± 0.60
HGDA ($W = 100$)	35.32 ± 0.81	$\mathbf{35.25} \pm 1.08$	$\mathbf{33.82} \pm 1.43$	$\mathbf{31.80} \pm 0.32$

Table A.37: Tiny ImageNet under Mimic. Best test accuracy (%).

Method	10%	20%	30%	40%
FedAvg (benign only)	36.02 ± 0.81	34.59 ± 0.23	33.50 ± 0.40	31.77 ± 0.58
FedAvg	35.54 ± 0.11	34.40 ± 0.49	32.73 ± 0.52	30.52 ± 1.45
Multi-Krum	34.73 ± 0.78	33.68 ± 0.19	30.74 ± 0.14	27.38 ± 0.91
Bulyan	35.62 ± 0.38	N/A	N/A	N/A
Coord. median	35.75 ± 0.40	34.43 ± 0.12	31.27 ± 0.70	28.55 ± 1.06
Trimmed mean	36.17 ± 0.49	34.62 ± 0.29	32.12 ± 0.89	30.04 ± 0.84
RFA	35.45 ± 0.15	32.93 ± 0.42	29.79 ± 0.41	25.83 ± 0.45
Fed-NGA	$\mathbf{36.57} \pm 0.16$	$\underline{34.85} \pm 0.12$	$\underline{33.33} \pm 0.58$	$\underline{30.63} \pm 0.56$
FLTrust-Early	34.84 ± 0.22	33.75 ± 0.33	32.05 ± 0.22	29.88 ± 0.46
FLTrust-Late	34.60 ± 0.22	33.30 ± 0.33	32.92 ± 0.24	29.99 ± 0.49
HGDA ($W = 100$)	$\underline{36.18} \pm 1.10$	$\mathbf{36.77} \pm 0.56$	$\mathbf{34.60} \pm 0.52$	$\mathbf{33.61} \pm 1.26$

Table A.38: Tiny ImageNet under IPM. Best test accuracy (%).

Method	10%	20%	30%	40%
FedAvg (benign only)	35.46 ± 0.61	35.07 ± 0.20	33.62 ± 0.32	31.38 ± 0.95
FedAvg	$\underline{36.51} \pm 0.32$	35.57 ± 0.33	$\underline{34.24} \pm 0.35$	$\underline{31.96} \pm 1.52$
Multi-Krum	35.35 ± 0.55	31.51 ± 0.72	11.31 ± 5.02	0.84 ± 0.14
Bulyan	35.59 ± 0.37	N/A	N/A	N/A
Coord. median	36.45 ± 0.23	35.15 ± 0.11	18.75 ± 2.86	0.99 ± 0.16
Trimmed mean	$\mathbf{36.65} \pm 0.09$	$\underline{35.75} \pm 0.21$	31.41 ± 0.62	19.32 ± 2.19
RFA	36.02 ± 0.16	32.85 ± 2.01	5.07 ± 1.20	0.93 ± 0.19
Fed-NGA	36.31 ± 0.32	32.21 ± 1.00	13.31 ± 0.90	0.93 ± 0.10
FLTrust-Early	34.28 ± 0.52	33.97 ± 0.49	32.62 ± 0.94	30.95 ± 0.61
FLTrust-Late	34.90 ± 0.56	33.60 ± 0.29	32.37 ± 0.53	30.38 ± 0.83
HGDA ($W = 100$)	$\underline{36.51} \pm 1.23$	$\mathbf{36.05} \pm 1.31$	$\mathbf{34.34} \pm 1.48$	$\mathbf{32.91} \pm 0.43$

C.6 REPEATED CIFAR-10 EVALUATIONS

The three-seed evaluations on MNIST, FEMNIST, Purchase-100 and Tiny ImageNet cover every clean and attacked condition in their respective grids (Appendices C.1 and C.3–C.5). Across these conditions, the standard deviation of HGDA's best accuracy stays below 0.7 percentage points on MNIST and FEMNIST and below 1.8 points on Purchase-100 and Tiny ImageNet. These results show limited run-to-run variation across different datasets, attacks and Byzantine fractions.

Table A.39 gives the individual CIFAR-10 results for all three seeds under sign-flip with 20% Byzantine clients. HGDA reaches 82.78–83.44% accuracy with $W = 10$ and 82.41–83.51% with $W = 100$. Both window settings outperform every evaluated robust baseline in each of the three runs, showing that the strong sign-flip performance persists across seeds.

Table A.39: CIFAR-10 with 20% Byzantine clients. Best test accuracy (%).

	Sign-flip				Label-flip
Method	Seed 2025	Seed 2026	Seed 2027	Mean	Mean
FedAvg (benign only)	84.92	84.33	85.36	84.87	84.87
Multi-Krum	75.97	73.93	77.82	75.91	81.51
Bulyan	72.76	70.69	72.64	72.03	76.14
Coord. median	76.21	76.42	75.21	75.95	82.47
Trimmed mean	77.17	76.86	75.70	76.57	83.10
RFA	78.16	80.03	76.68	78.29	83.86
Fed-NGA	78.46	80.76	76.76	78.66	84.54
FLTrust-Early	82.20	81.23	81.33	81.59	82.29
FLTrust-Late	71.33	67.33	70.76	69.81	80.89
BOBA	80.32	80.76	80.54	80.54	82.92
FLDetector+Median	76.72	76.47	75.39	76.19	82.38
HGDA ($W = 10$)	82.92	83.44	82.78	83.05	84.17
HGDA ($W = 100$)	83.51	82.41	82.77	82.90	84.36

C.7 LEARNING WITHIN SHORTER COMMUNICATION BUDGETS

We examine learning progress before the maximum evaluation horizon using shorter cutoffs of the main-experiment trajectories.

CIFAR-10. Under sign-flip with 20% Byzantine clients, HGDA reaches 95% of its best accuracy over 1500 rounds within 100 rounds for $W = 100$ and within 250 rounds for $W = 10$. Its best accuracies within 500 rounds are 80.66% and 81.47%, respectively, both exceeding 97% of their corresponding 1500-round best accuracies.

MNIST. We compare HGDA's best test accuracy within the first 100 and 200 rounds, using $W = 10$ and $W = 100$. The comparison covers clean training and all 43 attacked conditions over three seeds. With $W = 100$, HGDA reaches at least 97.50% mean best accuracy across the attacked conditions within 100 rounds and at least 97.59% within 200 rounds.

Table A.40: HGDA on MNIST. Best test accuracy within 100 and 200 rounds (%).

		100 rounds		200 rounds	
Attack	m (%)	$W = 10$	$W = 100$	$W = 10$	$W = 100$
Clean	0	97.80 ± 0.19	98.72 ± 0.08	98.31 ± 0.31	98.76 ± 0.07
Sign-flip	20	97.37 ± 0.25	98.66 ± 0.01	97.89 ± 0.42	98.68 ± 0.04
	40	97.33 ± 0.17	98.53 ± 0.09	97.85 ± 0.29	98.56 ± 0.08
	50	97.03 ± 0.32	98.32 ± 0.23	97.73 ± 0.34	98.33 ± 0.23
	60	97.03 ± 0.42	98.12 ± 0.28	97.60 ± 0.19	98.12 ± 0.28
	80	96.23 ± 0.52	97.50 ± 0.49	96.96 ± 0.46	97.59 ± 0.46
Label-flip	20	97.59 ± 0.32	98.65 ± 0.08	98.10 ± 0.45	98.67 ± 0.11
	40	97.68 ± 0.24	98.61 ± 0.03	98.26 ± 0.27	98.65 ± 0.06
	50	97.53 ± 0.31	98.45 ± 0.23	98.52 ± 0.06	98.47 ± 0.24
	60	97.47 ± 0.20	98.27 ± 0.11	98.30 ± 0.27	98.36 ± 0.07
	80	96.68 ± 0.42	97.72 ± 0.28	97.69 ± 0.22	97.83 ± 0.34
Gaussian	20	97.58 ± 0.34	98.68 ± 0.05	98.07 ± 0.35	98.69 ± 0.05
	40	97.61 ± 0.16	98.58 ± 0.10	98.27 ± 0.30	98.64 ± 0.09
	50	97.58 ± 0.18	98.43 ± 0.13	98.03 ± 0.19	98.49 ± 0.17
	60	97.23 ± 0.20	98.32 ± 0.08	97.74 ± 0.33	98.32 ± 0.08
	80	96.57 ± 0.53	97.85 ± 0.06	97.39 ± 0.28	97.88 ± 0.08

Table A.41: HGDA on MNIST. Best test accuracy within 100 and 200 rounds (%).

		100 rounds		200 rounds	
Attack	m (%)	$W = 10$	$W = 100$	$W = 10$	$W = 100$
SameValue	20	97.57 ± 0.33	98.67 ± 0.03	98.07 ± 0.38	98.71 ± 0.07
	40	97.61 ± 0.16	98.59 ± 0.11	98.25 ± 0.28	98.65 ± 0.08
	50	97.58 ± 0.17	98.43 ± 0.13	98.04 ± 0.19	98.49 ± 0.17
	60	97.24 ± 0.20	98.31 ± 0.09	97.78 ± 0.27	98.33 ± 0.06
	80	96.63 ± 0.46	97.85 ± 0.03	97.44 ± 0.20	97.87 ± 0.05
ALIE	20	97.58 ± 0.32	98.67 ± 0.01	98.19 ± 0.32	98.69 ± 0.01
	40	97.63 ± 0.18	98.62 ± 0.04	98.25 ± 0.30	98.63 ± 0.06
	50	97.64 ± 0.13	98.43 ± 0.17	98.10 ± 0.16	98.46 ± 0.17
Mimic	20	97.58 ± 0.32	98.67 ± 0.07	98.10 ± 0.34	98.69 ± 0.08
	40	97.63 ± 0.16	98.57 ± 0.05	98.27 ± 0.28	98.63 ± 0.07
	50	97.57 ± 0.21	98.41 ± 0.13	98.00 ± 0.13	98.48 ± 0.14
	60	97.27 ± 0.17	98.31 ± 0.08	97.77 ± 0.29	98.31 ± 0.07
	80	96.64 ± 0.57	97.79 ± 0.07	97.47 ± 0.17	97.82 ± 0.08

Table A.42: HGDA on MNIST. Best test accuracy within 100 and 200 rounds (%).

		100 rounds		200 rounds	
Attack	m (%)	$W = 10$	$W = 100$	$W = 10$	$W = 100$
Min-Sum	20	97.56 ± 0.30	98.66 ± 0.01	98.09 ± 0.41	98.72 ± 0.01
	40	97.67 ± 0.12	98.61 ± 0.04	98.17 ± 0.34	98.64 ± 0.03
	50	97.66 ± 0.23	98.44 ± 0.20	98.09 ± 0.32	98.48 ± 0.20
	60	97.32 ± 0.13	98.30 ± 0.15	97.79 ± 0.34	98.36 ± 0.10
	80	96.53 ± 0.57	97.81 ± 0.11	97.36 ± 0.31	97.86 ± 0.07
Min-Max	20	97.59 ± 0.32	98.67 ± 0.03	98.10 ± 0.33	98.69 ± 0.02
	40	97.69 ± 0.14	98.61 ± 0.03	98.28 ± 0.25	98.64 ± 0.03
	50	97.64 ± 0.15	98.41 ± 0.17	98.08 ± 0.26	98.48 ± 0.18
	60	97.28 ± 0.22	98.29 ± 0.08	97.84 ± 0.40	98.33 ± 0.01
	80	96.63 ± 0.42	97.81 ± 0.05	97.35 ± 0.19	97.84 ± 0.06
IPM	20	97.54 ± 0.38	98.68 ± 0.05	97.90 ± 0.44	98.70 ± 0.05
	40	97.62 ± 0.16	98.58 ± 0.10	98.12 ± 0.17	98.64 ± 0.09
	50	97.47 ± 0.33	98.42 ± 0.13	97.94 ± 0.16	98.45 ± 0.16
	60	97.29 ± 0.20	98.31 ± 0.08	97.53 ± 0.09	98.31 ± 0.08
	80	96.54 ± 0.45	97.86 ± 0.04	97.46 ± 0.28	97.87 ± 0.05

D ABLATIONS AND REFERENCE-ALIGNMENT OBSERVATIONS

CIFAR-10 ablations use full participation and 1500 rounds. Reference and grouping comparisons use $m = 80\%$ and $W = 10$, while window comparisons use $m = 50\%$ and both references. Lines and bands show the mean and minimum–maximum range within a centered 25-round smoothing window.

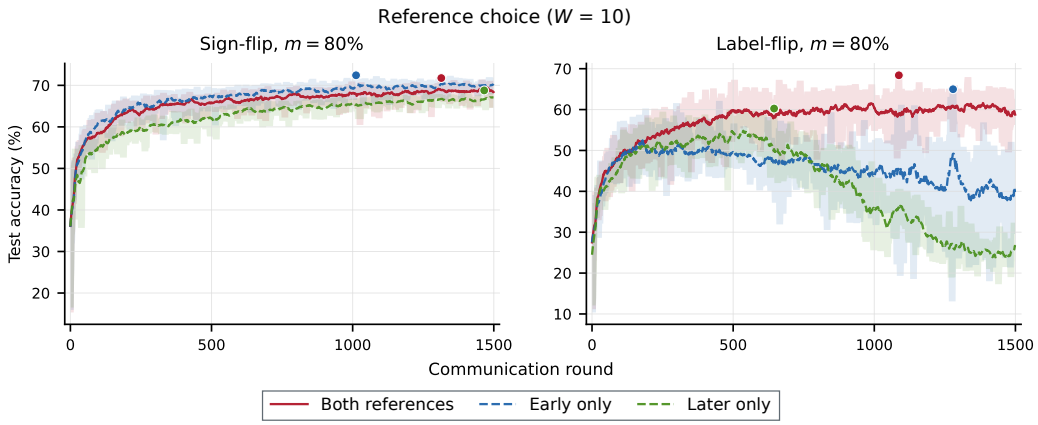


Figure A.1: Effect of reference choice on CIFAR-10 training.

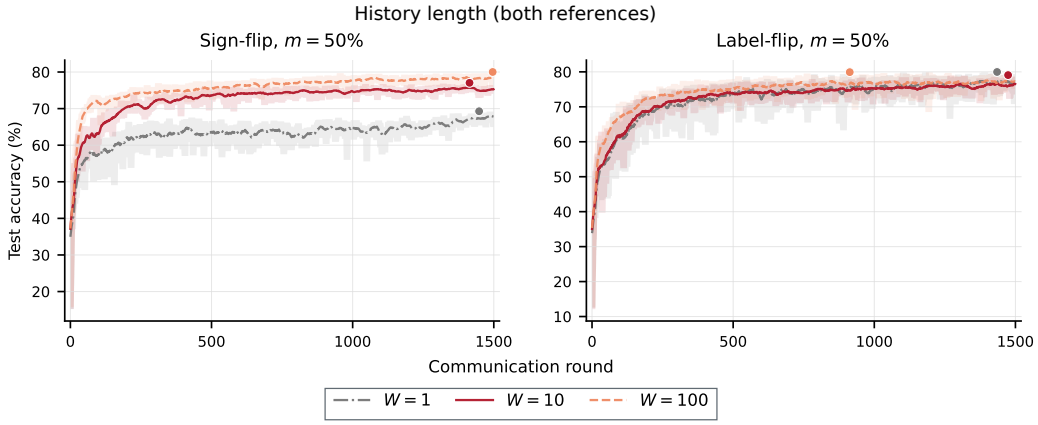


Figure A.2: Effect of history length on CIFAR-10 training.

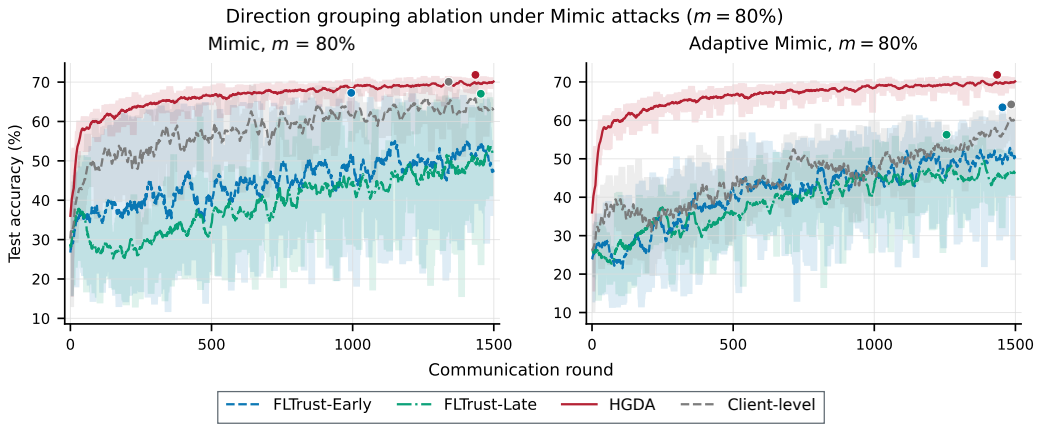


Figure A.3: Client-level aggregation versus HGDA's direction-level aggregation under Mimic and adaptive Mimic.

The client-level variant retains HGDA's scoring rule and server optimizer but aggregates clients separately. Copies of one direction therefore receive a total weight proportional to the sum of their scores, rather than the maximum used by HGDA. Its best accuracy is 70.08% under Mimic and 64.14% under adaptive Mimic, compared with the HGDA reference of 71.83% (Table 2(c)).

D.1 Client assessment with full and partial participation

We examine client assessment on the logged HGDA trajectories from two main-experiment settings in Appendix A.3. CIFAR-10 uses 100 clients with full participation, 50% Byzantine clients and 1500 rounds. Tiny ImageNet uses 100 clients with ten participants per round, 40% Byzantine clients and 500 rounds (three seeds). For each observed update, we evaluate Equation (4) using either the current round alone or the most recent 100 global rounds, with $\lambda = 1$. Early-only and later-only assessment use the corresponding reference score, while joint assessment uses their minimum before clipping at zero. The updates and server references are shared across these scoring comparisons.

Figure 3 separates the three reference choices and compares current-round with historical positive-score rates under two attacks. Rates use all 150000 observed updates on CIFAR-10 and all 5000 observations per seed on Tiny ImageNet, including the initial rounds before the window is full. They are computed separately for benign and Byzantine participations.

The references provide complementary checks. Under Min-Max, 78.5% of Byzantine updates have positive historical scores with the early reference, compared with 0.6% with the later reference. Under sign-flip, the early reference gives 1.8% positive scores, compared with 14.8% for the later reference. Taking the minimum retains both checks, giving 0.6% and 0.9%, respectively. With joint assessment, history raises the benign positive-score rate from 54.5% to 78.2% under sign-flip and lowers the Byzantine rate from 10.8% to 0.9%. Table A.43 gives all six attack comparisons.

Table A.43: Positive-score rates on CIFAR-10 with 50% Byzantine clients (%).

		Current round			With history ($W = 100$)		
Attack	Clients	Early	Later	Both	Early	Later	Both
Label-flip	Benign	70.0	68.3	54.1	88.6	82.0	78.0
	Byzantine	72.0	12.2	11.6	82.1	6.0	5.8
Sign-flip	Benign	70.6	68.9	54.5	90.7	81.5	78.2
	Byzantine	22.5	28.1	10.8	1.8	14.8	0.9
IPM	Benign	70.1	69.4	55.1	88.9	84.6	80.2
	Byzantine	10.2	15.8	3.9	0.0	0.0	0.0
Min-Max	Benign	72.7	71.7	57.8	91.7	87.9	83.2
	Byzantine	56.6	20.7	15.3	78.5	0.6	0.6
ALIE	Benign	73.7	69.8	57.2	95.0	84.3	83.5
	Byzantine	26.2	5.0	2.2	0.5	0.4	0.4
Mimic	Benign	70.1	68.7	54.8	89.2	83.5	79.8
	Byzantine	89.1	93.7	84.7	100.0	100.0	100.0

Table A.44 extends the comparison to partial participation on Tiny ImageNet. Under label-flip, the early reference alone gives positive scores to most Byzantine updates, while the later reference scores many of them negatively. With both references, history raises the benign positive-score rate from 74.9% to 88.8% and lowers the Byzantine rate from 50.6% to 41.7%.

Table A.44: Positive-score rates (%) on Tiny ImageNet with 40% Byzantine clients over three seeds. All scoring variants use the same logged updates.

		Current round			With history ($W = 100$)		
Attack	Clients	Early	Later	Both	Early	Later	Both
Label-flip	Benign	76.6 ± 0.9	92.2 ± 0.9	74.9 ± 0.6	89.0 ± 1.2	99.2 ± 0.3	88.8 ± 1.0
	Byzantine	86.9 ± 3.0	51.7 ± 4.4	50.6 ± 4.1	96.2 ± 0.7	42.9 ± 2.1	41.7 ± 1.9
Sign-flip	Benign	82.6 ± 5.7	89.0 ± 1.7	80.7 ± 4.9	92.0 ± 5.2	97.6 ± 1.2	91.6 ± 5.1
	Byzantine	0.9 ± 0.2	0.8 ± 0.1	0.2 ± 0.2	1.2 ± 0.3	0.5 ± 0.5	0.1 ± 0.1
IPM	Benign	80.4 ± 0.5	89.3 ± 0.7	78.9 ± 0.4	89.1 ± 0.9	97.6 ± 0.2	88.9 ± 1.0
	Byzantine	10.9 ± 0.4	5.0 ± 1.2	4.7 ± 0.9	1.5 ± 0.4	0.0 ± 0.1	0.0 ± 0.1
Mimic	Benign	81.4 ± 5.9	89.0 ± 1.3	79.6 ± 5.0	91.4 ± 6.9	97.3 ± 1.3	90.9 ± 6.8
	Byzantine	76.9 ± 8.2	87.5 ± 1.5	75.2 ± 6.9	90.2 ± 9.8	97.8 ± 1.3	89.9 ± 9.4

Under sign-flip, history raises the benign positive-score rate from 80.7% to 91.6% with both references, while the Byzantine rate remains below 0.2%. Under IPM, the benign rate increases from 78.9% to 88.9% and the Byzantine rate falls from 4.7% to below 0.1%. Mimic behaves differently because it copies a benign update. Such copies can receive positive assessments along with genuine benign contributions. Their repeated influence is controlled by direction-level weighting rather than by requiring the scoring stage to identify their origin (Section 3.3).

E Trusted-data coverage and sensitivity

E.1 Class coverage

The trusted sets in the three partial-participation tasks are not balanced by class. Table A.45 records their realized coverage for seeds 2025, 2026 and 2027.

Table A.45: Class coverage of the trusted sets.

			2025		2026		2027	
Task	Examples	Classes	Present	Absent	Present	Absent	Present	Absent
FEMNIST	620	62	57	5	57	5	59	3
Purchase-100	100	100	64	36	65	35	60	40
Tiny ImageNet	100	200	83	117	84	116	72	128

The Purchase class analysis uses the main training settings in Table A.3. We select each run's best overall-accuracy round, then average its logged class accuracies separately over classes present in and absent from D_s. Table A.46 covers clean training and all four attacks at 10–60% Byzantine clients over three seeds.

Table A.46: Class-conditioned accuracy on Purchase-100 (%).

Condition	Overall	Present classes	Absent classes
Clean	91.79 ± 0.45	91.88 ± 0.48	90.66 ± 0.10
Sign-flip 10%	90.95 ± 0.85	91.29 ± 0.89	89.78 ± 0.91
Label-flip 10%	89.28 ± 1.10	89.58 ± 1.07	87.78 ± 1.15
Mimic 10%	91.40 ± 0.99	91.99 ± 0.84	89.62 ± 0.90
IPM 10%	90.97 ± 1.14	91.49 ± 1.27	88.98 ± 1.06
Sign-flip 20%	89.92 ± 0.92	90.53 ± 0.51	88.20 ± 1.14
Label-flip 20%	87.76 ± 1.44	87.95 ± 1.77	85.57 ± 1.63
Mimic 20%	90.97 ± 0.25	91.28 ± 0.36	89.65 ± 0.23
IPM 20%	90.37 ± 0.67	90.56 ± 0.91	88.95 ± 0.58
Sign-flip 30%	89.31 ± 0.70	89.44 ± 1.07	88.14 ± 0.54
Label-flip 30%	86.58 ± 1.35	86.89 ± 1.51	84.35 ± 1.38
Mimic 30%	90.46 ± 0.51	90.66 ± 0.44	89.00 ± 0.12
IPM 30%	89.96 ± 0.21	90.42 ± 0.54	88.04 ± 0.60
Sign-flip 40%	88.03 ± 1.04	88.23 ± 0.88	85.80 ± 0.79
Label-flip 40%	85.89 ± 0.80	86.19 ± 0.76	83.34 ± 1.03
Mimic 40%	89.54 ± 0.61	89.93 ± 0.53	87.52 ± 0.61
IPM 40%	88.85 ± 0.93	89.14 ± 0.88	86.83 ± 0.92
Sign-flip 50%	86.90 ± 0.87	87.40 ± 0.79	84.90 ± 1.06
Label-flip 50%	84.84 ± 1.33	85.07 ± 1.85	81.81 ± 1.59
Mimic 50%	88.79 ± 0.59	89.07 ± 0.54	86.98 ± 0.61
IPM 50%	87.96 ± 0.79	88.29 ± 0.78	86.29 ± 1.54
Sign-flip 60%	85.40 ± 1.77	85.54 ± 1.41	83.68 ± 2.20
Label-flip 60%	83.27 ± 0.98	83.32 ± 0.68	80.39 ± 1.50
Mimic 60%	87.58 ± 0.56	87.96 ± 0.34	85.55 ± 0.99
IPM 60%	86.35 ± 0.54	87.28 ± 0.17	83.28 ± 0.81

E.2 History under sparse participation

The FEMNIST window comparison in Table 2(d) uses the main 500-round settings in Table A.3, with 100 of 3400 naturally partitioned writer clients per round and 10% sign-flip. We fix $\lambda = 1$ and $(E_s, E_l) = (1, 2)$ and vary W. The window spans global rounds and uses each client's available observations.

E.3 Score regularization

We vary λ under the same FEMNIST settings with $W = 100$ (Table 2(e)).

Table A.47: History-window and score-regularizer sensitivity on FEMNIST. Best test accuracy (%) over three seeds.

(a) Window length ($\lambda = 1$)

W	Accuracy (%)
10	84.94 ± 0.35
50	84.74 ± 0.29
100	84.77 ± 0.32
300	84.89 ± 0.29

(b) Regularizer ($W = 100$)

λ	Accuracy (%)
0.1	84.86 ± 0.22
1	84.77 ± 0.32
5	84.77 ± 0.34
10	85.03 ± 0.26

E.4 Reference checkpoints

Checkpoint placement on Purchase. On Purchase-100, we examine the placement of each reference checkpoint while keeping the other fixed at its default value. The default pair is $(E_s, E_l) = (1, 8)$, with the later checkpoint determined by the step-count rule in Appendix A.3. We vary E_l with $E_s = 1$ (Table A.48) and vary E_s with $E_l = 8$ (Table A.49). Each sweep considers 1, 2, 4, 5, 7 and 8, with $W = 100$, $\lambda = 1$ and the main training settings in Table A.3. The $(1, 1)$ and $(8, 8)$ settings reduce to single-reference assessment at the early and later default checkpoints, respectively.

Table A.48: Later-reference checkpoint sweep on Purchase-100, with $E_s = 1$. Best test accuracy (%) over three seeds.

(a) Sign-flip

(E_s, E_l)	$m = 40\%$	$m = 50\%$	$m = 60\%$	$m = 80\%$
$(1, 1)$	88.41 ± 1.44	87.15 ± 0.98	85.24 ± 0.71	79.84 ± 0.60
$(1, 2)$	87.94 ± 1.22	86.97 ± 0.75	85.00 ± 1.10	80.06 ± 0.35
$(1, 4)$	88.12 ± 1.66	86.99 ± 1.13	85.49 ± 1.10	79.83 ± 0.20
$(1, 5)$	88.01 ± 1.10	86.59 ± 0.89	85.15 ± 0.92	79.87 ± 0.79
$(1, 7)$	88.17 ± 1.23	86.97 ± 0.68	85.25 ± 1.43	79.90 ± 0.46
$(1, 8)$	88.03 ± 1.04	86.90 ± 0.87	85.40 ± 1.77	79.94 ± 0.59

(b) Label-flip

(E_s, E_l)	$m = 40\%$	$m = 50\%$	$m = 60\%$	$m = 80\%$
$(1, 1)$	86.11 ± 1.06	84.88 ± 1.00	83.25 ± 0.49	78.49 ± 0.32
$(1, 2)$	85.76 ± 1.19	84.69 ± 0.94	83.07 ± 0.41	77.97 ± 0.29
$(1, 4)$	85.81 ± 0.92	84.78 ± 1.01	83.25 ± 0.20	77.85 ± 0.75
$(1, 5)$	86.01 ± 1.13	84.77 ± 0.64	83.03 ± 0.61	78.23 ± 0.69
$(1, 7)$	85.84 ± 1.18	84.68 ± 1.29	83.33 ± 0.69	78.50 ± 0.17
$(1, 8)$	85.89 ± 0.80	84.84 ± 1.33	83.27 ± 0.98	78.27 ± 0.47

Table A.49: Early-reference checkpoint sweep on Purchase-100, with $E_l = 8$. Best test accuracy (%) over three seeds.

(a) Sign-flip

(E_s, E_l)	$m = 40\%$	$m = 50\%$	$m = 60\%$	$m = 80\%$
$(1, 8)$	88.03 ± 1.04	86.90 ± 0.87	85.40 ± 1.77	79.94 ± 0.59
$(2, 8)$	88.08 ± 1.13	86.95 ± 0.34	85.24 ± 1.02	79.38 ± 0.95
$(4, 8)$	87.84 ± 1.31	86.77 ± 0.44	85.40 ± 1.02	79.62 ± 0.36
$(5, 8)$	88.07 ± 1.29	86.94 ± 0.25	84.87 ± 0.62	79.61 ± 0.84
$(7, 8)$	87.96 ± 1.38	86.71 ± 0.58	84.99 ± 0.81	79.80 ± 0.28
$(8, 8)$	88.21 ± 0.93	86.71 ± 0.42	84.78 ± 1.15	79.48 ± 0.89

(b) Label-flip

(E_s, E_l)	$m = 40\%$	$m = 50\%$	$m = 60\%$	$m = 80\%$
$(1, 8)$	85.89 ± 0.80	84.84 ± 1.33	83.27 ± 0.98	78.27 ± 0.47
$(2, 8)$	86.06 ± 0.85	84.93 ± 1.12	83.54 ± 1.20	78.26 ± 0.22
$(4, 8)$	86.24 ± 0.97	84.95 ± 1.00	83.31 ± 0.78	78.26 ± 0.38
$(5, 8)$	85.95 ± 1.10	84.94 ± 1.18	83.20 ± 0.56	78.14 ± 0.42
$(7, 8)$	86.21 ± 1.48	85.07 ± 0.88	83.47 ± 0.63	78.51 ± 0.57
$(8, 8)$	86.28 ± 1.39	84.51 ± 0.66	83.08 ± 0.90	78.40 ± 0.78

Within each attack/fraction condition, mean best accuracy varies by 0.20–0.65 percentage points when changing E_l and 0.24–0.61 points when changing E_s. HGDA maintains similar accuracy across both checkpoint sweeps under sign-flip and label-flip at all four Byzantine fractions.

Single- and dual-reference assessment on CIFAR-10. We compare early-only, later-only and joint assessment with 80% Byzantine clients and $W = 10$. The references are taken after one and 20 server-training epochs. Table A.50 gives the best accuracies, and Figure A.1 shows the corresponding training trajectories.

Table A.50: Single- and dual-reference assessment on CIFAR-10. Best test accuracy (%).

References	(E_s, E_l)	Sign-flip	Label-flip
Early only	$(1, 1)$	72.42	64.98
Later only	$(20, 20)$	68.78	60.26
Both	$(1, 20)$	71.76	68.39

Under label-flip, using both references improves best accuracy by 3.41 percentage points over early-only assessment and 8.13 points over later-only assessment. Under sign-flip, it reaches 71.76%, close to 72.42% with the early reference alone and above 68.78% with the later reference alone. Joint assessment thus improves label-flip performance while retaining comparable sign-flip accuracy to the stronger single reference.

F Resilience to on-off attacks

On-off attacks test whether HGDA can continue learning as Byzantine clients alternate between benign training and poisoning. We expand the evaluation in Section 4.5 with full training trajectories and phase-specific accuracies, then examine how the sliding window updates client assessments across these changes in behavior. The experiments use CIFAR-10 with 50% Byzantine clients, full participation and the attack protocols in Appendix B.2. We compare $W = 10$ and $W = 100$ over 1500 rounds.

F.1 Resilience across attack schedules

With $W = 100$, HGDA reaches 79.38–85.29% best accuracy across the scheduled and adaptive settings (Table 2(f)). This resilience extends to attackers that use the preceding round's client scores to time their switches. Under adaptive sign-flip and label-flip, best accuracy reaches 82.10–84.53% and 84.56–84.80%, respectively, compared with 80.01% and 79.92% under continuous attacks. The adaptive trajectories in Figure A.4 show continued learning through repeated switches. Figure A.5 shows the corresponding training behavior for fixed schedules alongside continuous attacks.

To examine performance within each cycle, Table A.51 separates attack-on and attack-off rounds. On and Off are the mean test accuracies over their respective phases in rounds 1300–1499, and $\Delta = \mathrm{Off} - \mathrm{On}$ is their difference in percentage points. On share gives the fraction of rounds spent attacking within each cycle.

Under scheduled sign-flip, mean accuracies in the two phases differ by less than 0.6 percentage points. Label-flip produces larger within-cycle fluctuations, followed by repeated returns to high accuracy during benign phases. With $W = 100$ and the $(1, 2)$ schedule, mean accuracy is 38.00% during attack-on rounds and 83.81% during attack-off rounds. With equal on/off durations, the corresponding means are 77.54% and 80.44%. Figure A.6 shows the last 80 rounds, making these changes within individual cycles visible.

Table A.51: Accuracy during attack-on and attack-off phases on CIFAR-10 (%).

			Sign-flip				Label-flip			
Schedule (on, off)	On share	W	Best	Off	On	Δ	Best	Off	On	Δ
(1, 2)	33%	10	83.83	82.15	81.71	0.44	83.86	82.30	40.15	42.15
		100	83.82	82.15	81.98	0.17	85.29	83.81	38.00	45.81
(1, 1)	50%	10	80.01	78.25	78.14	0.12	83.01	81.23	64.37	16.86
		100	80.88	79.12	79.16	−0.04	82.39	80.44	77.54	2.90
(2, 1)	67%	10	79.14	77.38	77.30	0.08	82.24	80.58	58.38	22.20
		100	79.67	78.05	77.97	0.09	82.68	80.91	67.40	13.51
(4, 1)	80%	10	79.23	77.57	77.45	0.12	80.27	78.61	76.44	2.17
		100	79.38	77.96	77.40	0.56	80.39	78.91	74.73	4.18
Continuous	100%	10	77.04	—	—	—	79.08	—	—	—
		100	80.01	—	—	—	79.92	—	—	—

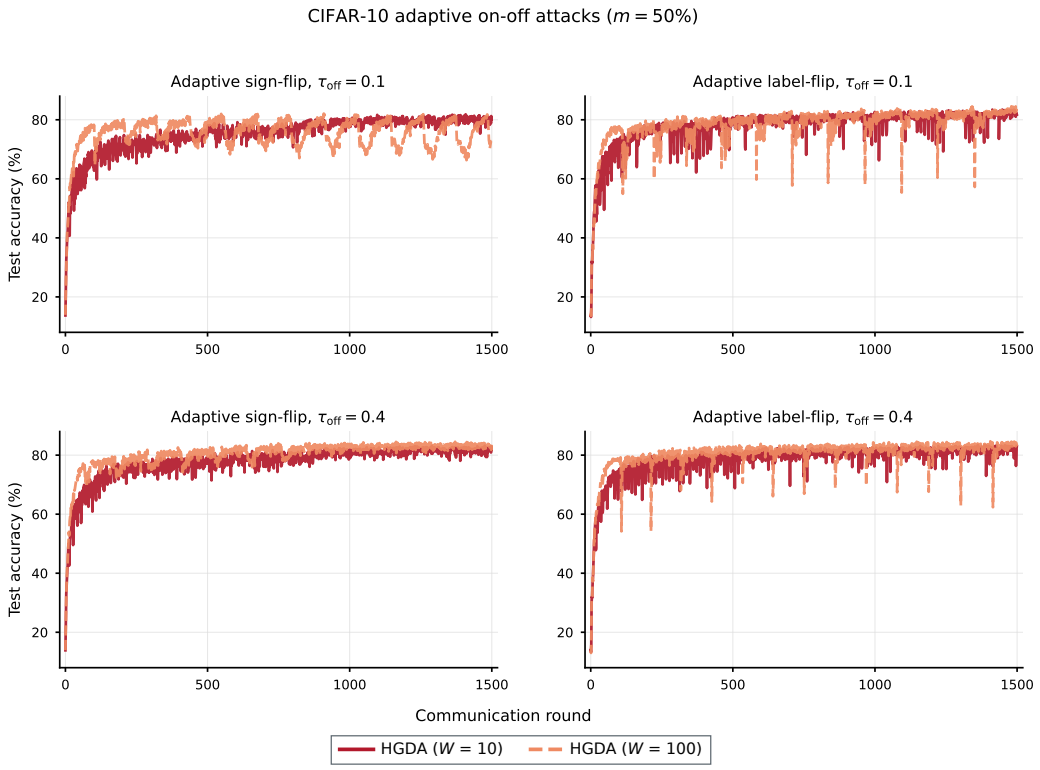


Figure A.4: Training under adaptive on-off sign-flip and label-flip attacks.

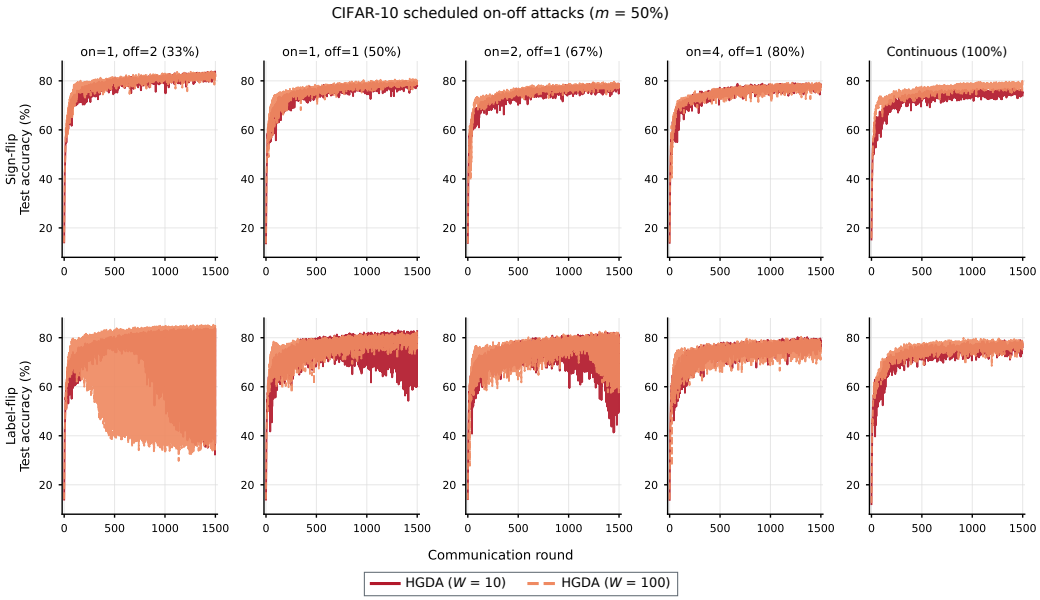


Figure A.5: Training under scheduled on-off attacks. The rightmost panels show continuous attacks.

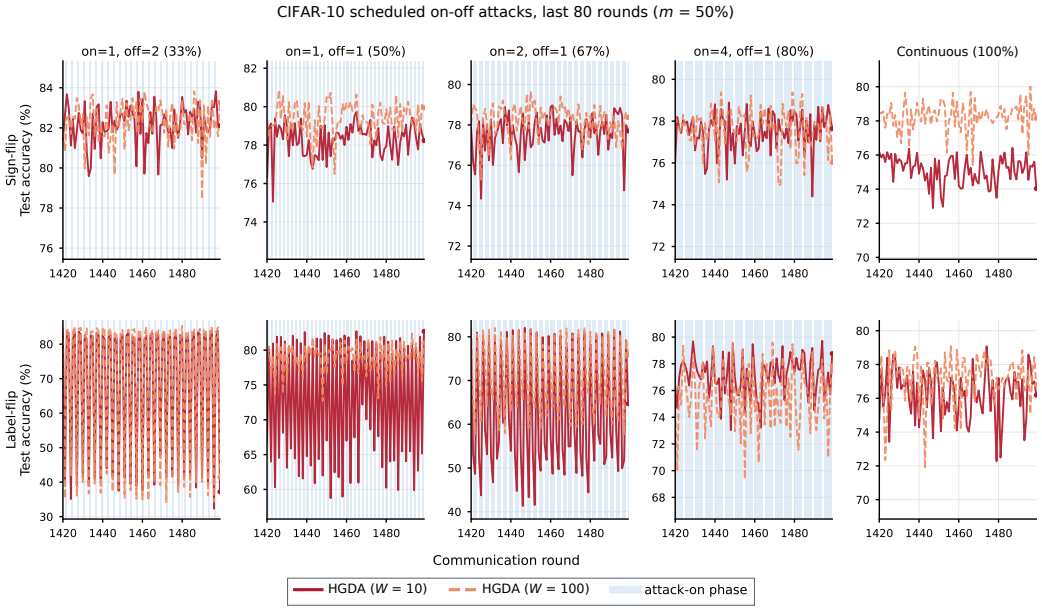


Figure A.6: The last 80 rounds of scheduled on-off training. Shading marks attack-on phases.

F.2 How the window responds to changing behavior

The sliding window carries recent evidence across a behavior switch while replacing older observations as training proceeds. To describe this process, fix a client and a reference $h \in \{s, l\}$. Write $M_h = P_h - N_h$ for the client's net support under that reference, using the evidence sums in Equation (3). Starting from any window, index subsequent global rounds by $j = 1, 2, \ldots$ and denote the balance after j rounds by $M_h^{(j)}$.

Let e_h^j be the alignment entering the window in round j, with $e_h^j = 0$ if the client does not participate in that round. Let o_h^j be the alignment leaving the window, with $o_h^j = 0$ if no observation was recorded in the expiring round. Advancing the window therefore gives

$$M_h^{(j)} = M_h^{(j-1)} + e_h^j - o_h^j.$$

New support and expiring opposition increase this balance. New opposition and expiring support decrease it. Accumulating these increases and decreases over k rounds gives

$$R_h(k) = \sum_{j=1}^{k} ([e_h^j]_+ + [-o_h^j]_+), \qquad C_h(k) = \sum_{j=1}^{k} ([-e_h^j]_+ + [o_h^j]_+).$$

Since $e - o = [e]_+ + [-o]_+ - [-e]_+ - [o]_+$, summing the per-round changes yields

$$M_h^{(k)} = M_h^{(0)} + R_h(k) - C_h(k).$$

The score in Equation (4) is positive only when both reference histories have positive net support. Thus, opposing alignments can offset support accumulated before a switch, and that earlier support also expires as the window advances. Conversely, new supporting alignments and the expiration of past opposition can restore a positive balance. The assessment reflects recent behavior through both the observations entering the window and those leaving it.

G Properties of scoring and directional aggregation

We first formalize how unit normalization removes update magnitude from aggregation and how groupwise maximum scores control the influence of repeated directions. We then examine how the number of observations within a window affects the sign of a client's score.

Proposition 1 (Positive-scale invariance)**.** *Fix the global model, server references and prior alignment histories. Replacing a nonzero current increment Δ_i^t by $c_i\Delta_i^t$, with $c_i > 0$, leaves its normalized direction and both current cosines unchanged. Consequently its score, direction-group assignment and the aggregate in Equation* (5) *are unchanged in exact arithmetic.*

Proof. Dividing $c_i\Delta_i^t$ by its norm cancels c_i. All subsequent scoring and grouping operations depend on that unit vector, its cosines and the fixed histories. □

Proposition 2 (Copy invariance at fixed group maxima)**.** *Fix the current directions and client scores. Adding submissions whose directions already occur in $\mathcal{G}_t$ leaves $\bar{u}_t$ unchanged if none increases the maximum score of its group and all other group maxima remain fixed.*

Proof. Both v_G^t and b_G^t in Equation (5) remain unchanged, hence so do their normalization and weighted sum. In the client-normalized implementation described in Appendix A.2, the common denominator in a_i^t cancels when the group maxima are normalized, yielding the same aggregate. □

The next result considers histories whose expected alignments remain positive under both references or negative under one. It relates the number of observed updates to whether a client receives a positive score.

Proposition 3 (Concentration for an observed window)**.** *Fix the observation indices $\mathcal{W}_{i,t}^{(W)}$, of cardinality $k \geq 1$. Conditional on these indices, suppose that the cosines for each reference are independent across observations and lie in $[-1, 1]$. If each observation has mean at least $\mu > 0$ under both references, then*

$$\Pr(s_i^t = 0) \leq 2\exp(-k\mu^2/2).$$

If there is a reference under which every observation has mean at most $-\mu$, then

$$\Pr(s_i^t > 0) \leq \exp(-k\mu^2/2).$$

Independence between the two references is not required.

Proof. Since the denominator in Equation (4) is positive, the sign of each reference score equals the sign of the sum of its k cosines. Hoeffding's inequality yields

$$\Pr\left(\sum_{\tau \in \mathcal{W}_{i,t}^{(W)}} r_{i,h}^{\tau} \leq 0\right) \leq \exp\left(-\frac{2(k\mu)^2}{4k}\right) = \exp(-k\mu^2/2)$$

under the positive-mean condition. A union bound over the two references gives the first result. Applying the same inequality to the negatives of the cosines for the negative-mean reference gives the second result, because a positive combined score requires that reference's sum to be positive. □

Under these conditions, the bounds decrease exponentially with the number of observations. More observations give a tighter upper bound on a zero score when mean alignment is positive under both references, and on a positive score when mean alignment is negative under one. Here k counts the updates observed within the window, rather than elapsed global rounds. It equals W under full participation once the window is filled, while partial participation can provide fewer observations within the same interval.

H Convergence analysis

HGDA uses trusted data to assess contributions, but learns from the directions supplied by clients. We analyze this distinction by following how signed evidence in each client's window determines direction weights and how the resulting aggregate advances the benign population's objective. The analysis first bounds the weight on directions supplied only by Byzantine clients. It then connects the retained benign directions to local training error and derives a stationarity bound for the server's Adam updates. The resulting bound separates the effects of client heterogeneity, residual adversarial influence and server optimization.

H.1 Objective and server-update sequence

Let $\mathcal{I}_{\mathrm{ben}}$ be the benign client population and define its learning objective as

$$F(x) = \sum_{i \in \mathcal{I}_{\mathrm{ben}}} p_i F_i(x), \qquad p_i \geq 0, \qquad \sum_{i \in \mathcal{I}_{\mathrm{ben}}} p_i = 1, \tag{6}$$

where F_i is client i's expected training loss and the population weights p_i are fixed. The variable $x \in \mathbb{R}^d$ comprises the trainable model parameters. We assume that F is bounded below by F_{inf} and has an L-Lipschitz gradient. This is the smooth nonconvex setting used in analyses of adaptive federated optimization (Reddi et al., 2021). All vector norms below are Euclidean.

We index the server updates that are actually applied by $k = 1, \ldots, T$. Let τ_k be the communication round of update k, let $x_{k-1} = \theta_{\tau_k}$ be the parameters before that update and write $U_k = \bar{u}_{\tau_k}$. Its learning rate is $\eta_k > 0$. Rounds without eligible updates leave the parameters and Adam state unchanged, so they do not advance k. Client histories continue to use communication rounds as defined in Equation (2).

The objective gradient and its unit descent direction before update k are

$$g_k = \nabla F(x_{k-1}), \qquad d_k = \begin{cases} -g_k/\|g_k\|, & g_k \neq 0, \\ 0, & g_k = 0. \end{cases} \tag{7}$$

The direction d_k provides a reference for the analysis. HGDA computes U_k from the client updates and their history-based scores.

H.2 From signed histories to direction weights

For a client observed in round t, let $n_i^t = |\mathcal{W}_{i,t}^{(W)}|$ be its number of observations in the current window. Define its smaller net support across the two references as

$$M_i^t = \left[\min_{h \in \{s,l\}} \left(P_{i,h}^{t,W} - N_{i,h}^{t,W} \right) \right]_+ . \tag{8}$$

This quantity retains the effect of opposition: M_i^t is zero if either reference has nonpositive net support. The next lemma translates this evidence into bounds on the client score and then on the weight assigned to adversarial directions.

At an applied round t, partition the eligible direction groups into $\mathcal{G}_t^{\mathrm{ben}}$, containing at least one benign client, and $\mathcal{G}_t^{\mathrm{adv}}$, containing only Byzantine clients. Define the score bounds

$$\underline{s}_i^t = \frac{M_i^t}{\sqrt{n_i^t + 2\lambda}}, \qquad \overline{s}_i^t = \frac{M_i^t}{\sqrt{M_i^t + 2\lambda}}, \tag{9}$$

and sum them over the corresponding directions:

$$\underline{B}_t = \sum_{G \in \mathcal{G}_t^{\mathrm{ben}}} \max_{i \in G \cap \mathcal{I}_{\mathrm{ben}}} \underline{s}_i^t, \qquad \overline{A}_t = \sum_{G \in \mathcal{G}_t^{\mathrm{adv}}} \max_{i \in G} \overline{s}_i^t. \tag{10}$$

Lemma 1 (Signed evidence and adversarial direction weight)**.** *Each observed client's score satisfies $\underline{s}_i^t \leq s_i^t \leq \overline{s}_i^t$. At every applied round, the total weight on directions supplied only by Byzantine clients satisfies*

$$\rho(t) := \sum_{G \in \mathcal{G}_t^{\mathrm{adv}}} \pi_G^t \leq \overline{\rho}_t := \frac{\overline{A}_t}{\overline{A}_t + \underline{B}_t}. \tag{11}$$

Proof. If $M_i^t = 0$, at least one reference score is nonpositive, so $s_i^t = 0$. Otherwise, both net supports are at least M_i^t. Since each alignment lies in $[-1, 1]$, the accumulated absolute evidence satisfies $P_{i,h}^{t,W} + N_{i,h}^{t,W} \leq n_i^t$. Both reference scores are therefore at least $M_i^t/\sqrt{n_i^t + 2\lambda}$. For a reference attaining the smaller net support M_i^t, its absolute evidence is at least M_i^t, giving the upper bound $M_i^t/\sqrt{M_i^t + 2\lambda}$ on that reference score and hence on s_i^t.

Let $A_t = \sum_{G \in \mathcal{G}_t^{\mathrm{adv}}} b_G^t$ and $B_t = \sum_{G \in \mathcal{G}_t^{\mathrm{ben}}} b_G^t$. Taking maxima within each group gives $A_t \le \overline{A}_t$ and $B_t \ge \underline{B}_t$. Every eligible client has positive score and hence positive M_i^t, so $\overline{A}_t + \underline{B}_t > 0$ at an applied round. The ratio $A_t/(A_t + B_t)$ is increasing in A_t and decreasing in B_t, proving Equation (11). □

The bound shows how the scoring choices enter aggregation. A positive score requires net support under both references within the window, and negative observations reduce that support. For example, n observations with mean alignment at least $\mu > 0$ under each reference give a score of at least $n\mu/\sqrt{n + 2\lambda}$. Under partial participation, n counts the observations actually retained within the global-round window. These statements concern the realized histories and apply as the window moves and client behavior changes.

The sums in Equation (10) run over distinct directions, with a maximum within each group. Several identities submitting one direction do not contribute a separate score term for every identity. A group containing a benign client also retains any larger score supplied by a Byzantine member through its actual weight b_G^t. The next part of the analysis accounts for how these weights distribute influence among benign directions.

H.3 Retained benign information and local training error

At applied update k, write $\rho_k = \rho(\tau_k)$ and combine the benign-supported directions as

$$c_k = \frac{\sum_{G \in \mathcal{G}_{\tau_k}^{\mathrm{ben}}} \pi_G^{\tau_k} v_G^{\tau_k}}{1 - \rho_k} \quad \text{if } \rho_k < 1, \qquad c_k = 0 \quad \text{if } \rho_k = 1. \tag{12}$$

Project this combined signal onto the population descent ray:

$$\chi_k = [\langle c_k, d_k \rangle]_+, \qquad \delta_k = \|c_k - \chi_k d_k\|. \tag{13}$$

Here χ_k measures the retained signal along the descent direction, while δ_k measures the residual after combining the benign directions. Individual benign updates can differ and their transverse components can offset one another. Since $\|c_k\| \le 1$, both χ_k and δ_k lie in $[0, 1]$.

Lemma 2 (Directional error of HGDA)**.** *For each applied update, define* $\alpha_k = (1 - \rho_k)\chi_k$ *and* $\xi_k = (1 - \rho_k)\delta_k + \rho_k$*. Then*

$$\|U_k\| \le 1, \qquad U_k = \alpha_k d_k + e_k, \qquad \|e_k\| \le \xi_k \le 1, \qquad 0 \le \alpha_k \le 1. \tag{14}$$

Proof. The norm bound follows because U_k is a convex combination of unit directions. Substituting Equation (13) into the group decomposition gives

$$e_k = (1 - \rho_k)(c_k - \chi_k d_k) + \sum_{G \in \mathcal{G}_{\tau_k}^{\mathrm{adv}}} \pi_G^{\tau_k} v_G^{\tau_k}.$$

The norm of the last sum is at most ρ_k, proving the error bound. The bounds on α_k and ξ_k follow from $\rho_k, \chi_k, \delta_k \in [0, 1]$. □

From local gradients to client directions. Choose one benign representative i_G from each $G \in \mathcal{G}_t^{\mathrm{ben}}$. For its local SGD steps, let $\eta_{i,\ell}^t \ge 0$ be the step sizes and $a_{i,\ell}^t$ the gradients actually applied, including clipping. If $\Gamma_i^t = \sum_\ell \eta_{i,\ell}^t > 0$, its effective local gradient is

$$h_i^t = -\frac{\Delta_i^t}{\Gamma_i^t} = \frac{1}{\Gamma_i^t} \sum_\ell \eta_{i,\ell}^t a_{i,\ell}^t, \qquad v_G^t = -\frac{h_{i_G}^t}{\|h_{i_G}^t\|}. \tag{15}$$

To relate these directions to the population gradient, for nonzero vectors g, h the triangle inequality gives

$$\|g\| \left\| \frac{g}{\|g\|} - \frac{h}{\|h\|} \right\| \le \|g - h\| + \big|\|h\| - \|g\|\big| \le 2\|g - h\|. \tag{16}$$

The same bound holds when a zero vector is assigned direction zero. For $\rho_k < 1$, define the direction-weighted local error

$$\sigma_k = \sum_{G \in \mathcal{G}^{\mathrm{ben}}_{\tau_k}} \frac{\pi_G^{\tau_k}}{1-\rho_k} \|h_{i_G}^{\tau_k} - g_k\|. \tag{17}$$

Set $\sigma_k = 0$ if $\rho_k = 1$. Projection onto the descent ray and Equation (16) imply

$$\|g_k\|\delta_k \le \|g_k\|\|c_k - d_k\| \le 2\sigma_k, \qquad \rho_k < 1. \tag{18}$$

Sources of local training error. The error in Equation (17) can be bounded using client heterogeneity, local drift and gradient estimation error. Let $y_{i,0}^t = x$ and $y_{i,\ell+1}^t = y_{i,\ell}^t - \eta_{i,\ell}^t a_{i,\ell}^t$ be a benign client's local trajectory from the current global model. Suppose F_i has an L_i-Lipschitz gradient along this trajectory and $\|a_{i,\ell}^t\| \le C$, as ensured by gradient clipping. Define

$$\begin{aligned} H_i^t &= \|\nabla F_i(x) - \nabla F(x)\|, \\ Z_i^t &= \frac{1}{\Gamma_i^t} \sum_\ell \eta_{i,\ell}^t \|a_{i,\ell}^t - \nabla F_i(y_{i,\ell}^t)\|, \\ E_i^t &= H_i^t + Z_i^t + \frac{L_i C}{2\Gamma_i^t} \left((\Gamma_i^t)^2 - \sum_\ell (\eta_{i,\ell}^t)^2 \right). \end{aligned} \tag{19}$$

Here H_i^t measures the difference between the client's objective and the population objective. The quantity Z_i^t includes stochastic gradient error and clipping bias, while the last term bounds the drift during local training.

Lemma 3 (Local training error bound)**.** *Under these local-training conditions,* $\|h_i^t - \nabla F(x)\| \le E_i^t$. *Consequently,* σ_k *is at most the benign direction-weighted mean of* $E_{i_G}^{\tau_k}$ *in Equation* (17)*.*

Proof. Add and subtract $\nabla F_i(y_{i,\ell}^t)$ and $\nabla F_i(x)$ inside the weighted sum defining h_i^t. The triangle inequality and local smoothness yield

$$\|h_i^t - \nabla F(x)\| \le Z_i^t + H_i^t + \frac{L_i}{\Gamma_i^t} \sum_\ell \eta_{i,\ell}^t \|y_{i,\ell}^t - x\|.$$

The clipped local steps give $\|y_{i,\ell}^t - x\| \le C \sum_{r<\ell} \eta_{i,r}^t$. Finally,

$$\sum_\ell \eta_{i,\ell}^t \sum_{r<\ell} \eta_{i,r}^t = \frac{1}{2} \left((\Gamma_i^t)^2 - \sum_\ell (\eta_{i,\ell}^t)^2 \right),$$

which gives E_i^t. Taking the weighted mean proves the claim for σ_k. □

Thus, the local error has explicit sources in the training protocol. When $g_k \neq 0$, these sources bound the directional residual by $\delta_k \le 2\sigma_k / \|g_k\|$. The actual residual can be smaller because it measures the combined directions rather than adding their individual discrepancies. This distinction preserves the role of useful differences among benign clients in the optimization analysis.

H.4 Optimization progress with server Adam

We analyze the bias-corrected Adam update used in Appendix A.2. With zero-initialized moments,

$$\begin{aligned} \widehat{\mathbf{m}}_k &= \sum_{j=1}^k \omega_{k,j} U_j, & \omega_{k,j} &= \frac{(1-\beta_1)\beta_1^{k-j}}{1-\beta_1^k}, \\ \widehat{\mathbf{v}}_k &= \sum_{j=1}^k \upsilon_{k,j} (U_j \odot U_j), & \upsilon_{k,j} &= \frac{(1-\beta_2)\beta_2^{k-j}}{1-\beta_2^k}, \\ x_k &= x_{k-1} + \eta_k D_k \widehat{\mathbf{m}}_k, & D_k &= \operatorname{diag}\big((\sqrt{\widehat{\mathbf{v}}_k} + \varepsilon)^{-1}\big). \end{aligned} \tag{20}$$

We use $0 < \beta_1 < 1$, $\beta_1^2 < \beta_2 < 1$ and $\varepsilon > 0$, as satisfied by the experimental values $(\beta_1, \beta_2) = (0.8, 0.9)$. The coefficients $\omega_{k,j}$ sum to one and describe server momentum. Client histories determine each U_j before it enters these moment estimates.

Step size and directional scaling. Let $R_k = \|D_k \widehat{\mathbf{m}}_k\|$, so the server step has length $\eta_k R_k$. To measure Adam's action on the objective gradient, define, for $g_k \neq 0$,

$$\nu_k = \frac{\|D_k g_k\|}{\|g_k\|}, \qquad \gamma_k = \frac{\langle g_k, D_k g_k \rangle}{\|g_k\| \|D_k g_k\|}. \tag{21}$$

Here $\nu_k > 0$ is the scaling along the gradient and $\gamma_k \in (0, 1]$ is the cosine between the gradient and its coordinate-scaled version. If $g_k = 0$, take $\gamma_k = 1$ and ν_k equal to the smallest diagonal entry of D_k. The following bounds retain the relationship between Adam's first and second moments.

Lemma 4 (Adam step and angular bounds)**.** *Define*

$$K_k = \frac{(1-\beta_1)^2(1-\beta_2^k)}{(1-\beta_1^k)^2(1-\beta_2)} \frac{1-(\beta_1^2/\beta_2)^k}{1-\beta_1^2/\beta_2}. \tag{22}$$

Then $R_k^2 \leq dK_k$, and K_k is uniformly bounded in k. If $d_{\min,k}$ and $d_{\max,k}$ are the smallest and largest diagonal entries of D_k, then

$$\gamma_k \geq \frac{2\sqrt{d_{\min,k} d_{\max,k}}}{d_{\min,k} + d_{\max,k}}. \tag{23}$$

Proof. For each coordinate r, weighted Cauchy–Schwarz gives

$$(\widehat{\mathbf{m}}_{k,r})^2 \leq \left(\sum_{j=1}^{k} \frac{\omega_{k,j}^2}{\upsilon_{k,j}} \right) \left(\sum_{j=1}^{k} \upsilon_{k,j} U_{j,r}^2 \right) = K_k \widehat{\mathbf{v}}_{k,r}.$$

Summing the geometric series gives Equation (22). Its factors are uniformly bounded because $1 - \beta_1^k \geq 1 - \beta_1 > 0$ and $\beta_1^2/\beta_2 < 1$. Dividing each coordinate inequality by $(\sqrt{\widehat{\mathbf{v}}_{k,r}} + \varepsilon)^2$ and summing proves $R_k^2 \leq dK_k$.

For the angular bound, let $z = g_k/\|g_k\|$, $a = d_{\min,k}$, $b = d_{\max,k}$ and $q = z^T D_k z$. Each diagonal entry d satisfies $d^2 \leq (a+b)d - ab$, so

$$\gamma_k^2 = \frac{q^2}{z^T D_k^2 z} \geq \frac{q^2}{(a+b)q - ab} \geq \frac{4ab}{(a+b)^2}.$$

The last inequality follows by expanding $((a+b)q - 2ab)^2 \geq 0$. The zero-gradient convention also satisfies the bound. □

The step bound is independent of ε^{-1} because the moments are generated by the same aggregate sequence. The angular bound describes the directional distortion due to coordinate scaling. We use the actual R_k, ν_k and γ_k in the progress inequality below.

Descent signal, aggregation error and momentum. Average the signal and error coefficients from Lemma 2 using Adam's momentum weights:

$$\widetilde{\alpha}_k = \sum_{j=1}^{k} \omega_{k,j} \alpha_j, \qquad \widetilde{\xi}_k = \sum_{j=1}^{k} \omega_{k,j} \xi_j. \tag{24}$$

Past directions were computed at earlier global models. To account for this change, let

$$A_k = \sum_{j=1}^{k} \omega_{k,j} \sum_{\ell=j}^{k-1} \eta_\ell R_\ell, \qquad S_T = \sum_{k=1}^{T} \eta_k \nu_k,$$
$$\mathcal{M}_T = \sum_{k=1}^{T} \eta_k \nu_k A_k, \qquad \mathcal{V}_T = \sum_{k=1}^{T} \eta_k^2 R_k^2. \tag{25}$$

The quantity A_k bounds the momentum-weighted distance traveled since those directions were computed. The sums $\mathcal{M}_T$ and $\mathcal{V}_T$ capture this lag and the squared server-step lengths, respectively.

Proposition 4 (Optimization progress under directional error)**.** *Under the objective assumptions in Appendix H.1, HGDA's applied server updates satisfy*

$$\sum_{k=1}^{T} \eta_k \nu_k (\gamma_k \widetilde{\alpha}_k - \widetilde{\xi}_k) \|g_k\| \le F(x_0) - F_{\inf} + 2L\mathcal{M}_T + \frac{L}{2}\mathcal{V}_T. \tag{26}$$

Proof. Decompose the bias-corrected first moment into the current descent signal, the change in past descent directions and the aggregation errors:

$$\widehat{\mathbf{m}}_k = \widetilde{\alpha}_k d_k + \sum_{j=1}^{k} \omega_{k,j} \alpha_j (d_j - d_k) + \sum_{j=1}^{k} \omega_{k,j} e_j.$$

After coordinate scaling, the first term has inner product $-\nu_k \gamma_k \widetilde{\alpha}_k \|g_k\|$ with g_k. The error term has inner product at most $\nu_k \widetilde{\xi}_k \|g_k\|$. For the middle term, Equation (16) gives $\|g_k\|\|d_j - d_k\| \le 2\|g_j - g_k\|$. Using $\alpha_j \le 1$ and smoothness,

$$\begin{aligned}
\left\langle g_k, D_k \sum_{j=1}^{k} \omega_{k,j} \alpha_j (d_j - d_k) \right\rangle &\le 2\nu_k \sum_{j=1}^{k} \omega_{k,j} \|g_j - g_k\| \\
&\le 2L\nu_k \sum_{j=1}^{k} \omega_{k,j} \|x_{j-1} - x_{k-1}\| \\
&\le 2L\nu_k A_k.
\end{aligned}$$

The smoothness inequality for the server step therefore yields

$$\begin{aligned}
F(x_k) \le F(x_{k-1}) &- \eta_k \nu_k (\gamma_k \widetilde{\alpha}_k - \widetilde{\xi}_k) \|g_k\| \\
&+ 2L\eta_k \nu_k A_k + \frac{L}{2} \eta_k^2 R_k^2.
\end{aligned} \tag{27}$$

Summing over the applied updates and using $F(x_T) \ge F_{\inf}$ proves the result. □

H.5 Stationarity from window evidence and local training

We now insert the evidence bound from Lemma 1 into the progress inequality. Because Adam retains past aggregates, the relevant adversarial weight is averaged over its momentum history:

$$r_* = \max_{1 \le k \le T} \sum_{j=1}^{k} \omega_{k,j} \overline{\rho}_{\tau_j}. \tag{28}$$

Let $\gamma_* > 0$ be a lower bound on γ_k over these updates, obtained directly from Equation (21) or from the optimizer-state bound in Equation (23). Define

$$\kappa = \gamma_* - (1 + \gamma_*) r_*. \tag{29}$$

The condition $\kappa > 0$ compares the residual adversarial score weight with the directional distortion introduced by Adam. It expresses the sufficient balance through window evidence assigned to each type of direction group and the optimizer's state.

Proposition 5 (Stationarity with signed-history control)**.** *Under the assumptions of Proposition 4, suppose $\kappa > 0$ and $\sigma_k \le \sigma$ for all applied updates. Then*

$$\min_{1 \le k \le T} \|\nabla F(x_{k-1})\| \le \frac{F(x_0) - F_{\inf} + 2L\mathcal{M}_T + (L/2)\mathcal{V}_T}{\kappa S_T} + \frac{2(1 + \gamma_*)(1 - r_*)}{\kappa} \sigma. \tag{30}$$

The local-error bound σ can be supplied by Lemma 3.

Proof. Let $r_T = \min_{1 \le k \le T} \|g_k\|$ and

$$R_\sigma = \frac{2(1 + \gamma_*)(1 - r_*)}{\kappa} \sigma.$$

If $r_T \le R_\sigma$, the conclusion follows. Otherwise, since $(1+\gamma_*)(1-r_*) = 1+\kappa$, we have $r_T > R_\sigma \ge 2\sigma$. Set $v = 2\sigma/r_T < 1$. Equation (18) gives $\|c_j - d_j\| \le v$ whenever $\rho_j < 1$, and hence $\chi_j \ge 1-v$ and $\delta_j \le v$. Thus

$$\alpha_j \ge (1-\rho_j)(1-v), \qquad \xi_j \le (1-\rho_j)v + \rho_j.$$

These two inequalities also hold when $\rho_j = 1$. Write $\widetilde{\rho}_k = \sum_j \omega_{k,j}\rho_j$. Lemma 1 and Equation (28) imply $\widetilde{\rho}_k \le r_*$. Averaging the preceding inequalities and using $\gamma_k \ge \gamma_*$ gives

$$\begin{aligned}\gamma_k\widetilde{\alpha}_k - \widetilde{\xi}_k &\ge \gamma_*(1-\widetilde{\rho}_k)(1-v) - (1-\widetilde{\rho}_k)v - \widetilde{\rho}_k \\ &\ge \kappa - (1+\gamma_*)(1-r_*)v = \kappa(1 - R_\sigma/r_T) > 0.\end{aligned}$$

The second inequality uses $v < 1$, which makes the preceding expression decreasing in $\widetilde{\rho}_k$. Inserting this positive lower bound into Proposition 4 yields

$$\kappa(r_T - R_\sigma)S_T \le F(x_0) - F_{\inf} + 2L\mathcal{M}_T + \frac{L}{2}\mathcal{V}_T,$$

and rearranging proves Equation (30). □

Learning-rate dependence. To make the optimization term explicit, suppose $0 < \nu_{\min} \le \nu_k \le \nu_{\max}$ and $R_k \le R_{\max}$. Such uniform bounds exist for fixed Adam parameters: $\|U_k\| \le 1$ bounds the diagonal scaling for $\varepsilon > 0$, and Lemma 4 bounds R_k. For $\ell < k$, the total momentum weight on updates through ℓ satisfies

$$\sum_{j=1}^{\ell} \omega_{k,j} = \beta_1^{k-\ell}\frac{1-\beta_1^{\ell}}{1-\beta_1^{k}} \le \beta_1^{k-\ell}.$$

Reversing the sums in A_k and using $2\eta_k\eta_\ell \le \eta_k^2 + \eta_\ell^2$ gives

$$\begin{aligned}\mathcal{M}_T &\le \nu_{\max}R_{\max}\sum_{1\le\ell<k\le T}\beta_1^{k-\ell}\eta_k\eta_\ell \\ &\le \nu_{\max}R_{\max}\frac{\beta_1}{1-\beta_1}\sum_{k=1}^{T}\eta_k^2.\end{aligned}$$

Also, $\mathcal{V}_T \le R_{\max}^2\sum_k \eta_k^2$ and $S_T \ge \nu_{\min}\sum_k \eta_k$. Define

$$C_{\mathrm{Adam}} = 2\nu_{\max}R_{\max}\frac{\beta_1}{1-\beta_1} + \frac{R_{\max}^2}{2}. \tag{31}$$

Equation (30) then gives

$$\min_{1\le k\le T}\|\nabla F(x_{k-1})\| \le \frac{F(x_0) - F_{\inf} + LC_{\mathrm{Adam}}\sum_{k=1}^{T}\eta_k^2}{\kappa\nu_{\min}\sum_{k=1}^{T}\eta_k} + R_\sigma. \tag{32}$$

For a budget of T applied updates and $\eta_k = \eta_0/\sqrt{T}$, this becomes

$$\min_{1\le k\le T}\|\nabla F(x_{k-1})\| \le \frac{(F(x_0) - F_{\inf})/\eta_0 + LC_{\mathrm{Adam}}\eta_0}{\kappa\nu_{\min}\sqrt{T}} + R_\sigma. \tag{33}$$

If κ is bounded away from zero and the local-error bound is uniform as T grows, the optimization term decreases as $O(T^{-1/2})$, leaving a stationarity neighborhood proportional to the local-training error. More generally, under the same uniform conditions, this term vanishes when $\sum_{k=1}^{T}\eta_k \to \infty$ and $\sum_{k=1}^{T}\eta_k^2/\sum_{k=1}^{T}\eta_k \to 0$. If $\sigma = 0$, the smallest gradient norm along the applied iterates tends to zero. A constant learning rate instead gives

$$\min_{1\le k\le T}\|\nabla F(x_{k-1})\| \le \frac{F(x_0) - F_{\inf}}{\kappa\nu_{\min}\eta T} + \frac{LC_{\mathrm{Adam}}\eta}{\kappa\nu_{\min}} + R_\sigma, \tag{34}$$

which separates the finite training budget from the learning-rate and local-error contributions.

Connection to HGDA's design. The bound follows the same division of responsibilities as the method. Signed evidence under both references controls the weight available to each client, and groupwise maxima transfer that evidence to distinct current directions. The combined benign directions determine the useful training signal, with heterogeneity, gradient error and local drift entering through Lemma 3. Adam then determines the parameter step through its own moment estimates. Throughout the analysis, the objective is the benign population loss in Equation (6). The trusted data influence the assessment and weighting of client information rather than replacing that objective with the server's local loss.