FUSE: FAIRNESS-AWARE UPDATE SCREENING AGAINST BYZANTINE EXPLOITATION IN q-FAIR FEDERATED LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Client-level fair federated learning, in the form of q-fair federated learning (q-FFL), agnostic federated learning, or tilted empirical risk minimization, has emerged as an attractive solution to equalize utility across heterogeneous clients. This paradigm grants high-loss clients greater relative influence, while a central server aggregates their loss-weighted updates into a single shared model. Yet, the Byzantine robustness of q-fair federated learning has been largely neglected, leaving the server unable to tell legitimate fairness-induced amplification from malicious update scaling. To address this critical gap, we formalize the problem of Byzantine-robust q-fair federated learning under an *update-only interface*, in which the server observes nothing but the fairness-scaled client updates. Our analysis under shared class-conditional label skew shows that fairness scaling preserves the nonnegative class-mixture structure of honest updates, while replacing their unit-sum constraint by a client-dependent radial scale. Moving beyond magnitude-based screening, we show that this warped geometry supplies two complementary tests, one radial and one structural, that a single norm cannot provide. We then introduce FUSE, a structure-aware robust aggregator that uses a small trusted labeled dataset for coefficient recovery, scale and cone-consistency screening, and post-screening step-size control. Across four benchmarks and seven Byzantine attacks, FUSE improves seed-wise worst-attack accuracy over the strongest baseline on each benchmark by 3.9–30.1 percentage points, while largely preserving the influence of the high-loss honest clients that fairness is meant to protect.

1 INTRODUCTION

Client-level fair federated learning aims to improve the utility of disadvantaged clients through objectives such as q-fair federated learning (q-FFL), agnostic federated learning, and tilted empirical risk minimization (Li et al., 2020; Mohri et al., 2019; Li et al., 2021). Recent multi-objective, proportional, and coalitional formulations extend this principle to broader notions of client welfare (Hu et al., 2022; Zhang et al., 2023; Chaudhury et al., 2022; Pan et al., 2024). A common implementation is to incorporate the loss-dependent weight locally: q-FedAvg forms loss-weighted updates at the client, PropFair transforms the client-side objective, and H-nobs directly transmits the fairness-aware gradient $\nabla H_i = F_i^q \nabla F_i$ (Li et al., 2020; Zhang et al., 2023; Zhou et al., 2023). This motivates an *update-only* interface in which the server observes the composite fairness-scaled update but not its loss and descent direction separately; avoiding an additional loss report also removes a data-dependent communication channel that can incur extra privacy cost (Ling et al., 2026; Zhu et al., 2019).

Robustness under this interface, however, remains poorly understood: q-FedAvg and PropFair do not model Byzantine clients, FedFDP focuses on fairness and privacy, while H-nobs and its extension primarily screen fairness-aware updates through their magnitudes (Li et al., 2020; Zhang et al., 2023; Ling et al., 2026; Zhou et al., 2023; Basharat et al., 2025). Such magnitude-based screening is intrinsically ambiguous in q-FFL, because large updates may arise either from legitimate amplification of high-loss clients or from adversarial scaling (Zhou et al., 2023; Basharat et al., 2025). Meanwhile, structure-aware robust aggregation exploits class-mixture geometry, but existing simplex-based formulations do not account for the radial distortion introduced by fairness scaling (Bao et al., 2024). We therefore study Byzantine-robust q-FFL under the update-only interface and

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.

ask what geometric structure remains observable after the loss weight and descent direction are fused into a single message (Zhou et al., 2023; Bao et al., 2024).

1.1 Main Contributions

Problem formulation and fairness-warped geometry. We formalize Byzantine-robust q-fair federated learning under an update-only interface, with up to $f < m/2$ of the m submissions arbitrarily corrupted at each round and a server that observes only the fairness-scaled updates. Under shared class-conditional label skew, we show that an honest submission stays a nonnegative combination of class-conditioned reference directions, but that its coefficients no longer sum to one: fairness scaling replaces the unit-sum simplex of prior structural methods (Bao et al., 2024) by a nonnegative cone whose coefficient mass is exactly the fairness scale F_i^q. A magnitude test alone cannot separate that legitimate radial growth from adversarial scaling.

Structure-aware robust aggregation. Building on this geometry, we propose FUSE, which recovers each submission's latent coefficients from a small trusted labeled set by nonnegative least squares. The coefficients supply two complementary statistics: their mass measures radial scale, and the residual outside the trusted cone measures structural incompatibility. FUSE spends its rejection budget on the two in sequence, and, since the normalized coefficients also reconstruct the client loss, rescales the retained mean to restore the step size that fairness weighting would inflate.

Empirical validation. We evaluate FUSE on four benchmarks (CIFAR-10, MNIST, EMNIST-Balanced, and AGNews) covering both vision and language modalities, against seven Byzantine attacks and seven baselines, over five registered seeds. Our approach improves seed-wise worst-attack accuracy over the strongest baseline on each benchmark by 3.9–30.1 percentage points, while rejecting $0 - 9.7\%$ of the high-loss honest submissions that fairness is meant to protect. Controlled geometry experiments, a full 2^3 ablation, and an adaptive cone-preserving attack isolate the mechanism and its boundary.

2 Problem of Robust q-Fair Federated Learning

We consider a synchronous federated learning system comprising m clients and a global model $w_t \in \mathbb{R}^d$. Each client i holds a local dataset $\mathcal{D}_i$ and has empirical loss

$$F_i(w) = \frac{1}{|\mathcal{D}_i|} \sum_{\xi \in \mathcal{D}_i} \ell(w; \xi).$$

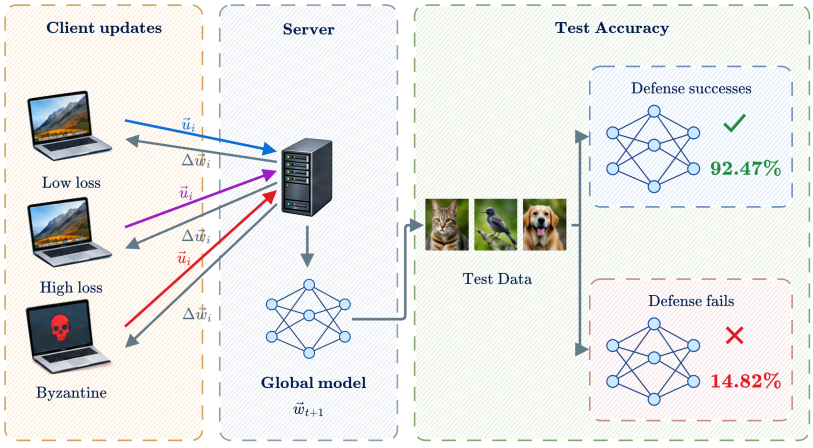


Figure 1: **Update-only Byzantine-robust q-fair federated learning.** Honest clients submit loss-scaled updates $u_i = F_i^q \Delta_i$, Byzantine clients may submit arbitrary vectors, and the server aggregates them into w_{t+1} with the help of a small trusted labeled dataset. Because q-fair learning legitimately enlarges the updates of high-loss honest clients, update magnitude alone cannot distinguish fairness-induced amplification from malicious scaling.

In this context, the goal of q-fair learning is to optimize one shared model while assigning greater relative influence to high-loss clients. Specifically, we seek a model minimizing

$$\mathcal{J}_q(w) = \frac{1}{m}\sum_{i=1}^{m} \frac{F_i(w)^{q+1}}{q+1}, \qquad q \geq 0.$$

Update-only q-fair FL. A typical realization of this objective weights each client descent direction by $F_i(w_t)^q$. Specifically, at round t, an honest client computes a one-step update and transmits only the resulting weighted message:

$$\Delta_i^t = -\eta \nabla F_i(w_t), \qquad u_i^t = F_i(w_t)^q \Delta_i^t. \tag{1}$$

The server therefore observes the product u_i^t, but not its loss and update factors separately. We refer to this setting as the *update-only interface*, and omit t when unambiguous.

Robust q-fair federated learning. We are interested in solving robust q-fair federated learning, wherein a fraction of client updates may be corrupted prior to aggregation. Specifically, Byzantine clients may coordinate and submit arbitrary vectors. Our goal is to design an aggregation rule that remains effective despite such corruptions while preserving the intended emphasis on high-loss clients.

Assumption 1 (Byzantine and observation model)**.** *At each round, $\mathcal{S}_t = \mathcal{H}_t \cup \mathcal{B}_t$ and $|\mathcal{B}_t| \leq f_t < |\mathcal{S}_t|/2$. Byzantine clients may coordinate and submit arbitrary vectors; the algorithm receives a declared rejection budget f_t. The server fixes a coordinate subset for screening. Screening uses only each submission restricted to these coordinates, the trusted reference, and quantities derived from them; it has no side information revealing client Byzantine status. Retained clients contribute their original full-model submissions to aggregation.*

3 From Simplex to Cone

Standard structure-aware aggregation under label skew models honest updates as class mixtures. We show that q-fair weighting preserves this mixture structure but removes its unit-sum constraint, changing the appropriate observation model from a simplex to a cone.

3.1 Fairness Scaling Breaks the Simplex

Magnitude ambiguity. For $F_i > 0$, $q > 0$, and any $\lambda > 0$,

$$(\lambda F_i)^q (\lambda^{-q}\Delta_i) = F_i^q \Delta_i.$$

The update-only message does not separately identify fairness scale and descent magnitude. Magnitude alone therefore cannot distinguish legitimate q-fair amplification from adversarial scaling.

Assumption 2 (Shared class-conditional label skew)**.** *For every honest client i,*

$$P_i = \sum_{c=1}^{C} p_{ic} Q_c, \qquad p_i \succeq 0, \qquad \mathbf{1}^\top p_i = 1,$$

where Q_c is the class-c conditional distribution shared by honest clients. At the broadcast model, the population class losses satisfy $0 < \ell_{\min} \leq \ell_c \leq \ell_{\max} < \infty$.

Here p_i contains client i's label proportions. The guarantees do not cover arbitrary violations of shared class-conditionals; covariate shifts are evaluated empirically in Appendix E.13.

Let $\ell = (\ell_1, \ldots, \ell_C)^\top$ and $G^\star = [g_1^\star, \ldots, g_C^\star]$ contain the population class losses and class-conditioned updates, respectively. Under Assumption 2, linearity of expectation gives

$$\begin{aligned} &\overline{F}_i = \ell^\top p_i, \qquad \overline{\Delta}_i = G^\star p_i, \\ &\overline{u}_i = (\ell^\top p_i)^q G^\star p_i = G^\star a_i^\star, \qquad a_i^\star := (\ell^\top p_i)^q p_i \succeq 0, \qquad \|a_i^\star\|_1 = (\ell^\top p_i)^q. \end{aligned} \tag{2}$$

Key consequence. Without fairness scaling, class-mixture coefficients satisfy $\mathbf{1}^\top p_i = 1$ and lie on a simplex. Fairness scaling preserves coefficient nonnegativity but changes their mass to F_i^q. Thus the honest observation model is no longer a unit-sum simplex but a radially varying nonnegative cone.

4 FUSE Robust Aggregator

In this section, we construct a robust aggregator from the cone structure identified above. To alleviate the ambiguity between legitimate scale and malicious deviation, FUSE combines trusted class references, coefficient recovery, two-stage screening, and post-screening step-size control, as illustrated in Figure 2.

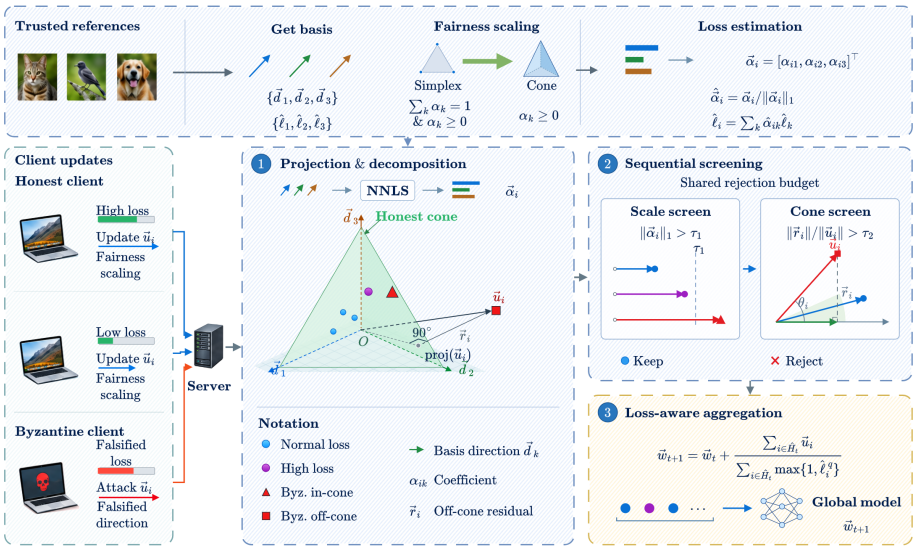


Figure 2: **FUSE overview.** Each fair update is decomposed over trusted class-conditioned directions by nonnegative least squares. The recovered mass $L_i = \|\widehat{a}_i\|_1$ tests for anomalous radial scaling; the normalized cone residual D_i^{cone} detects updates inconsistent with any nonnegative class mixture. FUSE removes scale outliers first, spends the remaining budget on the largest cone residuals, and reconstructs the client loss from the normalized coefficients for post-screening step-size control.

Trusted class reference. The server uses a small labeled set to construct class-conditioned reference directions $G = [g_1, \dots, g_C]$ and trusted class losses $\widehat{\ell}$ in the screening coordinates:

$$\begin{aligned} \widehat{\ell}_c &= \frac{1}{|\mathcal{D}_s^c|} \sum_{\xi \in \mathcal{D}_s^c} \ell(w_t; \xi), \\ g_c &= \left[-\eta \nabla_w \frac{1}{|\mathcal{D}_s^c|} \sum_{\xi \in \mathcal{D}_s^c} \ell(w_t; \xi) \right]_{\text{scr}}, \\ G &= [g_1, \dots, g_C], \qquad \widehat{\ell} = (\widehat{\ell}_1, \dots, \widehat{\ell}_C)^\top. \end{aligned} \tag{3}$$

Define the trusted cone $\mathcal{C}_G := \{Ga : a \succeq 0\}$.

Coefficient recovery and screening scores. Since honest submissions are approximately nonnegative combinations of the trusted directions, FUSE recovers their latent coefficients by nonnegative least squares. Specifically, for each submission it solves

$$\widehat{a}_i \in \arg\min_{a \succeq 0} \|v_i - Ga\|_2^2, \qquad L_i = \|\widehat{a}_i\|_1, \qquad D_i^{\text{cone}} = \frac{\|v_i - G\widehat{a}_i\|_2^2}{\max\{\|v_i\|_2^2, \epsilon\}}. \tag{4}$$

L_i measures radial scale, while D_i^{cone} measures structural incompatibility with the trusted nonnegative cone. Stage I removes up to f submissions with implausibly large coefficient mass $L_i > \kappa L_{\text{med}}$; Stage II spends the remaining rejection budget on the largest cone residuals, yielding the retained set $\widehat{\mathcal{H}}_t$.

Trusted loss reconstruction and aggregation. Finally, normalized coefficients estimate class composition and hence the client loss:

$$\widehat{p}_i = \frac{\widehat{a}_i}{\max\{L_i, \epsilon\}}, \qquad \widetilde{F}_i = \widehat{\ell}^\top \widehat{p}_i, \qquad w_{t+1} = w_t + \frac{\sum_{i \in \widehat{\mathcal{H}}_t} u_i}{\sum_{i \in \widehat{\mathcal{H}}_t} \max\{1, \widetilde{F}_i^q\}}. \tag{5}$$

Algorithm 1 FUSE at round t

Require: submissions $\{u_i\}_{i\in\mathcal{S}_t}$, trusted class data $\{\mathcal{D}_s^c\}_{c=1}^C$, budget f, q, κ, ϵ

1: Construct $(G, \widehat{\ell})$ using Eq. (3)
2: **for** each $i \in \mathcal{S}_t$ **do**
3: Recover $\widehat{a}_i$ by NNLS in Eq. (4)
4: Compute L_i and D_i^{cone} using Eq. (4)
5: **end for**
6: Reject scale anomalies using $L_i > \kappa L_{\text{med}}$
7: Use the remaining budget to reject the largest cone scores
8: For each retained client, compute $\widehat{p}_i$ and $\widetilde{F}_i$ using Eq. (5)
9: Update w_{t+1} using Eq. (5)

Ensure: w_{t+1}

In the ideal model, $\widehat{p}_i = p_i$ and $\widetilde{F}_i = \ell^\top p_i$. The denominator only rescales the retained mean and therefore does not change its direction. Additional implementation details are provided in Appendix B.3.

5 Experiments

This section summarizes the key results of our experiments. In all that follows, $q = 1$ and the declared rejection budget is $f = 4$. We evaluate FUSE against fair, robust, and structure-aware baselines in Section 5.2, and audit whether its robustness preserves the high-loss clients that q-FFL emphasizes in Section 5.3. We isolate the geometric mechanism and conduct a component ablation in Section 6; further results are reported in the appendix.

5.1 Experimental Setup

Datasets and models. We evaluate CIFAR-10/ResNet-18 (Krizhevsky, 2009; He et al., 2016), MNIST/MLP (LeCun et al., 1998), EMNIST-Balanced/CNN (Cohen et al., 2017), and AG-News/GRU (Zhang et al., 2015; Cho et al., 2014). Data are partitioned by Dirichlet label skew with $\alpha = 1$, and each round samples 10 of 100 honest clients. FUSE uses 20 trusted labeled examples per class. Screening uses the classifier block on the three vision benchmarks and all model parameters on AGNews; retained clients always contribute their original full-model submissions.

Threat model and baselines. We consider Sign Flip, IPM, Gaussian, ALIE, Min-Max, Min-Sum, and Magnitude Scaling (Xie et al., 2020; Baruch et al., 2019; Shejwalkar & Houmansadr, 2021). Four Byzantine Sybils are appended in attacked rounds. We compare with q-FFL (Li et al., 2020), H-nobs (Zhou et al., 2023), Trimmed Mean (Yin et al., 2018), Krum (Blanchard et al., 2017), FLTrust (Cao et al., 2021), BOBA (Bao et al., 2024), and FedAA (He et al., 2025) under the same fairness-scaled message whenever applicable.

Evaluation. We conduct each experiment with five registered seeds (40–44), and report global and worst-10% client accuracy (W10) averaged over the final ten rounds. Our primary statistic is the seed-wise minimum across the seven attacks, *i.e.*the accuracy an adversary can force by picking the best attack for that seed. Runs triggering the registered collapse criterion are not replaced by pre-collapse maxima, and the number of completed seeds is reported as $^{(k)}$. We include additional details in Appendix B.

5.2 Robustness under Byzantine Attacks

In Table 1, we observe that FUSE consistently attains the highest seed-wise worst-attack global and W10 accuracy on all four benchmarks, improving worst-attack global accuracy over the best baseline by 30.1, 3.9, 17.1, and 13.4 percentage points on CIFAR-10, MNIST, EMNIST-Balanced, and AGNews respectively. Beyond worst-case performance, it attains the highest accuracy in 13 of the 28 dataset–attack pairs; across the remaining 15 its largest deficit to the best method is 3.5 percentage points, under Sign Flip on AGNews. Relative to FLTrust, the strongest baseline overall, the paired global difference is positive on all five seeds of each vision benchmark, with bootstrap 95%

Table 1: **Robustness across datasets and attacks.** Mean global accuracy (%) over the seeds reaching the registered endpoint; $^{(k)}$ denotes k completed seeds. The last column is the seed-wise worst attack over the seven attacks, reported as global (G) / worst-10% client (W10) accuracy. Per-attack W10 results and across-seed standard deviations are in Appendix C.1.

Dataset/model	Method	Clean	Sign	IPM	Gauss.	ALIE	Min-Max	Min-Sum	Mag.	Worst attack (G / W10)
CIFAR-10 ResNet-18	q-FFL	76.96	76.47	–	13.81$^{(3)}$	63.43	14.76$^{(4)}$	14.05	43.45$^{(4)}$	–
	H-nobs	76.96	12.19$^{(3)}$	57.13	79.28	35.65	42.86	20.15	76.96	12.19 / 0.35$^{(3)}$
	Trim.	70.56	59.66	44.84	79.45	39.27	13.37	14.06	74.35	13.18 / 1.28
	Krum	63.66	44.69	62.78	60.21	22.24	61.58	45.62	63.56	21.67 / 9.78
	FLTrust	81.31	**81.29**	**81.31**	81.25	55.13	48.84	43.81	79.44	43.81 / 30.53
	BOBA (cls.)	81.93	72.68	10.93	81.58	72.00	17.31	17.83	73.64	10.93 / 0.66
	FedAA	75.07	45.32	74.07$^{(4)}$	72.60	61.44	22.54	43.07	66.07	22.15 / 10.27$^{(4)}$
	FUSE	**82.25**	79.29	80.59	**82.30**	**77.74**	**82.17**	**82.19**	**81.12**	**73.88 / 63.58**
MNIST MLP	q-FFL	94.90	93.24	–	62.99$^{(3)}$	94.29	48.54$^{(3)}$	56.03$^{(3)}$	68.76$^{(3)}$	–
	H-nobs	94.90	14.82$^{(3)}$	91.09	94.90	93.45	**94.80**	65.54$^{(3)}$	94.90	14.82 / 1.91$^{(3)}$
	Trim.	94.90	85.77	87.30$^{(4)}$	94.84	65.65$^{(3)}$	48.61$^{(3)}$	48.74$^{(3)}$	95.33	48.24 / 37.25$^{(3)}$
	Krum	93.99	90.45	**93.42**	93.99	63.99$^{(3)}$	94.00	93.76	93.99	63.99 / 56.14$^{(3)}$
	FLTrust	94.96	93.30	93.34	94.98	93.10	89.97	91.77	95.07	88.52 / 79.64
	BOBA (cls.)	96.41	**94.45**	11.23	**96.41**	95.70	80.55	82.26	**96.60**	11.23 / 0.16
	FedAA	**96.42**	51.53	85.86	96.29	**96.30**	74.04	60.57	96.22	43.00 / 31.26
	FUSE	93.83	93.15	92.42	93.83	93.81	93.84	**93.84**	93.83	**92.42 / 87.07**
EMNIST-Bal. CNN	q-FFL	63.79	74.93	–	2.13$^{(1)}$	53.66$^{(3)}$	44.08$^{(3)}$	27.01$^{(3)}$	65.81	–
	H-nobs	63.79	28.70$^{(2)}$	78.66	63.79	78.55$^{(3)}$	52.75$^{(3)}$	50.65$^{(3)}$	63.79	28.70 / 21.64$^{(2)}$
	Trim.	80.30	63.17	54.93$^{(4)}$	63.28	66.82$^{(4)}$	25.45$^{(3)}$	34.35$^{(3)}$	64.13	37.11 / 26.83$^{(2)}$
	Krum	78.49	7.35$^{(2)}$	78.64	78.49	71.34$^{(3)}$	47.92$^{(3)}$	4.12$^{(3)}$	78.49	5.12 / 0.97$^{(2)}$
	FLTrust	71.34	71.02	71.46	72.17	75.48	63.45	63.88	68.45	61.21 / 50.22
	BOBA (cls.)	**80.34**	73.19	–	**80.45**	78.91	72.64	75.60	**79.59**	–
	FedAA	78.15	47.32$^{(4)}$	71.70$^{(2)}$	78.26	71.43	75.34	60.20	78.39	30.43 / 23.53$^{(2)}$
	FUSE	79.12	**79.51**	**79.41**	79.67	**79.61**	**78.89**	**79.11**	79.12	**78.31 / 70.00**
AGNews GRU	q-FFL	81.52	64.00	–	27.06$^{(4)}$	73.23	40.49	45.40$^{(4)}$	76.35	–
	H-nobs	81.52	26.69$^{(3)}$	81.62	81.52	77.53	72.91	73.71	81.52	26.69 / 0.56$^{(3)}$
	Trim.	81.57	36.61	35.88	79.67	74.76	38.63	39.37	79.89	30.85 / 5.71
	Krum	72.69	77.17	76.62	72.69	72.18	77.42	74.77	72.69	64.43 / 48.20
	FLTrust	**83.92**	**83.92**	**83.92**	**83.98**	77.19	66.82	64.26	83.03	64.26 / 45.46
	BOBA (cls.)	83.82	60.57	27.65	83.82	**80.82**	56.44	59.43	**83.53**	27.65 / 1.65
	FedAA	78.36	38.32	57.68$^{(2)}$	77.83	75.29	41.18	40.34	79.67	31.29 / 4.44$^{(2)}$
	FUSE	82.29	80.46$^{(4)}$	81.12	82.17	78.62	**81.44**	**81.36**	82.29	**77.84 / 63.75**$^{(4)}$

intervals $[20.2, 37.8]$, $[1.6, 6.3]$ and $[12.3, 21.9]$. The advantage is particularly pronounced against the invalid-mixture attacks: on CIFAR-10 the baselines drop to 13–62% under Min-Max and Min-Sum while FUSE stays above 82%. The worst-attack column makes the general pattern visible: every baseline has at least one attack for which its accuracy floor falls well below its own clean performance. This robustness does not uniformly improve clean utility: FUSE trails the best clean baseline by 2.6, 1.2, and 1.6 percentage points on MNIST, EMNIST-Balanced, and AGNews, and leads it by 0.3 on CIFAR-10. We regard this as the price of a substantially higher adversarial floor.

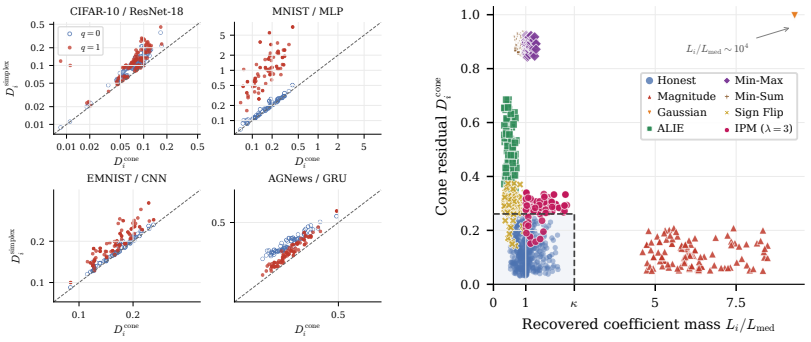
5.3 Preserving High-Loss Honest Clients

A robust aggregator should not obtain its robustness by discarding the high-loss honest clients that q-FFL is designed to emphasize, nor buy a clean fairness record by screening nothing at all. We therefore audit both sides of the screening decision under matched seeds, attacks, rejection budget $f = 4$, evaluation grid and round window. We focus on the two closest routes: H-nobs, which filters fair updates by norm, and BOBA, which fits a unit-sum class-mixture model on classifier coordinates.

In Table 2, we observe that the two baselines fail on opposite sides. Norm screening pays a steep fairness price: H-nobs rejects $22.2 - 35.7\%$ of the highest-loss honest submissions, and up to 91.7% of them under its worst attack, in exchange for a $61.1 - 68.6\%$ Byzantine rejection rate. BOBA barely touches the fair tail ($0.0 - 5.6\%$), but only because it rejects almost nothing: it removes $14.3 - 36.8\%$ of the Byzantine submissions, and nothing at all under Sign Flip and ALIE. FUSE is the only method that is conservative on both counts, rejecting $0.0 - 9.7\%$ of the high-loss honest submissions while removing $85.7 - 99.8\%$ of the Byzantine ones. Its robustness therefore does not arise from suppressing the q-fair tail. The one exception is ALIE on AGNews, which FUSE does not separate within this window and which produces the 0.00 worst-case entry; it is also the mildest attack in the suite, costing 3.7 percentage points of accuracy in Table 1. The complete per-attack audit is in Appendix E.

Table 2: **Whom each screen actually rejects, five seeds.** Left: the fraction (%) of the highest-loss 20% of honest submissions rejected, which fairness wants low. Right: the fraction (%) of Byzantine submissions rejected, which robustness wants high. Both pool the first 300 rounds and the seven attacks within a seed, reported as mean$_{\pm\text{std}}$ over seeds 40–44; "worst" is the attack at which that column is worst in each seed. Best in bold, runner-up underlined.

		High-loss honest rejected ↓		Byzantine rejected ↑	
Dataset/model	Method	Pooled	Worst	Pooled	Worst
CIFAR-10/ResNet-18	H-nobs	$34.95_{\pm 1.62}$	$90.67_{\pm 2.53}$	$\underline{63.31}_{\pm 1.09}$	$\underline{0.00}_{\pm 0.00}$
	BOBA (cls.)	$\underline{0.33}_{\pm 0.27}$	$\underline{1.67}_{\pm 1.18}$	$27.12_{\pm 0.79}$	$\underline{0.00}_{\pm 0.00}$
	FUSE	$\mathbf{0.19}_{\pm 0.20}$	$\mathbf{1.00}_{\pm 0.91}$	$\mathbf{98.71}_{\pm 0.70}$	$\mathbf{94.83}_{\pm 2.16}$
MNIST/MLP	H-nobs	$35.71_{\pm 1.40}$	$91.67_{\pm 3.33}$	$\underline{61.07}_{\pm 1.43}$	$\underline{0.00}_{\pm 0.00}$
	BOBA (cls.)	$\underline{5.62}_{\pm 0.87}$	$\underline{21.67}_{\pm 2.04}$	$36.81_{\pm 2.40}$	$\underline{0.00}_{\pm 0.00}$
	FUSE	$\mathbf{1.67}_{\pm 1.04}$	$\mathbf{5.67}_{\pm 2.53}$	$\mathbf{96.05}_{\pm 1.56}$	$\mathbf{85.83}_{\pm 5.14}$
EMNIST-Balanced/CNN	H-nobs	$\underline{23.95}_{\pm 1.06}$	$\underline{70.15}_{\pm 7.32}$	$\underline{66.13}_{\pm 0.12}$	$\underline{0.00}_{\pm 0.00}$
	BOBA (cls.)	$\mathbf{0.00}_{\pm 0.00}$	$\mathbf{0.00}_{\pm 0.00}$	$14.31_{\pm 0.04}$	$\underline{0.00}_{\pm 0.00}$
	FUSE	$\mathbf{0.00}_{\pm 0.00}$	$\mathbf{0.00}_{\pm 0.00}$	$\mathbf{99.81}_{\pm 0.25}$	$\mathbf{98.67}_{\pm 1.73}$
AGNews/GRU	H-nobs	$22.19_{\pm 3.58}$	$80.00_{\pm 11.79}$	$\underline{68.62}_{\pm 0.56}$	$0.00_{\pm 0.00}$
	BOBA (cls.)	$\mathbf{0.00}_{\pm 0.00}$	$\mathbf{0.00}_{\pm 0.00}$	$29.71_{\pm 0.64}$	$0.00_{\pm 0.00}$
	FUSE	$\underline{9.71}_{\pm 1.42}$	$\underline{68.00}_{\pm 9.96}$	$\mathbf{85.71}_{\pm 0.00}$	$0.00_{\pm 0.00}$



(a) Controlled simplex–cone mismatch. (b) Scale and cone statistics in training.

Figure 3: **Fairness-warped geometry in practice.** (a) Fairness scaling moves the simplex residual off the diagonal while leaving the cone residual nearly unchanged; hollow and filled points are $q = 0$ and $q = 1$. (b) On MNIST (seed 42, $q = 1$), radial and structural attacks occupy different regions of the (L_i, D_i^{cone}) plane, which is what motivates screening the two in sequence rather than with one test (Appendices E.2 and E.3).

6 Why Does Cone Geometry Matter?

The results above establish robustness, but do not show that the gain comes from the fairness-warped cone geometry rather than from some other ingredient of the aggregator. We therefore ask whether q-fair scaling breaks the simplex model in practice, whether the mismatch changes behavior in an otherwise identical pipeline, which components are responsible, and how the method behaves as its operating parameters vary.

6.1 Fairness Scaling Breaks the Simplex in Practice

Holding the checkpoint, honest submissions, and trusted basis fixed and varying only the coefficient constraint, the median cone/simplex residual on MNIST is $0.178/0.185$ at $q = 0$ but $0.178/1.077$ after $q = 1$ scaling: radial scaling preserves the cone, whereas the unit-sum simplex becomes mismatched (Figure 3(a)). The recovered coefficient mass tracks the fairness scale (Pearson $r = 0.940$ between L_i and F_i^q), and normalized coefficients reconstruct client losses with $r = 0.923$ and MAE 0.053.

Table 3: **Component ablation of FUSE.** CIFAR-10/ResNet-18, seed 42, 1000 rounds. Attack columns give global accuracy (%) over the final ten rounds; the worst-attack column gives global / worst-10% client accuracy and is the componentwise minimum over the seven attacks. Clean uses $f = 0$; "–" marks a run with no terminal window.

Variant	Scale	Cone	Loss norm.	Clean	Magnitude	Sign	IPM$_{\lambda=3}$	Gauss.	ALIE	Min-Max	Min-Sum	Worst attack
FUSE	✓	✓	✓	**79.03**	**78.97**	77.50	**77.83**	**78.82**	78.50	**79.28**	78.71	**77.50/68.37**
w/o scale	–	✓	✓	**79.03**	52.38	**78.20**	76.08	**78.82**	**78.97**	79.00	**79.03**	52.38/40.25
w/o cone	✓	–	✓	**79.03**	78.67	69.95	–	**78.82**	55.16	36.71	38.94	–
w/o scale and cone	–	–	✓	**79.03**	63.23	70.16	–	26.11	63.20	34.07	35.66	–
w/o loss norm.	✓	✓	–	78.04	70.06	73.33	74.01	77.89	76.98	76.04	76.02	70.06/59.88
cone only	–	✓	–	78.04	10.89	71.59	75.30	77.65	76.87	78.04	78.04	10.89/0.05
scale only	✓	–	–	78.04	73.14	71.66	–	77.74	46.45	13.87	–	–
no component	–	–	–	78.04	16.91	72.56	–	–	48.22	15.41	15.04	–

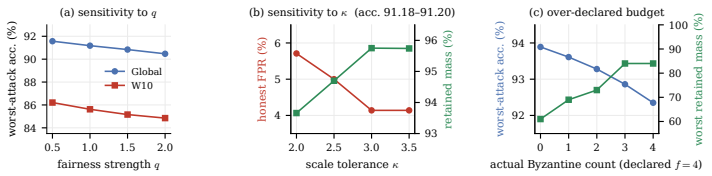


Figure 4: **Sensitivity and budget misspecification on MNIST/MLP, seed 42.** (a) Worst-attack accuracy over q. (b) Larger κ lowers honest FPR and raises retained fair mass, with accuracy within 91.18–91.20%. (c) When the declared budget stays at $f = 4$ while the true number of Byzantine submissions falls from 4 to 0, robustness is stable, but the surplus rejection budget is spent on honest clients: the worst retained fair mass drops from 84% to 61%. Over-declaring f therefore costs fairness rather than accuracy (Appendix E).

6.2 Matched Training Consequence of the Simplex Mismatch

To isolate the geometry, we replace only FUSE's Stage-II cone score by a unit-sum simplex score on MNIST. Both scores saturate on Gaussian and Magnitude Scaling, so the geometry is visible only where the structural test does the work. There the cone dominates: under IPM, Min-Max, and Min-Sum it lifts Byzantine recall from $0.45/0.56/0.62$ to $0.69/1.00/1.00$ and cuts the honest false-positive rate from $0.219/0.175/0.153$ to $0.124/0.001/0.000$, worth 2.44, 1.03, and 1.13 percentage points. Sign Flip is an informative counterexample, where the simplex score is the better detector (0.92 against 0.58). The cone is not universally superior; it is the representation matched to fairness-scaled honest mixtures (Appendix E.3).

6.3 Which Components Matter?

The full 2^3 ablation assigns each component the role our geometry predicts. Removing the scale screen exposes FUSE to Magnitude Scaling (worst attack 77.50 to 52.38) while leaving the other attacks intact, confirming it is the radial test. Removing the cone screen instead eliminates the IPM terminal window and breaks the invalid-mixture attacks (ALIE 78.50 to 55.16, Min-Max 79.28 to 36.71, Min-Sum 78.71 to 38.94); removing both also breaks Gaussian (26.11). The screens are thus complementary, while loss normalization is an optimization component, raising the floor from 70.06 to 77.50 (Appendix E.1).

6.4 Sensitivity to q, κ, and the Byzantine Budget

Worst-attack global/W10 accuracy varies smoothly from 91.57/86.21 to 90.47/84.85 across the tested q, so the method does not rely on a particular fairness strength. Increasing κ lowers the honest false-positive rate and raises the retained fair mass without materially changing accuracy, and over-declaring the budget costs fairness rather than accuracy (Appendix E). Neither statistic can certify maliciousness for a submission inside the trusted cone with a plausible recovered mass: the adaptive attack of Appendix E.14 reaches 65.6% on CIFAR-10 against 82.4% under clean training, an observable boundary of FUSE's two screening statistics rather than a missing threshold.

7 THEORETICAL ANALYSIS

We now characterize when the fairness-warped geometry exploited by FUSE is reliably observable, and when it suffices to exactly recover the honest set, thereby matching the post-screening aggregate obtained with oracle knowledge of the honest identities. The analysis proceeds in two steps: stable recovery of the latent fair coefficients, then conditions under which the two screening stages recover the honest set exactly.

7.1 RECOVERABILITY OF THE FAIRNESS-WARPED GEOMETRY

The cone identity in Section 3 is a population-level description. In practice, both client updates and the trusted class basis are estimated from finite data. For an honest client, write its screening-coordinate submission as

$$v_i = Ga_i + e_i, \qquad a_i := F_i^q p_i \succeq 0, \qquad \|a_i\|_1 = F_i^q, \tag{6}$$

where e_i collects the finite-sample client and trusted-basis approximation errors. We are now ready to state the corresponding recovery guarantee.

Theorem 1 (Stable recovery of the fairness-warped representation)**.** *If G has full column rank, the NNLS estimator in (4) satisfies* $\|\widehat{a}_i - a_i\|_2 \le \frac{2\|e_i\|_2}{\sigma_{\min}(G)}, \qquad |L_i - F_i^q| \le \frac{2\sqrt{C}\|e_i\|_2}{\sigma_{\min}(G)}$.

Recovery is therefore controlled by the finite-sample representation error and the conditioning of the trusted basis: when honest submissions stay close to the trusted cone, $L_i = \|\widehat{a}_i\|_1$ stably recovers the fairness scale F_i^q. The same argument yields stable composition and trusted-loss recovery; the full bounds and proof are in Appendix F, and the prediction matches the diagnostics in Figure 3.

7.2 ROBUSTNESS OF TWO-STAGE SCREENING

Stable recovery alone does not guarantee robust screening: a Byzantine submission must also be separated from the natural variation of honest submissions. Let $m_H := \mathrm{med}_{h\in\mathcal{H}_t} \log L_h$, $r_H := \max_{h\in\mathcal{H}_t} |\log L_h - m_H|$, $\epsilon_H^{\mathrm{cone}} := \max_{h\in\mathcal{H}_t} D_h^{\mathrm{cone}}$, and, for a vector c, let $(c)_- := \max\{-c, 0\}$ denote its coordinate-wise negative part. Stage I measures radial separation, whereas Stage II measures structural separation from the trusted nonnegative cone.

Theorem 2 (Oracle recovery under geometric separation)**.** *Assume that G has full column rank, honest submissions form a strict majority, $f_t = |\mathcal{B}_t|$, and $\log\kappa > 2r_H$. Any Byzantine submission satisfying $\log L_b - m_H > \log\kappa + r_H$ is rejected by Stage I; in particular, for a positive radial copy $v_b = \lambda v_h$ this holds whenever $\log\lambda > \log\kappa + 2r_H$. For any Byzantine submission surviving Stage I, write $v_b = Gc_b + r_b$ with $r_b \perp \mathrm{span}(G)$. If $\frac{\|r_b\|_2^2 + \sigma_{\min}^2(G)\|(c_b)_-\|_2^2}{\|v_b\|_2^2} > \epsilon_H^{\mathrm{cone}}$, then v_b ranks above every honest submission in Stage II. Consequently, if every Byzantine submission satisfies one of these two separation conditions, then $\widehat{\mathcal{H}}_t = \mathcal{H}_t$, and Eq. (5) returns exactly the aggregate that would be obtained by running the same post-screening rule with oracle knowledge of the honest identities.*

The theorem makes the roles of the two statistics explicit. Robustness is controlled by the honest radial spread r_H, the honest cone variation $\epsilon_H^{\mathrm{cone}}$, and the radial or structural deviation of the Byzantine submission. Stage I detects excessive radial scaling, while Stage II detects off-span energy or negative mixture mass beyond honest finite-sample variation. Their complementary behavior is observed empirically in Figure 3 and Appendix E.2. If neither separation condition holds, the theorem provides no certificate of maliciousness; the adaptive cone-preserving attack in Appendix E.14 probes this observable boundary.

8 CONCLUSION

We formalized Byzantine-robust q-fair federated learning under an update-only interface, and showed that fairness scaling relaxes the unit-sum simplex of honest class mixtures into a nonnegative cone with client-dependent radial mass. Building on this geometry, we proposed FUSE, which screens the recovered coefficient mass and the cone residual in sequence and rescales the retained mean by a reconstructed loss. Our experiments show that FUSE consistently improves worst-attack accuracy while preserving the high-loss honest clients, and our adaptive attack delimits the interface: only deviations the observable statistics separate can be resisted.

Reproducibility Statement

All details needed to reproduce our experiments are given in Section 5 and Appendix B: registered seeds, data partitions, hyperparameters, attacks, and the reporting protocol. Component, boundary, and geometry diagnostics are in Appendix E.

AI Use Statement

Generative AI assisted execution, debugging, and language polishing. The authors developed the research and verified all assisted content and results.

References

Youssef Allouah, Rachid Guerraoui, Nirupam Gupta, Rafaël Pinot, and Geovani Rizk. Robust distributed learning: Tight error bounds and breakdown point under data heterogeneity. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Wenxuan Bao, Jun Wu, and Jingrui He. BOBA: Byzantine-robust federated learning with label skewness. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *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 *Advances in Neural Information Processing Systems*, volume 32, 2019.

Alina Basharat, Yijun Bian, Ping Xu, and Zhi Tian. Towards trustworthy federated learning. *arXiv preprint arXiv:2503.03684*, 2025. doi: 10.48550/arXiv.2503.03684.

Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. Machine learning with adversaries: Byzantine tolerant gradient descent. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

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*, 2021. doi: 10.14722/ndss.2021.24434.

Bhaskar Ray Chaudhury, Linyi Li, Mintong Kang, Bo Li, and Ruta Mehta. Fairness in federated learning via core-stability. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Yudong Chen, Lili Su, and Jiaming Xu. Distributed statistical machine learning in adversarial settings: Byzantine gradient descent. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 1(2):44:1–44:25, 2017. doi: 10.1145/3154503.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724–1734, 2014.

Gregory Cohen, Saeed Afshar, Jonathan Tapson, and André van Schaik. EMNIST: Extending MNIST to handwritten letters. In *International Joint Conference on Neural Networks (IJCNN)*, pp. 2921–2926, 2017.

El-Mahdi El-Mhamdi, Rachid Guerraoui, and Sébastien Rouault. The hidden vulnerability of distributed learning in byzantium. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3521–3530. PMLR, 2018.

Jialuo He, Wei Chen, and Xiaojin Zhang. FedAA: A reinforcement learning perspective on adaptive aggregation for fair and robust federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 17085–17093, 2025. doi: 10.1609/aaai.v39i16.33878.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

Zeou Hu, Kiarash Shaloudegi, Guojun Zhang, and Yaoliang Yu. Federated learning meets multi-objective optimization. *IEEE Transactions on Network Science and Engineering*, 9(4):2039–2051, 2022. doi: 10.1109/TNSE.2022.3169117.

Sai Praneeth Reddy Karimireddy, Lie He, and Martin Jaggi. Byzantine-robust learning on heterogeneous datasets via bucketing. In *International Conference on Learning Representations*, 2022.

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

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86:2278–2324, 1998.

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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 1544–1551, 2019. doi: 10.1609/aaai.v33i01.33011544.

Tian Li, Maziar Sanjabi, Ahmad Beirami, and Virginia Smith. Fair resource allocation in federated learning. In *International Conference on Learning Representations*, 2020.

Tian Li, Ahmad Beirami, Maziar Sanjabi, and Virginia Smith. Tilted empirical risk minimization. In *International Conference on Learning Representations*, 2021.

Xinpeng Ling, Jie Fu, Kuncan Wang, Huifa Li, Tong Cheng, and Zhili Chen. Fedfdp: Fairness-aware federated learning with differential privacy. In *Applied Cryptography and Network Security*, volume 16572 of *Lecture Notes in Computer Science*, pp. 129–159. Springer Nature Switzerland, 2026. doi: 10.1007/978-3-032-32575-4_6.

Yuchen Liu, Chen Chen, Lingjuan Lyu, Fangzhao Wu, Sai Wu, and Gang Chen. Byzantine-robust learning on heterogeneous data via gradient splitting. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 21404–21425. PMLR, 2023.

Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4615–4625. PMLR, 2019.

Zibin Pan, Chi Li, Fangchen Yu, Shuyi Wang, Haijin Wang, Xiaoying Tang, and Junhua Zhao. FedLF: Layer-wise fair federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 14527–14535, 2024. doi: 10.1609/aaai.v38i13.29368.

Krishna Pillutla, Sham M. Kakade, and Zaid Harchaoui. Robust aggregation for federated learning. *IEEE Transactions on Signal Processing*, 70:1142–1154, 2022. doi: 10.1109/TSP.2022.3153135.

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*, 2021. doi: 10.14722/ndss.2021.24498.

Cong Xie, Oluwasanmi Koyejo, and Indranil Gupta. Fall of empires: Breaking byzantine-tolerant sgd by inner product manipulation. In *Proceedings of the 35th Uncertainty in Artificial Intelligence Conference*, volume 115 of *Proceedings of Machine Learning Research*, pp. 261–270. PMLR, 2020.

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

Guojun Zhang, Saber Malekmohammadi, Xi Chen, and Yaoliang Yu. Proportional fairness in federated learning. *Transactions on Machine Learning Research*, 2023.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, 2015.

Guanqiang Zhou, Ping Xu, Yue Wang, and Zhi Tian. H-nobs: Achieving certified fairness and robustness in distributed learning on heterogeneous datasets. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Ligeng Zhu, Zhijian Liu, and Song Han. Deep leakage from gradients. In *Advances in Neural Information Processing Systems*, volume 32, pp. 14747–14756, 2019.

A Related Work

Client-level fairness. Client-level fairness has been pursued through q-FFL, agnostic and tilted objectives, multi-objective optimization, and proportional, coalitional, or gradient-conflict principles (Li et al., 2020; Mohri et al., 2019; Li et al., 2021; Hu et al., 2022; Zhang et al., 2023; Chaudhury et al., 2022; Pan et al., 2024). Yet, the Byzantine robustness of these schemes has received limited attention. An initial preliminary work was conducted by Zhou et al. (2023), who combine fairness-aware learning with a norm filter, as does Basharat et al. (2025). However, a norm filter reads exactly the quantity that fairness is designed to inflate, so it suppresses the high-loss clients it is meant to protect. Closest in intent, FedAA learns aggregation weights with a reinforcement agent driven by pairwise distances between client models (He et al., 2025); its weights are still a function of how far a submission sits from the others, so it too never models the $F_i^q \Delta_i$ product that makes a high-loss honest client distant by construction. We instead treat that ambiguity as the object of study and exploit the class-mixture structure that survives fairness scaling.

Byzantine distributed learning. Byzantine robustness has become a central topic in distributed learning, where malicious clients may send arbitrary updates during training (Blanchard et al., 2017; Yin et al., 2018; Chen et al., 2017; Pillutla et al., 2022; El-Mhamdi et al., 2018; Cao et al., 2021; Shejwalkar & Houmansadr, 2021; Li et al., 2019; Karimireddy et al., 2022; Liu et al., 2023; Allouah et al., 2023). Robust aggregators in this line filter by distances, order statistics, trusted directions, spectral structure, regularization, bucketing, or gradient decomposition, all of which treat honest updates as concentrated around a common mean. Our setting differs in that every honest q-fair update undergoes a loss-dependent radial transformation before reaching the server, so honest submissions are deliberately dispersed in magnitude.

Structure-aware Byzantine aggregation. A parallel line of work exploits the structure of updates under label skew rather than their concentration. BOBA, our closest structural comparison, models unscaled updates with a unit-sum class-mixture simplex (Bao et al., 2024). These approaches, however, assume normalized honest coefficients, an assumption which does not hold here: fairness weighting preserves nonnegativity but scales the coefficients by F_i^q. FUSE therefore retains the radial mass as a screening statistic in its own right, and normalizes the coefficients only when class composition is needed.

B Experimental Setup and Attack Details

B.1 Datasets, models, and optimization

The benchmark contains 100 honest clients and samples 10 honest clients per communication round. Attacked rounds append four Byzantine Sybil submissions, while clean rounds append none. Client data follow a Dirichlet label-skew partition with concentration $\alpha = 1.0$ and class-availability probability 0.7. Each honest client reserves 2% of its local data for validation. All reported runs use one local epoch, a whole-client batch, SGD without momentum, and fairness strength $q = 1$ unless a sensitivity experiment states otherwise.

Table 4: Dataset, model, and optimization settings used by the main comparison.

Dataset	Model	LR	Rounds	Classes	Batch	Weight decay
CIFAR-10	ResNet-18	0.03	1500	10	whole client	0
MNIST	MLP (784–200–10)	0.10	1000	10	whole client	5×10^{-4}
EMNIST-Balanced	five-layer CNN	0.03	1000	47	whole client	5×10^{-4}
AGNews	one-layer GRU	0.05	2000	4	whole client	5×10^{-4}

The five-layer CNN contains three unpadded 3×3 convolutions with 32, 64, and 64 output channels, followed by fully connected layers of width 64 and the dataset-specific classifier. MNIST uses one hidden layer with 200 ReLU units. AGNews uses 300-dimensional embeddings and a one-layer unidirectional GRU with hidden dimension 256; tokens with frequency below two are removed and sequences are truncated to length 128.

B.2 ATTACK DEFINITIONS AND ADVERSARIAL INFORMATION

The attack suite contains seven Byzantine attacks. Sign Flip submits $-\bar{u}_H$, where $\bar{u}_H$ is the current honest submitted mean. IPM submits $-3\bar{u}_H$. Gaussian uses standard deviation 200. ALIE uses the original-z rule with configured value approximately 0.2533 for 10 honest and four Byzantine submissions. Min-Max and Min-Sum use $\gamma_0 = 10$, search tolerance 10^{-5}, and shared Sybils. Magnitude Scaling copies the sampled honest submission with the highest start loss and multiplies its magnitude by five.

The attack generator therefore receives the honest submitted mean for Sign Flip and IPM and simulator-side start losses for the Magnitude source choice. These quantities are not inputs to FUSE. The adversary knows the learning rule, model, defense, and hyperparameters but not the realized trusted examples or trusted class losses. Trusted server records are assumed clean.

B.3 FUSE IMPLEMENTATION

Each client uploads only $u_i = F_i(w_t)^q \Delta_i$. FUSE uses 20 unique trusted examples per class. CIFAR-10/ResNet-18 uses 24 stochastic augmentation views per trusted example; the other main dataset/model configurations use one view. The scale threshold is $\kappa = 2.5$. The algorithm receives a round-specific declared budget f_t: the main benchmark uses $f_t = 4$ in attacked conditions and $f_t = 0$ in clean conditions, matching the number of Byzantine submissions. Appendix E.12 evaluates the over-declared clean setting $f_t = 4$.

Screening uses the classifier parameters for ResNet-18, MNIST/MLP, and EMNIST/CNN, and all model parameters for AGNews/GRU. Accepted clients contribute their original full-parameter submissions to Eq. (5). To recover coefficients efficiently, the implementation forms $K = G^\top G, \qquad b_i = G^\top v_i$, removes numerically degenerate reference directions, and solves NNLS in the retained at-most-C-dimensional Gram space. Let $\widehat{a}_i$ denote this implementation-level NNLS solution. Forming the inner products still depends on the selected parameter dimension; only the constrained numerical solve is low dimensional.

The trusted start-of-round class-loss vector $\widehat{\ell}$ is evaluated on the same broadcast model as the client fairness losses. The final method uses $\widetilde{F}_i = \widehat{\ell}^\top \frac{\widehat{a}_i}{\max\{\|\widehat{a}_i\|_1, \epsilon\}}, \qquad d_i = \max\{1, \widetilde{F}_i^q\}$, No client loss or additional scalar is communicated.

B.4 BASELINES AND INFORMATION COMPATIBILITY

q-FFL-GD averages the same fairness-scaled gradient carrier used by the benchmark. H-nobs follows its norm-based screening mechanism: it removes the declared number of largest-ℓ_2 fair gradients and averages the retained vectors. Median, Trimmed Mean, Krum, Multi-Krum, and DnC operate on the same submitted fair vector. FLTrust additionally constructs a clean server root gradient. FedAA and Ditto retain their method-specific aggregation or personalization mechanisms.

All matched DnC runs triggered the registered collapse criterion before their final reporting windows and are therefore not included in the main endpoint tables. BOBA is evaluated over the matched five seeds with classifier-coordinate geometry; the original full-parameter implementation is reproduced for seed 42 only and is compared with the classifier-coordinate variant in Appendix B.5. Table 5 summarizes the information-interface compatibility of the evaluated methods under our benchmark.

B.5 BOBA IMPLEMENTATION AND FULL-SPACE FIDELITY CHECK

BOBA (Bao et al., 2024) fits an affine subspace and a class-mixture simplex to the received updates and accepts the submissions whose simplex fit is plausible. In the original implementation both fits are computed in the full parameter space; on ResNet-18 this costs 30–47 GPU-hours per 1500-round run (20 Stage-1 subspace iterations on a 14×11M matrix), which is why the matched five-seed benchmark uses classifier-coordinate BOBA: the Stage-1 subspace, the simplex coefficients, and the acceptance decision are computed on the classifier block ($5{,}130$ coordinates for ResNet-18, $2{,}010$ for the MLP, $68{,}655$ for the CNN, $1{,}028$ for the GRU), the classifier coordinates are aggregated by BOBA's projection, and all other parameters are averaged over the accepted clients. For the three vision benchmarks, this matches the coordinate choice used by FUSE for screening, so the two

Table 5: Baseline information compatibility. "Same fair update" indicates whether the method is applied to the fairness-scaled client submission $u_i = F_i^q \Delta_i$ used in our benchmark. A dash indicates that the corresponding detail is not documented or cannot be verified from the current manuscript repository.

Method	Same fair update u_i	Extra trusted server data	Output	Uses declared f
q-FFL-GD (mean)	Yes	No	Shared global	No
H-nobs	Yes	No	Shared global	Yes
Trimmed Mean	Yes	No	Shared global	–
Median	Yes	No	Shared global	No documented dependence
Krum	Yes	No	Shared global	–
Multi-Krum	Yes	No	Shared global	–
FLTrust	Yes	Yes, clean root gradient	Shared global	–
FedAA	–	–	Shared aggregation	–
Ditto	–	No	Personalized	–
DnC	Yes	No	Shared global	–
BOBA	Yes	Yes, per-class references from the same pool	Shared global	Yes
FUSE	Yes	Yes, clean labeled server data	Shared global	Yes

methods see the same geometry there. All other BOBA settings follow the reference implementation ($p_{\min} = -0.5$, 20 Stage-1 iterations, the same 20 trusted examples per class as FUSE and FLTrust).

Table 6 reports the full-parameter reproduction for the only seed on which it exists. It is a sensitivity and fidelity check, not the primary statistical comparison, and a single seed does not support a statistical claim in either direction. Where both CIFAR-10 variants are available for seed 42, the classifier-coordinate variant is not uniformly weaker: it is stronger under clean training (82.19 vs. 78.34), Sign Flip (72.50 vs. 11.01), Gaussian (81.66 vs. 76.29), ALIE (73.70 vs. 24.27), and Magnitude (72.02 vs. 62.50), while the full-space variant is stronger under Min-Max (45.57 vs. 17.53) and Min-Sum (65.26 vs. 20.15). Both fail severely under IPM (11.84 for the classifier variant; the full-space run collapsed at round 18), and the full-space MNIST and EMNIST IPM runs also collapsed or ended near chance (2.69). The five-seed BOBA rows in Tables 1 and 7–10 are the numbers the paper's comparison rests on.

Table 6: **Full-parameter BOBA, seed 42 (fidelity check).** Affine subspace and simplex fitted in the full parameter space; global / worst-10% accuracy (%) over the registered final window. "collapsed (r)" = terminated by the registered collapse rule at round r; "n/a" = not run; "incomplete" = did not reach the endpoint within the compute budget.

Dataset	Clean	Sign	IPM	Gauss.	ALIE	Min-Max	Min-Sum	Mag.	Worst attack
CIFAR-10 / ResNet-18	78.34/68.84	11.01/0.30	collapsed (18)	76.29/66.97	24.27/11.68	45.57/33.37	65.26/53.62	62.50/51.41	11.01/0.30
MNIST / MLP	n/a	92.57/87.26	collapsed (17)	n/a	collapsed (779)	n/a	n/a	94.73/90.52	–
EMNIST-Balanced / CNN	n/a	n/a	2.69/0.17	n/a	n/a	n/a	n/a	n/a	–
AGNews / GRU	n/a	n/a	incomplete (1047)	n/a	n/a	n/a	n/a	n/a	–

B.6 METRICS AND REPORTING

We report global test accuracy and worst-10% client accuracy. For each seed, both metrics are averaged over the final ten rounds: 1490–1499 for CIFAR-10, 990–999 for MNIST and EMNIST-Balanced, and 1990–1999 for AGNews. Main tables then average the seed-level window means over seeds 40–44 that reach the registered endpoint. Superscript (k) indicates that only k seeds reach the endpoint; a dash indicates that no valid complete result is available. Pre-collapse maxima are never substituted.

The main worst-attack result is seed-wise: for every seed completing all seven attacks, we first take the minimum attacked accuracy and then average these minima across seeds. This conservative statistic supports paired comparisons and differs from the minimum of the seven attack-column means. Clean runs use $f_t = 0$. Selection recall, honest FPR, retained fairness mass, and loss-quantile audits use simulator-side labels or losses only for offline evaluation; they are not selector inputs. The high-loss retention audit in Table 2 uses all five registered seeds and pools rounds r with $(r+1) \bmod 10 = 0$, the grid common to every logging cadence; the H-nobs fairness comparison and the loss-quantile breakdown in Table 24 remain seed-42 audits. A collapsed trajectory contributes diagnostic observations only through its registered stopping round.

Table 7: CIFAR-10 / ResNet-18, rounds 1490–1499. Entries are global / worst-10% client accuracy (%), mean over seeds 40, 41, 42, 43, 44 with the across-seed standard deviation as subscript; the worst-attack column is the seed-wise minimum over the seven attacks, also mean with std subscript, over seeds completing all seven attacks.

Method	Clean	Sign	IPM$_{\lambda=3}$	Gauss.	ALIE	Min-Max	Min-Sum	Magnitude	Worst attack
q-FFL-GD (mean)	$76.96_{\pm 8.6}$/$67.34_{\pm 9.9}$	$76.47_{\pm 2.1}$/$66.94_{\pm 2.7}$	-	$13.81_{\pm 3.3}$/$4.71_{\pm 4.2}$ $^{(3)}$	$63.43_{\pm 13.4}$/$51.90_{\pm 14.7}$	$14.76_{\pm 1.4}$/$2.24_{\pm 0.6}$ $^{(4)}$	$14.05_{\pm 1.9}$/$2.04_{\pm 0.9}$	$43.45_{\pm 31.2}$/$32.27_{\pm 31.3}$ $^{(4)}$	-
H-nobs	$76.96_{\pm 8.6}$/$67.34_{\pm 9.9}$	$12.19_{\pm 1.9}$/$0.35_{\pm 0.5}$ $^{(3)}$	$57.13_{\pm 24.6}$/$46.25_{\pm 24.8}$	$79.28_{\pm 3.7}$/$70.03_{\pm 4.4}$	$35.65_{\pm 20.3}$/$23.39_{\pm 19.8}$	$42.86_{\pm 32.1}$/$31.99_{\pm 33.3}$	$20.15_{\pm 11.2}$/$8.34_{\pm 9.5}$	$76.96_{\pm 8.6}$/$67.34_{\pm 9.9}$	$12.19_{\pm 1.9}$/$0.35_{\pm 0.5}$ $^{(3)}$
Trimmed Mean	$70.56_{\pm 18.4}$/$60.32_{\pm 20.4}$	$59.66_{\pm 7.7}$/$46.72_{\pm 9.3}$	$44.84_{\pm 13.6}$/$31.29_{\pm 15.2}$	$79.45_{\pm 3.5}$/$70.42_{\pm 3.4}$	$39.27_{\pm 9.6}$/$25.86_{\pm 8.8}$	$13.37_{\pm 2.1}$/$1.31_{\pm 1.2}$	$14.06_{\pm 1.8}$/$2.08_{\pm 1.1}$	$74.35_{\pm 7.3}$/$64.41_{\pm 8.6}$	$13.18_{\pm 1.9}$/$1.28_{\pm 1.1}$
Median	$45.02_{\pm 16.4}$/$31.89_{\pm 15.2}$	$27.99_{\pm 2.4}$/$14.01_{\pm 2.4}$	$23.83_{\pm 2.5}$/$10.14_{\pm 2.8}$	$41.65_{\pm 14.1}$/$28.89_{\pm 13.2}$	$24.71_{\pm 10.7}$/$12.87_{\pm 9.3}$	$12.85_{\pm 1.5}$/$1.32_{\pm 0.8}$	$12.73_{\pm 2.5}$/$1.22_{\pm 1.0}$	$47.89_{\pm 9.8}$/$34.11_{\pm 10.1}$	$12.34_{\pm 1.9}$/$1.10_{\pm 0.9}$
Krum	$63.66_{\pm 17.4}$/$51.97_{\pm 19.3}$	$44.69_{\pm 7.9}$/$29.35_{\pm 6.8}$	$62.78_{\pm 17.7}$/$51.13_{\pm 19.6}$	$60.21_{\pm 20.7}$/$48.53_{\pm 21.7}$	$22.24_{\pm 11.2}$/$10.14_{\pm 10.2}$	$61.58_{\pm 18.0}$/$50.09_{\pm 19.5}$	$45.62_{\pm 14.3}$/$32.93_{\pm 15.0}$	$63.56_{\pm 17.3}$/$51.66_{\pm 18.9}$	$21.67_{\pm 10.4}$/$9.78_{\pm 9.7}$
Multi-Krum	$80.15_{\pm 3.6}$/$71.05_{\pm 4.6}$	$15.49_{\pm 4.9}$/$3.79_{\pm 3.3}$ $^{(3)}$	$68.36_{\pm 3.1}$/$57.04_{\pm 2.4}$	$81.90_{\pm 1.5}$/$73.17_{\pm 1.8}$	$39.70_{\pm 10.9}$/$27.01_{\pm 10.2}$	$17.25_{\pm 5.9}$/$4.71_{\pm 4.7}$	$19.64_{\pm 4.9}$/$7.13_{\pm 3.8}$	$80.15_{\pm 3.6}$/$71.05_{\pm 4.6}$	$12.97_{\pm 2.6}$/$1.48_{\pm 1.3}$ $^{(3)}$
FLTrust	$81.31_{\pm 0.5}$/$72.22_{\pm 0.5}$	$81.29_{\pm 0.3}$/$72.26_{\pm 0.4}$	$81.31_{\pm 0.3}$/$72.27_{\pm 0.4}$	$81.25_{\pm 0.4}$/$72.16_{\pm 0.5}$	$55.13_{\pm 1.9}$/$42.07_{\pm 1.5}$	$48.84_{\pm 1.8}$/$35.47_{\pm 1.8}$	$43.81_{\pm 2.5}$/$30.53_{\pm 2.5}$	$79.44_{\pm 1.0}$/$70.05_{\pm 1.1}$	$43.81_{\pm 2.5}$/$30.53_{\pm 2.5}$
BOBA (cls.)	$81.93_{\pm 1.1}$/$73.13_{\pm 1.2}$	$72.68_{\pm 1.8}$/$62.12_{\pm 1.8}$	$10.93_{\pm 0.7}$/$0.66_{\pm 0.8}$	$81.58_{\pm 1.3}$/$72.73_{\pm 1.6}$	$72.00_{\pm 3.6}$/$60.85_{\pm 3.8}$	$17.31_{\pm 0.8}$/$5.27_{\pm 1.0}$	$17.83_{\pm 1.9}$/$5.35_{\pm 2.2}$	$73.64_{\pm 1.5}$/$62.39_{\pm 2.1}$	$10.93_{\pm 0.7}$/$0.66_{\pm 0.8}$
FedAA	$75.07_{\pm 1.1}$/$64.87_{\pm 1.3}$	$45.32_{\pm 19.1}$/$32.48_{\pm 19.3}$	$74.07_{\pm 2.6}$/$62.88_{\pm 3.7}$ $^{(4)}$	$72.60_{\pm 1.4}$/$61.23_{\pm 5.8}$	$61.44_{\pm 18.1}$/$49.78_{\pm 19.4}$	$22.54_{\pm 6.3}$/$10.34_{\pm 5.7}$	$43.07_{\pm 18.9}$/$30.00_{\pm 19.9}$	$66.07_{\pm 6.1}$/$53.75_{\pm 7.5}$	$22.15_{\pm 7.2}$/$10.27_{\pm 6.4}$ $^{(4)}$
Ditto	$81.73_{\pm 1.1}$/$68.94_{\pm 2.6}$	$75.10_{\pm 1.2}$/$24.59_{\pm 8.9}$	-	$15.55_{\pm 0.9}$/$5.25_{\pm 1.8}$	$75.21_{\pm 1.8}$/$35.95_{\pm 3.7}$	$33.69_{\pm 1.2}$/$16.73_{\pm 0.6}$	$34.39_{\pm 1.9}$/$19.40_{\pm 3.2}$	$75.52_{\pm 3.0}$/$63.50_{\pm 3.1}$	-
FUSE	$82.25_{\pm 0.5}$/$73.88_{\pm 0.8}$	$79.29_{\pm 4.3}$/$70.15_{\pm 5.9}$	$80.59_{\pm 1.1}$/$71.62_{\pm 1.4}$	$82.30_{\pm 0.5}$/$73.91_{\pm 0.8}$	$77.74_{\pm 10.2}$/$68.35_{\pm 12.1}$	$82.17_{\pm 0.5}$/$73.91_{\pm 0.9}$	$82.19_{\pm 0.5}$/$73.87_{\pm 0.8}$	$81.12_{\pm 2.4}$/$72.67_{\pm 3.1}$	$73.88_{\pm 9.1}$/$63.58_{\pm 10.7}$

Table 8: MNIST / MLP, rounds 990–999. Entries are global / worst-10% client accuracy (%), mean over seeds 40, 41, 42, 43, 44 with the across-seed standard deviation as subscript; the worst-attack column is the seed-wise minimum over the seven attacks, also mean with std subscript, over seeds completing all seven attacks.

Method	Clean	Sign	IPM$_{\lambda=3}$	Gauss.	ALIE	Min-Max	Min-Sum	Magnitude	Worst attack
q-FFL-GD (mean)	$94.90_{\pm 1.3}$/$90.72_{\pm 2.1}$	$93.24_{\pm 1.1}$/$88.31_{\pm 1.5}$	-	$62.99_{\pm 42.3}$/$54.32_{\pm 45.8}$ $^{(3)}$	$94.29_{\pm 1.1}$/$89.66_{\pm 1.6}$	$48.54_{\pm 33.0}$/$37.24_{\pm 32.5}$ $^{(3)}$	$56.03_{\pm 40.0}$/$46.60_{\pm 40.4}$ $^{(3)}$	$68.76_{\pm 49.7}$/$63.08_{\pm 54.6}$ $^{(3)}$	-
H-nobs	$94.90_{\pm 1.3}$/$90.72_{\pm 2.1}$	$14.82_{\pm 4.0}$/$1.91_{\pm 2.1}$ $^{(3)}$	$91.09_{\pm 2.2}$/$84.66_{\pm 2.3}$	$94.90_{\pm 1.3}$/$90.72_{\pm 2.1}$	$93.45_{\pm 1.3}$/$88.26_{\pm 1.3}$	$94.80_{\pm 1.3}$/$90.59_{\pm 2.0}$	$65.54_{\pm 44.3}$/$56.83_{\pm 49.0}$ $^{(3)}$	$94.90_{\pm 1.3}$/$90.72_{\pm 2.1}$	$14.82_{\pm 4.0}$/$1.91_{\pm 2.1}$ $^{(3)}$
Trimmed Mean	$94.90_{\pm 1.3}$/$90.71_{\pm 2.1}$	$85.77_{\pm 4.7}$/$75.80_{\pm 7.0}$	$87.30_{\pm 3.6}$/$78.31_{\pm 8.1}$ $^{(4)}$	$94.84_{\pm 1.4}$/$90.46_{\pm 2.3}$	$65.65_{\pm 48.6}$/$59.07_{\pm 51.2}$ $^{(3)}$	$48.61_{\pm 34.4}$/$37.79_{\pm 32.7}$ $^{(3)}$	$48.74_{\pm 33.7}$/$37.52_{\pm 32.5}$ $^{(3)}$	$95.33_{\pm 1.3}$/$91.29_{\pm 1.9}$	$48.24_{\pm 34.1}$/$37.25_{\pm 32.3}$ $^{(3)}$
Median	$66.04_{\pm 39.1}$/$55.57_{\pm 42.6}$ $^{(4)}$	$61.37_{\pm 44.3}$/$51.79_{\pm 45.0}$ $^{(3)}$	$86.33_{\pm 5.0}$/$78.61_{\pm 6.5}$ $^{(3)}$	$66.10_{\pm 48.8}$/$59.72_{\pm 51.7}$ $^{(3)}$	$67.35_{\pm 41.2}$/$58.41_{\pm 47.1}$ $^{(3)}$	$47.54_{\pm 32.9}$/$36.27_{\pm 31.4}$ $^{(3)}$	$47.25_{\pm 32.3}$/$35.64_{\pm 30.9}$ $^{(3)}$	$73.93_{\pm 42.4}$/$68.19_{\pm 45.5}$ $^{(4)}$	$65.88_{\pm 0.4}$/$53.38_{\pm 0.6}$ $^{(2)}$
Krum	$93.99_{\pm 1.4}$/$89.24_{\pm 2.5}$	$90.45_{\pm 1.5}$/$83.30_{\pm 2.7}$	$93.42_{\pm 1.3}$/$88.42_{\pm 2.3}$	$93.99_{\pm 1.4}$/$89.24_{\pm 2.5}$	$63.99_{\pm 46.7}$/$56.77_{\pm 49.2}$ $^{(3)}$	$94.00_{\pm 1.5}$/$89.24_{\pm 2.6}$	$93.76_{\pm 1.2}$/$88.61_{\pm 1.7}$	$93.99_{\pm 1.4}$/$89.24_{\pm 2.5}$	$63.99_{\pm 46.7}$/$56.14_{\pm 48.6}$ $^{(3)}$
Multi-Krum	$94.72_{\pm 1.4}$/$90.39_{\pm 2.3}$	$59.90_{\pm 43.4}$/$46.41_{\pm 40.2}$ $^{(3)}$	$90.27_{\pm 0.5}$/$83.57_{\pm 1.1}$	$94.72_{\pm 1.4}$/$90.39_{\pm 2.3}$	$66.22_{\pm 48.9}$/$59.88_{\pm 51.9}$ $^{(3)}$	$94.16_{\pm 1.0}$/$89.52_{\pm 1.4}$	$61.54_{\pm 44.4}$/$54.29_{\pm 47.1}$ $^{(3)}$	$94.72_{\pm 1.4}$/$90.39_{\pm 2.3}$	$59.90_{\pm 43.4}$/$46.41_{\pm 40.2}$ $^{(3)}$
FLTrust	$94.96_{\pm 0.6}$/$90.53_{\pm 0.8}$	$93.30_{\pm 0.6}$/$88.05_{\pm 1.7}$	$93.34_{\pm 0.8}$/$88.24_{\pm 1.6}$	$94.98_{\pm 0.5}$/$90.54_{\pm 0.8}$	$93.10_{\pm 3.4}$/$87.16_{\pm 5.7}$	$89.97_{\pm 3.0}$/$82.26_{\pm 6.4}$	$91.77_{\pm 2.4}$/$85.34_{\pm 3.4}$	$95.07_{\pm 0.6}$/$90.76_{\pm 0.8}$	$88.52_{\pm 3.3}$/$79.64_{\pm 5.2}$
BOBA (cls.)	$96.41_{\pm 0.1}$/$92.68_{\pm 0.4}$	$94.45_{\pm 0.2}$/$89.83_{\pm 0.3}$	$11.23_{\pm 0.5}$/$0.16_{\pm 0.4}$	$96.41_{\pm 0.1}$/$92.68_{\pm 0.4}$	$95.70_{\pm 0.1}$/$91.40_{\pm 0.5}$	$80.55_{\pm 3.8}$/$70.92_{\pm 5.8}$	$82.26_{\pm 5.4}$/$72.58_{\pm 7.2}$	$96.60_{\pm 0.1}$/$93.00_{\pm 0.2}$	$11.23_{\pm 0.5}$/$0.16_{\pm 0.4}$
FedAA	$96.42_{\pm 0.2}$/$92.71_{\pm 0.4}$	$51.53_{\pm 40.3}$/$39.47_{\pm 37.7}$	$85.86_{\pm 10.7}$/$77.56_{\pm 14.4}$	$96.29_{\pm 0.3}$/$92.31_{\pm 0.9}$	$96.30_{\pm 0.1}$/$92.39_{\pm 0.4}$	$74.04_{\pm 10.2}$/$62.88_{\pm 13.1}$	$60.57_{\pm 18.8}$/$47.28_{\pm 20.0}$	$96.22_{\pm 0.1}$/$92.41_{\pm 0.2}$	$43.00_{\pm 32.2}$/$31.26_{\pm 28.8}$
Ditto	$96.41_{\pm 0.1}$/$92.72_{\pm 0.6}$	$94.53_{\pm 0.1}$/$89.05_{\pm 1.2}$	-	$87.36_{\pm 0.5}$/$78.18_{\pm 3.4}$	$95.95_{\pm 0.1}$/$91.73_{\pm 1.0}$	$81.79_{\pm 1.4}$/$63.73_{\pm 3.9}$	$76.89_{\pm 4.6}$/$61.49_{\pm 1.4}$	$97.33_{\pm 0.2}$/$94.41_{\pm 0.8}$	-
FUSE	$93.83_{\pm 0.1}$/$89.16_{\pm 0.2}$	$93.15_{\pm 0.1}$/$88.30_{\pm 0.4}$	$92.42_{\pm 0.3}$/$87.07_{\pm 0.2}$	$93.83_{\pm 0.1}$/$89.16_{\pm 0.2}$	$93.81_{\pm 0.1}$/$89.17_{\pm 0.1}$	$93.84_{\pm 0.1}$/$89.14_{\pm 0.2}$	$93.84_{\pm 0.1}$/$89.15_{\pm 0.2}$	$93.83_{\pm 0.1}$/$89.19_{\pm 0.2}$	$92.42_{\pm 0.3}$/$87.07_{\pm 0.2}$

Table 9: EMNIST-Balanced / CNN, rounds 990–999. Entries are global / worst-10% client accuracy (%), mean over seeds 40, 41, 42, 43, 44 with the across-seed standard deviation as subscript; the worst-attack column is the seed-wise minimum over the seven attacks, also mean with std subscript, over seeds completing all seven attacks.

Method	Clean	Sign	IPM$_{\lambda=3}$	Gauss.	ALIE	Min-Max	Min-Sum	Magnitude	Worst attack
q-FFL-GD (mean)	$63.79_{\pm 34.5}$/$57.08_{\pm 32.0}$	$74.93_{\pm 2.5}$/$66.23_{\pm 3.0}$	-	$2.13_{\pm 0.0}$/$0.00_{\pm 0.0}$ $^{(1)}$	$53.66_{\pm 44.4}$/$46.71_{\pm 40.4}$ $^{(3)}$	$44.08_{\pm 36.5}$/$36.43_{\pm 31.8}$ $^{(3)}$	$27.01_{\pm 38.2}$/$21.24_{\pm 34.3}$ $^{(3)}$	$65.81_{\pm 35.6}$/$58.92_{\pm 33.0}$	-
H-nobs	$63.79_{\pm 34.5}$/$57.08_{\pm 32.0}$	$28.70_{\pm 37.6}$/$21.64_{\pm 30.6}$ $^{(2)}$	$78.66_{\pm 1.5}$/$70.66_{\pm 1.5}$	$63.79_{\pm 34.5}$/$57.08_{\pm 32.0}$	$78.55_{\pm 0.6}$/$69.11_{\pm 0.4}$ $^{(3)}$	$52.75_{\pm 43.8}$/$45.63_{\pm 39.5}$ $^{(3)}$	$50.65_{\pm 42.0}$/$43.19_{\pm 37.4}$ $^{(3)}$	$63.79_{\pm 34.5}$/$57.08_{\pm 32.0}$	$28.70_{\pm 37.6}$/$21.64_{\pm 30.6}$ $^{(2)}$
Trimmed Mean	$80.30_{\pm 0.5}$/$72.43_{\pm 0.6}$	$63.17_{\pm 1.7}$/$52.69_{\pm 2.0}$	$54.93_{\pm 4.5}$/$44.44_{\pm 3.9}$ $^{(4)}$	$63.28_{\pm 34.3}$/$56.28_{\pm 31.6}$	$66.82_{\pm 18.7}$/$57.29_{\pm 18.9}$ $^{(4)}$	$25.45_{\pm 20.2}$/$17.89_{\pm 15.6}$ $^{(3)}$	$34.35_{\pm 29.2}$/$26.51_{\pm 24.6}$ $^{(3)}$	$64.13_{\pm 34.7}$/$56.61_{\pm 31.7}$	$37.11_{\pm 1.8}$/$26.83_{\pm 2.4}$ $^{(2)}$
Median	$71.43_{\pm 10.3}$/$60.98_{\pm 11.1}$ $^{(4)}$	$45.35_{\pm 21.3}$/$34.82_{\pm 19.9}$	$58.80_{\pm 2.6}$/$47.41_{\pm 7.3}$	$59.14_{\pm 30.2}$/$49.74_{\pm 29.3}$ $^{(3)}$	$65.63_{\pm 9.7}$/$56.19_{\pm 9.3}$ $^{(4)}$	$17.57_{\pm 26.8}$/$12.35_{\pm 21.4}$ $^{(3)}$	$30.96_{\pm 25.1}$/$22.84_{\pm 20.0}$ $^{(3)}$	$78.51_{\pm 0.5}$/$68.41_{\pm 0.1}$ $^{(3)}$	$15.72_{\pm 23.6}$/$10.51_{\pm 18.2}$ $^{(3)}$
Krum	$78.49_{\pm 0.8}$/$70.12_{\pm 0.7}$	$7.35_{\pm 2.5}$/$1.24_{\pm 1.0}$ $^{(2)}$	$78.64_{\pm 0.7}$/$70.15_{\pm 0.7}$	$78.49_{\pm 0.8}$/$70.12_{\pm 0.7}$	$71.34_{\pm 4.5}$/$60.74_{\pm 4.7}$ $^{(3)}$	$47.92_{\pm 39.7}$/$39.93_{\pm 34.6}$ $^{(3)}$	$4.12_{\pm 3.5}$/$1.02_{\pm 1.8}$ $^{(3)}$	$78.49_{\pm 0.8}$/$70.12_{\pm 0.7}$	$5.12_{\pm 4.3}$/$0.97_{\pm 1.4}$ $^{(2)}$
Multi-Krum	$80.49_{\pm 0.5}$/$72.46_{\pm 0.6}$	$62.69_{\pm 2.9}$/$52.43_{\pm 3.1}$ $^{(3)}$	$78.71_{\pm 0.7}$/$70.46_{\pm 1.2}$ $^{(4)}$	$80.49_{\pm 0.5}$/$72.46_{\pm 0.6}$	$60.16_{\pm 22.4}$/$49.78_{\pm 23.1}$ $^{(4)}$	$38.19_{\pm 31.2}$/$30.24_{\pm 26.5}$ $^{(3)}$	$38.14_{\pm 31.2}$/$30.05_{\pm 26.1}$ $^{(3)}$	$80.49_{\pm 0.5}$/$72.46_{\pm 0.6}$	38.14/30.05
FLTrust	$71.34_{\pm 2.7}$/$62.53_{\pm 3.5}$	$71.02_{\pm 2.8}$/$62.29_{\pm 3.7}$	$71.46_{\pm 2.6}$/$62.52_{\pm 3.6}$	$72.17_{\pm 1.3}$/$63.44_{\pm 2.1}$	$75.48_{\pm 0.7}$/$65.71_{\pm 1.0}$	$63.45_{\pm 4.4}$/$52.76_{\pm 5.1}$	$63.88_{\pm 5.2}$/$53.36_{\pm 5.8}$	$68.45_{\pm 7.3}$/$59.25_{\pm 8.1}$	$61.21_{\pm 5.4}$/$50.22_{\pm 5.9}$
BOBA (cls.)	$80.34_{\pm 0.5}$/$72.19_{\pm 1.0}$	$73.19_{\pm 1.1}$/$64.32_{\pm 1.3}$	-	$80.45_{\pm 0.5}$/$72.31_{\pm 0.9}$	$78.91_{\pm 0.5}$/$69.21_{\pm 0.5}$	$72.64_{\pm 2.9}$/$63.86_{\pm 9.1}$	$75.60_{\pm 1.7}$/$66.98_{\pm 2.4}$	$79.59_{\pm 0.6}$/$70.85_{\pm 1.3}$	-
FedAA	$78.15_{\pm 0.6}$/$68.99_{\pm 0.6}$	$47.32_{\pm 28.0}$/$38.31_{\pm 25.1}$ $^{(4)}$	$71.70_{\pm 5.9}$/$61.89_{\pm 8.0}$ $^{(2)}$	$78.26_{\pm 0.6}$/$69.34_{\pm 0.4}$	$71.43_{\pm 4.0}$/$60.91_{\pm 4.8}$	$75.35_{\pm 4.4}$/$66.42_{\pm 5.1}$	$60.20_{\pm 31.9}$/$52.65_{\pm 29.3}$	$78.39_{\pm 0.6}$/$69.18_{\pm 1.0}$	$30.43_{\pm 38.4}$/$23.53_{\pm 32.5}$ $^{(2)}$
Ditto	$80.47_{\pm 0.6}$/$65.80_{\pm 1.5}$	$72.16_{\pm 1.9}$/$42.02_{\pm 2.5}$	-	$2.13_{\pm 0.0}$/$0.00_{\pm 0.0}$ $^{(1)}$	$78.07_{\pm 0.8}$/$61.66_{\pm 0.7}$	$61.34_{\pm 1.4}$/$53.18_{\pm 2.6}$	$62.93_{\pm 0.8}$/$54.56_{\pm 1.0}$	$82.58_{\pm 0.8}$/$74.83_{\pm 0.5}$	-
FUSE	$79.12_{\pm 2.1}$/$71.04_{\pm 2.1}$	$79.51_{\pm 0.8}$/$71.52_{\pm 1.2}$	$79.41_{\pm 0.5}$/$71.30_{\pm 0.9}$	$79.67_{\pm 1.0}$/$71.63_{\pm 1.1}$	$79.61_{\pm 1.0}$/$71.67_{\pm 1.2}$	$78.89_{\pm 2.6}$/$70.77_{\pm 2.8}$	$79.11_{\pm 2.1}$/$71.05_{\pm 2.1}$	$79.12_{\pm 2.1}$/$71.04_{\pm 2.1}$	$78.31_{\pm 2.3}$/$70.00_{\pm 2.4}$

B.7 THREE-COMPONENT ABLATION PROTOCOL

The factorial ablation traverses all 2^3 combinations of the scale screen, cone screen, and loss-normalized aggregation. All four dataset/model pairs use 1000 communication rounds under matched client partitions and attack realizations. The complete 2^3 factorial ablation for all four dataset/model pairs is reported in Appendix E.1; the main text shows the CIFAR-10 instance as a representative mechanism study. Auxiliary runs that reached round 299 were continued from their matched checkpoints. Clean trajectories are reused across selector-component rows within an aggregation mode because clean runs use $f = 0$ and therefore perform no screening rejection.

After this reuse, 58 CIFAR-10, 53 EMNIST-Balanced, and 60 AGNews conceptual conditions reach their rounds 990–999 reporting windows. A missing terminal window is retained as "–" rather than replaced by a pre-collapse maximum; all such dashes correspond to a registered algorithmic collapse, not an unfinished scheduler job.

C FIVE-SEED RESULTS AND TRAINING DYNAMICS

C.1 FULL FIVE-SEED ENDPOINT RESULTS

Tables 7–10 report all ten baselines and FUSE over seeds 40–44. Each attacked column first averages the final reporting window within a seed and then averages completed seeds. The last column instead computes the minimum across all seven attacks within each seed before averaging; it is therefore the seed-wise worst-attack statistic used throughout the main paper. Superscript (k) indicates that only k seeds reached the endpoint. The BOBA rows use classifier-coordinate geometry (Appendix B.5); full-space BOBA is available for seed 42 only. The AGNews FUSE rows screen all GRU parameters.

Table 11 mirrors the main-paper robustness table using worst-10% client accuracy instead of global accuracy. Table 12 then reports paired seed-wise comparisons for the global worst-attack metric.

Table 10: AGNews / GRU, rounds 1990–1999. Entries are global / worst-10% client accuracy (%), mean over seeds 40, 41, 42, 43, 44 with the across-seed standard deviation as subscript; the worst-attack column is the seed-wise minimum over the seven attacks, also mean with std subscript, over seeds completing all seven attacks.

Method	Clean	Sign	$\text{IPM}_{\lambda=3}$	Gauss.	ALIE	Min-Max	Min-Sum	Magnitude	Worst attack
q-FFL-GD (mean)	$81.52_{\pm 3.0}/68.95_{\pm 2.7}$	$64.00_{\pm 2.0}/49.40_{\pm 4.7}$	-	$27.06_{\pm 0.7}/16.92_{\pm 1.0}$ [4]	$73.23_{\pm 9.0}/59.00_{\pm 8.9}$	$40.49_{\pm 15.2}/23.90_{\pm 11.6}$	$45.40_{\pm 18.0}/29.31_{\pm 16.8}$ [4]	$76.35_{\pm 22.8}/64.38_{\pm 26.1}$	-
H-nobs	$81.52_{\pm 3.0}/68.95_{\pm 2.7}$	$26.69_{\pm 1.4}/0.56_{\pm 0.9}$ [3]	$81.62_{\pm 2.4}/67.99_{\pm 3.4}$	$81.52_{\pm 3.0}/68.95_{\pm 2.7}$	$77.53_{\pm 0.5}/62.83_{\pm 2.2}$	$72.91_{\pm 13.2}/58.16_{\pm 13.7}$	$73.71_{\pm 1.9}/58.29_{\pm 2.0}$	$81.52_{\pm 3.0}/68.95_{\pm 2.7}$	$26.69_{\pm 1.4}/0.56_{\pm 0.9}$ [3]
Trimmed Mean	$81.57_{\pm 2.9}/69.06_{\pm 2.5}$	$36.61_{\pm 2.9}/8.79_{\pm 6.4}$	$35.88_{\pm 2.7}/6.90_{\pm 6.6}$	$79.67_{\pm 5.9}/66.55_{\pm 7.0}$	$74.76_{\pm 0.9}/59.20_{\pm 2.3}$	$38.63_{\pm 13.4}/19.84_{\pm 9.6}$	$39.37_{\pm 13.5}/22.11_{\pm 8.3}$	$79.89_{\pm 3.0}/67.29_{\pm 2.2}$	$30.85_{\pm 2.9}/5.71_{\pm 4.7}$
Median	$66.83_{\pm 4.0}/46.50_{\pm 7.0}$	$40.53_{\pm 3.8}/15.32_{\pm 5.6}$	$37.58_{\pm 4.7}/12.00_{\pm 9.2}$	$72.53_{\pm 1.6}/54.11_{\pm 3.7}$	$70.64_{\pm 1.3}/55.17_{\pm 1.7}$	$49.33_{\pm 11.0}/30.59_{\pm 9.1}$	$49.01_{\pm 10.7}/29.42_{\pm 9.9}$	$64.82_{\pm 18.3}/48.88_{\pm 17.5}$	$33.50_{\pm 3.8}/9.20_{\pm 4.8}$
Krum	$72.69_{\pm 10.1}/56.90_{\pm 10.1}$	$77.17_{\pm 2.6}/60.86_{\pm 5.0}$	$76.62_{\pm 6.7}/58.96_{\pm 10.6}$	$72.69_{\pm 10.1}/56.90_{\pm 10.1}$	$72.18_{\pm 1.6}/56.64_{\pm 0.7}$	$77.42_{\pm 5.2}/62.50_{\pm 6.7}$	$74.77_{\pm 13.3}/59.40_{\pm 16.2}$	$72.69_{\pm 10.1}/56.90_{\pm 10.1}$	$64.43_{\pm 9.9}/48.20_{\pm 12.1}$
Multi-Krum	$76.02_{\pm 9.0}/62.24_{\pm 10.6}$	$66.96_{\pm 11.5}/46.28_{\pm 12.4}$	$73.08_{\pm 13.2}/56.59_{\pm 17.9}$	$76.02_{\pm 9.0}/62.24_{\pm 10.6}$	$74.87_{\pm 1.1}/57.84_{\pm 2.0}$	$78.23_{\pm 4.6}/64.39_{\pm 3.1}$	$66.27_{\pm 5.3}/49.22_{\pm 5.5}$	$76.02_{\pm 9.0}/62.24_{\pm 10.6}$	$58.34_{\pm 5.5}/38.31_{\pm 8.0}$
FLTrust	$83.92_{\pm 1.1}/70.63_{\pm 2.6}$	$83.92_{\pm 1.1}/70.63_{\pm 2.6}$	$83.92_{\pm 1.1}/70.63_{\pm 2.6}$	$83.98_{\pm 1.0}/70.79_{\pm 2.4}$	$77.19_{\pm 1.0}/62.57_{\pm 2.5}$	$66.82_{\pm 3.4}/49.24_{\pm 4.8}$	$64.26_{\pm 3.6}/45.46_{\pm 4.8}$	$83.03_{\pm 0.7}/69.91_{\pm 1.8}$	$64.26_{\pm 3.6}/45.46_{\pm 4.8}$
BOBA (cls.)	$83.82_{\pm 1.0}/70.71_{\pm 2.1}$	$60.57_{\pm 0.8}/44.07_{\pm 2.9}$	$27.65_{\pm 1.4}/1.65_{\pm 1.7}$	$83.82_{\pm 1.0}/70.71_{\pm 2.1}$	$80.82_{\pm 2.0}/66.48_{\pm 4.5}$	$56.44_{\pm 1.1}/38.29_{\pm 2.5}$	$59.43_{\pm 1.5}/41.98_{\pm 2.4}$	$83.53_{\pm 1.5}/69.52_{\pm 3.3}$	$27.65_{\pm 1.4}/1.65_{\pm 1.7}$
FedAA	$78.36_{\pm 1.4}/61.25_{\pm 3.1}$	$38.32_{\pm 12.0}/11.62_{\pm 12.9}$	$57.68_{\pm 22.8}/38.52_{\pm 30.1}$ [2]	$77.83_{\pm 3.0}/61.20_{\pm 6.5}$	$75.29_{\pm 4.7}/60.29_{\pm 6.7}$	$41.18_{\pm 22.2}/16.55_{\pm 26.3}$	$40.34_{\pm 11.4}/15.72_{\pm 13.2}$	$79.67_{\pm 1.4}/64.88_{\pm 3.9}$	$31.29_{\pm 0.8}/4.44_{\pm 0.7}$ [2]
Ditto	$83.82_{\pm 1.0}/62.56_{\pm 4.9}$	$63.17_{\pm 0.9}/38.63_{\pm 2.6}$	-	$25.05_{\pm 0.8}/16.74_{\pm 1.8}$ [3]	$80.94_{\pm 2.4}/62.93_{\pm 3.1}$	$39.97_{\pm 5.1}/29.18_{\pm 4.0}$	$56.25_{\pm 2.9}/39.36_{\pm 3.2}$	$86.53_{\pm 0.6}/77.52_{\pm 2.5}$	-
FUSE	$82.29_{\pm 1.5}/69.58_{\pm 2.1}$	$80.46_{\pm 1.0}/66.94_{\pm 2.8}$ [4]	$81.12_{\pm 1.0}/67.92_{\pm 1.8}$	$82.17_{\pm 1.1}/69.43_{\pm 2.5}$	$78.62_{\pm 4.1}/65.58_{\pm 6.4}$	$81.44_{\pm 1.7}/68.25_{\pm 3.1}$	$81.36_{\pm 1.5}/68.22_{\pm 3.0}$	$82.29_{\pm 1.5}/69.58_{\pm 2.1}$	$77.84_{\pm 4.4}/63.75_{\pm 6.6}$ [4]

Paired interpretation. Against FLTrust, FUSE improves the seed-wise global worst-attack accuracy by 30.07 points on CIFAR-10, 3.90 on MNIST, and 17.09 on EMNIST; the paired differences are positive on all five seeds and their bootstrap 95% intervals exclude zero. The exact two-sided sign-flip test is nevertheless limited to $p = 0.0625$ with five pairs. On AGNews the floor is 77.84 over the four seeds completing all attacks, 13.03 points above FLTrust ([+11.84, +14.22]) and 15.40 above Krum ([+4.62, +26.18]) on the same seeds; with four pairs the sign-flip test is limited to $p = 0.125$.

C.2 Training dynamics

Figures 5–8 show the seed-42 worst-10% client-accuracy trajectories for every dataset/model pair and all seven attacks, together with the clean condition. We retain three representative baselines spanning fair optimization, norm screening, and trusted-gradient aggregation. "Ours" denotes FUSE. A cross marks a trajectory that terminates before the registered training horizon.

Table 11: **Worst-10% client robustness across datasets and attacks.** Mean worst-10% client accuracy (W10, %) over seeds reaching the registered endpoint. Worst attack is the seed-wise minimum over seven attacks; $^{(k)}$ denotes k completed seeds.

Dataset/model	Method	Clean	Sign	IPM	Gauss.	ALIE	Min-Max	Min-Sum	Mag.	Worst attack
CIFAR-10 ResNet-18	q-FFL	67.34	66.94	–	4.71$^{(3)}$	51.90	2.24$^{(4)}$	2.04	32.27$^{(4)}$	–
	H-nobs	67.34	0.35$^{(3)}$	46.25	70.03	23.39	31.99	8.34	67.34	0.35$^{(3)}$
	Trim.	60.32	46.72	31.29	70.42	25.86	1.31	2.08	64.41	1.28
	Krum	51.97	29.35	51.13	48.53	10.14	50.09	32.93	51.66	9.78
	FLTrust	72.22	**72.26**	**72.27**	72.16	42.07	35.47	30.53	70.05	30.53
	BOBA (cls.)	73.13	62.12	0.66	72.73	60.85	5.27	5.35	62.39	0.66
	FedAA	64.87	32.48	62.88$^{(4)}$	61.23	49.78	10.34	30.00	53.75	10.27$^{(4)}$
	FUSE	**73.88**	70.15	71.62	**73.91**	**68.35**	**73.91**	**73.87**	**72.67**	**63.58**
MNIST MLP	q-FFL	90.72	88.31	–	54.32$^{(3)}$	89.66	37.24$^{(3)}$	46.60$^{(3)}$	63.08$^{(3)}$	–
	H-nobs	90.72	1.91$^{(3)}$	84.66	90.72	88.26	**90.59**	56.83$^{(3)}$	90.72	1.91$^{(3)}$
	Trim.	90.71	75.80	78.31$^{(4)}$	90.46	59.07$^{(3)}$	37.79$^{(3)}$	37.52$^{(3)}$	91.29	37.25$^{(3)}$
	Krum	89.24	83.30	**88.42**	89.24	56.77$^{(3)}$	89.24	88.61	89.24	56.14$^{(3)}$
	FLTrust	90.53	88.05	88.24	90.54	87.16	82.26	85.34	90.76	79.64
	BOBA (cls.)	92.68	**89.83**	0.16	**92.68**	91.40	70.92	72.58	**93.00**	0.16
	FedAA	**92.71**	39.47	77.56	92.31	**92.39**	62.88	47.28	92.41	31.26
	FUSE	89.16	88.30	87.07	89.16	89.17	89.14	**89.15**	89.19	**87.07**
EMNIST-Bal. CNN	q-FFL	57.08	66.23	–	0.00$^{(1)}$	46.71$^{(3)}$	36.43$^{(3)}$	21.24$^{(3)}$	58.92	–
	H-nobs	57.08	21.64$^{(2)}$	70.66	57.08	69.11$^{(3)}$	45.63$^{(3)}$	43.19$^{(3)}$	57.08	21.64$^{(2)}$
	Trim.	**72.43**	52.69	44.44$^{(4)}$	56.28	57.29$^{(4)}$	17.89$^{(3)}$	26.51$^{(3)}$	56.61	26.83$^{(2)}$
	Krum	70.12	1.24$^{(2)}$	70.15	70.12	60.74$^{(3)}$	39.93$^{(3)}$	1.02$^{(3)}$	70.12	0.97$^{(2)}$
	FLTrust	62.53	62.29	62.52	63.44	65.71	52.76	53.36	59.25	50.22
	BOBA (cls.)	72.19	64.32	–	**72.31**	69.21	63.86	66.98	70.85	–
	FedAA	68.99	38.31$^{(4)}$	61.89$^{(2)}$	69.34	60.91	66.42	52.65	69.18	23.53$^{(2)}$
	FUSE	71.04	**71.52**	**71.30**	71.63	**71.67**	**70.77**	**71.05**	**71.04**	**70.00**
AGNews GRU	q-FFL	68.95	49.40	–	16.92$^{(4)}$	59.00	23.90	29.31$^{(4)}$	64.38	–
	H-nobs	68.95	0.56$^{(3)}$	67.99	68.95	62.83	58.16	58.29	68.95	0.56$^{(3)}$
	Trim.	69.06	8.79	6.90	66.55	59.20	19.84	22.11	67.29	5.71
	Krum	56.90	60.86	58.96	56.90	56.64	62.50	59.40	56.90	48.20
	FLTrust	70.63	**70.63**	**70.63**	**70.79**	62.57	49.24	45.46	**69.91**	45.46
	BOBA (cls.)	**70.71**	44.07	1.65	70.71	**66.48**	38.29	41.98	69.52	1.65
	FedAA	61.25	11.62	38.52$^{(2)}$	61.20	60.29	16.55	15.72	64.88	4.44$^{(2)}$
	FUSE	69.58	66.94$^{(4)}$	67.92	69.43	65.58	**68.25**	**68.22**	69.58	**63.75**$^{(4)}$

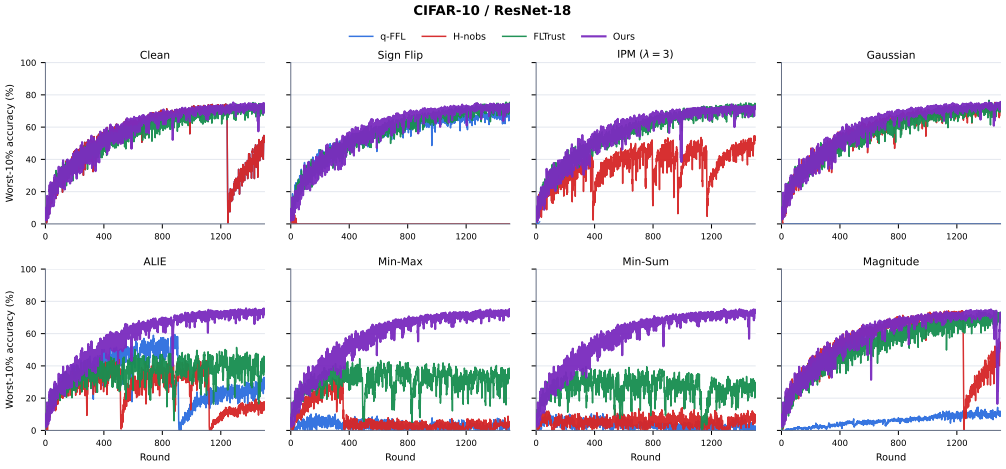


Figure 5: **CIFAR-10/ResNet-18 training dynamics.** Worst-10% client accuracy (%) over 1500 communication rounds.

Table 12: Seed-wise worst-attack accuracy of FUSE minus each baseline (global accuracy, percentage points). Worst-attack values are computed per seed as the minimum over the seven attacks; seeds in which either method collapsed under any attack are excluded from that pair (n is the number of paired seeds). CI: 95% percentile bootstrap over paired seeds; p: exact two-sided sign-flip test.

Dataset	Baseline	FUSE worst attack	Baseline worst attack	Δ [95% CI]	p (n)
CIFAR-10 / ResNet-18	q-FFL-GD (mean)	73.88	-	-	-
CIFAR-10 / ResNet-18	H-nobs	73.88	12.19	+63.89 [+56.93, +67.83]	0.250 (3)
CIFAR-10 / ResNet-18	Trimmed Mean	73.88	13.18	+60.70 [+52.08, +67.62]	0.062 (5)
CIFAR-10 / ResNet-18	Median	73.88	12.34	+61.53 [+53.51, +68.25]	0.062 (5)
CIFAR-10 / ResNet-18	Krum	73.88	21.67	+52.21 [+43.96, +61.29]	0.062 (5)
CIFAR-10 / ResNet-18	Multi-Krum	73.88	12.97	+63.12 [+56.20, +66.92]	0.250 (3)
CIFAR-10 / ResNet-18	FLTrust	73.88	43.81	+30.07 [+20.24, +37.83]	0.062 (5)
CIFAR-10 / ResNet-18	BOBA (cls.)	73.88	10.93	+62.95 [+55.02, +69.61]	0.062 (5)
CIFAR-10 / ResNet-18	FedAA	73.88	22.15	+55.31 [+45.62, +64.99]	0.125 (4)
CIFAR-10 / ResNet-18	Ditto	73.88	-	-	-
MNIST / MLP	q-FFL-GD (mean)	92.42	-	-	-
MNIST / MLP	H-nobs	92.42	14.82	+77.42 [+73.28, +81.00]	0.250 (3)
MNIST / MLP	Trimmed Mean	92.42	48.24	+44.00 [+24.25, +83.43]	0.250 (3)
MNIST / MLP	Median	92.42	65.88	+26.31 [+26.27, +26.34]	0.500 (2)
MNIST / MLP	Krum	92.42	63.99	+28.25 [+0.76, +82.26]	0.250 (3)
MNIST / MLP	Multi-Krum	92.42	59.90	+32.34 [+6.50, +82.55]	0.250 (3)
MNIST / MLP	FLTrust	92.42	88.52	+3.90 [+1.56, +6.25]	0.062 (5)
MNIST / MLP	BOBA (cls.)	92.42	11.23	+81.19 [+80.73, +81.75]	0.062 (5)
MNIST / MLP	FedAA	92.42	43.00	+49.42 [+24.50, +74.34]	0.062 (5)
MNIST / MLP	Ditto	92.42	-	-	-
EMNIST-Balanced / CNN	q-FFL-GD (mean)	78.31	-	-	-
EMNIST-Balanced / CNN	H-nobs	78.31	28.70	+50.66 [+24.77, +76.55]	0.500 (2)
EMNIST-Balanced / CNN	Trimmed Mean	78.31	37.11	+42.33 [+40.47, +44.19]	0.500 (2)
EMNIST-Balanced / CNN	Median	78.31	15.72	+63.46 [+37.12, +76.72]	0.250 (3)
EMNIST-Balanced / CNN	Krum	78.31	5.12	+73.64 [+70.74, +76.55]	0.500 (2)
EMNIST-Balanced / CNN	Multi-Krum	78.31	-	-	-
EMNIST-Balanced / CNN	FLTrust	78.31	61.21	+17.09 [+12.33, +21.89]	0.062 (5)
EMNIST-Balanced / CNN	BOBA (cls.)	78.31	-	-	-
EMNIST-Balanced / CNN	FedAA	78.31	30.43	+49.01 [+21.26, +76.76]	0.500 (2)
EMNIST-Balanced / CNN	Ditto	78.31	-	-	-
AGNews / GRU	q-FFL-GD (mean)	77.84	-	-	-
AGNews / GRU	H-nobs	77.84	26.69	+49.48 [+45.11, +53.86]	0.500 (2)
AGNews / GRU	Trimmed Mean	77.84	30.85	+47.96 [+41.57, +51.37]	0.125 (4)
AGNews / GRU	Median	77.84	33.50	+43.97 [+38.72, +49.22]	0.125 (4)
AGNews / GRU	Krum	77.84	64.43	+15.40 [+4.62, +26.18]	0.250 (4)
AGNews / GRU	Multi-Krum	77.84	58.34	+21.16 [+14.84, +27.08]	0.125 (4)
AGNews / GRU	FLTrust	77.84	64.26	+13.03 [+11.84, +14.22]	0.125 (4)
AGNews / GRU	BOBA (cls.)	77.84	27.65	+49.87 [+45.74, +52.88]	0.125 (4)
AGNews / GRU	FedAA	77.84	31.29	+49.14 [+49.14, +49.14]	1.000 (1)
AGNews / GRU	Ditto	77.84	-	-	-

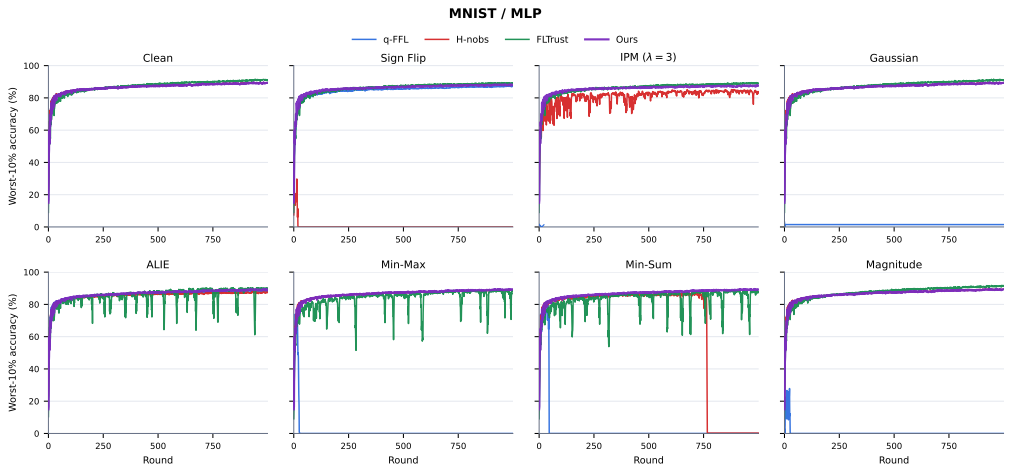


Figure 6: **MNIST/MLP training dynamics.** Worst-10% client accuracy (%) over 1000 communication rounds.

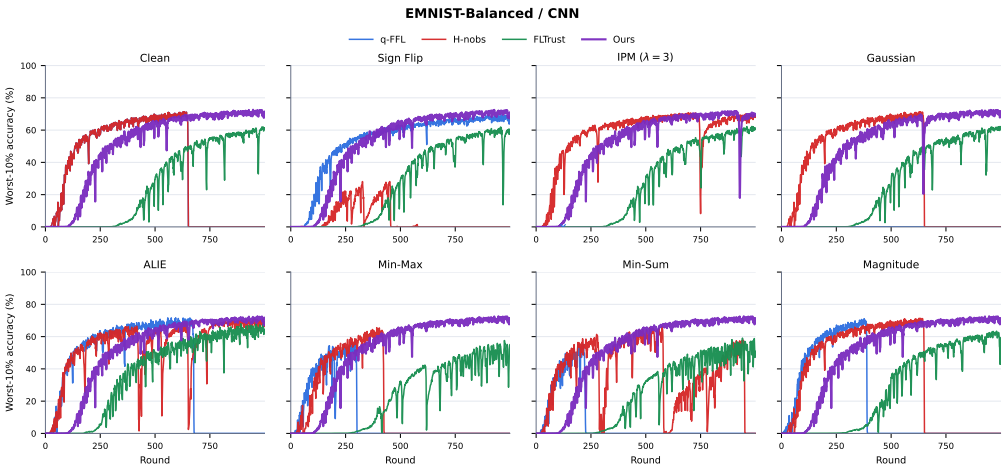


Figure 7: **EMNIST-Balanced/CNN training dynamics.** Worst-10% client accuracy (%) over 1000 communication rounds.

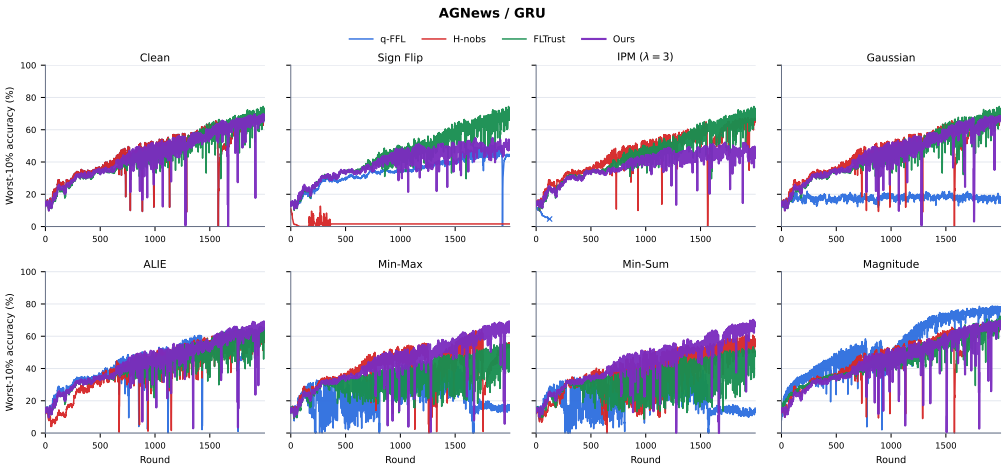


Figure 8: **AGNews/GRU training dynamics.** Worst-10% client accuracy (%) over 2000 communication rounds.

D EXPERIMENTAL VALIDATION OF THE GEOMETRIC ANALYSIS

We validate the two predictions underlying Theorems 1 and 2: coefficient mass should track the fairness scale, and scale and cone residual should respond to complementary deviations. The diagnostic uses MNIST, seed 42, and $q = 1$; complete stage-wise statistics are in Appendix E.

The recovered mass has Pearson correlation $r = 0.940$ with F_i^q; the trusted loss estimate has Pearson $r = 0.923$ and MAE 0.053 over 76,457 retained honest submissions. Gaussian and Magnitude Scaling primarily produce large mass, whereas ALIE, Min-Max, and Min-Sum primarily produce large cone residual. Sign Flip and IPM overlap more strongly with the honest region, which illustrates the observable boundary of the update-only interface.

Finally, the controlled simplex–cone study in Appendix E.3 holds the checkpoint, honest updates, and trusted basis fixed while changing only the coefficient constraint. It shows why BOBA's unit-sum simplex is mismatched after fairness scaling: the cone is invariant to the legitimate radial transformation, whereas a fixed simplex is not.

E ABLATIONS AND DIAGNOSTIC EVIDENCE

This appendix groups mechanism attribution, fairness audits, sensitivity controls, trusted-geometry diagnostics, computational cost, and explicit robustness boundaries. Unless otherwise stated, diagnostic statistics use seed 42 and are not presented as cross-seed uncertainty estimates.

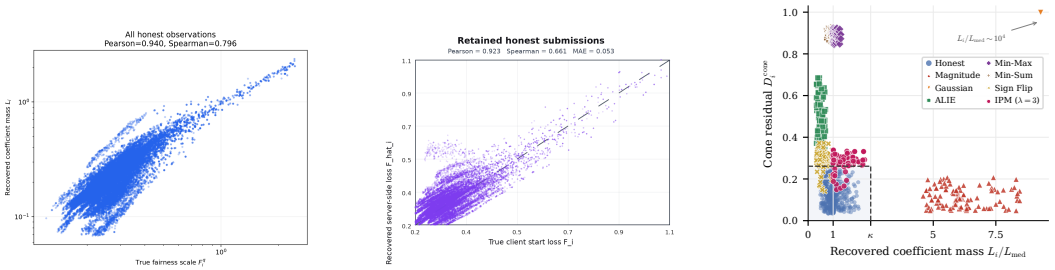
E.1 THREE-COMPONENT FACTORIAL ABLATION

The complete factorial ablation is reported here for all dataset/model pairs. Tables 13–16 use the identical 1000-round design on every dataset (including AGNews, whose main-table horizon is 2000 rounds) and report all seven attacks. The three binary components are the coefficient-scale screen, normalized cone screen, and loss-normalized aggregation. Runs that reached round 299 in the original factorial were continued from their matched checkpoints; configurations that had already met a registered collapse criterion retain a missing terminal window. Because clean training uses $f = 0$, its trajectory depends on the aggregation rule but not on the enabled screening stages, so the matched clean trajectory is reused within each aggregation mode.

Role of cone screening. Cone screening has the clearest structural-robustness effect. Holding the scale screen and loss-normalized aggregation fixed, removing the cone screen eliminates the IPM terminal window on CIFAR-10, MNIST, EMNIST-Balanced, and AGNews, whereas the complete FUSE configuration reaches the registered endpoint on all four datasets. The same pattern holds when both screens are removed. This supports the intended role of the cone stage against invalid signed mixtures and other structural violations, rather than a claim that it uniformly increases clean or per-attack endpoint accuracy.

Role of scale screening. With cone screening and loss normalization fixed, adding the scale stage changes the worst-attack result from 52.38%/40.25% to 77.50%/68.37% on CIFAR-10, from 91.83%/86.29% to 92.35%/87.01% on MNIST, from 79.21%/69.89% to 79.30%/69.89% on EMNIST-Balanced, and from 53.27%/36.15% to 53.34%/38.33% on AGNews. Thus the scale stage never lowers the paired worst-attack result and is decisive on CIFAR-10, although the small differences on the other datasets should not be over-interpreted from a single seed. Its role is also visible without loss normalization: it prevents the severe CIFAR-10 Magnitude failure and the MNIST cone-only Magnitude collapse. Conversely, removing the scale stage improves the Magnitude endpoint on EMNIST and AGNews. The evidence therefore supports scale screening as protection against radial worst cases, not as a component that uniformly raises every attack endpoint.

Role of loss normalization. Loss normalization is an optimization and stability component with a dataset-dependent effect. With both screens fixed, it raises the complete CIFAR-10 worst-attack result from 70.06%/59.88% to 77.50%/68.37%. On EMNIST-Balanced, the mean-u Sign-Flip and IPM runs have no terminal window, while the loss-normalized configuration completes all seven attacks. The MNIST worst-attack results are essentially unchanged (92.34%/87.06% for mean-u



(a) L_i versus F_i^q. (b) $\widetilde{F}_i$ versus F_i. (c) Scale versus cone score.

Figure 9: **Experimental validation of the geometric analysis.** Recovered mass tracks the fairness scale (a), normalized coefficients support loss reconstruction (b), and the two screening statistics separate complementary attack patterns (c).

Table 13: **Component ablation of FUSE.** CIFAR-10/ResNet-18, seed 42, 1000 rounds. Global / worst-10% client accuracy (%) over the final ten rounds; the worst-attack column is the componentwise minimum over the seven attacks. Clean uses $f = 0$; "–" marks a run with no terminal window.

Variant	Scale	Cone	Loss norm.	Clean	Magnitude	Sign	IPM$_{\lambda=3}$	Gauss.	ALIE	Min-Max	Min-Sum	Worst attack
FUSE	✓	✓	✓	**79.03/70.41**	**78.97/70.38**	77.50/68.37	**77.83/68.91**	**78.82/69.57**	78.50/69.28	**79.28/70.73**	78.71/69.93	**77.50/68.37**
w/o scale	–	✓	✓	**79.03/70.41**	52.38/40.25	**78.20/69.08**	76.08/66.91	**78.82/69.57**	**78.97/70.16**	79.00/70.29	**79.03/70.41**	52.38/40.25
w/o cone	✓	–	✓	**79.03/70.41**	78.67/70.09	69.95/59.40	–	**78.82/69.57**	55.16/42.83	36.71/24.49	38.94/27.18	–
w/o scale and cone	–	–	✓	**79.03/70.41**	63.23/51.72	70.16/59.92	–	26.11/13.69	63.20/50.88	34.07/22.76	35.66/23.99	–
w/o loss norm.	✓	✓	–	78.04/69.52	70.06/59.88	73.33/63.97	74.01/64.70	77.89/68.61	76.98/67.90	76.04/67.28	76.02/67.30	70.06/59.88
cone only	–	✓	–	78.04/69.52	10.89/0.05	71.59/62.06	75.30/65.97	77.65/69.14	76.87/68.40	78.04/69.52	78.04/69.52	10.89/0.05
scale only	✓	–	–	78.04/69.52	73.14/64.22	71.66/61.40	–	77.74/69.17	46.45/34.25	13.87/2.92	–	–
no component	–	–	–	78.04/69.52	16.91/5.99	72.56/63.11	–	–	48.22/36.59	15.41/5.06	15.04/3.21	–

Table 14: **Component ablation of FUSE.** MNIST/MLP, seed 42, 1000 rounds. Global / worst-10% client accuracy (%) over the final ten rounds; the worst-attack column is the componentwise minimum over the seven attacks. Clean uses $f = 0$; "–" marks a run with no terminal window.

Variant	Scale	Cone	Loss norm.	Clean	Magnitude	Sign	IPM$_{\lambda=3}$	Gauss.	ALIE	Min-Max	Min-Sum	Worst attack
FUSE	✓	✓	✓	93.95/89.26	93.94/89.23	93.20/88.09	**92.35/87.01**	93.95/89.26	93.94/89.14	93.94/89.23	93.96/89.26	**92.35/87.01**
w/o scale	–	✓	✓	93.95/89.26	94.40/89.91	93.22/88.23	91.83/86.29	93.95/89.26	93.95/89.27	93.95/89.26	93.95/89.26	91.83/86.29
w/o cone	✓	–	✓	93.95/89.26	93.96/89.27	92.43/87.07	–	93.95/89.26	93.97/89.38	89.06/83.15	55.42/41.64	–
w/o scale and cone	–	–	✓	93.95/89.26	**95.14/89.63**	92.39/87.11	–	52.87/39.71	93.96/89.21	89.78/84.07	79.52/69.42	–
w/o loss norm.	✓	✓	–	**93.97/89.29**	93.98/89.26	93.24/88.15	92.34/87.06	**93.97/89.29**	93.96/89.14	**93.98/89.29**	**93.98/89.29**	92.34/87.06
cone only	–	✓	–	**93.97/89.29**	–	**93.27/88.25**	92.09/86.73	**93.97/89.29**	**93.98/89.29**	93.97/89.29	93.97/89.29	–
scale only	✓	–	–	**93.97/89.29**	93.98/89.27	92.45/87.13	10.33/0.00	**93.97/89.29**	93.97/89.37	87.78/81.05	78.96/69.45	10.33/0.00
no component	–	–	–	**93.97/89.29**	–	92.40/87.15	–	–	93.96/89.23	51.72/34.59	53.41/40.21	–

Table 15: **Component ablation of FUSE.** EMNIST-Balanced/CNN, seed 42, 1000 rounds. Global / worst-10% client accuracy (%) over the final ten rounds; the worst-attack column is the componentwise minimum over the seven attacks. Clean uses $f = 0$; "–" marks a run with no terminal window.

Variant	Scale	Cone	Loss norm.	Clean	Magnitude	Sign	IPM$_{\lambda=3}$	Gauss.	ALIE	Min-Max	Min-Sum	Worst attack
FUSE	✓	✓	✓	79.59/70.34	79.59/70.34	**79.31/70.05**	**79.51/70.33**	**79.30/69.89**	**79.70/70.65**	79.59/70.34	79.59/70.34	**79.30/69.89**
w/o scale	–	✓	✓	79.59/70.34	**82.69/74.52**	**79.31/70.05**	79.21/70.12	**79.30/69.89**	**79.70/70.65**	79.59/70.34	79.59/70.34	79.21/69.89
w/o cone	✓	–	✓	79.59/70.34	79.59/70.34	71.50/62.16	–	**79.30/69.89**	77.99/68.66	58.36/46.70	59.18/47.88	–
w/o scale and cone	–	–	✓	79.59/70.34	82.47/73.37	71.50/62.16	–	–	77.99/68.66	58.36/46.70	59.18/47.88	–
w/o loss norm.	✓	✓	–	**80.34/71.73**	80.34/71.73	–	–	20.01/11.20	58.01/48.11	**80.34/71.73**	**80.34/71.73**	–
cone only	–	✓	–	**80.34/71.73**	–	–	–	20.01/11.20	58.01/48.11	**80.34/71.73**	**80.34/71.73**	–
scale only	✓	–	–	**80.34/71.73**	80.34/71.73	76.71/67.50	–	20.01/11.20	78.95/70.21	50.48/38.00	52.31/40.19	–
no component	–	–	–	**80.34/71.73**	78.21/68.79	76.71/67.50	–	–	78.95/70.21	50.48/38.00	52.31/40.19	–

and 92.35%/87.01% with loss normalization). On AGNews, mean-u instead has the higher worst-attack result, 55.45%/40.77% versus 53.34%/38.33%. Hence the factorial supports a long-horizon stabilization benefit on CIFAR-10 and EMNIST, a neutral effect on MNIST, and a trade-off on AGNews; it does not support a claim of uniform endpoint improvement from loss normalization.

Of the 64 conceptual conditions per auxiliary dataset, 58 CIFAR-10, 53 EMNIST-Balanced, and 60 AGNews conditions reach rounds 990–999 after clean trajectory reuse. Every remaining dash denotes a registered collapse rather than an unfinished scheduler job. In particular, two continued CIFAR-10 configurations and five continued EMNIST mean-u configurations collapse after round 299; all launched AGNews continuations reach the terminal window.

Table 16: **Component ablation of FUSE.** AGNews/GRU, seed 42, 2000 rounds. Global / worst-10% client accuracy (%) over the final ten rounds; the worst-attack column is the componentwise minimum over the seven attacks. Clean uses $f = 0$; "–" marks a run with no terminal window.

Variant	Scale	Cone	Loss norm.	Clean	Magnitude	Sign	IPM$_{\lambda=3}$	Gauss.	ALIE	Min-Max	Min-Sum	Worst attack
FUSE	✓	✓	✓	**83.66/68.09**	83.65/68.09	79.45/62.86	81.56/65.49	83.35/67.62	**81.31/68.09**	83.14/67.13	82.94/66.88	79.45/62.86
w/o scale	–	✓	✓	83.64/68.03	86.53/73.99	**83.64/68.03**	**83.64/68.03**	**83.65/68.06**	80.44/66.99	**83.64/68.03**	83.70/68.16	**80.44/66.99**
w/o cone	✓	–	✓	**83.66/68.09**	83.66/68.09	–	–	28.34/17.87	80.53/67.12	40.94/16.66	55.80/37.26	–
w/o scale and cone	–	–	✓	**83.66/68.09**	86.90/75.05	–	–	28.41/16.32	81.26/67.35	43.55/21.00	56.08/36.75	–
w/o loss norm.	✓	✓	–	81.93/66.28	81.93/66.28	62.69/46.97	68.08/50.74	35.92/25.22	71.58/58.13	76.38/62.02	**84.18/68.95**	35.92/25.22
cone only	–	✓	–	81.93/66.28	67.41/46.86	81.93/66.28	81.93/66.28	81.93/66.33	77.01/63.68	81.93/66.28	79.36/62.37	67.41/46.86
scale only	✓	–	–	81.93/66.28	81.93/66.28	65.85/46.21	–	24.38/14.31	76.93/62.97	25.72/3.36	48.17/23.84	–
no component	–	–	–	81.92/66.33	**87.14/76.64**	61.64/43.98	–	25.16/14.26	78.51/64.55	–	49.19/25.77	–

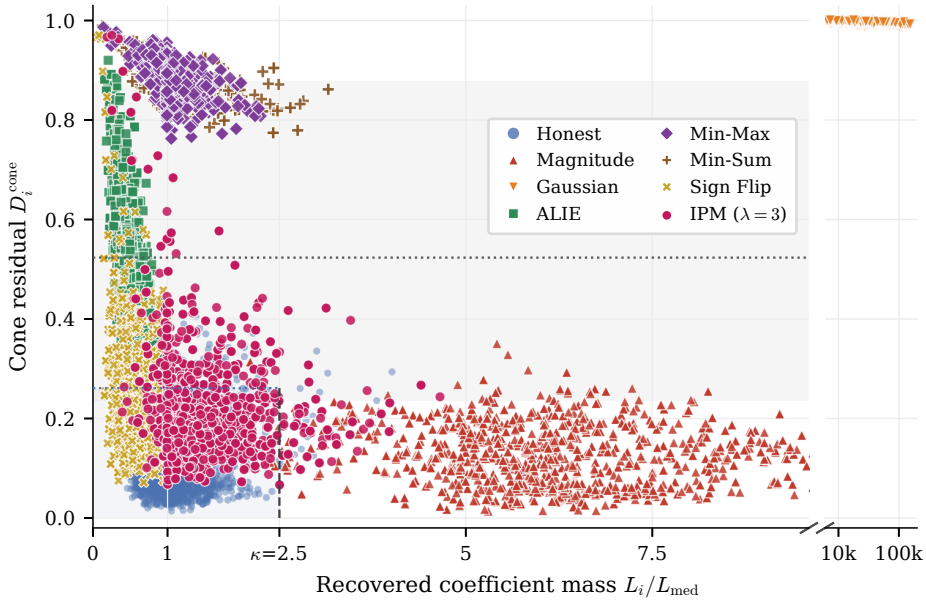


Figure 10: **Uncurated version of Figure 9(c).** MNIST, seed 42, $q = 1$, all 1000 rounds of the clean and seven attacked runs; a plain random subsample of 3,000 honest and 1,500 submissions per attack with no quantile filtering. The shaded box is bounded by $\kappa = 2.5$ and the honest 95th-percentile cone residual (0.26); the gray dotted line and band are the median and interquartile range of the per-round Stage II cutoff (0.52, [0.24, 0.88]). Gaussian submissions are on the broken right axis. Sign Flip and IPM overlap the honest box in 69.8% and 68.4% of rounds; Min-Max and Min-Sum coincide.

Table 17: **Stage-wise detection on MNIST, seed 42.** Statistics are pooled over all 1000 rounds. Recall, honest FPR, and fair mass are percentages; fair mass is the retained fraction of honest F_i^q mass.

Attack	Recall ↑	Honest FPR ↓	Fair mass ↑	Scale used	Cone used
Sign Flip	58.4	16.7	82.9	0.15	3.85
IPM ($\lambda = 3$)	59.4	16.2	83.6	0.06	3.94
Gaussian	**100.0**	**0.0**	**100.0**	**4.00**	0.00
ALIE	94.9	2.0	97.3	0.17	3.83
Min-Max	99.6	0.2	99.8	0.01	**3.98**
Min-Sum	99.7	0.1	99.8	0.03	3.97
Magnitude	99.5	0.2	99.8	3.98	0.02

E.2 Stage-wise screening behavior

Gaussian and Magnitude Scaling primarily consume scale-stage budget, whereas ALIE, Min-Max, and Min-Sum primarily consume cone-stage budget. Sign Flip and IPM are harder: their Byzantine recall is about 59%, and the selector also rejects about 16% of honest observations.

E.3 FROM SIMPLEX TO CONE: CONTROLLED FAIRNESS-WARP TESTS

This subsection separates two questions that the end-to-end BOBA comparison of Section 5.2 cannot separate: whether the unit-sum simplex is the right observation model for fairness-scaled messages, and whether a training run built on it is robust. The first question is answered by a fixed-checkpoint reconstruction diagnostic; the second by the training control of Table 19.

Fixed-checkpoint reconstruction residuals. Table 18 and Figure 11 use the same checkpoint, the same sampled honest clients, and the same trusted basis G for both geometries, so that only the feasible set of the coefficient fit differs. For $q = 0$ the message is the plain update and its nonnegative cone mass $\|\hat{a}_i\|_1$ is already close to one on MNIST and EMNIST (0.88 and 0.84), so the unit-sum and cone residuals nearly coincide there; on CIFAR-10 and AGNews the plain updates have mass 0.65 and 0.74 and the unit-sum fit is already worse. For $q = 1$ the cone residual is unchanged on every submission, because the cone is invariant to $u_i \mapsto F_i u_i$, whereas the unit-sum residual moves with the distance of the true mass from one: on MNIST ($F_i \approx 0.24$) it rises from 0.19 to 1.08, i.e. the simplex fit becomes worse than predicting zero; on EMNIST ($F_i \approx 0.77$) and CIFAR-10 ($F_i \approx 0.92$) the increase is moderate; on AGNews ($F_i \approx 1.10$) the scaling moves the mass toward one and the unit-sum penalty shrinks. The direction of the effect therefore follows $|\log \|\hat{a}_i\|_1|$, not the spread of the losses inside a round. The unit-sum residual is never smaller than the cone residual (the simplex is a subset of the cone), and the diagnostic makes no statement about attacks: a negatively scaled mean has zero nonnegative mass under either model, so this experiment does not by itself explain BOBA's IPM collapse.

Table 18: **Unit-sum simplex versus nonnegative cone on identical inputs.** For each dataset the FUSE run's seed-43 checkpoint at round 600 is loaded, the next ten rounds' honest clients are replayed with the production local step (10 clients per round, 100 submissions), and the production server basis G (20 trusted examples per class) is built once per round. Each honest update Δ_i is submitted as $u_i = F_i^q \Delta_i$ with $q = 0$ (plain update) and $q = 1$ (the paper's fairness-scaled vector, F_i = start-of-round loss). $D^{\text{cone}} = \min_{a \geq 0} \|Su_i - Ga\|^2 / \|Su_i\|^2$ and $D^{\text{simplex}} = \min_{p \geq 0, \mathbf{1}^\top p = 1} \|Su_i - Gp\|^2 / \|Su_i\|^2$ (energy ratios) are computed in the same screened coordinates S with the aggregator's own solvers; $\|\hat{a}_i\|_1$ is the mass of the cone fit. Medians over the 100 submissions; the gap is $D^{\text{simplex}} - D^{\text{cone}}$ and is positive in every case. No optimizer, trajectory, attack or aggregation rule is involved.

Dataset	Screened coords.	q	F_i med. [p10, p90]	$\|\hat{a}_i\|_1$ med.	D^{cone} med. / p90	D^{simplex} med. / p90	gap med. / p90	$\Pr[D^{\text{simplex}} > D^{\text{cone}}]$
CIFAR-10 / ResNet-18	classifier (5,130)	0	0.92 [0.81, 1.18]	0.65	0.080 / 0.111	0.102 / 0.175	+0.024 / +0.066	1.00
		1	0.92 [0.81, 1.18]	0.59	0.080 / 0.111	0.111 / 0.223	+0.035 / +0.113	1.00
	full (11,173,962)	0	0.92 [0.81, 1.18]	0.28	0.823 / 0.890	1.242 / 1.660	+0.427 / +0.799	1.00
		1	0.92 [0.81, 1.18]	0.26	0.823 / 0.890	1.307 / 2.079	+0.502 / +1.267	1.00
MNIST / MLP	classifier (2,010)	0	0.24 [0.20, 0.30]	0.88	0.178 / 0.297	0.185 / 0.306	+0.008 / +0.047	1.00
		1	0.24 [0.20, 0.30]	0.21	0.178 / 0.297	1.077 / 3.247	+0.899 / +3.057	1.00
	full (159,010)	0	0.24 [0.20, 0.30]	0.57	0.411 / 0.575	0.484 / 0.751	+0.066 / +0.269	1.00
		1	0.24 [0.20, 0.30]	0.14	0.411 / 0.575	3.233 / 8.688	+2.759 / +8.160	1.00
EMNIST / CNN	classifier (68,655)	0	0.77 [0.70, 0.83]	0.84	0.162 / 0.211	0.165 / 0.218	+0.006 / +0.019	1.00
		1	0.77 [0.70, 0.83]	0.64	0.162 / 0.211	0.188 / 0.281	+0.029 / +0.069	1.00
	full (124,975)	0	0.77 [0.70, 0.83]	0.84	0.161 / 0.212	0.169 / 0.224	+0.005 / +0.016	1.00
		1	0.77 [0.70, 0.83]	0.64	0.161 / 0.212	0.190 / 0.275	+0.030 / +0.069	1.00
AGNews / GRU	full (14,282,672)	0	1.10 [1.01, 1.17]	0.74	0.339 / 0.395	0.403 / 0.464	+0.061 / +0.078	1.00
		1	1.10 [1.01, 1.17]	0.82	0.339 / 0.395	0.366 / 0.433	+0.026 / +0.047	1.00

Table 19: Controlled fairness-warp mechanism test on MNIST/MLP, seed 42, 1000 rounds. Entries are global / worst-10% client accuracy (%) over rounds 990–999; "n/a" = condition not run. The two $q = 1$ screening rows differ only in the Stage II score (unit-sum simplex distance versus nonnegative-cone distance); everything else, including Stage I and the loss-normalized aggregation, is identical.

Structural model	q	Clean	Sign Flip	Magnitude	IPM	Min-Max	Min-Sum
No screening	1	93.95/89.26	92.39/87.11	95.20/90.06	n/a	n/a	n/a
Simplex screening	0	96.55/92.13	96.44/91.93	96.62/92.33	n/a	n/a	n/a
Simplex screening	1	93.95/89.26	93.85/89.08	93.95/89.26	89.91/83.98	92.91/87.46	92.83/87.40
Scale + cone	1	93.95/89.26	93.20/88.09	93.96/89.27	92.35/87.01	93.94/89.23	93.96/89.26

Training control. Table 19 replaces FUSE's nonnegative-cone fit by a unit-sum simplex fit inside the same pipeline. The $q = 0$ simplex row is a sanity check of the unwarped geometry: with plain updates the simplex is the appropriate model and performs well. It cannot, however, be compared directly with the $q = 1$ rows, because changing q alters both the upload geometry and

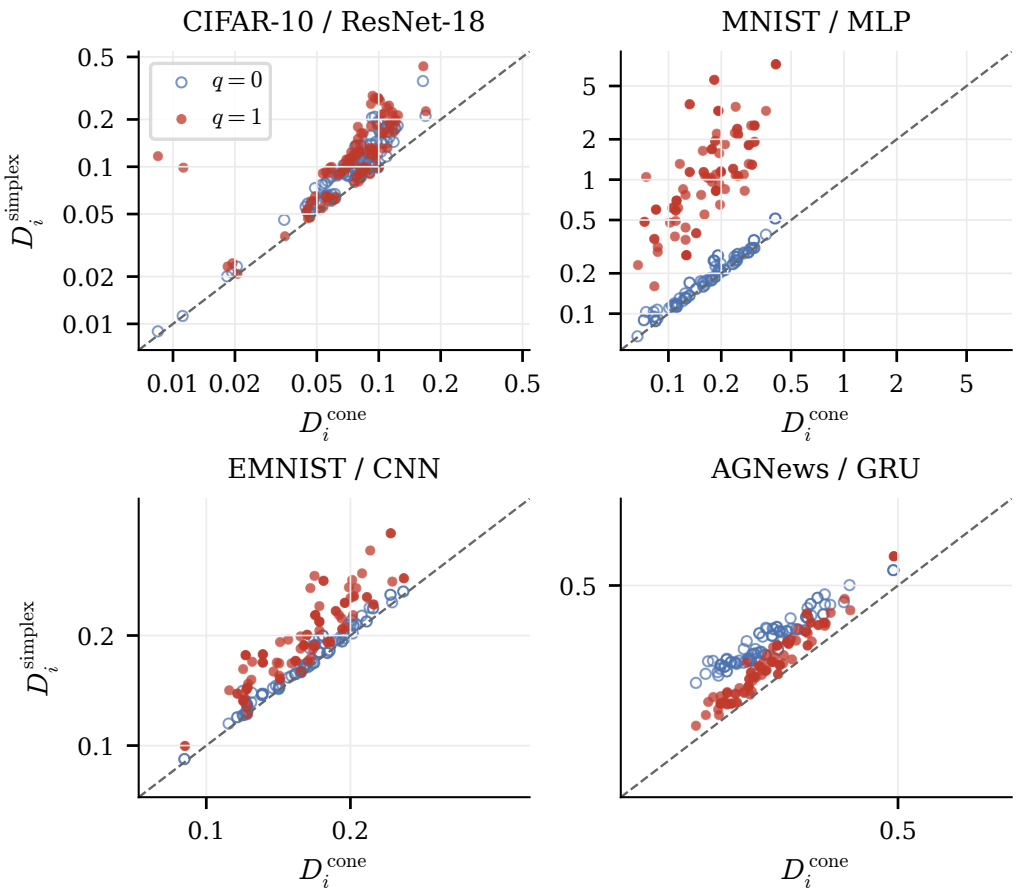


Figure 11: **Unit-sum versus cone residual for the same honest submissions.** Each panel uses that dataset's own screening coordinates (classifier block on the three vision benchmarks, all parameters on AGNews), seed-43 checkpoint at round 600, rounds 600–609 (100 submissions per dataset); plain updates ($q = 0$, hollow) and fairness-scaled submissions ($q = 1$, filled). Every point lies above the diagonal; log axes.

Table 20: Screening behavior behind the two $q = 1$ rows of Table 19: Byzantine recall and honest false-rejection rate, pooled over all 1000 rounds (4 Byzantine and 10 honest submissions per round).

	Simplex screening, $q = 1$			Scale + cone, $q = 1$		
Attack	Recall	Honest FPR	Global acc.	Recall	Honest FPR	Global acc.
Sign Flip	0.92	0.03	93.85	0.58	0.17	93.20
Magnitude	0.99	0.00	93.95	0.99	0.00	93.96
IPM	0.45	0.22	89.91	0.62	0.15	92.35
Min-Max	0.56	0.18	92.91	1.00	0.00	93.94
Min-Sum	0.62	0.15	92.83	1.00	0.00	93.96

the optimized fairness objective, so the accuracy difference does not isolate geometry. The matched $q = 1$ comparison is the relevant test of observation-model mismatch, and Table 20 shows where the two models separate. Under Clean and Magnitude the rows coincide (Stage I, shared by both, removes Magnitude). Under Sign Flip the simplex score is the *better* detector (recall 0.92 against 0.58, honest FPR 0.03 against 0.17): a negated honest mean has a negative coefficient sum, so the unit-sum constraint rejects it directly, whereas the cone only measures its residual; the MLP tolerates the leaked vectors and the accuracy gap is 0.65 points in the simplex's favor. Under the invalid-mixture attacks the picture reverses. For Min-Max and Min-Sum the cone score removes essentially every Byzantine submission (recall 1.00, FPR 0.00) while the simplex score removes 56–62% of them and rejects 15–18% of honest clients instead; for IPM the simplex recall drops to 0.45 with FPR 0.22 against 0.62 and 0.15 for the cone. The corresponding accuracy differences on MNIST are moderate (1.0–2.4 percentage points global, 1.8–3.0 worst-10%) because the MLP is forgiving, but the screening statistics separate the two geometries clearly: fairness-scaled honest updates have coefficient mass $F_i^q \approx 0.2$–0.3, so their unit-sum residuals overlap those of the attacks (Table 18) and the simplex ranking spends its budget on honest clients. These numbers do not license a claim of universal cone superiority (Sign Flip is a counterexample), but they do isolate the mechanism: on the attacks where end-to-end BOBA fails, replacing the nonnegative cone by the unit-sum simplex inside FUSE's own pipeline reproduces the loss of recall, with no other component changed.

E.4 Loss-scale recovery across fairness strengths

The ideal coefficient identity $L_i = F_i^q$ implies $\frac{1}{q} \log L_i = \log F_i, \qquad q > 0$. We test this identity in a matched MNIST/MLP checkpoint experiment at $q \in \{0.1, 0.5, 1.0, 1.5, 2.0, 3.0\}$. For every value of q, FUSE is trained from scratch with that same upload exponent to round 299, using the same seed, partition, optimizer, scale-and-cone selector, and previous server-loss q-FFL aggregation. Trusted-reference augmentation is fixed to one view throughout. We then freeze the corresponding checkpoint, form $u_i = F_i^q \Delta_i$ for all 100 honest clients, and rerun the full-parameter nonnegative-cone recovery to obtain $L_i = \|\widehat{a}_i\|_1$. Thus, each panel contains 100 clients evaluated at one matched checkpoint; no Byzantine submissions or pooled training-round observations enter this diagnostic.

Table 21: Numerical summary of Fig. 12. Perfect loss-scale recovery has Pearson and Spearman correlation one, slope one, and zero mean absolute log error.

q	Pearson r	Spearman ρ	OLS slope	MALE
0.1	0.153	0.179	1.88	3.997
0.5	0.400	0.355	1.11	0.695
1.0	0.608	0.537	1.02	0.301
1.5	0.736	0.670	**1.00**	0.180
2.0	<u>0.817</u>	<u>0.751</u>	**1.00**	<u>0.123</u>
3.0	**0.903**	**0.848**	<u>0.99</u>	**0.070**

The inverse transformation becomes progressively better calibrated as q increases in this checkpoint audit: the fitted slope is within 0.02 of one for $q \geq 1$, while MALE decreases from 0.301 at $q = 1$ to 0.070 at $q = 3$. Rank recovery also strengthens, reaching Spearman $\rho = 0.848$ at $q = 3$. In contrast, $q = 0.1$ is a clear failure regime. If $\log L_i = q \log F_i + e_i$, then the reconstructed log loss is $q^{-1} \log L_i = \log F_i + e_i/q$; hence coefficient-recovery error is amplified by $1/q$. The result supports using L_i directly as a radial screening statistic, while cautioning against interpreting $L_i^{1/q}$ as an accurate client-loss estimator when q is small.

E.5 Local-step sensitivity

The analysis of Section 3 and the main benchmark use one full-batch local step, for which $\Delta_i \approx G p_i$ is a first-order identity. Table 22 asks whether the geometry survives several local steps. Honest clients take $E \in \{2, 5\}$ full-batch steps (learning rate 0.1) and upload $F_i(w_t)^q (w_i - w_t)$, with $F_i(w_t)$ the train-mode loss of the first step; the server references g_c are built with the same E steps, so the basis and the submissions undergo the same local trajectory. Everything else follows the main protocol (MNIST/MLP, seed 42, $q = 1$, 1000 rounds; $E = 1$ is the main-table run).

The mixture structure degrades gracefully rather than breaking. The honest cone residual grows with E (median 0.139, 0.246, 0.358 for $E = 1, 2, 5$) but stays well below the residuals of the structural

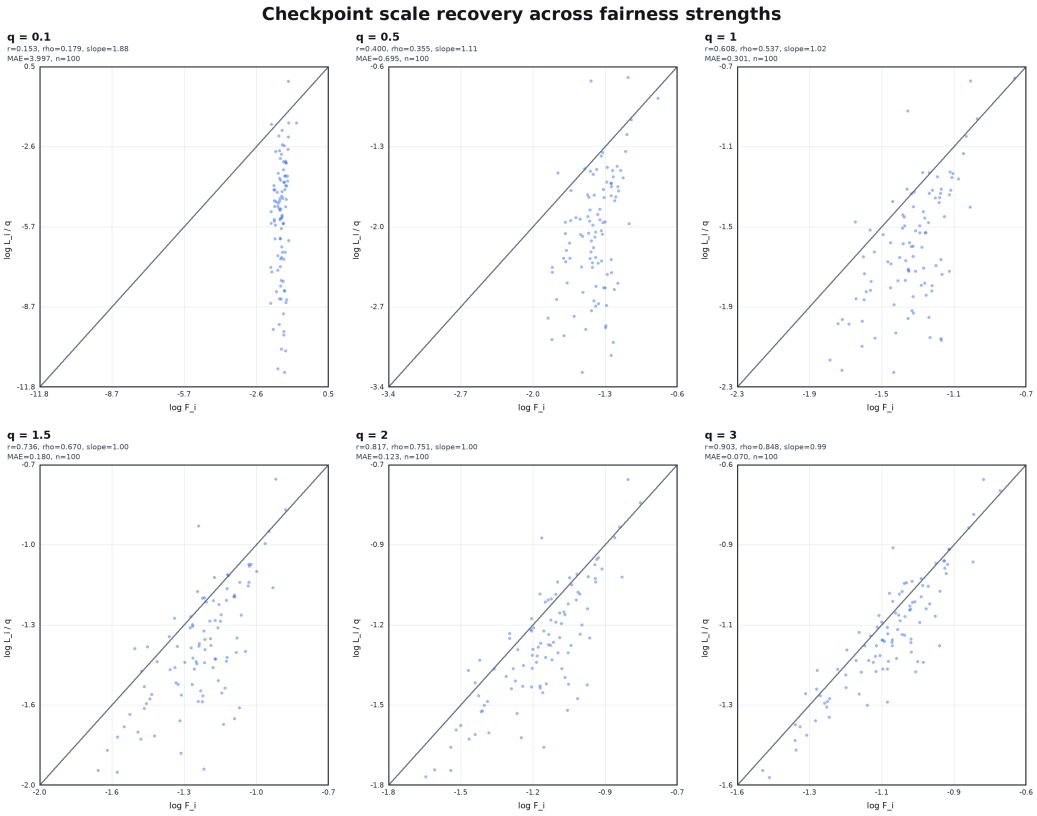


Figure 12: **Checkpoint recovery of client loss scale across q.** Each panel plots $x = \log F_i$ against $y = q^{-1} \log L_i = \log L_i^{1/q}$ for all 100 honest clients at the matched round-299 checkpoint; the diagonal is $y = x$. Pearson r, Spearman ρ, ordinary-least-squares slope, and mean absolute log error (MALE) use all points. Axes retain equal scaling and show the central 98% only for visualization.

attacks (0.49–1.00 in Figure 9c), and the scale identity is preserved: the correlation between L_i and F_i^q is 0.944, 0.981, and 0.976. Clean accuracy rises with more local computation (93.95, 94.98, 96.06), every attack except Sign Flip and IPM stays within 0.1 points of the corresponding clean run, and the seven-attack floor moves from 92.35/87.01 to 91.33/85.96 and 91.45/85.78, i.e. IPM remains the binding attack and its gap to clean widens from 1.6 to 3.7 and 4.6 points as the pooled recall over the seven attacks falls from 0.88 to 0.81 and 0.79 (honest FPR 0.05 to 0.08). The applicable regime is therefore not limited to a single step, but the guarantees should be read as small-local-step statements: with more local progress the honest updates drift away from the trusted class directions and the cone-preserving attacks gain room.

Table 22: Local-step sensitivity (MNIST/MLP, seed 42, $q = 1$, 1000 rounds). Clients take E full-batch local steps and upload $F_i(w_t)^q(w_i - w_t)$ with the start-of-round loss; server references use the same E steps. Accuracy is global / worst-10% over rounds 990–999; the worst attack is the minimum over the seven attacks. Honest cone residual and the Pearson correlation between $L_i = \|\hat{a}_i\|_1$ and F_i^q are pooled over all rounds of the clean run; recall and honest FPR are pooled over the seven attacked runs.

E	Clean	Sign	IPM$_{\lambda=3}$	Gauss.	ALIE	Min-Max	Min-Sum	Magnitude	Worst	honest D^{cone} med/p90	$r(L_i, F_i^q)$	recall / FPR
1	93.95/89.26	93.20/88.09	92.35/87.01	93.95/89.26	93.94/89.14	93.94/89.23	93.96/89.26	93.96/89.27	92.35/87.01	0.139/0.238	0.944	0.88 / 0.05
2	94.98/90.37	92.85/87.57	91.33/85.96	94.98/90.37	94.95/90.22	95.01/90.41	95.01/90.45	94.98/90.37	91.33/85.96	0.246/0.391	0.981	0.81 / 0.08
5	96.06/91.42	94.07/89.02	91.45/85.78	96.06/91.42	96.03/91.49	96.04/91.47	96.05/91.52	96.06/91.42	91.45/85.78	0.358/0.530	0.976	0.79 / 0.08

E.6 Operational sensitivity to fairness strength and scale tolerance

Theorem 2 predicts that larger q widens honest radial spread, while inverse scale recovery improves because log-scale error is divided by q. Across the tested κ values, the worst-attack accuracy is nearly unchanged; increasing κ from 2.0 to 3.0 reduces honest FPR and raises retained fair mass.

Table 23: Final-method MNIST/MLP sensitivity, seed 42, rounds 290–299. Worst-attack results use all seven attacks; FPR and retained fair mass are pooled across attacked conditions. Entries are percentages.

q	0.5	1.0	1.5	2.0
Worst-attack global accuracy	**91.57**	91.18	90.84	90.47
Worst-attack worst-10% accuracy	**86.21**	85.62	85.16	84.85
κ	2.0	2.5	3.0	3.5
Worst-attack global / worst-10%	**91.20**/85.60	91.18/**85.62**	91.19/85.61	91.18/85.61
Honest FPR	5.71	5.00	**4.14**	**4.14**
Retained fair mass	93.66	94.70	**95.75**	95.74

E.7 High-loss-client audit

Table 24: Honest-client rejection by within-round start-loss quantile, pooled over the seven attacked conditions. Rejection share is the fraction of all rejected honest observations contributed by the row.

Dataset	Honest loss quantile	Rejection rate	Retained fair mass	Rejection share
CIFAR-10/ResNet-18	0–50%	3.15%	97.45%	**58.41%**
	50–80%	**1.44%**	**98.99%**	16.05%
	80–100%	3.45%	96.64%	25.54%
MNIST	0–50%	**4.74%**	**95.42%**	**46.82%**
	50–80%	5.19%	95.20%	30.76%
	80–100%	5.67%	94.65%	22.41%
EMNIST-Balanced	0–50%	1.51%	99.26%	**61.00%**
	50–80%	**0.89%**	**99.55%**	21.64%
	80–100%	1.07%	99.48%	17.36%
AGNews	0–50%	**5.53%**	**95.09%**	**40.60%**
	50–80%	6.02%	94.30%	26.52%
	80–100%	11.19%	89.83%	32.87%

The AGNews tail is the least well preserved, consistent with the weaker IPM result in Table 10.

Table 25 extends the highest-loss-20% audit to the five registered seeds and splits it by attack. It pools the *full* training horizon, whereas the matched comparison of Table 2 pools the first 300 rounds, the window over which the H-nobs and BOBA audit runs were reproduced. The two windows are not interchangeable: early in training the honest losses are still close together, so every method rejects fewer high-loss clients and retention is higher. For FUSE the pooled retained mass over the first 300 rounds is 99.77/98.48/100.00/90.06% on CIFAR-10/MNIST/EMNIST-Balanced/AGNews against 95.23/94.27/99.65/91.40% over the full horizon. The ordering between methods is unaffected. The clean-side attacks (Gaussian, Magnitude, and on most datasets Min-Max and Min-Sum) never remove a high-loss honest client. The cost is concentrated in attacks whose Byzantine vectors are placed near the honest cone: Sign Flip on MNIST, ALIE on CIFAR-10, and ALIE on AGNews, where the cone stage removes the honest tail clients whose class mixtures are least typical.

Table 25: **Per-attack high-loss retention, five seeds.** Retained fair mass (%) of the highest-loss 20% of honest clients, mean$_{\pm\text{std}}$ over seeds 40–44, pooled over rounds r with $(r+1) \bmod 10 = 0$. The last column is the retained fair mass of all honest clients (pooled over attacks).

Dataset/model	Sign	IPM$_{\lambda=3}$	Gauss.	ALIE	Min-Max	Min-Sum	Magnitude	Pooled	All honest
CIFAR-10/ResNet-18	$89.09_{\pm 7.27}$	$99.17_{\pm 0.78}$	$100.00_{\pm 0.00}$	$81.43_{\pm 5.53}$	$98.22_{\pm 1.91}$	$99.86_{\pm 0.13}$	$100.00_{\pm 0.00}$	$95.23_{\pm 2.32}$	$97.63_{\pm 0.58}$
MNIST/MLP	$82.64_{\pm 5.71}$	$89.40_{\pm 2.91}$	$100.00_{\pm 0.00}$	$92.17_{\pm 5.93}$	$98.46_{\pm 3.24}$	$98.45_{\pm 3.47}$	$100.00_{\pm 0.00}$	$94.27_{\pm 2.87}$	$96.01_{\pm 1.07}$
EMNIST-Balanced/CNN	$99.18_{\pm 0.47}$	$99.26_{\pm 0.51}$	$100.00_{\pm 0.00}$	$99.13_{\pm 0.38}$	$100.00_{\pm 0.00}$	$100.00_{\pm 0.00}$	$100.00_{\pm 0.00}$	$99.65_{\pm 0.17}$	$99.21_{\pm 0.34}$
AGNews/GRU	$90.28_{\pm 10.73}$	$97.79_{\pm 2.34}$	$100.00_{\pm 0.00}$	$57.33_{\pm 6.03}$	$98.49_{\pm 2.00}$	$98.58_{\pm 1.82}$	$100.00_{\pm 0.00}$	$91.40_{\pm 1.92}$	$94.54_{\pm 0.66}$

E.8 Aggregation controls

The eight EMNIST $q+1$ runs did not fail because of the scheduler, an out-of-memory error, or a non-finite value. They all stopped at round 123 under the registered chance-level rule. For 47 classes, this rule declares collapse after 25 consecutive evaluated rounds at or below $1/47 + 0.005 = 2.628\%$, beginning no earlier than the 100th one-indexed round. The recorded accuracies from zero-indexed

Table 26: Post-screening aggregation controls, seed 42. Entries are global / worst-10% client accuracy (%). Worst attack is the componentwise minimum across the available attacked conditions. A dash indicates that at least one required attack lacks a final window.

Dataset	Aggregation	Clean	Worst attack
CIFAR-10/ResNet-18	server-loss q-FFL	69.13/59.12	63.38/52.19
	mean-u	79.25/70.60	57.01/45.18
	loss-normalized	**82.19/72.98**	**76.92/67.30**
MNIST/MLP	server-loss q-FFL	95.54/91.46	**93.86/88.96**
	mean-u	93.98/89.29	92.81/87.69
	recovered q-FFL	**95.56/91.47**	–
	loss-normalized	93.95/89.26	92.77/87.60
EMNIST-Balanced/CNN	server-loss q-FFL	73.14/64.03	73.14/63.56
	mean-u	60.96/50.97	9.14/2.83
	recovered q-FFL	69.65/60.30	–
	loss-normalized	**79.55/70.28**	**78.68/69.37**
AGNews/GRU	server-loss q-FFL	72.02/53.93	63.86/45.27
	mean-u	82.01/66.51	35.65/22.83
	recovered q-FFL	68.84/50.24	–
	loss-normalized	**83.79/68.08**	**66.29/45.70**

Table 27: Loss-normalization exponent control, seed 42. Entries are global / worst-10% client accuracy (%). CIFAR-10 uses the auxiliary five-layer CNN in this controlled comparison. A dash for EMNIST means that all eight conditions (Clean and seven attacks) met the registered collapse criterion.

Dataset	Denominator exponent	Clean	Worst attack
CIFAR-10/CNN	$q+1$	39.80/27.55	39.70/27.45
CIFAR-10/CNN	q	**46.53/34.15**	**46.35/33.93**
MNIST/MLP	$q+1$	93.93/89.23	92.75/87.64
MNIST/MLP	q	**93.95/89.26**	**92.77/87.60**
EMNIST/CNN	$q+1$	–	–
EMNIST/CNN	q	**79.55/70.28**	**78.68/69.37**
AGNews/GRU	$q+1$	82.75/66.31	65.23/45.87
AGNews/GRU	q	**83.79/68.08**	**66.29/45.70**

rounds 99 through 123 all satisfy this condition; at round 123 the clean run has 2.622% global accuracy and 0% worst-10% client accuracy. Hence round 123 is the first round at which the 25-round patience can be exhausted.

The underlying failure is excessive step attenuation in this historical denominator control. Because $q = 1$, using exponent $q + 1$ squares the reconstructed loss. At round 0 of the clean EMNIST run, where the estimated losses are close to $\log 47 \approx 3.85$, the $q + 1$ denominator is 148.74, compared with 38.57 for exponent q. The corresponding aggregate-update norm is 8.99×10^{-4} rather than 3.47×10^{-3}, a reduction by a factor of about 3.86. Test loss decreases only from 3.853 to 3.846 by round 123, so the model does not leave the chance-level region before the registered stop. The same outcome in the clean condition shows that this is an optimization failure caused by the denominator choice, rather than an attack-specific failure. With exponent q, the matched clean run reaches 10.31% accuracy at round 123 and later completes normally. This diagnostic illustrates the optimization role of Component III: normalization should suppress excessive post-screening steps without over-attenuating the update. The $q + 1$ denominator violates this balance on EMNIST, whereas the final exponent q preserves a substantially larger update and completes training normally.

E.9 Attack strength and budget

Table 28: CIFAR-10/ResNet-18 sign-reversal strength, seed 42. Entries are global / worst-10% client accuracy (%).

Method	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
FUSE	**81.73/73.35**	**80.26/71.32**	**81.43/72.18**
server-loss q-FFL aggregation	66.43/55.37	63.38/52.19	62.89/51.65
mean-u aggregation	78.86/69.87	76.64/68.34	72.31/61.88

Table 29: Declared-budget sensitivity for final FUSE on MNIST, seed 42, with four actual Byzantine submissions.

Declared f	Worst attack (global / worst-10% accuracy, %)
2	81.91/74.58
4	92.35/87.01
6	**92.86/87.63**

Under-declaring the budget is harmful under Gaussian perturbations. The slight advantage of $f = 6$ is a single-seed endpoint observation rather than evidence that over-rejection is generally beneficial.

E.10 Trusted data, heterogeneity, and identifiability

Table 30: Empirical validity of the trusted screening geometry at fixed CIFAR-10/ResNet-18 checkpoints. Captured energy is the fraction of honest update energy represented by the trusted class span; oracle overlap compares the trusted span with an all-client-data class-span oracle.

Checkpoint	Classifier energy	Classifier overlap	Full energy	Full overlap
Round 299	0.892	**0.978**	**0.249**	**0.703**
Round 599	0.901	0.972	0.161	0.595
Round 899	**0.918**	0.967	0.120	0.461
Round 1499	0.852	0.975	0.051	0.337

Table 31: Trusted augmentation repeats at fixed CIFAR-10/ResNet-18 checkpoints, classifier coordinates, seed 42.

Repeats	Projection energy	Negative coefficient mass	Loss MAE	Time (s)
1	0.9223	14.5973	1.7817	**0.408**
6	**0.9237**	**5.1632**	**1.2567**	1.061
12	0.9234	7.6181	1.3797	1.793
24	0.9236	6.6128	1.4823	3.358
64	0.9236	7.0401	1.4833	8.584

Table 32: Trusted-size sensitivity for final FUSE on MNIST, seed 42.

Samples/class	Worst attack (global / worst-10% accuracy, %)
1	90.30/84.44
20	92.35/87.01
50	**92.79/87.52**

Table 33: Label-skew sensitivity for final FUSE on MNIST, seed 42.

Dirichlet α	Worst attack (global / worst-10% accuracy, %)
0.1	92.70/86.93
0.5	**92.86/87.98**
1.0	92.35/87.01

Table 34: Numerical identifiability of the trusted basis over completed rounds and conditions. Values are median [5th, 95th percentile].

Dataset	Numerical rank	$\sigma_{\min}$	Condition number
CIFAR-10/ResNet-18	10 [10, 10]	0.0823 [0.0649, 0.0971]	6.78 [5.77, 9.22]
MNIST	10 [10, 10]	0.0590 [0.0529, 0.0785]	8.64 [7.60, 9.99]
EMNIST-Balanced	47 [47, 47]	0.0968 [0.0145, 0.1951]	14.26 [7.18, 34.02]
AGNews	4 [4, 4]	0.0286 [0.0157, 0.0456]	4.09 [1.73, 7.09]

E.11 Computation and communication

FUSE preserves the one-vector communication protocol but is approximately $2.9\times$ slower than BOBA in this measurement. The timing uses PyTorch on NVIDIA A100 resources and includes trusted-reference construction.

Table 35: Measured server-side time per attacked round with 14 submissions. All listed methods communicate one model-shaped vector per client.

Method	Server time/round	Client vectors	Extra client statistics
q-FFL-GD	**1.71** ms	1	none
H-nobs	3.05 ms	1	none
BOBA	478.18 ms	1	none
FUSE	1387.10 ms	1	none

E.12 Over-declared budgets and always-on screening

The main benchmark declares $f_t = 0$ in clean conditions and $f_t = 4$ when four Byzantine submissions are appended. To test deployment with an over-declared budget, we also run clean training with $f_t = 4$ for five seeds.

Table 36: Clean accuracy (%) with a matched budget ($f_t = 0$) and an over-declared always-on budget ($f_t = 4$), five seeds.

Dataset/model	Matched $f_t = 0$	Always-on $f_t = 4$
CIFAR-10/ResNet-18	82.25	79.68
MNIST/MLP	93.83	93.78
EMNIST-Balanced/CNN	79.12	80.06

Accuracy changes little, but the fairness cost is substantial. Because the fixed-budget rule rejects exactly four of ten honest submissions, its honest FPR is 40%; retained fair mass falls to 59–62%, and the highest-loss 20% of clients retain only 53–79% of their mass. A median-plus-three-MAD threshold alternative reduces clean FPR to 2.8–5.7% and retains 94.4–98.3% of fair mass, but trades away attack recall, which drops from 88.0–96.5% under the fixed budget to 70.7–78.7%. Thus an accurate fixed budget favors attack protection, whereas adaptive rejection better protects fairness when the attack count is unknown.

Declared budget above the actual attack count. The matched benchmark gives every fixed-budget rule the true Byzantine upper bound. Table 37 keeps the declared budget at $f = 4$ while only 1, 2, or 3 of the 14 submissions are Byzantine (MNIST/MLP, seed 42). Robustness is unaffected: accuracy stays within 0.7 points of the matched run for every attack, and Gaussian, Min-Max, Min-Sum, and Magnitude are still removed with recall 1.00. The cost is entirely on the fairness side and is monotone in the over-declaration: with one actual attacker the screens must still reject four submissions, so the honest false-rejection rate rises from 0.00–0.17 to 0.30–0.35 and the retained fair mass of the highest-loss 20% of honest clients falls from 0.84–1.00 to 0.69–0.75; with three attackers the values are 0.10–0.20 and 0.84–0.96. Sign Flip and IPM recall also degrades as the budget is over-declared (0.58/0.62 at four attackers, 0.77/0.78 at one), because the surplus budget is spent on honest clients whose cone scores overlap those attacks. Together with the clean $f = 4$ run (FPR 0.40, retained mass 0.61) this quantifies the price of not knowing the attack count: accuracy is preserved, fairness mass is not.

Table 37: Declared budget $f = 4$ with fewer actual Byzantine submissions (MNIST/MLP, seed 42, $q = 1$, 1000 rounds). Global / worst-10% accuracy over rounds 990–999; below each block, Byzantine recall, honest false-rejection rate, and the retained fair mass of the highest-loss 20% of honest clients, pooled over all rounds. Actual $= 0$ is the over-declared clean run; actual $= 4$ is the matched main-table run.

Actual Byzantine	Sign	IPM$_{\lambda=3}$	Gauss.	ALIE	Min-Max	Min-Sum	Magnitude
1 of 14: accuracy	93.76/88.86	93.61/88.76	93.95/89.32	93.92/89.32	93.95/89.30	93.95/89.29	93.93/89.32
recall / honest FPR	0.77 / 0.32	0.78 / 0.32	1.00 / 0.30	0.49 / 0.35	1.00 / 0.30	1.00 / 0.30	1.00 / 0.30
retained top-20% fair mass	0.71	0.73	0.75	0.69	0.75	0.75	0.75
2 of 14: accuracy	93.58/88.67	93.28/88.37	93.97/89.21	93.92/89.28	93.97/89.26	93.97/89.22	93.97/89.21
recall / honest FPR	0.74 / 0.25	0.75 / 0.25	1.00 / 0.20	0.45 / 0.31	1.00 / 0.20	1.00 / 0.20	1.00 / 0.20
retained top-20% fair mass	0.80	0.79	0.85	0.73	0.84	0.85	0.85
3 of 14: accuracy	93.40/88.40	92.86/87.68	93.98/89.34	93.90/89.01	93.91/89.12	93.97/89.33	93.98/89.34
recall / honest FPR	0.68 / 0.20	0.68 / 0.20	1.00 / 0.10	0.99 / 0.10	1.00 / 0.10	1.00 / 0.10	1.00 / 0.10
retained top-20% fair mass	0.84	0.88	0.96	0.90	0.95	0.96	0.96
4 of 14: accuracy	93.20/88.09	92.35/87.01	93.95/89.26	93.94/89.14	93.94/89.23	93.96/89.26	93.96/89.27
recall / honest FPR	0.58 / 0.17	0.62 / 0.15	1.00 / 0.00	0.95 / 0.02	1.00 / 0.00	1.00 / 0.00	0.99 / 0.00
retained top-20% fair mass	0.84	0.90	1.00	0.93	0.99	1.00	1.00
0 of 10 (clean, $f = 4$): accuracy 93.89/89.18, honest FPR 0.40, retained top-20% fair mass 0.61							

E.13 COVARIATE-SHIFT STRESS TESTS

We stress-test CIFAR-10/ResNet-18 with client-specific photometric changes, rotations, and CIFAR-10-C corruptions at severities 3 and 5. These experiments use a common 300-round diagnostic window. Shift lowers clean accuracy for all methods, but FUSE's worst-attack accuracy remains within one point of its own shifted-clean value in every condition; for example, severity-3 corruption gives 51.75% clean and 51.03% under the worst attack. The trusted basis is built from unshifted clean examples, yet shifted honest updates retain cone medians of 0.060–0.069, honest FPR of 1.4–2.0%, total fair mass of 95–96%, and high-loss fair mass of 98.5–99.5%.

Varying the trusted set from 5 to 50 examples per class changes the four-attack worst-attack accuracy by at most 1.5 points in both the clean and corrupted domains. Honest FPR falls from 3.2% to 0.9%, while retained fair mass rises from 97.5% to 99.2%. These results show empirical tolerance to the tested covariate shifts; they do not establish guarantees for arbitrary violations of shared class-conditionals.

E.14 ADAPTIVE CONE-PRESERVING ATTACKS

The adaptive attacker constructs a nonnegative trusted-cone vector with a coefficient mass chosen to remain plausible under Stage I. Its cone score can therefore be lower than the honest median, directly realizing the observable boundary of Theorem 2.

Table 38: CIFAR-10 diagnostic accuracy (%) under adaptive cone-preserving attacks. The 300-round values describe the mid-training window and are not terminal 1500-round conclusions.

Window	Clean	Surrogate	Oracle	Slack	Mixture
Rounds 200–299	54.1	20.9	13.6	12.7	19.3
Rounds 1490–1499	82.4	65.6	–	–	–

The 300-round diagnostic shows severe degradation, but the full-horizon surrogate runs partially recover to 65.6% versus 82.4% clean. We therefore interpret the available evidence as a severe transient degradation with a persistent terminal gap, not as universal permanent collapse. Short-horizon Krum and FLTrust values are not used because those baselines have not converged by round 300 on several dataset/model pairs. Cone-copy attacks, which exactly reproduce a high-loss honest vector, remain within one to two points of clean accuracy; harmful adaptive attacks require exploiting the allowed cone and radial degrees of freedom rather than merely duplicating an honest update.

E.15 OTHER OBSERVABLE-BOUNDARY EXPERIMENTS

A cone-copy attacker chooses a scale-plausible coefficient $a_b \succeq 0$ and submits Ga_b. Such a vector has no intrinsic cone-distance certificate of malicious origin. On CIFAR-10/ResNet-18, seed 42, the cone-copy condition completes with 78.32%/68.54% global / worst-10% accuracy for FUSE and 73.16%/63.55% for the unscreened q-FFL-GD control.

A parameter-selection attacker preserves the selected classifier coordinates (it submits the honest classifier mean there) while reversing and amplifying the unselected coordinates. On CIFAR-10/ResNet-18 both the classifier-selected FUSE run and its full-parameter control triggered the collapse criterion before the registered final window. The same attack on the classifier-screened MNIST/MLP and EMNIST/CNN benchmarks (seed 42, $f = 4$) collapses FUSE at rounds 539 and 889: in the screened coordinates the Byzantine submissions are indistinguishable from honest ones (honest cone residual median 0.05 and 0.02, Byzantine recall 0.29 and 0.37, honest false-rejection rate 0.28 and 0.25), so the screens spend their budget on honest clients while the unselected coordinates carry the attack. This is the sharpest empirical statement of the boundary in Section 4: FUSE is robust only to deviations that are observable in the screening coordinates; attacks confined to unobserved coordinates therefore fall outside the guarantees of the selected-coordinate interface.

F Additional Theoretical Analysis and Proofs

This appendix provides the full statements and proofs for the theoretical results in Sections 3 and 7, under the modeling assumptions introduced in Section 2. We restate only the notation and additional quantitative bounds needed by the proofs.

F.1 Proof notation

We fix one communication round and suppress the round index. For compactness, $u_i \in \mathbb{R}^{d_s}$ and $\Delta_i \in \mathbb{R}^{d_s}$ in this appendix denote the screening-coordinate representations (that is, u_i denotes v_i from the main text). The server constructs a trusted class basis $G = [g_1, \dots, g_C] \in \mathbb{R}^{d_s \times C}$. For any nonzero vector u, define its NNLS coefficient, cone projection, and normalized cone distance by $\widehat{a}(u) \in \arg\min_{a \succeq 0} \|u - Ga\|_2^2, \qquad z(u) := G\widehat{a}(u), \qquad D_G(u) := \frac{\|u - z(u)\|_2^2}{\|u\|_2^2}$. For client i, we write $\widehat{a}_i = \widehat{a}(v_i)$, $z_i = z(v_i)$, and $L_i = \|\widehat{a}_i\|_1$. All class losses and class-conditioned updates are evaluated at the same broadcast model, so both are expectations under the same mixture p_i. The trusted loss vector $\widehat{\ell}$ is used only after screening. For notational economy, full-model submissions are distinguished only when needed.

Assumption 3 (Bounded estimation error and nondegenerate trusted geometry)**.** *For every honest client i,*

$$|F_i - \overline{F}_i| \le \epsilon_F, \qquad \|\Delta_i - \overline{\Delta}_i\|_2 \le \epsilon_x.$$

The trusted estimates satisfy

$$\|G - G^\star\|_2 \le \epsilon_G, \qquad \|\widehat{\ell} - \ell\|_\infty \le \epsilon_\ell, \qquad \sigma_{\min}(G) \ge \sigma > 0.$$

F.2 Fairness-warped geometry and stable recovery

Theorem 3 (Full finite-sample coefficient-recovery guarantee)**.** *Under Assumption 3, the finite-sample model in Eq. (6) satisfies $\|e_i\|_2 \le F_i^q(\epsilon_x + \epsilon_G)$, and NNLS recovery obeys*

$$\|\widehat{a}_i - a_i\|_2 \le \frac{2\|e_i\|_2}{\sigma_{\min}(G)}, \qquad \big|\, \|\widehat{a}_i\|_1 - F_i^q \,\big| \le \frac{2\sqrt{C}\,\|e_i\|_2}{\sigma_{\min}(G)}.$$

Proof of Theorem 3. We prove the geometric and recovery claims together.

Population geometry. By Assumption 2 and linearity of expectation, an honest client with class proportions p_i has population loss and population one-step update $\overline{F}_i = \ell^\top p_i, \qquad \overline{\Delta}_i = G^\star p_i$. Therefore its population fairness-scaled update is

$$\overline{u}_i = (\ell^\top p_i)^q G^\star p_i = G^\star a_i^\star, \qquad a_i^\star := (\ell^\top p_i)^q p_i \succeq 0. \tag{7}$$

Because $\|p_i\|_1 = 1$, $\|a_i^\star\|_1 = (\ell^\top p_i)^q$. Thus the honest population family is a radially warped image of the class-mixture simplex: fairness changes the radial coefficient scale while preserving the nonnegative class-mixture structure.

For completeness, the same identities imply $\mathbf{1}^\top a_i^\star = (\ell^\top p_i)^q, \qquad \ell^\top a_i^\star = (\ell^\top p_i)^{q+1}$, and hence $(\mathbf{1}^\top a_i^\star)^{q+1} = (\ell^\top a_i^\star)^q$. At $q = 0$, this reduces to the usual simplex constraint $\mathbf{1}^\top a_i^\star = 1$.

Every population point in Eq. (7) belongs to $\mathrm{cone}(G^\star)$. Conversely, for any $a \succeq 0$, $a \neq 0$, let $s = \mathbf{1}^\top a$ and $p = a/s$. Since $\ell^\top p > 0$, $G^\star a = \frac{s}{(\ell^\top p)^q} G^\star[(\ell^\top p)^q p]$. Hence the conic hull of the fairness-warped honest population family is exactly $\{G^\star a : a \succeq 0\}$. This equality concerns the conic hull, not the family itself, which generally occupies only a subset of the cone and retains the fairness-dependent radial constraint above.

Finite-sample cone model. For an honest client, set $s_i := F_i^q, \qquad a_i := s_i p_i \succeq 0$. Using $u_i = s_i \Delta_i$, add and subtract $s_i G^\star p_i$ to obtain

$$u_i = Ga_i + e_i, \qquad e_i := s_i(\Delta_i - G^\star p_i) + s_i(G^\star - G)p_i. \tag{8}$$

Since $\|p_i\|_2 \le \|p_i\|_1 = 1$, Assumption 3 gives $\|e_i\|_2 \le s_i(\epsilon_x + \epsilon_G)$.

NNLS coefficient and scale recovery. The true coefficient a_i is feasible for the NNLS problem. Therefore, by optimality of $\widehat{a}_i$, $\|u_i - G\widehat{a}_i\|_2 \le \|u_i - Ga_i\|_2 = \|e_i\|_2$. Together with $u_i = Ga_i + e_i$, $\|G(\widehat{a}_i - a_i)\|_2 \le \|G\widehat{a}_i - u_i\|_2 + \|e_i\|_2 \le 2\|e_i\|_2$. The smallest-singular-value bound then yields $\|\widehat{a}_i - a_i\|_2 \le \frac{2\|e_i\|_2}{\sigma_{\min}(G)} \le \frac{2s_i(\epsilon_x+\epsilon_G)}{\sigma}$. Define $\delta_i := \frac{2\sqrt{C}\, s_i(\epsilon_x+\epsilon_G)}{\sigma}$. Since $\|v\|_1 \le \sqrt{C}\|v\|_2$ and $\|a_i\|_1 = s_i$, $|L_i - s_i| \le \|\widehat{a}_i - a_i\|_1 \le \delta_i$. This is the scale-recovery claim used by Component I.

Composition and loss recovery. When $\delta_i < s_i$, $L_i \ge s_i - \delta_i > 0$ and $\widehat{p}_i := \widehat{a}_i/L_i$ satisfies

$$\|\widehat{p}_i - p_i\|_1 \le \frac{\|\widehat{a}_i - a_i\|_1}{L_i} + \frac{|L_i - s_i|}{L_i} \le \frac{2\delta_i}{s_i - \delta_i}.$$

Consequently, with $\widehat{F}_i := \widehat{\ell}^\top \widehat{p}_i$,

$$\begin{aligned}|\widehat{F}_i - F_i| &\le |\widehat{\ell}^\top(\widehat{p}_i - p_i)| + |(\widehat{\ell} - \ell)^\top p_i| + |\ell^\top p_i - F_i| \\ &\le (\ell_{\max} + \epsilon_\ell)\frac{2\delta_i}{s_i - \delta_i} + \epsilon_\ell + \epsilon_F.\end{aligned}$$

Thus the same coefficient recovery that supplies the Stage-I scale also yields a stable class-composition estimate, and therefore a stable trusted-loss reconstruction whenever the coefficient error is small relative to the honest fairness scale. This completes the proof. □

F.3 Two-stage screening guarantee

Theorem 4 (Full two-stage oracle-recovery guarantee)**.** *Assume that G has full column rank, honest clients form a strict majority, and $f_t = |\mathcal{B}_t|$. If $\log\kappa > 2r_H$, then no honest submission enters the Stage-I candidate set. Moreover, any Byzantine submission b satisfying*

$$\log L_b - m_H > \log\kappa + r_H$$

is rejected by Stage I; in particular, for a positively scaled copy $v_b = \lambda v_h$ this condition is guaranteed whenever

$$\log\lambda > \log\kappa + 2r_H, \qquad r_H \le q\log\frac{\ell_{\max}}{\ell_{\min}} + 2\epsilon_{\log L}.$$

For a remaining Byzantine submission $v_b = Gc_b + r_b$ with $r_b \perp \operatorname{span}(G)$,

$$D_G(v_b) \ge \frac{\|r_b\|_2^2 + \sigma_{\min}^2(G)\|(c_b)_-\|_2^2}{\|v_b\|_2^2}.$$

If this lower bound exceeds the largest honest cone score, the submission ranks above every honest client in Stage II. Consequently, if every Byzantine submission satisfies either the Stage-I or the Stage-II separation condition, then

$$\widehat{\mathcal{H}}_t = \mathcal{H}_t.$$

Proof of Theorem 4. We analyze the radial and structural stages in the same proof because they are designed to detect complementary deviations.

Stage I: positive homogeneity and the contaminated median. For any $\lambda > 0$, the change of variables $a = \lambda b$ gives

$$\widehat{a}(\lambda u) = \lambda\widehat{a}(u), \qquad L(\lambda u) = \lambda L(u). \tag{9}$$

The cone projection scales in the same way, $z(\lambda u) = \lambda z(u)$, so $D_G(\lambda u) = D_G(u)$. Thus a positive radial copy changes the coefficient scale but not the normalized cone score.

Let $x_h = \log L_h$ for honest clients, let $m_H = \operatorname{median}_{h\in\mathcal{H}} x_h$, $\quad r_H = \max_{h\in\mathcal{H}}|x_h - m_H|$, and let $m_S = \log L_{\text{med}}$ be a median over all received submissions. Every honest score lies in $[m_H - r_H, m_H + r_H]$. Because honest submissions form a strict majority, any contaminated median must also lie in this interval. Therefore, for every honest client, $x_h - m_S \le 2r_H$. If $\log\kappa > 2r_H$, no honest client can satisfy $L_h > \kappa L_{\text{med}}$.

Conversely, for any Byzantine submission b satisfying $x_b - m_H > \log\kappa + r_H$, we have, since $m_S \le m_H + r_H$, $x_b - m_S \ge x_b - m_H - r_H > \log\kappa$. Hence b enters the Stage-I candidate set. Since no honest submission enters that set and $f_t = |\mathcal{B}_t|$, every such candidate is rejected.

Now consider a positive radial copy $u_b = \lambda u_h$. By Eq. (9), $x_b = x_h + \log\lambda$. Since $x_h \geq m_H - r_H$ and $m_S \leq m_H + r_H$, $x_b - m_S \geq \log\lambda - 2r_H$. Hence $\log\lambda > \log\kappa + 2r_H$ places b in the Stage-I candidate set. The same condition implies $x_b > m_H + r_H \geq \max_{j\in\mathcal{H}} x_j$, so such an attacker has larger recovered scale than every honest client and cannot be displaced by an honest candidate in the top-f truncation.

The fairness strength q enters this condition through the honest radial spread. In the exact population model, $L_h = (\ell^\top p_h)^q$, so $q\log\ell_{\min} \leq \log L_h \leq q\log\ell_{\max}, \qquad r_H \leq q\log\frac{\ell_{\max}}{\ell_{\min}}$. If every recovered log scale differs from its population value by at most $\epsilon_{\log L}$, the same argument gives $r_H \leq q\log\frac{\ell_{\max}}{\ell_{\min}} + 2\epsilon_{\log L}$. Thus stronger fairness weighting legitimately widens the honest radial range and correspondingly increases the separation margin required by a purely radial detector.

Stage II: distance to the trusted nonnegative cone. For any nonzero received vector u, write its orthogonal decomposition relative to $\mathrm{span}(G)$ as $u = Gc + r, \qquad c = G^\dagger u, \qquad r \perp \mathrm{span}(G)$. Here $\mathrm{span}(G)$ contains all linear combinations of the columns of G, including signed combinations, and $G^\dagger = (G^\top G)^{-1}G^\top$ is the Moore–Penrose pseudoinverse of the full-column-rank basis. For every $a \succeq 0$, orthogonality gives

$$\|u - Ga\|_2^2 = \|r\|_2^2 + \|G(c-a)\|_2^2. \tag{10}$$

Let $c_- := \max\{-c, 0\}$ coordinate-wise. The Euclidean distance from c to the nonnegative orthant is $\|c_-\|_2$. Applying the singular-value bounds to Eq. (10) yields

$$D_G(u) \geq \frac{\|r\|_2^2 + \sigma_{\min}^2(G)\|c_-\|_2^2}{\|u\|_2^2}. \tag{11}$$

The first term measures energy outside the trusted class span; the second measures a within-span violation of nonnegative class-mixture structure.

The NNLS projection also gives the geometric condition used in the method interpretation. Let $\rho = u - z(u)$. The Karush–Kuhn–Tucker (KKT) optimality conditions are $G^\top\rho \preceq 0, \qquad \widehat{a}_c(u)\, g_c^\top\rho = 0 \quad (c = 1,\dots,C)$. Hence ρ is orthogonal to every active generator g_c with $\widehat{a}_c(u) > 0$, and therefore to the active face containing the projection $z(u)$. It need not be orthogonal to the entire span of G. In particular, summing the complementary products gives $\langle u - z(u), z(u)\rangle = \widehat{a}(u)^\top G^\top\rho = 0$. Thus the residual is orthogonal to the fitted update, not to every vector in the cone.

For an honest client, Eq. (8) provides a feasible nonnegative coefficient a_i, and therefore $D_G(u_i) \leq \frac{\|e_i\|_2^2}{\|u_i\|_2^2}$. Define $\varepsilon_C := \max_{h\in\mathcal{H}} D_G(u_h)$. For any remaining Byzantine submission $u_b = Gc_b + r_b$, if

$$\frac{\|r_b\|_2^2 + \sigma_{\min}^2(G)\|(c_b)_-\|_2^2}{\|u_b\|_2^2} > \varepsilon_C, \tag{12}$$

then Eq. (11) implies $D_G(u_b) > D_G(u_h)$ for every honest client h. Thus every Byzantine vector satisfying the stated structural margin ranks ahead of all honest clients in Stage II. If Stage I has used rejection slots only on Byzantine clients and all remaining Byzantine clients satisfy Eq. (12), the remaining top-score budget removes them before any honest client. This completes the proof. □

Post-screening step-size control. The final aggregation does not alter the direction of the retained mean submission. Writing $\Delta_t^{\mathrm{SC}} = \alpha_t\bar{u}_t, \qquad \alpha_t = \frac{|\widehat{\mathcal{H}}_t|}{\sum_{i\in\widehat{\mathcal{H}}_t} d_i}$, and assuming $\widetilde{F}_i \leq \widetilde{F}_{\max}$ gives $\frac{1}{\max\{1,\widetilde{F}_{\max}^q\}} \leq \alpha_t \leq 1$. Therefore the post-screening rule prevents additional amplification while retaining a positive multiplicative lower bound on the step size.

$$\bar{u}_t = \frac{1}{|\widehat{\mathcal{H}}_t|}\sum_{i\in\widehat{\mathcal{H}}_t} u_i, \qquad \alpha_t = \frac{|\widehat{\mathcal{H}}_t|}{\sum_{i\in\widehat{\mathcal{H}}_t}\max\{1,\widetilde{F}_i^q\}}, \qquad w_{t+1} = w_t + \alpha_t\bar{u}_t.$$

F.4 Conditional one-step descent

Proposition 1 (Conditional one-step descent after screening). *Let $\bar{u}_t$ and $\alpha_t > 0$ be as defined above. Suppose $\mathcal{J}_q$ is L_J-smooth along the update segment, with $L_J > 0$, and the retained mean satisfies,*

for some $\gamma_t > 0$, $-\langle \nabla \mathcal{J}_q(w_t), \bar{u}_t \rangle \geq \gamma_t \|\nabla \mathcal{J}_q(w_t)\|_2 \|\bar{u}_t\|_2$. *Then* $w_{t+1} = w_t + \alpha_t \bar{u}_t$ *satisfies* $\mathcal{J}_q(w_{t+1}) \leq \mathcal{J}_q(w_t) - \alpha_t \gamma_t \|\nabla \mathcal{J}_q(w_t)\|_2 \|\bar{u}_t\|_2 + \frac{L_J}{2} \alpha_t^2 \|\bar{u}_t\|_2^2$. *Consequently,* $\mathcal{J}_q(w_{t+1}) < \mathcal{J}_q(w_t)$ *if* $\bar{u}_t \neq 0$ *and* $\alpha_t \|\bar{u}_t\|_2 < \frac{2\gamma_t}{L_J} \|\nabla \mathcal{J}_q(w_t)\|_2$.

Proof. Here $\bar{u}_t$ is the original full-model retained mean, not the selected representation used in the screening proofs. By L_J-smoothness along the update segment, $\mathcal{J}_q(w_t + \alpha_t \bar{u}_t) \leq \mathcal{J}_q(w_t) + \alpha_t \langle \nabla \mathcal{J}_q(w_t), \bar{u}_t \rangle + \frac{L_J}{2} \alpha_t^2 \|\bar{u}_t\|_2^2$. Since $\alpha_t > 0$, substituting the assumed alignment inequality gives the claimed bound. Its change from $\mathcal{J}_q(w_t)$ is at most $-\alpha_t \|\bar{u}_t\|_2 \left(\gamma_t \|\nabla \mathcal{J}_q(w_t)\|_2 - \frac{L_J}{2} \alpha_t \|\bar{u}_t\|_2 \right)$. For $\bar{u}_t \neq 0$, the prefactor $\alpha_t \|\bar{u}_t\|_2$ is positive; the stated step-size condition makes the parenthesis positive, yielding strict descent. If $\bar{u}_t = 0$, the model and objective remain unchanged. □

F.5 Attack implications and observable boundary

Representative attack implications. The two parts of Theorem 2 immediately explain the representative attack geometries used in the experiments. A positive magnitude copy $u_b = \lambda u_h$ satisfies $D_G(u_b) = D_G(u_h)$ but shifts the recovered log scale by exactly $\log \lambda$, so it is a Stage-I rather than Stage-II deviation. For an ideal sign reversal $u_b = -\lambda G a$ with $a \succeq 0$, the vector remains inside $\mathrm{span}(G)$ but has negative cone coordinates; consequently $D_G(u_b) \geq \frac{\sigma_{\min}^2(G) \|a\|_2^2}{\|Ga\|_2^2}$. An unconstrained subspace distance would be zero for the same vector, which is why the nonnegativity constraint matters. Finally, for an off-span perturbation $u = Ga + r$ with $a \succeq 0$ and $r \perp \mathrm{span}(G)$, $D_G(u) = \frac{\|r\|_2^2}{\|u\|_2^2}$. These are direct specializations of the general two-stage theorem rather than separate theoretical claims.

Observable boundary. The nonnegative cone is a relaxation of the fairness-warped honest family; cone consistency tests structural compatibility but does not certify honesty. An exact cone-preserving submission with a plausible coefficient scale satisfies $u_b = G a_b, \quad a_b \succeq 0, \quad \|a_b\|_1 \leq \kappa L_{\mathrm{med}} \implies D_G(u_b) = 0, \quad b \notin \mathcal{R}_{\mathrm{scale}}$. Likewise, parameter-selection equivalence implies $v_b = v_h \implies (\widehat{a}_b, L_b, D_G(u_b), \widetilde{F}_b) = (\widehat{a}_h, L_h, D_G(u_h), \widetilde{F}_h)$. Thus FUSE provides no score-based certificate unless an attack creates a separated scale or cone deviation in the selected screening representation.