When Aggregation Hides Directional Structures in Federated Learning

Anonymous authors
Paper under double-blind review

Abstract

Federated learning (FL) relies on aggregating decentralized client updates into a global update, yet this many-to-one aggregation can obscure important directional structures contained in individual client updates. In particular, distinct client directional configurations may yield similar aggregate updates because opposing directions partially cancel each other, making the underlying directional relationships less observable from the aggregate representation alone. We characterize this phenomenon as Hidden Directional Opposition (HDO), which captures client-level directional opposition that exists before aggregation but becomes partially unobservable after aggregation. To analyze such hidden structures in a shared coordinate system, we construct a geometry-aware tangent-space directional representation that reduces endpoint-dependent interference while retaining relative directional displacement. We further introduce a direction-preserving adaptive optimization mechanism that adapts update magnitudes without introducing additional directional rotation. Experiments on CIFAR-10 and CIFAR-100 show that HDO consistently emerges under heterogeneous federated training, becomes more pronounced with stronger client heterogeneity, and exhibits distinct temporal and layer-wise patterns. These results highlight that federated aggregation determines not only the global update but also which client-level directional structures remain observable at the server.

1 Introduction

Federated learning (FL) enables collaborative training across decentralized clients through local optimization and server-side aggregation without centralizing raw data (McMahan et al., 2017; Kairouz et al., 2021; Bonawitz et al., 2019). Prior work has extensively addressed optimization challenges under heterogeneous data, including client drift, objective inconsistency, and optimization stability (Li et al., 2020; Karimireddy et al., 2020; Wang et al., 2020; Acar et al., 2021). While these advances improve federated optimization under heterogeneity, the server ultimately acts on an aggregate update that compresses multiple client optimization states into a single representation. This raises a complementary question: which client-level directional structures remain observable from the aggregate representation?

We revisit federated aggregation from a representation perspective. Beyond determining the global update, aggregation defines a many-to-one mapping from multiple client updates to a single aggregate representation. Under conventional weighted aggregation, $G = \sum_i p_i \Delta_i$, distinct client update configurations can therefore produce identical or similar aggregate updates, making their underlying directional structures indistinguishable from the aggregate representation alone. We refer to this limitation as *aggregation ambiguity*. Importantly, this ambiguity is optimization-relevant: a small aggregate update may indicate either weak client directional activity or strong but mutually canceling client updates. These cases appear similar at the server while corresponding to fundamentally different local optimization states. The key question is therefore not whether averaging is non-injective, but which optimization-relevant directional structures become less observable through this mapping and how they can be characterized. This issue is particularly relevant under heterogeneous training, where

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.

non-IID client distributions induce diverse local optimization behaviors and representation shifts (Hsu et al., 2019; Zhao et al., 2018; Li et al., 2021a).

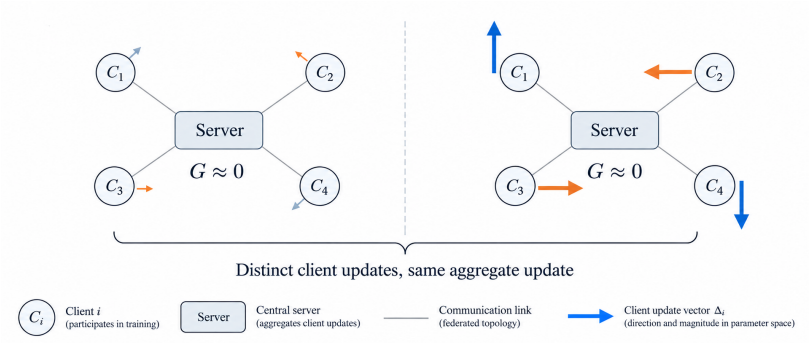


Figure 1: Aggregation ambiguity: distinct client directional structures can yield indistinguishable aggregate updates.

Figure 1 illustrates this ambiguity: clients with weak directional activity and clients with strong but mutually canceling directions may yield similar aggregate observations despite representing substantially different local optimization processes. Existing gradient-conflict studies primarily characterize or mitigate disagreement among client or task update directions before aggregation (Zheng et al., 2025; Yu et al., 2020; Liu et al., 2021; Sener & Koltun, 2018). We instead ask a complementary question: to what extent do these client-level directional relationships remain observable after the many-to-one aggregation mapping? To characterize this reduction in directional observability, we introduce Hidden Directional Opposition (HDO), which captures directional opposition that exists among client updates but becomes partially unobservable after aggregation. HDO therefore complements conventional disagreement analysis by explicitly characterizing how aggregation obscures client-level directional structure.

To analyze HDO in a shared coordinate system, we construct a geometry-aware tangent-space directional representation that reduces endpoint-dependent interference while retaining relative displacement information. Tangent-space and Riemannian formulations provide established mathematical tools for representing structured directional relationships (Absil et al., 2008; Bonnabel, 2013). We use this representation primarily as an analytical coordinate system for studying directional observability. Furthermore, once directional structures are treated as meaningful optimization information, subsequent adaptive transformations should avoid unnecessarily altering their directional semantics. We therefore introduce a direction-preserving adaptive optimization mechanism that adapts update magnitude without introducing additional directional rotation.

The main contributions of this work are:

1. We formulate federated aggregation from a directional observability perspective and introduce Hidden Directional Opposition (HDO) to characterize client-level directional opposition whose observability is reduced by aggregation. We further characterize this phenomenon across data heterogeneity, training dynamics, and network layers.
2. We construct a geometry-aware tangent-space directional representation that provides a shared analytical coordinate system for studying client directional relationships while reducing endpoint-dependent interference and retaining relative displacement information.
3. We introduce a direction-preserving adaptive optimization mechanism that preserves directional semantics during magnitude adaptation, providing an optimization realization for studying whether directional information can be retained during subsequent server updates.

2 Related Work

Federated aggregation and optimization. Federated optimization under heterogeneous client data has primarily focused on improving convergence and reducing optimization bias. FedAvg establishes the standard aggregation paradigm (McMahan et al., 2017), while FedProx, SCAFFOLD, FedNova, FedDyn, and FedAdam address client drift, objective inconsistency, or adaptive server optimization under non-IID training (Li et al., 2020; Karimireddy et al., 2020; Wang et al., 2020; Acar et al., 2021; Reddi et al., 2021); recent analyses further refine the characterization of federated heterogeneity (Zhu et al., 2021; Zeng et al., 2026). These studies mainly ask how aggregation affects optimization, whereas we ask what client-level directional structure remains observable in the aggregate representation.

Client disagreement and representation heterogeneity. A complementary line of work studies conflicting optimization directions, inter-client relationships, personalization, and representation inconsistency (Yu et al., 2020; Liu et al., 2021; Sener & Koltun, 2018; Yan et al., 2024; Li et al., 2021c; Dinh et al., 2020; Fallah et al., 2020; Li et al., 2021b; Collins et al., 2021; Li et al., 2021a). These approaches characterize or mitigate heterogeneity before or during client combination; HDO instead characterizes whether directional opposition present among client updates becomes partially unobservable through the many-to-one aggregation mapping.

Geometric representations. Tangent-space and Riemannian formulations provide established tools for representing and analyzing non-Euclidean optimization dynamics (Absil et al., 2008; Bonnabel, 2013). We use this geometry as an analytical representation for directional observability, rather than as a new manifold optimization objective.

3 Method

Our analysis proceeds in three steps. We first map client updates into a shared tangent space, then quantify negative directional interactions using HDO. Finally, we study A2 as a downstream optimization mechanism that adapts update magnitude while preserving the aggregated first-moment direction.

3.1 Aggregation Ambiguity in Federated Learning

Let $\Delta_{t,i} \in \mathbb{R}^d$ denote the local update generated by client i at communication round t. The server aggregates client updates through weighted averaging(McMahan et al., 2017):

$$G_t = \sum_{i=1}^{N} p_i \Delta_{t,i}$$

The aggregation weights satisfy $\sum_i p_i = 1$. From an optimization perspective, G_t is treated as the global update. However, from a representation perspective, the aggregation process defines a mapping from multiple client states to a single server-side representation:

$$\mathcal{A} : \{\Delta_{t,1}, \Delta_{t,2}, ..., \Delta_{t,N}\} \to G_t.$$

Such a many-to-one mapping implies that different client update configurations may produce identical or similar aggregate representations. For example, opposing updates $\Delta_{t,1} = u$, $\Delta_{t,2} = -u$ and inactive updates $\Delta_{t,1} = \Delta_{t,2} = 0$ can both lead to $G_t \approx 0$, despite representing fundamentally different client optimization states. We refer to this limitation as *aggregation ambiguity*, where the aggregate representation is insufficient to uniquely characterize underlying client directional structures.

3.2 Directional Representation Space for HDO Analysis

To analyze directional structures hidden by the aggregation mapping, we construct a geometry-aware directional representation. Let x_t denote the current global model, and let the endpoint of client i after local optimization be

$$z_{t,i} = x_t + \Delta_{t,i}.$$

Let $R_t = \|x_t\|_2$ denote the radius of the current global model. The projected state is defined as

$$y_{t,i} = R_t \frac{z_{t,i}}{\|z_{t,i}\|_2}.$$

Since the spherical space is non-Euclidean, we map the projected state into the tangent space at x_t through the logarithmic map (Absil et al., 2008; Bonnabel, 2013):

$$\xi_{t,i} = \mathrm{Log}_{x_t}(y_{t,i}).$$

The final directional representation is defined as

$$v_{t,i} = \frac{\xi_{t,i}}{R_t}.$$

Different from directly normalizing raw updates, this radius-normalized tangent representation retains relative displacement information in the tangent space while reducing endpoint-dependent interference. Therefore, it provides a shared analytical coordinate system for analyzing client directional structures and characterizing HDO.

3.3 Hidden Directional Opposition Characterization

Based on the above directional representation, we characterize directional structures that become partially unobservable after aggregation. The aggregated directional representation is

$$\bar{v}_t = \sum_i p_i v_{t,i}.$$

We first define the observable aggregate directional energy as

$$E_A = \|\bar{v}_t\|_2^2$$

which represents the net directional activity preserved after aggregation. In contrast, the overall client directional activity is defined as

$$E_Q = \sum_i p_i \|v_{t,i}\|_2^2$$

which captures the total directional activity before aggregation.

The gap $E_Q - E_A$ quantifies the overall difference between client-level directional activity and aggregate directional energy, which includes both aggregation-induced attenuation and pairwise directional interactions. To isolate the contribution of opposing directions, we define the negative directional opposition energy D_t^-:

$$D_t^- = 2 \sum_{i<j} p_i p_j \max(-\langle v_{t,i}, v_{t,j} \rangle, 0).$$

Unlike cosine-based disagreement measures used in prior conflict-based analyses (Yu et al., 2020; Zheng et al., 2025), which mainly describe angular relationships, D_t^- preserves tangent-space magnitude information and quantifies the directional opposition suppressed by aggregation.

To characterize the relative amount of hidden opposition, we further define the opposition ratio:

$$r_t^- = \frac{D_t^-}{E_A + D_t^- + \varepsilon_0}, \qquad \varepsilon_0 = 10^{-12}.$$

The pair (D_t^-, r_t^-) jointly defines Hidden Directional Opposition (HDO), which characterizes how aggregation transforms client-level directional relationships and reduces the observability of these directional structures in the aggregate representation.

3.4 Direction-Preserving Adaptive Optimization

Once client directional structures are treated as meaningful optimization information, subsequent adaptive transformations should avoid introducing unnecessary directional rotation. Conventional adaptive optimization methods such as AdaGrad and Adam, as well as federated adaptive optimizers, employ coordinate-wise adaptive scaling (Duchi et al., 2011; Kingma & Ba, 2015; Reddi et al., 2021), which may introduce directional rotation. A2 instead adapts magnitude through scalar factors. Let

$$\bar{\xi}_t = \sum_i p_i \xi_{t,i}$$

denote the aggregated tangent update at round t. A2 maintains a transported first moment

$$m_t = \beta_1 \widetilde{m}_{t-1} + (1-\beta_1)\bar{\xi}_t,$$

where $\widetilde{m}_{t-1}$ is the previous first moment transported to the current tangent space. The scalar second moment and bias-corrected moments are

$$a_t = \frac{\|\bar{\xi}_t\|_2^2}{R_t^2 + \varepsilon_0}, \quad v_t = \beta_2 v_{t-1} + (1-\beta_2) a_t,$$
$$\hat{m}_t = \frac{m_t}{1-\beta_1^t}, \quad \hat{v}_t = \frac{v_t}{1-\beta_2^t}.$$

Instead of applying coordinate-wise adaptive scaling, A2 uses the scalar gain

$$\gamma_t = \frac{\eta_{\text{server}}}{\sqrt{\hat{v}_t} + \varepsilon_{\text{A}}}.$$

yielding the raw tangent step

$$z_t^{\text{raw}} = \gamma_t \hat{m}_t.$$

When the angular step exceeds the prescribed bound, a scalar cap

$$\kappa_t = \min\left(1, \frac{\theta_{\max} R_t}{\|z_t^{\text{raw}}\|_2}\right)$$

is applied, giving

$$z_t = \kappa_t z_t^{\text{raw}} = \kappa_t \gamma_t \hat{m}_t.$$

Because both γ_t and κ_t are scalar factors, they adapt the step magnitude without introducing additional directional rotation. The server model is then updated through the exponential map,

$$x_{t+1} = \text{Exp}_{x_t}(z_t).$$

Thus, A2 separates adaptive magnitude modulation from directional semantics in the tangent update.

4 Experiments

Our experiments follow the paper's three steps. We evaluate whether HDO appears under heterogeneous client data and how it varies across training and network layers; compare tangent-space directions with endpoint and raw-update representations; and assess whether A2 preserves update direction while maintaining effective optimization.

4.1 Experimental Setup

We conduct the main experiments on CIFAR-10 and report additional CIFAR-100 results in the Appendix, using ResNet-18 as the backbone (Krizhevsky, 2009; He et al., 2016). We consider IID and Dirichlet-based non-IID partitions with $\alpha \in \{0.5, 0.1\}$, where smaller α indicates stronger client heterogeneity (Hsu et al., 2019). Unless otherwise specified, experiments use 10 clients, 100 communication rounds, full participation, and three random seeds.

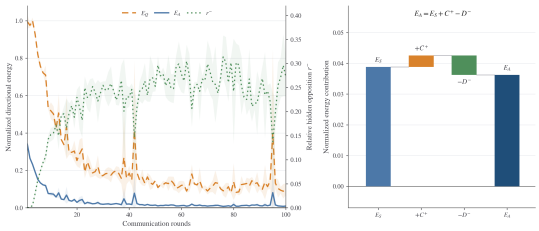
We compare FedAvg (McMahan et al., 2017), FedAdam (Reddi et al., 2021), the geometry-aware variant C2, and the proposed direction-preserving adaptive optimization mechanism A2. Here, C2 denotes non-adaptive common-tangent Log/Exp aggregation, $\bar{\xi}_t = \sum_i p_i \xi_{t,i}$ and $x_{t+1}^{\mathrm{C2}} = \mathrm{Exp}_{x_t}(\bar{\xi}_t)$, without server momentum or an adaptive second moment.For directional analysis, we report E_A, E_Q, $E_Q - E_A$, D^-, and r^- as defined in Section 3.3, together with final accuracy and AUC for optimization performance.

4.2 Evidence of Aggregation Ambiguity

Section 3.1 formulates federated aggregation as a many-to-one mapping

$$\mathcal{A} : \{\Delta_{t,i}\}_{i=1}^{N} \mapsto G_t.$$

raising the question of whether the aggregate representation preserves the directional activity contained in individual client updates.



(a) Directional activity and hidden opposition over training

(b) Directional energy decomposition

Figure 2: Evidence of aggregation ambiguity: client-level directional activity becomes partially unobservable after aggregation. Results are shown for C2 on CIFAR-10 with Dirichlet $\alpha = 0.1$, full participation, and seeds 3–5.

As shown in Figure 2, the client-level directional activity E_Q remains consistently larger than the observable aggregate energy E_A throughout training, indicating reduced directional observability in the aggregate representation. Figure 2(b) further decomposes the aggregate energy as $E_A = E_S + C^+ - D^-$, where D^- captures cancellation induced by opposing client directions. Hence, the observed gap is not explained by magnitude reduction from averaging alone, but also reflects directional interactions that become less observable through aggregation.

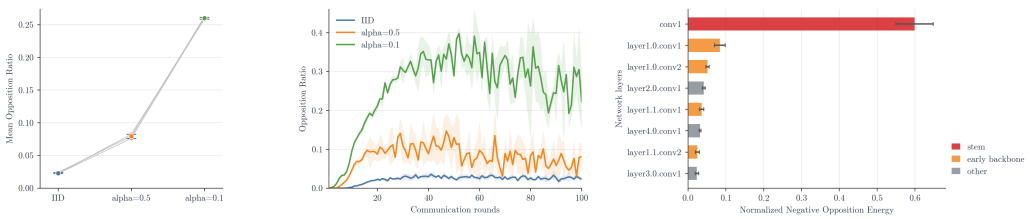
Importantly, Figure 2 is obtained with C2, which does not include the direction-preserving A2 mechanism. The observed aggregation ambiguity therefore arises from the aggregation representation rather than being introduced by A2.

4.3 Characterizing Hidden Directional Opposition

Having established aggregation ambiguity, we next examine whether Hidden Directional Opposition (HDO) is a systematic phenomenon rather than an isolated training effect. Figure 3 characterizes HDO from three complementary perspectives: heterogeneity dependence, temporal evolution, and layer-wise distribution.

As shown in Figure 3, hidden opposition becomes more pronounced under stronger client heterogeneity, HDO persists throughout training, and negative opposition energy is distributed non-uniformly across network layers. These observations indicate that HDO is neither confined to a particular convergence stage nor uniformly distributed across model components.

Table 1 provides the corresponding quantitative evidence. On CIFAR-10, for example, the opposition ratio r^- of C2 increases from 0.0310 under IID training to 0.0795 at $\alpha = 0.5$ and 0.2256 at $\alpha = 0.1$; A2 exhibits the same overall pattern. Similar increases are observed on CIFAR-100, showing that stronger heterogeneity is generally associated with a larger fraction of hidden directional opposition.



(a) Heterogeneity dependence of opposition ratio r^-

(b) Temporal evolution of opposition ratio r^-

(c) Layer-wise negative opposition energy D^-

Figure 3: Hidden Directional Opposition persists under heterogeneous federated training. Results are obtained with A2 on CIFAR-10 over seeds 3, 4, and 5.

Table 1: Quantitative characterization of Hidden Directional Opposition.

Dataset	Method	Setting	$E_Q - E_A$	D^-	r^-
CIFAR-10	C2	IID	4.349 ± 0.041	2.343 ± 0.055	0.0310 ± 0.0003
CIFAR-10	C2	Dir. $\alpha = 0.5$	3.016 ± 0.118	3.769 ± 0.575	0.0795 ± 0.0026
CIFAR-10	C2	Dir. $\alpha = 0.1$	1.509 ± 0.004	4.750 ± 0.464	0.2256 ± 0.0020
CIFAR-10	A2	IID	2.891 ± 0.039	1.139 ± 0.054	0.0229 ± 0.0009
CIFAR-10	A2	Dir. $\alpha = 0.5$	2.288 ± 0.065	2.395 ± 0.523	0.0793 ± 0.0032
CIFAR-10	A2	Dir. $\alpha = 0.1$	1.019 ± 0.002	2.307 ± 0.235	0.2599 ± 0.0016
CIFAR-100	C2	Dir. $\alpha = 0.5$	4.472 ± 0.061	5.977 ± 0.137	0.1072 ± 0.0030
CIFAR-100	C2	Dir. $\alpha = 0.1$	3.635 ± 0.111	5.959 ± 0.146	0.1613 ± 0.0006
CIFAR-100	A2	Dir. $\alpha = 0.5$	2.835 ± 0.056	1.361 ± 0.094	0.0593 ± 0.0037
CIFAR-100	A2	Dir. $\alpha = 0.1$	2.195 ± 0.077	1.713 ± 0.099	0.1192 ± 0.0031

HDO should not be conflated with conventional gradient-conflict analysis. Existing conflict-based methods primarily characterize whether client or task gradients disagree before aggregation (Zheng et al., 2025; Yu et al., 2020; Liu et al., 2021; Sener & Koltun, 2018). HDO instead asks whether such directional relationships remain observable in the aggregate representation after the many-to-one aggregation mapping

$$\mathcal{A} : \{\Delta_{t,i}\}_{i=1}^{N} \mapsto G_t.$$

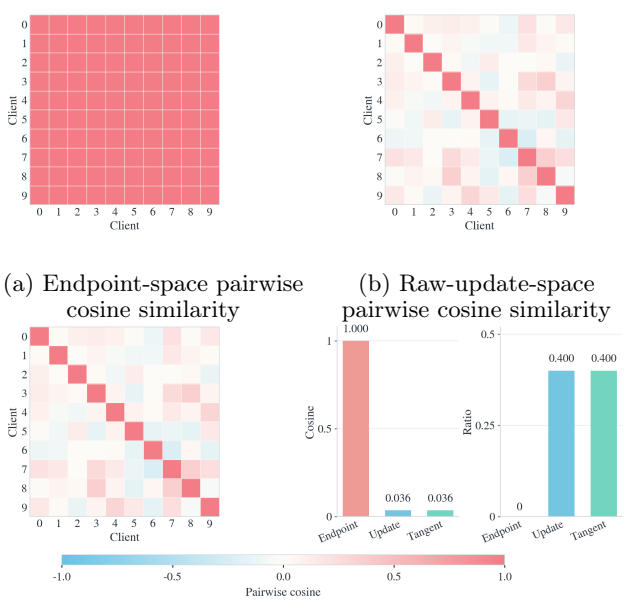
. It therefore complements disagreement measures by characterizing the reduction in directional observability induced by aggregation.

4.4 Geometry-aware Directional Observability

A key question is whether client directional structures can be analyzed directly in the original parameter space. As discussed in Section 3.2, endpoint representations mix model-state location with optimization-induced changes, which can obscure directional relationships. In the representative snapshot shown in Figure 4, we compare endpoint, raw-update, and tangent-space representations.

In this representative snapshot, endpoint states exhibit near-uniform similarity because their shared model-state location dominates the comparison, whereas raw updates expose the underlying client directional structure after this common component is removed. The tangent representation retains similar directional structure while providing a geometry-aware shared analytical coordinate system based on established tangent-space formulations (Absil et al., 2008; Bonnabel, 2013).

We therefore use tangent space as an analytical representation rather than claiming that it creates additional directional conflict or is universally superior to raw updates. Unlike simple unit normalization, the tangent representation retains relative displacement information in the tangent space while reducing endpoint-dependent interference, making it suitable for subsequent HDO analysis.



(a) Endpoint-space pairwise cosine similarity

(b) Raw-update-space pairwise cosine similarity

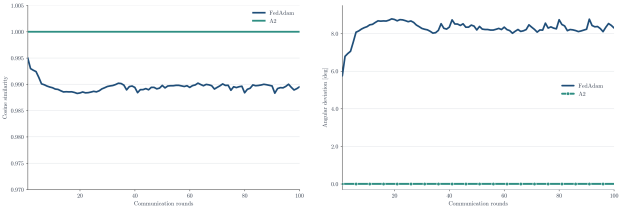
(c) Tangent-space pairwise cosine similarity

(d) Mean pairwise cosine and negative-pair ratio

Figure 4: Directional structure across endpoint, raw-update, and tangent representations. A representative CIFAR-10 A2 snapshot at Dirichlet $\alpha = 0.1$, seed 3, round 5 is shown.

4.5 Direction Preservation in Adaptive Optimization

Finally, we investigate whether adaptive optimization alters directional semantics. Although adaptive optimizers effectively adjust update magnitudes, coordinate-wise adaptive scaling used in methods such as Adam and FedAdam may rotate the bias-corrected first-moment direction (Kingma & Ba, 2015; Reddi et al., 2021). In general, this yields $z_t^{\mathrm{FedAdam}} \not\parallel \hat{m}_t$, because different coordinates may be rescaled by different factors.



(a) Cosine similarity to the first-moment direction

(b) Angular deviation from the first-moment direction

Figure 5: Direction-preserving optimization maintains directional semantics during adaptive updates.

We compare FedAdam and A2 in terms of directional changes during adaptive updates. FedAdam applies coordinate-wise scaling, which may rotate the bias-corrected first-moment direction, whereas A2 applies the scalar adaptive gain and angular cap:

$$z_t = \kappa_t \gamma_t \hat{m}_t.$$

Because γ_t and κ_t are scalar factors, A2 preserves the bias-corrected first-moment direction in the tangent adaptive step. The results demonstrate that adapting update magnitudes does not necessarily require changing directional semantics, and direction-preserving optimization can maintain directional structures while retaining optimization adaptability.

4.6 Optimization Compatibility

Finally, we examine whether preserving directional semantics remains compatible with effective federated optimization.

Table 2: Optimization compatibility (mean ± standard deviation over three seeds).

Method	IID Acc.	IID AUC	α=0.5 Acc.	α=0.5 AUC	α=0.1 Acc.	α=0.1 AUC
FedAvg	65.17 ± 1.07	63.48 ± 0.04	59.55 ± 0.28	56.72 ± 0.27	50.90 ± 0.89	49.38 ± 0.25
FedAdam	80.32 ± 0.17	63.82 ± 0.69	79.89 ± 0.64	68.70 ± 0.65	68.44 ± 0.95	62.16 ± 0.68
C2	65.46 ± 0.48	63.35 ± 0.27	60.10 ± 0.74	57.14 ± 0.08	51.41 ± 0.75	49.76 ± 0.29
A2	69.16 ± 0.26	64.52 ± 0.56	61.44 ± 0.26	58.70 ± 0.18	56.58 ± 0.19	53.11 ± 0.47

Across all evaluated partitions, A2 maintains effective optimization while preserving the directional property studied in Section 4.5. It remains stable as heterogeneity increases and outperforms FedAvg/C2 under the strongest non-IID setting, although FedAdam achieves higher predictive performance, indicating that direction preservation is compatible with effective federated optimization rather than establishing A2 as a universally superior optimizer.

5 Discussion

Our results support an observability perspective on federated aggregation. Conventional directional-disagreement analyses characterize relationships among client or task updates before aggregation (Zheng et al., 2025; Yu et al., 2020; Liu et al., 2021; Sener & Koltun, 2018), whereas HDO explicitly relates client directional opposition to the information retained by the aggregation mapping. It therefore provides a complementary view of heterogeneous federated optimization by distinguishing directional structure that exists at the client level from that which remains observable in the aggregate.

The tangent-space construction serves two roles: it provides a geometry-aware analytical space for HDO by reducing endpoint-dependent interference, and it supports A2 in adapting update magnitude while preserving the first-moment direction. We do not claim that tangent representations are universally superior to raw updates or that direction preservation necessarily improves predictive performance; rather, they provide controlled tools for analyzing how directional information is represented and transformed in federated optimization.

Several limitations remain. Our experiments focus primarily on vision benchmarks, and broader validation with larger client populations, models, modalities, and real-world federated systems is needed. The tangent-space analysis and pairwise HDO computation also introduce additional computational cost at larger scales. Moreover, direct HDO estimation requires access to individual client updates; under partial participation or secure aggregation, some directional information may be unavailable by design. Finally, we do not establish a universal causal relationship between HDO magnitude and predictive performance, leaving its connections to optimization stability and generalization as important directions for future work.

6 Conclusion

This work revisits federated aggregation from a directional observability perspective. We introduce Hidden Directional Opposition to characterize client-level directional opposition obscured by aggregation, use a geometry-aware tangent-space representation for analysis, and develop a direction-preserving adaptive mechanism. Experiments show that HDO persists under heterogeneous training and varies systematically across heterogeneity levels, training stages, and network layers. Overall, the aggregate update alone may not fully reflect the underlying client optimization state, and our results highlight directional information obscured through aggregation. Future work will extend this analysis to larger and privacy-constrained federated settings and further examine the relationship between HDO, optimization stability, and generalization.

References

Pierre-Antoine Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2008.

Durmus Alp Emre Acar, Yue Zhao, Jeff Matas, and Venkatesh Saligrama. Federated learning based on dynamic regularization. In *International Conference on Learning Representations*, 2021.

Keith Bonawitz, Hubert Eichner, Wolfgang Grieskamp, Dzmitry Huba, Alex Ingerman, Vladimir Ivanov, Chloe Kiddon, Jakub Konecný, Stefano Mazzocchi, Brendan McMahan, Timon Van Overveldt, David Petrou, Daniel Ramage, and Jason Roselander. Towards federated learning at scale: System design. In *Proceedings of Machine Learning and Systems*, volume 1, 2019.

Silvere Bonnabel. Stochastic gradient descent on riemannian manifolds. *IEEE Transactions on Automatic Control*, 58(9):2217–2229, 2013.

Liam Collins, Hamed Hassani, Aryan Mokhtari, and Sanjay Shakkottai. Exploiting shared representations for personalized federated learning. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 2089–2099, 2021.

Canh T. Dinh, Nguyen H. Tran, and Tuan Dung Nguyen. Personalized federated learning with moreau envelopes. In *Advances in Neural Information Processing Systems*, volume 33, pp. 21394–21405, 2020.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(61): 2121–2159, 2011.

Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. In *Advances in Neural Information Processing Systems*, volume 33, pp. 3557–3568, 2020.

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

Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. Measuring the effects of non-identical data distribution for federated visual classification. *arXiv preprint arXiv:1909.06335*, 2019.

Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurelien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista A. Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 14(1–2):1–210, 2021. doi: 10.1561/2200000083.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian U. Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 5132–5143, 2020.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

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

Fan Lai, Xiangfeng Zhu, Harsha V. Madhyastha, and Mosharaf Chowdhury. Oort: Efficient federated learning via guided participant selection. In *15th USENIX Symposium on Operating Systems Design and Implementation (OSDI 21)*, pp. 19–35. USENIX Association, 2021.

Qinbin Li, Bingsheng He, and Dawn Song. Model-contrastive federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10713–10722, 2021a.

Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In *Proceedings of Machine Learning and Systems*, volume 2, pp. 429–450, 2020.

Tian Li, Shengyuan Hu, Ahmad Beirami, and Virginia Smith. Ditto: Fair and robust federated learning through personalization. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 6357–6368, 2021b.

Xiaoxiao Li, Meirui Jiang, Xiaofei Zhang, Michael Kamp, and Qi Dou. FedBN: Federated learning on non-iid features via local batch normalization. In *International Conference on Learning Representations*, 2021c.

Yangxu Liao, Wenke Huang, Guancheng Wan, Jian Liang, Bin Yang, and Mang Ye. Splitting with importance-aware updating for heterogeneous federated learning with large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 37495–37510. PMLR, 2025. URL https://proceedings.mlr.press/v267/liao25c.html.

Bo Liu, Xianyuan Liu, Xiaojing Jin, Peter Stone, and Qiang Liu. Conflict-averse gradient descent for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pp. 1273–1282, 2017.

Takayuki Nishio and Ryo Yonetani. Client selection for federated learning with heterogeneous resources in mobile edge. In *2019 IEEE International Conference on Communications (ICC)*, pp. 1–7, 2019. doi: 10.1109/ICC.2019.8761315.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Sashank J. Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Karan Rush, Jakub Konecný, Sanjiv Kumar, and Brendan McMahan. Adaptive federated optimization. In *International Conference on Learning Representations*, 2021.

Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. In *Advances in Neural Information Processing Systems*, volume 31, pp. 525–536, 2018.

Zhenheng Tang, Yonggang Zhang, Shaohuai Shi, Xinmei Tian, Tongliang Liu, Bo Han, and Xiaowen Chu. Fedimpro: Measuring and improving client update in federated learning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/5b2fa23e4ef0f7ac6c4f01d7998e6237-Abstract-Conference.html.

Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H. Vincent Poor. Tackling the objective inconsistency problem in heterogeneous federated optimization. In *Advances in Neural Information Processing Systems*, volume 33, pp. 7611–7623, 2020.

Kunda Yan, Sen Cui, Abudukelimu Wuerkaixi, Jingfeng Zhang, Bo Han, Gang Niu, Masashi Sugiyama, and Changshui Zhang. Balancing similarity and complementarity for federated learning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 55739–55758. PMLR, 2024. URL https://proceedings.mlr.press/v235/yan24a.html.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 20063–20074, 2020.

Dun Zeng, Zenglin Xu, Shiyu Liu, Yu Pan, Qifan Wang, and Xiaoying Tang. On the power of adaptive weighted aggregation in heterogeneous federated learning and beyond. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 1081–1089. PMLR, 2025. URL `https://proceedings.mlr.press/v258/zeng25b.html`.

Wenkai Zeng, Nan Yang, Zhiyu Zhu, Zhibo Jin, and Dong Yuan. Disentangling federated learning heterogeneity: A dual-perspective analysis of quantifying skew versus scarcity. In *Proceedings of The 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pp. 5068–5076. PMLR, 2026. URL `https://proceedings.mlr.press/v300/zeng26b.html`.

Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Damon Civin, and Vikas Chandra. Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582*, 2018.

Hao Zheng, Zhigang Hu, Liu Yang, Meiguang Zheng, Aikun Xu, and Boyu Wang. ConFREE: Conflict-free client update aggregation for personalized federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 22875–22883, 2025. doi: 10.1609/aaai.v39i21.34449. URL `https://ojs.aaai.org/index.php/AAAI/article/view/34449`.

Hangyu Zhu, Jinjin Xu, Shiqing Liu, and Yaochu Jin. Federated learning on non-iid data: A survey. *Neurocomputing*, 465:371–390, 2021. doi: 10.1016/j.neucom.2021.07.098.

Appendix

A Experimental Details **14**

B Analysis of Aggregation Ambiguity and Directional Cancellation . . . **15**

C Additional Characterization of Hidden Directional Opposition **17**

C.1 Cross-dataset and Heterogeneity Validation 17

C.2 Temporal Evolution of Hidden Directional Opposition 17

C.3 Layer-wise Localization of Hidden Directional Opposition 18

C.4 Relationship with Directional Disagreement Measures 19

D Geometry-aware Representation Analysis **19**

E Partial Participation Extension: Information Recovery **21**

E.1 Directional Information Gap under Partial Participation 22

E.2 Geometry Stability under Partial Participation 23

E.3 Information Recovery through Auxiliary Observations 23

F Direction Preservation Analysis **24**

A Experimental Details

We conduct experiments on CIFAR-10 and CIFAR-100 image classification benchmarks (Krizhevsky, 2009). Both datasets consist of 32x32 RGB images with 50,000 training samples and 10,000 test samples. CIFAR-10 contains 10 classes, while CIFAR-100 contains 100 classes. We adopt a CIFAR-adapted ResNet-18 backbone for all experiments (He et al., 2016), where the original ImageNet-style stem is replaced with a 3x3 stride-1 convolution, the max-pooling operation is removed, and the final classifier is adjusted according to the number of target classes. The resulting models contain 10-way and 100-way classifiers for CIFAR-10 and CIFAR-100, respectively. The dataset and model configurations are summarized in Table A1.

Table A1: Dataset and model configurations.

Dataset	Input	Classes	Backbone	Training/Test Samples
CIFAR-10	32x32 RGB	10	CIFAR-adapted ResNet-18	50,000 / 10,000
CIFAR-100	32x32 RGB	100	CIFAR-adapted ResNet-18	50,000 / 10,000

All federated experiments use 10 clients and 100 communication rounds. Each client performs 5 local epochs per round using SGD with a learning rate of 0.01, momentum 0, and weight decay 0. The local batch size is set to 64. Models are evaluated on the complete test set after each communication round. We consider both IID and heterogeneous partitions. For heterogeneous settings, client data distributions are generated using label-skew Dirichlet partitions with concentration parameters $\alpha = 0.5$ and $\alpha = 0.1$ (Hsu et al., 2019). Unless otherwise specified, all multi-seed experiments are conducted with three random seeds (3, 4, and 5). The shared federated training configurations are summarized in Table A2.

Table A2: Federated training configurations.

Configuration	Value
Number of clients	10
Communication rounds	100
Local epochs	5
Batch size	64
Client optimizer	SGD
Client learning rate	0.01
Momentum	0
Weight decay	0
Seeds	3, 4, 5
Data partition	IID / Dirichlet non-IID
Dirichlet α	0.5, 0.1

For full-participation experiments, all 10 clients participate in each communication round. We evaluate FedAvg, FedAdam, C2, and A2 under the corresponding formal settings. For partial-participation analysis, we consider a controlled 50% participation scenario where 5 out of 10 clients are selected per round while keeping the same local training and evaluation protocols. A2PP, FMO, SW2, and RANDOM2 are evaluated under this setting. SW2 and RANDOM2 additionally query two auxiliary clients per round, resulting in 500 formal client computations and 200 auxiliary client computations over 100 rounds. These auxiliary-client evaluations are used to analyze directional information recovery under partial participation rather than introducing a new optimization objective.

We report standard optimization metrics including test accuracy and trajectory-level AUC. For directional analysis, we use the Hidden Directional Opposition (HDO) metrics defined in the main paper, including aggregate directional energy E_A, client directional activity E_Q, directional opposition energy D^-, and opposition ratio r^-. Direction-preservation experiments additionally evaluate momentum-update cosine similarity and angular deviation when corresponding diagnostic artifacts are available.

All experiments are implemented in PyTorch (Paszke et al., 2019), with torchvision used for standard vision dataset and preprocessing utilities. The geometry-aware aggregation, tangent-space transformation, HDO computation, and direction-preserving optimization components are implemented with repository-local code. No external manifold optimization toolbox, including Manopt or pymanopt, is used.

B Analysis of Aggregation Ambiguity and Directional Cancellation

A potential concern is whether the gap between client-level directional activity and observable aggregate energy is merely a trivial consequence of averaging. In this appendix, we provide a decomposition analysis to clarify that the proposed Hidden Directional Opposition (HDO) characterization does not rely on energy reduction alone, but specifically captures directional cancellation induced by opposing client updates. The weighted aggregation considered here is the standard federated averaging operation (McMahan et al., 2017).

Recall that the observable aggregate directional energy is defined as

$$E_A = \left\| \sum_i p_i v_i \right\|_2^2$$

while the overall client directional activity is

$$E_Q = \sum_i p_i \|v_i\|_2^2.$$

Expanding the aggregate energy gives

$$E_A = \sum_i p_i^2 \|v_i\|_2^2 + 2 \sum_{i<j} p_i p_j \langle v_i, v_j \rangle.$$

The first term reflects the weighted contribution of individual client directions after aggregation. The second term captures pairwise directional interactions among clients. When client directions are aligned, the interaction term remains positive and contributes to observable aggregate activity. In contrast, negatively correlated client directions introduce cancellation through negative inner products, reducing the observable energy after aggregation.

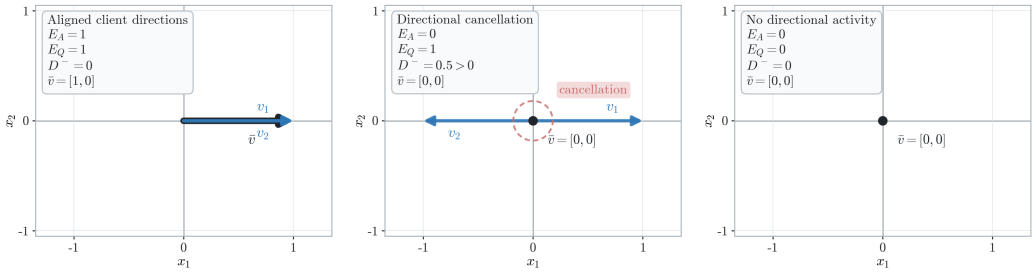
Therefore, the difference between client-level activity and aggregate observability contains two distinct effects: the inherent attenuation introduced by weighted aggregation and the directional cancellation caused by opposing client updates. The proposed negative directional opposition energy

$$D^- = 2 \sum_{i<j} p_i p_j \max(-\langle v_i, v_j \rangle, 0)$$

isolates the latter component by measuring only the contribution from negative directional interactions. Unlike simple norm reduction, D^- explicitly captures directional opposition that exists before aggregation but becomes less observable in the aggregate representation.

To illustrate this distinction, we provide a controlled synthetic example in Figure B1. Although different client configurations can produce similar aggregate representations, their underlying directional structures are fundamentally different. The example demonstrates that aggregation ambiguity arises not only from reduced update magnitude, but also from directional cancellation among heterogeneous clients.

This decomposition provides a mathematical basis for distinguishing HDO from ordinary averaging effects. The empirical observations in the main paper therefore reflect not merely



(a) Aligned client directions (b) Directional cancellation (c) No directional activity

Figure B1: Illustration of directional cancellation under federated aggregation.

reduced update magnitude, but the reduced observability of client-level directional structures caused by aggregation-induced directional cancellation.

C Additional Characterization of Hidden Directional Opposition

To further examine whether Hidden Directional Opposition (HDO) is a dataset-specific phenomenon or a general property induced by heterogeneous federated aggregation, we provide additional analyses across datasets, heterogeneity levels, training dynamics, and network components. These analyses complement the main experiments by evaluating the consistency and robustness of hidden directional structures under different federated learning conditions.

C.1 Cross-dataset and Heterogeneity Validation

We first evaluate HDO across different datasets and client heterogeneity levels. Table C1 summarizes the directional statistics under CIFAR-10 and CIFAR-100 with different data partitions. The results show that HDO consistently appears across datasets and heterogeneous settings. In particular, stronger distribution shifts generally lead to larger hidden directional opposition, indicating that heterogeneous client updates contain increasingly diverse directional structures that become partially unobservable after aggregation.

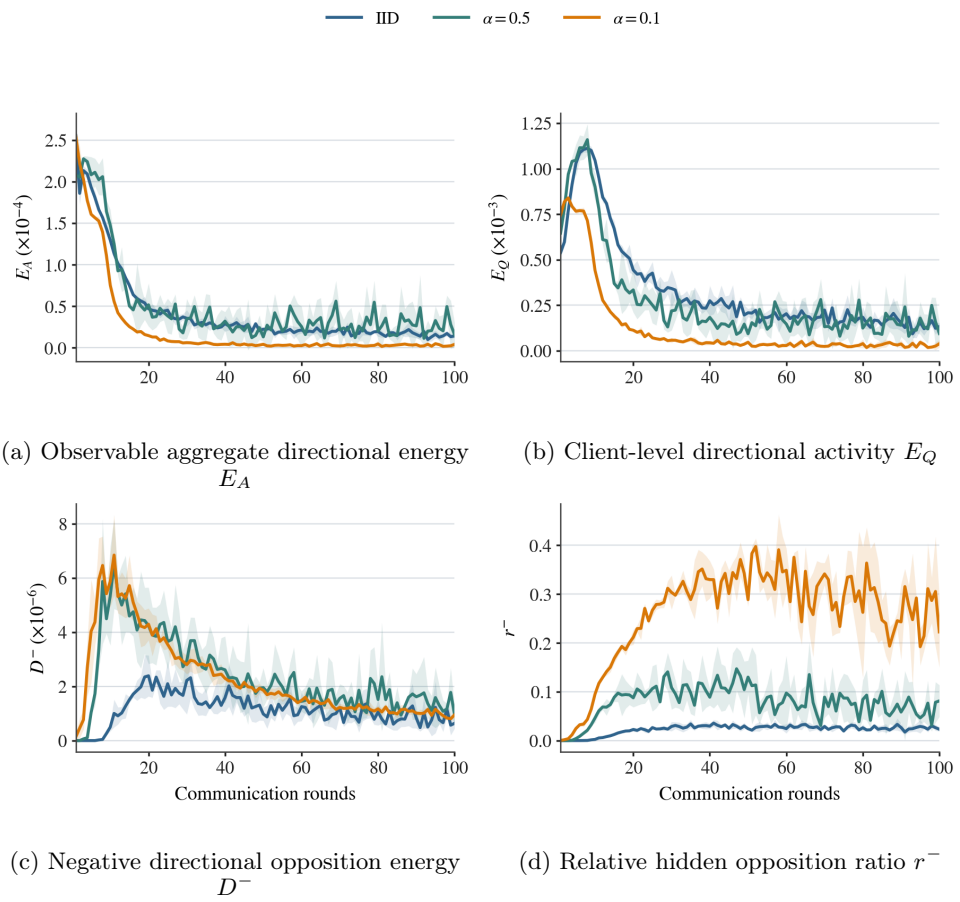
Table C1: Hidden directional opposition under different datasets and heterogeneity levels.

Dataset	Method	Setting	$E_Q - E_A$	D^-	r^-
CIFAR-10	C2	IID	4.349 ± 0.041	2.343 ± 0.055	0.0310 ± 0.0003
CIFAR-10	C2	Dirichlet $\alpha = 0.5$	3.016 ± 0.118	3.769 ± 0.575	0.0795 ± 0.0026
CIFAR-10	C2	Dirichlet $\alpha = 0.1$	1.509 ± 0.004	4.750 ± 0.464	0.2256 ± 0.0020
CIFAR-100	C2	Dirichlet $\alpha = 0.5$	4.472 ± 0.061	5.977 ± 0.137	0.1072 ± 0.0030
CIFAR-100	C2	Dirichlet $\alpha = 0.1$	3.635 ± 0.111	5.959 ± 0.146	0.1613 ± 0.0006
CIFAR-100	A2	Dirichlet $\alpha = 0.5$	2.835 ± 0.056	1.361 ± 0.094	0.0593 ± 0.0037
CIFAR-100	A2	Dirichlet $\alpha = 0.1$	2.195 ± 0.077	1.713 ± 0.099	0.1192 ± 0.0031

C.2 Temporal Evolution of Hidden Directional Opposition

To examine whether HDO only appears at specific training stages, we analyze the temporal evolution of directional opposition during federated optimization. Figure C1 presents the round-level trajectories of E_A, E_Q, D^-, and r^- under different levels of client heterogeneity.

The results show that HDO persists across the training process. Although the magnitude of directional opposition changes during optimization, non-zero HDO is consistently detected across training stages. This indicates that aggregation-induced directional cancellation is a persistent characteristic of heterogeneous federated optimization.



(a) Observable aggregate directional energy E_A

(b) Client-level directional activity E_Q

(c) Negative directional opposition energy D^-

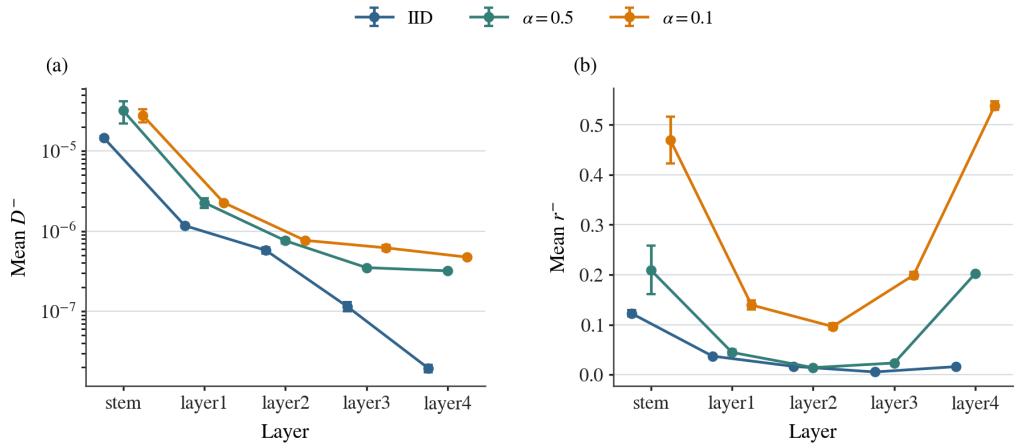
(d) Relative hidden opposition ratio r^-

Figure C1: Temporal evolution of directional statistics under different levels of client heterogeneity.

C.3 Layer-wise Localization of Hidden Directional Opposition

We further investigate whether hidden directional structures are uniformly distributed across model components. Figure C2 reports the layer-wise distribution of HDO statistics across different network blocks.

The results suggest that different network components contribute differently to client directional interactions. Rather than being uniformly distributed across all parameters, hidden directional structures exhibit layer-dependent characteristics, providing additional evidence that aggregation affects directional observability at the representation level.



(a) Mean negative directional opposition energy D^-

(b) Mean opposition ratio r^-

Figure C2: Layer-wise localization of Hidden Directional Opposition under different heterogeneity levels.

C.4 Relationship with Directional Disagreement Measures

Existing directional disagreement analyses typically characterize relationships among client updates before aggregation, such as pairwise angular similarity or negative directional interactions (Zheng et al., 2025). In contrast, HDO focuses on how these directional relationships are transformed by the aggregation mapping and whether they remain observable after aggregation.

To illustrate the complementary relationship between existing directional measures and HDO, Table C2 compares their analysis perspectives.

Table C2: Comparison between directional disagreement measures and HDO characterization.

Measure	Analysis Object	Before Aggregation	After Aggregation	Magnitude-aware
Pairwise cosine similarity	Client directional relationship	Yes	No	No
Negative pair ratio	Client directional opposition	Yes	No	Partial
HDO (D^-, r^-)	Aggregation-induced directional observability	Yes	Yes	Yes

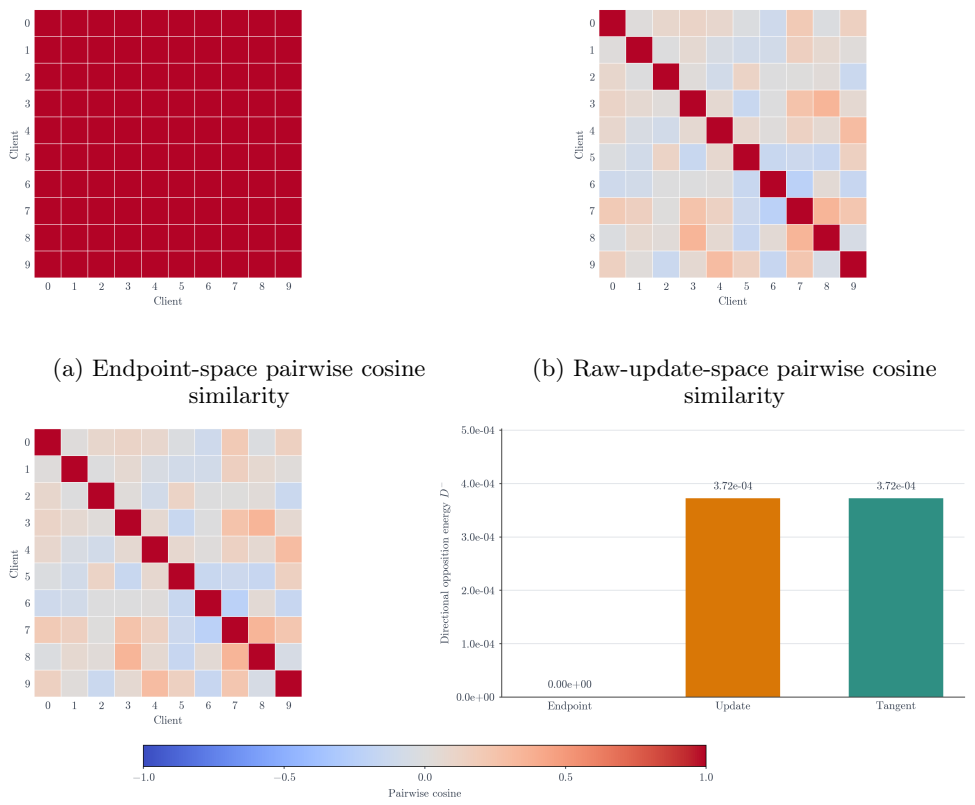
Unlike conventional disagreement metrics that focus on whether clients generate conflicting directions, HDO characterizes whether such directional relationships remain represented after the aggregation process. Therefore, HDO does not replace existing directional disagreement analysis, but provides a complementary perspective on the observability of client-level directional structures.

D Geometry-aware Representation Analysis

Different representations may expose different aspects of client directional structures after federated aggregation. In this appendix, we further analyze how representation choice influences the observability of client-level directional relationships. Specifically, we compare three representations: endpoint representation, raw update representation, and tangent-space directional representation.

The endpoint representation directly describes client model states after local optimization. However, endpoint states contain both optimization-induced changes and the geometric location of client models, which may entangle directional evolution with endpoint-dependent effects. The raw update representation removes the shared global model component by considering client displacements:

$$\Delta_{t,i} = z_{t,i} - x_t.$$



(a) Endpoint-space pairwise cosine similarity

(b) Raw-update-space pairwise cosine similarity

(c) Tangent-space pairwise cosine similarity

(d) Directional opposition strength across representations

Figure D1: Representation sensitivity of client directional structure across endpoint, raw-update, and tangent spaces.

Although update representations provide a more direct description of local optimization changes, they still reside in Euclidean parameter space, where directional relationships may be affected by the underlying geometry of model states. To obtain a geometry-aware directional representation, we map client states into a shared tangent space. Given the projected client state $y_{t,i}$, the tangent representation is obtained through the logarithmic map:

$$\xi_{t,i} = \mathrm{Log}_{x_t}(y_{t,i})$$

and the radius-normalized tangent representation is defined as:

$$v_{t,i} = \frac{\xi_{t,i}}{R_t}.$$

Tangent-space and Riemannian formulations provide established tools for analyzing optimization dynamics in non-Euclidean spaces (Absil et al., 2008; Bonnabel, 2013). Different from introducing a new geometric optimization framework, our goal is to construct an appropriate analytical space for studying how client directional structures are represented after aggregation.

The figure compares endpoint, raw-update, and tangent-space representations through client-level directional analysis. Endpoint representation contains strong location-dependent effects, resulting in highly aligned client states despite different underlying optimization behaviors. Raw-update representation removes the shared global model component and reveals more informative directional relationships. The tangent representation provides a geometry-

aware analytical space that reduces endpoint-dependent interference while retaining relative tangent-space displacement information.

Table D1: Quantitative comparison of directional representations.

Representation	Mean Pairwise Cosine	Mean Pairwise L2	Negative Pair Ratio	Opposition Strength
Endpoint	0.999988	0.151934	0.000	0.000000
Raw Update	0.035747	0.151934	0.400	0.034787
Tangent	0.035747	0.151934	0.400	0.034787

Mean pairwise cosine measures client directional similarity, while negative pair ratio and opposition strength quantify the presence of opposing directional structures. The L2 distances are computed through offline recomputation from the saved representation vectors and are included to analyze Euclidean distance sensitivity. All non-L2 metrics are directly extracted from existing representation artifacts.

The results demonstrate that representation choice strongly affects the observable client directional structures. Endpoint representations exhibit near-identical cosine similarity and no observable directional opposition, indicating that model-state location can dominate directional analysis. In contrast, raw update representations reveal substantially different client directional relationships after removing the shared global model component. Interestingly, endpoint and raw update representations have identical pairwise Euclidean distances because subtracting the same global model state corresponds to a translation operation, showing that Euclidean distance alone cannot characterize directional observability.

The tangent representation produces directional statistics consistent with raw updates while providing a geometry-aware coordinate system for analysis. This indicates that the tangent representation does not create additional directional conflicts; instead, it reduces endpoint-dependent interference and provides a more suitable space for characterizing Hidden Directional Opposition (HDO).

E Partial Participation Extension: Information Recovery

Partial participation and client selection are widely studied in federated learning to reduce communication and computation costs and to select resource-compatible or informative participants (Kairouz et al., 2021; Bonawitz et al., 2019; Nishio & Yonetani, 2019; Lai et al., 2021). However, from the perspective of directional observability, partial participation introduces an additional challenge: the server receives only incomplete observations of client-level optimization states. Consequently, the aggregate representation may preserve fewer directional structures than the full-participation setting.

In this appendix, we investigate how partial participation affects directional observability and whether additional client observations can recover part of the missing directional information. We consider a controlled 50% participation setting where 5 out of 10 clients are selected in each communication round. The local optimization protocol remains identical to the full-participation setting. We compare A2PP, RANDOM2, SW2, and FMO under this setting. RANDOM2 and SW2 additionally query two auxiliary clients per round with identical observation budgets, enabling a matched comparison of auxiliary observation strategies. FMO is included as an oracle intervention with access to full first-moment information and is therefore not directly resource matched with other methods.

Table E1 summarizes the client observation budget and optimization performance under partial participation. The table reports mean and standard deviation over three random seeds. SW2 and RANDOM2 use identical formal and auxiliary client budgets, allowing their comparison to focus on auxiliary selection effects rather than additional observation quantity.

The table summarizes client observation budgets and optimization performance under partial participation. SW2 and RANDOM2 use identical total client-observation budgets, enabling matched comparison of auxiliary client selection effects. FMO is reported as an oracle intervention with full first-moment information and is not resource matched with the

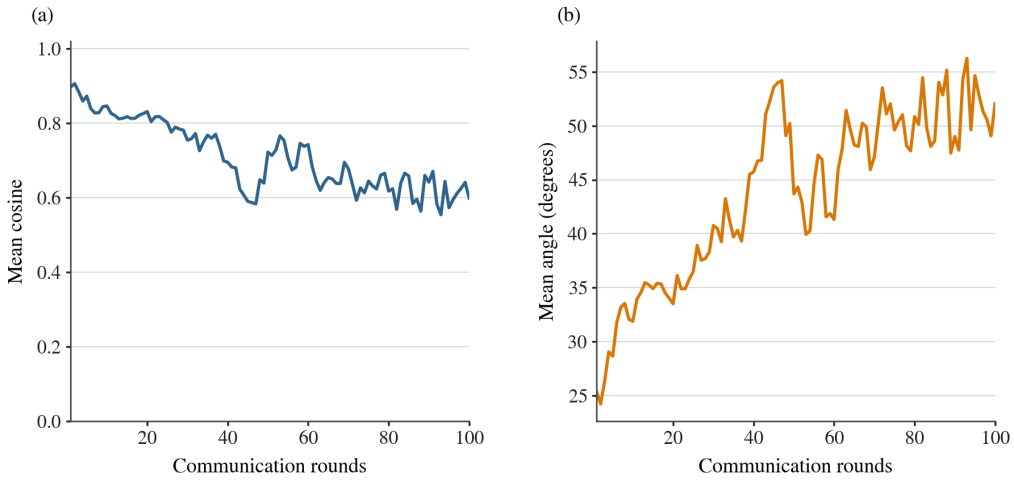
Table E1: Partial participation settings and optimization results.

Method	Partition	Formal Clients	Auxiliary Clients	Total Observations	Accuracy	AUC
Full A2	$\alpha = 0.5$	10/10	0	10/10	0.6144 ± 0.0026	0.5870 ± 0.0017
Full A2	$\alpha = 0.1$	10/10	0	10/10	0.5658 ± 0.0019	0.5311 ± 0.0047
A2PP	$\alpha = 0.5$	5/10	0	5/10	0.5712 ± 0.0127	0.5480 ± 0.0087
A2PP	$\alpha = 0.1$	5/10	0	5/10	0.4683 ± 0.0136	0.4477 ± 0.0020
RANDOM2	$\alpha = 0.5$	5/10	2/10	7/10	0.5887 ± 0.0340	0.5538 ± 0.0129
RANDOM2	$\alpha = 0.1$	5/10	2/10	7/10	0.4915 ± 0.0409	0.4605 ± 0.0017
SW2	$\alpha = 0.5$	5/10	2/10	7/10	0.5940 ± 0.0232	0.5534 ± 0.0078
SW2	$\alpha = 0.1$	5/10	2/10	7/10	0.5313 ± 0.0282	0.4779 ± 0.0026
FMO (oracle)	$\alpha = 0.5$	5/10	5/10	10/10	0.6081 ± 0.0206	0.5627 ± 0.0084
FMO (oracle)	$\alpha = 0.1$	5/10	5/10	10/10	0.5305 ± 0.0190	0.4873 ± 0.0014

other methods. Direction-recovery metrics are not included because existing artifacts do not provide a formally defined full-reference comparison for every method, partition, and seed.

E.1 Directional Information Gap under Partial Participation

To quantify the directional information gap introduced by partial participation, we compare the estimated directional state under 50% participation with the corresponding full-participation reference. Specifically, we analyze the cosine similarity and angular deviation between the partial directional estimate and the full-participation reference throughout training.



(a) Cosine similarity to the full-participation reference

(b) Angular deviation from the full-participation reference

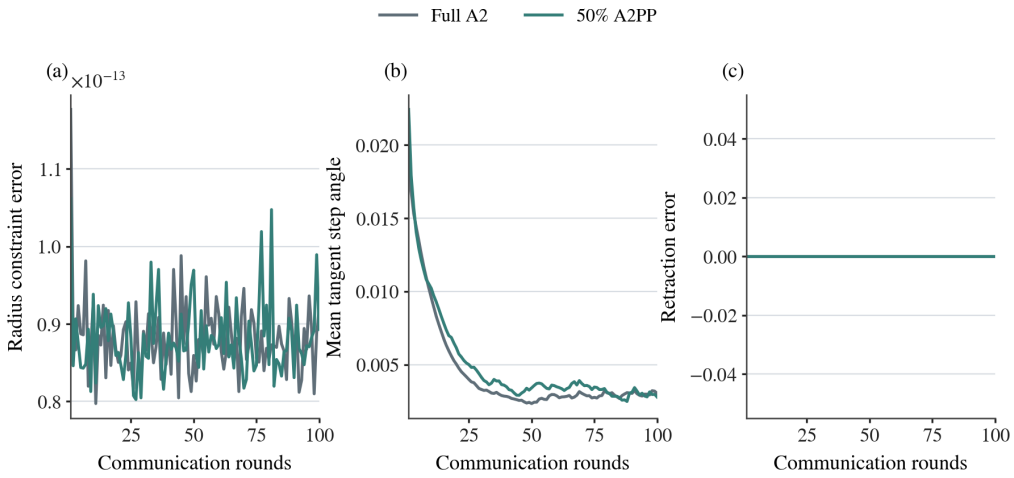
Figure E1: Direction estimation quality under partial participation. The recorded trajectory uses Dirichlet $\alpha = 0.5$, 50% participation, and 100 communication rounds.

The cosine similarity and angular deviation between partial-participation directional estimates and the full-participation reference under the recorded $\alpha = 0.5$ setting characterize the directional discrepancy introduced by incomplete client observation. The results show that partial participation introduces observable directional deviation during federated optimization.

The results indicate that partial participation does not merely reduce the number of available client updates, but also changes the directional state observable from the server perspective. This observation is consistent with the representation perspective of the main paper, where aggregation determines which client-level directional structures remain accessible after transformation.

E.2 Geometry Stability under Partial Participation

Since the geometry-aware representation is designed to analyze directional structures after aggregation transformation, we further verify whether partial client observation introduces numerical instability into the geometric operations. We evaluate radius constraint error, tangent step angle, and retraction error under full participation and 50% participation.



(a) Radius-constraint error (b) Mean tangent-step angle (c) Retraction error

Figure E2: Numerical stability of geometry-aware operations under partial participation.

The radius constraint error, tangent step angle, and retraction error of the geometry-aware representation are evaluated under full participation and 50% participation. The recorded retraction error remains zero throughout training, while the radius constraint and tangent-step dynamics remain stable, indicating that partial client observation does not compromise the numerical validity of the geometry-aware representation.

These results suggest that the directional discrepancy observed under partial participation is primarily caused by incomplete client observation rather than instability of the geometry-aware transformation itself.

E.3 Information Recovery through Auxiliary Observations

The previous analyses demonstrate that partial participation introduces a directional observation gap while preserving the numerical stability of the geometry-aware representation. We next investigate whether additional client observations can recover part of the missing directional information.

To separate the contribution of additional observations from auxiliary client selection, we decompose the improvement of SW2 relative to A2PP into two components:

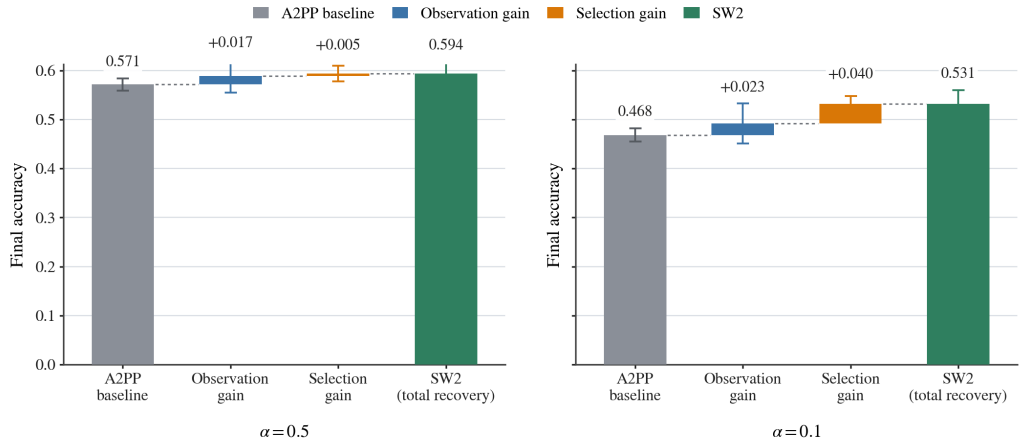
$$\Delta_{\text{obs}} = \text{RANDOM2} - \text{A2PP}$$

which measures the gain from acquiring additional auxiliary observations, and

$$\Delta_{\text{sel}} = \text{SW2} - \text{RANDOM2}$$

which measures the additional contribution of auxiliary client selection under the same observation budget.

The improvement obtained by auxiliary-client strategies is decomposed into additional observation gain and auxiliary selection gain. Since SW2 and RANDOM2 use identical client-



(a) Dirichlet $\alpha = 0.5$ (b) Dirichlet $\alpha = 0.1$

Figure E3: Decomposition of auxiliary-observation gains under partial participation. The bars report final-accuracy differences.

observation budgets, their difference isolates the contribution of auxiliary selection rather than additional computation. The analysis separates information acquisition effects from selection effects and highlights partial participation as an information observability limitation.

Under $\alpha = 0.5$, SW2 exceeds RANDOM2 in final accuracy by $+0.0053 \pm 0.0160$, while the AUC difference is -0.0004 ± 0.0074. Under stronger heterogeneity ($\alpha = 0.1$), the final-accuracy difference increases to $+0.0398 \pm 0.0166$, with an AUC difference of $+0.0174 \pm 0.0009$. These results should be interpreted as partition-dependent auxiliary-selection effects rather than unconditional optimization improvements, since SW2 and RANDOM2 use the same formal and auxiliary client budgets.

F Direction Preservation Analysis

Adaptive optimization methods adapt coordinate-wise or global update magnitudes using historical gradient statistics (Duchi et al., 2011; Kingma & Ba, 2015; Reddi et al., 2021). However, coordinate-wise adaptive transformations may also modify the relative scaling among update dimensions, potentially rotating the bias-corrected first-moment direction. In the context of this work, such directional changes are important because client update directions represent optimization trajectories that contain information beyond the final update magnitude. Therefore, we analyze whether adaptive transformations preserve the direction of the bias-corrected first moment.

A2 applies scalar adaptive modulation to the bias-corrected first moment, while preserving its tangent-space direction. Unlike coordinate-wise adaptive preconditioning, the A2 transformation uses a scalar adaptive gain followed, when necessary, by a scalar angular cap.

Direction-preservation property. Let $\hat{m}_t$ denote the bias-corrected first moment and define the A2 tangent update as

$$z_t = \kappa_t \gamma_t \hat{m}_t,$$

where $\gamma_t > 0$ is the scalar adaptive gain and $\kappa_t \in (0, 1]$ is the angular-cap factor. For any nonzero $\hat{m}_t$,

$$\frac{z_t}{\|z_t\|_2} = \frac{\hat{m}_t}{\|\hat{m}_t\|_2},$$

and therefore

$$\cos(z_t, \hat{m}_t) = 1.$$

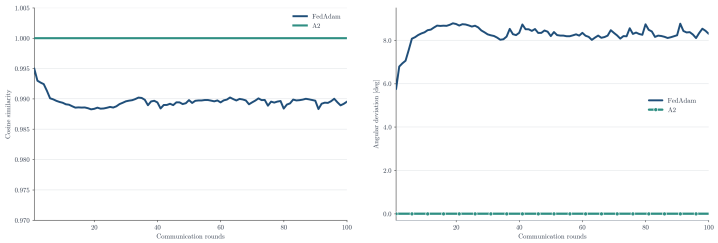
Hence, both scalar adaptivity and angular clipping modify only the tangent-step magnitude and do not introduce additional directional rotation.

To quantify directional consistency, we measure the cosine similarity and angular deviation between the adaptive tangent update and the bias-corrected first-moment direction. We compare A2 with FedAdam under the available direction-preservation diagnostic setting. The results are summarized in Table F1.

Table F1: Direction-preservation statistics between adaptive tangent updates and the bias-corrected first-moment direction.

Method	Setting	Mean cosine similarity	Mean angular deviation (degrees)
FedAdam	CIFAR-10, Dirichlet $\alpha = 0.5$	0.989548	8.281059
A2	CIFAR-10, Dirichlet $\alpha = 0.5$	1.000000	0.000003

The results show that A2 maintains near-perfect consistency with the bias-corrected first-moment direction under the evaluated setting, with an angular deviation close to zero. In contrast, FedAdam introduces measurable angular changes due to coordinate-wise adaptive transformations. These observations indicate that adaptive magnitude adjustment and directional preservation can be achieved simultaneously when the adaptive transformation is designed to avoid unnecessary directional rotation.



(a) Cosine similarity to the first-moment direction

(b) Angular deviation from the first-moment direction

Figure F1: Directional consistency of adaptive updates.

The cosine similarity and angular deviation between adaptive tangent updates and the bias-corrected first-moment direction are compared between FedAdam and A2 throughout training. A2 maintains directional consistency through scalar adaptive scaling and scalar angular clipping, whereas FedAdam introduces additional angular deviations through coordinate-wise adaptive transformations.

The purpose of A2 is not to replace existing adaptive optimization strategies or maximize predictive performance alone. Instead, A2 provides one realization of direction-preserving optimization by maintaining update directional semantics while adapting optimization magnitudes. This analysis further supports the main paper's perspective that directional structures can serve as meaningful optimization information beyond conventional magnitude-based update analysis.

AI Use Statement

The authors used generative AI tools for document-format conversion, reference-format checking, and language-level editing. All scientific claims, mathematical definitions, experimental results, figures, and final wording were reviewed and verified by the authors, who take responsibility for the final content.