THE AGREEMENT CONE: WHEN CAN ONE SERVER STEP IMPROVE EVERY CLIENT IN FEDERATED LEARNING?

Anonymous authors
Paper under double-blind review

ABSTRACT

Federated learning aggregates client updates into a single server step. In heterogeneous data, this step can decrease the average objective while increasing the loss of individual clients. Existing analyses describe heterogeneity with scalar summaries, such as mean pairwise gradient cosine, gradient variance, and bounded-dissimilarity constants. These summaries track progress on the average objective. They do not determine whether some direction decreases every client objective at once, nor how robust such a common improvement is to aggregation noise. We introduce the *agreement cone*, the set of directions that give a first-order decrease for every client's loss. We characterize it by its normalized Gaussian volume α_t and by the statistical dimension Δ_t of the cone generated by the client gradients. Both quantities depend only on the Gram matrix of the $n \times n$ client grade, not on the dimension of the model. We prove that scalar heterogeneity summaries can rank client populations in the wrong order: for every $n \geq 10$, they prefer a population with no common descent direction over one that admits it. Under spherically symmetric aggregation noise, we show that the probability of improving all clients is lower-bounded by the volume of a margin-shrunken agreement cone, and that this bound becomes exact in the noise-dominated limit. Because this volume can be exponentially small in n, we derive an unbiased Monte Carlo estimator of Δ_t whose sample complexity is independent of the model dimension. On FMNIST and CIFAR-10 with Dirichlet label skew, the cone statistics track the fraction of improved clients under strong heterogeneity, at a strength comparable to mean cosine, and add a small signal beyond it, as our analysis predicts. The proposed work of this paper provides a geometric complement to average-case heterogeneity measures and multi-objective feasibility diagnostics.

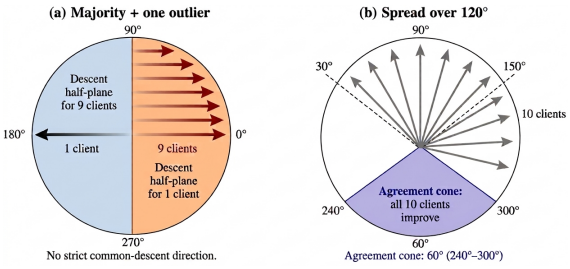
1 INTRODUCTION

Federated learning (FL) trains a shared model on clients that keep their data local (McMahan et al., 2017; Kairouz & McMahan, 2021). Client data are rarely identically distributed (Zhao et al., 2018; Hsu et al., 2019; Li et al., 2022), so one server step can lower the average loss while raising the loss of some clients. A client harmed round after round may stop participating, and fairness-oriented methods exist largely to prevent this (Li et al., 2019a; Mohri et al., 2019). Yet standard analyses measure heterogeneity with a single scalar, such as the bounded-dissimilarity constant (Li et al., 2020), gradient diversity (Yin et al., 2018), gradient variance, or the optimum gap Γ (Li et al., 2019b). These constants enter the convergence bounds of FedAvg and its variants (Karimireddy et al., 2020; Wang et al., 2020; Reddi et al., 2020; Acar et al., 2021), and empirical work uses pairwise cosine in the same spirit (Sattler et al., 2020; Luo et al., 2021). All of them describe how fast the *average* objective falls. None answers a different question: can one server step improve *every* participating client? The reason is structural. For unit-norm client gradients $_1, \ldots, _n$,

$$\Big\| \sum_{i=1}^{n} {}_i \Big\|^2 \;=\; n + n(n-1)\,\overline{\cos}. \tag{1}$$

Mean cosine, gradient variance, and bounded dissimilarity are all fixed by this one length. Each compresses the Gram matrix $G_{ij} = \langle {}_i, {}_j \rangle$ into one number. Yet a direction d improves client i only

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.



	(a)	(b)
mean pairwise cosine	**0.600**	0.584
$\|\sum_i i\|$	**8.000**	7.910
gradient variance	**0.360**	0.374
bounded dissimilarity B^2	**1.563**	1.599
agreement cone		60°

Figure 1: **Scalar measures can rank client populations backwards.** Ten unit gradients in $\mathbb{R}^2$. In (a), nine gradients point at $0°$ and one at $180°$, so no direction strictly improves all clients. In (b), the gradients span $120°$ and leave a $60°$ agreement cone, which lies opposite the gradients because descent needs $\langle g_i, d\rangle < 0$. Yet every scalar measure ranks (a) as more homogeneous. The reversal holds for all $n \geq 10$ and grows with n (Section 6.1).

if $\langle g_i, d\rangle < 0$, so each client rules out half of all directions. Whether any direction survives all n cuts depends on the joint geometry of the gradients, which a single number cannot record.

Figure 1 makes this concrete. In (a), every scalar reports low heterogeneity, yet no nonzero direction strictly improves every client. In (b), every scalar looks worse, yet all ten clients can improve at once. The average loss can thus fall steadily while a minority client is harmed in every round. Multi-objective methods address related conflicts. MGDA returns a common descent direction whenever one exists (Désidéri, 2009; Fliege & Svaiter, 2000; Sener & Koltun, 2018), FedMGDA+ brings this test to FL (Hu et al., 2022), and gradient surgery modifies conflicting updates (Yu et al., 2020; Liu et al., 2021). These tools are binary or algorithmic. They do not measure *how much* room for common descent remains, so they cannot distinguish a knife-edge round from a comfortable one.

In this work, we measure that room directly. We call the set of directions that improve every client the *agreement cone*. Its normalized volume α_t is a Gaussian orthant probability of the $n \times n$ Gram matrix (Slepian, 1962), independent of the model dimension. Because α_t can be exponentially small, we also use the statistical dimension Δ_t of the cone the gradients generate (Amelunxen et al., 2014) as a bounded surrogate, reported together with α_t and the min-norm test. Our goal is diagnostic, not a new optimizer. In theory, the cone captures what scalars cannot. In practice, on Dirichlet partitions, it tracks client improvement comparably to mean cosine and adds a small signal beyond it. This is consistent with our theory: these partitions are expected to yield near-exchangeable Gram matrices rather than the structured conflict of Figure 1. To sum up, the contributions of this study are as follows:

- We introduce the agreement cone and characterize it through the client-gradient Gram matrix alone, with volume α_t as the primary quantity and statistical dimension Δ_t as a bounded surrogate (Section 3).
- We prove that scalar summaries cannot characterize common-descent geometry. For every $n \geq 10$, they rank a population with an empty agreement cone above one with a $60°$ cone, and the gap widens with n (Theorem 4).
- We show that, under spherically symmetric aggregation noise, if the mean aggregate lies in a margin-shrunken cone, all clients improve with probability at least its volume; the bound becomes exact in the noise-dominated limit. The required margin scales with update and gradient magnitudes (Theorem 5, Proposition 6).
- We give an unbiased estimator of Δ_t whose sample complexity is independent of the model dimension (Section 4). On FMNIST and CIFAR-10, it matches closed-form values, and the cone statistics track client improvement under strong heterogeneity (Section 6).

2 Related Work

Heterogeneity as a scalar: FedAvg (McMahan et al., 2017) has been analysed under non-identical client data along several lines: bounded dissimilarity (Li et al., 2020), control variates for client

drift (Karimireddy et al., 2020), objective-inconsistency corrections (Wang et al., 2020), server-side adaptivity (Reddi et al., 2020), and dynamic regularization (Acar et al., 2021). In every case heterogeneity enters the bound through one scalar: the dissimilarity constant B (Yin et al., 2018), the gradient variance σ^2, or the optimum gap Γ (Li et al., 2019b). Surveys (Kairouz & McMahan, 2021; Wang et al., 2021a) treat these as standard vocabulary. Each summarizes the same Gram matrix, and by Equation (1) each is tied to the length of the average gradient. So each describes how fast the average loss falls, not who is helped along the way. A parallel empirical line uses pairwise cosine as a diagnostic (Zhao et al., 2018; Hsu et al., 2019; Sattler et al., 2020; Luo et al., 2021), but always pairwise or averaged. What the full set of client gradients permits jointly is not asked. Appendix B expands on both lines.

Gradient conflict and conic geometry: The closest work treats each client as a separate objective. MGDA (Désidéri, 2009; Fliege & Svaiter, 2000) computes the minimum-norm point of the convex hull of objective gradients, giving a common descent direction; FedMGDA+ (Hu et al., 2022) imports this test into federated learning. Hu et al. (2022) also measures the fraction of clients whose loss falls each round. Their test is binary: it detects whether common descent exists, not how much room surrounds it. Our measure instead comes from conic integral geometry. Amelunxen et al. (2014) define statistical dimension, while Slepian (1962) and Plackett (1954) establish the orthant-probability identity and correlation monotonicity. To our knowledge, conic geometry has not previously been used to quantify client heterogeneity in FL. Appendix B gives the full account.

3 Problem Setup

3.1 The round-t picture

There are n participating clients with losses $F_i : \mathbb{R}^d \to \mathbb{R}$, and the server minimizes

$$F(w) = \sum_{i=1}^{n} p_i F_i(w), \qquad p_i \geq 0, \quad \sum_i p_i = 1. \tag{2}$$

At round t the server holds w_t, broadcasts it, and each client returns an update. Let $g_i \neq 0$ denote the direction client i effectively pulls in: for one local step $g_i = \nabla F_i(w_t)$; for several local steps it is the negated local displacement, $g_i = (w_t - w_t^i)/\eta E$. Nothing below depends on which is meant, so we call g_i the client gradient throughout. We study one round and normalize,

$$_i = \frac{g_i}{\|g_i\|}, \qquad G_{ij} = \langle _i, _j \rangle. \tag{3}$$

So $G \in \mathbb{R}^{n \times n}$ is the client-gradient Gram matrix, and its off-diagonal entries are the pairwise cosine similarities that every scalar measure in Section 1 averages. Everything below depends on G, not on the model dimension d.

3.2 From gradients to the agreement cone

For a server direction d, the first-order change in client i's loss is

$$F_i(w_t + \eta d) \approx F_i(w_t) + \eta \langle g_i, d \rangle, \tag{4}$$

so strict first-order descent for every client requires $\langle g_i, d \rangle < 0$ for all i. This defines the *agreement cone*,

$$\mathcal{C}_t = \big\{ d \in \mathbb{R}^d : \langle g_i, d \rangle < 0 \text{ for all } i = 1, \ldots, n \big\}. \tag{5}$$

It depends only on the normalized directions, and it is an open convex cone. It lies opposite the gradients, since descent needs a negative inner product; Figure 1 illustrates this. Where closedness is needed, for polarity and statistical dimension, we use $\overline{\mathcal{C}_t}$, which changes neither the spherical volume nor any result below. Equation (4) is used only for intuition. Every result we prove uses L-smoothness instead, which replaces $\langle g_i, d \rangle < 0$ by a strict version carrying a margin fixed by the step size (Theorem 5); the geometry is the same, and the margin vanishes as $\eta \to 0$.

By Gordan's theorem, $\mathcal{C}_t$ is empty exactly when a nonzero nonnegative combination of the client gradients vanishes, equivalently when the origin lies in $\mathrm{conv}\{ _1, \ldots, _n \}$. Then no nonzero direction is

a strict first-order descent direction for every client at once, which rules out a common strict-descent step under the smoothness conditions used later. This is the feasibility condition the minimum-norm test of FedMGDA+ (Hu et al., 2022) detects, and it coincides with Pareto stationarity. Positive pairwise cosines are sufficient for nonemptiness (Lemma 7), but they are not necessary, and they say nothing about how much common-descent space remains.

3.3 Measuring the cone

Cone volume: The quantity we care about is the fraction of directions that help every client. This is the normalized spherical volume of the agreement cone,

$$\alpha_t = \frac{\sigma(\mathcal{C}_t)}{\sigma(\mathbb{S}^{d-1})} =_{z\sim\mathcal{N}(0,I_d)} \big[\langle_i, z\rangle < 0 \;\forall i\big]. \tag{6}$$

The second equality holds because a standard Gaussian vector points in a uniformly random direction. With $b = (\langle_i, z\rangle)_{i=1}^n \sim \mathcal{N}(0, G)$, α_t is a Gaussian orthant probability with correlation matrix G (Slepian, 1962). It depends on the round only through G, and never on the model dimension d (Proposition 3). Our guarantee is also stated in terms of α_t: Theorem 5 lower-bounds the probability that every client improves by a margin version of α_t. We therefore treat α_t as the primary measure of common-descent geometry. Its weakness is scale. For mutually orthogonal client directions, $\alpha_t = 2^{-n}$. This is about 10^{-6} at $n = 20$ and 10^{-30} at $n = 100$. Plain Monte Carlo then returns zero, and dedicated orthant estimators are required (Genz & Bretz, 2009; Botev, 2017). We work with $\log \alpha_t$ throughout.

Statistical dimension as a surrogate: A second measure avoids this numerical problem. Let $K = \{_1, \ldots, _n\}$ be the cone the client gradients generate, and define its statistical dimension (Amelunxen et al., 2014),

$$\Delta_t = \delta(K) = \mathbb{E}_{z\sim\mathcal{N}(0,I_d)} \big\|\Pi_K(z)\big\|^2, \tag{7}$$

where Π_K is the Euclidean projection onto K. The closed agreement cone is the polar of K, $\overline{\mathcal{C}_t} = K^\circ$. Hence $\delta(\overline{\mathcal{C}_t}) = d - \Delta_t$ (Amelunxen et al., 2014, Prop. 3.1). Δ_t thus describes the same pair of cones as α_t, but on a bounded scale. It counts the dimensions the gradients effectively occupy. It equals $1/2$ for identical gradients, $n/2$ for mutually orthogonal ones, and r when K is an r-dimensional subspace. In general $\Delta_t \in [\frac{1}{2}, \min(n, d)]$: small when clients agree, large when they conflict. Section 4 shows it can be estimated without ever sampling in $\mathbb{R}^d$.

When the surrogate is valid: Δ_t is not a function of α_t. The two are linked only when K is pointed, that is, when K contains no line. In that case large volume forces small Δ_t (Corollary 13). By Gordan's theorem, K is pointed exactly when $\mathcal{C}_t \neq$, or equivalently when the minimum-norm point of $\mathrm{conv}\{_i\}$ is nonzero. When K is not pointed, Δ_t can rank a population with no common descent direction as the less conflicted one. Configuration (a) of Figure 1 is such a case. The failure is cheap to detect, since the min-norm magnitude is exactly what FedMGDA+ computes (Hu et al., 2022). We therefore report α_t, Δ_t and the min-norm magnitude together, and read Δ_t as a surrogate for α_t only when the min-norm magnitude is positive.

4 Estimating the Statistics

The cone volume α_t is a Gaussian orthant probability. We compute $\log \alpha_t$ with the separation-of-variables estimator of Genz & Bretz (2009), which remains accurate at the scale of our experiments (Appendix F.2). The statistical dimension needs more care. Equation (7) is an expectation in $\mathbb{R}^d$, but its estimation reduces to an n-dimensional problem. Collect the normalized client gradients as the columns of $V = [_1, \ldots, _n] \in \mathbb{R}^{d\times n}$. Then $K = \{V\lambda : \lambda \geq 0\}$ and $G = V^\top V$.

Lemma 1 (Reduction to n dimensions) *Let $z \sim \mathcal{N}(0, I_d)$ and $b = V^\top z$. Then $b \sim \mathcal{N}(0, G)$, $\|\Pi_K(z)\|^2 = b^\top \lambda^*(b)$, and*

$$\Delta_t \;=\; \mathbb{E}_{b\sim\mathcal{N}(0,G)}\big[\, b^\top \lambda^*(b)\,\big], \qquad \lambda^*(b) \;=\; \arg\min_{\lambda\geq 0}\big\{\lambda^\top G\lambda - 2b^\top\lambda\big\}. \tag{8}$$

The ambient Gaussian is therefore never sampled. Each sample is a nonnegative least-squares (NNLS) problem in n variables with the fixed matrix G. This is a quadratic program of the same

Algorithm 1 GEOMETRYPROBE: estimating Δ_t at communication round t

Require: Client updates $\{\delta_i\}_{i \in S_t}$, sample budget M
Ensure: Estimate $\widehat{\Delta}_t$ of the statistical dimension
1: $_i \leftarrow \delta_i / \|\delta_i\|$ for $i \in S_t$; $n \leftarrow |S_t|$; $V \leftarrow [_1, \ldots, _n]$ ▷ keep direction, discard magnitude
2: $G \leftarrow V^\top V$; choose $A \in \mathbb{R}^{k \times n}$ with $A^\top A = G$, $k = \operatorname{rank}(G)$ ▷ only step involving the ambient dimension d
3: $s \leftarrow 0$
4: **for** $m = 1, \ldots, M$ **do**
5: $c_m \sim \mathcal{N}(0, I_k)$; $b_m \leftarrow A^\top c_m$ ▷ $b_m \sim \mathcal{N}(0, G)$ by Lemma 1
6: $\lambda_m \leftarrow \text{NNLS}(A, c_m)$; $s \leftarrow s + b_m^\top \lambda_m$ ▷ $= \|\Pi_K(z_m)\|^2$
7: **end for**
8: **return** $\widehat{\Delta}_t \leftarrow s/M$

size as the one FedMGDA+ solves once per round (Hu et al., 2022). The sample mean is unbiased for every budget M. Algorithm 1 summarizes the procedure.

Line 1 rescales each update to unit length. A client that merely took larger local steps thus gains no extra influence, matching the normalization of FedMGDA+. Line 6 is exact, since $\|c_m - A\lambda\|^2 = \lambda^\top G \lambda - 2 b_m^\top \lambda + \|c_m\|^2$. Forming G in line 2 is the only operation in $\mathbb{R}^d$, at cost $O(n^2 d)$. The M NNLS solves that follow cost $O(Mn^3)$ in the worst case, and they are independent and batchable. We factor G on its range space rather than by Cholesky. G is singular whenever n exceeds the gradient rank, and a regularized factor would bias the estimate upward by $1/2$ per null direction (Appendix E). Storing the updates needs $O(nd)$ memory. This is the binding cost at realistic model sizes, and it is incompatible with secure aggregation. Appendix E gives a sketching variant that stores only a low-dimensional projection of each update.

Proposition 2 (Sample complexity) *For every $\varepsilon > 0$ and $\delta \in (0, 1)$,*

$$\big(|\widehat{\Delta}_t - \Delta_t| \geq \varepsilon\big) \;\leq\; 2\exp\Big(-\frac{M\varepsilon^2}{8(\Delta_t + \varepsilon)}\Big). \tag{9}$$

Hence $M \geq 8(\Delta_t + \varepsilon)\log(2/\delta)/\varepsilon^2$ achieves additive error ε with probability at least $1 - \delta$. At relative error $\varepsilon = \rho\Delta_t$ it suffices that $M \geq 8(1 + \rho)\log(2/\delta)/(\rho^2 \Delta_t)$.

The bound involves neither d nor n, only Δ_t. The relative-error requirement *falls* as Δ_t grows. The expensive regime is thus the one where clients almost entirely agree, and precision matters least there. In our federated runs, $\Delta_t \approx 9$ with $n = 20$. For $\rho = 0.1$ and $\delta = 0.05$, the bound then asks for $M \approx 360$ samples per round. Empirically the bound is conservative but not vacuous. On 13 configurations with closed-form Δ_t, every 95% interval at $M = 50{,}000$ covers the exact value. Across 500 random Gram matrices with $n \leq 64$, every observed error lies inside the bound (Section 6.2).

What the estimator discards: Δ_t depends only on directions, by design. Gradients $g_1 = e$ and $g_2 = 1000\, e$ generate the same cone as $g_1 = g_2 = e$ (Remark 8). This stops a large local step from dominating the geometry. It also means Δ_t cannot see aggregate influence. Magnitudes matter for finite steps, and they re-enter the descent condition through the margin τ of Theorem 5, which is set by $\min_i \|g_i\|$. Section 5.3 makes this precise. It predicts where direction-only statistics lose power, which we observe in Section 6.3.

5 THEORETICAL ANALYSIS

We establish three things: that agreement-cone volume depends only on the client-gradient Gram matrix, that scalar heterogeneity summaries cannot characterize it, and that it controls the probability of simultaneous client improvement under directional aggregation noise. Proofs and extensions are in Appendix D.

5.1 Cone volume from the Gram matrix

Proposition 3 (Slepian, 1962, §2.1) *Let $_1, \ldots, _n$ be unit vectors in $\mathbb{R}^d$ with Gram matrix G. Then*

$$\alpha_t \;=\; \frac{\sigma(\mathcal{C}_t)}{\sigma(\mathbb{S}^{d-1})} \;=\; {}_{b\sim\mathcal{N}(0,\,G)}\big[\, b_i < 0 \;\; \forall i \,\big] \;=\; \Phi_n(\mathbf{0}; G), \tag{10}$$

independently of the ambient dimension d.

So agreement-cone volume is a Gaussian orthant probability determined by the $n \times n$ Gram matrix, and the numerical literature on such probabilities (Genz & Bretz, 2009; Botev, 2017) applies directly. It also increases coordinatewise with the pairwise gradient correlations, by Slepian's inequality (Proposition 9, Appendix D).

5.2 Scalar measures cannot see the cone

Mean pairwise cosine, the aggregate norm $\|\sum_i {}_i\|$, gradient variance and B^2 are all computed on unit directions. By Equation (1), each is a strictly monotone function of the others. They therefore always produce the same ranking of two populations. The next result shows that this common ranking can be the wrong one.

Theorem 4 (Separation) *Let $\mathcal{A}_n$ place $n-1$ unit gradients at 0° and one at 180°. Let $\mathcal{B}_n$ place n unit gradients evenly over $[-60^\circ, 60^\circ]$. Let $c_\infty = (3\sqrt{3}/2\pi)^2 \approx 0.684$.*

(i) For every $n \geq 10$, all four scalar measures rank $\mathcal{A}_n$ as the more homogeneous population. The gap in mean cosine satisfies $\overline{\cos}(\mathcal{A}_n) - \overline{\cos}(\mathcal{B}_n) \geq 1 - 4/n - c_\infty$, which is positive for $n \geq 13$ and tends to $1 - c_\infty \approx 0.316$.

(ii) For every $n \geq 3$, $\mathcal{C}_t(\mathcal{A}_n) =$, whereas $\mathcal{C}_t(\mathcal{B}_n)$ has angular width 60°, so that $\alpha_t(\mathcal{B}_n) = 1/6$.

Figure 1 is the case $n = 10$. There the scalars differ by under 3%, yet only $\mathcal{B}_{10}$ admits common descent. As n grows, the scalars do not merely fail to separate the two populations. They become more confident in the wrong ranking. For $3 \leq n \leq 9$ the ranking is correct, so the threshold $n \geq 10$ is sharp for this construction (Appendix D.11). The mechanism is Equation (1). Every scalar measure is a function of $\|\sum_i {}_i\|$. This length asks how much the gradients reinforce each other *on aggregate*. Whether one direction suits every client is a different question. It asks whether the origin lies in the convex hull of the gradients. The information is not missing from G, which fixes the configuration up to rotation. It is lost when G is compressed to one number. Because α_t is dimension-free (Proposition 3), the construction embeds unchanged into $\mathbb{R}^d$ for every $d \geq 2$. Δ_t does not share this behaviour. Here $K(\mathcal{A}_n)$ is a line, so it is not pointed, and $\Delta_t(\mathcal{A}_n) = 1 < 7/6 = \Delta_t(\mathcal{B}_n)$. This is the degenerate case flagged in Section 3.3, and the min-norm test detects it.

5.3 Simultaneous client improvement

We now connect the cone to what happens in a round. The server cannot choose its direction freely. The aggregate is whatever the clients produce, and it is random through client sampling and mini-batch noise. We therefore model it as

$$u \;=\; \bar{g} + \xi, \qquad \bar{g} = \sum_i p_i g_i, \qquad w_{t+1} = w_t - \eta\hat{u}, \quad \hat{u} = u/\|u\|, \tag{11}$$

where ξ is spherically symmetric with $[\xi = 0] = 0$. Gaussian noise is one example, but the main result does not need it. For a margin $\tau \in [0, 1)$, define the margin cone

$$\mathcal{K}_t^{(\tau)} \;=\; \big\{\, v \in \mathbb{R}^d \;:\; \langle {}_i, v\rangle \;\geq\; \tau\|v\| \;\; \forall i \,\big\}, \tag{12}$$

the directions that agree with every client by at least τ. Let $\alpha_t^{(\tau)}$ be its normalized spherical volume. Since $\mathcal{K}_t^{(0)} = -\overline{\mathcal{C}_t}$, we have $\alpha_t^{(0)} = \alpha_t$. Each constraint in (12) is a second-order cone constraint, so $\mathcal{K}_t^{(\tau)}$ is convex. The theorem rests on this fact.

Theorem 5 (Simultaneous descent under aggregation noise) *Assume each F_i is L-smooth, $\eta < 2\min_i \|g_i\|/L$, and*

$$\tau \;=\; \frac{L\eta}{2\min_i \|g_i\|} \;\in\; (0,1). \tag{13}$$

(i) ***Sufficiency.*** *If $\hat{u} \in \operatorname{int} \mathcal{K}_t^{(\tau)}$, then $F_i(w_{t+1}) < F_i(w_t)$ for every client i.*

(ii) ***Aligned case.*** *If $\bar{g} \in \mathcal{K}_t^{(\tau)}$, then at every noise level*

$$\big[\, F_i(w_{t+1}) < F_i(w_t) \;\; \forall i \,\big] \;\geq\; \alpha_t^{(\tau)}. \tag{14}$$

(iii) ***Noise-dominated limit.*** *If $\xi = \sigma\zeta$ for a fixed spherically symmetric ζ, then $[\hat{u} \in \mathcal{K}_t^{(\tau)}] \to \alpha_t^{(\tau)}$ as $\|\bar{g}\|/\sigma \to 0$.*

Moreover, $\alpha_t^{(\tau)} \uparrow \alpha_t$ as $\tau \downarrow 0$.

The proof is short. If $\bar{g}$ and a noise draw both lie in the convex cone, so does their sum. A spherically symmetric vector lands in a cone with probability equal to the cone's normalized volume. The volume is therefore a lower bound in the aligned case, and it is exact once noise dominates. The guarantee is per round, not a convergence result. It is silent when $\bar{g}$ lies outside the margin cone, and it is loose at low noise, where a union bound is stronger. Appendix D.13 discusses these limits, and Appendix D gives the proofs.

Magnitude sets the margin: Theorem 5 fixes τ through the step size and the smallest client gradient. The next result makes this dependence explicit for the unnormalized step $w_{t+1} = w_t - \eta u$ that FedAvg takes.

Proposition 6 (Finite-step outcome of each client) *Assume F_i is L-smooth, and let $\tau_i = L\eta\|u\|/(2\|g_i\|)$. For the step $w_{t+1} = w_t - \eta u$,*

$$\langle i, \hat{u}\rangle > \tau_i \;\Rightarrow\; F_i(w_{t+1}) < F_i(w_t), \qquad \langle i, \hat{u}\rangle < -\tau_i \;\Rightarrow\; F_i(w_{t+1}) > F_i(w_t). \tag{15}$$

Within the band $|\langle i, \hat{u}\rangle| \leq \tau_i$, direction alone does not determine whether client i improves.

The proof applies the two-sided smoothness bound $|F_i(w - \eta u) - F_i(w) + \eta\langle g_i, u\rangle| \leq L\eta^2\|u\|^2/2$ (Appendix D.14). The proposition shows where geometry decides a client's outcome and where it cannot. The band width τ_i grows with the update norm $\|u\|$ and shrinks with the client's own gradient norm $\|g_i\|$. Both are magnitudes, and α_t, Δ_t and mean cosine all discard them. When client gradients are large relative to $L\eta\|u\|$, the band is narrow. Direction then decides who improves, and direction-only statistics can track the improved-client fraction. When client gradients are small relative to the step, the band is wide. Many clients then fall inside it, where curvature and magnitude decide the outcome, so directional statistics should lose power. This is consistent with what we observe as heterogeneity weakens, and where on CIFAR-10 at $\beta = 100$ a magnitude-carrying measure is the strongest predictor (Section 6.3). We do not measure τ_i directly.

6 Experiments

The theory makes claims about geometry. The experiments ask three separate questions. **(Q1)** Does the separation of Theorem 4 hold numerically, and how does it scale with n? **(Q2)** Are α_t and Δ_t computed correctly, and is Proposition 2 honest? **(Q3)** In federated training, do the cone statistics track how many clients a round improves, and do they add information beyond mean cosine?

Federated setup: We train on FMNIST and CIFAR-10 for 300 rounds, sampling $n = 20$ clients per round with 5 local epochs and learning rate 0.01. Data follow Dirichlet label-skew partitions with $\beta \in \{0.05, 0.1, 0.5, 100\}$, where smaller β indicates stronger heterogeneity. Each setting uses five seeds. Before each aggregation, we compute mean cosine, the FedMGDA+ min-norm magnitude (Hu et al., 2022), $\log \alpha_t$, and $\widehat{\Delta}_t$. The target is the improved-client fraction IUF_t, the fraction of sampled clients whose loss decreases. Each measure is oriented a priori, and we report the mean



Figure 2: **Scalar measures rank the populations of Theorem 4 backwards, increasingly so as n grows.** (a) Mean pairwise cosine and (b) B^2 for $\mathcal{A}_n$ and $\mathcal{B}_n$. Both rank $\mathcal{A}_n$ as more homogeneous for every $n \geq 10$. (c) α_t is exactly 0 for $\mathcal{A}_n$ and $1/6$ for $\mathcal{B}_n$ at every n. The Δ_t panel and all values are in Appendix F.1.

per-seed Spearman ρ with a 95% t-interval across seeds. On unit directions, variance and B^2 are rank-identical to mean cosine (Equation (1)), so we report them only in raw form (Appendix F.5). Client updates are nearly orthogonal ($\widehat{\Delta}_t \approx 9$ against $n/2 = 10$), and majority-harm rounds are common only at $\beta \leq 0.1$ (Appendix F.4).

6.1 The separation grows with the number of clients

Figure 2 evaluates the construction for $n \in [3, 100]$. For $n \leq 9$ the scalars rank the populations correctly. From $n = 10$ onward, every scalar prefers $\mathcal{A}_n$, as Theorem 4 states. The gap then widens: at $n = 100$, mean cosine is 0.960 against 0.675, and B^2 is 1.041 against 1.474. The scalars thus become more confident in the wrong ranking as clients are added. Meanwhile $\alpha_t(\mathcal{A}_n) = 0$ and $\alpha_t(\mathcal{B}_n) = 1/6$ at every n, matching Proposition 3 to six decimals. $\widehat{\Delta}_t$ recovers 1 and $7/6$, and so ranks $\mathcal{A}_n$ as less conflicted. This is the non-pointed case of Section 3.3, and the min-norm test flags it.

6.2 The estimators are accurate and the bound is conservative

We test $\widehat{\Delta}_t$ on 13 configurations with closed-form statistical dimension: identical, orthogonal and antipodal-basis gradients, with $\Delta_t \in [\frac{1}{2}, 25]$. At $M = 50{,}000$, every 95% interval covers the exact value. Standardized errors satisfy $|z| \leq 1.36$, and relative errors are at most 1.2% (Table 2). For Proposition 2, we use 500 (Gram matrix, M) cells with n up to 64. Every observed error falls inside the $\delta = 0.05$ bound, so coverage is 1.0 against the nominal 0.95. The bound exceeds the observed error by a median factor of 5.5 and a minimum of 1.15 (Table 4). It is conservative but not vacuous. The orthant estimator for $\log \alpha_t$ matches the closed form of Lemma 15 within 1.5% for $n \leq 20$ (Table 3).

6.3 Association with client improvement: three regimes

Figure 3 shows three regimes, each matching a different part of the theory. Detailed results are in Appendix F.6.

Strong heterogeneity ($\beta \leq 0.1$): Both cone statistics show the predicted association in all four cells. Their strength ($\rho \in [0.23, 0.31]$) is comparable to mean cosine and the min-norm magnitude, although mean cosine is the largest measure in three of four cells. Agreement geometry is therefore relevant here, but not superior to mean cosine.

Moderate heterogeneity ($\beta = 0.5$): The association vanishes on FMNIST and reverses on CIFAR-10. Mean cosine reverses as well, so the effect reflects the regime rather than the cone. Proposition 6 suggests why. When a client's alignment falls inside its band $|\langle_i, \hat{u}\rangle| \leq \tau_i$, magnitude and curvature decide the outcome, not direction. We did not measure τ_i, so this explanation is consistent with the theory but not tested.

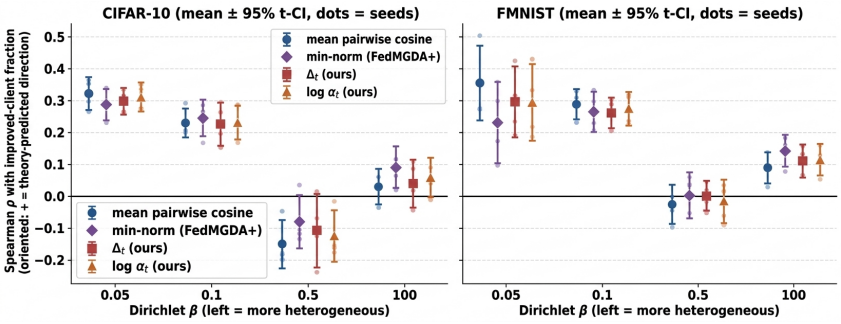


Figure 3: **Oriented Spearman ρ between each measure and the improved-client fraction** IUF_t, within each heterogeneity level (mean and 95% t-interval over five seeds; dots denote seeds; positive is the direction predicted by theory). Full values are in Appendix F.5.

Near-IID ($\beta = 100$): Associations are small. On CIFAR-10, the magnitude-carrying raw B^2 is the strongest measure (0.256). This again points to magnitude, which every directional statistic discards by construction (Remark 8).

6.4 Information beyond cosine: a predicted redundancy

The theory also predicts how much the cone should add to cosine. When the pairwise cosines are nearly equal, α_t is pinned to a monotone function of mean cosine (Remark 16), so the cone adds little. It can add more only under structured conflict, where some clients systematically oppose others. Our symmetric Dirichlet partitions and uniform client sampling do not induce structured conflict, so we expect little residual signal. Three tests confirm this (Appendix F.8). After controlling for cosine and round, Δ_t shows no significant residual in six of eight cells, while $\log \alpha_t$ and the min-norm magnitude retain a small residual in two. Adding any measure to a cosine-plus-round classifier changes AUROC by at most 0.018. Within cosine quintiles, rounds ranked as more agreeing improve up to 3 percentage points more clients. This gap is significant in $6/8$ cells for $\log \alpha_t$, $5/8$ for the min-norm magnitude, and $3/8$ for Δ_t. The small residual is thus carried mainly by $\log \alpha_t$, the quantity in Theorem 5. We did not record the spread of G, so this mechanism is predicted rather than measured.

7 Conclusion

We introduced the agreement cone, the set of server directions that improve every participating client. Its volume α_t is a Gaussian orthant probability determined by the client-gradient Gram matrix. Its statistical dimension Δ_t is a bounded surrogate that we estimate without sampling in the model dimension. Our theory shows that scalar heterogeneity measures can rank client populations backwards (Theorem 4) and that cone volume lower-bounds the probability that a noisy aggregate improves all clients (Theorem 5). Update magnitudes set the margin within which direction alone cannot determine a client's outcome (Proposition 6). On Dirichlet partitions, the cone statistics track client improvement under strong heterogeneity comparably to mean cosine. Beyond cosine, they add only a small signal, as predicted by the near-exchangeable Gram matrices induced by these partitions (Remark 16).

Limitations and future work: The proposed methodology guarantees per-round and does not imply convergence. The misaligned case of Theorem 5 remains open. Computing the statistics also requires individual client updates and is therefore incompatible with secure aggregation. Most importantly, our federated experiments do not include structured conflict, such as a minority of clients opposing a majority. This is the regime in which theory predicts the cone and mean cosine to disagree. Testing this setting is an immediate priority. Using cone statistics to adapt server updates and combining them with magnitude-aware measures are natural directions for future work.

REFERENCES

Durmus Alp Emre Acar, Yue Zhao, Ramon Matas Navarro, Matthew Mattina, Paul N Whatmough, and Venkatesh Saligrama. Federated learning based on dynamic regularization. *arXiv preprint arXiv:2111.04263*, 2021.

Dimitris Achlioptas. Database-friendly random projections. In *Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pp. 274–281, 2001.

Dennis Amelunxen, Martin Lotz, Michael B McCoy, and Joel A Tropp. Living on the edge: Phase transitions in convex programs with random data. *Information and Inference: A Journal of the IMA*, 3(3):224–294, 2014.

Zdravko I Botev. The normal law under linear restrictions: simulation and estimation via minimax tilting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(1):125–148, 2017.

Venkat Chandrasekaran, Benjamin Recht, Pablo A Parrilo, and Alan S Willsky. The convex geometry of linear inverse problems. *Foundations of Computational Mathematics*, 12(6):805–849, 2012.

Zhao Chen, Jiquan Ngiam, Yanping Huang, Thang Luong, Henrik Kretzschmar, Yuning Chai, and Dragomir Anguelov. Just pick a sign: Optimizing deep multitask models with gradient sign dropout. *Advances in Neural Information Processing Systems*, 33:2039–2050, 2020.

Sanjoy Dasgupta and Anupam Gupta. An elementary proof of a theorem of johnson and lindenstrauss. *Random Structures & Algorithms*, 22(1):60–65, 2003.

Jean-Antoine Désidéri. Multiple-gradient descent algorithm (mgda). 2009.

A Fallah, A Mokhtari, and E Asuman. Ozdaglar,"personalized federated learning with theoretical guarantees: a modelagnostic meta-learning approach,". *Neural Information Proceesing System*, 2020.

Jörg Fliege and Benar Fux Svaiter. Steepest descent methods for multicriteria optimization. *Mathematical Methods of Operations Research*, 51(3):479–494, 2000.

Alan Genz and Frank Bretz. *Computation of multivariate normal and t probabilities*. Springer Science & Business Media, 2009.

Yehoram Gordon. On Milman's inequality and random subspaces which escape through a mesh in

Farzin Haddadpour and Mehrdad Mahdavi. On the convergence of local descent methods in federated learning. *arXiv preprint arXiv:1910.14425*, 2019.

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.

Zeou Hu, Kiarash Shaloudegi, Guojun Zhang, and Yaoliang Yu. Federated learning meets multiobjective optimization. *IEEE Transactions on Network Science and Engineering*, 9(4):2039–2051, 2022.

Peter Kairouz and H Brendan McMahan. Advances and open problems in federated learning. *Foundations and trends in machine learning*, 14(1-2):1–210, 2021.

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

Ahmed Khaled, Konstantin Mishchenko, and Peter Richtárik. Tighter theory for local sgd on identical and heterogeneous data. In *International conference on artificial intelligence and statistics*, pp. 4519–4529. PMLR, 2020.

Charles L. Lawson and Richard J. Hanson. *Solving Least Squares Problems*, volume 15 of *Classics in Applied Mathematics*. SIAM, 1995. Reprint of the 1974 Prentice-Hall edition.

Qinbin Li, Yiqun Diao, Quan Chen, and Bingsheng He. Federated learning on non-iid data silos: An experimental study. In *2022 IEEE 38th international conference on data engineering (ICDE)*, pp. 965–978. IEEE, 2022.

Tian Li, Maziar Sanjabi, Ahmad Beirami, and Virginia Smith. Fair resource allocation in federated learning. *arXiv preprint arXiv:1905.10497*, 2019a.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450, 2020.

Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of fedavg on non-iid data. *arXiv preprint arXiv:1907.02189*, 2019b.

Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-averse gradient descent for multi-task learning. *Advances in neural information processing systems*, 34:18878–18890, 2021.

Mi Luo, Fei Chen, Dapeng Hu, Yifan Zhang, Jian Liang, and Jiashi Feng. No fear of heterogeneity: Classifier calibration for federated learning with non-IID data. *Advances in Neural Information Processing Systems*, 34:5972–5984, 2021.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. Pmlr, 2017.

Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning. In *International conference on machine learning*, pp. 4615–4625. PMLR, 2019.

Peter Orlik and Hiroaki Terao. *Arrangements of hyperplanes*. Springer Science & Business Media, 2013.

Zibin Pan, Shuyi Wang, Chi Li, Haijin Wang, Xiaoying Tang, and Junhua Zhao. Fedmdfg: Federated learning with multi-gradient descent and fair guidance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 9364–9371, 2023.

Robin L Plackett. A reduction formula for normal multivariate integrals. *Biometrika*, 41(3/4):351–360, 1954.

Sashank Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečnỳ, Sanjiv Kumar, and H Brendan McMahan. Adaptive federated optimization. *arXiv preprint arXiv:2003.00295*, 2020.

Felix Sattler, Klaus-Robert Müller, and Wojciech Samek. Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints. *IEEE transactions on neural networks and learning systems*, 32(8):3710–3722, 2020.

Rolf Schneider and Wolfgang Weil. *Stochastic and integral geometry*. Springer, 2008.

Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. *Advances in neural information processing systems*, 31, 2018.

David Slepian. The one-sided barrier problem for gaussian noise. *Bell System Technical Journal*, 41(2):463–501, 1962.

Richard P Stanley et al. An introduction to hyperplane arrangements. *Geometric combinatorics*, 13: 389–496, 2007.

Sebastian U Stich. Local sgd converges fast and communicates little. *arXiv preprint arXiv:1805.09767*, 2018.

Yung Liang Tong. *The multivariate normal distribution*. Springer Science & Business Media, 2012.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H Vincent Poor. Tackling the objective inconsistency problem in heterogeneous federated optimization. *Advances in neural information processing systems*, 33:7611–7623, 2020.

Jianyu Wang, Zachary Charles, Zheng Xu, Gauri Joshi, H Brendan McMahan, Maruan Al-Shedivat, Galen Andrew, Salman Avestimehr, Katharine Daly, Deepesh Data, et al. A field guide to federated optimization. *arXiv preprint arXiv:2107.06917*, 2021a.

Zheng Wang, Xiaoliang Fan, Jianzhong Qi, Chenglu Wen, Cheng Wang, and Rongshan Yu. Federated learning with fair averaging. *arXiv preprint arXiv:2104.14937*, 2021b.

Blake Woodworth, Kumar Kshitij Patel, Sebastian Stich, Zhen Dai, Brian Bullins, Brendan Mcmahan, Ohad Shamir, and Nathan Srebro. Is local sgd better than minibatch sgd? In *International Conference on Machine Learning*, pp. 10334–10343. PMLR, 2020.

Haibo Yang, Zhuqing Liu, Jia Liu, Chaosheng Dong, and Michinari Momma. Federated multi-objective learning. *Advances in neural information processing systems*, 36:39602–39625, 2023.

Dong Yin, Ashwin Pananjady, Max Lam, Dimitris Papailiopoulos, Kannan Ramchandran, and Peter Bartlett. Gradient diversity: a key ingredient for scalable distributed learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 1998–2007. PMLR, 2018.

Hao Yu, Sen Yang, and Shenghuo Zhu. Parallel restarted sgd with faster convergence and less communication: Demystifying why model averaging works for deep learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 5693–5700, 2019.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *Advances in neural information processing systems*, 33: 5824–5836, 2020.

Thomas Zaslavsky. Facing up to arrangements. *Mem. Amer. Math. Soc*, 1(154), 1975.

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.

A Overview of the Appendix

The appendix contains

- extended related work (Appendix B),
- a summary of mathematical notation (Appendix C),
- complete proofs, supporting lemmas, and technical remarks (Appendix D),
- numerical implementation details for the estimators, including singular Gram-matrix factorization, the NNLS solver, and a memory-saving sketching variant (Appendix E), and
- additional experiments on the synthetic separation, estimator accuracy, sample complexity, federated-training regimes, full correlation and AUROC results, and conditional analyses beyond mean cosine (Appendix F).

B Extended Related Work

Heterogeneity in federated convergence theory: Analyses of local SGD (Stich, 2018; Khaled et al., 2020; Woodworth et al., 2020; Yu et al., 2019; Gordon, 1988) follow the same pattern as the FedAvg analyses discussed in Section 2: heterogeneity enters through a single constant, and that constant is a function of $\| \sum_i {}_i \|$.

Measuring and diagnosing heterogeneity: Zhao et al. (2018) and Hsu et al. (2019) quantify how label skew degrades accuracy, and Li et al. (2022) benchmark partitioning schemes systematically. Clustered federated learning (Sattler et al., 2020) partitions clients by pairwise cosine, and Luo et al. (2021) ttrace damage to specific layers.

Multi-objective federated learning: MGDA was brought to multi-task learning by Sener & Koltun (2018) and extended by gradient-surgery methods (Yu et al., 2020; Liu et al., 2021; Chen et al., 2020). FedMGDA+ (Hu et al., 2022) normalizes client gradients before solving its quadratic program, and later work develops the federated multi-objective setting further (Yang et al., 2023; Pan et al., 2023; Wang et al., 2021b). Fairness-motivated methods (Li et al., 2019a; Mohri et al., 2019) and personalization (Fallah et al., 2020) address related concerns by reweighting rather than by geometry.

Conic geometry and Gaussian orthant probabilities: The theory underlying Amelunxen et al. (2014) is developed in Schneider & Weil (2008). The descent-cone viewpoint entered machine learning through Chandrasekaran et al. (2012) and the Gaussian-process arguments of Gordon (1988). Textbook statements of the orthant identity appear in Tong (2012). Computing such probabilities is a mature numerical problem (Tong, 2012; Genz & Bretz, 2009), with minimax tilting (Botev, 2017) remaining accurate in the negatively correlated regime where separation-of-variables estimators fail, precisely the regime federated gradients occupy. We also considered the combinatorial invariants of the induced hyperplane arrangement (Zaslavsky, 1975; Stanley et al., 2007; Orlik & Terao, 2013), but these are unchanged by continuous angular deformation and so cannot bound disagreement.

C Notations Used

Symbol	Description
$g_i,\ _i$	Client i's gradient at round t, and its normalized version
G	Gram matrix, $G_{ij} = \langle {}_i, {}_j \rangle$
$\mathcal{C}_t$	Agreement cone (Eq. (5))
K	Cone generated by the gradients, $\{{}_1, \ldots, {}_n\}$
α_t	Normalized spherical volume of $\mathcal{C}_t$ (Eq. (6))
Δ_t	Statistical dimension of K (Eq. (7))
$\mathcal{K}_t^{(\tau)}$	Shrunken reflected cone with margin τ (Eq. (12))
$\alpha_t^{(\tau)}$	Normalized volume of $\mathcal{K}_t^{(\tau)}$

Lemma 7 (Positive pairwise cosines suffice) *If $G_{ij} > 0$ for all $i \neq j$, then $\mathcal{C}_t \neq$.*

Suppose $\sum_i \lambda_{ii} = 0$ for some $\lambda \geq 0$ with $\lambda \neq 0$. Choose j with $\lambda_j > 0$ and take the inner product with ${}_j$. Every term with $i \neq j$ satisfies $\lambda_i G_{ij} \geq 0$, and the term $i = j$ equals λ_j since $G_{jj} = 1$, so $0 = \sum_i \lambda_i G_{ij} \geq \lambda_j > 0$, a contradiction. By Gordan's theorem the cone is nonempty.

The condition is only sufficient, and it is weak in practice. In real federated training some pair of clients has negative cosine within the first few rounds, after which the test says nothing, and it never distinguishes a barely nonempty cone from a comfortably wide one.

D Proofs

Throughout, ${}_1, \ldots, {}_n$ are unit vectors in $\mathbb{R}^d$ with Gram matrix $G_{ij} = \langle {}_i, {}_j \rangle$, and $\mathcal{C}_t = \{d : \langle {}_i, d \rangle < 0\ \forall i\}$.

D.1 Proof of Lemma 1

The vector $b = V^\top z$ is a linear image of a standard Gaussian with covariance $V^\top I_d V = G$, so $b \sim \mathcal{N}(0, G)$. The projection is $\Pi_K(z) = V\lambda^*$, where λ^* minimizes $\|z - V\lambda\|^2 = \|z\|^2 - 2b^\top\lambda + \lambda^\top G\lambda$ over $\lambda \geq 0$, which is the program in (8). The KKT conditions give $G\lambda^* - b = \mu$ with $\mu \geq 0$ and $\mu^\top\lambda^* = 0$, hence $\lambda^{*\top} G\lambda^* - \lambda^{*\top} b = \lambda^{*\top}\mu = 0$. Therefore $\|\Pi_K(z)\|^2 = \|V\lambda^*\|^2 = \lambda^{*\top} G\lambda^* = b^\top\lambda^*$, and taking expectations gives (8).

D.2 Proof of Proposition 2

Amelunxen et al. (2014, Sublemma E.3) give

$$\mathbb{E}\exp\big(\theta(\|\Pi_K(z)\|^2 - \Delta_t)\big) \ \leq\ \exp\big(2\theta^2\Delta_t/(1 - 4|\theta|)\big) \qquad \text{for } |\theta| < 1/4, \tag{16}$$

which is sub-gamma with variance factor $v = 4\Delta_t$ and scale $c = 4$. For an average of M independent samples the two-sided Bernstein bound gives $(|\widehat{\Delta}_t - \Delta_t| \geq \varepsilon) \leq 2\exp\big(-M\varepsilon^2/(2(v + c\varepsilon))\big)$, and $2(v + c\varepsilon) = 8(\Delta_t + \varepsilon)$, which is (9). Solving for M gives the stated budgets, and substituting $\varepsilon = \rho\Delta_t$ gives the relative-error form.

D.3 What Δ_t discards

Remark 8 (What Δ_t measures, and what it discards) *Δ_t is the statistical dimension of the cone generated by the* normalized *client gradients. It is not a measure of client disagreement in any general sense. The normalization is a deliberate design choice inherited from FedMGDA+ (Hu et al., 2022): the agreement cone is defined by directions, and a client that merely takes larger local steps should not dominate the geometry.*

The cost is explicit. Two clients with gradients $g_1 = e$ and $g_2 = 1000\,e$ produce the same Δ_t as two clients with $g_1 = g_2 = e$, because both generate the ray $\mathbb{R}_+ e$. A client that agrees in direction but dominates in magnitude is invisible to Δ_t, even though it determines the aggregate. Magnitude information is not lost to the framework as a whole: it enters Lemma 10 through $\min_i \|g_i\|$ and Theorem 5 through τ. It is, however, absent from Δ_t itself, so comparisons based on Δ_t alone should be read accordingly.

D.4 Monotonicity in the correlations

Proposition 9 (Monotonicity; Slepian, 1962, Lemma 1; strict form Plackett, 1954) *Let G and G' be correlation matrices with $G_{ij} \geq G'_{ij}$ for all $i \neq j$. Then $\Phi_n(\mathbf{0}; G) \geq \Phi_n(\mathbf{0}; G')$. If both matrices are positive definite, the inequality is strict whenever they differ.*

Slepian's lemma needs only positive semidefiniteness, which matters here: whenever n exceeds the rank of the client gradients, routine at large n, G is singular. Bracketing G by the constant-correlation matrices built from $\rho_- = \min_{i\neq j} G_{ij}$ and $\rho_+ = \max_{i\neq j} G_{ij}$ gives two-sided bounds on α_t. For $\rho \geq 0$ the constant-correlation orthant probability has the closed form $\Psi_n(\rho) = \int \Phi\big(z\sqrt{\rho/(1-\rho)}\big)^n \varphi(z)\,dz$ (Tong, 2012, Thm. 5.3.3). The bracketing is folklore (Tong, 2012, Problem 7.2); we record it as Lemma 15 because it makes precise in what sense the extremal pairwise cosines are the loose endpoints of the quantity we actually want. The closed form applies only for $\rho_- \geq 0$, and federated gradients routinely have $\rho_- < 0$, so the lower bracket must usually be evaluated numerically.

D.5 Deterministic descent

If the server could choose its direction freely, the relevant quantity would be the inradius of the agreement cone rather than its volume.

Lemma 10 (Deterministic descent) *Assume each F_i is L-smooth. Let $m_t = \max_{\|d\|=1} \min_i \langle -_i, d\rangle$ be the inradius of $\mathcal{C}_t$, attained at d^*. If $m_t > 0$ then for $\eta = m_t \min_i \|g_i\|/L$,*

$$F_i(w_t + \eta d^*) \;\leq\; F_i(w_t) - \tfrac{1}{2L}\, m_t^2 \min_j \|g_j\|^2 \qquad \text{for every } i. \tag{17}$$

If $m_t = 0$ then $\mathcal{C}_t$ has empty interior, so no nonzero direction is a strict first-order descent direction for all clients simultaneously.

By definition of m_t we have $\langle_i, d^*\rangle \leq -m_t$ for every i, hence $\langle g_i, d^*\rangle \leq -m_t\|g_i\| \leq -m_t \min_j \|g_j\|$. L-smoothness gives $F_i(w_t + \eta d^*) \leq F_i(w_t) + \eta\langle g_i, d^*\rangle + L\eta^2/2$. Substituting the bound and minimizing the right-hand side over η gives the stated step size and (17). If $m_t = 0$ then no unit d satisfies $\langle_i, d\rangle < 0$ for all i with a uniform margin, so $\mathcal{C}_t$ has empty interior.

The inradius and the volume are not interchangeable: a thin sliver has positive volume but tiny inradius. Lemma 10 also assumes the server can select d^*, which it cannot. Theorem 5 closes both gaps. An earlier form of that theorem assumed the realized direction had density at least $1/\beta$ on the agreement cone. That is an assumption about the aggregation rather than a property of FedAvg, and the noise model of Equation (11) replaces it.

D.6 SMALL-NOISE AND UNNORMALIZED VARIANTS

The two remarks below are the only places Gaussianity of ξ is used; both rely on Gaussian norm concentration, which spherical symmetry alone does not provide.

Remark 11 (Small-noise regime) *Let $\xi \sim \mathcal{N}(0, \sigma^2 I_d)$. If $\langle_i, \bar{g}\rangle \geq \gamma$ for all i and $c = \tau\big(\|\bar{g}\| + \sigma(\sqrt{d} + s)\big) < \gamma$, then*

$$[\textit{all clients improve}] \;\geq\; 1 - n\,\overline{\Phi}\big((\gamma - c)/\sigma\big) - e^{-s^2/2}, \tag{18}$$

where $\overline{\Phi}$ is the standard normal upper tail. The bound tends to 1 as $\sigma \to 0$ and becomes vacuous as σ grows, so it is complementary to Theorem 5: the union bound governs low noise, and $\alpha_t^{(\tau)}$ governs high noise, where it is exact by part (iii). Requiring $c < \gamma$ forces $\eta = O(1/\sqrt{d})$; no such restriction is needed for Theorem 5(ii).

Remark 12 (Unnormalized steps) *For $w_{t+1} = w_t - \eta u$ the same argument applies on the event $\{\|u\| \leq R\}$, with τ replaced by $L\eta R/(2\min_i \|g_i\|)$. Taking $R = \|\bar{g}\| + \sigma(\sqrt{d} + s)$ and Gaussian norm concentration (Vershynin, 2018, Thm. 3.1.1) costs an additive $e^{-s^2/2}$ in every bound above.*

D.7 THE COMPUTED STATISTIC AND THE DEGENERATE CASE

Corollary 13 (Interpretation of the computed statistic) *Assume that*

$$K = \{_1, \ldots, _n\}$$

is pointed, equivalently, that $\overline{\mathcal{C}_t}$ is full-dimensional. Then α_t is the top conic intrinsic volume of $\overline{\mathcal{C}_t}$. Moreover, by Amelunxen et al. (2014, Thm. 6.1), the conic intrinsic volumes of $\overline{\mathcal{C}_t}$ concentrate around

$$\delta(\overline{\mathcal{C}_t}) = d - \Delta_t.$$

Consequently, α_t can be non-negligible only when Δ_t is small.

The corollary is qualitative: it does not provide a two-sided bound relating α_t and Δ_t. It explains why Δ_t can be informative about α_t when K is pointed. It also explains why α_t is numerically unusable at moderate n, since it is one entry in the far tail of a concentrated intrinsic-volume distribution.

The pointedness assumption is essential. In Configuration $\mathcal{A}$ of Theorem 4, the generated cone is a line. Thus, K is not pointed and $\overline{\mathcal{C}_t}$ has empty interior. In this case, $\alpha_t = 0$, whereas

$$\Delta_t = 1 < \frac{7}{6} = \Delta_t(\mathcal{B}).$$

Therefore, Δ_t ranks the population with no common strict-descent direction as less conflicted than $\mathcal{B}$. This failure mode is easy to detect: K is not pointed if and only if the minimum-norm point of

$$\mathrm{conv}\{_1, \ldots, _n\}$$

is zero, precisely the feasibility diagnostic computed by FedMGDA+ (Hu et al., 2022). We therefore report α_t, Δ_t, and the min-norm diagnostic together, and interpret Δ_t as a surrogate for α_t only when the min-norm magnitude is positive. Section 6 validates all three quantities empirically.

D.8 RELATION TO FEDERATED CONVERGENCE BOUNDS

Can Δ_t be substituted into an existing convergence rate? Not into the heterogeneity constant $\Gamma = F^* - \sum_k p_k F_k^*$ of Li et al. (2019b), which is defined at the optima and is not a per-round quantity. Round-t geometry does, however, appear in the client-drift term. Li et al. (2019b, Lemma 3) control $\sum_k p_k \|\bar{w}_t - w_t^k\|^2$ using only the uniform bound $\|\nabla F_k\|^2 \leq G^2$, treating each client in isolation. The dispersion of the round's local displacements admits an exact expression instead.

Proposition 14 (Round-update dispersion) *Let δ_k be client k's local displacement and $\bar{\delta} = \sum_k p_k \delta_k$. Writing $\delta_k = -\eta E g_k$,*

$$\sum_k p_k \|\delta_k - \bar{\delta}\|^2 \;=\; \eta^2 E^2 \Big(\textstyle\sum_k p_k \|g_k\|^2 - \big\| \sum_k p_k g_k \big\|^2 \Big). \tag{19}$$

For unit-norm client gradients and $p_k = 1/n$ this equals $\eta^2 E^2 \big(1 - \frac{1}{n} - \frac{n-1}{n}\,\overline{\cos}\big)$, a per-round quantity decreasing in the mean pairwise cosine.

Expand $\sum_k p_k \|\delta_k - \bar{\delta}\|^2 = \sum_k p_k \|\delta_k\|^2 - \|\bar{\delta}\|^2$ and substitute $\delta_k = -\eta E g_k$ to obtain (19). For $\|g_k\| = 1$ and $p_k = 1/n$, Equation (1) gives $\| \sum_k p_k g_k \|^2 = \frac{1}{n} + \frac{n-1}{n}\overline{\cos}$, and the stated form follows.

Equation (19) is an algebraic identity for single-round displacements. It suggests replacing the worst-case constant G^2 by a per-round quantity, but we do not re-derive the multi-step bound of Li et al. (2019b) under this substitution, and we make no convergence claim from it.

One point deserves emphasis because it bears on how our contributions should be read. The right-hand side of (19) is gradient diversity (Yin et al., 2018), hence by Equation (1) mean pairwise cosine. The place where per-round geometry enters standard FedAvg analysis therefore recovers the baseline statistic, not the cone. This is not a defect: averaged-loss convergence is governed by $\|\bar{g}\|$, while simultaneous improvement is governed by the cone. The two are complementary, and we keep them separate.

D.9 Proof of Proposition 3

Let $z \sim \mathcal{N}(0, I_d)$. By rotational invariance, $z/\|z\|$ is uniform on $\mathbb{S}^{d-1}$, and since $\mathcal{C}_t$ is a cone the event $z \in \mathcal{C}_t$ depends only on $z/\|z\|$. Hence $[z \in \mathcal{C}_t] = \sigma(\mathcal{C}_t)/\sigma(\mathbb{S}^{d-1}) = \alpha_t$. Set $b_i = \langle_i, z\rangle$. The vector $b = V^\top z$ with $V = [_1, \ldots, _n]$ is Gaussian with mean zero and covariance $V^\top V = G$. The event $z \in \mathcal{C}_t$ is exactly $\{b_i < 0 \;\forall i\}$, giving $\alpha_t = \Phi_n(\mathbf{0}; G)$. No step involves d beyond the rotational invariance argument.

D.10 The correlation sandwich

Lemma 15 *Let $\rho_- = \min_{i \neq j} G_{ij}$ and $\rho_+ = \max_{i \neq j} G_{ij}$, and suppose $\rho_- > -1/(n-1)$ so that both constant-correlation matrices are positive semidefinite. Then $\Phi_n(\mathbf{0}; \rho_- \mathbf{1}\mathbf{1}^\top + (1 - \rho_-)I) \le \alpha_t \le \Phi_n(\mathbf{0}; \rho_+ \mathbf{1}\mathbf{1}^\top + (1-\rho_+)I)$. For $\rho \ge 0$ the bracketing values equal $\Psi_n(\rho) = \int_{\mathbb{R}} \Phi\big(z\sqrt{\rho/(1-\rho)}\big)^n \varphi(z)\, dz$.*

Both inequalities are Slepian's Lemma 1 (Slepian, 1962) applied to the constant-correlation matrices, which dominate and are dominated by G entrywise by construction. The closed form for $\rho \ge 0$ follows by writing the exchangeable normal vector as $b_i = \sqrt{\rho}\, z_0 + \sqrt{1-\rho}\, z_i$ with z_0, z_i i.i.d. standard normal, conditioning on z_0, and using independence (Tong, 2012, Thm. 5.3.3). Positive semidefiniteness of the constant-correlation matrix requires $\rho \ge -1/(n-1)$ (Tong, 2012, Prop. 5.2.2), which at $n = 100$ permits only $\rho \ge -0.0101$; the lower bracket must therefore be computed numerically in most federated settings.

Remark 16 (Near-exchangeable Gram matrices) *Suppose every off-diagonal entry of G lies in $[\bar{\rho} - \epsilon,\ \bar{\rho} + \epsilon]$, where $\bar{\rho} = \overline{\cos}$ and $\bar{\rho} - \epsilon > -1/(n-1)$. Write $\Psi_n(\rho) = \Phi_n(\mathbf{0}; \rho \mathbf{1}\mathbf{1}^\top + (1-\rho)I)$. Lemma 15 then gives*

$$\Psi_n(\bar{\rho} - \epsilon) \;\le\; \alpha_t \;\le\; \Psi_n(\bar{\rho} + \epsilon). \tag{20}$$

The function Ψ_n is continuous and increasing in ρ (Proposition 9). As $\epsilon \to 0$, α_t is therefore pinned to $\Psi_n(\overline{\cos})$, a monotone function of mean cosine. In this regime, α_t and mean cosine produce nearly the same ranking of rounds, and the cone can add little beyond cosine. Substantial departures from this near-exchangeable regime, such as structured conflict (Figure 1), can allow the two measures to differ.

D.11 Proof of Theorem 4

One condition controls all four scalars: For unit vectors let $m = \|\frac{1}{n}\sum_i {}_i\|^2$. By Equation (1), $\overline{\cos} = (nm - 1)/(n-1)$. The aggregate norm is $n\sqrt{m}$. The gradient variance on unit directions is

$\frac{1}{n}\sum_i \|_i - \|^2 = 1 - m$, and the dissimilarity constant is $B^2 = 1/m$. At fixed n, all four are strictly monotone in m, and so in $\overline{\cos}$. It therefore suffices to compare mean cosines.

Population $\mathcal{A}_n$. Of the $\binom{n}{2}$ pairs, $\binom{n-1}{2}$ have cosine $+1$ and $n-1$ have cosine -1. Hence

$$\overline{\cos}(\mathcal{A}_n) = \frac{\binom{n-1}{2} - (n-1)}{\binom{n}{2}} = 1 - \frac{4}{n}. \tag{21}$$

Population $\mathcal{B}_n$. Let $a = \pi/3$ and $x = a/(n-1)$. The angles are $\theta_k = -a + 2kx$ for $k = 0, \ldots, n-1$. A geometric sum gives $|\sum_k e^{\mathrm{i}\theta_k}| = \sin(nx)/\sin x$. By Equation (1),

$$\overline{\cos}(\mathcal{B}_n) = \frac{1}{n(n-1)}\Big(\frac{\sin^2(nx)}{\sin^2 x} - n\Big). \tag{22}$$

Lemma 17 *$\overline{\cos}(\mathcal{B}_n) < c_\infty = (\sin a/a)^2 = 27/(4\pi^2)$ for every $n \geq 10$.*

Write $f(y) = \sin^2 y/y$, so that $c_\infty = f(a)/a$. Substitute $n = (a+x)/x$ and $n - 1 = a/x$ into (22). After multiplying out, the claim becomes

$$\Big(\frac{x}{\sin x}\Big)^2 f(a+x) \;<\; x + f(a). \tag{23}$$

We bound the two factors on the left. First, $f'(y) = \sin y\,(2y\cos y - \sin y)/y^2$. The bracket is decreasing for $y \geq a$, because its derivative $\cos y - 2y\sin y$ is negative there. Hence $f'(y) \leq (2a\cos a - \sin a)/a^2 < 0.2$ on $[a, 10a/9]$. For $n \geq 10$ we have $x \leq a/9$, so the mean value theorem gives $f(a+x) \leq f(a) + 0.2\,x$. Second, $\sin x \geq x - x^3/6$ implies $(x/\sin x)^2 \leq (1 - x^2/6)^{-2}$. It thus suffices that $f(a) + 0.2\,x \leq (x + f(a))(1 - x^2/6)^2$. Since $(1 - x^2/6)^2 \geq 1 - x^2/3$, this follows from $0.8\,x \geq (x + f(a))\,x^2/3$, that is, from $(x + f(a))\,x \leq 2.4$. This holds with a wide margin, because $x < 0.12$ and $f(a) < 0.72$.

Part (i). For $n \geq 13$, $\overline{\cos}(\mathcal{A}_n) = 1 - 4/n \geq 0.692$. By Lemma 17, this exceeds $c_\infty > \overline{\cos}(\mathcal{B}_n)$, and the gap is at least $1 - 4/n - c_\infty$. For $n \in \{10, 11, 12\}$, direct evaluation of (22) gives 0.584, 0.594 and 0.603, against 0.600, 0.636 and 0.667 for $\mathcal{A}_n$. For $3 \leq n \leq 9$, the same evaluation gives the opposite order, so the threshold is sharp (Table 1). The ranking transfers to the other three scalars by the monotonicity shown above.

Part (ii). In $\mathcal{A}_n$, the clients at 0° require $d_x < 0$, and the client at 180° requires $d_x > 0$. No direction satisfies both, so $\mathcal{C}_t =$. Equivalently, $\frac{1}{2}\mathbf{1} + \frac{1}{2}\mathbf{n} = 0$ lies in the convex hull. In $\mathcal{B}_n$, a direction at angle φ is a descent direction for client k exactly when φ differs from θ_k by more than 90°. With θ_k covering $[-60^\circ, 60^\circ]$, this holds for all k exactly when $\varphi \in (150^\circ, 210^\circ)$. The arc has width 60°, so $\alpha_t = 60/360 = 1/6$. In $\mathbb{R}^d$ the descent condition depends only on the projection of z onto the plane of the gradients. That projection is isotropic, so α_t is unchanged.

The statistical dimension: $K(\mathcal{A}_n)$ is a line, so $\Delta_t = 1$. $K(\mathcal{B}_n)$ is a planar cone of angle 120°. A planar cone of angle $\theta \leq \pi$ has statistical dimension $\frac{1}{2} + \theta/\pi$, so $\Delta_t(\mathcal{B}_n) = \frac{1}{2} + \frac{2}{3} = 7/6$.

D.12 Proof of Theorem 5

$\mathcal{K}_t^{(\tau)}$ is a convex cone. Fix i. For $\tau > 0$ the constraint $\langle_i, v\rangle \geq \tau\|v\|$ reads $\|v\| \leq \langle_i, v\rangle/\tau$. This is a second-order (Lorentz) cone, hence convex. Both sides are positively homogeneous of degree one, so it is a cone. For $\tau = 0$ it is a half-space. $\mathcal{K}_t^{(\tau)}$ is a finite intersection of such sets, hence a closed convex cone. Its interior is $\operatorname{int}\mathcal{K}_t^{(\tau)} = \{v : \langle_i, v\rangle > \tau\|v\|\ \forall i\}$. Two facts follow. First, for $v \in \mathcal{K}_t^{(\tau)}$ and $v' \in \operatorname{int}\mathcal{K}_t^{(\tau)}$, we have $v + v' \in \operatorname{int}\mathcal{K}_t^{(\tau)}$. Indeed, $\langle_i, v + v'\rangle > \tau(\|v\| + \|v'\|) \geq \tau\|v + v'\|$. Second, $\mathcal{K}_t^{(\tau)} \cap \mathbb{S}^{d-1}$ is a finite intersection of spherical caps. Its boundary lies in finitely many small spheres and has zero spherical measure. Hence $\operatorname{int}\mathcal{K}_t^{(\tau)}$ and $\mathcal{K}_t^{(\tau)}$ have the same normalized volume $\alpha_t^{(\tau)}$.

(i) Sufficiency. By L-smoothness, for the unit step $-\eta\hat{u}$,

$$F_i(w_t - \eta\hat{u}) - F_i(w_t) \;\le\; -\eta\,\|g_i\|\,\langle \imath_i, \hat{u}\rangle + \tfrac{L\eta^2}{2}. \tag{24}$$

The right-hand side is negative exactly when $\langle \imath_i, \hat{u}\rangle > L\eta/(2\|g_i\|)$. Since $\|g_i\| \ge \min_j \|g_j\|$, we have $L\eta/(2\|g_i\|) \le \tau$. The condition $\langle \imath_i, \hat{u}\rangle > \tau$ therefore suffices for every i at once. As $\|\hat{u}\| = 1$, this is exactly $\hat{u} \in \operatorname{int} \mathcal{K}_t^{(\tau)}$.

(ii) Aligned case. Suppose $\bar{g} \in \mathcal{K}_t^{(\tau)}$. On the event $\{\xi \in \operatorname{int} \mathcal{K}_t^{(\tau)}\}$, the first fact above gives $u = \bar{g} + \xi \in \operatorname{int} \mathcal{K}_t^{(\tau)}$. Since the interior is a cone, $\hat{u} \in \operatorname{int} \mathcal{K}_t^{(\tau)}$ as well. Part (i) then yields strict descent for every client, so

$$[\text{all clients improve}] \;\ge\; \big[\xi \in \operatorname{int} \mathcal{K}_t^{(\tau)}\big]. \tag{25}$$

ξ is spherically symmetric with $[\xi = 0] = 0$, so $\xi/\|\xi\|$ is uniform on $\mathbb{S}^{d-1}$. The event $\{\xi \in \operatorname{int} \mathcal{K}_t^{(\tau)}\}$ depends on ξ only through $\xi/\|\xi\|$, because the set is a cone. Its probability is therefore the normalized spherical measure of $\operatorname{int} \mathcal{K}_t^{(\tau)}$, which is $\alpha_t^{(\tau)}$. This holds at every noise level and needs no assumption on the law of $\hat{u}$.

(iii) Noise-dominated limit. Let $\xi = \sigma\zeta$ with ζ spherically symmetric and $[\zeta = 0] = 0$. Then $\hat{u} = (\bar{g}/\sigma + \zeta)/\|\bar{g}/\sigma + \zeta\|$. As $\|\bar{g}\|/\sigma \to 0$, the numerator converges to $\zeta \neq 0$. Normalization is continuous away from the origin, so $\hat{u} \to \zeta/\|\zeta\|$ almost surely. The limit is uniform on $\mathbb{S}^{d-1}$. The boundary of $\mathcal{K}_t^{(\tau)} \cap \mathbb{S}^{d-1}$ is null under this law. The portmanteau theorem therefore gives $[\hat{u} \in \mathcal{K}_t^{(\tau)}] \to \alpha_t^{(\tau)}$.

Monotonicity in τ. $\mathcal{K}_t^{(\tau)} \subseteq \mathcal{K}_t^{(\tau')}$ whenever $\tau \ge \tau'$. Moreover, $\bigcup_{\tau>0} \mathcal{K}_t^{(\tau)} = \{v : \min_i \langle \imath_i, v\rangle > 0\} \cup \{0\} = -\mathcal{C}_t \cup \{0\}$. Any $v \neq 0$ in the middle set lies in $\mathcal{K}_t^{(\tau)}$ for $\tau = \min_i \langle \imath_i, v\rangle/\|v\| > 0$. Continuity from below of the spherical measure gives $\alpha_t^{(\tau)} \uparrow \sigma(-\mathcal{C}_t)/\sigma(\mathbb{S}^{d-1}) = \alpha_t$ as $\tau \downarrow 0$.

Remark 18 (The misaligned case) *Suppose $\bar{g} \notin \mathcal{K}_t^{(\tau)}$. Let $q = \Pi_{\mathcal{K}_t^{(\tau)}}(\bar{g})$ and $\delta = \|\bar{g} - q\|$. Write $\bar{g} + \xi = q + (\xi - (q - \bar{g}))$. Since $q \in \mathcal{K}_t^{(\tau)}$, the first fact in the proof gives $[\hat{u} \in \operatorname{int} \mathcal{K}_t^{(\tau)}] \ge [\xi \in (q - \bar{g}) + \operatorname{int} \mathcal{K}_t^{(\tau)}]$. This is the probability that the noise lands in a translated cone. With $\xi = \sigma\zeta$, it depends on $(\bar{g}, \sigma)$ only through the vector $(q - \bar{g})/\sigma$, whose norm is $\rho = \delta/\sigma$. As $\rho \to 0$ it tends to $\alpha_t^{(\tau)}$, by the same boundary argument as in part (iii). We do not give a quantitative lower bound for fixed $\rho > 0$. The natural route inscribes a circular cone of fixed half-angle, but that bound decays exponentially in d. It is therefore uninformative at the dimensions of federated models. This is the main gap left in the stochastic analysis.*

D.13 Scope and limits of Theorem 5

Why volume, not inradius?: If the server could pick any direction, the right quantity would be the inradius of $\mathcal{C}_t$ (Lemma 10, Appendix D.5). The server cannot choose, however. The aggregate is produced by the clients, and it is random through client sampling and local minibatch noise. Under the noise model (11), the probability that the realized direction lands in the cone is governed by the cone's volume. This is why Theorem 5 is stated in terms of $\alpha_t^{(\tau)}$.

Three limits. First, the guarantee concerns simultaneous improvement within one round. It is not a convergence result for the averaged objective. Second, part (ii) requires the mean aggregate $\bar{g}$ to lie in the margin cone. When it does not, only noise can move the step into the cone, and we give no quantitative bound (Remark 18). Third, the bound is loose at low noise, where a union bound over clients is stronger (Remark 11). The two bounds are complementary. The union bound governs low noise, and $\alpha_t^{(\tau)}$ governs high noise, where it is exact by part (iii).

Link to the computed statistic: α_t and Δ_t are linked only when $K = \{_1, \ldots, _n\}$ is pointed. This holds exactly when $\overline{\mathcal{C}_t}$ is full-dimensional. Then α_t is the top conic intrinsic volume of $\overline{\mathcal{C}_t}$. These intrinsic volumes concentrate around $\delta(\overline{\mathcal{C}_t}) = d - \Delta_t$ (Amelunxen et al., 2014, Thm. 6.1). Hence α_t is non-negligible only when Δ_t is small (Corollary 13). When K is not pointed, the min-norm point of $\mathrm{conv}\{_i\}$ is zero. This is detectable at no extra cost (Section 3.3).

Relation to Proposition 6. Theorem 5 uses the normalized step $\eta\hat{u}$. There $\|\hat{u}\| = 1$, and the per-client margins reduce to $\tau_i = L\eta/(2\|g_i\|) \le \tau$. Proposition 6 covers the unnormalized FedAvg step, where the margins scale with $\|u\|$. Remark 12 gives the corresponding version of Theorem 5 on the event $\{\|u\| \le R\}$.

D.14 Proof of Proposition 6

L-smoothness of F_i gives, for any w and any step s,

$$\left|F_i(w+s) - F_i(w) - \langle \nabla F_i(w), s\rangle\right| \;\le\; \tfrac{L}{2}\|s\|^2. \tag{26}$$

Take $w = w_t$, $s = -\eta u$ and $\nabla F_i(w_t) = g_i$, and write $\langle g_i, u\rangle = \|g_i\|\,\|u\|\,\langle _i, \hat{u}\rangle$. The upper half of the bound gives

$$F_i(w_{t+1}) - F_i(w_t) \;\le\; -\eta\|g_i\|\|u\|\,\langle _i, \hat{u}\rangle + \tfrac{L}{2}\eta^2\|u\|^2. \tag{27}$$

The right-hand side is negative exactly when $\langle _i, \hat{u}\rangle > L\eta\|u\|/(2\|g_i\|) = \tau_i$. This proves the first implication. The lower half gives

$$F_i(w_{t+1}) - F_i(w_t) \;\ge\; -\eta\|g_i\|\|u\|\,\langle _i, \hat{u}\rangle - \tfrac{L}{2}\eta^2\|u\|^2, \tag{28}$$

whose right-hand side is positive exactly when $\langle _i, \hat{u}\rangle < -\tau_i$. This proves the second implication. Inside the band, both bounds allow either sign of $F_i(w_{t+1}) - F_i(w_t)$, so direction alone does not decide the outcome.

Inexact client directions. In practice g_i is often the negated local displacement, not the exact gradient (Section 3.1). Let $\tilde{g}_i = \nabla F_i(w_t)$, and suppose $\|_i - \tilde{g}_i/\|\tilde{g}_i\|\| \le \epsilon_i$. Since $\|\hat{u}\| = 1$, Cauchy–Schwarz gives $|\langle _i, \hat{u}\rangle - \langle \tilde{g}_i/\|\tilde{g}_i\|, \hat{u}\rangle| \le \epsilon_i$. Both implications therefore hold with τ_i computed from $\|\tilde{g}_i\|$ and widened to $\tau_i + \epsilon_i$.

D.15 Proof of Remark 11

Here $\xi \sim \mathcal{N}(0, \sigma^2 I_d)$, as the remark assumes. Set $Z_i = \langle _i, \xi\rangle \sim \mathcal{N}(0, \sigma^2)$ and $\gamma_i = \langle _i, \bar{g}\rangle \ge \gamma$, so that $\langle _i, u\rangle = \gamma_i + Z_i$. By Vershynin (2018, Thm. 3.1.1), $\|\xi\| \le \sigma(\sqrt{d} + s)$ with probability at least $1 - e^{-s^2/2}$. On that event, $\|u\| \le \|\bar{g}\| + \sigma(\sqrt{d} + s)$. The requirement $\langle _i, u\rangle > \tau\|u\|$ is then implied by $\gamma_i + Z_i > c$, with $c = \tau(\|\bar{g}\| + \sigma(\sqrt{d} + s))$. Moreover, $[\gamma_i + Z_i \le c] \le \overline{\Phi}((\gamma - c)/\sigma)$. A union bound over the n clients, together with the norm event, gives the stated inequality.

E Numerical Implementation

E.1 Factorizing a singular Gram matrix

Line 3 of Algorithm 1 needs A with $A^\top A = G$. Cholesky requires $G \succ 0$, which fails whenever $n > \mathrm{rank}(V)$. The natural repair factoring $G + \tau I_n$ for small $\tau > 0$ is incorrect, and the reason is worth recording.

Sampling $b = \mathrm{chol}(G + \tau I)\,u$ with $u \sim \mathcal{N}(0, I_n)$ gives $b \sim \mathcal{N}(0, G + \tau I)$, so b carries a component outside $\mathrm{col}(G)$. Along any direction v in the null space of G the objective $\lambda^\top G\lambda - 2b^\top\lambda$ reduces to $-2b^\top\lambda$, which is unbounded below on $\lambda \ge 0$; only the implicit $\tau\|\lambda\|^2$ arrests the minimization, at $\lambda \sim \langle b, v\rangle/\tau$. Since $\langle b, v\rangle \sim \mathcal{N}(0, \tau)$, the resulting contribution to $b^\top\lambda^*$ has expectation $1/2$ per null direction, independently of τ. A rank deficiency of r therefore inflates $\widehat{\Delta}_t$ by exactly $r/2$, and the bias does not vanish as $\tau \to 0$.

We instead take the eigendecomposition $G = U\,\mathrm{diag}(w)U^\top$, retain the k eigenvalues exceeding $10^{-10}\max(w)$, and set $A = \mathrm{diag}(\sqrt{w_{1:k}})\,U^\top_{:,1:k} \in \mathbb{R}^{k\times n}$. Then $A^\top A = G$ exactly, and $b =$

$A^\top c$ with $c \sim \mathcal{N}(0, I_k)$ satisfies both $b \sim \mathcal{N}(0, G)$ and $b \in \mathrm{col}(G)$ by construction, with no regularization anywhere.

Section 6 validates this against configurations of known statistical dimension: identical gradients ($\Delta_t = 1/2$), mutually orthogonal gradients ($n/2$), and rank-r cancellation (r, for $r \in \{1, 2, 3, 5\}$). All ten cases agree with the analytic value to within 0.010.

E.2 NNLS SOLVER

We use the active-set method of Lawson & Hanson (1995); any NNLS or projected-gradient solver suffices, the problem being convex with fixed Hessian G. Consecutive Gaussian samples usually activate the same constraints, so warm-starting each solve from the previous solution lets most terminate after a few active-set exchanges rather than the worst-case $O(n^3)$.

E.3 REDUCING SERVER MEMORY BY SKETCHING

When the server cannot hold $V \in \mathbb{R}^{d\times n}$, apply a Johnson–Lindenstrauss sketch $S \in \mathbb{R}^{k\times d}$ to each update on arrival and retain only S_i, reducing storage to nk floats. With $k = O(\varepsilon^{-2} \log n)$ the sketched Gram matrix $\widetilde{G}$ satisfies $|\widetilde{G}_{ij} - G_{ij}| \le \varepsilon$ for all i, j with high probability (Dasgupta & Gupta, 2003; Achlioptas, 2001); a structured transform keeps the per-client cost at $O(d \log d)$. Sketching buys memory, not time: since k typically exceeds n, applying the sketch costs more arithmetic than forming G directly, so it is worthwhile only when V does not fit.

Transferring the entrywise guarantee to one on Δ_t requires a stability result. The statistical dimension is rotation-invariant (Amelunxen et al., 2014, Prop. 3.1), hence a function $\Delta(G)$ of the Gram matrix alone, and a Cauchy–Schwarz argument together with non-expansiveness of the projection gives

$$\left|\Delta(G) - \Delta(\widetilde{G})\right| \;\le\; 2n\,\theta(K, \widetilde{K}), \tag{29}$$

with θ the conic Hausdorff distance between the two cones. Bounding θ by the generator perturbation carries a conditioning factor that degenerates for nearly flat cones, so we do not claim a bound uniform over all configurations, and instead measure the realized error against exact values for $n \le 12$ in Section 6. All experiments reported here compute G directly; sketching was not required at our scale.

F ADDITIONAL EXPERIMENTS

This appendix gives the full results behind Section 6. Sections F.1–F.3 concern the geometry and the estimators (Q1–Q2); Sections F.4–F.8 concern federated training (Q3). Unless stated otherwise, intervals are 95% t-intervals over 5 seeds, and all federated statistics are computed per seed within one (dataset, β) cell before aggregation.

F.1 THE SEPARATION ACROSS THE NUMBER OF CLIENTS

Figure 4 extends Figure 2 with the Δ_t panel, and Table 1 lists every value. Mean cosine and B^2 (both computed on unit directions) rank $\mathcal{A}_n$ as more homogeneous from $n = 10$ onward, and never for $n \le 8$. α_t is computed both in closed form and numerically; the numerical value for $\mathcal{B}_n$ is 0.166666 at every n, against the exact $1/6$. $\widehat{\Delta}_t$ uses $M = 20{,}000$ samples. Its exact values are 1 for $\mathcal{A}_n$, whose generated cone is a line, and $7/6$ for $\mathcal{B}_n$, a planar cone of angle 120° with $\delta = \frac{1}{2} + \frac{120^\circ}{180^\circ}$. All 16 estimates lie within 3 standard errors of these values. The ordering $\Delta_t(\mathcal{A}_n) < \Delta_t(\mathcal{B}_n)$ is the non-pointed failure discussed after Corollary 13.

F.2 THE ESTIMATORS RECOVER KNOWN VALUES

Statistical dimension: We validate Algorithm 1 on three families with closed-form Δ_t: n identical gradients ($\Delta_t = \frac{1}{2}$); n mutually orthogonal gradients ($\Delta_t = n/2$); and the antipodal basis $\{\pm e_1, \ldots, \pm e_r\}$, whose generated cone is the r-dimensional subspace $\mathrm{span}\{e_1, \ldots, e_r\}$ ($\Delta_t = r$). The antipodal family has a singular, rank-deficient G and non-pointed K, so it exercises the range-space factorization of Appendix E. A regularized Cholesky factor would inflate these estimates by

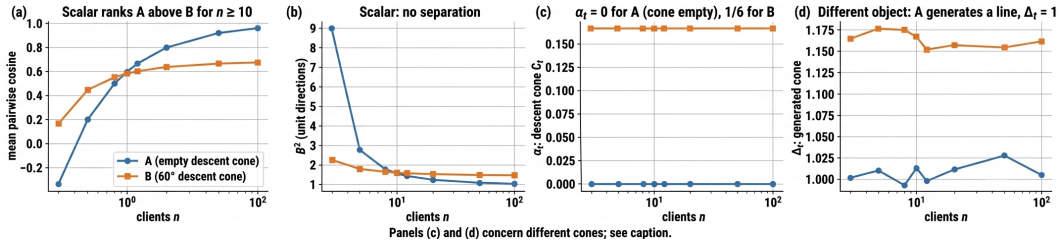


Figure 4: Separation of Theorem 4 for $n \in \{3, 5, 8, 10, 12, 20, 50, 100\}$. (a) Mean pairwise cosine; (b) B^2 on unit directions; (c) α_t of the agreement cone $\mathcal{C}_t$; (d) Δ_t of the generated cone K. Panels (c) and (d) concern different cones: α_t measures the room for common descent, while Δ_t measures the dimension the gradients occupy, and in the non-pointed case (a) they disagree (Corollary 13).

Table 1: Scalar measures and $\widehat{\Delta}_t$ for $\mathcal{A}_n$ and $\mathcal{B}_n$. ✓ marks n at which the scalar ranks $\mathcal{A}_n$ (empty agreement cone) as more homogeneous than $\mathcal{B}_n$ (60° cone). $\alpha_t(\mathcal{A}_n) = 0$ and $\alpha_t(\mathcal{B}_n) = 1/6$ at every n. $\widehat{\Delta}_t$ is mean $\pm$ standard error, $M = 20{,}000$; exact values are 1 and $7/6 \approx 1.167$.

	mean pairwise cosine			B^2 (unit)			$\widehat{\Delta}_t$	
n	$\mathcal{A}_n$	$\mathcal{B}_n$	wrong	$\mathcal{A}_n$	$\mathcal{B}_n$	wrong	$\mathcal{A}_n$	$\mathcal{B}_n$
3	−0.333	0.167	✗	9.000	2.250	✗	1.001 ±0.010	1.165 ±0.012
5	0.200	0.446	✗	2.778	1.795	✗	1.010 ±0.010	1.177 ±0.012
8	0.500	0.554	✗	1.778	1.641	✗	0.992 ±0.010	1.175 ±0.012
10	0.600	0.584	✓	1.563	1.599	✓	1.012 ±0.010	1.167 ±0.012
12	0.667	0.603	✓	1.440	1.572	✓	0.997 ±0.010	1.152 ±0.012
20	0.800	0.638	✓	1.235	1.525	✓	1.011 ±0.010	1.157 ±0.012
50	0.920	0.666	✓	1.085	1.486	✓	1.027 ±0.010	1.154 ±0.012
100	0.960	0.675	✓	1.041	1.474	✓	1.004 ±0.010	1.162 ±0.012

$r/2$. With $M = 50{,}000$, every 95% interval covers the exact value, and the standardized errors $z = (\widehat{\Delta}_t - \Delta_t)/\mathrm{SE}$ lie in $[-1.06, 1.36]$, consistent with an unbiased estimator (Figure 5, Table 2).

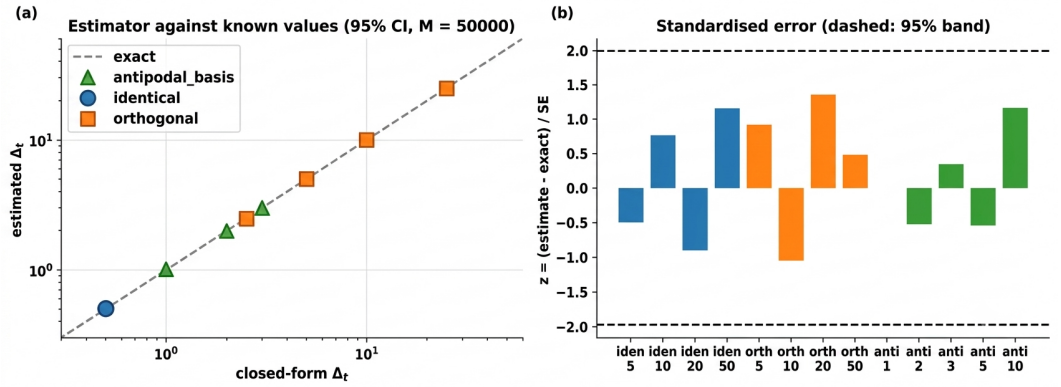


Figure 5: (a) $\widehat{\Delta}_t$ against the closed-form value for 13 anchor configurations ($M = 50{,}000$, 95% intervals). (b) Standardized error; dashed lines mark ± 1.96.

Cone volume. $\log \alpha_t$ is computed with the separation-of-variables orthant estimator of Genz & Bretz (2009) using $N = 16{,}384$ samples. For constant-correlation matrices with $\rho \geq 0$, Lemma 15 gives $\alpha_t = \Psi_n(\rho)$ in closed form. Table 3 compares the two for $n \in \{5, 10, 20\}$: the relative error is at most 1.44%, and eight of nine cases lie within two standard errors. This check covers

Table 2: Anchor configurations for $\widehat{\Delta}_t$, $M = 50{,}000$. Every 95% interval covers the exact value.

Family	Size	Exact Δ_t	$\widehat{\Delta}_t$	SE	rel. err. (%)	z
Identical	$n = 5$	0.5	0.4975	0.0050	0.50	−0.50
	$n = 10$	0.5	0.5039	0.0050	0.77	0.77
	$n = 20$	0.5	0.4955	0.0050	0.91	−0.91
	$n = 50$	0.5	0.5058	0.0050	1.17	1.16
Orthogonal	$n = 5$	2.5	2.5104	0.0112	0.41	0.93
	$n = 10$	5	4.9834	0.0157	0.33	−1.06
	$n = 20$	10	10.0303	0.0224	0.30	1.35
	$n = 50$	25	25.0173	0.0353	0.07	0.49
Antipodal basis	$r = 1$	1	1.0000	0.0063	0.00	−0.00
	$r = 2$	2	1.9953	0.0089	0.24	−0.53
	$r = 3$	3	3.0039	0.0110	0.13	0.36
	$r = 5$	5	4.9923	0.0141	0.15	−0.54
	$r = 10$	10	10.0234	0.0200	0.23	1.17

only nonnegative correlations. In the federated runs, where some pairwise cosines are negative, the estimator's relative standard error stayed below 0.5 in every round (Table 5).

Table 3: Orthant estimator against the closed form $\Psi_n(\rho)$ for constant-correlation Gram matrices ($N = 16{,}384$).

n	ρ	$\log \alpha_t$ (estimate)	$\log \alpha_t$ (exact)	rel. err. (%)	rel. SE (%)
5	0.05	−3.1775	−3.1785	0.10	0.09
	0.2	−2.5650	−2.5586	0.64	0.27
	0.5	−1.7912	−1.7918	0.06	0.41
10	0.05	−5.7950	−5.7937	0.13	0.25
	0.2	−3.9984	−3.9974	0.11	0.63
	0.5	−2.3979	−2.3979	0.00	0.72
20	0.05	−9.9549	−9.9594	0.45	0.66
	0.2	−5.7780	−5.7923	1.44	1.36
	0.5	−3.0381	−3.0445	0.64	1.16

F.3 Sample complexity

For $n \in \{4, 8, 16, 32, 64\}$ we draw 20 random Gram matrices per n. We estimate a reference value Δ_t^{ref} of each with $M = 20{,}000$, then re-estimate at $M \in \{10^2, 5{\times}10^2, 10^3, 5{\times}10^3, 10^4\}$. This gives 500 cells. For each we compare $|\widehat{\Delta}_t - \Delta_t^{\mathrm{ref}}|$ with the error ε that Proposition 2 guarantees at $\delta = 0.05$. The bound held in all 500 cells, so empirical coverage is 1.00 against the nominal 0.95. The ratio $\varepsilon/|\text{error}|$ has median 5.5 and minimum 1.15, and even the 95th-percentile error stays below ε in every (n, M) group (Figure 6, Table 4). Relative error falls as n, and with it Δ_t, grows, as the relative form of Proposition 2 predicts: at $M = 10^4$ it is 0.9% for $n = 4$ and 0.2% for $n = 64$.

F.4 Federated runs: setup and regime

Each (dataset, β) cell comprises 5 seeds of 300 rounds with $n = 20$ clients sampled per round, 5 local epochs and learning rate 0.01. Table 5 summarizes the regime. "Top-class share" is the mean fraction of a client's data in its most frequent class and confirms that β controls label skew. A round is *majority-harm* if more than half of the sampled clients' losses increase. Three features of the regime matter for Section 6.3. Majority-harm rounds are frequent only at $\beta \le 0.1$. Client updates are nearly orthogonal throughout (mean cosine at most 0.046, $\widehat{\Delta}_t$ close to $n/2 = 10$). And α_t is of order e^{-12} to e^{-13}, so it can only be handled on the log scale, as Corollary 13 anticipates. In no round did the orthant estimator have relative standard error above 0.5, and no round produced an undefined $\log \alpha_t$.

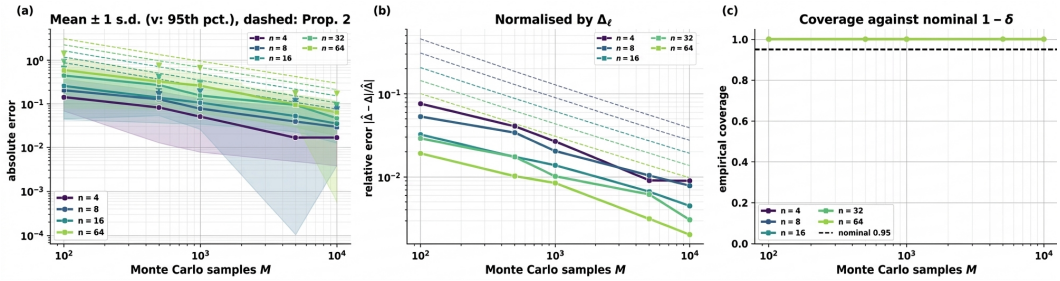


Figure 6: (a) Absolute error (mean ± 1 s.d.; triangles: 95th percentile) against the Proposition 2 bound (dashed). (b) Relative error $|\widehat{\Delta}_t - \Delta_t|/\Delta_t$ against the relative bound. (c) Empirical coverage against the nominal $1 - \delta = 0.95$.

Table 4: Sample complexity over 20 Gram matrices per (n, M). ε is the Proposition 2 bound at $\delta = 0.05$; "slack" is $\varepsilon/|\text{error}|$.

n	$\bar{\Delta}_t^{\text{ref}}$	M	mean err.	p95 err.	ε	coverage	median slack	min slack
4	2.00	100	0.146	0.258	0.925	1.00	8.1	3.28
		500	0.084	0.169	0.372	1.00	4.7	1.25
		1,000	0.052	0.115	0.257	1.00	6.3	1.68
		5,000	0.017	0.040	0.111	1.00	8.1	2.42
		10,000	0.017	0.039	0.078	1.00	6.1	1.90
8	3.96	100	0.209	0.471	1.233	1.00	7.4	2.26
		500	0.129	0.221	0.511	1.00	3.8	2.13
		1,000	0.080	0.197	0.355	1.00	5.6	1.67
		5,000	0.040	0.123	0.155	1.00	7.6	1.15
		10,000	0.030	0.075	0.109	1.00	5.9	1.33
16	8.21	100	0.264	0.680	1.708	1.00	8.2	2.04
		500	0.141	0.301	0.725	1.00	5.8	2.36
		1,000	0.109	0.205	0.506	1.00	4.2	2.35
		5,000	0.054	0.107	0.223	1.00	4.9	1.97
		10,000	0.036	0.081	0.157	1.00	4.6	1.61
32	16.05	100	0.458	0.982	2.326	1.00	6.6	1.91
		500	0.275	0.434	1.002	1.00	3.6	2.11
		1,000	0.158	0.312	0.702	1.00	5.3	2.16
		5,000	0.097	0.183	0.310	1.00	2.6	1.63
		10,000	0.048	0.095	0.219	1.00	4.7	1.69
64	31.90	100	0.610	1.492	3.217	1.00	6.7	1.50
		500	0.331	0.795	1.401	1.00	6.2	1.47
		1,000	0.270	0.694	0.984	1.00	4.0	1.19
		5,000	0.098	0.186	0.436	1.00	4.5	1.92
		10,000	0.064	0.181	0.308	1.00	6.7	1.36

F.5 Spearman correlation: all measures

Table 6 reports the oriented Spearman correlation between every measure and IUF_t. Orientation is fixed before looking at the data: cosine, min-norm and $\log \alpha_t$ enter as they are, while variance, B^2 and Δ_t are negated, so a positive value is the predicted direction. The unit-normalized variance and B^2 are not listed separately. By Equation (1) they are strictly monotone in mean cosine at fixed n, and their Spearman correlations coincide with cosine's to every reported digit.

Two features of the magnitude-carrying measures deserve mention. Raw variance is negative in five of eight cells, significantly so in four: rounds whose updates vary more in magnitude tend to improve more clients, strongly so at $\beta = 100$ (-0.309 and -0.382). Raw B^2 has the predicted sign in six of eight cells and is the strongest measure on CIFAR-10 at $\beta = 100$ (0.256). Both effects come from update magnitudes, which every directional measure, including ours, discards (Remark 8). This

Table 5: Regime of the federated runs (means over 5 seeds $\times$ 300 rounds; accuracy is final-round test accuracy $\pm$ 95% half-width). $\max \mathrm{SE}$ is the largest per-round standard error of $\widehat{\Delta}_t$.

Dataset	β	top-class share	final acc.	majority-harm rate	$\overline{\mathrm{IUF}}_t$	$\overline{\cos}$	$\widehat{\Delta}_t$	$\max \mathrm{SE}$	median $\log \alpha_t$
CIFAR-10	0.05	0.793	0.589 $\pm$0.028	0.136	0.642	0.046	8.22	0.245	-11.49
	0.1	0.661	0.596 $\pm$0.009	0.131	0.643	0.035	8.56	0.253	-11.65
	0.5	0.388	0.609 $\pm$0.011	0.017	0.738	0.016	9.25	0.251	-12.80
	100	0.133	0.623 $\pm$0.013	0.006	0.961	0.012	9.54	0.251	-13.28
FMNIST	0.05	0.783	0.853 $\pm$0.033	0.209	0.621	0.035	8.58	0.245	-13.09
	0.1	0.674	0.880 $\pm$0.006	0.161	0.626	0.031	8.82	0.257	-12.93
	0.5	0.384	0.893 $\pm$0.004	0.053	0.688	0.020	9.24	0.255	-12.98
	100	0.134	0.901 $\pm$0.003	0.000	0.951	0.015	9.50	0.251	-13.36

supports the reading in Section 6.3 that magnitude, not direction, drives harm in the near-IID and moderate regimes.

Table 6: Oriented Spearman ρ with IUF_t (mean $\pm$ 95% half-width over 5 seeds, 300 rounds each). Positive is the direction predicted by theory. $^\circ$: interval contains zero. Bold: largest among the four directional measures (cosine, min-norm, Δ_t, $\log \alpha_t$). "Raw" variance and B^2 use unnormalized updates.

Dataset	β	cosine	variance (raw)	B^2 (raw)	min-norm	Δ_t (ours)	$\log \alpha_t$ (ours)
CIFAR-10	0.05	**0.323** $\pm$0.051	-0.152 ± 0.067	0.253 ± 0.065	0.287 ± 0.048	0.298 ± 0.042	0.311 ± 0.045
	0.1	0.230 ± 0.045	$-0.033 \pm 0.072^\circ$	0.159 ± 0.050	**0.246** $\pm$0.057	0.226 ± 0.068	0.231 ± 0.052
	0.5	-0.151 ± 0.076	-0.161 ± 0.064	-0.181 ± 0.086	$\mathbf{-0.081} \pm 0.084^\circ$	$-0.108 \pm 0.116^\circ$	-0.126 ± 0.081
	100	$0.030 \pm 0.056^\circ$	-0.309 ± 0.066	0.256 ± 0.126	**0.091** $\pm$0.065	$0.040 \pm 0.075^\circ$	$0.059 \pm 0.061^\circ$
FMNIST	0.05	**0.356** $\pm$0.117	$0.043 \pm 0.075^\circ$	0.190 ± 0.102	0.232 ± 0.128	0.297 ± 0.111	0.295 ± 0.120
	0.1	**0.289** $\pm$0.047	$0.089 \pm 0.093^\circ$	0.131 ± 0.063	0.265 ± 0.063	0.262 ± 0.048	0.275 ± 0.052
	0.5	$-0.025 \pm 0.061^\circ$	$0.004 \pm 0.094^\circ$	-0.089 ± 0.083	**0.003** $\pm$0.072°	$0.001 \pm 0.047^\circ$	$-0.016 \pm 0.068^\circ$
	100	0.089 ± 0.049	-0.382 ± 0.055	0.079 ± 0.040	**0.143** $\pm$0.049	0.111 ± 0.051	0.115 ± 0.049

F.6 Association by regime: detailed results

This subsection gives the detailed results behind the three regimes in Section 6.3. All values are reported in Table 6.

Strong heterogeneity ($\beta \leq 0.1$): Both cone statistics carry the predicted association in all four cells. Their intervals exclude zero, and their signs are correct in every seed. The correlations lie in $[0.23, 0.30]$ for Δ_t and $[0.23, 0.31]$ for $\log \alpha_t$. Mean cosine ($[0.23, 0.36]$) and the min-norm magnitude ($[0.23, 0.29]$) span comparable ranges, and all intervals overlap. Majority-harm AUROC agrees, lying in $[0.60, 0.68]$ for all four directional measures (Appendix F.7).

Moderate heterogeneity ($\beta = 0.5$): On CIFAR-10, every directional measure is negatively associated with IUF_t, significantly so for mean cosine (-0.151) and $\log \alpha_t$ (-0.126). Only 1.7% of CIFAR-10 rounds harm a majority of clients in this setting (Table 5). On FMNIST, no directional measure is significant.

Near-IID ($\beta = 100$): On FMNIST, all four directional measures remain significant, led by the min-norm magnitude (0.143). On CIFAR-10, only the min-norm magnitude is significant (0.091), while the magnitude-carrying raw B^2 reaches 0.256, the largest value among all measures in that cell.

F.7 Majority-harm AUROC

For each seed we compute the AUROC of the score $-$(oriented measure) for classifying majority-harm rounds. The orientation is fixed a priori, so values below 0.5 are genuine reversals and not label

flips. Because the event is rare outside $\beta \leq 0.1$, AUROC is estimable only where a seed contains at least one majority-harm round. That leaves 3 seeds for CIFAR-10 at $\beta = 0.5$, a single seed for CIFAR-10 at $\beta = 100$, and no seeds for FMNIST at $\beta = 100$. We therefore interpret AUROC only at $\beta \leq 0.1$. There all agreement measures lie in $[0.60, 0.68]$ with overlapping intervals (Figure 7, Table 7). The isolated 0.957 for raw B^2 on CIFAR-10 at $\beta = 100$ comes from one seed with 0.6% positive rounds and should not be read as a result.

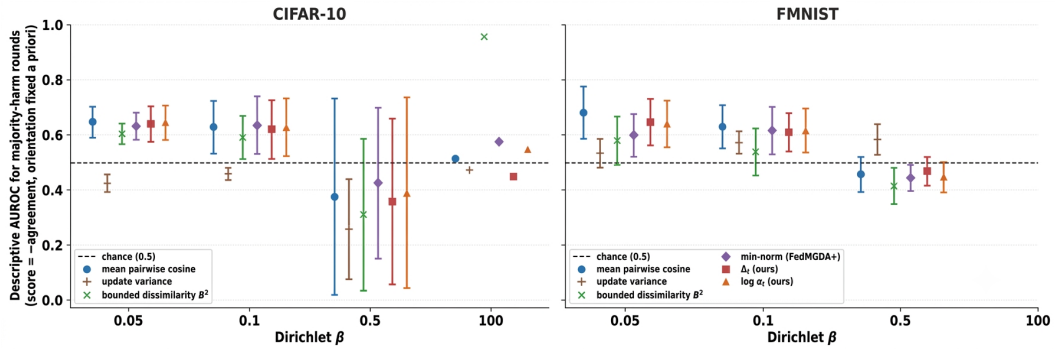


Figure 7: Descriptive majority-harm AUROC by measure and β (score $= -$agreement, orientation fixed a priori; 95% intervals over the seeds in which the event occurs). Only $\beta \leq 0.1$ has enough positive rounds for stable estimates.

Table 7: Descriptive majority-harm AUROC (mean $\pm$ 95% half-width). "Rate": fraction of majority-harm rounds; "seeds": seeds with at least one such round.

Dataset	β	rate	seeds	cosine	variance (raw)	B^2 (raw)	min-norm	Δ_t	$\log \alpha_t$
CIFAR-10	0.05	0.136	5	0.647 $\pm$0.056	0.425 $\pm$0.032	0.604 $\pm$0.038	0.632 $\pm$0.049	0.640 $\pm$0.065	0.646 $\pm$0.062
	0.1	0.131	5	0.629 $\pm$0.095	0.459 $\pm$0.022	0.591 $\pm$0.078	0.636 $\pm$0.104	0.619 $\pm$0.107	0.629 $\pm$0.105
	0.5	0.017	3	0.376 $\pm$0.357	0.258 $\pm$0.182	0.311 $\pm$0.276	0.425 $\pm$0.275	0.359 $\pm$0.301	0.389 $\pm$0.346
	100	0.006	1	0.515	0.473	0.957	0.576	0.449	0.548
FMNIST	0.05	0.209	5	0.681 $\pm$0.095	0.535 $\pm$0.052	0.580 $\pm$0.087	0.600 $\pm$0.078	0.648 $\pm$0.084	0.641 $\pm$0.085
	0.1	0.161	5	0.631 $\pm$0.079	0.573 $\pm$0.041	0.539 $\pm$0.085	0.617 $\pm$0.086	0.610 $\pm$0.069	0.617 $\pm$0.080
	0.5	0.053	5	0.458 $\pm$0.063	0.585 $\pm$0.056	0.415 $\pm$0.066	0.444 $\pm$0.047	0.468 $\pm$0.052	0.447 $\pm$0.054
	100	0.000	0	–	–	–	–	–	–

F.8 Conditional analysis: information beyond cosine

We use three complementary tests of whether a measure X carries information about IUF_t that mean cosine does not. All are computed per seed and aggregated over seeds.

- **Partial Spearman correlation:** We rank-transform X, IUF_t, cosine and the round index. We regress the first two on the controls (cosine alone, or cosine and round) and correlate the residuals. Round index is included because both the measures and IUF_t drift over training.
- **Leave-one-seed-out AUROC gain:** Fit a logistic regression for majority-harm on (cosine, round) and on (cosine, round, X) using four seeds, and evaluate AUROC on the held-out seed. A fold is valid only if the held-out seed contains both classes.
- **Cosine-stratified gap:** Split each seed's rounds into cosine quintiles. Within each quintile, split at the median of X, and record mean IUF_t of the more-agreeing half minus that of the less-agreeing half. Average over quintiles. This test makes no linearity assumption.

Table 8 reports all three. Controlling for round in addition to cosine matters. On FMNIST at $\beta = 100$, for example, the partial correlation of min-norm drops from 0.347 (cosine only) to 0.172 (cosine and round), so part of the apparent residual signal is a shared training-time trend. After both controls, the residual association is significant mainly for $\log \alpha_t$ and min-norm, in CIFAR-10 at $\beta = 0.5$ and FMNIST at $\beta = 100$. No measure improves out-of-seed classification. The stratified

gap shows a small but consistent residual effect of 0.3–2.9 percentage points. It is strongest and most reproducible on CIFAR-10 at $\beta \leq 0.1$, where all three measures are positive in every seed. It is weakest for Δ_t, which is also the only measure with a significantly negative partial correlation (FMNIST, $\beta = 0.05$). Per-seed, per-stratum values are provided in the supplementary material.

Table 8: Information beyond cosine. Partial ρ with IUF_t given cosine and round, and given cosine only; leave-one-seed-out AUROC gain over the (cosine, round) baseline; cosine-stratified IUF_t gap. Brackets are 95% intervals; bold marks intervals excluding zero. †: a single valid fold. "–": no majority-harm rounds. The last column counts seeds with a positive gap.

Dataset, β	Measure	partial ρ (cos, round)	partial ρ (cos only)	CV AUROC gain	stratified gap	seeds > 0
CIFAR-10, 0.05	Δ_t	0.022 [−0.022, 0.067]	0.013 [−0.033, 0.060]	−0.001 [−0.004, 0.002]	**0.014** [0.007, 0.020]	5/5
	$\log \alpha_t$	0.041 [−0.030, 0.113]	0.045 [−0.026, 0.116]	−0.001 [−0.007, 0.004]	**0.018** [0.008, 0.029]	5/5
	min-norm	0.051 [−0.026, 0.128]	0.068 [−0.005, 0.141]	0.001 [−0.009, 0.010]	**0.014** [0.003, 0.024]	5/5
CIFAR-10, 0.1	Δ_t	0.045 [−0.080, 0.171]	0.030 [−0.086, 0.146]	−0.010 [−0.035, 0.015]	**0.014** [0.003, 0.026]	5/5
	$\log \alpha_t$	0.039 [−0.083, 0.162]	0.035 [−0.083, 0.153]	−0.005 [−0.015, 0.005]	**0.012** [0.001, 0.022]	5/5
	min-norm	0.081 [−0.042, 0.205]	0.087 [−0.032, 0.205]	0.007 [−0.010, 0.024]	**0.017** [0.006, 0.027]	5/5
CIFAR-10, 0.5	Δ_t	0.024 [−0.097, 0.145]	0.029 [−0.092, 0.151]	0.001 [−0.013, 0.015]	0.010 [−0.009, 0.030]	4/5
	$\log \alpha_t$	**0.089** [0.064, 0.114]	**0.153** [0.129, 0.177]	−0.000 [−0.004, 0.004]	**0.015** [0.000, 0.030]	4/5
	min-norm	**0.057** [0.037, 0.077]	**0.155** [0.125, 0.185]	−0.000 [−0.002, 0.002]	**0.029** [0.020, 0.037]	5/5
CIFAR-10, 100	Δ_t	−0.005 [−0.099, 0.089]	0.028 [−0.053, 0.108]	−0.018†	0.008 [−0.010, 0.025]	3/5
	$\log \alpha_t$	0.112 [−0.014, 0.238]	**0.313** [0.205, 0.421]	−0.007†	**0.014** [0.010, 0.018]	5/5
	min-norm	0.074 [−0.018, 0.167]	**0.316** [0.229, 0.403]	0.005†	**0.025** [0.011, 0.038]	5/5
FMNIST, 0.05	Δ_t	**−0.041** [−0.081, −0.001]	**−0.056** [−0.102, −0.009]	−0.000 [−0.002, 0.002]	0.007 [−0.012, 0.026]	3/5
	$\log \alpha_t$	−0.034 [−0.086, 0.019]	−0.040 [−0.095, 0.015]	0.000 [−0.003, 0.003]	0.013 [−0.008, 0.033]	4/5
	min-norm	−0.051 [−0.146, 0.044]	−0.044 [−0.139, 0.052]	0.003 [−0.008, 0.014]	0.003 [−0.015, 0.021]	2/5
FMNIST, 0.1	Δ_t	−0.004 [−0.044, 0.036]	−0.018 [−0.062, 0.026]	−0.006 [−0.030, 0.018]	0.011 [−0.008, 0.030]	3/5
	$\log \alpha_t$	0.034 [−0.018, 0.086]	0.020 [−0.032, 0.073]	−0.002 [−0.004, 0.001]	**0.017** [0.002, 0.031]	5/5
	min-norm	0.053 [−0.018, 0.125]	0.052 [−0.020, 0.125]	−0.003 [−0.009, 0.003]	0.018 [−0.005, 0.040]	4/5
FMNIST, 0.5	Δ_t	**0.052** [0.030, 0.074]	**0.055** [0.035, 0.076]	−0.004 [−0.032, 0.025]	0.005 [−0.012, 0.023]	2/5
	$\log \alpha_t$	0.035 [−0.062, 0.132]	0.048 [−0.068, 0.164]	0.005 [−0.039, 0.049]	0.007 [−0.007, 0.020]	4/5
	min-norm	0.067 [−0.035, 0.168]	0.082 [−0.045, 0.208]	−0.003 [−0.012, 0.005]	0.014 [−0.002, 0.030]	4/5
FMNIST, 100	Δ_t	−0.002 [−0.059, 0.055]	**0.066** [0.008, 0.124]	–	**0.010** [0.003, 0.017]	5/5
	$\log \alpha_t$	**0.168** [0.130, 0.206]	**0.299** [0.239, 0.359]	–	**0.017** [0.007, 0.026]	5/5
	min-norm	**0.172** [0.144, 0.200]	**0.347** [0.301, 0.393]	–	**0.024** [0.018, 0.030]	5/5

Figure 8 summarizes the incremental value of each measure beyond mean cosine. Panel (a) reports partial Spearman correlations with IUF_t after controlling for mean cosine and communication round. Panel (b) reports the leave-one-seed-out AUROC gain from adding each measure to a cosine-plus-round baseline.

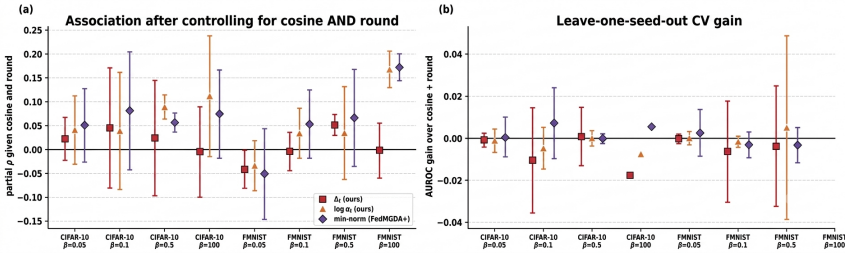


Figure 8: **Incremental value over cosine.** (a) Partial Spearman ρ with IUF_t after controlling for mean cosine and round index. (b) Leave-one-seed-out gain in majority-harm AUROC from adding each measure to a cosine-plus-round baseline. Intervals are 95% over seeds; missing points are cells with too few harmful rounds. Values are in Table 8.

Summary: After controlling for cosine and round, Δ_t shows no significant residual in six of eight cells (Figure 8a). $\log \alpha_t$ and the min-norm magnitude retain a significant residual in two cells: CIFAR-10 at $\beta = 0.5$ (0.089 and 0.057, respectively) and FMNIST at $\beta = 100$ (0.168 and 0.172, respectively). Adding any measure to a cosine-plus-round classifier changes AUROC by at most 0.018, and every estimable interval contains zero (Figure 8b). Within cosine quintiles, the stratified gap is significant in $6/8$ cells for $\log \alpha_t$, $5/8$ for the min-norm magnitude, and $3/8$ for Δ_t (Table 8).

Our partitions are not designed to produce the structured conflict of Figure 1. The regime in which theory predicts the cone to outperform cosine therefore remains untested in federated training.