ESCAPING SILENT FAILURE: AUDITABLE ADAPTIVE CLIPPING UNDER CLIENT-LEVEL LOCAL DIFFERENTIAL PRIVACY

Anonymous authors
Paper under double-blind review

ABSTRACT

Client-level local differential privacy (LDP) trusts no party: each client noises its own update before transmission. At $(\varepsilon, \delta) = (8, 10^{-5})$ over $T = 100$ rounds without amplification, the released update carries Gaussian noise of norm $\kappa_\ell C\sqrt{d}$ with $\kappa_\ell \approx 14$, so the clipping radius C sets the noise scale outright. We show that recent local-model designs fail silently, and how to build one whose failure modes are auditable. For the released-norm rules we analyse (scalar per-client statistics of noised released norms) the three natural routes fail: a capped statistic saturates and the radius follows a data-independent schedule (the plug-in map freezes it at a constant); an uncapped one is pinned at the largest admissible radius; uncapped debiased squared-norm estimation has prohibitive variance at our calibration. Their closed-form target omits the model dimension and, relative to the error bound, is too large by $d^{1/4}$. We derive three design laws for auditable scale tracking: pay for a privatized pre-noise statistic; set its cap at the tracked update scale, not at a small multiple of the radius; keep cohort size times averaging window above an explicit threshold. A reference algorithm meets all three at radius-independent Rényi-DP cost. Every prediction was fixed before its run and the failed ones are reported; the failure modes occur in real training (saturation on two architectures). The design settles above $C_{\min}$ in every CNN and MLP run at $m \geq 100$. Under one accounting, the released-norm plug-in rule diverges in every cell (6.9–12.4% final accuracy against 19.9–38.5%), and the design leads a budget-matched randomized-response quantile tracker and a DP-FedPUAC port in every cell when paired ($+0.6$ to $+5.5$ points, none significant after Holm correction at five seeds). Because the regime is fully clipped, the radius matters only to an order of magnitude: three statistic-only releases before training shed an over-estimated prior in the tested range, and continued tracking adds only $+0.3$ and $+0.9$ points over freezing the calibrated radius. In-budget sweeps cost 4–14 points; an offline oracle sweep the threat model does not grant is 1–5.3 points ahead. We distill the results into deployment checks and audit signals computable from released messages alone.

1 INTRODUCTION

Federated learning (FL) (McMahan et al., 2017) keeps data on the clients, but their updates leak (Zhu et al., 2019; Geiping et al., 2020). Differential privacy (DP) (Dwork & Roth, 2014; Abadi et al., 2016) is the standard remedy, in two trust models: central DP assumes a curator who sees raw updates and noises their aggregate; local DP (LDP) (Kasiviswanathan et al., 2011; Duchi et al., 2018) assumes nothing, each client noising its own update before transmission. The stronger model has a quantifiable price: central noise averages down as $1/m^2$ over a cohort of m clients, local noise only as $1/m$. Sample-level (per-example-clipped) federated fine-tuning runs at noise multiplier $\sigma = 1$ for $\varepsilon \in [1, 3]$ (Liu et al., 2026b); client-level LDP needs $\kappa_\ell \approx 14$ at the laxer $\varepsilon = 8$. The released update then carries noise of norm $\kappa_\ell C\sqrt{d}$ for model dimension d, and a single parameter, the clipping radius C, sets that scale.

Adaptive clipping sets C online from the observed update scale; the local model has its own designs: client-local threshold schedules (ALDP-FL; Cui & Wu, 2025), randomized-response norm compar-

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.

isons (DP-FedPUAC; Yuan et al., 2026), plaintext norm reports (AdaCT-DPFL; Liu et al., 2026a), and plug-in moment rules that read the scale off the released norms (FedDimDP; Wang et al., 2026), following central-DP ancestors (Andrew et al., 2021; Pichapati et al., 2019; Wei et al., 2025). We identify three structural problems beneath them and exhibit each experimentally.

Problem 1: the scale is estimated from noised releases, and the estimate stops responding to the data. The server sees $\widetilde{\Delta}_k = \bar{\Delta}_k + z_k$ with $\|z_k\| \approx \kappa_\ell C\sqrt{d}$, orders of magnitude above any admissible radius: a bounded statistic of these norms sits at its cap for every client in every round and the radius freezes at a data-independent constant; an unbounded one drives it to the largest admissible value. *Evidence:* at $d = 5.8 \times 10^5$, $m = 100$, $\varepsilon = 8$ saturation occurs in every client-round, the plug-in radius freezes at 8.5104 and training diverges, and a rule of FedDimDP's uncapped form is pinned to $C_{\max}$ within one round (Table 1).

Problem 2: the radius target omits the model dimension. The bias–noise balance behind plug-in rules carries a noise constant κ_ℓ^2/m, while the injected noise has power $d\kappa_\ell^2 C^2/m$: the bound-minimizing radius is smaller than the plug-in target by $(d/(p-1))^{1/(2p)}$, i.e. $d^{1/4}$ at $p = 2$ ($28\times$ for our CNN, $58\times$ for our ResNet). This corrects the bound, not the accuracy optimum (Section 3.1). *Evidence:* fed the same saturated statistic, the two maps freeze at 8.5104 and 0.3081 (Table 1); at the former, fixed-radius training is at chance (Appendix D.8).

Problem 3: the natural repairs are unstable. Debiasing the released norms is free but closed-loop unstable. Paying for a privatized pre-noise statistic and capping it at a small multiple of the radius collapses the radius to $C_{\min}$ under the target of Problem 2. *Evidence:* the debiased arm (arm: one experimental condition) reaches $C_{\max}$ in three rounds, the radius-coupled cap collapses in one (Table 1).

This paper turns the three problems into a design (*auditable*: checkable from released messages alone, Section 7). Our contributions:

1. **Failure of the released-norm routes** of Definition 1 (scalar per-client statistics of noised released norms). Capped statistics saturate (Theorem 2: the plug-in radius freezes, and any rule under (S1)–(S3) follows a data-independent schedule), the uncapped power map is pinned (Corollary 2), the uncapped debiased route carries a $\sqrt{2d}\,\kappa_\ell^2$ error floor (Corollary 3); each occurs in real training, saturation on two architectures and across $245\times$ in d. A lower bound over all rules of the class is open.
2. **A bound-level target and its limits.** A three-source error bound (Theorem 1) restores the dimension factor the plug-in target omits; the plug-in radius diverges because its statistic saturates, not for want of the factor, and for accuracy the bound-level target is 13–$20\times$ conservative at $m = 100$ (Section 3.1).
3. **Three design laws.** Pay for a pre-noise statistic (Proposition 3); set its cap at the estimated update scale, since an $O(1)$ multiple of the radius collapses (Theorem 3), which yields a stable tracker with a quantified latency (Lemma 1); keep the effective sample size mW (cohort size m times averaging window W) above an explicit threshold (Proposition 4).
4. **A private, auditable scale calibration, and its price.** Algorithm 1 (*the tracker*) composes both channels at radius-independent Rényi-DP cost (Theorem 4). It trains where the released-norm plug-in rule diverges (19.9–38.5% against 6.9–12.4%) and leads the RR-quantile tracker and the DP-FedPUAC port in every cell when paired ($+0.6$ to $+5.5$ pp, n.s. at five seeds). As the radius matters only to an order of magnitude (Finding 3), a box floor and three statistic-only releases do the calibrating; continued tracking adds only $+0.3$ and $+0.9$ pp. In-budget sweeps cost 4–14 pp; an offline oracle sweep is 1–5.3 pp ahead (Section 6; checks and audit signals in Section 7).

2 Setting: Federated Training under Client-Level LDP

Protocol. N clients hold datasets $D_1, \ldots, D_N$ and minimize $F(w) = \sum_k p_k F_k(w)$ over $w \in \mathbb{R}^d$, non-IID ($p_k = 1/N$: full participation, equal-weight average). In round t the server broadcasts the model w_t, the radius C_t, and a statistic cap S_t to a cohort $\mathcal{U}_t$ of m clients. Each client runs plain local SGD (no per-step DP operation), forms $\Delta_k = w_{k,E} - w_t$, and releases two messages, the

update channel and the *statistic channel*:

$$\widetilde{\Delta}_k = \underbrace{\Delta_k \min\{1, C_t/\|\Delta_k\|\}}_{=\bar{\Delta}_k} + z_k, \qquad z_k \sim \mathcal{N}(0, \kappa_\ell^2 C_t^2 I_d); \tag{1}$$

$$\tilde{a}_k = \min\{\|\Delta_k\|^p, S_t\} + \zeta_k, \qquad \zeta_k \sim \mathcal{N}(0, \kappa_C^2 S_t^2),$$

with noise multipliers κ_ℓ, κ_C set by the accountant. The server aggregates $\tilde{g}_t = \frac{1}{m}\sum_k \widetilde{\Delta}_k$, takes a server-side adaptive step (Reddi et al., 2021) with a delayed preconditioner (Li et al., 2023), and updates C_{t+1}, S_{t+1} from the released statistics. $\mathcal{F}_t$ is the σ-algebra of messages released before round t and $\mu_t = \mathbb{E}[\Delta]$ the population mean update; C_t and S_t are deterministic functions of released messages, hence $\mathcal{F}_t$-measurable and public.

Threat model and privacy semantics. The adversary is any party observing a client's transcript, white-box (architecture, hyper-parameters, the rule, the public $\{(C_t, S_t)\}$, (κ_ℓ, κ_C), participation); the protected unit is a client's entire local dataset. A client's transcript $R(D) = (M_1(D), \ldots, M_{T_i}(D))$, the adaptively composed per-round releases $(\widetilde{\Delta}, \tilde{a})$, is (ε, δ)-LDP if $\Pr[R(D) \in \mathcal{O}] \le e^\varepsilon \Pr[R(D') \in \mathcal{O}] + \delta$ for all local datasets D, D' and measurable $\mathcal{O}$. We use replacement semantics (Kifer & Machanavajjhala, 2011; Vadhan, 2017) (update-channel sensitivity $2C_t$), claim no subsampling amplification, and compose with Rényi DP (Mironov, 2017; Balle & Wang, 2018). A statistic is *free* when it is post-processing of released quantities and *paid* when it needs a separate release with its own budget share.

The released-norm class and the plug-in rule. The plug-in rules of the local-model line assume $\mathbb{E}\|\Delta\|^p \le M_p$ and bound the per-round error of the clipped-and-noised mean by

$$\mathcal{B}(C) = K_1 C^{2-2p} + K_2 C^{2-p} + K_3 C^2, \qquad K_1 = M_p^2,\ K_2 = M_p,\ K_3 = \kappa_\ell^2/m, \tag{2}$$

a Markov clipping-bias bound (Chen et al., 2020; Koloskova et al., 2023; Zhang et al., 2022), a tail-truncation term, and a noise term whose constant contains no d (Andrew et al., 2021; Cui & Wu, 2025; Yuan et al., 2026). Balancing bias against noise and substituting a per-round estimate of the moment gives the *plug-in rule*

$$A_{p,t} = \tfrac{1}{m}\textstyle\sum_{k\in\mathcal{U}_t} \min\{\|\widetilde{\Delta}_k\|^p, C_{\max}^p\}, \qquad C_{t+1} = \Pi_{[C_{\min}, C_{\max}]}\big[(A_{p,t}^2 m/\kappa_\ell^2)^{1/(2p)}\big], \tag{3}$$

computed from the *noised* releases, the simplest member of the class below.

Definition 1 (Released-norm adaptation rule)**.** *A rule is a* released-norm rule *if* $s_{k,t} = \varphi_t(\|\widetilde{\Delta}_k\|)$, $\bar{s}_t = \frac{1}{m}\sum_{k\in\mathcal{U}_t} s_{k,t}$, *and* $C_{t+1} = U(\bar{s}_t, \bar{s}_{t-1}, \ldots; C_t) \in [C_{\min}, C_{\max}]$ *for fixed statistic maps* φ_t *and a deterministic update* U. *Hypotheses used where stated:* ***(S1)*** *each* φ_t *is non-decreasing and constant above a cap point* $x_{\mathrm{cap}}(C_t)$*, fixed (e.g.* $C_{\max}$*) or radius-proportional (e.g.* γC*);* ***(S2)*** *noise dominance:* $C(\kappa_\ell\sqrt{d/2} - 1) \ge x_{\mathrm{cap}}(C)$ *for every admissible* C*;* ***(S3)*** U *is deterministic. The cap* binds *when a client's statistic sits at its cap value. The radius* freezes *when it stays at a data-independent constant inside the box* $[C_{\min}, C_{\max}]$*, is* pinned *when the projection onto the box is active, and* collapses *when it decays geometrically to* $C_{\min}$*. All three are* silent*: training runs and privacy holds, but the radius carries no data.*

3 Why the Current Designs Fail Silently

3.1 The target omits the dimension

Assumption 1 (Finite p-th moment; sampling dispersion)**.** *(i) There are* $p \ge 2$ *and* $M_p < \infty$ *with* $\mathbb{E}\|\Delta_k\|^p \le M_p$ *for every client and round. (ii) The dispersion of per-client clipped means under cohort sampling is bounded by* σ_ν^2*. (*$p = 2$*, used in all experiments, is admissible;* $p < 2$ *is outside scope.)*

Theorem 1 (Three-source error bound)**.** *Under Assumption 1, for every* $C > 0$*,*

$$\mathbb{E}\|\tilde{g}_t - \mu_t\|^2 \le \underbrace{M_p^2 C^{2-2p}}_{\text{clipping bias}^2} + \underbrace{\tfrac{M_p^{2/p} + \sigma_\nu^2}{m}}_{\text{variance (}C\text{-free)}} + \underbrace{\tfrac{d\kappa_\ell^2}{m} C^2}_{\text{LDP noise}}.$$

Corollary 1 (Bound-optimal radius)**.** *The C-dependent part of Theorem 1 is minimized at* $C^\star_{\mathrm{bnd}} = \big((p-1)M_p^2 m/(d\kappa_\ell^2)\big)^{1/(2p)}$.

The plug-in rule (3) balances bias against the d-free constant κ_ℓ^2/m, Corollary 1 against the true $d\kappa_\ell^2/m$; at $A_{p,t} = M_p$ the radii differ by $(d/(p-1))^{1/(2p)}$, i.e. $d^{1/4}$ at $p = 2$: $28\times$ (CNN), $58\times$ (ResNet). The bound is flat near its minimum ($E(\theta C^\star_{\mathrm{bnd}})/E(C^\star_{\mathrm{bnd}}) = (\theta^{2-2p} + (p-1)\theta^2)/p$ for its C-dependent part E, 2.1 at $\theta = 2$; Proposition 5 in Appendix A), so the moment need not be precise, but the $28\times$ and $58\times$ overestimates inflate it by $\approx 3.8 \times 10^2$ and 1.7×10^3. *Evidence.* Table 1 drives one saturated released-norm statistic ($A_{p,t} \equiv C^p_{\max}$) through the plug-in map (3) and the d-aware map of Corollary 1: the frozen radii 8.5104 and 0.3081 are the two maps at the cap, and their ratio $27.62 = d^{1/4}$ is the dimension factor realized in training; a GroupNorm ResNet-18 ($d = 1.1 \times 10^7$) gives the ratio $57.8 = d^{1/4}$ and five CNN widths ($245\times$ in d) the $d^{-1/4}$ ladder (Fig. 1b). Once the statistic saturates these values follow by arithmetic; the consequence does not: at the plug-in radius training diverges within the first rounds in every utility cell (Table 4) and in the CNN arm itself (Appendix D.8).

Bound versus accuracy. Corollary 1 corrects the bound, not the accuracy optimum: on a fixed-radius sweep $C^\star_{\mathrm{bnd}} \approx 0.015$ scores $8.1 \pm 0.9\%$, below chance, while the plateau peak (13.6%) lies 13–$20\times$ higher. The d-free map fed the oracle moment returns 0.41 (0.27 capped) and scores 13.2 and 13.6, leading $C^\star_{\mathrm{bnd}}$ by 5.1 and 5.6 pp paired: the better *fixed* radius. No closed loop reaches it, the map's output inflating the moment it reads ($68\times$ tracked, $2.2\times$ with an oracle moment): a corrected target must be stable under its own feedback, which is open (Appendix D.9).

3.2 The statistic saturates, is pinned, or is infeasible

Theorem 2 (Saturation)**.** *Let* $\tilde{a}_k = \min\{\|\widetilde{\Delta}_k\|^p, C^p_{\max}\}$ *and* $C_{t+1} = \Pi_{[C_{\min}, C_{\max}]}[(A^2_{p,t} m/\kappa_\ell^2)^{1/(2p)}]$ *with* $A_{p,t} = \frac{1}{m}\sum_k \tilde{a}_k$. *If*

$$\kappa_\ell \sqrt{d/2} \;\geq\; 1 + C_{\max}/C_{\min}, \tag{4}$$

then with probability at least $1 - mTe^{-d/16}$ *every client in every round has* $\tilde{a}_k = C^p_{\max}$, *and* $C_t \equiv \bar{c} := \Pi_{[C_{\min}, C_{\max}]}[C_{\max}(m/\kappa_\ell^2)^{1/(2p)}]$ *for* $1 \leq t \leq T$*: the adaptive rule degenerates to fixed clipping at a data-independent constant.*

Condition (4) is not an edge case: at $d = 5.8 \times 10^5$, $\sqrt{d/2} \approx 539$, so it holds whenever $\kappa_\ell \geq 0.5$ and $C_{\max}/C_{\min} \leq 268$ ($\kappa_\ell \geq 3.71$ for our box), and it contains no m. It is only sufficient: the freeze is self-consistent whenever $\min\{C_0, \bar{c}\}(\kappa_\ell\sqrt{d/2} - 1) \geq C_{\max}$, which covers $d = 5{,}130$, where (4) fails and the rule still froze. On the saturation event any rule under (S1)–(S3) follows a data-independent schedule, and an EMA *looks* adaptive for $O(1/(1-\rho_A))$ rounds, encoding nothing.

Corollary 2 (Boundary pinning for uncapped rules)**.** *Drop the cap in (S1) and let* $s_{k,t} = \|\widetilde{\Delta}_k\|^p$ *with the memoryless update (C_{t+1} depends on $\bar{s}_t$ only)* $C_{t+1} = \Pi_{[C_{\min}, C_{\max}]}[(\bar{s}_t^2 m/\kappa_\ell^2)^{1/(2p)}]$. *With probability at least* $1 - 2mTe^{-d/16}$ *every released norm lies in a data-independent band* $\kappa_\ell C_t \sqrt{d}\,[b_-, b_+]$, *so before projection* $g_- C_t \leq C_{t+1} \leq g_+ C_t$ *with data-free gains* $g_\pm = b_\pm \kappa_\ell \sqrt{d}\,(m/\kappa_\ell^2)^{1/(2p)}$*; whenever* $g_- > 1$ *the radius reaches* $C_{\max}$ *within* $\lceil \log(C_{\max}/C_0)/\log g_- \rceil$ *rounds and stays there (at our calibration* $g_- \approx 6.3 \times 10^3$*: one round).*

Corollary 3 (No free path)**.** *For a released-norm rule with radius-proportional cap* $\gamma^* = x_{\mathrm{cap}}(C)/C$*: (i) if* $\gamma^* < \kappa_\ell\sqrt{d/2} - 1$ *the statistic saturates and the radius follows a data-independent schedule; (ii) if the cap is removed, the power maps are pinned at* $C_{\max}$ *when* $g_- > 1$ *(Corollary 2); (iii) for* $p = 2$ *the uncapped debiased estimator of Proposition 1 has leading-order relative sd* $\sqrt{2d}\,\kappa_\ell^2/\sqrt{mW}$ *on the* C^2 *scale, so relative accuracy* τ *needs* $mW \gtrsim 2d\kappa_\ell^4/\tau^2$ *($\approx 4 \times 10^{10}$ at $\tau = 1$; $W \gtrsim 2\kappa_\ell^2/\tau^2$ on the moment scale at $C = C^\star_{\mathrm{bnd}}$): statistical infeasibility; a cap* $\gamma^* \geq b_+ \kappa_\ell \sqrt{d}$ *coincides with it off an event of probability* $\leq mTe^{-d/16}$*, without an unconditional risk bound. Caps between the thresholds of (i) and (iii) are not covered; a minimax statement over all rules of Definition 1, and free statistics outside it (e.g. cross-client inner products), are open.*

Table 1: Ablation of the statistic source and the map $A \mapsto C$ in training ($m = 100$, MNIST/CNN, $\varepsilon = 8$, $d = 5.8 \times 10^5$; predictions frozen before the runs). Rows 1–3 and 5 are mechanism checks (given the cited event the value follows by arithmetic), rows 4 and 6 statistical predictions. Last row: frozen coefficient $(m/(d\kappa_\ell^2))^{1/4} = 0.0291$; $M_2 = 0.254$ is the oracle uncapped second moment of an instrumented run (the capped moment the tracker sees is $2.2\times$ smaller and predicts 0.0099); observed: mean over rounds 40–100 (five further runs: 0.0084–0.0143, Appendix D.5).

Arm: statistic → map (result)	Prediction	Observed
released norm, capped → plug-in (Thm. 2)	freeze at 8.510	8.5104; all at cap
released norm, capped → d-aware (Thm. 2)	freeze at 0.3081	0.3081; all at cap
released norm, uncapped → plug-in (Cor. 2)	$C_{\max}$ within 1 round	1 round, pinned
debiased norm, free → d-aware (Sec. 4.1)	rel. err. 2.1×10^4	1.7×10^4; $C_{\max}$ in 3 rounds
pre-noise, radius-coupled cap → d-aware, $\times 2$ (Thm. 3)	$C_{\min}$ within 2 rounds	1 round, pinned
pre-noise, decoupled cap → d-aware (Alg. 1, Lem. 1)	$0.0291\sqrt{M_2} = 0.0147$	0.0128, unpinned

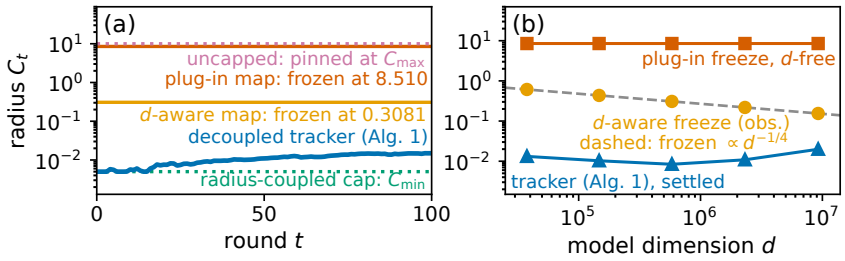


Figure 1: (a) Radius trajectories of five arms of Table 1 (freeze at the predicted constants, pinning at $C_{\max}$, collapse to $C_{\min}$, unpinned tracker). (b) Width-scaled CNN family ($245\times$ in d): saturation at every width makes the plug-in freeze flat in d and the d-aware freeze follow the pre-specified $d^{-1/4}$ ladder; the tracker, whose radius no identity fixes, settles above $C_{\min}$ at every d.

Which published rules fall in scope. FedDimDP (Wang et al., 2026) (record-level per-step noise; median of released norms, no upper bound) takes the median rather than the mean of Definition 1, but every released norm lies in the band of Corollary 2, hence so does their median, and the same envelopes apply with gain $\eta\sigma\sqrt{Ed}$ (Appendix D.7). ALDP-FL, AdaCT-DPFL and DP-FedPUAC leave the class only by relocating the statistic, DP-LAC and PAC-DP need public data the local model does not grant, and central-DP rules see raw norms (Appendix C).

Evidence. Table 1 and Fig. 1a show the three free configurations at $m = 100$ in real training: the capped arms saturate in every client-round ($A_{p,t} \equiv C_{\max}^p$ for all 101 rounds), the uncapped arm is pinned in one round, and the wide-cap debiased route shows a relative error of 1.7×10^4 against 2.1×10^4 predicted and a closed-loop instability (Section 4.1); saturation of both capped maps reproduces on the ResNet, at every width of Fig. 1b and, (4) being cohort-free, in a mechanism-level simulation to $m = 10^6$ (Fig. 3a); the uncapped and debiased arms were run on the CNN only.

4 Three Design Laws for Auditable Adaptation

4.1 Law 1: Pay for the Statistic

Two repairs suggest themselves: subtract the known noise contribution from the released norms (free; cf. Tang et al., 2024), or report a separately privatized pre-noise statistic (paid).

Proposition 1 (Debiased post-processing, $p = 2$)**.** $\hat{a}_k = \|\widetilde{\Delta}_k\|^2 - d\kappa_\ell^2 C_t^2$ *satisfies* $\mathbb{E}[\hat{a}_k \mid \bar{\Delta}_k] = \|\bar{\Delta}_k\|^2$ *and* $\mathrm{Var}(\hat{a}_k \mid \bar{\Delta}_k) = 4\kappa_\ell^2 C_t^2 \|\bar{\Delta}_k\|^2 + 2d\kappa_\ell^4 C_t^4$*; averaged over* m *clients, the relative error on the* C_t^2 *scale is* $\approx \sqrt{2d}\,\kappa_\ell^2/\sqrt{m}$.

Proposition 2 (Privatized pre-noise channel)**.** *For any* $\mathcal{F}_t$*-measurable public cap* $S_t > 0$ *and the releases* $\tilde{a}_k$ *of* (1)*,* $\bar{a}_t = \frac{1}{m}\sum_k \tilde{a}_k$ *is unbiased for the capped moment with* $\mathrm{sd}(\bar{a}_t) \leq$

$S_t\sqrt{1/4+\kappa_C^2}/\sqrt{m} \approx \kappa_C S_t/\sqrt{m}$*; at the operating cap* $S_t = \gamma^p A_t$ *of the tracker below this is a relative error* $\approx \kappa_C\gamma^p/\sqrt{m}$ *on the moment scale.*

Proposition 3 (Crossover law, leading order; $p = 2$)**.** *At* $C = C^\star_{\mathrm{bnd}}$*, on the moment scale, the per-round relative error of the debiased route (Proposition 1) is* $\sqrt{2}\,\kappa_\ell$*, independent of* m *and* d*, and that of the paid channel is* $\kappa_C\gamma^p/\sqrt{m}$*; the paid channel has the smaller error precisely when* $m > (\kappa_C\gamma^p)^2/(2\kappa_\ell^2)$*, with advantage ratio* $\sqrt{2m}\,\kappa_\ell/(\kappa_C\gamma^p)$*:* $m \geq 3$ *and* $6.3\times$ *at* $m = 100$ *at the experimental calibration (*$\kappa_C = \kappa_\ell$*,* $\gamma = 1.5$*,* $p = 2$*).*

The noise comparison is the weaker argument: the free estimand is the clipped norm, C_t^2 in the fully-clipped regime of Section 6, and says nothing about M_p; the paid share (Theorem 4) raises the update-noise power by 2.8% at $\kappa_C = 3\kappa_\ell$ and 25% at $\kappa_C = \kappa_\ell$. Free debiasing is moreover closed-loop unstable: the clamp $\max\{0, \cdot\}$ adds bias $\approx 0.4\sqrt{2d}\,\kappa_\ell^2 C_t^2$, the radius rule amplifies it, and once it dominates the tracked moment the per-round gain exceeds one (Table 1, row 4).

4.2 Law 2: cap at the update scale, not at an $O(1)$ multiple of the radius

Theorem 3 (Collapse of an $O(1)$ radius-proportional cap)**.** *Let* $\|\Delta_k\| \geq L > 0$ *a.s. (a model of concentrated updates at norm scale* L*), let the statistic channel use* $S_t = (\gamma C_t)^p$*, and let the radius follow Corollary 1 driven by the released statistic mean. With loop gain* $g_{\mathrm{cpl}} := \gamma\big((p-1)m/(d\kappa_\ell^2)\big)^{1/(2p)}$*: on any round with* $\gamma C_t < L$ *the cap binds for every client and* $C_{t+1} = \Pi_{[C_{\min}, C_{\max}]}[g_{\mathrm{cpl}} C_t |1 + \bar\zeta_t|^{1/p}]$*,* $\bar\zeta_t \sim \mathcal{N}(0, s^2)$*,* $s := \kappa_C/\sqrt{m}$*. Fix* $a > 0$ *with* $q := g_{\mathrm{cpl}}(1+a)^{1/p} < 1$*. On the event* $\max_{t<T}|\bar\zeta_t| \leq a$*, of probability at least* $1 - 2Te^{-a^2/(2s^2)}$*, every such round gives* $C_{t+1} \leq \max\{C_{\min}, qC_t\}$*; hence if* $\gamma C_0 < L$ *and* $\gamma C_{\min} < L$ *the radius stays in the binding region for all* $t < T$ *and decays geometrically to* $C_{\min}$*, independently of the data.*

Remark (what the collapse says). Write $c_d := ((p-1)m/(d\kappa_\ell^2))^{1/(2p)}$, so $C(A) = c_d A^{1/p}$ and $g_{\mathrm{cpl}} = \gamma c_d$ ($1/c_d \approx 33$ on our CNN at the theory calibration). A cap at an $O(1)$ multiple of the radius binds for every client and collapses ($\gamma = 2$: $g_{\mathrm{cpl}} \approx 0.061$; Appendix A); avoiding it needs $\gamma \gtrsim 1/c_d$, a cap at the scale of the pre-noise update, which under Corollary 1 lies $1/c_d$ above the radius. Algorithm 1's cap $S_t = \gamma^p A_t = (\gamma C_t/c_d)^p$ is exactly such a cap, with radius multiplier $\gamma/c_d \approx 50$ (experimental calibration) and the cost of Proposition 4 at the moment-relative multiplier γ, not at γ/c_d. The failure is one of scale, not of which public variable the cap is written in.

The repair sets the cap at the scale of the statistic itself, carried as its own history. Let $\bar a_t = \frac{1}{m}\sum_k \tilde a_k$ and let $\bar A_t$ be its scale-normalized, unclamped EMA,

$$N_t = \rho_A N_{t-1} + (1-\rho_A)\bar a_t/S_t, \quad D_t = \rho_A D_{t-1} + (1-\rho_A)/S_t, \quad \bar A_t = N_t/D_t. \tag{5}$$

The geometric weights act on the normalized releases $r_j = \bar a_j/S_j$, whose variance $(\mathrm{Var}(a)/S_j^2 + \kappa_C^2)/m$ is nearly round-independent; on the raw $\bar a_j$ this is weight $\propto 1/S_j$, not $1/S_j^2$, which over-trusts censored rounds. With $\Pi_A := \Pi_{[A(C_{\min}), A(C_{\max})]}$, $A(C) := (C^{2p} d\kappa_\ell^2/((p-1)m))^{1/2}$, set $A_{t+1} = \Pi_A[\bar A_t]$, $S_{t+1} = \gamma^p A_{t+1}$, $C_{t+1} = C(A_{t+1})$: the radius is a read-only map *downstream* of the statistic.

Lemma 1 (The self-referencing cap is a stable tracker)**.** *Assume a fixed update distribution with* $\Pr(\|\Delta\| > 0) = 1$ *and* $\gamma > 1$*, and let* $\varphi(A) := \mathbb{E}\min\{\|\Delta\|^p, \gamma^p A\}$*. (i)* φ *is concave, non-decreasing,* $\varphi(0) = 0$*,* $\varphi'(0^+) = \gamma^p > 1$*,* $\varphi \leq M_p$*; hence it has a unique positive fixed point* $A^\star$*, the capped moment at* $S^\star = \gamma^p A^\star$*, which can lie far below* M_p *under heavy tails (Appendix A). (ii) The noiseless recursion converges to* $A^\star$ *from any* $A_0 > 0$ *if* $A^\star \in [A(C_{\min}), A(C_{\max})]$*, else to the nearer box edge (check D6), at a rate set by the history the EMA carries.* Cold start *(*$N_{-1} = D_{-1} = 0$*): the first update returns* $\bar A_0 = \bar a_0$ *whatever* A_0 *is, so in the noiseless recursion an overestimate is shed in one round (to* $\varphi(A_0) \leq M_p$*; with noise see (iii)), and an underestimate expands at transient gains* $\gamma^p, \ldots$ *decreasing toward the asymptotic saturated-regime gain* $g_{\mathrm{sat}} = \gamma^p/(1+\rho_A(\gamma^p - 1)) > 1$*.* Mid-run *(history full): an underestimate expands at* $\approx g_{\mathrm{sat}}$ *and an overestimate contracts at* $\approx \rho_A$ *per round as the stale history decays, so either correction takes* $O(1/(1-\rho_A))$ *rounds. (iii) The released noise enters* N_t *linearly with conditional mean zero given* $\mathcal{F}_t$*, so at a differentiable interior fixed point a small-noise linearization puts the tracker's sd at* $R(\rho_A, b)$ *(*$\to (1-b)^{-1/2}$ *as* $\rho_A \to 1$*) times the frozen-cap benchmark of Proposition 4,* $b = \varphi'(A^\star)$ *(Appendix A; a local approximation,*

Algorithm 1 Auditable tracker (server view; client steps inlined)

Require: T, m, $\kappa_\ell, \kappa_C, \gamma, p$, $[C_{\min}, C_{\max}]$, ρ_A, A_0; server optimizer state

1: $N \leftarrow D \leftarrow 0$; $A_0 \leftarrow \Pi_A[A_0]$, $C_0 \leftarrow C(A_0)$, $S_0 \leftarrow \gamma^p A_0$

2: optional warm-up: K times run lines 4–7, 10, 11 at w_0, releasing only $\tilde{a}_k$ (no model step), updating N, D, A, C, S in place; the K releases are composed

3: **for** $t = 0, \dots, T-1$ **do**

4: broadcast (w_t, C_t, S_t) to cohort $\mathcal{U}_t$

5: **for** each client $k \in \mathcal{U}_t$ in parallel **do**

6: $\Delta_k \leftarrow \textsc{LocalSGD}(w_t)$; $\widetilde{\Delta}_k \leftarrow \Delta_k \min\{1, C_t/\|\Delta_k\|\} + \mathcal{N}(0, \kappa_\ell^2 C_t^2 I_d)$ ▷ update channel

7: $\tilde{a}_k \leftarrow \min\{\|\Delta_k\|^p, S_t\} + \mathcal{N}(0, \kappa_C^2 S_t^2)$ ▷ paid channel, decoupled cap

8: **end for**

9: $\tilde{g}_t \leftarrow \frac{1}{m}\sum_k \widetilde{\Delta}_k$; $w_{t+1} \leftarrow$ server adaptive step on $\tilde{g}_t$ with delayed preconditioner (Li et al., 2023)

10: $\bar{a}_t \leftarrow \frac{1}{m}\sum_k \tilde{a}_k$; $N \leftarrow \rho_A N + (1-\rho_A)\bar{a}_t/S_t$; $D \leftarrow \rho_A D + (1-\rho_A)/S_t$ ▷ scale-normalized EMA (5)

11: $A_{t+1} \leftarrow \Pi_A[N/D]$; $C_{t+1} \leftarrow \big((p-1)A_{t+1}^2 m/(d\kappa_\ell^2)\big)^{1/(2p)}$; $S_{t+1} \leftarrow \gamma^p A_{t+1}$ ▷ Cor. 1; Thm. 3

12: **end for**

not a guarantee); N_t/D_t after Π_A is not unbiased (Appendix D.6), and on the A scale the round-0 noise sd is $\kappa_C \gamma^p A_0/\sqrt{m} \propto A_0$, so an overestimate is shed only up to the sign of one draw (Appendix D.10).

At the defaults $(\rho_A, \gamma) = (0.98, 1.5)$, $g_{\text{sat}} = 1.011$: a $10\times$ cold-start underestimate reaches $0.9A^\star$ after 47–117 updates, a mid-run $25\times$ drop needs 138–180 (Appendix A.8). *Evidence.* The radius-coupled cap of Table 1 collapsed to $C_{\min}$ in one round in both arms and at round 0 on the ResNet. The decoupled tracker spent 16 rounds near $C_{\min}$ after a negative round-0 draw, then climbed to 0.0128 against the oracle-moment prediction 0.0147 (Appendix D.5), and stayed above $C_{\min}$ at every width of Fig. 1b; a from-scratch ResNet sat at the floor (D6, flagged by A2; Appendix D.3). After a mid-run $\times 5$ local-step shift (MLP, $m = 300$, five seeds) the radius recovered $1.54\times$ in 50 rounds (recursion: $1.55\times$); after a $\times 0.2$ shift it remains $1.64\times$ above target: latency is asymmetric (Appendix D.11).

4.3 Law 3: the effective sample size mW must exceed a threshold

Proposition 4 (Statistical feasibility)**.** *To drive the leading-order relative error of Proposition 2 below τ, the pre-noise channel requires effective sample size $mW \geq (\kappa_C \gamma^p/\tau)^2$, where W is the number of averaged releases (EMA window $W \approx (1+\rho_A)/(1-\rho_A)$ asymptotically at equal caps; after n releases $W_n = \frac{1+\rho_A}{1-\rho_A} \cdot \frac{1-\rho_A^n}{1+\rho_A^n}$, 76 of 99 at $n = 101$; or accumulated sparse releases).*

Evidence. In the mechanism simulation the tracker's relative error over $(m, \rho_A, \kappa_C, \gamma)$ collapses onto one curve against $\kappa_C \gamma^p/\sqrt{mW}$ and diverges above 1 (Fig. 3b). In training at $m = 10^4$ the tracker stays unpinned in every seed with a relative fluctuation of 1–2% in four seeds (4.8% in one) against the benchmark's 3.5%; the pre-specified rms error against the *instantaneous* oracle moment failed (0.08–0.22 against 0.07): the moment doubles during training and the EMA lags it (Appendix D.3).

5 The Algorithm and Its Privacy Accounting

Algorithm 1 assembles the three laws; the map $A_{t+1} \mapsto C_{t+1}$ is one replaceable line encoding the target $C_{\text{bnd}}^\star$.

Theorem 4 (End-to-end LDP guarantee)**.** *Conditioned on $\mathcal{F}_t$ the per-round RDP costs are $\varepsilon_{\text{upd}}(\alpha) = 2\alpha/\kappa_\ell^2$ and $\varepsilon_{\text{stat}}(\alpha) = \alpha/(2\kappa_C^2)$, independent of the realized C_t, S_t (Lemma 2, Appendix A). The full transcript of a client i participating in T_i rounds of Algorithm 1 satisfies (ε, δ)-LDP under replacement semantics with $\varepsilon = \Lambda_i + 2\sqrt{\Lambda_i \log(1/\delta)}$, $\Lambda_i := T_i\big(2/\kappa_\ell^2 + 1/(2\kappa_C^2)\big) +$*

$K/(2\kappa_C^2)$ *(K warm-up releases, Algorithm 1); the statistic channel's per-round budget share is* $\kappa_\ell^2/(4\kappa_C^2+\kappa_\ell^2)$.

Proof. Lemma 2 makes the per-round RDP bound uniform over the adaptive C_t; adaptive composition (Mironov, 2017) gives $(\alpha, \Lambda_i\alpha)$-RDP (the K warm-up releases contribute $\alpha/(2\kappa_C^2)$ each), converted at $\alpha^\star = 1+\sqrt{\log(1/\delta)/\Lambda_i}$. □

At $T_i = 100$ (the runs execute 101 rounds), $\delta = 10^{-5}$, $\varepsilon = 8$ the solver gives the two calibrations used here: *theory* ($\kappa_C = 3\kappa_\ell$, $\gamma = 2$, $\kappa_\ell \approx 14.0$, share 2.7%) and *experimental* ($\kappa_C = \kappa_\ell = 15.44$, $\gamma = 1.5$, share 20%).

6 Experiments

Setup. Mechanism experiments (Table 1, Section 6.1) use MNIST with a FedAvg CNN ($d = 582{,}026$; $m = 100$; Dirichlet(0.1); $T = 100$) and a CIFAR-10 replication with a GroupNorm ResNet-18 ($d = 1.1\times10^7$; $T = 30$). The utility campaign (Section 6.2) uses MNIST and FashionMNIST with a 79,510-parameter MLP at $m = 300$, Dirichlet(0.3), twenty paired seeds for the headline arms and five otherwise. Every run uses full participation ($N = m$; no amplification). Multipliers are solved from Theorem 4 at $T = 100$, $\delta = 10^{-5}$ ($\varepsilon = 8$: $\kappa_\ell = \kappa_C = 15.44$ at a 20% share, 13.81 without); every reported guarantee is the accountant's value for the $T{+}1$ rounds run, $(8.045, 10^{-5})$ and $(16.10, 10^{-5})$ (Appendix D.1). Utility is end-of-training accuracy, modest at these cohort sizes (Algorithm 1 20–39%, Table 2; 91.4/84.7% noise-free: the LDP noise sets the level; Appendix D.2).

6.1 Tracking across calibrations

One run per calibration in the Table 1 configuration (*settled radius* := mean C_t over rounds 40–100): $\varepsilon \in \{4, 8, 16\}$ gives $\kappa_\ell = \{29.0, 15.4, 8.5\}$, for which $C^\star_{\mathrm{bnd}} \propto \kappa_\ell^{-1/2}$ predicts $\{0.0107, 0.0147, 0.0198\}$; observed $\{0.0103, 0.0128, 0.0213\}$ (−4%, −13%, +8%; the $\varepsilon = 4$ radius is still climbing 0.95% per round). Across calibrations the ratios match in direction and magnitude, not in the exponent (fitted −0.59; Appendix D.6). Against the trailing oracle capped moment (what an EMA can track) the tracker's offset over rounds 40–100 averages −0.15 to −0.17 (eight and three runs): about 15% on the moment scale, 8% in radius, with within-run fluctuation 0.06–0.40, the order of the frozen-cap benchmark of Law 3; at $m = 10^4$ the offset is −0.08 to +0.08 (Appendix D.6).

6.2 Utility: the price of not sweeping

Under client-level LDP a radius sweep is an oracle outside the threat model or private runs composing against the same budget; we compare with both. *In-model arms*, accounted to the matched ε: (1) Algorithm 1; (2) a budget-matched randomized-response quantile tracker, an LDP analogue of Andrew et al. (2021); (3) the plug-in rule; (4) a port of DP-FedPUAC's clipping component (Yuan et al., 2026); (5) H1, five full arms composed under Theorem 4, best of five; (6) H2, a five-arm 20-round private pilot choosing the radius for one 100-round run. *Oracle arms*, radius chosen outside the budget: (7) fixed $C = 0.01$ (a single-run offline sweep, carried to all cells) and $C = 0.1$ (retrospective grid best); (8) LDP-FedAvg at $C = 0.01$. Table 2 reports paired differences (Appendix D.2).

Finding 1: the in-model sweeps lose to the tracker in every cell. H1 pays $\sqrt{5} \approx 2.2\times$ the noise multiplier and trails Algorithm 1 by 4–14 pp (paired throughout); the pilot H2 trails by 5–9 pp. The margin is the cost of composition; the tracker avoids it without a guessed radius (Appendix C).

Finding 2: against an oracle sweep the threat model does not grant, the deficit is bounded and priced by the paid share. The tracker trails the sweep's selected radius by 1–3.4 pp, non-inferior at the 5 pp margin (Holm-corrected; Holm, 1979), and the retrospective grid best by 1.7–5.3 pp (uncertified at $\varepsilon = 16$). Every arm of Table 2 with $C \le 0.1$ clips ≥99.97% of client-rounds (Appendix D.3), so the fixed arms are the normalized-update baseline of Bu et al. (2023); the gap is consistent with, not identified as, the 20% statistic share. At the share (D5) allows (2.7%) the gap to $C = 0.01$ is $-0.7\,[-4.8, 3.4]$ and $+1.0\,[-0.7, 2.8]$ pp (MNIST, five seeds; Appendix D.10).

Table 2: Absolute accuracy of Algorithm 1 (%, mean±std, twenty seeds; all arms in Table 4) and paired differences (pp, 95% CIs; $m = 300$, MLP). Middle: Algorithm 1 minus the in-budget sweeps (five seeds; twenty for H1 at MNIST $\varepsilon = 8$). Right: oracle fixed radius minus Algorithm 1 (twenty seeds); every $C = 0.01$ cell is non-inferior at the pre-set 5 pp margin after Holm ($p \leq 0.019$).

	Alg. 1	Alg. 1 − in-model sweep		Oracle radius − Alg. 1	
	acc. (%)	H1 (full sweep)	H2 (pilot)	C=0.01 (sweep pick)	C=0.1 (grid best)
MNIST, ε=8	19.9±3.8	+4.3 [3.0, 5.6]	+5.0 [0.7, 9.4]	+1.0 [−0.4, 2.3]	+1.7 [0.5, 2.9]
MNIST, ε=16	30.7±5.0	+12.6 [10.2, 15.1]	+9.2 [0.6, 17.7]	+2.7 [1.4, 4.1]	+5.3 [4.3, 6.3]
FMNIST, ε=8	25.3±5.3	+8.7 [2.5, 15.0]	+5.3 [0.6, 10.1]	+2.3 [0.9, 3.7]	+2.5 [0.6, 4.5]
FMNIST, ε=16	38.5±5.0	+13.7 [11.8, 15.6]	+8.0 [5.5, 10.5]	+3.4 [2.0, 4.9]	+4.8 [2.8, 6.7]

Table 3: Pre-deployment checks (D1–D6) and runtime audit signals (A1–A3).

(D1) Sample	$mW \geq (\kappa_C \gamma^p/\tau)^2$ (Prop. 4; $\tau \approx 0.35$ at $m = 100$; frozen-cap noise benchmark, local closed-loop sd $\times R$, Lem. 1(iii)); else fixed clipping.
(D2) Channel	Pay for the pre-noise channel if $m \geq (\kappa_C \gamma^p)^2/(2\kappa_\ell^2)$ at $C = C^\star_{\mathrm{bnd}}$ (Prop. 3); the free route sees clipped norms only.
(D3) Cap	Cap at the update scale, $S_t = \gamma^p A_t$, i.e. $\gamma/c_d \gg 1$ radii; an $O(1)$ multiple collapses (Thm. 3).
(D4) γ	Smaller γ relaxes (D1) as γ^{2p} but slows recovery (Lem. 1(ii)): the smallest tolerable.
(D5) Budget	Both channels compose at radius-independent cost (Thm. 4); share from the (D1) benchmark: $\kappa_C = \tau\sqrt{mW}/\gamma^p$.
(D6) Box	$A(C_{\min})$ below the smallest moment expected, within $\sim$10$\times$ at $T \approx 100$ (the floor catches a low start); start A_0 low.
(A1) Freeze	C_t constant at machine precision at $\bar{c}$ of Thm. 2 after $O(1/(1-\rho_A))$ rounds: saturation.
(A2) Pinning	At $C_{\max}$ early: an uncapped statistic (Cor. 2); at $C_{\min}$: an $O(1)$ radius-coupled cap (Thm. 3) or a high floor (D6).
(A3) Shadow	Offline: shadow runs at two budgets settle in ratio $(\kappa'_\ell/\kappa_\ell)^{-1/2}$ up to τ (Sec. 6.1: direction and magnitude only).

Finding 3: the radius must be right to an order of magnitude; an over-estimate is corrected by a private start-up calibration, not by online tracking. For a fixed radius only γ'/C matters under server Adam (mechanism check: Appendix D.10). A prior 100$\times$ too small is caught by the box floor of (D6), for the tracker and fixed $C = C_{\min}$ alike (-0.5, -0.8 pp), not by tracking (644–738 updates). ≈100$\times$ too large ($A_0 = 251$: $C_0 = 1.0$ and 1.3), fixed $C = 1$ loses 6–18 pp, as does the tracker started there; three statistic-only releases before round 0 ($\varepsilon : 8.045 \to 8.072$) shed it in all ten runs, $+9.0\,[1.9, 16.0]$ and $+22.0\,[19.1, 24.8]$ pp over fixed $C = 1$, and freezing the radius after those releases scores within 0.3 and 0.9 pp of continued tracking (five seeds; Appendix D.10).

Finding 4: under one accounting, the released-norm rules fail and the tracker leads the remaining adaptive baselines, not significantly. The plug-in rule diverges in every cell (12.4/12.0/7.7/6.9% against 19.9/30.7/25.3/38.5%), and the re-accounted ALDP-FL port stays at chance. Paired over five seeds, Algorithm 1 exceeds the RR-quantile tracker and the DP-FedPUAC port in every cell ($+0.6$ to $+5.5$ pp, none significant after Holm; the twenty-seed means of Table 4 put the port 0.7 pp ahead in one cell); the port never lowers its radius (Appendix D.2).

Limitations. The tracker verifiably tracks $C(A^\star)$, the bound-level radius at the capped moment (Appendix D.6), not the accuracy-optimal radius (13–20$\times$ higher; Appendix D.8). Against a well-guessed radius it trails by 1.8–3.9 pp at $m = 10^4$ and in head-only CIFAR-10 fine-tuning (Appendix D.3). Every reported radius is a window mean: at $T = 300$ neither radius nor target stops moving (Appendix D.11).

7 Deployment Guidance and Conclusion

A signal of Table 3 is *auditable* if it is computable from the transcript $\{(C_t, S_t)\}$ alone under a proved failure mode: on the events of Theorems 2, 3 and Corollary 2 the radius sits at $\bar{c}$ or a box edge, so A1–A2 have no false negatives there; silence certifies nothing more, and no signal detects manipulated reports (Appendix B). *Why adapt, then?* Fixed clipping needs a radius within an order of magnitude of the optimizer's constants, unverifiable from the transcript; the tracker needs public constants and the floor of (D6), and with the warm-up no upper prior in the tested range ($A_0 \leq 251$, $K = 3$). The evidence supports the start-up calibration: freezing the radius after it came within 0.3 and 0.9 pp of continued tracking. We keep online tracking because the moment drifts (it doubled at $m = 10^4$) and a calibrated radius may leave its order of magnitude over longer horizons, which we have not shown. A1–A2 flag those two failure modes, not the accuracy of the target.

REPRODUCIBILITY STATEMENT

All proofs are in Appendix A; hyper-parameters in Appendix D.1; a notation table in Appendix E. The implementation, the time-stamped prediction file, the per-run ledger of $1{,}000$+ runs, and all simulation, figure, and launch scripts are provided as anonymized supplementary material.

AI USE STATEMENT

In this work, we used a generative AI tool (ChatGPT & Claude) for the tasks below with required disclosure.

Mathematical claims and proofs: The assistant helped formulate and verify statements and proofs (identifying errors in earlier drafts and assisting with constant derivation, including the scoping of Proposition 3 and Lemma 1(ii)). All theorems, proofs and constants in the paper were cross-checked by the authors, and closed-form constants were recomputed via scripts provided in the supplementary material. Experimental design, implementation, and result interpretation: The assistant provided feedback on experimental design, assisted with revising experimental code, and aided in interpreting experimental results. The authors reviewed and tested all AI-assisted code and validated all reported numerical values against per-run experimental logs. We did not use generative AI to generate synthetic datasets (the synthetic norms for mechanism simulations are produced by a fixed-seed script), clean or reformat datasets, or conduct qualitative data analysis. No translation work was performed for this paper.

Additionally, we leveraged the tool for tasks under recommended disclosure: revising the manuscript to improve readability and manage length; suggesting paper structure and title; preparing internal mock reviews of the draft. All AI-assisted outputs have been reviewed by the authors. The authors take full responsibility for the final content of this work, including text, claims, code and other artifacts produced with assistance from generative AI.

REFERENCES

Martín Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pp. 308–318, 2016.

Naman Agarwal, Peter Kairouz, and Ziyu Liu. The Skellam mechanism for differentially private federated learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2021.

Galen Andrew, Om Thakkar, H. Brendan McMahan, and Swaroop Ramaswamy. Differentially private learning with adaptive clipping. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2021.

Hilal Asi, Vitaly Feldman, and Kunal Talwar. Optimal algorithms for mean estimation under local differential privacy. In *International Conference on Machine Learning (ICML)*, pp. 1046–1056, 2022.

Borja Balle and Yu-Xiang Wang. Improving the Gaussian mechanism for differential privacy: Analytical calibration and optimal denoising. In *International Conference on Machine Learning (ICML)*, pp. 394–403, 2018.

Abhishek Bhowmick, John Duchi, Julien Freudiger, Gaurav Kapoor, and Ryan Rogers. Protection against reconstruction and its applications in private federated learning. *arXiv preprint arXiv:1812.00984*, 2018.

Zhiqi Bu, Yu-Xiang Wang, Sheng Zha, and George Karypis. Automatic clipping: Differentially private deep learning made easier and stronger. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.

Xiangyi Chen, Steven Z. Wu, and Mingyi Hong. Understanding gradient clipping in private SGD: A geometric perspective. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

Lixin Cui and Xu Wu. ALDP-FL for adaptive local differential privacy in federated learning. *Scientific Reports*, 15:26679, 2025.

John C. Duchi, Michael I. Jordan, and Martin J. Wainwright. Minimax optimal procedures for locally private estimation. *Journal of the American Statistical Association*, 113(521):182–201, 2018.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4):211–407, 2014.

Jonas Geiping, Hartmut Bauermeister, Hannah Dröge, and Michael Moeller. Inverting gradients – how easy is it to break privacy in federated learning? In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 16937–16947, 2020.

Antonious Girgis, Deepesh Data, Suhas Diggavi, Peter Kairouz, and Ananda Theertha Suresh. Shuffled model of differential privacy in federated learning. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 2521–2529, 2021.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

Peter Kairouz, Ziyu Liu, and Thomas Steinke. The distributed discrete Gaussian mechanism for federated learning with secure aggregation. In *International Conference on Machine Learning (ICML)*, pp. 5201–5212, 2021.

Shiva Prasad Kasiviswanathan, Homin K. Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. What can we learn privately? *SIAM Journal on Computing*, 40(3):793–826, 2011.

Daniel Kifer and Ashwin Machanavajjhala. No free lunch in data privacy. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 193–204, 2011.

Anastasia Koloskova, Hadrien Hendrikx, and Sebastian U. Stich. Revisiting gradient clipping: Stochastic bias and tight convergence guarantees. In *International Conference on Machine Learning (ICML)*, pp. 17343–17363, 2023.

Béatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302–1338, 2000.

Tian Li, Manzil Zaheer, Ken Ziyu Liu, Sashank J. Reddi, H. Brendan McMahan, and Virginia Smith. Differentially private adaptive optimization with delayed preconditioners. In *International Conference on Learning Representations (ICLR)*, 2023.

Jianhua Liu, Yanglin Zeng, Zhongmei Wang, Weiqing Zhang, and Yao Tong. Differentially private federated learning with adaptive clipping thresholds. *Future Internet*, 18(3):148, 2026a.

Jin Liu, Ning Xi, Yinbin Miao, and Junkang Liu. DP-FedAdamW: An efficient optimizer for differentially private federated large models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3358–3368, 2026b. arXiv:2602.19945.

Jingcheng Liu and Kunal Talwar. Private selection from private candidates. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing (STOC)*, pp. 298–309, 2019.

Gábor Lugosi and Shahar Mendelson. Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics*, 19(5):1145–1190, 2019.

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

Haaris Mehmood, Jie Xu, Karthikeyan Saravanan, Rogier Van Dalen, and Mete Ozay. DP-LAC: Lightweight adaptive clipping for differentially private federated fine-tuning of language models. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2931–2935, 2026. arXiv:2605.10272.

Ilya Mironov. Rényi differential privacy. In *IEEE Computer Security Foundations Symposium (CSF)*, pp. 263–275, 2017.

Nicolas Papernot and Thomas Steinke. Hyperparameter tuning with Rényi differential privacy. In *International Conference on Learning Representations (ICLR)*, 2022.

Venkatadheeraj Pichapati, Ananda Theertha Suresh, Felix X. Yu, Sashank J. Reddi, and Sanjiv Kumar. AdaCliP: Adaptive clipping for private SGD. *arXiv preprint arXiv:1908.07643*, 2019.

Tiberiu Popoviciu. Sur les équations algébriques ayant toutes leurs racines réelles. *Mathematica (Cluj)*, 9:129–145, 1935.

Sashank J. Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečný, Sanjiv Kumar, and H. Brendan McMahan. Adaptive federated optimization. In *International Conference on Learning Representations (ICLR)*, 2021.

Qiaoyue Tang, Frederick Shpilevskiy, and Mathias Lécuyer. DP-AdamBC: Your DP-Adam is actually DP-SGD (unless you apply bias correction). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 15276–15283, 2024.

Salil Vadhan. The complexity of differential privacy. In *Tutorials on the Foundations of Cryptography*, pp. 347–450. Springer, 2017.

Jin Wang, Hui Ma, Yajun Zhang, Xinjun Pei, Ming Yan, Fei Xing, and Yikun Chen. An adaptive differentially private federated learning framework. In *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2026. arXiv:2602.06838.

Chengkun Wei, Weixian Li, Chen Gong, and Wenzhi Chen. DC-SGD: Differentially private SGD with dynamic clipping through gradient norm distribution estimation. *IEEE Transactions on Information Forensics and Security*, 20:4498–4511, 2025.

Dong Yin, Yudong Chen, Kannan Ramchandran, and Peter L. Bartlett. Byzantine-robust distributed learning: Towards optimal statistical rates. In *International Conference on Machine Learning (ICML)*, pp. 5650–5659, 2018.

Jiangyong Yuan, Yong Chen, Zheyi Wang, ChenXiao Wang, Xiaoyue Hu, and Zihao Zeng. DP-FedPUAC: Federated learning with differential privacy via adaptive gradient clipping and local iteration optimization. *Information Sciences*, 733:122981, 2026. doi: 10.1016/j.ins.2025.122981.

Xinwei Zhang, Xiangyi Chen, Mingyi Hong, Steven Wu, and Jinfeng Yi. Understanding clipping for federated learning: Convergence and client-level differential privacy. In *International Conference on Machine Learning (ICML)*, 2022.

Hao Zhou, Siqi Cai, Hua Dai, Geng Yang, Jing Luo, and Hui Cai. PAC-DP: Personalized adaptive clipping for differentially private federated learning. *arXiv preprint arXiv:2603.24003*, 2026.

Ligeng Zhu, Zhijian Liu, and Song Han. Deep leakage from gradients. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

A Proofs

A.1 Proof of Theorem 1

Write $\tilde{g}_t - \mu_t = \big(\frac{1}{m}\sum_k(\bar{\Delta}_k - \nu_k)\big) + \big(\frac{1}{m}\sum_k \nu_k - \mu_t\big) + \bar{z}_t$ with $\nu_k = \mathbb{E}[\bar{\Delta}_k]$ and $\bar{z}_t = \frac{1}{m}\sum_k z_k$; the three parts are conditionally uncorrelated. *Bias:* $\|\nu_k - \mathbb{E}\Delta_k\| \le \mathbb{E}[\|\Delta\|\mathbf{1}_{\|\Delta\|>C}] \le C^{1-p}\mathbb{E}\|\Delta\|^p \le M_p C^{1-p}$, with no condition on C. *Variance:* per client, $\mathrm{Var}(\bar{\Delta}_k) \le \mathbb{E}\|\bar{\Delta}_k\|^2 \le \mathbb{E}\|\Delta_k\|^2 \le (\mathbb{E}\|\Delta_k\|^p)^{2/p} \le M_p^{2/p}$ by the power-mean inequality ($p \ge 2$); conditional independence across clients divides by m, and the dispersion of the ν_k under sampling contributes σ_ν^2/m. *Noise:* $\mathbb{E}\|\bar{z}_t\|^2 = d\kappa_\ell^2 C^2/m$ by independence. □

On the surrogate's middle term. The term K_2C^{2-p} in (2) bounds $\mathbb{E}\|\Delta - \bar{\Delta}\|^2$. As a variance bound it is invalid for $p > 2$: $\Delta = \pm M_p^{1/p} e_1$ has C-free variance at every $C \geq M_p^{1/p}$. Sharpening the variance bound to $\min\{C^2, M_p^{2/p}\}$ leaves Corollary 1 unchanged to leading order when $d\kappa_\ell^2 \gg 1$.

A.2 Proof of Corollary 1; the cost of a misspecified radius

$\frac{d}{dC}\left(M_p^2 C^{2-2p} + (d\kappa_\ell^2/m)C^2\right) = 0$ gives $(p-1)M_p^2 C^{-2p} = d\kappa_\ell^2/m$. □

Proposition 5 (Cost of a misspecified radius)**.** *Let $E(C) = K_1C^{2-2p} + K_3C^2$ be the C-dependence of Theorem 1. For any $\theta > 0$, $E(\theta C^\star_{\mathrm{bnd}})/E(C^\star_{\mathrm{bnd}}) = \Gamma(\theta, p) = \left(\theta^{2-2p} + (p-1)\theta^2\right)/p$; in particular $\Gamma(2, 2) = 2.125$.*

Proof. At the optimum $(p-1)K_1(C^\star)^{2-2p} = K_3(C^\star)^2$, hence $E(C^\star) = K_3(C^\star)^2 p/(p-1)$; substitute $C = \theta C^\star$ and divide. □

A.3 Proof of Theorem 2

$\|z_k\|^2/(\kappa_\ell^2 C_t^2) \sim \chi_d^2$; the Laurent–Massart lower tail (Laurent & Massart, 2000), $\Pr(\chi_d^2 \leq d - 2\sqrt{dx}) \leq e^{-x}$ with $x = d/16$, gives $\|z_k\| \geq \kappa_\ell C_t\sqrt{d/2}$ with probability $\geq 1 - e^{-d/16}$. On this event, by the triangle inequality and $\|\bar{\Delta}_k\| \leq C_t$, $\|\widetilde{\Delta}_k\| \geq C_t(\kappa_\ell\sqrt{d/2} - 1) \geq C_{\min} \cdot C_{\max}/C_{\min} = C_{\max}$ using (4) and $C_t \geq C_{\min}$; hence $\tilde{a}_k = C_{\max}^p$. A union bound over the mT pairs gives the stated probability, and substituting $A_{p,t} \equiv C_{\max}^p$ into the update yields $\bar{c}$. On the same event every client's statistic under (S1) equals the cap value $\varphi_t(x_{\mathrm{cap}}(C_t))$, which depends only on C_t, so any rule satisfying (S1)–(S3) follows a data-independent schedule. Condition (4) is only sufficient: it bounds C_t below by $C_{\min}$, whereas on the event the radius takes just two values, C_0 and $\bar{c}$, so the same argument goes through whenever $\min\{C_0, \bar{c}\}(\kappa_\ell\sqrt{d/2} - 1) \geq C_{\max}$ (the freeze is then self-consistent). At $d = 5{,}130$, $\kappa_\ell = 13.807$, condition (4) would need $\kappa_\ell \geq 39.5$ and fails, but $\min\{0.85, 8.51\} \times 698 = 594 \geq 10$, and the plug-in rule froze at 8.5104 there as well; the same holds at our smallest width ($d = 37{,}610$: $1893 < 2001$ in (4), but $0.061 \times 1892 = 115 \geq 10$). □

A.4 Proof of Corollary 2

Two-sided Laurent–Massart at $x = d/16$ gives $d/2 \leq \chi_d^2 \leq 13d/8$ with probability $\geq 1 - 2e^{-d/16}$, i.e. $\|z_k\| \in \kappa_\ell C_t\sqrt{d}\,[1/\sqrt{2}, \sqrt{13/8}]$. The triangle inequality with $\|\bar{\Delta}_k\| \leq C_t$ widens each endpoint by $C_t = \kappa_\ell C_t\sqrt{d} \cdot (\kappa_\ell\sqrt{d})^{-1}$, giving the band with endpoints $b_- = 1/\sqrt{2} - 1/(\kappa_\ell\sqrt{d})$ and $b_+ = \sqrt{13/8} + 1/(\kappa_\ell\sqrt{d})$ and gains $g_\pm = b_\pm\kappa_\ell\sqrt{d}\,(m/\kappa_\ell^2)^{1/(2p)}$ (linear in the endpoint by 1-homogeneity of the update in the norm scale); a union bound covers all mT pairs. On this event $\bar{s}_t \in (\kappa_\ell C_t\sqrt{d})^p[b_-^p, b_+^p]$, and substituting into the update, $C_{t+1} \in C_t\kappa_\ell\sqrt{d}\,(m/\kappa_\ell^2)^{1/(2p)}[b_-, b_+]$ before projection: a lower and an upper envelope, $C_{t+1}^\pm = g_\pm C_t^\pm$, that contain neither data nor noise. If $g_- > 1$ the lower envelope grows geometrically until the projection binds at $C_{\max}$, after which both envelopes are pinned there. The data never enters: its entire influence is confined to the $(\kappa_\ell\sqrt{d})^{-1}$ widening of the band, already absorbed in $b_\pm$. Corollary 3(iii) is Proposition 1 averaged over mW releases at a fixed scale ($p = 2$). For $\gamma^* \geq b_+\kappa_\ell\sqrt{d}$ the cap binds only where the upper edge of Corollary 2's band fails (probability $\leq mTe^{-d/16}$), and elsewhere the capped estimator coincides with the uncapped one; coincidence off a small event does not by itself transfer Proposition 1's unconditional moments to the capped estimator, which is why (iii) is stated for the uncapped one. In the intermediate band $\kappa_\ell\sqrt{d/2} - 1 \leq \gamma^* < b_+\kappa_\ell\sqrt{d}$ our concentration bounds determine neither a uniform saturation nor a non-binding regime; finite capping can trade variance for the truncation bias $-\mathbb{E}[(\|\widetilde{\Delta}_k\|^2 - \gamma^{*2}C_t^2)_+]$, and we give no uniform risk bound there. Corollary 3(i) is the argument of Theorem 2 with cap $\gamma^* C_t$: on the same event $\|\widetilde{\Delta}_k\| \geq (\kappa_\ell\sqrt{d/2} - 1)C_t > \gamma^* C_t$ for every client, so $\bar{s}_t = (\gamma^* C_t)^p$ and C_{t+1} is a data-free function of C_t; for the power maps it is geometric, with ratio $0.851\gamma^*$ (plug-in) or $0.031\gamma^*$ (d-aware) at the calibration of Table 1 ($\kappa_\ell = 13.807$), so the radius decays to $C_{\min}$ or grows to $C_{\max}$ (it stays put only at ratio exactly 1). Corollary 3(ii) is Corollary 2. □

A.5 Proofs of Propositions 1–2 and 3

Expand $\|\bar{\Delta} + z\|^2$; the cross term has variance $4\sigma^2\|\bar{\Delta}\|^2$ and $\|z\|^2$ has variance $2d\sigma^4$ (χ^2_d), uncorrelated by the symmetry of z. For the paid channel, unbiasedness is immediate; $\mathrm{Var}(\bar{a}_t) = \frac{1}{m}(\mathrm{Var}(a) + \kappa_C^2 S_t^2)$ with $\mathrm{Var}(a) \le S_t^2/4$ by Popoviciu's inequality (Popoviciu, 1935) on $[0, S_t]$; the capping bias $\mathbb{E}[(\|\Delta\|^p - S_t)_+]$ is the same one-sided, Markov-controlled truncation already priced into the surrogate. Both errors carry $1/\sqrt{m}$ on their own scales. At $C = C^\star_{\mathrm{bnd}}$ and $p = 2$, $C^2 = M_2\sqrt{m}/(\kappa_\ell\sqrt{d})$, so the debiased route's absolute error $\sqrt{2d}\,\kappa_\ell^2 C^2/\sqrt{m}$ equals $\sqrt{2}\,\kappa_\ell M_2$: a relative error $\sqrt{2}\,\kappa_\ell$ on the moment scale, against $\kappa_C\gamma^p/\sqrt{m}$ for the paid channel; comparing the two gives the crossover law. At the theory calibration ($\kappa_C = 3\kappa_\ell$, $\gamma = 2$, $p = 2$) the paid channel wins for $m > 72$, by $1.2\times$ at $m = 100$. □

A.6 Proof of Theorem 3

On the binding event $a_k = S_t$ for every k, so $\frac{1}{m}\sum_k \tilde{a}_k = S_t(1 + \bar{\zeta}_t)$ with $\bar{\zeta}_t \sim \mathcal{N}(0, \kappa_C^2/m)$; substituting $S_t = (\gamma C_t)^p$ into the radius rule gives the stated recursion (a per-round clamp only accelerates the descent; the EMA-driven arm of Table 1 collapsed as well, an observation we do not prove here). On $\{|\bar{\zeta}_t| \le a\}$, $|1 + \bar{\zeta}_t|^{1/p} \le (1 + a)^{1/p}$, so the pre-projection factor is at most $q < 1$ and $C_{t+1} \le \max\{C_{\min}, qC_t\}$; with $\gamma C_{\min} < L$ the next round binds again, and induction over $t < T$ on the intersection of the T events, whose complement has probability at most $2Te^{-a^2/(2s^2)}$ by the Gaussian tail bound and a union bound, gives the geometric decay. A negative expected log-drift alone would not suffice: with $g_{\mathrm{cpl}} = 0.8$, $s = 0.5$, $p = 2$ the drift is negative yet a trajectory started at $\gamma C_0 = 0.99L$ leaves the binding region in one round with probability $\approx 12\%$; the finite-horizon event statement is what the calibration actually delivers (at $g_{\mathrm{cpl}} \approx 0.061$, $s \approx 4.2$ the exit needs a $\ge 60\sigma$ draw, and at $a = 200$ the failure probability is below 10^{-30}). Entry is generic: the rule's own target satisfies $\gamma C^\star_{\mathrm{bnd}} < L$ exactly when $\gamma < L(d\kappa_\ell^2/((p-1)M_p^2 m))^{1/(2p)}$, so the better the estimate, the faster the collapse. The Remark's threshold $\gamma \gtrsim 1/c_d$ ($1/c_d \approx 33$ on the CNN and 68 on the ResNet at the theory calibration) is the negation of the entry condition with L at the norm scale $M_p^{1/p}$ of concentrated updates (a leading-order, not a rigorous, statement); the identity $S_t = \gamma^p A_t = (\gamma C_t/c_d)^p$ is $C(A) = c_d A^{1/p}$ inverted. □

A.7 Proof of Lemma 1

(i) For each fixed x, $A \mapsto \min\{x, \gamma^p A\}$ is the minimum of a linear and a constant function, hence concave and non-decreasing; expectation preserves both. $\varphi(0) = 0$, and $\varphi(A)/A \to \gamma^p$ as $A \downarrow 0$ by monotone convergence under $\Pr(\|\Delta\| > 0) = 1$; boundedness is Assumption 1. A concave function with $\varphi(0) = 0$, initial slope > 1, and a finite upper bound crosses the identity exactly once on $(0, \infty)$. (ii) With geometric weights $c_j \propto \rho_A^{t-j}(1 - \rho_A)$, $\bar{A}_t = \sum_j c_j r_j / \sum_j c_j/S_j$ with $r_j = \bar{a}_j/S_j$. Saturated, $r_j \equiv 1$ and $\bar{A}_t = (\sum_j c_j/S_j)^{-1}\sum_j c_j$, the ρ_A-harmonic mean of the caps. Substituting the growth ansatz $S_j = S_t g^{-(t-j)}$ gives $\bar{A}_t = S_t(1 - \rho_A g)/(1 - \rho_A)$, and the closure $S_{t+1} = \gamma^p \bar{A}_t = gS_t$ yields $g(1 - \rho_A) = \gamma^p(1 - \rho_A g)$, i.e. g_{sat} as stated. The ansatz presumes an infinite geometric history, so g_{sat} is the asymptotic gain; with the finite history of a cold start the harmonic mean runs over fewer, smaller caps and the gain is larger ($2.25, 1.39, 1.23, 1.17, \ldots$ at the defaults), decreasing to g_{sat} as the history fills; at $t = 0$ the ratio reduces to $\bar{A}_0 = \bar{a}_0 = S_0$, whence the γ^p first-round gain. *Convergence.* Without projection, (5) gives $\bar{A}_t = (1 - \lambda_t)\bar{A}_{t-1} + \lambda_t\varphi(A_t)$ with $\lambda_t := ((1 - \rho_A)/S_t)/D_t \in (0, 1]$ ($\lambda_0 = 1$ at a cold start), so with $A_{t+1} = \bar{A}_t$ the noiseless recursion is $A_{t+1} = A_t + \lambda_t(\varphi(A_t) - A_t)$. By (i), $\varphi(A) - A$ has the sign of $A^\star - A$, and φ is non-decreasing with $\varphi(A^\star) = A^\star$, so A_{t+1} lies between A_t and $\varphi(A_t)$, on the same side of $A^\star$: the sequence is monotone and bounded by $A^\star$, hence converges. Since $D_t \le 1/\min_{j \le t} S_j$, $\lambda_t \ge (1 - \rho_A)A_0/A^\star$ on an increasing path and $\lambda_t \ge 1 - \rho_A$ on a decreasing one, so $\varphi(A_t) - A_t \to 0$ and, φ being γ^p-Lipschitz, the limit is the unique positive fixed point $A^\star$. With projection: A_0 is projected into the box (Algorithm 1, line 1), so when $A^\star \in [A(C_{\min}), A(C_{\max})]$ every iterate lies between A_0 and $A^\star$, Π_A never binds and the argument applies unchanged; when $A^\star$ lies outside the box the iterates move monotonically toward it until Π_A binds, and the recursion stays at the nearer box edge. The one-round correction of an overestimate is a noiseless cold-start property: with $N_{-1} = D_{-1} = 0$, $\bar{A}_0 = \bar{a}_0 = \varphi_m(S_0)$, which depends on A_0 only through the cap. With noise, $\bar{a}_0 = \varphi_m(S_0) + (\kappa_C S_0/\sqrt{m})Z$: the normalized release has sd $\kappa_C/\sqrt{m}$ whatever A_0, but on the A

scale the sd is $\kappa_C \gamma^p A_0/\sqrt{m}$ ($\approx$500 at $A_0 = 251$, $m = 300$, $\varepsilon = 8$), so for an inflated A_0 the sign of $\bar{A}_0$ is close to a fair coin; a negative draw is floored by Π_A, a positive one is not (Appendix D.10). Mid-run, the stale history decays only geometrically: new releases enter N_t with weight $(1-\rho_A)$ per round, so an overestimate contracts by $\approx\rho_A$ per round (0.95–0.98 numerically) and an underestimate expands at a gain near g_{sat}; Appendix A.8 tabulates the round counts. (iii) is Proposition 2 plus linearity of the EMA; no per-round clamp is applied, since it would re-introduce the clamping bias of Section 4.1 (for $X \sim \mathcal{N}(\mu, \sigma^2)$, $\mathbb{E}\max\{0, X\} - \mu = \sigma\phi(\mu/\sigma) - \mu\Phi(-\mu/\sigma)$, an upward bias of $\approx 0.94\bar{A}$ per round at the experimental calibration's single-round SNR $\mu/\sigma = \sqrt{m}/(\kappa_C\gamma^p) \approx 0.29$), and deferring all clamping to Π_A keeps the per-round innovation conditionally mean-zero; the ratio N_t/D_t and the projection are not unbiased for $A^\star$, and Appendix D.6 measures the resulting offset. Assume $A^\star$ lies strictly inside the projection interval, $\Pr(\|\Delta\|^p = S^\star) = 0$ (so that φ is differentiable at $A^\star$), and a local small-noise stationary regime in which projection is negligible. Then $S_t \approx S^\star$, $\lambda_t \approx 1 - \rho_A$, and (5) linearizes to $e_{t+1} = [\rho_A + (1-\rho_A)b]e_t + (1-\rho_A)\eta_t$ with $e_t := A_t - A^\star$, $b := \varphi'(A^\star) = \gamma^p \Pr(\|\Delta\|^p > S^\star) < 1$ (concavity), and η_t the release noise of $\bar{a}_t$; the stationary variance $(1-\rho_A)^2\sigma_\eta^2/(1-[\rho_A+(1-\rho_A)b]^2)$ is the frozen-cap value of Proposition 4 times $R(\rho_A, b)^2$, $R(\rho, b) := \sqrt{(1-\rho^2)/(1-[\rho+(1-\rho)b]^2)} \to (1-b)^{-1/2}$ as $\rho \to 1$ (1.220 vs. 1.222 at $b = 0.33$, 3.148 vs. 3.162 at $b = 0.9$, $\rho = 0.98$). Noisy runs of (5) match it ($b = 0.33$ for the lognormal law of Appendix A.8: $\times 1.22$ in sd; a two-point law with $b = 0.9$: $\times 3.2$; script in the artifact). Proposition 4 is therefore the frozen-cap release-noise benchmark, which cap feedback inflates; it does not bound the total closed-loop error. Finally, $A^\star$ is the capped moment, not M_p: for a two-point law with mass 0.99 at 1 and 0.01 at $B > \gamma^p A^\star$, $A^\star = 0.99/(1 - 0.01\gamma^p) \approx 1.01$ at $\gamma^p = 2.25$ while $M_2 = 0.99 + 0.01B$ is unbounded in B, so the tracker converges to $C(A^\star)$, the bound-level radius at the capped moment, and the ratio $C(A^\star)/C^\star_{\mathrm{bnd}}$ is a property of the tail (Appendix D.6 measures it: 0.66–0.95 here). □

A.8 Latency of the deterministic tracker recursion

Iterating (5) with the noiseless release $\bar{a}_t = \varphi(S_t)$ at the defaults $(\rho_A, \gamma, p) = (0.98, 1.5, 2)$ (recursion script in the artifact). *Cold start* ($N_{-1} = D_{-1} = 0$), point-mass norms ($A^\star = 1$), $A_0 = 0.1$: gains $2.25, 1.39, 1.23, 1.17, 1.13, 1.08, 1.06, \ldots$; $\bar{A} \geq 0.9A^\star$ after 47 updates (round index 46) and $\geq 0.99A^\star$ after 155; $A_0 = 0.01$: 232 updates. Lognormal norms ($\log\|\Delta\|^2 \sim \mathcal{N}(0,1)$; $A^\star = 1.28$ against $M_2 = 1.65$, 14.5% of clients capped at $S^\star$): $A_0 = A^\star/10$ needs 117 updates (295 to within 1%), $A^\star/100$ needs 326. A cold-start overestimate ($A_0 = 10^3 A^\star$) returns to M_2 in one update ($A^\star$ for the point mass, $1.29A^\star$ lognormal) and to $1.08A^\star$ in two. *Mid-run* (recursion first run to its fixed point): dividing the moment by 25 (norms $\div 5$), the error contracts by 0.95–0.98 per round and $\bar{A}$ is within 10% of the new fixed point after 138 (point mass) or 180 (lognormal) updates, within 50% after 95–108; multiplying the moment by 25 (the $\times 5$ local-step shift of Sec. 4.2), the radius ratio after 50 updates is 1.54–1.55 and 10% accuracy needs 322–416 updates; a $10\times$ rise needs 241–333. At $\rho_A = 0.9$ the recursion predicts a ratio 4.3 after 50 updates for the $25\times$ rise; the realized 2.89 (Sec. 4.2) reflects a realized norm rise below $5\times$. The shift experiment of Sec. 4.2 multiplies the local step by 5 at round 50 (MLP, $m = 300$, five seeds); its recovery ratio is the round-100 radius over the pre-shift mean of rounds 40–49: per seed $1.46/1.32/1.72/1.60/1.61$ (mean 1.54) at $\rho_A = 0.98$ and $3.42/2.55/3.62/2.70/2.18$ (mean 2.89) at $\rho_A = 0.9$. The naive count $\log 9/\log g_{\mathrm{sat}} \approx 197$ describes neither regime: cold start is faster (transient gains), mid-run slower (memory). The mid-run latency is $O(1/(1-\rho_A))$ rounds in either direction, the window $W \approx (1+\rho_A)/(1-\rho_A) = 99$ of Proposition 4; at $\rho_A = 0.9$ the same $25\times$ shifts settle to 10% in 27–82 updates.

A.9 Clipping-independence of the per-round cost

Lemma 2 (Clipping-independence of the per-round cost)**.** *Conditioned on $\mathcal{F}_t$, C_t and S_t are constants, and for every $\alpha > 1$ the per-round RDP costs $\varepsilon_{\mathrm{upd}}(\alpha) = \alpha(2C_t)^2/(2\kappa_\ell^2 C_t^2) = 2\alpha/\kappa_\ell^2$ and $\varepsilon_{\mathrm{stat}}(\alpha) = \alpha S_t^2/(2\kappa_C^2 S_t^2) = \alpha/(2\kappa_C^2)$ are independent of the realized C_t.*

Proof. C_t and S_t are deterministic functions of previously released privatized statistics (for S_t this is exactly the decoupled construction of Section 4.2), hence $\mathcal{F}_t$-measurable; conditioning makes them constants, and the Gaussian RDP bound $\alpha\Delta^2/(2\sigma^2)$ (Mironov, 2017) cancels the C_t- and S_t-factors. The bound holds for every realization, hence unconditionally. □

B Manipulation of the Statistic Channel

The tracker is driven by an unweighted mean of client reports, and under LDP the server cannot inspect a report's plausibility. This appendix bounds a coalition's effect under a *bounded-report model*, in which manipulated reports stay in $[0, S_t]$. The model is a design assumption, not an enforceable one: an honest report is $\min\{\|\Delta\|^p, S_t\}+\mathcal{N}(0, \kappa_C^2 S_t^2)$, with standard deviation $\kappa_C S_t \approx 15.4\, S_t$ at the experimental calibration, so clamping received reports to $[0, S_t]$ would discard the honest signal (the tracker's EMA is deliberately unclamped, Section 4.2). Independently of the model, Π_A and the radius box bound the worst-case radius at $C_{\max}$, and Lemma 2 is unaffected (accounting never depends on report values).

Proposition 6 (Breakdown point, bounded-report model)**.** *Suppose f of the m clients report the cap S_t in every round while honest reports have saturated at the true capped moment (φ locally constant at $A^\star$). The noiseless tracker fixed point inflates by $A_{\mathrm{adv}}/A^\star = (1 - f/m)/(1 - f\gamma^p/m)$, finite precisely when $f < f^\star := m/\gamma^p$; the radius inflates by the square root at $p = 2$. The report-zero attack shrinks the scale by at least the factor $1 - f/m$ (exactly while φ stays constant) and collapses the tracker to the floor $A(C_{\min})$ once $f > m(1 - \gamma^{-p})$ (56 of 100 at the experimental calibration): under-reporting has a breakdown point too (above that of over-reporting when $\gamma^p > 2$, as here).*

Proof. The round mean is $\bar{a}_t = (1 - f/m)A^\star + (f/m)S_t$ with $S_t = \gamma^p A_t$; imposing the fixed point $A = \bar{a}$ and solving gives the factor, positive iff $f\gamma^p < m$. With f zero reports the fixed point solves $A = (1 - f/m)\varphi(A) \le (1 - f/m)A^\star$, since $\varphi \le A^\star$ on $[0, A^\star]$; φ being concave with $\varphi'(0^+) = \gamma^p$, a positive root exists if $(1 - f/m)\gamma^p > 1$ and none if $(1 - f/m)\gamma^p < 1$; at equality it depends on the law (a point mass gives an interval of fixed points). □

At the experimental calibration $f^\star = 44$ of 100 clients; $f = 5$ inflates the tracked scale by 7% and $f = 30$ by 2.2×, and a noisy simulation of the full recursion tracks the formula within ∼20% up to $f = 30$. Outside the saturated premise the factor is a lower bound: while the cap still truncates honest norms, $\varphi(S_t)$ grows with S_t, so honest reports rise with the adversarial ones (with lognormal norms the simulated inflation exceeds the formula by 7–22% at $f = 5$–20, noisy simulation).

Reports outside the model. At the operating noise level the bound does not hold. A report at $+3$ standard deviations ($3\kappa_C S_t \approx 46\, S_t$) is indistinguishable from an honest one and shifts the round mean by $3\kappa_C S_t/m \approx 0.46\, S_t$, 46× the bounded-model shift S_t/m; through the cap feedback $S_t = \gamma^p A_t$ this already exceeds the fixed point ($1 - 3\kappa_C\gamma^p/m < 0$), so one unclamped client removes the fixed point and the tracked scale grows without bound, though slowly: in the noiseless recursion the radius rises 1.6× over 101 rounds and reaches $C_{\max}$ only after ∼10^4 rounds. The least invasive enforcement, clamping reports at $c\,\kappa_C S_t$, replaces the breakdown point by $f^\star = m/(\gamma^p c\,\kappa_C)$: 2.9 clients at $c = 1$ and 0.96 at $c = 3$, one to three of 100 clients (4.5 at $\gamma = 1.2$), and $c = 1$ attenuates the honest mean itself by about 30% ($\mathrm{erf}(1/\sqrt{2}) = 0.68$). The channel therefore limits manipulation only within the bounded-report model, and no audit signal of Table 3 detects the attack: a coalition below $f^\star$ inflates the radius without a signal, and one above it drives the radius toward $C_{\max}$, but slowly (noiseless recursion from steady state: 2.1× in radius over 100 rounds even if every client reports the cap), and may remain below $C_{\max}$ throughout the evaluated horizon, so (A2) need not fire. Robust aggregation of the noisy reports (trimmed mean, coordinate-wise median, median-of-means (Yin et al., 2018; Lugosi & Mendelson, 2019)) is the natural repair; it forfeits the unbiasedness of Proposition 2, so its bias, its breakdown point under adaptive attacks, and its interaction with the accounting must be re-derived. We list this as open.

C Scope Relative to Other Rules, Mechanisms, and Trust Models

Published local-model rules. ALDP-FL (Cui & Wu, 2025) moves the statistic to client-local pre-noise histories, which makes the released update's sensitivity data-dependent; AdaCT-DPFL (Liu et al., 2026a) to pre-noise norms sent in the clear, a release the accountant does not see; DP-FedPUAC (Yuan et al., 2026) to a separately paid binary channel, the closest prior remedy. DP-LAC (Mehmood et al., 2026) is client-level central DP (the server noises the sum of clipped updates; threshold adapted from a server-side validation loss on public data) and PAC-DP (Zhou et al., 2026) is record-level with a public proxy set. DP-FedPUAC's privacy unit could not be verified from the published version and we make no claim about it.

Mechanisms and trust models. The laws constrain the statistic, not the mechanism. Optimal LDP mean estimators (Bhowmick et al., 2018; Asi et al., 2022) improve the noise constant, not its $\sqrt{d}$ scaling (Duchi et al., 2018); PrivUnit releases a unit vector of data-independent norm, so released-norm rules then carry *no* scale information, and Bhowmick et al. (2018) pay for the magnitude in a separate scalar channel, an instance of Law 1. Secure-aggregation (Kairouz et al., 2021; Agarwal et al., 2021) and shuffle (Girgis et al., 2021) hybrids buy near-central constants by trusting an aggregator or shuffler; the three laws apply unchanged to any hybrid whose statistic channel stays locally privatized, with different constants. Two sweep substitutes are not run in Section 6.2: private selection (Liu & Talwar, 2019; Papernot & Steinke, 2022), which composes a random number of full runs at a constant-factor cost, and automatic clipping (Bu et al., 2023), under which the radius becomes a learning-rate factor; the fully-clipped fixed-radius arms of Finding 2 realize the latter's normalized-update baseline without its per-sample smoothing, which addresses a different privacy unit.

D Experimental Details

D.1 Hyper-parameters

Every run uses full participation (join ratio 1): $N = m$, i.e. 100 clients on the CNN partition, 300 on the MLP partitions, 10^4 in the large-cohort run, so each client's T_i equals the executed round count. Reported guarantees use the solver's unrounded multipliers (15.437, 13.807), which give 8.045 at 101 rounds; the rounded 15.44/13.81 give 8.043. The $T = 30$ ResNet replication was solved at $\kappa_\ell = 7.56$ (8.45 for its two-channel tracker arm) and reports 8.150 over its 31 executed rounds; the composed H1 and H2 values are given in Appendix D.2. All cells share $\gamma = 1.5$, $\rho_A = 0.98$, $A_0 = 1$, $[C_{\min}, C_{\max}] = [0.005, 10]$, server Adam $(\beta_1, \beta_2, \gamma') = (0.9, 0.99, 10^{-3})$ with the delayed preconditioner of Li et al. (2023), server rate 0.01, local SGD step 0.005, local epochs 1 (mechanism tier) or 2 (utility tier), local batch 10. Standard federated BatchNorm is incompatible with the mechanism, hence GroupNorm on the ResNet. The FL framework we build on re-seeds `torch.manual_seed(0)` at every client construction; we re-seed once per repetition after client setup so that noise, selection, and statistic draws vary across repetitions, and flag this for other users.

D.2 Full accuracy table, H1, and H2

Table 4: Final accuracy (%), mean±sample std (20 paired seeds for the headline arms and for H1 at MNIST $\varepsilon = 8$, 5 otherwise; $m = 300$, MLP, matched ε). †: early divergence, model frozen by the NaN guard.

	MNIST		FashionMNIST	
	ε=8	ε=16	ε=8	ε=16
In-model arms:				
Alg. 1 (tracker)	19.9±3.8	30.7±5.0	25.3±5.3	38.5±5.0
RR-quantile	16.3±4.8	30.2±4.3	21.1±5.6	37.4±2.5
DP-FedPUAC-clip (port)	18.0±4.1	28.5±4.8	26.0±4.9	38.0±5.0
Plug-in rule	12.4†	12.0†	7.7†	6.9†
H1 sweep (best of five)	15.6±3.7	18.3±4.1	18.7±2.1	26.7±2.4
H2 pilot sweep (main run)	14.3±3.5	21.8±6.3	22.1±2.1	32.4±1.6
Oracle-radius arms:				
Fixed C=0.01 (sweep pick)	20.8±4.3	33.5±4.6	27.6±4.2	41.9±3.5
Fixed C=0.1 (grid best)	21.6±4.0	36.0±4.5	27.9±4.4	43.3±3.5
LDP-FedAvg (C=0.01)	18.9±4.1	27.4±3.7	31.1±4.7	40.7±3.5
Noise-free controls, identical pipeline (5 seeds):				
No clip, no noise	91.4±0.2		84.7±0.1	
Clip only (C=0.01)	84.5±0.4		76.5±0.1	

LDP-FedAvg. Plain averaging at $C = 0.01$ leads Algorithm 1 in both FashionMNIST cells (+5.8, +2.2 pp in means; paired five-seed difference $-4.6\,[-7.7, -1.5]$ at $\varepsilon = 8$, or $-3.7\,[-6.1, -1.3]$

against seeds 0–4 of the twenty-seed family; pairing as under *Ports* below) and trails in both MNIST cells. We had attributed this to the server preconditioner adding variance; the direct test does not support that reading (the tracker with a plain-averaging server scores $-0.5\,[-3.5, +2.6]$ pp at FashionMNIST $\varepsilon = 8$, Appendix D.10), and we leave the lead of that arm (a fixed radius, averaging, and no statistic share) unexplained.

Statistic share. The tracker at a 2.7% share ($\kappa_C = 3\kappa_\ell$, $\gamma = 1.5$; MNIST $\varepsilon = 16$, three seeds) scored $29.2/37.6/28.6\%$ (mean 31.8) against $27.8/35.6/25.7$ (mean 29.7) for the 20%-share tracker on the same seeds: paired $+2.0$ pp $[0.2, 3.9]$, and $+1.0$ pp against the twenty-seed mean 30.7. A gain of $+0.5$–1.5 pp was pre-specified; the direction held, the size exceeds it, and three seeds certify nothing; the five-seed repetition in two cells is in Appendix D.10.

Ports. Each ported baseline is moved from its original threat model to the unified client-level replacement accounting, favoring the original where ambiguous; we claim nothing about behavior under the original models. The RR-quantile tracker replaces the curator-visible norm comparisons of Andrew et al. (2021) by a randomized-response bit at $\varepsilon_b = 1/\kappa_C$ (matched budget). The DP-FedPUAC port keeps its clipping component and its guard $C = \min(C_0, G_{\text{eval}})$ with $C_0 = C(A_0)$, the shared initial radius (0.063 at $\varepsilon = 8$, 0.085 at $\varepsilon = 16$); G_{eval} never fell below C_0 in any run. Paired five-seed differences, Algorithm 1 minus port (MNIST $\varepsilon = 8/16$, FashionMNIST $\varepsilon = 8/16$): DP-FedPUAC $+1.5\,[-0.9, 3.8]$, $+2.9\,[-0.1, 5.9]$, $+0.6\,[-4.6, 5.7]$, $+1.7\,[-3.3, 6.6]$; RR-quantile $+3.1\,[0.3, 5.9]$, $+1.3\,[0.1, 2.4]$, $+5.5\,[-3.8, 14.7]$, $+2.3\,[-3.9, 8.5]$: two nominally significant, none after Holm correction over the eight comparisons. These differences, like the LDP-FedAvg difference above, pair against the five-seed tracker family of the same campaign (means $19.4/31.4/26.6/39.7$), which shares seeds 0, 1, 2, 4 with the twenty-seed family of Table 4 (whose seeds 0–4 average $19.4/30.9/27.4/40.4$) and differs at seed 3; against that family's seeds 0–4 the differences are $+1.4/+2.4/+1.5/+2.4$ (DP-FedPUAC; MNIST $\varepsilon = 16$: $[0.2, 4.6]$) and $+3.1/+0.8/+6.3/+3.1$ (RR-quantile; MNIST $\varepsilon = 8$: $[0.2, 5.9]$), same signs, again two nominally significant and none after Holm. The released ALDP-FL implementation does not realize client-level replacement accounting, and the re-accounted port does not train under it; it is shown for dynamics only (Fig. 2) and excluded from the utility comparison.

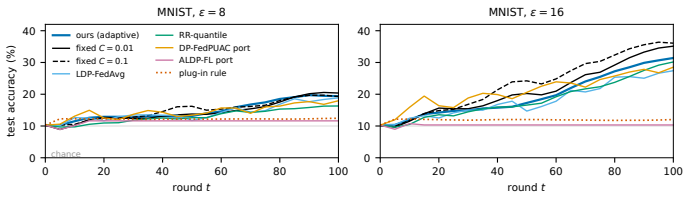


Figure 2: Test accuracy during training (MNIST, $m = 300$; seed means): the plug-in rule freezes, the re-accounted ALDP-FL port stays at chance, the DP-FedPUAC port is held up by its $\min(C_0, G_{\text{eval}})$ guard, the tracker (thick) follows both oracle fixed radii.

H1: the offline sweep inside the budget. Five fixed-clip arms on the oracle sweep grid $\{0.003, 0.01, 0.03, 0.1, 0.3\}$, each a full $T = 100$ run without a statistic channel, are composed under Theorem 4 to the cell's budget: $\kappa_\ell = 30.87$ ($\varepsilon = 8$) and 17.07 ($\varepsilon = 16$) per arm. The arm with the best final accuracy is selected per seed (the same retrospective rule as the oracle sweep) and paired by repetition index with Algorithm 1's seeds. Per-arm means (MNIST $\varepsilon = 8/16$, FashionMNIST $\varepsilon = 8/16$): $C = 0.003$: $14.0/15.2/18.0/25.8$; 0.01: $14.3/16.4/18.0/26.0$; 0.03: $14.0/17.0/17.8/25.6$; 0.1: $13.7/17.2/17.4/25.6$; 0.3: $13.3/16.9/14.3/24.2$. The selected radius varied across seeds in every cell. The pre-specified prediction (selected arm below Algorithm 1 by ≥ 3 pp in all four cells) held: paired differences $+4.3/+12.6/+8.7/+13.7$ pp; composed guarantees 8.046 and 16.100 for 5×101 rounds (multipliers solved for 5×100). The MNIST $\varepsilon = 8$ cell was extended to twenty paired seeds as pre-specified. Against Algorithm 1's twenty-seed means of Table 4 rather than the paired seeds 0–4, the gaps are $4.3/12.4/6.6/11.8$ pp for H1 and $5.5/9.0/3.2/6.1$ pp for H2 (MNIST $\varepsilon = 8/16$, FashionMNIST $\varepsilon = 8/16$).

H2: a private pilot at equal budget. Five pilot arms of $T_p = 20$ rounds on the same grid, the best post-pilot test accuracy selected per seed, then one $T = 100$ run at the selected radius, all composed to the cell's budget over $5 \times 21 + 101 = 206$ round-equivalents: $\kappa_\ell = 19.82$ ($\varepsilon = 8$) / 10.96 ($\varepsilon = 16$), realized 7.999/15.998. A first run calibrated for 200 round-equivalents closed at 8.133/16.286, over the pre-specified 1% tolerance, and was rerun; its pilot picks were identical in all 20 seeds. H2 thus runs 0.6% tighter in ε than the arms it is compared with ($\kappa_\ell = 19.82$ against 19.72 if matched to 8.045, a 1% difference in noise power). Results (five paired seeds): main-run accuracy $14.3/21.8/22.1/32.4$; Algorithm $1 - \text{H2} = +5.0/+9.2/+5.3/+8.0$ pp, every CI excluding zero; the pilot's pick coincided with the full sweep's pick in 5 of 20 seeds (chance 4). H2 selects on the pilot's test accuracy, which reuses the evaluation set: it is an optimistic in-budget baseline, and a deployable protocol would need a validation split whose access is itself accounted. Pilot length is squeezed from both sides: at $T_p = 20$ the pilots sit at 6–24%, half within three points of chance, while κ_ℓ grows as $\sqrt{5(T_p + 1) + 101}$ ($T_p = 40$ costs $1.5\times$ the noise power of $T_p = 20$; $T_p = 80$, $2.5\times$).

Statistical power. At five seeds a one-sided non-inferiority test at a 5 pp margin does not reach significance in any cell (minimal detectable effect ≈8 pp). The twenty-seed paired extension (equal repetition index shares initialization and data order; paired sd ≈3.0 vs unpaired ≈6.5) gives the deficits of Table 2, right: significant in three cells, non-inferior within 5 pp in all four (largest Holm-adjusted $p = 0.019$). The margin was fixed in the analysis plan for the twenty-seed extension, after the five-seed stage and before the extension was run, at 1.5–$1.7\times$ the paired seed-sd.

D.3 Clip fractions and two higher-utility settings

Clip fractions. The ledger logs the update-channel clip fraction each round. Averaged over the last 50 rounds it is 1.000 in every seed of every tracker cell at $m = 300$ (seed-median settled radius 0.019–0.027 per cell, per-seed range 0.0085–0.047; over rounds 80–100, 2–3 seeds per $\varepsilon = 8$ cell still had over half of their statistic reports at the cap, seeds 0 and 9 at 86–94% with radius 0.010–0.011 climbing 0.4–1.0% per round, and 1–2 seeds per $\varepsilon = 16$ cell; the round-0 minimum is 0.987), 1.000 for fixed $C \in \{0.003, 0.01, 0.03\}$, 0.9997–1.000 for $C = 0.1$, and 0.92 for $C = 0.3$; 1.000 for the $m = 100$ CNN tracker of Table 1 and for the head-only CIFAR-10 tracker below. At $m = 10^4$ it is 1.000 over the last 50 rounds (0.988 over all rounds; the pre-settling rounds are unclipped) for the tracker and 0.93 (0.77 over all rounds) for fixed $C = 0.1$. Every $m = 300$ utility arm with $C \leq 0.1$ therefore operates in the fully-clipped regime of Finding 2. *Training at $m = 10^4$.* MNIST, the same MLP, 10^4 clients each holding two classes and four training samples, full participation, $T = 100$, $\varepsilon = 8$ (8.045 logged), the E3 calibrations, five paired seeds: Algorithm 1 $27.5 \pm 3.2\%$; fixed $C = 0.01$ and 0.1 carried over from $m = 300$, 31.4 ± 2.7 and 31.1 ± 3.7 (Algorithm 1 minus the better fixed arm -3.9 pp $[-7.7, -0.1]$, inside the 5 pp margin but uncertified at five seeds); plug-in rule frozen at $C_{\max} = 10$ from round 0, 12.7%; noise-free control 87.2%. The tracker stayed unpinned in every seed (settled radius 0.0135–0.0152, cap-hit fraction 0). The pre-specified tracking criterion failed: the rms relative error of $\bar{A}_t$ against the instantaneous capped oracle moment $\frac{1}{m}\sum_k \min\{\|\Delta_k\|^p, S_t\}$ over rounds 40–100 was $0.222/0.188/0.210/0.078/0.202$ against the frozen $2\tau = 0.070$. Post hoc (not pre-specified) it is almost entirely bias ($-0.22/-0.19/-0.21/-0.06/-0.20$) with fluctuation sd $0.015/0.020/0.020/0.048/0.010$, below $\tau = 0.035$ in four seeds: the oracle moment doubles over training (0.0062 at round 10, 0.0125 at round 100) and the EMA tracks its trailing W-round average, against which the tracker is within -0.08 to $+0.08$ (mean -0.03). The criterion should have been stated against that average; we record the mis-specification rather than repair it. (Two classes and four samples per client is a cross-device scale, chosen so that 10^4 clients fit one dataset.) Per-coordinate noise $\kappa_\ell C/\sqrt{m}$ fell $8.6\times$ from $m = 300$ (mean radii 0.0215 and 0.0143), but the four-sample local updates shrank $3.7\times$, so the signal-to-noise ratio improved only $2.3\times$: cohort scaling buys utility only relative to the per-client signal. *Head-only CIFAR-10 fine-tuning.* A GroupNorm ResNet-18 backbone pretrained centrally on CIFAR-100 is frozen and only the $512 \to 10$ multinomial-logistic head ($d = 5{,}130$) is trained privately ($m = 100$, Dirichlet(0.5), $\varepsilon = 8$, five seeds): Algorithm 1 $16.2 \pm 2.4\%$ (settled radius 0.065), fixed $C = 0.01$ 18.9 ± 4.1, $C = 0.1$ 18.0 ± 3.1, LDP-FedAvg 20.2 ± 4.8, RR-quantile 12.9 ± 2.9, plug-in 9.7 ± 2.9 (frozen at 8.5104 although (S2) fails at this d: the logged statistic equalled its cap in every round); noise-free probe 65.8%; paired differences against the fixed radii -2.7 $[-6.3, 0.9]$ and -1.8 $[-5.0, 1.3]$. Full-model

private fine-tuning at $d = 1.1 \times 10^7$ collapsed to 9.8–10.0% (its tracker stayed unpinned, settled radius 0.017, no round at $C_{\min}$); in from-scratch ResNet training at the same d (where the saturated plug-in and d-aware maps freeze at 11.50, clamped to $C_{\max} = 10$, and 0.1989) the tracked moment fell below the floor $A(C_{\min}) = 0.071$ ($\bar{A} = 0.017$–0.070 at $\kappa_\ell = 8.45$) and the tracker was pinned at $C_{\min}$ for 19 of 31 rounds in the Table 1 replication and for 18–89 of 101 rounds in six $T = 100$ seeds at $\varepsilon \in \{8, 16\}$. In both settings no in-model adaptive arm beats fixed clipping; the tracker's advantage is the certified absence of a sweep and the audit signals, not accuracy.

D.4 Mechanism simulation

Fig. 3 shows the mechanism-only simulation referenced in Sections 3.2 and 4.3 (script in the artifact).

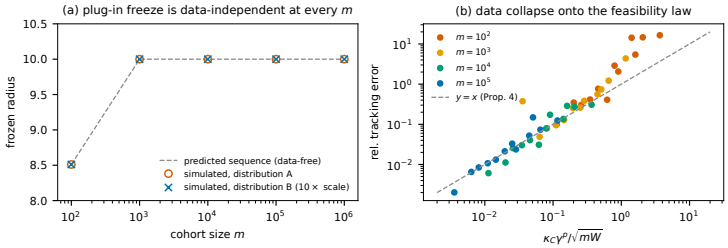


Figure 3: Mechanism-only simulation (synthetic heavy-tailed norms drive the radius recursions). (a) Mechanism check: plug-in freeze vs. cohort size 10^2–10^6 for two norm distributions $10\times$ apart. (b) Tracker relative error versus the frozen-cap noise benchmark $\kappa_C \gamma^p / \sqrt{mW}$ (dashed: $y = x$) across $(m, \rho_A, \kappa_C, \gamma)$; divergence above 1 is the infeasible regime.

D.5 Cold-start trajectories

Fig. 4 shows the cold-start probes at $m = 100$ (MNIST/CNN, the configuration of Table 1; single runs; the tracked $\bar{A}$ is of order 0.2 here). *Uniform EMA, $A_0 = 10$:* the round-0 cap $S_0 = \gamma^p A_0 = 22.5$ inflates the round-0 release noise (sd $\kappa_C S_0/\sqrt{m} \approx 35$ on the moment scale); the draw $\bar{a}_0 = -28.6$ enters the EMA at full weight, decays only as ρ_A^t, and keeps $\bar{A}_t < 0$ and the radius at $C_{\min}$ for all 100 rounds. *Normalized tracker, $A_0 = 170$* ($\approx 10^3\times$ its eventual value): $S_0 = 382$ and the round-0 mean release was $\bar{a}_0 = -84.1$, a -0.14-sd draw at noise sd ≈ 590; because the normalized tracker weights round j by $1/S_j$, the poisoned round carries $S_1/S_0 \approx 1/5800$ of the weight of round 1 and $\bar{A}_1 = +0.17$: shed in one round, as Lemma 1(ii) states for a cold start. *Normalized tracker, A_0 floored at $A(C_{\min}) = 0.029$:* it left the Π_A floor at its first update and reached $\bar{A} \approx 0.2$ by round 5, faster than the noiseless recursion because positive noise excursions assist the escape. *Normalized tracker, $A_0 = 1$* (the arm of Table 1): its round-0 draw was negative ($\bar{a}_0 = -3.8$ at $S_0 = 2.25$, -1.1 sd), so it was floored and, at the single-round SNR ≈ 0.29 of $m = 100$, needed 16 rounds to leave the neighbourhood of $C_{\min}$. All three normalized probes then wander in the band of the frozen-cap benchmark at $\tau \approx 0.35$ (Law 3): settled radii (rounds 40–100) of 0.0128, 0.0115 and 0.0097 for $A_0 = 1$, 170 and floored (0.0145, 0.0118, 0.0084 over rounds 80–100). With the three instrumented seeds of Appendix D.8 (0.0084, 0.0143, 0.0123), six runs of this configuration settle at 0.0084–0.0143, -43% to -3% of the uncapped-moment value 0.0147 and all below it, as censoring of the tracked moment implies (Appendix D.8).

D.6 Tracking error: bias and fluctuation

Across the three calibrations of Section 6.1, the settled radii relative to the $\varepsilon = 8$ value are 0.805 and 1.664 against the predicted 0.729 and 1.346 ($\kappa_\ell^{-1/2}$ scaling; fitted exponent -0.59). Proposition 4 supplies a frozen-cap release-noise benchmark for the *fluctuation* of the tracked scale, not a systematic offset, and the two are separated here (post hoc, existing runs, rounds 40–100). The reference is the quantity an EMA can track: the bias-corrected trailing average, at the tracker's own ρ_A, of the oracle capped moment $\frac{1}{m}\sum_k \min\{\|\Delta_k\|^p, S_t\}$. In the Table 1 configuration ($m = 100$,

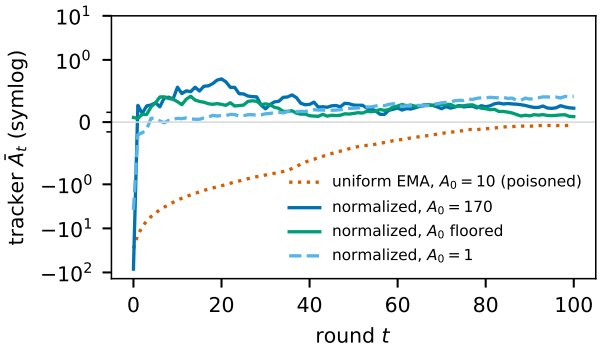


Figure 4: Cold start (tracker $\bar{A}_t$, symlog scale).

law $\tau \approx 0.35$) the three tracker seeds of Appendix D.9 have a mean relative offset of -0.36, -0.19, $+0.04$ (mean -0.17, i.e. -9% in radius at $p = 2$) with within-run standard deviation 0.06, 0.32, 0.22; the eight oracle-target d-aware runs, whose tracker is unchanged, give -0.38 to $+0.09$ (mean -0.15) and 0.07–0.40. The observed fluctuation is of the same order as the frozen-cap benchmark; the downward offset is outside that benchmark's scope. Two sources are visible. Censoring: the capped moment the tracker targets is 0.43–0.91 of the uncapped second moment that Corollary 1 is written for, which accounts for most of the -43% to -3% offset against the closed-form prediction (Table 1). Weighting: the normalized EMA weights round j by $1/S_j$, which favours rounds with a small cap and a small capped mean; we have not isolated its share. In the twenty-seed $m = 300$ arms, 2–3 seeds per $\varepsilon = 8$ cell still have more than half of their reports at the cap over rounds 80–100 (radius 0.010–0.011, climbing 0.4–1.0% per round). At $m = 300$ (MLP, $\varepsilon = 8$ at $T = 300$, $\tau \approx 0.35$) the offset is -0.49 to -0.14; at $m = 10^4$ it is -0.08 to $+0.08$ (Appendix D.3), so it shrinks as the feasibility margin grows. A diagnostic that needs no extra budget: alternating the cap between two values $S_1 < S_2$ across rounds measures $\mathbb{E}\min\{X, S_2\} - \mathbb{E}\min\{X, S_1\} = \int_{S_1}^{S_2} \Pr(X > u)\, du$ for $X = \|\Delta\|^p$, the tail mass between the caps, with the channel's form and sensitivity, hence Theorem 4, unchanged; it says nothing about the tail above S_2, so an extrapolation to M_p needs a tail model and is not a budget-free bias correction. We have not run it.

D.7 FedDimDP's rule in the mechanism simulation

We simulate the rule of FedDimDP (Wang et al., 2026) ($C^{(t+1)} = \max\{\text{median}_i \|\Delta w_i\|, C_{\min}\}$, updates accumulating E locally-noised steps) at the norm level with a leading-order surrogate: the E noise vectors add, so $\|\Delta w_i\| = \eta\sqrt{s_i^2 + E\sigma^2 C^2 \chi_d^2}$ with $s_i = \min(\text{raw}_i, C)E$ (mean-matched; the signal–noise cross term and the dependence of later local gradients on earlier noise are dropped; the simulation uses $E = 1$, where this is the form actually sampled). This is a consistency check of the norm recursion, not an experiment: with $d = 5.8 \times 10^5$ the χ_d^2 factor concentrates to 0.2% and the signal term is negligible, so the realized per-round gains $\{1.5232, 3.8079, 7.6157, 15.2308\}$ necessarily reproduce the analytic $g = \eta\sigma\sqrt{Ed}$. Its content is the dichotomy of Corollary 2 in FedDimDP's own parameters: $\sigma \geq 0.5$ overflows 10^{30} mid-horizon and $\sigma = 0.2$ reaches 1.9×10^{18} by $T = 100$ (no upper bound exists); below the boundary $\sigma^* = 1/(\eta\sqrt{Ed}) \approx 0.131$ the threshold pins at $C_{\min}$ at the predicted round; and two norm distributions $10\times$ apart give the same trajectories throughout, the data-independence the corollary asserts. Training-level interactions are outside the simulation. FedDimDP's mechanism adds Gaussian noise to every local step at record level (DP-SGD inside the client), a privacy unit different from our client-level one; the simulation keeps that per-step noise model (E noised steps at multiplier σ) and claims nothing about FedDimDP's guarantee: pinning follows from noise dominance in the released norm alone.

D.8 The bound-to-accuracy gap

Fixed-radius runs at $m = 100$, $\varepsilon = 8$ (MNIST/CNN; ten radii with three seeds each, plus single runs at 0.005, 1.5, and 5) trace a broad plateau (Fig. 5): $8.1 \pm 0.9\%$ near $C^\star_{\text{bnd}} \approx 0.015$, a high-

variance transition band over $[0.025, 0.08]$, and a peak of 13.6 ± 1.6 at $C \in [0.2, 0.3]$; radii ≥ 1.5 are catastrophic and the plug-in radius 8.5 lies far off the right edge (in the CNN arm of Table 1 the final training loss at that radius is 1.4×10^5 against 7 for the tracker). The d-free map (3) evaluated at the oracle second moment of Table 1 ($M_2 = 0.254$, $\sqrt{M_2} \approx 0.50$, uncapped; the mean over rounds 40–100 of the instrumented seed below, whose two companion seeds gave 0.53 and 0.26) returns 0.41, between the grid points 0.3 (13.6%) and 0.5 ($10.9\pm1.7\%$, three seeds); at the capped moment the tracker sees ($2.2\times$ smaller, below) it returns 0.27, inside the peak band. In accuracy terms the d-free target is thus within a factor 1.4 of the best grid point while $C^\star_{\mathrm{bnd}}$ is 13–$20\times$ too small; the plug-in rule fails in training because saturation freezes it at 8.5 (Theorem 2), not because of the missing factor. The tracker's settled radius lies 3–43% below the uncapped-moment value of $C^\star_{\mathrm{bnd}}$ over six runs of this configuration (Appendix D.5) and within -13% to $+8\%$ of it across the three calibrations of Section 6.1, so the gap is a property of the target, not of the estimator. Two cautions on reading the sweep: every point lies within 4 pp of chance, and with three seeds (std 0.9–1.7 pp at the quoted points) the optimum is located only coarsely, so "13–$20\times$" names a broad plateau rather than a peak. At $d = 8\times10^4$ (MLP, $m = 300$) the gap is small at $\varepsilon = 8$ and grows with the budget: at $\varepsilon = 8$ the offline sweep's basin was flat over $[0.003, 0.03]$ (single run, $17.1/17.7/17.6/17.1/15.9\%$ at $C \in \{0.003, 0.01, 0.03, 0.1, 0.3\}$; the pick $C = 0.01$ was made before the multi-seed campaign and carried to all cells) and the grid best $C = 0.1$ leads the sweep's pick $C = 0.01$ by only 0.3–0.8 pp (Table 4); at $\varepsilon = 16$ it leads by 1.4–2.5 pp and the accuracy optimum sits $\approx5\times$ above the bound optimum 0.021. The gap is unchanged over $\beta_2 \in \{0.9, 0.99, 0.999\}$, a $4\times$ larger local step, and Dirichlet(0.5) heterogeneity: d and the noise level move it, the optimizer and heterogeneity settings do not. Two oracle diagnostics (three instrumented seeds of this configuration, rounds 40–100) decompose part of it: the cap binds in $44\pm26\%$ of client-rounds and the tracked capped moment sits $2.2\pm0.8\times$ below the uncapped second moment (3.0, 2.1, $1.5\times$ per seed; a $\approx1.5\times$ radius underestimate at $p = 2$, seed-dependent: the Table 1 run had the cap binding in 18% of reports), and the round-to-round cosine of the mean clip-residual vector is 0.995 ± 0.001, so under a surrogate in which the clipping bias accumulates coherently the end-of-training optimum scales as $T^{1/(2p)}$ ($3.2\times$ at $T = 100$). Together they account for $\approx4.7\times$ (52–60% in log terms); a static multiplicative correction of the target does not close the remainder, because raising the radius raises the noise, the released update norm, and the tracked moment in turn (a closed-loop amplification any corrected target must model). The landscape of Fig. 5 is moreover open-loop: in a rescaled-map experiment (the map of Corollary 1 multiplied by 15 inside the same tracker; three seeds; final accuracy $10.0\pm2.2\%$; the accountant's value 8.045 in every run, as Lemma 2 requires), the seed whose radius settled in the plateau band ($C = 0.21$) finished at 8.1% and the seed that settled at $C = 1.52$ finished at 12.3%, so fixed-radius accuracy does not transfer to an adaptive arm at the same settled radius. Appendix D.9 runs that comparison directly (fixed radii at each map's oracle-moment value, then the same tracker and an oracle moment driving each map) and reports what it settles and what it leaves open.

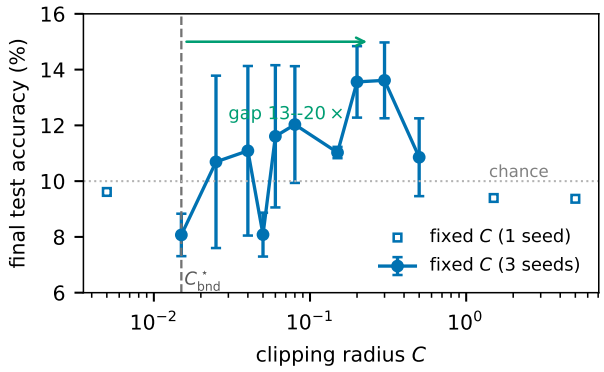


Figure 5: Final accuracy vs. fixed radius ($m = 100$, MNIST/CNN, $\varepsilon = 8$; error bars over 3 seeds, open markers single seed).

D.9 WHICH MAP TO PLUG IN: FIXED RADII, THE TRACKER, AND AN ORACLE MOMENT

Predictions frozen before these runs (ledger block "R1"); MNIST/CNN, $d = 582{,}026$, $m = 100$, $\varepsilon = 8$, seven arms of three seeds, final accuracy, no failures.

Fixed radii at each map's oracle-moment value (no statistic channel, $\kappa_\ell = 13.807$, identical to the sweep arms of Fig. 5). The d-free map returns 0.41 at the uncapped oracle moment and 0.27 at the capped moment the tracker sees; the d-aware map returns 0.0147 and 0.0099. Observed: 13.2 ± 1.6 at $C{=}0.41$, 13.6 ± 1.7 at $C{=}0.27$, 10.5 ± 2.8 at $C{=}0.0099$, against 8.1 ± 0.9 at $C{=}0.015$ from the sweep. Paired over repetition indices, the two d-free radii lead $C^\star_{\mathrm{bnd}}$ by $+5.1\,[0.6, 9.5]$ and $+5.6\,[0.8, 10.3]$ pp (both CIs exclude zero) — the prediction frozen as P-R1a, hit. $C{=}0.27$ reproduces the independently run plateau peak (13.6) exactly. As a *fixed* radius, the d-free target is the better one. (P-R1b, that $C{=}0.0099$ would land within 2 pp of $C{=}0.015$, failed at $+2.4$ pp: the sub-optimal branch is noisier than the frozen window assumed, one seed reaching 13.7.)

The same tracker driving each map ($\kappa_\ell = \kappa_C = 15.437$, only `-rrule` differs). The d-aware arm settles at 0.0116 $[0.0084, 0.0143, 0.0123]$, so the open-loop expectation for the d-free map is $0.0116\,d^{1/4} = 0.322$; the d-free arm instead settles at 1.846 $[1.858, 3.647, 0.033]$, $5.7\times$ higher (P-R1c, hit). The loop is visible in the quantity the tracker reads: the oracle uncapped moment over rounds 40–100 is 0.349 in the d-aware arm and 23.7 in the d-free arm, a $68\times$ inflation (42, 114, $0.6\times$ per seed). The third seed went the other way, collapsing to $C = 0.033$ with the cap binding in every round: under the d-free map the closed loop shows two qualitatively different outcomes across seeds, amplifying or saturating. Accuracy 11.2 ± 1.8 vs 10.4 ± 2.6, paired $+0.85\,[-4.7, +6.4]$ pp (P-R1e, hit: no 3 pp win for adaptation on the d-free map).

An oracle moment driving each map (diagnostic arm, not private: the radius map is fed the logged oracle capped moment while the statistic channel is unchanged; eight seeds, the last five added by a dated amendment after the three-seed outcome $+1.50\,[-0.85, +3.86]$ pp was read, with the analysis frozen before they ran). The inflation falls from $68\times$ to $2.2\times$ (ratio of means; per seed 3.1, 4.7, 3.0, 2.1, 1.3, 2.3, 1.1, 1.0), so almost all of the amplification runs through the *estimate* — a larger radius makes the releases noisier, which raises the tracked moment, which raises the radius — and a small residual through the trajectory itself. The split resolves mostly downward: two seeds amplify (settled 0.53, 0.83) and six collapse to 0.022–0.049 with the cap binding in every round, against 0.0116–0.0180 for the d-aware map. The decisive paired test (P-R1g) stays unresolved at eight seeds: oracle d-free − oracle d-aware $= -0.28\,[-2.82, +2.26]$ pp (9.7 ± 2.2 vs 10.0 ± 2.5); no further extension was run.

What this settles. The d-free target is the better fixed radius, but no closed loop we can build reaches it: the tracked and oracle-moment arms trail the fixed radius at their own target by $-2.4\,[-3.5, -1.4]$ and $-3.3\,[-7.9, +1.2]$ pp (the three repetitions that arm shares with them). Substituting a map does not move the operating point to the better radius, because the map's output re-enters the moment it reads. A corrected target must therefore be stable under its own feedback; naming one is open. For the transcript-adapted (private) arms the accountant reports $\varepsilon = 8.0452$ irrespective of the realized radius (settled radii span $430\times$, 0.0084 to 3.65), as Lemma 2 requires (P-R1d, pass); the oracle-driven arms are non-private diagnostics and carry no guarantee, whatever the accountant prints.

D.10 A WRONG RADIUS PRIOR, THE STATISTIC SHARE, AND THE SERVER OPTIMIZER

Three follow-ups on the utility pipeline of Section 6.2 (MLP, $m = 300$; five repetitions per arm, paired by repetition index with seeds 0–4 of the twenty-seed families; predictions frozen before the runs as ledger block "R2"; 60 runs, none failed; every arm reports the accountant's 8.045 or 16.10).

A wrong prior. Fixed clipping and the tracker are handed a prior $\approx 100\times$ off the sweep's $C = 0.01$ in two cells (MNIST $\varepsilon = 8$ / FashionMNIST $\varepsilon = 16$; default tracker 19.4 / 40.4 on these seeds); the tracker receives the same A_0 in both cells, which the map $C(A)$ turns into $C_0 = 1.00$ at $\varepsilon = 8$ and 1.34 at $\varepsilon = 16$ ($c_d = 0.063$ and 0.085), while fixed clipping runs $C = 1$ in both. *Too small*: fixed $C = 10^{-4}$ scores 10.5 ± 1.9 / 16.2 ± 4.9, the tracker started at $A_0 = 10^{-6}$ (floored by Π_A) 17.7 ± 3.5 / 42.2 ± 2.9: the default tracker leads the fixed arm by $+8.8\,[5.6, 12.1]$ and $+24.2\,[14.3, 34.1]$ pp, and the low-started tracker is within $-1.7\,[-3.4, 0.0]$ and $+1.8\,[-0.7, 4.3]$ pp of the default, its radius back at 0.010–0.026 (frozen prediction, hit in both cells). This is *not* evidence that tracking corrects the prior. $A_0 = 10^{-6}$ maps to $C_0 = 6.3 \times 10^{-5}$, but Π_A lifts it at once to $C_{\min} = 0.005$,

fifty times above the fixed arm, and by the recursion of Appendix A.8 an unfloored $100\times$ radius underestimate needs 644 (point mass) to 738 (lognormal) updates to reach $0.9A^\star$ (232–326 for $10\times$), far beyond $T = 100$. Three controls (ledger block "R3", frozen before the runs; 30 runs) separate the effects. Fixed $C = C_{\min}$ scores 18.1 ± 4.6 / 43.0 ± 2.3: it does not fail, and the floored tracker is indistinguishable from it ($-0.5\,[-2.6, 1.7]$, $-0.8\,[-3.1, 1.5]$ pp); the floor of (D6), which both methods can be given, is what catches the low prior. How low the floor may sit depends on the horizon: by the same recursion a $10\times$ gap in the moment leaves the radius at 94–98% of its target after 100 updates, a $100\times$ gap at 60–69%, so "within $\sim10\times$" in (D6) is a statement about $T \approx 100$. Fixed $C = 10^{-4}$ with the server Adam constant scaled with it, $\gamma' = 10^{-5}$, reproduces fixed $C = 0.01$ to the last digit in all ten runs (19.2 / 44.4; paired difference 0.00). This is a mechanism check in the sense of Table 1, an identity rather than an empirical finding: with every update clipped, $\tilde{g}_t(C) = C\,\tilde{g}_t(1)$ holds path by path at fixed seeds, and the Adam step is unchanged when γ'/C is held fixed. So for a fixed radius the server-Adam dynamics depend on C only through γ'/C (a common rescaling of a whole radius schedule together with γ' and the optimizer state is equivariant in the same way; a change in the *shape* of the schedule enters the optimizer history, as the $A_0 = 251$ arm below shows), and the failure at $C = 10^{-4}$ above is the unscaled constant (every update clipped to 10^{-4}, the aggregate below $\gamma' = 10^{-3}$, preconditioner entries 875–1000 $\approx 1/\gamma'$). Under a plain-averaging server the radius is a step factor and $C = 10^{-4}$ does not train at all (10.3 ± 0.6 / 12.1 ± 4.0). In the fully-clipped regime the radius therefore has to be right to an order of magnitude relative to the optimizer's constants and no more precisely than that, which is also why a loose tracker ($\tau \approx 0.55$, below) loses nothing. *Too large*: fixed $C = 1$ scores 13.4 ± 4.2 / 22.8 ± 3.4 (degraded, not at chance as we had predicted), and the tracker started at $A_0 = 251$ ($C_0 = 1.00$ / 1.34) does *not* rescue it: 15.0 ± 3.1 / 22.2 ± 7.1, i.e. $-4.3\,[-8.7, 0.1]$ and $-18.2\,[-27.0, -9.4]$ pp against the default and indistinguishable from fixed $C = 1$ (frozen prediction, failed). The tracker itself behaves as Lemma 1(ii)–(iii) say: its radius is at the box floor after one update in eight of ten runs (one extra round at 1.78 in the other two) and at 0.010–0.033 over rounds 40–100. But round 0 is run at $C_0 = 1.00$ / 1.34, and that single aggregate (per-coordinate noise sd $\approx$0.9 / 0.66) enters the server's second-moment state, which forgets at $1-\beta_2 = 0.01$ per round: the smallest preconditioner entry is 0.5–2.9 for the whole run against 84–118 in the low-started arm (post hoc). The exposure is the optimizer state, not the radius, which suggests shedding the prior before any update is noised at it.

A statistic-only warm-up (ledger block "R4", frozen before the runs; 30 runs). Before round 0 the clients release only the privatized statistic, K times; the model and the server optimizer are untouched, the multipliers are unchanged, and only the statistic channel's cost is composed, so the reported guarantee is the accountant's value for 101 rounds plus K statistic releases: 8.054 ($K = 1$) and 8.072 ($K = 3$) at nominal $\varepsilon = 8$, 16.12 and 16.16 at 16. With $K = 3$ the tracker started at $A_0 = 251$ scores 22.3 ± 4.5 / 44.8 ± 2.9: $+9.0\,[1.9, 16.0]$ and $+22.0\,[19.1, 24.8]$ pp over fixed $C = 1$, $+7.3\,[3.5, 11.1]$ and $+22.6\,[13.8, 31.4]$ pp over the same tracker without the warm-up, and $+3.0\,[-2.0, 7.9]$ and $+4.4\,[2.4, 6.4]$ pp against the default tracker (the frozen prediction asked for within ±3 pp of the default and is missed from above at FashionMNIST; we do not explain that excess). Every seed enters round 0 at a radius of 0.005–0.012. With $K = 1$ the outcome is a per-seed dichotomy, as predicted: the first release is pure noise at the inflated cap (sd $\approx$500), so a negative draw is floored by Π_A and a positive one is not; four of five seeds in each cell enter round 0 at the floor and recover, the fifth enters at $C = 1.78$ and stays degraded (14.4 / 13.7%), for means of 16.5 ± 1.2 / 36.7 ± 13.2 (the frozen improvement bar is met at FashionMNIST, $+14.5\,[4.3, 24.7]$, and missed at MNIST, $+1.5$). On the default arm one warm-up release is harmless ($-1.5\,[-5.7, 2.7]$, $+1.1\,[-2.8, 5.0]$ pp). Two cautions. The correction from above also passes through the box: at an inflated cap a release is a coin flip between the floor and a radius of order one, and only from the floor, where the cap is small, do the releases become informative; because the EMA numerator carries the sign of earlier draws (the warm-up does not reset N, D), the K releases are not independent trials and we do not calibrate a failure rate from them: with $K = 3$ none of ten runs stayed high and with $K = 1$ two of ten did (0 of 10 bounds the rate by 26% at 95% confidence), so K is a parameter to validate per deployment, at 0.009 in ε per release here. And the comparison concedes the tracker that extra 0.027.

Warm-up, then freeze (ledger block "R5", frozen before the runs; 10 runs). To separate the start-up calibration from continued tracking, a control arm runs the same $K = 3$ warm-up from $A_0 = 251$ and then holds the tracker state for all 101 rounds (radius and cap fixed at their post-warm-up values; the statistic is still released and composed, so the accountant's value is the same 8.072 /

16.16). Its frozen radii are 0.0050–0.0124 (MNIST) and 0.0050–0.0109 (FashionMNIST), and it scores 22.0±5.0 / 43.9±3.6 against the tracked arm's 22.3±4.5 / 44.8±2.9: tracked minus frozen $+0.3\,[-0.4, 1.1]$ and $+0.9\,[-0.5, 2.4]$ pp (frozen prediction of $|\cdot| \leq 3$ pp in both cells, hit). The secondary prediction that the frozen arm would sit within 3 pp of fixed $C = C_{\min}$ held at FashionMNIST ($+0.8\,[-2.9, 4.6]$) and failed at MNIST ($+3.9\,[-4.0, 11.9]$, mean outside the window, CI covering zero). Post hoc, the frozen arm is $+2.7\,[-3.0, 8.3]$ and $+3.5\,[0.9, 6.0]$ pp above the default tracker: a radius fixed by three noisy releases scored no worse on average than the tracked one (one CI covers zero), consistent with the order-of-magnitude reading above. The two directions are therefore not symmetric, and neither is corrected by tracking: an under-estimate is not corrected within the horizon and needs the floor, which fixed clipping can be given too; an over-estimate is shed by the statistic channel in one to three releases before any update carries the wrong radius, which fixed clipping cannot do, and what corrects it is that start-up calibration, not the tracking that follows.

Statistic share. At $\kappa_C = 3\kappa_\ell$ (share 2.7%; (D1) gives $\tau \approx 0.55$ here against 0.20 at the 20% share) the tracker gains $+0.5\,[-1.0, 2.0]$ and $+1.8\,[1.1, 2.6]$ pp over the 20% share (MNIST $\varepsilon = 8$ / 16) and the oracle gap to fixed $C = 0.01$ becomes $-0.7\,[-4.8, 3.4]$ and $+1.0\,[-0.7, 2.8]$ pp (from -0.2 and $+2.9$ on the same seeds); both frozen predictions hit. Settled radii spread more widely (0.0066–0.030), without an accuracy cost in this fully-clipped regime: the share should be the smallest (D1) allows, which is check (D5).

Server optimizer. With a plain-averaging server step the tracker scores 27.0±6.7 against 27.4±5.1 (FashionMNIST $\varepsilon = 8$; $-0.5\,[-3.5, 2.6]$ pp) and 18.0±4.4 against 19.4±2.8 (MNIST $\varepsilon = 8$; $-1.3\,[-4.2, 1.5]$), at the same radii (0.010–0.030). The prediction that averaging would lift the tracker by 3 pp at FashionMNIST failed, which withdraws our explanation of LDP-FedAvg's lead there (Appendix D.2); what survives is that in the two cells tested a plain-averaging server moved the tracker's accuracy by -0.5 and -1.3 pp, neither significant, at the same radii.

D.11 A DOWNWARD SCALE SHIFT, AND WHAT "SETTLED" MEANS AT $T = 300$

Same configuration as the upward shift of Section 4.2 (MLP, $m = 300$, five seeds), with the local step multiplied by 0.2 at round 50. The frozen predictions (P-R1h, P-R1i) assumed the update moment would fall by the full $(0.2)^{-2} = 25\times$; it fell by $4.3\times$ (measured M_2 ratio 0.230±0.010), so the realised target ratio is $\sqrt{0.230} = 0.480$, not 0.200, and both predictions failed as frozen — the same mis-specification (frozen against an assumed rather than a realised moment) recorded for the $m = 10^4$ tracking prediction. Against the realised target, $\rho_A = 0.98$ reaches 0.787±0.164, i.e. $1.64\times$ its target, and $\rho_A = 0.9$ reaches 0.534±0.148, i.e. $1.11\times$. The asymmetry is the finding (P-R1j, hit): the same $\rho_A = 0.98$ tracker matched its *upward* prediction to within 1% but remains 64% above its realised *downward* target after the same 50 rounds. Utility separates no better than under the upward shift: 18.3±2.2 (ρ_A=0.98), 18.3±1.7 (ρ_A=0.9), 18.0±4.1 (fixed 0.01), 16.1±4.4 (fixed 0.002); paired, $+0.31\,[-2.7, +3.3]$ pp against fixed 0.01.

Three $T = 300$ arms (multipliers re-solved for $T = 300$, $\kappa_\ell = 26.737$, reported $\varepsilon = 8.015$ over 301 rounds, so their accuracy is not comparable to the $T = 100$ arms) test whether the radius ever stops moving. It does not: over the last 50 rounds $|\Delta C|/C$ averages 2.0–2.4% per round with no shift, 2.7–2.8% after a downward shift at round 100, and 1.3–1.6% after an upward one (P-R1k, failed). Neither does the quantity it tracks: the oracle capped moment drifts $+5$ to $+9$% over the last 25 rounds of the no-shift arm and $+12$ to $+25$% in the up-shift arm. The reviewer-visible symptom — a radius still rising at round 100 — is therefore not tracker latency but a non-stationary target, and we report an operating radius (a window mean) rather than a limit throughout. The downward-shifted arm ends at 0.917 of its pre-shift radius against a predicted 0.20 (P-R1l, failed), because that upward drift of the moment undoes most of the shift over the following 200 rounds; the upward-shifted arm reaches 3.56 against a predicted 3.7–4.3.

E NOTATION

Table 5 lists the symbols by the section that introduces them; letters that carry more than one meaning are listed together so that each meaning is visible.

Table 5: Notation.

Protocol and threat model (Section 2)	
$N, m, \mathcal{U}_t$	number of clients; cohort size; cohort of round t (every experiment uses full participation, $m = N$)
T, T_i, T_p	rounds; rounds in which client i participates; pilot length of H2
d, w_t, F, F_k, p_k	model dimension; global model; global and client objectives; client weights
$\Delta_k, \bar{\Delta}_k, \tilde{\Delta}_k, z_k$	raw, clipped, and released update of client k; update-channel noise
$C_t, \quad [C_{\min}, C_{\max}], \quad \Pi_{[C_{\min}, C_{\max}]}$	clipping radius (public); its box; projection onto the box
$S_t, \tilde{a}_k, \zeta_k, \bar{a}_t$	statistic cap; released statistic of client k; its noise; cohort mean of the $\tilde{a}_k$
κ_ℓ, κ_C	noise multipliers of the update and statistic channels
$\tilde{g}_t, \mu_t, \nu_k$	noised aggregate; population mean update; $\mathbb{E}[\bar{\Delta}_k]$
$\mathcal{F}_t$	σ-algebra of messages released before round t
p, M_p, σ_ν^2	moment order; p-th moment bound; sampling dispersion of the ν_k (Assumption 1)
$\varepsilon, \delta, \mathcal{O}$	privacy budget; measurable output set in the LDP definition
Released-norm class and bounds (Sections 2–3)	
$\mathcal{B}(C), K_1, K_2, K_3$	plug-in surrogate bound (2) and its constants
$A_{p,t}$	capped mean of released p-th-power norms in the plug-in rule (3)
$s_{k,t}, \bar{s}_t, \varphi_t, U, x_{\text{cap}}$	per-client statistic, its mean, statistic map, update map, cap point (Definition 1)
$C^\star_{\text{bnd}}, E(C), \Gamma(\theta, p)$	bound-optimal radius; C-dependent part of the bound; cost of the misspecification $C = \theta C^\star_{\text{bnd}}$ (Proposition 5)
$\bar{c}$	frozen radius of the saturated plug-in rule (Theorem 2)
$b_\pm, g_\pm$	band endpoints and recursion gains of the uncapped rule (Corollary 2)
γ^*	cap-to-radius ratio $x_{\text{cap}}(C)/C$ of a radius-proportional cap (Corollary 3)
c_d	scale coefficient $((p-1)m/(d\kappa_\ell^2))^{1/(2p)}$ of the map $C(A) = c_d A^{1/p}$ (Remark after Theorem 3; distinct from the Rényi order α)
$\hat{a}_k$	debiased released norm (Proposition 1)
τ, W, mW	target relative error; averaging window; effective sample size (Proposition 4)
Tracker (Sections 4–5)	
γ	cap multiplier, $S_t = \gamma^p A_t$ (distinct from Adam's constant γ' below)
$\rho_A, N_t, D_t, \bar{A}_t$	EMA factor; normalized numerator and denominator; scale-normalized EMA (5)
$A_t, A_0, A(C), C(A), \Pi_A$	tracked scale; its initial value; the scale-to-radius maps; projection onto $[A(C_{\min}), A(C_{\max})]$
$\varphi(A), A^\star, S^\star$	capped-moment map, its fixed point, and the fixed-point cap (Lemma 1)
$g_{\text{sat}}, g_{\text{cpl}}$	saturated-recursion gain (Lemma 1); loop gain of the radius-coupled cap (Theorem 3)
$L, \bar{\zeta}_t, s, a, q$	a.s. lower bound on $\|\Delta_k\|$; normalized statistic-noise mean, $\mathcal{N}(0, s^2)$ with $s = \kappa_C/\sqrt{m}$; deviation level and contraction factor $q = g_{\text{cpl}}(1+a)^{1/p}$ (Theorem 3)
K, W_n	number of statistic-only warm-up releases (Algorithm 1, Appendix D.10); finite-history EMA window (Proposition 4)
$f, f^\star$	coalition size; breakdown point m/γ^p of the bounded-report model (Appendix B)
Accounting (Section 5)	
$\alpha, \alpha^\star$	Rényi order; optimal order in the conversion
$\varepsilon_{\text{upd}}(\alpha), \varepsilon_{\text{stat}}(\alpha)$	per-round RDP costs of the two channels (Lemma 2)
Λ_i	client i's accumulated RDP slope $T_i(2/\kappa_\ell^2 + 1/(2\kappa_C^2)) + K/(2\kappa_C^2)$ (Theorem 4)
ε_b	per-bit budget of the RR-quantile port (Appendix D.2)
Other	
$\beta_1, \beta_2, \gamma'$	server Adam momenta and numerical constant (Appendix D.1)
η, σ, E	FedDimDP's step size, noise multiplier, and local step count (Section 3, Appendix D.7); σ also denotes the sample-level noise multiplier quoted in Section 1
distributions A, B	the two synthetic norm distributions of the mechanism simulation (legend of Fig. 3a), unrelated to A_t
H1, H2	in-budget full sweep and private pilot sweep (Section 6.2)