# CLIPPED OPTIMIZATION OVER HIERARCHICAL NETWORKS UNDER GENERALIZED SMOOTHNESS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Hierarchical learning systems combine federated aggregation within groups of clients with decentralized communication across hubs. We study optimization in this setting under $(H_0, H_1)$-smoothness, where the local smoothness level may grow with function suboptimality. This model is particularly convenient for hierarchical objectives because averaging preserves the curvature-growth parameter $H_1$, with heterogeneity entering through an additive interpolation defect. The main difficulty is locality, since the corresponding descent inequality is valid only for sufficiently small displacements. We therefore use clipping at both levels of the hierarchy: clients clip gradients before aggregation, while hubs clip the optimization step and leave the gradient tracker unchanged. We prove nonconvex convergence guarantees for the resulting federated, decentralized, and hierarchical methods, with accelerated dependence on the network spectral gap. For convex objectives, the same mechanism leads to two phases: linear convergence while clipping is active, followed by an $O(1/K)$ regime. Under strong convexity, both phases are linear.

## 1 INTRODUCTION

Collaborative learning is often organized in two levels. Devices report to edge servers or institutional silos (Liu et al., 2020; Abad et al., 2020). The silos communicate over a peer-to-peer backbone without a central coordinator (Castiglia et al., 2021; Sun et al., 2022). Links from a device to its hub are cheap, and links between hubs are expensive. We study the problem

$$\min_{x\in\mathbb{R}^d} f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x), \qquad f_i(x) = \frac{1}{m}\sum_{j\in\mathcal{C}_i} f_{ij}(x), \qquad f_{ij}(x) = \mathbb{E}_{\xi\sim\mathcal{D}_{ij}}[F_{ij}(x;\xi)], \tag{1}$$

over such a network. Hub $i$ serves $m$ clients $\mathcal{C}_i$ with local data $\mathcal{D}_{ij}$, and the $n$ hubs exchange messages by gossip. Figure 1 compares this structure with the flat federated and decentralized settings. Every client message is clipped *before* aggregation. This order bounds the sensitivity of each client for differential privacy (Abadi et al., 2016), and a hub can enforce it without trusting its clients. It is also harder to analyze, because an average of clipped vectors differs from the clipped average (Shulgin et al., 2025).

Relaxed smoothness makes these objectives realistic. For LSTMs (Zhang et al., 2020b) and transformers (Crawshaw et al., 2022) the local smoothness constant varies by orders of magnitude and grows with the gradient norm. The $(L_0, L_1)$ model captures this, and clipping provably helps in it (Zhang et al., 2020a; Koloskova et al., 2023; Gorbunov et al., 2025). An average of $(L_0, L_1)$-smooth functions need not be $(L_0, L_1)$-smooth with the same constants (Luo et al., 2026, Prop. 3). Distributed analyses therefore add dissimilarity assumptions. These inflate the smoothness constants of $f$. We work in the function-gap class $(H_0, H_1)$ (Alimisis et al., 2026; Liu et al., 2025b; Lobanov, 2026). There the curvature is bounded by the suboptimality gap, and gaps average linearly. The class has a price: its descent inequality holds only for displacements up to $1/\sqrt{H_1}$. Every displacement of the method, including the consensus shift from gossip, must stay within this radius. Bounded updates certify this, and clipping bounds the updates.

**Contributions.**



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.

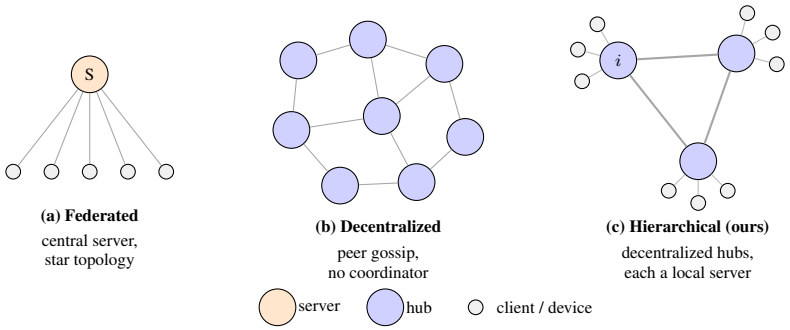


Figure 1: Three collaborative-learning topologies. **(a)** Federated learning: clients report to a single server (a star). **(b)** Decentralized learning: peers exchange messages over a sparse graph with no coordinator. **(c)** The two-level hierarchy studied here. Hubs form a decentralized backbone *and* each hub is a federated server for its own client fleet (drawn around it).

1. *A smoothness class for networks.* Averages of $(H_0, H_1)$-smooth functions are $(H_0 + H_1\delta, H_1)$-smooth, so heterogeneity enters only through the interpolation defect $\delta$ (Proposition 2). The descent inequality of the class is local. Clipping keeps every client and network displacement inside the locality radius, and unclipped constant-step tracking can diverge (Proposition 3).
2. *Two phases at every level.* For convex objectives and exact gradients, HUBCLIP, GT-CLIP and H$^2$GT-CLIP reach accuracy $\varepsilon$ after a clipped phase of length logarithmic in $F_0$ and a smooth phase of $\mathcal{O}(1/\varepsilon)$ rounds. Under strong convexity both phases are linear (Theorems 1, 3 and 4). With a step set by the curvature $\widehat{H}_0$ at the minimum, the initial curvature $H_1 F_0$ enters only through a logarithm (Theorem 5, Figure 3).
3. *Beyond convexity.* Nonconvex guarantees for all three methods are in Appendix A. At the federated level clipping also handles noise with infinite variance (Theorem 2). For explicit Laplacian weights, flattening the hierarchy into one graph increases the certified message count by $\Omega(m^2)$ (Proposition 1).

Experiments on logistic, exponential and quadratic losses illustrate these effects (§7). All proofs are in the appendix.

## 2 RELATED WORK

**Generalized smoothness and clipping.** The $(L_0, L_1)$ model comes from Zhang et al. (2020b;a). Convex theory for clipped and normalized methods includes Koloskova et al. (2023); Gorbunov et al. (2025); Vankov et al. (2025); Gaash et al. (2025); Chezhegov et al. (2025). Lobanov et al. (2024) make explicit the linear phase at large gradients followed by a sublinear phase; see also Gorbunov et al. (2025). The function-gap class was introduced to explain warmup (Alimisis et al., 2026; Liu et al., 2025b) and admits accelerated methods (Lobanov, 2026). It is related to the $\ell(\cdot)$-smoothness of Li et al. (2023); Tyurin (2025). The two conditions control curvature through different quantities, the suboptimality gap and the gradient norm, and neither class contains the other (Lobanov, 2026, App. G). To the best of our knowledge, existing results for the function-gap class are centralized. Koloskova et al. (2023) show the bias floor of clipped SGD, and Lobanov & Koloskova (2026) remove it under strong growth with batches of size $\mathcal{O}(\rho)$. We extend this to averages of individually clipped vectors. Related are clipping in private SGD (Chen et al., 2020; Das et al., 2023), variance-reduced clipping (Reisizadeh et al., 2025) and parameter-agnostic methods (Hübler et al., 2024).

**Distributed methods.** Error feedback (Richtárik et al., 2021), adapted to clipping and normalization (Khirirat et al., 2023; 2025; Shulgin et al., 2025), removes their bias for arbitrary heterogeneity at the cost of client memory. Decentralized analyses under $L$-smoothness are standard (Koloskova et al., 2020), including gradient tracking (Nedić et al., 2017; Qu & Li, 2018) with multi-consensus (Ye et al., 2023; Scaman et al., 2017; Lu & De Sa, 2021) or local steps (Metelev et al., 2024; Liu et al., 2025a; Nguyen et al., 2023a). Federated analyses under relaxed smoothness cover client sampling and local steps (Crawshaw et al., 2023; Demidovich et al., 2025). Decentralized analyses under $(L_0, L_1)$

assume bounded dissimilarity (Luo et al., 2026; Sun et al., 2023; Jiang et al., 2025). Bo & Wang (2026) combine gradient tracking with clipping over directed graphs under an affine dissimilarity bound. Our decentralized level needs no hub-dissimilarity condition and adds the factor $1/\sqrt{1-\lambda}$ and the convex two-phase rates. Yu et al. (2026) treat symmetric heavy-tailed noise with smoothed clipping and error feedback for strongly convex objectives. Hierarchical learning (Castiglia et al., 2021; Wang et al., 2022; Sun et al., 2022; Rodio et al., 2026) is analyzed under $L$-smoothness with unclipped updates. We are not aware of hierarchical analyses with pre-aggregation clipping or under the function-gap class.

## 3 Problem setup and structural results

**Notation.** Stacked network variables are bold, e.g. $\mathbf{X}^k = [x_1^k; \dots; x_n^k] \in \mathbb{R}^{n\times d}$ with row mean $\bar{x}^k$. We write $\mathrm{clip}_\tau(g) = \min\{1, \tau/\|g\|\}g$, $\Phi_\tau(g) := \|g\| \min\{\|g\|, \tau\}$ and $F_0 = f(x^0) - f^\star$. For a convex function $h$ with a minimizer $x^\star$, $R := \sup\{\|x - x^\star\| : h(x) \le h(x^0)\}$ is the sublevel radius. The hubs communicate over an undirected graph with a symmetric doubly stochastic matrix $W$ and $\lambda := \max\{|\lambda_2(W)|, |\lambda_n(W)|\} < 1$. Hub $i$ averages the messages of a cohort $\mathcal{S}_i^k \subseteq \mathcal{C}_i$ of size $s$. Throughout, when $H_1 = 0$, the locality restriction is void: conditions involving $1/\sqrt{H_1}$ are omitted and upper bounds such as $M/\sqrt{H_1}$ are interpreted as $+\infty$. This convention recovers the usual $L$-smooth case.

**Proposition 1** (hierarchy vs. flat gossip on the same links)**.** *Let the backbone $G$ be connected and $d$-regular with Laplacian $L_G$, and let $G_{\mathrm{flat}}$ be $G$ with $m$ leaves attached to every hub. Use the weights $W_G := I - L_G/(2d)$ and $W_{\mathrm{flat}} := I - L_{\mathrm{flat}}/(2(d+m))$. Both are symmetric, doubly stochastic and positive semidefinite, and their spectral gaps satisfy*

$$1 - \lambda_2(W_{\mathrm{flat}}) \;\le\; \frac{d}{(d+m)(m+1)}\big(1 - \lambda_2(W_G)\big).$$

*For a fixed target accuracy, the round count of Lemma 6 on the flat graph is therefore larger by the factor $\sqrt{(d+m)(m+1)/d}$. One flat round uses $nm + nd/2$ links and one backbone round $nd/2$ links. The certified message count of the flat scheme is larger by $\Omega(m^2)$ for constant $d$. The comparison concerns these two weight matrices and is not a lower bound for arbitrary flat algorithms.*

The proof in Appendix B is a Rayleigh-quotient argument. A backbone disagreement vector is lifted to the flat graph by giving every leaf the value of its hub. The leaf edges then carry no energy, and the squared norm grows by the factor $m + 1$. A flat scheme that zeroes the leaf weights recovers the backbone rate; that scheme is the hierarchy.

### 3.1 Assumptions

Let $\phi_{ij}(x) := f_{ij}(x) - f_{ij}^{\inf}$.

**Assumption 1** ($(H_0, H_1)$-smoothness)**.** Each $f_{ij}$ is bounded below by $f_{ij}^{\inf}$ and is $(H_0, H_1)$-smooth: writing $\phi_{ij}(x) := f_{ij}(x) - f_{ij}^{\inf}$, for all $x, y$ with $\|y - x\| \le 1/\sqrt{H_1}$,

$$\|\nabla f_{ij}(y) - \nabla f_{ij}(x)\| \le \big(H_0 + H_1\,\phi_{ij}(x)\big)\,\|y - x\|\,.$$

**Assumption 2** (averaged multiplicative client dissimilarity)**.** There exists $\nu \in [0, 1)$ such that for all $i$ and all $x$,

$$\frac{1}{m}\sum_{j\in\mathcal{C}_i} \|\nabla f_{ij}(x) - \nabla f_i(x)\|^2 \;\le\; \nu^2\,\|\nabla f_i(x)\|^2\,. \tag{2}$$

**Assumption 3** (pointwise multiplicative hub heterogeneity)**.** There exists $\nu_h \ge 0$ such that for all $i$ and all $x$, $\|\nabla f_i(x) - \nabla f(x)\| \le \nu_h\,\|\nabla f(x)\|$.

**Assumption 4** (strong growth, interpolation form)**.** For all $i, j, x$,

$$\mathbb{E}_\xi[\nabla F_{ij}(x;\xi)] = \nabla f_{ij}(x), \qquad \mathbb{E}_\xi\,\|\nabla F_{ij}(x;\xi) - \nabla f_{ij}(x)\|^2 \le (\rho_\ell - 1)\,\|\nabla f_{ij}(x)\|^2\,.$$

Assumption 1 contains every $L$-smooth function ($H_1 = 0$) and every $(L_0, L_1)$-smooth function with $H_0 = 3L_0$, $H_1 = 9L_1^2$; the inclusion is strict (Appendix B). Assumption 2 is the $(G, B)$-dissimilarity

of Karimireddy et al. (2020) with $G = 0$ and $B^2 = 1+\nu^2$. Its deviations vanish where $f_i$ is stationary, as for an interpolating hub. Assumption 3 is a pointwise analogue for hubs and is used only in §6. Assumption 4 is strong growth (Vaswani et al., 2019). Additive noise is not covered, since cohort averaging leaves the bias of clipping unchanged. A heavy-tailed variant is treated in §4.

### 3.2 Closure under averaging

Define the interpolation defects

$$\delta_i := f_i^{\inf} - \frac{1}{m}\sum_{j\in\mathcal{C}_i} f_{ij}^{\inf} \ \geq 0, \qquad \delta := f^\star - \frac{1}{nm}\sum_{i,j} f_{ij}^{\inf} \ \geq 0, \tag{3}$$

They vanish under interpolation if the component infima are attained at a common point.

**Proposition 2** (closure under averaging)**.** *Let $f_{ij}$ satisfy Assumption 1. Then (i) each hub objective $f_i$ is $(H_0+H_1\delta_i,\ H_1)$-smooth, and (ii) the global objective $f$ is $(\widehat{H}_0, H_1)$-smooth with $\widehat{H}_0 := H_0+H_1\delta$. No dissimilarity assumption is used, and the second constant $H_1$ is unchanged.*

The proof averages the moduli and uses $\frac{1}{m}\sum_j \phi_{ij} = \phi_i + \delta_i$. The dependence on the interpolation defect is unavoidable up to universal constants in the gradient-form definition. For $f_{1,2}(x) = \cosh(x \mp 1) - 1$ the Hessian modulus of the average equals $(1+\delta) + \bar{\phi}$ at every point (Appendix B). Along a descent trajectory the gap of $f$ does not grow. The modulus of $f$ at the iterates is then at most the *trajectory modulus*

$$M \ := \ \widehat{H}_0 + H_1 F_0 \ = \ H_0 + H_1(F_0 + \delta), \tag{4}$$

Heterogeneity enters $M$ only through $\delta$. The gradient is bounded by the gap as well (Lemma 3). This bound controls the local moduli in §5.

### 3.3 Convexity and two phases

All guarantees in the main text rest on one mechanism. Let $f$ be convex with a minimizer $x^\star$. The methods below do not increase $f$, so the iterates stay in the sublevel set of radius $R$, and convexity gives $F(x) \leq R\,\|\nabla f(x)\|$. Each method also satisfies a descent inequality $F^{k+1} \leq F^k - q\,\Phi_\tau(\nabla f(\bar{x}^k))$ for some $q > 0$. Together they give the scalar recursion

$$F^{k+1} \ \leq \ F^k - q\,\min\Big\{\frac{(F^k)^2}{R^2},\ \frac{\tau F^k}{R}\Big\}. \tag{5}$$

While $F^k \geq \tau R$, convexity forces $\big\|\nabla f(x^k)\big\| \geq \tau$. Clipping is then active, and the gap contracts by the factor $1 - q\tau/R$. This *clipped phase* takes $\frac{R}{q\tau}\log\frac{F_0}{\tau R}$ rounds. In the *smooth phase* that follows, $F^k$ decreases as $R^2/(qk)$, the rate of gradient descent. Clipping may still act in this phase, and the rate holds regardless. Under $\mu$-strong convexity the smooth phase is linear with factor $1 - q\min\{\tau/R, 2\mu\}$. Lemma 7 in Appendix B makes this precise.

## 4 Federated level: clipping before aggregation

At hub $i$, each sampled client computes a mini-batch gradient, clips it at radius $\tau$ and sends it (Algorithm 1). The hub averages the messages. An average of clipped vectors can point away from $\nabla f_i$ when a few clients carry large gradients in other directions. Under Assumption 2 such clients are few, and clipping caps each of them at $\tau$.

**Lemma 1** (alignment under the averaged condition)**.** *Let Assumption 2 hold with $\nu \leq \frac{1}{6}$ and $u_j = \mathrm{clip}_\tau(\nabla f_{ij}(x))$. Then $\frac{1}{m}\sum_j \langle \nabla f_i(x), u_j\rangle \geq \frac{1}{8}\Phi_\tau(\nabla f_i(x))$ and $\frac{1}{m}\sum_j \|u_j\|^2 \leq \frac{3}{2}\Phi_\tau(\nabla f_i(x))$, where $\Phi_\tau(g) := \|g\|\min\{\|g\|, \tau\}$. With mini-batches of size $b \geq 2304(\rho_\ell - 1)$ and $\nu \leq \frac{1}{12}$ the same holds in expectation for the cohort average $d$ with constants $\frac{1}{16}$ and $2$ (Lemma 11).*

Appendix C proves it as Lemmas 10 and 11. The proof splits the clients by their deviation. Each good client contributes at least $\frac{1}{3}\Phi_\tau(\nabla f_i)$. Bad clients form at most a $4\nu^2$ fraction, and each costs at most $\tau\,\|\nabla f_i\|$.

**Algorithm 1** HUBCLIP$(x; \mathcal{S}, \tau, b)$ — one federated round at hub $i$

**input:** iterate $x$, cohort $\mathcal{S} \subseteq \mathcal{C}_i$ of size $s$, clipping radius $\tau$, batch size $b$
hub broadcasts $x$ to every client $j \in \mathcal{S}$ ▷ server → clients
**for** each client $j \in \mathcal{S}$ **in parallel do**
    draw i.i.d. $\xi_{j,1}, \dots, \xi_{j,b} \sim \mathcal{D}_{ij}$; $g_j \leftarrow \frac{1}{b}\sum_{r=1}^{b} \nabla F_{ij}(x; \xi_{j,r})$
    transmit $u_j \leftarrow \mathrm{clip}_\tau(g_j)$ to the hub ▷ clip before aggregation
**end for**
**return** $d \leftarrow \frac{1}{s}\sum_{j\in\mathcal{S}} u_j$ ▷ stand-alone FL: $x^+ = x - \gamma d$

**Theorem 1** (HUBCLIP, convex)**.** *Let every $f_{ij}$ be convex, and let Assumption 1 and Assumption 2 with $\nu \le \frac{1}{6}$ hold. Let all $m$ clients send $\mathrm{clip}_\tau(\nabla f_{ij}(x^k))$. Let the sublevel radius $R$ of $f_i$ be finite, $F^k := f_i(x^k) - f_i^\star$, $M_f := H_0 + H_1(\delta_i + F_0)$ and*

$$\gamma \;=\; \min\Big\{\frac{1}{16 M_f},\; \frac{1}{6\sqrt{H_1}\tau},\; \frac{R}{\tau}\Big\}.$$

*Then $F^k$ is nonincreasing, and:*

*(a) $F^K \le \varepsilon$ as soon as*

$$K \;\ge\; \frac{16R}{\gamma\tau}\log^+\frac{F_0}{\tau R} + \frac{16R^2}{\gamma\varepsilon} + 1;$$

*(b) if $f_i$ is $\mu$-strongly convex, then $F^K \le \varepsilon$ as soon as*

$$K \;\ge\; \frac{16R}{\gamma\tau}\log^+\frac{F_0}{\tau R} + \max\Big\{\frac{16R}{\gamma\tau}, \frac{8}{\gamma\mu}\Big\}\log^+\frac{\tau R}{\varepsilon} + 1.$$

In both bounds the first term is the clipped phase and the rest is the smooth phase. Take $\tau = M_f/\sqrt{H_1}$ and $R \ge 1/(16\sqrt{H_1})$. Then $\gamma = 1/(16M_f)$ and $\gamma\tau = 1/(16\sqrt{H_1})$. The clipped phase lasts $256\sqrt{H_1}R\log^+\frac{\sqrt{H_1}F_0}{M_f R}$ rounds. Its length does not depend on $\varepsilon$. The smooth phase needs $256\, M_f R^2/\varepsilon$ rounds. Under strong convexity it is linear and needs $\max\{256\sqrt{H_1}R, 128M_f/\mu\}\log^+\frac{\tau R}{\varepsilon}$ rounds. For convex clients Assumption 2 forces $\delta_i = 0$ (Appendix C). Stochastic gradients and partial participation are covered by the nonconvex Theorem 6 in Appendix A. It has no bias floor and allows any cohort size.

### 4.1 HEAVY-TAILED NOISE

Gradient noise in attention models is heavy-tailed (Zhang et al., 2020c). Clipping is the standard remedy (Gorbunov et al., 2020; Sadiev et al., 2023; Nguyen et al., 2023b), and it is already part of HUBCLIP. We replace Assumption 4 by a moment condition of order $\alpha \le 2$, so the variance may be infinite.

**Assumption 5** (heavy-tailed local noise)**.** For all $i, j, x$: $\mathbb{E}_\xi[\nabla F_{ij}(x;\xi)] = \nabla f_{ij}(x)$ and, for some $\alpha \in (1, 2]$, $\mathbb{E}_\xi \|\nabla F_{ij}(x;\xi) - \nabla f_{ij}(x)\|^\alpha \le \sigma_\alpha^\alpha$.

**Theorem 2** (heavy-tailed noise, dominated radius)**.** *Let Assumption 1, Assumption 2 with $\nu \le \frac{1}{2}$ and Assumption 5 hold. Run* HUBCLIP *with mini-batches of size $b$ and fix $\beta \in (0,1)$. Let $M_f := H_0 + H_1(\delta_i + 4F_0/\beta)$, $\gamma = \min\{[32M_f]^{-1}, [\sqrt{H_1}\tau]^{-1}\}$, and assume the radius dominates the client gradients on a sublevel set:*

$$\tau \;\ge\; 2\sup\big\{\|\nabla f_{ij}(x)\| :\; j \in \mathcal{C}_i,\; \phi_i(x) \le 4F_0/\beta\big\}.$$

*Let $\varepsilon := F_0/(2\gamma K)$ and*

$$b^{\alpha-1} \;\ge\; \sigma_\alpha^\alpha \max\Big\{\frac{32}{\tau^{\alpha-1}\sqrt{\varepsilon}},\; \frac{80\,\tau^{2-\alpha}}{\varepsilon}\Big\}.$$

*Then, with probability at least $1-\beta$, $\min_{k<K}\big\|\nabla f_i(x^k)\big\|^2 \le 8F_0/(\beta\gamma K)$.*

**Algorithm 2** GT-CLIP (decentralized top level)

**input:** $\mathbf{X}^0 = \mathbf{1}(x^0)^\top$; step $\gamma$, radius $\tau$, batch $b$, inner rounds $K_{\text{in}}$ and $K_{\text{in}}^0$
$\mathbf{G}^0 \leftarrow$ row-wise mini-batch gradients of $\{f_i\}$ at $\mathbf{X}^0$
$\mathbf{V}^0 \leftarrow \textsc{FastMix}(\mathbf{G}^0, K_{\text{in}}^0)$ ▷ Algorithm 4
**for** $k = 0, 1, \ldots, K-1$ **do**
  $\mathbf{X}^{k+1} \leftarrow \textsc{FastMix}\big(\mathbf{X}^k - \gamma\ \mathrm{clip}_\tau(\mathbf{V}^k),\ K_{\text{in}}\big)$ ▷ clipped *step*: perturbation $\le \gamma\tau$
  $\mathbf{G}^{k+1} \leftarrow$ row-wise mini-batch gradients of $\{f_i\}$ at $\mathbf{X}^{k+1}$
  $\mathbf{V}^{k+1} \leftarrow \textsc{FastMix}\big(\mathbf{V}^k + \mathbf{G}^{k+1} - \mathbf{G}^k,\ K_{\text{in}}\big)$ ▷ *unclipped* tracker: $\bar{v}^k = \bar{g}^k$
**end for**

The supremum is finite by Lemma 3, since $\phi_{ij} \le m(\phi_i + \delta_i)$. The condition says that the radius exceeds every client gradient on the sublevel set the run cannot leave. Then clipping never acts on the mean, and its bias decays as $\tau^{1-\alpha}$ (Lemma 12). No variance appears in the bound. A radius below the client gradients leaves the bias of clipping uncontrolled, and that regime is open. A prior decentralized treatment, for strongly convex objectives, needs symmetric noise and smoothed clipping with error feedback (Yu et al., 2026).

## 5 DECENTRALIZED LEVEL: GRADIENT TRACKING WITH A CLIPPED STEP

Here the hubs have exact gradients $\nabla f_i$; clients enter in §6. Gradient tracking keeps a variable $v_i^k$ that follows the network-average gradient (Nedić et al., 2017; Qu & Li, 2018). The identity $\bar{v}^k = \bar{g}^k$ makes it insensitive to heterogeneity. GT-CLIP clips the step and leaves the tracker unclipped, which preserves this identity (Algorithm 2). Each per-round perturbation is then at most $\gamma\tau$. FASTMIX runs $K_{\text{in}}$ gossip exchanges of a normalized Chebyshev filter (Algorithm 4). It contracts the deviation from the mean by a factor $\theta$, and $\theta \le \theta_0$ once $K_{\text{in}} = \mathcal{O}(\log(1/\theta_0)/\sqrt{1-\lambda})$ (Lemma 6). This is the source of the factor $1/\sqrt{1-\lambda}$.

Without clipping, no constant step keeps the iterates local.

**Proposition 3** (bounded steps are necessary for a uniform guarantee)**.** *Let every $f_i(x) = \cosh(x) - 1$, which is convex and $(e, e)$-smooth. For every constant step size $\gamma > 0$ there is $x^0$ with $F_0 < \infty$ from which unclipped gradient tracking diverges doubly exponentially on any network.* GT-CLIP *converges from the same $x^0$ by Theorem 3. Hence no constant-step guarantee uniform over the class holds for unclipped tracking.*

With identical hubs the consensus subspace is invariant, and the dynamics reduce to gradient descent on $\cosh$ (Appendix D).

**Theorem 3** (GT-CLIP, convex)**.** *Let every $f_i$ be convex, let Assumption 1 hold, let hubs use exact gradients, and assume $\nabla f(x^0) \neq 0$ and $R < \infty$. Let $\tau \le M/\sqrt{H_1}$, $\gamma \le \min\{[64M]^{-1}, R/\tau\}$ and $K_{\text{in}} = \lceil \log(16n)/\sqrt{2(1-\lambda)} \rceil$. Mix the initial tracker once to accuracy $\theta_0 \le 1/(80\chi_0)$, where*

$$\chi_0 \ := \ \frac{\max_i \left\|\nabla f_i(x^0)\right\|}{\min\{\left\|\nabla f(x^0)\right\|, \tau\}}.$$

*Then $f(\bar{x}^k)$ is nonincreasing, and:*

*(a) $f(\bar{x}^K) - f^\star \le \varepsilon$ as soon as*

$$K \ \ge \ \frac{2R}{\gamma\tau} \log^+ \frac{F_0}{\tau R} + \frac{2R^2}{\gamma\varepsilon} + 1;$$

*(b) if $f$ is $\mu$-strongly convex, then $f(\bar{x}^K) - f^\star \le \varepsilon$ as soon as*

$$K \ \ge \ \frac{2R}{\gamma\tau} \log^+ \frac{F_0}{\tau R} + \max\Big\{\frac{2R}{\gamma\tau}, \frac{1}{\gamma\mu}\Big\} \log^+ \frac{\tau R}{\varepsilon} + 1.$$

*Each round costs $K_{\text{in}}$ gossip steps, and the initial mixing costs $\mathcal{O}(\log \chi_0/\sqrt{1-\lambda})$ steps once.*

**Algorithm 3** H$^2$GT-CLIP (hierarchical: decentralized top, federated bottom)

**input:** $\mathbf{X}^0 = \mathbf{1}(x^0)^\top$; step $\gamma$, radii $\tau_\ell, \tau$, batch $b$, inner rounds $K_{\text{in}}$ and $K_{\text{in}}^0$
$d_i^0 \leftarrow \text{HUBCLIP}(x^0; \mathcal{S}_i^0, \tau_\ell, b)$ at every hub $i$ ▷ Algorithm 1
$\mathbf{V}^0 \leftarrow \text{FASTMIX}(\mathbf{D}^0, K_{\text{in}}^0)$, where $\mathbf{D}^k := [d_1^k; \dots; d_n^k]$ ▷ Algorithm 4
**for** $k = 0, 1, \dots, K-1$ **do**
  $\mathbf{X}^{k+1} \leftarrow \text{FASTMIX}\big(\mathbf{X}^k - \gamma\ \text{clip}_\tau(\mathbf{V}^k),\ K_{\text{in}}\big)$ ▷ clipped step, as in Algorithm 2
  $d_i^{k+1} \leftarrow \text{HUBCLIP}(x_i^{k+1}; \mathcal{S}_i^{k+1}, \tau_\ell, b)$ at every hub $i$ ▷ Algorithm 1
  $\mathbf{V}^{k+1} \leftarrow \text{FASTMIX}\big(\mathbf{V}^k + \mathbf{D}^{k+1} - \mathbf{D}^k,\ K_{\text{in}}\big)$ ▷ *unclipped* tracker: $\bar{v}^k = \bar{d}^k$
**end for**

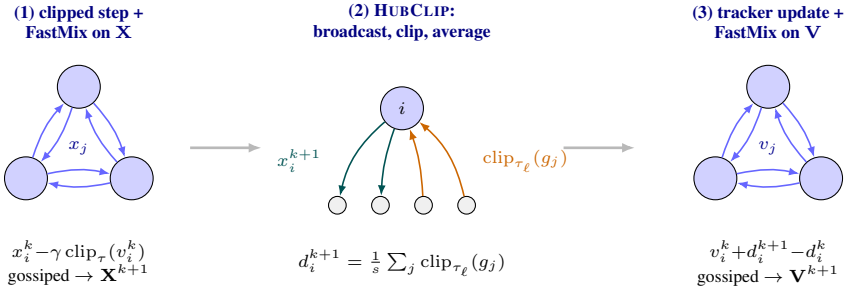


Figure 2: One round of H$^2$GT-CLIP (Algorithm 3), left to right. **(1)** Each hub forms the clipped step $x_i^k - \gamma\,\text{clip}_\tau(v_i^k)$ and the hubs run FASTMIX over the backbone. Blue arrows are the $K_{\text{in}}$ gossip exchanges, along which the local iterates $x_j$ are averaged into $\mathbf{X}^{k+1}$. Only the *step* is clipped, so each per-round displacement is bounded. **(2)** Each hub runs HUBCLIP on its updated iterate. It *broadcasts* $x_i^{k+1}$ to its clients (teal, down), and each client returns its *clipped* gradient $\text{clip}_{\tau_\ell}(g_j)$ (orange, up). The hub averages them into the surrogate direction $d_i^{k+1}$. **(3)** The tracker increment $v_i^k + d_i^{k+1} - d_i^k$ is gossiped by FASTMIX over the backbone, *unclipped*, which preserves the tracking invariant $\bar{v}^k = \bar{d}^k$ (the surrogate average). In the rigorous statements the two radii coincide, $\tau_\ell = \tau$.

If $R \geq 1/(64\sqrt{H_1})$, the choice $\tau = M/\sqrt{H_1}$ and $\gamma = 1/(64M)$ is admissible. The clipped phase then lasts $128\sqrt{H_1}R\log^+ \frac{\sqrt{H_1}F_0}{MR}$ rounds and the smooth phase $128\,MR^2/\varepsilon$ rounds. Each round costs $\mathcal{O}(\log n/\sqrt{1-\lambda})$ gossip steps. No hub-dissimilarity assumption is needed; hub heterogeneity enters only through $\chi_0$, once. The proof first propagates by induction the certificate

$$M\left\|\mathbf{X}^k - \mathbf{1}\bar{x}^k\right\|_F \leq \tfrac{m^k}{40}, \qquad \tfrac{1}{\sqrt{n}}\left\|\mathbf{V}^k - \mathbf{1}\bar{v}^k\right\|_F \leq \tfrac{m^k}{40}, \qquad m^k := \min\{\left\|\nabla f(\bar{x}^k)\right\|, \tau\}. \quad (6)$$

Consensus and tracking errors scale with the current gradient. The certificate yields the pathwise descent $F^{k+1} \leq F^k - \frac{\gamma}{2}\Phi_\tau(\nabla f(\bar{x}^k))$, which holds without convexity (Theorem 7 in Appendix A). The local moduli at the hub iterates sum to at most $\frac{5}{4}nM$. This requires $\theta n \leq \frac{1}{8}$ and costs the factor $\log n$. At $H_1 = 0$ the factor $1/\sqrt{1-\lambda}$ matches the optimal one for smooth decentralized optimization (Scaman et al., 2017; Lu & De Sa, 2021).

## 6 THE HIERARCHICAL METHOD

H$^2$GT-CLIP runs GT-CLIP over the backbone and replaces each hub gradient by the HUBCLIP surrogate $d_i$ (Algorithm 3, Figure 2). Hubs gossip clipped surrogates. Their average is aligned with $\nabla f$ under the pointwise hub condition $\nu_h \leq \frac{1}{48}$ (Lemma 15). If hubs may exchange unclipped cohort averages, the condition is not needed and §5 applies directly.

**Theorem 4** (H$^2$GT-CLIP, convex)**.** *Let every $f_{ij}$ be convex, and let Assumption 1, Assumption 2 with $\nu \leq \frac{1}{6}$ and Assumption 3 with $\nu_h \leq \frac{1}{48}$ hold, and let $R < \infty$. Let all clients send exact clipped gradients with $\tau_\ell = \tau \leq M/\sqrt{H_1}$, let $\gamma \leq \min\{[256M]^{-1}, R/\tau\}$ and $K_{\text{in}} = \lceil \log(64n)/\sqrt{2(1-\lambda)} \rceil$. Mix the initial tracker once with $K_{\text{in}}^0 = \lceil \log(1000)/\sqrt{2(1-\lambda)} \rceil$ rounds. Then $f(\bar{x}^k)$ is nonincreasing, and:*

*(a)* $f(\bar{x}^K) - f^\star \le \varepsilon$ *as soon as*

$$K \;\ge\; \frac{25R}{\gamma\tau}\log^+\frac{F_0}{\tau R} + \frac{25R^2}{\gamma\varepsilon} + 1;$$

*(b) if* $f$ *is* $\mu$*-strongly convex, then* $f(\bar{x}^K) - f^\star \le \varepsilon$ *as soon as*

$$K \;\ge\; \frac{25R}{\gamma\tau}\log^+\frac{F_0}{\tau R} + \max\Big\{\frac{25R}{\gamma\tau}, \frac{25}{2\gamma\mu}\Big\}\log^+\frac{\tau R}{\varepsilon} + 1.$$

The two phases are those of Theorem 3, with $q = \gamma/25$. Because the surrogates are clipped, the initial mixing costs a constant number of rounds. With stochastic clients the nonconvex guarantee holds with high probability for a batch that grows with the horizon (Corollary 1 in Appendix A).

**A step set by the curvature at the minimum.** In Theorems 1–4 the step is $\gamma \asymp 1/M$, where $M = \widehat{H}_0 + H_1F_0$ is the largest modulus on the sublevel set. The accuracy-dependent phase therefore pays for the curvature at the start. Convexity gives the sharper bound $\widehat{H}_0 + H_1R\,\|\nabla f\|$ on the modulus along the trajectory. With a radius $\tau \asymp \widehat{H}_0/(H_1R)$, clipping pays for the second term and the step can be set by $\widehat{H}_0$ alone. For convex clients, Assumptions 2 and 3 force a common minimizer, so $\delta = 0$ and $\widehat{H}_0 = H_0$. We keep $\widehat{H}_0$ to match the general notation. The next theorem uses exact gradients and full participation.

**Theorem 5** (convex: refined two-phase rate)**.** *Let Assumption 1 hold with* $H_1 > 0$ *and* $\widehat{H}_0 > 0$*, Assumption 2 with* $\nu \le \frac{1}{6}$ *and Assumption 3 with* $\nu_h \le \frac{1}{48}$*, and let every* $f_{ij}$ *be convex. Let all clients send exact clipped gradients with full participation and* $\tau_\ell = \tau$*, and mix the initial tracker as in Theorem 4. Let* $R < \infty$*,* $\widetilde{R} := \max\{R, 1/\sqrt{H_1}\}$ *and* $\Gamma := \max\{1, M/(180\widehat{H}_0)\}$*. Run Algorithm 3 with*

$$\tau = \frac{\widehat{H}_0}{H_1\widetilde{R}}, \qquad \gamma = \frac{1}{180\,\widehat{H}_0}, \qquad K_{\mathrm{in}} = \Big\lceil\frac{\log(2048\,n\Gamma)}{\sqrt{2(1-\lambda)}}\Big\rceil.$$

*Then* $f(\bar{x}^k)$ *is nonincreasing, and* $f(\bar{x}^K) - f^\star \le \varepsilon$ *as soon as*

$$K \;\ge\; 3600\,H_1\widetilde{R}^2\log^+\frac{H_1F_0}{\widehat{H}_0} \;+\; \frac{3600\,\widehat{H}_0\widetilde{R}^2}{\varepsilon} \;+\; 1.$$

*Each round costs* $K_{\mathrm{in}}$ *gossip steps.*

The first term counts the rounds of the clipped phase. There the gap contracts geometrically, and the curvature $H_1F_0$ at the start enters only through the logarithm. The second term counts the rounds of the smooth phase. It is the rate of gradient descent on an $\widehat{H}_0$-smooth function. An analysis with a single smoothness constant must use $M$ and needs $MR^2/\varepsilon$ rounds. For small $\varepsilon$ and $R \ge 1/\sqrt{H_1}$ the two bounds differ by the factor $1 + H_1F_0/\widehat{H}_0$, up to absolute constants. The shape matches centralized bounds for clipped gradient descent under $(L_0, L_1)$-smoothness, with $\widehat{H}_0$ in place of $L_0$ and $H_1$ in place of $L_1^2$ (Koloskova et al., 2023; Gorbunov et al., 2025). Figure 3 shows both phases on real data with tuned parameters.

**Nonconvex objectives and privacy.** Without convexity all three methods converge to stationary points in the clipped measure $\Phi_\tau$ (Appendix A). Gaussian noise on the clipped messages, re-clipped at the hub, gives per-round differential privacy with sensitivity $2\tau$ (Proposition 4).

## 7 Numerical experiments

**Setup.** The backbone is a ring of $n = 8$ hubs with Metropolis weights ($1 - \lambda = 0.195$), and each hub has $m = 8$ clients. Data are split non-IID by label and by the leading principal component. We use logistic and exponential losses on breast-cancer ($N$=569, $d$=30) and digits ($N$=1797, $d$=64, digits 0–4 vs. 5–9). The Hessian of the exponential loss is bounded by the loss value, so it satisfies Assumption 1 with small $H_0$. We also use a heterogeneous quadratic with $d = 20$ and condition number 50. The baselines are hierarchical decentralized SGD (Sun et al., 2022; Castiglia et al., 2021),

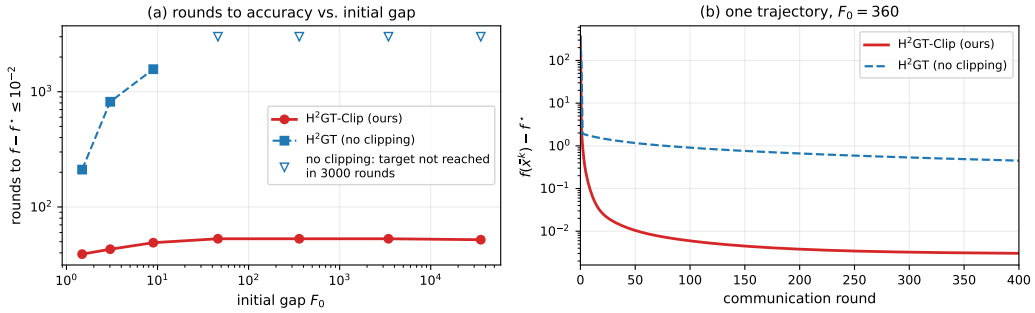


Figure 3: Theorem 5 on real data. Exponential loss on breast-cancer, $n = m = 8$, $K_{\text{in}} = 8$, starts $x^0 = -s\,w$ for $s = 1, \dots, 7$, where $w$ is the normalized mean of $y_i x_i$. Both methods are tuned separately for each $F_0$. **(a)** Rounds to $f - f^\star \le 10^{-2}$. For H$^2$GT-CLIP the count grows from 39 to 53 while $F_0$ grows from 1.5 to $3.5 \cdot 10^4$. Without clipping the count grows linearly in $F_0$. For $F_0 \ge 46$ no step in the grid reaches the target within 3000 rounds. **(b)** The two phases on one trajectory, $F_0 = 360$.

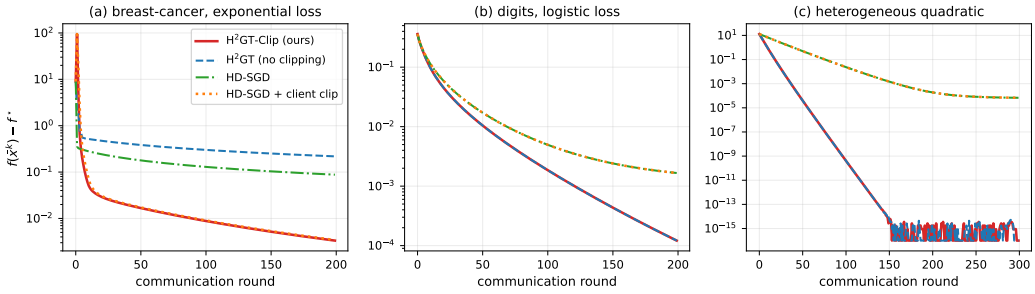


Figure 4: Hierarchical training with $n = m = 8$, a ring backbone and a non-IID split. Each method uses its own tuned $(\gamma, \tau)$; the metric is $f(\bar{x}^k) - f^\star$. **(a)** Exponential loss on breast-cancer from a distant start. The unclipped methods diverge at our step $\gamma = 16$. At their best stable steps they end 26 and 65 times worse. **(b)** Logistic loss on digits with $K_{\text{in}} = 2$. Tracking removes the heterogeneity floor of hierarchical D-SGD. **(c)** Heterogeneous quadratic with $K_{\text{in}} = 2$. Tracking reaches machine precision, and clipping does not slow it down.

with and without client clipping, and H$^2$GT without clipping. Each method is tuned on a log grid (Appendix F).

**Findings.** Figure 3 illustrates Theorem 5: after tuning for each start, the number of rounds of our method barely depends on $F_0$. The tuned parameters show the mechanism, not the theorem constants. On the exponential loss from a distant start, both unclipped methods diverge at our step size. At their best stable steps they end 26 and 65 times worse (Figure 4a). With cheap gossip ($K_{\text{in}} = 2$), tracking removes the heterogeneity floor of hierarchical D-SGD (Figure 4b,c). To reach $f - f^\star \le 10^{-2}$, the hierarchy uses 38 times fewer messages than GT-CLIP on the flat graph. With one label-flipped client whose gradients are scaled by 50, HUBCLIP reaches $\min_k \|\nabla f\| = 0.017$. Clipping after aggregation reaches 0.027, and no clipping reaches 0.034 (Figure 5). Along our trajectories the measured $\nu$ and $\nu_h$ are far above the thresholds of the theorems. The alignment ratios used in the proofs still stay above 0.81 for clients and 0.51 for hubs, against the certified $\frac{1}{8}$ and $\frac{1}{12}$ (Appendix F). The thresholds are sufficient, not necessary.

**Limitations.** The constants and the thresholds on $\nu$ and $\nu_h$ are not optimized. Additive noise is not covered. The decentralized and hierarchical theorems use exact client gradients or large batches. The averaged hub condition, heavy tails in the large-gradient regime and network lower bounds for the class are open.

### ETHICS STATEMENT

This work is theoretical and methodological. It studies optimization algorithms for hierarchical federated and decentralized learning and does not involve human subjects, personal data or new data collection. The experiments use two public benchmark datasets, breast-cancer and digits, distributed with scikit-learn, and a synthetic quadratic. Clipping before aggregation bounds the influence of each client. This is the property required by differential privacy. Proposition 4 gives a per-round guarantee only. A deployed system would still need a full privacy accounting over all rounds and an audit of its threat model. We are not aware of direct harmful applications beyond those of distributed learning in general.

### AI USE STATEMENT

Large language models were used throughout this project. They helped to draft and edit sections of the paper, to develop and check proofs and numerical constants, to search the literature, and to write and debug the experiment scripts. All results, proofs and experimental claims were verified by the authors, who take full responsibility for the content of the paper. No language model is an author of this work.

### REPRODUCIBILITY STATEMENT

All assumptions are stated in Section 3, and complete proofs of every result are in Appendices A–E. Each proof traces its numerical constants through the stated lemmas. The algorithms are given in full as pseudocode, including the consensus subroutine (Algorithms 1–4). Appendix F lists the data preprocessing, the network topology, the tuning grids, the selected parameters and the measured constants. The supplementary material contains the code that reproduces every figure and table, with instructions. All experiments use a fixed random seed and run on a CPU in a few minutes.

## REFERENCES

Mehdi Salehi Heydar Abad, Emre Ozfatura, Deniz Gündüz, and Özgür Erçetin. Hierarchical federated learning across heterogeneous cellular networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8866–8870, 2020.

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *ACM CCS*, 2016.

Foivos Alimisis, Rustem Islamov, and Aurélien Lucchi. Why do we need warm-up? A theoretical perspective. In *International Conference on Machine Learning (ICML)*, 2026.

Yanan Bo and Yongqiang Wang. Provably convergent decentralized optimization over directed graphs under generalized smoothness. *arXiv preprint arXiv:2601.03566*, 2026.

Timothy Castiglia, Anirban Das, and Stacy Patterson. Multi-level local SGD: Distributed SGD for heterogeneous hierarchical networks. In *ICLR*, 2021.

Xiangyi Chen, Zhiwei Steven Wu, and Mingyi Hong. Understanding gradient clipping in private SGD: A geometric perspective. In *Advances in Neural Information Processing Systems*, volume 33, pp. 13773–13782, 2020.

Savelii Chezhegov, Aleksandr Beznosikov, Samuel Horváth, and Eduard Gorbunov. Convergence of clipped-SGD for convex $(L_0, L_1)$-smooth optimization with heavy-tailed noise. *arXiv:2505.20817*, 2025.

Michael Crawshaw, Mingrui Liu, Francesco Orabona, Wei Zhang, and Zhenxun Zhuang. Robustness to unbounded smoothness of generalized SignSGD. In *NeurIPS*, 2022.

Michael Crawshaw, Yajie Bao, and Mingrui Liu. Federated learning with client subsampling, data heterogeneity, and unbounded smoothness: A new algorithm and lower bounds. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Rudrajit Das, Satyen Kale, Zheng Xu, Tong Zhang, and Sujay Sanghavi. Beyond uniform Lipschitz condition in differentially private optimization. In *ICML*, 2023.

Yury Demidovich, Petr Ostroukhov, Grigory Malinovsky, Samuel Horváth, Martin Takáč, Peter Richtárik, and Eduard Gorbunov. Methods with local steps and random reshuffling for generally smooth non-convex federated optimization. In *ICLR*, 2025.

Ofir Gaash, Kfir Yehuda Levy, and Yair Carmon. Convergence of clipped SGD on convex $(L_0, L_1)$-smooth functions. In *NeurIPS*, 2025.

Eduard Gorbunov, Marina Danilova, and Alexander Gasnikov. Stochastic optimization with heavy-tailed noise via accelerated gradient clipping. In *NeurIPS*, 2020.

Eduard Gorbunov, Nazarii Tupitsa, Sayantan Choudhury, Alen Aliev, Peter Richtárik, Samuel Horváth, and Martin Takáč. Methods for convex $(L_0, L_1)$-smooth optimization: Clipping, acceleration, and adaptivity. In *ICLR*, 2025.

Florian Hübler, Junchi Yang, Xiang Li, and Niao He. Parameter-agnostic optimization under relaxed smoothness. In *AISTATS*, 2024.

Zhanhong Jiang, Aditya Balu, and Soumik Sarkar. Decentralized relaxed smooth optimization with gradient descent methods. *arXiv:2508.08413*, 2025.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. SCAFFOLD: Stochastic controlled averaging for federated learning. In *ICML*, 2020.

Sarit Khirirat, Eduard Gorbunov, Samuel Horváth, Rustem Islamov, Fakhri Karray, and Peter Richtárik. Clip21: Error feedback for gradient clipping. *arXiv:2305.18929*, 2023.

Sarit Khirirat, Abdurakhmon Sadiev, Artem Riabinin, Eduard Gorbunov, and Peter Richtárik. Error feedback under $(L_0, L_1)$-smoothness: Normalization and momentum. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Anastasia Koloskova, Nicolas Loizou, Sadra Boreiri, Martin Jaggi, and Sebastian Stich. A unified theory of decentralized SGD with changing topology and local updates. In *ICML*, 2020.

Anastasia Koloskova, Hadrien Hendrikx, and Sebastian U Stich. Revisiting gradient clipping: Stochastic bias and tight convergence guarantees. In *ICML*, 2023.

Haochuan Li, Jian Qian, Yi Tian, Alexander Rakhlin, and Ali Jadbabaie. Convex and non-convex optimization under generalized smoothness. In *Advances in Neural Information Processing Systems*, volume 36, pp. 40238–40271, 2023.

Ji Liu and A. Stephen Morse. Accelerated linear iterations for distributed averaging. *Annual Reviews in Control*, 35(2):160–165, 2011.

Lumin Liu, Jun Zhang, Shenghui Song, and Khaled B Letaief. Client-edge-cloud hierarchical federated learning. In *IEEE International Conference on Communications (ICC)*, pp. 1–6, 2020.

Yue Liu, Tao Lin, Anastasia Koloskova, and Sebastian U. Stich. Decentralized gradient tracking with local steps. *Optimization Methods and Software*, 40(5):1153–1180, 2025a.

Yuxing Liu, Yuze Ge, Rui Pan, An Kang, and Tong Zhang. Theoretical analysis on how learning rate warmup accelerates convergence. *arXiv preprint arXiv:2509.07972*, 2025b.

Aleksandr Lobanov. A few accelerated algorithms for convex optimization under $(h_0, h_1)$-smoothness. *arXiv preprint arXiv:2608.04884*, 2026.

Aleksandr Lobanov and Anastasia Koloskova. Avoiding bias in clipped SGD for overparameterized models under generalized smoothness. *arXiv:2605.14800*, 2026.

Aleksandr Lobanov, Alexander Gasnikov, Eduard Gorbunov, and Martin Takáč. Linear convergence rate in convex setup is possible! Gradient descent method variants under $(l_0, l_1)$-smoothness. *arXiv preprint arXiv:2412.17050*, 2024.

Yucheng Lu and Christopher De Sa. Optimal complexity in decentralized training. In *ICML*, 2021.

Luo Luo, Xue Cui, Tingkai Jia, and Cheng Chen. Decentralized stochastic nonconvex optimization under the $(L_0, L_1)$-smoothness. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, pp. 3470–3479, 2026. doi: 10.1145/3770855.3817843.

Dmitry Metelev, Savelii Chezhegov, Alexander Rogozin, Aleksandr Beznosikov, Alexander Sholokhov, Alexander Gasnikov, and Dmitry Kovalev. Decentralized finite-sum optimization over time-varying networks. *arXiv preprint arXiv:2402.02490*, 2024.

Ilya Mironov, Kunal Talwar, and Li Zhang. Rényi differential privacy of the sampled Gaussian mechanism. arXiv:1908.10530, 2019.

Eduardo Montijano, Juan Ignacio Montijano, and Carlos Sagüés. Chebyshev polynomials in distributed consensus applications. *IEEE Transactions on Signal Processing*, 61(3):693–706, 2013.

Angelia Nedić, Alex Olshevsky, and Wei Shi. Achieving geometric convergence for distributed optimization over time-varying graphs. *SIAM Journal on Optimization*, 27(4):2597–2633, 2017.

Edward Duc Hien Nguyen, Sulaiman A. Alghunaim, Kun Yuan, and César A. Uribe. On the performance of gradient tracking with local updates. In *62nd IEEE Conference on Decision and Control (CDC)*, pp. 4309–4313, 2023a.

Ta Duy Nguyen, Thien Hang Nguyen, Alina Ene, and Huy L. Nguyen. Improved convergence in high probability of clipped gradient methods with heavy tailed noise. In *NeurIPS*, 2023b.

Iosif Pinelis. Multidimensional probability inequalities via spherical symmetry. arXiv:2210.04391, 2022.

Guannan Qu and Na Li. Harnessing smoothness to accelerate distributed optimization. *IEEE Transactions on Control of Network Systems*, 5(3):1245–1260, 2018.

Amirhossein Reisizadeh, Haochuan Li, Subhro Das, and Ali Jadbabaie. Variance-reduced clipping for non-convex optimization. In *IEEE ICASSP*, 2025.

Peter Richtárik, Igor Sokolov, and Ilyas Fatkhullin. EF21: A new, simpler, theoretically better, and practically faster error feedback. In *NeurIPS*, 2021.

Angelo Rodio, Giovanni Neglia, Zheng Chen, and Erik G. Larsson. A unified convergence analysis for semi-decentralized learning: sampled-to-sampled vs. sampled-to-all communication. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 25151–25159, 2026.

Abdurakhmon Sadiev, Marina Danilova, Eduard Gorbunov, Samuel Horváth, Gauthier Gidel, Pavel Dvurechensky, Alexander Gasnikov, and Peter Richtárik. High-probability bounds for stochastic optimization and variational inequalities: the case of unbounded variance. In *ICML*, 2023.

Kevin Scaman, Francis Bach, Sébastien Bubeck, Yin Tat Lee, and Laurent Massoulié. Optimal algorithms for smooth and strongly convex distributed optimization in networks. In *ICML*, 2017.

Egor Shulgin, Sarit Khirirat, and Peter Richtárik. Private federated learning with provable convergence via smoothed normalization. In *CFAgentic Workshop at ICML*, 2025.

Tao Sun, Qingsong Wang, Dongsheng Li, and Bao Wang. Clipping for nonconvex DSGD under weak smoothness assumptions. Preprint, `https://www.researchgate.net/publication/376614522`, 2023.

Yuchang Sun, Jiawei Shao, Yuyi Mao, Jessie Hui Wang, and Jun Zhang. Semi-decentralized federated edge learning for fast convergence on non-IID data. In *IEEE Wireless Communications and Networking Conference (WCNC)*, 2022.

Alexander Tyurin. Toward a unified theory of gradient descent under generalized smoothness. In *International Conference on Machine Learning (ICML)*, volume 267 of *PMLR*, pp. 60493–60514, 2025.

Daniil Vankov, Anton Rodomanov, Angelia Nedich, Lalitha Sankar, and Sebastian U Stich. Optimizing $(L_0, L_1)$-smooth functions by gradient methods. In *ICLR*, 2025.

Sharan Vaswani, Francis Bach, and Mark Schmidt. Fast and faster convergence of SGD for over-parameterized models and an accelerated perceptron. In *AISTATS*, 2019.

Jiayi Wang, Shiqiang Wang, Rong-Rong Chen, and Mingyue Ji. Demystifying why local aggregation helps: Convergence analysis of hierarchical SGD. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 8548–8556, 2022.

Yu-Xiang Wang, Borja Balle, and Shiva Prasad Kasiviswanathan. Subsampled rényi differential privacy and analytical moments accountant. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 1226–1235. PMLR, 2019.

Haishan Ye, Luo Luo, Ziang Zhou, and Tong Zhang. Multi-consensus decentralized accelerated gradient descent. *Journal of Machine Learning Research*, 24(306):1–50, 2023.

Shuhua Yu, Dušan Jakovetić, and Soummya Kar. Smoothed gradient clipping and error feedback for decentralized optimization under symmetric heavy-tailed noise. *SIAM Journal on Optimization*, 36(2):703–728, 2026.

Bohang Zhang, Jikai Jin, Cong Fang, and Liwei Wang. Improved analysis of clipping algorithms for non-convex optimization. In *NeurIPS*, 2020a.

Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. In *ICLR*, 2020b.

Jingzhao Zhang, Sai Praneeth Karimireddy, Andreas Veit, Seungyeon Kim, Sashank Reddi, Sanjiv Kumar, and Suvrit Sra. Why are adaptive methods good for attention models? In *NeurIPS*, 2020c.

# A NONCONVEX GUARANTEES

The main text states the convex guarantees. This appendix states the nonconvex ones and explains how they are proved. The proofs are in Appendices C–E. The convex theorems of the main text reuse the descent inequalities established in these proofs.

**Common structure.** All guarantees are stated for the clipped stationarity measure $\Phi_\tau(g) = \|g\| \min\{\|g\|, \tau\}$. It equals $\|g\|^2$ for small gradients and $\tau \|g\|$ for large ones. Each proof has three steps. First, an alignment lemma shows that the applied direction $u$ satisfies

$$\langle \nabla f, u \rangle \;\geq\; c\, \Phi_\tau(\nabla f), \qquad \|u\|^2 \;\leq\; C\, \Phi_\tau(\nabla f)$$

for absolute constants $c, C$. Second, clipping keeps every displacement inside the locality radius $1/\sqrt{H_1}$. The descent lemma then holds with the modulus at the current point. Third, the resulting inequality $F^{k+1} \leq F^k - c'\gamma\, \Phi^k$ is summed over $k$. Table 1 lists the results.

Table 1: Nonconvex guarantees. All oracles are exact unless stated otherwise.

| Result | Method | Setting | Guarantee on $\min_{k<K} \Phi_\tau(\nabla f)$ | Proof |
|---|---|---|---|---|
| Thm. 6 | HUBCLIP | stochastic, any cohort | $64F_0/(\beta\gamma K)$ w.p. $1-\beta$ | App. C |
| Thm. 7 | GT-CLIP | exact hub gradients | $2F_0/(\gamma K)$ | App. D |
| Thm. 8 | H$^2$GT-CLIP | exact clients | $25F_0/(\gamma K)$ | App. E |
| Cor. 1 | H$^2$GT-CLIP | stochastic clients | $25F_0/(\gamma K)$ w.p. $1-\beta$ | App. E |
| Thm. 2 | HUBCLIP | $\alpha$-moment noise | $\|\nabla f_i\|^2 \leq 8F_0/(\beta\gamma K)$ w.p. $1-\beta$ | App. C.4 |

### A.1 FEDERATED LEVEL

**Theorem 6** (HUBCLIP, nonconvex)**.** *Let Assumptions 1, 2 ($\nu \le \frac{1}{12}$) and 4 hold at hub $i$, with any cohort size $s \ge 1$ and batch size $b \ge 2304(\rho_\ell - 1)$. Fix $\beta \in (0,1)$, write $F_0 := \phi_i(x^0)$ for the hub gap, let $M_f := H_0 + H_1(\delta_i + 2F_0/\beta)$ and $\gamma = \min\{[32M_f]^{-1}, [\sqrt{H_1}\tau]^{-1}\}$. Then with probability at least $1-\beta$,*

$$\min_{k<K} \Phi_\tau\big(\nabla f_i(x^k)\big) \;\le\; \frac{64\,F_0}{\beta\,\gamma\,K}, \qquad \textit{hence} \qquad \min_{k<K} \big\|\nabla f_i(x^k)\big\| \le \varepsilon \;\textit{ once }\; K \ge \frac{64F_0}{\beta\gamma\varepsilon^2} \;\; (\varepsilon \le \tau).$$

The alignment of the cohort average follows from the averaged condition of Assumption 2 by counting good and bad clients (Lemmas 10 and 11). The modulus of $f_i$ at the random iterate is controlled by a stopping time at the exit from a sublevel set. A supermartingale argument bounds the probability of an early exit. The bound has no floor, the cohort size is arbitrary, and the constants do not depend on $m$ or $d$.

### A.2 DECENTRALIZED LEVEL

**Theorem 7** (GT-CLIP, nonconvex, exact hub gradients)**.** *Let Assumption 1 hold, let hubs use exact gradients $\nabla f_i$, and assume $\nabla f(x^0) \ne 0$. Let $\tau \le M/\sqrt{H_1}$, $\gamma \le [64M]^{-1}$ and $K_{\mathrm{in}} = \lceil \log(16n)/\sqrt{2(1-\lambda)} \rceil$, so that $\theta n \le \frac{1}{8}$. Let*

$$\chi_0 := \frac{\max_i \big\|\nabla f_i(x^0)\big\|}{\min\{\|\nabla f(x^0)\|, \tau\}},$$

*and mix the initial tracker once to accuracy $\theta_0 \le 1/(80\chi_0)$, which takes $K_{\mathrm{in}}^0 = \mathcal{O}(\log \chi_0/\sqrt{1-\lambda})$ rounds. Then $f(\bar{x}^k)$ is nonincreasing and*

$$\min_{k<K} \Phi_\tau\big(\nabla f(\bar{x}^k)\big) \;\le\; \frac{2F_0}{\gamma K}.$$

*The same bound holds at every hub iterate up to the factor $(\frac{41}{40})^2$. For $\varepsilon \le \tau$, choosing the largest curvature-based step $\gamma = 1/(64M)$ gives*

$$K \cdot K_{\mathrm{in}} + K_{\mathrm{in}}^0 = \mathcal{O}\left(\frac{MF_0\varepsilon^{-2}\log n + \log \chi_0}{\sqrt{1-\lambda}}\right)$$

*hub-to-hub gossip exchanges.*

The proof propagates the certificate (6) by induction. It keeps the applied direction aligned with $\nabla f(\bar{x}^k)$ and every hub displacement inside the locality radius. The local moduli at the hub iterates may differ across hubs. Their sum is controlled by the self-bounding property of the gradient (Lemma 14).

### A.3 HIERARCHICAL LEVEL

**Theorem 8** ($\mathrm{H}^2$GT-CLIP, nonconvex)**.** *Let Assumptions 1, 2 ($\nu \le \frac{1}{6}$) and 3 ($\nu_h \le \frac{1}{48}$) hold. Let all clients send exact clipped gradients with $\tau_\ell = \tau \le M/\sqrt{H_1}$, let $\gamma \le [256M]^{-1}$ and $K_{\mathrm{in}} = \lceil \log(64n)/\sqrt{2(1-\lambda)} \rceil$. Mix the initial tracker once with $K_{\mathrm{in}}^0 = \lceil \log(1000)/\sqrt{2(1-\lambda)} \rceil$ rounds. Then*

$$\min_{k<K} \Phi_\tau\big(\nabla f(\bar{x}^k)\big) \;\le\; \frac{25F_0}{\gamma K}.$$

*Choosing $\gamma = 1/(256M)$ gives communication complexity $\mathcal{O}(MF_0\varepsilon^{-2}\log n/\sqrt{1-\lambda})$, and each hub uses $m$ client gradients per round.*

The surrogates are Lipschitz in the iterate with the averaged client modulus (Lemma 16). Under the pointwise hub condition their average is aligned with $\nabla f$ (Lemma 15). One lemma controls a single round under the certificate (Lemma 17). It is used for Theorem 8, for Corollary 1 and for the convex Theorems 4 and 5.

With stochastic clients the surrogate at hub $i$ differs from the exact one by a vector $\xi_i^k$. If every mini-batch gradient is close to the exact one in relative terms, these vectors are small compared with $m^k$.

**Corollary 1** (stochastic clients, large batch)**.** *Let the assumptions of Theorem 8 and Assumption 4 hold, and let every client compute mini-batch gradients of size*

$$b \;\geq\; 1.6 \cdot 10^5 \,(\rho_\ell - 1)\, \frac{(K+1)\, nm}{\beta}.$$

*Then, with probability at least* $1-\beta$*, the conclusion of Theorem 8 holds for* H$^2$GT-CLIP *with stochastic clients.*

The batch grows with the horizon. Fresh noise enters the tracker in every round, and the proof requires all $(K+1)nm$ draws to be accurate at once.

# B PRELIMINARIES AND STRUCTURAL RESULTS

This appendix collects the notation and the technical lemmas used by all later proofs. It then proves the structural results of Section 3. Each lemma is preceded by a short comment on its content and its role.

## B.1 NOTATION AND TECHNICAL LEMMAS

**Notation.** For a point $x$ we write $g_{ij} = \nabla f_{ij}(x)$, $g_i = \nabla f_i(x)$ and $g = \nabla f(x)$. The gaps are $\phi_{ij} = f_{ij} - f_{ij}^{\inf}$, $\phi_i = f_i - f_i^{\inf}$ and $F = f - f^\star$. The moduli are

$$M_{ij}(x) = H_0 + H_1\phi_{ij}(x), \qquad M_i(x) = H_0 + H_1(\delta_i + \phi_i(x)), \qquad M(x) = \widehat{H}_0 + H_1 F(x),$$

with $\widehat{H}_0 = H_0 + H_1\delta$. Clipping $\mathrm{clip}_\tau$ is the Euclidean projection onto the ball of radius $\tau$. Hence it is non-expansive, $\|\mathrm{clip}_\tau(u) - \mathrm{clip}_\tau(v)\| \leq \|u - v\|$, and $\langle u, \mathrm{clip}_\tau(u)\rangle = \Phi_\tau(u)$. The map $t \mapsto t\min\{t, \tau\}$ is nondecreasing on $[0,\infty)$. For $c \geq 1$ we use $\min\{\tau, ct\} \leq c\min\{\tau, t\}$.

The first lemma is the basic tool of the class. It bounds the function along a short step by a quadratic whose curvature is the modulus at the starting point. The bound holds only for steps of length at most $1/\sqrt{H_1}$. Every descent inequality in the paper rests on it. It applies to $f_{ij}$ with $H_0' = H_0$, to $f_i$ with $H_0' = H_0 + H_1\delta_i$ and to $f$ with $H_0' = \widehat{H}_0$ (Proposition 2).

**Lemma 2** (descent)**.** *Let* $h$ *satisfy the gradient bound of Assumption 1 with constants* $(H_0', H_1)$*, gap* $\phi_h$ *and modulus* $M_h(x) = H_0' + H_1\phi_h(x)$*. Then for every* $x$ *and every* $s$ *with* $\|s\| \leq 1/\sqrt{H_1}$*,*

$$h(x+s) \;\leq\; h(x) + \langle \nabla h(x), s\rangle + \frac{M_h(x)}{2}\,\|s\|^2\,.$$

*Proof.* For $t \in [0,1]$ the point $x + ts$ is within distance $1/\sqrt{H_1}$ of $x$. The gradient bound at the base point $x$ gives

$$h(x+s) - h(x) - \langle \nabla h(x), s\rangle = \int_0^1 \langle \nabla h(x+ts) - \nabla h(x), s\rangle \; dt \leq \int_0^1 M_h(x)\, t\, \|s\|^2 \; dt = \frac{M_h(x)}{2}\,\|s\|^2\,.$$

□

The next lemma bounds the gradient by the gap. In the class the modulus depends on the gap, so this bound also controls how fast the modulus can change between nearby points. It is used to compare the moduli at different hub iterates (Lemma 14) and to show that the radius condition of Theorem 2 is finite.

**Lemma 3** (self-bounding gradient)**.** *Under the conditions of Lemma 2 with* $H_1 > 0$*,*

$$\|\nabla h(x)\| \;\leq\; \sqrt{2M_h(x)\phi_h(x)} + \sqrt{H_1}\,\phi_h(x) + \frac{M_h(x)}{2\sqrt{H_1}}.$$

*For* $H_1 = 0$ *one has* $\|\nabla h(x)\| \leq \sqrt{2H_0'\phi_h(x)}$*.*

*Proof.* Let $u = \nabla h(x)/\|\nabla h(x)\|$, $t = \min\{\|\nabla h(x)\|/M_h(x), 1/\sqrt{H_1}\}$ and $s = -tu$. Since $h(x+s) \geq h^{\inf}$, Lemma 2 gives

$$\phi_h(x) \;\geq\; h(x) - h(x+s) \;\geq\; t\,\|\nabla h(x)\| - \frac{M_h(x)}{2}t^2.$$

If $t = \|\nabla h(x)\|/M_h(x)$, this reads $\phi_h \geq \|\nabla h\|^2/(2M_h)$, so $\|\nabla h\| \leq \sqrt{2M_h\phi_h}$. If $t = 1/\sqrt{H_1}$, then $M_h/\sqrt{H_1} < \|\nabla h\|$ and $\phi_h \geq \|\nabla h\|/\sqrt{H_1} - M_h/(2H_1)$, so $\|\nabla h\| \leq \sqrt{H_1}\phi_h + M_h/(2\sqrt{H_1})$. The sum of the two bounds covers both cases. For $H_1 = 0$ only the first case occurs. □

The angle lemma treats a single clipped vector. If $v$ is close to $u$ in relative terms, then $\mathrm{clip}_\tau(v)$ still makes an acute angle with $u$, and its inner product with $u$ is of order $\Phi_\tau(u)$. The alignment lemmas of Appendix C apply it to each client whose gradient is close to the hub gradient.

**Lemma 4** (angle lemma)**.** *If $\|v-u\| \leq \nu\|u\|$ with $\nu \in [0,1)$, then*

$$\langle u, \mathrm{clip}_\tau(v)\rangle \;\geq\; \frac{1-\nu}{1+\nu}\,\Phi_\tau(u), \qquad \|\mathrm{clip}_\tau(v)\| \;\leq\; \min\{\tau, (1+\nu)\|u\|\}.$$

*Proof.* We have $\langle u, v\rangle \geq \|u\|^2 - \|u\|\,\|v-u\| \geq (1-\nu)\|u\|^2$ and $\|v\| \leq (1+\nu)\|u\|$. If $\|v\| \leq \tau$, then $\mathrm{clip}_\tau(v) = v$ and $\langle u, v\rangle \geq (1-\nu)\|u\|^2 \geq \frac{1-\nu}{1+\nu}\Phi_\tau(u)$, since $\Phi_\tau(u) \leq \|u\|^2$. If $\|v\| > \tau$, then

$$\langle u, \mathrm{clip}_\tau(v)\rangle = \frac{\tau\langle u, v\rangle}{\|v\|} \;\geq\; \frac{\tau(1-\nu)\|u\|^2}{(1+\nu)\|u\|} = \frac{1-\nu}{1+\nu}\,\tau\|u\| \;\geq\; \frac{1-\nu}{1+\nu}\,\Phi_\tau(u),$$

since $\Phi_\tau(u) \leq \tau\|u\|$. The norm bound follows from $\|\mathrm{clip}_\tau(v)\| \leq \min\{\tau, \|v\|\}$. □

The next lemma compares the clipped versions of two close vectors. The error is at most $3\min\{\|a\|, \tau\}$ times the relative perturbation. Near stationarity it is therefore as small as the clipped gradient. All certified errors must scale in this way. It controls the mini-batch noise in the hub surrogates (Corollary 1).

**Lemma 5** (clipped differences)**.** *If $a \neq 0$ and $\|b-a\| \leq \frac{1}{16}\|a\|$, then*

$$\|\mathrm{clip}_\tau(a) - \mathrm{clip}_\tau(b)\| \;\leq\; 3\min\{\|a\|, \tau\}\,\frac{\|a-b\|}{\|a\|}.$$

*Proof.* If $\|a\| < \tau$, the factor $3\min\{\|a\|, \tau\}/\|a\|$ is at least 1, and the claim is non-expansiveness. Let $\|a\| \geq \tau$. If also $\|b\| \geq \tau$, then $\|\mathrm{clip}_\tau(a) - \mathrm{clip}_\tau(b)\| = \tau\|a/\|a\| - b/\|b\|\| \leq 2\tau\|a-b\|/\|a\|$. If $\|b\| < \tau$, then $\mathrm{clip}_\tau(b) = b$ and $\|\tau a/\|a\| - b\| \leq (\|a\| - \tau) + \|a-b\| \leq 2\|a-b\|$. Moreover $\|a\| \leq \|b\| + \|a-b\| < \tau + \frac{1}{16}\|a\|$, so $\|a\| < \frac{16}{15}\tau$ and $2\|a-b\| \leq \frac{32}{15}\tau\|a-b\|/\|a\|$. □

**Algorithm 4** FastMix$(\mathbf{X}, K_{\mathrm{in}})$ — normalized Chebyshev consensus filter

**input:** $\mathbf{X} \in \mathbb{R}^{n\times d}$, symmetric doubly stochastic $W$ with $\lambda = \max\{|\lambda_2(W)|, |\lambda_n(W)|\} \in (0,1)$, inner rounds $K_{\mathrm{in}} \geq 1$
$\mathbf{Z}^{(0)} \leftarrow \mathbf{X}$, $\quad \mathbf{Z}^{(1)} \leftarrow W\mathbf{X}/\lambda$
**for** $t = 1, 2, \ldots, K_{\mathrm{in}} - 1$ **do**
  $\mathbf{Z}^{(t+1)} \leftarrow 2W\mathbf{Z}^{(t)}/\lambda - \mathbf{Z}^{(t-1)}$ ▷ one gossip exchange with neighbors
**end for**
**return** $\mathbf{Z}^{(K_{\mathrm{in}})}/T_{K_{\mathrm{in}}}(1/\lambda)$ ▷ $T_k$: Chebyshev polynomial of the first kind

The next lemma gives the contraction factor $\theta$ of one call of FastMix. The algorithm computes $p(W)\mathbf{X}$ for the normalized polynomial $p(t) = T_{K_{\mathrm{in}}}(t/\lambda)/T_{K_{\mathrm{in}}}(1/\lambda)$. This polynomial is bounded on the whole interval $[-\lambda, \lambda]$, so $W$ need not be positive semidefinite. Chebyshev filtering for accelerated consensus is standard (Montijano et al., 2013; Scaman et al., 2017); see also the two-step acceleration of Liu & Morse (2011). If $\lambda = 0$, one multiplication by $W$ averages exactly, and we set FastMix$(\mathbf{X}, K_{\mathrm{in}}) = W\mathbf{X}$. The number of inner rounds grows as $1/\sqrt{1-\lambda}$. This is the source of the factor in all communication bounds. The lemma is used in the consensus and tracking steps of Appendices D and E.

**Lemma 6** (accelerated consensus)**.** *Let $W$ be symmetric and doubly stochastic with $\lambda = \max\{|\lambda_2(W)|, |\lambda_n(W)|\} < 1$. Then* FastMix$(\mathbf{X}, K_{\text{in}})$ *preserves the row mean and*

$$\left\|\textsc{FastMix}(\mathbf{X}, K_{\text{in}}) - \mathbf{1}\bar{x}^\top\right\|_F \;\le\; \theta\,\left\|\mathbf{X} - \mathbf{1}\bar{x}^\top\right\|_F, \qquad \theta := \frac{1}{T_{K_{\text{in}}}(1/\lambda)} \;\le\; 2e^{-K_{\text{in}}\,\operatorname{arcosh}(1/\lambda)}.$$

*In particular $\theta \le \theta_0$ if $K_{\text{in}} \ge \lceil \log(2/\theta_0)/\sqrt{2(1-\lambda)} \rceil$.*

*Proof.* Let $J = \mathbf{1}\mathbf{1}^\top/n$ and $p$ be the polynomial above. Since $W\mathbf{1} = \mathbf{1}$ and $p(1) = 1$, we have $p(W)J = J$ and $Jp(W) = J$. The second identity shows that the row mean is preserved. The first gives $p(W)\mathbf{X} - \mathbf{1}\bar{x}^\top = p(W)(I - J)\mathbf{X}$. On the orthogonal complement of $\mathbf{1}$ every eigenvalue $\mu$ of $W$ satisfies $|\mu| \le \lambda$, hence $|T_{K_{\text{in}}}(\mu/\lambda)| \le 1$. Spectral calculus gives $\|p(W)(I - J)\|_2 \le \theta$, and the Frobenius bound follows. For $x > 1$,

$$T_{K_{\text{in}}}(x) = \cosh\bigl(K_{\text{in}} \operatorname{arcosh} x\bigr) \;\ge\; \tfrac{1}{2}e^{K_{\text{in}} \operatorname{arcosh} x}.$$

Let $s = \operatorname{arcosh}(1/\lambda)$, so that $\lambda = 1/\cosh s$. From $\cosh s \le e^{s^2/2}$ we get $\lambda \ge e^{-s^2/2} \ge 1 - s^2/2$, that is $s \ge \sqrt{2(1-\lambda)}$. Hence $\theta \le 2e^{-K_{\text{in}}\sqrt{2(1-\lambda)}} \le \theta_0$ for the stated $K_{\text{in}}$. □

The next lemma solves the scalar recursion (5). It turns a descent inequality in the clipped measure into an iteration count. Part (i) covers the convex case, and part (ii) adds strong convexity. All convex theorems end with an application of this lemma.

**Lemma 7** (two-phase recursion)**.** *Let $\mu_0 \ge \mu_1 \ge \cdots \ge 0$, $0 < q \le R/\tau$ and $K_1 := \lceil \frac{R}{q\tau} \log^+ \frac{\mu_0}{\tau R} \rceil$.*

*(i) If $\mu_{k+1} \le \mu_k - q\min\{\mu_k^2/R^2,\ \tau\mu_k/R\}$ for all $k$, then $\mu_k \le \max\{(1 - q\tau/R)^k\mu_0, \tau R\}$ for all $k$, $\mu_{K_1} \le \tau R$, and $\mu_k \le R^2/(q(k - K_1))$ for $k > K_1$. In particular, $\mu_K \le \varepsilon$ as soon as $K \ge K_1 + R^2/(q\varepsilon)$.*

*(ii) Let $\rho := \min\{\tau/R, 2\mu\}$ with $2q\mu \le 1$. If in addition $\mu_{k+1} \le (1 - q\rho)\mu_k$ for all $k$, then $\mu_k \le \tau R\,(1 - q\rho)^{k-K_1}$ for $k \ge K_1$. In particular, $\mu_K \le \varepsilon$ as soon as $K \ge K_1 + \frac{1}{q\rho}\log^+ \frac{\tau R}{\varepsilon}$.*

*Proof.* Let $k^* := \min\{k : \mu_k \le \tau R\}$.

*Step 1: clipped phase.* For $j < k^*$ we have $\mu_j > \tau R$. The minimum then equals $\tau\mu_j/R$, and $\mu_{j+1} \le (1 - q\tau/R)\mu_j$. The factor lies in $[0, 1)$ because $q\tau \le R$. Hence $\mu_k \le (1 - q\tau/R)^k\mu_0$ for $k \le k^*$. For $k > k^*$ monotonicity gives $\mu_k \le \mu_{k^*} \le \tau R$. Therefore

$$\mu_k \;\le\; \max\Bigl\{\Bigl(1 - \frac{q\tau}{R}\Bigr)^k\mu_0,\ \tau R\Bigr\} \qquad \text{for all } k. \tag{7}$$

*Step 2: length of the clipped phase.* By the choice of $K_1$, $(1 - q\tau/R)^{K_1}\mu_0 \le e^{-q\tau K_1/R}\mu_0 \le \tau R$. By (7), $\mu_{K_1} \le \tau R$. Hence $k^* \le K_1$, and $\mu_k \le \tau R$ for all $k \ge K_1$.

*Step 3: smooth phase, convex case.* For $j \ge k^*$ we have $\mu_j \le \tau R$, and the minimum equals $\mu_j^2/R^2$. If some $\mu_{j+1} = 0$, all later terms vanish. Otherwise, dividing by $\mu_j\mu_{j+1}$ and using $\mu_{j+1} \le \mu_j$,

$$\frac{1}{\mu_{j+1}} - \frac{1}{\mu_j} \;=\; \frac{\mu_j - \mu_{j+1}}{\mu_j\mu_{j+1}} \;\ge\; \frac{q\mu_j^2/R^2}{\mu_j^2} \;=\; \frac{q}{R^2}.$$

Summing from $k^*$ to $k - 1$ gives $\mu_k \le R^2/(q(k - k^*)) \le R^2/(q(k - K_1))$ for $k > K_1$. This quantity is at most $\varepsilon$ once $k \ge K_1 + R^2/(q\varepsilon)$.

*Step 4: smooth phase, strongly convex case.* The factor $1 - q\rho$ lies in $[0, 1)$ because $q\tau \le R$ and $2q\mu \le 1$. Iterating from $K_1$ and using Step 2 gives $\mu_k \le \tau R\,(1 - q\rho)^{k-K_1} \le \tau R\,e^{-q\rho(k-K_1)}$ for $k \ge K_1$. If $\varepsilon \ge \tau R$, then $\mu_{K_1} \le \varepsilon$. Otherwise the right-hand side is at most $\varepsilon$ once $k - K_1 \ge \frac{1}{q\rho}\log\frac{\tau R}{\varepsilon}$. □

The next lemma links the clipped measure to the gap. For a convex function with a bounded sublevel set, a large gap forces a large gradient. With a descent inequality this yields the recursion of Lemma 7.

**Lemma 8** (convexity lower bound)**.** *Let $h$ be convex with a minimizer $x^\star$, let $R_h := \sup\{\|x - x^\star\| : h(x) \le h(x^0)\} < \infty$, and let $h(x) \le h(x^0)$. Then $h(x) - h(x^\star) \le R_h \|\nabla h(x)\|$ and, for every $\tau > 0$,*

$$\Phi_\tau(\nabla h(x)) \;\ge\; \min\Big\{ \frac{(h(x) - h(x^\star))^2}{R_h^2},\; \frac{\tau\,(h(x) - h(x^\star))}{R_h} \Big\}.$$

*Proof.* Convexity gives $h(x) - h(x^\star) \le \langle \nabla h(x), x - x^\star \rangle \le \|\nabla h(x)\|\,\|x - x^\star\|$, and $\|x - x^\star\| \le R_h$ because $x$ lies in the sublevel set. The second claim follows because $t \mapsto t \min\{t, \tau\}$ is nondecreasing. □

The last tool is used only under strong convexity. It shows that the strong-convexity constant cannot exceed the modulus at the minimum, and it recalls the Polyak–Łojasiewicz inequality. The first fact keeps the linear factors of the convex theorems in $[0, 1)$. The second gives the linear rate of the smooth phase.

**Lemma 9** (strong convexity and the modulus)**.** *If $f$ is $\mu$-strongly convex and satisfies Assumption 1 with constants $(\widehat{H}_0, H_1)$, then $\mu \le \widehat{H}_0$ and $\|\nabla f(x)\|^2 \ge 2\mu F(x)$ for all $x$.*

*Proof.* At the minimizer $x^\star$ we have $\nabla f(x^\star) = 0$ and $F(x^\star) = 0$, so the modulus there equals $\widehat{H}_0$. For $\|s\| \le 1/\sqrt{H_1}$, strong convexity and Lemma 2 give

$$\frac{\mu}{2}\,\|s\|^2 \;\le\; f(x^\star + s) - f(x^\star) \;\le\; \frac{\widehat{H}_0}{2}\,\|s\|^2,$$

so $\mu \le \widehat{H}_0$. The second claim is the Polyak–Łojasiewicz inequality. It follows by minimizing the lower bound $f(y) \ge f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|y - x\|^2$ over $y$. □

### B.2 Proof of Proposition 2

Fix $i$ and points with $\|y - x\| \le 1/\sqrt{H_1}$. Averaging the gradient bound of Assumption 1 over $j \in \mathcal{C}_i$ gives

$$\|\nabla f_i(y) - \nabla f_i(x)\| \;\le\; \frac{1}{m}\sum_j \big(H_0 + H_1 \phi_{ij}(x)\big)\,\|y - x\|.$$

By the definition (3) of $\delta_i$,

$$\frac{1}{m}\sum_j \phi_{ij}(x) \;=\; f_i(x) - \frac{1}{m}\sum_j f_{ij}^{\inf} \;=\; \phi_i(x) + \delta_i.$$

Hence $f_i$ is $(H_0 + H_1\delta_i, H_1)$-smooth. The same computation over all pairs $(i, j)$ uses $\frac{1}{nm}\sum_{i,j}\phi_{ij}(x) = F(x) + \delta$ and shows that $f$ is $(H_0 + H_1\delta, H_1)$-smooth.

*The defect cannot be removed.* Let $f_{1,2}(x) = \cosh(x \mp 1) - 1$. Then $f''_{1,2} = 1 + \phi_{1,2}$. The average is $\bar{f}(x) = \cosh 1 \cdot \cosh x - 1$ with $\delta = \cosh 1 - 1$, and

$$\bar{f}''(x) \;=\; \cosh 1 \cdot \cosh x \;=\; (1 + \delta) + \bar{\phi}(x)$$

at every point. This is an equality for the pointwise Hessian modulus. In the gradient form of Assumption 1 each $f_{1,2}$ is $(e, e)$-smooth (proof of Proposition 3 in Appendix D). Hence the example matches the closure constants up to the factor $e$. □

### B.3 Relation to the gradient-based class

*Remark* 1. Every $(L_0, L_1)$-smooth function is $(3L_0, 9L_1^2)$-smooth in the sense of Assumption 1. The inclusion is strict.

*Proof.* Let $\big\|\nabla^2 h(x)\big\| \le L_0 + L_1\|\nabla h(x)\|$ for all $x$, and write $g = \nabla h(x)$. Along a segment $y = x + tv$ with a unit vector $v$, the function $u(t) = \|\nabla h(x + tv)\| + L_0/L_1$ satisfies $u' \le L_1 u$ almost everywhere. Grönwall's inequality gives $u(t) \le u(0)e^{L_1 t}$ (cf. Zhang et al., 2020a, Lemma A.2), so the Hessian along a segment of length $r$ is at most $(L_0 + L_1\|g\|)e^{L_1 r}$.

*Gradient bound.* Take $s = -tg/\|g\|$ with $t = \|g\|/(2(L_0 + L_1\|g\|)) \le 1/(2L_1)$, so that $e^{L_1 t} \le e^{1/2} \le 1.65$. Since $h(x+s) \ge h^{\inf}$, the second-order Taylor bound gives

$$\phi_h(x) \;\ge\; h(x) - h(x+s) \;\ge\; t\,\|g\| - \frac{1.65}{2}\,t^2(L_0 + L_1\|g\|) \;=\; \Big(\frac{1}{2} - \frac{1.65}{8}\Big)\frac{\|g\|^2}{L_0 + L_1\|g\|} \;\ge\; \frac{\|g\|^2}{4(L_0 + L_1\|g\|)};$$

compare Gorbunov et al. (2025, Lemma 2.2). Solving the quadratic inequality in $\|g\|$ gives $\|g\| \le 4L_1\phi_h(x) + 2\sqrt{L_0\phi_h(x)}$.

*Hessian bound.* Take $r \le 1/(3L_1)$, so that $e^{L_1 r} \le 1.4$. For $y$ on the segment,

$$\big\|\nabla^2 h(y)\big\| \;\le\; 1.4\,(L_0 + L_1\|g\|) \;\le\; 1.4\,L_0 + 5.6\,L_1^2\phi_h(x) + 2.8\,L_1\sqrt{L_0\phi_h(x)} \;\le\; 2.8\,L_0 + 7\,L_1^2\phi_h(x),$$

where the last step uses $L_1\sqrt{L_0\phi_h} \le L_0/2 + L_1^2\phi_h/2$. Integrating the Hessian along the segment gives the gradient bound with $H_0 = 3L_0$, $H_1 = 9L_1^2$ and radius $1/\sqrt{H_1} = 1/(3L_1)$.

*Strictness.* Let $h(x) = x^2 + 2\sin(x^2)$. Then $h \ge x^2 - 2$ and $h(0) = 0$, so $x^2 \le \phi_h(x) + 2$. The derivatives are

$$h'(x) = 2x\big(1 + 2\cos x^2\big), \qquad h''(x) = 2 + 4\cos x^2 - 8x^2\sin x^2,$$

so $|h''(x)| \le 6 + 8x^2$. For $|z - x| \le \frac{1}{4}$ we have $z^2 \le 2x^2 + \frac{1}{8}$ and

$$|h''(z)| \;\le\; 7 + 16x^2 \;\le\; 39 + 16\,\phi_h(x).$$

Integrating along the segment shows that $h$ is $(39, 16)$-smooth, with radius $\frac{1}{4} = 1/\sqrt{16}$. At the critical points $\cos x^2 = -\frac{1}{2}$ we have $h'(x) = 0$ and $|h''(x)| = 4\sqrt{3}\,x^2$, which is unbounded. Hence no pair $(L_0, L_1)$ satisfies $|h''| \le L_0 + L_1|h'|$. □

### B.4 Proof of Proposition 1

*Proof.* Let $\lambda_2(L)$ denote the second-smallest eigenvalue of a graph Laplacian. Because $G$ is $d$-regular,

$$\Delta_G = 1 - \lambda_2(W_G) = \frac{\lambda_2(L_G)}{2d}.$$

The maximum degree of $G_{\text{flat}}$ is $d + m$, so $W_{\text{flat}} = I - L_{\text{flat}}/[2(d+m)]$ is symmetric, doubly stochastic, and positive semidefinite, with

$$\Delta_{\text{flat}} = \frac{\lambda_2(L_{\text{flat}})}{2(d+m)}.$$

Let $z \perp \mathbf{1}$ be a unit Fiedler vector of $L_G$. Construct $\widetilde{z} \in \mathbb{R}^{n(m+1)}$ by assigning the value $z_i$ to hub $i$ and to all $m$ leaves attached to it. Then $\widetilde{z} \perp \mathbf{1}$,

$$\|\widetilde{z}\|^2 = m + 1, \qquad \widetilde{z}^\top L_{\text{flat}}\widetilde{z} = z^\top L_G z = \lambda_2(L_G),$$

because every leaf edge has zero difference. The Rayleigh principle yields

$$\lambda_2(L_{\text{flat}}) \le \frac{\lambda_2(L_G)}{m+1}, \qquad \Delta_{\text{flat}} \le \frac{d}{(d+m)(m+1)}\Delta_G.$$

For a fixed target contraction, Lemma 6 prescribes a degree of order $\Delta^{-1/2}$. Thus the ratio of the certified round counts is at least $\Omega(\sqrt{(d+m)(m+1)/d})$. Multiplying by the edge-count ratio $(nm + nd/2)/(nd/2) = (d + 2m)/d$ proves the message-count claim. The result concerns these explicit matrices and does not preclude a better flat weighting or a different consensus method. □

## C Proofs for the federated level

This appendix proves Lemma 1 and Theorems 1 and 2 of Section 4, and the nonconvex Theorem 6.

Throughout this appendix the hub $i$ is fixed. At the current iterate $x$ the clients send $u_j = \mathrm{clip}_\tau(g_j)$, where $g_j$ is a mini-batch gradient, and the hub forms $d = \frac{1}{s}\sum_{j\in\mathcal{S}} u_j$. We write $D_j := \|g_{ij} - g_i\|$.

### C.1 ALIGNMENT

The next lemma is the key step at the federated level. It shows that an average of individually clipped vectors stays aligned with $u$ when the vectors deviate from $u$ only on average. Under Assumption 2 it applies with $u = g_i$ and $v_j = g_{ij}$. The proof counts good and bad clients: the bad ones are few, and clipping caps the damage from each of them.

**Lemma 10** (deterministic alignment)**.** *Let $u \neq 0$ and $v_1, \dots, v_m$ satisfy $\frac{1}{m}\sum_j \|v_j - u\|^2 \le \nu^2 \|u\|^2$ with $\nu \le \frac{1}{6}$, and let $u_j = \mathrm{clip}_\tau(v_j)$. Then*

$$\frac{1}{m}\sum_j \langle u, u_j\rangle \;\ge\; \frac{1}{8}\,\Phi_\tau(u), \qquad \frac{1}{m}\sum_j \|u_j\|^2 \;\le\; \min\Big\{\tau^2, \frac{3}{2}\|u\|^2\Big\} \;\le\; \frac{3}{2}\,\Phi_\tau(u).$$

*Proof.* Write $D_j = \|v_j - u\|$. Call index $j$ *good* if $D_j \le \frac{1}{2}\|u\|$ and *bad* otherwise. Let $B$ be the set of bad indices and $k := |B|$. Markov's inequality applied to $\sum_j D_j^2 \le m\nu^2\|u\|^2$ gives

$$\frac{k}{m} \;\le\; 4\nu^2 \;\le\; \frac{1}{9}.$$

For a good index, Lemma 4 with ratio $\frac{1}{2}$ gives $\langle u, u_j\rangle \ge \frac{1}{3}\Phi_\tau(u)$. The good indices therefore contribute at least $(1 - \frac{k}{m})\frac{1}{3}\Phi_\tau(u) \ge \frac{8}{27}\Phi_\tau(u)$ to the average.

*Large gradients*, $\|u\| \ge \tau$. Here $\Phi_\tau(u) = \tau\|u\|$. For a bad index, $\langle u, u_j\rangle \ge -\|u\|\,\|u_j\| \ge -\tau\|u\|$. Hence

$$\frac{1}{m}\sum_j \langle u, u_j\rangle \;\ge\; \Big(\frac{8}{27} - \frac{1}{9}\Big)\Phi_\tau(u) \;=\; \frac{5}{27}\Phi_\tau(u) \;\ge\; \frac{1}{8}\Phi_\tau(u).$$

*Small gradients*, $\|u\| < \tau$. Here $\Phi_\tau(u) = \|u\|^2$. For a bad index, $\langle u, u_j\rangle \ge -\|u\|\,\|v_j\| \ge -\|u\|\,(\|u\| + D_j)$. By Cauchy–Schwarz, $\sum_{j\in B} D_j \le \sqrt{k}\,(\sum_j D_j^2)^{1/2} \le \sqrt{k}\,\nu\sqrt{m}\,\|u\|$. Therefore

$$\frac{1}{m}\sum_{j\in B}\langle u, u_j\rangle \;\ge\; -\|u\|^2\Big(\frac{k}{m} + \nu\sqrt{\frac{k}{m}}\Big) \;\ge\; -6\nu^2\|u\|^2 \;\ge\; -\frac{1}{6}\|u\|^2,$$

and the total is at least $(\frac{8}{27} - \frac{1}{6})\|u\|^2 \ge \frac{1}{8}\|u\|^2$.

*Second moment.* We have $\|u_j\|^2 \le \min\{\tau^2, \|v_j\|^2\}$. By the triangle inequality and Cauchy–Schwarz,

$$\frac{1}{m}\sum_j \|v_j\|^2 \;\le\; \|u\|^2 + 2\|u\|\,\frac{1}{m}\sum_j D_j + \frac{1}{m}\sum_j D_j^2 \;\le\; (1+\nu)^2\|u\|^2 \;\le\; \frac{3}{2}\|u\|^2.$$

Finally, $\min\{\tau^2, \frac{3}{2}\|u\|^2\} \le \frac{3}{2}\min\{\tau^2, \|u\|^2\} \le \frac{3}{2}\Phi_\tau(u)$. □

The margin in the small-gradient case is $\frac{8}{27} - \frac{1}{6} \approx 0.13$ against $\frac{1}{8}$. The condition $\nu \le \frac{1}{6}$ cannot be relaxed within this argument.

The next lemma adds client sampling and mini-batch noise to Lemma 10. The same conclusions hold in expectation, with smaller constants. A batch of size proportional to $\rho_\ell$ suffices, and the cohort size is arbitrary. It gives the one-round descent in Theorem 6 and in Proposition 4.

**Lemma 11** (stochastic alignment)**.** *Let Assumption 2 hold with $\nu \le \frac{1}{12}$, and let Assumption 4 hold. Let every sampled client use a mini-batch of size $b \ge 2304(\rho_\ell - 1)$, so that $\mathbb{E} g_j = g_{ij}$ and $\mathbb{E}\|g_j - g_{ij}\|^2 \le \eta^2\|g_{ij}\|^2$ with $\eta \le \frac{1}{48}$. Then for a cohort drawn uniformly without replacement,*

$$\mathbb{E}\,\langle g_i, d\rangle \;\ge\; \frac{1}{16}\,\Phi_\tau(g_i), \qquad \mathbb{E}\,\|d\|^2 \;\le\; 2\,\Phi_\tau(g_i).$$

*Proof.* Every client is in the cohort with the same probability. Hence $\mathbb{E}\,\langle g_i, d\rangle = \frac{1}{m}\sum_{j\in\mathcal{C}_i}\mathbb{E}\,\langle g_i, \mathrm{clip}_\tau(g_j)\rangle$, and the same holds for $\mathbb{E}\frac{1}{s}\sum_{j\in\mathcal{S}}\|\mathrm{clip}_\tau(g_j)\|^2$.

*Small gradients*, $\|g_i\| < \tau$. Non-expansiveness gives $\langle g_i, \mathrm{clip}_\tau(g_j)\rangle \geq \langle g_i, \mathrm{clip}_\tau(g_{ij})\rangle - \|g_i\|\,\|g_j - g_{ij}\|$. Moreover $\mathbb{E}\,\|g_j - g_{ij}\| \leq \eta\,\|g_{ij}\|$ and $\frac{1}{m}\sum_j \|g_{ij}\| \leq (1+\nu)\,\|g_i\|$. Lemma 10 then gives

$$\mathbb{E}\,\langle g_i, d\rangle \;\geq\; \frac{1}{8}\,\|g_i\|^2 - \eta(1+\nu)\,\|g_i\|^2 \;\geq\; \Big(\frac{1}{8} - \frac{13}{12\cdot 48}\Big)\,\|g_i\|^2 \;\geq\; \frac{1}{16}\,\|g_i\|^2 .$$

*Large gradients*, $\|g_i\| \geq \tau$. Call client $j$ good if $D_j \leq \frac{1}{4}\,\|g_i\|$. By Markov's inequality the fraction of bad clients is at most $16\nu^2 \leq \frac{1}{9}$. By Chebyshev's inequality the event $\|g_j - g_{ij}\| \leq \frac{1}{4}\,\|g_{ij}\|$ has probability at least $1 - 16\eta^2 \geq 1 - \frac{1}{144}$. On this event, for a good client, $\|g_{ij}\| \leq \frac{5}{4}\,\|g_i\|$ and

$$\|g_j - g_i\| \;\leq\; \tfrac{1}{4}\,\|g_{ij}\| + D_j \;\leq\; \tfrac{5}{16}\,\|g_i\| + \tfrac{4}{16}\,\|g_i\| \;=\; \tfrac{9}{16}\,\|g_i\| .$$

Lemma 4 with ratio $\frac{9}{16}$ gives $\langle g_i, \mathrm{clip}_\tau(g_j)\rangle \geq \frac{7}{25}\tau\,\|g_i\|$. In all other cases $\langle g_i, \mathrm{clip}_\tau(g_j)\rangle \geq -\tau\,\|g_i\|$. Hence

$$\mathbb{E}\,\langle g_i, d\rangle \;\geq\; \Big[\frac{8}{9}\Big(\frac{143}{144}\cdot\frac{7}{25} - \frac{1}{144}\Big) - \frac{1}{9}\Big]\tau\,\|g_i\| \;\geq\; 0.12\,\tau\,\|g_i\| \;\geq\; \frac{1}{16}\Phi_\tau(g_i).$$

*Second moment.* By Jensen's inequality and concavity of $t \mapsto \min\{\tau^2, t\}$,

$$\mathbb{E}\,\|d\|^2 \;\leq\; \min\Big\{\tau^2,\; \frac{1}{m}\sum_j \mathbb{E}\,\|g_j\|^2\Big\} \;\leq\; \min\Big\{\tau^2,\; (1+\eta^2)\,\frac{3}{2}\,\|g_i\|^2\Big\} \;\leq\; 2\,\Phi_\tau(g_i),$$

where $\mathbb{E}\,\|g_j\|^2 = \|g_{ij}\|^2 + \mathbb{E}\,\|g_j - g_{ij}\|^2$ and the second step of Lemma 10 were used. □

### C.2 Proof of Theorem 6

*Idea.* The descent lemma holds with the modulus at the current random point. We stop the process when the gap leaves a sublevel set, where the modulus is bounded. A supermartingale bounds the probability of stopping early.

*Notation.* Let $F^k := \phi_i(x^k)$, $\Phi^k := \Phi_\tau(\nabla f_i(x^k))$ and $M_k := M_i(x^k)$. Fix $\beta \in (0,1)$ and define the stopping time

$$T := \min\{k :\; F^k > 2F_0/\beta\} \qquad (T = \infty \text{ if the set is empty}).$$

For $k < T$ we have $M_k \leq M_f = H_0 + H_1(\delta_i + 2F_0/\beta)$. The event $\{k < T\}$ is determined by the history up to round $k$.

*Step 1: one round.* The step has length $\gamma\,\|d^k\| \leq \gamma\tau \leq 1/\sqrt{H_1}$. Lemma 2 for $f_i$ and Lemma 11 give, on $\{k < T\}$,

$$\mathbb{E}_k F^{k+1} \;\leq\; F^k - \gamma\,\mathbb{E}_k\,\langle g_i^k, d^k\rangle + \frac{\gamma^2 M_k}{2}\,\mathbb{E}_k\,\|d^k\|^2 \;\leq\; F^k - \gamma\Phi^k\Big(\frac{1}{16} - \gamma M_k\Big) \;\leq\; F^k - \frac{\gamma}{32}\Phi^k,$$

where the last step uses $\gamma M_f \leq \frac{1}{32}$.

*Step 2: supermartingale.* Let

$$\Psi^k \;:=\; F^{k\wedge T} + \frac{\gamma}{32}\sum_{j<k\wedge T}\Phi^j.$$

It is nonnegative and $\Psi^0 = F_0$. On $\{k < T\}$ Step 1 gives $\mathbb{E}_k\Psi^{k+1} \leq \Psi^k$. On $\{k \geq T\}$ the process is constant. Hence $(\Psi^k)$ is a supermartingale.

*Step 3: two bad events.* On $\{T \leq K\}$ we have $\Psi^T \geq F^T > 2F_0/\beta$. Ville's maximal inequality gives

$$\mathbb{P}(T \leq K) \;\leq\; \mathbb{P}\Big(\sup_k \Psi^k \geq \frac{2F_0}{\beta}\Big) \;\leq\; \frac{\beta}{2}.$$

Moreover $\mathbb{E}\Psi^K \leq F_0$, so $\mathbb{E}\big[\sum_{j<K\wedge T}\Phi^j\big] \leq 32F_0/\gamma$. Markov's inequality gives

$$\mathbb{P}\Big(\sum_{j<K\wedge T}\Phi^j \;\geq\; \frac{64F_0}{\beta\gamma}\Big) \;\leq\; \frac{\beta}{2}.$$

*Step 4: conclusion.* Outside both events, $T > K$ and $\sum_{j<K}\Phi^j < 64F_0/(\beta\gamma)$. Hence $\min_{k<K}\Phi^k \leq 64F_0/(\beta\gamma K)$ with probability at least $1-\beta$. If $\varepsilon \leq \tau$ and $\Phi_\tau(g) \leq \varepsilon^2$, then $\|g\| \leq \varepsilon$. □

### C.3 Proof of Theorem 1

With exact gradients and all clients participating, the iteration is deterministic. Lemma 10 applies at every iterate with $u = g_i^k$ and $v_j = g_{ij}^k$. Assumption 2 implies $\nabla f_{ij}(x) = 0$ for all $j$ whenever $\nabla f_i(x) = 0$, since the averaged deviation must vanish. For convex clients every minimizer of $f_i$ therefore minimizes each $f_{ij}$, and $\delta_i = 0$.

*Proof of part (a).* Suppose $F^k \le F_0$, so that $M_k \le M_f$. The step has length at most $\gamma\tau \le 1/(6\sqrt{H_1})$. Lemma 2 and Lemma 10 give

$$F^{k+1} \;\le\; F^k - \gamma\Phi^k\Big(\frac{1}{8} - \frac{3}{4}\gamma M_k\Big) \;\le\; F^k - \frac{\gamma}{16}\Phi^k,$$

because $\gamma M_k \le \frac{1}{16}$. In particular $F^{k+1} \le F^k \le F_0$, and by induction this holds for all $k$. Every iterate lies in the sublevel set, and Lemma 8 gives $\Phi^k \ge \min\{(F^k)^2/R^2, \tau F^k/R\}$. The sequence $\mu_k := F^k$ satisfies Lemma 7(i) with $q = \gamma/16$. The condition $q \le R/\tau$ holds because $\gamma \le R/\tau$. The lemma gives $F^K \le \varepsilon$ once $K \ge \lceil \frac{16R}{\gamma\tau}\log^+\frac{F_0}{\tau R}\rceil + \frac{16R^2}{\gamma\varepsilon}$, since $R/(q\tau) = 16R/(\gamma\tau)$ and $R^2/q = 16R^2/\gamma$. This is implied by the claimed condition. □

*Proof of part (b).* Strong convexity bounds the sublevel set: $\frac{\mu}{2}\|x - x^\star\|^2 \le f_i(x) - f_i^\star \le F_0$, so $R \le \sqrt{2F_0/\mu}$. The descent inequality of the previous proof holds. If $\|g_i^k\| \ge \tau$, then $\Phi^k = \tau\|g_i^k\| \ge \tau F^k/R$ by Lemma 8, and $F^{k+1} \le (1 - \frac{\gamma\tau}{16R})F^k$. If $\|g_i^k\| < \tau$, then $\Phi^k = \|g_i^k\|^2 \ge 2\mu F^k$ by Lemma 9 applied to $f_i$, and $F^{k+1} \le (1 - \frac{\gamma\mu}{8})F^k$. With $q = \gamma/16$ both cases give $F^{k+1} \le (1 - q\min\{\tau/R, 2\mu\})F^k$. Moreover $2q\mu = \gamma\mu/8 \le \gamma M_f \le 1$ by Lemma 9. The recursion of part (a) also holds, so Lemma 7(ii) applies with $\rho = \min\{\tau/R, 2\mu\}$. The claim follows from $\frac{1}{q\rho} = \max\{\frac{16R}{\gamma\tau}, \frac{8}{\gamma\mu}\}$. □

### C.4 Heavy-tailed noise: proof of Theorem 2

*Remark* 2 (scope). All transmitted vectors are clipped, so their second moments exist for every $\alpha \in (1, 2]$. The bias of clipping is controlled only when the radius is at least twice the mean (Lemma 12). We therefore prove the heavy-tailed guarantee when the radius dominates the client gradients. The large-gradient heavy-tailed regime is open.

The next lemma quantifies what clipping does to a heavy-tailed vector whose mean lies well inside the ball. The bias decays as $\tau^{1-\alpha}$, and the variance is finite although the noise may have none. The condition $\tau \ge 2\|\mu\|$ is where the dominated-radius assumption of Theorem 2 enters.

**Lemma 12** (clipping bias). *Let $\mathbb{E}g = \mu$, $\mathbb{E}\|g - \mu\|^\alpha \le \sigma^\alpha$ with $\alpha \in (1, 2]$, and $\tau \ge 2\|\mu\|$. Then*

$$\|\mathbb{E}\,\mathrm{clip}_\tau(g) - \mu\| \;\le\; \frac{8\sigma^\alpha}{\tau^{\alpha-1}}, \qquad \mathbb{E}\|\mathrm{clip}_\tau(g) - \mathbb{E}\,\mathrm{clip}_\tau(g)\|^2 \;\le\; 17\,\sigma^\alpha\tau^{2-\alpha}.$$

*Proof.* Let $A := \{\|g - \mu\| \le \tau/2\}$. On $A$ we have $\|g\| \le \tau$, so $\mathrm{clip}_\tau(g) = g$. On the complement $A^c$ we have $\|g - \mu\| > \tau/2 \ge \|\mu\|$, hence

$$\|\mathrm{clip}_\tau(g) - g\| \;\le\; \tau + \|\mu\| + \|g - \mu\| \;\le\; 4\,\|g - \mu\|\,.$$

Markov's inequality gives $\mathbb{P}(A^c) \le (2/\tau)^\alpha\sigma^\alpha$. Hölder's inequality with exponents $\alpha$ and $\alpha/(\alpha - 1)$ gives

$$\|\mathbb{E}\,\mathrm{clip}_\tau(g) - \mu\| \;\le\; 4\,\mathbb{E}\big[\|g - \mu\|\,\mathbf{1}_{A^c}\big] \;\le\; 4\sigma\,\mathbb{P}(A^c)^{(\alpha-1)/\alpha} \;\le\; \frac{2^{\alpha+1}\sigma^\alpha}{\tau^{\alpha-1}} \;\le\; \frac{8\sigma^\alpha}{\tau^{\alpha-1}}.$$

For the second moment, the variance is at most $\mathbb{E}\|\mathrm{clip}_\tau(g) - \mu\|^2$. On $A$ this is $\|g - \mu\|^2 \le (\tau/2)^{2-\alpha}\|g - \mu\|^\alpha$. On $A^c$ both $\mathrm{clip}_\tau(g)$ and $\mu$ lie in the ball of radius $\tau$, so the squared distance is at most $4\tau^2$. Hence

$$\mathbb{E}\|\mathrm{clip}_\tau(g) - \mu\|^2 \;\le\; \sigma^\alpha\tau^{2-\alpha} + 4\tau^2\cdot\frac{2^\alpha\sigma^\alpha}{\tau^\alpha} \;\le\; 17\,\sigma^\alpha\tau^{2-\alpha}.$$

□

*Proof of Theorem 2. Idea.* The radius exceeds every client gradient, so the method is always in the small-gradient regime. The bias and the variance of the clipped messages are then small, and the stopping-time argument of Theorem 6 applies with an additive error.

*Notation.* Let $F^k := \phi_i(x^k)$, $\Phi^k := \left\|\nabla f_i(x^k)\right\|^2$, and let $T$ be the first $k$ with $F^k > 4F_0/\beta$. On $\{k < T\}$ we have $M_i(x^k) \le M_f$ and $\|g_{ij}\| \le \tau/2$ for every client, hence $\|g_i\| \le \tau/2$ and $\Phi_\tau(g_i) = \Phi^k$.

*Step 1: bias and variance.* The mini-batch mean has $\alpha$-th central moment at most $2\sigma_\alpha^\alpha b^{1-\alpha}$ by the Hilbert-space von Bahr–Esseen inequality (Pinelis, 2022). Lemma 12 with $\mu = g_{ij}$ gives $\mathbb{E}_k \operatorname{clip}_\tau(g_j) = g_{ij} + \beta_j$ with

$$\|\beta_j\| \;\le\; \frac{16\sigma_\alpha^\alpha}{b^{\alpha-1}\tau^{\alpha-1}} \;\le\; \frac{\sqrt{\varepsilon}}{2}, \qquad \mathbb{E}_k \left\|\operatorname{clip}_\tau(g_j) - \mathbb{E}_k \operatorname{clip}_\tau(g_j)\right\|^2 \;\le\; \frac{34\,\sigma_\alpha^\alpha \tau^{2-\alpha}}{b^{\alpha-1}} \;\le\; \frac{\varepsilon}{2}.$$

Both bounds use the batch condition.

*Step 2: one round.* Each client is in the cohort with the same probability, so $\mathbb{E}_k d = g_i + \bar\beta$ with $\left\|\bar\beta\right\| \le \sqrt{\varepsilon}/2$. Hence

$$\mathbb{E}_k \langle g_i, d\rangle \;\ge\; \|g_i\|^2 - \|g_i\| \frac{\sqrt{\varepsilon}}{2} \;\ge\; \frac{15}{16}\Phi^k - \varepsilon.$$

For the second moment, $\mathbb{E}_k \|d\|^2 \le \frac{1}{m}\sum_j \mathbb{E}_k \left\|\operatorname{clip}_\tau(g_j)\right\|^2$. Each term is at most $(\|g_{ij}\| + \sqrt{\varepsilon}/2)^2 + \varepsilon/2 \le 2\|g_{ij}\|^2 + \varepsilon$. With $\frac{1}{m}\sum_j \|g_{ij}\|^2 \le (1+\nu^2)\|g_i\|^2$ this gives $\mathbb{E}_k \|d\|^2 \le \frac{5}{2}\Phi^k + \varepsilon$. The step has length at most $\gamma\tau \le 1/\sqrt{H_1}$. Lemma 2 and $\gamma M_f \le \frac{1}{32}$ give, on $\{k < T\}$,

$$\mathbb{E}_k F^{k+1} \;\le\; F^k - \gamma\Phi^k\Big(\frac{15}{16} - \frac{5}{128}\Big) + \gamma\varepsilon\Big(1 + \frac{1}{64}\Big) \;\le\; F^k - \frac{\gamma}{2}\Phi^k + 2\gamma\varepsilon.$$

*Step 3: supermartingale.* Let

$$\Psi^k \;:=\; F^{k\wedge T} + \frac{\gamma}{2}\sum_{j<k\wedge T} \Phi^j + 2\gamma\varepsilon\,(K - k\wedge T), \qquad k \le K.$$

It is nonnegative, $\Psi^0 = F_0 + 2\gamma\varepsilon K = 2F_0$, and Step 2 shows that it is a supermartingale. On $\{T \le K\}$ we have $\Psi^T \ge F^T > 4F_0/\beta = \frac{2}{\beta}\Psi^0$. Ville's inequality gives $\mathbb{P}(T \le K) \le \beta/2$. Markov's inequality applied to $\mathbb{E}\Psi^K \le 2F_0$ gives $\sum_{j<K\wedge T}\Phi^j < 8F_0/(\beta\gamma)$ with probability at least $1 - \beta/2$. Outside both events, $\min_{k<K}\Phi^k \le 8F_0/(\beta\gamma K)$. □

## D PROOFS FOR THE DECENTRALIZED LEVEL

This appendix proves Proposition 3 and Theorem 3 of Section 5, and the nonconvex Theorem 7.

**Setting and notation.** The hubs use exact gradients $g_i^k = \nabla f_i(x_i^k)$. We write $\bar x^k$ for the row mean of $\mathbf{X}^k$, $g^k := \nabla f(\bar x^k)$ and $\bar g^k := \frac{1}{n}\sum_i g_i^k$. The consensus and tracking errors are

$$\varepsilon_x^k := \left\|\mathbf{X}^k - \mathbf{1}(\bar x^k)^\top\right\|_F, \qquad \varepsilon_v^k := \left\|\mathbf{V}^k - \mathbf{1}(\bar v^k)^\top\right\|_F.$$

The applied directions are $u_i^k := \operatorname{clip}_\tau(v_i^k)$ with mean $\bar u^k$, so that $\bar x^{k+1} = \bar x^k - \gamma\bar u^k$. FastMix preserves row means, and the tracker update adds $\mathbf{G}^{k+1} - \mathbf{G}^k$. Hence the tracking invariant $\bar v^k = \bar g^k$ holds for all $k$. Finally $F^k := F(\bar x^k)$, $M_k := M(\bar x^k)$, $\Phi^k := \Phi_\tau(g^k)$ and $m^k := \min\{\left\|g^k\right\|, \tau\}$. The constant $M = \widehat H_0 + H_1 F_0$ bounds $M_k$ whenever $F^k \le F_0$. Since the clients do not appear here, $\delta$ may be replaced by $f^\star - \frac{1}{n}\sum_i f_i^{\inf} \le \delta$.

The certificates below bound the network errors by multiples of $m^k = \min\{\left\|g^k\right\|, \tau\}$. The next lemma shows that one step changes the gradient norm only by a small factor. Hence $m^{k+1}$ is comparable to $m^k$, and the certificate can be carried from round $k$ to round $k+1$.

**Lemma 13** (gradient-norm stability). *If $\gamma\left\|\bar u^k\right\| \le 1/\sqrt{H_1}$, then $\left|\left\|g^{k+1}\right\| - \left\|g^k\right\|\right| \le \gamma M_k \left\|\bar u^k\right\|$.*

*Proof.* This is the gradient bound of Assumption 1 for $f$ at the base point $\bar{x}^k$ (Proposition 2). □

The next lemma collects the three recursions of gradient tracking. Part (i) compares the mean of the hub gradients with the gradient at the mean point. Part (ii) propagates the consensus error through one clipped step and one call of FASTMIX. Part (iii) propagates the tracking error. The local moduli differ across hubs, and part (iii) shows that their sum exceeds $nM_k$ by at most the factor $\frac{5}{4}$. This is where the condition $\theta n \le \frac{1}{8}$ and the factor $\log n$ in $K_{\mathrm{in}}$ come from.

**Lemma 14** (error recursions). *Let* $e^k := \|\bar{g}^k - g^k\|$ *and* $S^k := \|\mathrm{clip}_\tau(\mathbf{V}^k)\|_F$.

*(i) If* $\varepsilon_x^k \le 1/\sqrt{H_1}$, *then* $e^k \le M_k\, \varepsilon_x^k$.

*(ii)* $\varepsilon_x^{k+1} \le \theta(\varepsilon_x^k + \gamma S^k)$ *and* $S^k \le \min\{\sqrt{n}\tau,\ \sqrt{n}(\|g^k\| + e^k) + \varepsilon_v^k\}$.

*(iii) Let* $D^k := \max_i \|x_i^{k+1} - x_i^k\|$. *If* $\varepsilon_x^k \le 1/(18\sqrt{H_1})$ *and* $D^k \le 1/(3\sqrt{H_1})$, *then*

$$\varepsilon_v^{k+1} \ \le\ \theta\Big(\varepsilon_v^k + \frac{5}{4}\, nM_k\, D^k\Big). \tag{8}$$

*Proof. (i)* Let $L_i := M_i(\bar{x}^k)$. The gradient bound for $f_i$ at base point $\bar{x}^k$ gives $\|\nabla f_i(x_i^k) - \nabla f_i(\bar{x}^k)\| \le L_i \|x_i^k - \bar{x}^k\|$. Each row norm is at most $\varepsilon_x^k$, so $e^k \le \big(\frac{1}{n}\sum_i L_i\big)\varepsilon_x^k$. The identity $\frac{1}{n}\sum_i(\delta_i + \phi_i(x)) = F(x) + \delta$ gives

$$\frac{1}{n}\sum_i L_i \ =\ H_0 + H_1\big(F^k + \delta\big) \ =\ M_k.$$

*(ii)* The matrix $\mathbf{X}^k - \gamma\,\mathrm{clip}_\tau(\mathbf{V}^k)$ deviates from its mean by at most $\varepsilon_x^k + \gamma S^k$ in Frobenius norm, and FASTMIX contracts this deviation by $\theta$. Each row satisfies $\|u_i^k\| \le \min\{\tau, \|v_i^k\|\}$. Moreover $\|\mathbf{V}^k\|_F \le \sqrt{n}\,\|\bar{v}^k\| + \varepsilon_v^k$ and $\|\bar{v}^k\| = \|\bar{g}^k\| \le \|g^k\| + e^k$.

*(iii)* Centering does not increase the Frobenius norm, so $\varepsilon_v^{k+1} \le \theta(\varepsilon_v^k + \|\mathbf{G}^{k+1} - \mathbf{G}^k\|_F)$. Let $L_i' := M_i(x_i^k)$. The gradient bound at base point $x_i^k$ and the inequality $\|\cdot\|_2 \le \|\cdot\|_1$ give

$$\|\mathbf{G}^{k+1} - \mathbf{G}^k\|_F \ \le\ \Big(\sum_i {L_i'}^2\, \|x_i^{k+1} - x_i^k\|^2\Big)^{1/2} \ \le\ \Big(\sum_i L_i'\Big) D^k.$$

It remains to show $\sum_i L_i' \le \frac{5}{4}\sum_i L_i = \frac{5}{4} nM_k$. Write $r := \|x_i^k - \bar{x}^k\| \le \varepsilon_x^k \le 1/(18\sqrt{H_1})$. Lemma 2 for $f_i$ at base point $\bar{x}^k$ gives

$$L_i' - L_i \ =\ H_1\big(\phi_i(x_i^k) - \phi_i(\bar{x}^k)\big) \ \le\ H_1\,\|\nabla f_i(\bar{x}^k)\|\, r + \frac{H_1 L_i}{2}\, r^2.$$

By Lemma 3 for $f_i$, and since $H_1\phi_i(\bar{x}^k) \le L_i$ and $\sqrt{2L_i H_1\phi_i} \le L_i + H_1\phi_i \le 2L_i$,

$$H_1\,\|\nabla f_i(\bar{x}^k)\|\, r \ \le\ \frac{1}{18}\Big(\sqrt{2L_i H_1\phi_i} + H_1\phi_i + \frac{L_i}{2}\Big) \ \le\ \frac{7}{36} L_i.$$

The second term is at most $L_i/648$. Summing over $i$ gives $\sum_i L_i' \le (1 + \frac{7}{36} + \frac{1}{648})\sum_i L_i \le \frac{5}{4}\sum_i L_i$. □

### D.1 PROOF OF THEOREM 7

*Idea.* We show by induction that consensus and tracking errors stay proportional to the current gradient scale $m^k$. This keeps the applied direction aligned with $g^k$ and every displacement local. The descent inequality then holds pathwise.

The parameters satisfy $\gamma \le 1/(64M)$, $\tau \le M/\sqrt{H_1}$ and $\theta n \le \frac{1}{8}$. In particular $\gamma\tau \le 1/(64\sqrt{H_1})$. We prove by induction the hypothesis

$$\mathcal{H}_k: \qquad M\,\varepsilon_x^k \le \frac{m^k}{40}, \qquad \frac{\varepsilon_v^k}{\sqrt{n}} \le \frac{m^k}{40}, \qquad F^k \le F_0. \tag{9}$$

Under $\mathcal{H}_k$ we have $M_k \le M$ and, by $m^k \le \tau \le M/\sqrt{H_1}$, $\varepsilon_x^k \le 1/(40\sqrt{H_1})$. Lemma 14(i) then gives $e^k \le m^k/40$.

*Step 0: base.* $\mathbf{X}^0 = \mathbf{1}(x^0)^\top$ gives $\varepsilon_x^0 = 0$, and $F^0 = F_0$. The initial tracker is mixed with accuracy $\theta_0 \le m^0/(80 \max_i \|g_i^0\|)$. Since $\|\mathbf{G}^0 - \mathbf{1}(\bar{g}^0)^\top\|_F \le \sqrt{n} \max_i \|g_i^0 - \bar{g}^0\| \le 2\sqrt{n} \max_i \|g_i^0\|$,

$$\frac{\varepsilon_v^0}{\sqrt{n}} \;\le\; 2\theta_0 \max_i \|g_i^0\| \;\le\; \frac{m^0}{40}.$$

*Step 1: alignment and second moment.* Assume $\mathcal{H}_k$. Non-expansiveness of clipping row by row, the invariant $\bar{v}^k = \bar{g}^k$ and Cauchy–Schwarz give

$$\langle g^k, \bar{u}^k \rangle \;\ge\; \langle g^k, \mathrm{clip}_\tau(\bar{g}^k) \rangle - \|g^k\| \frac{\varepsilon_v^k}{\sqrt{n}} \;\ge\; \Phi^k - \|g^k\| \Big(e^k + \frac{\varepsilon_v^k}{\sqrt{n}}\Big) \;\ge\; \Big(1 - \frac{1}{20}\Big)\Phi^k.$$

For the second moment, Jensen's inequality and $\frac{1}{n}\sum_i \|v_i^k\|^2 = \|\bar{v}^k\|^2 + (\varepsilon_v^k)^2/n$ give

$$\|\bar{u}^k\|^2 \;\le\; \min\Big\{\tau^2,\; (\|g^k\| + e^k)^2 + \frac{(\varepsilon_v^k)^2}{n}\Big\} \;\le\; \min\{\tau^2,\, 2\|g^k\|^2\} \;\le\; 2\Phi^k.$$

In particular $\|\bar{u}^k\| \le \sqrt{2}\, m^k$.

*Step 2: descent.* The step has length $\gamma\|\bar{u}^k\| \le \gamma\tau \le 1/\sqrt{H_1}$. Lemma 2 for $f$ at $\bar{x}^k$ and Step 1 give

$$F^{k+1} \;\le\; F^k - \gamma\Big(1 - \frac{1}{20}\Big)\Phi^k + \gamma^2 M_k \Phi^k \;\le\; F^k - \frac{\gamma}{2}\Phi^k,$$

because $\gamma M_k \le \frac{1}{64}$. In particular $F^{k+1} \le F_0$.

*Step 3: gradient scale.* Lemma 13 and $\|\bar{u}^k\| \le \sqrt{2}\|g^k\|$ give $\|g^{k+1}\| \ge (1 - \frac{\sqrt{2}}{64})\|g^k\| \ge \frac{31}{32}\|g^k\|$. Hence $m^{k+1} \ge \frac{31}{32} m^k$.

*Step 4: consensus.* By Lemma 14(ii) and $\mathcal{H}_k$, $S^k \le \sqrt{n}\min\{\tau,\; \frac{21}{20}\|g^k\|\} \le \frac{21}{20}\sqrt{n}\, m^k$. With $\theta \le \frac{1}{8n}$ and $\gamma M \le \frac{1}{64}$,

$$M\varepsilon_x^{k+1} \;\le\; \theta\big(M\varepsilon_x^k + \gamma M S^k\big) \;\le\; \frac{1}{8n}\Big(\frac{1}{40} + \frac{21\sqrt{n}}{20 \cdot 64}\Big) m^k \;\le\; 0.006\, m^k \;\le\; \frac{m^{k+1}}{40}.$$

*Step 5: displacement.* Each hub moves by at most

$$D^k \;\le\; \varepsilon_x^{k+1} + \gamma\|\bar{u}^k\| + \varepsilon_x^k \;\le\; \Big(\frac{1}{20} + \frac{\sqrt{2}}{64}\Big)\frac{m^k}{M} \;\le\; \frac{m^k}{13\, M}.$$

Since $m^k/M \le 1/\sqrt{H_1}$, the conditions of Lemma 14(iii) hold.

*Step 6: tracking.* By (8), $M_k \le M$ and $\frac{5}{4} \cdot \frac{1}{13} = \frac{5}{52}$,

$$\frac{\varepsilon_v^{k+1}}{\sqrt{n}} \;\le\; \frac{1}{8n\sqrt{n}}\Big(\frac{\sqrt{n}}{40} + \frac{5n}{52}\Big) m^k \;\le\; \Big(\frac{1}{320\, n} + \frac{5}{416\sqrt{n}}\Big) m^k \;\le\; 0.016\, m^k \;\le\; \frac{m^{k+1}}{40}.$$

Steps 2, 4 and 6 give $\mathcal{H}_{k+1}$.

*Step 7: conclusion.* Summing the descent inequality of Step 2 gives $\min_{k<K} \Phi^k \le 2F_0/(\gamma K)$. For every hub, $\|\nabla f(x_i^k) - g^k\| \le M_k \varepsilon_x^k \le m^k/40$, so $\|\nabla f(x_i^k)\| \le \frac{41}{40}\|g^k\|$. Since $\Phi_\tau(cg) \le c^2 \Phi_\tau(g)$ for $c \ge 1$, the same bound holds at every hub up to the factor $(\frac{41}{40})^2$. □

### D.2 Proof of Theorem 3

The proof of Theorem 7 gives, for every $k$, the pathwise inequality

$$F^{k+1} \;\le\; F^k - \frac{\gamma}{2}\,\Phi^k. \tag{10}$$

Hence $F^k \le F_0$ and $\bar{x}^k$ lies in the sublevel set $\{f \le f(x^0)\}$. The function $f$ is convex as an average of convex $f_i$.

*Proof of part (a).* Lemma 8 gives $\Phi^k \geq \min\{(F^k)^2/R^2, \tau F^k/R\}$. With (10), the sequence $\mu_k := F^k$ satisfies Lemma 7(i) with $q = \gamma/2$. Its condition $q \leq R/\tau$ holds because $\gamma\tau \leq R$. The lemma gives $F^K \leq \varepsilon$ once $K \geq \lceil \frac{2R}{\gamma\tau} \log^+ \frac{F_0}{\tau R} \rceil + \frac{2R^2}{\gamma\varepsilon}$. □

*Proof of part (b).* A strongly convex function has bounded sublevel sets, so $R < \infty$. If $\|g^k\| \geq \tau$, then $\Phi^k = \tau \|g^k\| \geq \tau F^k/R$ by Lemma 8, and (10) gives $F^{k+1} \leq (1 - \frac{\gamma\tau}{2R})F^k$. If $\|g^k\| < \tau$, then $\Phi^k = \|g^k\|^2 \geq 2\mu F^k$ by Lemma 9, and $F^{k+1} \leq (1 - \gamma\mu)F^k$. With $q = \gamma/2$ both cases give $F^{k+1} \leq (1 - q\min\{\tau/R, 2\mu\})F^k$, and $2q\mu = \gamma\mu \leq \gamma M \leq \frac{1}{64}$ by Lemma 9. Lemma 7(ii), together with the recursion of part (a), gives the claim, since $\frac{1}{q\rho} = \max\{\frac{2R}{\gamma\tau}, \frac{1}{\gamma\mu}\}$. □

### D.3 Proof of Proposition 3

*Proof. The function is in the class.* Let $\phi(x) = \cosh x - 1$. For $|y - x| \leq 1$ we have $\cosh z \leq e \cosh x$ on the segment. Hence, for $|y - x| \leq 1/\sqrt{e}$,

$$|\phi'(y) - \phi'(x)| \;\leq\; \sup_{|z-x|\leq 1} \cosh z\, |y - x| \;\leq\; e\,(1 + \phi(x))\,|y - x|,$$

so $\phi$ is $(e, e)$-smooth.

*Reduction.* All hubs hold the same function and start at the same point. Mixing preserves equal rows, and the tracker equals the common gradient. The dynamics of unclipped tracking reduce to $x^{k+1} = x^k - \gamma \sinh x^k$.

*Divergence.* Let $x_* := \max\{2\ln(8/\gamma), 4\}$ and $|x^k| \geq x_*$. We use $\sinh t \geq e^t/4$ for $t \geq 1$ and $te^{-t/2} \leq 2/e < 1$. Then

$$|x^{k+1}| \;\geq\; \gamma \sinh|x^k| - |x^k| \;\geq\; e^{|x^k|/2}\Big(\frac{\gamma}{4}e^{|x^k|/2} - |x^k|e^{-|x^k|/2}\Big) \;\geq\; e^{|x^k|/2},$$

because $\frac{\gamma}{4}e^{|x^k|/2} \geq \frac{\gamma}{4} \cdot \frac{8}{\gamma} = 2$. Since $e^{t/2} \geq t$ for $t \geq 4$, the condition $|x^k| \geq x_*$ is preserved. Iterating $|x^{k+1}| \geq e^{|x^k|/2}$ shows that $|x^k|$ grows at least like a tower of exponentials. Hence $F^k = \cosh x^k - 1$ grows at least doubly exponentially.

*Convergence with clipping.* The function $\phi$ is convex, its sublevel set at $x_*$ is $[-x_*, x_*]$, so $R = x_* \geq 4$. All hubs share the gradient $\sinh x_* \neq 0$, so $\chi_0 = \sinh x_*/\min\{\sinh x_*, \tau\} < \infty$. Under the step rule of Theorem 3, $\gamma\tau \leq 1/(64\sqrt{e}) \leq R$. Hence Theorem 3 applies from $x^0 = x_*$. □

The example reduces to centralized gradient descent. It rules out a constant-step guarantee that is uniform over the class. It does not rule out other step-size schedules.

## E Proofs for the hierarchical method

This appendix proves Theorems 4 and 5 of Section 6, the nonconvex Theorem 8, Corollary 1 and the privacy guarantee of Proposition 4.

**Setting and notation.** We keep the notation of Appendix D. The exact surrogate of hub $i$ at a point $x$ is

$$d_i(x) \;:=\; \frac{1}{m}\sum_{j\in\mathcal{C}_i} \mathrm{clip}_\tau\big(\nabla f_{ij}(x)\big), \qquad \bar{d}(x) := \frac{1}{n}\sum_i d_i(x).$$

In round $k$ hub $i$ uses $d_i^k = d_i(x_i^k) + \xi_i^k$. With exact client gradients $\xi_i^k = 0$. With stochastic clients $\xi_i^k$ collects the effect of the mini-batch noise; $d_i^k$ is still an average of vectors of norm at most $\tau$. The tracker update adds $\mathbf{D}^{k+1} - \mathbf{D}^k$, so the invariant $\bar{v}^k = \bar{d}^k := \frac{1}{n}\sum_i d_i^k$ holds. We use the certificate

$$\mathcal{C}_k: \qquad M\,\varepsilon_x^k \leq \frac{m^k}{160}, \qquad \frac{\varepsilon_v^k}{\sqrt{n}} \leq \frac{m^k}{160}, \qquad F^k \leq F_0.$$

### E.1 HUB-LEVEL ALIGNMENT AND THE SURROGATES

The next lemma lifts the alignment of Lemma 10 from one hub to the network. Each hub surrogate is aligned with its own gradient $g_i$. The pointwise hub condition then aligns the average of the surrogates with $g$.

**Lemma 15** (hub alignment)**.** *Let Assumption 2 hold with $\nu \le \frac{1}{6}$ and Assumption 3 with $\nu_h \le \frac{1}{48}$. Then at every point $x$,*

$$\big\langle g, \bar{d}(x) \big\rangle \;\ge\; \frac{1}{12}\,\Phi_\tau(g), \qquad \frac{1}{n}\sum_i \|d_i(x)\|^2 \;\le\; 2\,\Phi_\tau(g), \qquad g = \nabla f(x).$$

*Proof.* By Lemma 10 and Jensen's inequality, each surrogate satisfies

$$\langle g_i, d_i\rangle \ge \frac{1}{8}\Phi_\tau(g_i), \qquad \|d_i\|^2 \le \min\{\tau^2, \tfrac{3}{2}\,\|g_i\|^2\}.$$

Assumption 3 gives $\|g_i - g\| \le \nu_h\,\|g\|$, so $\|g_i\|$ lies between $(1-\nu_h)\,\|g\|$ and $(1+\nu_h)\,\|g\|$. Since $\Phi_\tau$ is monotone in the norm, $\Phi_\tau(g_i) \ge (1-\nu_h)^2\Phi_\tau(g)$. Moreover,

$$\|d_i\| \le \min\{\tau, \sqrt{3/2}(1+\nu_h)\,\|g\|\} \le \sqrt{3/2}(1+\nu_h)\min\{\tau, \|g\|\}.$$

Hence

$$\langle g, d_i\rangle \;\ge\; \langle g_i, d_i\rangle - \|g_i - g\|\,\|d_i\| \;\ge\; \Big[\frac{1}{8}(1-\nu_h)^2 - \sqrt{\tfrac{3}{2}}\,\nu_h(1+\nu_h)\Big]\Phi_\tau(g) \;\ge\; 0.09\,\Phi_\tau(g).$$

Averaging over $i$ gives the first claim. The second follows from $\|d_i\|^2 \le \frac{3}{2}(1+\nu_h)^2\min\{\tau^2, \|g\|^2\} \le 2\Phi_\tau(g)$. □

At the client level the averaged condition suffices, because a good client contributes $\frac{1}{3}\Phi_\tau$ and a bad one costs at most $\tau\,\|g_i\|$. At the hub level a good hub contributes only $\frac{1}{8}\Phi_\tau$, the constant delivered by the client level. The counting argument of Lemma 10 then does not close, and we use the pointwise hub condition.

In the hierarchy the tracker follows the surrogates $d_i$. The next lemma gives the properties of the surrogates that replace smoothness in the tracking argument. They are bounded by $\tau$ and Lipschitz with the averaged client modulus. The sum of these moduli over the hubs obeys the same bound as in Lemma 14(iii). Part (iv) is used for the mini-batch noise.

**Lemma 16** (surrogates)**.** *Let $L_i^c(x) := \frac{1}{m}\sum_j M_{ij}(x)$. For every hub $i$ and all $x, y$ with $\|x-y\| \le 1/\sqrt{H_1}$:*

*(i)* $\|d_i(x)\| \le \tau$ *and* $\|d_i(x) - d_i(y)\| \le L_i^c(x)\,\|x-y\|$*;*

*(ii)* $\frac{1}{n}\sum_i L_i^c(x) = M(x)$*;*

*(iii) if* $\|x_i - \bar{x}\| \le 1/(18\sqrt{H_1})$ *for all $i$, then* $\sum_i L_i^c(x_i) \le \frac{5}{4}\,nM(\bar{x})$*;*

*(iv) under Assumption 2,* $\frac{1}{m}\sum_j \min\{\|\nabla f_{ij}(x)\|, \tau\} \le (1+\nu)\min\{\|\nabla f_i(x)\|, \tau\}$.

*Proof. (i)* The first bound holds because $d_i$ averages vectors of norm at most $\tau$. The second follows from non-expansiveness of clipping and the gradient bound of Assumption 1 for each $f_{ij}$ at base point $x$. *(ii)* This is the identity $\frac{1}{nm}\sum_{i,j}\phi_{ij}(x) = F(x) + \delta$. *(iii)* Apply the computation in the proof of Lemma 14(iii) to each $f_{ij}$, with modulus $M_{ij}$ and gap $\phi_{ij}$ in place of $M_i$ and $\phi_i$. It gives $M_{ij}(x_i) \le \frac{5}{4}M_{ij}(\bar{x})$. Average over $j$, sum over $i$ and use (ii). *(iv)* By concavity of $t \mapsto \min\{t, \tau\}$ and Assumption 2,

$$\frac{1}{m}\sum_j \min\{\|g_{ij}\|, \tau\} \;\le\; \min\Big\{\frac{1}{m}\sum_j \|g_{ij}\|, \tau\Big\} \;\le\; \min\{(1+\nu)\,\|g_i\|, \tau\} \;\le\; (1+\nu)\min\{\|g_i\|, \tau\}.$$

□

### E.2 One round under the certificate

The next lemma summarizes one round of H$^2$GT-Clip under the certificate $\mathcal{C}_k$. It gives the alignment of the applied direction, its second moment and a bound on the clipped tracker. The noise level $\zeta$ covers stochastic clients. The lemma is shared by Theorem 8, Corollary 1 and Theorem 5.

**Lemma 17** (one round)**.** *Let Assumption 1, Assumption 2 with $\nu \le \frac{1}{6}$ and Assumption 3 with $\nu_h \le \frac{1}{48}$ hold, and let $\tau \le M/\sqrt{H_1}$. Suppose $\mathcal{C}_k$ holds and $\frac{1}{n}\sum_i \|\xi_i^k\| \le \zeta m^k$ with $\zeta \in [0, \frac{1}{100}]$. Then:*

*(a)* $\|\bar{d}^k - \bar{d}(\bar{x}^k)\| \le (\frac{1}{160} + \zeta)\, m^k$;

*(b)* $\langle g^k, \bar{u}^k \rangle \ge c_\zeta \Phi^k$ *with* $c_\zeta := \frac{1}{12} - \frac{1}{80} - \zeta$; *in particular* $c_\zeta \ge \frac{1}{17}$ *and* $c_0 \ge \frac{1}{15}$;

*(c)* $\|\bar{u}^k\|^2 \le 3\Phi^k$ *and* $\|\bar{u}^k\| \le \min\{\tau, \sqrt{3}\, m^k\}$;

*(d)* $S^k := \|\mathrm{clip}_\tau(\mathbf{V}^k)\|_F \le 1.44\sqrt{n}\, m^k$.

*Proof.* Under $\mathcal{C}_k$ we have $M_k \le M$ and $\varepsilon_x^k \le \tau/(160M) \le 1/(160\sqrt{H_1})$.

*(a)* Split $\bar{d}^k - \bar{d}(\bar{x}^k)$ into $\frac{1}{n}\sum_i [d_i(x_i^k) - d_i(\bar{x}^k)]$ and $\frac{1}{n}\sum_i \xi_i^k$. By Lemma 16(i)–(ii), the first part has norm at most $M_k \varepsilon_x^k \le m^k/160$. The second has norm at most $\zeta m^k$.

*(b)* Non-expansiveness row by row, the invariant $\bar{v}^k = \bar{d}^k$ and Cauchy–Schwarz give $\langle g^k, \bar{u}^k \rangle \ge \langle g^k, \mathrm{clip}_\tau(\bar{d}^k) \rangle - \|g^k\|\, \varepsilon_v^k/\sqrt{n}$. Since $\|\bar{d}^k\| \le \tau$, clipping does not change $\bar{d}^k$. By Lemma 15 and part (a),

$$\langle g^k, \bar{u}^k \rangle \;\ge\; \frac{1}{12}\Phi^k - \|g^k\| \Big(\frac{1}{160} + \zeta + \frac{1}{160}\Big) m^k \;=\; c_\zeta\, \Phi^k.$$

*(c)* By Jensen's inequality and $\frac{1}{n}\sum_i \|v_i^k\|^2 = \|\bar{v}^k\|^2 + (\varepsilon_v^k)^2/n$, $\|\bar{u}^k\|^2 \le \min\{\tau^2, \|\bar{v}^k\|^2 + (\varepsilon_v^k)^2/n\}$. By Lemma 15, $\|\bar{d}(\bar{x}^k)\| \le \sqrt{2\Phi^k}$. Part (a) and $(m^k)^2 \le \Phi^k$ give

$$\|\bar{v}^k\|^2 + \frac{(\varepsilon_v^k)^2}{n} \;\le\; \Big[\Big(\sqrt{2} + \frac{1}{160} + \zeta\Big)^2 + \frac{1}{160^2}\Big]\Phi^k \;\le\; 3\,\Phi^k.$$

If $\|g^k\| \ge \tau$ then $\|\bar{u}^k\| \le \tau = m^k$. Otherwise $\Phi^k = (m^k)^2$ and $\|\bar{u}^k\| \le \sqrt{3}\, m^k$.

*(d)* We have $S^k \le \min\{\sqrt{n}\tau,\ \sqrt{n}\,\|\bar{v}^k\| + \varepsilon_v^k\}$. If $\|g^k\| \ge \tau$, then $S^k \le \sqrt{n}\, m^k$. Otherwise the bound on $\|\bar{v}^k\|$ from (c) gives $S^k \le (\sqrt{2} + \frac{2}{160} + \zeta)\sqrt{n}\, m^k$. □

### E.3 Proof of Theorem 8 and Corollary 1

*Idea.* The proof follows that of Theorem 7. The one-round lemma replaces Step 1, and Lemma 16 replaces the tracking bound for exact gradients. For the corollary the mini-batch noise enters as small perturbations $\xi_i^k$.

The parameters satisfy $\gamma \le 1/(256M)$, $\tau \le M/\sqrt{H_1}$ and $\theta n \le \frac{1}{32}$. For Theorem 8 we set $\zeta = 0$. For the corollary we set $\zeta = \frac{1}{100}$ and work on the event of Step 0.

*Step 0: the noise event (corollary only).* Let $\eta' = \frac{1}{400}$. By Assumption 4 and Chebyshev's inequality, a mini-batch gradient $g_j$ satisfies $\|g_j - g_{ij}\| \le \eta' \|g_{ij}\|$ with probability at least $1 - 160000(\rho_\ell - 1)/b$. The run draws $(K+1)nm$ mini-batches. By the batch condition and a union bound, all of them satisfy this bound with probability at least $1 - \beta$. We work on this event. If $g_{ij} = 0$, then $g_j = g_{ij}$. Otherwise Lemma 5 gives $\|\mathrm{clip}_\tau(g_j) - \mathrm{clip}_\tau(g_{ij})\| \le 3\eta' \min\{\|g_{ij}\|, \tau\}$. By Lemma 16(iv),

$$\|\xi_i^k\| \;\le\; 3\eta'(1+\nu)\min\big\{\|\nabla f_i(x_i^k)\|, \tau\big\}.$$

If $M\varepsilon_x^k \le m^k/160$, then $\min\{\|\nabla f_i(x_i^k)\|, \tau\} \le (1+\nu_h)m^k + M_i(\bar{x}^k)\varepsilon_x^k$. Averaging over $i$ and using $\frac{1}{n}\sum_i M_i(\bar{x}^k) = M_k \le M$,

$$\frac{1}{n}\sum_i \|\xi_i^k\| \;\le\; 3\eta' \cdot \frac{7}{6}\Big(\frac{49}{48} + \frac{1}{160}\Big) m^k \;\le\; 0.009\, m^k \;\le\; \zeta m^k. \tag{11}$$

At $k = 0$ all hubs share $x^0$, and the same bound holds for each hub separately.

*Step 1: base.* $\varepsilon_x^0 = 0$ and $F^0 = F_0$. By Lemma 15, $\|d_i(x^0)\| \le \sqrt{3/2}(1+\nu_h)m^0 \le 1.26\, m^0$. With Step 0, $\|d_i^0\| \le 1.27\, m^0$ for every hub. Hence $\|\mathbf{D}^0 - \mathbf{1}(\bar{d}^0)^\top\|_F \le 2.54\sqrt{n}\, m^0$. The initial mixing with $\theta_0 \le \frac{1}{500}$ gives $\varepsilon_v^0/\sqrt{n} \le 0.0051\, m^0 \le m^0/160$.

*Step 2: descent.* Assume $\mathcal{C}_k$. By Step 0 the noise condition of Lemma 17 holds. The step has length at most $\gamma\tau \le 1/(256\sqrt{H_1})$. Lemma 2 and Lemma 17(b)–(c) give

$$F^{k+1} \;\le\; F^k - \gamma c_\zeta \Phi^k + \frac{\gamma^2 M_k}{2}\cdot 3\Phi^k \;\le\; F^k - \gamma\Phi^k\Big(\frac{1}{17} - \frac{3}{512}\Big) \;\le\; F^k - \frac{\gamma}{25}\Phi^k.$$

*Step 3: gradient scale.* Lemma 13 and $\|\bar{u}^k\| \le \sqrt{3}\,\|g^k\|$ give

$$\Big(1 - \frac{\sqrt{3}}{256}\Big)m^k \;\le\; m^{k+1} \;\le\; \Big(1 + \frac{\sqrt{3}}{256}\Big)m^k,$$

so $m^{k+1} \ge \frac{63}{64}m^k$ and $m^{k+1} \le 1.007\, m^k$.

*Step 4: consensus.* By Lemma 14(ii) and Lemma 17(d),

$$M\varepsilon_x^{k+1} \;\le\; \frac{1}{32n}\Big(\frac{m^k}{160} + \frac{1.44\sqrt{n}\, m^k}{256}\Big) \;\le\; 0.0004\, m^k \;\le\; \frac{m^{k+1}}{160}.$$

*Step 5: displacement.* Each hub moves by at most

$$D^k \;\le\; \varepsilon_x^{k+1} + \varepsilon_x^k + \gamma\,\|\bar{u}^k\| \;\le\; \Big(0.0004 + \frac{1}{160} + \frac{\sqrt{3}}{256}\Big)\frac{m^k}{M} \;\le\; 0.014\,\frac{m^k}{M}.$$

Since $m^k/M \le 1/\sqrt{H_1}$, both $\varepsilon_x^k \le 1/(18\sqrt{H_1})$ and $D^k \le 1/(3\sqrt{H_1})$ hold.

*Step 6: tracking.* Centering does not increase the Frobenius norm, so $\varepsilon_v^{k+1} \le \theta(\varepsilon_v^k + \|\mathbf{D}^{k+1} - \mathbf{D}^k\|_F)$. The exact part of the increment is bounded by Lemma 16(i),(iii) and $\|\cdot\|_2 \le \|\cdot\|_1$. The noise part is bounded by (11) at rounds $k$ and $k+1$; the latter applies by Step 4. Hence

$$\|\mathbf{D}^{k+1} - \mathbf{D}^k\|_F \;\le\; \frac{5}{4}\,nMD^k + n\zeta\big(m^{k+1} + m^k\big) \;\le\; n\,m^k\big(0.0175 + 2.01\,\zeta\big),$$

and

$$\frac{\varepsilon_v^{k+1}}{\sqrt{n}} \;\le\; \frac{1}{32n\sqrt{n}}\Big(\frac{\sqrt{n}\, m^k}{160} + 0.038\, n\, m^k\Big) \;\le\; 0.0014\, m^k \;\le\; \frac{m^{k+1}}{160}.$$

Steps 2, 4 and 6 give $\mathcal{C}_{k+1}$.

*Step 7: conclusion.* Summing the inequality of Step 2 gives $\min_{k<K}\Phi^k \le 25F_0/(\gamma K)$. For the corollary this holds on the event of Step 0, which has probability at least $1-\beta$. □

### E.4 Proof of Theorem 4

The proof above gives, for every $k$, the pathwise inequality $F^{k+1} \le F^k - \frac{\gamma}{25}\Phi^k$. Hence $F^k \le F_0$, and $\bar{x}^k$ lies in the sublevel set. The function $f$ is convex as an average of convex $f_{ij}$.

*Proof of part (a).* Lemma 8 gives $\Phi^k \ge \min\{(F^k)^2/R^2, \tau F^k/R\}$. The sequence $\mu_k := F^k$ satisfies Lemma 7(i) with $q = \gamma/25$. Its condition $q \le R/\tau$ holds because $\gamma\tau \le R$. The lemma gives $F^K \le \varepsilon$ once $K \ge \lceil \frac{25R}{\gamma\tau}\log^+\frac{F_0}{\tau R}\rceil + \frac{25R^2}{\gamma\varepsilon}$. □

*Proof of part (b).* If $\|g^k\| \ge \tau$, then $\Phi^k \ge \tau F^k/R$ and $F^{k+1} \le (1 - \frac{\gamma\tau}{25R})F^k$. If $\|g^k\| < \tau$, then $\Phi^k \ge 2\mu F^k$ by Lemma 9, and $F^{k+1} \le (1 - \frac{2\gamma\mu}{25})F^k$. With $q = \gamma/25$ both cases give $F^{k+1} \le (1 - q\min\{\tau/R, 2\mu\})F^k$, and $2q\mu \le \gamma M \le \frac{1}{256}$ by Lemma 9. Lemma 7(ii), together with the recursion of part (a), gives the claim, since $\frac{1}{q\rho} = \max\{\frac{25R}{\gamma\tau}, \frac{25}{2\gamma\mu}\}$. □

### E.5 Proof of Theorem 5

*Idea.* The nonconvex proof bounds the modulus along the trajectory by $M$. Here convexity gives the sharper bound $M_k \le \widehat{H}_0 + H_1 \widetilde{R} \left\|g^k\right\|$. With $\tau = \widehat{H}_0/(H_1\widetilde{R})$ the curvature term $H_1 F^k$ is paid by the clipping radius. The step can then be set by $\widehat{H}_0$ instead of $M$.

The exact client gradients give $\xi_i^k = 0$, so Lemma 17 applies with $\zeta = 0$. Let $\Gamma = \max\{1, \gamma M\}$. The choice of $K_{\mathrm{in}}$ and Lemma 6 give $\theta n \le 1/(1024\,\Gamma)$. Since $\widetilde{R} \ge 1/\sqrt{H_1}$,

$$\tau \;\le\; \frac{\widehat{H}_0}{\sqrt{H_1}} \;\le\; \frac{M}{\sqrt{H_1}}, \qquad \gamma\tau \;=\; \frac{1}{180\,H_1\widetilde{R}} \;\le\; \frac{1}{180\sqrt{H_1}}, \qquad H_1\widetilde{R}\,\tau \;=\; \widehat{H}_0. \tag{12}$$

We prove $\mathcal{C}_k$ by induction. The base case is Step 1 of the previous proof with $\zeta = 0$.

*Step 1: modulus along the trajectory.* Assume $\mathcal{C}_k$. Then $\bar{x}^k$ lies in the sublevel set, and Lemma 8 with $R \le \widetilde{R}$ gives $F^k \le \widetilde{R}\left\|g^k\right\|$. Using $\left\|\bar{u}^k\right\|^2 \le \tau\left\|\bar{u}^k\right\|$, Lemma 17(c) and (12),

$$M_k\left\|\bar{u}^k\right\|^2 \;\le\; \widehat{H}_0\left\|\bar{u}^k\right\|^2 + H_1\widetilde{R}\left\|g^k\right\|\,\tau\,\left\|\bar{u}^k\right\| \;\le\; 3\widehat{H}_0\Phi^k + \sqrt{3}\,H_1\widetilde{R}\tau\,\left\|g^k\right\| m^k \;=\; (3+\sqrt{3})\,\widehat{H}_0\,\Phi^k.$$

In the same way, $M_k\left\|\bar{u}^k\right\| \le (\sqrt{3}+1)\widehat{H}_0\left\|g^k\right\|$.

*Step 2: descent.* The step has length at most $\gamma\tau \le 1/\sqrt{H_1}$. Lemma 2, Lemma 17(b) and Step 1 give

$$F^{k+1} \;\le\; F^k - \gamma\Phi^k\Big(\frac{1}{15} - \frac{3+\sqrt{3}}{360}\Big) \;\le\; F^k - \frac{\gamma}{20}\Phi^k.$$

*Step 3: gradient scale.* Lemma 13 and Step 1 give $\left\|g^{k+1}\right\| \ge (1 - \frac{\sqrt{3}+1}{180})\left\|g^k\right\|$, so $m^{k+1} \ge \frac{63}{64}m^k$.

*Step 4: consensus.* By Lemma 14(ii), Lemma 17(d) and $\gamma M \le \Gamma$,

$$M\varepsilon_x^{k+1} \;\le\; \frac{1}{1024\,n\Gamma}\Big(\frac{m^k}{160} + 1.44\,\Gamma\sqrt{n}\,m^k\Big) \;\le\; 0.0015\,m^k \;\le\; \frac{m^{k+1}}{160}.$$

*Step 5: displacement.* By $\left\|\bar{u}^k\right\| \le \sqrt{3}\,m^k$ and $\Gamma \ge 1$,

$$D^k \;\le\; \varepsilon_x^{k+1} + \varepsilon_x^k + \gamma\left\|\bar{u}^k\right\| \;\le\; 0.008\,\frac{m^k}{M} + \sqrt{3}\,\gamma M\,\frac{m^k}{M} \;\le\; 1.75\,\Gamma\,\frac{m^k}{M}.$$

By (12), $\varepsilon_x^k \le 1/(160\sqrt{H_1})$ and $D^k \le 0.008/\sqrt{H_1} + \sqrt{3}\gamma\tau \le 0.02/\sqrt{H_1}$.

*Step 6: tracking.* As in Step 6 of the previous proof with $\zeta = 0$,

$$\frac{\varepsilon_v^{k+1}}{\sqrt{n}} \;\le\; \frac{1}{1024\,n\Gamma\sqrt{n}}\Big(\frac{\sqrt{n}\,m^k}{160} + \frac{5}{4}\,n\cdot 1.75\,\Gamma\,m^k\Big) \;\le\; 0.0022\,m^k \;\le\; \frac{m^{k+1}}{160}.$$

This completes the induction.

*Step 7: two phases.* Lemma 8 gives $\Phi^k \ge \min\{(F^k)^2/\widetilde{R}^2, \tau F^k/\widetilde{R}\}$. With Step 2, the sequence $\mu_k := F^k$ satisfies Lemma 7(i) with $R$ replaced by $\widetilde{R}$ and $q = \gamma/20 = 1/(3600\widehat{H}_0)$. Its condition $q \le \widetilde{R}/\tau$ holds, since $q\tau = 1/(3600H_1\widetilde{R}) \le \widetilde{R}$. The lemma gives $F^K \le \varepsilon$ once

$$K \;\ge\; \Big\lceil \frac{\widetilde{R}}{q\tau}\log^+\frac{F_0}{\tau\widetilde{R}}\Big\rceil + \frac{\widetilde{R}^2}{q\varepsilon}.$$

Finally $\widetilde{R}/(q\tau) = 3600H_1\widetilde{R}^2$, $\tau\widetilde{R} = \widehat{H}_0/H_1$ and $\widetilde{R}^2/q = 3600\widehat{H}_0\widetilde{R}^2$. □

### E.6 Differential privacy: proof of Proposition 4

Each hub runs HubClip on its clients, so the privacy guarantee is a statement about one hub. Clipping before aggregation bounds the sensitivity of each client message. The hub re-clips the noisy average to keep the step local.

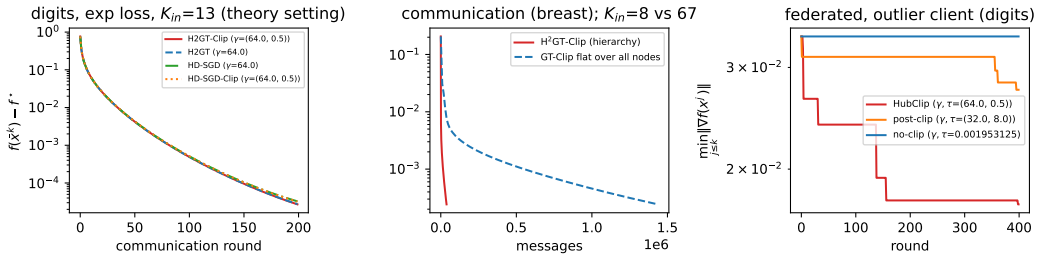


Figure 5: **(a)** Exponential loss on digits with $K_{\text{in}} = 8$, the value required by Theorem 7. All four methods nearly coincide. **(b)** Messages to accuracy for the hierarchy and for GT-CLIP over the star-augmented graph (72 nodes, spectral gap $0.0056$ against $0.195$ for the backbone). **(c)** One hub with $m = 16$ clients, cohort $s = 4$ and mini-batch $b = 8$. Client 0 has flipped labels and gradients scaled by 50.

**Proposition 4** (differential privacy)**.** *Let the conditions of Theorem 6 hold, let $\epsilon_0 \in (0, 1)$, and let each client add independent noise $\zeta_j \sim \mathcal{N}(0, \sigma_{\text{DP}}^2 I_d)$ to its clipped message, with $\sigma_{\text{DP}} \geq 2\tau\sqrt{2\log(1.25/\delta_0)}/\epsilon_0$. The hub steps along $\tilde{d} := \text{clip}_{2\tau}(d + \bar{\zeta})$, $\bar{\zeta} = \frac{1}{s}\sum_{j\in\mathcal{S}} \zeta_j$, with $\gamma = \min\{[128M_f]^{-1}, [2\sqrt{H_1}\tau]^{-1}\}$ and $M_f = H_0 + H_1(\delta_i + 4F_0/\beta)$. Then every round is $(\epsilon_0, \delta_0)$-differentially private for each client. If moreover $\varsigma := \sigma_{\text{DP}}\sqrt{d/s}$ satisfies $\varsigma \leq \tau/64$ and $17\varsigma^2 \leq F_0/(\gamma K)$, then with probability at least $1 - \beta$*

$$\min_{k<K} \Phi_\tau\big(\nabla f_i(x^k)\big) \;\leq\; \frac{256\, F_0}{\beta\gamma K}.$$

*Proof. Privacy.* Changing the data of one client changes its clipped message by at most $2\tau$. The Gaussian mechanism with the stated $\sigma_{\text{DP}}$ is $(\epsilon_0, \delta_0)$-private. The re-clipping at the hub is post-processing.

*Locality.* The step has length at most $2\gamma\tau \leq 1/\sqrt{H_1}$. Without re-clipping the noisy aggregate would be unbounded.

*One round.* Since $\|d\| \leq \tau$, re-clipping does not change $d$, and non-expansiveness gives

$$\left\langle g_i, \tilde{d} \right\rangle \;\geq\; \langle g_i, d\rangle - \|g_i\|\,\|\bar{\zeta}\|, \qquad \left\|\tilde{d}\right\|^2 \;\leq\; 2\|d\|^2 + 2\|\bar{\zeta}\|^2.$$

The noises are independent, so $\mathbb{E}\|\bar{\zeta}\|^2 = \varsigma^2$ and $\mathbb{E}\|\bar{\zeta}\| \leq \varsigma$. If $\|g_i\| \geq \tau$, then $\|g_i\|\varsigma \leq \frac{1}{64}\Phi_\tau(g_i)$. If $\|g_i\| < \tau$, then $\|g_i\|\varsigma \leq \frac{1}{64}\|g_i\|^2 + 16\varsigma^2$. With Lemma 11 and $\gamma M_f \leq \frac{1}{128}$, on $\{k < T\}$,

$$\mathbb{E}_k F^{k+1} \;\leq\; F^k - \gamma\Phi^k\Big(\frac{1}{16} - \frac{1}{64} - 2\gamma M_k\Big) + 16\gamma\varsigma^2 + \gamma^2 M_k\varsigma^2 \;\leq\; F^k - \frac{\gamma}{64}\Phi^k + 17\gamma\varsigma^2.$$

*Conclusion.* Step 3 of the proof of Theorem 2 applies with $\frac{\gamma}{64}$ in place of $\frac{\gamma}{2}$ and $17\varsigma^2$ in place of $2\varepsilon$. Here $\Psi^0 \leq 2F_0$, and Markov's inequality gives the constant $4 \cdot 64 = 256$. □

The privacy of the whole run follows by composing the $K$ subsampled Gaussian mechanisms in Rényi DP and converting the result to $(\epsilon, \delta)$-DP (Wang et al., 2019; Mironov et al., 2019). The composed value depends on the sampling scheme, so we state only the per-round guarantee.

## F EXPERIMENTAL DETAILS

**Protocol.** The hierarchical runs use exact client gradients and full participation, as in Theorems 7 and 8. FASTMIX is the normalized Chebyshev filter of Algorithm 4. Data rows are scaled to norm at most 3, and an $\ell_2$-regularization of $10^{-4}$ is added. For every method the step $\gamma$ is chosen from $\{2^{-9}, \ldots, 2^6\}$; clipped methods also choose $\tau \in \{0.5, 2, 8, 32\}$. The criterion is the best suboptimality over the last 10 rounds, and divergent runs are discarded. The selected values $(\gamma, \tau)$

for Figure 4 are as follows. Breast-cancer, exponential loss: $(16, 2)$ for both clipped methods and $\gamma = 0.125$ for H$^2$GT and $\gamma = 0.5$ for HD-SGD. Digits, logistic loss: $(64, 0.5)$ for the clipped methods and $\gamma = 64$ for the unclipped ones. Quadratic: $(0.5, 8)$ for H$^2$GT-CLIP, $\gamma = 0.5$ for H$^2$GT, $\gamma = 0.125$ for HD-SGD and $(0.125, 32)$ for HD-SGD with client clipping.

**Measured constants.** Every 20 rounds along the trajectory of H$^2$GT-CLIP we evaluate four quantities at $\bar{x}^k$. They are the client dispersion $\nu_i$ of each hub, the hub ratio $\nu_h$, and the two alignment ratios used in the proofs:

$$\nu_i = \frac{\left(\frac{1}{m}\sum_j \|\nabla f_{ij} - \nabla f_i\|^2\right)^{1/2}}{\|\nabla f_i\|}, \quad \nu_h = \max_i \frac{\|\nabla f_i - \nabla f\|}{\|\nabla f\|}, \quad A_i = \frac{\langle \nabla f_i, d_i\rangle}{\Phi_\tau(\nabla f_i)}, \quad A_h = \frac{\langle \nabla f, \bar{d}\rangle}{\Phi_\tau(\nabla f)}.$$

Lemma 10 certifies $A_i \geq \frac{1}{8}$ when $\nu \leq \frac{1}{6}$, and Lemma 15 certifies $A_h \geq \frac{1}{12}$ when in addition $\nu_h \leq \frac{1}{48}$. The table reports the median and the range over the trajectory.

| Run | $\nu_i$ | $\nu_h$ | $A_i$ | $A_h$ |
|---|---|---|---|---|
| breast-cancer, exponential | 1.08 [0.35, 2.8] | 38 [6.3, 68] | 1.00 [0.86, 1] | 1.00 [0.52, 1] |
| digits, logistic | 0.71 [0.47, 1.4] | 50 [4.2, 187] | 1.00 [1, 1] | 1.00 [1, 1] |
| digits, exponential | 0.96 [0.51, 1.6] | 172 [3.7, 861] | 1.00 [0.81, 1] | 1.00 [0.92, 1] |

The dispersion ratios exceed the thresholds of the theorems by one to four orders of magnitude. The alignment ratios stay well above the certified values. The median 1 reflects rounds in which no client is clipped, so that $d_i = \nabla f_i$. The assumptions are thus sufficient conditions for alignment, and alignment itself persists far beyond them.

**Messages.** One round of the hierarchy costs $2nK_{\text{in}} + 2nm$ messages: backbone gossip plus client uploads and downloads. One round of flat gossip costs $2|E(G_{\text{flat}})|K_{\text{in}}^{\text{flat}}$. Both $K_{\text{in}} = 8$ and $K_{\text{in}}^{\text{flat}} = 67$ follow the rule of Theorem 7 for the respective graph. The measured ratio of 38 reflects both the larger edge count and the smaller spectral gap of the flat graph. The flat graph uses Metropolis weights, so Proposition 1 does not apply to it directly.

**Two-phase experiment (Figure 3).** Exponential loss on breast-cancer with $n = m = 8$, $K_{\text{in}} = 8$ and exact gradients. For each start $x^0 = -s\,w$ with $s \in \{1, \dots, 7\}$, both methods are tuned over $\gamma \in \{2^{-6}, \dots, 2^6\}$; the clipped method also chooses $\tau \in \{0.5, 2\}$. The reported count is the smallest number of rounds to $f - f^\star \leq 10^{-2}$ over the grid, with a budget of 3000 rounds. The clipped method selects $(32, 2)$ for $s \leq 3$ and $(64, 0.5)$ for $s \geq 4$. The unclipped method selects $\gamma = 8, 2, 1$ for $s = 1, 2, 3$. In the right panel the unclipped method uses $\gamma = 2^{-5}$, its best value after 3000 rounds at $s = 5$. The step sizes are tuned and are much larger than the conservative choice of Theorem 5. The experiment therefore illustrates the two phases and does not verify the constants.

**Caveats.** In Figure 4(b,c), $K_{\text{in}} = 2$ is below the value required by Theorem 7. The measured $\nu$ and $\nu_h$ violate the thresholds of the theorems, so the experiments do not validate Assumptions 2 and 3 quantitatively. All results use one random seed.