DECENTRALIZED NONCONVEX OPTIMIZATION UNDER HEAVY-TAILED NOISE WITH FIXED-BATCH LOCAL UPDATES

Anonymous authors
Paper under double-blind review

ABSTRACT

Multiple local updates can reduce communication in decentralized optimization, but under heavy-tailed gradient noise, obtaining reliable gradient estimates from a fixed batch is challenging. We study smooth nonconvex decentralized optimization under conditionally centered gradient noise with a bounded conditional p-th moment, $p \in (1, 2)$, using a fixed batch size per local update. We first establish a dimension-independent lower bound on the trade-off between uniform bias and worst-case mean squared error (MSE) in fixed-sample mean estimation. We then augment a three-block coordinatewise median-of-means (MoM) estimator with a clipped correction toward the pooled mean computed from the same batch, and show that the resulting estimator matches this lower bound up to constants depending only on p. Building on this estimator, we propose CRAFT, a decentralized method combining corrected robust estimation, recursive filtering, and local gradient tracking. We establish finite-time stationarity and consensus guarantees and show that, with a fixed batch size per local update, increasing the number of local updates between communications with the horizon reduces the sufficient communication complexity, while the total numbers of local updates and samples required per agent remain of the same order as with a fixed local period. Experiments on image classification, language modeling, and controlled synthetic problems evaluate communication efficiency, robustness under heavy-tailed perturbations, and the roles of estimator correction and recursive filtering.

1 INTRODUCTION

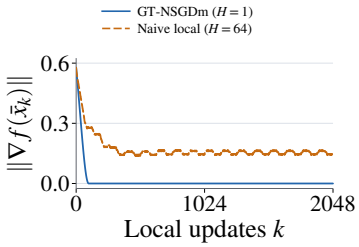
Decentralized learning allows agents to optimize a shared objective using local data and neighbor communication. When communication is more costly than local computation, multiple local updates between exchanges can reduce communication frequency (Koloskova et al., 2020; Wang & Joshi, 2021). However, heterogeneous local objectives cause models to drift between communications. Gradient tracking can mitigate this drift, but its interaction with local updates requires careful control (Nguyen et al., 2023; Liu et al., 2025; Ge & Chang, 2023).

Heavy-tailed gradient noise creates a complementary statistical challenge. A bounded central p-th moment with $1 < p < 2$ allows rare large perturbations and even infinite variance, consistent with empirical observations of stochastic gradients (Simsekli et al., 2019; Yang et al., 2022). With a fixed batch size, the sample mean remains unbiased but may have infinite mean squared error (MSE), whereas robust estimators can have finite MSE at the cost of bias under asymmetric noise. Reliable local updates therefore require controlling both bias and stochastic fluctuations without increasing the sampling cost per update.

These communication and estimation challenges are coupled. Heavy-tailed decentralized methods such as GT-NSGDm achieve convergence with a fixed-sample oracle using normalized updates, momentum, and gradient tracking (Yu et al., 2026), but simply removing intermediate communications need not preserve their finite-time behavior. This motivates a joint treatment of fixed-batch robust estimation and multiple local updates.

Figure 1 illustrates these challenges. Panel (a) shows that periodic communication can change the finite-time behavior of a normalized tracking method even with exact gradients. Panel (b) shows the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) Periodic communication

Estimator	Bias	MSE
Three-block MoM	May be nonzero	Uniformly bounded
Sample mean	Zero	May be infinite

At a fixed sample size, no estimator is uniformly unbiased with finite worst-case MSE over the moment class with $\sigma > 0$ and $1 < p < 2$ (Theorem 1).

(b) Fixed-sample estimation under $1 < p < 2$

Figure 1: Two challenges for local-update optimization under heavy-tailed noise. (a) Deterministic normalized-tracking diagnostic comparing communication every update ($H = 1$) with every 64 updates ($H = 64$). The runs use exact gradients and differ only in communication period; the $H = 64$ run retains a nonzero gradient residual over the displayed horizon. See Appendix A. (b) Fixed-sample behavior of the three-block coordinatewise MoM estimator and sample mean over the bounded central p-th-moment class. The trade-off is uniform over the distribution class; a single distribution need not exhibit both difficulties.

fixed-sample statistical tension: a three-block coordinatewise median-of-means (MoM) estimator has finite worst-case MSE but may be biased, whereas the sample mean is unbiased but may have infinite MSE. This motivates our central question: *how can we control estimation bias and stochastic fluctuations with a fixed batch size while allowing multiple local updates between communications?*

Contributions. To address this question, we first characterize the bias–MSE trade-off in mean estimation at a fixed sample size and construct a corrected robust estimator that matches the resulting lower bound up to constants. Building on this estimator, we propose CRAFT (*Corrected Robust Anchors for Filtered Tracking*), a decentralized method combining corrected robust estimation, recursive filtering, and local gradient tracking. Our main contributions are as follows.

1. **A fixed-sample bias–MSE lower bound.** For distributions with central p-th moment bounded by $\sigma^p > 0$, $1 < p < 2$, we establish a dimension-independent lower bound relating uniform bias and worst-case MSE at a fixed sample size. In particular, uniform unbiasedness is incompatible with finite worst-case MSE over this moment class. This is a statistical estimation lower bound, not an optimization or communication lower bound (Theorem 1).
2. **A matching corrected estimator.** We augment a three-block coordinatewise median-of-means (MoM) estimator with a clipped correction toward the pooled mean computed from the same batch of N samples. The correction radius κ provides a tunable bias–MSE trade-off. For $\kappa \geq \sigma N^{-(p-1)/p}$, the resulting estimator matches the lower-bound trade-off up to constants depending only on p (Theorem 2).
3. **Decentralized convergence with communication-efficient local updates.** We establish finite-time stationarity and consensus guarantees for CRAFT under smoothness, network-mixing, and conditional finite-p-moment assumptions, with a fixed batch size per local update. For expected gradient-norm accuracy ε, our sufficient bound gives $O(\varepsilon^{-(3p-2)/(p-1)})$ communication rounds when the number of local updates per round is fixed. Allowing this number to grow with the horizon reduces the sufficient communication complexity to $O(\varepsilon^{-(5p-4)/(2(p-1))})$, while the total numbers of local updates and samples required per agent remain of the same order (Theorem 3 and Corollary 1).
4. **Experiments and ablations.** Experiments on image classification, language modeling, and synthetic problems evaluate CRAFT across communication budgets and noise regimes. Matched-budget ablations examine the roles of estimator correction and recursive filtering.

Related work. MoM and related estimators provide robust mean estimation under weak moment assumptions (Hsu & Sabato, 2016; Lugosi & Mendelson, 2019), but median aggregation can introduce bias (Karimireddy et al., 2021). Our fixed-sample lower bound builds on change-of-expectation arguments for bias–variance trade-offs (Derumigny & Schmidt-Hieber, 2023), while our estimator adds a bounded same-batch correction from a robust median anchor toward the pooled mean. In

contrast, Sun et al. (2026) study an online MoM–mean interpolation for strongly monotone games. Clipping and normalization offer alternative approaches to heavy-tailed optimization (Hübler et al., 2025; Liu & Zhou, 2025).

In decentralized optimization, GT-NSGDm combines momentum, gradient tracking, and normalization with one sample per update (Yu et al., 2026), whereas DNSGD uses accuracy-dependent batches and multiple consensus steps (Wang et al., 2026). Under finite variance, K-GT and LSGT combine gradient tracking with multiple local updates (Liu et al., 2025; Ge & Chang, 2023). CRAFT instead combines fixed-batch robust estimation with local updates under conditional finite-p-moment noise. Section 5 compares the guarantees; see Appendix B for details.

2 Problem Setting

Consider n agents connected by a fixed communication graph. Agent i has a differentiable, possibly nonconvex objective $f_i : \mathbb{R}^d \to \mathbb{R}$ and access to a stochastic first-order oracle. The agents seek to minimize the global objective $f(x) := n^{-1} \sum_{i=1}^n f_i(x)$ despite heterogeneous local objectives (Koloskova et al., 2020; Liu et al., 2025). Introducing a local model x_i at each agent gives the equivalent consensus problem

$$\min_{\boldsymbol{X} \in \mathbb{R}^{n \times d}} F(\boldsymbol{X}) := \frac{1}{n} \sum_{i=1}^n f_i(x_i) \qquad \text{s.t.} \qquad \Pi \boldsymbol{X} = 0, \tag{1}$$

where $\boldsymbol{X} = [x_1^\top; \dots; x_n^\top]$, $J := \mathbf{1}_n \mathbf{1}_n^\top / n$, and $\Pi := I_n - J$. Write $\bar{x} := n^{-1} \sum_i x_i$ and $G(\boldsymbol{X}) := [\nabla f_1(x_1)^\top; \dots; \nabla f_n(x_n)^\top]$, and use the normalized Frobenius norm $\|V\|_n := \|V\|_F / \sqrt{n}$. Then $\|\Pi \boldsymbol{X}\|_n$ measures consensus disagreement, while $\|JG(J\boldsymbol{X})\|_n = \|\nabla f(\bar{x})\|$ measures stationarity of the global objective at the network-average model.

2.1 Assumptions

Assumption 1 (L-smoothness and lower boundedness)**.** *There exists $L > 0$ such that, for every $i \in [n]$ and $x, y \in \mathbb{R}^d$,*

$$\|\nabla f_i(x) - \nabla f_i(y)\| \le L \|x - y\| .$$

Moreover, $f_{\inf} := \inf_{x \in \mathbb{R}^d} f(x) > -\infty$.

Assumption 2 (Heavy-tailed stochastic gradients)**.** *At each query, let $\mathcal{F}$ be the σ-field generated by the history before drawing a fresh batch, and let x be any $\mathcal{F}$-measurable query point. There exist $p \in (1, 2]$ and $\sigma \ge 0$, common to all agents and queries, such that, for every $i \in [n]$, almost surely,*

$$\mathbb{E}[g_i(x; \xi) \mid \mathcal{F}] = \nabla f_i(x), \qquad \mathbb{E}[\|g_i(x; \xi) - \nabla f_i(x)\|^p \mid \mathcal{F}] \le \sigma^p.$$

Within each agent's fresh batch, samples are conditionally independent given $\mathcal{F}$.

Assumption 3 (Network mixing)**.** *The fixed mixing matrix $W \in \mathbb{R}^{n \times n}$ is nonnegative, compatible with the communication graph, and doubly stochastic:*

$$W \mathbf{1}_n = \mathbf{1}_n, \qquad \mathbf{1}_n^\top W = \mathbf{1}_n^\top.$$

Moreover, $\lambda := \|W - J\|_2 < 1$. Thus W preserves the network average and contracts disagreement by at most λ.

Assumptions 1 and 3 are standard in nonconvex decentralized gradient tracking (Liu et al., 2025; Yu et al., 2026), while Assumption 2 follows the finite-p heavy-tailed oracle setting (Liu & Zhou, 2025; Yu et al., 2026). The conditional formulation allows adaptive query points and supports the MoM analysis; the precise filtration is given in Appendix D. We require neither noise symmetry nor independence across agents or coordinates, and impose no uniform bound on gradients or gradient heterogeneity. The mixing matrix need not be symmetric, although double stochasticity is an additional restriction for directed graphs.

2.2 Fixed batch size and local updates

Multiple local updates between neighbor exchanges can reduce communication, but each update also incurs a sampling cost. We assume that every agent uses a fixed batch of N fresh samples per

oracle query, independent of the horizon and target accuracy. Thus, communication savings from local updates do not rely on increasing the sampling cost per update.

Fixed-sample oracles are already used in heavy-tailed decentralized methods such as GT-NSGDm (Yu et al., 2026). For robust mean estimation, however, controlling both bias and MSE from a fixed batch is challenging under heavy-tailed noise; MoM-based approaches may instead increase the batch size to improve estimation accuracy (Sun et al., 2026). Here N remains fixed, motivating the fixed-sample estimation problem studied next.

3 Fixed-Batch Gradient Estimation

We first isolate the mean-estimation problem underlying each gradient query. For $p \in (1, 2]$ and $\sigma > 0$, define

$$\mathcal{P}_{p,\sigma}^{d} := \{P : \ \mu_P := \mathbb{E}_P V \text{ exists}, \quad \mathbb{E}_P \|V - \mu_P\|^p \le \sigma^p\}, \tag{2}$$

where $V \sim P$ on $\mathbb{R}^d$. Given N i.i.d. samples from $P \in \mathcal{P}_{p,\sigma}^{d}$, we estimate μ_P; conditionally on the algorithmic history, this models a local gradient query with $\mu_P = \nabla f_i(x)$.

For $p < 2$, the sample mean may have infinite MSE, whereas finite-MSE robust estimators may be biased. For any measurable estimator $T_N = T_N(V_1, \dots, V_N)$ integrable over $\mathcal{P}_{p,\sigma}^{d}$, define

$$B(T_N) := \sup_{P \in \mathcal{P}_{p,\sigma}^{d}} \|\mathbb{E}_P T_N - \mu_P\|, \qquad Q(T_N) := \sup_{P \in \mathcal{P}_{p,\sigma}^{d}} \mathbb{E}_P \|T_N - \mu_P\|^2. \tag{3}$$

Jensen's inequality gives

$$B(T_N)^2 \le Q(T_N). \tag{4}$$

Theorem 1 (Fixed-sample bias–MSE lower bound)**.** *Let $p \in (1, 2)$, $\sigma > 0$, and $N, d \ge 1$. Any estimator T_N with $0 < B(T_N) < \infty$ satisfies*

$$B(T_N)^{2-p} Q(T_N)^{p-1} \ge 8^{-p} \sigma^p N^{1-p}. \tag{5}$$

If $B(T_N) = 0$, then $Q(T_N) = \infty$.

The proof uses a two-point change-of-expectation argument (Derumigny & Schmidt-Hieber, 2023; Duchi & Ruan, 2021); see Appendix C. Thus no fixed-batch estimator is both uniformly unbiased and finite-MSE over $\mathcal{P}_{p,\sigma}^{d}$ for $1 < p < 2$. This is a worst-case estimation result, not an optimization lower bound. For $p = 2$, the sample mean is unbiased with worst-case MSE at most σ^2/N.

Corrected estimator. For $N = 3m$, split the observations into three blocks and define

$$\bar{V}_b := \frac{1}{m} \sum_{r=1}^{m} V_{b,r} \quad (b = 1, 2, 3), \qquad R := \mathrm{med}_{b=1,2,3} \bar{V}_b, \qquad U := \frac{1}{3} \sum_{b=1}^{3} \bar{V}_b, \tag{6}$$

where the median is coordinatewise and U is the pooled mean. With $\mathrm{clip}_a(v) := v \min\{1, a/\|v\|\}$ and $\mathrm{clip}_a(0) := 0$, define

$$Z_\kappa := R + \mathrm{clip}_\kappa(U - R), \qquad \kappa > 0. \tag{7}$$

The robust anchor R has finite MSE (Lemma 3), whereas U is unbiased but may have infinite MSE. The same-batch correction moves R toward U by at most κ; since $U - R$ is translation invariant, it clips their discrepancy rather than the unknown mean.

Matching guarantee. Write $u \asymp_p v$ if their ratio is bounded above and below by positive constants depending only on p.

Theorem 2 (Matching bias and MSE bounds)**.** *Let $p \in (1, 2)$, $\sigma > 0$, and $N = 3m$ for integers $m, d \ge 1$. For Z_κ in (7), if $\kappa \ge \sigma N^{-(p-1)/p}$, then*

$$B(Z_\kappa) \asymp_p \sigma^p N^{1-p} \kappa^{1-p}, \qquad Q(Z_\kappa) \asymp_p \sigma^p N^{1-p} \kappa^{2-p}.$$

Consequently,

$$B(Z_\kappa)^{2-p} Q(Z_\kappa)^{p-1} \asymp_p \sigma^p N^{1-p}.$$

The constants depend only on p, uniformly over d, N, σ, and κ in the stated range.

Thus Z_κ matches Theorem 1 up to p-dependent constants. Increasing κ decreases the guaranteed bias scale and increases the MSE scale, although the exact risks need not be monotone or maximized by the same distribution. See Appendix D.4 for the proof and minimax characterization, and Appendix E for the limitations of R and U alone.

Conditional bounds for optimization. At an $\mathcal{F}$-measurable query x of agent i, substitute $V_{b,r} = g_i(x; \xi_{i,b,r})$ in (6)–(7) and denote the estimator by $Z_{i,\kappa}(x)$. Let $C_p > 0$ be the constant in Lemma 4, depending only on p, and define

$$b_\kappa := C_p \sigma^p N^{1-p} \kappa^{1-p}, \qquad q_\kappa := C_p \left(\sigma^2 N^{-2(p-1)/p} + \sigma^p N^{1-p} \kappa^{2-p} \right). \tag{8}$$

Lemma 4 gives, almost surely,

$$\|\mathbb{E}[Z_{i,\kappa}(x) - \nabla f_i(x) \mid \mathcal{F}]\| \le b_\kappa, \qquad \mathbb{E}[\|Z_{i,\kappa}(x) - \nabla f_i(x)\|^2 \mid \mathcal{F}] \le q_\kappa.$$

These bounds hold for $N = 3m$, $\kappa > 0$, and also $\sigma = 0$; they allow sample reuse between R and U, nonidentical conditional laws within a batch, and dependence across agents. In the convergence analysis, b_κ, q_κ denote the deterministic bounds (8); b_κ decreases with κ, while q_κ increases for $p < 2$ and is constant in κ for $p = 2$.

4 CRAFT: Recursive Filtering and Local Gradient Tracking

We propose CRAFT for Problem (1), with agents communicating every H local updates and all parameters fixed within each run. Let $\boldsymbol{X}$, M, and Y stack the local models, filtered gradient estimates, and gradient trackers, respectively. For $\boldsymbol{X} = [x_1^\top; \dots; x_n^\top]$, let $Z_\kappa(\boldsymbol{X})$ have i-th row $Z_{i,\kappa}(x_i)^\top$. The map clip_C is applied row-wise, with $\mathrm{clip}_\infty(v) := v$.

Algorithm 1 CRAFT for Problem (1)

Require: W; deterministic $x_0 \in \mathbb{R}^d$; integers $T, H, m \ge 1$; $\eta, \kappa > 0$; $C \in (0, \infty]$; $\theta \in (0, 1]$

$\boldsymbol{X}_0 \leftarrow \mathbf{1}_n x_0^\top$; $M_0 \leftarrow 0$; $Y_0 \leftarrow 0$
for $t = 0, \dots, T-1$ **do**
  $\boldsymbol{X}_{t,0} \leftarrow W\boldsymbol{X}_t$; $Y_{t,0} \leftarrow WY_t$; $M_{t,0} \leftarrow M_t$ ▷ neighbor exchange
  for $h = 0, \dots, H-1$ **do**
    $\boldsymbol{X}_{t,h+1} \leftarrow \boldsymbol{X}_{t,h} - \eta\, \mathrm{clip}_C(Y_{t,h})$
    $Z_{t,h+1} \leftarrow Z_\kappa(\boldsymbol{X}_{t,h+1})$ ▷ $N = 3m$ fresh samples per agent
    $M_{t,h+1} \leftarrow (1-\theta) M_{t,h} + \theta Z_{t,h+1}$
    $Y_{t,h+1} \leftarrow Y_{t,h} + M_{t,h+1} - M_{t,h}$
  end for
  $(\boldsymbol{X}_{t+1}, Y_{t+1}, M_{t+1}) \leftarrow (\boldsymbol{X}_{t,H}, Y_{t,H}, M_{t,H})$
end for

Corrected queries and recursive filtering. Each $Z_{t,h+1}$ estimates the local gradients at the updated models, with conditional bias and MSE bounded by b_κ and q_κ, respectively. Since $M_0 = Y_0 = 0$, the first model update is zero, after which the first query initializes both states without an extra batch. The recursion $M^+ = (1-\theta)M + \theta Z$ exponentially smooths successive corrected queries. After removing their conditional biases, the query errors form a martingale-difference sequence; filtering bounds the second moment of this centered component by $\theta q_\kappa / (2-\theta)$, while its accumulated bias remains bounded by b_κ (Lemma 5). Changes in the target gradients create an additional filtering lag handled in the convergence analysis. The correction radius κ may depend on the prescribed horizon but is fixed within a run.

Gradient tracking through local updates. At each round boundary, agents mix $\boldsymbol{X}$ and Y, while M remains local. During the following local updates, Y accumulates changes in M, giving

$$Y_{t,h} = WY_t + M_{t,h} - M_t, \qquad JY_{t,h} = JM_{t,h}, \qquad 0 \le h \le H.$$

The second identity follows from double stochasticity and $JY_0 = JM_0 = 0$. Thus the network average of the tracker exactly follows the average filtered estimate, while disagreement is contracted at communication boundaries and can accumulate during local updates. This mechanism is related to local-update gradient tracking (Nguyen et al., 2023; Liu et al., 2025; Ge & Chang, 2023).

Clipped updates and resource cost. For finite C, row-wise clipping limits each local displacement to ηC; $C = \infty$ gives the unclipped update. The direction radius C is distinct from the correction radius κ, which controls the bias–MSE trade-off of the gradient estimator.

Each communication round exchanges $\boldsymbol{X}$ and Y; M and raw samples remain local. Over T rounds, CRAFT performs TH local gradient queries, uses $3mTH$ raw samples per agent, and communicates for T rounds. Appendix F.5 gives the detailed resource accounting.

5 Convergence and Communication Complexity

Let $(\widehat{t}, \widehat{h})$ be uniform on $\{0, \dots, T-1\} \times \{0, \dots, H-1\}$ and I uniform on $[n]$, independently of the algorithm. Writing $x_{i,t,h}$ for row i of $\boldsymbol{X}_{t,h}$, define $\widehat{\boldsymbol{X}} := \boldsymbol{X}_{\widehat{t},\widehat{h}}$ and $\widehat{x} := x_{I,\widehat{t},\widehat{h}}$. Then $\mathbb{E}\,\|\nabla f(\widehat{x})\|$ and $\mathbb{E}\left\|\Pi\widehat{\boldsymbol{X}}\right\|_n^2$ measure stationarity and consensus at a random pre-update iterate.

Let $\Delta_0 := f(x_0) - f_{\inf}$ and let b_κ, q_κ be defined in (8). Define

$$\mathcal{R}_{T,H} := \frac{\Delta_0}{TH\eta} + \frac{\|G(\boldsymbol{X}_0)\|_n^2}{TH\theta} + b_\kappa^2 + \left(\theta + H^2\eta^2 + H\theta^2\right) q_\kappa.$$

Theorem 3 (Finite-time convergence)**.** *Under Assumptions 1–3, initialize Algorithm 1 with $\boldsymbol{X}_0 = \mathbf{1}_n x_0^\top$ and $M_0 = Y_0 = 0$, where x_0 is deterministic. There exists $c = c(L, \lambda) \in (0, 1/2]$ such that, for $T, H, m \geq 1$, $\eta, \kappa > 0$, $\theta \in (0, 1]$, and $C \in (0, \infty]$ satisfying*

$$\eta \leq c\theta, \qquad H\theta \leq c, \qquad H^2\eta^2 \leq c\theta, \tag{9}$$

the random output satisfies

$$\mathbb{E}\,\|\nabla f(\widehat{x})\| \lesssim_{L,\lambda} \sqrt{\mathcal{R}_{T,H}} + \frac{\mathcal{R}_{T,H}}{C}, \tag{10}$$

$$\mathbb{E}\left\|\Pi\widehat{\boldsymbol{X}}\right\|_n^2 \lesssim_{L,\lambda} H^3\eta^2\theta^2\left(H\mathcal{R}_{T,H} + q_\kappa\right). \tag{11}$$

Here $1/\infty := 0$. For $C = \infty$, additionally,

$$\mathbb{E}\,\|\nabla f(\widehat{x})\|^2 \lesssim_{L,\lambda} \mathcal{R}_{T,H}. \tag{12}$$

Here $\lesssim_{L,\lambda}$ hides constants depending only on L, λ; see Corollary 3 for explicit constants. The proof controls squared update energy uniformly in C, so neither $\mathcal{R}_{T,H}$ nor (9) depends on C. The term $\mathcal{R}_{T,H}/C$ arises only when converting the clipped descent certificate to gradient norm; the result neither favors a finite C nor identifies an optimal radius and also covers $C = \infty$.

Corollary 1 (Growing local blocks)**.** *Under Theorem 3, let $a := p/(3p-2)$ and $r := 2(p-1)/(3p-2) = 1 - a$. Fix $m \geq 1$, so $N = 3m$ is fixed, and let $C \in (0, \infty]$ be horizon-independent. Choose $0 \leq \chi \leq p/(5p-4)$ and $c_\kappa, c_\theta, c_\eta > 0$ with $c_\theta \leq c/4$ and $c_\eta \leq c c_\theta$. Set*

$$H = \lceil T^\chi \rceil, \qquad \eta = c_\eta (TH)^{-a}, \qquad \theta = c_\theta (TH)^{-a}, \qquad \kappa = c_\kappa (TH)^{1/(3p-2)}. \tag{13}$$

Then

$$\mathbb{E}\,\|\nabla f(\widehat{x})\| = O\left((TH)^{-r/2}\right), \tag{14}$$

$$\mathbb{E}\left\|\Pi\widehat{\boldsymbol{X}}\right\|_n^2 = O\Big(H^3(TH)^{-(1+2a)}\Big) = O\Big((TH)^{-(1+a/2)}\Big), \tag{15}$$

with constants independent of T, H.

For fixed H, the same asymptotic rates hold under (13) as $T \to \infty$. More generally, $H^2(TH)^{-a} \leq 4$ controls the additional local-update errors and holds for the stated range of χ; see Appendix F.5. Indeed,

$$b_\kappa^2 = O((TH)^{-r}), \qquad q_\kappa = O((TH)^{(2-p)/(3p-2)}), \qquad \theta q_\kappa = O((TH)^{-r}).$$

To achieve $\mathbb{E}\,\|\nabla f(\widehat{x})\| \leq \varepsilon$, it suffices that

$$TH = O\Big(\varepsilon^{-\frac{3p-2}{p-1}}\Big), \qquad T = O\Big(\varepsilon^{-\frac{3p-2}{(p-1)(1+\chi)}}\Big), \qquad N_{\mathrm{raw}} = 3mTH \quad \text{per agent.}$$

At $\chi = p/(5p-4)$, the communication complexity becomes $O(\varepsilon^{-(5p-4)/(2(p-1))})$. Thus growing the local period reduces the sufficient communication complexity relative to fixed H, while the total local updates TH and samples per agent retain the same accuracy order.

Table 1: Sufficient costs for expected global gradient norm at a uniformly sampled local model to be at most ε, with n and other problem and network parameters fixed. Here $s_p = (3p-2)/(p-1)$ and $\widetilde{O}$ suppresses logarithmic factors. Batch size is per agent and query; communications count neighbor exchanges up to constants.

Method	Batch size	Samples per agent	Communications
GT-NSGDm (Yu et al., 2026)	1	$O(\varepsilon^{-s_p})$	$O(\varepsilon^{-s_p})$
DNSGD (Wang et al., 2026)	$\Theta(\varepsilon^{-p/(p-1)})$	$O(\varepsilon^{-s_p})$	$\widetilde{O}(\varepsilon^{-2})$
CRAFT, $\chi = p/(5p-4)$	fixed $3m$	$O(\varepsilon^{-s_p})$	$O(\varepsilon^{-(5p-4)/(2(p-1))})$

Comparison with related guarantees. Table 1 compares sufficient costs under a common expected-gradient-norm criterion, fixing n and other problem and network parameters. GT-NSGDm uses one sample and communicates at every update (Yu et al., 2026); DNSGD uses accuracy-dependent batches and accelerated gossip (Wang et al., 2026); CRAFT uses a fixed batch and growing local blocks. DNSGD's system-wide sample count is converted to a per-agent count. The methods differ in oracle, network, and communication assumptions, so the table does not imply dominance across protocols or a speedup in n; see Appendix B.2.

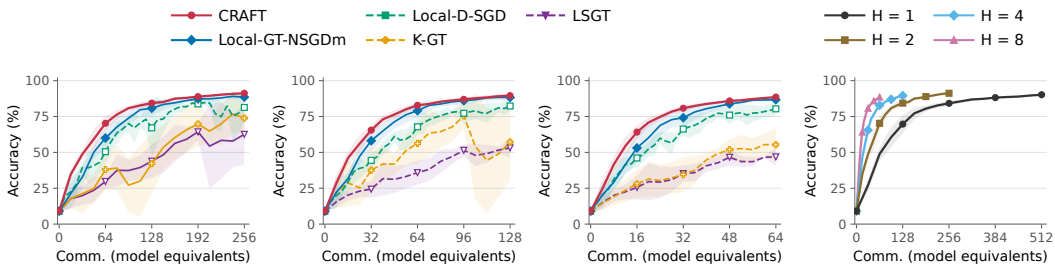
6 Experiments

We evaluate CRAFT along four dimensions: communication efficiency at a fixed per-update sampling cost, robustness to increasing noise intensity, performance on a larger model across fixed network topologies, and the contributions of estimator correction and recursive filtering. We consider MNIST and CIFAR-10 for image classification and GPT-2 (124M) on WikiText-103 for language modeling (LeCun et al., 1998; Krizhevsky, 2009; Radford et al., 2019; Merity et al., 2017). We compare CRAFT with D-SGD, GT-NSGDm (Yu et al., 2026), K-GT (Liu et al., 2025), and LSGT (Ge & Chang, 2023). For $H > 1$, Local-D-SGD and Local-GT-NSGDm denote variants that perform H local updates between consecutive communication rounds.

Protocol and budgets. We evaluate the mean model and measure communication in directed-link bytes, counting model and tracker payloads separately. At matched endpoint communication, D-SGD performs twice as many updates and uses twice as many training samples or tokens as CRAFT; matched-sample views appear in Appendix H. For MNIST and CIFAR-10, losses are weighted by $1 + \rho\sigma_0 z$, with independent $z \sim t_{1.5}$ and task-specific σ_0. These perturbations have finite p-th moments for $1 < p < 1.5$ but do not imply a uniform moment bound over model parameters. GPT-2 uses clean fine-tuning. Unless stated otherwise, results are mean $\pm$ sample SD over five seeds; full protocols and limitations are in Appendix G.

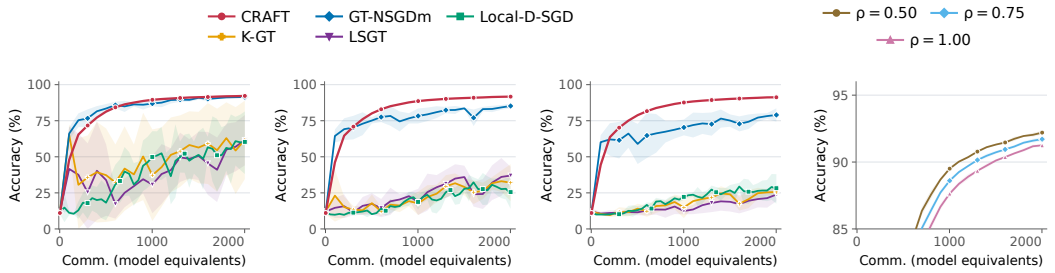
Communication–accuracy trade-off. MNIST uses 16 ring-connected agents, label-Dirichlet partitions ($\alpha = 0.3$), and 126 samples per agent and update. With 256 updates per agent, increasing H from 1 to 8 reduces CRAFT's communication from 512 to 64 model-equivalent units at the same 516,096 system samples (Figure 2). Terminal accuracy is $90.21 \pm 0.31\%$, $91.23 \pm 0.41\%$, and $88.57 \pm 0.81\%$ at $H = 1, 2, 8$, respectively. Configurations are tuned separately, with $\theta = 1$ at $H = 1, 2$ (Table 2), so this is a comparison of operating points rather than an isolated effect of H. Among completed-run statistics, CRAFT has the highest terminal mean at $H = 2, 4, 8$. CRAFT and Local-GT-NSGDm complete all five seeds; other completion counts are reported in Appendix G.1. Figure 6 gives the matched-sample view.

Robustness to perturbation intensity. CIFAR-10 uses 16 agents, IID partitions, a pretrained ResNet18 variant, and $H = 1$. With CRAFT's configuration fixed across noise levels, terminal accuracy decreases from $92.20 \pm 0.13\%$ at $\rho = 0.5$ to $91.25 \pm 0.10\%$ at $\rho = 1$, versus $91.40 \pm 0.53\%$ to $79.01 \pm 4.63\%$ for GT-NSGDm (Figure 3). Under the reported configurations, the other baselines have lower terminal means and larger seed variation; all runs complete (Table 4). At the matched endpoint, CRAFT uses 1,025,024 samples versus 2,050,048 for D-SGD. CRAFT's batch of 64 uses a $22/21/21$ split with size-weighted pooling, an empirical extension not covered directly by the



(a) $H = 2$ (b) $H = 4$ (c) $H = 8$ (d) CRAFT across H

Figure 2: MNIST test accuracy versus directed-link communication. Each curve uses a fixed cohort of completed seeds; bands show sample SD, and hollow markers denote cohorts with fewer than five completed runs. Panels (a–c) compare methods at fixed H; panel (d) compares separately tuned CRAFT configurations. At matched endpoint communication, D-SGD uses twice as many training samples as CRAFT. See Appendix G.1.



(a) $\rho = 0.50$ (b) $\rho = 0.75$ (c) $\rho = 1.00$ (d) CRAFT across noise

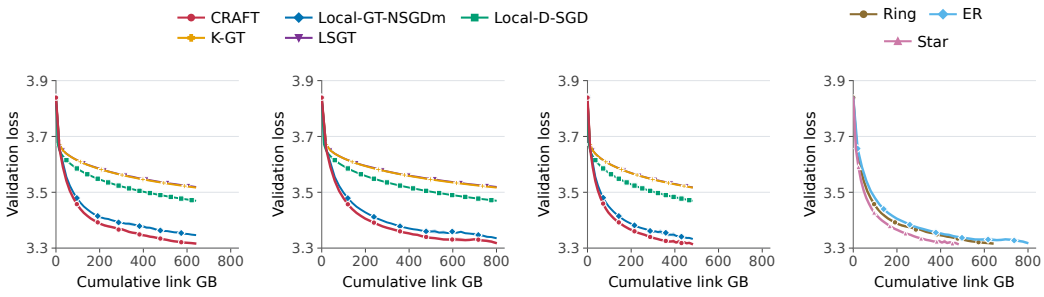
Figure 3: CIFAR-10 robustness at $H = 1$ under matched endpoint communication. Curves show five-seed mean accuracy with sample-SD bands clipped to $[0, 100]\%$; panel (d) uses a magnified accuracy axis.

equal-block theory. Thus these runs are finite-budget robustness tests rather than a direct verification of the oracle assumption.

Language modeling across topologies. We fine-tune GPT-2 (124M) on WikiText-103 with four agents, $H = 16$, and no injected noise. Each method uses one configuration across Ring, ER, and Star, although configurations differ across methods. At matched directed-link bytes within each graph, CRAFT attains terminal validation losses 3.3165 ± 0.0039, 3.3186 ± 0.0043, and 3.3150 ± 0.0018, respectively, the lowest terminal means among the five methods (Figure 4). Local-D-SGD reaches approximately 3.4699 while using twice as many updates and training tokens. Because edge counts differ, the cross-graph results assess consistency across topologies rather than topology effects at a common communication budget. This clean-data experiment evaluates communication efficiency rather than heavy-tail robustness.

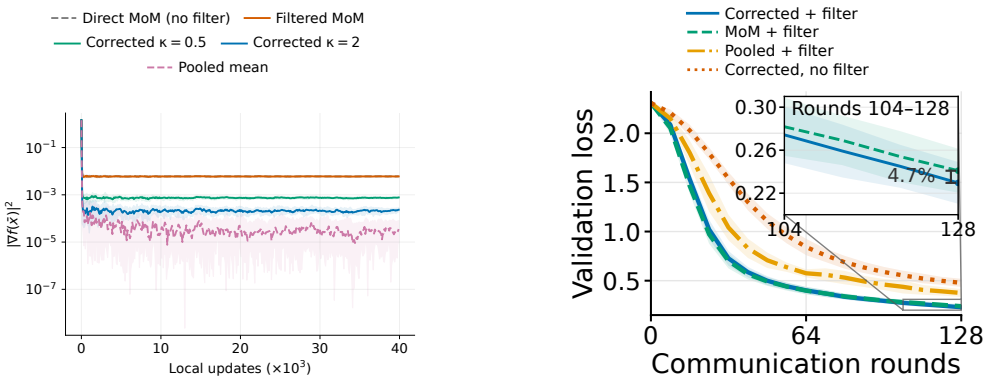
Correction and filtering ablations. The paired MNIST ablation matches initialization, sample streams, stepsize, direction clipping, tracking, and communication. Corrected MoM with filtering attains endpoint validation loss 0.2296 ± 0.0194, versus 0.2409 ± 0.0205 for filtered MoM and 0.4781 ± 0.0416 without filtering (Figure 5b). Correction lowers loss in all five paired seeds and the endpoint mean by 4.7%; full metrics, including the pooled-mean control, are in Table 3.

On a scalar quadratic with centered Pareto noise and matched 24-sample queries, the median terminal squared gradient is 5.98×10^{-3} for filtered MoM and 7.62×10^{-4} and 2.26×10^{-4} for correction radii 0.5 and 2 (Figure 5a). The pooled mean attains a lower finite-run median despite infinite population MSE. On a separate nonconvex problem, mean endpoint squared gradients over 20 paired seeds are 0.002315 for CRAFT, 0.075969 without filtering, and 0.381656 without correction. Full



(a) Ring (b) ER (c) Star (d) CRAFT across graphs

Figure 4: Clean GPT-124M fine-tuning at $H = 16$: mean-model validation loss versus cumulative directed-link GB. Curves show unsmoothed five-seed means with sample-SD bands. Graph-specific edge counts yield different communication endpoints.



(a) Centered Pareto noise

(b) MNIST query/filter ablation

Figure 5: Component diagnostics. (a) Medians with 10th–90th percentiles over 20 paired runs using the same 24 samples per query. (b) Means with sample-SD bands over five paired runs under matched sampling and communication budgets. "Filter" denotes recursive exponential smoothing; all variants retain fixed-radius direction clipping.

controls appear in Table 5 and Figure 9. Tracking and direction clipping remain fixed and are not separately ablated.

7 Discussion and Conclusion

Our results highlight a fundamental tension in fixed-batch heavy-tailed optimization: reducing uniform bias comes at the cost of larger worst-case MSE over the moment class. CRAFT makes this trade-off tunable through the correction radius κ, while recursive filtering attenuates the resulting stochastic fluctuations and local updates reduce communication without increasing the sampling cost per update. This communication gain is specific to the fixed-batch setting and does not imply dominance over methods that increase batch sizes or use additional consensus steps. The current analysis also requires knowledge of the moment index and a prescribed horizon, assumes a fixed doubly stochastic network, and provides expected guarantees for a randomly sampled iterate rather than terminal-iterate or high-probability guarantees. It remains open to adapt the estimator and local-update schedule to unknown tail behavior and time-varying networks, and to obtain sharper communication and high-probability guarantees.

AI USE STATEMENT

We used generative AI tools to assist with language editing, organization, LaTeX formatting, and checking the clarity and consistency of mathematical exposition. AI-assisted text and technical suggestions are subject to author review before submission. The mathematical claims, proofs, experimental results, conclusions, and final content remain the authors' responsibility.

ETHICS STATEMENT

This work develops statistical and optimization methods for decentralized learning under heavy-tailed stochastic noise. The experiments use standard public benchmark datasets and synthetic objectives and do not involve human-subject studies or the collection of personal or sensitive data. We are not aware of direct ethical risks specific to the proposed methodology beyond those generally associated with machine-learning and distributed optimization systems. In particular, the theoretical guarantees concern optimization and communication properties and should not be interpreted as guarantees of fairness, privacy, security, or safety in downstream applications. We report experimental protocols, incomplete runs, and limitations of the empirical evidence explicitly to reduce the risk of overstating the scope of the results.

REPRODUCIBILITY STATEMENT

We provide the definitions and proofs needed to assess the theoretical results, together with the available protocols and aggregate results for the experiments. The problem setting, assumptions, corrected estimator, and CRAFT algorithm are specified in Sections 3–4, while complete proofs of the statistical and convergence results, including explicit constants and intermediate lemmas, are given in Appendices C–F. Experimental protocols, including datasets, model and network configurations, noise constructions, hyperparameters, seeds, evaluation procedures, resource accounting, and known limitations of the available experimental record, are documented in Appendix G. Additional results and controlled ablations are reported in Appendix H. Where empirical protocols differ from the exact setting analyzed in the theory, these differences are stated explicitly in the corresponding experimental sections. The main implementation is available in an anonymous repository at `https://anonymous.4open.science/r/CRAFT-code-EF27/`. The corresponding experimental settings are provided in Appendix G.

REFERENCES

Aleksandar Armacki, Shuhua Yu, Pranay Sharma, Gauri Joshi, Dragana Bajovic, Dusan Jakovetic, and Soummya Kar. High-probability convergence bounds for online nonlinear stochastic gradient descent under heavy-tailed noise. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 1774–1782. PMLR, 2025.

Lawrence D. Brown and Mark G. Low. A constrained risk inequality with applications to nonparametric functional estimation. *The Annals of Statistics*, 24(6):2524–2535, 1996.

Alexis Derumigny and Johannes Schmidt-Hieber. On lower bounds for the bias-variance trade-off. *The Annals of Statistics*, 51(4):1510–1533, 2023. doi: 10.1214/23-AOS2279. URL `https://arxiv.org/abs/2006.00278`.

John Duchi and Feng Ruan. A constrained risk inequality for general losses. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 802–810. PMLR, 2021.

Songyang Ge and Tsung-Hui Chang. Gradient tracking with multiple local SGD for decentralized non-convex learning. In *Proceedings of the 62nd IEEE Conference on Decision and Control*, pp. 133–138, 2023. URL `https://arxiv.org/abs/2302.01537`.

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

Daniel Hsu and Sivan Sabato. Loss minimization and parameter estimation with heavy tails. *Journal of Machine Learning Research*, 17(18):1–40, 2016.

Florian Hübler, Ilyas Fatkhullin, and Niao He. From gradient clipping to normalization for heavy tailed SGD. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 2413–2421, 2025.

Dušan Jakovetić, Dragana Bajović, Anit Kumar Sahu, Soummya Kar, Nemanja Milošević, and Dušan Stamenković. Nonlinear gradient mappings and stochastic optimization: A general framework with applications to heavy-tail noise. *SIAM Journal on Optimization*, 33(2):394–423, 2023.

Sai Praneeth Karimireddy, Lie He, and Martin Jaggi. Learning from history for Byzantine robust optimization. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5311–5319. PMLR, 2021.

Anastasia Koloskova, Nicolas Loizou, Sadra Boreiri, Martin Jaggi, and Sebastian U. Stich. A unified theory of decentralized SGD with changing topology and local updates. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5381–5393. PMLR, 2020.

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

Kasper Green Larsen, Rasmus Pagh, and Jakub Tětek. CountSketches, feature hashing and the median of three. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 6011–6020. PMLR, 2021. URL https://proceedings.mlr.press/v139/larsen21a.html.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

Su Hyeong Lee, Manzil Zaheer, and Tian Li. Efficient distributed optimization under heavy-tailed noise. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 33833–33882. PMLR, 2025.

Yue Liu, Tao Lin, Anastasia Koloskova, and Sebastian U. Stich. Decentralized gradient tracking with local steps. *Optimization Methods and Software*, 40(5):1153–1180, 2025. doi: 10.1080/10556788.2024.2322095.

Zijian Liu and Zhengyuan Zhou. Nonconvex stochastic optimization under heavy-tailed noises: Optimal convergence without gradient clipping. In *International Conference on Learning Representations*, 2025.

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

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Byj72udxe.

Edward Duc Hien Nguyen, Sulaiman A. Alghunaim, Kun Yuan, and César A. Uribe. On the performance of gradient tracking with local updates. In *2023 62nd IEEE Conference on Decision and Control (CDC)*, pp. 4309–4313. IEEE, 2023. doi: 10.1109/CDC49753.2023.10383873.

Nikita Puchkin, Eduard Gorbunov, Nickolay Kutuzov, and Alexander Gasnikov. Breaking the heavy-tailed noise barrier in stochastic optimization problems. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 856–864. PMLR, 2024.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Umut Simsekli, Levent Sagun, and Mert Gurbuzbalaban. A tail-index analysis of stochastic gradient noise in deep neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5827–5837. PMLR, 2019.

Chao Sun, Huiming Zhang, Bo Chen, Jianzheng Wang, Zheming Wang, and Li Yu. Clipping-free Nash equilibrium seeking in heavy-tailed games via median-of-means. *arXiv preprint arXiv:2604.26588v3*, 2026. URL https://arxiv.org/abs/2604.26588v3.

Jianyu Wang and Gauri Joshi. Cooperative SGD: A unified framework for the design and analysis of local-update SGD algorithms. *Journal of Machine Learning Research*, 22(213):1–50, 2021.

Menglian Wang, Zhuanghua Liu, and Luo Luo. Near-optimal decentralized stochastic nonconvex optimization with heavy-tailed noise. *arXiv preprint arXiv:2601.11435v1*, 2026. URL https://arxiv.org/abs/2601.11435v1.

Tongle Wu, Zhize Li, and Ying Sun. The effectiveness of local updates for decentralized learning under data heterogeneity. *IEEE Transactions on Signal Processing*, 73:751–765, 2025. doi: 10.1109/TSP.2025.3533208.

Haibo Yang, Peijie Qiu, and Jia Liu. Taming fat-tailed noise in federated learning. In *Advances in Neural Information Processing Systems*, 2022.

Shuhua Yu, Dusan Jakovetic, and Soummya Kar. Decentralized nonconvex optimization under heavy-tailed noise: Normalization and optimal convergence. In *International Conference on Learning Representations*, 2026.

APPENDIX CONTENTS

A A Fixed-Period Communication Diagnostic **14**

B Related Work and Theoretical Comparisons **15**
B.1 Connections to prior work 15
B.2 Stationarity Criteria and Complexity Conversions 16

C Fixed-Sample Bias–MSE Lower Bound **18**
C.1 A two-point inequality 18
C.2 Proof of Theorem 1. 19
C.3 Scope of the Gradient-Oracle Implication 19

D Corrected Estimation and Recursive Filtering **20**
D.1 Conditional sampling and integrability 20
D.2 Conditional MSE of the median anchor. 20
D.3 Conditional bias and MSE bounds 22
D.4 Matching bounds and minimax MSE 23
D.5 Recursive filtering 25

E Why the Median and Pooled Mean Alone Are Insufficient **26**

F Finite-Time Convergence and Communication Bounds **27**
F.1 Proof sketches for Theorem 3 and Corollary 1 28
F.2 Geometry of the clipped direction 29
F.3 Filter, consensus, and descent bounds 30
F.4 Closing the coupled inequalities with explicit constants 37
F.5 Horizon-dependent rates and output guarantees 41

G Experimental Details **45**
G.1 MNIST: local steps and budget accounting 46
G.2 CIFAR-10 and GPT-124M: equal-communication protocols 48
G.3 Controlled mechanism experiments 50
G.4 Clean Multi30K diagnostics 52

H Additional Experimental Results **53**
H.1 MNIST operating points 53
H.2 MNIST estimator and recursive-filter ablation 53
H.3 CIFAR-10: final accuracy and sample-budget comparisons. 54
H.4 GPT-124M: clean fine-tuning results 54
H.5 Estimator scaling and asymmetric-noise diagnostics 56
H.6 Skewed-noise component ablation 57
H.7 Hyperparameter sensitivity 57
H.8 Local-step wall-clock comparisons 57
H.9 CRAFT–NR consensus and accuracy trajectories 59

A A FIXED-PERIOD COMMUNICATION DIAGNOSTIC

Figure 1(a) isolates the effect of periodic communication in a deterministic setting, without stochastic estimation. The example illustrates that suppressing intermediate communication can alter the finite-time dynamics of a normalized tracking method; it is not an impossibility result for local updates.

Objective and recurrence. Consider four agents with scalar variables. Let $u := \max\{x, 0\}$ and define

$$f(x) := u - \arctan(u) + 0.1(1 - \cos u)^2,$$
$$f_i(x) := f(x) + 4b_i \tanh(2x), \qquad (b_1, b_2, b_3, b_4) := (1, -1, 2, -2).$$

Since $\sum_{i=1}^4 b_i = 0$, the global objective satisfies

$$\frac{1}{4}\sum_{i=1}^4 f_i(x) = f(x).$$

Each f_i is bounded below and has a Lipschitz continuous gradient. The global objective is nonconvex; for example,

$$f''(\pi) = \frac{2\pi}{(1+\pi^2)^2} - 0.4 < 0.$$

Moreover, $f(x) = 0$ for $x \le 0$ and $f(x) > 0$ for $x > 0$, so its minimizer set is $\{x \in \mathbb{R} : x \le 0\}$. Its derivative, together with the local gradients, is

$$f'(x) = \begin{cases} 0, & x \le 0, \\ \dfrac{x^2}{1+x^2} + 0.2(1-\cos x)\sin x, & x > 0, \end{cases} \qquad f_i'(x) = f'(x) + 8b_i\big(1 - \tanh^2(2x)\big).$$

In particular, $|f''(x)| \le 2$ and $|f_i''(x)| \le 2 + 32|b_i| \le 66$ wherever the second derivatives are evaluated, so $L = 66$ is a valid common Lipschitz constant for the local gradients.

Set

$$W := \frac{\mathbf{1}_4\mathbf{1}_4^\top}{4}, \qquad \boldsymbol{X}^0 := \mathbf{1}_4, \qquad V^{-1} := 0, \qquad Y^{-1} := 0.$$

For a communication period $H \in \{1, 64\}$, define

$$W_k := \begin{cases} W, & k \equiv 0 \pmod H, \\ I, & \text{otherwise.} \end{cases}$$

We run the deterministic recursion

$$\begin{aligned} V^k &= 0.5V^{k-1} + 0.5G(\boldsymbol{X}^k), \\ Y^k &= W_k\big(Y^{k-1} + V^k - V^{k-1}\big), \\ \boldsymbol{X}^{k+1} &= W_k\big(\boldsymbol{X}^k - 0.01\ \mathrm{sign}(Y^k)\big), \end{aligned} \tag{16}$$

where

$$G(\boldsymbol{X})_i := f_i'(x_i), \qquad \mathrm{sign}(0) := 0,$$

and sign is applied componentwise. Thus, $\mathrm{sign}(Y^k)$ gives the normalized scalar direction at each agent, with no normalization regularizer or dead zone.

The two runs differ only in the communication period H. Over K updates, both evaluate $4K$ exact local gradients. Communication occurs every H updates, with both the model and tracker exchanged at each communication event. The stepsize and momentum parameters are fixed specifically for this diagnostic and are not chosen according to the theoretical tuning of GT-NSGDm.

Trajectory and interpretation. Let

$$\bar{x}_k := \frac{1}{4}\sum_{i=1}^{4} x_i^k, \qquad \mathcal{C}_k := \left(\frac{1}{4}\sum_{i=1}^{4}(x_i^k - \bar{x}_k)^2\right)^{1/2},$$

where $\mathcal{C}_k$ denotes the consensus root-mean-square (RMS) error. Figure 1(a) plots $g_k := |f'(\bar{x}_k)|$ at every update $0 \le k \le 2048$, without smoothing or restricting evaluation to communication times. A float64 continuation through 131,072 updates gives the following temporal statistics over $k = 98{,}305, \ldots, 131{,}072$:

H	Mean g_k	Minimum g_k	Maximum g_k	Mean $\mathcal{C}_k$
1	0	0	0	0
64	0.154318	0.140893	0.166965	0.237861

These are temporal summaries of deterministic trajectories, not seed-based uncertainty estimates. An independent implementation of (16) reproduces the reported values to the shown precision.

For $H = 1$, complete averaging preserves exact consensus. Under the stated initialization and recurrence, the average momentum remains positive while the common iterate approaches the flat minimizer set, so the normalized update moves the iterate by -0.01 per step and reaches this set after 100 updates. Subsequent motion within the flat region is compatible with $f'(\bar{x}_k) = 0$ and therefore should not be interpreted as convergence of the parameter iterates. By contrast, for $H = 64$, both the global-gradient residual g_k and the consensus error $\mathcal{C}_k$ remain nonzero over the reported late window. Thus, under this fixed-step deterministic configuration, periodic communication produces different finite-time dynamics from communication at every update. The example does not imply that such residuals persist under other parameter choices, diminishing stepsizes, or local-update schemes.

Selection and sensitivity. The reported example was selected through an exploratory search consisting of 1,296 screening trajectories followed by 72 longer verification trajectories. As a numerical precision check, the selected recurrence was also evaluated through 8,192 updates using both 40- and 80-digit arithmetic, with agreement over the reported trajectory.

To assess local sensitivity around the selected configuration, we evaluated 90 combinations of initialization, stepsize, and momentum, formed from

$$\{0.95, 0.99, 1, 1.01, 1.05, 1.9, 1.98, 2, 2.02, 2.1\}, \quad \{0.0095, 0.01, 0.0105\}, \quad \{0.45, 0.5, 0.55\},$$

respectively. Of these 90 configurations, 70 satisfied the prescribed separation criteria over the final quarter of 32,768 updates:

$$\max_k g_k^{H=1} \le 10^{-10}, \qquad \min_k g_k^{H=64} > 0.01, \qquad \mathrm{mean}_k\, g_k^{H=64} > 0.05,$$

where the extrema and mean are taken over that late window.

These counts describe local sensitivity around the selected example and should not be interpreted as statistical confidence statements. Larger changes in the initialization or stepsize, including the use of diminishing stepsizes, can remove the observed residual. The diagnostic therefore documents a finite-time difference between the two communication schedules under the stated constant-step configurations; it does not establish a general failure of periodic local updates or a phenomenon specific to heavy-tailed noise.

B Related Work and Theoretical Comparisons

B.1 Connections to prior work

Robust mean estimation and estimation bias. Median-of-means (MoM) and related estimators enable mean estimation under weak moment assumptions (Hsu & Sabato, 2016; Lugosi & Mendelson, 2019). In optimization, their bias also matters: median-based aggregation can introduce systematic error even without adversarial observations (Karimireddy et al., 2021). Puchkin et al. (2024)

analyze a smoothed MoM estimator under additional regularity conditions on the noise density. More closely related, Sun et al. (2026) use the online linear correction

$$\widetilde{g}_k = (1-\eta_k)\widehat{g}_k + \eta_k \bar{g}_k = \widehat{g}_k + \eta_k(\bar{g}_k - \widehat{g}_k),$$

where $\bar{g}_k$ averages the current block means. Their correction uses an untruncated residual with $\eta_k \to 0$, and their almost-sure convergence analysis for strongly monotone games uses a growing batch size.

Our estimator instead uses the bounded same-batch correction

$$Z_\kappa = R + \mathrm{clip}_\kappa(U - R), \qquad \|Z_\kappa - R\| \le \kappa,$$

where R and U are computed from the same fixed batch of $N = 3m$ observations. The distinction is therefore not the general idea of combining a robust anchor with mean information, but the bounded nonlinear correction and its fixed-sample bias–MSE guarantee.

Median-of-three aggregation has also been analyzed for CountSketch coordinate estimates (Larsen et al., 2021). In our setting, the three inputs are independent sample block means, and their coordinatewise median admits a Euclidean MSE bound without a dimension factor. Our lower bound adapts change-of-expectation and constrained-risk arguments (Brown & Low, 1996; Duchi & Ruan, 2021; Derumigny & Schmidt-Hieber, 2023) to the finite-p moment class considered here. The corrected estimator matches this lower bound up to constants depending only on p over the range in Theorem 2.

Stochastic optimization under heavy-tailed noise. Clipping and normalization are widely used to control heavy-tailed stochastic gradients. Hübler et al. (2025) analyze normalized SGD under finite-p moments, including parameter-free and tuned regimes, while Liu & Zhou (2025) establish nonconvex guarantees for normalized stochastic methods without gradient clipping. Methods based on clipping, sign, and normalization have also been studied under other noise models, including settings with symmetry assumptions (Jakovetić et al., 2023; Armacki et al., 2025). In federated settings, Yang et al. (2022) study heavy-tailed noise with local computation, and Lee et al. (2025) consider adaptive local updates with server-based aggregation. CRAFT instead uses neighbor-to-neighbor communication and separately tracks the conditional bias and MSE of its corrected gradient estimator through the decentralized dynamics.

Decentralized optimization and local gradient tracking. GT-NSGDm combines normalized updates, momentum, and gradient tracking under conditionally centered finite-p noise (Yu et al., 2026). Its known-p result achieves an expected-gradient-norm rate of $T^{-(p-1)/(3p-2)}$ using one stochastic gradient sample and a constant number of neighbor exchanges per update. Thus, increasing the batch size is not necessary for expected-gradient-norm convergence under heavy-tailed noise. Our focus is instead on combining a robust estimator with an explicit bias–MSE trade-off with multiple local updates between communications. The normalized directions in GT-NSGDm and the fixed-radius clipped directions in CRAFT also require different descent analyses.

Under finite-variance assumptions, LU-GT, K-GT, and LSGT develop gradient-tracking schemes with multiple local updates between communication rounds (Nguyen et al., 2023; Liu et al., 2025; Ge & Chang, 2023); subsequent work further studies the effects of local updates and data heterogeneity (Wu et al., 2025). Recent heavy-tailed decentralized methods use normalization together with multiple consensus steps per update to obtain refined sample and communication bounds (Wang et al., 2026). Multiple consensus steps and multiple local updates address different communication–computation trade-offs: the former performs additional communication within an update, whereas the latter performs additional local computation between communications. CRAFT studies the latter setting with a fixed batch size per update, allowing the local period to grow with the prescribed horizon.

B.2 Stationarity Criteria and Complexity Conversions

Table 1 compares orders in the target accuracy ε, with the number of agents and all problem- and network-dependent parameters fixed. Its common stationarity criterion is the expected norm of the global gradient ∇f at a uniformly sampled local model. For CRAFT, (14) uses an independent uniform agent I and uniform pre-update index $(\widehat{t}, \widehat{h})$. This criterion differs from a squared-gradient

guarantee, a network-average stationarity measure, and a terminal-iterate guarantee. We next state the corresponding criteria and complexity conversions for the methods compared in Section 5.

GT-NSGDm. Under smooth local objectives, a lower-bounded global objective, conditionally centered finite-p noise, and the oracle-independence conditions stated in Yu et al. (2026), Theorem 1 gives, with its prescribed horizon-dependent tuning and for sufficiently large T_G,

$$\frac{1}{nT_G}\sum_{t=0}^{T_G-1}\sum_{i=1}^{n}\mathbb{E}\left\|\nabla f(x_i^t)\right\| = O\Big(T_G^{-\frac{p-1}{3p-2}}\Big). \tag{17}$$

The left-hand side is exactly the expected gradient norm at an independently and uniformly sampled agent–time pair, matching the output criterion used for CRAFT. GT-NSGDm uses one fresh sample and a constant number of neighbor exchanges per update. Hence, with $s_p := (3p-2)/(p-1)$, (17) gives the sufficient orders $O(\varepsilon^{-s_p})$ for updates, samples per agent, and communications, with the remaining parameters fixed. Its mixing matrix is primitive and doubly stochastic, but need not be symmetric.

DNSGD on an undirected network. Theorem 2 of Wang et al. (2026) assumes smooth local objectives, a lower-bounded global objective, centered finite-p noise, independent oracle samples, and a symmetric positive-semidefinite doubly stochastic mixing matrix. For each agent i, it guarantees $\mathbb{E}\left\|\nabla f(\widehat{x}_i)\right\| \le \varepsilon$, where $\widehat{x}_i$ is uniformly sampled from that agent's iterates. With the network contraction $\beta_D < 1$ and the remaining parameters fixed, the prescribed choices have orders

$$T_D = \Theta(\varepsilon^{-2}), \qquad B_D = \Theta\Big(n^{-1}\varepsilon^{-\frac{p}{p-1}}\Big), \qquad G_D = \widetilde{\Theta}\Big((1-\beta_D)^{-1/2}\Big),$$

where B_D is the batch size per agent and G_D is the number of accelerated gossip steps per update. For fixed n and sufficiently small ε, this yields per-agent sample complexity $O(n^{-1}\varepsilon^{-s_p})$ and communication complexity $\widetilde{O}(\varepsilon^{-2})$ on a fixed network. The source theorem reports the system-wide sample count; Table 1 reports the corresponding per-agent order. Unlike CRAFT, DNSGD uses an accuracy-dependent batch size and multiple consensus steps per update.

CRAFT. For CRAFT, Corollary 1 with $H = \lceil T^{\chi} \rceil$ and $\chi = p/(5p-4)$ gives

$$TH = O(\varepsilon^{-s_p}), \qquad T = O\Big(\varepsilon^{-\frac{5p-4}{2(p-1)}}\Big), \qquad N_{\mathrm{raw}} = 3mTH \quad \text{per agent.}$$

Here the batch size $3m$ remains fixed, while T counts communication rounds and TH counts local gradient queries. Oracle samples need not be independent across agents, and the doubly stochastic mixing matrix need not be symmetric. In addition to the randomly sampled-agent guarantee used in Table 1, (97) gives, for fixed n,

$$\mathbb{E}\max_{i\in[n]}\left\|\nabla f(x_{i,\widehat{t},\widehat{h}})\right\| = O\Big((TH)^{-\frac{p-1}{3p-2}}\Big),$$

with an n-dependent constant. Thus, the three table entries share a comparable expected local-model stationarity criterion, but differ in their oracle assumptions, network conditions, batching strategies, and communication protocols.

Finite-variance local gradient tracking. K-GT and LSGT are discussed separately because their stationarity criteria differ from the local-model criterion used in Table 1. Under bounded oracle variance, Theorem 3.2 of Liu et al. (2025) controls a round-averaged squared global gradient at the network-average model. With the remaining problem and network parameters fixed, converting its squared-gradient target to expected gradient-norm accuracy ε yields the communication order

$$O\Big(H^{-1}\varepsilon^{-4} + H^{-1/2}\varepsilon^{-3} + \varepsilon^{-2}\Big).$$

This result also uses the network and initialization conditions stated in Liu et al. (2025), including a symmetric doubly stochastic mixing matrix.

Under bounded variance, Corollary 1 of Ge & Chang (2023) instead bounds the squared aggregate local-gradient residual at the round-start models y_i^t:

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(y_i^t)\right\|^2 = O\Big((nHT)^{-1/2}\Big), \qquad H \le (T/n^5)^{1/3},$$

under its prescribed stepsize and horizon conditions. Neither criterion directly yields the expected global-gradient norm at a uniformly sampled local model used in Table 1; we therefore do not include K-GT or LSGT as rows in that table.

C Fixed-Sample Bias–MSE Lower Bound

Throughout this section, let $p \in (1, 2)$, $\sigma > 0$, and $N, d \geq 1$. Recall the moment class $\mathcal{P}_{p,\sigma}^d$ and the uniform bias and worst-case MSE, $B(T_N)$ and $Q(T_N)$, defined in (2)–(3). We consider measurable estimators that are integrable under every $P \in \mathcal{P}_{p,\sigma}^d$. Any auxiliary randomization is independent of the observations and has a law that does not depend on P. For brevity, write $B := B(T_N)$ and $Q := Q(T_N)$.

C.1 A two-point inequality

Lemma 1 (Two-point bias–MSE inequality)**.** *Suppose $B < \infty$. For any $\varepsilon \in (0, 1/2]$, define*

$$\Delta_\varepsilon := 2^{-1/p}\sigma\varepsilon^{(p-1)/p}, \qquad r_\varepsilon := 1 - (1-\varepsilon)^N.$$

Then

$$Q \geq \frac{1 - r_\varepsilon}{r_\varepsilon}\big[\Delta_\varepsilon - 2B\big]_+^2, \qquad [z]_+ := \max\{z, 0\}. \tag{18}$$

Proof. The claim is immediate if $Q = \infty$, so suppose $Q < \infty$. Fix a unit vector $u \in \mathbb{R}^d$ and set $a_\varepsilon := 2^{-1/p}\sigma\varepsilon^{-1/p}$. Consider

$$P_0 := \delta_0, \qquad P_\varepsilon := (1-\varepsilon)\delta_0 + \varepsilon\delta_{a_\varepsilon u}.$$

Their means are 0 and $\Delta_\varepsilon u$, respectively. Moreover, $P_0, P_\varepsilon \in \mathcal{P}_{p,\sigma}^d$, since

$$\begin{aligned}\mathbb{E}_{P_\varepsilon}\|V - \Delta_\varepsilon u\|^p &= a_\varepsilon^p\left[(1-\varepsilon)\varepsilon^p + \varepsilon(1-\varepsilon)^p\right] \\ &\leq 2\varepsilon a_\varepsilon^p = \sigma^p.\end{aligned}$$

Let $V_1, \ldots, V_N$ be i.i.d. samples from P_ε, and let A be the event that at least one observation is nonzero. Then $\Pr(A) = 1 - (1-\varepsilon)^N = r_\varepsilon =: r$. Conditional on A^c, all observations are zero, so the conditional distribution of T_N is the same as its distribution under N i.i.d. samples from P_0, including any independent auxiliary randomization.

Let

$$S := \langle T_N, u\rangle, \qquad t_0 := \mathbb{E}_{P_0} S, \qquad t_\varepsilon := \mathbb{E}_{P_\varepsilon} S, \qquad t_A := \mathbb{E}_{P_\varepsilon}[S \mid A].$$

By the definition of B, $|t_0| \leq B$ and $|t_\varepsilon - \Delta_\varepsilon| \leq B$, so

$$|t_\varepsilon - t_0| \geq \big[\Delta_\varepsilon - 2B\big]_+. \tag{19}$$

Since $t_\varepsilon = (1-r)t_0 + rt_A$, the law of total variance yields

$$\begin{aligned}Q &\geq \mathbb{E}_{P_\varepsilon}(S - \Delta_\varepsilon)^2 \geq \mathrm{Var}_{P_\varepsilon}(S) \\ &\geq r(1-r)(t_A - t_0)^2 = \frac{1-r}{r}(t_\varepsilon - t_0)^2.\end{aligned}$$

Combining this inequality with (19) proves (18). Since the two distributions are supported on the one-dimensional subspace spanned by u, the bound is independent of the ambient dimension. □

This two-point argument is a direct instance of constrained-risk methods for bias–variance lower bounds (Brown & Low, 1996; Duchi & Ruan, 2021; Derumigny & Schmidt-Hieber, 2023).

C.2 Proof of Theorem 1

Proof. Let $s_N := \sigma N^{-(p-1)/p}$ and $A_N := s_N^p = \sigma^p N^{1-p}$. First suppose $B = 0$. By Lemma 1,

$$Q \geq \frac{1 - r_\varepsilon}{r_\varepsilon} \Delta_\varepsilon^2.$$

As $\varepsilon \downarrow 0$, $r_\varepsilon \sim N\varepsilon$, and hence

$$\frac{1 - r_\varepsilon}{r_\varepsilon} \Delta_\varepsilon^2 \sim \frac{2^{-2/p}\sigma^2}{N} \varepsilon^{(p-2)/p} \longrightarrow \infty,$$

because $p < 2$. Therefore, $Q = \infty$.

Now suppose $0 < B < \infty$. If $Q = \infty$, the result is immediate, so assume $Q < \infty$. We consider two cases.

(i) If $B \leq s_N/8$, choose

$$\varepsilon := \left(\frac{4\, 2^{1/p} B}{\sigma} \right)^{p/(p-1)}.$$

Then $\varepsilon \leq 1/(2N)$, $\Delta_\varepsilon = 4B$, and $r_\varepsilon \leq N\varepsilon \leq 1/2$. Therefore, Lemma 1 gives

$$Q \geq \frac{2B^2}{N\varepsilon} = 2(4\, 2^{1/p})^{-p/(p-1)} \frac{\sigma^{p/(p-1)}}{N B^{(2-p)/(p-1)}}.$$

Raising this inequality to the power $p - 1$ and multiplying by B^{2-p} yields

$$B^{2-p} Q^{p-1} \geq 2^{-p-2} \sigma^p N^{1-p} \geq 8^{-p} \sigma^p N^{1-p}.$$

(ii) If $B > s_N/8$, Jensen's inequality (4) gives $Q \geq B^2$, and therefore

$$B^{2-p} Q^{p-1} \geq B^p > 8^{-p} s_N^p = 8^{-p} \sigma^p N^{1-p}.$$

This proves (5). Finally, if $B = \infty$, (4) implies $Q = \infty$. □

C.3 Scope of the Gradient-Oracle Implication

The mean-estimation problem above can be embedded directly into smooth optimization with a lower-bounded objective. Fix $x_0 \in \mathbb{R}^d$ and $L > 0$. For any $P \in \mathcal{P}_{p,\sigma}^d$, define

$$f_P(x) := \frac{L}{2} \|x - x_0\|^2 + \langle \mu_P, x - x_0 \rangle, \qquad g_P(x; V) := L(x - x_0) + V, \qquad V \sim P.$$

Then $\nabla f_P(x) = L(x - x_0) + \mu_P$, so f_P is L-smooth and bounded below, while

$$g_P(x; V) - \nabla f_P(x) = V - \mu_P$$

has p-th moment at most σ^p. At x_0, $g_P(x_0; V) = V$, so a fresh oracle batch consists exactly of i.i.d. observations from P.

Consequently, any common gradient-estimation rule that uses only a fresh batch of N oracle samples, without prior knowledge of P or f_P, together with independent auxiliary randomness, is subject to Theorem 1. The same conclusion applies to rules using at most N fresh samples, since additional samples may simply be ignored. This implication is specific to batch-only estimation: if an estimator also uses informative observations accumulated from previous queries, its full information must be included in the statistical model, and the theorem cannot in general be applied with N equal only to the number of fresh samples.

Finally, for any P under which T_N has finite MSE,

$$\mathbb{E}_P \|T_N - \mu_P\|^2 = \mathbb{E}_P \|T_N - \mathbb{E}_P T_N\|^2 + \|\mathbb{E}_P T_N - \mu_P\|^2.$$

Hence, if the estimator has uniform bias at most b and uniform centered second moment at most v, then its worst-case MSE is at most $v + b^2$. Theorem 1 therefore implies

$$b^{2-p}(v + b^2)^{p-1} \geq 8^{-p} \sigma^p N^{1-p}.$$

This remains a lower bound for fixed-sample estimation; it is not an optimization or communication lower bound.

D CORRECTED ESTIMATION AND RECURSIVE FILTERING

D.1 CONDITIONAL SAMPLING AND INTEGRABILITY

Let the algorithm be defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$. The problem data, mixing matrix W, algorithmic parameters, and initial point x_0 are deterministic, with $\boldsymbol{X}_0 = \mathbf{1}_n x_0^\top$ and $M_0 = Y_0 = 0$.

For each communication round t and local step h, let $\mathcal{F}_{t,h}$ denote the σ-field generated by all algorithmic states and oracle samples available before the fresh batch at step (t, h) is drawn. In particular, the boundary mixing at the beginning of round t is measurable with respect to $\mathcal{F}_{t,0}$, and

$$\boldsymbol{X}_{t,h+1} = \boldsymbol{X}_{t,h} - \eta \operatorname{clip}_C(Y_{t,h})$$

is $\mathcal{F}_{t,h}$-measurable. Thus, conditional on $\mathcal{F}_{t,h}$, the query point is fixed before the new samples are drawn, and Assumption 2 applies with $\mathcal{F} = \mathcal{F}_{t,h}$. At each agent, the $N = 3m$ fresh samples are conditionally independent given $\mathcal{F}_{t,h}$. After adjoining these samples and performing the estimator, filter, and tracker updates, the resulting states are included in $\mathcal{F}_{t,h+1}$. At the end of a local block, the corresponding states generate the history for the next communication round. Random indices used to select the output are sampled independently of the algorithmic history.

All conditional bounds below concern oracle and estimator errors centered at the true gradient at the current query point. Assumption 2 ensures conditional integrability of the raw centered errors, without requiring an unconditional moment bound on gradients evaluated at the random iterates. The results below establish the corresponding integrability and moment bounds for the median and corrected estimator before their conditional biases are invoked.

D.2 CONDITIONAL MSE OF THE MEDIAN ANCHOR

For a batch of $N = 3m$ stochastic gradients at agent i and query point x, partition the samples into three blocks of size m and define

$$\bar{g}_{i,b}(x) := \frac{1}{m} \sum_{r=1}^{m} g_i(x; \xi_{i,b,r}), \qquad b = 1, 2, 3.$$

Let $R_i(x)$ be the coordinatewise median of the three block means, and define their pooled mean by

$$U_i(x) := \frac{1}{3} \sum_{b=1}^{3} \bar{g}_{i,b}(x).$$

Thus, $R_i(x)$ and $U_i(x)$ are the gradient-query counterparts of R and U in (6)–(7).

Lemma 2 (Conditional vector moment bound)**.** *Let $p \in (1, 2]$ and $\sigma \geq 0$. Suppose $V_1, \ldots, V_N \in \mathbb{R}^d$ are conditionally independent given $\mathcal{F}$ and satisfy, almost surely,*

$$\mathbb{E}[V_r \mid \mathcal{F}] = 0, \qquad \mathbb{E}[\|V_r\|^p \mid \mathcal{F}] \leq \sigma^p, \qquad r = 1, \ldots, N.$$

Then

$$\mathbb{E}\left[\left\| \frac{1}{N} \sum_{r=1}^{N} V_r \right\|^p \middle| \mathcal{F} \right] \leq 2^p \sigma^p N^{1-p}.$$

Proof. On an extension of the probability space, let $V_1', \ldots, V_N'$ be conditional copies of $V_1, \ldots, V_N$ such that, given $\mathcal{F}$, the variables V_r' are independent of each other and of the original batch, with V_r' having the same conditional law as V_r. Let $\epsilon_1, \ldots, \epsilon_N$ be independent Rademacher variables, independent of $\mathcal{F}$ and of both batches.

Since $\mathbb{E}[V_r' \mid \mathcal{F}] = 0$, conditional Jensen's inequality gives

$$\begin{aligned} \mathbb{E}\left[\left\| \sum_{r=1}^{N} V_r \right\|^p \middle| \mathcal{F} \right] &= \mathbb{E}\left[\left\| \mathbb{E}\left[\sum_{r=1}^{N} (V_r - V_r') \middle| \mathcal{F}, V_1, \ldots, V_N \right] \right\|^p \middle| \mathcal{F} \right] \\ &\leq \mathbb{E}\left[\left\| \sum_{r=1}^{N} (V_r - V_r') \right\|^p \middle| \mathcal{F} \right]. \end{aligned}$$

Conditional on $\mathcal{F}$, the differences $V_r - V_r'$ are independent and symmetric. Hence their joint conditional law is invariant under independent sign changes, and therefore

$$\mathbb{E}\left[\left\|\sum_{r=1}^{N}(V_r - V_r')\right\|^p \middle| \mathcal{F}\right] = \mathbb{E}\left[\mathbb{E}_\epsilon\left\|\sum_{r=1}^{N}\epsilon_r(V_r - V_r')\right\|^p \middle| \mathcal{F}\right].$$

For deterministic vectors $w_1, \ldots, w_N$, since $p/2 \le 1$, Jensen's inequality gives

$$\mathbb{E}_\epsilon\left\|\sum_{r=1}^{N}\epsilon_r w_r\right\|^p = \mathbb{E}_\epsilon\left(\left\|\sum_{r=1}^{N}\epsilon_r w_r\right\|^2\right)^{p/2} \le \left(\mathbb{E}_\epsilon\left\|\sum_{r=1}^{N}\epsilon_r w_r\right\|^2\right)^{p/2}.$$

Expanding the squared norm and using $\mathbb{E}_\epsilon[\epsilon_r\epsilon_s] = 0$ for $r \ne s$ gives

$$\mathbb{E}_\epsilon\left\|\sum_{r=1}^{N}\epsilon_r w_r\right\|^2 = \sum_{r=1}^{N}\|w_r\|^2.$$

Consequently,

$$\mathbb{E}_\epsilon\left\|\sum_{r=1}^{N}\epsilon_r w_r\right\|^p \le \left(\sum_{r=1}^{N}\|w_r\|^2\right)^{p/2} \le \sum_{r=1}^{N}\|w_r\|^p, \tag{20}$$

where the last inequality follows from $(\sum_r a_r)^{p/2} \le \sum_r a_r^{p/2}$ for $a_r \ge 0$ and $p/2 \le 1$.

Applying (20) with $w_r = V_r - V_r'$ and using

$$\|u - v\|^p \le 2^{p-1}\big(\|u\|^p + \|v\|^p\big)$$

yields

$$\begin{aligned}\mathbb{E}\left[\left\|\sum_{r=1}^{N}V_r\right\|^p \middle| \mathcal{F}\right] &\le \sum_{r=1}^{N}\mathbb{E}\left[\|V_r - V_r'\|^p \middle| \mathcal{F}\right] \\ &\le 2^{p-1}\sum_{r=1}^{N}\big(\mathbb{E}[\|V_r\|^p \mid \mathcal{F}] + \mathbb{E}[\|V_r'\|^p \mid \mathcal{F}]\big) \\ &\le 2^p N\sigma^p.\end{aligned}$$

Dividing by N^p proves the claim. The argument uses only vector norms and conditional independence across samples; no independence among coordinates is required. □

Lemma 3 (Conditional MSE of the median anchor)**.** *Under Assumption 2, let $m \ge 1$ and let x be any $\mathcal{F}$-measurable query point. Then, for every agent i,*

$$\mathbb{E}\left[\|R_i(x) - \nabla f_i(x)\|^2 \middle| \mathcal{F}\right] \le q_{\mathrm{MoM}} := 12\sigma^2 m^{-2(p-1)/p}. \tag{21}$$

Proof. Fix an agent i and condition on $\mathcal{F}$. Define the three block errors

$$B_b := \bar{g}_{i,b}(x) - \nabla f_i(x), \qquad b = 1, 2, 3.$$

For any $a, b, c \in \mathbb{R}$,

$$|\operatorname{med}(a, b, c)|^2 \le |ab| + |ac| + |bc|.$$

Indeed, if the median is nonzero, at least two of a, b, c have the same sign as the median and absolute value at least its absolute value; the case of a zero median is immediate. Applying this inequality coordinatewise and then Cauchy–Schwarz gives

$$\|R_i(x) - \nabla f_i(x)\|^2 \le \|B_1\|\,\|B_2\| + \|B_1\|\,\|B_3\| + \|B_2\|\,\|B_3\|. \tag{22}$$

Each B_b is the average of m conditionally independent, conditionally centered gradient errors. Hence, by Lemma 2,

$$\mathbb{E}[\|B_b\|^p \mid \mathcal{F}] \le 2^p\sigma^p m^{1-p}, \qquad b = 1, 2, 3.$$

Conditional Jensen's inequality gives

$$\mathbb{E}[\|B_b\| \mid \mathcal{F}] \le (\mathbb{E}[\|B_b\|^p \mid \mathcal{F}])^{1/p} \le 2\sigma m^{-(p-1)/p}.$$

Because the three blocks are conditionally independent given $\mathcal{F}$, their errors B_1, B_2, B_3 are also conditionally independent. Therefore, for $b \neq b'$,

$$\mathbb{E}[\|B_b\| \, \|B_{b'}\| \mid \mathcal{F}] = \mathbb{E}[\|B_b\| \mid \mathcal{F}] \, \mathbb{E}[\|B_{b'}\| \mid \mathcal{F}] \le 4\sigma^2 m^{-2(p-1)/p}.$$

Taking conditional expectations in (22) and summing the three cross terms yields

$$\mathbb{E}\left[\|R_i(x) - \nabla f_i(x)\|^2 \middle| \mathcal{F}\right] \le 12\sigma^2 m^{-2(p-1)/p} = q_{\mathrm{MoM}},$$

which proves (21).

The argument uses independence across the three sample blocks, but not across coordinates. It also does not require the samples or blocks to be identically distributed, provided the conditional centering and p-th-moment bound hold uniformly. In particular, the bound is independent of the dimension d. □

D.3 Conditional bias and MSE bounds

Lemma 4 (Conditional bias and MSE of the corrected query)**.** *Under Assumption 2, let $m \ge 1$, $N = 3m$, and $\kappa > 0$. Define*

$$C_{q,p} := 2^2\, 3^{2(p-1)/p+1}, \qquad C_{a,p} := 2^{p-1}\big(2^p + C_{q,p}^{p/2}\big), \qquad C_p := 2(C_{q,p} + C_{a,p}), \tag{23}$$

and let b_κ and q_κ be defined by (8) *with this choice of C_p. For the intermediate bounds, set*

$$q_{\mathrm{MoM}} := 12\sigma^2 m^{-2(p-1)/p}, \qquad a_m := 2^{p-1}\left(2^p \sigma^p N^{1-p} + q_{\mathrm{MoM}}^{p/2}\right).$$

Then, for every agent i and every $\mathcal{F}$-measurable query point x, almost surely,

$$\|\mathbb{E}[Z_{i,\kappa}(x) - \nabla f_i(x) \mid \mathcal{F}]\| \le a_m \kappa^{1-p} \le b_\kappa, \tag{24}$$

$$\mathbb{E}[\|Z_{i,\kappa}(x) - \nabla f_i(x)\|^2 \mid \mathcal{F}] \le 2q_{\mathrm{MoM}} + 2a_m \kappa^{2-p} \le q_\kappa. \tag{25}$$

The samples within each batch need only be conditionally independent with the common conditional mean and moment bound in Assumption 2; identical conditional distributions are not required.

Proof. Fix an agent i and condition on $\mathcal{F}$. Write $g := \nabla f_i(x)$ and

$$V := U_i(x) - R_i(x).$$

Since $U_i(x)$ is the average of all $N = 3m$ samples, Lemma 2 gives

$$\mathbb{E}[\|U_i(x) - g\|^p \mid \mathcal{F}] \le 2^p \sigma^p N^{1-p}.$$

By Lemma 3, $\mathbb{E}[\|R_i(x) - g\|^2 \mid \mathcal{F}] \le q_{\mathrm{MoM}}$. Since $p \le 2$, conditional Jensen's inequality therefore yields

$$\mathbb{E}[\|R_i(x) - g\|^p \mid \mathcal{F}] \le \Big(\mathbb{E}[\|R_i(x) - g\|^2 \mid \mathcal{F}]\Big)^{p/2} \le q_{\mathrm{MoM}}^{p/2}.$$

Using $\|a - b\|^p \le 2^{p-1}(\|a\|^p + \|b\|^p)$ with $V = (U_i - g) - (R_i - g)$, we obtain

$$\mathbb{E}[\|V\|^p \mid \mathcal{F}] \le 2^{p-1}\left(2^p \sigma^p N^{1-p} + q_{\mathrm{MoM}}^{p/2}\right) = a_m. \tag{26}$$

We first bound the conditional bias. Since the pooled mean is conditionally unbiased, $\mathbb{E}[U_i(x) - g \mid \mathcal{F}] = 0$, and

$$Z_{i,\kappa}(x) - g = U_i(x) - g + \mathrm{clip}_\kappa(V) - V.$$

Hence

$$\|\mathbb{E}[Z_{i,\kappa}(x) - g \mid \mathcal{F}]\| = \|\mathbb{E}[\mathrm{clip}_\kappa(V) - V \mid \mathcal{F}]\| \le \mathbb{E}[\|\mathrm{clip}_\kappa(V) - V\| \mid \mathcal{F}].$$

By the definition of clip_κ,

$$\|\mathrm{clip}_\kappa(v) - v\| = (\|v\| - \kappa)_+ \le \kappa^{1-p} \|v\|^p, \qquad v \in \mathbb{R}^d.$$

Combining this inequality with (26) gives

$$\|\mathbb{E}[Z_{i,\kappa}(x) - g \mid \mathcal{F}]\| \le a_m \kappa^{1-p}. \tag{27}$$

For the MSE, write $Z_{i,\kappa}(x) - g = (R_i(x) - g) + \mathrm{clip}_\kappa(V)$. Then

$$\begin{aligned}\mathbb{E}[\|Z_{i,\kappa}(x) - g\|^2 \mid \mathcal{F}] &\le 2\mathbb{E}[\|R_i(x) - g\|^2 \mid \mathcal{F}] + 2\mathbb{E}[\|\mathrm{clip}_\kappa(V)\|^2 \mid \mathcal{F}] \\ &\le 2q_{\mathrm{MoM}} + 2a_m \kappa^{2-p},\end{aligned} \tag{28}$$

where the second inequality uses

$$\|\mathrm{clip}_\kappa(v)\|^2 \le \kappa^{2-p} \|v\|^p, \qquad v \in \mathbb{R}^d.$$

Indeed, if $\|v\| \le \kappa$, then $\|v\|^2 \le \kappa^{2-p} \|v\|^p$, while if $\|v\| > \kappa$, then $\|\mathrm{clip}_\kappa(v)\|^2 = \kappa^2 \le \kappa^{2-p} \|v\|^p$.

It remains to compare the intermediate bounds with b_κ and q_κ. Since $N = 3m$,

$$q_{\mathrm{MoM}} = C_{q,p} \sigma^2 N^{-2(p-1)/p}, \qquad a_m = C_{a,p} \sigma^p N^{1-p}.$$

Therefore, by the definition $C_p = 2(C_{q,p} + C_{a,p})$, (27)–(28) imply (24)–(25).

No independence between $R_i(x)$ and $U_i(x)$ is used; both are computed from the same batch. □

The preceding bounds extend directly to the stacked estimator. For $\boldsymbol{X} = [x_1^\top; \ldots; x_n^\top]$, let $Z_\kappa(\boldsymbol{X})$ denote the matrix whose ith row is $Z_{i,\kappa}(x_i)^\top$. Averaging the row-wise bounds gives

$$\|\mathbb{E}[Z_\kappa(\boldsymbol{X}) - G(\boldsymbol{X}) \mid \mathcal{F}]\|_n \le b_\kappa, \qquad \mathbb{E}[\|Z_\kappa(\boldsymbol{X}) - G(\boldsymbol{X})\|_n^2 \mid \mathcal{F}] \le q_\kappa.$$

No independence across agents is needed for these inequalities. Moreover, since J and Π are orthogonal projections, $\|JA\|_n \le \|A\|_n$ and $\|\Pi A\|_n \le \|A\|_n$. Hence, the same MSE upper bound applies after either projection.

For the convergence analysis, define the predictable query bias and centered residual by

$$\mu_{t,h+1} := \mathbb{E}[Z_{t,h+1} - G(\boldsymbol{X}_{t,h+1}) \mid \mathcal{F}_{t,h}], \qquad \xi_{t,h+1} := Z_{t,h+1} - G(\boldsymbol{X}_{t,h+1}) - \mu_{t,h+1}.$$

Then

$$\|\mu_{t,h+1}\|_n \le b_\kappa, \qquad \mathbb{E}[\xi_{t,h+1} \mid \mathcal{F}_{t,h}] = 0, \qquad \mathbb{E}[\|\xi_{t,h+1}\|_n^2 \mid \mathcal{F}_{t,h}] \le q_\kappa.$$

The last inequality follows from the conditional variance decomposition: centering by the conditional mean cannot increase the conditional second moment. These are the bounds used in Appendix F. In particular, conditional centering of the raw oracle samples does not imply that the median or corrected estimator is conditionally unbiased.

D.4 Matching bounds and minimax MSE

We first give explicit constants for Theorem 2. Let $p \in (1, 2)$, $\sigma > 0$, and $N = 3m$ with $m, d \ge 1$. Define

$$s_N := \sigma N^{-(p-1)/p}, \qquad A_N := \sigma^p N^{1-p} = s_N^p, \qquad c_p := 8^{-p},$$

and recall $C_{q,p}$, $C_{a,p}$, and C_p from (23). For every finite $\kappa \ge s_N$,

$$\left(\frac{c_p}{C_p^{p-1}}\right)^{1/(2-p)} A_N \kappa^{1-p} \le B(Z_\kappa) \le C_{a,p} A_N \kappa^{1-p}, \tag{29}$$

$$\left(\frac{c_p}{C_{a,p}^{2-p}}\right)^{1/(p-1)} A_N \kappa^{2-p} \le Q(Z_\kappa) \le C_p A_N \kappa^{2-p}. \tag{30}$$

Consequently,

$$c_p A_N \le B(Z_\kappa)^{2-p} Q(Z_\kappa)^{p-1} \le C_{a,p}^{2-p} C_p^{p-1} A_N. \tag{31}$$

Proof of Theorem 2. Since $N = 3m$, Lemma 4 gives

$$q_{\mathrm{MoM}} = C_{q,p} s_N^2, \qquad a_m = C_{a,p} A_N.$$

Therefore,

$$B(Z_\kappa) \le C_{a,p} A_N \kappa^{1-p}, \qquad Q(Z_\kappa) \le 2C_{q,p} s_N^2 + 2C_{a,p} A_N \kappa^{2-p}.$$

If $\kappa \ge s_N$, then

$$s_N^2 \le A_N \kappa^{2-p},$$

because $A_N = s_N^p$. Hence

$$Q(Z_\kappa) \le C_p A_N \kappa^{2-p}, \qquad C_p = 2(C_{q,p} + C_{a,p}),$$

which proves the upper bounds in (29)–(30).

These upper bounds show that $Q(Z_\kappa) < \infty$. By Theorem 1, this rules out $B(Z_\kappa) = 0$ and gives

$$B(Z_\kappa)^{2-p} Q(Z_\kappa)^{p-1} \ge c_p A_N. \tag{32}$$

Set

$$b_* := A_N \kappa^{1-p}, \qquad q_* := A_N \kappa^{2-p}.$$

Since $b_*^{2-p} q_*^{p-1} = A_N$, (32) becomes

$$\left(\frac{B(Z_\kappa)}{b_*}\right)^{2-p} \left(\frac{Q(Z_\kappa)}{q_*}\right)^{p-1} \ge c_p.$$

Combining this inequality first with $B(Z_\kappa) \le C_{a,p} b_*$ and then with $Q(Z_\kappa) \le C_p q_*$ yields

$$Q(Z_\kappa) \ge \left(\frac{c_p}{C_{a,p}^{2-p}}\right)^{1/(p-1)} q_*, \qquad B(Z_\kappa) \ge \left(\frac{c_p}{C_p^{p-1}}\right)^{1/(2-p)} b_*.$$

These are the lower bounds in (29)–(30). Finally, multiplying the two upper bounds with exponents $2-p$ and $p-1$ gives the upper inequality in (31). □

The quantities $B(Z_\kappa)$ and $Q(Z_\kappa)$ are separate suprema over $\mathcal{P}_{p,\sigma}^d$ and need not be attained, or approached, by the same distributions. The theorem also does not imply that Z_κ is biased under every distribution. For example, if $P = \delta_\mu$, then $Z_\kappa = \mu$ almost surely. More generally, if the centered joint batch distribution is invariant under simultaneous sign reversal, then translation equivariance of R and U, together with the oddness of the coordinatewise median and radial clipping, makes $Z_\kappa - \mu$ an odd function of the centered batch. Whenever it is integrable, this symmetry implies $\mathbb{E}Z_\kappa = \mu$. Thus, the nonzero uniform bias arises from the worst case over the full moment class, which includes asymmetric distributions.

Corollary 2 (Bias-constrained minimax MSE)**.** *Let $p \in (1,2)$, $\sigma > 0$, $d \ge 1$, $N \ge 3$, and $b > 0$. Define*

$$\mathcal{R}_N(b) := \inf\{Q(T_N) : T_N \text{ admissible},\ B(T_N) \le b\}, \qquad S_N(b) := \frac{\sigma^{p/(p-1)}}{N b^{(2-p)/(p-1)}}.$$

Then

$$\max\left\{\frac{s_N^2}{64},\ 8^{-p/(p-1)} S_N(b)\right\} \le \mathcal{R}_N(b). \tag{33}$$

If $N = 3m$, then

$$\mathcal{R}_N(b) \le C_p s_N^2 + 2C_{a,p}^{1/(p-1)} S_N(b), \tag{34}$$

and the upper bound is attained up to the displayed constants by Z_κ with

$$\kappa := \max\left\{s_N, \left(\frac{C_{a,p} A_N}{b}\right)^{1/(p-1)}\right\}. \tag{35}$$

For arbitrary $N \ge 3$, the same upper bound holds up to an additional factor of 2.

Proof. Consider any admissible T_N with $B := B(T_N) \le b$ and $Q := Q(T_N)$. If $Q = \infty$, the lower bound is immediate. If $Q < \infty$, Theorem 1 implies $0 < B < \infty$ and

$$B^{2-p} Q^{p-1} \ge 8^{-p} A_N. \tag{36}$$

Since $B^2 \le Q$,

$$B^{2-p} \le Q^{(2-p)/2},$$

and therefore

$$Q^{p/2} \ge 8^{-p} A_N.$$

Using $A_N = s_N^p$ gives

$$Q \ge \frac{s_N^2}{64}.$$

Alternatively, using $B \le b$ in (36) yields

$$Q \ge 8^{-p/(p-1)} A_N^{1/(p-1)} b^{-(2-p)/(p-1)} = 8^{-p/(p-1)} S_N(b).$$

Taking the infimum over all admissible estimators with $B(T_N) \le b$ proves (33).

Now suppose $N = 3m$ and choose κ as in (35). Since

$$B(Z_\kappa) \le C_{a,p} A_N \kappa^{1-p},$$

this choice ensures $B(Z_\kappa) \le b$. Moreover, Lemma 4 gives

$$\begin{aligned} Q(Z_\kappa) &\le 2C_{q,p} s_N^2 + 2C_{a,p} A_N \kappa^{2-p} \\ &\le 2(C_{q,p} + C_{a,p}) s_N^2 + 2C_{a,p}^{1/(p-1)} S_N(b) \\ &= C_p s_N^2 + 2C_{a,p}^{1/(p-1)} S_N(b), \end{aligned}$$

which proves (34).

For arbitrary $N \ge 3$, let $N' := 3\lfloor N/3 \rfloor$ and apply the construction to the first N' observations, ignoring the remainder. Since $N/2 \le N' \le N$,

$$s_{N'}^2 \le 2s_N^2, \qquad S_{N'}(b) \le 2S_N(b).$$

Applying (34) with N' therefore gives the stated upper bound up to an additional factor of 2. □

D.5 Recursive filtering

CRAFT recursively averages its corrected gradient estimates according to

$$M_{t,h+1} = (1-\theta) M_{t,h} + \theta Z_{t,h+1}, \qquad 0 < \theta \le 1.$$

Because M is retained across communication rounds, the local queries can be indexed consecutively over the entire run. The following lemma isolates the effect of this recursion on the estimator bias and centered stochastic fluctuations. Variation in the underlying gradient targets is treated separately in the convergence analysis.

Lemma 5 (Exponential filtering of bias and centered fluctuations)**.** *Consider $K \ge 1$ successive corrected queries, indexed by $\ell = 1, \dots, K$, and let*

$$M_0 = 0, \qquad M_\ell = (1-\theta) M_{\ell-1} + \theta Z_\ell, \qquad 0 < \theta \le 1.$$

Let $\mathcal{F}_{\ell-1}$ denote the history before query ℓ and $\mathcal{F}_\ell$ the history after its fresh batch is observed. Under Assumption 2 and Lemma 4, write

$$Z_\ell = G_\ell + \mu_\ell + \xi_\ell,$$

where G_ℓ is the stacked local-gradient matrix at the query point, μ_ℓ is the conditional bias, and ξ_ℓ is the centered residual. Then, almost surely,

$$\|\mu_\ell\|_n \le b_\kappa, \qquad \mathbb{E}[\xi_\ell \mid \mathcal{F}_{\ell-1}] = 0, \qquad \mathbb{E}[\|\xi_\ell\|_n^2 \mid \mathcal{F}_{\ell-1}] \le q_\kappa.$$

For every $\ell = 1, \dots, K$,

$$M_\ell = \theta \sum_{j=1}^{\ell} (1-\theta)^{\ell-j} Z_j. \tag{37}$$

Moreover, the filtered bias satisfies, almost surely,

$$\left\| \theta \sum_{j=1}^{\ell} (1-\theta)^{\ell-j} \mu_j \right\|_n \le \left[1 - (1-\theta)^{\ell}\right] b_\kappa, \tag{38}$$

and the centered component satisfies

$$\mathbb{E} \left\| \theta \sum_{j=1}^{\ell} (1-\theta)^{\ell-j} \xi_j \right\|_n^2 \le \frac{\theta}{2-\theta} \left[1 - (1-\theta)^{2\ell}\right] q_\kappa. \tag{39}$$

No independence across agents or across query times is required.

Proof. Unrolling the recursion gives (37). The weights are nonnegative and satisfy

$$\theta \sum_{j=1}^{\ell} (1-\theta)^{\ell-j} = 1 - (1-\theta)^{\ell}.$$

Together with $\|\mu_j\|_n \le b_\kappa$, the triangle inequality therefore gives (38).

For the centered component, let

$$S_\ell := \theta \sum_{j=1}^{\ell} (1-\theta)^{\ell-j} \xi_j.$$

If $j < k$, then ξ_j is $\mathcal{F}_{k-1}$-measurable, and the martingale-difference property gives

$$\mathbb{E}\langle \xi_j, \xi_k \rangle_F = \mathbb{E}\left[\langle \xi_j, \mathbb{E}[\xi_k \mid \mathcal{F}_{k-1}] \rangle_F\right] = 0.$$

Thus all cross terms vanish in expectation, and

$$\begin{aligned} \mathbb{E}\,\|S_\ell\|_n^2 &= \theta^2 \sum_{j=1}^{\ell} (1-\theta)^{2(\ell-j)} \mathbb{E}\,\|\xi_j\|_n^2 \\ &\le \theta^2 q_\kappa \sum_{j=1}^{\ell} (1-\theta)^{2(\ell-j)} \\ &= \frac{\theta}{2-\theta} \left[1 - (1-\theta)^{2\ell}\right] q_\kappa. \end{aligned}$$

This proves (39). □

E Why the Median and Pooled Mean Alone Are Insufficient

The following examples isolate two limitations of the median and pooled mean that motivate the same-batch correction. They concern fixed-batch mean estimation and do not constitute lower bounds for optimization algorithms that may reuse information across queries.

A biased three-block median. Fix $p \in (1, 2)$, $\sigma > 0$, and an integer block size $m \ge 1$. Let $0 < \varepsilon \le 1/(24m^2)$ and define

$$c_\varepsilon := \sigma \left[(1-\varepsilon) + (1-\varepsilon)^p \varepsilon^{1-p}\right]^{-1/p}.$$

Let

$$\zeta = \begin{cases} -c_\varepsilon, & \text{with probability } 1-\varepsilon, \\ c_\varepsilon(1-\varepsilon)/\varepsilon, & \text{with probability } \varepsilon. \end{cases}$$

Then $\mathbb{E}\zeta = 0$ and $\mathbb{E}|\zeta|^p = \sigma^p$.

Form three independent block means, each from m i.i.d. copies of ζ, and let R_m denote their median. If a block contains no rare positive observation, its mean is $-c_\varepsilon$. If it contains at least one, its mean

is positive because $m\varepsilon < 1$. A given block contains a rare observation with probability at most $m\varepsilon$; hence, by a union bound over the three pairs of blocks, the probability that at least two blocks contain a rare observation is at most $3m^2\varepsilon^2$. On the complement of this event, $R_m = -c_\varepsilon$, while on the event itself $R_m \le c_\varepsilon/\varepsilon$. Therefore,

$$\mathbb{E}R_m \le -c_\varepsilon + 3m^2\varepsilon^2 c_\varepsilon \left(1+\varepsilon^{-1}\right) \le -\frac{3}{4}c_\varepsilon < 0.$$

Thus, even for a centered and bounded distribution, the three-block median can be biased despite having finite MSE.

An unbiased pooled mean with infinite MSE. Choose $\alpha \in (p, 2)$, let R have Pareto tail

$$\Pr(R > r) = r^{-\alpha}, \qquad r \ge 1,$$

and let S be an independent Rademacher variable. Define

$$\zeta := \sigma \left(\frac{\alpha - p}{\alpha}\right)^{1/p} SR.$$

Then ζ is centered and symmetric, with

$$\mathbb{E}|\zeta|^p = \sigma^p, \qquad \mathbb{E}\zeta^2 = \infty.$$

For any finite $N \ge 1$ and i.i.d. copies $\zeta_1, \dots, \zeta_N$, the pooled mean

$$\bar{\zeta}_N := \frac{1}{N}\sum_{j=1}^{N} \zeta_j$$

is unbiased but has infinite MSE:

$$\mathbb{E}\bar{\zeta}_N = 0, \qquad \mathbb{E}|\bar{\zeta}_N|^2 = \infty.$$

Indeed, with positive probability the sum of $\zeta_2, \dots, \zeta_N$ is bounded, while ζ_1 retains an infinite truncated second moment. Consequently, for a fresh independent batch, applying a linear recursive filter with a finite pre-query state and a nonzero pre-query-measurable gain does not restore a finite conditional MSE.

This obstruction concerns the pooled-mean estimator and its linear filtering. It does not preclude optimization guarantees obtained by subsequently clipping or normalizing the update direction and analyzing the resulting method without requiring a finite-MSE gradient estimator.

F Finite-Time Convergence and Communication Bounds

This appendix proves Theorem 3 and Corollary 1. Throughout, Assumptions 1–3 hold, $C \in (0, \infty]$, and Algorithm 1 is initialized at

$$\boldsymbol{X}_0 = \mathbf{1}_n x_0^\top, \qquad M_0 = Y_0 = 0, \tag{40}$$

where x_0 is deterministic. All algorithmic parameters are fixed within each run, and we write $K := TH$ for the total number of local updates. Define

$$s := 1+\lambda, \quad \ell := \max\{1, L\}, \quad \rho_\lambda := \frac{1+\lambda^2}{2}, \quad c_\lambda := \frac{1+\lambda^2}{1-\lambda^2}, \quad \Gamma := \frac{c_\lambda}{1-\rho_\lambda} = \frac{2(1+\lambda^2)}{(1-\lambda^2)^2}.$$

In particular, $\Gamma \ge 2$. These constants depend on the smoothness and network-contraction parameters as indicated, but not explicitly on the number of agents or the model dimension.

Stationarity and consensus quantities. For the matrix formulation (1), we perform descent on the global objective evaluated at the consensus representative:

$$F(J\boldsymbol{X}_{t,h}) = f(\bar{x}_{t,h}).$$

Thus, only the lower bound $f_{\inf} > -\infty$ is needed; no lower bound on $F(\boldsymbol{X})$ away from the consensus subspace is used. Under the standard Frobenius inner product, $\nabla_{\boldsymbol{X}} F(\boldsymbol{X}) = G(\boldsymbol{X})/n$, and the constrained first-order conditions are

$$\Pi\boldsymbol{X} = 0, \qquad JG(\boldsymbol{X}) = 0.$$

Accordingly, the analysis controls the stationarity and consensus residuals through $JG(\boldsymbol{X})$ and $\Pi\boldsymbol{X}$, measured in the normalized Frobenius norm.

Corrected-query decomposition. For each local query, write

$$Z_{t,h+1} = G(\boldsymbol{X}_{t,h+1}) + \mu_{t,h+1} + \xi_{t,h+1}, \tag{41}$$

where $\mu_{t,h+1}$ is the conditional bias given the pre-query σ-field $\mathcal{F}_{t,h}$ and $\xi_{t,h+1}$ is the corresponding centered residual. By Lemma 4 and the conditional variance decomposition,

$$\|\mu_{t,h+1}\|_n \le b_\kappa, \qquad \mathbb{E}[\xi_{t,h+1} \mid \mathcal{F}_{t,h}] = 0, \qquad \mathbb{E}[\|\xi_{t,h+1}\|_n^2 \mid \mathcal{F}_{t,h}] \le q_\kappa. \tag{42}$$

These matrix bounds follow by averaging the row-wise bounds and require no independence across agents. Moreover, centered residuals from different query times are orthogonal in expectation: if one query precedes another, the earlier residual is measurable with respect to the later pre-query history, and the tower property gives a zero expected Frobenius inner product.

Adaptivity and integrability. As detailed in Appendix D, every query model is measurable with respect to its pre-query history. In particular,

$$\boldsymbol{X}_{t,h+1} = \boldsymbol{X}_{t,h} - \eta \operatorname{clip}_C(Y_{t,h})$$

is computed before the batch used to form $Z_{t,h+1}$ is drawn. Hence Assumption 2 applies conditionally at these adaptive query points; no independence between the current iterate and the past is assumed.

The model, filter, tracker, corrected-query, and gradient-error matrices used below are square integrable over every finite horizon. This follows inductively from the deterministic initialization, the nonexpansiveness

$$\|\operatorname{clip}_C(Y)\|_n \le \|Y\|_n,$$

which also holds for $C = \infty$, the at-most-linear growth of $G(\boldsymbol{X})$ implied by smoothness, and the conditional second-moment bound in (42). The model, filter, and tracker recursions therefore preserve square integrability at each finite time. Moreover, smoothness gives at most quadratic growth of the objective relative to a fixed reference point, so the objective values are integrable along these iterates. Hence all conditional and unconditional expectations used below are well defined.

F.1 Proof sketches for Theorem 3 and Corollary 1

Proof sketch of Theorem 3. Each corrected query is decomposed into the local gradient, a predictable conditional bias bounded by b_κ, and a conditionally centered residual with second moment bounded by q_κ. The clipped-direction inequalities in Lemma 6 yield descent in the certificate

$$\phi_C(g) := \min\{\|g\|^2, C\|g\|\}$$

for finite C, with $\phi_\infty(g) := \|g\|^2$, while also controlling the squared update direction by the certificate and the tracking mismatch. The filter, tracker, and consensus recursions then form a coupled system of error inequalities. Martingale orthogonality is used to retain the $H\theta^2 q_\kappa$ dependence of the direct tracker-noise term rather than incurring an additional factor of H.

Under the smallness conditions in (9), the feedback terms in this coupled system can be absorbed, yielding bounds on the averaged descent certificate, filter error, tracker disagreement, and model disagreement in terms of $\mathcal{R}_{T,H}$. Retaining the gain in the tracker recursion gives the refined consensus estimate used in Theorem 3. For a uniformly sampled pre-update model, the certificate conversion

$$\|g\| \le \sqrt{\phi_C(g)} + \frac{\phi_C(g)}{C}$$

controls the gradient at the network-average model, and smoothness transfers this bound to a uniformly sampled local model through the consensus error. This yields (10). When $C = \infty$, the certificate directly equals the squared gradient norm, giving the corresponding squared-gradient guarantee. The proof controls model motion through the average squared update energy rather than the pathwise bound ηC, so the stability conditions and the remainder $\mathcal{R}_{T,H}$ are uniform in C.

Proof sketch of Corollary 1. Let

$$a := \frac{p}{3p-2}, \qquad r := 1-a = \frac{2(p-1)}{3p-2}, \qquad K := TH.$$

Under the tuning in (13), the estimator bounds satisfy

$$b_\kappa^2 = O(K^{-r}), \qquad q_\kappa = O\Big(K^{\frac{2-p}{3p-2}}\Big) = O(K^{2a-1}).$$

For $H = \lceil T^\chi \rceil$ with

$$0 \le \chi \le \frac{a}{2-a} = \frac{p}{5p-4},$$

we have $H \le 2T^\chi$ and $H^2K^{-a} \le 4$. This condition ensures that the local-drift terms remain of the same or smaller order as the leading terms in $\mathcal{R}_{T,H}$. In particular,

$$H^2\eta^2 q_\kappa = O\left(\frac{H^2}{K}\right) = O(K^{-r}), \qquad H\theta^2 q_\kappa = O\left(\frac{H}{K}\right) = O(K^{-r}),$$

and hence

$$\mathcal{R}_{T,H} = O(K^{-r}).$$

Theorem 3 then gives

$$\mathbb{E}\,\|\nabla f(\widehat{x})\| = O(K^{-r/2}).$$

Substituting the same tuning into the refined consensus bound yields

$$\mathbb{E}\left\|\Pi\widehat{\boldsymbol{X}}\right\|_n^2 = O\Big(H^3K^{-(1+2a)}\Big) = O\Big(K^{-(1+a/2)}\Big).$$

Finally, $H = \lceil T^\chi \rceil$ implies $K = \Theta(T^{1+\chi})$. Therefore, achieving $\mathbb{E}\,\|\nabla f(\widehat{x})\| \le \varepsilon$ requires the sufficient budgets

$$K = O\Big(\varepsilon^{-2/r}\Big), \qquad T = O\Big(\varepsilon^{-2/[r(1+\chi)]}\Big).$$

Since the batch size per local update is fixed, the per-agent sample complexity has the same order as K. The remainder of this appendix gives the detailed inequalities and explicit constants underlying these bounds.

F.2 Geometry of the clipped direction

For $C \in (0,\infty)$, clip_C is the Euclidean projection onto the closed ball $\{v \in \mathbb{R}^d : \|v\| \le C\}$. For convenience, define $\mathrm{clip}_\infty(v) := v$.

Lemma 6 (Clipped-direction geometry and certificate conversion)**.** *Let $C \in (0,\infty]$ and define*

$$\phi_C(g) := \begin{cases} \min\{\|g\|^2, C\,\|g\|\}, & C < \infty, \\ \|g\|^2, & C = \infty, \end{cases} \qquad \frac{1}{\infty} := 0.$$

Then, for all $g, y, z \in \mathbb{R}^d$,

$$\langle g, \mathrm{clip}_C(y)\rangle \ge \frac{1}{2}\phi_C(g) - \frac{1}{2}\,\|y-g\|^2, \tag{43}$$

$$\|\mathrm{clip}_C(y)\|^2 \le 2\phi_C(g) + 2\,\|y-g\|^2, \tag{44}$$

$$\|\mathrm{clip}_C(y) - \mathrm{clip}_C(z)\| \le \|y-z\|, \tag{45}$$

$$\|g\| \le \sqrt{\phi_C(g)} + \frac{\phi_C(g)}{C}. \tag{46}$$

Proof. For $C = \infty$, clip_∞ is the identity and $\phi_\infty(g) = \|g\|^2$. In this case, (43) follows from

$$2\,\langle g, y\rangle = \|g\|^2 + \|y\|^2 - \|y-g\|^2 \ge \|g\|^2 - \|y-g\|^2,$$

while (44) follows from $\|y\|^2 \le 2\,\|g\|^2 + 2\,\|y-g\|^2$. Equations (45) and (46) are immediate.

Now suppose $C < \infty$. Write

$$r := \|g\|, \qquad u := \|y\|, \qquad a := \begin{cases} \min\{1, C/u\}, & u > 0, \\ 1, & u = 0. \end{cases}$$

Then $\mathrm{clip}_C(y) = ay$, and polarization gives

$$2\langle g, \mathrm{clip}_C(y)\rangle = a\big(r^2 + u^2 - \|y-g\|^2\big). \tag{47}$$

If $u \le C$, then $a = 1$ and

$$a(r^2 + u^2) \ge r^2 \ge \phi_C(g).$$

If $u > C$, then $a = C/u$ and the arithmetic–geometric mean inequality gives

$$a(r^2 + u^2) = C\left(\frac{r^2}{u} + u\right) \ge 2Cr \ge \phi_C(g).$$

Since $0 < a \le 1$, $-a\|y-g\|^2 \ge -\|y-g\|^2$. Substituting these bounds into (47) proves (43).

Because clip_C is the Euclidean projection onto a closed convex ball, it is nonexpansive:

$$\|\mathrm{clip}_C(y) - \mathrm{clip}_C(z)\| \le \|y - z\|,$$

which proves (45). Moreover,

$$\|\mathrm{clip}_C(g)\|^2 = \min\{r^2, C^2\} \le \min\{r^2, Cr\} = \phi_C(g).$$

Using nonexpansiveness and the squared triangle inequality,

$$\begin{aligned} \|\mathrm{clip}_C(y)\|^2 &\le 2\|\mathrm{clip}_C(g)\|^2 + 2\|\mathrm{clip}_C(y) - \mathrm{clip}_C(g)\|^2 \\ &\le 2\phi_C(g) + 2\|y-g\|^2, \end{aligned}$$

which proves (44).

Finally, if $r \le C$, then $\phi_C(g) = r^2$, so $\sqrt{\phi_C(g)} = r$. If $r > C$, then $\phi_C(g) = Cr$, so $\phi_C(g)/C = r$. In either case,

$$\|g\| \le \sqrt{\phi_C(g)} + \frac{\phi_C(g)}{C},$$

proving (46). □

The quantity $\phi_C(g)$ is the descent certificate induced by direction clipping. For $\|g\| \le C$, it coincides with $\|g\|^2$, whereas for $\|g\| > C$ it grows linearly as $C\|g\|$. Inequality (46) converts control of this clipped certificate back to control of the ordinary gradient norm.

F.3 Filter, consensus, and descent bounds

We begin with the filter error and the disagreement quantities used throughout the proof. At the beginning of each communication round, the model is mixed according to $\boldsymbol{X}_{t,0} = W\boldsymbol{X}_t$, while the filter state is retained, $M_{t,0} = M_t$. To preserve telescoping of the filter error across rounds, we define the boundary error relative to the pre-mixing model:

$$E_{t,0} := M_t - G(\boldsymbol{X}_t), \qquad E_{t,h} := M_{t,h} - G(\boldsymbol{X}_{t,h}), \quad 1 \le h \le H.$$

Since $(\boldsymbol{X}_{t+1}, M_{t+1}) = (\boldsymbol{X}_{t,H}, M_{t,H})$, this convention gives $E_{t,H} = E_{t+1,0}$. Defining the boundary error instead relative to $G(W\boldsymbol{X}_t)$ would introduce an additional mixing-induced jump into the filter recursion.

Define the model disagreement, tracker disagreement, filter error, and squared update-direction norm by

$$\begin{aligned} A_t &:= \mathbb{E}\|\Pi\boldsymbol{X}_t\|_n^2, & A_{t,h} &:= \mathbb{E}\|\Pi\boldsymbol{X}_{t,h}\|_n^2, \\ B_t &:= \mathbb{E}\|\Pi Y_t\|_n^2, & B_{t,h} &:= \mathbb{E}\|\Pi Y_{t,h}\|_n^2, \\ e_{t,h} &:= \mathbb{E}\|E_{t,h}\|_n^2, & D_{t,h} &:= \mathbb{E}\|\mathrm{clip}_C(Y_{t,h})\|_n^2. \end{aligned}$$

For quantities indexed by the $K = TH$ pre-update local steps, write

$$\overline{A} := \frac{1}{K}\sum_{t=0}^{T-1}\sum_{h=0}^{H-1} A_{t,h}, \quad \overline{B} := \frac{1}{K}\sum_{t=0}^{T-1}\sum_{h=0}^{H-1} B_{t,h},$$
$$\overline{E} := \frac{1}{K}\sum_{t=0}^{T-1}\sum_{h=0}^{H-1} e_{t,h}, \quad \overline{D} := \frac{1}{K}\sum_{t=0}^{T-1}\sum_{h=0}^{H-1} D_{t,h}.$$

We also use

$$\overline{A}_{\mathrm{ep}} := \frac{1}{T}\sum_{t=0}^{T-1} A_t, \qquad \overline{\Phi} := \frac{1}{K}\sum_{t=0}^{T-1}\sum_{h=0}^{H-1} \mathbb{E}\phi_C(\nabla f(\bar{x}_{t,h})).$$

Here $\overline{A}_{\mathrm{ep}}$ measures disagreement immediately before communication, whereas $\overline{A}$ uses the actual within-round models; in particular, $A_{t,0}$ corresponds to the post-mixing model $\boldsymbol{X}_{t,0} = W\boldsymbol{X}_t$.

Lemma 7 (Tracker identities)**.** *For every $0 \le t < T$ and $0 \le h \le H$,*

$$Y_{t,h} = WY_t + M_{t,h} - M_t, \qquad JY_{t,h} = JM_{t,h}.$$

Proof. Since $Y_{t,0} = WY_t$ and $M_{t,0} = M_t$, summing

$$Y_{t,r+1} - Y_{t,r} = M_{t,r+1} - M_{t,r}, \qquad r = 0, \dots, h-1,$$

gives

$$Y_{t,h} = WY_t + M_{t,h} - M_t.$$

For the second identity, note that $JY_0 = JM_0 = 0$. Suppose $JY_t = JM_t$. Since $JW = J$, the first identity gives

$$JY_{t,h} = JY_t + JM_{t,h} - JM_t = JM_{t,h}.$$

At $h = H$, this yields $JY_{t+1} = JM_{t+1}$, so the identity follows for all rounds by induction. □

Lemma 8 (Accumulated filter error)**.** *Under the standing assumptions, if $0 < \theta \le 1$, then*

$$\overline{E} \le \frac{\|G(\boldsymbol{X}_0)\|_n^2}{K\theta} + \frac{4L^2 s^2}{H\theta^2}\overline{A}_{\mathrm{ep}} + \frac{4L^2\eta^2}{\theta^2}\overline{D} + 2b_\kappa^2 + \theta q_\kappa, \qquad s := 1 + \lambda. \tag{48}$$

Proof. To handle the model mixing at the beginning of a round, define

$$\boldsymbol{X}_{t,0}^{\circ} := \boldsymbol{X}_t, \qquad \boldsymbol{X}_{t,h}^{\circ} := \boldsymbol{X}_{t,h}, \quad 1 \le h \le H.$$

Using $M_{t,h+1} = (1-\theta)M_{t,h} + \theta Z_{t,h+1}$ and the query decomposition (41), we obtain

$$\begin{aligned} E_{t,h+1} = (1-\theta)E_{t,h} &+ (1-\theta)\big(G(\boldsymbol{X}_{t,h}^{\circ}) - G(\boldsymbol{X}_{t,h+1})\big) \\ &+ \theta\mu_{t,h+1} + \theta\xi_{t,h+1}. \end{aligned} \tag{49}$$

The first three terms on the right are measurable with respect to the pre-query σ-field $\mathcal{F}_{t,h}$, whereas $\mathbb{E}[\xi_{t,h+1} \mid \mathcal{F}_{t,h}] = 0$. Hence the expected cross term between $\xi_{t,h+1}$ and the remaining terms vanishes.

Let

$$\Delta_{t,h} := G(\boldsymbol{X}_{t,h}^{\circ}) - G(\boldsymbol{X}_{t,h+1}).$$

The deterministic part of (49) can be written as

$$(1-\theta)E_{t,h} + \theta\left(\frac{1-\theta}{\theta}\Delta_{t,h} + \mu_{t,h+1}\right).$$

By convexity of the squared norm,

$$\begin{aligned} &\left\|(1-\theta)E_{t,h} + \theta\left(\frac{1-\theta}{\theta}\Delta_{t,h} + \mu_{t,h+1}\right)\right\|_n^2 \\ &\le (1-\theta)\|E_{t,h}\|_n^2 + \theta\left\|\frac{1-\theta}{\theta}\Delta_{t,h} + \mu_{t,h+1}\right\|_n^2 \\ &\le (1-\theta)\|E_{t,h}\|_n^2 + \frac{2}{\theta}\|\Delta_{t,h}\|_n^2 + 2\theta\|\mu_{t,h+1}\|_n^2, \end{aligned}$$

where we used $1-\theta \le 1$ and $\|U+V\|_n^2 \le 2\|U\|_n^2 + 2\|V\|_n^2$. Taking expectations, applying L-smoothness row-wise, and using (42) therefore gives

$$e_{t,h+1} \le (1-\theta)e_{t,h} + \frac{2L^2}{\theta}\mathbb{E}\left\|\boldsymbol{X}_{t,h}^{\circ} - \boldsymbol{X}_{t,h+1}\right\|_n^2 + 2\theta b_\kappa^2 + \theta^2 q_\kappa. \tag{50}$$

For the first local update of round t,

$$\boldsymbol{X}_t - \boldsymbol{X}_{t,1} = (I-W)\boldsymbol{X}_t + \eta\,\mathrm{clip}_C(Y_{t,0}).$$

Because $(I-W)J = 0$ and $\|(I-W)\Pi\|_2 \le 1+\lambda = s$,

$$\begin{aligned}\mathbb{E}\,\|\boldsymbol{X}_t - \boldsymbol{X}_{t,1}\|_n^2 &\le 2\mathbb{E}\,\|(I-W)\Pi\boldsymbol{X}_t\|_n^2 + 2\eta^2 D_{t,0}\\ &\le 2s^2 A_t + 2\eta^2 D_{t,0}.\end{aligned} \tag{51}$$

For $h \ge 1$, no communication occurs between $\boldsymbol{X}_{t,h}$ and $\boldsymbol{X}_{t,h+1}$, so

$$\mathbb{E}\,\|\boldsymbol{X}_{t,h} - \boldsymbol{X}_{t,h+1}\|_n^2 = \eta^2 D_{t,h}.$$

Finally, rearrange (50) as

$$\theta e_{t,h} \le e_{t,h} - e_{t,h+1} + \frac{2L^2}{\theta}\mathbb{E}\left\|\boldsymbol{X}_{t,h}^{\circ} - \boldsymbol{X}_{t,h+1}\right\|_n^2 + 2\theta b_\kappa^2 + \theta^2 q_\kappa.$$

Summing over all $K = TH$ local updates, the error terms telescope because $E_{t,H} = E_{t+1,0}$. Since

$$e_{0,0} = \|G(\boldsymbol{X}_0)\|_n^2$$

and the terminal error is nonnegative,

$$\begin{aligned}\theta K\overline{E} \le\ &\|G(\boldsymbol{X}_0)\|_n^2 + \frac{4L^2 s^2}{\theta}\sum_{t=0}^{T-1} A_t + \frac{4L^2\eta^2}{\theta}\sum_{t=0}^{T-1}\sum_{h=0}^{H-1} D_{t,h}\\ &+ 2\theta K b_\kappa^2 + \theta^2 K q_\kappa.\end{aligned}$$

Using $\sum_t A_t = T\overline{A}_{\mathrm{ep}}$, $\sum_{t,h} D_{t,h} = K\overline{D}$, and $K = TH$, and then dividing by $K\theta$, proves (48). □

Lemma 9 (Model disagreement)**.** *Under the standing assumptions, for any $\eta > 0$ and integers $T, H \ge 1$,*

$$\overline{A}_{\mathrm{ep}} \le \Gamma H^2\eta^2\overline{B}, \qquad \overline{A} \le \Gamma H^2\eta^2\overline{B}, \qquad \Gamma := \frac{2(1+\lambda^2)}{(1-\lambda^2)^2}. \tag{52}$$

Proof. We first record two inequalities used below.

Step 1: row-wise clipping does not increase disagreement. Let $\bar{y} := n^{-1}\sum_{i=1}^n y_i$ denote the row average of $Y = [y_1^\top; \dots; y_n^\top]$. Since $J\,\mathrm{clip}_C(Y)$ is the orthogonal projection of $\mathrm{clip}_C(Y)$ onto the row-constant subspace,

$$\begin{aligned}\|\Pi\,\mathrm{clip}_C(Y)\|_n &= \min_{z\in\mathbb{R}^d}\left\|\mathrm{clip}_C(Y) - \mathbf{1}_n z^\top\right\|_n\\ &\le \left\|\mathrm{clip}_C(Y) - \mathbf{1}_n\,\mathrm{clip}_C(\bar{y})^\top\right\|_n\\ &\le \left\|Y - \mathbf{1}_n\bar{y}^\top\right\|_n = \|\Pi Y\|_n\,.\end{aligned} \tag{53}$$

The second inequality follows from the 1-Lipschitz property of clip_C, applied separately to each row. This argument also covers $C = \infty$.

Step 2: a weighted network contraction. Recall

$$\rho_\lambda := \frac{1+\lambda^2}{2}, \qquad c_\lambda := \frac{1+\lambda^2}{1-\lambda^2}.$$

For arbitrary matrices V, Z with $JV = 0$, $\|(W-J)V\|_F \le \lambda\|V\|_F$. Hence

$$\|(W-J)V + Z\|_n \le \lambda\|V\|_n + \|Z\|_n\,.$$

Weighted Cauchy–Schwarz gives

$$\begin{aligned}\left(\lambda \|V\|_n + \|Z\|_n\right)^2 &\le \left(\rho_\lambda \|V\|_n^2 + c_\lambda \|Z\|_n^2\right)\left(\frac{\lambda^2}{\rho_\lambda} + \frac{1}{c_\lambda}\right)\\ &= \rho_\lambda \|V\|_n^2 + c_\lambda \|Z\|_n^2,\end{aligned}$$

because $\lambda^2/\rho_\lambda + 1/c_\lambda = 1$. Therefore,

$$\|(W-J)V + Z\|_n^2 \le \rho_\lambda \|V\|_n^2 + c_\lambda \|Z\|_n^2. \tag{54}$$

We now apply these inequalities to the model recursion. Since $\boldsymbol{X}_{t,0} = W\boldsymbol{X}_t$ and

$$\boldsymbol{X}_{t,h} = W\boldsymbol{X}_t - \eta \sum_{r=0}^{h-1} \mathrm{clip}_C(Y_{t,r}), \qquad 0 \le h \le H,$$

double stochasticity of W gives

$$\Pi\boldsymbol{X}_{t,h} = (W-J)\Pi\boldsymbol{X}_t - \eta \sum_{r=0}^{h-1} \Pi\,\mathrm{clip}_C(Y_{t,r}).$$

Applying (54), followed by (53) and $\left\|\sum_{r<h} V_r\right\|_n^2 \le h \sum_{r<h} \|V_r\|_n^2$, yields

$$A_{t,h} \le \rho_\lambda A_t + c_\lambda \eta^2 h \sum_{r=0}^{h-1} B_{t,r}, \qquad 0 \le h \le H. \tag{55}$$

At $h = H$, $\boldsymbol{X}_{t,H} = \boldsymbol{X}_{t+1}$, so

$$A_{t+1} \le \rho_\lambda A_t + c_\lambda \eta^2 H \sum_{r=0}^{H-1} B_{t,r}.$$

Summing over $t = 0, \dots, T-1$ and using $A_0 = 0$ gives

$$(1-\rho_\lambda) \sum_{t=0}^{T-1} A_t \le c_\lambda \eta^2 H \sum_{t=0}^{T-1} \sum_{r=0}^{H-1} B_{t,r},$$

where the nonnegative terminal term A_T has been dropped. Since $K = TH$,

$$\overline{A}_{\mathrm{ep}} \le \frac{c_\lambda}{1-\rho_\lambda} H^2\eta^2 \overline{B} = \Gamma H^2 \eta^2 \overline{B}. \tag{56}$$

Finally, averaging (55) over $t = 0, \dots, T-1$ and $h = 0, \dots, H-1$ gives

$$\overline{A} \le \rho_\lambda \overline{A}_{\mathrm{ep}} + c_\lambda H^2 \eta^2 \overline{B}.$$

Combining this with (56) and using

$$\rho_\lambda \Gamma + c_\lambda = \Gamma$$

yields

$$\overline{A} \le \Gamma H^2 \eta^2 \overline{B}.$$

This proves both inequalities in (52). □

Lemma 10 (Tracker disagreement)**.** *Under the standing assumptions, if* $0 < \theta \le 1$*, then*

$$\overline{B} \le \Gamma\big[6H^2\theta^2\overline{E} + 12L^2H^2\theta^2\eta^2\overline{D} + 6H^2\theta^2 b_\kappa^2 + 2H\theta^2 q_\kappa + 12L^2 s^2 H\theta^2 \overline{A}_{\mathrm{ep}}\big], \tag{57}$$

where $s := 1 + \lambda$ *and* $\Gamma := 2(1+\lambda^2)/(1-\lambda^2)^2$*. If, in addition,*

$$\tau := 12\Gamma^2 L^2 s^2 H^3 \theta^2 \eta^2 < 1,$$

then

$$\overline{B} \le \frac{6\Gamma H^2\theta^2}{1-\tau}\left(\overline{E} + 2L^2\eta^2\overline{D} + b_\kappa^2\right) + \frac{2\Gamma H\theta^2}{1-\tau} q_\kappa. \tag{58}$$

Proof. We first bound the change in the filter state within a communication round. Define

$$\Delta G_{t,1} := G(\boldsymbol{X}_{t,1}) - G(\boldsymbol{X}_t), \qquad \Delta G_{t,r} := G(\boldsymbol{X}_{t,r}) - G(\boldsymbol{X}_{t,r-1}), \quad r \geq 2.$$

By the definition of $E_{t,r-1}$ and the corrected-query decomposition,

$$Z_{t,r} - M_{t,r-1} = -E_{t,r-1} + \Delta G_{t,r} + \mu_{t,r} + \xi_{t,r}.$$

Since $M_{t,r} - M_{t,r-1} = \theta(Z_{t,r} - M_{t,r-1})$, summing over $r = 1, \ldots, h$ gives

$$M_{t,h} - M_t = \theta \sum_{r=1}^{h} \big(-E_{t,r-1} + \Delta G_{t,r} + \mu_{t,r}\big) + \theta \sum_{r=1}^{h} \xi_{t,r}. \tag{59}$$

The centered residuals form a martingale-difference sequence with respect to the query filtration. Hence, for $r < r'$,

$$\mathbb{E}\langle \xi_{t,r}, \xi_{t,r'} \rangle_F = 0,$$

and therefore

$$\mathbb{E} \left\| \sum_{r=1}^{h} \xi_{t,r} \right\|_n^2 = \sum_{r=1}^{h} \mathbb{E} \left\| \xi_{t,r} \right\|_n^2 \leq h q_\kappa.$$

The predictable sum and the martingale sum in (59) need not be orthogonal, because later predictable terms may depend on earlier residuals. We therefore use $\|U+V\|_n^2 \leq 2\|U\|_n^2 + 2\|V\|_n^2$. For the predictable sum, Cauchy–Schwarz gives

$$\begin{aligned}
&\mathbb{E} \left\| \sum_{r=1}^{h} \big(-E_{t,r-1} + \Delta G_{t,r} + \mu_{t,r}\big) \right\|_n^2 \\
&\qquad \leq 3h \sum_{r=1}^{h} \Big(e_{t,r-1} + \mathbb{E} \left\| \Delta G_{t,r} \right\|_n^2 + \mathbb{E} \left\| \mu_{t,r} \right\|_n^2 \Big) \\
&\qquad \leq 3h \sum_{r=0}^{h-1} e_{t,r} + 3h \sum_{r=1}^{h} \mathbb{E} \left\| \Delta G_{t,r} \right\|_n^2 + 3h^2 b_\kappa^2.
\end{aligned}$$

It remains to control the gradient increments. By L-smoothness,

$$\left\| \Delta G_{t,r} \right\|_n \leq L \left\| \boldsymbol{X}_{t,r} - \boldsymbol{X}_{t,r-1}^{\circ} \right\|_n,$$

where $\boldsymbol{X}_{t,0}^{\circ} = \boldsymbol{X}_t$ and $\boldsymbol{X}_{t,r-1}^{\circ} = \boldsymbol{X}_{t,r-1}$ for $r \geq 2$. Using (51) for $r = 1$ and the local-update identity for $r \geq 2$ gives

$$\sum_{r=1}^{h} \mathbb{E} \left\| \Delta G_{t,r} \right\|_n^2 \leq 2L^2 s^2 A_t + 2L^2 \eta^2 \sum_{r=0}^{h-1} D_{t,r}. \tag{60}$$

Combining (59)–(60) yields

$$\begin{aligned}
\mathbb{E} \left\| M_{t,h} - M_t \right\|_n^2 \leq{}& 6h\theta^2 \sum_{r=0}^{h-1} e_{t,r} + 12L^2 s^2 h \theta^2 A_t \\
&+ 12L^2 h \theta^2 \eta^2 \sum_{r=0}^{h-1} D_{t,r} + 6h^2 \theta^2 b_\kappa^2 + 2h\theta^2 q_\kappa.
\end{aligned}$$

We now transfer this bound to the tracker disagreement. By Lemma 7,

$$\Pi Y_{t,h} = (W - J)\Pi Y_t + \Pi(M_{t,h} - M_t).$$

Applying (54) and using $\|\Pi V\|_n \leq \|V\|_n$ gives

$$B_{t,h} \leq \rho_\lambda B_t + c_\lambda \mathbb{E} \left\| M_{t,h} - M_t \right\|_n^2. \tag{61}$$

For convenience, define

$$V_0 := 6H^2\theta^2\overline{E} + 12L^2H^2\theta^2\eta^2\overline{D} + 6H^2\theta^2 b_\kappa^2 + 2H\theta^2 q_\kappa + 12L^2 s^2 H\theta^2\overline{A}_{\mathrm{ep}}. \tag{62}$$

Both the endpoint average and the within-round average of $\mathbb{E}\,\|M_{t,h} - M_t\|_n^2$ are bounded by V_0. Indeed, for example,

$$h\sum_{r=0}^{h-1} e_{t,r} \le H\sum_{r=0}^{H-1} e_{t,r},$$

and summing this inequality over $h = 0, \dots, H-1$ produces the factor H^2 in (62); the same argument applies to the $D_{t,r}$ terms. The remaining terms follow directly from $h \le H$ and $h^2 \le H^2$.

At $h = H$, $Y_{t,H} = Y_{t+1}$, so (61) gives

$$B_{t+1} \le \rho_\lambda B_t + c_\lambda \mathbb{E}\,\|M_{t,H} - M_t\|_n^2 .$$

Since $B_0 = 0$, summing over communication rounds yields an endpoint-average bound

$$\frac{1}{T}\sum_{t=0}^{T-1} B_t \le \frac{c_\lambda}{1-\rho_\lambda} V_0 = \Gamma V_0.$$

Averaging the path inequality (61) over all $t < T$ and $h < H$ then gives

$$\overline{B} \le \rho_\lambda \Gamma V_0 + c_\lambda V_0.$$

Since $\rho_\lambda\Gamma + c_\lambda = \Gamma$, we obtain

$$\overline{B} \le \Gamma V_0,$$

which is exactly (57).

Finally, Lemma 9 gives

$$\overline{A}_{\mathrm{ep}} \le \Gamma H^2\eta^2\overline{B}.$$

Substituting this into (57), the last term on its right-hand side becomes

$$12\Gamma^2 L^2 s^2 H^3\theta^2\eta^2\overline{B} = \tau\overline{B}.$$

When $\tau < 1$, moving this term to the left and collecting the remaining terms yields

$$(1-\tau)\overline{B} \le 6\Gamma H^2\theta^2\left(\overline{E} + 2L^2\eta^2\overline{D} + b_\kappa^2\right) + 2\Gamma H\theta^2 q_\kappa,$$

which proves (58). □

Lemma 11 (Mismatch and objective descent)**.** *Under the standing assumptions, define*

$$R_{t,h} := \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\|Y_{i,t,h} - \nabla f(\bar{x}_{t,h})\|^2\right], \qquad \overline{R} := \frac{1}{K}\sum_{t=0}^{T-1}\sum_{h=0}^{H-1} R_{t,h}.$$

Then

$$\overline{R} \le \overline{B} + 2\overline{E} + 2L^2\left(\overline{A} + \frac{\overline{A}_{\mathrm{ep}}}{H}\right), \tag{63}$$

$$\overline{D} \le 2\overline{\Phi} + 2\overline{R}. \tag{64}$$

If, in addition, $L\eta \le 1/4$ and $\Delta_0 := f(x_0) - f_{\inf}$, then

$$\overline{\Phi} \le \frac{4\Delta_0}{K\eta} + 3\overline{R}, \qquad \overline{D} \le \frac{8\Delta_0}{K\eta} + 8\overline{R}. \tag{65}$$

Proof. We first relate the tracker mismatch to tracker disagreement, filter error, and model disagreement. For any matrix Y and vector $g \in \mathbb{R}^d$, orthogonality of the consensus and disagreement subspaces gives

$$\left\|Y - \mathbf{1}_n g^\top\right\|_n^2 = \|\Pi Y\|_n^2 + \left\|JY - \mathbf{1}_n g^\top\right\|_n^2. \tag{66}$$

For $h \geq 1$, Lemma 7 gives $JY_{t,h} = JM_{t,h}$. Using $M_{t,h} = G(\boldsymbol{X}_{t,h}) + E_{t,h}$, we obtain

$$\begin{aligned}\left\|JY_{t,h} - \mathbf{1}_n \nabla f(\bar{x}_{t,h})^\top\right\|_n^2 &= \left\|JE_{t,h} + JG(\boldsymbol{X}_{t,h}) - \mathbf{1}_n \nabla f(\bar{x}_{t,h})^\top\right\|_n^2 \\ &\leq 2\left\|E_{t,h}\right\|_n^2 + 2\left\|JG(\boldsymbol{X}_{t,h}) - \mathbf{1}_n \nabla f(\bar{x}_{t,h})^\top\right\|_n^2.\end{aligned}$$

Since

$$JG(\boldsymbol{X}_{t,h}) = \mathbf{1}_n \left(\frac{1}{n}\sum_{i=1}^n \nabla f_i(x_{i,t,h})\right)^\top$$

and $\nabla f(\bar{x}_{t,h}) = n^{-1}\sum_i \nabla f_i(\bar{x}_{t,h})$, L-smoothness gives

$$\begin{aligned}\left\|JG(\boldsymbol{X}_{t,h}) - \mathbf{1}_n \nabla f(\bar{x}_{t,h})^\top\right\|_n^2 &= \left\|\frac{1}{n}\sum_{i=1}^n \big(\nabla f_i(x_{i,t,h}) - \nabla f_i(\bar{x}_{t,h})\big)\right\|^2 \\ &\leq \frac{1}{n}\sum_{i=1}^n \left\|\nabla f_i(x_{i,t,h}) - \nabla f_i(\bar{x}_{t,h})\right\|^2 \\ &\leq L^2 \left\|\Pi \boldsymbol{X}_{t,h}\right\|_n^2. \end{aligned} \tag{67}$$

Combining (66)–(67) and taking expectations yields, for $h \geq 1$,

$$R_{t,h} \leq B_{t,h} + 2e_{t,h} + 2L^2 A_{t,h}. \tag{68}$$

At $h = 0$, the tracker has already been mixed, $Y_{t,0} = WY_t$, whereas the boundary filter error is defined relative to the pre-mixing model:

$$E_{t,0} = M_t - G(\boldsymbol{X}_t).$$

Because mixing preserves the row average, $\bar{x}_{t,0} = \bar{x}_t$, and the same argument gives

$$R_{t,0} \leq B_{t,0} + 2e_{t,0} + 2L^2 A_t. \tag{69}$$

Averaging (68) and (69) over all $K = TH$ pre-update indices gives

$$\overline{R} \leq \overline{B} + 2\overline{E} + \frac{2L^2}{K}\left(\sum_{t=0}^{T-1} A_t + \sum_{t=0}^{T-1}\sum_{h=1}^{H-1} A_{t,h}\right).$$

Using $\sum_t A_t = T\overline{A}_{\mathrm{ep}}$ and bounding the second sum by $K\overline{A}$ proves (63).

Next, apply (44) row-wise with

$$g = \nabla f(\bar{x}_{t,h}), \qquad y = Y_{i,t,h}.$$

Since the same g is used for every row,

$$\begin{aligned} D_{t,h} &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \left\|\mathrm{clip}_C(Y_{i,t,h})\right\|^2\right] \\ &\leq 2\mathbb{E}\phi_C(\nabla f(\bar{x}_{t,h})) + 2R_{t,h}. \end{aligned}$$

Averaging over t and h proves (64).

It remains to establish descent. Since mixing is doubly stochastic, it preserves the row average. During a local update,

$$\bar{x}_{t,h+1} = \bar{x}_{t,h} - \eta \bar{d}_{t,h}, \qquad \bar{d}_{t,h} := \frac{1}{n}\sum_{i=1}^n \mathrm{clip}_C(Y_{i,t,h}).$$

By L-smoothness of f,

$$f(\bar{x}_{t,h+1}) \leq f(\bar{x}_{t,h}) - \eta\left\langle \nabla f(\bar{x}_{t,h}), \bar{d}_{t,h}\right\rangle + \frac{L\eta^2}{2}\left\|\bar{d}_{t,h}\right\|^2. \tag{70}$$

Applying (43) to each row and averaging gives

$$\left\langle \nabla f(\bar{x}_{t,h}), \bar{d}_{t,h} \right\rangle \geq \frac{1}{2}\phi_C(\nabla f(\bar{x}_{t,h})) - \frac{1}{2n}\sum_{i=1}^{n} \left\| Y_{i,t,h} - \nabla f(\bar{x}_{t,h}) \right\|^2. \tag{71}$$

Similarly, Jensen's inequality and (44) give

$$\begin{aligned} \left\| \bar{d}_{t,h} \right\|^2 &\leq \frac{1}{n}\sum_{i=1}^{n} \left\| \mathrm{clip}_C(Y_{i,t,h}) \right\|^2 \\ &\leq 2\phi_C(\nabla f(\bar{x}_{t,h})) + \frac{2}{n}\sum_{i=1}^{n} \left\| Y_{i,t,h} - \nabla f(\bar{x}_{t,h}) \right\|^2. \end{aligned} \tag{72}$$

Substituting (71) and (72) into (70), and then taking expectations, yields

$$\begin{aligned} \mathbb{E} f(\bar{x}_{t,h+1}) \leq \mathbb{E} f(\bar{x}_{t,h}) &- \eta\left(\frac{1}{2} - L\eta\right)\mathbb{E}\phi_C(\nabla f(\bar{x}_{t,h})) \\ &+ \eta\left(\frac{1}{2} + L\eta\right) R_{t,h}. \end{aligned}$$

Because mixing preserves the row average, these inequalities telescope across both local steps and communication rounds. Summing over all $K = TH$ updates and using $f \geq f_{\inf}$ gives

$$\eta\left(\frac{1}{2} - L\eta\right) K\overline{\Phi} \leq \Delta_0 + \eta\left(\frac{1}{2} + L\eta\right) K\overline{R}.$$

If $L\eta \leq 1/4$, then $\frac{1}{2} - L\eta \geq 1/4$ and $\frac{1}{2} + L\eta \leq 3/4$, so

$$\overline{\Phi} \leq \frac{4\Delta_0}{K\eta} + 3\overline{R}.$$

Finally, substituting this bound into (64) gives

$$\overline{D} \leq 2\overline{\Phi} + 2\overline{R} \leq \frac{8\Delta_0}{K\eta} + 8\overline{R},$$

which proves (65). □

F.4 Closing the coupled inequalities with explicit constants

We now combine the preceding filter, disagreement, and descent estimates. The following result gives explicit constants for the finite-time bounds. The constant C_* below is a proof constant and is unrelated to the direction-clipping radius C.

Theorem 4 (Finite-time bound with explicit constants)**.** *Under Assumptions 1–3 and the initialization* (40)*, define*

$$\Gamma := \frac{2(1+\lambda^2)}{(1-\lambda^2)^2}, \quad \ell := \max\{1, L\}, \quad s := 1 + \lambda, \quad c_* := \frac{1}{64\Gamma\ell s}, \quad C_* := 1024\Gamma^2\ell^2 s^2.$$

Let $T, H, m \geq 1$ be integers, $\eta, \kappa > 0$, $C \in (0, \infty]$, and suppose

$$0 < \theta \leq \frac{1}{2}, \qquad L\eta \leq \frac{1}{4}, \qquad \frac{\eta}{\theta} \leq c_*, \qquad H\theta \leq c_*, \qquad \frac{H^2\eta^2}{\theta} \leq c_*. \tag{73}$$

For $K := TH$, define

$$\mathcal{R}_{T,H} := \frac{\Delta_0}{K\eta} + \frac{\left\| G(\boldsymbol{X}_0) \right\|_n^2}{K\theta} + b_\kappa^2 + \left(\theta + H^2\eta^2 + H\theta^2\right) q_\kappa.$$

Then

$$\overline{\Phi} + \overline{E} + \overline{B} \leq C_* \mathcal{R}_{T,H}, \tag{74}$$

$$\max\{\overline{A}_{\mathrm{ep}}, \overline{A}\} \leq \Gamma C_* H^2\eta^2 \mathcal{R}_{T,H}, \tag{75}$$

$$\max\{\overline{A}_{\mathrm{ep}}, \overline{A}\} \leq 6\Gamma^2 C_* H^4\eta^2\theta^2 \mathcal{R}_{T,H} + 4\Gamma^2 H^3\eta^2\theta^2 q_\kappa. \tag{76}$$

More precisely, with

$$C_R := 128\Gamma^2\ell^2 s^2 = \frac{C_*}{8},$$

we have

$$\overline{R} \le C_R\mathcal{R}_{T,H}, \quad \overline{E} \le C_R\mathcal{R}_{T,H}, \quad \overline{D} \le 16C_R\mathcal{R}_{T,H}, \quad \overline{B} \le 3C_R\mathcal{R}_{T,H}, \quad \overline{\Phi} \le 4C_R\mathcal{R}_{T,H}. \tag{77}$$

All displayed constants depend only on L and λ and are independent of $T, H, m, \eta, \theta, \kappa, C, n,$ and d.

Proof. We close the coupled inequalities in three steps.

Step 1: reduction to a scalar system. Define

$$\delta_* := \frac{\Delta_0}{K\eta}, \qquad E_* := \frac{\|G(\boldsymbol{X}_0)\|_n^2}{K\theta} + 2b_\kappa^2 + \theta q_\kappa, \qquad \tau := 12\Gamma^2L^2s^2H^3\theta^2\eta^2.$$

The smallness conditions in (73) imply

$$H^3\eta^2 \le c_*^2, \qquad H^2\eta^2 \le c_*^4, \qquad \eta^2 \le \frac{c_*^2}{4}, \qquad \tau \le 3\Gamma^2\ell^2s^2c_*^2 = \frac{3}{4096} < \frac{1}{2}.$$

Here the first inequality follows from $(H^2\eta^2/\theta)(H\theta) \le c_*^2$, the second from $(H\eta)^2 \le c_*^4$, and the third from $\eta \le c_*\theta$ and $\theta \le 1/2$.

Introduce

$$\alpha := \frac{4\Gamma L^2s^2H\eta^2}{\theta^2}, \qquad \beta := \frac{4L^2\eta^2}{\theta^2},$$
$$U := \frac{6\Gamma H^2\theta^2}{1-\tau}, \qquad V := 2L^2\eta^2U,$$
$$W_1 := Ub_\kappa^2 + \frac{2\Gamma H\theta^2}{1-\tau}q_\kappa, \qquad Z_1 := 1 + 2\Gamma L^2H^2\eta^2\left(1 + \frac{1}{H}\right).$$

Lemmas 8–11 then imply

$$\overline{E} \le E_* + \alpha\overline{B} + \beta\overline{D}, \qquad \overline{B} \le U\overline{E} + V\overline{D} + W_1, \qquad \overline{R} \le 2\overline{E} + Z_1\overline{B}, \qquad \overline{D} \le 8\delta_* + 8\overline{R}. \tag{78}$$

Indeed, the first inequality follows from Lemma 8 after substituting $\overline{A}_{\mathrm{ep}} \le \Gamma H^2\eta^2\overline{B}$; the second is (58); the third follows from (63) and Lemma 9; and the last is (65).

The choice of c_* ensures

$$U \le 12\Gamma c_*^2 \le 1, \qquad \alpha U \le 48\Gamma^2\ell^2s^2c_*^2 \le \frac{1}{4}, \qquad 1 \le Z_1 \le 2. \tag{79}$$

Define

$$a_1 := \frac{2 + Z_1U}{1 - \alpha U}, \qquad a_2 := \frac{2\alpha + Z_1}{1 - \alpha U}, \qquad k_1 := 8(a_1\beta + a_2V).$$

Using

$$\beta \le 4\ell^2c_*^2, \qquad V \le \frac{1}{2}\ell^2c_*^2, \qquad \alpha V \le \frac{1}{8}\ell^2c_*^2,$$

and (79), we have $a_1 \le 16/3$ and $a_2V \le (5/3)\ell^2c_*^2$. Consequently,

$$k_1 \le 184\ell^2c_*^2 \le \frac{1}{2}.$$

Step 2: absorption and bounds for the main quantities. Eliminating $\overline{E}$ and $\overline{B}$ from the first three inequalities in (78) gives

$$\overline{R} \le a_1E_* + a_2W_1 + (a_1\beta + a_2V)\overline{D}.$$

Using $\overline{D} \le 8(\delta_* + \overline{R})$ and the definition of k_1,

$$\overline{R} \le a_1E_* + a_2W_1 + k_1(\delta_* + \overline{R}),$$

and hence

$$\overline{R} \le \frac{a_1 E_* + a_2 W_1 + k_1 \delta_*}{1 - k_1}. \tag{80}$$

The coefficient bounds above imply $a_1 \le 8$ and, using $1/(1-\tau) \le 2$,

$$a_2 W_1 \le 5 b_\kappa^2 + \left(64\Gamma^2 L^2 s^2 H^2 \eta^2 + 16\Gamma H \theta^2\right) q_\kappa.$$

Since $k_1 \le 1/2$, (80) therefore yields

$$\begin{aligned}\overline{R} &\le \delta_* + 16 E_* + 10 b_\kappa^2 \\ &\quad + \left(128\Gamma^2 L^2 s^2 H^2 \eta^2 + 32\Gamma H \theta^2\right) q_\kappa \\ &\le C_R \mathcal{R}_{T,H},\end{aligned}$$

where $C_R = 128\Gamma^2 \ell^2 s^2$. By Lemma 11,

$$\overline{D} \le 8\delta_* + 8\overline{R} \le 16 C_R \mathcal{R}_{T,H}, \qquad \overline{\Phi} \le 4\delta_* + 3\overline{R} \le 4 C_R \mathcal{R}_{T,H}. \tag{81}$$

We next recover the filter and tracker errors. Combining the first two inequalities in (78) gives

$$\overline{E} \le \frac{E_* + \alpha W_1}{1 - \alpha U} + \frac{\beta + \alpha V}{1 - \alpha U} \overline{D}.$$

The coefficient bounds imply

$$E_* \le 2\mathcal{R}_{T,H}, \qquad \alpha W_1 \le \left(\frac{1}{4} + \frac{C_R}{8}\right) \mathcal{R}_{T,H}, \qquad \frac{\beta + \alpha V}{1 - \alpha U} \le \frac{k_1}{16}.$$

Together with (81), this gives

$$\overline{E} \le \left(3 + \frac{2C_R}{3}\right) \mathcal{R}_{T,H} \le C_R \mathcal{R}_{T,H},$$

where the last inequality uses $C_R \ge 512$.

Finally, $V \le k_1/8$, $W_1 \le 4\Gamma \mathcal{R}_{T,H}$, $U \le 1$, and the already established bounds imply

$$\overline{B} \le U\overline{E} + V\overline{D} + W_1 \le 3 C_R \mathcal{R}_{T,H}. \tag{82}$$

Since $C_* = 8C_R$, the bounds (81)–(82) imply (77) and, in particular, (74).

Step 3: consensus bounds. Lemma 9 and $\overline{B} \le C_* \mathcal{R}_{T,H}$ immediately give

$$\max\{\overline{A}_{\mathrm{ep}},\, \overline{A}\} \le \Gamma C_* H^2 \eta^2 \mathcal{R}_{T,H},$$

which proves (75).

For the sharper bound, retain the factor $H^2\theta^2$ in (58). Since $\tau < 1/2$,

$$\begin{aligned}\overline{B} &\le 12\Gamma H^2 \theta^2 \left(\overline{E} + 2L^2 \eta^2 \overline{D} + b_\kappa^2\right) + 4\Gamma H \theta^2 q_\kappa \\ &\le 6\Gamma C_* H^2 \theta^2 \mathcal{R}_{T,H} + 4\Gamma H \theta^2 q_\kappa.\end{aligned}$$

For the second inequality, we used

$$\overline{E} \le C_R \mathcal{R}_{T,H}, \qquad \overline{D} \le 16 C_R \mathcal{R}_{T,H}, \qquad b_\kappa^2 \le \mathcal{R}_{T,H},$$

together with $L\eta \le 1/4$ and

$$C_R + 32 L^2 \eta^2 C_R + 1 \le 4 C_R = \frac{C_*}{2}.$$

Applying Lemma 9 once more gives

$$\max\{\overline{A}_{\mathrm{ep}},\, \overline{A}\} \le \Gamma H^2 \eta^2 \overline{B} \le 6\Gamma^2 C_* H^4 \eta^2 \theta^2 \mathcal{R}_{T,H} + 4\Gamma^2 H^3 \eta^2 \theta^2 q_\kappa,$$

which is (76). □

Corollary 3 (Finite-time random-output bounds)**.** *Under the conditions of Theorem 4, let $(\widehat{t},\widehat{h})$ be uniform on $\{0,\dots,T-1\}\times\{0,\dots,H-1\}$ and let I be uniform on $[n]$, independently of each other and of the algorithm. Define*

$$\widehat{\boldsymbol{X}} := \boldsymbol{X}_{\widehat{t},\widehat{h}}, \qquad \widehat{x} := x_{I,\widehat{t},\widehat{h}}.$$

Then

$$\mathbb{E}\,\|\nabla f(\widehat{x})\| \le \sqrt{C_*}\big(1+L\sqrt{\Gamma}\,c_*^2\big)\sqrt{\mathcal{R}_{T,H}} + \frac{C_*\mathcal{R}_{T,H}}{C}, \tag{83}$$

$$\mathbb{E}\left\|\Pi\widehat{\boldsymbol{X}}\right\|_n^2 \le 6\Gamma^2 C_* H^3\eta^2\theta^2\big(H\mathcal{R}_{T,H}+q_\kappa\big). \tag{84}$$

Here $1/\infty := 0$. For $C=\infty$, the expected squared gradient norm additionally satisfies

$$\mathbb{E}\,\|\nabla f(\widehat{x})\|^2 \le 2C_*\big(1+L^2\Gamma c_*^4\big)\mathcal{R}_{T,H}. \tag{85}$$

Proof. Let

$$\widehat{\bar{x}} := \bar{x}_{\widehat{t},\widehat{h}}.$$

By uniform sampling of $(\widehat{t},\widehat{h})$,

$$\mathbb{E}\phi_C(\nabla f(\widehat{\bar{x}})) = \overline{\Phi}, \qquad \mathbb{E}\left\|\Pi\widehat{\boldsymbol{X}}\right\|_n^2 = \overline{A}. \tag{86}$$

Moreover, conditional on $\widehat{\boldsymbol{X}}$, the uniform choice of I gives

$$\mathbb{E}\left[\left\|\widehat{x}-\widehat{\bar{x}}\right\|^2 \,\middle|\, \widehat{\boldsymbol{X}}\right] = \frac{1}{n}\sum_{i=1}^n \left\|x_{i,\widehat{t},\widehat{h}}-\widehat{\bar{x}}\right\|^2 = \left\|\Pi\widehat{\boldsymbol{X}}\right\|_n^2.$$

Therefore,

$$\mathbb{E}\left\|\widehat{x}-\widehat{\bar{x}}\right\|^2 = \overline{A}. \tag{87}$$

By L-smoothness,

$$\|\nabla f(\widehat{x})\| \le \left\|\nabla f(\widehat{\bar{x}})\right\| + L\left\|\widehat{x}-\widehat{\bar{x}}\right\|.$$

Applying the certificate conversion (46) to $\nabla f(\widehat{\bar{x}})$, followed by Jensen's inequality and (87), yields

$$\begin{aligned}\mathbb{E}\,\|\nabla f(\widehat{x})\| &\le \mathbb{E}\sqrt{\phi_C(\nabla f(\widehat{\bar{x}}))} + \frac{1}{C}\mathbb{E}\phi_C(\nabla f(\widehat{\bar{x}})) + L\mathbb{E}\left\|\widehat{x}-\widehat{\bar{x}}\right\| \\ &\le \sqrt{\overline{\Phi}} + \frac{\overline{\Phi}}{C} + L\sqrt{\overline{A}}.\end{aligned} \tag{88}$$

Theorem 4 gives $\overline{\Phi} \le C_*\mathcal{R}_{T,H}$ and, from the coarse consensus bound,

$$\overline{A} \le \Gamma C_* H^2\eta^2\mathcal{R}_{T,H}.$$

The smallness conditions $\eta/\theta \le c_*$ and $H\theta \le c_*$ imply

$$H\eta = (H\theta)\frac{\eta}{\theta} \le c_*^2.$$

Consequently,

$$L\sqrt{\overline{A}} \le L\sqrt{\Gamma C_*}\,H\eta\sqrt{\mathcal{R}_{T,H}} \le L\sqrt{\Gamma C_*}\,c_*^2\sqrt{\mathcal{R}_{T,H}}.$$

Substituting these bounds into (88) proves (83).

For the consensus error, uniform sampling and the sharp bound in Theorem 4 give

$$\begin{aligned}\mathbb{E}\left\|\Pi\widehat{\boldsymbol{X}}\right\|_n^2 = \overline{A} &\le 6\Gamma^2 C_* H^4\eta^2\theta^2\mathcal{R}_{T,H} + 4\Gamma^2 H^3\eta^2\theta^2 q_\kappa \\ &\le 6\Gamma^2 C_* H^3\eta^2\theta^2\big(H\mathcal{R}_{T,H}+q_\kappa\big),\end{aligned}$$

where the last step uses $C_* \ge 1$. This proves (84).

Finally, suppose $C = \infty$. Then $\phi_\infty(g) = \|g\|^2$, and L-smoothness gives

$$\|\nabla f(\widehat{x})\|^2 \le 2\left\|\nabla f(\widehat{\widehat{x}})\right\|^2 + 2L^2\left\|\widehat{x} - \widehat{\widehat{x}}\right\|^2.$$

Taking expectations and using (86)–(87),

$$\mathbb{E}\,\|\nabla f(\widehat{x})\|^2 \le 2\overline{\Phi} + 2L^2\overline{A}.$$

Applying the master and coarse consensus bounds, followed by $H\eta \le c_*^2$, gives

$$\mathbb{E}\,\|\nabla f(\widehat{x})\|^2 \le 2C_*\big(1 + L^2\Gamma c_*^4\big)\mathcal{R}_{T,H},$$

which proves (85). □

Recovery of Theorem 3. Set $c := c_*$. Since $H \ge 1$, the condition $H\theta \le c_*$ implies $\theta \le c_* \le 1/2$. Moreover, $\eta \le c_*\theta$ and $\theta \le c_*$ give

$$L\eta \le Lc_*^2 \le \frac{1}{4}.$$

Thus the conditions in (9) imply all the smallness conditions in (73). Applying Corollary 3 then gives (10) and (11), with constants depending only on L and λ. For finite C, the expected-gradient-norm guarantee is obtained directly from the clipped certificate and does not require a bound on the expected squared norm of the unclipped gradient.

Role of the direction radius. For finite C, clipping gives the pathwise bound

$$\|x_{i,t,h+1} - x_{i,t,h}\| \le \eta C.$$

The convergence proof, however, does not use this worst-case bound to control the accumulated model motion. Instead, it uses the squared direction energy $D_{t,h}$. By (65) and Theorem 4,

$$\frac{1}{K}\sum_{t=0}^{T-1}\sum_{h=0}^{H-1}\mathbb{E}\,\|\boldsymbol{X}_{t,h+1} - \boldsymbol{X}_{t,h}\|_n^2 = \eta^2\overline{D} \le 16C_R\eta^2\mathcal{R}_{T,H}.$$

Hence the filter, tracker, and consensus bounds close without replacing $D_{t,h}$ by the coarser estimate C^2. The dependence on C in the finite-C stationarity bound arises only when the clipped certificate is converted back to the ordinary gradient norm through (46). For $C = \infty$, the same analysis applies directly with clip_∞ equal to the identity map. This is a separate admissible case of the analysis and does not rely on taking a limit of the algorithmic trajectories as $C \to \infty$; the corrected gradient estimator still uses a finite correction radius κ.

Why the tracker gain is retained. The coefficient α introduced in the proof of Theorem 4 need not itself be small, since it contains the factor $H(\eta/\theta)^2$. What enters the feedback loop is the product αU, whose scaling retains the compensating factor from the tracker recursion and is controlled by the stated smallness conditions. Requiring α alone to be small would impose a stronger restriction and could exclude the growing-local-period regime analyzed in Corollary 1. Similarly, martingale orthogonality yields the term $H\theta^2 q_\kappa$ in the tracker bound; applying Cauchy–Schwarz directly to the sum of centered residuals would instead introduce an additional factor of H. The explicit constants above are chosen to close these inequalities and are not intended as practical parameter recommendations.

F.5 Horizon-dependent rates and output guarantees

Corollary 4 (Explicit horizon-dependent rates)**.** *Under the assumptions and initialization of Theorem 4, define*

$$a := \frac{p}{3p-2}, \qquad r := \frac{2(p-1)}{3p-2} = 1 - a, \qquad 0 \le \chi \le \frac{p}{5p-4}.$$

Fix an integer $m \ge 1$, a direction radius $C \in (0, \infty]$ independent of the horizon, and constants

$$c_\kappa > 0, \qquad 0 < c_\theta \le \frac{c_*}{4}, \qquad 0 < c_\eta \le c_*c_\theta.$$

For any integer $T \geq 1$, set

$$H := \lceil T^{\chi} \rceil, \qquad K := TH, \qquad \eta := c_\eta K^{-a}, \qquad \theta := c_\theta K^{-a}, \qquad \kappa := c_\kappa K^{1/(3p-2)}. \tag{89}$$

Let

$$N := 3m, \qquad s_N := \sigma N^{-(p-1)/p}, \qquad A_N := \sigma^p N^{1-p},$$

and, with C_p from Lemma 4, define

$$\begin{aligned} Q_* &:= C_p \left(s_N^2 + A_N c_\kappa^{2-p} \right), \\ R_* &:= \frac{\Delta_0}{c_\eta} + \frac{\|G(\boldsymbol{X}_0)\|_n^2}{c_\theta} + C_p^2 A_N^2 c_\kappa^{-2(p-1)} + \left(c_\theta + 4c_\eta^2 + 4c_\theta^2 \right) Q_*, \\ C_A &:= 4\Gamma^2 c_\eta^2 c_\theta^2 \left(6 C_* R_* + Q_* \right). \end{aligned}$$

Then

$$\overline{\Phi} \leq C_* R_* K^{-r}, \tag{90}$$

$$\max\{\overline{A}_{\mathrm{ep}}, \overline{A}\} \leq C_A H^3 K^{-(1+2a)} \leq 8 C_A K^{-(1+a/2)}. \tag{91}$$

Alternatively, let $H \geq 1$ be any fixed integer and use the same formulas for η, θ, and κ in (89)*, with $K = TH$. The same bounds, with the same constants, hold whenever*

$$H^2 K^{-a} \leq 4.$$

Proof. We first verify the smallness conditions of Theorem 4. Since $T^{\chi} \leq H \leq 2T^{\chi}$,

$$\begin{aligned} H^2 K^{-a} &= H^{2-a} T^{-a} \\ &\leq 2^{2-a} T^{\chi(2-a)-a} \leq 4. \end{aligned} \tag{92}$$

The last inequality follows from

$$\chi \leq \frac{a}{2-a} = \frac{p}{5p-4};$$

at the endpoint the exponent of T is zero, so no strict inequality is required. Consequently,

$$H\theta = c_\theta H K^{-a} \leq c_\theta H^2 K^{-a} \leq 4c_\theta \leq c_*,$$

where we used $H \geq 1$. Similarly,

$$\frac{H^2 \eta^2}{\theta} = \frac{c_\eta^2}{c_\theta} H^2 K^{-a} \leq \frac{4c_\eta^2}{c_\theta} \leq 4c_*^2 c_\theta \leq c_*^3 \leq c_*.$$

Also,

$$\frac{\eta}{\theta} = \frac{c_\eta}{c_\theta} \leq c_*.$$

Together with $H\theta \leq c_*$ and $H \geq 1$, this gives $\theta \leq c_* \leq 1/2$; moreover, $L\eta \leq 1/4$ by the choice of c_*. Thus Theorem 4 applies.

We next bound the estimator terms. From (8) and the choice of κ,

$$b_\kappa^2 = C_p^2 A_N^2 c_\kappa^{-2(p-1)} K^{-r}, \tag{93}$$

because $2(p-1)/(3p-2) = r$. Likewise,

$$\begin{aligned} q_\kappa &= C_p \left(s_N^2 + A_N c_\kappa^{2-p} K^{(2-p)/(3p-2)} \right) \\ &\leq Q_* K^{(2-p)/(3p-2)}, \end{aligned}$$

where the last step uses $K \geq 1$.

We now verify the order of each term in $\mathcal{R}_{T,H}$. Since $K = TH$,

$$\begin{aligned} \frac{\Delta_0}{K\eta} &= \frac{\Delta_0}{c_\eta} K^{-(1-a)} = \frac{\Delta_0}{c_\eta} K^{-r}, \\ \frac{\|G(\boldsymbol{X}_0)\|_n^2}{K\theta} &= \frac{\|G(\boldsymbol{X}_0)\|_n^2}{c_\theta} K^{-r}, \\ \theta q_\kappa &\leq c_\theta Q_* K^{-a+(2-p)/(3p-2)} = c_\theta Q_* K^{-r}, \end{aligned}$$

where

$$a - \frac{2-p}{3p-2} = r.$$

For the local-drift terms, (92) gives

$$\begin{aligned} H^2\eta^2 q_\kappa &\le c_\eta^2 Q_* H^2 K^{-2a+(2-p)/(3p-2)} = c_\eta^2 Q_* H^2 K^{-1} \\ &= c_\eta^2 Q_* \big(H^2 K^{-a}\big) K^{-r} \le 4c_\eta^2 Q_* K^{-r}, \end{aligned}$$

because $2a - (2-p)/(3p-2) = 1$ and $1 - a = r$. Similarly,

$$\begin{aligned} H\theta^2 q_\kappa &\le c_\theta^2 Q_* H K^{-2a+(2-p)/(3p-2)} = c_\theta^2 Q_* H K^{-1} \\ &= c_\theta^2 Q_* \big(H K^{-a}\big) K^{-r} \le 4c_\theta^2 Q_* K^{-r}. \end{aligned}$$

Combining these estimates with (93) yields

$$\mathcal{R}_{T,H} \le R_* K^{-r}. \tag{94}$$

The master bound therefore gives

$$\overline{\Phi} \le C_* R_* K^{-r},$$

which proves (90).

For consensus, the refined estimate (76) and (94) give

$$\max\{\overline{A}_{\mathrm{ep}}, \overline{A}\} \le 6\Gamma^2 C_* c_\eta^2 c_\theta^2 R_* H^4 K^{-(1+3a)} + 4\Gamma^2 c_\eta^2 c_\theta^2 Q_* H^3 K^{-(1+2a)}.$$

Since $HK^{-a} \le H^2 K^{-a} \le 4$, the first term is at most

$$24\Gamma^2 C_* c_\eta^2 c_\theta^2 R_* H^3 K^{-(1+2a)}.$$

Thus, by the definition of C_A,

$$\max\{\overline{A}_{\mathrm{ep}}, \overline{A}\} \le C_A H^3 K^{-(1+2a)}.$$

Finally, (92) implies $H^2 \le 4K^a$, hence

$$H^3 \le 8K^{3a/2}.$$

Therefore,

$$H^3 K^{-(1+2a)} \le 8K^{-(1+a/2)},$$

which proves the second inequality in (91).

If H is fixed, the assumed condition $H^2 K^{-a} \le 4$ provides the same smallness and rate estimates, so the argument above applies without change. □

Corollary 5 (Random-time output)**.** *Under the conditions of Corollary 4, draw $(\widehat{t}, \widehat{h})$ uniformly from $\{0, \dots, T-1\} \times \{0, \dots, H-1\}$ and I uniformly from $[n]$, independently of each other and of the algorithm. Define*

$$\widehat{\boldsymbol{X}} := \boldsymbol{X}_{\widehat{t},\widehat{h}}, \qquad \widehat{x} := x_{I,\widehat{t},\widehat{h}}, \qquad \widehat{\bar{x}} := \bar{x}_{\widehat{t},\widehat{h}}.$$

With $1/\infty := 0$, let

$$C_g := \sqrt{C_* R_*} + \frac{C_* R_*}{C} + L\sqrt{8C_A}.$$

Then the sampled local model satisfies

$$\mathbb{E}\,\|\nabla f(\widehat{x})\| \le C_g K^{-r/2}, \tag{95}$$

$$\mathbb{E}\,\big\|\widehat{x} - \widehat{\bar{x}}\big\|^2 \le C_A H^3 K^{-(1+2a)} \le 8C_A K^{-(1+a/2)}. \tag{96}$$

At the same random time, the local models jointly satisfy

$$\mathbb{E} \max_{i\in[n]} \Big\|\nabla f(x_{i,\widehat{t},\widehat{h}})\Big\| \le \left(\sqrt{C_* R_*} + \frac{C_* R_*}{C} + L\sqrt{8nC_A}\right) K^{-r/2}. \tag{97}$$

The corresponding matrix output satisfies

$$\mathbb{E}\left\|JG(\widehat{\boldsymbol{X}})\right\|_n \le C_g K^{-r/2}, \qquad \mathbb{E}\left\|\Pi\widehat{\boldsymbol{X}}\right\|_n^2 \le 8C_A K^{-(1+a/2)}. \tag{98}$$

In particular, for $\chi = p/(5p-4)$, the sufficient budgets for $\mathbb{E}\,\|\nabla f(\widehat{x})\| \le \varepsilon$ are

$$K = O\Big(\varepsilon^{-\frac{3p-2}{p-1}}\Big), \qquad T = O\Big(\varepsilon^{-\frac{5p-4}{2(p-1)}}\Big), \tag{99}$$

where $K = TH$ is the number of local updates per agent and T is the number of communication rounds.

Proof. Uniform sampling gives

$$\mathbb{E}\phi_C(\nabla f(\widehat{\widehat{x}})) = \overline{\Phi}, \qquad \mathbb{E}\left\|\Pi\widehat{\boldsymbol{X}}\right\|_n^2 = \overline{A}, \qquad \mathbb{E}\left\|\widehat{x} - \widehat{\widehat{x}}\right\|^2 = \overline{A}.$$

The last identity follows by conditioning on $\widehat{\boldsymbol{X}}$ and averaging over the independent uniform agent I.

By (88),

$$\mathbb{E}\,\|\nabla f(\widehat{\widehat{x}})\| \le \sqrt{\overline{\Phi}} + \frac{\overline{\Phi}}{C} + L\sqrt{\overline{A}}. \tag{100}$$

Corollary 4 gives

$$\overline{\Phi} \le C_* R_* K^{-r}, \qquad \overline{A} \le 8C_A K^{-(1+a/2)}.$$

Since $K \ge 1$ and

$$1 + \frac{a}{2} \ge r = 1 - a,$$

we have

$$K^{-(1+a/2)/2} \le K^{-r/2}, \qquad K^{-r} \le K^{-r/2}.$$

Substituting these inequalities into (100) proves (95), while (96) follows directly from (91).

For the all-agent bound, smoothness gives, at the sampled time,

$$\begin{aligned}\max_{i\in[n]}\left\|\nabla f(x_{i,\widehat{t},\widehat{h}})\right\| &\le \left\|\nabla f(\widehat{\widehat{x}})\right\| + L\max_{i\in[n]}\left\|x_{i,\widehat{t},\widehat{h}} - \widehat{\widehat{x}}\right\| \\ &\le \left\|\nabla f(\widehat{\widehat{x}})\right\| + L\sqrt{n}\,\left\|\Pi\widehat{\boldsymbol{X}}\right\|_n.\end{aligned}$$

Applying (46), taking expectations, and using Jensen's inequality together with the bounds above yields (97). The factor $\sqrt{n}$ arises from converting the normalized aggregate disagreement into a maximum over agents.

For the matrix stationarity residual, write

$$\begin{aligned}\|JG(\boldsymbol{X})\|_n &\le \|JG(J\boldsymbol{X})\|_n + \big\|J\big(G(\boldsymbol{X}) - G(J\boldsymbol{X})\big)\big\|_n \\ &\le \|\nabla f(\bar{x})\| + L\,\|\Pi\boldsymbol{X}\|_n,\end{aligned}$$

where $\|JG(J\boldsymbol{X})\|_n = \|\nabla f(\bar{x})\|$ and the second term is bounded by smoothness and Jensen's inequality. Taking expectations and applying the same bounds as in (95) proves the first inequality in (98); the second is exactly the consensus bound from Corollary 4.

Finally, $\mathbb{E}\,\|\nabla f(\widehat{\widehat{x}})\| \le C_g K^{-r/2}$ implies the sufficient local-update budget

$$K = O(\varepsilon^{-2/r}) = O\Big(\varepsilon^{-\frac{3p-2}{p-1}}\Big).$$

For $H = \lceil T^\chi \rceil$,

$$T^\chi \le H \le 2T^\chi \quad \Longrightarrow \quad K = TH = \Theta(T^{1+\chi}),$$

and hence

$$T = \Theta\Big(K^{1/(1+\chi)}\Big).$$

At $\chi = p/(5p-4)$,

$$\frac{2}{r(1+\chi)} = \frac{5p-4}{2(p-1)},$$

which gives the communication order in (99). □

Resource counts and interpretation. For Algorithm 1 as written, the per-agent resource counts are

$$N_{\mathrm{rounds}} = T, \qquad N_{\mathrm{local}} = K = TH, \qquad N_{\mathrm{raw}} = 3mK,$$

where one communication round consists of exchanging both the model and tracker with neighboring agents. These counts include the initial communication, the first no-op model update caused by $Y_0 = 0$, and the query after the final model update. If only a model iterate is returned, the final query, which updates only the filter and tracker states, can be omitted; this changes the exact sample count by one batch but not its asymptotic order.

For any $0 \le \chi \le p/(5p-4)$, the growing-block schedule gives the sufficient orders

$$N_{\mathrm{local}} = O\left(\varepsilon^{-\frac{3p-2}{p-1}}\right), \qquad N_{\mathrm{raw}} = O\left(\varepsilon^{-\frac{3p-2}{p-1}}\right),$$
$$N_{\mathrm{rounds}} = O\left(\varepsilon^{-\frac{3p-2}{(p-1)(1+\chi)}}\right).$$

for the expected gradient-norm criterion in (95). Thus increasing χ reduces the sufficient communication order while leaving the local-update and sample orders unchanged, since the per-update batch size $3m$ is fixed.

These guarantees concern a uniformly sampled pre-update local model rather than the terminal iterate. The simultaneous all-agent bound (97) additionally incurs a $\sqrt{n}$ factor when aggregate disagreement is converted to a maximum over agents. The analysis does not establish linear speedup in n or optimal communication complexity. Its constants depend on the initial local gradients and network contraction, but no uniform bound on gradient dissimilarity across agents is assumed.

G Experimental Details

This appendix describes the experimental protocols and resource accounting used in the reported results. Learning curves are evaluated at the arithmetic mean of the agents' parameters. Communication is measured in directed-link bytes: transmitting a parameter-sized vector across one directed edge costs db bytes, where d is the number of transmitted parameters and $b = 4$ for float32. Model and tracker payloads are counted separately. Communication rounds, transmitted bytes, gradient queries, training samples, and tokens are therefore distinct resource measures. Unless stated otherwise, uncertainty bands show sample standard deviations across seeds rather than confidence intervals.

Relation to the heavy-tail assumptions. For the image-classification robustness experiments, stochastic gradients are perturbed through multiplicative loss weights of the form

$$1 + \rho\sigma z, \qquad z \sim t_{1.5},$$

where the draws of z are independent and centered. Conditional on the unperturbed per-example gradient, multiplication by this independent mean-one factor preserves its conditional expectation. Since a Student-t random variable with 1.5 degrees of freedom has finite absolute moments of order $p < 1.5$ but infinite variance, these perturbations provide a controlled source of heavy-tailed randomness. They do not, however, establish the uniform conditional p-moment bound over all model parameters required by Assumption 2. Accordingly, the neural-network experiments evaluate finite-budget robustness rather than directly verify the theorem's oracle assumptions; the synthetic additive-noise experiments provide the controlled moment examples.

In these perturbation experiments, ρ controls the noise intensity and σ is a calibrated experimental scale. This σ is not identified with the uniform conditional moment bound in Assumption 2. The empirical hyperparameter schedules are also selected for the experiments and need not satisfy the sufficient tuning conditions used in the convergence theory.

Method names and implementation controls. GT-NSGDm denotes the normalized momentum and gradient-tracking method of Yu et al. (2026). For $H > 1$, Local-GT-NSGDm denotes our implemented variant that omits intermediate communication; it should not be interpreted as the original method or as covered by its published guarantee. The label Local-D-SGD is used analogously. K-GT and LSGT already incorporate local updates. A duplicate implementation labeled Local-D-SGT/LSGT is reported only once under LSGT.

Some figure legends retain the historical labels "E" or "A3-NL-E" for CRAFT with fixed-radius direction clipping, and "filter" denotes the recursive exponential smoothing of the corrected gradient estimates. The supplementary direction-map controls are

$$P(y) := y, \qquad NR(y) := \frac{y}{\tau + \|y\|}, \qquad \tau > 0.$$

These controls replace only the direction map unless otherwise stated. The clipping radius C and the normalization parameter τ are distinct, even where the implementation uses the variable name `tau` for C. The convergence analysis includes the identity map $P(y) = y$ through the case $C = \infty$, whereas the reported CRAFT configurations use finite direction clipping.

Batch accounting and reproducibility. Throughout the experiments, batch size refers to the number of training examples or tokens used by one logical agent for one gradient query at one local update; it is distinct from a system-wide or evaluation batch. CRAFT uses the same batch to construct its three block estimates and pooled mean, so the correction requires no additional samples. MNIST and GPT-2 (124M) use three equal blocks. CIFAR-10 and Multi30K use nearly equal blocks with size-weighted pooling; these are empirical extensions of the exact $N = 3m$ estimator analyzed in the theory.

The available experimental record includes the reported figures and aggregate statistics but not complete per-run logs or all training code. The protocol below records the available settings, while some baseline tuning grids and architecture or preprocessing details are unavailable. These limitations should be taken into account when interpreting reproducibility. The theoretical guarantees concern the estimator and algorithm specified in the main text and do not depend on these empirical extensions.

G.1 MNIST: LOCAL STEPS AND BUDGET ACCOUNTING

The MNIST communication- and sample-budget comparisons in Figures 2 and 6 use 16 agents.

Data, model, and perturbations. We use 55,000 MNIST training examples and a 5,000-example holdout set. Training data are partitioned across agents using a label-Dirichlet distribution with concentration parameter $\alpha = 0.3$ and partition seed 20260824. At each local query, an agent samples 126 examples with replacement from its fixed local shard. Inputs use the standard dataset normalization without training-time augmentation.

The convolutional model has 21,840 parameters and no batch-normalization layers. Evaluation uses the arithmetic mean of the 16 agent models on all 10,000 MNIST test examples. Agents communicate over a fixed 16-node ring with self and neighbor weights $1/3$ and no additional consensus steps.

For the heavy-tail perturbation, each per-example loss is multiplied by

$$1 + \sigma z, \qquad z \sim t_{1.5}, \qquad \sigma = 0.6149446274571448,$$

corresponding to the calibrated intensity $\rho = 1$. These are multiplicative loss-weight perturbations rather than additive Gaussian gradient noise. CRAFT partitions each 126-example batch into three 42-example blocks and computes both the median anchor and pooled correction from the same batch; the correction therefore requires no additional samples.

Communication and sample budgets. For this model and ring topology, one model-equivalent communication unit corresponds to

$$2ndb = 2 \times 16 \times 21{,}840 \times 4 = 2{,}795{,}520$$

directed-link bytes, where $d = 21{,}840$ and float32 uses $b = 4$ bytes per parameter. CRAFT exchanges both the model and tracker at each communication boundary and therefore uses two model-equivalent payloads per round. Local-D-SGD exchanges only the model and uses one.

With 256 local updates per agent, CRAFT uses $256/H$ communication rounds and $512/H$ model-equivalent communication units. To match this endpoint communication budget, the Local-D-SGD

reruns use 512 local updates and $512/H$ communication rounds. Consequently, the cumulative system-wide training-sample counts are

$$16 \times 256 \times 126 = 516{,}096$$

for CRAFT and

$$16 \times 512 \times 126 = 1{,}032{,}192$$

for Local-D-SGD.

The lower-row display limit in Figure 6 is a plotting crop rather than a shortened rerun. Recorded initialization samples are included on the LSGT sample axis, and its final displayed segment crosses the crop boundary; that segment is clipped visually without adding a synthetic evaluation or endpoint.

For K-GT at $H = 2$, an independently versioned dense rerun records rounds $0, 8, \ldots, 128$ for each of its three completed seeds. These runs reproduce the previously recorded terminal metrics. Seeds 20261003 and 20261004 again reach non-finite states at iterations 116 and 64, respectively. The displayed K-GT mean and band therefore use the same three completed seeds at every evaluation point; prefixes from the two incomplete runs are not included, and no smoothing or imputation is applied.

Hyperparameters and interpretation. Table 2 reports the constant-step CRAFT configurations used for each communication period. The implementation argument `tau` denotes the direction-clipping radius C, not the regularization parameter of the normalized direction map NR. At $H = 1$ and $H = 2$, $\theta = 1$, so the recursive filter reduces to the current corrected query without temporal averaging.

These configurations represent separately tuned operating points rather than a controlled ablation of the local period, and they are not intended to reproduce the asymptotic tuning in the convergence theorem. In the equal-communication comparisons, Local-D-SGD uses learning rate 0.01 for $H = 2, 4, 8$. Thus, increasing H directly reduces communication at a fixed CRAFT sample budget, but differences in accuracy across H should also be interpreted in light of the separately chosen hyperparameters and finite training horizon.

Seeds and incomplete runs. The five planned seeds are 20261001–20261005. CRAFT, including its $H = 1$ configuration in the cross-H comparison, and Local-GT-NSGDm complete all five seeds. Local-D-SGD and LSGT complete four seeds at each displayed H, while K-GT completes three seeds at $H = 2$ and $H = 4$ and four at $H = 8$. Local-D-SGD seed 20261004 is incomplete for every displayed H; for K-GT at $H = 2$, seeds 20261003 and 20261004 are incomplete.

For each method and setting, the displayed trajectory uses a fixed cohort of completed runs at all evaluation points. Bands are sample standard deviations over that cohort, not confidence intervals, and incomplete runs are not encoded as zero accuracy. The resulting statistics are therefore conditional on completion and should be interpreted together with the reported completion counts. Local-D-SGT coincides with LSGT in this implementation and is not counted as a separate baseline. No smoothing or splicing of trajectories from different runs is used.

Matched estimator and recursive-filter ablation. Figure 9(e) uses the same heavy-tail MNIST task with 16 agents, $H = 2$, and 128 communication rounds. All four variants use the same carried-tracker recurrence, direction-clipping radius $C = 0.5$, learning rate 0.2, and batch size 126. When filtering is enabled, $\theta = 0.3$; the corrected-MoM variants use correction radius $\kappa = 3$.

The five paired seeds are 20260910 and 20261001–20261004. Within each seed, initialization and the raw sample, perturbation, and evaluation streams are shared across variants. Each run therefore uses

$$256 \times 126 = 32{,}256$$

training samples per agent and

$$256 \times 2{,}795{,}520 = 715{,}653{,}120$$

cumulative directed-link bytes system-wide. All 20 configuration–seed runs complete. Validation loss is evaluated every eight communication rounds, and Figure 9(e) displays all recorded evaluations without smoothing or interpolation.

The uncorrected-MoM control replaces Z_κ by the three-block coordinatewise median R. The effective pooled-mean control sets $\kappa = 10^9$; the correction is never observed to activate at the recorded post-update evaluations, so this configuration behaves as the pooled mean on those recorded states. The no-filter control changes only θ to 1. All four variants retain the same direction clipping and carried gradient tracker.

G.2 CIFAR-10 AND GPT-124M: EQUAL-COMMUNICATION PROTOCOLS

Run completion and uncertainty. The CIFAR-10 D-SGD supplements contain 15 completed configuration–seed runs across the three matched communication budgets. The clean GPT-124M benchmark contains all 75 planned method–topology–seed runs: five methods, three graph topologies, and five seeds. Each run starts from the stated pretrained initialization. The CIFAR runner additionally records the final update when it falls outside the regular 50-update evaluation grid.

Within each displayed CIFAR-10 or GPT-124M group, all five seeds are used at every recorded evaluation point. Curves show arithmetic means with sample-standard-deviation bands, using denominator $n-1$; the bands are not confidence intervals. No smoothing, interpolation, extrapolation, or selective removal of finite trajectories is applied. These complete cohorts are distinct from the incomplete baseline cohorts reported for MNIST.

G.2.1 CIFAR-10 SETTINGS AND ACCOUNTING

The CIFAR-10 experiment (Krizhevsky, 2009) uses IID data partitions across 16 logical agents and a pretrained ResNet18_M variant (He et al., 2016) with image size 56. The batch size is 64 per agent and gradient query. CRAFT partitions each batch into blocks of sizes 22, 21, and 21 and forms the pooled mean using the corresponding block-size weights. This is an empirical extension of the equal-block estimator analyzed in the theory.

All methods use $H = 1$, learning rate 0.1, warmup length 60, no additional optimizer momentum or weight decay, and learning-rate decay milestones at 2400 and 4800, both beyond the reported horizons. CRAFT uses direction radius $C = 0.1$ (stored as `tau` in the implementation), filter weight $\theta = 0.1$, and correction radius $\kappa = 1$. GT-NSGDm uses momentum coefficient 0.9683772233983162. The seeds are 209, 666, 756, 866, and 930.

For the robustness experiment, each per-example loss is multiplied by

$$1 + \sigma z, \qquad z \sim t_{1.5}, \qquad \sigma = 0.625600248298246\,\rho, \qquad \rho \in \{0.50, 0.75, 1.00\}.$$

CRAFT uses the same hyperparameters, including learning rate 0.1, at all three perturbation levels.

The model has $d = 11{,}181{,}642$ parameters. On the 16-node ring, one model-equivalent communication unit corresponds to

$$2ndb = 2 \times 16 \times 11{,}181{,}642 \times 4 = 1{,}431{,}250{,}176$$

directed-link bytes. Every method is run to an endpoint budget of 2002 such units, corresponding to

$$2{,}865{,}362{,}852{,}352$$

directed-link bytes. D-SGD transmits one model payload per synchronization and therefore performs 2002 updates, whereas the tracking methods transmit both model and tracker and perform 1001 updates at the same payload budget. Thus synchronization counts, payload counts, and update counts are not interchangeable.

At batch size 64, the corresponding system-wide training-sample counts are

$$16 \times 1001 \times 64 = 1{,}025{,}024$$

for CRAFT and the other 1001-update methods, and

$$16 \times 2002 \times 64 = 2{,}050{,}048$$

for D-SGD. LSGT additionally counts 1024 initialization samples. D-SGD has 42 recorded evaluations at indices $0, 50, \ldots, 2000, 2001$, whereas the tracking methods have 21 at $0, 50, \ldots, 1000$. These are zero-based update indices, so the final index corresponds to one more completed update

than its numerical value alone might suggest. Evaluation uses the mean-parameter model, worker-0 buffers, and the same validation aggregation procedure across methods.

Figure 7 displays the common 1,025,024-sample range. This is a plotting viewport rather than a claim that all runs terminate at the same sample budget; segments crossing the boundary are clipped without adding synthetic endpoints.

The $\rho = 0.50$ and $\rho = 0.75$ D-SGD supplements were run on A100 GPUs, whereas the $\rho = 1.00$ supplement was run on H100 GPUs. The remaining CIFAR-10 runs use A100 GPUs. We therefore do not use these experiments for uniform-hardware timing comparisons or attribute every numerical difference across perturbation levels solely to ρ. Table 4 reports all five displayed methods.

We retain LSGT as the sole representative of the implemented D-SGT/LSGT pair in the $H = 1$ comparison. In these implementations,

$$W(\boldsymbol{X} - \eta Y) = W\boldsymbol{X} - \eta WY,$$

and the tracker updates are otherwise the same. Although floating-point evaluation order can produce different numerical trajectories, we do not treat the two implementations as independent algorithmic baselines. Accordingly, only LSGT is plotted and tabulated. Local-D-SGD is a distinct non-tracking baseline and remains included.

G.2.2 GPT-124M SETTINGS AND TOPOLOGY BUDGETS

We fine-tune the pretrained GPT-2 small initialization (Radford et al., 2019) on WikiText-103 (Merity et al., 2017). The experiment uses four logical agents, context length 256, 12 raw sequences per agent and gradient query, microbatch size one for computational accumulation, local period $H = 16$, and float32 arithmetic. CRAFT divides each 12-sequence query into three equal four-sequence blocks and computes the pooled correction from the same samples. No synthetic heavy-tailed perturbation is injected.

Seeds 20261001–20261005 are run on each fixed graph. Validation uses the same frozen set of 931 sequences and is performed every two communication blocks. The five seeds therefore represent independent training realizations on each fixed topology, not independent graph realizations.

CRAFT uses learning rate 0.01, direction radius $C = 0.4$, filter weight $\theta = 0.02$, and correction radius $\kappa = 10$. Local-GT-NSGDm uses learning rate 0.003, while Local-D-SGD, K-GT, and LSGT use 0.0003. Each method uses one configuration across the Ring, ER, and Star graphs, but configurations differ across methods. These experiments therefore compare separately configured methods rather than matched effective stepsizes or equal tuning budgets.

Ring seed 20261001 was included in CRAFT's eight-candidate parameter selection and is therefore not fully held out. As a sensitivity check, the terminal validation loss over the other four Ring seeds is 3.3156 ± 0.0038, close to the corresponding five-seed summary.

The Ring, ER, and Star graphs have 8, 10, and 6 directed edges, respectively. With $d = 124{,}439{,}808$ parameters and four bytes per float32 scalar, the matched endpoint budgets are 637.131817, 796.414771, and 477.848863 GB, respectively, in decimal units. These quantities count algorithmic model and tracker payloads over directed edges; diagnostic and evaluation collectives are excluded, so they should not be interpreted as measured hardware-network traffic.

Local-D-SGD sends one parameter-sized payload per communication block and runs for 160 blocks, corresponding to 2560 local updates. Each tracking method sends both model and tracker payloads and therefore runs for 80 blocks, corresponding to 1280 local updates, at the matched endpoint byte budget. Evaluations include the initial model and then occur every second block, giving 81 recorded points for Local-D-SGD and 41 for each tracking method.

The resulting system-wide training-token counts are 31,457,280 for Local-D-SGD; 15,728,640 for CRAFT, Local-GT-NSGDm, and K-GT; and 15,740,928 for LSGT, including its initialization cost. Evaluation tokens are excluded. Equal endpoint communication bytes therefore do not imply equal numbers of training tokens, local updates, or synchronization rounds.

Verified full-state continuations are concatenated without smoothing or approximate restart reconstruction. Some continuations changed between A100 and H100 hardware, so the reported results are not used to make wall-clock comparisons or claims of bitwise reproducibility across hardware.

G.3 CONTROLLED MECHANISM EXPERIMENTS

Exact query calculations. For $p = 1.5$ and $\sigma = 1$, define

$$c_\varepsilon := \left[(1-\varepsilon) + (1-\varepsilon)^p \varepsilon^{1-p}\right]^{-1/p}.$$

The centered scalar noise is

$$\zeta = \begin{cases} -c_\varepsilon, & \text{with probability } 1-\varepsilon, \\ c_\varepsilon(1-\varepsilon)/\varepsilon, & \text{with probability } \varepsilon, \end{cases}$$

so that $\mathbb{E}\zeta = 0$ and $\mathbb{E}|\zeta|^p = 1$. Each individual law is bounded, although the rare-event magnitude increases as ε decreases.

We evaluate 481 logarithmically spaced values of $\varepsilon \in [10^{-12}, 0.5]$ with $m = 8$, together with 81 logarithmically spaced values of

$$\frac{\kappa}{s_N} \in [1, 1000], \qquad s_N = 24^{-1/3}.$$

For each parameter pair, all $(m+1)^3 = 729$ triples of block counts are enumerated exactly. The worst-case bias and MSE are maximized separately over the same ε grid, so their maximizing distributions need not coincide. The normalized product formed from these two maxima lies between 0.610 and 0.620 for $\kappa/s_N \geq 32$. This calculation is a finite-grid diagnostic and does not replace the supremum over the full moment class in Theorem 2.

Paired asymmetric-noise optimization. Eight agents communicate over a ring with self and neighbor weights $1/3$. Their local objectives are

$$f_i(x) := \frac{1}{2}(x - a_i)^2, \qquad a_i := 0.35 \cos\left(\frac{2\pi i}{8}\right),$$

so the global gradient is $\nabla f(x) = x$. All agents start from $x = 1.2$ with $M = Y = 0$. We use $H = 4$, $m = 8$, and 10,000 communication rounds, corresponding to 40,000 local updates per agent.

For the bounded asymmetric noise, we set $\varepsilon = 1/64$ in the two-point construction above. We also consider Pareto noise

$$\zeta := a(V - 3), \qquad \Pr(V > v) = v^{-1.5}, \quad v \geq 1, \qquad a := \left[2^{1/4}(6 + 3^{5/4})\right]^{-4/5}.$$

Since $\mathbb{E}V = 3$, this noise is centered. It has $\mathbb{E}|\zeta|^{1.25} \leq 1$ and infinite variance. The raw noise is neither truncated nor annealed.

All methods use the same 24 raw samples per local query, initialization, and round-indexed stepsize

$$\eta_t := 0.05(1 + t/50)^{-0.75}.$$

Filtered variants use $\theta = 0.2$, and all displayed methods use direction radius $C = 1$. Direct MoM sets $\theta = 1$, while corrected-query variants use fixed $\kappa \in \{0.5, 2\}$. Filtered MoM and filtered pooled-mean controls are also included. These fixed-θ, fixed-κ configurations are designed to isolate mechanisms rather than reproduce the horizon-dependent tuning of the convergence theorem; in particular, a fixed correction radius need not make the estimator bias vanish.

The 20 paired seeds are 2026091101–2026091120, and all are retained. Within each seed, methods share the same underlying sample streams. Trajectories are recorded every 100 local updates without smoothing. Terminal statistics first average the final 1,000 pre-update iterates within each seed before aggregating across seeds.

Supplementary skewed-heavy-tail nonconvex ablation. A separate component ablation uses the nonconvex objective

$$f(x) := \frac{1}{2}\|x\|^2 + 2\sum_{j=1}^{2}(1 - \cos x_j), \qquad \nabla f(x) = x + 2\sin x. \tag{101}$$

Its Hessian is diagonal with entries $1 + 2\cos x_j \in [-1, 3]$, so f is globally 3-smooth, nonconvex, and bounded below.

Sixteen agents communicate over the fixed ring with self and nearest-neighbor weights $1/3$, using $H = 4$ and

$$f_i(x) := f(x) + b_i^\top x, \qquad b_i := 0.3\left(\cos\frac{2\pi i}{16}, \sin\frac{2\pi i}{16}\right).$$

The vectors b_i are explicitly recentered in the implementation to have zero numerical mean. All agents start from $(4, -3)$ with $M = Y = 0$.

Each raw oracle coordinate receives independent additive noise

$$0.5(V - 3), \qquad \Pr(V > v) = v^{-1.5}, \quad v \geq 1.$$

The noise is centered and has infinite variance; the finite moment order used for this diagnostic is $p = 1.25$. Samples are fresh and independent, with no additional shocks, truncation, or noise annealing. Each agent query uses three blocks of $m = 2$ samples, for six raw vector samples per query.

Matched updates and schedules. At each communication boundary, the model and tracker are mixed according to

$$\boldsymbol{X} \leftarrow W\boldsymbol{X}, \qquad Y \leftarrow WY,$$

while M remains local and is carried across the boundary. Each local update first applies the clipped tracker direction to the model, then queries the new model and updates M and Y as in Algorithm 1. Both correction and direction clipping use the Euclidean norm.

For zero-based communication round t, define

$$a_t := 1 + \frac{t}{100}.$$

The four local updates in that round use

$$\eta_t := 0.01a_t^{-5/7}, \qquad \theta_t := 0.1a_t^{-5/7}, \qquad \kappa_t := 2a_t^{4/7}, \qquad C := 1.$$

The no-filter control changes only θ_t to 1, while the no-correction control replaces Z_{κ_t} by the MoM anchor R and retains filtering. All variants retain direction clipping. Two additional no-filter controls multiply η_t by $1/2$ and 2, respectively, while leaving the remaining settings unchanged.

Within each seed, all configurations use the same raw noise for every agent, local-update index, sample, and coordinate. Each formal run contains 50,000 communication rounds and 200,000 local updates per agent. Since each query uses six raw vectors, this corresponds to 1,200,000 raw vector samples per agent and 19,200,000 across the 16-agent network. The model and tracker are communicated identically across configurations.

Seeds, recording, and uncertainty. Five pilot seeds, 2026091401–2026091405, are run for 10,000 rounds only to check numerical finiteness and the tracker-mean identity. The pilot gate contains no performance-ranking or residual threshold, and all algorithmic parameters are fixed before the pilot. The 20 disjoint formal seeds are 2026091501–2026091520. All 100 formal configuration–seed runs complete and are retained.

The maximum observed numerical error in the tracker-mean identity is below 8.7×10^{-14}, and the recorded trajectories enter regions of negative curvature. At every completed communication boundary, we compute

$$G_t := \|\nabla f(\bar{x}_t)\|^2.$$

The reported trajectories retain every 50th completed boundary, including the initialization and endpoint, with no temporal smoothing, running minimum, or interpolation of missing evaluations. Bands are pointwise 95% percentile bootstrap intervals for the seed mean, based on 2,000 whole-trajectory resamples of the 20 seeds. They are descriptive pointwise intervals rather than simultaneous bands, and no finite-sample coverage claim under heavy-tailed noise is made.

Spectral-screen display subset. Figure 9(f) compares neighbor weights

$$\alpha \in \{0.10, 0.20, 1/3\}$$

on the same 16-node ring, with self-weight $1 - 2\alpha$. The corresponding spectral gaps $\delta := 1 - \|W - J\|_2$ are approximately 0.01522, 0.03045, and 0.05075, respectively. The three displayed settings use seed 666, $H = 2$, 501 communication rounds, and identical sample and communication-byte budgets. These are single-seed, finite-budget diagnostics and do not establish monotonic dependence of optimization performance on the spectral gap.

Original nonconvex toy and sensitivity. The original toy used in Figure 9(g–h) and Table 6 is distinct from the supplementary skewed-noise experiment above. It uses 16 agents, a fixed ring with neighbor weights $1/3$, $H = 2$, and

$$f_i(x) := 0.05\,\|x\|^2 + (1 - \cos x_1) + (1 - \cos x_2) + 0.1\big(1 - \cos(x_1 - x_2)\big) + b_i^\top x, \qquad (102)$$

where the fixed b_i have zero mean and heterogeneity scale 0.5. All agents start from $(4, -3)$. The additive noise is symmetric Student-t with 1.5 degrees of freedom and scale 0.5, and each query uses three blocks of eight samples.

The baseline configuration uses

$$\eta_0 = 0.03, \qquad \theta = 0.20, \qquad C = 1, \qquad \kappa = 0.50,$$

with local-update stepsizes

$$\eta_s = \eta_0(1 + s/100)^{-0.75}$$

over 100,000 local updates. The 20 paired seeds are 20261001–20261020. The implementation variable `tau` denotes the direction-clipping radius C in these runs; it is distinct from the regularization parameter τ used for the supplementary normalized direction map NR.

Wall-clock local-step ablation. The wall-clock experiment compares CRAFT, D-SGD, GT-NSGDm, K-GT, and LSGT on the original symmetric-noise nonconvex toy in (102), rather than on the supplementary skewed-heavy-tail objective (101). The ring, initialization, Student-$t_{1.5}$ oracle, stepsize schedule, and 20 paired seeds are shared across methods. Each run performs 10,000 system-level local updates and records a checkpoint every 100 updates. The baseline methods use the mean of the same three sample blocks, whereas CRAFT uses its corrected and recursively filtered query.

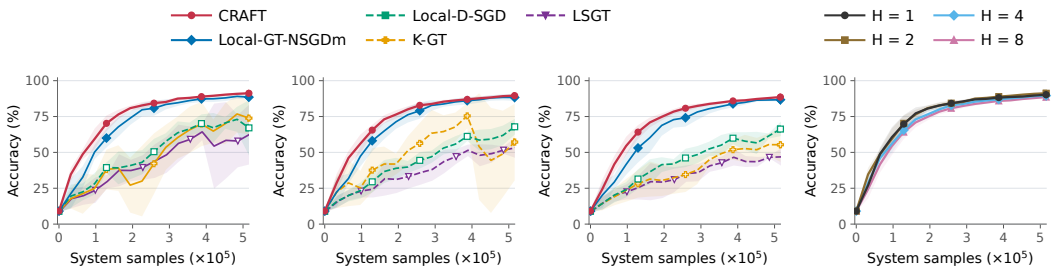
Panels (a)–(c) of Figure 10 compare the five methods at $H \in \{1, 4, 8\}$. Panel (d) separately compares CRAFT at $H \in \{1, 2, 4, 8\}$ under the same protocol and paired seeds. Each panel plots the global gradient norm against measured per-run CPU wall-clock time. Timing begins before method-specific state initialization. At each checkpoint, the curve reports the mean elapsed time and mean gradient norm across seeds, with empirical 10th–90th seed-percentile bands for the gradient norm. All panels use the same horizontal time range, and each trajectory ends at its measured completion time. No smoothing or interpolation of the time axis is applied.

D-SGD and GT-NSGDm are originally specified as single-step communication methods in the corresponding comparison. Their $H > 1$ implementations are therefore labeled Local-D-SGD and Local-GT-NSGDm to distinguish these engineering extensions from the original recurrences. K-GT and LSGT natively support local steps. Under the local extension used in this toy, D-SGT and LSGT have the same state recurrence; we therefore retain only the LSGT trajectories and measured times and do not pool the duplicate D-SGT timing runs.

G.4 Clean Multi30K diagnostics

G.4.1 Paired temporal protocol

Figure 11 reports a paired temporal comparison on the clean Multi30K same-position token-reconstruction diagnostic. The experiment uses GPT-3M and eight agents connected by a fixed ring with self and neighbor weights $1/3$. The nominal per-agent query size is 64 sentences, the local period is $H = 8$, and each run contains 684 local optimizer updates.



(a) $H = 2$ (b) $H = 4$ (c) $H = 8$ (d) CRAFT across H

Figure 6: Sample-budget view of the MNIST runs in Figure 2. All panels use the same 516,096-sample display range; D-SGD continues beyond this range in the matched-communication comparison. Fixed completed-seed cohorts, sample-SD bands, and hollow markers follow the conventions of Figure 2. Initialization samples are included for LSGT. Segments crossing the display boundary are clipped without adding synthetic evaluations or endpoints.

Each epoch contains 56 full 64-sentence queries followed by one retained 41-sentence tail query. For CRAFT, the full queries are partitioned into blocks of sizes $22/21/21$, while the tail query is partitioned as $14/14/13$. The pooled term is weighted by the number of non-padding tokens in each block. These unequal-block constructions are empirical extensions of the equal-block estimator analyzed in the theory. No synthetic Student-t perturbation is injected.

The 684 local updates form 86 communication blocks: 85 full blocks of eight updates and a final block of four updates. All 86 block-end evaluations are retained without smoothing or interpolation. Consensus RMS is measured after the local updates of each block and before the mixing step that begins the next block. The accompanying accuracy metric is validation token accuracy evaluated at the arithmetic mean of the agent parameters.

The two methods use the same five formal seeds, 20260926–20260930, which are disjoint from the tuning-probe seed 20260925; the data-partition seed is 20260824. CRAFT uses learning rate 0.45, filter weight $\theta = 0.035$, direction-clipping radius $C = 0.5$, and correction radius $\kappa = 10$. The normalized-direction reference (NR) uses learning rate 0.3, filter weight $\theta = 0.1$, normalization parameter $\tau = 0.003$, and $\kappa = 10$.

The comparison is paired through the common seeds and experimental protocol, but the methods use separately selected hyperparameters. Accordingly, it is intended as a finite-budget comparison of the two configured methods rather than a one-factor ablation of the direction map.

H Additional Experimental Results

This appendix reports additional sample-budget comparisons, endpoint statistics, and mechanism experiments for the protocols described in Appendix G.

H.1 MNIST operating points

Table 2 summarizes the CRAFT configurations used in the MNIST local-step comparison and their terminal test accuracy. All configurations use the same number of local updates and training samples, while the communication budget decreases with H. Protocol details and run-completion counts are given in Appendix G.1.

H.2 MNIST estimator and recursive-filter ablation

Table 3 reports the paired endpoint statistics for the component ablation in Figure 9(e). The shared experimental protocol is described in Appendix G.1.

Because initialization, sample streams, stepsize, direction clipping, tracking, and communication are matched across variants, this experiment provides a controlled comparison of the query and filtering

Table 2: CRAFT configurations and terminal mean-model test accuracy on MNIST. All rows use 256 local updates per agent and 516,096 training samples system-wide. Communication is reported in model-equivalent units, and accuracy is reported as mean ± sample standard deviation over five seeds.

H	Learning rate	θ	C	κ	Communication units	Accuracy (%)
1	0.25	1.0	0.5	20	512	90.21 ± 0.31
2	0.40	1.0	0.5	20	256	91.23 ± 0.41
4	0.20	0.6	0.5	3	128	89.56 ± 1.06
8	0.11	0.2	0.5	2	64	88.57 ± 0.81

Table 3: Paired five-seed MNIST query/filter ablation at $H = 2$ and communication round 128. Entries are mean ± sample SD across seeds. Accuracy and loss are evaluated using the mean model on the validation set.

Query and filter	Accuracy (%)	Loss	Consensus RMS
Pooled mean + filter	89.66 ± 0.56	0.3747 ± 0.0399	0.9377 ± 0.4717
MoM + filter	92.66 ± 0.53	0.2409 ± 0.0205	0.3774 ± 0.0062
Corrected MoM + filter	93.22 ± 0.45	0.2296 ± 0.0194	0.3530 ± 0.0034
Corrected MoM, no filter	86.81 ± 0.76	0.4781 ± 0.0416	0.2057 ± 0.0028

choices under the shared settings. It does not compare separately tuned configurations for each variant. The carried gradient tracker and direction clipping are used in all four configurations and are therefore not individually ablated. In particular, consensus RMS should be interpreted together with the optimization metrics: lower disagreement alone does not imply lower loss or higher accuracy.

H.3 CIFAR-10: FINAL ACCURACY AND SAMPLE-BUDGET COMPARISONS

Table 4 reports terminal validation accuracy at the matched communication budget, while Figure 7 provides the complementary view over a common sample-budget range. Experimental settings and resource-accounting conventions are given in Appendix G.2.1.

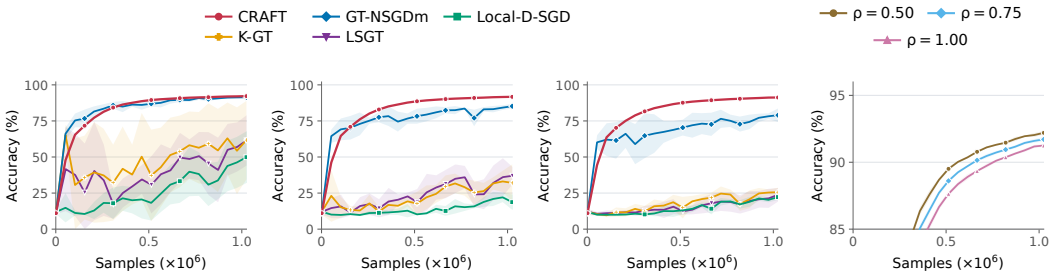
Table 4: Terminal CIFAR-10 validation accuracy (%) at a matched communication budget of 2002 model-equivalent units. Entries are mean ± sample SD over five seeds.

Method	$\rho = 0.50$	$\rho = 0.75$	$\rho = 1.00$
CRAFT	92.20 ± 0.13	91.71 ± 0.16	91.25 ± 0.10
GT-NSGDm	91.40 ± 0.53	85.22 ± 1.65	79.01 ± 4.63
Local-D-SGD	60.31 ± 21.32	25.56 ± 10.86	28.27 ± 10.15
K-GT	61.72 ± 27.22	32.06 ± 12.82	25.64 ± 3.05
LSGT	61.54 ± 17.43	36.98 ± 6.14	23.83 ± 4.67

H.4 GPT-124M: CLEAN FINE-TUNING RESULTS

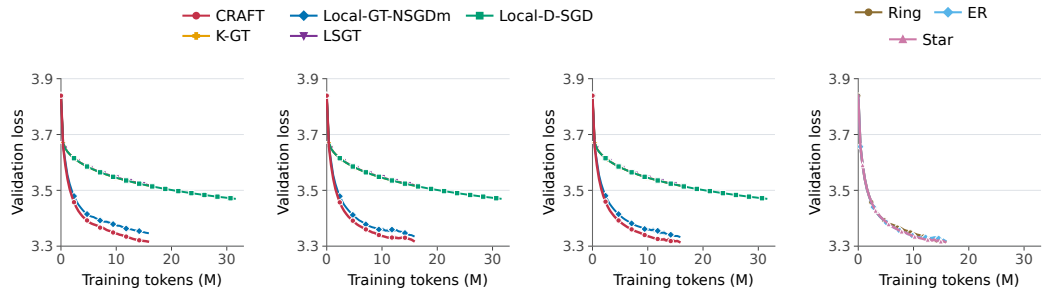
Figure 8 presents the GPT-124M runs from Figure 4 as a function of cumulative training tokens. The fixed-graph protocol, hyperparameters, and communication accounting are given in Appendix G.2.2.

Interpretation and limitations. At the matched endpoint communication budget within each graph, CRAFT achieves the lowest mean validation loss among the five compared methods while using half as many training tokens as Local-D-SGD. These results provide a finite-budget comparison on clean data among separately configured methods. They do not evaluate heavy-tail robustness, identify a causal effect of network topology, or imply superiority across other metrics, budgets, or hyperparameter choices.



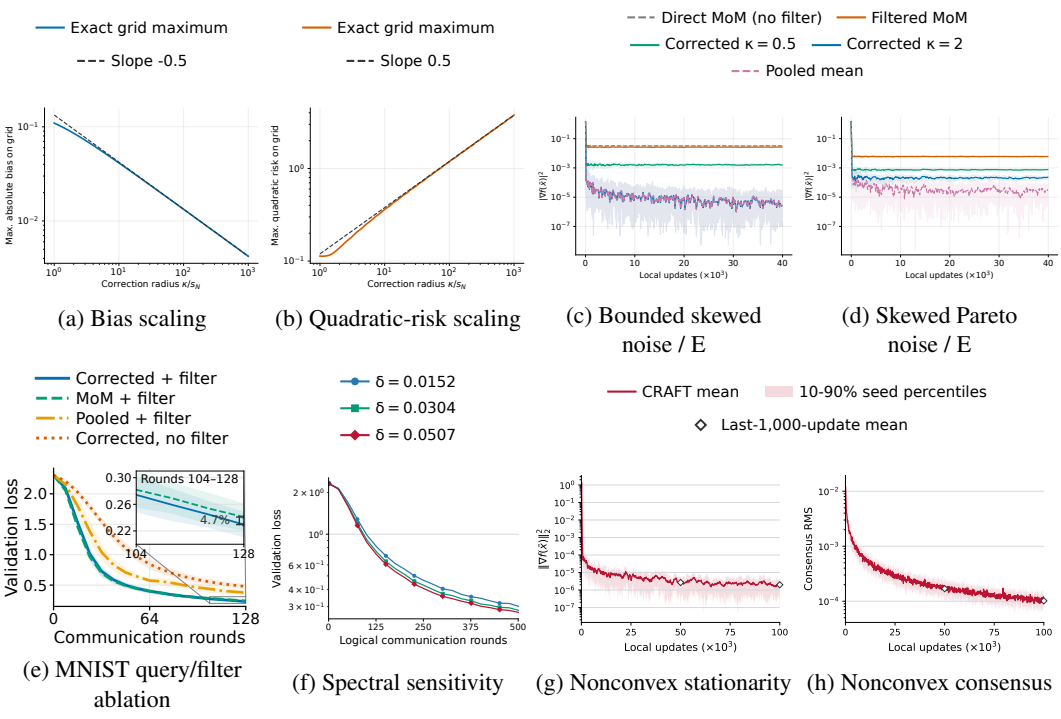
(a) $\rho = 0.50$ (b) $\rho = 0.75$ (c) $\rho = 1.00$ (d) CRAFT across noise

Figure 7: Sample-budget view of the CIFAR-10 runs in Figure 3. All panels use the same display range of 1,025,024 cumulative system-wide training samples. D-SGD continues beyond this range in the matched-communication comparison, so its terminal equal-communication accuracy is not the accuracy shown at the sample-budget boundary. Initialization samples are included in the logged counts.



(a) Ring (b) ER (c) Star (d) CRAFT across graphs

Figure 8: Training-token view of the clean GPT-124M runs in Figure 4. The horizontal axis reports cumulative raw training tokens across the four agents, excluding evaluation tokens. Local-D-SGD continues to approximately 31.46M tokens, whereas the tracking methods end near 15.73M. Curves are not extrapolated; equal-token comparisons are therefore restricted to their shared token range.



(a) Bias scaling (b) Quadratic-risk scaling (c) Bounded skewed noise / E (d) Skewed Pareto noise / E

(e) MNIST query/filter ablation (f) Spectral sensitivity (g) Nonconvex stationarity (h) Nonconvex consensus

Figure 9: Query-scaling and mechanism diagnostics. (a–b) Exact finite-grid maxima; dashed lines show reference slopes. (c–d) Medians with 10th–90th seed percentiles over 20 paired runs. (e) MNIST $H = 2$ query/filter ablation under Student-$t_{1.5}$ loss perturbations. Curves show mean validation loss with pointwise sample-SD bands over five paired seeds; the horizontal axis counts communication rounds. All four configurations use identical sample and directed-link communication budgets. The inset magnifies rounds 104–128 for filtered MoM and corrected MoM using the same unsmoothed statistics; the endpoint bracket indicates the 4.7% relative difference between their five-seed mean endpoint losses. (f) Selected CIFAR-10 spectral settings for seed 666. (g–h) Means with 10th–90th seed percentiles; diamonds indicate tail averages. Panels (c–d) and (g–h) use different synthetic objectives, whereas panel (e) uses MNIST. See Appendices G.1 and G.3 for the corresponding protocols.

H.5 Estimator scaling and asymmetric-noise diagnostics

Figure 9(a–b) evaluates the two-point family described in Appendix G.3. For $p = 1.5$, $\sigma = 1$, and $m = 8$, all 729 block-count triples are enumerated for each of 481 noise laws. Over the range $\kappa/s_N \geq 32$, log–log fits to the separate finite-grid maxima of absolute bias and MSE have slopes -0.498 and 0.504, respectively, close to the theoretical exponents $1 - p = -0.5$ and $2 - p = 0.5$. The distributions maximizing the two quantities need not coincide, and these finite-grid maxima should not be identified with the suprema over the full moment class.

Panels (c–d) instead fix either a bounded asymmetric noise law or a centered Pareto noise law and examine finite-time optimization behavior. Under Pareto noise, the median terminal squared gradient norm is 5.98×10^{-3} for filtered MoM, 7.62×10^{-4} for the corrected estimator with $\kappa = 0.5$, and 2.26×10^{-4} with $\kappa = 2$. The pooled-mean control attains a smaller median residual in these finite runs despite having infinite population MSE under the stated noise law. These observations concern finite-time behavior; in particular, a fixed correction radius can retain estimator bias.

Panels (g–h) report stationarity and consensus trajectories for a separate smooth nonconvex toy with symmetric noise. They are mechanism diagnostics and are not used to estimate the convergence exponent in the theory.

H.6 SKEWED-NOISE COMPONENT ABLATION

Table 5 reports endpoint and tail statistics for the supplementary skewed-noise component ablation, including two stepsize controls for the no-filter variant. The matched protocol is specified in Appendix G.3.

Table 5: Supplementary matched-budget component ablation under synthetic skewed heavy-tailed noise. Entries are mean ± sample SD over 20 formal seeds. The endpoint is communication round 50,000. For each seed, the tail statistic averages rounds 45,001–50,000 before aggregation across seeds. The final two rows vary only the stepsize of the no-filter control.

Configuration	Endpoint G_t	Tail mean G_t
Full CRAFT	0.002315 ± 0.000680	0.002524 ± 0.000291
No filtering	0.075969 ± 0.001726	0.076458 ± 0.000772
No correction	0.381656 ± 0.003058	0.382083 ± 0.001226
No filtering, $\eta_t/2$	0.076165 ± 0.001084	0.076464 ± 0.000743
No filtering, $2\eta_t$	0.075816 ± 0.002546	0.076433 ± 0.000782

Interpretation and relation to the theory. On this instance, full CRAFT has a lower endpoint squared gradient norm than each control in every paired formal seed. Together with the matched settings, this supports complementary empirical contributions from correction and recursive filtering in this experiment. It does not establish that either component is universally necessary, and direction clipping is not ablated here.

The schedule in Appendix G.3 is an empirical time-varying schedule rather than the constant-within-run tuning analyzed in Theorem 3; in particular, $\eta_t/\theta_t = 0.1$. Moreover, the reported quantity

$$G_t = \|\nabla f(\bar{x}_t)\|^2$$

is evaluated at communication boundaries, whereas the theory controls a random pre-update output through the clipped descent certificate

$$\phi_C(g) = \min\{\|g\|^2, C\|g\|\}.$$

These quantities coincide pointwise when $\|g\| \le C$, but the experimental protocol does not provide a direct test of the theorem's finite-time rate. The correction radius κ_t also increases over time, so the nonzero finite-budget residual should not be interpreted as a permanent fixed-κ bias floor.

The mean tail consensus RMS is 4.61×10^{-5} for full CRAFT, 7.17×10^{-4} without filtering, and 1.27×10^{-5} without correction. Thus the no-correction control attains smaller disagreement than full CRAFT while retaining a substantially larger squared gradient norm. This illustrates that consensus and stationarity are distinct criteria; the experiment does not establish superiority of one configuration on every metric. The symmetric-noise diagnostic uses a different objective and is reported separately.

H.7 HYPERPARAMETER SENSITIVITY

Table 6 reports a one-parameter-at-a-time sensitivity analysis on the original symmetric-noise nonconvex toy described in Appendix G.3. The first row is the baseline configuration. Within each subsequent block, only the shaded parameter is varied, while the remaining parameters are held at their baseline values.

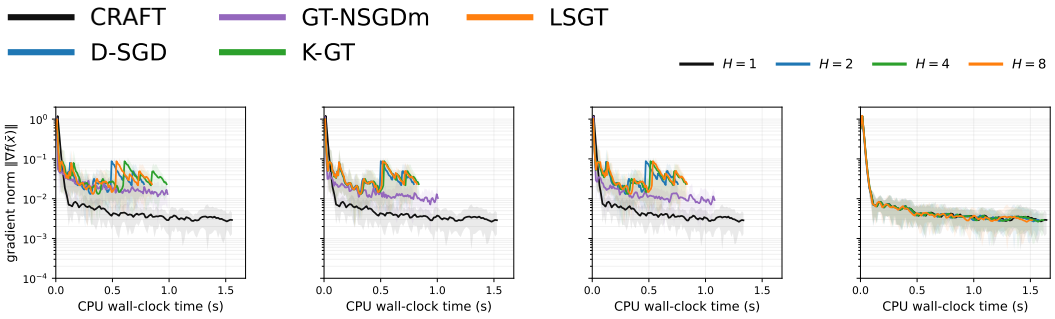
H.8 LOCAL-STEP WALL-CLOCK COMPARISONS

Figure 10 complements the update-based comparisons with measured CPU wall-clock time under the protocol described in Appendix G.3.

At the common endpoint of 10,000 local updates, CRAFT uses 1.33–1.56 mean CPU seconds and attains mean gradient norms of 2.86×10^{-3}–2.92×10^{-3} across the three displayed values of H. GT-NSGDm uses 0.99–1.08 seconds and attains 9.26×10^{-3}–1.31×10^{-2}, while the remaining

Table 6: Hyperparameter sensitivity on the original symmetric-noise nonconvex toy. The first row is the baseline configuration; shaded cells indicate the parameter varied from the baseline. For each seed, each metric is first averaged over the final 1,000 local updates. Entries report the mean $\pm$ sample SD of these averages over 20 seeds. Here $G := \|\nabla f(\bar{x})\|^2$ and $\mathcal{C} := \left(n^{-1}\sum_i \|x_i - \bar{x}\|^2\right)^{1/2}$.

η_0	θ	C	κ	G (10^{-6})	$\mathcal{C}$ (10^{-4})
0.03	0.20	1.00	0.50	1.994 ± 1.582	1.018 ± 0.054
0.01	0.20	1.00	0.50	0.842 ± 0.638	0.340 ± 0.018
0.02	0.20	1.00	0.50	1.449 ± 1.168	0.679 ± 0.036
0.04	0.20	1.00	0.50	2.584 ± 1.943	1.356 ± 0.072
0.05	0.20	1.00	0.50	3.228 ± 2.284	1.695 ± 0.090
0.03	0.10	1.00	0.50	1.995 ± 1.592	0.869 ± 0.054
0.03	0.15	1.00	0.50	1.993 ± 1.588	0.958 ± 0.054
0.03	0.25	1.00	0.50	1.997 ± 1.578	1.064 ± 0.053
0.03	0.30	1.00	0.50	2.000 ± 1.571	1.105 ± 0.053
0.03	0.20	0.20	0.50	1.891 ± 1.514	0.948 ± 0.048
0.03	0.20	0.40	0.50	1.989 ± 1.565	1.010 ± 0.053
0.03	0.20	0.60	0.50	1.997 ± 1.574	1.015 ± 0.053
0.03	0.20	0.80	0.50	1.996 ± 1.580	1.017 ± 0.053
0.03	0.20	1.00	0.10	1.587 ± 1.287	0.897 ± 0.046
0.03	0.20	1.00	0.30	1.718 ± 1.377	0.947 ± 0.049
0.03	0.20	1.00	0.70	2.302 ± 1.886	1.084 ± 0.058
0.03	0.20	1.00	0.90	2.551 ± 2.196	1.143 ± 0.061



(a) $H = 1$ (b) $H = 4$ (c) $H = 8$ (d) CRAFT: H sweep

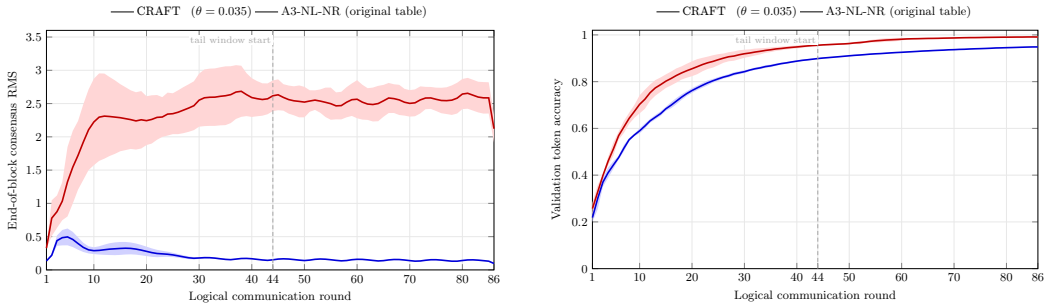
Figure 10: Gradient-norm wall-clock comparison on the smooth nonconvex toy. Panels (a)–(c) compare five methods at fixed $H = 1, 4, 8$, respectively, while panel (d) compares CRAFT across $H \in \{1, 2, 4, 8\}$. The left legend applies to panels (a)–(c), and the right legend to panel (d). The horizontal axis reports per-run CPU wall-clock time, and the vertical axis reports $\|\nabla f(\bar{x})\|$. Curves show mean checkpoint time against mean gradient norm over 20 paired seeds; bands show the empirical 10th–90th seed percentiles of the gradient norm. All panels use the same time range, each curve ends at its measured completion time, and no smoothing or time-axis interpolation is applied. Panel (d) uses a separate paired timing sweep under the same protocol. For D-SGD and GT-NSGDm, the $H > 1$ curves denote Local extensions rather than the original single-step recurrences. Only LSGT is shown for the duplicate D-SGT/LSGT state recurrence, using its original measured times. These results are single-machine CPU timing diagnostics, not distributed-network latency measurements.

reported methods have endpoint means of approximately 2.31×10^{-2}–2.34×10^{-2}. Thus, under this implementation and fixed-update budget, the methods exhibit different computation–accuracy trade-offs rather than a uniform ordering in both runtime and stationarity.

The separate CRAFT sweep in panel (d) gives endpoint gradient norms between 2.86×10^{-3} and 2.92×10^{-3} and mean completion times between 1.51 and 1.64 seconds for $H \in \{1, 2, 4, 8\}$. These measurements come from a single-machine implementation and may reflect Python overhead, CPU scheduling, and thermal variation. They should therefore be interpreted as implementation-level timing diagnostics rather than distributed-network latency measurements or general runtime comparisons.

H.9 CRAFT–NR CONSENSUS AND ACCURACY TRAJECTORIES

Figure 11 reports the paired temporal diagnostic described in Appendix G.4.1. These runs are separate from the cross-horizon experiments reported above.



(a) End-of-block consensus RMS

(b) Validation token accuracy

Figure 11: CRAFT–NR temporal comparison on the clean copy diagnostic with $H = 8$. Blue denotes the tuned CRAFT configuration and red the retained NR reference. Curves show means over five paired formal seeds; bands are pointwise 95% t-intervals for the seed mean with four degrees of freedom, rather than sample-SD bands. The original unsmoothed block-level trajectories are shown. The dashed line at communication round 44 marks the start of the tail-analysis window and does not indicate a change in the algorithm or schedule. Over the reported finite horizon, CRAFT exhibits lower consensus RMS, whereas NR attains higher token accuracy.

At the terminal evaluation, the mean consensus RMS is 0.09673 for CRAFT and 2.12192 for NR, while the corresponding validation token accuracies are 0.94851 and 0.99098. The terminal token cross-entropies are 0.46140 and 0.06528, respectively. Thus, in this experiment, CRAFT maintains substantially smaller model disagreement, whereas NR attains better task loss and accuracy. The comparison illustrates that consensus and predictive performance are distinct criteria: lower disagreement does not by itself imply lower task loss or higher accuracy.

Over communication rounds 44–86, the fitted tail slope is negative for each of the five CRAFT seeds, whereas the interval estimate for the mean NR tail slope includes zero. This provides a finite-time, setting-specific difference in the observed consensus trajectories. It does not establish pointwise monotonicity, asymptotic convergence, or a general limitation of NR, and it does not exclude different behavior under other NR hyperparameters or stepsize schedules.