What's in a Smoothness Constant? Tighter Rates for Local SGD with Bounded Second-order Heterogeneity

Kumar Kshitij Patel[1,*,†] **Rustem Islamov**[2,*] **Sebastian Stich**[3,†]
Aurelien Lucchi[2] **Eduard Gorbunov**[4] **Lingxiao Wang**[5]
[1]Yale University [2]University of Basel [3]CISPA Helmholtz Center
[4]MBZUAI [5]New Jersey Institute of Technology

Abstract

Local SGD, also known as Federated Averaging, is a widely used distributed optimization algorithm. Although Local SGD often outperforms alternatives such as Mini-batch SGD in practice, theory still only partially explains when and why local updates help under realistic data heterogeneity. Recent work by Patel et al. [1] shows that a bounded second-order heterogeneity assumption captures the efficiency of Local SGD for strongly convex objectives and conjectures that the same principle extends to the general convex setting. In this paper, we prove this conjecture by establishing an improved convergence guarantee for Local SGD on general convex objectives under bounded second-order heterogeneity. We also improve the best-known lower bounds for Local SGD in this setting, showing that our upper bounds are nearly tight. Together, these results provide a sharper, more fine-grained convergence theory for Local SGD. As a further application of our techniques, we provide a lower bound for serial SGD with replacement, showing how second-order heterogeneity captures the impact of rare high-curvature clients.

1 Introduction

In large-scale distributed training of machine learning algorithms, communication is often the primary bottleneck, outweighing computation in both runtime and energy costs [2, 3, 4]. This raises a key question: how much local computation can each client perform between communication rounds while preserving convergence guarantees? To formalize this trade-off, we study the following distributed optimization problem:

$$\min_{x\in\mathbb{R}^d}\left\{F(x) \coloneqq \frac{1}{M}\sum_{m=1}^{M} F_m(x)\right\}, \tag{1}$$

where $F_m(x) \coloneqq \mathbb{E}_{z_m\sim\mathcal{D}_m}[f(x;z_m)]$ is the population loss on machine m, defined using a loss function $f(\cdot;\, z \in \mathcal{Z})$ and a local data distribution $\mathcal{D}_m \in \Delta(\mathcal{Z})$. Optimization problems of the above form arise throughout machine learning, from multi-GPU training in data centers [4, 5] to decentralized training across millions of edge devices [3, 6]. A simple and widely studied model of communication for such problems is that of *intermittent communication* (IC) [7]. In the IC setting, machines communicate for R rounds, and within each round, each machine performs K units of computation. The pair (K, R) captures the trade-off between local computation and communication.

While numerous algorithms [8, 9, 10, 11, 12, 13, 14] have been proposed to solve Problem (1) in the IC setting, in practice, the dominant approaches remain first-order methods that compute K stochastic gradients per client between communication rounds. This naturally raises a central question: how should these K stochastic gradients be used? In particular, should each machine use them to perform *sequential local updates*, or instead *average* them to obtain a lower-variance estimate of the true

*Equal contribution. †Part of this work was supported by and carried out while KKP and SS were visiting the Simons Institute for the Theory of Computing.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

local gradient? A clean way to study this question is to contrast **Local SGD** (Federated Averaging) [15, 6, 3], which emphasizes local updates, with **Mini-batch SGD** [16], which emphasizes averaging. These canonical baselines isolate the two key design primitives underlying local computation.

Early analyses of Local SGD identified regimes in which it can outperform Mini-batch SGD, but these results rely on restrictive assumptions about data heterogeneity. In particular, many works either focus on the homogeneous setting, where all clients share the same data distribution [17, 18], or assume bounded *first-order heterogeneity*, which controls how much local gradients differ across clients. Concretely, these assumptions require that the gradients $\nabla F_m(x)$ remain uniformly close to the global gradient $\nabla F(x)$ for all x, effectively limiting the variability between client objectives. While analytically convenient, such conditions are quite stylized, as they rule out realistic settings in which data distributions can differ substantially across clients (see further discussion after Assumption 2).

In parallel, a separate line of work has shown that *second-order heterogeneity*, which instead controls the similarity of local curvature (e.g., Hessians) across clients, can provably reduce the communication complexity of related distributed methods [19, 20, 21, 22, 23, 12, 24, 25], without requiring restrictive bounds on gradient dissimilarity.[3] Building on this evidence, Patel et al. [26] conjectured that Local SGD should also benefit from bounded second-order heterogeneity. If true, this would identify a significantly broader and more realistic regime in which Local SGD can provably outperform Mini-batch SGD. While Patel et al. [1] recently confirmed this conjecture in the strongly convex setting, the general convex case has remained largely open; we close this gap in this paper.

1.1 Our Contributions

We establish the first convergence guarantees for Local SGD in the general convex setting that *explicitly capture the benefits of bounded second-order heterogeneity* (see Table 1, Theorems 1 and 2). We also prove a complementary upper bound under uniformly bounded first-order heterogeneity (Theorem 7; see Table 1). Our second-order heterogeneity results show that Local SGD can provably improve upon Mini-batch SGD without relying on potentially restrictive uniform first-order heterogeneity assumptions, thereby identifying a broader and more realistic regime in which local updates are fundamentally advantageous.

The novelty of our analysis is a trajectory-dependent control of heterogeneity. Prior analyses of Local SGD bound client drift using a uniform heterogeneity parameter that controls gradient disagreement globally across the parameter space. Instead, we show that the consensus error depends only on the gradient disagreement encountered along the trajectory of the average iterate. We then couple this trajectory-wise heterogeneity with the iterate error itself via a self-bounding recursion, thereby allowing the dynamics of Local SGD to control its own heterogeneity.

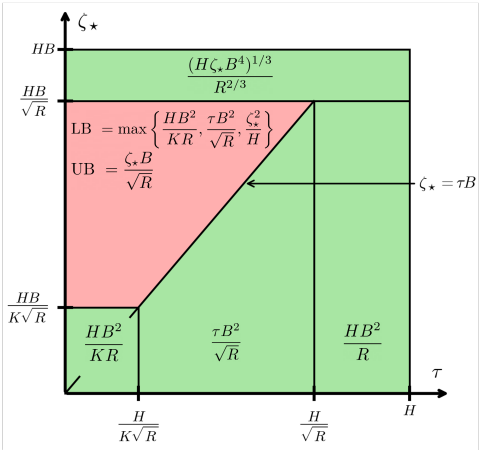

Figure 1: The tightness of our upper and lower bounds, highlighted in the setting with $\sigma = 0$ and $\zeta_\star \leq HB$. The green areas indicate regimes where the upper and lower bounds match, along with the dominant term in each, while the red area indicates the regime where the upper (UB) and lower (LB) bounds differ. See Appendix H for details.

We further provide new lower bounds (Theorems 3 and 10 and Corollaries 1 and 2) by developing sharper constructions that disentangle the distinct roles of first- and second-order heterogeneity [1]. These results reveal that our upper bound is nearly tight and isolate the precise regimes in which second-order effects yield gains (see Figure 1). Together, our results resolve the conjecture of Patel et al. [26] in the general convex setting, proving that second-order heterogeneity is not merely a technical artifact but a key structural property governing the efficiency of local update methods. A by-product of our lower-bound ideas is a refined lower-bound construction for multi-epoch SGD with replacement (Theorem 4), which shows that the standard dependence on worst-case client smoothness is tight even in heterogeneous settings, where worst-case smoothness is much larger than the average.

[3] Note that τ-second-order heterogeneity (Assumption 3) always holds with $\tau \leq 2H$ if local objectives are H-smooth.

1.2 Related work

The contrast between mini-batch and Local SGD has driven over a *decade of work* on establishing tight non-asymptotic convergence guarantees for Local SGD [15, 27, 28, 17, 29, 30, 31, 10, 18, 32, 33, 34, 35, 36, 37, 38, 39, 26, 1, 40, 41, 42, 43]. This effort has proved remarkably fruitful: it has motivated new primitives such as control variates [10, 44], revealed separations between the min-max complexities of serial and distributed optimization [7, 34], helped discover a novel form of acceleration [13], motivated algorithms for training large language models [4, 45] and more broadly supplied numerous analytical tools that extend to settings such as decentralized optimization [31], communication compression [46, 47, 48, 49, 50], quantization [51, 52], asynchronous updates [53, 54, 55], differential privacy [56, 57, 58, 59, 60, 61, 62], and Byzantine robustness [63, 64, 65, 66]. Our paper contributes to this broader discourse, providing both novel guarantees and analytical tools.

2 Setting and Preliminaries

Notation. For $i \le j \in \mathbb{Z}_{\ge 0}$, we use $[i, j]$ to denote $\{i, i+1, \dots, j-1, j\}$, and when $i = 1$, we denote it by $[j]$. We use $\lesssim$, $\gtrsim$, and $\cong$ for comparisons up to universal numerical constants. We also use $\delta(t) \coloneqq t - t \bmod K$ to denote the most recent communication round at or before time t.

2.1 Regularity Assumptions and the Algorithms

An instance of Problem (1) can be characterized by the client distributions $\{\mathcal{D}_m \in \Delta(\mathcal{Z})\}_{m\in[M]}$ and a differentiable[4] loss function $f(\cdot; z \in \mathcal{Z}) : \mathbb{R}^d \to \mathbb{R}$. We denote the set of all problem instances by $\mathcal{P}$. To further restrict the problems we study, we assume throughout that for all $m \in [M]$, the objective function F_m is differentiable, convex, and H-smooth, i.e., for all $x, y \in \mathbb{R}^d$,

$$0 \le F_m(y) - F_m(x) - \langle \nabla F_m(x), y - x\rangle \le \frac{H}{2}\|x - y\|_2^2. \tag{2}$$

Each machine $m \in [M]$ computes stochastic gradients of F_m by sampling data from its distribution $z \sim \mathcal{D}_m$ independent of other machines and time steps. These gradients satisfy for all $x \in \mathbb{R}^d$,

$$\mathbb{E}_{z\sim\mathcal{D}_m}\left[\nabla f(x; z)\right] = \nabla F_m(x) \quad \text{and} \quad \mathbb{E}_{z\sim\mathcal{D}_m}\left[\|\nabla f(x; z) - \nabla F_m(x)\|_2^2\right] \le \sigma^2. \tag{3}$$

Finally, we assume that the average objective F has at least one bounded minimizer, i.e.,

$$\exists\, x^\star \in \arg\min_{x\in\mathbb{R}^d} F(x) \quad \text{s.t.} \quad \|x^\star\|_2 \le B. \tag{4}$$

We will denote the class of all problems satisfying the above assumptions (2), (3), and (4) by $\mathcal{P}^{H,B,\sigma}$.

We next describe our algorithms, i.e., Local SGD and Mini-batch SGD, with intermittent communication (see Figure A.1 for an illustration). The key difference between these algorithms is in how each client utilizes its local stochastic gradients between communication rounds.

Local SGD. Local SGD runs K SGD steps independently on each machine, and then averages and synchronizes the resulting models across machines at each communication round. Specifically, at each time step $t \in [0, T-1]$, machine $m \in [M]$ samples $z_t^m \sim \mathcal{D}_m$ and updates as follows:

$$x_{t+1}^m \coloneqq \begin{cases} x_t^m - \eta\, \nabla f(x_t^m; z_t^m) & (t+1) \bmod K \ne 0, \\ \frac{1}{M}\sum_{n=1}^M \left[x_t^n - \eta\, \nabla f(x_t^n; z_t^n)\right] & (t+1) \bmod K = 0. \end{cases} \tag{5}$$

We assume all the machines are initialized at $x_0 = 0$. The above algorithm is often called "vanilla local SGD" to distinguish it from another popular variant of Local SGD with inner and outer step sizes [10]. Algorithm 1 presents this two-step-size version of Local SGD, which often performs better in practice [67], and has motivated an entire class of algorithms with an inner and outer optimization loop[5]. In Algorithm 1, $x_{r,k}^m$ is the kth local model on machine m, leading up to the rth round of communication, while x_r is the consensus (global) model at the end of the rth communication. For

[4] Our analysis can be extended to sub-differentiable loss functions using standard changes to our assumptions and analysis.

[5] For instance, see DILOCO for LLM optimization [4, 45] and the variance-reduced local optimization algorithms for distributed non-convex optimization [23, 12, 60].

Local SGD, η is referred to as the inner (local) step-size, while β is the outer (server) step-size. Note that setting $\beta = 1$ recovers *"vanilla Local SGD"* with a single step-size.

Mini-batch SGD. For Mini-batch SGD (see Algorithm 2), within each communication round, all machines compute stochastic gradients at the same shared iterate (the current global model), rather than evolving local models independently. The algorithm then aggregates these gradients into a single update to the global model. Because this algorithm uses an effective batch size of MK, it is often called "Large" Mini-batch SGD [68]. This synchronized evaluation removes client drift and ensures that each round directly optimizes the global objective F (see Woodworth et al. [32] for further discussion). Specifically, for optimization problems in the class $\mathcal{P}^{H,B,\sigma}$, mini-batch SGD achieves the following tight convergence rate, which cannot be improved under heterogeneity assumptions[6].

$$\mathbb{E}\left[F(\hat{x}^{\text{MB-SGD}})\right] - F(x^\star) \cong \frac{HB^2}{R} + \min\left\{HB^2, \frac{\sigma B}{\sqrt{MKR}}\right\} \quad . \tag{6}$$

The above rate also explains why Local SGD can outperform Mini-batch SGD in regimes with ample local computation and when σ is small: Local SGD uses the K gradient evaluations to move the model K times, rather than to form a single averaged update direction. In particular, even with $K \to \infty$, the optimization error of Mini-batch SGD does not improve. In practice, we observe that sequential local updates often use available computation more effectively than pure mini-batching [3, 69, 68, 4]. This gap motivates the central question in the theory of Local SGD: under what notions of data heterogeneity can Local SGD provably dominate Mini-batch SGD in a non-trivial regime?

2.2 A Hierarchy of Data Heterogeneity Assumptions

Homogeneous setting. When data heterogeneity is low, the Local SGD iterates across machines remain close between communication rounds, so Local SGD incurs only a small *consensus error* [17, 1]. In the extreme homogeneous case $\mathcal{D}_1 = \cdots = \mathcal{D}_M$, the client objectives coincide, and the local iterates on different machines evolve identically in expectation. This case, therefore, serves as a natural starting point, and much of the early work studied Local SGD under homogeneity.

We denote the homogeneous problem class by $\mathcal{P}^{H,B,\sigma}_{\text{hom}} \subseteq \mathcal{P}^{H,B,\sigma}$. As indicated by the bottom two rows of Table 1, the minimax-optimal rate of Local SGD over $\mathcal{P}^{H,B,\sigma}_{\text{hom}}$ is already known [18, 36]. However, homogeneity also hides the central difficulty in federated and distributed learning, as it ignores differences among clients. While Local SGD can outperform mini-batch SGD in this setting, this advantage arises only when even single-machine SGD also outperforms mini-batch SGD [34].

Approximate simultaneous realizability. A peculiarity of the homogeneous setting is that, since all clients share the same minimizer set, each client can, in principle, identify this set on its own in low-noise regimes. Thus, collaboration primarily helps by reducing variance. This perspective makes collaboration look rather narrow: communication primarily reduces noise rather than enabling qualitatively better optimization under heterogeneity. One could instead ask what happens in problems where clients are allowed to have different data distributions but share some minimizers, i.e., they are simultaneously realizable [70, 71]. The following first-order assumption formalizes this idea.

Assumption 1. *Objectives $\{F_m\}_{m\in[M]}$ satisfy $\zeta_\star$-first-order heterogeneity at the optima if there exists a minimizer of the average objective $x^\star \in \arg\min_{x\in\mathbb{R}^d} F(x)$ such that $\|x^\star\|_2 \le B$ and,*

$$\tfrac{1}{M}\textstyle\sum_{m\in[M]} \|\nabla F_m(x^\star)\|_2^2 \le \zeta_\star^2.$$

We denote the class of problems that satisfy the above assumptions by $\mathcal{P}^{H,B,\sigma}_{\zeta_\star}$. Tight convergence rates for Local SGD on this problem class are known from Koloskova et al. [31] and Patel et al. [26], as summarized in the first two rows of Table 1. Unfortunately, Local SGD cannot strictly dominate Mini-batch SGD for this problem class for any value of $\zeta_\star$. This indicates that stronger assumptions are needed to establish a provable advantage for Local SGD.

[6] A folklore argument showing that mini-batch SGD is agnostic to data heterogeneity, is to put the same data distribution on each machine and allude to lower bounds in serial stochastic optimization (see Patel et al. [26, Remark 10]). Having said that, it is important to note that when we drop the assumption that the smoothness parameter is identical across machines, the second-order heterogeneity assumption essentially captures the gap between the worst smoothness constant across clients and the smoothness of the average objective. We discuss this nuance further in Section 5 while analyzing SGD with replacement.

Paper (Result type)	Function sub-optimality: $\mathbb{E}\left[F(\hat{x}^{\text{Local-SGD}})\right] - F(x^\star) \lesssim$ or $\gtrsim$
Problem class $\mathcal{P}^{H,B,\sigma}_{\zeta_\star}$: *Assumption 1*	
Koloskova et al. [31] (Upper bound)	$\frac{HB^2}{R} + \frac{(H\sigma^2B^4)^{1/3}}{K^{1/3}R^{2/3}} + \frac{\sigma B}{\sqrt{MKR}} + \frac{(H\zeta_\star^2B^4)^{1/3}}{R^{2/3}}$
Patel et al. [26] (Lower Bound)	$\frac{HB^2}{R} + \frac{(H\sigma^2B^4)^{1/3}}{K^{1/3}R^{2/3}} + \frac{\sigma B}{\sqrt{MKR}} + \frac{(H\zeta_\star^2B^4)^{1/3}}{R^{2/3}}$
Problem class $\mathcal{P}^{H,B,\sigma}_{\zeta_\star,\tau}$: *Assumptions 1 and 3*	
Theorem 1 (Upper Bound)	$\frac{HB^2}{KR} + \frac{(H\tau^2)^{1/3}B^2}{R^{2/3}} + \frac{(H\sigma^2B^4)^{1/3}}{K^{1/3}R^{2/3}} + \frac{\sigma B}{\sqrt{MKR}} + \frac{(H\zeta_\star^2B^4)^{1/3}}{R^{2/3}}$
Theorem 2 (Upper Bound)	$\frac{HB^2}{KR} + \frac{\tau B^2}{\sqrt{R}} + \frac{\sigma B}{\sqrt{KR}} + \frac{\zeta_\star B}{\sqrt{R}}$
Theorem 3† (Lower Bound)	$\frac{HB^2}{KR} + \min\left\{\frac{HB^2}{R}, \frac{\tau B^2}{\sqrt{R}}\right\} + \min\left\{\frac{\sigma B}{\sqrt{KR}}, \frac{(H\sigma^2B^4)^{1/3}}{K^{1/3}R^{2/3}}\right\} + \frac{\sigma B}{\sqrt{MKR}} + \min\left\{\frac{\zeta_\star^2}{H}, \frac{(H\zeta_\star^2B^4)^{1/3}}{R^{2/3}}\right\}$
Patel et al. [1] (Lower Bound)	$\frac{HB^2}{KR} + \frac{\tau B^2}{R} + \min\left\{\frac{\sigma B}{\sqrt{KR}}, \frac{(H\sigma^2B^4)^{1/3}}{K^{1/3}R^{2/3}}\right\} + \frac{\sigma B}{\sqrt{MKR}} + \frac{(\tau\zeta_\star^2B^4)^{1/3}}{R^{2/3}}$
Problem class $\mathcal{P}^{H,B,\sigma}_{\zeta}$: *Assumption 2*	
Woodworth et al. [32], Luo et al. [42] (Upper bound)	$\frac{HB^2}{KR} + \frac{(H\sigma^2B^4)^{1/3}}{K^{1/3}R^{2/3}} + \frac{\sigma B}{\sqrt{MKR}} + \frac{(H\zeta^2B^4)^{1/3}}{R^{2/3}}$
Theorem 7 (Upper Bound)	$\frac{HB^2}{KR} + \frac{\sigma B}{\sqrt{KR}} + \frac{\zeta B}{\sqrt{R}}$
Corollaries 1 and 2† (Lower Bound)	$\frac{HB^2}{KR} + \min\left\{\frac{\sigma B}{\sqrt{KR}}, \frac{(H\sigma^2B^4)^{1/3}}{K^{1/3}R^{2/3}}\right\} + \frac{\sigma B}{\sqrt{MKR}} + \min\left\{\frac{\zeta^2}{H}, \frac{(H\zeta^2B^4)^{1/3}}{R^{2/3}}\right\}$
Problem class $\mathcal{P}^{H,B,\sigma}_{hom}$: *Homogeneous setting*	
Woodworth et al. [18]‡ (Upper bound)	$\frac{HB^2}{KR} + \min\left\{\frac{\sigma B}{\sqrt{KR}}, \frac{(H\sigma^2B^4)^{1/3}}{K^{1/3}R^{2/3}}\right\} + \frac{\sigma B}{\sqrt{MKR}}$
Glasgow et al. [36] (Lower Bound)	$\frac{HB^2}{KR} + \min\left\{\frac{\sigma B}{\sqrt{KR}}, \frac{(H\sigma^2B^4)^{1/3}}{K^{1/3}R^{2/3}}\right\} + \frac{\sigma B}{\sqrt{MKR}}$

Table 1: Convergence guarantees of Local SGD for general convex functions across different data heterogeneity assumptions. **From top to bottom, we go from more heterogeneous problem classes to less heterogeneous ones**. All lower-bound expressions $\mathcal{L}$ are understood as $\min\{HB^2, \mathcal{L}\}$, accounting for the trivial guarantee $F(x_0) - F(x^\star) \le HB^2/2$ obtained by returning the initialization. †The lower bounds assume $K, M \ge 2$. All the terms of the lower bound hold for the averaged as well as the final local SGD iterate, except for the red terms which we only show for the last iterate. ‡This rate is the minimum of the two convergence rates reported by Woodworth et al. [18].

Uniformly bounded first-order heterogeneity. Much of the early analysis of Local SGD considers a stronger variant of $\mathcal{P}^{H,B,\sigma}_{\zeta_\star}$ by requiring the local gradients to track the global gradient *uniformly over the domain* [30, 32, 10]. This yields the following more stringent notion of first-order heterogeneity.

Assumption 2. *Objectives $\{F_m\}_{m\in[M]}$ satisfy ζ-first-order heterogeneity if for all $x \in \mathbb{R}^d$,*

$$\frac{1}{M}\sum_{m\in[M]}\left\|\nabla F_m(x) - \nabla F(x)\right\|_2^2 \le \zeta^2.$$

We denote[7] the corresponding problem class by $\mathcal{P}_{\zeta}^{H,B,\sigma}$. Under Assumption 2, Woodworth et al. [32], Luo et al. [42] established regimes where vanilla Local SGD improves over mini-batch SGD (see Table 1). Furthermore, the limit $\zeta \to 0$ smoothly recovers the homogeneous setting, making this interpolation between heterogeneity and homogeneity particularly appealing. At the same time, $\mathcal{P}_{\zeta}^{H,B,\sigma}$ is a very restricted problem class. For instance, in the case of quadratic objectives, a uniform bound as in Assumption 2 forces all clients to share the same Hessian (see, e.g., Patel et al. [26, Remarks 2, 3]), which makes the problem essentially homogeneous.

This motivates a natural question: is there an intermediate heterogeneity model that is weaker than Assumption 2 but stronger than Assumption 1, under which Local SGD provably outperforms mini-batch SGD while still allowing meaningful heterogeneity? Since Assumption 2 controls global geometric heterogeneity across clients, the intermediate heterogeneity class must also impose such global control without requiring gradients to remain uniformly close across clients.

Controlling global geometry using bounded second-order heterogeneity. Patel et al. [26] conjectured that the appropriate assumption to capture a nontrivial regime in which Local SGD can outperform mini-batch SGD is *bounded second-order heterogeneity* (see Propositions 1 and 2)

Assumption 3. *Objectives $\{F_m\}_{m\in[M]}$ satisfy τ-second-order heterogeneity if, for all $x, y \in \mathbb{R}^d$,*

$$\tfrac{1}{M}\sum_{m\in[M]}\left\|\nabla F_m(x) - \nabla F_m(y) - (\nabla F(x) - \nabla F(y))\right\|_2^2 \le \tau^2 \left\|x - y\right\|_2^2 .$$

We denote by $\mathcal{P}_{\zeta_\star,\tau}^{H,B,\sigma}$ the class of problems satisfying Assumptions 1 and 3.

The conjecture of Patel et al. [26] states that, over the problem class $\mathcal{P}_{\zeta_\star,\tau}^{H,B,\sigma}$, Local SGD should provably outperform mini-batch SGD while still allowing meaningful heterogeneity. Their conjecture is supported by three sources of evidence: (i) their own analysis of Local SGD under second-order heterogeneity together with the restrictive Assumption 2; (ii) near-optimal analyses of related local-update methods in nonconvex settings [12]; and (iii) an independent line of work on distributed proximal methods (for example, DANE and its variants [19, 20, 21, 24, 25]) whose communication complexity improves under the same Hessian-similarity condition.

Recently, Patel et al. [1] verified this conjecture in the strongly convex subclass of $\mathcal{P}_{\zeta_\star,\tau}^{H,B,\sigma}$, but their result does not extend to general convex objectives via a black-box convex-to-strongly-convex reduction as their guarantees impose restrictions on the strong convexity parameter. At the same time, they established a lower bound for the general convex class $\mathcal{P}_{\zeta_\star,\tau}^{H,B,\sigma}$ that improves with τ and $\zeta_\star$ and interpolates to the homogeneous setting as these parameters vanish. In the next section, we show that our new upper bound for Local SGD on $\mathcal{P}_{\zeta_\star,\tau}^{H,B,\sigma}$ matches this lower bound (up to constants) in the regime of small τ, while retaining the desirable interpolation properties that previous upper bounds achieved only on the much smaller problem class $\mathcal{P}_{\zeta}^{H,B,\sigma}$ (see Table 1).

3 Improved Analyses of Local SGD for Classes $\mathcal{P}_{\zeta_\star,\tau}^{H,B,\sigma}$

In this section, we will prove three upper bounds for Local SGD. The first analyzes perturbations against dense mini-batch SGD [18], i.e., KR steps of mini-batch SGD (as opposed to R steps). The second analyzes perturbations to KR steps of SGD on a single machine under the additional assumption of uniform second-order heterogeneity. The third gives a complementary single-machine-style analysis under uniformly bounded first-order heterogeneity. We will denote by $\bar{x}_t := \frac{1}{M}\sum_{m=1}^{M} x_t^m$ the average iterate for all time steps $t \in [0, T]$. This "ghost sequence" is useful for analysis even at the time steps where the average is not actually computed.

3.1 Perturbing Against Dense Mini-batch SGD

Step 1: A canonical recursion that isolates consensus error. We begin with the standard smoothness-based one-step recursion for the average iterate, which appears in several analyses of Local SGD [17, 18, 32]. This inequality mirrors the descent lemma for dense mini-batch SGD,

[7] With a slight abuse of notation, we use $\zeta_\star$ and ζ as both scalar values and to demarcate the problem classes implied by Assumptions 1 and 2, respectively. The usage should be clear from the context.

with an additional term capturing the deviation of local iterates from their average. As a result, it decomposes the error into three components: optimization error, stochastic noise, and *consensus error*, which measures the client drift due to the lack of complete synchrony.

Lemma 1 (Average iterate recursion with consensus error)**.** *Consider the problems in $\mathcal{P}^{H,B,\sigma}$. If we run vanilla Local SGD with constant step-size $\eta \le 1/(10H)$, then for every $t \in [0, T-1]$,*

$$\mathbb{E}[F(\bar{x}_t) - F^\star] \le \frac{\mathbb{E}[\|\bar{x}_t - x^\star\|_2^2] - \mathbb{E}[\|\bar{x}_{t+1} - x^\star\|_2^2]}{\eta} + \frac{3\eta\sigma^2}{M} + \frac{2H}{M}\sum_{m=1}^M \mathbb{E}[\|\bar{x}_t - x_t^m\|_2^2]. \tag{7}$$

The consensus error $C(t) \coloneqq \frac{1}{M}\sum_{m=1}^M \mathbb{E}[\|\bar{x}_t - x_t^m\|_2^2]$ is the main difficulty in heterogeneous settings. In prior works [32, 42], the consensus error is directly bound using the uniform heterogeneity Assumption 2. Our goal is to replace this ζ-based control with a bound relying on $(\zeta_\star, \tau)$. A key observation is that the same recursion also controls the expected iterate error of the average trajectory. Specifically, denote $A_{\max} \coloneqq \max_{t\in[0,T]} \mathbb{E}[\|\bar{x}_t - x^\star\|_2^2]$. By unrolling recursion (7), one can show

$$A_{\max} \le B^2 + \frac{3\eta^2\sigma^2 T}{M} + 2H\eta T C_{\max}, \qquad \text{where} \qquad C_{\max} \coloneqq \max_{s\in[0,T]} C(s). \tag{8}$$

Thus, controlling the consensus error directly yields control over the expected iterate error of the average trajectory. Further details of this derivation are provided in Lemma 9 in Appendix D.

Step 2: Controlling consensus error via trajectory-wise gradient dissimilarity. Our next step is to relate the consensus error to the gradient heterogeneity *along the trajectory* of the average iterates, rather than uniformly over the entire domain. In particular, denote $V_{\max} \coloneqq \max_{s\in[0,T]} \frac{1}{M}\sum_{m\in[M]} \mathbb{E}[\|\nabla F_m(\bar{x}_s) - \nabla F(\bar{x}_s)\|_2^2]$. Then in Lemma 7 in Appendix D we show that

$$C_{\max} \lesssim \eta^2 K^2 V_{\max} + \eta^2 K \sigma^2. \tag{9}$$

This bound shows that the consensus error is governed by the *largest gradient dissimilarity encountered along the average trajectory*. In contrast to prior analyses (e.g., Woodworth et al. [32]), which directly upper-bound $V_{\max}$ using the uniform Assumption 2, we defer control of $V_{\max}$ to the next step, which uses a trajectory-dependent control, thus avoiding the dependence on ζ.

Step 3: Controlling $V_{\max}$ using the iterate error. We bound $V_{\max}$ using Assumptions 1 and 3 (see Lemma 8 in Appendix D):

$$V_{\max} \le 2\tau^2 A_{\max} + 2\zeta_\star^2. \tag{10}$$

Step 4: Self-bounding loop and the main theorem. We now close the argument by combining the bounds from the previous steps. Equation (8) shows that $A_{\max}$ is controlled by the largest consensus error $C_{\max}$. On the other hand, combining equations (9) with (10) shows that the largest consensus error itself can be bounded in terms of $A_{\max}$. Together, these relations form a self-bounding loop:

$$A_{\max} \longrightarrow C_{\max} \longrightarrow V_{\max} \longrightarrow A_{\max}.$$

Under a sufficiently small step-size η, this recursion can be closed, allowing us to simultaneously bound $A_{\max}$ and $C_{\max}$ purely in terms of $(\zeta_\star, \tau)$ (see Lemmas 9 and 10 in Appendix D for details).

This self-bounding argument is the key technical ingredient of the proof: it replaces the need for Assumption 2 by controlling the entire trajectory of Local SGD through its own dynamics. Substituting the resulting consensus bound into Lemma 1 yields our new upper bound (see Appendix E).

Theorem 1 (Informal version of Theorem 5)**.** *For problems in the class $\mathcal{P}^{H,B,\sigma}_{\zeta_\star,\tau}$, using an appropriate step-size, the Local SGD output $\hat{x} = \frac{1}{T}\sum_{t=0}^{T-1} \bar{x}_t$ satisfies,*

$$\mathbb{E}\left[F(\hat{x}) - F(x^\star)\right] \lesssim \frac{HB^2}{KR} + \frac{(H\tau^2)^{1/3}B^2}{R^{2/3}} + \frac{\sigma B}{\sqrt{MKR}} + \frac{(H\zeta_\star^2 B^4)^{1/3}}{R^{2/3}} + \frac{(H\sigma^2 B^4)^{1/3}}{K^{1/3}R^{2/3}}.$$

Comparing the above theorem to that of Woodworth et al. [32] for the class $\mathcal{P}^{H,B,\sigma}_{\zeta_\star}$, we note that replacing ζ by $\zeta_\star + \tau B$ recovers our theorem (up to numerical constants). This is not a mere coincidence, and in fact, for the step-size used in Theorem 1 we show that $A_{\max} \lesssim B^2$ (see Lemmas 9 and 10), which allows us to bound the gradient heterogeneity on the trajectory of Local SGD, i.e., $V_{\max} \lesssim \tau^2 B^2 + \zeta_\star^2$. Notably, this connects our proof to the insight of Patel et al. [26, Proposition 13] and the guarantees of Patel et al. [1] which rely on a similar idea, albeit with strong convexity.

Remark 1 (Beating Mini-batch SGD)**.** Comparing our bound in Theorem 1 with the mini-batch SGD rate in (6), we see that Local SGD improves over mini-batch SGD whenever

$$K \gg 1, \qquad \tau \lesssim \frac{H}{\sqrt{R}}, \qquad \zeta_\star \lesssim \frac{HB}{\sqrt{R}}, \qquad \sigma \lesssim HB\sqrt{\frac{K}{R}}.$$

Compared with the analysis under Assumption 2 by Woodworth et al. [32], this is a significantly larger regime, highlighting the role of second-order heterogeneity (cf. Patel et al. [26, Theorem 15]). Note that this result does not guarantee that the improvement holds for arbitrary R: the region where Local SGD outperforms Mini-batch SGD shrinks with R for fixed problem constants.

3.2 Perturbing Against Single-Machine SGD

We now proceed to prove our second new upper bound under the second-order heterogeneity Assumption 3. Our starting point is the observation that, even in the homogeneous setting, the optimal upper bound for Local SGD (see Table 1) due to Woodworth et al. [18] requires two analyses, and specifically one of them shows that Local SGD is never worse than single-machine SGD. In the following theorem, we extend this analysis from the class $\mathcal{P}^{H,B,\sigma}_{hom}$ to $\mathcal{P}^{H,B,\sigma}_{\zeta_\star,\tau}$.

Theorem 2 (Informal version of Theorem 6)**.** *For problems in the class $\mathcal{P}^{H,B,\sigma}_{\zeta_\star,\tau}$, using an appropriate step-size, the Local SGD output $\hat{x} = \frac{1}{T}\sum_{t=0}^{T-1} \bar{x}_t$ satisfies,*

$$\mathbb{E}\left[F(\hat{x}) - F(x^\star)\right] \lesssim \frac{HB^2}{KR} + \frac{\sigma B}{\sqrt{KR}} + \frac{\tau B^2}{\sqrt{R}} + \frac{\zeta_\star B}{\sqrt{R}}.$$

Note that when τ and $\zeta_\star$ are zero, the above theorem recovers the convergence rate of single-machine SGD, highlighting that the last two terms indeed capture the cost of having different data distributions across machines. The proof of the above theorem relies on a fundamentally different one-step recursion (see Lemma 13 in Appendix D), but also uses a self-bounding argument that couples consensus error with the following alternative iterate error (Lemmas 12 and 14 in Appendix D),

$$B_{\max} := \max_{0\le s\le T} \frac{1}{M}\sum_{m\in[M]} \mathbb{E}\left[\|x_s^m - x^\star\|_2^2\right] \ge A_{\max}.$$

We defer the proof to the appendix, as it is qualitatively similar to the proof of Theorem 1. While Theorem 2 is crucial to characterize the tight rates for local SGD (see Section 4), it does not characterize a wider regime than Theorem 1 where local SGD beats mini-batch SGD. In Appendix F, we additionally provide an improved bound under Assumption 2.

4 A New Lower Bound and the Tightness of Our Results

In this section, we prove a new lower bound for Local SGD, which builds on existing lower bounds by Glasgow et al. [36] and Patel et al. [1]. The lower bound is crucial for decoupling the effects of first- and second-order heterogeneity and relies on two orthogonal mechanisms of client drift.

Theorem 3 (Informal version of Theorems 8 to 10)**.** *If $K, M \ge 2$, then there exists a problem in the class $\mathcal{P}^{H,B,\sigma}_{\zeta_\star,\tau}$, such that for any step-size η, $\hat{x} = \frac{1}{T}\sum_{t=0}^{T-1} \bar{x}_t$ must have,*

$$\mathbb{E}\left[F(\hat{x})\right] - F(x^\star) \gtrsim \min\left\{\frac{HB^2}{R}, \frac{\tau B^2}{\sqrt{R}}\right\} + \min\left\{HB^2, \frac{\zeta_\star^2}{H}, \frac{(H\zeta_\star^2 B^4)^{1/3}}{R^{2/3}}\right\}.$$

To obtain the final lower bound in Table 1, we combine our construction with the homogeneous lower bound of Glasgow et al. [36] by placing their hard instance on disjoint coordinates. Since this instance has $\zeta_\star = \tau = 0$, it satisfies all heterogeneity assumptions and does not restrict our parameters. We will also use two distinct hard instances to obtain the two terms in the lower bound. Note that we lose only numerical constants by using extra coordinates for different hard instances. We next discuss the key ideas in proving both components of the lower bound.

Second-order lower bound. Our construction (see Theorem 8) builds on Patel et al. [1, Theorem 1]: two clients share an optimizer but recover it only on their respective subspaces in isolation, while together they recover it globally. By tuning the angle between these subspaces, we induce drift across rounds so that the joint dynamics reduce to R steps of gradient descent on the average objective with an effective step size depending on K and η, yielding a lower bound that does not improve with K.

The key improvement in our analysis is that we decouple the "scale" of the client objectives from the size of their second-order heterogeneity, which allows us to obtain a sharper tuning of our hard instance, leading to the improved lower bound versus the $\frac{\tau B^2}{R}$ lower bound of Patel et al. [1].

First-order lower bound. This construction (see Theorem 9) uses three coordinates to separate optimization and heterogeneity effects. On the active coordinate, we define a convex H-smooth function with a curvature change at the origin,

$$g(s) = \mathbf{1}_{s<0} \cdot \tfrac{H}{4} s^2 + \mathbf{1}_{s\geq 0} \cdot \tfrac{H}{2} s^2,$$

and add opposite linear shifts $\pm\zeta_\star s$ across clients. This yields $\zeta_\star$-heterogeneity at the optimum while keeping $\tau = 0$. Local updates then place clients in different curvature regimes, inducing a persistent bias of order $\zeta_\star/H$ per round, which leads to the $R^{-2/3}$ rate after balancing step-size effects. Crucially, our construction uses a piecewise-quadratic function g whose curvature changes abruptly at the origin. It is an open question whether this failure of twice differentiability is necessary for the lower bound. The other two coordinates control extreme step sizes and prevent trivial convergence.

Using the ideas above in Theorems 8 and 9, we first prove the lower bound for the final iterate of Local SGD $\bar{x}_{KR}$, and then extend it to the averaged iterate $\hat{x}$ in Theorem 10. The same applies to the other terms in the lower bound mentioned in Table 1, except for the red term, which we currently prove only for the final Local SGD iterate.

Min-max optimal analysis of Local SGD. We can now compare our lower bound in Theorem 3 along with the homogeneous lower bound of Glasgow et al. [36] to the best of three upper bounds: (i) from Koloskova et al. [31] (see Table 1); (ii) our consensus-error based upper bound in Theorem 1; and (iii) our single-machine mimicking upper bound in Theorem 2. For simplicity, we look at the problem class $\mathcal{P}^{H,B,0}_{\zeta_\star,\tau}$ in the noiseless setting. We summarize the conclusion of this exercise (see calculations in Appendix H) in Figure 1 and note that, in most regimes, we characterize the min-max optimal convergence guarantee for Local SGD.[8] We believe new ideas are needed to close the gap in the red regime of Figure 1.

Comparison to Existing Upper and Lower Bounds. In Figure I.1, we provide illustrative phase diagrams of both the lower bound of Patel et al. [1] and the upper bound of Koloskova et al. [31]. We observe that there is no regime where the rates match, demonstrating the looseness of prior analyses for Local SGD over the class $\mathcal{P}^{H,B,\sigma}_{\zeta_\star,\tau}$. We derive both of these phase diagrams in Appendix I.

Although Theorem 3 is stated for the class $\mathcal{P}^{H,B,\sigma}_{\zeta_\star,\tau}$, it also implies a lower bound for the more restrictive uniformly bounded first-order heterogeneity class $\mathcal{P}^{H,B,\sigma}_{\zeta}$ (see Corollaries 1 and 2). The key observation is the inclusion $\mathcal{P}^{H,B,\sigma}_{\zeta_\star=\zeta,\tau=0} \subseteq \mathcal{P}^{H,B,\sigma}_{\zeta}$, which follows from Proposition 2.

5 Multi-pass SGD with Replacement under Second-Order Heterogeneity

The preceding sections focused on Local SGD, where second-order heterogeneity controls the drift created by multiple local updates between communication rounds. We now ask whether the same second-order viewpoint is useful even in an algorithmic setting with no local-updates. To this end, we study serial fresh-sample SGD with replacement. This algorithm does not create local client trajectories, but it still samples gradients from heterogeneous clients, and so second-order heterogeneity can affect how the client smoothness parameters enter the convergence rate.

To capture this, we track client-specific smoothness constants separately from the smoothness of the average objective. Each client objective F_m is H_m-smooth, and we define $\widehat{H} := \max_{m\in[M]} H_m$, which remains a uniform smoothness bound across clients. We also assume that the average objective F is H-smooth. Note that $H \leq \hat{H}$, but the ratio $\hat{H}/H$ can be as large as M. All other assumptions remain unchanged, including convexity, bounded optimum, and Assumptions 1 and 3. We denote this problem class by $\mathcal{P}^{H,\hat{H},B,\sigma}_{\zeta_\star,\tau}$.

Specifically, we consider the multi-pass SGD with replacement algorithm for $N := TM$ iterations, where T denotes the number of "effective epochs". At each iteration $t \in [0, N-1]$, the algorithm

[8] We compare previous upper bound of Koloskova et al. [31] and the lower bound of Patel et al. [1] in Figure I.1 and show that no regime exists in the $\zeta_\star$–τ plane where they match, highlighting the necessity of a more refined analysis.

samples $m_t \sim \mathrm{Unif}([M])$, draws $z_t \sim \mathcal{D}_{m_t}$, and performs $x_{t+1} = x_t - \eta \nabla f(x_t; z_t)$, with initialization $x_0 = 0$. The output is the average iterate $\bar{x}_N := \frac{1}{N}\sum_{t=0}^{N-1} x_t$. The sampled gradient is unbiased for $\nabla F(x_t)$. Moreover, by Assumption 1 and the stochastic-gradient assumption in (3), we have $\mathbb{E}\left[\|\nabla f(x^\star; z_t)\|_2^2\right] \le \sigma^2 + \zeta_\star^2$. Thus, standard SGD analyses for convex smooth functions give,

$$\mathbb{E}\big[F(\bar{x}_N) - F(x^\star)\big] \lesssim \frac{\hat{H}B^2}{TM} + \frac{(\sigma+\zeta_\star)B}{\sqrt{TM}} \quad . \tag{11}$$

The dependence on $\hat{H}$ in (11) is already necessary on homogeneous instances: all clients coincide, $\tau = 0$, and $H = \hat{H}$, so the standard lower bound $\Omega(\hat{H}B^2/(TM))$ applies. However, this argument does not establish the same dependence when $H \ll \hat{H}$. A homogeneous hard instance with average smoothness H yields only an $\Omega(HB^2/(TM))$ lower bound, which can be much smaller than the $\hat{H}$-dependent upper bound.

By Assumption 3 and the triangle inequality, every client objective is $(H + \sqrt{M}\tau)$-smooth. Applying the standard SGD guarantee with this smoothness bound gives the optimization term

$$\frac{(H+\sqrt{M}\tau)B^2}{TM} = \frac{HB^2}{TM} + \frac{\tau B^2}{T\sqrt{M}} \le \frac{HB^2}{TM} + \frac{\tau B^2}{\sqrt{TM}}, \qquad T \ge 1.$$

Thus, this H- and τ-dependent rate already follows from the standard SGD analysis. We next ask whether the full $\hat{H}$-dependent optimization term remains necessary on genuinely heterogeneous instances with $H \ll \hat{H}$. The following lower bound shows that the $\hat{H}B^2/(TM)$ optimization term remains tight even in this heterogeneous regime. The key point is that a large $\hat{H}$ can arise from rare high-curvature directions, with $\hat{H} \cong \sqrt{M}\,\tau$, rather than from large curvature across all clients.

Theorem 4 (Informal version of Theorem 11)**.** *Fix $M, T \ge 2$, $B > 0$, and parameters satisfying $\tau \cong H \cong \frac{\hat{H}}{\sqrt{M}}$. Then there exists an instance in $\mathcal{P}_{\zeta_\star,\tau}^{H,\hat{H},B,\sigma}$ such that for every constant stepsize $\eta > 0$, fresh-sample SGD with replacement satisfies*

$$\mathbb{E}\big[F(\bar{x}_{TM}) - F(x^\star)\big] \gtrsim \frac{\hat{H}B^2}{TM} + \min\left\{HB^2, \frac{(\sigma+\zeta_\star)B}{\sqrt{TM}}\right\} \quad .$$

The construction, given formally in Appendix J, follows the same guiding principle as our second-order-heterogeneity lower bound for Local SGD: it uses a rare-curvature direction carried by heterogeneous clients. In particular, the construction has clients whose curvature is much larger than that of the average objective, yielding the scaling $\tau \cong H \cong \hat{H}/\sqrt{M}$.

The key difference from Local SGD is that, for serial fresh-sample SGD, this rare curvature does not create a new local-drift effect. Instead, it realizes the usual worst-client smoothness barrier. An orthogonal stochastic and first-order-heterogeneity block gives the usual noise lower bound.

The lower bound should therefore be interpreted as follows. The standard $\hat{H}B^2/(TM)$ barrier is not merely an artifact of the homogeneous $\hat{H}$-smooth hard instance. Even in a genuinely heterogeneous regime where $H \ll \hat{H}$, serial fresh-sample SGD can still suffer the usual $\hat{H}$-dependent rate. In this sense, for serial SGD with replacement, second-order heterogeneity explains the source of worst-client smoothness: $\hat{H} \cong \sqrt{M}\,\tau$. This contrasts with the Local SGD results in the main part of the paper, where bounded second-order heterogeneity leads to genuinely new local-drift and communication effects that cannot be reduced to the standard serial $\hat{H}$-smooth analysis.

Limitations. Our upper and lower bounds match in parameter dependence in the green noiseless regimes of Figure 1, but a gap remains in the red regime. Our lower bounds for the time-averaged output cover the noiseless terms; extending the stochastic client-drift lower bound to this output remains open. Our first-order lower-bound construction uses a function that is not twice differentiable, and we do not know whether this feature is necessary. It also remains open whether separate local and server step sizes can yield a single guarantee that recovers the best of our three Local SGD upper bounds. Finally, our multi-pass results concern fresh-sample SGD with replacement. Further work could identify alternative second-order heterogeneity assumptions that improve its dependence on worst-client smoothness and develop a corresponding theory for random reshuffling.

Acknowledgments

Rustem Islamov and Aurelien Lucchi acknowledge the financial support of the Swiss National Science Foundation, SNSF grant No 207392. The authors used LLM based tools to help refine and tighten some proof arguments and to assist with paraphrasing and revising the manuscript.

References

[1] Kumar Kshitij Patel, Ali Zindari, Sebastian U Stich, and Lingxiao Wang. Revisiting consensus error: A fine-grained analysis of local sgd under second-order data heterogeneity. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. (Cited on pages 1, 2, 3, 4, 5, 6, 7, 8, 9, 17, 26, 37, 62, 63, and 64)

[2] Brendan McMahan and Daniel Ramage. Federated learning: Collaborative machine learning without centralized training data, Apr 2017. URL https://ai.googleblog.com/2017/04/federated-[]learning-[]collaborative.html. (Cited on page 1)

[3] H Brendan McMahan, Eider Moore, Daniel Ramage, S Hampson, and B Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data (2016). *arXiv preprint arXiv:1602.05629*, 2016. (Cited on pages 1, 2, and 4)

[4] Arthur Douillard, Qixuan Feng, Andrei A Rusu, Rachita Chhaparia, Yani Donchev, Adhiguna Kuncoro, Marc'Aurelio Ranzato, Arthur Szlam, and Jiajun Shen. Diloco: Distributed low-communication training of language models. *arXiv preprint arXiv:2311.08105*, 2023. (Cited on pages 1, 3, 4, and 18)

[5] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. (Cited on page 1)

[6] H Brendan McMahan, Eider Moore, Daniel Ramage, and Blaise Agüera y Arcas. Federated learning of deep networks using model averaging. *arXiv preprint arXiv:1602.05629*, 2016. (Cited on pages 1 and 2)

[7] Blake E Woodworth, Jialei Wang, Adam Smith, Brendan McMahan, and Nati Srebro. Graph oracle models, lower bounds, and gaps for parallel stochastic optimization. *Advances in neural information processing systems*, 31, 2018. (Cited on pages 1 and 3)

[8] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Keith Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. corr. *arXiv preprint arXiv:1912.04977*, 2019. (Cited on page 1)

[9] Jianyu Wang, Zachary Charles, Zheng Xu, Gauri Joshi, H Brendan McMahan, Maruan Al-Shedivat, Galen Andrew, Salman Avestimehr, Katharine Daly, Deepesh Data, et al. A field guide to federated optimization. *arXiv preprint arXiv:2107.06917*, 2021. (Cited on page 1)

[10] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, pages 5132–5143. PMLR, 2020. (Cited on pages 1, 3, and 5)

[11] Brian Bullins, Kumar Kshitij Patel, Ohad Shamir, Nathan Srebro, and Blake E Woodworth. A stochastic newton algorithm for distributed convex optimization. *Advances in Neural Information Processing Systems*, 34, 2021. (Cited on page 1)

[12] Kumar Kshitij Patel, Lingxiao Wang, Blake Woodworth, Brian Bullins, and Nathan Srebro. Towards optimal communication complexity in distributed non-convex optimization. In *Advances in Neural Information Processing Systems*, 2022. (Cited on pages 1, 2, 3, 6, and 37)

[13] Konstantin Mishchenko, Grigory Malinovsky, Sebastian Stich, and Peter Richtárik. Proxskip: Yes! local gradient steps provably lead to communication acceleration! finally! In *International Conference on Machine Learning*, pages 15750–15769. PMLR, 2022. (Cited on pages 1 and 3)

[14] Reese Pathak and Martin J Wainwright. Fedsplit: An algorithmic framework for fast federated optimization. *Advances in neural information processing systems*, 33:7057–7066, 2020. (Cited on page 1)

[15] Ryan Mcdonald, Mehryar Mohri, Nathan Silberman, Dan Walker, and Gideon Mann. Efficient large-scale distributed training of conditional maximum entropy models. *Advances in neural information processing systems*, 22, 2009. (Cited on pages 2 and 3)

[16] Ofer Dekel, Ran Gilad-Bachrach, Ohad Shamir, and Lin Xiao. Optimal distributed online prediction using mini-batches. *Journal of Machine Learning Research*, 13(1), 2012. (Cited on page 2)

[17] Sebastian U Stich. Local sgd converges fast and communicates little. *arXiv preprint arXiv:1805.09767*, 2018. (Cited on pages 2, 3, 4, and 6)

[18] Blake Woodworth, Kumar Kshitij Patel, Sebastian Stich, Zhen Dai, Brian Bullins, Brendan Mcmahan, Ohad Shamir, and Nathan Srebro. Is local sgd better than minibatch sgd? In *International Conference on Machine Learning*, pages 10334–10343. PMLR, 2020. (Cited on pages 2, 3, 4, 5, 6, 8, 22, 23, 31, and 37)

[19] Ohad Shamir, Nati Srebro, and Tong Zhang. Communication-efficient distributed optimization using an approximate newton-type method. In *International conference on machine learning*, pages 1000–1008. PMLR, 2014. (Cited on pages 2 and 6)

[20] Ying Sun, Gesualdo Scutari, and Amir Daneshmand. Distributed optimization based on gradient tracking revisited: Enhancing convergence rate via surrogation. In *SIAM Journal on Optimization*, 2022. (Cited on pages 2 and 6)

[21] Dmitry Kovalev, Aleksandr Beznosikov, Ekaterina Dmitrievna Borodich, Alexander Gasnikov, and Gesualdo Scutari. Optimal gradient sliding and its application to optimal distributed optimization under similarity. In *Advances in Neural Information Processing Systems*, 2022. (Cited on pages 2 and 6)

[22] Sai Praneeth Karimireddy, Martin Jaggi, Satyen Kale, Mehryar Mohri, Sashank J Reddi, Sebastian U Stich, and Ananda Theertha Suresh. Mime: Mimicking centralized stochastic algorithms in federated learning. *arXiv preprint arXiv:2008.03606*, 2020. (Cited on page 2)

[23] Tomoya Murata and Taiji Suzuki. Bias-variance reduced local sgd for less heterogeneous federated learning. *arXiv preprint arXiv:2102.03198*, 2021. (Cited on pages 2 and 3)

[24] Xiaowen Jiang, Anton Rodomanov, and Sebastian U Stich. Federated optimization with doubly regularized drift correction. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 21912–21945. PMLR, 21–27 Jul 2024. (Cited on pages 2 and 6)

[25] Xiaowen Jiang, Anton Rodomanov, and Sebastian U Stich. Stabilized proximal-point methods for federated optimization. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. (Cited on pages 2 and 6)

[26] Kumar Kshitij Patel, Margalit Glasgow, Ali Zindari, Lingxiao Wang, Sebastian U Stich, Ziheng Cheng, Nirmit Joshi, and Nathan Srebro. The limits and potentials of local sgd for distributed heterogeneous learning with intermittent communication. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 4115–4157. PMLR, 2024. (Cited on pages 2, 3, 4, 5, 6, 7, 8, 21, and 37)

[27] Martin Zinkevich, Markus Weimer, Lihong Li, and Alex Smola. Parallelized stochastic gradient descent. *Advances in neural information processing systems*, 23, 2010. (Cited on page 3)

[28] Jian Zhang, Christopher De Sa, Ioannis Mitliagkas, and Christopher Ré. Parallel sgd: When does averaging help? *arXiv preprint arXiv:1606.07365*, 2016. (Cited on page 3)

[29] Aymeric Dieuleveut and Kumar Kshitij Patel. Communication trade-offs for local-sgd with large step size. *Advances in Neural Information Processing Systems*, 32, 2019. (Cited on page 3)

[30] Ahmed Khaled, Konstantin Mishchenko, and Peter Richtárik. Tighter theory for local sgd on identical and heterogeneous data. In *International Conference on Artificial Intelligence and Statistics*, pages 4519–4529. PMLR, 2020. (Cited on pages 3 and 5)

[31] Anastasia Koloskova, Nicolas Loizou, Sadra Boreiri, Martin Jaggi, and Sebastian Stich. A unified theory of decentralized sgd with changing topology and local updates. In *International Conference on Machine Learning*, pages 5381–5393. PMLR, 2020. (Cited on pages 3, 4, 5, 9, 17, 62, 63, and 64)

[32] Blake E Woodworth, Kumar Kshitij Patel, and Nati Srebro. Minibatch vs local sgd for heterogeneous distributed learning. *Advances in Neural Information Processing Systems*, 33:6281–6292, 2020. (Cited on pages 3, 4, 5, 6, 7, 8, 16, 21, 22, 24, 33, 36, and 37)

[33] Honglin Yuan and Tengyu Ma. Federated accelerated stochastic gradient descent. *Advances in Neural Information Processing Systems*, 33:5332–5344, 2020. (Cited on page 3)

[34] Blake E Woodworth, Brian Bullins, Ohad Shamir, and Nathan Srebro. The min-max complexity of distributed stochastic convex optimization with intermittent communication. In *Conference on Learning Theory*, pages 4386–4437. PMLR, 2021. (Cited on pages 3, 4, 23, and 59)

[35] Blake Woodworth. The minimax complexity of distributed optimization. *arXiv preprint arXiv:2109.00534*, 2021. (Cited on pages 3 and 37)

[36] Margalit R Glasgow, Honglin Yuan, and Tengyu Ma. Sharp bounds for federated averaging (local sgd) and continuous perspective. In *International Conference on Artificial Intelligence and Statistics*, pages 9050–9090. PMLR, 2022. (Cited on pages 3, 4, 5, 8, 9, 37, and 49)

[37] Jianyu Wang, Rudrajit Das, Gauri Joshi, Satyen Kale, Zheng Xu, and Tong Zhang. On the unreasonable effectiveness of federated averaging with heterogeneous data. *arXiv preprint arXiv:2206.04723*, 2022. (Cited on page 3)

[38] Kumar Kshitij Patel, Lingxiao Wang, Aadirupa Saha, and Nathan Srebro. Federated online and bandit convex optimization. In *International Conference on Machine Learning*, pages 27439–27460. PMLR, 2023. (Cited on page 3)

[39] Kumar Kshitij Patel, Margalit Glasgow, Lingxiao Wang, Nirmit Joshi, and Nathan Srebro. On the still unreasonable effectiveness of federated averaging for heterogeneous distributed learning. In *Federated Learning and Analytics in Practice: Algorithms, Systems, Applications, and Opportunities*, 2023. (Cited on pages 3 and 37)

[40] Paul Mangold, Alain Oliviero Durmus, Aymeric Dieuleveut, Sergey Samsonov, and Eric Moulines. Refined analysis of constant step size federated averaging and federated richardson-romberg extrapolation. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258, pages 5023–5031, 2025. (Cited on page 3)

[41] Kumar Kshitij Patel. What makes local updates effective: The role of data heterogeneity and smoothness. *arXiv preprint arXiv:2507.00195*, 2025. (Cited on pages 3 and 37)

[42] Ruichen Luo, Sebastian U Stich, Samuel Horvath, and Martin Takac. Revisiting localsgd and scaffold: Improved rates and missing analysis. *arXiv preprint arXiv:2501.04443*, 2025. (Cited on pages 3, 5, 6, 7, 16, 21, 22, 24, and 25)

[43] Ahmed Khaled, Satyen Kale, Arthur Douillard, Chi Jin, Rob Fergus, and Manzil Zaheer. Understanding outer optimizers in local sgd: Learning rates, momentum, and acceleration. *arXiv preprint arXiv:2509.10439*, 2025. (Cited on page 3)

[44] Eduard Gorbunov, Filip Hanzely, and Peter Richtárik. Local sgd: Unified theory and new efficient methods. In *International Conference on Artificial Intelligence and Statistics*, pages 3556–3564. PMLR, 2021. (Cited on page 3)

[45] Zachary Charles, Gabriel Teston, Lucio Dery, Keith Rush, Nova Fallen, Zachary Garrett, Arthur Szlam, and Arthur Douillard. Communication-efficient language model training scales reliably and robustly: Scaling laws for diloco. *arXiv preprint arXiv:2503.09799*, 2025. (Cited on pages 3 and 18)

[46] Sebastian U Stich, Jean-Baptiste Cordonnier, and Martin Jaggi. Sparsified sgd with memory. In *Proceedings of Advances in Neural Information Processing Systems*, 2018. (Cited on page 3)

[47] Sai Praneeth Karimireddy, Quentin Rebjock, Sebastian Stich, and Martin Jaggi. Error feedback fixes signsgd and other gradient compression schemes. In *Proceedings of the 36th International Conference on Machine Learning (ICML 2019)*, 2019. (Cited on page 3)

[48] Yuan Gao, Rustem Islamov, and Sebastian U Stich. EControl: Fast distributed optimization with compression and error control. In *The Twelfth International Conference on Learning Representations*, 2024. (Cited on page 3)

[49] Eduard Gorbunov, Dmitry Kovalev, Dmitry Makarenko, and Peter Richtárik. Linearly converging error compensated SGD. *Advances in Neural Information Processing Systems*, 33: 20889–20900, 2020. (Cited on page 3)

[50] Abdurakhmon Sadiev, Grigory Malinovsky, Eduard Gorbunov, Igor Sokolov, Ahmed Khaled, Konstantin Burlachenko, and Peter Richtárik. Don't compress gradients in random reshuffling: compress gradient differences. *Advances in Neural Information Processing Systems*, 37:84523–84605, 2024. (Cited on page 3)

[51] Dan Alistarh, Torsten Hoefler, Mikael Johansson, Nikola Konstantinov, Sarit Khirirat, and Cédric Renggli. The convergence of sparsified gradient methods. In *Proceedings of Advances in Neural Information Processing Systems*, 2018. (Cited on page 3)

[52] Sindri Magnússon, Hossein Shokri-Ghadikolaei, and Na Li. On maintaining linear convergence of distributed learning and optimization under limited communication. In *2019 53rd Asilomar Conference on Signals, Systems, and Computers*, pages 432–436, 2019. doi: 10.1109/IEEECONF44664.2019.9049052. (Cited on page 3)

[53] Min Ye and Emmanuel Abbe. Communication-computation efficient gradient coding. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5610–5619. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/ye18a.html. (Cited on page 3)

[54] Sebastian U Stich and Sai Praneeth Karimireddy. The error-feedback framework: Better rates for sgd with delayed gradients and compressed updates. *Journal of Machine Learning Research*, 2020. (Cited on page 3)

[55] Rustem Islamov, Mher Safaryan, and Dan Alistarh. Asgrad: A sharp unified analysis of asynchronous-sgd algorithms. In *International Conference on Artificial Intelligence and Statistics*, pages 649–657. PMLR, 2024. (Cited on page 3)

[56] Kang Wei, Jun Li, Ming Ding, Chuan Ma, Howard H Yang, Farhad Farokhi, Shi Jin, Tony QS Quek, and H Vincent Poor. Federated learning with differential privacy: Algorithms and performance analysis. *IEEE transactions on information forensics and security*, 15:3454–3469, 2020. (Cited on page 3)

[57] Antonious Girgis, Deepesh Data, Suhas Diggavi, Peter Kairouz, and Ananda Theertha Suresh. Shuffled model of differential privacy in federated learning. In *International Conference on Artificial Intelligence and Statistics*, pages 2521–2529. PMLR, 2021. (Cited on page 3)

[58] Andrew Lowy, Ali Ghafelebashi, and Meisam Razaviyayn. Private non-convex federated learning without a trusted server. In *International Conference on Artificial Intelligence and Statistics*, pages 5749–5786. PMLR, 2023. (Cited on page 3)

[59] Tomoya Murata and Taiji Suzuki. Diff2: Differential private optimization via gradient differences for nonconvex distributed learning. In *International Conference on Machine Learning*, pages 25523–25548. PMLR, 2023. (Cited on page 3)

[60] Lingxiao Wang, Xingyu Zhou, Kumar Kshitij Patel, Lawrence Tang, and Aadirupa Saha. Efficient private federated non-convex optimization with shuffled model. In *Privacy Regulation and Protection in Machine Learning*, 2024. (Cited on page 3)

[61] Rustem Islamov, Samuel Horvath, Aurelien Lucchi, Peter Richtarik, and Eduard Gorbunov. Double momentum and error feedback for clipping with fast rates and differential privacy. *arXiv preprint arXiv:2502.11682*, 2025. (Cited on page 3)

[62] Egor Shulgin, Sarit Khirirat, and Peter Richtárik. Smoothed normalization for efficient distributed private optimization. *arXiv preprint arXiv:2502.13482*, 2025. (Cited on page 3)

[63] Dan Alistarh, Zeyuan Allen-Zhu, and Jerry Li. Byzantine stochastic gradient descent. *Advances in neural information processing systems*, 31, 2018. (Cited on page 3)

[64] Sai Praneeth Karimireddy, Lie He, and Martin Jaggi. Learning from history for byzantine robust optimization. In *International conference on machine learning*, pages 5311–5319. PMLR, 2021. (Cited on page 3)

[65] Rustem Islamov, Grigory Malinovsky, Alexander Gaponov, Aurelien Lucchi, Peter Richtárik, and Eduard Gorbunov. Byzantine-robust and differentially private federated optimization under weaker assumptions. In Emilija Perković and Daniel Malinsky, editors, *Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence*, volume 337 of *Proceedings of Machine Learning Research*, pages 2378–2441. PMLR, 17–21 Aug 2026. URL https://proceedings.mlr.press/v337/islamov26a.html. (Cited on page 3)

[66] Grigory Malinovsky, Peter Richtárik, Samuel Horváth, and Eduard Gorbunov. Byzantine robustness and partial participation can be achieved at once: Just clip gradient differences. *Advances in Neural Information Processing Systems*, 37:34900–34979, 2024. (Cited on page 3)

[67] Zachary Charles and Jakub Konecny. On the outsized importance of learning rates in local update methods. *arXiv preprint arXiv:2007.00878*, 2020. (Cited on pages 3 and 18)

[68] Tao Lin, Sebastian U Stich, Kumar Kshitij Patel, and Martin Jaggi. Don't use large mini-batches, use local sgd. *arXiv preprint arXiv:1808.07217*, 2018. (Cited on page 4)

[69] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017. (Cited on page 4)

[70] Avrim Blum, Nika Haghtalab, Ariel D Procaccia, and Mingda Qiao. Collaborative pac learning. *Advances in Neural Information Processing Systems*, 30, 2017. (Cited on page 4)

[71] Minbiao Han, Kumar Kshitij Patel, Han Shao, and Lingxiao Wang. On the effect of defections in federated learning and how to prevent them. *arXiv preprint arXiv:2311.16459*, 2023. (Cited on page 4)

[72] Jean-Bernard Baillon and Georges Haddad. Quelques propriétés des opérateurs angle-bornés et n-cycliquement monotones. *Israel Journal of Mathematics*, 26(2):137–150, 1977. (Cited on page 22)

[73] Heinz H Bauschke and Patrick L Combettes. The baillon-haddad theorem revisited. *arXiv preprint arXiv:0906.0807*, 2009. (Cited on page 22)

[74] Charlie Hou, Kiran K Thekumparampil, Giulia Fanti, and Sewoong Oh. Fedchain: Chained algorithms for near-optimal communication cost in federated learning. *arXiv preprint arXiv:2108.06869*, 2021. (Cited on pages 49 and 50)

Appendix

Appendix Contents

1 Introduction **1**

1.1 Our Contributions 2

1.2 Related work 3

2 Setting and Preliminaries **3**

2.1 Regularity Assumptions and the Algorithms 3

2.2 A Hierarchy of Data Heterogeneity Assumptions 4

3 Improved Analyses of Local SGD for Classes $\mathcal{P}^{H,B,\sigma}_{\zeta_\star,\tau}$ **6**

3.1 Perturbing Against Dense Mini-batch SGD 6

3.2 Perturbing Against Single-Machine SGD 8

4 A New Lower Bound and the Tightness of Our Results **8**

5 Multi-pass SGD with Replacement under Second-Order Heterogeneity **9**

A Missing Details from Section 2 **18**

B Additional Notation for the Appendix **20**

C Useful Prior Results and Technical Context **21**

C.1 A Canonical Descent Lemma and the Upper Bound of Woodworth et al. [32] . . . 21

C.2 The Empirical Variance Trick of Luo et al. [42] 21

C.3 A Co-coercivity Property on Each Machine 22

C.4 Local SGD is Never Worse than Single Machine SGD in the Homogeneous Setting 23

C.5 Other Useful Analytical Identities 23

D Useful New Technical Lemmas for Our Upper Bounds **24**

D.1 Technical Lemmas for the Consensus Error-based Proof 24

D.2 Technical Lemmas for the Single-Machine Mimicking Proof 27

E Proofs of Our Upper Bounds for $\mathcal{P}^{H,B,\sigma}_{\zeta_\star,\tau}$ **31**

E.1 Proof of Theorem 5 31

E.2 Proof of Theorem 6 32

F A Single-Machine-Style Upper Bound for $\mathcal{P}^{H,B,\sigma}_{\zeta}$ **33**

F.1 A One-Step Recursion using $B(t)$ and $D(t)$ 34

F.2 Proof of Theorem 7 36

G Proofs of Our Lower Bounds **37**
G.1 A Sharper Lower Bound When $\zeta_\star = 0$ . . . 37
G.2 A Sharper Lower Bound When $\tau = 0$ . . . 41
G.3 A New Lower Bound for $\mathcal{P}_\zeta^{H,B,\sigma}$ . . . 49
G.4 Lower Bounds for the Time-Averaged Output . . . 50

H Phase Diagram: Comparison of New Lower and Upper Bounds **59**
H.1 Analysis of Upper Bound's Dominant Term . . . 59
H.2 Analysis of Lower Bound's Dominant Term . . . 60
H.3 Combined Phase Diagram . . . 62

I Phase Diagram: Comparison of Prior Lower and Upper Bounds **62**
I.1 Analysis of Dominant Term of the Upper Bound in Koloskova et al. [31] . . . 63
I.2 Analysis of Dominant Term of the Lower Bound in Patel et al. [1] . . . 63

J Fresh-sample SGD Lower Bound in the Heterogeneous Smoothness Regime **64**
J.1 A rare-curvature block . . . 64
J.2 Calibration of the parameters . . . 70
J.3 Adding the noise block . . . 70
J.4 Proof of the theorem . . . 70

A Missing Details from Section 2

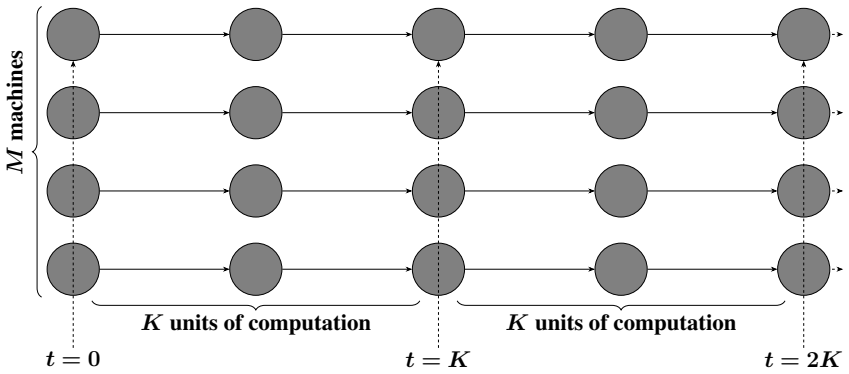

Figure A.1: Illustration of the IC setting: M clients communicate R times with K units of computation in between. The baselines we consider in the paper all require K stochastic gradient computations between communication rounds on each client.

Local SGD with Two Step-sizes. Algorithm 1 presents a popular version of Local SGD which often performs better in practice [67], and has motivated an entire class of algorithms with an inner and outer optimization loop (for instance, see DILOCO [4, 45]). In Algorithm 1, $x_{r,k}^m$ is the kth local model on machine m, leading up to the rth round of communication, while x_r is the consensus (global) model at the end of the rth communication. For Local SGD, η is referred to as the inner (local) step-size, while β is the outer (server) step-size. Setting $\beta = 1$ recovers *"vanilla Local SGD"* with a single step-size.

Proposition 1 (Bounded Hessian heterogeneity implies Assumption 3)**.** *Suppose that each F_m is twice differentiable and that, for every $z \in \mathbb{R}^d$,*

$$\frac{1}{M} \sum_{m \in [M]} \left\| \nabla^2 F_m(z) - \nabla^2 F(z) \right\|_2^2 \leq \tau^2 \ .$$

Algorithm 1 Local SGD / Federated Averaging with server step-size β

Input: communication rounds R, local steps K, step-size η, server step-size β, client distributions $\{\mathcal{D}_1, \ldots, \mathcal{D}_M\}$
Initialize: global model $x_0 = 0$
for $r = 1$ **to** R **do**
  for $m = 1$ **to** M **do**
    $x_{r,0}^m \leftarrow x_{r-1}$
    for $k = 0$ **to** $K - 1$ **do**
      Sample $z_{r,k}^m \overset{\text{i.i.d.}}{\sim} \mathcal{D}_m$
      $x_{r,k+1}^m \leftarrow x_{r,k}^m - \eta \nabla f(x_{r,k}^m; z_{r,k}^m)$
    end for
  end for
  $x_r \leftarrow x_{r-1} + \frac{\beta}{M} \sum_{m=1}^M (x_{r,K}^m - x_{r-1})$
end for

Algorithm 2 Mini-batch SGD in the IC setting

Input: communication rounds R, local batch size K, step-size γ, client distributions $\{\mathcal{D}_1, \ldots, \mathcal{D}_M\}$
Initialize: global model $x_0 = 0$
for $r = 1$ **to** R **do**
  $G_r \leftarrow 0$
  for $m = 1$ **to** M **do**
    for $k = 0$ **to** $K - 1$ **do**
      Sample $z_{r,k}^m \overset{\text{i.i.d.}}{\sim} \mathcal{D}_m$
      $g_{r,k}^m \leftarrow \nabla f(x_{r-1}; z_{r,k}^m)$
      $G_r \leftarrow G_r + g_{r,k}^m$
    end for
  end for
  $x_r \leftarrow x_{r-1} - \frac{\gamma}{MK} G_r$
end for

Figure A.2: Two canonical distributed optimization baselines implemented in the intermittent communication setting: Local SGD with an inner and outer step-size and "large" mini-batch SGD. Note that for η defined as $\gamma = K\eta$ in Algorithm 2, step size η corresponds to the per-gradient step size used by each of the sequential Local SGD updates.

Then the objectives satisfy τ-second-order heterogeneity in the sense of Assumption 3. Indeed, for any $x, y \in \mathbb{R}^d$, letting $v = x - y$, the fundamental theorem of calculus gives

$$\nabla F_m(x) - \nabla F_m(y) - \big(\nabla F(x) - \nabla F(y)\big) = \int_0^1 \big(\nabla^2 F_m(y + tv) - \nabla^2 F(y + tv)\big) v \, dt \ .$$

Therefore, by Jensen's inequality and the assumed Hessian heterogeneity bound,

$$\begin{aligned}
&\frac{1}{M} \sum_{m \in [M]} \big\| \nabla F_m(x) - \nabla F_m(y) - \big(\nabla F(x) - \nabla F(y)\big) \big\|_2^2 \\
&\qquad \le \int_0^1 \frac{1}{M} \sum_{m \in [M]} \big\| \nabla^2 F_m(y + tv) - \nabla^2 F(y + tv) \big\|_2^2 \, \|x - y\|_2^2 \, dt \le \tau^2 \, \|x - y\|_2^2 \ .
\end{aligned}$$

This is why we refer to Assumption 3 as a second-order heterogeneity assumption.

Proposition 2 ($\mathcal{P}^{H,B,\sigma}_{\zeta_\star,\tau}$ interpolates between problem classes)**.** *Assumption 3 controls how different the "second-order" geometry can be across clients, as opposed to the "first-order" geometry controlled by Assumption 2. It also interpolates between the problem classes discussed above. At one extreme, if each F_m is H-smooth, then F is also H-smooth, and*

$$\big\| \nabla F_m(x) - \nabla F_m(y) - \big(\nabla F(x) - \nabla F(y)\big) \big\|_2 \le 2H \, \|x - y\|_2 \ .$$

Thus, Assumption 3 is vacuous when $\tau = 2H$, and $\boldsymbol{\mathcal{P}^{H,B,\sigma}_{\zeta_\star,\tau=2H}} \subseteq \boldsymbol{\mathcal{P}^{H,B,\sigma}_{\zeta_\star}}$.

At the other extreme, when $\tau = 0$, the gradient gaps $\nabla F_m(x) - \nabla F(x)$ are constant in x implying,

$$\nabla F_m(x) - \nabla F(x) = \nabla F_m(x^\star) - \nabla F(x^\star) = \nabla F_m(x^\star) \ ,$$

and therefore

$$\frac{1}{M} \sum_{m \in [M]} \big\| \nabla F_m(x) - \nabla F(x) \big\|_2^2 = \frac{1}{M} \sum_{m \in [M]} \big\| \nabla F_m(x^\star) \big\|_2^2 \le \zeta_\star^2 \ .$$

Thus when $\tau = 0$, we have

$$\boldsymbol{\mathcal{P}^{H,B,\sigma}_{\zeta_\star,\tau=0}} \subseteq \boldsymbol{\mathcal{P}^{H,B,\sigma}_{\zeta=\zeta_\star}}.$$

This is precisely why $\mathcal{P}^{H,B,\sigma}_{\zeta_\star,\tau}$ sits between the two problem classes $\mathcal{P}^{H,B,\sigma}_{\zeta_\star}$ and $\mathcal{P}^{H,B,\sigma}_{\zeta}$ in Table 1.

Example 1 (Demonstration of $H \gg \tau$)**.** *For each $m \in [M]$, let*

$$F_m(x) = \frac{1}{2}(h + \delta s_m)x^2 + b_m x,$$

where $s_m \in \{-1, 1\}$, $\frac{1}{M}\sum_{m=1}^M s_m = 0$, and $\delta \le h$. Then Assumptions 2 and 3 hold with $H = h$ and $\tau = \delta$, respectively. The coefficients $\{b_m\}_{m=1}^M$ may also differ substantially, producing different local gradients and minimizers without affecting τ. In particular, when $\delta = 0$, we have $\tau = 0$ even though the client objectives need not be identical.

Proof. To prove this statement, one needs to do simple computations of derivatives and plug them into Assumptions 2 and 3. □

This example can be further extended to the case with

$$F_m(x) = \frac{1}{2n} \sum_{i=1}^n (a_{m,i}^\top x - y_{m,i})^2.$$

In the case of Gaussian isotropic local covariates, one can obtain the relation

$$\frac{\tau}{H} = \tilde{\mathcal{O}}(\sqrt{d/n}).$$

B Additional Notation for the Appendix

Throughout the appendix, we analyze the vanilla variant of Local SGD, which uses the outer step-size $\beta = 1$ (see Algorithm 1). In particular, for brevity, we will use the notation in Equation (5), along with the following *"ghost iterate"* which might not be computed at each time step $t \in [0, KR]$,

$$\bar{x}_t = \frac{1}{M} \sum_{m \in [M]} x_t^m.$$

It would also be useful to define the following sequences, which appear many times throughout our analyses, for all $t \in [0, KR]$

$$A(t) \coloneqq \mathbb{E}\left[\|\bar{x}_t - x^\star\|_2^2\right], \qquad \text{(Expected ghost iterate error)}$$

$$B(t) \coloneqq \frac{1}{M} \sum_{m \in [M]} \mathbb{E}\left[\|x_t^m - x^\star\|_2^2\right] \geq A(t), \qquad \text{(Expected average iterate error)}$$

$$\Xi_t \coloneqq \frac{1}{M} \sum_{m \in [M]} \|x_t^m - \bar{x}_t\|_2^2, \qquad \text{(Consensus error)}$$

$$C(t) \coloneqq \mathbb{E}\left[\Xi_t\right] = \frac{1}{M} \sum_{m \in [M]} \mathbb{E}\left[\|x_t^m - \bar{x}_t\|_2^2\right], \qquad \text{(Expected consensus error)}$$

$$D(t) \coloneqq \mathbb{E}\left[F(\bar{x}_t) - F(x^\star)\right] \geq 0, \qquad \text{(Expected function sub-optimality)}$$

$$E(t) \coloneqq \frac{1}{M} \sum_{m \in [M]} \mathbb{E}\left[F(x_t^m) - F(x^\star)\right], \qquad \text{(Expected average function sub-optimality)}$$

$$V(t) \coloneqq \frac{1}{M} \sum_{m \in [M]} \mathbb{E}\left[\|\nabla F_m(\bar{x}_t) - \nabla F(\bar{x}_t)\|_2^2\right], \qquad \text{(Expected gradient heterogeneity)}$$

$$R(t) \coloneqq \frac{1}{M} \sum_{m \in [M]} \mathbb{E}[\mathcal{D}(x_t^m, \bar{x}_t)], \qquad \text{(Expected average Bregman divergence)}$$

where $\mathcal{D}_{F_m}(x_t^m, \bar{x}_t) \coloneqq F_m(x_t^m) - F_m(\bar{x}_t) - \langle \nabla F_m(\bar{x}_t), x_t^m - \bar{x}_t \rangle$

$$Q(t) \coloneqq \frac{1}{M} \sum_{m \in [M]} \mathbb{E}[\langle \nabla F_m(\bar{x}_t) - \nabla F(\bar{x}_t), x_t^m - \bar{x}_t \rangle],$$

$$S(t) \coloneqq \frac{1}{M} \sum_{m \in [M]} \mathbb{E}[F_m(x_t^m) - F_m(x^\star)], \qquad \text{(Expected local function sub-optimality).}$$

where throughout we fix $x^\star \in \arg\min_{x \in \mathbb{R}^d} F(x)$ to be the optimizer of the average objective which satisfies Assumption 1. We will also denote the maximum of the first two sequences by,

$$A_{\max} \coloneqq \max_{t \in [0, KR]} A(t),$$

$$B_{\max} \coloneqq \max_{t \in [0, KR]} B(t),$$

$$C_{\max} \coloneqq \max_{t \in [0, KR]} C(t),$$

$$V_{\max} \coloneqq \max_{t \in [0, KR]} V(t).$$

It would also be useful to define the following filtration for all $t \in [KR]$, which allows easily dealing with the stochastic noise on each client,

$$\mathcal{H}_t \coloneqq \sigma\left(z_0^1, \ldots, z_0^M, \ldots, z_{t-1}^1, \ldots, z_{t-1}^M\right),$$

where $\sigma(S)$ represents the sigma algebra[9] generated by a set $S \in \mathcal{Z}^\star$. We write $x \in \mathcal{H}$ to mean that x is measurable under the sigma algebra $\mathcal{H}$. It would be useful to note that,

$$x_t^m, \bar{x}_t, \Xi_t \in \mathcal{H}_t,$$

for all $m \in [M]$ and $t \in [KR]$.

[9] We apologize for the abuse of notation w.r.t. Equation (3).

C Useful Prior Results and Technical Context

C.1 A Canonical Descent Lemma and the Upper Bound of Woodworth et al. [32]

Many of our proofs rely on the following standard lemma, which, for Local SGD, imitates the one-step recursion usual in SGD analysis by introducing an expected consensus error term.

Lemma 2 (Lemma 7, Woodworth et al. [32])**.** *Consider problems in $\mathcal{P}^{H,B,\sigma}$ then for $\eta \leq \frac{1}{2H}$, the Local SGD ghost iterate satisfies for all $t \in [0, T-1]$*

$$\mathbb{E}[F(\bar{x}_t) - F^\star] \leq \frac{1}{\eta}\mathbb{E}[\|\bar{x}_t - x^*\|_2^2] - \frac{1}{\eta}\mathbb{E}[\|\bar{x}_{t+1} - x^\star\|_2^2] + \frac{3\eta\sigma^2}{M} + \frac{2H}{M}\sum_{m=1}^{M}\mathbb{E}\left[\|\bar{x}_t - x_t^m\|_2^2\right]. \quad (12)$$

In the standard analysis due to Woodworth et al. [32], Assumption 2 allows bounding the consensus error as follows,

$$C(t) \leq 3\eta^2 K\sigma^2 + 6\eta^2 K^2\zeta^2. \quad (13)$$

Plugging this upper bound, unrolling the recursion, telescoping appropriately, and tuning the step-size results in the upper bound stated in Table 1 for problem class $\mathcal{P}_{\zeta}^{H,B,\sigma}$. Our analysis in this paper effectively replaces the ζ^2 term in the above upper bound by $\zeta_\star^2 + \tau^2 B^2$ up to numerical constants, thus also allowing us to replace the dependence on Assumption 2 by Assumptions 1 and 3, i.e., extending their paper to $\mathcal{P}_{\zeta_\star,\tau}^{H,B,\sigma}$. Our strategy builds on the following observation due to Patel et al. [26], which bounds the uniform gradient heterogeneity in terms of the existing iterate error.

Lemma 3. *Consider problems in $\mathcal{P}_{\zeta_\star,\tau}^{H,B,\sigma}$ which also satisfy twice differentiability on each client, then for all $m \in [M]$,*

$$\sup_{x\in\mathbb{B}_2(D)}\frac{1}{M}\sum_{m\in[M]}\|\nabla F_m(x) - \nabla F(x)\|_2^2 \leq (\zeta_\star + \tau D)^2.$$

In Lemma 8, we refine the above lemma in the following ways:

1. we require it to only hold in expectation over the trajectory of Local SGD, i.e., we only want to upper bound $V_{\max}$;
2. we remove the twice-differentiability requirement;
3. up to numerical constants, we replace the D in the upper bound by B using a careful analysis of the Local SGD iterates, and bounding $A_{\max}$; and
4. we replace the supremum by an average over $m \in [M]$.

C.2 The Empirical Variance Trick of Luo et al. [42]

The consensus error upper bound of Woodworth et al. [32] mentioned in Equation (13) proceeds by applying Jensen's inequality as follows,

$$\begin{aligned}
C(t) &= \frac{1}{M}\sum_{m\in[M]}\mathbb{E}\left[\|x_t^m - \bar{x}_t\|_2^2\right], \\
&= \frac{1}{M}\sum_{m\in[M]}\mathbb{E}\left[\left\|x_t^m - \frac{1}{M}\sum_{n\in[M]}x_t^n\right\|_2^2\right], \\
&\leq \frac{1}{M^2}\sum_{m,n\in[M]}\mathbb{E}\left[\|x_t^m - x_t^n\|_2^2\right],
\end{aligned}$$

and then upper bound $\mathbb{E}\left[\|x_t^m - x_t^n\|_2^2\right]$ for every pair of machines $m, n \in [M]$. This is also why they need the stronger version of Assumption 2 which takes a supremum over $m \in [M]$, instead of an average. Luo et al. [42] instead use the following series of calculations,

$$\Xi_t = \frac{1}{M}\sum_{m\in[M]}\|x_t^m - \bar{x}_t\|_2^2,$$

$$= \min_{y \in \mathbb{R}^d} \frac{1}{M} \sum_{m \in [M]} \|x_t^m - y\|_2^2 ,$$

$$\leq \frac{1}{M} \sum_{m \in [M]} \|x_t^m - (\bar{x}_{t-1} - \eta \nabla F(\bar{x}_{t-1}))\|_2^2 , \tag{14}$$

which is simply a restatement of the fact that the empirical mean of M vectors minimizes their empirical variance. While quite simple, this trick is quite useful for our analyses, as the reader will note in the next section.

C.3 A Co-coercivity Property on Each Machine

We note the following property on each machine, which is often used in Local SGD analysis to bound the consensus error [18, 32, 42], and more generally relies on the folklore co-coercivity result of Baillon-Haddad [72, 73].

Lemma 4. *For problems in the class $\mathcal{P}^{H,B,\sigma}$ and with $\eta \leq \frac{2}{H}$ each machine $m \in [M]$ satisfies for all $x, y \in \mathbb{R}^d$,*

$$\|x - \eta \nabla F_m(x) - (y - \eta \nabla F_m(y))\|_2^2 \leq \|x - y\|_2^2 .$$

Proof. We first prove the following auxiliary inequality: for any $x, y \in \mathbb{R}^d$,

$$\frac{1}{2H} \|\nabla F_m(x) - \nabla F_m(y)\|_2^2 \leq F_m(x) - F_m(y) - \langle \nabla F_m(y), x - y \rangle . \tag{15}$$

To show this, define

$$g(z) \coloneqq F_m(z) - \langle \nabla F_m(y), z \rangle .$$

Then g is convex and H-smooth, and

$$\nabla g(z) = \nabla F_m(z) - \nabla F_m(y), \qquad \nabla g(y) = 0.$$

Since g is H-smooth, for any $u, v \in \mathbb{R}^d$,

$$g(u) \leq g(v) + \langle \nabla g(v), u - v \rangle + \frac{H}{2} \|u - v\|_2^2 .$$

Apply this with $u = x - \frac{1}{H} \nabla g(x)$ and $v = x$ to get

$$g\left(x - \frac{1}{H} \nabla g(x)\right) \leq g(x) + \left\langle \nabla g(x), -\frac{1}{H} \nabla g(x) \right\rangle + \frac{H}{2} \left\| \frac{1}{H} \nabla g(x) \right\|_2^2$$

$$= g(x) - \frac{1}{2H} \|\nabla g(x)\|_2^2 .$$

Since g is convex and $\nabla g(y) = 0$, the point y is a global minimizer of g, hence

$$g(y) \leq g\left(x - \frac{1}{H} \nabla g(x)\right) .$$

Combining the last two displays yields

$$g(x) - g(y) \geq \frac{1}{2H} \|\nabla g(x)\|_2^2 .$$

Substituting back the definition of g gives (15).

Now apply (15) twice: once as written, and once with x and y swapped. This gives

$$\frac{1}{2H} \|\nabla F_m(x) - \nabla F_m(y)\|_2^2 \leq F_m(x) - F_m(y) - \langle \nabla F_m(y), x - y \rangle ,$$

$$\frac{1}{2H} \|\nabla F_m(x) - \nabla F_m(y)\|_2^2 \leq F_m(y) - F_m(x) - \langle \nabla F_m(x), y - x \rangle .$$

Adding these inequalities, the function-value terms cancel, and we obtain

$$\frac{1}{H}\left\|\nabla F_m(x)-\nabla F_m(y)\right\|_2^2 \le \left\langle \nabla F_m(x)-\nabla F_m(y), x-y\right\rangle . \tag{16}$$

Now note that

$$\begin{aligned}
&\left\|x-\eta\nabla F_m(x)-(y-\eta\nabla F_m(y))\right\|_2^2 \\
&\quad = \left\|x-y-\eta\left(\nabla F_m(x)-\nabla F_m(y)\right)\right\|_2^2, \\
&\quad = \|x-y\|_2^2+\eta^2\left\|\nabla F_m(x)-\nabla F_m(y)\right\|_2^2-2\eta\left\langle x-y,\nabla F_m(x)-\nabla F_m(y)\right\rangle .
\end{aligned}$$

Using the bound just proved,

$$-2\eta\left\langle x-y,\nabla F_m(x)-\nabla F_m(y)\right\rangle \le -\frac{2\eta}{H}\left\|\nabla F_m(x)-\nabla F_m(y)\right\|_2^2 .$$

Therefore,

$$\begin{aligned}
&\left\|x-\eta\nabla F_m(x)-(y-\eta\nabla F_m(y))\right\|_2^2 \\
&\quad \le \|x-y\|_2^2+\eta^2\left\|\nabla F_m(x)-\nabla F_m(y)\right\|_2^2-\frac{2\eta}{H}\left\|\nabla F_m(x)-\nabla F_m(y)\right\|_2^2 \\
&\quad = \|x-y\|_2^2+\eta\left(\eta-\frac{2}{H}\right)\left\|\nabla F_m(x)-\nabla F_m(y)\right\|_2^2 \\
&\quad \le \|x-y\|_2^2,
\end{aligned}$$

where the last step uses $\eta \le \frac{2}{H}$. □

C.4 Local SGD is Never Worse than Single Machine SGD in the Homogeneous Setting

A classic result from Woodworth et al. [18] is to show that in the homogeneous setting, that is, for problems in the problem class $\mathcal{P}_{hom}^{H,B,\sigma}$, Local SGD is no worse than single machine SGD. Specifically, they presented an alternative convergence analysis of Local SGD that is required to obtain the rate in Table 1.

Lemma 5. *For problems in the problem class $\mathcal{P}_{hom}^{H,B,\sigma}$ choosing the step-size as $\eta = \min\left\{\frac{1}{2H}, \frac{B}{\sigma\sqrt{KR}}\right\}$, the Local SGD iterate $\hat{x}^{\text{Local-SGD}} = \frac{1}{TM}\sum_{t\in[0,T-1],m\in[M]} x_t^m$ satisfies,*

$$\mathbb{E}\left[F(\hat{x}^{\text{Local-SGD}})-F(x^\star)\right] \le 2\cdot\left(\frac{HB^2}{KR}+\frac{\sigma B}{\sqrt{KR}}\right).$$

In the regime when σ is small, in fact Woodworth et al. [34] showed that accelerated single machine SGD is mini-max optimal, and in that regime, using a similar argument as the above result, one can show that accelerated local SGD is optimal as well. In Theorem 6, we extend the above guarantee to the heterogeneous setting, specifically to the problem class $\mathcal{P}_{\zeta_\star,\tau}^{H,B,\sigma}$.

C.5 Other Useful Analytical Identities

We will also use the following inequality several times; it is essentially a variant of the A.M.-G.M. inequality.

Lemma 6. *For any $a, b \in \mathbb{R}$ and $\gamma > 0$ we have,*

$$(a+b)^2 \le \left(1+\frac{1}{\gamma}\right)a^2+(1+\gamma)\,b^2.$$

Proof. Note the following,

$$\begin{aligned}
(a+b)^2 &= a^2+b^2+2ab, \\
&= a^2+b^2+2\left(\frac{a}{\sqrt{\gamma}}\right)\left(\sqrt{\gamma}b\right),
\end{aligned}$$

$$\overset{(i)}{\leq} a^2 + b^2 + \frac{a^2}{\gamma} + \gamma b^2,$$
$$\leq \left(1 + \frac{1}{\gamma}\right) a^2 + (1+\gamma)\, b^2,$$

where (i) uses A.M.-G.M. Inequality. This proves the first statement of the lemma. □

D Useful New Technical Lemmas for Our Upper Bounds

We will prove several important technical lemmas in this section.

D.1 Technical Lemmas for the Consensus Error-based Proof

In this sub-section, we develop lemmas that would be used in the proof of Theorem 5 in the next section.

D.1.1 A Refined Upper Bound for $C(t)$

We begin by proving the following refined consensus error upper bound by refining the proof of Woodworth et al. [32] based on some tricks in Luo et al. [42] and classical convex-smooth identities.

Lemma 7. *For problems in the class $\mathcal{P}^{H,B,\sigma}$ and for $\eta \leq \frac{2}{H}$, the iterates of Local SGD satisfy for all $t \in [0, T]$,*

$$C(t) = \mathbb{E}\left[\frac{1}{M}\sum_{m=1}^{M} \|\bar{x}_t - x_t^m\|_2^2\right] \leq 3(t - \delta(t)) \cdot \left[\eta^2 K V_{\max} + \eta^2 \sigma^2\right].$$

In particular, this implies $C_{\max} \leq 3\eta^2 K^2 V_{\max} + 3\eta^2 K \sigma^2$.

Proof. For $t = \delta(t)$, i.e., at the time steps where communication happens, the upper bound in the lemma is trivially true as both the left and right-hand sides are zero. For the rest of the proof, assume $t \neq \delta(t)$ and note the following,

$$\begin{aligned}
C(t) &= \mathbb{E}\left[\Xi_t\right], \\
&= \frac{1}{M}\sum_{m\in[M]} \mathbb{E}\left[\|x_t^m - \bar{x}_t\|_2^2\right], \\
&\overset{(i)}{\leq} \frac{1}{M}\sum_{m=1}^{M} \mathbb{E}\left[\|x_t^m - (\bar{x}_{t-1} - \eta\nabla F(\bar{x}_{t-1}))\|_2^2\right], \\
&= \frac{1}{M}\sum_{m=1}^{M} \mathbb{E}\left[\|x_{t-1}^m - \eta\nabla f_m(x_{t-1}^m; z_{t-1}^m) - (\bar{x}_{t-1} - \eta\nabla F(\bar{x}_{t-1}))\|_2^2\right], \\
&= \frac{1}{M}\sum_{m=1}^{M} \mathbb{E}\left[\|x_{t-1}^m \pm \nabla F_m(x_{t-1}^m) - \eta\nabla f_m(x_{t-1}^m; z_{t-1}^m) - (\bar{x}_{t-1} - \eta\nabla F(\bar{x}_{t-1}))\|_2^2\right], \\
&\overset{(ii)}{\leq} \frac{1}{M}\sum_{m=1}^{M} \mathbb{E}\left[\|x_{t-1}^m - \eta\nabla F_m(x_{t-1}^m) - (\bar{x}_{t-1} - \eta\nabla F(\bar{x}_{t-1}))\|_2^2\right] + \eta^2\sigma^2, \\
&= \frac{1}{M}\sum_{m=1}^{M} \mathbb{E}\left[\|x_{t-1}^m - \eta\nabla F_m(x_{t-1}^m) - (\bar{x}_{t-1} \pm \eta\nabla F_m(\bar{x}_{t-1}) - \eta\nabla F(\bar{x}_{t-1}))\|_2^2\right] + \eta^2\sigma^2, \\
&\overset{(iii)}{\leq} \frac{1}{M}\sum_{m=1}^{M}(1 + \frac{1}{K-1})\mathbb{E}\left[\|x_{t-1}^m - \eta\nabla F_m(x_{t-1}^m) - (\bar{x}_{t-1} - \eta\nabla F_m(\bar{x}_{t-1}))\|_2^2\right] \\
&\qquad + \frac{\eta^2}{M}\sum_{m=1}^{M}(1 + (K-1))\,\mathbb{E}\left[\|\nabla F(\bar{x}_{t-1}) - \nabla F_m(\bar{x}_{t-1})\|_2^2\right] + \eta^2\sigma^2,
\end{aligned}$$

$$\overset{(iv)}{\leq} \left(1+\frac{1}{K-1}\right)\cdot\frac{1}{M}\sum_{m=1}^{M}\mathbb{E}\left[\|x_{t-1}^m-\bar{x}_{t-1}\|_2^2\right]+\eta^2KV(t-1)+\eta^2\sigma^2,$$
$$\leq \left(1+\frac{1}{K-1}\right)\cdot C(t-1)+\eta^2KV_{\max}+\eta^2\sigma^2,$$

where (i) uses the variance trick of [42] mentioned in Equation (14); (ii) uses conditioning on $\mathcal{H}_{t-1}$, applying tower rule, noting the independence of noise across machines and then using the variance bound in Equation (3); (iii) uses modified AM-GM inequality, i.e., Lemma 6 with $\gamma=\frac{1}{K-1}$; (iv) uses the definition of $V(t-1)$ and Lemma 4. Now we unroll the above recursion up to the last communication round,

$$C(t)\leq\left(1+\frac{1}{K-1}\right)^{t-\delta(t)}\cdot C(\delta(t))+\left(\eta^2KV_{\max}+\eta^2\sigma^2\right)\cdot\sum_{k=0}^{t-\delta(t)-1}\left(1+\frac{1}{K-1}\right)^k,$$
$$\leq\left(\eta^2KV_{\max}+\eta^2\sigma^2\right)\cdot\sum_{k=0}^{t-\delta(t)-1}\left(1+\frac{1}{K-1}\right)^{K-1},$$
$$\leq\left(\eta^2KV_{\max}+\eta^2\sigma^2\right)\cdot\sum_{k=0}^{t-\delta(t)-1}e,$$
$$\leq 3(t-\delta(t))\left(\eta^2KV_{\max}+\eta^2\sigma^2\right).$$

This proves the first statement, and for the second statement, we use $t-\delta(t)\leq K$. □

D.1.2 A Refined Upper Bound for $V_{\max}$

We refine Lemma 2 in the following lemma and allow it to hold over $\mathcal{P}_{\zeta_\star,\tau}^{H,B,\sigma}$.

Lemma 8. *For all problems in the class $\mathcal{P}_{\zeta_\star,\tau}^{H,B,\sigma}$ we have,*

$$V_{\max}\leq 2\tau^2A_{\max}+2\zeta_\star^2.$$

Proof. Using the definition of $V(t)$, we have for all $t\in[0,T]$,

$$\begin{aligned}
V(t)&=\frac{1}{M}\sum_{m=1}^{M}\mathbb{E}\left[\|\nabla F_m(\bar{x}_t)-\nabla F(\bar{x}_t)\|_2^2\right],\\
&=\frac{1}{M}\sum_{m=1}^{M}\mathbb{E}\left[\|\nabla F_m(\bar{x}_t)-\nabla F(\bar{x}_t)\pm(\nabla F_m(x^\star)-\nabla F(x^\star))\|_2^2\right],\\
&\overset{(i)}{\leq}\frac{2}{M}\sum_{m=1}^{M}\mathbb{E}\left[\|\nabla F_m(\bar{x}_t)-\nabla F(\bar{x}_t)-(\nabla F_m(x^\star)-\nabla F(x^\star))\|_2^2\right]\\
&\qquad\qquad+\frac{2}{M}\sum_{m=1}^{M}\mathbb{E}\left[\|\nabla F_m(x^\star)-\nabla F(x^\star)\|_2^2\right],\\
&\overset{(ii)}{\leq}2\tau^2\mathbb{E}[\|\bar{x}_t-x^\star\|_2^2]+2\zeta_\star^2,\\
&\leq 2\tau^2A_{\max}+2\zeta_\star^2,
\end{aligned}\tag{17}$$

where (i) uses Lemma 6 with $\gamma=1$; (ii) uses Assumption 1 and Assumption 3. Since the bound holds for any t, then we have

$$V_{\max}=\max_{0\leq s\leq T}V(s)\leq 2\tau^2A_{\max}+2\zeta_\star^2,$$

which proves the lemma. □

D.1.3 A New Upper Bound for $A_{\max}$

Lemma 9. *For all problems in the problem class $\mathcal{P}^{H,B,\sigma}_{\zeta_\star,\tau}$ and with step-size $\eta \leq \min\left\{\frac{1}{2H}, \sqrt[3]{\frac{1}{24H\tau^2K^3R}}\right\}$ Local SGD's iterates are such that,*

$$A_{\max} \leq 2B^2 + \frac{6\eta^2\sigma^2 T}{M} + 24\eta^3 HK^3 R\zeta_\star^2 + 12\eta^3 HK^2 R\sigma^2.$$

Proof. Starting from a re-arrangement of Lemma 2 we get that for all $t \in [0, T-1]$,

$$\begin{aligned}
A(t+1) &\leq A(t) - \eta D(t) + \frac{3\eta^2\sigma^2}{M} + 2\eta H C(t), \\
&\overset{(i)}{\leq} A(t) + \frac{3\eta^2\sigma^2}{M} + 2\eta H C(t), \\
&\leq A(0) + (t+1)\cdot\frac{3\eta^2\sigma^2}{M} + 2\eta H\sum_{k=0}^{t} C(k), \\
&\overset{(ii)}{\leq} B^2 + (t+1)\cdot\left(\frac{3\eta^2\sigma^2}{M} + 2\eta H C_{\max}\right), \\
&\overset{(iii)}{\leq} B^2 + (t+1)\cdot\left(\frac{3\eta^2\sigma^2}{M} + 2\eta H\left(3\eta^2K^2V_{\max} + 3\eta^2K\sigma^2\right)\right), \\
&= B^2 + (t+1)\cdot\left(\frac{3\eta^2\sigma^2}{M} + 6\eta^3HK^2V_{\max} + 6\eta^3HK\sigma^2\right), \\
&\overset{(iv)}{\leq} B^2 + T\cdot\left(\frac{3\eta^2\sigma^2}{M} + 12\eta^3H\tau^2K^2A_{\max} + 12\eta^3HK^2\zeta_\star^2 + 6\eta^3HK\sigma^2\right),
\end{aligned}$$

where in (i) we use that the function-suboptimality is always non-negative, i.e., $D(t) \geq 0$; in (ii) we use that $\bar{x}_0 = 0$, $\|x^\star\|_2 \leq B$ and the definition of $C_{\max}$; in (iii) we use Lemma 7; in (iv) we use Lemma 8 and the fact that $t+1 \leq T$. Since the above bound holds for all $t \in [0, T-1]$ as well as trivially for $t = -1$ we can take the maximum over time to get,

$$\begin{aligned}
A_{\max} &= \max_{t\in[-1,T-1]} A(t+1), \\
&\leq B^2 + KR\cdot\left(\frac{3\eta^2\sigma^2}{M} + 12\eta^3H\tau^2K^2A_{\max} + 12\eta^3HK^2\zeta_\star^2 + 6\eta^3HK\sigma^2\right).
\end{aligned}$$

Re-arranging the above we get,

$$\begin{aligned}
\left(1 - 12\eta^3H\tau^2K^3R\right)A_{\max} &\leq B^2 + KR\cdot\left(\frac{3\eta^2\sigma^2}{M} + 12\eta^3HK^2\zeta_\star^2 + 6\eta^3HK\sigma^2\right), \\
\Rightarrow \frac{1}{2}A_{\max} &\overset{(i)}{\leq} B^2 + KR\cdot\left(\frac{3\eta^2\sigma^2}{M} + 12\eta^3HK^2\zeta_\star^2 + 6\eta^3HK\sigma^2\right), \\
\Rightarrow A_{\max} &\leq 2B^2 + \frac{6\eta^2\sigma^2T}{M} + 24\eta^3HK^3R\zeta_\star^2 + 12\eta^3HK^2R\sigma^2,
\end{aligned}$$

where in (i) we used that $12\eta^3H\tau^2K^3R \leq \frac{1}{2}$ which follows from our assumption in the lemma $\eta \leq \sqrt[3]{\frac{1}{24H\tau^2K^3R}}$. □

D.1.4 An Even More Refined Upper Bound for $C_{\max}$!

The observations so far allow us to prove the key technical upper-bound result of our paper, which complements Patel et al. [1] by providing a refined upper bound on the consensus error of Local SGD.

Lemma 10. *For all problems in the problem class $\mathcal{P}^{H,B,\sigma}_{\zeta_\star,\tau}$ and with step-size $\eta \leq \min\left\{\frac{1}{2H}, \sqrt[3]{\frac{1}{24H\tau^2K^3R}}\right\}$ Local SGD's iterates are such that,*

$$C_{\max} \leq \frac{B^2}{\eta HKR} + \frac{3\eta\sigma^2}{HM} + 18\eta^2K^2\zeta_\star^2 + 9\eta^2K\sigma^2.$$

Proof. This lemma follows trivially from combining Lemmas 7 to 9,

$$\begin{aligned}
C_{\max} &\overset{(i)}{\leq} 3\eta^2K^2V_{\max} + 3\eta^2K\sigma^2,\\
&\overset{(ii)}{\leq} 6\eta^2K^2\tau^2A_{\max} + 6\eta^2K^2\zeta_\star^2 + 3\eta^2K\sigma^2,\\
&\overset{(iii)}{\leq} \frac{A_{\max}}{2\eta HKR} + 6\eta^2K^2\zeta_\star^2 + 3\eta^2K\sigma^2,\\
&\overset{(iv)}{\leq} \frac{B^2}{\eta HKR} + \frac{3\eta\sigma^2}{HM} + 12\eta^2K^2\zeta_\star^2 + 6\eta^2K\sigma^2 + 6\eta^2K^2\zeta_\star^2 + 3\eta^2K\sigma^2,\\
&= \frac{B^2}{\eta HKR} + \frac{3\eta\sigma^2}{HM} + 18\eta^2K^2\zeta_\star^2 + 9\eta^2K\sigma^2,
\end{aligned}$$

where in (i) we use Lemma 7; in (ii) we use Lemma 8; in (iii) we use the assumption on the step-size that $\eta \leq \sqrt[3]{\frac{1}{24H\tau^2K^3R}}$ which implies, $6\eta^2K^2\tau^2 \leq \frac{1}{2\eta HKR}$; in (iv) we use Lemma 9.

This proves the lemma. □

D.2 Technical Lemmas for the Single-Machine Mimicking Proof

Lemma 11. *For all problems in the class $\mathcal{P}^{H,B,\sigma}_{\zeta_\star,\tau}$ we have,*

$$V_{\max} \leq 2\tau^2B_{\max} + 2\zeta_\star^2.$$

Proof. We continue the inequality $V(t) \leq 2\tau^2\mathbb{E}[\|\bar{x}_t - x^\star\|_2^2] + 2\zeta_\star^2$ obtained in Lemma 8 as follows

$$\begin{aligned}
V(t) &\overset{(i)}{\leq} 2\tau^2A(t) + 2\zeta_\star^2\\
&\overset{(ii)}{\leq} 2\tau^2B(t) + 2\zeta_\star^2\\
&\overset{(iii)}{\leq} 2\tau^2B_{\max} + 2\zeta_\star^2,
\end{aligned} \tag{18}$$

where (i) is obtained in Lemma 8, (ii) follows from $B(t) \geq A(t)$, and (iii) is by the definition of $B_{\max}$. Since the inequality is true for any $t \in [0, KR]$, then we have $V_{\max} \leq 2\tau^2B_{\max} + 2\zeta_\star^2$, which concludes the proof. □

Lemma 12. *For all problems in the class $\mathcal{P}^{H,B,\sigma}_{\zeta_\star,\tau}$ and $\eta \leq \frac{2}{H}$ we have,*

$$C_{\max} \leq 6\eta^2K^2\tau^2B_{\max} + 6\eta^2K^2\zeta_\star^2 + 3\eta^2K\sigma^2.$$

Proof. We need to use the statement of Lemma 11 in Lemma 7:

$$\begin{aligned}
C(t) &\overset{(i)}{\leq} 3\eta^2K^2V_{\max} + 3\eta^2K\sigma^2\\
&\overset{(ii)}{\leq} 6\eta^2K^2\tau^2B_{\max} + 6\eta^2K^2\zeta_\star^2 + 3\eta^2K\sigma^2,
\end{aligned}$$

where (i) is obtained in Lemma 7, (ii) uses Lemma 11. □

D.2.1 An Alternative One Step Descent Lemma Using $B(t)$ and $D(t)$

Lemma 13. *Consider problem in $\mathcal{P}^{H,B,\sigma}_{\zeta_\star,\tau}$, then for $\eta \le \frac{1}{4H}$, the Local SGD iterates satisfy for all $t \in [0, T-1]$,*

$$D(t) \le \frac{B(t) - B(t+1)}{\eta} + 2\eta V(t) + \eta\sigma^2 + 2\sqrt{V(t)C(t)}.$$

Moreover,

$$B(t+1) \le B(t) + 2\eta^2 V(t) + \eta^2\sigma^2 + 2\eta\sqrt{V(t)C(t)}.$$

Proof. Since each F_m is convex, then $R(t) \ge 0$ for all $t \in [0, T]$. Now we decompose the term $S(t) = \frac{1}{M}\sum_{m=1}^M \mathbb{E}[F_m(x_t^m) - F_m(x^\star)]$ as follows

$$\begin{aligned} S(t) &= \frac{1}{M}\sum_{m=1}^M \mathbb{E}[F_m(x_t^m) - F_m(\bar{x}_t) - \langle \nabla F_m(\bar{x}_t), x_t^m - \bar{x}_t\rangle] \\ &\quad + \frac{1}{M}\sum_{m=1}^M \mathbb{E}[F_m(\bar{x}_t) - F_m(x^\star) + \langle \nabla F_m(\bar{x}_t), x_t^m - \bar{x}_t\rangle] \\ &= R(t) + D(t) + \frac{1}{M}\sum_{m=1}^M \mathbb{E}[\langle \nabla F_m(\bar{x}_t), x_t^m - \bar{x}_t\rangle]. \end{aligned}$$

Since $\frac{1}{M}\sum_{m=1}^M \mathbb{E}[\langle \nabla F(\bar{x}_t), x_t^m - \bar{x}_t\rangle = 0$, then we continue as follows

$$S(t) = \frac{1}{M}\sum_{m=1}^M \mathbb{E}[F_m(x_t^m) - F_m(x^\star)] = R(t) + D(t) + Q(t). \tag{19}$$

By Cauchy-Schwarz inequality, we have

$$\begin{aligned} |Q(t)| &= \left| \frac{1}{M}\sum_{m=1}^M \mathbb{E}[\nabla F_m(\bar{x}_t) - \nabla F(\bar{x}_t), x_t^m - \bar{x}_t] \right| \\ &\le \frac{1}{M}\sum_{m=1}^M \mathbb{E}[\|\nabla F_m(\bar{x}_t) - \nabla F(\bar{x}_t)\|_2^2 \cdot \|x_t^m - \bar{x}_t\|_2^2] \\ &\le \left(\frac{1}{M}\sum_{m=1}^M \mathbb{E}[\|\nabla F_m(\bar{x}_t) - \nabla F(\bar{x}_t)\|_2^2] \right)^{1/2} \cdot \left(\frac{1}{M}\sum_{m=1}^M \mathbb{E}[\|x_t^m - \bar{x}_t\|_2^2] \right)^{1/2} \\ &= \sqrt{V(t)C(t)}. \end{aligned}$$

Fix any $m \in [M]$ and $t \in [0, T-1]$. Define the tentative local update

$$\tilde{x}_{t+1}^m \coloneqq x_t^m - \eta \nabla f(x_t^m; z_t^m).$$

As before, if $t+1 \ne \delta(t+1)$, then $\tilde{x}_{t+1}^m = x_{t+1}^m$, whereas if $t+1 = \delta(t+1)$, then $\tilde{x}_{t+1}^m$ is the pre-communication iterate.

We begin with the identity

$$\left\|\tilde{x}_{t+1}^m - x^\star\right\|_2^2 = \|x_t^m - x^\star\|_2^2 - 2\eta \left\langle x_t^m - x^\star, \nabla f(x_t^m; z_t^m)\right\rangle + \eta^2 \left\|\nabla f(x_t^m; z_t^m)\right\|_2^2. \tag{20}$$

Conditioning on $\mathcal{H}_t$ and using $x_t^m \in \mathcal{H}_t$, we have

$$\mathbb{E}[\nabla f(x_t^m; z_t^m) \mid \mathcal{H}_t] = \nabla F_m(x_t^m).$$

Thus

$$\mathbb{E}\left[\left\|\tilde{x}_{t+1}^m - x^\star\right\|_2^2 \mid \mathcal{H}_t\right] = \|x_t^m - x^\star\|_2^2 - 2\eta \left\langle x_t^m - x^\star, \nabla F_m(x_t^m)\right\rangle$$

$$
\begin{aligned}
&\quad + \eta^2 \mathbb{E}\left[\|\nabla f(x_t^m; z_t^m)\|_2^2 \mid \mathcal{H}_t\right] \\
&\le \|x_t^m - x^\star\|_2^2 - 2\eta(F_m(x_t^m) - F_m(x^\star)) \\
&\quad + \eta^2 \mathbb{E}\left[\|\nabla f(x_t^m; z_t^m)\|_2^2 \mid \mathcal{H}_t\right], \qquad (21)
\end{aligned}
$$

where the inequality follows from convexity. Rearranging gives

$$
\begin{aligned}
&F_m(x_t^m) - F_m(x^\star) \\
&\quad = \frac{1}{2\eta}\left(\|x_t^m - x^\star\|_2^2 - \mathbb{E}\left[\left\|\tilde{x}_{t+1}^m - x^\star\right\|_2^2 \mid \mathcal{H}_t\right]\right) + \frac{\eta}{2}\mathbb{E}\left[\|\nabla f(x_t^m; z_t^m)\|_2^2 \mid \mathcal{H}_t\right]. \qquad (22)
\end{aligned}
$$

We next upper bound the stochastic second moment. Since

$$
\nabla f(x_t^m; z_t^m) = \nabla F_m(x_t^m) + (\nabla f(x_t^m; z_t^m) - \nabla F_m(x_t^m)),
$$

expanding the square and conditioning on $\mathcal{H}_t$ gives

$$
\begin{aligned}
\mathbb{E}\left[\|\nabla f(x_t^m; z_t^m)\|_2^2 \mid \mathcal{H}_t\right] &= \|\nabla F_m(x_t^m)\|_2^2 + \mathbb{E}\left[\|\nabla f(x_t^m; z_t^m) - \nabla F_m(x_t^m)\|_2^2 \mid \mathcal{H}_t\right] \\
&\le \|\nabla F_m(x_t^m)\|_2^2 + \sigma^2,
\end{aligned}
$$

where the cross term vanishes because the stochastic gradient is unbiased.

Averaging (22) over $m \in [M]$ and taking full expectation, we obtain

$$
\begin{aligned}
S(t) \le \frac{1}{2\eta}\left(B(t) - \frac{1}{M}\sum_{m\in[M]} \mathbb{E}\left[\left\|\tilde{x}_{t+1}^m - x^\star\right\|_2^2\right]\right) + \frac{\eta}{2M}\sum_{m\in[M]} \mathbb{E}\left[\|\nabla F_m(x_t^m)\|_2^2\right] \\
+ \frac{\eta\sigma^2}{2}. \qquad (23)
\end{aligned}
$$

First, note that

$$
\frac{1}{M}\sum_{m\in[M]} \mathbb{E}\left[\left\|\tilde{x}_{t+1}^m - x^\star\right\|_2^2\right] \ge \mathbb{E}\left[\|\bar{x}_{t+1} - x^\star\|_2\right]^2,
$$

therefore from (23) we have

$$
S(t) \le \frac{1}{2\eta}(B(t) - B(t+1)) + \frac{\eta}{2M}\sum_{m\in[M]} \mathbb{E}\left[\|\nabla F_m(x_t^m)\|_2^2\right] + \frac{\eta\sigma^2}{2} \qquad (24)
$$

Second, we have by Young's inequality in Lemma 6 that

$$
\begin{aligned}
\frac{1}{M}\sum_{m\in[M]} \mathbb{E}\left[\|\nabla F_m(x_t^m)\|_2^2\right] &\overset{(i)}{\le} \frac{2}{M}\sum_{m\in[M]} \mathbb{E}\left[\|\nabla F_m(x_t^m) - \nabla F_m(\bar{x}_t)\|_2^2\right] \\
&\quad + \frac{2}{M}\sum_{m\in[M]} \mathbb{E}\left[\|\nabla F_m(\bar{x}_t)\|_2^2\right] \\
&\overset{(ii)}{\le} \frac{4H}{M}\sum_{m\in[M]} \mathbb{E}[F_m(x_t^m) - F_m(\bar{x}_t) - \langle \nabla F_m(\bar{x}_t), x_t^m - \bar{x}_t\rangle] \\
&\quad + \frac{2}{M}\sum_{m\in[M]} \mathbb{E}\left[\|\nabla F(\bar{x}_t)\|_2^2\right] \\
&\quad + \frac{2}{M}\sum_{m\in[M]} \mathbb{E}\left[\|\nabla F_m(\bar{x}_t) - \nabla F(\bar{x}_t)\|_2^2\right] \\
&\overset{(iii)}{\le} 4HR(t) + 4HD(t) + 2V(t), \qquad (25)
\end{aligned}
$$

where (i) follows from Lemma 6, (ii) follows from (15) and equality $\frac{1}{M}\sum_{m\in[M]}\|a_m\|_2^2 = \left\|\frac{1}{M}\sum_{m\in[M]}a_m\right\|_2^2 + \frac{1}{M}\sum_{m\in[M]}\left\|a_m - \frac{1}{M}\sum_{m\in[M]}a_m\right\|_2^2$ for any $\{a_m\}_{m\in[M]}$, (iii) from H-smoothness inequality $\|\nabla F(\bar{x}_t)\|_2^2 \leq 2H(F(\bar{x}_t) - F(x^\star))$ and definitions of $R(t), D(t), V(t)$. We use (25) and (19) in (24) and obtain

$$S(t) = R(t) + D(t) + Q(t) \leq \frac{1}{2\eta}(B(t) - B(t+1)) + \frac{\eta\sigma^2}{2} + 2\eta H(R(t) + D(t)) + \eta V(t). \quad (26)$$

This implies that

$$(1 - 2\eta H)(D(t) + R(t)) \leq \frac{B(t) - B(t+1)}{2\eta} + \eta V(t) + \frac{\eta\sigma^2}{2} - Q(t). \quad (27)$$

Using $\eta \leq \frac{1}{4H}$ and $|Q(t)| \leq \sqrt{V(t)C(t)}$, we obtain

$$D(t) + R(t) \leq \frac{B(t) - B(t+1)}{\eta} + 2\eta V(t) + \eta\sigma^2 + 2\sqrt{V(t)C(t)},$$

which proves the first statement of the lemma. At the same time, from (27) we have

$$\begin{aligned} B(t+1) &\leq B(t) + 2\eta^2 V(t) + \eta^2\sigma^2 - 2\eta(1 - 2\eta H)(D(t) + R(t)) - 2\eta Q(t) \\ &\leq B(t) + 2\eta^2 V(t) + \eta^2\sigma^2 + 2\eta\sqrt{V(t)C(t)}, \end{aligned}$$

which proves the second statement of the lemma. □

Lemma 14. *Let*

$$\eta = \left\{\frac{1}{4H}, \frac{1}{10\tau K\sqrt{R}}, \frac{B}{10\zeta_\star K\sqrt{R}}, \frac{B}{10\sigma\sqrt{KR}}\right\}.$$

Then,

$$\begin{aligned} &B_{\max} \leq 2B^2, \quad V_{\max} \leq 4\tau^2 B^2 + 2\zeta_\star^2 \leq 4(\tau^2 B^2 + \zeta_\star^2), \\ &C_{\max} \leq 12\eta^2 K^2\tau^2 B^2 + 6\eta^2 K^2\zeta_\star^2 + 3\eta^2 K\sigma^2. \end{aligned}$$

Proof. Let $U^2 := \tau^2 B_{\max} + \zeta_\star^2$. From Lemma 11, we have $V(t) \leq 2U^2$, and from Lemma 12 we have $C_{\max} \leq 6\eta^2 K^2 U^2 + 3\eta^2 K\sigma^2$. This implies that

$$\sqrt{C_{\max}} \leq \sqrt{6}\eta K U + \sqrt{3}\eta\sqrt{K}\sigma.$$

From Lemma 13 we obtain

$$\begin{aligned} B(t+1) &\leq B(t) + 2\eta^2 V(t) + \eta^2\sigma^2 + 2\eta\sqrt{V(t)C(t)} \\ &\leq B(t) + 4\eta^2 U^2 + \eta^2\sigma^2 + 2\eta\sqrt{2}U \cdot \sqrt{C_{\max}} \\ &\leq B(t) + 4\eta^2 U^2 + \eta^2\sigma^2 + 2\sqrt{2}\eta U(\sqrt{6}\eta K U + \sqrt{3}\eta\sqrt{K}\sigma) \\ &= B(t) + 4\eta^2 U^2 + \eta^2\sigma^2 + 4\sqrt{3}\eta^2 K U^2 + 2\sqrt{6}\eta^2 U\sqrt{K}\sigma \\ &\leq B(t) + 4\eta^2 U^2 + \eta^2\sigma^2 + 4\sqrt{3}\eta^2 K U^2 + \sqrt{6}\eta^2 U^2 K + \sqrt{6}\eta^2\sigma^2 \\ &= B(t) + (4 + 4\sqrt{3}K + \sqrt{6}K)U^2\eta^2 + \eta^2\sigma^2(1 + \sqrt{6}) \\ &\leq B(t) + 14K\eta^2 U^2 + 4\eta^2\sigma^2. \end{aligned} \quad (28)$$

Using the definition of U^2, we then have

$$B(t+1) \leq B(t) + 14\eta^2 K\tau^2 B_{\max} + 14\eta^2 K\zeta_\star^2 + 4\eta^2\sigma^2.$$

We sum the inequality above for all iterations between 0 and $t-1$ and obatin

$$\begin{aligned} B(t) &\leq B(0) + 14\eta^2 Kt\tau^2 B_{\max} + 14\eta^2 Kt\zeta_\star^2 + 4\eta^2\sigma^2 t \\ &\leq B(0) + 14\eta^2 KT\tau^2 B_{\max} + 14\eta^2 KT\zeta_\star^2 + 4\eta^2\sigma^2 T \\ &\leq B^2 + 14\eta^2 K^2 R\tau^2 B_{\max} + 14\eta^2 K^2 R\zeta_\star^2 + 4\eta^2\sigma^2 KR, \end{aligned} \quad (29)$$

where we use $B(0) \leq B^2$. Using the choice of the step size η, we get

$$14\eta^2 K^2 R\tau^2 \leq 14K^2 R\tau^2 \cdot \frac{1}{100\tau^2 K^2 R} = \frac{14}{100},$$
$$14\eta^2 K^2 R\zeta_\star^2 \leq 14K^2 R\zeta_\star^2 \cdot \frac{B^2}{100\zeta_\star^2 K^2 R} = \frac{14B^2}{100},$$
$$4\eta^2 KR\sigma^2 \leq 4KR\sigma^2 \cdot \frac{B^2}{100\sigma^2 KR} = \frac{4B^2}{100}.$$

From (29) we then obtain

$$B_{\max} \leq B^2 + \frac{14B_{\max}}{100} + \frac{18B^2}{100} \Rightarrow B_{\max} \leq \frac{118}{86}B^2 < 2B^2.$$

□

E Proofs of Our Upper Bounds for $\mathcal{P}^{H,B,\sigma}_{\zeta_\star,\tau}$

In this section, we will prove our main upper bounds.

The first upper bound we prove combines the canonical one-step result in Lemma 2, with our refined consensus error upper bound in Lemma 10 followed by carefully tuning the step-size.

Theorem 5 (An Upper Bound Based on Consensus Error; Full Statement of Theorem 1)**.** *For problems in the class $\mathcal{P}^{H,B,\sigma}_{\zeta_\star,\tau}$, assuming the step-size is chosen as follows,*

$$\eta = \min\left\{\frac{1}{2H}, \sqrt[3]{\frac{1}{24H\tau^2K^3R}}, \sqrt{\frac{B^2M}{3\sigma^2KR}}, \sqrt[3]{\frac{B^2}{12H\zeta_\star^2K^3R}}, \sqrt[3]{\frac{B^2}{6H\sigma^2K^2R}}\right\},$$

Local SGD generates the output $\hat{x}^{\text{Local-SGD}} = \frac{1}{TM}\sum_{t\in[0,T-1],m\in[M]} x_t^m$ which satisfies,

$$\mathbb{E}\left[F(\hat{x}^{\text{Local-SGD}}) - F(x^\star)\right] \leq 18 \cdot \left(\frac{HB^2}{KR} + \frac{(H\tau^2)^{1/3}B^2}{R^{2/3}} + \frac{\sigma B}{\sqrt{MKR}} + \frac{(H\zeta_\star^2 B^4)^{1/3}}{R^{2/3}} + \frac{(H\sigma^2B^4)^{1/3}}{K^{1/3}R^{2/3}}\right).$$

The second upper bound we prove follows the observation of Woodworth et al. [18] that Local SGD is never worse than single-machine SGD in the homogeneous setting, extending it over the problem class $\mathcal{P}^{H,B,\sigma}_{\zeta_\star,\tau}$ using the tools we developed in Appendix D.

Theorem 6 (An Upper Bound Based on Comparison to Single Machine SGD; Full Statement of Theorem 2)**.** *For problems in the class $\mathcal{P}^{H,B,\sigma}_{\zeta_\star,\tau}$, assuming the step-size is chosen as follows,*

$$\eta = \min\left\{\frac{1}{4H}, \frac{1}{10\tau K\sqrt{R}}, \frac{B}{10\zeta_\star K\sqrt{R}}, \frac{B}{10\sigma\sqrt{KR}}\right\},$$

Local SGD generates the output $\hat{x}^{\text{Local-SGD}} = \frac{1}{TM}\sum_{t\in[0,T-1],m\in[M]} x_t^m$ which satisfies,

$$\mathbb{E}[F(\hat{x}) - F(x^\star)] \leq 13 \cdot \left(\frac{HB^2}{KR} + \frac{\sigma B}{\sqrt{KR}} + \frac{\zeta_\star B}{\sqrt{R}} + \frac{\tau B^2}{\sqrt{R}}\right).$$

E.1 Proof of Theorem 5

Proof. We begin by re-stating the one-step recursion from Lemma 2 for all $t \in [0, T-1]$

$$D(t) \leq \frac{A(t) - A(t+1)}{\eta} + \frac{3\eta\sigma^2}{M} + 2HC(t),$$
$$\leq \frac{A(t) - A(t+1)}{\eta} + \frac{3\eta\sigma^2}{M} + 2HC_{\max}$$

Averaging this over time steps $t \in [0, T-1]$ and applying Jensen's inequality we get,

$$\begin{aligned}
\mathbb{E}\left[F(\hat{x}^{\text{Local-SGD}}) - F(x^\star)\right] &\le \frac{1}{T}\sum_{t=0}^{T-1} D(t), \\
&\le \frac{A(0) - A(T)}{\eta T} + \frac{3\eta\sigma^2}{M} + 2HC_{\max}, \\
&\overset{(i)}{\le} \frac{B^2}{\eta T} + \frac{3\eta\sigma^2}{M} + \frac{2B^2}{\eta T} + \frac{6\eta\sigma^2}{M} + 36\eta^2 HK^2\zeta_\star^2 + 18\eta^2 HK\sigma^2, \\
&\le \frac{3B^2}{\eta T} + \frac{9\eta\sigma^2}{M} + 36\eta^2 HK^2\zeta_\star^2 + 18\eta^2 HK\sigma^2,
\end{aligned}$$

where in (i) we use Lemma 10. Now to get the final bound we will balance the second, third, and fourth term above with the first-term by setting the step-size η. In particular, combining this with the constraints on the step-size from Lemma 10 we get,

$$\eta = \min\left\{\frac{1}{2H}, \sqrt[3]{\frac{1}{24H\tau^2K^3R}}, \sqrt{\frac{B^2M}{3\sigma^2KR}}, \sqrt[3]{\frac{B^2}{12H\zeta_\star^2K^3R}}, \sqrt[3]{\frac{B^2}{6H\sigma^2K^2R}}\right\}.$$

With this choice of step-size we get,

$$\begin{aligned}
&\mathbb{E}\left[F(\hat{x}^{\text{Local-SGD}}) - F(x^\star)\right] \\
&\quad\le \frac{6HB^2}{KR} + \frac{6\sqrt[3]{24}(H\tau^2)^{1/3}B^2}{R^{2/3}} + \frac{6\sigma B}{\sqrt{MKR}} + \frac{6\sqrt[3]{12}(H\zeta_\star^2B^4)^{1/3}}{R^{2/3}} + \frac{6\sqrt[3]{6}(H\sigma^2B^4)^{1/3}}{K^{1/3}R^{2/3}}, \\
&\quad\le \frac{6HB^2}{KR} + \frac{18(H\tau^2)^{1/3}B^2}{R^{2/3}} + \frac{6\sigma B}{\sqrt{MKR}} + \frac{18(H\zeta_\star^2B^4)^{1/3}}{R^{2/3}} + \frac{12(H\sigma^2B^4)^{1/3}}{K^{1/3}R^{2/3}}, \\
&\quad\le 18\cdot\left(\frac{HB^2}{KR} + \frac{(H\tau^2)^{1/3}B^2}{R^{2/3}} + \frac{\sigma B}{\sqrt{MKR}} + \frac{(H\zeta_\star^2B^4)^{1/3}}{R^{2/3}} + \frac{(H\sigma^2B^4)^{1/3}}{K^{1/3}R^{2/3}}\right),
\end{aligned}$$

which proves the theorem. □

E.2 Proof of Theorem 6

Proof. Let $U_0^2 = \tau^2B^2 + \zeta_\star^2$. Then from Lemma 14 we have

$$V_{\max} \le 4U_0^2, \quad C_{\max} \le 12\eta^2K^2\tau^2B^2 + 6\eta^2K^2\zeta_\star^2 + 3\eta^2K\sigma^2 \le 12\eta^2K^2U_0^2 + 3\eta^2K\sigma^2.$$

This implies that $\sqrt{C_{\max}} \le \sqrt{12}\eta KU_0 + \sqrt{3}\eta\sqrt{K}\sigma \le 4\eta KU_0 + 2\eta\sqrt{K}\sigma$. From Lemma 13 we then obtain

$$\begin{aligned}
D(t) &\le \frac{B(t) - B(t+1)}{\eta} + \eta\sigma^2 + 2\eta V(t) + 2\sqrt{V(t)C(t)} \\
&\le \frac{B(t) - B(t+1)}{\eta} + \eta\sigma^2 + 8\eta U_0^2 + 2\cdot 2U_0(4\eta KU_0 + 2\eta\sqrt{K}\sigma) \\
&\le \frac{B(t) - B(t+1)}{\eta} + 5\eta\sigma^2 + 28\eta KU_0^2 \\
&= \frac{B(t) - B(t+1)}{\eta} + 5\eta\sigma^2 + 28\eta(\tau^2B^2 + \zeta_\star^2).
\end{aligned}$$

Averaging this inequality over all iterations $t \in [0, T-1]$, we obtain

$$\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1} D(t) &\le \frac{B^2}{\eta T} + 5\eta\sigma^2 + 28\eta K(\tau^2B^2 + \zeta_\star^2) \\
&= \frac{B^2}{\eta KR} + 5\eta\sigma^2 + 28\eta K(\tau^2B^2 + \zeta_\star^2).
\end{aligned} \tag{30}$$

From the choice of the step size, we have

$$\frac{1}{\eta} \leq 4H + 10\tau K\sqrt{R} + \frac{10\zeta_\star K\sqrt{R}}{B} + \frac{10\sigma\sqrt{KR}}{B}.$$

Hence,

$$\frac{B^2}{\eta KR} \leq \frac{4HB^2}{KR} + \frac{10\tau B^2}{\sqrt{R}} + \frac{10\zeta_\star B}{\sqrt{R}} + \frac{10\sigma B}{\sqrt{KR}}. \tag{31}$$

Moreover, we have

$$28\eta K\tau^2 B^2 \leq 28K\tau^2 B^2 \cdot \frac{1}{10\tau K\sqrt{R}} = 2.8\frac{\tau B^2}{\sqrt{R}}, \tag{32}$$

and

$$28\eta K\zeta_\star^2 \leq 28K\zeta_\star^2 \cdot \frac{B}{10\zeta_\star K\sqrt{R}} = 2.8\frac{\zeta_\star B}{\sqrt{R}}. \tag{33}$$

Finally, we also have

$$5\eta\sigma^2 \leq 5\sigma^2 \frac{B}{10\sigma\sqrt{KR}} \leq 0.5\frac{\sigma B}{\sqrt{KR}}. \tag{34}$$

Combining (31), (32), (33), and (34) in (30), we obtain

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1} D(t) &\leq \frac{4HB^2}{KR} + \frac{12.8\tau B^2}{\sqrt{R}} + \frac{12.8\zeta_\star B}{\sqrt{R}} + \frac{10.5\sigma B}{\sqrt{KR}} \\ &\leq 13 \cdot \left(\frac{HB^2}{KR} + \frac{\tau B^2}{\sqrt{R}} + \frac{\zeta_\star B}{\sqrt{R}} + \frac{\sigma B}{\sqrt{KR}}\right).\end{aligned}$$

It remains to use Jensen's inequality to obtain the bound for $\hat{x}^{\text{Local-SGD}}$. □

F A Single-Machine-Style Upper Bound for $\mathcal{P}_\zeta^{H,B,\sigma}$

In this section, we give a second upper bound for Local SGD under the uniformly bounded first-order heterogeneity condition in Assumption 2. The guarantee is complementary to the bound of Woodworth et al. [32] reported in Table 1. Rather than comparing the ghost iterate to dense mini-batch SGD, we use the average client distance $B(t)$ as the potential, as in a single-machine SGD analysis, while directly controlling the global function sub-optimality $D(t)$. A local Bregman-divergence term then absorbs the part of the stochastic gradient second moment caused by evaluating the client gradients at different local iterates.

We begin by stating the following resulting upper bound.

Theorem 7 (A Single-Machine-Style Upper Bound under Uniform First-Order Heterogeneity)**.** *For problems in the class $\mathcal{P}_\zeta^{H,B,\sigma}$, assume the step-size is chosen as follows,*

$$\eta = \min\left\{\frac{1}{4H}, \frac{B}{\sigma\sqrt{KR}}, \frac{B}{\zeta K\sqrt{R}}\right\},$$

Local SGD generates the output

$$\hat{x}^{\textit{Local-SGD}} = \frac{1}{TM}\sum_{t\in[0,T-1],\, m\in[M]} x_t^m = \frac{1}{T}\sum_{t=0}^{T-1}\bar{x}_t$$

which satisfies

$$\mathbb{E}\left[F\left(\hat{x}^{\textit{Local-SGD}}\right) - F(x^\star)\right] \leq 9 \cdot \left(\frac{HB^2}{KR} + \frac{\sigma B}{\sqrt{KR}} + \frac{\zeta B}{\sqrt{R}}\right).$$

F.1 A One-Step Recursion using $B(t)$ and $D(t)$

The following lemma is the main new ingredient. It is stated for the base class $\mathcal{P}^{H,B,\sigma}$ in terms of the trajectory-dependent gradient heterogeneity $V(t)$. The uniform bound from Assumption 2 will only be used afterward.

Lemma 15. *For problems in the class $\mathcal{P}^{H,B,\sigma}$ and with $\eta \le \frac{1}{4H}$, the iterates of Local SGD satisfy, for all $t \in [0, T-1]$,*

$$D(t) \le \frac{B(t) - B(t+1)}{\eta} + \eta\sigma^2 + 2\eta V(t) + 2\sqrt{V(t)C(t)}.$$

Proof. Fix any $t \in [0, T-1]$. For brevity, define

$$\Delta_m(x) \coloneqq \nabla F_m(x) - \nabla F(x),$$

and recall that we denote the average local Bregman divergence between x_t^m and $\bar{x}_t$ by

$$R(t) = \frac{1}{M} \sum_{m\in[M]} \mathbb{E}\left[F_m(x_t^m) - F_m(\bar{x}_t) - \langle \nabla F_m(\bar{x}_t), x_t^m - \bar{x}_t\rangle\right] \ge 0,$$

where the inequality follows from the convexity of each F_m. Expanding every $F_m(x_t^m)$ around the common point $\bar{x}_t$ gives

$$\begin{aligned}\frac{1}{M} \sum_{m\in[M]} &\mathbb{E}\left[F_m(x_t^m) - F_m(x^\star)\right] \\ &= D(t) + \mathcal{R}(t) + \frac{1}{M} \sum_{m\in[M]} \mathbb{E}\left[\langle \Delta_m(\bar{x}_t), x_t^m - \bar{x}_t\rangle\right]. \end{aligned} \tag{35}$$

Indeed, the contribution of the global gradient vanishes because

$$\frac{1}{M} \sum_{m\in[M]} \langle \nabla F(\bar{x}_t), x_t^m - \bar{x}_t\rangle = 0.$$

Moreover, Cauchy–Schwarz on the product space consisting of the random trajectory and a uniformly sampled machine gives

$$\left|\frac{1}{M} \sum_{m\in[M]} \mathbb{E}\left[\langle \Delta_m(\bar{x}_t), x_t^m - \bar{x}_t\rangle\right]\right| \le \sqrt{V(t)C(t)}. \tag{36}$$

We next relate the left-hand side of (35) to the potential $B(t)$. For every $m \in [M]$, define the tentative pre-synchronization update

$$\widetilde{x}_{t+1}^m \coloneqq x_t^m - \eta \nabla f(x_t^m; z_t^m).$$

Conditioning on $\mathcal{H}_t$, expanding the squared distance to $x^\star$, and using the unbiasedness condition in Equation (3), we obtain

$$\begin{aligned}\mathbb{E}&\left[\left\|\widetilde{x}_{t+1}^m - x^\star\right\|_2^2 \mid \mathcal{H}_t\right] \\ &= \|x_t^m - x^\star\|_2^2 - 2\eta \left\langle x_t^m - x^\star, \nabla F_m(x_t^m)\right\rangle + \eta^2 \mathbb{E}\left[\|\nabla f(x_t^m; z_t^m)\|_2^2 \mid \mathcal{H}_t\right].\end{aligned}$$

By convexity of F_m,

$$F_m(x_t^m) - F_m(x^\star) \le \left\langle x_t^m - x^\star, \nabla F_m(x_t^m)\right\rangle.$$

Therefore, averaging over $m \in [M]$ and taking full expectation yields

$$\frac{1}{M} \sum_{m\in[M]} \mathbb{E}\left[F_m(x_t^m) - F_m(x^\star)\right]$$

$$\leq \frac{1}{2\eta}\left(B(t) - \frac{1}{M}\sum_{m\in[M]}\mathbb{E}\left[\left\|\widetilde{x}_{t+1}^m - x^\star\right\|_2^2\right]\right) + \frac{\eta}{2M}\sum_{m\in[M]}\mathbb{E}\left[\left\|\nabla f(x_t^m; z_t^m)\right\|_2^2\right]. \quad (37)$$

The stochastic-gradient variance bound gives

$$\frac{1}{M}\sum_{m\in[M]}\mathbb{E}\left[\left\|\nabla f(x_t^m; z_t^m)\right\|_2^2\right] \leq \frac{1}{M}\sum_{m\in[M]}\mathbb{E}\left[\left\|\nabla F_m(x_t^m)\right\|_2^2\right] + \sigma^2. \quad (38)$$

It remains to control the deterministic-gradient second moment while retaining the non-negative term $R(t)$. We have

$$\begin{aligned}
&\frac{1}{M}\sum_{m\in[M]}\mathbb{E}\left[\left\|\nabla F_m(x_t^m)\right\|_2^2\right] \\
&\quad\leq \frac{2}{M}\sum_{m\in[M]}\mathbb{E}\left[\left\|\nabla F_m(x_t^m) - \nabla F_m(\bar{x}_t)\right\|_2^2\right] + \frac{2}{M}\sum_{m\in[M]}\mathbb{E}\left[\left\|\nabla F_m(\bar{x}_t)\right\|_2^2\right] \\
&\quad\overset{(i)}{\leq} 4HR(t) + 2\mathbb{E}\left[\left\|\nabla F(\bar{x}_t)\right\|_2^2\right] + 2V(t) \\
&\quad\overset{(ii)}{\leq} 4HR(t) + 4HD(t) + 2V(t). \quad (39)
\end{aligned}$$

Here, (i) uses the one-sided co-coercivity inequality (15) and the variance identity

$$\frac{1}{M}\sum_{m\in[M]}\left\|\nabla F_m(\bar{x}_t)\right\|_2^2 = \left\|\nabla F(\bar{x}_t)\right\|_2^2 + \frac{1}{M}\sum_{m\in[M]}\left\|\nabla F_m(\bar{x}_t) - \nabla F(\bar{x}_t)\right\|_2^2.$$

For (ii), note that F is also convex and H-smooth, and $\nabla F(x^\star) = 0$. Applying the same one-sided co-coercivity inequality as in (15) to F, with $x = \bar{x}_t$ and $y = x^\star$, gives

$$\mathbb{E}\left[\left\|\nabla F(\bar{x}_t)\right\|_2^2\right] \leq 2HD(t).$$

Substituting (38) and (39) into (37), and using the fact that synchronization can only decrease the average squared distance to $x^\star$, gives

$$\frac{1}{M}\sum_{m\in[M]}\mathbb{E}\left[F_m(x_t^m) - F_m(x^\star)\right] \leq \frac{B(t) - B(t+1)}{2\eta} + 2H\eta\left(D(t) + R(t)\right) \quad (40)$$

$$+ \eta V(t) + \frac{\eta\sigma^2}{2}. \quad (41)$$

For completeness, at a communication time Jensen's inequality gives

$$B(t+1) \leq \frac{1}{M}\sum_{m\in[M]}\mathbb{E}\left[\left\|\widetilde{x}_{t+1}^m - x^\star\right\|_2^2\right],$$

and away from communication times the two sides are equal.

Combining (35) and (41), and then using (36), yields

$$(1 - 2H\eta)\left(D(t) + R(t)\right) \leq \frac{B(t) - B(t+1)}{2\eta} + \eta V(t) + \frac{\eta\sigma^2}{2} + \sqrt{V(t)C(t)}.$$

Since $R(t) \geq 0$ and $\eta \leq \frac{1}{4H}$, we have $1 - 2H\eta \geq \frac{1}{2}$. Consequently,

$$D(t) \leq \frac{B(t) - B(t+1)}{\eta} + 2\eta V(t) + \eta\sigma^2 + 2\sqrt{V(t)C(t)},$$

which proves the lemma. □

F.2 Proof of Theorem 7

Proof. Under Assumption 2, the definition of $V(t)$ immediately gives, for every $t \in [0, T]$,

$$V(t) = \frac{1}{M} \sum_{m \in [M]} \mathbb{E}\left[\|\nabla F_m(\bar{x}_t) - \nabla F(\bar{x}_t)\|_2^2\right] \le \zeta^2.$$

In particular, $V_{\max} \le \zeta^2$. Since the step-size in the theorem satisfies $\eta \le \frac{1}{4H} \le \frac{2}{H}$, Lemma 7 implies

$$C_{\max} \le 3\eta^2 K^2 \zeta^2 + 3\eta^2 K \sigma^2. \tag{42}$$

Consequently,

$$\sqrt{C_{\max}} \le \sqrt{3}\eta K \zeta + \sqrt{3}\eta\sqrt{K}\sigma.$$

Applying Lemma 15, and using $V(t) \le \zeta^2$ and $C(t) \le C_{\max}$, gives

$$\begin{aligned} D(t) &\le \frac{B(t) - B(t+1)}{\eta} + \eta\sigma^2 + 2\eta\zeta^2 + 2\zeta\sqrt{C_{\max}} \\ &\le \frac{B(t) - B(t+1)}{\eta} + \eta\sigma^2 + 2\eta\zeta^2 + 2\sqrt{3}\eta K\zeta^2 + 2\sqrt{3}\eta\sqrt{K}\zeta\sigma. \end{aligned}$$

Using the A.M.-G.M. inequality,

$$2\sqrt{3}\eta\sqrt{K}\zeta\sigma \le \sqrt{3}\eta\left(K\zeta^2 + \sigma^2\right).$$

Since $K \ge 1$, the preceding bounds imply

$$D(t) \le \frac{B(t) - B(t+1)}{\eta} + 8\eta\left(K\zeta^2 + \sigma^2\right). \tag{43}$$

Summing (43) over $t \in [0, T-1]$, telescoping the potential, and using $B(0) = \|x^\star\|_2^2 \le B^2$ and $B(T) \ge 0$, we obtain

$$\frac{1}{T}\sum_{t=0}^{T-1} D(t) \le \frac{B^2}{\eta T} + 8\eta\left(K\zeta^2 + \sigma^2\right). \tag{44}$$

By convexity of F,

$$\mathbb{E}\left[F\left(\hat{x}^{\text{Local-SGD}}\right) - F(x^\star)\right] \le \frac{1}{T}\sum_{t=0}^{T-1} D(t).$$

It remains to substitute the step-size. Since $T = KR$,

$$\begin{aligned} \frac{B^2}{\eta T} &\le 4\frac{HB^2}{KR} + \frac{\sigma B}{\sqrt{KR}} + \frac{\zeta B}{\sqrt{R}}, \\ 8\eta\sigma^2 &\le 8\frac{\sigma B}{\sqrt{KR}}, \\ 8\eta K\zeta^2 &\le 8\frac{\zeta B}{\sqrt{R}}. \end{aligned}$$

Combining these inequalities with (44) proves

$$\mathbb{E}\left[F\left(\hat{x}^{\text{Local-SGD}}\right) - F(x^\star)\right] \le 4\frac{HB^2}{KR} + 9\frac{\sigma B}{\sqrt{KR}} + 9\frac{\zeta B}{\sqrt{R}},$$

and hence the claimed result. □

The bound in Theorem 7 is complementary to the result of Woodworth et al. [32] in Table 1. In particular, the heterogeneity contribution $\zeta B/\sqrt{R}$ is no larger than $(H\zeta^2 B^4)^{1/3}/R^{2/3}$ whenever $\zeta \le HB/\sqrt{R}$. Similarly, $\sigma B/\sqrt{KR}$ is no larger than $(H\sigma^2 B^4)^{1/3}/(K^{1/3}R^{2/3})$ whenever $\sigma \le HB\sqrt{K/R}$. Moreover, in the limit $\zeta = 0$, the theorem recovers the single-machine SGD term $\sigma B/\sqrt{KR}$ without requiring any second-order heterogeneity assumption.

G Proofs of Our Lower Bounds

Key Simplifications. In this section, we will prove the lower bound mentioned in Theorem 3. We will make some standard simplifying assumptions in the proof, which do not affect the lower bound beyond numerical constants:

- Our constructions are all noiseless, i.e., $\sigma = 0$, as we do not need to rely on the stochastic gradient noise to prove any of the lower bounds.
- We will consider the $\tau = 0$ and $\zeta_\star = 0$ settings separately, i.e., we will provide two distinct constructions for these settings. This is without loss of generality, because both of our constructions lie in the class $\mathcal{P}^{H,B,0}_{\zeta_\star,\tau}$ and can be combined by putting the hard functions on disjoint coordinates. Since both constructions isolate a single form of data heterogeneity, the second-order construction does not constrain the choice of $\zeta_\star$, and vice versa. The reason is that the second-order construction ensures a shared optimizer across clients, so it does not constrain the choice of $\zeta_\star$ in the first-order construction for the other coordinates. Similarly, the first-order construction ensures that the clients have the same second-order derivative wherever it exists, which means it does not restrict τ in any way for the second-order construction. Combining the constructions leads to the desired lower bound. This idea of using disjoint coordinates is also how we combine the lower bound in Theorem 3 with the homogeneous lower bound of Glasgow et al. [36] to get the final lower bound in Table 1. For more context on using disjoint co-ordinates to prove distributed optimization lower bounds, see Woodworth [35] and Patel [41].
- As the reader will notice, the constructions in Theorems 8 and 9 use only $M = 2$, i.e., two machines. This is without loss of generality. If there is an even number of machines, we can duplicate these two machines and generate the same lower bound. If there is an odd number of machines, we have to be a bit more careful, and we show how to handle this case in Section G.4, where we also extend our lower bounds to the averaged iterate.

All of the above simplifications are usual in this literature [18, 32, 36, 12, 39, 26, 1]. With these caveats in mind, we are ready to state our lower bounds.

G.1 A Sharper Lower Bound When $\zeta_\star = 0$

In this section, we prove a lower bound that refines the basic $\Omega(\tau B^2/R)$ construction due to Patel et al. [1]. The key point is to separate the hard block's curvature scale from the instance's actual second-order heterogeneity.

Theorem 8. *There exist two deterministic convex quadratic client objectives $F_1, F_2 : \mathbb{R}^2 \to \mathbb{R}$ in the problem class $\mathcal{P}^{H,B,0}_{0,\tau}$, such that for Local SGD (see update (5)) initialized at $x_0 = 0$, with any number of local steps $K \geq 1$, any step-size $\eta > 0$, the final iterate $\bar{x}_R$ after R communication rounds satisfies*

$$F(\bar{x}_R) - F(x^\star) \geq c \cdot \min\left\{\frac{HB^2}{R}, \frac{\tau B^2}{\sqrt{R}}\right\},$$

where $c > 0$ is a universal numerical constant.

Proof. Let

$$\kappa \coloneqq 12R, \qquad \text{and} \qquad \alpha \coloneqq \frac{\kappa - 1}{\kappa + 1}.$$

Then

$$\frac{1+\alpha}{1-\alpha} = \kappa.$$

Define for some $\alpha \in [0, 1]$ to be determined later,

$$e_1 \coloneqq (1, 0)^\top, \qquad \text{and} \qquad v \coloneqq \left(\alpha, \sqrt{1-\alpha^2}\right)^\top.$$

These vectors give the following rank-one matrices,

$$\hat{A}_1 \coloneqq e_1 e_1^\top, \qquad \text{and} \qquad \hat{A}_2 \coloneqq vv^\top.$$

We define the hard-block curvature scale as follows

$$\lambda \coloneqq \min\{H, \sqrt{3R}\,\tau\}$$

and use it to define the actual Hessians on both our clients as

$$A_1 \coloneqq \lambda \hat{A}_1, \qquad \text{and} \qquad A_2 \coloneqq \lambda \hat{A}_2.$$

Specifically, our two client objectives are

$$F_m(x) \coloneqq \frac{1}{2}(x - x^\star)^\top A_m (x - x^\star), \qquad \text{for } m \in \{1, 2\},$$

where $x^\star$ will be chosen below.

Since $\|\hat{A}_m\|_2 = 1$, we have

$$\|A_m\|_2 = \lambda \le H.$$

Thus, both the clients are convex and H-smooth.

We next compute the second-order heterogeneity. Since

$$\hat{A}_1 - \hat{A}_2 = \begin{bmatrix} 1-\alpha^2 & -\alpha\sqrt{1-\alpha^2} \\ -\alpha\sqrt{1-\alpha^2} & -(1-\alpha^2) \end{bmatrix},$$

the eigenvalues of $\hat{A}_1 - \hat{A}_2$ are

$$\pm\sqrt{1-\alpha^2}.$$

Therefore

$$\|A_1 - A_2\|_2 = \lambda\sqrt{1-\alpha^2}.$$

Now

$$1 - \alpha^2 = (1-\alpha)(1+\alpha) \le \frac{4}{\kappa},$$

so

$$\sqrt{1-\alpha^2} \le \frac{2}{\sqrt{\kappa}} = \frac{1}{\sqrt{3R}}.$$

By the choice of λ,

$$\|A_1 - A_2\| \le \frac{\lambda}{\sqrt{3R}} \le \tau.$$

Thus, the second-order heterogeneity (the stronger Assumption 3) is at most τ.

Moreover, the average Hessian is

$$A \coloneqq \frac{A_1 + A_2}{2} = \frac{\lambda}{2}(\hat{A}_1 + \hat{A}_2).$$

The two nonzero eigenvalues of A are

$$\Lambda_{\max} = \frac{\lambda}{2}(1+\alpha), \qquad \text{and} \qquad \Lambda_{\min} = \frac{\lambda}{2}(1-\alpha).$$

Hence

$$\frac{\Lambda_{\max}}{\Lambda_{\min}} = \frac{1+\alpha}{1-\alpha} = \kappa = 12R.$$

Let $u_{\max}$ and $u_{\min}$ be unit eigenvectors of A associated with $\Lambda_{\max}$ and $\Lambda_{\min}$, respectively. Set

$$x^\star \coloneqq -\frac{B}{\sqrt{2}}(u_{\max} + u_{\min}).$$

Then

$$\|x^\star\|_2 = B.$$

We now analyze the Local SGD dynamics in error coordinates, where $r \in [0, T]$

$$\tilde{x}_t \coloneqq \bar{x}_t - x^\star.$$

Thus $\tilde{x}_{qK}$ denotes the synchronized iterate after q communication rounds.

For client m, starting from the synchronized iterate $\tilde{x}_{qK}$, the next K local gradient steps give

$$\tilde{x}^m_{qK+K} = (I - \eta A_m)^K \tilde{x}_{qK}.$$

Since

$$A_m = \lambda p_m p_m^\top, \qquad p_1 = e_1, \qquad p_2 = v,$$

is rank one and satisfies $p_m^\top p_m = 1$, we have

$$A_m^2 = \lambda A_m.$$

Therefore,

$$(I - \eta A_m)^K = I - \frac{1 - (1 - \eta\lambda)^K}{\lambda} A_m.$$

Define the effective local step-size

$$\tilde{\eta} \coloneqq \frac{1 - (1 - \eta\lambda)^K}{\lambda}.$$

Then

$$\tilde{x}^m_{qK+K} = (I - \tilde{\eta} A_m)\tilde{x}_{qK}.$$

Since we take the standard outer step-size $\beta = 1$, the server simply averages the two client endpoints:

$$\tilde{x}_{qK+K} = \frac{1}{2}\left(\tilde{x}^1_{qK+K} + \tilde{x}^2_{qK+K}\right).$$

Substituting the local-update formula gives

$$\begin{aligned}\tilde{x}_{qK+K} &= \frac{1}{2}\left[(I - \tilde{\eta}A_1)\tilde{x}_{qK} + (I - \tilde{\eta}A_2)\tilde{x}_{qK}\right] \\ &= \left(I - \tilde{\eta}\frac{A_1 + A_2}{2}\right)\tilde{x}_{qK}.\end{aligned}$$

Let

$$A \coloneqq \frac{A_1 + A_2}{2}.$$

Then the communication-round dynamics is exactly

$$\tilde{x}_{qK+K} = (I - \tilde{\eta}A)\tilde{x}_{qK}.$$

Iterating over $R = T/K$ communication rounds,

$$\tilde{x}_T = (I - \tilde{\eta}A)^R \tilde{x}_0.$$

Thus Local SGD on this hard instance is exactly gradient descent for R iterations on the average quadratic with Hessian A, using the effective step-size $\tilde{\eta}$.

We now lower bound this gradient descent trajectory. Recall that A has eigenvalues

$$\Lambda_{\max} = \frac{\lambda}{2}(1 + \alpha), \qquad \Lambda_{\min} = \frac{\lambda}{2}(1 - \alpha),$$

and condition number

$$\kappa \coloneqq \frac{\Lambda_{\max}}{\Lambda_{\min}} = \frac{1 + \alpha}{1 - \alpha} = 12R.$$

Thus, the eigenvalues are

$$\boxed{\Lambda_{\max} = \lambda \cdot \frac{12R}{12R + 1}, \qquad \text{and} \qquad \Lambda_{\min} = \frac{\lambda}{12R + 1}.}$$

Let $u_{\max}$ and $u_{\min}$ be corresponding unit eigenvectors. We choose

$$x^\star \coloneqq -\frac{B}{\sqrt{2}}(u_{\max} + u_{\min}),$$

so that $\|x^\star\| = B$. Since the algorithm starts from $\bar{x}_0 = 0$, we have

$$\tilde{x}_0 = -x^\star = \frac{B}{\sqrt{2}}(u_{\max} + u_{\min}).$$

Hence

$$\tilde{x}_T = \frac{B}{\sqrt{2}}(1 - \tilde{\eta}\Lambda_{\max})^R u_{\max} + \frac{B}{\sqrt{2}}(1 - \tilde{\eta}\Lambda_{\min})^R u_{\min}.$$

If

$$\tilde{\eta} \geq \frac{3}{\Lambda_{\max}},$$

then

$$|1 - \tilde{\eta}\Lambda_{\max}| \geq 2,$$

and so the high-curvature component gives

$$\begin{aligned}
F(\bar{x}_T) - F(x^\star) &= \frac{1}{2}\tilde{x}_T^\top A \tilde{x}_T, \\
&\geq \frac{1}{2}\Lambda_{\max}\left(\frac{B}{\sqrt{2}}|1 - \tilde{\eta}\Lambda_{\max}|^R\right)^2, \\
&\geq \frac{\Lambda_{\max} B^2 2^R}{4}, \\
&= \frac{3R^2 2^R}{12R+1} \cdot \frac{\lambda B^2}{R}, \\
&\overset{(R\geq 1)}{\geq} \frac{1}{4} \cdot \frac{\lambda B^2}{R}.
\end{aligned}$$

It remains to now consider the more interesting case

$$\tilde{\eta} < \frac{3}{\Lambda_{\max}}.$$

Then

$$\tilde{\eta}\Lambda_{\min} \leq \frac{3\Lambda_{\min}}{\Lambda_{\max}} = \frac{3}{\kappa}.$$

Therefore

$$|1 - \tilde{\eta}\Lambda_{\min}| \geq 1 - \frac{3}{\kappa}.$$

The low-curvature component gives

$$\begin{aligned}
F(\bar{x}_T) - F(x^\star) &\geq \frac{1}{2}\Lambda_{\min}\left(\frac{B}{\sqrt{2}}\left(1 - \frac{3}{\kappa}\right)^R\right)^2, \\
&= \frac{\Lambda_{\min} B^2}{4}\left(1 - \frac{3}{\kappa}\right)^{2R}.
\end{aligned}$$

Since $\kappa = 12R$,

$$\left(1 - \frac{3}{\kappa}\right)^{2R} = \left(1 - \frac{1}{4R}\right)^{2R} \overset{(R\geq 1)}{\geq} \frac{9}{16}.$$

Hence

$$\begin{aligned}
F(\bar{x}_T) - F(x^\star) &\geq \frac{9R}{64(12R+1)} \cdot \frac{\lambda B^2}{R}, \\
&\overset{(R\geq 1)}{\geq} \frac{3}{256} \cdot \frac{\lambda B^2}{R}.
\end{aligned}$$

Combining the two cases,

$$F(\bar{x}_T) - F(x^\star) \geq \frac{3}{256} \cdot \frac{\lambda B^2}{R},$$

Finally, since

$$\lambda \coloneqq \min\{H, \sqrt{3R}\tau\},$$

we have

$$\frac{\lambda B^2}{R} = \Theta\left(\min\left\{\frac{HB^2}{R}, \frac{\tau B^2}{\sqrt{R}}\right\}\right).$$

This proves the claim. □

G.2 A Sharper Lower Bound When $\tau = 0$

We first isolate three elementary lemmas that will be used in the construction.

Lemma 16 (Active-coordinate lower bound)**.** *Fix $L > 0$, $\zeta_\star > 0$, $R \geq 1$, an integer $K \geq 2$, and recall that $T = KR$. Define*

$$g(s) = \begin{cases} \frac{L}{4}s^2, & s < 0, \\ \frac{L}{2}s^2, & s \geq 0, \end{cases} \qquad G_1(s) = g(s) - \zeta_\star s, \qquad G_2(s) = g(s) + \zeta_\star s.$$

Run vanilla Local SGD on G_1, G_2 with $M = 2$, step-size $\eta \leq 1/L$, and initialization

$$s_0^1 = s_0^2 = 0.$$

Let

$$\bar{s}_t \coloneqq \frac{s_t^1 + s_t^2}{2}.$$

Then

$$g(\bar{s}_T) - g(0) \gtrsim \min\left\{\frac{\zeta_\star^2}{L},\ LK^2\eta^2\zeta_\star^2 \min\{(RK\eta L)^2, 1\}\right\}.$$

Proof. We compare the above dynamics to a simpler quadratic system. Define

$$\widetilde{G}_1(s) \coloneqq \frac{L}{2}s^2 - \zeta_\star s, \qquad \widetilde{G}_2(s) \coloneqq \frac{L}{4}s^2 + \zeta_\star s.$$

Let $\widetilde{s}_t^1, \widetilde{s}_t^2$ denote vanilla Local SGD on $\widetilde{G}_1, \widetilde{G}_2$, with the same K, η and initialization

$$\widetilde{s}_0^1 = \widetilde{s}_0^2 = 0.$$

Let

$$\widetilde{\bar{s}}_t \coloneqq \frac{\widetilde{s}_t^1 + \widetilde{s}_t^2}{2}.$$

We first show that, at all communication times,

$$\bar{s}_{rK} \leq \widetilde{\bar{s}}_{rK} \leq 0, \qquad r = 0, 1, \ldots, R. \tag{45}$$

For the true dynamics, the one-step maps are

$$T_1(s) \coloneqq s - \eta(G_1'(s)) = s - \eta(g'(s) - \zeta_\star),$$

and

$$T_2(s) \coloneqq s - \eta(G_2'(s)) = s - \eta(g'(s) + \zeta_\star).$$

For the comparison dynamics, the one-step maps are

$$\widetilde{T}_1(s) \coloneqq s - \eta(\widetilde{G}_1'(s)) = (1 - \eta L)s + \eta\zeta_\star,$$

and

$$\widetilde{T}_2(s) \coloneqq s - \eta(\widetilde{G}_2'(s)) = \left(1 - \frac{\eta L}{2}\right)s - \eta\zeta_\star.$$

Since $\eta \leq 1/L$, all four maps are nondecreasing.

We now compare them pointwise. For client 1, if $s < 0$, then

$$T_1(s) = \left(1 - \frac{\eta L}{2}\right)s + \eta\zeta_\star \leq (1 - \eta L)s + \eta\zeta_\star = \widetilde{T}_1(s),$$

because $s < 0$. If $s \geq 0$, then $T_1(s) = \widetilde{T}_1(s)$. Hence

$$T_1(s) \leq \widetilde{T}_1(s) \qquad \forall s \in \mathbb{R}.$$

For client 2, if $s < 0$, then $T_2(s) = \widetilde{T}_2(s)$. If $s \geq 0$, then

$$T_2(s) = (1 - \eta L)s - \eta\zeta_\star \leq \left(1 - \frac{\eta L}{2}\right) s - \eta\zeta_\star = \widetilde{T}_2(s).$$

Therefore

$$T_2(s) \leq \widetilde{T}_2(s) \qquad \forall s \in \mathbb{R}.$$

We next prove the comparison at communication times by induction. At $t = 0$, the two systems agree. Suppose

$$\bar{s}_{rK} \leq \widetilde{\bar{s}}_{rK}.$$

At time rK, all clients are synchronized, so

$$s^1_{rK} = s^2_{rK} = \bar{s}_{rK}, \qquad \widetilde{s}^1_{rK} = \widetilde{s}^2_{rK} = \widetilde{\bar{s}}_{rK}.$$

Because T_m and $\widetilde{T}_m$ are nondecreasing and $T_m \leq \widetilde{T}_m$ pointwise, induction over the K local steps gives

$$s^m_{rK+k} \leq \widetilde{s}^m_{rK+k}, \qquad m \in \{1, 2\}, \quad k = 0, 1, \ldots, K.$$

Averaging at the communication step gives

$$\bar{s}_{(r+1)K} = \frac{s^1_{(r+1)K} + s^2_{(r+1)K}}{2} \leq \frac{\widetilde{s}^1_{(r+1)K} + \widetilde{s}^2_{(r+1)K}}{2} = \widetilde{\bar{s}}_{(r+1)K}.$$

This proves the first inequality (45) by induction.

We now compute the comparison recursion. Let

$$a \coloneqq 1 - \frac{\eta L}{2}, \qquad b \coloneqq 1 - \eta L.$$

Then $a, b \in [0, 1]$. Starting from a synchronized value s, comparison client 1 satisfies

$$\widetilde{s}^1_{k+1} = b\, \widetilde{s}^1_k + \eta\zeta_\star.$$

Thus after K local steps,

$$\widetilde{s}^1_K = b^K s + \eta\zeta_\star \sum_{j=0}^{K-1} b^j = b^K s + \frac{\zeta_\star}{L}(1 - b^K).$$

Similarly, comparison client 2 satisfies

$$\widetilde{s}^2_{k+1} = a\, \widetilde{s}^2_k - \eta\zeta_\star,$$

and hence

$$\widetilde{s}^2_K = a^K s - \eta\zeta_\star \sum_{j=0}^{K-1} a^j = a^K s - \frac{2\zeta_\star}{L}(1 - a^K).$$

Therefore the synchronized comparison iterate obeys

$$\widetilde{\bar{s}}_{(r+1)K} = A_K \widetilde{\bar{s}}_{rK} - \Delta_K,$$

where

$$A_K \coloneqq \frac{a^K + b^K}{2}, \qquad \Delta_K \coloneqq \frac{\zeta_\star}{2L}\left(1 + b^K - 2a^K\right).$$

Since $\widetilde{\bar{s}}_0 = 0$, iterating the affine recursion gives

$$\widetilde{\bar{s}}_{rK} = -\Delta_K \sum_{j=0}^{r-1} A_K^j \qquad \forall r \in [0, R].$$

We also have $A_K \in [0,1]$ because $a, b \in [0,1]$. Moreover,

$$\Delta_K = \frac{\zeta_\star}{2L}\left(1 + b^K - 2a^K\right) \geq 0.$$

To see this, define

$$\phi(x) \coloneqq 1 + (1-x)^K - 2(1-x/2)^K, \qquad x \in [0,1].$$

Then

$$1 + b^K - 2a^K = \phi(\eta L).$$

Since

$$\phi(0) = 0, \qquad \phi'(x) = K\left[(1-x/2)^{K-1} - (1-x)^{K-1}\right] \geq 0,$$

we have $\phi(x) \geq 0$ for all $x \in [0,1]$. Hence $\Delta_K \geq 0$.

Therefore

$$\widetilde{\bar{s}}_T = -\Delta_K \sum_{r=0}^{R-1} A_K^r \leq 0.$$

By the comparison inequality at communication times, already proved above,

$$\bar{s}_T \leq \widetilde{\bar{s}}_T \leq 0.$$

It remains to lower bound the right-hand side. Define

$$\theta \coloneqq K\eta L.$$

We first show

$$1 + b^K - 2a^K \gtrsim \min\{\theta^2, 1\}.$$

Let

$$f(x) \coloneqq (1-x)^K.$$

Then

$$1 + b^K - 2a^K = f(0) + f(\eta L) - 2f(\eta L/2).$$

Case 1. If $\boldsymbol{\theta \leq 1}$, then $\eta L \leq 1/K$. Since

$$f''(x) = K(K-1)(1-x)^{K-2},$$

for all $x \in [0, \eta L]$ we have

$$f''(x) \geq K(K-1)\left(1 - \frac{1}{K}\right)^{K-2} \gtrsim K^2,$$

where the last inequality holds for all $K \geq 2$. Now, let $h \coloneqq \eta L$ and $m \coloneqq \inf_{x \in [0,h]} f''(x)$. Since f is m-strongly convex on $[0,h]$,

$$f(0) + f(h) - 2f(h/2) \geq \frac{mh^2}{4}.$$

Using the above lower bound on m we get,

$$f(0) + f(\eta L) - 2f(\eta L/2) \gtrsim K^2\eta^2 L^2 = \theta^2.$$

Case 2. If $\boldsymbol{\theta \geq 1}$, define

$$\phi(x) \coloneqq 1 + (1-x)^K - 2(1-x/2)^K.$$

Then

$$\phi'(x) = K\left[(1-x/2)^{K-1} - (1-x)^{K-1}\right] \geq 0 \qquad x \in [0,1].$$

Since $\eta L \geq 1/K$,

$$1 + b^K - 2a^K = \phi(\eta L) \geq \phi(1/K) \gtrsim 1,$$

where the last inequality follows from the previous case applied at $\theta = 1$.

Therefore, combining the two cases above, we get,

$$\boxed{\Delta_K \gtrsim \frac{\zeta_\star}{L} \min\{\theta^2, 1\}.}$$

Next, we lower bound the geometric sum. Since $A_K \in [0, 1]$,

$$\sum_{r=0}^{R-1} A_K^r = 1 + A_K + \cdots + A_K^{R-1}.$$

We next lower bound the geometric sum. Recall that

$$A_K = \frac{a^K + b^K}{2}, \qquad a = 1 - \frac{\eta L}{2}, \qquad b = 1 - \eta L.$$

Since $\eta \le 1/L$, we have $a, b \in [0, 1]$, and therefore $A_K \in [0, 1]$.

We claim that

$$\sum_{r=0}^{R-1} A_K^r \gtrsim \begin{cases} \min\{R, 1/\theta\}, & \theta \le 1, \\ 1, & \theta \ge 1. \end{cases}$$

First suppose $\theta \ge 1$. Since the first term of the sum is $A_K^0 = 1$, we immediately have

$$\sum_{r=0}^{R-1} A_K^r \ge 1.$$

Now suppose $\theta \le 1$. We first upper bound $1 - A_K$. Since

$$1 - A_K = 1 - \frac{a^K + b^K}{2} = \frac{1 - a^K}{2} + \frac{1 - b^K}{2},$$

and since $1 - x^K \le K(1 - x)$ for every $x \in [0, 1]$, which follows from

$$1 - x^K = (1 - x)(1 + x + \cdots + x^{K-1}) \le K(1 - x),$$

we get

$$1 - A_K \le \frac{K(1 - a)}{2} + \frac{K(1 - b)}{2}.$$

Substituting $1 - a = \eta L/2$ and $1 - b = \eta L$ gives

$$1 - A_K \le \frac{K\eta L}{4} + \frac{K\eta L}{2} = \frac{3}{4} K\eta L = \frac{3}{4}\theta \le \theta.$$

We now use this to show that the first $\asymp \min\{R, 1/\theta\}$ terms of the geometric sum are bounded below by a constant. If $\theta = 0$, then $A_K = 1$ and the sum is exactly R, so the claim is trivial. Hence assume $\theta > 0$.

Let

$$N := \min\left\{R, \left\lfloor \frac{1}{2\theta} \right\rfloor\right\}.$$

If $N = 0$, then $\theta > 1/2$, and since $\theta \le 1$, the desired lower bound is only a universal constant:

$$\min\{R, 1/\theta\} \le 2.$$

In that case,

$$\sum_{r=0}^{R-1} A_K^r \ge 1 \gtrsim \min\{R, 1/\theta\}.$$

So assume $N \ge 1$.

For every $0 \le r \le N - 1$, Bernoulli's inequality gives

$$A_K^r = (1 - (1 - A_K))^r \ge 1 - r(1 - A_K).$$

Using $1 - A_K \le \theta$ and $r \le N - 1 \le 1/(2\theta)$, we get

$$A_K^r \ge 1 - r\theta \ge \frac{1}{2}.$$

Therefore

$$\sum_{r=0}^{R-1} A_K^r \ge \sum_{r=0}^{N-1} A_K^r \ge \frac{N}{2}.$$

Finally, N is within a universal constant of $\min\{R, 1/\theta\}$, except in the already handled case $N = 0$. More explicitly, when $\min\{R, 1/(2\theta)\} \ge 1$, we have

$$\left\lfloor \frac{1}{2\theta} \right\rfloor \ge \frac{1}{4\theta},$$

and hence

$$N = \min\left\{R, \left\lfloor \frac{1}{2\theta} \right\rfloor\right\} \gtrsim \min\{R, 1/\theta\}.$$

Thus

$$\sum_{r=0}^{R-1} A_K^r \gtrsim \min\{R, 1/\theta\}.$$

Combining the drift and geometric-sum bounds gives

$$-\widetilde{\bar{s}}_T = \Delta_K \sum_{r=0}^{R-1} A_K^r.$$

We analyze this expression by splitting into four regimes depending on $\theta = K\eta L$ and $R\theta$.

Case A: $\theta \le 1$ and $R\theta \le 1$. In this regime,

$$\Delta_K \gtrsim \frac{\zeta_\star}{L}\theta^2, \qquad \sum_{r=0}^{R-1} A_K^r \gtrsim R.$$

Therefore

$$-\widetilde{\bar{s}}_T \gtrsim \frac{\zeta_\star}{L} R\theta^2.$$

Case B: $\theta \le 1$ and $R\theta \ge 1$. In this regime,

$$\Delta_K \gtrsim \frac{\zeta_\star}{L}\theta^2, \qquad \sum_{r=0}^{R-1} A_K^r \gtrsim \frac{1}{\theta}.$$

Thus

$$-\widetilde{\bar{s}}_T \gtrsim \frac{\zeta_\star}{L}\,\theta.$$

Case C: $\theta \ge 1$. Here we use the crude bounds

$$\Delta_K \gtrsim \frac{\zeta_\star}{L}, \qquad \sum_{r=0}^{R-1} A_K^r \ge 1,$$

which yield

$$-\widetilde{\bar{s}}_T \gtrsim \frac{\zeta_\star}{L}.$$

Conclusion. Combining the three regimes, we obtain

$$-\widetilde{\bar{s}}_T \gtrsim \frac{\zeta_\star}{L} \min\{1, \theta, R\theta^2\}.$$

Since $\bar{s}_T \le \widetilde{\bar{s}}_T \le 0$,

$$|\bar{s}_T| \ge |\widetilde{\bar{s}}_T| \gtrsim \frac{\zeta_\star}{L} \min\{1, \theta, R\theta^2\}.$$

Finally, because $\bar{s}_T \le 0$,

$$g(\bar{s}_T) - g(0) = \frac{L}{4}\bar{s}_T^2.$$

Therefore

$$g(\bar{s}_T) - g(0) \gtrsim \frac{\zeta_\star^2}{L} \min\{1, \theta^2, R^2\theta^4\}.$$

Using $\theta = K\eta L$, this is exactly

$$g(\bar{s}_T) - g(0) \gtrsim \min\left\{\frac{\zeta_\star^2}{L},\ LK^2\eta^2\zeta_\star^2 \min\{(RK\eta L)^2, 1\}\right\}.$$

□

Lemma 17 (Auxiliary-coordinate lower bound)**.** *Fix $\mu > 0$ and define*

$$q(u) = \frac{\mu}{2}\left(u + \frac{B}{3}\right)^2.$$

Suppose Local SGD is run on this coordinate with identical client objectives, initialized at $u_0^1 = u_0^2 = 0$. If

$$\eta \le \frac{1}{2\mu KR},$$

then at time $T = KR$,

$$q(\bar{u}_T) - \min_u q(u) \gtrsim \mu B^2.$$

Proof. Since both clients have the same deterministic quadratic objective and start from the same point, their iterates agree at all times. Let

$$e_t \coloneqq u_t + \frac{B}{3}.$$

Then

$$e_{t+1} = (1 - \eta\mu)e_t, \qquad e_0 = \frac{B}{3}.$$

The assumption $\eta \le 1/(2\mu KR)$ gives

$$0 \le \eta\mu \le \frac{1}{2KR}.$$

Hence

$$|e_T| = (1 - \eta\mu)^{KR}\frac{B}{3} \ge \left(1 - \frac{1}{2KR}\right)^{KR}\frac{B}{3} \gtrsim B.$$

Therefore

$$q(\bar{u}_T) - \min_u q(u) = \frac{\mu}{2}e_T^2 \gtrsim \mu B^2.$$

□

Lemma 18 (Guard-coordinate lower bound)**.** *Define*

$$h(v) = \frac{H}{2}\left(v + \frac{B}{3}\right)^2.$$

Suppose Local SGD is run on this coordinate with identical client objectives, initialized at $v_0^1 = v_0^2 = 0$. If

$$\eta > \frac{2}{H},$$

then at time $T = KR$,

$$h(\bar{v}_T) - \min_v h(v) \gtrsim HB^2.$$

Proof. Since both clients have the same deterministic quadratic objective and start from the same point, their iterates agree at all times. Let

$$w_t := v_t + \frac{B}{3}.$$

Then

$$w_{t+1} = (1 - \eta H) w_t, \qquad w_0 = \frac{B}{3}.$$

If $\eta > 2/H$, then $|1 - \eta H| > 1$. Hence

$$|w_T| \geq |w_0| = \frac{B}{3}.$$

Therefore

$$h(\bar{v}_T) - \min_v h(v) = \frac{H}{2} w_T^2 \gtrsim H B^2.$$

□

Theorem 9 ($\tau = 0$ lower bound for Local SGD)**.** *There exist two deterministic convex client objectives $F_1, F_2 : \mathbb{R}^3 \to \mathbb{R}$ in the class $\mathcal{P}^{H,B,0}_{\zeta_\star,0,0}$, such that vanilla Local SGD with $M = 2$ clients, $K \geq 2$ local steps per round, arbitrary step-size $\eta > 0$ and initialized at $x_0^1 = x_0^2 = 0$, must have its final averaged iterate*

$$\bar{x}_T := \frac{1}{2}(x_T^1 + x_T^2)$$

such that it satisfies

$$F(\bar{x}_T) - F(x^\star) \gtrsim \min\left\{ H B^2,\ \frac{\zeta_\star^2}{H},\ \frac{(H \zeta_\star^2 B^4)^{1/3}}{R^{2/3}} \right\}.$$

Proof. Set

$$L := \frac{H}{2}.$$

We construct a deterministic instance on $\mathbb{R}^3$, writing

$$x = (s, u, v).$$

Use the active objectives G_1, G_2 from Lemma 16 with this value of L. Let $\mu > 0$ be chosen later, and define

$$q(u) = \frac{\mu}{2}\left(u + \frac{B}{3}\right)^2, \qquad h(v) = \frac{H}{2}\left(v + \frac{B}{3}\right)^2.$$

Define the two client objectives

$$F_1(s, u, v) = G_1(s) + q(u) + h(v),$$

and

$$F_2(s, u, v) = G_2(s) + q(u) + h(v).$$

Then

$$F(s, u, v) := \frac{F_1(s, u, v) + F_2(s, u, v)}{2} = g(s) + q(u) + h(v).$$

The minimizer is

$$x^\star = \left(0, -\frac{B}{3}, -\frac{B}{3}\right),$$

so

$$\|x^\star\|_2 = \frac{\sqrt{2} B}{3} \leq B.$$

The initialization is

$$x_0^1 = x_0^2 = (0, 0, 0).$$

The instance is deterministic, so $\sigma = 0$. The smoothness of the active coordinate is $L = H/2$, the smoothness of the auxiliary coordinate is μ, and the smoothness of the guard coordinate is H. We

will choose $\mu \le H$, so each client objective is H-smooth, convex, and satisfies the bounded optima assumption in (4).

At $x^\star$,

$$\nabla F_1(x^\star) = (-\zeta_\star, 0, 0), \qquad \nabla F_2(x^\star) = (\zeta_\star, 0, 0).$$

Therefore

$$\frac{1}{2}\sum_{m=1}^{2} \|\nabla F_m(x^\star)\|_2^2 = \zeta_\star^2.$$

Moreover, the two clients differ only by the linear terms $\mp\zeta_\star s$. Hence, for all x, y,

$$\nabla F_1(x) - \nabla F_1(y) = \nabla F_2(x) - \nabla F_2(y) = \nabla F(x) - \nabla F(y).$$

Thus Assumptions 3 holds with $\tau = 0$.

Now coming back to the choice of μ, define

$$\mathcal{T} \coloneqq \min\left\{HB^2,\ \frac{\zeta_\star^2}{H},\ \frac{(H\zeta_\star^2 B^4)^{1/3}}{R^{2/3}}\right\}, \qquad \mu \coloneqq \frac{\mathcal{T}}{4B^2}.$$

Since $\mathcal{T} \le HB^2$, we have

$$\mu \le \frac{H}{4} \le H.$$

Thus, all of this shows that the construction lies in $\mathcal{P}^{H,B,0}_{\zeta_\star,0,0}$.

We now split into three step-size regimes.

Small step-sizes. If

$$\eta \le \frac{1}{2\mu KR},$$

then Lemma 17 gives

$$F(\bar{x}_T) - F(x^\star) \ge q(\bar{u}_T) - \min_u q(u) \gtrsim \mu B^2 = \frac{\mathcal{T}}{4} \gtrsim \mathcal{T}.$$

Large step-sizes. If

$$\eta > \frac{2}{H},$$

then Lemma 18 gives

$$F(\bar{x}_T) - F(x^\star) \ge h(\bar{v}_T) - \min_v h(v) \gtrsim HB^2 \ge \mathcal{T}.$$

Intermediate step-sizes. It remains to consider

$$\frac{1}{2\mu KR} < \eta \le \frac{2}{H}.$$

Since $L = H/2$, this is

$$\frac{1}{2\mu KR} < \eta \le \frac{1}{L}.$$

Let

$$\eta_0 \coloneqq \frac{1}{2\mu KR}.$$

The active-coordinate lower bound in Lemma 16 is nondecreasing in η on $[0, 1/L]$. Since $\eta > \eta_0$, it suffices to evaluate the bound at η_0.

We have

$$RK\eta_0 L = \frac{L}{2\mu}.$$

Because $\mu \le H/4 = L/2$, we have

$$RK\eta_0 L \ge 1.$$

Therefore Lemma 16 gives

$$g(\bar{s}_T) - g(0) \gtrsim \min\left\{\frac{\zeta_\star^2}{L},\ LK^2\eta_0^2\zeta_\star^2\right\}.$$

Substituting $\eta_0 = 1/(2\mu KR)$,

$$LK^2\eta_0^2\zeta_\star^2 = \frac{L\zeta_\star^2}{4\mu^2R^2}.$$

Thus

$$g(\bar{s}_T) - g(0) \gtrsim \min\left\{\frac{\zeta_\star^2}{L},\ \frac{L\zeta_\star^2}{\mu^2R^2}\right\}.$$

The first term satisfies

$$\frac{\zeta_\star^2}{L} = \frac{2\zeta_\star^2}{H} \geq 2\mathcal{T}.$$

For the second term, using $\mu = \mathcal{T}/(4B^2)$ and $L = H/2$,

$$\frac{L\zeta_\star^2}{\mu^2R^2} = \frac{16L\zeta_\star^2B^4}{\mathcal{T}^2R^2} = \frac{8H\zeta_\star^2B^4}{\mathcal{T}^2R^2}.$$

Since

$$\mathcal{T} \leq \frac{(H\zeta_\star^2B^4)^{1/3}}{R^{2/3}},$$

we have

$$\mathcal{T}^3 \leq \frac{H\zeta_\star^2B^4}{R^2}.$$

Therefore

$$\frac{8H\zeta_\star^2B^4}{\mathcal{T}^2R^2} \geq 8\mathcal{T}.$$

Thus the intermediate regime also gives

$$F(\bar{x}_T) - F(x^\star) \geq g(\bar{s}_T) - g(0) \gtrsim \mathcal{T}.$$

Combining the three regimes gives

$$F(\bar{x}_T) - F(x^\star) \gtrsim \mathcal{T},$$

which proves the theorem. □

G.3 A New Lower Bound for $\mathcal{P}_\zeta^{H,B,\sigma}$

Although Theorem 3 is stated for the class $\mathcal{P}_{\zeta_\star,\tau}^{H,B,\sigma}$, it also implies a lower bound for the more restrictive uniformly bounded first-order heterogeneity class $\mathcal{P}_\zeta^{H,B,\sigma}$. The key observation is the inclusion

$$\mathcal{P}_{\zeta_\star=\zeta,\tau=0}^{H,B,\sigma} \subseteq \mathcal{P}_\zeta^{H,B,\sigma},$$

which follows from the calculation in Proposition 2.

Therefore, specializing the hard instance in Theorem 3 to $\tau = 0$ and $\zeta_\star = \zeta$ gives the following corollary.

Corollary 1. *If $K, M \geq 2$, then there exists a problem in the class $\mathcal{P}_\zeta^{H,B,\sigma}$ such that, for any constant step-size $\eta > 0$, the final Local SGD iterate $\bar{x}_{KR}$ satisfies*

$$\mathbb{E}[F(\bar{x}_{KR}) - F(x^\star)] \gtrsim \min\left\{HB^2,\ \frac{\zeta^2}{H},\ \frac{(H\zeta^2B^4)^{1/3}}{R^{2/3}}\right\}.$$

To our knowledge, Corollary 1 is the first lower bound for Local SGD that makes the dependence on the averaged uniform heterogeneity parameter ζ explicit.[10] A related algorithm-independent lower bound was proved by Hou et al. [74], but under a localized clientwise heterogeneity condition

[10] The homogeneous lower bound of Glasgow et al. [36] also applies to the class $\mathcal{P}_\zeta^{H,B,\sigma}$, but it corresponds to the case $\zeta = 0$.

rather than the global averaged condition in Assumption 2.[11] The main difficulty in proving lower bounds under Assumption 2 is that this assumption rules out many standard quadratic hard instances with different client Hessians: for such instances, the average gradient disagreement typically grows unboundedly with $\|x\|_2$. Our first-order block avoids this obstruction. It has $\tau = 0$, and the clients differ only through opposite linear perturbations. Hence the average gradient disagreement is constant over the entire domain, so the bound at $x^\star$ becomes a uniform bound over all x (see Proposition 2). Applying Theorem 3 with $\tau = 0$ and $\zeta_\star = \zeta$ proves the corollary.

G.4 Lower Bounds for the Time-Averaged Output

Theorems 8 and 9 concern the final iterate. We now extend their noiseless lower bounds to the same time-averaged output used in our upper bounds. Throughout this subsection, Local SGD follows (5), with a constant local step size $\eta > 0$, server step size $\beta = 1$, and initialization $x_0^m = 0$. Writing $T = KR$, its output is

$$\widehat{x} := \frac{1}{T}\sum_{t=0}^{T-1} \bar{x}_t, \qquad \bar{x}_t := \frac{1}{M}\sum_{m=1}^{M} x_t^m. \tag{46}$$

The proof directly controls the error of this average. In particular, the hard trajectories retain a common sign in the relevant coordinates, while an additional homogeneous coordinate handles large step sizes.

Theorem 10 (Noiseless lower bound for the time-averaged output)**.** *Fix $H, B > 0$, $\zeta_\star, \tau \geq 0$, and integers $K, M \geq 2$ and $R \geq 1$. There exists a deterministic instance in $\mathcal{P}^{H,B,0}_{\zeta_\star,\tau}$ such that, for every constant step size $\eta > 0$, the output in (46) satisfies*

$$F(\widehat{x}) - F(x^\star) \gtrsim \frac{HB^2}{KR} + \min\left\{\frac{HB^2}{R}, \frac{\tau B^2}{\sqrt{R}}\right\} + \min\left\{HB^2, \frac{\zeta_\star^2}{H}, \frac{(H\zeta_\star^2 B^4)^{1/3}}{R^{2/3}}\right\}. \tag{47}$$

The implicit constant is universal, and the instance does not depend on η.

Proof. We construct three blocks and then place them at disjoint coordinates. Each block works for every constant step size and uses a minimizer of norm at most B. At the end, we split the radius budget across the blocks.

A common guard coordinate. We first record an elementary fact used below. Suppose every client has a coordinate defining the same scalar quadratic function

$$h(v) = \frac{H}{2}(v - v^\star)^2.$$

Since all clients start at the same point, their iterations on this coordinate coincide, and synchronization leaves them unchanged. Their common iterate satisfies

$$v_{t+1} = v_t - \eta H(v_t - v^\star).$$

Thus, the error $e_t := v_t - v^\star$ obeys

$$e_{t+1} = (1 - \eta H)e_t, \qquad e_t = q^t e_0, \qquad q := 1 - \eta H.$$

When $\eta > 4/H$, write $a = -q > 3$.

We next consider the parity of T. If $T = 2s$, group consecutive terms into pairs:

$$\sum_{t=0}^{2s-1} q^t = \sum_{j=0}^{s-1}\left(a^{2j} - a^{2j+1}\right) = -(a-1)\sum_{j=0}^{s-1} a^{2j}.$$

Every pair is negative, so taking the absolute value gives

$$\left|\sum_{t=0}^{2s-1} q^t\right| = (a-1)\sum_{j=0}^{s-1} a^{2j} \geq 2s = T,$$

[11] Hou et al. [74] bound $\sup_{x\in A, m\in[M]} \|\nabla F_m(x) - \nabla F(x)\|_2$ over an algorithm-dependent region A. If one instead compares the global clientwise condition $\sup_{x\in\mathbb{R}^d, m\in[M]} \|\nabla F_m(x) - \nabla F(x)\|_2 \leq \bar{\zeta}$ with Assumption 2, then the former implies Assumption 2 with $\zeta \leq \bar{\zeta}$, while the reverse implication can lose a factor of order $\sqrt{M}$ in the heterogeneity parameter.

where we used $a - 1 \geq 2$ and $a^{2j} \geq 1$.

If $T = 2s + 1$, keep the initial term $q^0 = 1$ separate and pair the remaining terms:

$$\sum_{t=0}^{2s} q^t = 1 + \sum_{j=0}^{s-1} \left(-a^{2j+1} + a^{2j+2}\right) = 1 + (a-1) \sum_{j=0}^{s-1} a^{2j+1}.$$

Every pair is now positive and at least 2, so

$$\left| \sum_{t=0}^{2s} q^t \right| = \sum_{t=0}^{2s} q^t \geq 1 + 2s = T.$$

Thus, in either case, $\left| \sum_{t=0}^{T-1} q^t \right| \geq T$.

Consequently, writing $\widehat{v} = T^{-1} \sum_{t=0}^{T-1} v_t$, we obtain

$$|\widehat{v} - v^\star| = \left| \frac{1}{T} \sum_{t=0}^{T-1} e_t \right| = |e_0| \left| \frac{1}{T} \sum_{t=0}^{T-1} q^t \right| \geq |e_0| = |v_0 - v^\star|, \qquad \eta > 4/H. \tag{48}$$

Therefore,

$$h(\widehat{v}) - h(v^\star) = \frac{H}{2} |\widehat{v} - v^\star|^2 \geq \frac{H}{2} |v_0 - v^\star|^2.$$

The homogeneous optimization term. Set $\mu_0 = H/(12T)$ and give every client the same objective

$$Q(u, v) = \frac{\mu_0}{2} \left(u + \frac{B}{\sqrt{2}} \right)^2 + \frac{H}{2} \left(v + \frac{B}{\sqrt{2}} \right)^2.$$

This objective is convex and H-smooth, and its minimizer has norm B. Both heterogeneity parameters are zero. Since all clients coincide, their trajectories agree at every local step. If $\eta \leq 4/H$, then $\mu_0 = H/(12T)$ gives $0 \leq \eta\mu_0 \leq 1/(3T)$. Since all clients follow the same trajectory, the error in the low-curvature coordinate satisfies

$$u_{t+1} + \frac{B}{\sqrt{2}} = (1 - \eta\mu_0) \left(u_t + \frac{B}{\sqrt{2}} \right).$$

Starting from $u_0 = 0$ and applying Bernoulli's inequality, we obtain

$$u_t + \frac{B}{\sqrt{2}} = (1 - \eta\mu_0)^t \frac{B}{\sqrt{2}} \geq (1 - t\eta\mu_0) \frac{B}{\sqrt{2}} \geq \left(1 - \frac{t}{3T} \right) \frac{B}{\sqrt{2}} \geq \frac{2B}{3\sqrt{2}}, \qquad 0 \leq t < T.$$

Thus, every iterate remains on the same side of the minimizer and at least $2B/(3\sqrt{2})$ away in this coordinate. Averaging gives

$$\widehat{u} + \frac{B}{\sqrt{2}} = \frac{1}{T} \sum_{t=0}^{T-1} \left(u_t + \frac{B}{\sqrt{2}} \right) \geq \frac{2B}{3\sqrt{2}}.$$

Since the other quadratic term is nonnegative, the objective gap satisfies

$$Q(\widehat{u}, \widehat{v}) - \min Q \geq \frac{\mu_0}{2} \left(\widehat{u} + \frac{B}{\sqrt{2}} \right)^2 \geq \frac{\mu_0 B^2}{9} = \frac{HB^2}{108T}.$$

If $\eta > 4/H$, consider the second coordinate of Q. Its minimizer is $v^\star = -B/\sqrt{2}$, while $v_0 = 0$, so its initial error is

$$e_0 = v_0 - v^\star = \frac{B}{\sqrt{2}}.$$

Applying (48) and noting that the first quadratic term in Q is nonnegative gives us,

$$Q(\widehat{u}, \widehat{v}) - \min Q \geq \frac{H}{2} \left(\widehat{v} + \frac{B}{\sqrt{2}} \right)^2 \geq \frac{HB^2}{4}.$$

Together with the case $\eta \leq 4/H$, this proves $\Omega(HB^2/T) = \Omega(HB^2/(KR))$ for every $\eta > 0$.

When $R = 1$ and $K = 2$, each heterogeneity term in (47) is at most HB^2. Since $HB^2/(KR) = HB^2/2$, the homogeneous block proves the entire claim in this case. Below, we assume $R \geq 2$ or $K \geq 3$.

The second-order heterogeneity term. Assume $\tau > 0$, since otherwise this term vanishes. We use a rescaled version of the rank-one construction in Theorem 8, together with the guard coordinate. Define

$$\kappa = 12R, \qquad \alpha = \frac{\kappa - 1}{\kappa + 1}, \qquad p_1 = (1, 0)^\top, \qquad p_2 = (\alpha, \sqrt{1 - \alpha^2})^\top,$$

and

$$\ell = \frac{1}{4} \min\{H, \sqrt{3R}\,\tau\}, \qquad A_i = \ell p_i p_i^\top \quad (i = 1, 2).$$

The factor $1/4$ ensures that $\eta\ell \le 1$ whenever $\eta \le 4/H$. This will keep the relevant error multipliers nonnegative; the guard coordinate will handle $\eta > 4/H$.

The average Hessian. To handle every $M \ge 2$, use $\lfloor M/2 \rfloor$ clients of the first type and $\lceil M/2 \rceil$ of the second. Write

$$\pi = \frac{\lfloor M/2 \rfloor}{M} \in [1/3, 1/2], \qquad A = \pi A_1 + (1 - \pi) A_2.$$

When M is even, this is the equal-weight average from Theorem 8. For arbitrary M, its matrix is

$$A = \ell \begin{bmatrix} \pi + (1 - \pi)\alpha^2 & (1 - \pi)\alpha\sqrt{1 - \alpha^2} \\ (1 - \pi)\alpha\sqrt{1 - \alpha^2} & (1 - \pi)(1 - \alpha^2) \end{bmatrix}.$$

Therefore,

$$\operatorname{tr}(A) = \ell, \qquad \det(A) = \ell^2 \pi (1 - \pi)(1 - \alpha^2).$$

Solving the characteristic equation gives the smaller eigenvalue

$$\lambda_- = \frac{\ell}{2} \left(1 - \sqrt{1 - 4\pi(1 - \pi)(1 - \alpha^2)}\right).$$

Rationalizing this expression yields

$$\lambda_- = \frac{2\ell\pi(1 - \pi)(1 - \alpha^2)}{1 + \sqrt{1 - 4\pi(1 - \pi)(1 - \alpha^2)}} \ge \ell\pi(1 - \pi)(1 - \alpha^2),$$

because the denominator is at most 2. Moreover, $\pi(1 - \pi) \ge 2/9$ and $1 - \alpha^2 = 4\kappa/(\kappa + 1)^2$, so

$$\ell\pi(1 - \pi)(1 - \alpha^2) \ge \frac{8\ell\kappa}{9(\kappa + 1)^2} \ge \frac{\ell}{2(\kappa + 1)},$$

where the last inequality uses $\kappa = 12R \ge 12$. For the upper bound, $\pi(1 - \pi) \le 1/4$ implies

$$\sqrt{1 - 4\pi(1 - \pi)(1 - \alpha^2)} \ge \alpha,$$

and hence $\lambda_- \le \ell(1 - \alpha)/2 = \ell/(\kappa + 1)$. Together,

$$\frac{\ell}{2(\kappa + 1)} \le \lambda_- \le \frac{\ell}{\kappa + 1}. \tag{49}$$

Thus, allowing unequal numbers of the two client types changes the smallest eigenvalue by at most a constant factor.

The objectives and their heterogeneity. Let u_- be a unit eigenvector of A associated with λ_-, and set

$$z^\star = -\frac{B}{\sqrt{2}} u_-, \qquad w^\star = -\frac{B}{\sqrt{2}}.$$

A client of type i has objective

$$F_i(z, w) = \frac{1}{2}(z - z^\star)^\top A_i (z - z^\star) + \frac{H}{2}(w - w^\star)^2.$$

Its Hessian is block diagonal with blocks A_i and H. Since $0 \preceq A_i \preceq \ell I$ and $\ell \le H/4$, every client is convex and H-smooth. All clients share the minimizer $x^\star = (z^\star, w^\star)$, and

$$\|x^\star\|_2^2 = \|z^\star\|_2^2 + |w^\star|^2 = B^2.$$

Their gradients vanish at this minimizer, so this block has $\zeta_\star = 0$. The average objective is

$$F(z, w) = \frac{1}{2}(z - z^\star)^\top A (z - z^\star) + \frac{H}{2}(w - w^\star)^2.$$

As calculated in the proof of Theorem 8, the eigenvalues of $p_1 p_1^\top - p_2 p_2^\top$ are $\pm\sqrt{1-\alpha^2}$. Consequently,

$$\|A_1 - A_2\|_{\mathrm{op}} = \ell\sqrt{1-\alpha^2} \le \frac{\ell}{\sqrt{3R}} \le \tau,$$

where we used $1-\alpha^2 = 4\kappa/(\kappa+1)^2 \le 4/\kappa = 1/(3R)$ and the definition of ℓ.

To verify Assumption 3 explicitly, observe that

$$A_1 - A = (1-\pi)(A_1 - A_2), \qquad A_2 - A = -\pi(A_1 - A_2).$$

Thus, for any $x = (z, w)$ and $y = (z', w')$,

$$\begin{aligned}
&\frac{1}{M}\sum_{m=1}^{M} \|\nabla F_m(x) - \nabla F_m(y) - (\nabla F(x) - \nabla F(y))\|_2^2 \\
&\quad = \pi(1-\pi)\,\|(A_1 - A_2)(z - z')\|_2^2 \\
&\quad \le \pi(1-\pi)\|A_1 - A_2\|_{\mathrm{op}}^2\,\|z - z'\|_2^2 \le \tau^2\,\|x - y\|_2^2 .
\end{aligned}$$

Here the guard coordinate cancels because it is identical across clients. The construction therefore belongs to $\mathcal{P}_{0,\tau}^{H,B,0}$.

Step sizes $\eta \le 4/H$. Let $P_i = p_i p_i^\top$, so $P_i^2 = P_i$ and $A_i = \ell P_i$. As in the proof of Theorem 8, the local update acts by multiplication by $1-\eta\ell$ on the range of P_i and by the identity on its orthogonal complement. Hence

$$(I - \eta A_i)^k = (I - P_i) + (1-\eta\ell)^k P_i = I - \gamma_k A_i, \qquad \gamma_k := \frac{1-(1-\eta\ell)^k}{\ell}.$$

Since $0 \le 1 - \eta\ell \le 1$, we have $0 \le \gamma_k \le 1/\ell$.

Let $e_r := \bar{z}_{rK} - z^\star$ denote the synchronized error at the start of round r. Starting from e_r, a client of type i has error

$$e_{r,k}^{(i)} = (I - \eta A_i)^k e_r = (I - \gamma_k A_i) e_r$$

after k local gradient steps, before the endpoint synchronization. Averaging over the two client types gives

$$\bar{e}_{r,k} := \pi e_{r,k}^{(1)} + (1-\pi) e_{r,k}^{(2)} = (I - \gamma_k A) e_r.$$

In particular, synchronization at the end of the round gives $e_{r+1} = \bar{e}_{r,K} = (I - \gamma_K A) e_r$. Since all clients start at zero,

$$e_0 = -z^\star = \frac{B}{\sqrt{2}} u_-,$$

and therefore

$$\bar{e}_{r,k} = (I - \gamma_k A)(I - \gamma_K A)^r e_0.$$

Because $A u_- = \lambda_- u_-$, the projection onto the slow eigendirection is

$$\langle \bar{e}_{r,k}, u_- \rangle = \frac{B}{\sqrt{2}}(1 - \gamma_k\lambda_-)(1 - \gamma_K\lambda_-)^r.$$

By (49) and $\gamma_k \le 1/\ell$, every factor satisfies

$$1 - \gamma_k\lambda_- \ge 1 - \frac{\lambda_-}{\ell} \ge 1 - \frac{1}{12R+1} > 0.$$

Thus, for $0 \le r < R$ and $0 \le k < K$,

$$\begin{aligned}
\langle \bar{e}_{r,k}, u_- \rangle &\ge \frac{B}{\sqrt{2}}\left(1 - \frac{1}{12R+1}\right)^{r+1} \\
&\ge \frac{B}{\sqrt{2}}\left(1 - \frac{1}{12R+1}\right)^{R} \\
&\ge \frac{B}{\sqrt{2}}\left(1 - \frac{R}{12R+1}\right) \ge \frac{11B}{12\sqrt{2}},
\end{aligned}$$

where the penultimate inequality is Bernoulli's inequality.

The projected errors therefore retain the same sign and remain bounded away from zero throughout the averaging window. Since $\bar{e}_{r,k} = \bar{z}_{rK+k} - z^\star$ for $0 \le k < K$, we obtain

$$\langle \widehat{z} - z^\star, u_- \rangle = \frac{1}{KR} \sum_{r=0}^{R-1} \sum_{k=0}^{K-1} \langle \bar{e}_{r,k}, u_- \rangle \ge \frac{11B}{12\sqrt{2}}.$$

The quadratic objective then gives

$$\begin{aligned} F(\widehat{x}) - F(x^\star) &\ge \frac{1}{2}(\widehat{z} - z^\star)^\top A(\widehat{z} - z^\star) \\ &\ge \frac{\lambda_-}{2} \langle \widehat{z} - z^\star, u_- \rangle^2 \\ &\ge \frac{121\lambda_- B^2}{576} \gtrsim \frac{\ell B^2}{R} \\ &\gtrsim \min\left\{ \frac{HB^2}{R}, \frac{\tau B^2}{\sqrt{R}} \right\}. \end{aligned}$$

Here we used the lower bound in (49), the inequality $12R + 1 \le 13R$, and the definition of ℓ.

Step sizes $\eta > 4/H$. The guard coordinate has minimizer $w^\star = -B/\sqrt{2}$ and initial iterate $w_0 = 0$. Its error therefore satisfies

$$w_t - w^\star = (1 - \eta H)^t (w_0 - w^\star) = (1 - \eta H)^t \frac{B}{\sqrt{2}}.$$

Applying (48) yields

$$|\widehat{w} - w^\star| \ge |w_0 - w^\star| = \frac{B}{\sqrt{2}}.$$

Consequently,

$$F(\widehat{x}) - F(x^\star) \ge \frac{H}{2} |\widehat{w} - w^\star|^2 \ge \frac{HB^2}{4}.$$

Since $\min\{HB^2/R, \tau B^2/\sqrt{R}\} \le HB^2$, this also proves the desired second-order term.

The first-order heterogeneity term. Assume $\zeta_\star > 0$, since otherwise this term vanishes. Recall that the homogeneous construction already handles $(R, K) = (1, 2)$, so we may assume $R \ge 2$ or $K \ge 3$. Define

$$E = \min\left\{ HB^2, \frac{\zeta_\star^2}{H}, \frac{(H\zeta_\star^2 B^4)^{1/3}}{R^{2/3}} \right\}, \qquad \mu = \frac{E}{8B^2}, \qquad L = \frac{H}{4}.$$

Since $E \le HB^2$, we have $\mu \le H/8 = L/2$.

The objectives and their heterogeneity. We use the same three-coordinate structure as Theorem 9, with a smaller curvature scale on the active coordinate. Define

$$g(s) = \begin{cases} Ls^2/4, & s < 0, \\ Ls^2/2, & s \ge 0, \end{cases} \qquad q(u) = \frac{\mu}{2}\left(u + \frac{B}{3}\right)^2, \qquad h(v) = \frac{H}{2}\left(v + \frac{B}{3}\right)^2.$$

Use $n = \lfloor M/2 \rfloor$ clients with objective $g(s) - \zeta_\star s + q(u) + h(v)$, another n with objective $g(s) + \zeta_\star s + q(u) + h(v)$, and, when M is odd, one neutral client with objective $g(s) + q(u) + h(v)$. Write $\rho = n/M \in [1/3, 1/2]$.

The linear terms cancel in the average, so

$$F(s, u, v) = g(s) + q(u) + h(v), \qquad x^\star = (0, -B/3, -B/3), \qquad \|x^\star\|_2 = \frac{\sqrt{2}B}{3} \le B.$$

The function g is convex and L-smooth. Since the three coordinates are separate and $\mu, L \le H$, every client is convex and H-smooth. The clients differ only through linear terms, so their gradient

increments coincide and the second-order heterogeneity is zero. At $x^\star$, the two nonneutral types have gradients $(-\zeta_\star, 0, 0)$ and $(\zeta_\star, 0, 0)$, respectively, while the neutral client has zero gradient. Hence

$$\frac{1}{M}\sum_{m=1}^{M}\|\nabla F_m(x^\star)\|_2^2 = 2\rho\zeta_\star^2 \le \zeta_\star^2.$$

Thus this block belongs to $\mathcal{P}_{\zeta_\star,0}^{H,B,0}$.

Small and large step sizes. Set $T = KR$ and define

$$\eta_0 = \frac{1}{2\mu T}.$$

The u coordinate is identical across clients and has minimizer $u^\star = -B/3$. If $\eta \le \eta_0$, then $\eta\mu \le 1/(2T)$, and its error satisfies

$$u_t - u^\star = (1-\eta\mu)^t\frac{B}{3} \ge (1-t\eta\mu)\frac{B}{3} \ge \frac{B}{6}, \qquad 0 \le t < T.$$

Averaging preserves this lower bound, giving

$$q(\widehat{u}) - q(u^\star) = \frac{\mu}{2}\left(\widehat{u} + \frac{B}{3}\right)^2 \ge \frac{\mu B^2}{72} = \frac{E}{576}.$$

If $\eta > 4/H$, the guard coordinate has minimizer $v^\star = -B/3$ and initial error $v_0 - v^\star = B/3$. By (48),

$$|\widehat{v} - v^\star| \ge \frac{B}{3}, \qquad h(\widehat{v}) - h(v^\star) \ge \frac{HB^2}{18} \ge \frac{E}{18}.$$

It remains to consider $\eta_0 < \eta \le 4/H = 1/L$. If this interval is empty, the preceding two cases already suffice.

Comparing the active-coordinate trajectories. We first analyze the active coordinate for any $0 < \eta \le 1/L$. As in Lemma 16, compare the objectives

$$G_-(s) = g(s) - \zeta_\star s, \qquad G_+(s) = g(s) + \zeta_\star s$$

with the quadratics

$$\widetilde{G}_-(s) = \frac{L}{2}s^2 - \zeta_\star s, \qquad \widetilde{G}_+(s) = \frac{L}{4}s^2 + \zeta_\star s.$$

For the neutral client, when present, compare $G_0(s) = g(s)$ with $\widetilde{G}_0(s) = Ls^2/4$.

Let $T_i(s) = s - \eta G_i'(s)$ and $\widetilde{T}_i(s) = s - \eta\widetilde{G}_i'(s)$ denote the corresponding update maps. The proof of Lemma 16 establishes $T_- \le \widetilde{T}_-$ and $T_+ \le \widetilde{T}_+$ pointwise. The neutral client satisfies the same ordering because

$$g'(s) \ge \frac{L}{2}s \quad\Longrightarrow\quad T_0(s) = s - \eta g'(s) \le \left(1 - \frac{\eta L}{2}\right)s = \widetilde{T}_0(s).$$

All these maps are nondecreasing: their slopes are either $1-\eta L$ or $1-\eta L/2$, both nonnegative. Consequently, if $s \le \widetilde{s}$, then

$$T_i(s) \le \widetilde{T}_i(s) \le \widetilde{T}_i(\widetilde{s}).$$

Starting both systems at zero, induction therefore preserves the ordering at every local step. Averaging at synchronization also preserves it.

Write

$$\xi = \eta L, \qquad a = 1 - \xi/2, \qquad b = 1 - \xi,$$

so $a, b \in [0, 1]$. Let $\widetilde{s}_r$ denote the synchronized comparison iterate at time rK. Starting from $\widetilde{s}_r$, the comparison trajectories after k local steps are

$$\widetilde{s}_{r,k}^{(-)} = b^k\widetilde{s}_r + \eta\zeta_\star\sum_{j=0}^{k-1}b^j = b^k\widetilde{s}_r + \frac{\zeta_\star}{L}(1-b^k),$$

$$\widetilde{s}_{r,k}^{(+)} = a^k \widetilde{s}_r - \eta\zeta_\star \sum_{j=0}^{k-1} a^j = a^k \widetilde{s}_r - \frac{2\zeta_\star}{L}(1-a^k),$$
$$\widetilde{s}_{r,k}^{(0)} = a^k \widetilde{s}_r.$$

These are the geometric-sum calculations from Lemma 16, applied at every local step. The neutral type has weight $1-2\rho$, which is zero when M is even. Thus the mean comparison iterate is

$$\widetilde{\bar{s}}_{r,k} = \rho\widetilde{s}_{r,k}^{(-)} + \rho\widetilde{s}_{r,k}^{(+)} + (1-2\rho)\widetilde{s}_{r,k}^{(0)} = A_k \widetilde{s}_r - \Delta_k,$$

where

$$A_k = \rho b^k + (1-\rho)a^k, \qquad \Delta_k = \frac{\rho\zeta_\star}{L}(1 + b^k - 2a^k).$$

We have $A_k \in [0,1]$. To check the sign of Δ_k, define

$$\phi_k(x) = 1 + (1-x)^k - 2(1-x/2)^k.$$

For $k \geq 1$ and $x \in [0,1]$,

$$\phi_k(0) = 0, \qquad \phi_k'(x) = k\big((1-x/2)^{k-1} - (1-x)^{k-1}\big) \geq 0.$$

Hence $\Delta_k = (\rho\zeta_\star/L)\phi_k(\xi) \geq 0$; also $\Delta_0 = 0$. The synchronized comparison iterates therefore satisfy

$$\widetilde{s}_{r+1} = A_K \widetilde{s}_r - \Delta_K, \qquad \widetilde{s}_0 = 0,$$

and remain nonpositive. Since $\widetilde{\bar{s}}_{r,k} = A_k \widetilde{s}_r - \Delta_k$, the within-round means are also nonpositive. Combining this with the clientwise comparison gives

$$\bar{s}_{rK+k} \leq \widetilde{\bar{s}}_{r,k} \leq 0, \qquad 0 \leq r < R, \quad 0 \leq k < K.$$

Thus the true and comparison means have the same sign throughout the averaging window.

An exact expression for the average bias. Unrolling the synchronized recursion gives

$$\widetilde{s}_r = -\Delta_K \sum_{j=0}^{r-1} A_K^j,$$

where the sum is zero for $r = 0$. Define

$$\widehat{s} = \frac{1}{KR}\sum_{r=0}^{R-1}\sum_{k=0}^{K-1} \bar{s}_{rK+k}, \qquad \widehat{s}_{\mathrm{cmp}} = \frac{1}{KR}\sum_{r=0}^{R-1}\sum_{k=0}^{K-1} \widetilde{\bar{s}}_{r,k},$$

and

$$\bar{A} = \frac{1}{K}\sum_{k=0}^{K-1} A_k, \qquad \bar{\Delta} = \frac{1}{K}\sum_{k=0}^{K-1} \Delta_k, \qquad G_R(A) = \frac{1}{R}\sum_{r=0}^{R-1}\sum_{j=0}^{r-1} A^j.$$

Averaging first within each round and then across rounds yields

$$\begin{aligned}\widehat{s}_{\mathrm{cmp}} &= \frac{1}{R}\sum_{r=0}^{R-1}\big(\bar{A}\widetilde{s}_r - \bar{\Delta}\big) \\ &= -\bar{\Delta} - \bar{A}\Delta_K G_R(A_K).\end{aligned}$$

Since $\widehat{s} \leq \widehat{s}_{\mathrm{cmp}} \leq 0$, we obtain

$$-\widehat{s} \geq -\widehat{s}_{\mathrm{cmp}} = \bar{\Delta} + \bar{A}\Delta_K G_R(A_K). \tag{50}$$

The first term captures the bias generated within a round, while the second captures the bias accumulated over preceding rounds. Both terms are nonnegative.

A quantitative lower bound on the average bias. We use the second-difference estimate from the proof of Lemma 16, with K replaced by any integer $k \geq 2$:

$$\phi_k(x) \geq \frac{1}{8\mathrm{e}}\min\{(kx)^2, 1\}, \qquad 0 \leq x \leq 1.$$

To recall the argument, when $x \le 1/k$, the function $f(u) = (1-u)^k$ satisfies

$$f''(u) = k(k-1)(1-u)^{k-2} \ge k(k-1)(1-1/k)^{k-2} \ge \frac{k^2}{2\mathrm{e}}, \qquad 0 \le u \le x.$$

The midpoint inequality for strongly convex functions then gives

$$\phi_k(x) = f(0) + f(x) - 2f(x/2) \ge \frac{k^2x^2}{8\mathrm{e}}.$$

For $x \ge 1/k$, monotonicity gives $\phi_k(x) \ge \phi_k(1/k) \ge 1/(8\mathrm{e})$. Since $\rho \ge 1/3$, it follows that

$$\Delta_k \ge \frac{\zeta_\star}{24\mathrm{e}L} \min\{(k\xi)^2, 1\}, \qquad k \ge 2. \tag{51}$$

Set $\theta = K\xi = K\eta L$. We now bound the two contributions in (50).

First suppose $\theta \le 1$ and $R \ge 2$. For $k < K$, Bernoulli's inequality gives

$$a^k = (1 - \xi/2)^k \ge 1 - k\xi/2 \ge 1/2.$$

Since $1 - \rho \ge 1/2$, this implies $A_k \ge 1/4$ and hence $\bar{A} \ge 1/4$. Also,

$$\begin{aligned} 1 - A_K &= \rho(1 - b^K) + (1-\rho)(1 - a^K) \\ &\le \rho K\xi + (1-\rho)K\xi/2 = \frac{1+\rho}{2}\theta \le \frac{3\theta}{4}. \end{aligned}$$

To lower-bound $G_R(A_K)$, let $J = \lfloor R/2 \rfloor$. For every $r \ge J$, the inner sum defining G_R contains at least the first J powers. Since $R - J \ge R/2$,

$$G_R(A_K) \ge \frac{1}{2} \sum_{j=0}^{J-1} A_K^j.$$

Writing $\delta = 1 - A_K > 0$ and using $A_K^J = (1-\delta)^J \le \exp(-J\delta)$, we obtain

$$G_R(A_K) \ge \frac{1 - \exp(-J\delta)}{2\delta} \gtrsim \min\{J, \delta^{-1}\} \gtrsim \min\{R, \theta^{-1}\}.$$

Here we used $1 - \exp(-u) \ge (1 - \mathrm{e}^{-1}) \min\{u, 1\}$, $J \ge R/3$, and $\delta \le 3\theta/4$. Combining this with (51) at $k = K$ gives

$$-\widehat{s} \ge \bar{A}\Delta_K G_R(A_K) \gtrsim \frac{\zeta_\star}{L}\theta^2 \min\{R, \theta^{-1}\} = \frac{\zeta_\star}{L} \min\{R\theta^2, \theta\}.$$

Next suppose $K \ge 3$. There are $\lfloor K/2 \rfloor \ge K/3$ indices in the range $\lceil K/2 \rceil \le k \le K-1$. Each such index satisfies $k \ge 2$ and $k\xi \ge \theta/2$, so

$$\begin{aligned} \bar{\Delta} &\ge \frac{1}{K} \sum_{k=\lceil K/2 \rceil}^{K-1} \Delta_k \\ &\gtrsim \frac{\zeta_\star}{L} \min\{\theta^2, 1\}. \end{aligned}$$

When $R = 1$, this is the desired bound because $\min\{1, \theta, R\theta^2\} = \min\{1, \theta^2\}$. When $\theta \ge 1$, it gives $-\widehat{s} \ge \bar{\Delta} \gtrsim \zeta_\star/L$ for every R.

The remaining case is $K = 2$, $R \ge 2$, and $\theta \ge 1$. Here (51) gives $\Delta_K \gtrsim \zeta_\star/L$. Moreover,

$$\bar{A} = \frac{A_0 + A_1}{2} \ge \frac{1}{2}, \qquad G_R(A_K) \ge \frac{R-1}{R} \ge \frac{1}{2},$$

because $A_0 = 1$ and each inner sum with $r \ge 1$ contains the term $A_K^0 = 1$. Thus $-\widehat{s} \ge \bar{A}\Delta_K G_R(A_K) \gtrsim \zeta_\star/L$.

Together, these cases establish

$$-\widehat{s} \gtrsim \frac{\zeta_\star}{L} \min\{1, \theta, R\theta^2\}.$$

Since $\widehat{s} \le 0$, the active part of the average objective satisfies

$$g(\widehat{s}) - g(0) = \frac{L}{4}\widehat{s}^2 \gtrsim \frac{\zeta_\star^2}{L} \min\{1, \theta^2, R^2\theta^4\}. \tag{52}$$

Completing the intermediate step-size case. Return to $\eta_0 < \eta \le 1/L$. The right-hand side of (52) is nondecreasing in $\theta = K\eta L$, so it is at least its value at

$$\theta_0 = K\eta_0 L = \frac{L}{2\mu R}.$$

Since $\mu \le L/2$, we have $R\theta_0 \ge 1$ and therefore $R^2\theta_0^4 \ge \theta_0^2$. Consequently,

$$\begin{aligned} g(\widehat{s}) - g(0) &\gtrsim \frac{\zeta_\star^2}{L} \min\{1, \theta_0^2, R^2\theta_0^4\} \\ &= \min\left\{\frac{\zeta_\star^2}{L}, \frac{L\zeta_\star^2}{4\mu^2 R^2}\right\}. \end{aligned}$$

Both terms inside this minimum are at least $4E$. Indeed, $E \le \zeta_\star^2/H$ gives

$$\frac{\zeta_\star^2}{L} = \frac{4\zeta_\star^2}{H} \ge 4E.$$

Also, substituting $\mu = E/(8B^2)$ and $L = H/4$ gives

$$\frac{L\zeta_\star^2}{4\mu^2 R^2} = \frac{4H\zeta_\star^2 B^4}{E^2 R^2} \ge 4E,$$

where the last inequality follows from $E^3 \le H\zeta_\star^2 B^4/R^2$.

The other two coordinates contribute nonnegative objective gaps, so $F(\widehat{x}) - F(x^\star) \ge g(\widehat{s}) - g(0) \gtrsim E$ throughout the intermediate range. Together with the small- and large-step-size cases, this proves the first-order term for every $\eta > 0$ on an instance chosen independently of η.

Combining the blocks. Construct each nonzero block above with radius $B/\sqrt{3}$, and add its client objectives on disjoint coordinates. The combined minimizer has norm at most B. Each client remains convex and H-smooth, because the blocks depend on disjoint coordinates. Only the second-order block contributes to Assumption 3, and only the first-order block contributes to Assumption 1. Thus the combined instance belongs to $\mathcal{P}^{H,B,0}_{\zeta_\star,\tau}$. Local SGD evolves independently on the blocks, and their objective gaps add. Replacing B by $B/\sqrt{3}$ reduces each target term by at most a universal constant factor. Since every block is fixed independently of η and works for every $\eta > 0$, their sum proves (47) with the stated quantifiers. □

Corollary 2 (Uniform first-order heterogeneity)**.** *Fix $H, B > 0$, $\zeta \ge 0$, and integers $K, M \ge 2$ and $R \ge 1$. There exists a deterministic instance in $\mathcal{P}^{H,B,0}_{\zeta}$ such that, for every constant step size $\eta > 0$, the output in (46) satisfies*

$$F(\widehat{x}) - F(x^\star) \gtrsim \frac{HB^2}{KR} + \min\left\{HB^2, \frac{\zeta^2}{H}, \frac{(H\zeta^2 B^4)^{1/3}}{R^{2/3}}\right\}.$$

Proof. Apply Theorem 10 with $\tau = 0$ and $\zeta_\star = \zeta$. Write $\rho = \lfloor M/2 \rfloor / M$. Its first-order block differs across clients only through linear terms, so

$$\frac{1}{M}\sum_{m=1}^{M} \|\nabla F_m(x) - \nabla F(x)\|_2^2 = 2\rho\zeta^2 \le \zeta^2$$

for every x; the homogeneous blocks add no disagreement. Thus the construction satisfies Assumption 2. When $\zeta = 0$ or $(R, K) = (1, 2)$, the homogeneous construction alone gives the claimed bound. □

Standard statistical term. The statistical lower bound

$$\Omega\left(\min\left\{HB^2, \frac{\sigma B}{\sqrt{MKR}}\right\}\right)$$

holds for arbitrary output rules and already arises on homogeneous instances; see Woodworth et al. [34, Theorem 1 and its proof sketch]. It therefore applies to (46) as well. For every fixed step size, the worst-case error over $\mathcal{P}^{H,B,\sigma}_{\zeta_\star,\tau}$ is at least the maximum of this statistical bound and the noiseless bound in (47), and hence at least a constant times their sum. This argument allows the stochastic hard instance to depend on the chosen step size. The separate stochastic local-drift term

$$\min\left\{\frac{\sigma B}{\sqrt{KR}}, \frac{(H\sigma^2 B^4)^{1/3}}{K^{1/3}R^{2/3}}\right\}$$

is not covered by the averaged-output argument above.

H Phase Diagram: Comparison of New Lower and Upper Bounds

In this section, we determine in which regimes in the $\zeta_\star$-τ diagram our new upper and lower bounds match, and when a gap remains.

To facilitate comparison, we divide the $\zeta_\star$-τ diagram into multiple segments and identify the dominant terms in the upper and lower bounds. In this section, we consider only a practical setting $\zeta_\star \le HB$ in the noiseless regime $\sigma = 0$.

In the aforementioned setting, our upper and lower bounds can be written as follows

$$\mathbb{E}[F(\hat{x}_t) - F^\star] \lesssim \min\left\{R_1(\tau,\zeta_\star), R_2(\tau,\zeta_\star), R_3(\tau,\zeta_\star)\right\}, \quad \text{where}$$

$$R_1(\tau,\zeta_\star) = \frac{HB^2}{KR} + \frac{\tau B^2}{\sqrt{R}} + \frac{\zeta_\star B}{\sqrt{R}}, \qquad R_2(\tau,\zeta_\star) = \frac{HB^2}{KR} + \frac{(H\tau^2)^{1/3}B^2}{R^{2/3}} + \frac{(H\zeta_\star^2 B^4)^{1/3}}{R^{2/3}},$$

$$R_3(\tau,\zeta_\star) = \frac{HB^2}{R} + \frac{(H\zeta_\star^2 B^4)^{1/3}}{R^{2/3}}, \quad \text{and}$$

$$\mathbb{E}[F(\hat{x}_t) - F^\star] \gtrsim \frac{HB^2}{KR} + \min\left\{\frac{HB^2}{R}, \frac{\tau B^2}{\sqrt{R}}\right\} + \min\left\{HB^2, \frac{\zeta_\star^2}{H}, \frac{(H\zeta_\star^2 B^4)^{1/3}}{R^{2/3}}\right\}.$$

Both bounds concern the time-averaged output in (46); the lower bound follows from Theorem 10.

We use the following notation $\alpha = \frac{HB^2}{R}$, $x_0 = \frac{\tau\sqrt{R}}{H}$, and $y_0 = \frac{\zeta_\star\sqrt{R}}{HB}$, then the bounds can be written as follows

$$U(x_0,y_0) = \min\left\{\frac{\alpha}{K} + \alpha x_0 + \alpha y_0, \frac{\alpha}{K} + \alpha x_0^{2/3} + \alpha y_0^{2/3}, \alpha + \alpha y_0^{2/3}\right\},$$

$$L(x_0,y_0) = \frac{\alpha}{K} + \min\left\{\alpha, \alpha x_0\right\} + \min\left\{\alpha R, \alpha y_0^2, \alpha y_0^{2/3}\right\},$$

where $U(x_0,y_0)$ is the upper bound and $L(x_0,y_0)$ is the lower bound.

H.1 Analysis of Upper Bound's Dominant Term

Case 1: $0 \le x_0 \le \frac{1}{K}, 0 \le y_0 \le \frac{1}{K}$**.** In this case, we have $R_1, R_2 \ge \frac{\alpha}{K}$ and $R_3 \ge \alpha$. So $U(x_0,y_0) \ge \frac{\alpha}{K}$. Moreover, $R_1 = \frac{\alpha}{K} + \alpha x_0 + \alpha y_0 \le \frac{3\alpha}{K}$, so in this regime $U(x_0,y_0) \sim \frac{\alpha}{K}$.

Case 2: $\frac{1}{K} \le x_0 \le 1, y_0 \le x_0$**.** In this case, we have

$$R_1 = \frac{\alpha}{K} + \alpha x_0 + \alpha y_0 \ge \alpha x_0, \quad R_2 = \frac{\alpha}{K} + \alpha x_0^{2/3} + \alpha y_0^{2/3} \ge \alpha x_0^{2/3} \ge \alpha x_0,$$

and

$$R_3 = \alpha + \alpha y_0^{2/3} \ge \alpha \ge \alpha x_0.$$

In other words, each term is at least αx_0, so $Y(x_0,y_0) \ge \alpha x_0$. At the same time, $R_1 \le 3\alpha x_0$. Therefore, in this regime $U(x_0,y_0) \sim \alpha x_0$.

Case 3: $\frac{1}{K} \leq y_0 \leq 1, x_0 \leq y_0$**.** This case is symmetric to Case 2, when swapping x_0 and y_0. Therefore, $U(x_0, y_0) \sim \alpha y_0$.

Case 4: $x_0 \geq 1, y_0 \leq 1$**.** In this case, each of the rates R_1, R_2, R_3 is at least α. Indeed, we have

$$R_1 = \frac{\alpha}{K} + \alpha x_0 + \alpha y_0 \geq \alpha x_0 \geq \alpha, \qquad R_2 = \frac{\alpha}{K} + \alpha x_0^{2/3} + \alpha y_0^{2/3} \geq \alpha x_0^{2/3} \geq \alpha,$$

and

$$R_3 = \alpha + \alpha y_0^{2/3} \geq \alpha.$$

Therefore, $U(x_0, y_0) \geq \alpha$ as well. At the same time, $R_3 \leq \alpha + \alpha = 2\alpha$. Thus, we obtain that in this regime $U(x_0, y_0) \sim \alpha$.

Case 5: $1 \leq y_0 \leq \sqrt{R}$**.** Finally, in this regime, we have

$$R_1 = \frac{\alpha}{K} + \alpha x_0 + \alpha y_0 \geq \alpha y_0 \geq \alpha y_0^{2/3}, \qquad R_2 = \frac{\alpha}{K} + \alpha x_0^{2/3} + \alpha y_0^{2/3} \geq \alpha y_0^{2/3},$$

and

$$R_3 = \alpha + \alpha y_0^{2/3} \geq \alpha y_0^{2/3}.$$

At the same time, $R_3 \leq 2\alpha y_0^{2/3}$. Thus, $U(x_0, y_0) \sim \alpha y_0^{2/3}$.

Final split. Summarizing the above 5 cases together, the upper bound has the following phase diagram

$$U(x_0, y_0) = \begin{cases} \frac{\alpha}{K}, & \text{if } 0 \leq x_0 \leq \frac{1}{K}, 0 \leq y_0 \leq \frac{1}{K}, \\ \alpha x_0, & \text{if } \frac{1}{K} \leq x_0 \leq 1, y_0 \leq x_0, \\ \alpha y_0, & \text{if } \frac{1}{K} \leq y_0 \leq 1, x_0 \leq y_0, \\ \alpha, & \text{if } x_0 \geq 1, y_0 \leq 1, \\ \alpha y_0^{2/3}, & \text{if } 1 \leq y_0 \leq \sqrt{R}. \end{cases}$$

In the original notation, the phase diagram of the upper bound has the following form

$$U(\tau, \zeta_\star) = \begin{cases} \frac{HB^2}{KR}, & \text{if } 0 \leq \tau \leq \frac{H}{K\sqrt{R}}, 0 \leq \zeta_\star \leq \frac{HB}{K\sqrt{R}}, \\ \frac{\tau B^2}{\sqrt{R}}, & \text{if } \frac{H}{K\sqrt{R}} \leq \tau \leq \frac{H}{\sqrt{R}}, \zeta_\star \leq \tau B, \\ \frac{\zeta_\star B}{\sqrt{R}}, & \text{if } \frac{HB}{K\sqrt{R}} \leq \zeta_\star \leq \frac{HB}{\sqrt{R}}, \tau B \leq \zeta_\star, \\ \frac{HB^2}{R}, & \text{if } \tau \geq \frac{H}{\sqrt{R}}, \zeta_\star \leq \frac{HB}{\sqrt{R}}, \\ \frac{(H\zeta_\star B^4)^{1/3}}{R^{2/3}}, & \text{if } \frac{HB}{\sqrt{R}} \leq \zeta_\star \leq HB. \end{cases}$$

H.2 Analysis of Lower Bound's Dominant Term

Note that the lower bound has the following form:

$$L(x_0, y_0) = \frac{\alpha}{K} + \min\{\alpha, \alpha x_0\} + \min\left\{\alpha R, \alpha y_0^2, \alpha y_0^{2/3}\right\}.$$

Case 1: $0 \leq x_0 \leq \frac{1}{K}, 0 \leq y_0 \leq \frac{1}{\sqrt{K}}$**.** In this case, we have $x_0 \leq \frac{1}{K}$ and $y_0^2 \leq \frac{1}{K}$. Therefore, $y_0^2 \leq y_0^{2/3}$, and the lower bound is

$$L(x_0, y_0) = \frac{\alpha}{K} + \alpha x_0 + \alpha y_0^2.$$

This implies that

$$\frac{\alpha}{K} \leq L(x_0, y_0) \leq \frac{3\alpha}{K}$$

in this regime, i.e., $L(x_0, y_0) \sim \frac{\alpha}{K}$.

Case 2: $\frac{1}{K} \leq x_0 \leq 1, y_0 \leq \sqrt{x_0}$**.** Note that in this regime we have $y_0^2 \leq x_0 \leq 1$ and $y_0^2 \leq y_0^{2/3}$. Thus, the lower bound is

$$L(x_0, y_0) = \frac{\alpha}{K} + \alpha x_0 + \alpha y_0^2.$$

This implies that

$$\alpha x_0 \leq L(x_0, y_0) \leq 3\alpha x_0$$

in this regime, i.e., $L(x_0, y_0) \sim \alpha x_0$.

Case 3: $\frac{1}{\sqrt{K}} \leq y_0 \leq 1, x_0 \leq y_0^2$**.** In this regime, we have $\frac{1}{K} \leq y_0^2 \leq y_0^{2/3} \geq 1$ and $x_0 \leq y_0^2 \geq 1$. This implies that the lower bound is

$$L(x_0, y_0) = \frac{\alpha}{K} + \alpha x_0 + \alpha y_0^2.$$

Therefore, we have

$$\alpha y_0^2 \leq L(x_0, y_0) \leq 3\alpha y_0^2,$$

i.e., $L(x_0, y_0) \sim \alpha y_0^2$.

Case 4: $1 \leq x_0, y_0 \leq 1$**.** Since $y_0^2 \leq y_0^{2/3} \leq 1$ in this regime, then the lower bound is

$$L(x_0, y_0) = \frac{\alpha}{K} + \alpha + \alpha y_0^2.$$

This implies that

$$\alpha \leq L(x_0, y_0) \leq 3\alpha,$$

i.e., $L(x_0, y_0) \sim \alpha$.

Case 5: $1 \leq y_0 \leq \sqrt{R}$**.** Note that now $y_0^{2/3} \leq y_0^2$. Therefore, in this regime, we have that the lower bound is

$$L(x_0, y_0) = \frac{\alpha}{K} + \min\{\alpha, \alpha x_0\} + \alpha y_0^{2/3}.$$

Since $\min\{\alpha, \alpha x_0\} \leq \alpha \leq \alpha y_0^{2/3}$ and $\frac{\alpha}{K} \leq \alpha \leq \alpha y_0^{2/3}$, then we have

$$\alpha y_0^{2/3} \leq L(x_0, y_0) \leq 3\alpha y_0^{2/3},$$

i.e., $L(x_0, y_0) \sim 3\alpha y_0^{2/3}$ in this regime.

Final split. Summarizing the above 5 cases together, the lower bound has the following phase diagram

$$L(x_0, y_0) = \begin{cases} \frac{\alpha}{K}, & \text{if } 0 \leq x_0 \leq \frac{1}{K}, 0 \leq y_0 \leq \frac{1}{\sqrt{K}}, \\ \alpha x_0, & \text{if } \frac{1}{K} \leq x_0 \leq 1, y_0 \leq \sqrt{x_0}, \\ \alpha y_0^2, & \text{if } \frac{1}{\sqrt{K}} \leq y_0 \leq 1, x_0 \leq y_0^2, \\ \alpha, & \text{if } 1 \leq x_0, y_0 \leq 1, \\ \alpha y_0^{2/3}, & \text{if } 1 \leq y_0 \leq \sqrt{R}. \end{cases}$$

In the original notation, the phase diagram for the lower bound has the following form

$$L(\tau, \zeta_\star) = \begin{cases} \frac{HB^2}{KR}, & \text{if } 0 \leq \tau \leq \frac{H}{K\sqrt{R}}, 0 \leq \zeta_\star \leq \frac{HB}{\sqrt{KR}}, \\ \frac{\tau B^2}{\sqrt{R}}, & \text{if } \frac{H}{K\sqrt{R}} \leq \tau \leq \frac{H}{\sqrt{R}}, \zeta_\star^2 \leq \frac{HB^2\tau}{\sqrt{R}}, \\ \frac{\zeta_\star^2}{H}, & \text{if } \frac{HB}{\sqrt{RK}} \leq \zeta_\star \leq \frac{HB}{\sqrt{R}}, \frac{HB^2\tau}{\sqrt{R}} \leq \zeta_\star^2, \\ \frac{HB^2}{R}, & \text{if } \frac{H}{\sqrt{R}} \leq \tau, \zeta_\star \leq \frac{HB}{\sqrt{R}}, \\ \frac{(H\zeta_\star^2 B^4)^{1/3}}{R^{2/3}}, & \text{if } \frac{HB}{\sqrt{R}} \leq \zeta_\star \leq HB. \end{cases}$$

H.3 Combined Phase Diagram

Knowing the phase diagram for both the lower and upper bounds, we can merge them into one to show the areas where the lower and upper bounds match (up to numerical constants) or do not. We have

$$(U(x_0,y_0),L(x_0,y_0)) = \begin{cases} (\frac{\alpha}{K},\frac{\alpha}{K}), & \text{if } 0\le x_0\le \frac{1}{K}, 0\le y_0\le \frac{1}{K},\\ (\alpha x_0,\alpha x_0) & \text{if } \frac{1}{K}\le x_0\le 1, y_0\le x_0,\\ (\alpha,\alpha), & \text{if } 1\le x_0\le \sqrt{R}, 0\le y_0\le 1,\\ (\alpha y_0^{2/3},\alpha y_0^{2/3}), & \text{if } 1\le y_0\le \sqrt{R},\\ (\alpha y_0,\max\{\frac{\alpha}{K},\alpha x_0,\alpha y_0^2\}), & \text{if } x_0\le y_0, \frac{1}{K}\le y_0\le 1. \end{cases}$$

In the notation $\tau,\zeta_\star$ the bounds become

$$(U(\tau,\zeta_\star),L(\tau,\zeta_\star)) = \begin{cases} (\frac{HB^2}{KR},\frac{HB^2}{KR}), & \text{if } 0\le \tau\le \frac{H}{K\sqrt{R}}, 0\le y_0\le \frac{HB}{K\sqrt{R}},\\ (\frac{\tau B^2}{\sqrt{R}},\frac{\tau B^2}{\sqrt{R}}) & \text{if } \frac{H}{K\sqrt{R}}\le \tau\le \frac{H}{\sqrt{R}}, \zeta_\star\le \tau B,\\ (\frac{HB^2}{R},\frac{HB^2}{R}), & \text{if } \frac{H}{\sqrt{R}}\le \tau\le H, 0\le \zeta_\star\le \frac{HB}{\sqrt{R}},\\ (\frac{(H\zeta_\star^2B^4)^{1/3}}{R^{2/3}},\frac{(H\zeta_\star^2B^4)^{1/3}}{R^{2/3}}), & \text{if } \frac{HB}{\sqrt{R}}\le \zeta_\star\le HB,\\ (\frac{\zeta_\star B}{\sqrt{R}},\max\{\frac{HB^2}{KR},\frac{\tau B^2}{\sqrt{R}},\frac{\zeta_\star^2}{H}), & \text{if } \tau B\le \zeta_\star, \frac{HB}{K\sqrt{R}}\le \zeta_\star\le \frac{HB}{\sqrt{R}}. \end{cases}$$

As we can see, the only regime when our lower and upper bounds do not match is $\tau B\le\zeta_\star$, $\frac{HB}{K\sqrt{R}}\le \zeta_\star\le \frac{HB}{\sqrt{R}}$.

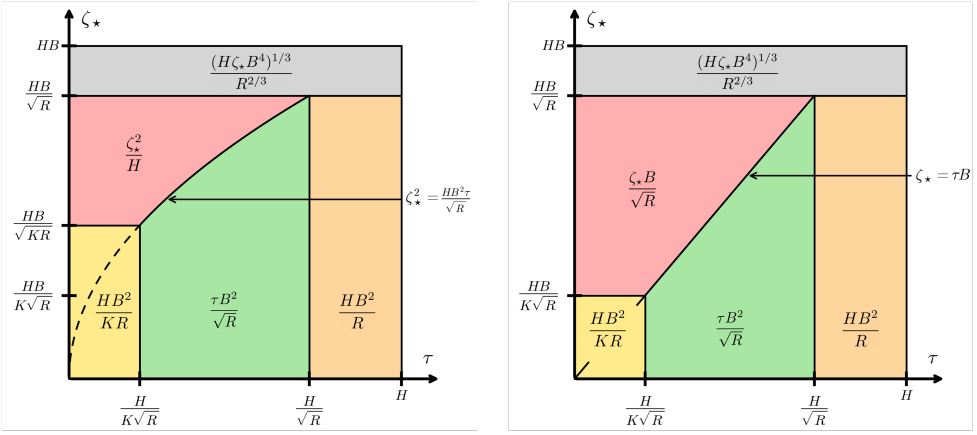

Figure H.1: Phase diagrams of the lower (left) and upper (right) bounds provided in this work: see Table 1 for the exact rates and Section H.1 and H.2 for derivations of the diagrams. The merged phase diagram is reported in Figure 1.

I Phase Diagram: Comparison of Prior Lower and Upper Bounds

In this section, we provide phase diagrams in $\zeta_\star$-τ diagram for the previous upper bound of Koloskova et al. [31] and lower bound of Patel et al. [1]. Again, we consider the noiseless regime $\sigma=0$, and make a restriction $\zeta_\star\le HB$.

In the aforementioned setting, the upper bound of Koloskova et al. [31] is

$$\mathbb{E}[F(\hat{x}_t)-F^\star]\lesssim \frac{HB^2}{R}+\frac{(H\zeta_\star^2B^4)^{1/3}}{R^{2/3}},$$

and the lower bound of Patel et al. [1] is

$$\mathbb{E}[F(\hat{x}_t)-F^\star]\gtrsim \frac{HB^2}{KR}+\frac{\tau B^2}{R}+\frac{(\tau\zeta_\star^2B^4)^{1/3}}{R^{2/3}},$$

for some $\hat{x}_t$.

We use the following notation $x_0 = \frac{\tau}{H}, y_0 = \frac{\zeta_\star \sqrt{R}}{HB}$, and $\alpha = \frac{HB^2}{R}$, then the bounds become

$$U_K = \alpha(1 + y_0^{2/3}), \quad \text{and} \quad L_P = \alpha(1/K + x_0 + x_0^{1/3} y_0^{2/3}).$$

I.1 Analysis of Dominant Term of the Upper Bound in Koloskova et al. [31]

The phase diagram for the upper bound is simple since it consists of only two terms, and it does not depend on τ at all. The switching point between the two terms is when they are equal:

$$\alpha = \alpha y_0^{2/3} \iff y_0 = 1.$$

Therefore, we have the following phase diagram for the upper bound

$$U_K(x_0, y_0) = \begin{cases} \alpha, & \text{if } y_0 \leq 1, \\ \alpha y_0^{2/3}, & \text{if } y_0 \geq 1. \end{cases}$$

In the standard notation $\zeta_\star$-τ, we have

$$U_K(\tau, \zeta_\star) = \begin{cases} \frac{HB^2}{R}, & \text{if } \zeta_\star \leq \frac{HB}{\sqrt{R}}, \\ \frac{(H\zeta_\star^2 B^4)^{1/3}}{R^{2/3}}, & \text{if } \zeta_\star \geq \frac{HB}{\sqrt{R}}. \end{cases}$$

I.2 Analysis of Dominant Term of the Lower Bound in Patel et al. [1]

Case 1: $0 \leq x_0 \leq \frac{1}{K}, x_0 y_0^2 \leq \frac{1}{K^3}$**.** In this case, we have $x_0 \leq \frac{1}{K}$ and $x_0^{1/3} y_0^{2/3} = (x_0 y_0^2)^{1/3} \leq \frac{1}{K}$. Therefore, we have that the lower bound satisfies

$$\frac{\alpha}{K} \leq L_P(x_0, y_0) = \frac{\alpha}{K} + \alpha x_0 + \alpha x_0^{1/3} y_0^{2/3} \leq \frac{3\alpha}{K},$$

i.e., $L_P \sim \frac{\alpha}{K}$ in this regime.

Case 2: $\frac{1}{K} \leq x_0, y_0 \leq x_0$**.** In this regime, we have $x_0 \geq \frac{1}{K}$ and $x_0^{1/3} y_0^{2/3} \leq x_0^{1/3} x_0^{2/3} = x_0$. Therefore, we have that the lower bound satisfies

$$\alpha x_0 \leq L_P(x_0, y_0) = \frac{\alpha}{K} + \alpha x_0 + \alpha x_0^{1/3} y_0^{2/3} \leq 3\alpha x_0,$$

i.e., $L_P \sim \alpha x_0$ in this regime.

Case 3: $x_0 \leq y_0, \frac{1}{K^3} \leq x_0 y_0^2$**.** In this regime, we have that $x_0 \leq x_0^{1/3} y_0^{2/3}$ and $\frac{1}{K} \leq (x_0 y_0^2)^{1/3} = x_0^{1/3} y_0^{2/3}$. Therefore, we have that the lower bound satisfies

$$\alpha x_0 y_0^{2/3} \leq L_P(x_0, y_0) = \frac{\alpha}{K} + \alpha x_0 + \alpha x_0^{1/3} y_0^{2/3} \leq 3\alpha x_0^{1/3} y_0^{2/3},$$

i.e., $L_P \sim x_0^{1/3} y_0^{2/3}$.

Final split. Summarizing the above 3 cases together, the lower bound has the following phase diagram

$$L_P(x_0, y_0) = \begin{cases} \frac{\alpha}{K}, & \text{if } 0 \leq x_0 \leq \frac{1}{K}, x_0 y_0^2 \leq \frac{1}{K^3}, \\ \alpha x_0, & \text{if } \frac{1}{K} \leq x_0, y_0 \leq x_0, \\ \alpha x_0^{1/3} y_0^{2/3}, & \text{if } x_0 \leq y_0, \frac{1}{K^3} \leq x_0 y_0^2. \end{cases}$$

In the original notation $\zeta_\star$-τ, the diagram has the following form

$$L_P(\tau, \zeta_\star) = \begin{cases} \frac{HB^2}{KR}, & \text{if } 0 \leq \tau \leq \frac{H}{K}, \tau\zeta_\star^2 \leq \frac{H^3 B^2}{K^3 R}, \\ \frac{\tau B^2}{R}, & \text{if } \frac{H}{K} \leq \tau, \zeta_\star \leq \frac{B\tau}{\sqrt{R}}, \\ \frac{(\tau\zeta_\star^2 B^4)^{1/3}}{R^{2/3}}, & \text{if } \frac{B\tau}{\sqrt{R}} \leq \zeta_\star, \frac{H^3 B^2}{K^3 R} \leq \tau\zeta_\star^2. \end{cases}$$

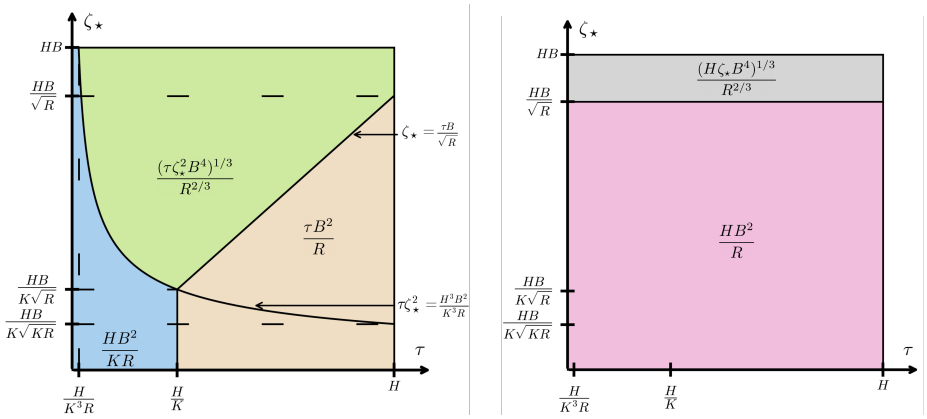

Figure I.1: Phase diagrams of the lower bound of Patel et al. [1] (left) and the upper bound of Koloskova et al. [31] (right): see Table 1 for the exact rates and Section I.1 and I.2 for derivations of the diagrams.

J Fresh-sample SGD Lower Bound in the Heterogeneous Smoothness Regime

In this appendix, we prove the lower bound stated informally in Theorem 4. The proof has two independent ingredients. First, we construct a deterministic rare-curvature quadratic block that gives the optimization lower bound $\hat{H}B^2/(TM)$ while satisfying

$$H \cong \tau \cong \frac{\hat{H}}{\sqrt{M}}.$$

Second, we add the standard stochastic and first-order heterogeneity lower-bound block on orthogonal coordinates to obtain the noise term.

J.1 A rare-curvature block

We begin with the deterministic block. Let $M \geq 2$, $T \geq 1$, and let $\lambda > 0$. Later we will set $\hat{H} = M\lambda$, and the lower bound will scale as $\lambda B^2/T = \hat{H}B^2/(TM)$.

Let u, v be two orthonormal vectors. Define the positive semidefinite matrix A by

$$Au = \lambda u, \qquad Av = \mu v, \qquad \mu := \frac{\lambda}{12T}.$$

On this block, define

$$R_1(y) := \frac{M}{2} y^\top A y, \qquad R_m(y) := 0 \quad \text{for } m = 2, \dots, M.$$

Then the average objective is

$$R(y) := \frac{1}{M}\sum_{m=1}^{M} R_m(y) = \frac{1}{2} y^\top A y.$$

Thus R is λ-smooth and minimized at 0, while the largest component smoothness on this block is $M\lambda$.

The averaged second-order heterogeneity of this block is

$$\frac{1}{M}\sum_{m=1}^{M} \left\| \big(\nabla R_m(y) - \nabla R_m(z)\big) - \big(\nabla R(y) - \nabla R(z)\big) \right\|_2^2 = (M-1)\left\|A(y-z)\right\|_2^2$$
$$\leq (M-1)\lambda^2 \left\|y-z\right\|_2^2.$$

Hence the block has second-order heterogeneity parameter

$$\tau_R = \sqrt{M-1}\,\lambda.$$

We initialize the rare-curvature block at[12]

$$y_0 := \frac{B_R}{\sqrt{2}}(u+v),$$

where B_R is the radius allocated to this block. We now lower bound the error of fresh-sample SGD with replacement on this block.

For an eigendirection of A with eigenvalue $a \in \{\lambda, \mu\}$, define the scalar coordinate

$$\theta_t := \langle y_t, w \rangle,$$

where $w \in \{u, v\}$ is the corresponding eigenvector. At each step, the algorithm samples the unique nonzero component with probability $1/M$. If that component is sampled, then the coordinate is multiplied by

$$q(a) := 1 - \eta M a;$$

otherwise the coordinate is unchanged.

Let $B_s \in \{0, 1\}$ be the indicator that the nonzero component is sampled at step s, and define

$$S_t := \sum_{s=0}^{t-1} B_s.$$

Thus S_t is the number of times the nonzero component has been sampled before iterate t. Since the samples are drawn with replacement, the variables B_s are i.i.d. Bernoulli with mean $1/M$. Along an eigenvalue a, we have

$$\theta_t = \theta_0 q(a)^{S_t}.$$

Therefore the coordinate of the averaged iterate

$$\bar{y}_N := \frac{1}{N} \sum_{t=0}^{N-1} y_t, \qquad N := TM,$$

equals

$$\langle \bar{y}_N, w \rangle = \theta_0 Z_N(q(a)), \qquad Z_N(q) := \frac{1}{N} \sum_{t=0}^{N-1} q^{S_t}.$$

We need the following elementary estimate.

Lemma 19 (Averaged multiplier lower bound)**.** *Let $M \geq 2$, $T \geq 1$, and $N = MT$. Let $B_0, \ldots, B_{N-1}$ be independent Bernoulli random variables with*

$$\mathbb{P}(B_t = 1) = \frac{1}{M}.$$

Define

$$S_t := \sum_{r=0}^{t-1} B_r, \qquad Z_N(q) := \frac{1}{N} \sum_{t=0}^{N-1} q^{S_t}.$$

Then the following two claims hold.

1. *If $q \in [1 - 1/4T, 1]$, then*
$$\mathbb{E}[Z_N(q)^2] \geq c_1.$$
2. *If $q \leq -2$, then*
$$\mathbb{E}[Z_N(q)^2] \geq \frac{c_2}{T}.$$

Here $c_1, c_2 > 0$ are universal numerical constants.

Proof. We prove the two claims separately.

[12] Note that the assumption about bounded optima in (4) can without loss of generality be thought of as an assumption bounding $\|x_0 - x^\star\|_2$. This is standard in the literature.

Case 1: $q \in [1 - \frac{1}{4T}, 1]$**.** Since $q \in [0, 1]$, we have

$$\mathbb{E}[q^{S_t}] = \left(1 - \frac{1-q}{M}\right)^t.$$

Moreover,

$$1 - q \le \frac{1}{4T},$$

and hence

$$1 - \frac{1-q}{M} \ge 1 - \frac{1}{4TM} = 1 - \frac{1}{4N}.$$

Therefore, for every $t \in \{0, \dots, N-1\}$,

$$\mathbb{E}[q^{S_t}] \ge \left(1 - \frac{1}{4N}\right)^N \ge c$$

for a universal constant $c > 0$. Thus

$$\mathbb{E}[Z_N(q)] = \frac{1}{N} \sum_{t=0}^{N-1} \mathbb{E}[q^{S_t}] \ge c.$$

By Jensen's inequality,

$$\mathbb{E}[Z_N(q)^2] \ge \big(\mathbb{E}[Z_N(q)]\big)^2 \ge c^2.$$

This proves the first claim.

Case 2: $q \le -2$**.** Let

$$Y_t := q^{B_t}.$$

Then

$$q^{S_t} = \prod_{r=0}^{t-1} Y_r.$$

Define

$$P_t := q^{S_t} = \prod_{r=0}^{t-1} Y_r, \qquad A_N := \sum_{t=0}^{N-1} P_t.$$

Thus

$$Z_N(q) = \frac{A_N}{N}.$$

Let

$$\beta := \mathbb{E}[Y_t] = 1 + \frac{q-1}{M}, \qquad \alpha := \mathbb{E}[Y_t^2] = 1 + \frac{q^2-1}{M}.$$

Also let

$$v := \mathrm{Var}(Y_t).$$

Since $Y_t = 1$ with probability $1 - \frac{1}{M}$ and $Y_t = q$ with probability $\frac{1}{M}$, we have

$$v = \frac{1}{M}\left(1 - \frac{1}{M}\right)(q-1)^2.$$

Let $\mathcal{F}_s := \sigma(Y_0, \dots, Y_{s-1})$, where $Y_r := q^{B_r}$, and recall that

$$P_t = q^{S_t} = \prod_{r=0}^{t-1} Y_r, \qquad Z_N(q) = \frac{1}{N} \sum_{t=0}^{N-1} P_t.$$

For $s = 0, \dots, N-2$, set

$$D_s := \mathbb{E}[Z_N(q) \mid \mathcal{F}_{s+1}] - \mathbb{E}[Z_N(q) \mid \mathcal{F}_s].$$

We first compute D_s. The variable Y_s can affect only those products P_t with $t \geq s+1$. For such t,

$$P_t = P_s Y_s \prod_{r=s+1}^{t-1} Y_r.$$

Conditioning on $\mathcal{F}_{s+1}$, the future variables $Y_{s+1}, \ldots, Y_{t-1}$ remain independent and have mean β, so

$$\mathbb{E}[P_t \mid \mathcal{F}_{s+1}] = P_s Y_s \beta^{t-s-1}.$$

Conditioning only on $\mathcal{F}_s$, the variable Y_s has not yet been revealed and has mean β, hence

$$\mathbb{E}[P_t \mid \mathcal{F}_s] = P_s \beta^{t-s}.$$

Therefore,

$$\mathbb{E}[P_t \mid \mathcal{F}_{s+1}] - \mathbb{E}[P_t \mid \mathcal{F}_s] = P_s(Y_s - \beta)\beta^{t-s-1}.$$

The terms P_t with $t \leq s$ do not depend on Y_s, so they cancel in the difference of conditional expectations. Thus

$$\begin{aligned} D_s &= \frac{1}{N} \sum_{t=s+1}^{N-1} P_s(Y_s - \beta)\beta^{t-s-1} \\ &= \frac{P_s(Y_s - \beta)}{N} \sum_{r=0}^{N-s-2} \beta^r. \end{aligned}$$

Since P_s is $\mathcal{F}_s$-measurable and Y_s is independent of $\mathcal{F}_s$,

$$\begin{aligned} \mathbb{E}[D_s^2] &= \frac{\mathbb{E}[P_s^2] \operatorname{Var}(Y_s)}{N^2} \left(\sum_{r=0}^{N-s-2} \beta^r \right)^2 \\ &= \frac{\alpha^s v}{N^2} \left(\sum_{r=0}^{N-s-2} \beta^r \right)^2. \end{aligned}$$

To relate these increments to the variance of $Z_N(q)$, define

$$M_s := \mathbb{E}[Z_N(q) \mid \mathcal{F}_s], \qquad s = 0, \ldots, N.$$

This is the Doob martingale obtained by revealing the variables $Y_0, Y_1, \ldots, Y_{N-1}$ one at a time. Indeed, M_s is $\mathcal{F}_s$-measurable, and by the tower property,

$$\mathbb{E}[M_{s+1} \mid \mathcal{F}_s] = \mathbb{E}[\mathbb{E}[Z_N(q) \mid \mathcal{F}_{s+1}] \mid \mathcal{F}_s] = \mathbb{E}[Z_N(q) \mid \mathcal{F}_s] = M_s.$$

Thus $(M_s)_{s=0}^N$ is a martingale. By the definition of D_s,

$$D_s = \mathbb{E}[Z_N(q) \mid \mathcal{F}_{s+1}] - \mathbb{E}[Z_N(q) \mid \mathcal{F}_s] = M_{s+1} - M_s.$$

Hence $\mathbb{E}[D_s \mid \mathcal{F}_s] = 0$. If $r < s$, then D_r is $\mathcal{F}_s$-measurable, so

$$\mathbb{E}[D_r D_s] = \mathbb{E}[D_r \, \mathbb{E}[D_s \mid \mathcal{F}_s]] = 0.$$

Thus the D_s's are pairwise orthogonal. Also, $D_{N-1} = 0$, since $Z_N(q)$ depends only on $Y_0, \ldots, Y_{N-2}$. Therefore

$$Z_N(q) - \mathbb{E}[Z_N(q)] = \sum_{s=0}^{N-2} D_s,$$

and by orthogonality,

$$\operatorname{Var}(Z_N(q)) = \sum_{s=0}^{N-2} \mathbb{E}[D_s^2].$$

Combining this identity with the formula for $\mathbb{E}[D_s^2]$ gives

$$\operatorname{Var}(Z_N(q)) = \frac{v}{N^2} \sum_{s=0}^{N-2} \alpha^s \left(\sum_{r=0}^{N-s-2} \beta^r \right)^2.$$

Since $\alpha \geq 1$, we can drop the factor α^s and re-index $\ell = N - s - 2$ to obtain

$$\mathrm{Var}(Z_N(q)) \geq \frac{v}{N^2} \sum_{\ell=0}^{N-2} \left(\sum_{r=0}^{\ell} \beta^r \right)^2.$$

We now lower bound the last sum. Define

$$\delta := 1 - \beta = \frac{1-q}{M}.$$

Since $q \leq -2$, we have

$$\delta \geq \frac{3}{M}.$$

Also, because $N = MT$,

$$N\delta \geq 3T \geq 3.$$

We claim that

$$\delta^2 \sum_{\ell=0}^{N-2} \left(\sum_{r=0}^{\ell} \beta^r \right)^2 \geq cN$$

for a universal constant $c > 0$. To prove the claim, consider two cases.

If $0 \leq \beta < 1$, then

$$\sum_{r=0}^{\ell} \beta^r = \frac{1-\beta^{\ell+1}}{1-\beta} = \frac{1-\beta^{\ell+1}}{\delta}.$$

For all $\ell + 1 \geq 1/\delta$, we have

$$\beta^{\ell+1} \leq e^{-\delta(\ell+1)} \leq e^{-1}.$$

Thus

$$\delta \sum_{r=0}^{\ell} \beta^r \geq 1 - e^{-1}.$$

Since $\delta \geq 3/M$, we have

$$\frac{1}{\delta} \leq \frac{M}{3} \leq \frac{N}{3}.$$

Hence a constant fraction of the indices $\ell \in \{0, \ldots, N-2\}$ satisfy $\ell + 1 \geq 1/\delta$, and the claimed bound follows.

If $\beta < 0$, then for every even ℓ, the integer $\ell + 1$ is odd, so $\beta^{\ell+1} < 0$. Therefore

$$\sum_{r=0}^{\ell} \beta^r = \frac{1-\beta^{\ell+1}}{1-\beta} \geq \frac{1}{\delta}.$$

At least a constant fraction of the indices $\ell \in \{0, \ldots, N-2\}$ are even, and again

$$\delta^2 \sum_{\ell=0}^{N-2} \left(\sum_{r=0}^{\ell} \beta^r \right)^2 \geq cN.$$

This proves the claim.

It remains to lower bound v. Since

$$q - 1 = -M\delta,$$

we have

$$\begin{aligned} v &= \frac{1}{M} \left(1 - \frac{1}{M} \right) (q-1)^2 \\ &= (M-1)\delta^2 \geq \frac{M}{2}\delta^2, \end{aligned}$$

where we used $M \geq 2$. Combining this with the claim gives

$$\begin{aligned}\mathrm{Var}(Z_N(q)) &\geq \frac{v}{N^2} \sum_{\ell=0}^{N-2} \left(\sum_{r=0}^{\ell} \beta^r \right)^2 \\ &\geq \frac{cM\delta^2}{N^2} \cdot \frac{N}{\delta^2} = \frac{cM}{N} = \frac{c}{T}.\end{aligned}$$

Finally,

$$\mathbb{E}[Z_N(q)^2] \geq \mathrm{Var}(Z_N(q)) \geq \frac{c}{T}.$$

This proves the second claim and completes the proof. □

We now prove the lower bound for the rare-curvature block.

Lemma 20 (Rare-curvature lower bound for fresh-sample SGD)**.** *For the block R defined above, for every constant stepsize $\eta > 0$, fresh-sample SGD with replacement satisfies*

$$\mathbb{E}\big[R(\bar{y}_N) - R(0)\big] \geq c\frac{\lambda B_R^2}{T}, \qquad N = TM,$$

where $c > 0$ is a universal constant.

Proof. We split into two stepsize regimes.

First suppose that $\eta M\lambda < 3$. Consider the low-curvature direction v, whose eigenvalue is $\mu = \lambda/(12T)$. Then

$$1 - \eta M\mu > 1 - \frac{1}{4T}.$$

Thus $q(\mu) \in [1 - \frac{1}{4T}, 1]$, and Lemma 19 gives

$$\mathbb{E}[Z_N(q(\mu))^2] \geq c.$$

The initial coordinate in the v-direction is $B_R/\sqrt{2}$. Hence

$$\begin{aligned}\mathbb{E}[R(\bar{y}_N) - R(0)] &\geq \frac{\mu}{2}\mathbb{E}[\langle \bar{y}_N, v\rangle^2] \\ &= \frac{\mu}{2} \cdot \frac{B_R^2}{2}\mathbb{E}[Z_N(q(\mu))^2] \\ &\geq c\mu B_R^2 = c\frac{\lambda B_R^2}{T}.\end{aligned}$$

Now suppose that $\eta M\lambda \geq 3$. Consider the high-curvature direction u. Then

$$q(\lambda) = 1 - \eta M\lambda \leq -2.$$

By Lemma 19,

$$\mathbb{E}[Z_N(q(\lambda))^2] \geq \frac{c}{T}.$$

Since the initial u-coordinate is $B_R/\sqrt{2}$, we get

$$\begin{aligned}\mathbb{E}[R(\bar{y}_N) - R(0)] &\geq \frac{\lambda}{2}\mathbb{E}[\langle \bar{y}_N, u\rangle^2] \\ &= \frac{\lambda}{2} \cdot \frac{B_R^2}{2}\mathbb{E}[Z_N(q(\lambda))^2] \\ &\geq c\frac{\lambda B_R^2}{T}.\end{aligned}$$

The two regimes cover all $\eta > 0$, proving the claim. □

J.2 Calibration of the parameters

We now calibrate the block to the heterogeneous smoothness regime. Let

$$\lambda := \frac{\hat{H}}{M}.$$

Then the rare block has maximum component smoothness $\hat{H}$, average-objective smoothness $\lambda = \hat{H}/M$, and second-order heterogeneity

$$\tau_R = \sqrt{M-1}\lambda \cong \frac{\hat{H}}{\sqrt{M}}.$$

In the regime of Theorem 4, we assume

$$H \cong \tau \cong \frac{\hat{H}}{\sqrt{M}}.$$

Since $\lambda = \hat{H}/M \leq H$, the rare block is smoother than the allowed average-objective smoothness H.

If one wants the full instance to have average-objective smoothness exactly of order H, we add an orthogonal calibration block

$$C_m(s) := \frac{H}{2}s^2, \qquad m = 1, \dots, M,$$

and initialize this coordinate at its minimizer $s = 0$. This block makes the average objective H-smooth, but contributes no error because its gradient is identically zero along the trajectory. It also does not change τ, since it is identical across clients. The maximum client smoothness of the full instance remains of order $\hat{H}$.

J.3 Adding the noise block

We also add an orthogonal standard stochastic/first-order-heterogeneity block. We use the following classical lower bound.

Lemma 21 (Standard stochastic and first-order heterogeneity lower bound)**.** *For every $N \geq 1$, $B > 0$, and $\sigma, \zeta_\star \geq 0$, there exists a convex H-smooth stochastic optimization instance with optimum norm at most B, within-client noise level at most σ, and first-order heterogeneity at the optimum at most $\zeta_\star$, such that any fresh-sample SGD method using N samples and outputting an averaged iterate satisfies*

$$\mathbb{E}[F_{\mathrm{noise}}(\bar{x}_N) - F_{\mathrm{noise}}(x^\star)] \geq c \cdot \min\left\{HB^2, \frac{(\sigma+\zeta_\star)B}{\sqrt{N}}\right\}.$$

This is the standard mean-estimation lower bound; the $\zeta_\star$ term is obtained by taking client-dependent linear perturbations at the optimum, and the σ term is obtained by adding independent zero-mean stochastic gradient noise.

J.4 Proof of the theorem

Theorem 11 (Fresh-sample SGD lower bound in the heterogeneous smoothness regime)**.** *Fix $M, T \geq 2$, $B > 0$, and parameters satisfying*

$$\tau \cong H \cong \frac{\hat{H}}{\sqrt{M}}.$$

There exists an instance in $\mathcal{P}^{H,\hat{H},B,\sigma}_{\zeta_\star,\tau}$ such that, for every constant stepsize $\eta > 0$, fresh-sample SGD with replacement satisfies

$$\mathbb{E}\big[F(\bar{x}_{TM}) - F(x^\star)\big] \geq c\left(\frac{\hat{H}B^2}{TM} + \min\left\{HB^2, \frac{(\sigma+\zeta_\star)B}{\sqrt{TM}}\right\}\right).$$

Proof. Set $N = TM$. Place the rare-curvature block, the calibration block, and the noise block on orthogonal coordinates. Allocate a constant fraction of the radius budget B to the rare-curvature block and another constant fraction to the noise block. The calibration block is initialized at its optimum and uses no radius.

By Lemma 20 with $\lambda = \hat{H}/M$, the rare-curvature block contributes

$$c\frac{\lambda B^2}{T} = c\frac{\hat{H}B^2}{TM}.$$

By Lemma 21, the noise block (which is also ensured to be smooth) contributes

$$c \cdot \min\left\{HB^2, \frac{(\sigma + \zeta_\star)B}{\sqrt{TM}}\right\}.$$

The calibration block ensures that the average objective has smoothness of order H, and it does not affect the error or the heterogeneity.

The full objective is the direct sum of the three blocks, and the algorithm evolves independently on the orthogonal coordinates. Therefore the suboptimalities add. The full instance has average-objective smoothness of order H, maximum client smoothness of order $\hat{H}$, and averaged second-order heterogeneity of order

$$\tau \cong \frac{\hat{H}}{\sqrt{M}} \cong H.$$

This proves the theorem, after absorbing the constant factors from the radius split into the universal constant c. □

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: This work studies Local SGD and presents a novel, tighter analysis. The main contributions are summarized in a separate paragraph in Section 1, highlighting the key results and their differences from prior work.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of our results and open questions in the paragraph titled "Limitations" at the end of the main text.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: All assumptions and their discussion are presented in Section 2. The main convergence results are stated as theorems in Section 3 and Section 4, with proofs deferred to the appendix. The proofs contain all necessary steps to obtain the convergence rates stated in the theorems.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [N/A]

Justification: [N/A]

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [N/A]

   Justification: [N/A]

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [N/A]

   Justification: [N/A]

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [N/A]

   Justification: [N/A]

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [N/A]

   Justification: [N/A]

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: We have reviewed the code of ethics, and the paper abides by it.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: While not non-standard we do discuss LLM usage in the acknowledgements.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.