

Supplementary material

A Supplementary experimental details

In this section we provide the details on the language model pretraining experiments discussed in the main text.

A.1 Language model pretraining

We study the impact of using various outer optimizers on large language model pretraining. We utilized Chinchilla-style decoder transformer architectures (Hoffmann et al., 2022) trained on the C4 dataset (Raffel et al., 2020), consistent with common practices in large-scale model training (Douillard, Feng, Rusu, Chhaparia, et al., 2023). The following subsections detail the specific hyperparameters, variations in training configurations (such as the number of inner steps and replicas/clients), and analyses of optimizer behavior, including learning rate scheduling and observed gradient cosine similarities.

A.1.1 Hyperparameters details

We show in Table 1 the hyperparameters considered and kept, and in Table 2 the architectural hyperparameters. We tuned all our optimizers on a separate validation set. We also considered using the Schedule-Free Optimizer with Nesterov acceleration on top but it was hard to tune and unstable.

Table 2: **Model Configuration** for the three evaluated sizes. All are based on the transformer architecture, chinchilla-style (Hoffmann et al., 2022).

Hyperparameter	150M	400M	1B
Number of layers	12	12	24
Hidden dim	896	1536	2048
Number of heads	16	12	16
K/V size	64	128	128
Vocab size	32,000		

A.1.2 Varying inner steps

In Figure 3, we compare the stability of different outer optimizers when varying the synchronization frequency. We experiments a different amount of inner steps, from 50, to 2000. All experiments are run in pretraining from scratch, with 150 millions (150M) parameters. We note that as the synchronization frequency decreases (number of inner/local steps increases), performance decreases. Notably, averaging (in orange), is relatively constant w.r.t the synchronization frequency: its performance stay stable from $H = 250$ to $H = 2000$. On the other hand, using Nesterov with high outer learning rate (in light green) is particularly unstable, its performance decreases by 10.7%, this indicates that the learning rate should be tuned alongside the synchronization frequency. On the hand, SF-SGD (in blue) has minimal degradation of performance (4.2%), highlighting the *schedule-free* property when varying hyperparameters.

A.1.3 Varying replicas / flops budget

When increasing the number of distributed replicas, two options are possible: (a) Keeping the local per-replica batch size constant and thus increasing global batch size and flops budget, and (b) Keeping the global batch size/flops budget constant and thus reducing the local per-replica batch size.

We present in Figure 4 results of the first option with x-axis the flops budget for a single model size (150M). It is worth noting that increasing the number of replicas improves the performance of Nesterov (in green) and SF-SGD (in blue) but the gain quickly plateau. On the other hand, increasing the batch size for data-parallel (at the cost of more communication, because more DP replicas) or the number of steps (at the cost of longer training) still rapidly improves perplexity. Therefore, we

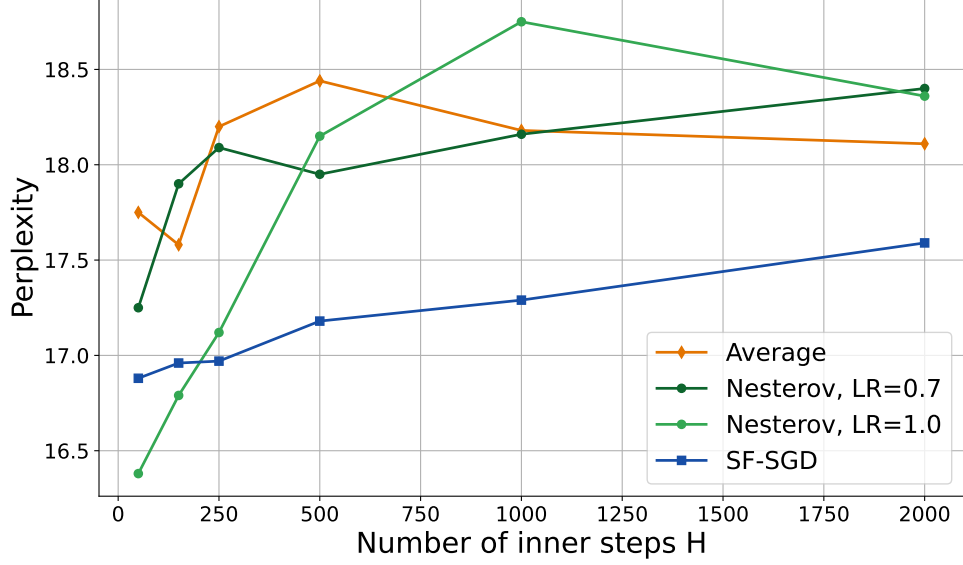


Figure 3: **Varying the communication frequency**, i.e. number of inner steps H , when pretraining from scratch at 150M parameters.

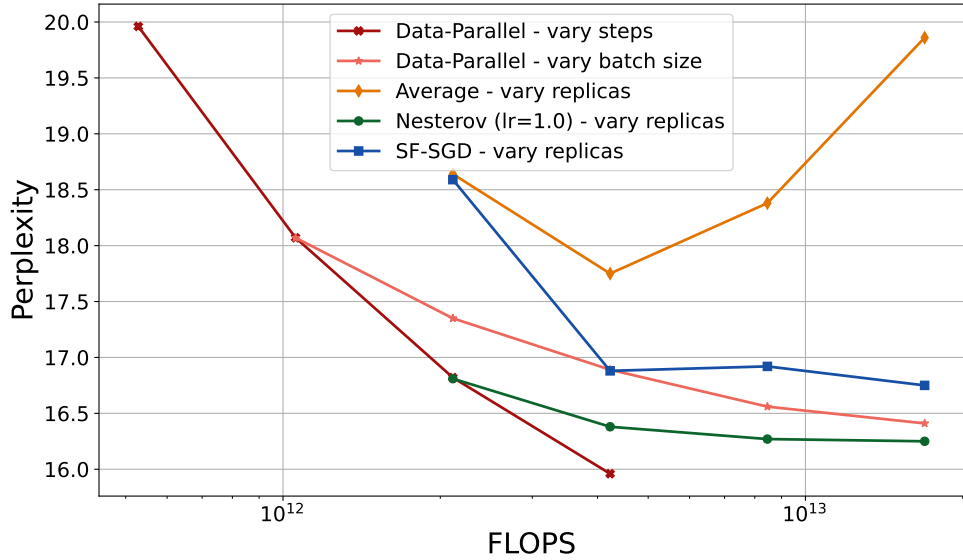


Figure 4: **Pareto front** of the flops vs perplexity, comparing various approach scaling the flops budget: increasing the number of steps, increasing the batch size in data-parallel, and increasing the number of replicas for federated learning.

950 wish to highlight here a disadvantage of federated learning methods seldom mentioned: while those
 951 methods are extremely communication-efficient, and can be made flops-efficient, their flops-efficiency
 952 disappear as the number of replicas increases.

953 To this problem, several hypotheses could be raised, such as the decreasing cosine similarity between
 954 outer gradients as the number of replicas increase, even when using an *i.i.d.* data split across replicas.
 955 In Figure 5, we report the average similarity across a whole training for different number of replicas.
 956 For momentum-based methods (Nesterov, SF-SGD), the similarity decreases from 30% at $M = 2$
 957 replicas to 10% at $M = 16$ replicas. Full details across training steps can be found in the appendix.

958 Finally, note that we didn't investigate further the second option of keeping the global batch size/flops
 959 budget constant and thus reducing the local per-replica batch size. We found that dividing the batch
 960 size by the number of replicas leads quickly to a local per-replica batch size that is critically low, and
 961 further reduces the flops-efficiency. More investigations should be pushed in that direction.

A.1.4 Schedule-free but not tuning-free

The *schedule-free* method of Defazio et al., 2024 enables not doing any learning rate scheduling, greatly simplifying training configuration. However, it doesn't mean it is *hyperparameters-tuning-free*. Indeed, we found out that we had to extensively tune the initial learning rate (to 2.0), remove learning rate warm-up contrarily to what is advised, and use a particularly low $b1$ decay: 0.2, as illustrated in Figure 6.

A.1.5 Pretraining: outer learning rate scheduling

Schedule-free SGD enables not having to manually scheduling the outer learning rate. Therefore, we wondered if we could improve the SotA federated learning baseline, DiLoCo (Nesterov outer optimizer), with an outer learning rate schedule. We investigate in Figure 7 three schedules: *constant* as in (Douillard, Feng, Rusu, Chhaparia, et al., 2023), *cosine decay*, and *linear after a plateau*. For the latter we consider a constant plateau for 10% and 25% of the total steps. For each method, we also tuned the peak outer learning rate. We don't use any warm-up in the outer optimization as we always found it to be harmful.

We find that constant outer learning rate is the best performing schedule. It's unclear how the other schedules are interacting with the inner learning rate scheduling. A possible solution, not investigated in this report, would be to increase the number of inner steps H as the inner learning rate decreases (Gu, Lyu, Arora, et al., 2024).

A.1.6 Cosine similarity between outer gradients

We display the cosine similarity between outer gradients, across scales (150M, 400M, and 1B) in Figure 8, and across replicas (for 150M, from 2 to 16 replicas) in Figure 9. The solid line represent the mean, and the shaded area the standard deviation. We normalize the x-axis as a percentage of the training in order to compare models which have done different amount of steps (e.g. 24,000 steps for 150M vs 30,000 for 400M).

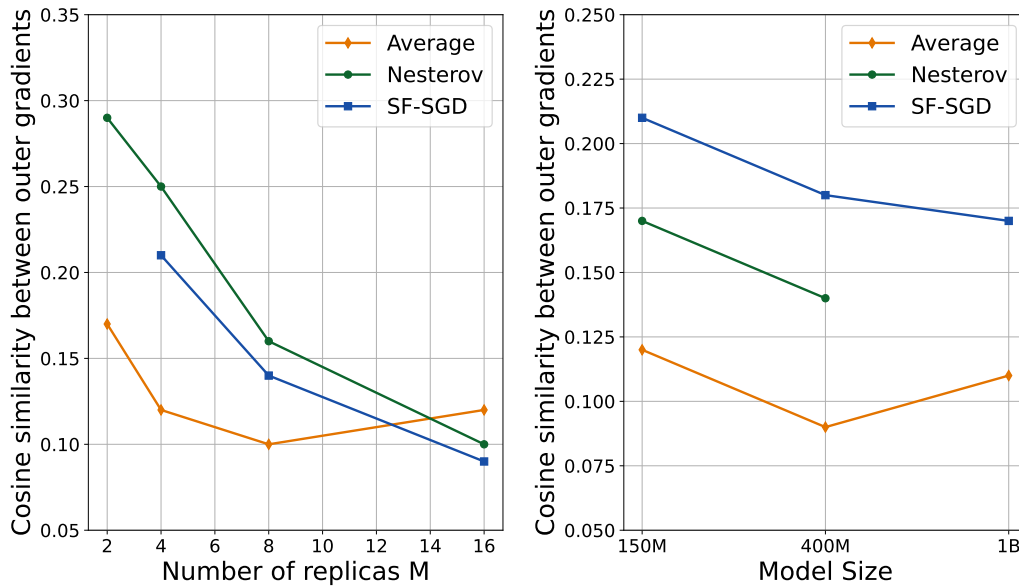


Figure 5: **Cosine similarity** between outer gradients across different number of replicas (*left*) and model scales (*right*). We average the similarity across the middle 50% of the training.

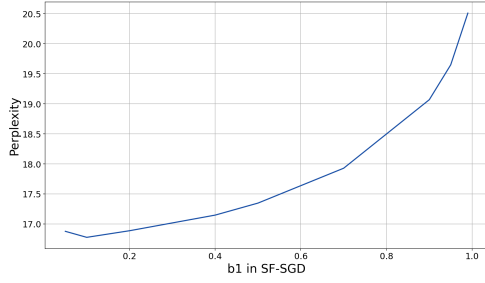


Figure 6: **Tuning b1 decay** has a major impact on performance, and its value must be very low.

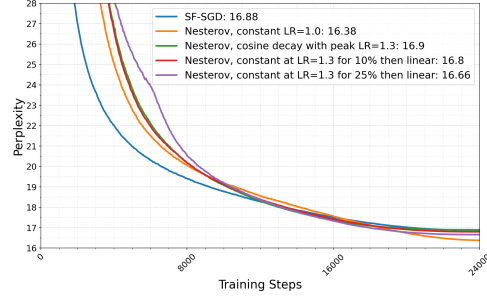


Figure 7: Which outer **learning rate schedule** to use?

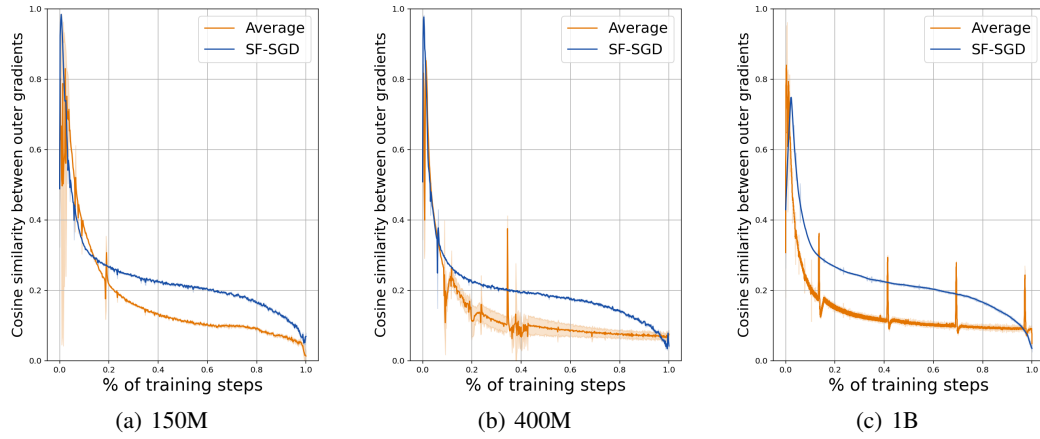


Figure 8: **Similarity** between outer gradients across steps and scales.

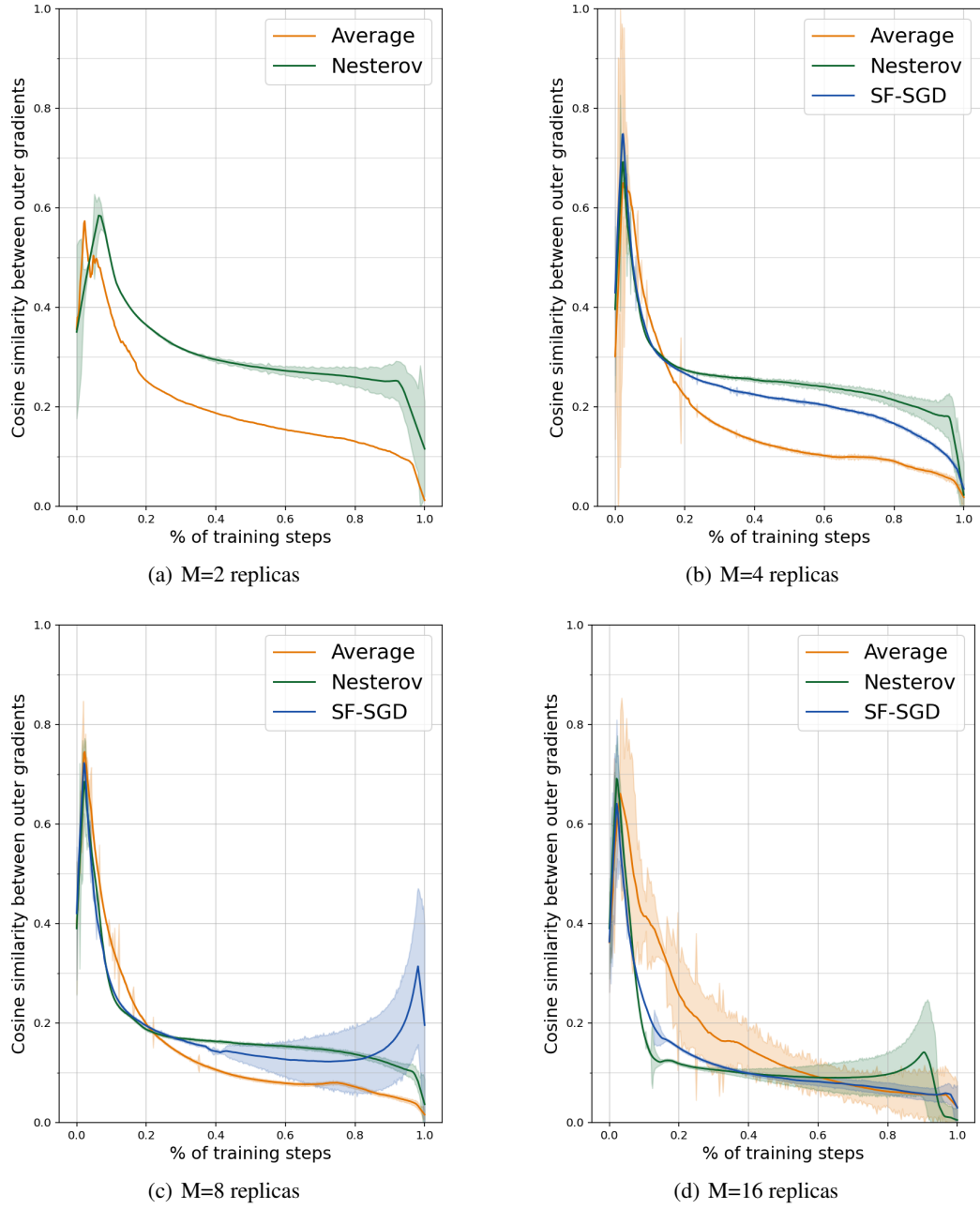


Figure 9: **Cosine similarity** between outer gradients across steps and number of replicas.

Theory

B Guarantees for Local SGD

First, we recall our setting and define some notation. We consider the problem of minimizing a function f in a distributed setting with M workers performing Local SGD. Let x_r denote the global model parameters at the beginning of round r . Each worker m initializes its local parameters as $y_{m,r,0} = x_r$ and performs H local SGD steps according to

$$y_{m,r,h+1} = y_{m,r,h} - \eta g_{m,r,h},$$

where $g_{m,r,h} = \nabla f(y_{m,r,h}) + n_{m,r,h}$ is the stochastic gradient with noise $n_{m,r,h}$, and $\bar{g}_{m,r,h} = \nabla f(y_{m,r,h})$ is the true gradient. By Assumption 3.2 we have $\mathbb{E}[g_{m,r,h}] = \bar{g}_{m,r,h}$. After H local steps, the global model update can be equivalently written as $x_{r+1} = x_r - \gamma \eta \sum_{h=0}^{H-1} g_{r,h}$ where $g_{r,h} = \frac{1}{M} \sum_{m=1}^M g_{m,r,h}$ is the average gradient across workers and $y_{r,h} = \frac{1}{M} \sum_{m=1}^M y_{m,r,h}$ is the average model. Note that these two last sequences are virtual sequences and not actually computed. We also define $x_{r,h} = x_r - \gamma \eta \sum_{h=0}^{H-1} g_{r,h}$ as an intermediate quantity used in the analysis.

B.1 Algorithm-independent results

Lemma B.1. (Karimireddy et al., 2020, Lemma 6) *Let f be a convex and L -smooth function. Suppose that $\eta \leq \frac{2}{L}$, let $T_\eta(x) = x - \eta \nabla f(x)$. Then*

$$\|T_\eta(x) - T_\eta(y)\|^2 \leq \|x - y\|^2.$$

Proof. The proof is provided for completeness only. We have

$$\|T_\eta(x) - T_\eta(y)\|^2 = \|x - y\|^2 + \eta^2 \|\nabla f(x) - \nabla f(y)\|^2 - 2\eta \langle x - y, \nabla f(x) - \nabla f(y) \rangle. \quad (8)$$

By the Baillon-Haddad theorem (Bauschke and Combettes, 2009) we have

$$\langle x - y, \nabla f(x) - \nabla f(y) \rangle \geq \frac{1}{L} \|\nabla f(x) - \nabla f(y)\|^2.$$

Using this in Equation (8) gives

$$\|T_\eta(x) - T_\eta(y)\|^2 \leq \|x - y\|^2 - \eta \left(\frac{2}{L} - \eta \right) \|\nabla f(x) - \nabla f(y)\|^2.$$

If $\eta \leq \frac{1}{L}$ then $\frac{2}{L} - \eta \geq 0$ and therefore $\|T_\eta(x) - T_\eta(y)\|^2 \leq \|x - y\|^2$. \square

Lemma B.2. *Let y_1, \dots, y_n be real numbers. Then,*

$$\frac{1}{n} \sum_{k=1}^n |y_i| \leq \sqrt{\frac{1}{n} \sum_{k=1}^n y_i^2}.$$

Proof. This is just the arithmetic mean-root mean square inequality and we include the proof solely for completeness. Let Y be a random variable that takes the value y_i^2 with probability $\frac{1}{n}$, and let $g(x) = \sqrt{x}$. Observe that

$$\frac{1}{n} \sum_{k=1}^n |y_i| = \mathbb{E}[g(Y)].$$

Since g is a concave function, by Jensen's inequality we have that $\mathbb{E}[g(Y)] \leq g(\mathbb{E}[Y])$. Therefore,

$$\frac{1}{n} \sum_{k=1}^n |y_i| = \mathbb{E}[g(Y)] \leq g(\mathbb{E}[Y]) = \sqrt{\frac{1}{n} \sum_{k=1}^n y_i^2}.$$

1006 \square

1007 **Lemma B.3.** (Variance of Sum of Conditionally Independent Random Variables). Let Z_1, \dots, Z_n be
 1008 random variables such that Z_i satisfies

$$\mathbb{E}_{i-1}[Z_i] = 0, \quad \text{and}, \quad \mathbb{E}[\|Z_i\|^2] = \sigma_i^2,$$

1009 where $\mathbb{E}_i[\cdot]$ denotes expectation conditional on Z_1, Z_2, \dots, Z_i . Then,

$$\mathbb{E}\left[\left\|\sum_{i=1}^n Z_i\right\|^2\right] = \sum_{i=1}^n \sigma_i^2.$$

Proof.

$$\begin{aligned} \mathbb{E}\left[\left\|\sum_{i=1}^n Z_i\right\|^2\right] &= \mathbb{E}\left[\mathbb{E}_{n-1}\left[\left\|\sum_{i=1}^n Z_i\right\|^2\right]\right] \\ &= \mathbb{E}\left[\mathbb{E}_{n-1}\left[\left\|\sum_{i=1}^{n-1} Z_i\right\|^2 + \|Z_n\|^2 + 2\left\langle\sum_{i=1}^{n-1} Z_i, Z_n\right\rangle\right]\right] \\ &= \mathbb{E}\left[\mathbb{E}_{n-1}\left[\left\|\sum_{i=1}^{n-1} Z_i\right\|^2\right] + \sigma_n^2\right]. \end{aligned}$$

1010 The cross-term $\mathbb{E}_{n-1}\left[2\left\langle\sum_{i=1}^{n-1} Z_i, Z_n\right\rangle\right]$ vanishes because $\mathbb{E}_{n-1}[Z_n] = 0$ and $\sum_{i=1}^{n-1} Z_i$ is mea-
 1011 surable with respect to the sigma-algebra generated by Z_1, \dots, Z_{n-1} . Continuing,

$$\mathbb{E}\left[\left\|\sum_{i=1}^n Z_i\right\|^2\right] = \mathbb{E}\left[\left\|\sum_{i=1}^{n-1} Z_i\right\|^2\right] + \sigma_n^2.$$

1012 Recursing we get,

$$\mathbb{E}\left[\left\|\sum_{i=1}^n Z_i\right\|^2\right] = \sum_{i=1}^n \sigma_i^2.$$

1013 This completes the proof. \square

1014 **Lemma B.4.** (Ivgy, Hinder, and Carmon, 2023, Lemma 7). Let S be the set of nonnegative and
 1015 nondecreasing sequences. Let y_1, y_2, \dots be a sequence in S . Let $C_t \in \mathcal{F}_{t-1}$ for all $t = 1, 2, \dots, T$
 1016 and let X_t be a martingale difference sequence adapted to \mathcal{F}_t such that $|X_t| \leq C_t$ with probability 1
 1017 for $t = 1, 2, \dots, T$. Then for all $\delta \in (0, 1)$ and $\hat{X}_t \in \mathcal{F}_{t-1}$ such that $|\hat{X}_t| \leq C_t$ with probability 1,
 1018 we have that with probability at least $1 - \delta - \text{Prob}(\exists t \leq T \mid C_t > c)$ that for all $c > 0$

$$\left|\sum_{i=1}^t y_i X_i\right| \leq 8y_t \sqrt{\theta_{t,\delta} \sum_{i=1}^t (X_i - \hat{X}_i)^2 + c^2 \theta_{t,\delta}^2},$$

1019 where $\theta_{t,\delta} = \log \frac{60 \log 6t}{\delta}$.

1020 **Lemma B.5.** Suppose we have

$$r_{k+1} \leq (1+a)r_k - b\delta_k + c$$

1021 Then,

$$\min_j \delta_j \leq \frac{r_0 e^{aK}}{bK} + \frac{c}{b}.$$

1022 *Proof.* Let $w_{k+1} = \frac{w_k}{1+a}$. We have

$$\begin{aligned} w_{k+1}r_{k+1} &\leq (1+a)w_{k+1}r_k - bw_{k+1}\delta_k + cw_{k+1} \\ &= w_k r_k - bw_{k+1}\delta_k + cw_{k+1}. \end{aligned}$$

1023 Telescoping,

$$w_K r_K \leq w_0 r_0 - b \sum_{j=0}^{K-1} w_{j+1} \delta_j + c \sum_{j=0}^{K-1} w_{j+1}.$$

1024 Rearranging,

$$\frac{1}{\sum_{j=0}^{K-1} w_{j+1}} \sum_{j=0}^{K-1} w_{j+1} \delta_j \leq \frac{w_0 r_0}{b \sum_{j=0}^{K-1} w_{j+1}} + \frac{c}{b}.$$

1025 We have $w_k = \frac{w_{k-1}}{1+a} = \frac{w_0}{(1+a)^k}$. Therefore,

$$\begin{aligned} \sum_{j=0}^{K-1} w_{j+1} &= \sum_{j=0}^{K-1} \frac{w_0}{(1+a)^{j+1}} \\ &\geq \sum_{j=0}^{K-1} \frac{w_0}{(1+a)^K} \\ &= \frac{w_0 K}{(1+a)^K}. \end{aligned}$$

1026 Therefore,

$$\frac{1}{\sum_{j=0}^{K-1} w_{j+1}} \sum_{j=0}^{K-1} w_{j+1} \delta_j \leq \frac{r_0 (1+a)^K}{bK} + \frac{c}{b}.$$

1027 Finally, it remains to use that $1+a \leq e^a$. □

1028 B.2 Non-adaptive guarantee without momentum

1029 We begin with a lemma that establishes the regret of the local optimizer. Often the regret is measured
1030 against the optimal point (like x_*) but here we instead utilize it against the *initial* point $y_{r,0} = x_r$.

1031 **Lemma B.6** (Regret against starting point). *For any learning rate $\eta > 0$, the inner product between*
1032 *the displacement from the initial average iterate and the average gradient satisfies,*

$$\sum_{h=0}^{H-1} \langle y_{r,h} - y_{r,0}, g_{r,h} \rangle \leq \frac{\eta}{2} \sum_{h=0}^{H-1} \|g_{r,h}\|^2 - \frac{1}{2\eta} \|y_{r,H} - y_{r,0}\|^2.$$

1033 *Proof.* We begin by using that $y_{r,h+1} = y_{r,h} - \eta g_{r,h}$ and expanding the square as

$$\begin{aligned} \|y_{r,h+1} - y_{r,0}\|^2 &= \|y_{r,h} - \eta g_{r,h} - y_{r,0}\|^2 \\ &= \|y_{r,h} - y_{r,0}\|^2 + \eta^2 \|g_{r,h}\|^2 - 2\eta \langle y_{r,h} - y_{r,0}, g_{r,h} \rangle. \end{aligned}$$

1034 Rearranging to isolate the inner product term, we obtain

$$\langle y_{r,h} - y_{r,0}, g_{r,h} \rangle = \frac{\|y_{r,h} - y_{r,0}\|^2 - \|y_{r,h+1} - y_{r,0}\|^2}{2\eta} + \frac{\eta}{2} \|g_{r,h}\|^2.$$

1035 Summing over h from 0 to $H-1$,

$$\begin{aligned} \sum_{h=0}^{H-1} \langle y_{r,h} - y_{r,0}, g_{r,h} \rangle &= \sum_{h=0}^{H-1} \left(\frac{\|y_{r,h} - y_{r,0}\|^2 - \|y_{r,h+1} - y_{r,0}\|^2}{2\eta} + \frac{\eta}{2} \|g_{r,h}\|^2 \right) \\ &= \frac{1}{2\eta} \sum_{h=0}^{H-1} (\|y_{r,h} - y_{r,0}\|^2 - \|y_{r,h+1} - y_{r,0}\|^2) + \frac{\eta}{2} \sum_{h=0}^{H-1} \|g_{r,h}\|^2. \end{aligned}$$

1036 The first sum telescopes

$$\begin{aligned} \sum_{h=0}^{H-1} (\|y_{r,h} - y_{r,0}\|^2 - \|y_{r,h+1} - y_{r,0}\|^2) &= \|y_{r,0} - y_{r,0}\|^2 - \|y_{r,H} - y_{r,0}\|^2 \\ &= -\|y_{r,H} - y_{r,0}\|^2. \end{aligned}$$

1037 Therefore,

$$\begin{aligned} \sum_{h=0}^{H-1} \langle y_{r,h} - y_{r,0}, g_{r,h} \rangle &= -\frac{\|y_{r,H} - y_{r,0}\|^2}{2\eta} + \frac{\eta}{2} \sum_{h=0}^{H-1} \|g_{r,h}\|^2 \\ &\leq \frac{\eta}{2} \sum_{h=0}^{H-1} \|g_{r,h}\|^2 - \frac{\|y_{r,H} - y_{r,0}\|^2}{2\eta}. \end{aligned}$$

1038 □

1039 **Lemma B.7.** (Local client drift bound). Suppose that Assumptions 3.1 and 3.2 hold. Then in
1040 Algorithm GEN-LOC-SGD for all r and h , if $\eta \leq \frac{1}{L}$, then

$$\mathbb{E} \left[\frac{1}{M^2} \sum_{m,s=1}^M \|y_{m,r,h} - y_{s,r,h}\|^2 \right] \leq 2\eta^2 \sigma^2 h.$$

1041 *Proof.* Let $\tilde{T}_\eta(y_{m,r,h}) = y_{m,r,h} - \eta g_{m,r,h}$ where $g_{m,r,h}$ is the stochastic gradient, and $T_\eta(y_{m,r,h}) =$
1042 $y - \eta \bar{g}_{m,r,h}$ is the corresponding expected gradient update. We have

$$\begin{aligned} y_{m,r,h+1} - y_{s,r,h+1} &= \tilde{T}_\eta(y_{m,r,h}) - \tilde{T}_\eta(y_{s,r,h}) \\ &= T_\eta(y_{m,r,h}) - T_\eta(y_{s,r,h}) + [\tilde{T}_\eta(y_{m,r,h}) - \tilde{T}_\eta(y_{s,r,h}) - (T_\eta(y_{m,r,h}) - T_\eta(y_{s,r,h}))] \\ &= T_\eta(y_{m,r,h}) - T_\eta(y_{s,r,h}) + [\xi_{m,r,h} - \xi_{s,r,h}], \end{aligned}$$

1043 where $\xi_{m,r,h} = \tilde{T}_\eta(y_{m,r,h}) - T_\eta(y_{m,r,h}) = -\eta n_{m,r,h}$ is the noise term. Define $\mathcal{V}_{r,h} =$
1044 $\frac{1}{M^2} \sum_{m,s=1}^M \|y_{m,r,h} - y_{s,r,h}\|^2$. It follows that

$$\begin{aligned} \mathcal{V}_{r,h+1} &= \frac{1}{M^2} \sum_{m,s=1}^M \|y_{m,r,h+1} - y_{s,r,h+1}\|^2 \\ &= \frac{1}{M^2} \sum_{m,s=1}^M \left[\|T_\eta(y_{m,r,h}) - T_\eta(y_{s,r,h})\|^2 + \|\xi_{m,r,h} - \xi_{s,r,h}\|^2 \right. \\ &\quad \left. + 2 \langle T_\eta(y_{m,r,h}) - T_\eta(y_{s,r,h}), \xi_{m,r,h} - \xi_{s,r,h} \rangle \right]. \end{aligned}$$

1045 Taking conditional expectation gives

$$\mathbb{E}_{r,h} [\mathcal{V}_{r,h+1}] = \frac{1}{M^2} \sum_{m,s=1}^M \left[\|T_\eta(y_{m,r,h}) - T_\eta(y_{s,r,h})\|^2 + \mathbb{E}_h [\|\xi_{m,r,h} - \xi_{s,r,h}\|^2] \right].$$

1046 Finally, using the fact that $\|T_\eta(x) - T_\eta(y)\|^2 \leq \|x - y\|^2$ whenever $\eta \leq \frac{2}{L}$ (Lemma B.1) and
1047 Assumption 3.2, we get

$$\begin{aligned} \mathbb{E}_{r,h} [\mathcal{V}_{r,h+1}] &\leq \frac{1}{M^2} \sum_{m,s=1}^M \left[\|y_{m,r,h} - y_{s,r,h}\|^2 + 2\eta^2 \sigma^2 \right] \\ &= \mathcal{V}_{r,h} + 2\eta^2 \sigma^2. \end{aligned}$$

1048 Therefore by taking unconditional expectation and recursing from $h = 0$ where all local iterates are
1049 equal to x_r (so $\mathcal{V}_{r,0} = 0$), we get $\mathbb{E} [\mathcal{V}_{r,h}] \leq 2\eta^2 \sigma^2 h$. □

1050 *Proof of Theorem 3.3.* We begin by analyzing how the squared distance to the optimal solution
 1051 changes after one round of communication. From the update rule, we have,

$$\|x_{r+1} - x_*\|^2 = \|x_r - x_*\|^2 - 2\eta\gamma \sum_{h=0}^{H-1} \langle x_r - x_*, g_{r,h} \rangle + \eta^2\gamma^2 \left\| \sum_{h=0}^{H-1} g_{r,h} \right\|^2. \quad (9)$$

1052 We rewrite the inner product term as

$$\begin{aligned} -\langle x_r - x_*, g_{r,h} \rangle &= \langle x_* - x_r, g_{r,h} \rangle \\ &= \langle x_* - y_{r,h}, g_{r,h} \rangle + \langle y_{r,h} - x_r, g_{r,h} \rangle. \end{aligned}$$

1053 Summing over all local steps we obtain

$$-\sum_{h=0}^{H-1} \langle x_r - x_*, g_{r,h} \rangle = \sum_{h=0}^{H-1} \langle x_* - y_{r,h}, g_{r,h} \rangle + \sum_{h=0}^{H-1} \langle y_{r,h} - x_r, g_{r,h} \rangle.$$

1054 Applying Lemma B.6 we get

$$-\sum_{h=0}^{H-1} \langle x_r - x_*, g_{r,h} \rangle = \sum_{h=0}^{H-1} \langle x_* - y_{r,h}, g_{r,h} \rangle - \frac{\|y_{r,H} - y_{r,0}\|^2}{2\eta} + \frac{\eta}{2} \sum_{h=0}^{H-1} \|g_{r,h}\|^2. \quad (10)$$

1055 Observe that since $y_{r,H} - y_{r,0} = -\eta \sum_{h=0}^{H-1} g_{r,h}$, Equation (10) becomes,

$$-\sum_{h=0}^{H-1} \langle x_r - x_*, g_{r,h} \rangle = \sum_{h=0}^{H-1} \langle x_* - y_{r,h}, g_{r,h} \rangle - \frac{\eta}{2} \left\| \sum_{h=0}^{H-1} g_{r,h} \right\|^2 + \frac{\eta}{2} \sum_{h=0}^{H-1} \|g_{r,h}\|^2.$$

1056 Plugging this back into Equation (9),

$$\begin{aligned} \|x_{r+1} - x_*\|^2 &\leq \|x_r - x_*\|^2 + 2\eta\gamma \sum_{h=0}^{H-1} \langle x_* - y_{r,h}, g_{r,h} \rangle \\ &\quad + \gamma\eta^2 \sum_{h=0}^{H-1} \|g_{r,h}\|^2 + \eta^2\gamma(\gamma - 1) \left\| \sum_{h=0}^{H-1} g_{r,h} \right\|^2. \end{aligned}$$

1057 Let us take expectation conditional on x_1, \dots, x_r ,

$$\begin{aligned} \mathbb{E}_r \left[\|x_{r+1} - x_*\|^2 \right] &\leq \|x_r - x_*\|^2 + 2\eta\gamma \sum_{h=0}^{H-1} \mathbb{E}_r [\langle x_* - y_{r,h}, g_{r,h} \rangle] \\ &\quad + \gamma\eta^2 \sum_{h=0}^{H-1} \mathbb{E}_r [\|g_{r,h}\|^2] + \eta^2\gamma(\gamma - 1) \mathbb{E}_r \left[\left\| \sum_{h=0}^{H-1} g_{r,h} \right\|^2 \right]. \end{aligned} \quad (11)$$

1058 For the squared norm of the average gradient:

$$\begin{aligned} \mathbb{E}_r [\|g_{r,h}\|^2] &= \mathbb{E}_r [\mathbb{E}_{r,h-1} [\|g_{r,h}\|^2]] \\ &= \mathbb{E}_r [\mathbb{E}_{r,h-1} [\|g_{r,h} - \bar{g}_{r,h}\|^2] + \|\bar{g}_{r,h}\|^2] \\ &= \frac{\sigma^2}{M} + \mathbb{E}_r [\|\bar{g}_{r,h}\|^2], \end{aligned}$$

1059 where we use $\mathbb{E}_{r,h-1} [\cdot]$ to denote expectation conditional on the σ -algebra generated by all the
 1060 stochastic gradients up to and including step $h - 1$. Substituting this into Equation (11),

$$\begin{aligned} \mathbb{E}_r [\|x_{r+1} - x_*\|^2] &\leq \|x_r - x_*\|^2 + 2\eta\gamma \sum_{h=0}^{H-1} \mathbb{E}_r [\langle x_* - y_{r,h}, g_{r,h} \rangle] + \frac{\gamma\eta^2 H \sigma^2}{M} \\ &\quad + \gamma\eta^2 \sum_{h=0}^{H-1} \mathbb{E}_r [\|\bar{g}_{r,h}\|^2] + \eta^2\gamma(\gamma - 1) \mathbb{E}_r \left[\left\| \sum_{h=0}^{H-1} g_{r,h} \right\|^2 \right]. \end{aligned} \quad (12)$$

1061 Now we bound the inner product term:

$$\begin{aligned}
\mathbb{E}_r [\langle x_* - y_{r,h}, g_{r,h} \rangle] &= \mathbb{E}_r [\mathbb{E}_{h-1} [\langle x_* - y_{r,h}, g_{r,h} \rangle]] \\
&= \mathbb{E}_r [\langle x_* - y_{r,h}, \bar{g}_{r,h} \rangle] \\
&= \frac{1}{M} \sum_{m=1}^M \mathbb{E}_r [\langle x_* - y_{r,h}, \bar{g}_{m,r,h} \rangle] \\
&= \frac{1}{M} \sum_{m=1}^M \mathbb{E}_r [\langle x_* - y_{m,r,h} + y_{m,r,h} - y_{r,h}, \bar{g}_{m,r,h} \rangle] \\
&= \frac{1}{M} \sum_{m=1}^M \mathbb{E}_r [\langle x_* - y_{m,r,h}, \bar{g}_{m,r,h} \rangle] + \frac{1}{M} \sum_{m=1}^M \mathbb{E}_r [\langle y_{m,r,h} - y_{r,h}, \bar{g}_{m,r,h} \rangle].
\end{aligned}$$

1062 Using Young's inequality for the second term,

$$\mathbb{E}_r [\langle x_* - y_{r,h}, g_{r,h} \rangle] \tag{13}$$

$$\begin{aligned}
&\leq \frac{1}{M} \sum_{m=1}^M \mathbb{E}_r [\langle x_* - y_{m,r,h}, \bar{g}_{m,r,h} \rangle] + \frac{1}{M} \sum_{m=1}^M \mathbb{E}_r \left[\frac{\|y_{m,r,h} - y_{r,h}\|^2}{2\alpha} + \frac{\alpha}{2} \|\bar{g}_{m,r,h}\|^2 \right] \\
&= \frac{1}{M} \sum_{m=1}^M \mathbb{E}_r [\langle x_* - y_{m,r,h}, \bar{g}_{m,r,h} \rangle] + \frac{V_{r,h}}{2\alpha} + \frac{\alpha}{2M} \sum_{m=1}^M \mathbb{E}_r [\|\bar{g}_{m,r,h}\|^2],
\end{aligned} \tag{14}$$

1063 where $V_{r,h} = \frac{1}{M} \sum_{m=1}^M \mathbb{E}_r [\|y_{m,r,h} - y_{r,h}\|^2]$ by definition. By the convexity of f ,

$$\begin{aligned}
\langle x_* - y_{m,r,h}, \bar{g}_{m,r,h} \rangle &= \langle x_* - y_{m,r,h}, \nabla f(y_{m,r,h}) \rangle \\
&\leq f(x_*) - f(y_{m,r,h}) \\
&= -(f(y_{m,r,h}) - f(x_*)).
\end{aligned} \tag{15}$$

1064 For the variance term, when $\eta \leq \frac{1}{L}$ we use Lemma B.7

$$\begin{aligned}
V_{r,h} &= \frac{1}{M} \sum_{m=1}^M \mathbb{E}_r [\|y_{m,r,h} - y_{r,h}\|^2] \\
&\leq \frac{1}{M} \sum_{m=1}^M \frac{1}{M} \sum_{s=1}^M \mathbb{E}_r [\|y_{m,r,h} - y_{s,r,h}\|^2] \\
&= \frac{1}{M^2} \sum_{m=1}^M \sum_{s=1}^M \mathbb{E}_r [\|y_{m,r,h} - y_{s,r,h}\|^2] \\
&\leq 2\eta^2 \sigma^2 h \leq 2\eta^2 \sigma^2 H.
\end{aligned} \tag{16}$$

1065 By smoothness,

$$\|\bar{g}_{m,r,h}\|^2 = \|\nabla f(y_{m,r,h})\|^2 \leq 2L(f(y_{m,r,h}) - f(x_*)). \tag{17}$$

1066 Plugging Equations (15) to (17) back into Equation (14) we get

$$\mathbb{E}_r [\langle x_* - y_{r,h}, g_{r,h} \rangle] \leq \frac{-(1-\alpha L)}{M} \sum_{m=1}^M (\mathbb{E}_r [f(y_{m,r,h})] - f(x_*)) + \frac{\eta^2 \sigma^2 H}{\alpha}. \tag{18}$$

1067 Substituting (18) back into our main recursion (Equation (11)),

$$\begin{aligned}
\mathbb{E}_r [\|x_{r+1} - x_*\|^2] &\leq \|x_r - x_*\|^2 - \frac{2\eta\gamma(1-\alpha L)}{M} \sum_{h=0}^{H-1} \sum_{m=1}^M (\mathbb{E}_r [f(y_{m,r,h})] - f(x_*)) + \frac{2\eta^3 \gamma \sigma^2 H^2}{\alpha} \\
&\quad + \frac{\gamma \eta^2 H \sigma^2}{M} + \gamma \eta^2 \sum_{h=0}^{H-1} \mathbb{E}_r [\|\bar{g}_{r,h}\|^2] + \eta^2 \gamma (\gamma - 1) \mathbb{E}_r \left[\left\| \sum_{h=0}^{H-1} g_{r,h} \right\|^2 \right].
\end{aligned} \tag{19}$$

1068 We now have two cases. **Case 1.** If $\gamma \geq 1$, then we have by Lemma B.3 and Jensen's inequality
 1069 applied to $\|\cdot\|^2$,

$$\begin{aligned}
 \mathbb{E}_r \left[\left\| \sum_{h=0}^{H-1} g_{r,h} \right\|^2 \right] &= \mathbb{E}_r \left[\left\| \sum_{h=0}^{H-1} (g_{r,h} - \mathbb{E}_r [g_{r,h}]) \right\|^2 \right] + \left\| \sum_{h=0}^{H-1} (\mathbb{E}_r [g_{r,h}]) \right\|^2 \\
 &= \mathbb{E}_r \left[\left\| \sum_{h=0}^{H-1} (g_{r,h} - \mathbb{E}_r [g_{r,h}]) \right\|^2 \right] + \left\| \sum_{h=0}^{H-1} (\mathbb{E}_r [\mathbb{E}_{r,h-1} [g_{r,h}]]) \right\|^2 \\
 &= \mathbb{E}_r \left[\left\| \sum_{h=0}^{H-1} (g_{r,h} - \mathbb{E}_r [g_{r,h}]) \right\|^2 \right] + \left\| \sum_{h=0}^{H-1} \mathbb{E}_r [\bar{g}_{r,h}] \right\|^2 \\
 &\leq \frac{\sigma^2 H}{M} + \mathbb{E}_r \left[\left\| \sum_{h=0}^{H-1} \bar{g}_{r,h} \right\|^2 \right] \\
 &\leq \frac{\sigma^2 H}{M} + H \sum_{h=0}^{H-1} \mathbb{E}_r [\|\bar{g}_{r,h}\|^2]. \tag{20}
 \end{aligned}$$

1070 Using Jensen's inequality and smoothness we have

$$\begin{aligned}
 \mathbb{E}_r [\|\bar{g}_{r,h}\|^2] &= \mathbb{E}_r \left[\left\| \frac{1}{M} \sum_{m=1}^M \nabla f(y_{m,r,h}) \right\|^2 \right] \\
 &\leq \frac{1}{M} \sum_{m=1}^M \mathbb{E}_r [\|\nabla f(y_{m,r,h})\|^2] \\
 &\leq \frac{2L}{M} \sum_{m=1}^M \mathbb{E}_r [f(y_{m,r,h}) - f(x_*)]. \tag{21}
 \end{aligned}$$

1071 Using Equations (20) and (21) into Equation (19) we get

$$\begin{aligned}
 \mathbb{E}_r [\|x_{r+1} - x_*\|^2] &\leq \|x_r - x_*\|^2 \\
 &\quad - \frac{2\eta\gamma(1 - \alpha L) - 2L\gamma\eta^2(1 + (\gamma - 1)H)}{M} \sum_{h=0}^{H-1} \sum_{m=1}^M (\mathbb{E}_r [f(y_{m,r,h})] - f(x_*)) + \frac{2\eta^3\gamma\sigma^2 H^2}{\alpha} \\
 &\quad + \frac{\gamma^2\eta^2 H\sigma^2}{M}. \\
 &= \|x_r - x_*\|^2 - \frac{2\eta\gamma[1 - \alpha L - L\eta(1 + (\gamma - 1)H)]}{M} \sum_{h=0}^{H-1} \sum_{m=1}^M (\mathbb{E}_r [f(y_{m,r,h})] - f(x_*)) \\
 &\quad + \frac{2\eta^3\gamma\sigma^2 H^2}{\alpha} + \frac{\gamma^2\eta^2 H\sigma^2}{M}. \\
 &= \|x_r - x_*\|^2 - 2\eta\gamma H(1 - \alpha L - L\eta(1 + (\gamma - 1)H)) \mathbb{E}_r [\hat{\delta}_{r+1}] + \frac{2\eta^3\gamma\sigma^2 H^2}{\alpha} + \frac{\eta^2\gamma^2 H\sigma^2}{M}, \tag{22}
 \end{aligned}$$

1072 where in the last line we defined

$$\hat{\delta}_{r+1} = \frac{1}{MH} \sum_{h=0}^{H-1} \sum_{m=1}^M (f(y_{m,r,h}) - f(x_*)) \tag{23}$$

1073 **Case 2.** If $\gamma \leq 1$, then we can simply drop the last term in Equation (19) and use Equation (17) to get

$$\begin{aligned} \mathbb{E}_r \left[\|x_{r+1} - x_*\|^2 \right] &\leq \|x_r - x_*\|^2 - \frac{2\eta\gamma(1 - \alpha L - \eta L)}{M} \sum_{h=0}^{H-1} \sum_{m=1}^M (\mathbb{E}_r [f(y_{m,r,h})] - f(x_*)) \\ &\quad + \frac{2\eta^3\gamma\sigma^2 H^2}{\alpha} + \frac{\gamma\eta^2 H\sigma^2}{M} \\ &= \|x_r - x_*\|^2 - 2\eta\gamma H(1 - \alpha L - \eta L)\mathbb{E}_r [\hat{\delta}_{r+1}] + \frac{2\eta^3\gamma\sigma^2 H^2}{\alpha} + \frac{\gamma\eta^2 H\sigma^2}{M}, \end{aligned} \quad (24)$$

1074 where in Equation (24) we again used the definition in Equation (23). Looking at both Equations (22)
1075 and (24) and taking the maximum we get that for *any* γ ,

$$\begin{aligned} \mathbb{E}_r \left[\|x_{r+1} - x_*\|^2 \right] &\leq \|x_r - x_*\|^2 - 2\eta\gamma H(1 - \alpha L - \eta L(1 + (\gamma - 1)_+ H))\mathbb{E}_r [\hat{\delta}_{r+1}] \\ &\quad + \frac{2\eta^3\gamma\sigma^2 H^2}{\alpha} + \frac{\eta^2 \max\{\gamma^2, \gamma\} H\sigma^2}{M}, \end{aligned}$$

1076 where $(x)_+ = \max(x, 0)$ is the ReLU function. Putting $\alpha = \frac{1}{2L}$ we get

$$\begin{aligned} \mathbb{E}_r \left[\|x_{r+1} - x_*\|^2 \right] &\leq \|x_r - x_*\|^2 - \eta\gamma H(1 - 2\eta L(1 + (\gamma - 1)_+ H))\mathbb{E}_r [\hat{\delta}_{r+1}] \\ &\quad + 4L\eta^3\gamma\sigma^2 H^2 + \frac{\eta^2 \max\{\gamma^2, \gamma\} H\sigma^2}{M}. \end{aligned}$$

1077 Under the requirement that the stepsizes η, γ satisfy

$$\eta L(1 + (\gamma - 1)_+ H) \leq \frac{1}{4},$$

1078 we obtain our recursion

$$\mathbb{E}_r \left[\|x_{r+1} - x_*\|^2 \right] \leq \|x_r - x_*\|^2 - \frac{\eta\gamma H}{2} \mathbb{E}_r [\hat{\delta}_{r+1}] + 4L\eta^3\gamma\sigma^2 H^2 + \frac{\eta^2 \max\{\gamma^2, \gamma\} H\sigma^2}{M}.$$

1079 Taking unconditional expectations and rearranging we obtain,

$$\mathbb{E} [\hat{\delta}_{r+1}] \leq \frac{2}{\gamma\eta H} \left[\mathbb{E} [\|x_r - x_*\|^2] - \mathbb{E} [\|x_{r+1} - x_*\|^2] \right] + 8L\eta^2\sigma^2 H + \frac{2\eta \max(\gamma, 1)\sigma^2}{M}.$$

1080 Summing up both sides as r varies from 0 to $R - 1$ and dividing by $1/R$ we get

$$\frac{1}{R} \sum_{r=0}^{R-1} \mathbb{E} [\hat{\delta}_{r+1}] \leq \frac{2}{\gamma\eta RH} \left[\|x_0 - x_*\|^2 - \mathbb{E} [\|x_R - x_*\|^2] \right] + 8L\eta^2\sigma^2 H + \frac{2\eta \max(\gamma, 1)\sigma^2}{M}.$$

1081 Observe that we can write $\max(\gamma, 1) = 1 + (\gamma - 1)_+$. Dropping the negative term and using Jensen's
1082 inequality gives

$$\begin{aligned} \mathbb{E} \left[f \left(\frac{1}{MRH} \sum_{r=0}^{R-1} \sum_{h=0}^{H-1} \sum_{m=1}^M f(y_{m,r,h}) \right) \right] - f(x_*) &\leq \frac{1}{R} \sum_{r=0}^{R-1} \mathbb{E} [\hat{\delta}_{r+1}] \\ &\leq \frac{2\|x_0 - x_*\|^2}{\gamma\eta RH} + 8L\eta^2\sigma^2 H + \frac{2\eta \max(\gamma, 1)\sigma^2}{M}, \end{aligned}$$

1083 and this is the statement of our theorem. \square

1084 B.3 Non-adaptive guarantee with momentum

1085 We present two guarantees. One is the proof of Theorem 3.5 as it is, and a second is **a new proof**
1086 **without the bounded iterates assumption**. The latter is new and wasn't mentioned in the main text,
1087 but is strictly superior to the claim in the main text. We start by presenting the new proof first.

1088 B.3.1 Main momentum guarantee

1089 **Theorem B.8.** *Let f be an L -smooth convex function. Consider Local SGD with momentum*
 1090 *parameter $\mu \in [0, 1)$ and communication interval H . Assume the stochastic gradients satisfy the*
 1091 *σ^2 -bounded variance assumption. Let the step sizes η, γ satisfy*

$$\eta L \left(1 + \left(\frac{\gamma}{1-\mu} - 1 \right)_+ H \right) \leq \frac{1}{4}, \quad \frac{\eta \gamma \mu L H}{1-\mu} \leq \frac{1}{16}.$$

1092 *Then after R rounds of communication, the averaged iterate satisfies*

$$\begin{aligned} \mathbb{E}[f(y_{\text{out}})] - f(x_*) &\leq \frac{4(1-\mu)\|z_0 - x_*\|^2}{\eta \gamma H R} + 16L\eta^2 \sigma^2 H \\ &\quad + \frac{4\eta \sigma^2}{M} \max\left(\frac{\gamma}{1-\mu}, 1\right) + \frac{8\eta \gamma \mu \sigma^2}{1-\mu M}. \end{aligned}$$

1093 *Proof.* We analyze the momentum variant of Local SGD:

$$x_{r+1} = x_r - \eta \gamma \left(\sum_{h=0}^{H-1} g_{r,h} \right) + \mu(x_r - x_{r-1}).$$

1094 Define

$$z_r = x_r + \frac{\mu}{1-\mu}(x_r - x_{r-1}).$$

1095 Then

$$z_{r+1} = z_r - \frac{\eta \gamma}{1-\mu} \sum_{h=0}^{H-1} g_{r,h}.$$

1096 We have

$$\begin{aligned} \|z_{r+1} - x_*\|^2 &= \|z_r - x_*\|^2 + \frac{\eta^2 \gamma^2}{(1-\mu)^2} \left\| \sum_{h=0}^{H-1} g_{r,h} \right\|^2 - \frac{2\eta \gamma}{1-\mu} \sum_{h=0}^{H-1} \langle z_r - x_*, g_{r,h} \rangle \\ &= \|z_r - x_*\|^2 + \frac{\eta^2 \gamma^2}{(1-\mu)^2} \left\| \sum_{h=0}^{H-1} g_{r,h} \right\|^2 - \frac{2\eta \gamma}{1-\mu} \sum_{h=0}^{H-1} \langle x_r - x_*, g_{r,h} \rangle \\ &\quad - \frac{2\eta \gamma \mu}{1-\mu} \sum_{h=0}^{H-1} \langle x_r - x_{r-1}, g_{r,h} \rangle. \end{aligned} \tag{25}$$

1097 Following the same proof as Theorem 3.3, we can bound (in expectation)

$$\begin{aligned} -\frac{2\eta \gamma}{1-\mu} \sum_{h=0}^{H-1} \mathbb{E}_r[\langle x_r - x_*, g_{r,h} \rangle] &+ \frac{\eta^2 \gamma^2}{(1-\mu)^2} \mathbb{E}_r \left[\left\| \sum_{h=0}^{H-1} g_{r,h} \right\|^2 \right] \leq -\frac{\eta \gamma H}{2(1-\mu)} \mathbb{E}_r[\hat{\delta}_{r+1}] \\ &+ 4L\eta^3 \frac{\gamma}{1-\mu} \sigma^2 H^2 + \frac{\eta^2 H \sigma^2}{M} \max \left(\left(\frac{\gamma}{1-\mu} \right)^2, \frac{\gamma}{1-\mu} \right), \end{aligned} \tag{26}$$

1098 because the local optimization procedure is the same– the same analysis holds line-by-line, only
 1099 replacing γ by $\frac{\gamma}{1-\mu}$, and requiring instead that

$$\eta L \left(1 + \left(\frac{\gamma}{1-\mu} - 1 \right)_+ H \right) \leq \frac{1}{4}. \tag{27}$$

1100 Using Equation (26) in Equation (25) (after taking expectation in the latter) we obtain

$$\begin{aligned} \mathbb{E}_r \left[\|z_{r+1} - x_*\|^2 \right] &\leq \|z_r - x_*\|^2 - \frac{\eta\gamma H}{2(1-\mu)} \mathbb{E}_r \left[\hat{\delta}_{r+1} \right] + 4L\eta^3 \frac{\gamma\sigma^2 H^2}{1-\mu} \\ &\quad + \frac{\eta^2 H\sigma^2}{M} \max \left(\left(\frac{\gamma}{1-\mu} \right)^2, \frac{\gamma}{1-\mu} \right) - \frac{2\eta\gamma\mu}{1-\mu} \sum_{h=0}^{H-1} \langle x_r - x_{r-1}, \bar{g}_{r,h} \rangle. \end{aligned} \quad (28)$$

1101 In the following, we use the shorthand $G_r \stackrel{\text{def}}{=} \sum_{h=0}^{H-1} g_{r,h}$. We now proceed to bound
 1102 $\sum_{h=0}^{H-1} \langle x_{r-1} - x_r, g_{r,h} \rangle = \langle x_{r-1} - x_r, G_r \rangle$ without using the bounded iterates assumption. We
 1103 note that by definition:

$$x_r - x_{r-1} = -\eta\gamma G_{r-1} + \mu(x_{r-1} - x_{r-2}).$$

1104 Expanding this out recursively, we get the following formula:

$$x_r - x_{r-1} = -\eta\gamma \sum_{s=0}^{r-1} \mu^{r-1-s} G_s.$$

1105 For our analysis, we'll bound the inner product

$$\begin{aligned} \langle x_{r-1} - x_r, G_r \rangle &= \left\langle \eta\gamma \sum_{s=0}^{r-1} \mu^{r-1-s} G_s, G_r \right\rangle \\ &= \eta\gamma \sum_{s=0}^{r-1} \mu^{r-1-s} \langle G_s, G_r \rangle \end{aligned}$$

1106 We will actually bound the sum of the momentum terms over r , i.e. $\sum_r \langle x_{r-1} - x_r, G_r \rangle$. We have

$$\begin{aligned} \sum_r \langle x_{r-1} - x_r, G_r \rangle &= \frac{\eta\gamma}{\mu} \sum_r \sum_{s < r} \langle \mu^{r-s} G_s, G_r \rangle \\ &= \frac{\eta\gamma}{2\mu} \left[\sum_r \sum_s \langle \mu^{|r-s|} G_s, G_r \rangle - \sum_r \|G_r\|^2 \right]. \end{aligned}$$

1107 To bound the first term above, let A be the $R \times R$ matrix whose (r, s) th entry equals $\mu^{|r-s|}$, and let
 1108 $\Gamma = [G_1 | G_2 | \dots | G_R]$. Then

$$\sum_r \sum_s \langle \mu^{|r-s|} G_s, G_r \rangle = \text{Tr}(\Gamma A \Gamma^\top).$$

1109 We now apply the Gershgorin circle theorem to bound this sum, observe that largest sum of absolute
 1110 values of entries in a row satisfy

$$1 + 2 \sum_{r=1}^{(R-1)/2} \mu^r = 1 + 2\mu \frac{1 - \mu^{(R-1)/2}}{1 - \mu} = \frac{1 + \mu - 2\mu^{(R+1)/2}}{1 - \mu} \leq \frac{1 + \mu}{1 - \mu}.$$

1111 Then, we have

$$\text{Tr}(\Gamma A \Gamma^\top) \leq \frac{1 + \mu}{1 - \mu} \text{Tr}(\Gamma \Gamma^\top) = \frac{1 + \mu}{1 - \mu} \sum_r \|G_r\|^2.$$

1112 Therefore, taking expectations we have

$$\begin{aligned} -\frac{2\eta\gamma\mu}{1-\mu} \sum_{r=0}^{R-1} \sum_{h=0}^{H-1} \mathbb{E} [\langle x_r - x_{r-1}, g_{r,h} \rangle] &= \frac{2\eta\gamma\mu}{1-\mu} \sum_{r=0}^{R-1} \mathbb{E} [\langle x_{r-1} - x_r, G_r \rangle] \\ &\leq \frac{2\eta\gamma\mu}{1-\mu} \frac{\eta\gamma}{1-\mu} \sum_{r=0}^{R-1} \mathbb{E} \left[\left\| \sum_{h=0}^{H-1} g_{r,h} \right\|^2 \right]. \end{aligned} \quad (29)$$

1113 Using Lemma B.3 we have

$$\begin{aligned}\mathbb{E} \left[\left\| \sum_{h=0}^{H-1} g_{r,h} \right\|^2 \right] &\leq \frac{\sigma^2 H}{M} + \mathbb{E} \left[\left\| \sum_{h=0}^{H-1} \bar{g}_{r,h} \right\|^2 \right] \\ &\leq \frac{\sigma^2 H}{M} + H \sum_{h=0}^{H-1} \mathbb{E} [\|\bar{g}_{r,h}\|^2] \\ &\leq \frac{\sigma^2 H}{M} + 2LH^2 \mathbb{E} [\hat{\delta}_{r+1}],\end{aligned}$$

1114 where in the last line we used Jensen's inequality and smoothness. Using this result in Equation (29)
1115 we get

$$-\frac{2\eta\gamma\mu}{1-\mu} \sum_{r=0}^{R-1} \sum_{h=0}^{H-1} \mathbb{E} [\langle x_r - x_{r-1}, g_{r,h} \rangle] \leq \frac{\eta\gamma}{2(1-\mu)} \frac{4\eta\gamma\mu}{1-\mu} \left[\frac{\sigma^2 RH}{M} + 2LH^2 \sum_{r=0}^{R-1} \mathbb{E} [\hat{\delta}_{r+1}] \right]. \quad (30)$$

1116 Rearranging and summing up Equation (28) then using Equation (30) we have

$$\begin{aligned}\mathbb{E} [\|z_R - x_*\|^2] &\leq \|z_0 - x_*\|^2 - \frac{\eta\gamma H}{2(1-\mu)} \left[1 - \frac{8\eta\gamma\mu LH}{1-\mu} \right] \sum_{r=0}^{R-1} \mathbb{E} [\hat{\delta}_{r+1}] \\ &\quad + 4L\eta^3 \frac{\gamma\sigma^2 H^2}{1-\mu} R + \frac{\eta^2 H\sigma^2}{M} \max \left(\left(\frac{\gamma}{1-\mu} \right)^2, \frac{\gamma}{1-\mu} \right) R + \frac{\eta\gamma H}{1-\mu} \frac{2\eta\gamma\mu}{1-\mu} \frac{\sigma^2 R}{M}.\end{aligned}$$

1117 Observe that under the condition

$$\frac{\eta\gamma\mu LH}{1-\mu} \leq \frac{1}{16}$$

1118 the last inequality becomes

$$\begin{aligned}\mathbb{E} [\|z_R - x_*\|^2] &\leq \|z_0 - x_*\|^2 - \frac{\eta\gamma H}{4(1-\mu)} \sum_{r=0}^{R-1} \mathbb{E} [\hat{\delta}_{r+1}] \\ &\quad + 4L\eta^3 \frac{\gamma\sigma^2 H^2}{1-\mu} R + \frac{\eta^2 H\sigma^2}{M} \max \left(\left(\frac{\gamma}{1-\mu} \right)^2, \frac{\gamma}{1-\mu} \right) R + \frac{\eta\gamma H}{1-\mu} \frac{2\eta\gamma\mu}{1-\mu} \frac{\sigma^2 R}{M}.\end{aligned}$$

1119 Continuing the proof and rearranging we get

$$\frac{1}{R} \sum_{r=0}^{R-1} \mathbb{E} [\hat{\delta}_{r+1}] \leq \frac{4(1-\mu)\|z_0 - x_*\|^2}{\eta\gamma HR} + 16L\eta^2\sigma^2 H + \frac{4\eta\sigma^2}{M} \max \left(\frac{\gamma}{1-\mu}, 1 \right) + \frac{8\eta\gamma\mu}{1-\mu} \frac{\sigma^2}{M}.$$

1120 It remains to use Jensen's inequality. □

1121 B.3.2 Legacy guarantee

1122 *Proof of Theorem 3.5.* We analyze the momentum variant of Local SGD,

$$x_{r+1} = x_r - \eta\gamma \left(\sum_{h=0}^{H-1} g_{r,h} \right) + \mu(x_r - x_{r-1}).$$

1123 Define

$$z_r = x_r + \frac{\mu}{1-\mu} (x_r - x_{r-1}).$$

1124 Then

$$z_{r+1} = z_r - \frac{\eta\gamma}{1-\mu} \sum_{h=0}^{H-1} g_{r,h}.$$

1125 We have

$$\begin{aligned} \|z_{r+1} - x_*\|^2 &= \|z_r - x_*\|^2 + \frac{\eta^2 \gamma^2}{(1-\mu)^2} \left\| \sum_{h=0}^{H-1} g_{r,h} \right\|^2 - \frac{2\eta\gamma}{1-\mu} \sum_{h=0}^{H-1} \langle x_r - x_*, g_{r,h} \rangle \\ &\quad - \frac{2\eta\gamma\mu}{1-\mu} \sum_{h=0}^{H-1} \langle x_r - x_{r-1}, g_{r,h} \rangle. \end{aligned} \quad (31)$$

1126 Following the same proof as Theorem 3.3 we can bound (in expectation)

$$\begin{aligned} -\frac{2\eta\gamma}{1-\mu} \sum_{h=0}^{H-1} \mathbb{E}_r [\langle x_r - x_*, g_{r,h} \rangle] &+ \frac{\eta^2 \gamma^2}{(1-\mu)^2} \mathbb{E}_r \left[\left\| \sum_{h=0}^{H-1} g_{r,h} \right\|^2 \right] \leq -\frac{\eta\gamma H}{2(1-\mu)} \mathbb{E}_r [\hat{\delta}_{r+1}] \\ &+ 4L\eta^3 \frac{\gamma}{1-\mu} \sigma^2 H^2 + \frac{\eta^2 H \sigma^2}{M} \max \left(\left(\frac{\gamma}{1-\mu} \right)^2, \frac{\gamma}{1-\mu} \right), \end{aligned} \quad (32)$$

1127 because the local optimization procedure is the same– the same analysis holds line-by-line, only
1128 replacing γ by $\frac{\gamma}{1-\mu}$, and requiring instead that

$$\eta L \left(1 + \left(\frac{\gamma}{1-\mu} - 1 \right)_+ H \right) \leq \frac{1}{4}. \quad (33)$$

1129 To bound the last inner product in Equation (31), observe that if the domain is D -bounded,

$$\begin{aligned} \langle x_{r-1} - x_r, \bar{g}_{r,h} \rangle &\leq \frac{D^2}{2\rho} + \frac{\rho}{2} \|\bar{g}_{r,h}\|^2 \\ &\leq \frac{D^2}{2\rho} + \frac{L\rho}{M} \sum_{m=1}^M (f(y_{m,r,h}) - f_*). \end{aligned}$$

1130 Summing up over H we get

$$\sum_{h=0}^{H-1} \mathbb{E}_r [\langle x_{r-1} - x_r, \bar{g}_{r,h} \rangle] \leq \frac{D^2 H}{2\rho} + L\rho H \mathbb{E}_r [\hat{\delta}_{r+1}]. \quad (34)$$

1131 Plugging Equations (32) and (34) into Equation (31) we get

$$\begin{aligned} \mathbb{E}_r [\|z_{r+1} - x_*\|^2] &\leq \|z_r - x_*\|^2 - \left[\frac{\eta\gamma H}{2(1-\mu)} - \frac{2\eta\gamma\mu L\rho H}{(1-\mu)} \right] \mathbb{E}_r [\hat{\delta}_{r+1}] + \frac{4L\eta^3 \gamma \sigma^2 H^2}{1-\mu} \\ &\quad + \frac{\eta^2 H \sigma^2}{M} \max \left(\left(\frac{\gamma}{1-\mu} \right)^2, \frac{\gamma}{1-\mu} \right) + \frac{\eta\gamma\mu H D^2}{(1-\mu)\rho}. \end{aligned}$$

1132 Setting $\rho = \frac{1}{8\mu L}$, we get

$$\begin{aligned} \|z_{r+1} - x_*\|^2 &\leq \|z_r - x_*\|^2 - \frac{\eta\gamma H}{4(1-\mu)} \hat{\delta}_{r+1} + \frac{4L\eta^3 \gamma \sigma^2 H^2}{(1-\mu)} \\ &\quad + \frac{\eta^2 H \sigma^2}{M} \max \left(\left(\frac{\gamma}{1-\mu} \right)^2, \frac{\gamma}{1-\mu} \right) + \frac{4\eta\gamma\mu^2 L H D^2}{1-\mu}. \end{aligned}$$

1133 We then continue exactly as in the proof of Theorem 3.3 to obtain

$$\begin{aligned} \mathbb{E} \left[f \left(\frac{1}{MRH} \sum_{r=0}^{R-1} \sum_{h=0}^{H-1} \sum_{m=1}^M y_{m,r,h} \right) \right] - f(x_*) &\leq \frac{4(1-\mu)}{\eta\gamma RH} \|x_0 - x_*\|^2 + \frac{16L\eta^2 \gamma \sigma^2 H}{1-\mu} \\ &\quad + \frac{4\eta\sigma^2}{M} \max \left(\frac{\gamma}{1-\mu}, 1 \right) + 16\mu^2 L D^2. \end{aligned}$$

1134

□

1135 B.4 Data-dependent guarantees

1136 **Lemma B.9.** *Let f be a convex and L -smooth function. Suppose that we run SGD on f on M*
 1137 *parallel nodes as follows*

$$y_{m,r,0} = x_r,$$

$$y_{m,r,h+1} = y_{m,r,h} - \eta g_{m,r,h},$$

1138 *where $m = 1, 2, \dots, M$, $h = 0, 1, \dots, H-1$, and $g_{1,r,h}, g_{2,r,h}, \dots, g_{M,r,h}$ are i.i.d. stochastic gra-*
 1139 *dient estimates such that $\mathbb{E}_{r,h}[g_{m,r,h}] = \nabla f(y_{m,r,h})$, where $\mathbb{E}_{r,h}[\cdot]$ denotes expectation conditional*
 1140 *on all information up to and including round r and local step h , and $\|g_{m,r,h} - \nabla f(y_{m,r,h})\| \leq \sigma$.*
 1141 *Define further $y_{r,h} = \frac{1}{M} \sum_{m=1}^M y_{m,r,h}$. Let $V_{r,h} = \frac{1}{M} \sum_{m=1}^M \|y_{m,r,h} - y_{r,h}\|^2$. Then for all $\eta \leq \frac{1}{L}$*
 1142 *we have with probability at least $1 - \delta$ that for all $h = 0, 1, \dots, H$*

$$V_{r,h} \leq 4104\eta^2\sigma^2(h+1)\theta_{h-1,\delta}^2,$$

1143 *where $\theta_{h,\delta} = \log \frac{60 \log 6h}{\delta}$.*

1144 *Proof.* Define

$$\Lambda_{r,h+1} = \frac{1}{M^2} \sum_{m=1}^M \sum_{s=1}^M \|y_{m,r,h+1} - y_{s,r,h+1}\|^2. \quad (35)$$

1145 We will bound $\Lambda_{r,h}$ first, and then use it to bound $V_{r,h}$ later. We have

$$\begin{aligned} y_{m,r,h+1} - y_{s,r,h+1} &= y_{m,r,h} - \eta g_{m,r,h} - [y_{s,r,h} - \eta g_{s,r,h}] \\ &= y_{m,r,h} - \eta \nabla f(y_{m,r,h}) - \eta [g_{m,r,h} - \nabla f(y_{m,r,h})] - [y_{s,r,h} - \eta \nabla f(y_{s,r,h}) - \eta [g_{s,r,h} - \nabla f(y_{s,r,h})]] \\ &= [y_{m,r,h} - \eta \nabla f(y_{m,r,h}) - [y_{s,r,h} - \eta \nabla f(y_{s,r,h})]] - \eta [(g_{m,r,h} - g_{s,r,h}) - [\nabla f(y_{m,r,h}) - \nabla f(y_{s,r,h})]]. \end{aligned}$$

1146 Therefore

$$\begin{aligned} \|y_{m,r,h+1} - y_{s,r,h+1}\|^2 &= \|T_\eta(y_{m,r,h}) - T_\eta(y_{s,r,h})\|^2 \\ &\quad + \eta^2 \|(g_{m,r,h} - g_{s,r,h}) - (\nabla f(y_{m,r,h}) - \nabla f(y_{s,r,h}))\|^2 \\ &\quad - 2\eta \langle T_\eta(y_{m,r,h}) - T_\eta(y_{s,r,h}), (g_{m,r,h} - g_{s,r,h}) - (\nabla f(y_{m,r,h}) - \nabla f(y_{s,r,h})) \rangle \end{aligned} \quad (36)$$

1147 We define $\rho_{m,r,h}$ as the stochastic gradient noise on node m at round r , step h : $\rho_{m,r,h} = g_{m,r,h} -$
 1148 $\nabla f(y_{m,r,h})$. Then we can write Equation (36) as

$$\begin{aligned} \|y_{m,r,h+1} - y_{s,r,h+1}\|^2 &= \|T_\eta(y_{m,r,h}) - T_\eta(y_{s,r,h})\|^2 + \eta^2 \|\rho_{m,r,h} - \rho_{s,r,h}\|^2 \\ &\quad - 2\eta \langle T_\eta(y_{m,r,h}) - T_\eta(y_{s,r,h}), \rho_{m,r,h} - \rho_{s,r,h} \rangle. \end{aligned} \quad (37)$$

1149 We now use the inequality $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$ to get

$$\begin{aligned} \|y_{m,r,h+1} - y_{s,r,h+1}\|^2 &\leq \|T_\eta(y_{m,r,h}) - T_\eta(y_{s,r,h})\|^2 + 2\eta^2 \|\rho_{m,r,h}\|^2 + 2\eta^2 \|\rho_{s,r,h}\|^2 \\ &\quad - 2\eta \langle T_\eta(y_{m,r,h}) - T_\eta(y_{s,r,h}), \rho_{m,r,h} - \rho_{s,r,h} \rangle. \end{aligned}$$

1150 By Lemma B.1, we have

$$\begin{aligned} \|y_{m,r,h+1} - y_{s,r,h+1}\|^2 &\leq \|y_{m,r,h} - y_{s,r,h}\|^2 + 2\eta^2 \|\rho_{m,r,h}\|^2 + 2\eta^2 \|\rho_{s,r,h}\|^2 \\ &\quad - 2\eta \langle T_\eta(y_{m,r,h}) - T_\eta(y_{s,r,h}), \rho_{m,r,h} - \rho_{s,r,h} \rangle. \end{aligned}$$

1151 Now, we consider the inner product term, observe

$$\begin{aligned} &\langle T_\eta(y_{m,r,h}) - T_\eta(y_{s,r,h}), \rho_{m,r,h} - \rho_{s,r,h} \rangle \\ &= \langle T_\eta(y_{m,r,h}) - T_\eta(y_{r,h}) + T_\eta(y_{r,h}) - T_\eta(y_{s,r,h}), \rho_{m,r,h} - \rho_{s,r,h} \rangle \\ &= \langle T_\eta(y_{m,r,h}) - T_\eta(y_{r,h}), \rho_{m,r,h} - \rho_{s,r,h} \rangle + \langle T_\eta(y_{r,h}) - T_\eta(y_{s,r,h}), \rho_{m,r,h} - \rho_{s,r,h} \rangle \\ &= \langle T_\eta(y_{m,r,h}) - T_\eta(y_{r,h}), \rho_{m,r,h} - \rho_{s,r,h} \rangle + \langle -(T_\eta(y_{s,r,h}) - T_\eta(y_{r,h})), -(\rho_{s,r,h} - \rho_{m,r,h}) \rangle \\ &= \langle T_\eta(y_{m,r,h}) - T_\eta(y_{r,h}), \rho_{m,r,h} - \rho_{s,r,h} \rangle + \langle T_\eta(y_{s,r,h}) - T_\eta(y_{r,h}), \rho_{s,r,h} - \rho_{m,r,h} \rangle. \end{aligned}$$

1152 Averaging with respect to s and m

$$\begin{aligned}
& \frac{1}{M^2} \sum_{m=1}^M \sum_{s=1}^M \langle T_\eta(y_{m,r,h}) - T_\eta(y_{r,h}) + T_\eta(y_{r,h}) - T_\eta(y_{s,r,h}), \rho_{m,r,h} - \rho_{s,r,h} \rangle \\
&= \frac{1}{M^2} \sum_{m=1}^M \sum_{s=1}^M \langle T_\eta(y_{m,r,h}) - T_\eta(y_{r,h}), \rho_{m,r,h} - \rho_{s,r,h} \rangle \\
&\quad + \frac{1}{M^2} \sum_{m=1}^M \sum_{s=1}^M \langle T_\eta(y_{s,r,h}) - T_\eta(y_{r,h}), \rho_{s,r,h} - \rho_{m,r,h} \rangle \\
&= \frac{2}{M^2} \sum_{m=1}^M \sum_{s=1}^M \langle T_\eta(y_{m,r,h}) - T_\eta(y_{r,h}), \rho_{m,r,h} - \rho_{s,r,h} \rangle. \tag{38}
\end{aligned}$$

1153 Averaging Equation (37) with respect to m and s and using Equation (38) we get

$$\begin{aligned}
\frac{1}{M^2} \sum_{m=1}^M \sum_{s=1}^M \|y_{m,r,h+1} - y_{s,r,h+1}\|^2 &\leq \frac{1}{M^2} \sum_{m=1}^M \sum_{s=1}^M \|y_{m,r,h} - y_{s,r,h}\|^2 + \frac{4\eta^2}{M} \sum_{m=1}^M \|\rho_{m,r,h}\|^2 \\
&\quad - \frac{2\eta}{M^2} \sum_{m=1}^M \sum_{s=1}^M \langle T_\eta(y_{m,r,h}) - T_\eta(y_{r,h}), \rho_{m,r,h} - \rho_{s,r,h} \rangle.
\end{aligned}$$

1154 Using $\Lambda_{r,h}$ as defined in Equation (35) we obtain the recursion

$$\Lambda_{r,h+1} \leq \Lambda_{r,h} + \frac{4\eta^2}{M} \sum_{m=1}^M \|\rho_{m,r,h}\|^2 - \frac{2\eta}{M^2} \sum_{m=1}^M \sum_{s=1}^M \langle T_\eta(y_{m,r,h}) - T_\eta(y_{r,h}), \rho_{m,r,h} - \rho_{s,r,h} \rangle.$$

1155 Now observe that $\|\rho_{m,r,h}\|^2 \leq \sigma^2$ by assumption, therefore

$$\Lambda_{r,h+1} \leq \Lambda_{r,h} + 4\eta^2 \sigma^2 - \frac{2\eta}{M^2} \sum_{m=1}^M \sum_{s=1}^M \langle T_\eta(y_{m,r,h}) - T_\eta(y_{r,h}), \rho_{m,r,h} - \rho_{s,r,h} \rangle.$$

1156 Recursing the above inequality we get

$$\begin{aligned}
\Lambda_{r,h} &\leq \Lambda_{r,0} + 4\eta^2 \sigma^2 h - \frac{2\eta}{M^2} \sum_{k=0}^{h-1} \sum_{m=1}^M \sum_{s=1}^M \langle T_\eta(y_{m,r,k}) - T_\eta(y_{r,k}), \rho_{m,r,k} - \rho_{s,r,k} \rangle \\
&= 4\eta^2 \sigma^2 h - \frac{2\eta}{M^2} \sum_{k=0}^{h-1} \sum_{m=1}^M \sum_{s=1}^M \langle T_\eta(y_{m,r,k}) - T_\eta(y_{r,k}), \rho_{m,r,k} - \rho_{s,r,k} \rangle, \tag{39}
\end{aligned}$$

1157 where we used the fact that since $y_{m,r,0} = y_{s,r,0} = x_r$ for all m, s then $\Lambda_{r,0} = 0$. Define

$$\mu_{r,h} = \frac{1}{M} \sum_{m=1}^M \|y_{m,r,h} - y_{r,h}\|, \quad \bar{\mu}_{r,h} = \max_{k \leq h} \mu_{r,k}, \tag{40}$$

$$X_{r,h} = \frac{1}{\bar{\mu}_{r,h}} \frac{1}{M^2} \sum_{m=1}^M \sum_{s=1}^M \langle T_\eta(y_{m,r,h}) - T_\eta(y_{r,h}), \rho_{m,r,h} - \rho_{s,r,h} \rangle. \tag{41}$$

1158 Let $\mathbb{E}_{r,h}[\cdot]$ denote the expectation conditional on all information up to and including round r and
1159 local step h . Then,

$$\mathbb{E}_{r,h}[X_{r,h}] = 0.$$

1160 Furthermore, we have by the triangle inequality, then our assumption on the noise followed by
1161 Lemma B.1 that almost surely

$$\begin{aligned}
|\langle T_\eta(y_{m,r,h}) - T_\eta(y_{r,h}), \rho_{m,r,h} - \rho_{s,r,h} \rangle| &\leq \|T_\eta(y_{m,r,h}) - T_\eta(y_{r,h})\| \|\rho_{m,r,h} - \rho_{s,r,h}\| \\
&\leq \|T_\eta(y_{m,r,h}) - T_\eta(y_{r,h})\| (\|\rho_{m,r,h}\| + \|\rho_{s,r,h}\|) \\
&\leq 2\sigma \|T_\eta(y_{m,r,h}) - T_\eta(y_{r,h})\| \\
&\leq 2\sigma \|y_{m,r,h} - y_{r,h}\|. \tag{42}
\end{aligned}$$

1162 By the definition of $X_{r,h}$ (Equation (41)), the triangle inequality, Equation (42), and the definition of
 1163 $\bar{\mu}_{r,h}$ (Equation (40)) we have almost surely

$$\begin{aligned}
 |X_{r,h}| &= \frac{1}{\bar{\mu}_{r,h}} \left| \frac{1}{M^2} \sum_{m=1}^M \sum_{s=1}^M \langle T_\eta(y_{m,r,h}) - T_\eta(y_{r,h}), \rho_{m,r,h} - \rho_{s,r,h} \rangle \right| \\
 &\leq \frac{1}{\bar{\mu}_{r,h}} \frac{1}{M^2} \sum_{m=1}^M \sum_{s=1}^M |\langle T_\eta(y_{m,r,h}) - T_\eta(y_{r,h}), \rho_{m,r,h} - \rho_{s,r,h} \rangle| \\
 &\leq \frac{2\sigma}{\bar{\mu}_{r,h}} \frac{1}{M^2} \sum_{m=1}^M \sum_{s=1}^M \|y_{m,r,h} - y_{r,h}\| \\
 &= 2\sigma \frac{\frac{1}{M} \sum_{m=1}^M \|y_{m,r,h} - y_{r,h}\|}{\bar{\mu}_{r,h}} \\
 &\leq 2\sigma.
 \end{aligned}$$

1164 Then by Lemma B.4 with $y_h = \bar{\mu}_{r,h}$ we have with probability at least $1 - \delta$

$$\begin{aligned}
 \left| \sum_{k=0}^{h-1} \bar{\mu}_{r,k} X_{r,k} \right| &\leq 8\bar{\mu}_{r,h-1} \sqrt{\theta_{h-1,\delta} \sum_{k=0}^{h-1} X_{r,k}^2 + 4\sigma^2 \theta_{h,\delta}^2} \\
 &\leq 8\bar{\mu}_{r,h-1} \sqrt{\theta_{h-1,\delta} 4h\sigma^2 + 4\sigma^2 \theta_{h,\delta}^2} \\
 &\leq 16\bar{\mu}_{r,h-1} \theta_{h-1,\delta} \sigma \sqrt{h+1}.
 \end{aligned} \tag{43}$$

1165 Observe that

$$\sum_{k=0}^{h-1} \bar{\mu}_{r,k} X_{r,k} = \frac{1}{M^2} \sum_{k=0}^{h-1} \sum_{m=1}^M \sum_{s=1}^M \langle T_\eta(y_{m,r,k}) - T_\eta(y_{r,k}), \rho_{m,r,k} - \rho_{s,r,k} \rangle.$$

1166 Using this and Equation (43) to upper bound the right hand side of Equation (39) we obtain

$$\begin{aligned}
 \Lambda_{r,h} &\leq 4\eta^2 \sigma^2 h + 32\eta \bar{\mu}_{r,h-1} \theta_{h-1,\delta} \sigma \sqrt{h+1} \\
 &\leq 4\eta^2 \sigma^2 h + 2\alpha (32\eta \theta_{h-1,\delta} \sigma \sqrt{h+1})^2 + \frac{\bar{\mu}_{r,h-1}^2}{2\alpha} \\
 &= \eta^2 \sigma^2 (h+1) \theta_{h-1,\delta}^2 (4 + 2048\alpha) + \frac{\bar{\mu}_{r,h-1}^2}{2\alpha},
 \end{aligned} \tag{44}$$

1167 where we used that $2ab \leq \alpha a^2 + \frac{1}{\alpha} b^2$ in the second step. Let $\bar{\Lambda}_{r,h} = \max_{k \leq h} \Lambda_{r,k}$. Observe that the
 1168 right hand side of Equation (44) is increasing in h , therefore

$$\bar{\Lambda}_{r,h} \leq \eta^2 \sigma^2 (h+1) \theta_{h-1,\delta}^2 (4 + 2048\alpha) + \frac{\bar{\mu}_{r,h-1}^2}{2\alpha}. \tag{45}$$

1169 Observe that by the triangle inequality followed by Lemma B.2

$$\begin{aligned}
 \mu_{r,h} &= \frac{1}{M} \sum_{m=1}^M \|y_{m,r,h} - y_{r,h}\| \\
 &\leq \frac{1}{M^2} \sum_{m=1}^M \sum_{s=1}^M \|y_{m,r,h} - y_{s,r,h}\| \\
 &\leq \sqrt{\frac{1}{M^2} \sum_{m=1}^M \sum_{s=1}^M \|y_{m,r,h} - y_{s,r,h}\|^2} \\
 &= \sqrt{\Lambda_{r,h}}.
 \end{aligned}$$

1170 It follows that $\bar{\mu}_{r,h} \leq \sqrt{\bar{\Lambda}_{r,h}}$. Using this in Equation (45) we get

$$\begin{aligned}\bar{\Lambda}_{r,h} &\leq \eta^2 \sigma^2 (h+1) \theta_{h-1,\delta}^2 (4 + 2048\alpha) + \frac{\bar{\Lambda}_{r,h-1}}{2\alpha} \\ &\leq \eta^2 \sigma^2 (h+1) \theta_{h-1,\delta}^2 (4 + 2048\alpha) + \frac{\bar{\Lambda}_{r,h}}{2\alpha}.\end{aligned}$$

1171 Rearranging we get

$$\left(1 - \frac{1}{2\alpha}\right) \bar{\Lambda}_{r,h} \leq \eta^2 \sigma^2 (h+1) \theta_{h-1,\delta}^2 (4 + 2048\alpha)$$

1172 Put $\alpha = 1$, then

$$\bar{\Lambda}_{r,h} \leq 4104 \eta^2 \sigma^2 (h+1) \theta_{h-1,\delta}^2. \quad (46)$$

1173 Now that we have our bound on $\bar{\Lambda}_{r,h}$, we can use it to bound $V_{r,h}$ as follows

$$V_{r,h} = \frac{1}{M} \sum_{m=1}^M \|y_{m,r,h} - y_{r,h}\|^2. \quad (47)$$

1174 Observe that by Jensen's inequality

$$\begin{aligned}\|y_{m,r,h} - y_{r,h}\|^2 &= \left\| y_{m,r,h} - \frac{1}{M} \sum_{s=1}^M y_{s,r,h} \right\|^2 \\ &= \left\| \frac{1}{M} (y_{m,r,h} - y_{s,r,h}) \right\|^2 \\ &\leq \frac{1}{M} \sum_{s=1}^M \|y_{m,r,h} - y_{s,r,h}\|^2.\end{aligned} \quad (48)$$

1175 Combining Equations (47) and (48) we have

$$V_{r,h} \leq \frac{1}{M^2} \sum_{m=1}^M \sum_{s=1}^M \|y_{m,r,h} - y_{s,r,h}\|^2 = \Lambda_{r,h}.$$

1176 Combining this with Equation (46) yields the lemma's statement. \square

1177 **Lemma B.10.** (Per-round regret). In Algorithm 1, the iterates in a single communication round
1178 satisfy

$$\begin{aligned}\|x_{r+1} - x_*\|^2 &\leq \|x_r - x_*\|^2 + \gamma^2 \eta^2 \sum_{h=0}^{H-1} \|g_{r,h}\|^2 + 2\gamma\eta |1 - \gamma| \zeta_2 \sum_{h=0}^{H-1} \|g_{r,h}\| \\ &\quad + \frac{\gamma\zeta_3 H}{\alpha} + \frac{\alpha\gamma\eta^2}{2} \frac{1}{M} \sum_{m=1}^M \sum_{h=0}^{H-1} \|g_{m,r,h}\|^2 - \frac{2\gamma\eta}{M} \sum_{h=0}^{H-1} \sum_{m=1}^M \langle y_{m,r,h} - x_*, g_{m,r,h} \rangle,\end{aligned}$$

1179 where $\alpha > 0$ is arbitrary and

$$\zeta_2 = \max_h \|y_{r,h} - y_{r,0}\|, \quad \zeta_3 = \max_h \frac{1}{M} \sum_{m=1}^M \|y_{m,r,h} - y_{r,h}\|^2.$$

1180 *Proof.* Define the virtual sequences

$$g_{r,h} = \frac{1}{M} \sum_{m=1}^M g_{m,r,h}, \quad x_{r,0} = x_r, \quad x_{r,h+1} = x_{r,h} - \gamma\eta g_{r,h}.$$

1181 We have

$$\|x_{r,h+1} - x_*\|^2 = \|x_{r,h} - x_*\|^2 + \gamma^2 \eta^2 \|g_{r,h}\|^2 - 2\gamma\eta \langle x_{r,h} - x_*, g_{r,h} \rangle \quad (49)$$

1182 The inner product term can be decomposed as

$$-\langle x_{r,h} - x_*, g_{r,h} \rangle = -\langle x_{r,h} - y_{r,h}, g_{r,h} \rangle - \langle y_{r,h} - x_*, g_{r,h} \rangle. \quad (50)$$

1183 Observe that $x_{r,h} = x_r - \gamma\eta \sum_{s=0}^{h-1} g_{r,s}$ and $y_{r,h} = x_r - \eta \sum_{s=0}^{h-1} g_{r,s}$. Therefore,

$$\begin{aligned} \|x_{r,h} - y_{r,h}\| &= \left\| (\gamma - 1)\eta \sum_{s=0}^{h-1} g_{r,s} \right\| \\ &= |\gamma - 1| \|y_{r,h} - y_{r,0}\| \\ &\leq |\gamma - 1| \zeta_2, \end{aligned}$$

1184 where $\zeta_2 = \max_h \|y_{r,h} - y_{r,0}\|$. Using this in Equation (50)

$$-\langle x_{r,h} - y_{r,h}, g_{r,h} \rangle \leq \|x_{r,h} - y_{r,h}\| \|g_{r,h}\| \leq |1 - \gamma| \zeta_2 \|g_{r,h}\|. \quad (51)$$

1185 Plugging Equation (51) into Equation (50) we get

$$\begin{aligned} -\langle x_{r,h} - x_*, g_{r,h} \rangle &\leq |1 - \gamma| \zeta_2 \|g_{r,h}\| - \langle y_{r,h} - x_*, g_{r,h} \rangle \\ &= |1 - \gamma| \zeta_2 \|g_{r,h}\| - \frac{1}{M} \sum_{m=1}^M \langle y_{r,h} - x_*, g_{m,r,h} \rangle \\ &= |1 - \gamma| \zeta_2 \|g_{r,h}\| - \frac{1}{M} \sum_{m=1}^M \langle y_{r,h} - y_{m,r,h}, g_{m,r,h} \rangle - \frac{1}{M} \sum_{m=1}^M \langle y_{m,r,h} - x_*, g_{m,r,h} \rangle. \quad (52) \end{aligned}$$

1186 For the second term in Equation (52) we have

$$\begin{aligned} -\frac{1}{M} \sum_{m=1}^M \langle y_{r,h} - y_{m,r,h}, g_{m,r,h} \rangle &\leq \frac{1}{M} \sum_{m=1}^M \|y_{r,h} - y_{m,r,h}\| \|g_{m,r,h}\| \\ &\leq \frac{1}{M} \sum_{m=1}^M \left[\frac{\|y_{r,h} - y_{m,r,h}\|^2}{2\alpha\eta} + \frac{\alpha\eta}{2} \|g_{m,r,h}\|^2 \right] \\ &\leq \frac{\zeta_3}{2\alpha\eta} + \frac{\alpha\eta}{2} \frac{1}{M} \sum_{m=1}^M \|g_{m,r,h}\|^2. \quad (53) \end{aligned}$$

1187 Plugging Equation (53) into Equation (52) we get

$$\begin{aligned} -\langle x_{r,h} - x_*, g_{r,h} \rangle &\leq |1 - \gamma| \zeta_2 \|g_{r,h}\| + \frac{\zeta_3}{2\alpha\eta} + \frac{\alpha\eta}{2} \frac{1}{M} \sum_{m=1}^M \|g_{m,r,h}\|^2 \\ &\quad - \frac{1}{M} \sum_{m=1}^M \langle y_{m,r,h} - x_*, g_{m,r,h} \rangle. \quad (54) \end{aligned}$$

1188 Plug Equation (54) back into Equation (49) to get

$$\begin{aligned} \|x_{r,h+1} - x_*\|^2 &\leq \|x_{r,h} - x_*\|^2 + \gamma^2 \eta^2 \|g_{r,h}\|^2 + 2\gamma\eta |1 - \gamma| \zeta_2 \|g_{r,h}\| \\ &\quad + \frac{\gamma\zeta_3}{\alpha} + \frac{\alpha\gamma\eta^2}{M} \sum_{m=1}^M \|g_{m,r,h}\|^2 - \frac{2\gamma\eta}{M} \sum_{m=1}^M \langle y_{m,r,h} - x_*, g_{m,r,h} \rangle. \end{aligned}$$

1189 Recursing we get

$$\begin{aligned} \|x_{r+1} - x_*\|^2 &\leq \|x_r - x_*\|^2 + \gamma^2 \eta^2 \sum_{h=0}^{H-1} \|g_{r,h}\|^2 + 2\gamma\eta |1 - \gamma| \zeta_2 \sum_{h=0}^{H-1} \|g_{r,h}\| \\ &\quad + \frac{\gamma\zeta_3 H}{\alpha} + \frac{\alpha\gamma\eta^2}{2} \frac{1}{M} \sum_{m=1}^M \sum_{h=0}^{H-1} \|g_{m,r,h}\|^2 - \frac{2\gamma\eta}{M} \sum_{h=0}^{H-1} \sum_{m=1}^M \langle y_{m,r,h} - x_*, g_{m,r,h} \rangle. \end{aligned}$$

1190 □

1191 *Proof of Theorem 3.6.* Starting with the per-round recursion lemma, we have

$$\begin{aligned} \|x_{r+1} - x_*\|^2 &\leq \|x_r - x_*\|^2 + \gamma^2 \eta^2 \sum_{h=0}^{H-1} \|g_{r,h}\|^2 + 2\gamma |1 - \gamma| \zeta_2 \sum_{h=0}^{H-1} \|g_{r,h}\| \\ &\quad + \frac{\gamma \zeta_3 H}{\alpha} + \frac{\alpha \gamma \eta^2}{2} \frac{1}{M} \sum_{m=1}^M \sum_{h=0}^{H-1} \|g_{m,r,h}\|^2 - \frac{2\gamma \eta}{M} \sum_{h=0}^{H-1} \sum_{m=1}^M \langle y_{m,r,h} - x_*, g_{m,r,h} \rangle. \end{aligned}$$

1192 Observe that

$$\begin{aligned} \|y_{r,h} - y_{r,0}\| &= \eta \left\| \sum_{k=0}^{h-1} g_{r,k} \right\| \\ &\leq \eta \sum_{k=0}^{h-1} \|g_{r,k}\| \\ &\leq \eta \sum_{k=0}^{H-1} \|g_{r,k}\|. \end{aligned} \tag{55}$$

1193 Since this holds for any h , we have that $\zeta_2 \leq \eta \sum_{k=0}^{H-1} \|g_{r,k}\|$, where ζ_2 is defined in Lemma B.10.
 1194 Moreover, by Lemma B.9 we have that with probability $1 - \delta$ and an application of the union bound
 1195 that for all r, h

$$\frac{1}{M} \sum_{m=1}^M \|y_{m,r,h} - y_{r,h}\|^2 \leq 4104\iota \eta^2 \sigma^2 H, \tag{56}$$

1196 where $\iota = 2 \cdot \log \frac{60 \log 6RH}{\delta}$ and we used that $H + 1 \leq 2H$. Since this bound holds for all h , we have

$$\zeta_3 = \max_h \frac{1}{M} \sum_{m=1}^M \|y_{m,r,h} - y_{r,h}\|^2 \leq 4104\iota \eta^2 \sigma^2 H.$$

1197 Therefore by Equation (55) and Lemma B.9

$$\begin{aligned} \|x_{r+1} - x_*\|^2 &\leq \|x_r - x_*\|^2 + \gamma^2 \eta^2 \sum_{h=0}^{H-1} \|g_{r,h}\|^2 + 2\gamma |1 - \gamma| \eta^2 \left(\sum_{h=0}^{H-1} \|g_{r,h}\| \right)^2 \\ &\quad + \frac{4104\gamma \eta^2 \sigma^2 H^2}{\alpha} \iota + \frac{\alpha \gamma \eta^2}{2} \frac{1}{M} \sum_{m=1}^M \sum_{h=0}^{H-1} \|g_{m,r,h}\|^2 - \frac{2\gamma \eta}{M} \sum_{h=0}^{H-1} \sum_{m=1}^M \langle y_{m,r,h} - x_*, g_{m,r,h} \rangle. \end{aligned}$$

1198 Let $\xi_{m,r,h} = g_{m,r,h} - \nabla f(y_{m,r,h})$. Then,

$$\begin{aligned} \|x_{r+1} - x_*\|^2 &\leq \|x_r - x_*\|^2 + \gamma^2 \eta^2 \sum_{h=0}^{H-1} \|g_{r,h}\|^2 + 2\gamma |1 - \gamma| \eta^2 \left(\sum_{h=0}^{H-1} \|g_{r,h}\| \right)^2 + \frac{4104\gamma \eta^2 \sigma^2 H^2}{\alpha} \iota \\ &\quad + \frac{\alpha \gamma \eta^2}{2} \frac{1}{M} \sum_{m=1}^M \sum_{h=0}^{H-1} \|g_{m,r,h}\|^2 - \frac{2\gamma \eta}{M} \sum_{h=0}^{H-1} \sum_{m=1}^M \langle y_{m,r,h} - x_*, \nabla f(y_{m,r,h}) \rangle \\ &\quad - \frac{2\gamma \eta}{M} \sum_{h=0}^{H-1} \sum_{m=1}^M \langle y_{m,r,h} - x_*, \xi_{m,r,h} \rangle, \end{aligned} \tag{57}$$

1199 where $\xi_{m,r,h} = g_{m,r,h} - \nabla f(y_{m,r,h})$. Define

$$\nu_{r,h} = \frac{1}{M} \sum_{m=1}^M \|y_{m,r,h} - x_*\|, \quad \bar{\nu}_{r,h} = \max_{p \leq r, s \leq h} \nu_{p,s}.$$

1200 Let

$$X_{r,h} = \frac{1}{\bar{\nu}_{r,h}} \frac{1}{M} \sum_{m=1}^M \langle y_{m,r,h} - x_*, \xi_{m,r,h} \rangle$$

1201 Let $\mathcal{F}_{r,h-1}$ denote the sigma algebra generated by all randomness up to and including step $r, h-1$.
 1202 Note that

$$\begin{aligned} \mathbb{E}_{\mathcal{F}_{r,h-1}} [X_{r,h}] &= \frac{1}{\bar{\nu}_{r,h}} \frac{1}{M} \sum_{m=1}^M \mathbb{E}_{\mathcal{F}_{r,h}} [\langle y_{m,r,h} - x_*, \xi_{m,r,h} \rangle] \\ &= \frac{1}{\bar{\nu}_{r,h}} \frac{1}{M} \sum_{m=1}^M \langle y_{m,r,h} - x_*, \mathbb{E}_{\mathcal{F}_{r,h}} [\xi_{m,r,h}] \rangle \\ &= 0, \end{aligned}$$

1203 where we used that $\nu_{r,h}$ and $y_{m,r,h}$ are both $\mathcal{F}_{r,h-1}$ -measurable and that the noise has mean zero.
 1204 The edge cases $X_{r,0}$ are handled similarly. Moreover, using the assumption that $\|\xi_{m,r,h}\| \leq \sigma$ almost
 1205 surely and the definition of $\bar{\nu}_{r,h}$,

$$\begin{aligned} \|X_{r,h}\| &= \left\| \frac{1}{\bar{\nu}_{r,h}} \frac{1}{M} \sum_{m=1}^M \langle y_{m,r,h} - x_*, \xi_{m,r,h} \rangle \right\| \\ &\leq \frac{1}{M} \sum_{m=1}^M \frac{\|y_{m,r,h} - x_*\| \|\xi_{m,r,h}\|}{\bar{\nu}_{r,h}} \\ &\leq \frac{1}{M} \sum_{m=1}^M (1 \cdot \sigma) \\ &= \sigma. \end{aligned}$$

1206 Applying Lemma B.4 on $X_{r,h}$ with $y_{r,h} = \bar{\nu}_{r,h}$, $C_{r,h} = \sigma$, $\hat{X}_{r,h} = 0$ we have

$$\left| \frac{1}{M} \sum_{r=0}^{R-1} \sum_{h=0}^{H-1} \sum_{m=1}^M \langle y_{m,r,h} - x_*, \xi_{m,r,h} \rangle \right| \leq 16\bar{\nu}_{R,H}\iota\sigma\sqrt{RH}, \quad (58)$$

1207 where ι is defined as before. Using Equation (58) in Equation (57)

$$\begin{aligned} \frac{2\gamma\eta}{M} \sum_{m,r,h} \langle y_{m,r,h} - x_*, \nabla f(y_{m,r,h}) \rangle &\leq \|x_0 - x_*\|^2 - \|x_R - x_*\|^2 + \gamma^2\eta^2 \sum_{r,h} \|g_{r,h}\|^2 \\ &\quad + 2\gamma|1-\gamma|\eta^2 \sum_{r=0}^{R-1} \left(\sum_{h=0}^{H-1} \|g_{r,h}\| \right)^2 + R \cdot \frac{4104\gamma\eta^2\sigma^2H^2}{\alpha} \iota \\ &\quad + \frac{\alpha\gamma\eta^2}{2} \frac{1}{M} \sum_{m,r,h} \|g_{m,r,h}\|^2 + 2\gamma\eta \left[16\bar{\nu}_{R,H}\iota\sigma\sqrt{RH} \right]. \end{aligned} \quad (59)$$

1208 Let

$$\begin{aligned} \Omega &= \gamma^2\eta^2 \sum_{r,h} \|g_{r,h}\|^2 + 2\gamma|1-\gamma|\eta^2 \sum_{r=0}^{R-1} \left(\sum_{h=0}^{H-1} \|g_{r,h}\| \right)^2 + R \cdot \frac{4104\gamma\eta^2\sigma^2H^2}{\alpha} \iota \\ &\quad + \frac{\alpha\gamma\eta^2}{2} \frac{1}{M} \sum_{m,r,h} \|g_{m,r,h}\|^2 \end{aligned} \quad (60)$$

1209 Then by convexity and Equation (59) we get

$$\begin{aligned} \|x_R - x_*\|^2 &\leq \|x_0 - x_*\|^2 + \Omega + 2\gamma\eta \left[16\bar{\nu}_{R,H}\iota\sigma\sqrt{RH} \right] - \frac{2\gamma\eta}{M} \sum_{m,r,h} \langle y_{m,r,h} - x_*, \nabla f(y_{m,r,h}) \rangle \\ &\leq \|x_0 - x_*\|^2 + \Omega + 2\gamma\eta \left[16\bar{\nu}_{R,H}\iota\sigma\sqrt{RH} \right], \end{aligned} \quad (61)$$

1210 where in the second line we used that x_* is the minimizer of f and therefore
 1211 $\langle y_{m,r,h} - x_*, \nabla f(y_{m,r,h}) \rangle \geq 0$ by convexity. It is not difficult to see that this guarantee in fact
 1212 applies not just on $\|x_R - x_*\|^2$ but on any x_r . Let $d_r = \|x_r - x_*\|$ and $\bar{d}_r = \max_{r' \leq r} d_{r'}$. Observe

$$\begin{aligned} \nu_{r,h} &= \frac{1}{M} \sum_{m=1}^M \|y_{m,r,h} - x_*\| \leq \frac{1}{M} \sum_{m=1}^M [\|y_{m,r,h} - y_{m,r,0}\| + \|x_r - x_*\|] \\ &\leq \left[\frac{\eta}{M} \sum_{m=1}^M \sum_{k=0}^{h-1} \|g_{m,r,k}\| \right] + \|x_r - x_*\| \\ &\leq \left[\frac{\eta}{M} \sum_{m=1}^M \sum_{k=0}^{H-1} \|g_{m,r,k}\| \right] + \|x_r - x_*\|. \end{aligned} \quad (62)$$

1213 Using Equation (62) in Equation (61) we get

$$\begin{aligned} \bar{d}_R^2 &\leq d_0^2 + \Omega + 32\gamma\eta\iota\sigma\sqrt{RH}\bar{\nu}_{R,H} \\ &\leq d_0^2 + \Omega + 32\gamma\eta\iota\sigma\sqrt{RH} \left[\frac{\eta}{M} \sum_{m,h} \|g_{m,r,h}\| \right] + 32\gamma\eta\iota\sigma\sqrt{RH}\bar{d}_R \\ &\leq d_0^2 + \Omega + 2 \left(32\gamma\eta\iota\sigma\sqrt{RH} \right)^2 + \eta^2 \left(\frac{1}{M} \sum_{m,h} \|g_{m,r,h}\| \right)^2 + \frac{\bar{d}_R^2}{2}. \end{aligned}$$

1214 Therefore

$$\bar{d}_R^2 \leq 2d_0^2 + 2\Omega + 4096\gamma^2\eta^2\iota^2\sigma^2RH + 2\eta^2 \left(\frac{1}{M} \max_r \sum_{m,h} \|g_{m,r,h}\| \right)^2. \quad (63)$$

1215 By the triangle inequality applied twice and the definition of \bar{d}_R ,

$$\begin{aligned} \|y_{m,r,s} - x_*\| &\leq \|y_{m,r,0} - y_{m,r,s}\| + \|y_{m,r,0} - x_*\| \\ &= \eta \left\| \sum_{h=0}^{s-1} g_{m,r,h} \right\| + \|y_{m,r,0} - x_*\| \\ &\leq \eta \sum_{h=0}^{s-1} \|g_{m,r,h}\| + \|y_{m,r,0} - x_*\| \\ &\leq \eta \sum_{h=0}^{s-1} \|g_{m,r,h}\| + \bar{d}_R \\ &\leq \eta \sum_{h=0}^{H-1} \|g_{m,r,h}\| + \bar{d}_R. \end{aligned}$$

1216 Therefore

$$\frac{1}{M} \sum_{m=1}^M \|y_{m,r,s} - x_*\| \leq \eta \left(\frac{1}{M} \sum_{m=1}^M \sum_{h=0}^{H-1} \|g_{m,r,h}\| \right) + \bar{d}_R$$

1217 We now use the inequality $(a + b)^2 \leq 2a^2 + 2b^2$ to get

$$\begin{aligned}\nu_{r,s}^2 &= \left(\frac{1}{M} \sum_{m=1}^M \|y_{m,r,s} - x_*\| \right)^2 \\ &\leq 2 \left(\eta \left(\frac{1}{M} \sum_{m=1}^M \sum_{h=0}^{H-1} \|g_{m,r,h}\| \right) \right)^2 + 2\bar{d}_R^2 \\ &= 2\eta^2 \left(\frac{1}{M} \sum_{m,h} \|g_{m,r,h}\| \right)^2 + 2\bar{d}_R^2.\end{aligned}$$

1218 Finally, using our bound on \bar{d}_R^2 given by equation (63)

$$\nu_{r,s}^2 \leq 4d_0^2 + 4\Omega + 8192\gamma^2\eta^2\iota^2\sigma^2RH + 6\eta^2 \left(\frac{1}{M} \sum_{m,h} \|g_{m,r,h}\| \right)^2,$$

1219 Therefore

$$\begin{aligned}\bar{\nu}_{R,H}^2 &= \max_{r,s} \nu_{r,s}^2 \\ &\leq 4d_0^2 + 4\Omega + 8192\gamma^2\eta^2\iota^2\sigma^2RH + 6\eta^2 \left(\frac{1}{M} \max_r \sum_{m,h} \|g_{m,r,h}\| \right)^2.\end{aligned}$$

1220 By Equations (59) and (60) and the last equation,

$$\begin{aligned}\frac{2\gamma\eta}{M} \sum_{m,r,h} \langle y_{m,r,h} - x_*, \nabla f(y_{m,r,h}) \rangle &\leq \|x_0 - x_*\|^2 - \|x_R - x_*\|^2 + \Omega + 2\gamma\eta \left[16\bar{\nu}_{R,H}\iota\sigma\sqrt{RH} \right] \\ &\leq d_0^2 - d_R^2 + \Omega + \frac{(32\gamma\eta\iota\sigma\sqrt{RH})^2}{2} + 4 \left[d_0^2 + \Omega + 2048\gamma^2\eta^2\iota^2\sigma^2RH \right] + 6\eta^2 R \left(\frac{1}{M} \max_r \sum_{m,h} \|g_{m,r,h}\| \right)^2 \\ &= d_0^2 - d_R^2 + \Omega + \frac{(32\gamma\eta\iota\sigma\sqrt{RH})^2}{2} + 4 \left[d_0^2 + \Omega + 2048\gamma^2\eta^2\iota^2\sigma^2RH \right] + 6\eta^2 R \left(\frac{1}{M} \max_r \sum_{m,h} \|g_{m,r,h}\| \right)^2 \\ &\leq d_0^2 - d_R^2 + 6\Omega + 8704\gamma^2\eta^2\iota^2\sigma^2RH + 4d_0^2 + 6\eta^2 R \left(\frac{1}{M} \max_r \sum_{m,h} \|g_{m,r,h}\| \right)^2. \quad (64)\end{aligned}$$

1221 Dropping the $-d_R^2$ term, we get

$$\begin{aligned}\frac{2\gamma\eta}{M} \sum_{m,r,h} \langle y_{m,r,h} - x_*, \nabla f(y_{m,r,h}) \rangle &\leq 5d_0^2 + 6\Omega + 8704\gamma^2\eta^2\iota^2\sigma^2RH + 6\eta^2 R \left(\frac{1}{M} \max_r \sum_{m,h} \|g_{m,r,h}\| \right)^2 \\ &\leq 5d_0^2 + 6\gamma^2\eta^2 \sum_{r,h} \|g_{r,h}\|^2 + 12\gamma|1 - \gamma|\eta^2 \sum_{r=0}^{R-1} \left(\sum_{h=0}^{H-1} \|g_{r,h}\| \right)^2 + RH \frac{24624\gamma\eta^2\sigma^2H\iota}{\alpha} \\ &\quad + \frac{3\alpha\gamma\eta^2}{M} \sum_{m,r,h} \|g_{m,r,h}\|^2 + 8704\gamma^2\eta^2\iota^2\sigma^2RH + 6\eta^2 R \left(\frac{1}{M} \max_r \sum_{m,h} \|g_{m,r,h}\| \right)^2.\end{aligned}$$

1222 Dividing both sides by $2\gamma\eta RH$ gives

$$\begin{aligned}
\frac{1}{MRH} \sum_{m,r,h} \langle y_{m,r,h} - x_*, \nabla f(y_{m,r,h}) \rangle &\leq \frac{5d_0^2}{2\gamma\eta RH} + \frac{3\gamma\eta}{RH} \sum_{r,h} \|g_{r,h}\|^2 \\
&+ \frac{6|1-\gamma|\eta}{RH} \sum_{r=0}^{R-1} \left(\sum_{h=0}^{H-1} \|g_{r,h}\| \right)^2 + \frac{24624\eta\sigma^2 H\iota}{\alpha} \\
&+ \frac{3\alpha\eta}{MRH} \sum_{m,r,h} \|g_{m,r,h}\|^2 + 8704\gamma\eta\iota^2\sigma^2 + \frac{6\eta}{\gamma H} \left(\frac{1}{M} \max_r \sum_{m,h} \|g_{m,r,h}\| \right)^2.
\end{aligned} \tag{65}$$

1223 Observe that by optimizing over α we have

$$\begin{aligned}
\frac{24624\eta\sigma^2 H\iota}{\alpha} + \frac{3\alpha\eta}{MRH} \sum_{m,r,h} \|g_{m,r,h}\|^2 &\leq 2 \sqrt{(24624\eta\sigma^2 H\iota) \left(\frac{3\eta}{MRH} \sum_{m,r,h} \|g_{m,r,h}\|^2 \right)} \\
&\leq 544\eta\sigma\iota \sqrt{\frac{1}{MR} \sum_{m,r,h} \|g_{m,r,h}\|^2}.
\end{aligned}$$

1224 Using this in Equation (65) followed by convexity completes the proof. \square