Reference	Setting	Best Convergence rate (i.e., $\mathbb{E}\left[F(x^{output}) - F(x^\star)\right] \lesssim$)
Stich (2018)	SC	$\frac{\sigma^2}{\lambda MKR} + \frac{H\sigma^2}{\lambda^2 MK^2R^2} + \frac{H(H^2B^2+\sigma^2)}{\lambda^2R^2} + \frac{H^3(H^2B^2+\sigma^2)}{\lambda^4K^3R^3} + \frac{H^2B^2+\sigma^2}{\lambda R^3}$
	Non-SC	$\frac{\sigma B}{(MKR)^{1/2}} + \frac{HB^2\left(1+(H^{-1}B^{-1}\sigma)^{2/3}\right)}{R^{2/3}} + \frac{HB^2\left(1+(H^{-1}B^{-1}\sigma)^{2/5}\right)}{(KR)^{3/5}} + \frac{HB^2+B\sigma}{R^{3/2}}$
Stich and Karimireddy (2019)	SC	$HKMB^2\exp\left(-\frac{\lambda R}{10HM}\right) + \frac{\sigma^2}{\lambda MKR}$
	Non-SC	$\frac{HMB^2}{R} + \frac{\sigma B}{\sqrt{MKR}}$
Khaled et al. (2019)	SC	$\frac{HB^2}{K^2R^2} + \frac{H\sigma^2}{\lambda^2MKR} + \frac{H^2\sigma^2}{\lambda^3KR^2}$
	Non-SC	$\frac{HB^2}{\sqrt{KRM}} + \frac{\sigma^2}{H\sqrt{KRM}} + \frac{\sigma^2M}{HR}$

Table 2. Best convergence rates up to constants in previous analyses under our assumptions.

A. Comparisons Between Existing Local SGD Analyses and Minibatch SGD

In this section, we describe the derivation of the entries in Table 1 for the cases in which it is not obvious. In particular, these previous analyses were stated based on different assumptions (stronger as well as weaker) which need to be reconciled with ours. Since local SGD is often analyzed in the strongly convex setting (or with weaker assumptions that are implied by strong convexity), we will make use of the following fact: If an algorithm guarantees error at most $\epsilon(\lambda)$ when applied to a λ-strongly convex function, then we can apply the algorithm to $F(x) + \frac{\lambda}{2}\|x\|^2$ in order to ensure error $\epsilon(\lambda) + \frac{\lambda}{2}\|x^*\|^2$. This applies for any $\lambda > 0$, so we can actually infer that the algorithm, in fact, guarantees error at most $\min_{\lambda>0}\epsilon(\lambda) + \frac{\lambda}{2}\|x^*\|^2$.

Since our purpose is to show that these analyses are dominated by minibatch SGD, the entries in the table are, in some sense, the most optimistic interpretation of the bounds stated in the paper. For example, if error $\epsilon_1(\lambda) + \epsilon_2(\lambda)$ is guaranteed for strongly convex functions, we actually enter $\frac{1}{2}\min_{\lambda>0}\epsilon_1(\lambda) + \frac{\lambda}{2}\|x^*\|^2 + \frac{1}{2}\min_{\lambda>0}\epsilon_2(\lambda) + \frac{\lambda}{2}\|x^*\|^2$ into the table, which is a lower bound on the actual guarantee.

For reference, we restate the worst-case guarantee of minibatch SGD:

$$\epsilon_{\text{MB-SGD}} \asymp \frac{HB^2}{R} + \frac{\sigma B}{\sqrt{MKR}} \tag{13}$$

A.1. Stich (2018)

The paper makes the same assumptions as us but, in addition, assumes that the stochastic gradients are uniformly bounded, i.e. $\mathbb{E}_{z\sim\mathcal{D}}\left[\|\nabla f(x;z)\|^2\right] \leq G^2$, $\forall x$. We relax this assumption by noting the following,

$$\begin{aligned}
\mathbb{E}_{z\sim\mathcal{D}}\left[\|\nabla f(x;z)\|^2\right] &= \mathbb{E}_{z\sim\mathcal{D}}\left[\|\nabla f(x;z) - \nabla f(x^\star;z) + \nabla f(x^\star;z) - \nabla F(x^\star)\|^2\right] && (14)\\
&\lesssim \mathbb{E}_{z\sim\mathcal{D}}\left[\|\nabla f(x;z) - \nabla f(x^\star;z)\|^2\right] + \mathbb{E}_{z\sim\mathcal{D}}\left[\|\nabla f(x^\star;z) - \nabla F(x^\star)\|^2\right] && (15)\\
&\lesssim H^2\|x - x^\star\|^2 + \sigma^2 && (16)\\
&\lesssim H^2\|x^\star\|^2 + \sigma^2 && (17)\\
&\leq H^2B^2 + \sigma^2 && (18)
\end{aligned}$$

In the last step we make the optimistic assumption that the iterates stray no farther from x^* than they were at initialization, i.e. $\|x_0 - x^*\| \leq B$. This *may not be true*, so this bound is optimistic. On the other hand, it is clear that one cannot generally upper bound $\mathbb{E}_{z\sim\mathcal{D}}\left[\|\nabla f(x;z)\|^2\right]$ any tighter than this in our setting. Since our goal is anyways to show that the analysis of Stich (2018) is deficient, we continue using the bound (18). This immediately gives the result for the strongly-convex setting in appendix A. For the non-strongly setting we extend their result by optimizing each term separately as $\epsilon(\lambda) + \frac{\lambda}{2}B^2$ and ignore the constants.

A.2. Stich and Karimireddy (2019)

The paper relaxes the convexity assumption, by assuming F is $\lambda^\star$-quasi convex, i.e., $\forall x\; F(x^\star) \leq F(x) + \langle \nabla F,\, x^\star - x\rangle + \frac{\lambda^\star}{2}\|x - x^\star\|^2$. This condition can also hold for certain non-convex functions and is implied by $\lambda^\star$-strong convexity. Besides they assume H-smoothness of F and multiplicative noise for the stochastic gradients, i.e., $\mathop{\mathbb{E}}_{z\sim\mathcal{D}}\left[\|\nabla f(x;z) - \nabla F(x)\|^2\right] \leq N\|x - x^\star\|^2 + \sigma_\star^2$. The latter assumption is a relaxation of the uniform upper bound on the variance of the stochastic gradients, which we have assumed. Thus to compare to their result we set $N = 0$ upper bounding the stochastic variance by σ^2 and use the strong convexity constant λ instead of $\lambda^\star$. For the non-strongly convex setting we use their rate, along with our uniform variance bound. Besides they use specific learning rate and averaging schedules to optimize their rates. Both these rates are given in Appendix A. For the general convex setting, we believe their dependence in M is poor and is improved upon by our upper bound in Section 4.

A.3. Khaled et al. (2019)

The relevant analysis from Khaled et al. (2019) is given in their Corollary 2, which is their only analysis that upper bounds the error in terms of the objective function suboptimality and in the setting where each machine receives i.i.d. stochastic gradients. Their Corollary 2 states that when $M \leq KR$, the error is bounded by[5]

$$\epsilon_{\text{L-SGD}} \leq \frac{HB^2}{\sqrt{MKR}} + \frac{\sigma^2}{H\sqrt{MKR}} + \frac{\sigma^2 M}{HR} \tag{19}$$

In the case where $H = B = \sigma^2 = 1$, it is clear that this is strictly worse than minibatch SGD since $\frac{M}{R} > \frac{1}{R}$. However, consider the case of arbitrary H, B and σ^2 and suppose Khaled et al. (2019)'s guarantee is less than $\frac{\sigma B}{\sqrt{KR}}$, in which case

$$\frac{HB^2}{\sqrt{MKR}} \leq \frac{\sigma B}{\sqrt{KR}} \implies M \geq \frac{H^2B^2}{\sigma^2} \implies \frac{\sigma^2 M}{HR} \geq \frac{HB^2}{R} \tag{20}$$

Consequently, (19) is either greater than $\frac{\sigma B}{\sqrt{KR}}$ or greater than $\frac{HB^2}{R}$. This does not mean that their upper bound is worse than minibatch SGD. However, it *is* worse than minibatch SGD unless $\frac{\sigma B}{\sqrt{KR}} \leq \frac{HB^2}{R}$.

If we interrogate what this regime corresponds to, we see that it is actually a trivial regime where KR steps of serial SGD, which achieves error $\frac{HB^2}{KR} + \frac{\sigma B}{\sqrt{KR}} \leq \frac{HB^2}{R}$, is actually better than minibatch SGD. That is, rather than implementing minibatch SGD distributed across the M machines, we are actually better off just ignoring $M-1$ of the available machines and doing serial SGD. If this is really the right thing to do, then there was never any need for parallelism in the first place, and thus there is no reason to use local SGD, which performs no better than serial SGD in this case anyways.

B. Proofs from Section 3

Theorem 1. *Let $\mathcal{A}$ be a linear update algorithm which, when executed for T iterations on any quadratic $(f,\mathcal{D}) \in \mathcal{F}(H,\lambda,B,\sigma^2)$, guarantees $\mathbb{E}F(x_T) - F^* \leq \epsilon(T,\sigma^2)$. Then, local-$\mathcal{A}$'s averaged final iterate $\bar{x}_{KR} = \frac{1}{M}\sum_{m=1}^M x_{KR}^m$ will satisfy $\mathbb{E}F(\bar{x}_{KR}) - F^* \leq \epsilon(KR, \frac{\sigma^2}{M})$.*

Proof. We will show that the average of the iterates at any particular time $\bar{x}_t = \frac{1}{M}\sum_{m=1}^M x_t^m$ evolves according to $\mathcal{A}$ with a lower variance stochastic gradient, even though this average iterate is not explicitly computed by the algorithm at every step. It is easily confirmed from (6) that

$$\bar{x}_{t+1} = \frac{1}{M}\sum_{m'=1}^M \mathcal{L}_2^{(t)}\left(x_1^{m'},\dots,x_t^{m'},\nabla f\left(\mathcal{L}_1^{(t)}\left(x_1^{m'},\dots,x_t^{m'}\right);z_t^{m'}\right)\right) \tag{21}$$

$$= \mathcal{L}_2^{(t)}\left(\bar{x}_1,\dots,\bar{x}_t,\frac{1}{M}\sum_{m'=1}^M \nabla f\left(\mathcal{L}_1^{(t)}\left(x_1^{m'},\dots,x_t^{m'}\right);z_t^{m'}\right)\right) \tag{22}$$

[5] There is a typo in their statement which omits the factor of H (L in their notation) from the numerator of the first term.

where we used that $\mathcal{L}_2^{(t)}$ is linear. We will now show that $\frac{1}{M}\sum_{m'=1}^{M}\nabla f\left(\mathcal{L}_1^{(t)}\left(x_1^{m'},\dots,x_t^{m'}\right);z_t^{m'}\right)$ is an unbiased estimate of $\nabla F\left(\mathcal{L}_1^{(t)}(\bar{x}_1,\dots,\bar{x}_t)\right)$ with variance bounded by $\frac{\sigma^2}{M}$. Therefore, $\bar{x}_{t+1}$ is updated exactly according to $\mathcal{A}$ with a lower variance stochastic gradient.

By the linearity of $\mathcal{L}_1^{(t)}$ and ∇F

$$\mathbb{E}\left[\frac{1}{M}\sum_{m'=1}^{M}\nabla f\left(\mathcal{L}_1^{(t)}\left(x_1^{m'},\dots,x_t^{m'}\right);z_t^{m'}\right)\right]=\frac{1}{M}\sum_{m'=1}^{M}\nabla F\left(\mathcal{L}_1^{(t)}\left(x_1^{m'},\dots,x_t^{m'}\right)\right)=\nabla F\left(\mathcal{L}_1^{(t)}(\bar{x}_1,\dots,\bar{x}_t)\right) \quad (23)$$

Furthermore, since the z_t^m on each machine are independent, and $\sup_x \mathbb{E}\|\nabla f(x;z)-\nabla F(x)\|^2\le\sigma^2$,

$$\begin{aligned}\mathbb{E}\left\|\frac{1}{M}\sum_{m'=1}^{M}\nabla f\left(\mathcal{L}_1^{(t)}\left(x_1^{m'},\dots,x_t^{m'}\right);z_t^{m'}\right)-\mathbb{E}\left[\frac{1}{M}\sum_{m'=1}^{M}\nabla f\left(\mathcal{L}_1^{(t)}\left(x_1^{m'},\dots,x_t^{m'}\right);z_t^{m'}\right)\right]\right\|^2 \\ =\frac{1}{M^2}\sum_{m=1}^{M}\mathbb{E}\left\|\nabla f\left(\mathcal{L}_1^{(t)}(x_1^m,\dots,x_t^m);z_t^m\right)-\nabla F\left(\mathcal{L}_1^{(t)}(x_1^m,\dots,x_t^m)\right)\right\|^2\le\frac{\sigma^2}{M} \quad (24)\end{aligned}$$

□

Corollary 1. *For any quadratic $(f,\mathcal{D})\in\mathcal{F}(H,\lambda=0,B,\sigma^2)$, there are constants c_1 and c_2 such that local-SGD returns a point $\hat{x}$ such that*

$$\mathbb{E}F(\hat{x})-F^*\le c_1\left(\frac{HB^2}{KR}+\frac{\sigma B}{\sqrt{MKR}}\right),$$

and local-AC-SA returns a point $\tilde{x}$ such that

$$\mathbb{E}F(\tilde{x})-F^*\le c_2\left(\frac{HB^2}{K^2R^2}+\frac{\sigma B}{\sqrt{MKR}}\right).$$

In particular, local-AC-SA is minimax optimal for quadratic objectives.

Proof. It is easily confirmed that SGD and AC-SA (Ghadimi and Lan, 2013) are linear update algorithms, which allows us to apply Theorem 1. In addition, Simchowitz (2018) shows that any randomized algorithm that accesses an *deterministic* first order oracle at most T times will have error at least $\frac{cHB^2}{T^2}$ in the worst case for an H-smooth, convex quadratic objective, for some universal constant c. Therefore, the first term of local-AC-SA's guarantee cannot be improved. The second term of the guarantee also cannot be improved (Nemirovsky and Yudin, 1983)—in fact, this term cannot be improved even by an algorithm which is allowed to make MKR *sequential* calls to a stochastic gradient oracle. □

C. Proof of Theorem 2

Before we prove Theorem 2, we will introduce some notation. Recall that the objective is of the form $F(x) := \mathbb{E}_{z\sim\mathcal{D}}[f(x;z)]$. Let η_t denote the stepsize used for the tth overall iteration. Let x_t^m denote the tth iterate on the mth machine, and let $\bar{x}_t=\frac{1}{M}\sum_{m=1}^{M}x_t^m$ denote the averaged tth iterate. The vector $\bar{x}_t$ may not actually be computed by the algorithm, but it will be central to our analysis. We will use $\nabla f(x_t^m;z_t^m)$ to denote the stochastic gradient computed at x_t^m by the mth machine at iteration t, and $g_t=\frac{1}{M}\sum_{m=1}^{M}\nabla f(x_t^m;z_t^m)$ will denote the average of the stochastic gradients computed at time t. Finally, let $\bar{g}_t=\frac{1}{M}\sum_{m=1}^{M}\nabla F(x_t^m)$ denote the average of the full gradients computed at the individual iterates.

Lemma 1 (See Lemma 3.1 (Stich, 2018))**.** *Let F be H-smooth and λ-strongly convex, let $\sup_x \mathbb{E}\|\nabla f(x;z)-\nabla F(x)\|^2\le\sigma^2$, and let $\eta_t\le\frac{1}{4H}$, then the iterates of local SGD satisfy*

$$\mathbb{E}[F(\bar{x}_t)-F^*]\le\left(\frac{2}{\eta_t}-2\lambda\right)\mathbb{E}\|\bar{x}_t-x^*\|^2-\frac{2}{\eta_t}\mathbb{E}\|\bar{x}_{t+1}-x^*\|^2+\frac{2\eta_t\sigma^2}{M}+\frac{4H}{M}\sum_{m=1}^{M}\mathbb{E}\|\bar{x}_t-x_t^m\|^2$$

Proof. This proof is nearly identical to the proof of Lemma 3.1 due to Stich (2018), and we claim no technical innovation here. We include it in order to be self-contained.

We begin by analyzing the distance of $\bar{x}_{t+1}$ from the optimum. Below, expectations are taken over the all of the random variables $\{z_t^m\}$ which determine the iterates $\{x_t^m\}$.

$$\mathbb{E}\|\bar{x}_{t+1} - x^*\|^2$$

$$= \mathbb{E}\|\bar{x}_t - \eta_t g_t - x^*\|^2 \tag{25}$$

$$= \mathbb{E}\|\bar{x}_t - x^*\| + \eta_t^2 \mathbb{E}\|\bar{g}_t\|^2 + \eta_t^2 \mathbb{E}\|g_t - \bar{g}_t\|^2 - 2\eta_t \mathbb{E}\langle \bar{x}_t - x^*, \bar{g}_t\rangle \tag{26}$$

$$\leq \mathbb{E}\|\bar{x}_t - x^*\| + \eta_t^2 \mathbb{E}\|\bar{g}_t\|^2 + \frac{\eta_t^2\sigma^2}{M} - \frac{2\eta_t}{M}\sum_{m=1}^M \mathbb{E}\langle \bar{x}_t - x^*, \nabla f(x_t^m; z_t^m)\rangle \tag{27}$$

$$= \mathbb{E}\|\bar{x}_t - x^*\| + \eta_t^2 \mathbb{E}\|\bar{g}_t\|^2 + \frac{\eta_t^2\sigma^2}{M} - \frac{2\eta_t}{M}\sum_{m=1}^M \left[\mathbb{E}\langle x_t^m - x^*, \nabla F(x_t^m)\rangle + \mathbb{E}\langle \bar{x}_t - x_t^m, \nabla F(x_t^m)\rangle\right] \tag{28}$$

For the second equality, we used that $\mathbb{E}[g_t - \bar{g}_t] = 0$; for the first inequality, we used that $\mathbb{E}\|g_t - \bar{g}_t\|^2 = \mathbb{E}\left\|\frac{1}{M}\sum_{m=1}^M \nabla f(x_t^m; z_t^m) - \nabla F(x_t^m)\right\|^2 \leq \frac{\sigma^2}{M}$ since the individual stochastic gradient estimates are independent; and for the final equality, we used that z_t^m is independent of $\bar{x}_t$.

For any vectors v_m, $\left\|\sum_{m=1}^M v_m\right\|^2 \leq M\sum_{m=1}^M \|v_m\|^2$. In addition, for any point x and H-smooth F, $\|\nabla F(x)\|^2 \leq 2H(F(x) - F(x^*))$, thus

$$\eta_t^2 \mathbb{E}\|\bar{g}_t\|^2 \leq \eta_t^2 M \sum_{m=1}^M \left\|\frac{1}{M}\nabla F(x_t^m)\right\|^2 \leq \frac{2H\eta_t^2}{M}\sum_{m=1}^M F(x_t^m) - F(x^*) \tag{29}$$

By the λ-strong convexity of F, we have that

$$-\frac{2\eta_t}{M}\sum_{m=1}^M \langle x_t^m - x^*, \nabla F(x_t^m)\rangle \leq -\frac{2\eta_t}{M}\sum_{m=1}^M \left[F(x_t^m) - F(x^*) + \frac{\lambda}{2}\|x_t^m - x^*\|^2\right]$$

$$\leq -\frac{2\eta_t}{M}\sum_{m=1}^M [F(x_t^m) - F(x^*)] - \lambda\eta_t \|\bar{x}_t - x^*\|^2 \tag{30}$$

Finally, using the fact that for any vectors a, b and any $\gamma > 0$, $2\langle a, b\rangle \leq \gamma\|a\|^2 + \gamma^{-1}\|b\|^2$ we have

$$-2\eta_t\langle \bar{x}_t - x_t^m, \nabla F(x_t^m)\rangle \leq \eta_t\gamma\|\bar{x}_t - x_t^m\|^2 + \frac{\eta_t}{\gamma}\|\nabla F(x_t^m)\|^2 \leq \eta_t\gamma\|\bar{x}_t - x_t^m\|^2 + \frac{2H\eta_t}{\gamma}[F(x_t^m) - F(x^*)] \tag{31}$$

Combining these with (28), we conclude that for $\gamma = 2H$

$$\mathbb{E}\|\bar{x}_{t+1} - x^*\|^2 \leq (1 - \lambda\eta_t)\mathbb{E}\|\bar{x}_t - x^*\| - \frac{2\eta_t(1 - H\eta_t)}{M}\sum_{m=1}^M \mathbb{E}[F(x_t^m) - F(x^*)] + \frac{\eta_t^2\sigma^2}{M}$$

$$+ \frac{\eta_t}{M}\sum_{m=1}^M \left[2H\mathbb{E}\|\bar{x}_t - x_t^m\|^2 + \mathbb{E}[F(x_t^m) - F(x^*)]\right] \tag{32}$$

$$= (1 - \lambda\eta_t)\mathbb{E}\|\bar{x}_t - x^*\| - \frac{\eta_t(1 - 2H\eta_t)}{M}\sum_{m=1}^M \mathbb{E}[F(x_t^m) - F(x^*)]$$

$$+ \frac{\eta_t^2\sigma^2}{M} + \frac{2H\eta_t}{M}\sum_{m=1}^M \mathbb{E}\|\bar{x}_t - x_t^m\|^2 \tag{33}$$

By the convexity of F and the fact that $\eta_t \leq \frac{1}{4H}$, this implies

$$\mathbb{E}\|\bar{x}_{t+1} - x^*\|^2 \leq (1 - \lambda\eta_t)\mathbb{E}\|\bar{x}_t - x^*\| - \frac{\eta_t}{2}\mathbb{E}[F(\bar{x}_t) - F(x^*)] + \frac{\eta_t^2\sigma^2}{M} + \frac{2H\eta_t}{M}\sum_{m=1}^{M}\mathbb{E}\|\bar{x}_t - x_t^m\|^2 \tag{34}$$

Rearranging completes the proof. □

We will proceed to bound the final term in Lemma 1 more tightly than was done by Stich (2018), which allows us to improve on their upper bound. To do so, we will use the following technical lemmas:

Lemma 2 (Co-Coercivity of the Gradient)**.** *For any H-smooth and convex F, and any x, and y*

$$\|\nabla F(x) - \nabla F(y)\|^2 \leq H\left\langle \nabla F(x) - \nabla F(y),\, x - y\right\rangle$$

and

$$\|\nabla F(x) - \nabla F(y)\|^2 \leq 2H(F(x) - F(y) - \langle \nabla F(y),\, x - y\rangle)$$

Proof. This proof follows closely from (Vandenberghe, 2019). Define the H-smooth, convex functions

$$F_x(z) = F(z) - \langle \nabla F(x),\, z\rangle \qquad \text{and} \qquad F_y(z) = F(z) - \langle \nabla F(y),\, z\rangle \tag{35}$$

By setting the gradients of these convex functions equal to zero, it is clear that x minimizes F_x and y minimizes F_y. For any H-smooth and convex F, for any z, $\|\nabla F(z)\|^2 \leq 2H(F(z) - \min_x F(x))$, therefore,

$$\begin{align} F(y) - F(x) - \langle \nabla F(x),\, y - x\rangle &= F_x(y) - F_x(x) \tag{36}\\ &\geq \frac{1}{2H}\|\nabla F_x(y)\|^2 \tag{37}\\ &= \frac{1}{2H}\|\nabla F(y) - \nabla F(x)\|^2 \tag{38}\end{align}$$

Similarly,

$$F(x) - F(y) - \langle \nabla F(y),\, x - y\rangle \geq \frac{1}{2H}\|\nabla F(y) - \nabla F(x)\|^2 \tag{39}$$

This is the second claim of the Lemma, and combining these last two inequalities proves the first claim. □

Lemma 3 (See Lemma 6 (Karimireddy et al., 2019))**.** *Let F be any H-smooth and λ-strongly convex function, and let $\eta \leq \frac{1}{H}$. Then for any x, y*

$$\|x - \eta\nabla F(x) - y + \eta\nabla F(y)\|^2 \leq (1 - \lambda\eta)\|x - y\|^2$$

Proof. This Lemma and its proof are essentially identical to (Karimireddy et al., 2019, Lemma 6), we include it here in order to keep our results self-contained, and we are more explicit about the steps used.

$$\begin{align} \|x - \eta\nabla F(x) - y + \eta\nabla F(y)\|^2 &= \|x - y\|^2 + \eta^2\|\nabla F(x) - \nabla F(y)\|^2 - 2\eta\left\langle \nabla F(x) - \nabla F(y),\, x - y\right\rangle \tag{40}\\ &\leq \|x - y\|^2 + \eta^2 H\left\langle \nabla F(x) - \nabla F(y),\, x - y\right\rangle - 2\eta\left\langle \nabla F(x) - \nabla F(y),\, x - y\right\rangle \tag{41}\end{align}$$

where the inequality follows from Lemma 2. Since $\eta H \leq 1$, we further conclude that

$$\|x - \eta\nabla F(x) - y + \eta\nabla F(y)\|^2 \leq \|x - y\|^2 - \eta\left\langle \nabla F(x) - \nabla F(y),\, x - y\right\rangle \tag{42}$$

Finally, by the λ-strong convexity of F

$$\begin{align} \langle \nabla F(x),\, x - y\rangle &\geq F(x) - F(y) + \frac{\lambda}{2}\|x - y\|^2 \tag{43}\\ -\langle \nabla F(y),\, x - y\rangle &\geq F(y) - F(x) + \frac{\lambda}{2}\|x - y\|^2 \tag{44}\end{align}$$

Combining these, we conclude

$$\begin{align} \|x - \eta\nabla F(x) - y + \eta\nabla F(y)\|^2 &\leq \|x - y\|^2 - \eta\left\langle \nabla F(x) - \nabla F(y),\, x - y\right\rangle \tag{45}\\ &\leq \|x - y\|^2 - \eta\lambda\|x - y\|^2 \tag{46}\end{align}$$

which completes the proof. □

Lemma 4. *For any t and $m \neq m'$*

$$\mathbb{E}\|x_t^m - \bar{x}_t\|^2 \leq \frac{M-1}{M}\mathbb{E}\left\|x_t^m - x_t^{m'}\right\|^2$$

Proof. First, we note that $x_t^1, \dots, x_t^M$ are identically distributed. Therefore,

$$\mathbb{E}\|x_t^m - \bar{x}_t\|^2 = \mathbb{E}\left\|x_t^m - \frac{1}{M}\sum_{m'=1}^{M} x_t^{m'}\right\|^2 \tag{47}$$

$$= \frac{1}{M^2}\mathbb{E}\left\|\frac{1}{M}\sum_{m'\neq m} x_t^m - x_t^{m'}\right\|^2 \tag{48}$$

$$= \frac{1}{M^2}\left[\sum_{m'\neq m}\mathbb{E}\left\|x_t^m - x_t^{m'}\right\|^2 + \sum_{m'\neq m, m''\neq m, m'\neq m''}\mathbb{E}\left\langle x_t^m - x_t^{m'},\, x_t^m - x_t^{m''}\right\rangle\right] \tag{49}$$

$$\leq \frac{1}{M^2}\left[(M-1)\mathbb{E}\left\|x_t^m - x_t^{m'}\right\|^2 + \sum_{m'\neq m, m''\neq m, m'\neq m''}\sqrt{\mathbb{E}\left\|x_t^m - x_t^{m'}\right\|^2\mathbb{E}\left\|x_t^m - x_t^{m''}\right\|^2}\right] \tag{50}$$

$$= \frac{1}{M^2}\left[(M-1)\mathbb{E}\left\|x_t^m - x_t^{m'}\right\|^2 + 2\binom{M-1}{2}\mathbb{E}\left\|x_t^m - x_t^{m'}\right\|^2\right] \tag{51}$$

$$= \frac{(M-1)^2}{M^2}\mathbb{E}\left\|x_t^m - x_t^{m'}\right\|^2 \tag{52}$$

$$\leq \frac{M-1}{M}\mathbb{E}\left\|x_t^m - x_t^{m'}\right\|^2 \tag{53}$$

□

Lemma 5. *Under the conditions of Lemma 1, with the additional condition that the sequence of stepsizes $\eta_1, \eta_2, \dots$ is non-increasing and $\eta_t \leq \frac{1}{H}$ for all t, for any t and any m*

$$\mathbb{E}\|x_t^m - \bar{x}_t\|^2 \leq \frac{2(M-1)(K-1)\eta_{t-K+1\wedge 0}^2\sigma^2}{M}$$

If $\eta_t = \frac{2}{\lambda(a+t+1)}$, then it further satisfies

$$\mathbb{E}\|x_t^m - \bar{x}_t\|^2 \leq \frac{2(M-1)(K-1)\eta_{t-1}^2\sigma^2}{M}$$

Proof. By Lemma 4, we can upper bound

$$\mathbb{E}\|x_t^m - \bar{x}_t\|^2 \leq \frac{M-1}{M}\mathbb{E}\left\|x_t^m - x_t^{m'}\right\|^2 \tag{54}$$

for all t and $m \neq m'$. In addition,

$$\mathbb{E}\left\|x_t^m - x_t^{m'}\right\|^2 = \mathbb{E}\left\|x_{t-1}^m - \eta_{t-1}\nabla f(x_{t-1}^m; z_{t-1}^m) - x_{t-1}^{m'} + \eta_{t-1}\nabla f(x_{t-1}^{m'}; z_{t-1}^{m'})\right\|^2 \tag{55}$$

$$\leq \mathbb{E}\left\|x_{t-1}^m - \eta_{t-1}\nabla F(x_{t-1}^m) - x_{t-1}^{m'} + \eta_{t-1}\nabla F(x_{t-1}^{m'})\right\|^2 + 2\eta_{t-1}^2\sigma^2 \tag{56}$$

$$\leq (1-\lambda\eta_{t-1})\mathbb{E}\left\|x_{t-1}^m - x_{t-1}^{m'}\right\|^2 + 2\eta_{t-1}^2\sigma^2 \tag{57}$$

where for the final inequality we used Lemma 3 and the fact that the stepsizes are less than $\frac{1}{H}$,. Since the iterates are averaged every K iterations, for each t, there must be a t_0 with $0 \leq t - t_0 \leq K-1$ such that $x_{t_0}^m = x_{t_0}^{m'}$. Therefore, we can unroll the recurrence above to conclude that

$$\mathbb{E}\left\|x_t^m - x_t^{m'}\right\|^2 \leq \sum_{i=t_0}^{t-1} 2\eta_i^2\sigma^2 \prod_{j=i+1}^{t-1}(1-\lambda\eta_j) \leq 2\sigma^2\sum_{i=t_0}^{t-1}\eta_i^2 \tag{58}$$

where we define $\sum_{i=a}^{b} c_i = 0$ and $\prod_{i=a}^{b} c_i = 1$ for all $a > b$ and all $\{c_i\}_{i \in \mathbb{N}}$. Therefore, for any non-increasing stepsizes, we conclude

$$\mathbb{E}\|x_t^m - \bar{x}_t\|^2 \leq \frac{2\eta_{t-K+1\wedge 0}^2 \sigma^2 (M-1)(K-1)}{M} \tag{59}$$

This implies the first claim.

In the special case $\eta_t = \frac{2}{\lambda(a+t+1)}$, we have

$$\begin{aligned}
\mathbb{E}\left\|x_t^m - x^{m'}\right\|^2 &\leq 2\sigma^2 \sum_{i=t_0}^{t-1} \eta_i^2 \prod_{j=i+1}^{t-1} (1 - \lambda\eta_j) && (60)\\
&= 2\sigma^2 \sum_{i=t_0}^{t-1} \eta_i^2 \prod_{j=i+1}^{t-1} \left(\frac{a+j-1}{a+j+1}\right) && (61)\\
&= 2\sigma^2\eta_{t-1}^2 + \frac{2\sigma^2\eta_{t-2}^2(a+t-2)}{a+t} + 2\sigma^2 \sum_{i=t_0}^{t-3} \eta_i^2 \frac{(a+i)(a+i+1)}{(a+t-1)(a+t)} && (62)\\
&= 2\sigma^2\eta_{t-1}^2 \left(1 + \frac{(a+t)(a+t-2)}{(a+t-1)^2} + \sum_{i=t_0}^{t-3} \frac{(a+i)(a+t)}{(a+t-1)(a+i+1)}\right) && (63)\\
&\leq 2\sigma^2\eta_{t-1}^2 (t - t_0) && (64)\\
&\leq 2(K-1)\sigma^2\eta_{t-1}^2 && (65)
\end{aligned}$$

This implies the second claim. □

Next, we show that Local SGD is always at least as good as KR steps of sequential SGD. To do so, we use the following result from Stich (2019):

Lemma 6 (Lemma 3 (Stich, 2019))**.** *For any recurrence of the form*

$$r_{t+1} \leq (1 - a\gamma_t) r_t - b\gamma_t s_t + c\gamma_t^2$$

with $a, b > 0$, there exists a sequence $0 < \gamma_t \leq \frac{1}{d}$ and weights $w_t > 0$ such that

$$\frac{b}{W_T} \sum_{t=0}^{T} [s_t w_t + a r_{t+1}] \leq 32 d r_0 \exp\left(-\frac{aT}{2d}\right) + \frac{36c}{aT}$$

where $W_T := \sum_{t=0}^{T} w_t$.

We now argue that Local SGD is never worse than KR steps of sequential SGD:

Lemma 7. *Let $(f, \mathcal{D}) \in \mathcal{F}(H, \lambda, B, \sigma^2)$. When $\lambda = 0$, an appropriate average of the iterates of Local SGD with an optimally tuned constant stepsize satisfies for a universal constant c*

$$F(\hat{x}) - F^* \leq c \cdot \frac{HB^2}{KR} + c \cdot \frac{\sigma B}{\sqrt{KR}}$$

In the case $\lambda > 0$, then an appropriate average of the iterates of Local SGD with decreasing stepsize $\eta_t \asymp (\lambda t)^{-1}$ satisfies for a universal constant c

$$F(\hat{x}) - F^* \leq c \cdot HB^2 \exp\left(-\frac{\lambda KR}{4H}\right) + c \cdot \frac{\sigma^2}{\lambda KR}$$

Proof. Define $T := KR$ and consider the $(t+1)$st iterate on some machine m, x_{t+1}^m. If $t + 1 \mod K \neq 0$, then

$x_{t+1}^m = x_t^m - \eta_t \nabla f(x_t^m; z_t^m)$. In this case, for $\eta_t \leq \frac{1}{2H}$

$$\begin{aligned}
\mathbb{E}\left\|x_{t+1}^m - x^*\right\|^2 &= \mathbb{E}\|x_t^m - \eta_t \nabla f(x_t^m; z_t^m) - x^*\|^2 &(66)\\
&= \mathbb{E}\|x_t^m - x^*\|^2 + \eta_t^2 \mathbb{E}\|\nabla f(x_t^m; z_t^m)\|^2 - 2\eta_t \mathbb{E}\left\langle \nabla f(x_t^m; z_t^m),\, x_t^m - x^*\right\rangle &(67)\\
&\leq \mathbb{E}\|x_t^m - x^*\|^2 + \eta_t^2\sigma^2 + \eta_t^2 \mathbb{E}\|\nabla F(x_t^m)\|^2 - 2\eta_t \mathbb{E}\left\langle \nabla F(x_t^m),\, x_t^m - x^*\right\rangle &(68)\\
&\leq \mathbb{E}\|x_t^m - x^*\|^2 + \eta_t^2\sigma^2 + 2H\eta_t^2 \mathbb{E}[F(x_t^m) - F^*] - 2\eta_t \mathbb{E}\left[F(x_t^m) - F^* + \frac{\lambda}{2}\|x_t^m - x^*\|^2\right] &(69)\\
&= (1 - \lambda\eta_t)\mathbb{E}\|x_t^m - x^*\|^2 + \eta_t^2\sigma^2 - 2\eta_t(1 - H\eta_t)\mathbb{E}[F(x_t^m) - F^*] &(70)\\
\implies \mathbb{E}[F(x_t^m) - F^*] &\leq \left(\frac{1}{\eta_t} - \lambda\right)\mathbb{E}\|x_t^m - x^*\|^2 - \frac{1}{\eta_t}\mathbb{E}\left\|x_{t+1}^m - x^*\right\|^2 + \eta_t\sigma^2 &(71)
\end{aligned}$$

Here, for the first inequality we used the variance bound on the stochastic gradients; for the second inequality we used the H-smoothness and λ-strong convexity of F; and for the final inequality we used that $H\eta_t \leq \frac{1}{2}$ and rearranged.

If, on the other hand, $t + 1 \mod K = 0$, then $x_{t+1}^m = \frac{1}{M}\sum_{m'=1}^M x_t^{m'} - \eta_t \nabla f(x_t^{m'}; z_t^{m'})$. Since the local iterates on the different machines are *identically distributed*,

$$\begin{aligned}
\mathbb{E}\left\|x_{t+1}^m - x^*\right\|^2 &= \mathbb{E}\left\|\frac{1}{M}\sum_{m'=1}^M x_t^{m'} - \eta_t \nabla f(x_t^{m'}; z_t^{m'}) - x^*\right\|^2 &(72)\\
&\leq \frac{1}{M}\sum_{m'=1}^M \mathbb{E}\left\|x_t^{m'} - \eta_t \nabla f(x_t^{m'}; z_t^{m'}) - x^*\right\|^2 &(73)\\
&= \mathbb{E}\|x_t^m - \eta_t \nabla f(x_t^m; z_t^m) - x^*\|^2 &(74)
\end{aligned}$$

Where for the first inequality we used Jensen's inequality, and for the final equality we used that the local iterates are identically distributed. From here, using the same computation as above, we conclude that in either case

$$\mathbb{E}[F(x_t^m) - F^*] \leq \left(\frac{1}{\eta_t} - \lambda\right)\mathbb{E}\|x_t^m - x^*\|^2 - \frac{1}{\eta_t}\mathbb{E}\left\|x_{t+1}^m - x^*\right\|^2 + \eta_t\sigma^2 \tag{75}$$

Weakly Convex Case $\lambda = 0$: Choose a constant learning rate $\eta_t = \eta = \min\left\{\frac{1}{2H}, \frac{B}{\sigma\sqrt{T}}\right\}$ and define the averaged iterate

$$\hat{x} = \frac{1}{MT}\sum_{m=1}^M \sum_{t=1}^T x_t^m \tag{76}$$

Then, by the convexity of F:

$$\begin{aligned}
\mathbb{E}F(\hat{x}) - F^* &\leq \frac{1}{MT}\sum_{m=1}^M \sum_{t=1}^T \mathbb{E}[F(x_t^m) - F^*] &(77)\\
&\leq \frac{1}{MT}\sum_{m=1}^M \sum_{t=1}^T \frac{1}{\eta}\mathbb{E}\|x_t^m - x^*\|^2 - \frac{1}{\eta}\mathbb{E}\left\|x_{t+1}^m - x^*\right\|^2 + \eta\sigma^2 &(78)\\
&= \frac{\|x_0 - x^*\|^2}{T\eta} + \eta\sigma^2 &(79)\\
&= \max\left\{\frac{2H\|x_0 - x^*\|^2}{T}, \frac{\sigma\|x_0 - x^*\|}{\sqrt{T}}\right\} + \frac{\sigma\|x_0 - x^*\|}{\sqrt{T}} &(80)\\
&\leq \frac{2H\|x_0 - x^*\|^2}{T} + \frac{2\sigma\|x_0 - x^*\|}{\sqrt{T}} &(81)
\end{aligned}$$

Strongly Convex Case $\lambda > 0$: Rearranging (75), we see that it has the same form as the recurrence analyzed in Lemma 6 with $r_t = \mathbb{E}\|x_t^m - x^*\|^2$, $s_t = \mathbb{E}[F(x_t^m) - F^*]$, $a = \lambda$, $c = \sigma^2$, and $\gamma_t = \eta_t$ with the requirement that $\eta_t \leq \frac{1}{2H}$, i.e. $d = 2H$. Consequently, by Lemma 6, we conclude that there is a sequence of stepsizes and weights w_t such that

$$\mathbb{E}\left[F\left(\frac{1}{M\sum_{t=0}^{KR} w_t}\sum_{m=1}^{M}\sum_{t=0}^{KR} w_t x_t^m\right) - F^*\right] \leq \frac{1}{M\sum_{t=0}^{KR} w_t}\sum_{m=1}^{M}\sum_{t=0}^{KR}\mathbb{E}[F(w_t x_t^m) - F^*] \tag{82}$$

$$\leq 64H\mathbb{E}\|x_0 - x^*\|^2 \exp\left(-\frac{\lambda KR}{4H}\right) + \frac{36\sigma^2}{\lambda KR} \tag{83}$$

The stepsizes and weights are chosen as follows: If $KR \leq \frac{2H}{\lambda}$, then $\eta_t = \frac{1}{2H}$ and $w_t = (1-\lambda\eta)^{-t-1}$. If $KR > \frac{2H}{\lambda}$ and $t < KR/2$, then $\eta_t = \frac{1}{2H}$ and $w_t = 0$. If $KR > \frac{2H}{\lambda}$ and $t \geq KR/2$, then $\eta_t = \frac{2}{4H + \lambda(t - KR/2)}$ and $w_t = (4H/\lambda + t - KR/2)^2$. This completes the proof. □

Finally, we prove our main analysis of Local SGD. Portions of the analysis of the strongly convex case follow closely the proof of (Stich, 2019, Lemma 3).

Theorem 2. *Let $(f, \mathcal{D}) \in \mathcal{F}(H, \lambda, B, \sigma^2)$. When $\lambda = 0$, an appropriate average of the iterates of Local SGD with an optimally tuned constant stepsize satisfies for a universal constant c*

$$\begin{aligned}&\mathbb{E}[F(\hat{x}) - F(x^*)]\\ &\quad\leq c \cdot \min\left\{\frac{HB^2}{KR} + \frac{\sigma B}{\sqrt{MKR}} + \frac{\left(H\sigma^2 B^4\right)^{\frac{1}{3}}}{K^{1/3}R^{2/3}},\right.\\ &\qquad\qquad\left.\frac{HB^2}{KR} + \frac{\sigma B}{\sqrt{KR}}\right\}\end{aligned}$$

If $\lambda > 0$, then an appropriate average of the iterates of Local SGD with decaying stepsizes satisfies for a universal constant c

$$\begin{aligned}&\mathbb{E}[F(\hat{x}) - F(x^*)]\\ &\quad\leq c \cdot \min\left\{HB^2\exp\left(-\frac{\lambda KR}{4H}\right) + \frac{\sigma^2}{\lambda MKR}\right.\\ &\qquad\qquad + \frac{H\sigma^2\log\left(9 + \frac{\lambda KR}{H}\right)}{\lambda^2 KR^2},\\ &\qquad\qquad\left. HB^2\exp\left(-\frac{\lambda KR}{4H}\right) + \frac{\sigma^2}{\lambda KR}\right\}.\end{aligned}$$

Proof. We will prove the first terms in the $\min$'s in Theorem in two parts, first for the convex case $\lambda = 0$, then for the strongly convex case $\lambda > 0$. Then, we conclude by invoking Lemma 7 showing that Local SGD is never worse than KR steps of SGD on a single machine, which corresponds to the second terms in the $\min$'s in the Theorem statement.

Convex Case $\lambda = 0$: By Lemma 1 and the first claim of Lemma 5, the mean iterate satisfies

$$\mathbb{E}[F(\bar{x}_t) - F^*] \leq \frac{2}{\eta_t}\mathbb{E}\|\bar{x}_t - x^*\|^2 - \frac{2}{\eta_t}\mathbb{E}\|\bar{x}_{t+1} - x^*\|^2 + \frac{2\eta_t\sigma^2}{M} + \frac{8H(M-1)(K-1)\eta_{t-K+2\wedge 0}^2\sigma^2}{M} \tag{84}$$

Consider a fixed stepsize $\eta_t = \eta$ which will be chosen later, and consider the average of the iterates

$$\hat{x} = \frac{1}{KR}\sum_{t=1}^{KR}\bar{x}_t \tag{85}$$

By the convexity of F,

$$\mathbb{E}[F(\hat{x}) - F^*] \leq \frac{1}{KR}\sum_{t=1}^{KR}\mathbb{E}[F(\bar{x}_t) - F^*] \tag{86}$$

$$\leq \frac{1}{KR}\sum_{t=1}^{KR}\left[\frac{2}{\eta}\mathbb{E}\|\bar{x}_t - x^*\|^2 - \frac{2}{\eta}\mathbb{E}\|\bar{x}_{t+1} - x^*\|^2 + \frac{2\eta\sigma^2}{M} + \frac{8H(M-1)(K-1)\eta^2\sigma^2}{M}\right] \tag{87}$$

$$\leq \frac{2B^2}{\eta KR} + \frac{2\eta\sigma^2}{M} + \frac{8H(M-1)(K-1)\eta^2\sigma^2}{M} \tag{88}$$

Choose as a stepsize

$$\eta = \begin{cases} \min\left\{\frac{1}{4H},\, \frac{B\sqrt{M}}{\sigma\sqrt{KR}}\right\} & K = 1 \text{ or } M = 1 \\ \min\left\{\frac{1}{4H},\, \frac{B\sqrt{M}}{\sigma\sqrt{KR}},\, \left(\frac{B^2}{H\sigma^2K^2R}\right)^{\frac{1}{3}}\right\} & \text{Otherwise} \end{cases} \tag{89}$$

Then,

$$\mathbb{E}[F(\hat{x}) - F^*] \leq \frac{2B^2}{\eta KR} + \frac{2\eta\sigma^2}{M} + \frac{8H(M-1)(K-1)\eta^2\sigma^2}{M} \tag{90}$$

$$\leq \max\left\{\frac{8HB^2}{KR},\, \frac{2\sigma B}{\sqrt{MKR}},\, \frac{2\left(H\sigma^2B^4\right)^{\frac{1}{3}}}{K^{1/3}R^{2/3}}\right\} + \frac{2\sigma B}{\sqrt{MKR}} + \frac{8\left(H\sigma^2B^4\right)^{\frac{1}{3}}}{K^{1/3}R^{2/3}} \tag{91}$$

$$\leq \frac{8HB^2}{KR} + \frac{4\sigma B}{\sqrt{MKR}} + \frac{10\left(H\sigma^2B^4\right)^{\frac{1}{3}}}{K^{1/3}R^{2/3}} \tag{92}$$

Strongly Convex Case $\lambda > 0$: For the strongly convex case, following Stich (2019)'s proof of Lemma 6, we choose stepsizes according to the following set of cases: If $KR \leq \frac{2H}{\lambda}$, then $\eta_t = \frac{1}{4H}$ and $w_t = (1-\lambda\eta)^{-t-1}$. If $KR > \frac{2H}{\lambda}$ and $t \leq KR/2$, then $\eta_t = \frac{1}{4H}$ and $w_t = 0$. If $KR > \frac{2H}{\lambda}$ and $t > KR/2$, then $\eta_t = \frac{2}{8H+\lambda(t-KR/2)}$ and $w_t = (8H/\lambda + t - KR/2)$. We note that in the second and third cases, the stepsize is either constant or equal to $\eta_t = \frac{2}{\lambda(a+t-KR/2)}$ (for $a = \frac{8H}{\lambda}$) within each individual round of communication.

By Lemma 1 and the first claim of Lemma 5, during the rounds of communication for which the stepsize is constant, we have the recurrence:

$$\mathbb{E}\|\bar{x}_{t+1} - x^*\|^2 \leq (1-\lambda\eta_t)\mathbb{E}\|\bar{x}_t - x^*\|^2 - \frac{\eta_t}{2}\mathbb{E}[F(\bar{x}_t) - F^*] + \frac{\eta_t^2\sigma^2}{M} + 4HK\eta_t^3\sigma^2 \tag{93}$$

On the other hand, during the rounds of communication in which the stepsize is decreasing, we have by Lemma 1 and the second claim of Lemma 5 that:

$$\mathbb{E}\|\bar{x}_{t+1} - x^*\|^2 \leq (1-\lambda\eta_t)\mathbb{E}\|\bar{x}_t - x^*\|^2 - \frac{\eta_t}{2}\mathbb{E}[F(\bar{x}_t) - F^*] + \frac{\eta_t^2\sigma^2}{M} + 4HK\eta_t\eta_{t-1}^2\sigma^2 \tag{94}$$

Furthermore, during the rounds (i.e. when $t > KR$) where the stepsize is decreasing,

$$\eta_{t-1}^2 = \eta_t^2\frac{(a+t-KR/2)^2}{(a-1+t-KR/2)^2} \leq 4\eta_t^2 \tag{95}$$

So, for every t we conclude

$$\mathbb{E}\|\bar{x}_{t+1} - x^*\|^2 \leq (1-\lambda\eta_t)\mathbb{E}\|\bar{x}_t - x^*\|^2 - \frac{\eta_t}{2}\mathbb{E}[F(\bar{x}_t) - F^*] + \frac{\eta_t^2\sigma^2}{M} + 16HK\eta_t^3\sigma^2 \tag{96}$$

First, suppose $KR > \frac{2H}{\lambda}$, and consider the steps during which $\eta_t = \frac{1}{4H}$:

$$
\begin{aligned}
\mathbb{E}\left\|\bar{x}_{KR/2} - x^*\right\|^2 &\le \left(1 - \frac{\lambda}{4H}\right)\mathbb{E}\|\bar{x}_t - x^*\|^2 - \frac{1}{8H}\mathbb{E}[F(\bar{x}_t) - F^*] + \frac{\sigma^2}{16H^2M} + \frac{K\sigma^2}{4H^2} && (97)\\
&\le \left(1 - \frac{\lambda}{4H}\right)\mathbb{E}\|\bar{x}_t - x^*\|^2 + \frac{\sigma^2}{16H^2M} + \frac{K\sigma^2}{4H^2} && (98)\\
&\le \left(1 - \frac{\lambda}{4H}\right)^{KR/2}\mathbb{E}\|\bar{x}_0 - x^*\|^2 + \left(\frac{\sigma^2}{16H^2M} + \frac{K\sigma^2}{4H^2}\right)\sum_{t=0}^{KR/2-1}\left(1 - \frac{\lambda}{4H}\right)^t && (99)\\
&\le \left(1 - \frac{\lambda}{4H}\right)^{KR/2}\mathbb{E}\|\bar{x}_0 - x^*\|^2 + \frac{4H}{\lambda}\left(\frac{\sigma^2}{16H^2M} + \frac{K\sigma^2}{4H^2}\right) && (100)\\
&\le \mathbb{E}\|\bar{x}_0 - x^*\|^2\exp\left(-\frac{\lambda KR}{8H}\right) + \frac{\sigma^2}{4H\lambda M} + \frac{K\sigma^2}{H\lambda} && (101)
\end{aligned}
$$

Now, consider the remaining steps. Rearranging, we have

$$
\mathbb{E}[F(\bar{x}_t) - F^*] \le \left(\frac{2}{\eta_t} - \frac{\lambda}{2}\right)\mathbb{E}\|\bar{x}_t - x^*\|^2 - \frac{2}{\eta_t}\mathbb{E}\|\bar{x}_{t+1} - x^*\|^2 + \frac{2\eta_t\sigma^2}{M} + 32HK\eta_t^2\sigma^2 \qquad (102)
$$

So, since $\eta_t = \frac{2}{\lambda(a+t)}$ where $a = \frac{8H}{\lambda} - \frac{KR}{2}$ and $w_t = (a+t)$, we have

$$
\begin{aligned}
&\frac{1}{W_T}\sum_{t=KR/2}^{KR} w_t\mathbb{E}[F(\bar{x}_t) - F^*]\\
&\le \frac{1}{W_T}\sum_{t=KR/2}^{KR} w_t\left[\left(\frac{2}{\eta_t} - 2\lambda\right)\mathbb{E}\|\bar{x}_t - x^*\|^2 - \frac{2}{\eta_t}\mathbb{E}\|\bar{x}_{t+1} - x^*\|^2 + \frac{2\eta_t\sigma^2}{M} + 32HK\eta_t^2\sigma^2\right] && (103)\\
&= \frac{1}{W_T}\sum_{t=KR/2}^{KR} \lambda(a+t)(a+t-2)\mathbb{E}\|\bar{x}_t - x^*\|^2 - \lambda(a+t)^2\mathbb{E}\|\bar{x}_{t+1} - x^*\|^2 + \frac{2\sigma^2}{\lambda M} + \frac{32HK\eta_t\sigma^2}{\lambda} && (104)\\
&\le \frac{1}{W_T}\sum_{t=KR/2}^{KR} \lambda(a+t-1)^2\mathbb{E}\|\bar{x}_t - x^*\|^2 - \lambda(a+t)^2\mathbb{E}\|\bar{x}_{t+1} - x^*\|^2 + \frac{2\sigma^2}{\lambda M} + \frac{32HK\eta_t\sigma^2}{\lambda} && (105)\\
&\le \frac{\lambda(a + KR/2 - 1)^2}{W_T}\mathbb{E}\left\|\bar{x}_{KR/2} - x^*\right\|^2 + \frac{2\sigma^2(KR/2)}{W_T\lambda M} + \frac{64HK\sigma^2}{W_T\lambda^2}\sum_{t=KR/2}^{KR}\frac{1}{a+t} && (106)\\
&= \frac{\lambda\left(\frac{8H}{\lambda} - 1\right)^2}{W_T}\mathbb{E}\left\|\bar{x}_{KR/2} - x^*\right\|^2 + \frac{2\sigma^2(KR/2)}{W_T\lambda M} + \frac{64HK\sigma^2}{W_T\lambda^2}\sum_{t'=1}^{KR/2}\frac{1}{\frac{8H}{\lambda} + t'} && (107)\\
&\le \frac{64H^2}{W_T\lambda}\mathbb{E}\left\|\bar{x}_{KR/2} - x^*\right\|^2 + \frac{2\sigma^2(KR/2)}{W_T\lambda M} + \frac{64HK\sigma^2}{W_T\lambda^2}\log\left(e + \frac{\lambda KR}{4H}\right) && (108)
\end{aligned}
$$

Finally, we recall (101), $KR > \frac{2H}{\lambda}$, and note that $W_T = \sum_{t=KR/2}^{KR} a + t \geq \frac{3K^2R^2}{8} + \frac{aKR}{2} = \frac{K^2R^2}{8} + \frac{4HKR}{\lambda} \geq \frac{8H^2}{\lambda^2}$ thus

$$\begin{aligned}
&\frac{1}{W_T}\sum_{t=KR/2}^{KR} w_t\mathbb{E}[F(\bar{x}_t) - F^*] \\
&\leq \frac{64H^2}{W_T\lambda}\left(\mathbb{E}\|\bar{x}_0 - x^*\|^2 \exp\left(-\frac{\lambda KR}{8H}\right) + \frac{\sigma^2}{4H\lambda M} + \frac{K\sigma^2}{H\lambda}\right) + \frac{2\sigma^2(KR/2)}{W_T\lambda M} + \frac{64HK\sigma^2}{W_T\lambda^2}\log\left(e + \frac{\lambda KR}{4H}\right) && (109)\\
&\leq \frac{64H^2}{W_T\lambda}\mathbb{E}\|\bar{x}_0 - x^*\|^2 \exp\left(-\frac{\lambda KR}{8H}\right) + \frac{16H\sigma^2}{\lambda^2 MW_T} + \frac{64HK\sigma^2}{\lambda^2 W_T} + \frac{8\sigma^2}{\lambda MKR} + \frac{512H\sigma^2}{\lambda^2K R^2}\log\left(e + \frac{\lambda KR}{4H}\right) && (110)\\
&\leq 8\lambda\mathbb{E}\|\bar{x}_0 - x^*\|^2 \exp\left(-\frac{\lambda KR}{8H}\right) + \frac{4\sigma^2}{\lambda MKR} + \frac{512H\sigma^2}{\lambda^2KR^2} + \frac{8\sigma^2}{\lambda MKR} + \frac{512H\sigma^2}{\lambda^2KR^2}\log\left(e + \frac{\lambda KR}{4H}\right) && (111)\\
&\leq 8\lambda\mathbb{E}\|\bar{x}_0 - x^*\|^2 \exp\left(-\frac{\lambda KR}{8H}\right) + \frac{12\sigma^2}{\lambda MKR} + \frac{512H\sigma^2}{\lambda^2KR^2}\log\left(9 + \frac{\lambda KR}{H}\right) && (112)
\end{aligned}$$

This concludes the proof for the case $KR > \frac{2H}{\lambda}$.

If $KR \leq \frac{2H}{\lambda}$, we use the constant stepsize $\eta_t = \eta$ and weights $w_t = (1-\lambda\eta)^{-t-1}$. Rearranging (93) therefore gives

$$\mathbb{E}[F(\bar{x}_t) - F^*] \leq \frac{2}{\eta}(1-\lambda\eta)\mathbb{E}\|\bar{x}_t - x^*\|^2 - \frac{2}{\eta}\mathbb{E}\|\bar{x}_{t+1} - x^*\|^2 + \frac{2\eta\sigma^2}{M} + 8HK\eta^2\sigma^2 \tag{113}$$

so

$$\begin{aligned}
&\frac{1}{W_T}\sum_{t=1}^{KR} w_t\mathbb{E}[F(\bar{x}_t) - F^*] \\
&\leq \frac{1}{W_T}\sum_{t=1}^{KR} w_t\left[\frac{2}{\eta}(1-\lambda\eta)\mathbb{E}\|\bar{x}_t - x^*\|^2 - \frac{2}{\eta}\mathbb{E}\|\bar{x}_{t+1} - x^*\|^2 + \frac{2\eta\sigma^2}{M} + 8HK\eta^2\sigma^2\right] && (114)\\
&= \frac{1}{W_T}\sum_{t=1}^{KR}\left[\frac{2}{\eta}(1-\lambda\eta)^{-t}\mathbb{E}\|\bar{x}_t - x^*\|^2 - \frac{2}{\eta}(1-\lambda\eta)^{-(t+1)}\mathbb{E}\|\bar{x}_{t+1} - x^*\|^2\right] + \frac{2\eta\sigma^2}{M} + 8HK\eta^2\sigma^2 && (115)\\
&\leq \frac{2\mathbb{E}\|\bar{x}_0 - x^*\|^2}{\eta W_T} + \frac{2\eta\sigma^2}{M} + 8HK\eta^2\sigma^2 && (116)
\end{aligned}$$

Finally, we note that $W_T \geq (1-\lambda\eta)^{-KR-1}$ so

$$\frac{1}{W_T}\sum_{t=1}^{KR} w_t\mathbb{E}[F(\bar{x}_t) - F^*] \leq \frac{2\mathbb{E}\|\bar{x}_0 - x^*\|^2}{\eta}\exp(-\lambda\eta(KR+1)) + \frac{2\eta\sigma^2}{M} + 8HK\eta^2\sigma^2 \tag{117}$$

We also observe that $2H \geq \lambda KR$ so with $\eta = \frac{1}{4H} \leq \frac{1}{2\lambda KR}$ we have

$$\frac{1}{W_T}\sum_{t=1}^{KR} w_t\mathbb{E}[F(\bar{x}_t) - F^*] \leq 8H\mathbb{E}\|\bar{x}_0 - x^*\|^2 \exp\left(-\frac{\lambda KR}{4H}\right) + \frac{\sigma^2}{\lambda MKR} + \frac{2H\sigma^2}{\lambda^2KR^2} \tag{118}$$

□

D. Proofs from Section 5

Here, we will prove the lower bound in Theorem 3. Recall the objective and stochastic gradient estimator for the hard instance are defined by

$$F(x) = \frac{\mu}{2}(x_1 - b)^2 + \frac{H}{2}(x_2 - b)^2 + \frac{L}{2}\left((x_3 - c)^2 + [x_3 - c]_+^2\right) \tag{119}$$

and

$$\nabla f(x;z) = \nabla F(x) + \begin{bmatrix} 0 \\ 0 \\ z \end{bmatrix} \qquad \text{where} \qquad \mathbb{P}[z = \sigma] = \mathbb{P}[z = -\sigma] = \frac{1}{2} \tag{120}$$

Due to the structure of the objective (119), which decomposes as a sum over three terms which each depend only on a single coordinate, the local-SGD dynamics on each coordinate of the optimization variable are independent of each other. For this reason, we are able to analyze local-SGD on each coordinate separately.

Define the $2L$-smooth and L-strongly convex function

$$g_L(x) = \frac{L}{2}x^2 + \frac{L}{2}[x]_+^2 \tag{121}$$

Define a stochastic gradient estimator for g_L via

$$g_L'(x,z) = g_L'(x) + z \tag{122}$$

for $z \sim \text{Uniform}(\pm\sigma)$. Observe that the third coordinate of local-SGD on F evolves exactly the same as local-SGD on the univariate function g_L. In the next three lemmas, we analyze the behavior of local-SGD on g_L:

Lemma 8. *Fix $L, \eta, \sigma > 0$ such that $L\eta \leq \frac{1}{2}$. Let x_0 denote a random initial point with $\mathbb{E}x_0 \leq 0$, and let $x_2 = x_0 - \eta g_L'(x_0, z_0) - \eta g_L'(x_0 - \eta g_L'(x_0, z_0), z_1)$ be the second iterate of stochastic gradient descent with fixed stepsize η intialized at x_0, and let $x_3 = x_2 - \eta g_L'(x_2, z_2)$ be the third iterate. Then*

$$\mathbb{E}x_2 \leq \begin{cases} \frac{-\eta\sigma}{48} & \mathbb{E}x_0 \leq \frac{-\eta\sigma}{48} \\ \frac{-\eta\sigma}{4} + (1 - L\eta)\big(\mathbb{E}x_0 + \frac{\eta\sigma}{4}\big) & \mathbb{E}x_0 \in \big(\frac{-\eta\sigma}{48}, 0\big] \end{cases}$$

$$\mathbb{E}x_3 \leq \begin{cases} \frac{-\eta\sigma}{48} & \mathbb{E}x_0 \leq \frac{-\eta\sigma}{48} \\ \frac{-\eta\sigma}{4} + (1 - L\eta)^2\big(\mathbb{E}x_0 + \frac{\eta\sigma}{4}\big) & \mathbb{E}x_0 \in \big(\frac{-\eta\sigma}{48}, 0\big] \end{cases}$$

Proof. Consider the 2nd iterate of SGD with fixed stepsize η:

$$\begin{aligned} x_2 &= x_1 - \eta g_L'(x_1, z_1) && (123) \\ &= (1 - L\eta)x_1 - L\eta[x_1]_+ - \eta z_1 && (124) \\ &= (1 - L\eta)(x_0 - \eta g_L'(x_0, z_0)) - L\eta[x_0 - \eta g_L'(x_0, z_0)]_+ - \eta z_1 && (125) \\ &= (1 - L\eta)^2 x_0 - L\eta(1 - L\eta)[x_0]_+ - L\eta\big[(1 - L\eta)x_0 - L\eta[x_0]_+ - \eta z_0\big]_+ - \eta(1 - \eta)z_0 - \eta z_1 && (126) \end{aligned}$$

Thus,

$$\mathbb{E}x_2 = (1 - L\eta)^2\mathbb{E}x_0 - L\eta(1 - L\eta)\mathbb{E}[x_0]_+ - L\eta\mathbb{E}\big[(1 - L\eta)x_0 - L\eta[x_0]_+ - \eta z_0\big]_+ \tag{127}$$

Define $y := (1 - L\eta)x_0 - L\eta[x_0]_+$, then

$$\begin{aligned} \mathbb{E}\big[(1 - L\eta)x_0 - L\eta[x_0]_+ - \eta z_0\big]_+ &= \mathbb{E}[y - \eta z_0]_+ && (128) \\ &= \frac{1}{2}\mathbb{E}[y - \eta\sigma]_+ + \frac{1}{2}\mathbb{E}[y + \eta\sigma]_+ && (129) \\ &= \mathbb{E}\begin{cases} y & y > \eta\sigma \\ \frac{y + \eta\sigma}{2} & |y| \leq \eta\sigma \\ 0 & y < -\eta\sigma \end{cases} && (130) \end{aligned}$$

The function

$$z \mapsto \begin{cases} z & z > \eta\sigma \\ \frac{z + \eta\sigma}{2} & |z| \leq \eta\sigma \\ 0 & z < -\eta\sigma \end{cases} \tag{131}$$

is convex, so by Jensen's inequality

$$\begin{aligned}
\mathbb{E}x_2 &= (1-L\eta)\mathbb{E}y - L\eta\mathbb{E}\begin{cases} y & y > \eta\sigma \\ \frac{y+\eta\sigma}{2} & |y| \leq \eta\sigma \\ 0 & y < -\eta\sigma \end{cases} && (132) \\
&\leq (1-L\eta)\mathbb{E}y - L\eta\begin{cases} \mathbb{E}y & \mathbb{E}y > \eta\sigma \\ \frac{\mathbb{E}y+\eta\sigma}{2} & |\mathbb{E}y| \leq \eta\sigma \\ 0 & \mathbb{E}y < -\eta\sigma \end{cases} && (133) \\
&= \begin{cases} (1-2L\eta)\mathbb{E}y & \mathbb{E}y > \eta\sigma \\ \left(1-\frac{3}{2}L\eta\right)\mathbb{E}y - \frac{L\eta^2\sigma}{2} & |\mathbb{E}y| \leq \eta\sigma \\ (1-L\eta)\mathbb{E}y & \mathbb{E}y < -\eta\sigma \end{cases} && (134) \\
&\leq \begin{cases} (1-2L\eta)\mathbb{E}y & \mathbb{E}y > \eta\sigma \\ \left(1-\frac{3}{2}L\eta\right)\mathbb{E}y - \frac{L\eta^2\sigma}{2} & |\mathbb{E}y| \leq \eta\sigma \\ \frac{-\eta\sigma}{2} & \mathbb{E}y < -\eta\sigma \end{cases} && (135)
\end{aligned}$$

where we used that $L\eta \leq \frac{1}{2}$ for the final inequality. Suppose $\mathbb{E}x_0 \leq \frac{-\eta\sigma}{48}$ which implies $\mathbb{E}y \leq \frac{-(1-L\eta)\eta\sigma}{48}$. Then we are in either the second or third case of (135). If we are in the third case then

$$\mathbb{E}x_2 \leq \frac{-\eta\sigma}{2} \leq \frac{-\eta\sigma}{48} \tag{136}$$

If we are in the second case, then

$$\begin{aligned}
\mathbb{E}x_2 &\leq \left(1-\frac{3}{2}L\eta\right)\mathbb{E}y - \frac{L\eta^2\sigma}{2} && (137) \\
&\leq \left(1-\frac{3}{2}L\eta\right)\frac{-(1-L\eta)\eta\sigma}{48} - \frac{L\eta^2\sigma}{2} && (138) \\
&= \frac{-\eta\sigma}{48} + \frac{3(1-L\eta)L\eta^2\sigma}{96} + \frac{L\eta^2\sigma}{48} - \frac{L\eta^2\sigma}{2} && (139) \\
&\leq \frac{-\eta\sigma}{48} && (140)
\end{aligned}$$

Either way, $\mathbb{E}x_2 \leq \frac{-\eta\sigma}{48}$.

Suppose instead that $\mathbb{E}x_0 \in \left(\frac{-\eta\sigma}{48}, 0\right]$. Then,

$$\begin{aligned}
\mathbb{E}x_2 &\leq \left(1-\frac{3}{2}L\eta\right)\mathbb{E}y - \frac{L\eta^2\sigma}{2} && (141) \\
&\leq (1-L\eta)\mathbb{E}x_0 - \frac{3L\eta(1-L\eta)}{2}\mathbb{E}x_0 - \frac{L\eta^2\sigma}{2} && (142) \\
&\leq (1-L\eta)\mathbb{E}x_0 + \frac{3L\eta}{2}\cdot\frac{\eta\sigma}{48} - \frac{L\eta^2\sigma}{2} && (143) \\
&\leq (1-L\eta)\mathbb{E}x_0 - \frac{L\eta^2\sigma}{4} && (144) \\
&= -\frac{\eta\sigma}{4} + (1-L\eta)\left(\mathbb{E}x_0 + \frac{\eta\sigma}{4}\right) && (145)
\end{aligned}$$

We conclude that

$$\mathbb{E}x_2 \leq \begin{cases} \frac{-\eta\sigma}{48} & \mathbb{E}x_0 \leq \frac{-\eta\sigma}{48} \\ \frac{-\eta\sigma}{4} + (1-L\eta)\left(\mathbb{E}x_0 + \frac{\eta\sigma}{4}\right) & \mathbb{E}x_0 \in \left(\frac{-\eta\sigma}{48}, 0\right] \end{cases} \tag{146}$$

Now, consider the third iterate of SGD, x_3:

$$\begin{aligned}
\mathbb{E}x_3 &= \mathbb{E}x_2 - \eta\mathbb{E}g_L'(x_2, z_2) \tag{147}\\
&= (1 - L\eta)\mathbb{E}x_2 - L\eta\mathbb{E}[x_2]_+ \tag{148}\\
&= (1 - L\eta)\mathbb{E}x_2 - L\eta\mathbb{E}[\mathbb{E}[x_2 \,|\, x_1] - \eta z_1]_+ \tag{149}\\
&\leq (1 - L\eta)\mathbb{E}x_2 - \frac{L\eta}{2}\mathbb{E}[\mathbb{E}[x_2 \,|\, x_1] + \eta\sigma]_+ \tag{150}
\end{aligned}$$

Since $z \mapsto [z]_+$ is convex, by Jensen's inequality

$$\begin{aligned}
\mathbb{E}x_3 &\leq (1 - L\eta)\mathbb{E}x_2 - \frac{L\eta}{2}[\mathbb{E}x_2 + \eta\sigma]_+ \tag{151}\\
&\leq \begin{cases} \left(1 - \frac{3L\eta}{2}\right)\mathbb{E}x_2 - \frac{L\eta^2\sigma}{2} & \mathbb{E}x_2 > -\eta\sigma \\ (1 - L\eta)\mathbb{E}x_2 & \mathbb{E}x_2 \leq -\eta\sigma \end{cases} \tag{152}\\
&\leq \begin{cases} \left(1 - \frac{3L\eta}{2}\right)\mathbb{E}x_2 - \frac{L\eta^2\sigma}{2} & \mathbb{E}x_2 > -\eta\sigma \\ \frac{-\eta\sigma}{2} & \mathbb{E}x_2 \leq -\eta\sigma \end{cases} \tag{153}
\end{aligned}$$

To complete the proof, we must show that

$$\mathbb{E}x_3 \leq \begin{cases} \frac{-\eta\sigma}{48} & \mathbb{E}x_0 \leq \frac{-\eta\sigma}{48} \\ \frac{-\eta\sigma}{4} + (1 - L\eta)^2\left(\mathbb{E}x_0 + \frac{\eta\sigma}{4}\right) & \mathbb{E}x_0 \in \left(\frac{-\eta\sigma}{48}, 0\right] \end{cases} \tag{154}$$

Returning to (153), note that if $\mathbb{E}x_2 \leq -\eta\sigma$ then $\mathbb{E}x_3 \leq \frac{-\eta\sigma}{2}$ implies (154). Therefore, we only need to consider the first case of (153).

Suppose first that $\mathbb{E}x_0 \leq \frac{-\eta\sigma}{48}$, then by (146) we have $\mathbb{E}x_2 \leq \frac{-\eta\sigma}{48}$, thus

$$\begin{aligned}
\mathbb{E}x_3 &\leq \left(1 - \frac{3L\eta}{2}\right)\mathbb{E}x_2 - \frac{L\eta^2\sigma}{2} \tag{155}\\
&\leq \left(1 - \frac{3L\eta}{2}\right)\frac{-\eta\sigma}{48} - \frac{L\eta^2\sigma}{2} \tag{156}\\
&\leq \frac{-\eta\sigma}{48} \tag{157}
\end{aligned}$$

If instead $\mathbb{E}x_0 \in \left(\frac{-\eta\sigma}{48}, 0\right]$, then by (146) we have $\mathbb{E}x_2 \leq \frac{-\eta\sigma}{4} + (1 - L\eta)\left(\mathbb{E}x_0 + \frac{\eta\sigma}{4}\right)$, thus

$$\begin{aligned}
\mathbb{E}x_3 &\leq \left(1 - \frac{3L\eta}{2}\right)\mathbb{E}x_2 - \frac{L\eta^2\sigma}{2} \tag{158}\\
&\leq \left(1 - \frac{3L\eta}{2}\right)\frac{-\eta\sigma}{4} + \left(1 - \frac{3L\eta}{2}\right)(1 - L\eta)\left(\mathbb{E}x_0 + \frac{\eta\sigma}{4}\right) - \frac{L\eta^2\sigma}{2} \tag{159}\\
&\leq \frac{-\eta\sigma}{4} + \frac{3L\eta^2\sigma}{8} - \frac{L\eta^2\sigma}{2} + (1 - L\eta)^2\left(\mathbb{E}x_0 + \frac{\eta\sigma}{4}\right) \tag{160}\\
&\leq \frac{-\eta\sigma}{4} + (1 - L\eta)^2\left(\mathbb{E}x_0 + \frac{\eta\sigma}{4}\right) \tag{161}
\end{aligned}$$

This completes both cases of (154). □

Lemma 9. *Fix $L, \eta, \sigma > 0$ such that $L\eta \leq \frac{1}{2}$ and let $k \geq 2$. Let x_0 denote a random initial point with $\mathbb{E}x_0 \leq 0$ and let x_k denote the kth iterate of stochastic gradient descent on g_L with fixed stepsize η intialized at x_0. Then*

$$\mathbb{E}x_k \leq \begin{cases} \frac{-\eta\sigma}{48} & \mathbb{E}x_0 \leq \frac{-\eta\sigma}{48} \\ \frac{-\eta\sigma}{4} + (1 - L\eta)^{k/2}\left(\mathbb{E}x_0 + \frac{\eta\sigma}{4}\right) & \mathbb{E}x_0 \in \left(\frac{-\eta\sigma}{48}, 0\right] \end{cases}$$

Proof. The idea of this proof is simple: k steps of SGD initialized at some point x_0 is equivalent to doing two steps of SGD initialized at x_0 to get x_2, then doing two more steps initialized at x_2 to get x_4, and so forth until k steps have been completed. The only minor complication is if k is odd, in which case we start by doing three steps initialized at x_0 to get x_3 and continue in steps of two.

We will consider two cases, either $\mathbb{E}x_0 \leq \frac{-\eta\sigma}{48}$ or $\mathbb{E}x_0 \in \left(\frac{-\eta\sigma}{48}, 0\right]$. In the first case, $\mathbb{E}x_0 \leq \frac{-\eta\sigma}{48}$, if k is even then by Lemma 8

$$\mathbb{E}x_0 \leq \frac{-\eta\sigma}{48} \implies \mathbb{E}x_2 \leq \frac{-\eta\sigma}{48} \implies \mathbb{E}x_4 \leq \frac{-\eta\sigma}{48} \implies \cdots \implies \mathbb{E}x_k \leq \frac{-\eta\sigma}{48} \tag{162}$$

If k is odd then

$$\mathbb{E}x_0 \leq \frac{-\eta\sigma}{48} \implies \mathbb{E}x_3 \leq \frac{-\eta\sigma}{48} \implies \mathbb{E}x_5 \leq \frac{-\eta\sigma}{48} \implies \cdots \implies \mathbb{E}x_k \leq \frac{-\eta\sigma}{48} \tag{163}$$

In the second case, $\mathbb{E}x_0 \in \left(\frac{-\eta\sigma}{48}, 0\right]$. Then, when k is even, by repeatedly invoking Lemma 8 we get

$$\mathbb{E}x_2 \leq \frac{-\eta\sigma}{4} + (1 - L\eta)\left(\mathbb{E}x_0 + \frac{\eta\sigma}{4}\right) \tag{164}$$

$$\mathbb{E}x_4 \leq \frac{-\eta\sigma}{4} + (1 - L\eta)\left(\mathbb{E}x_2 + \frac{\eta\sigma}{4}\right) \leq \frac{-\eta\sigma}{4} + (1 - L\eta)^2\left(\mathbb{E}x_0 + \frac{\eta\sigma}{4}\right) \tag{165}$$

$$\mathbb{E}x_6 \leq \frac{-\eta\sigma}{4} + (1 - L\eta)\left(\mathbb{E}x_4 + \frac{\eta\sigma}{4}\right) \leq \frac{-\eta\sigma}{4} + (1 - L\eta)^3\left(\mathbb{E}x_0 + \frac{\eta\sigma}{4}\right) \tag{166}$$

$$\vdots \tag{167}$$

$$\mathbb{E}x_k \leq \frac{-\eta\sigma}{4} + (1 - L\eta)^{k/2}\left(\mathbb{E}x_0 + \frac{\eta\sigma}{4}\right) \tag{168}$$

The same argument applies when k is odd (using the bound on $\mathbb{E}x_3$) to prove

$$\mathbb{E}x_k \leq \frac{-\eta\sigma}{4} + (1 - L\eta)^{(k+1)/2}\left(\mathbb{E}x_0 + \frac{\eta\sigma}{4}\right) \leq \frac{-\eta\sigma}{4} + (1 - L\eta)^{k/2}\left(\mathbb{E}x_0 + \frac{\eta\sigma}{4}\right) \qquad \square \tag{169}$$

Lemma 10. *Let $K \geq 2$ and let $\hat{x}$ be the output of local-SGD(K, R, M) on F using a fixed stepsize $\eta \leq \frac{1}{2L}$ and initialized at zero. Then*

$$\mathbb{E}\left[\frac{L}{2}\left((\hat{x}_3 - c)^2 + [\hat{x}_3 - c]_+^2\right)\right] \geq \frac{L\eta^2\sigma^2}{4608}\mathbb{1}_{\left\{\eta\leq\frac{1}{2L}\right\}}\mathbb{1}_{\left\{c\geq\frac{\eta\sigma}{48}\vee\eta\geq\frac{2}{LRK}\right\}}$$

Proof. Since each coordinate evolves independently when optimizing F using local-SGD, we can ignore the first two coordinates and focus only on the third. Observe that using local-SGD(K, R, M) on F with a fixed stepsize η and initialized at zero to obtain $\hat{x}_3$ is exactly equivalent to using local-SGD(K, R, M) on g_L with the same fixed stepsize η and initialized at $-c$. The different initialization is due to the fact that the local-SGD dynamics do not change with the change of variables $x - c \to x$. Let $\bar{x}_r$ denote the averaged iterate of local-SGD(K, R, M) initialized at $-c$ with stepsize η after the rth round of communication and let $x_{r,k,m}$ denote its kth iterate during the rth round of communication on the mth machine. We will start by proving that when $\eta \leq \frac{1}{2L}$ and *either* $c \geq \frac{\eta\sigma}{8}$ *or* $\eta \geq \frac{2}{LRK}$ then

$$\mathbb{E}\hat{x}_3 - c = \mathbb{E}\bar{x}_R \leq \frac{-\eta\sigma}{48} \tag{170}$$

Consider first the case $\mathbb{E}x_0 = -c \leq \frac{-\eta\sigma}{48}$. Then by Lemma 9

$$\mathbb{E}x_0 = -c \leq \frac{-\eta\sigma}{48} \implies \mathbb{E}x_{1,K,m} \leq \frac{-\eta\sigma}{48} \quad \forall m \tag{171}$$

therefore

$$\mathbb{E}\bar{x}_1 = \mathbb{E}\left[\frac{1}{M}\sum_{m=1}^{M} x_{1,K,m}\right] \leq \frac{-\eta\sigma}{48} \tag{172}$$

Repeatedly applying Lemma 9 shows that for each r

$$\mathbb{E}\bar{x}_r \leq \frac{-\eta\sigma}{48} \implies \mathbb{E}x_{r+1,K,m} \leq \frac{-\eta\sigma}{48} \implies \mathbb{E}\bar{x}_{r+1} = \mathbb{E}\left[\frac{1}{M}\sum_{m=1}^{M} x_{r+1,K,m}\right] \leq \frac{-\eta\sigma}{48} \tag{173}$$

We conclude $\mathbb{E}\bar{x}_R \leq \frac{-\eta\sigma}{48}$.

Consider instead the case that $\mathbb{E}x_0 = -c \in \left(\frac{-\eta\sigma}{48}, 0\right]$ and $\eta \geq \frac{2}{LRK}$. Then, by Lemma 9

$$\mathbb{E}x_0 = -c \in \left(\frac{-\eta\sigma}{48}, 0\right] \implies \mathbb{E}x_{1,K,m} \leq \frac{-\eta\sigma}{4} + (1-L\eta)^{K/2}\left(\frac{\eta\sigma}{4} - c\right) \ \forall m \tag{174}$$

and so

$$\mathbb{E}\bar{x}_1 = \mathbb{E}\left[\frac{1}{M}\sum_{m=1}^{M} x_{1,K,m}\right] \leq \frac{-\eta\sigma}{4} + (1-L\eta)^{K/2}\left(\frac{\eta\sigma}{4} - c\right) \tag{175}$$

Again, we can repeatedly apply Lemma 9 to show

$$\mathbb{E}\bar{x}_2 \leq \frac{-\eta\sigma}{4} + (1-L\eta)^{K/2}\left(\mathbb{E}\bar{x}_1 + \frac{\eta\sigma}{4}\right) \leq \frac{-\eta\sigma}{4} + (1-L\eta)^{2K/2}\left(\frac{\eta\sigma}{4} - c\right) \tag{176}$$

$$\mathbb{E}\bar{x}_3 \leq \frac{-\eta\sigma}{4} + (1-L\eta)^{K/2}\left(\mathbb{E}\bar{x}_2 + \frac{\eta\sigma}{4}\right) \leq \frac{-\eta\sigma}{4} + (1-L\eta)^{3K/2}\left(\frac{\eta\sigma}{4} - c\right) \tag{177}$$

$$\vdots \tag{178}$$

$$\mathbb{E}\bar{x}_R \leq \frac{-\eta\sigma}{4} + (1-L\eta)^{RK/2}\left(\frac{\eta\sigma}{4} - c\right) \tag{179}$$

$$\leq -\left(1 - (1-L\eta)^{RK/2}\right)\frac{\eta\sigma}{4} \tag{180}$$

$$\leq -\left(1 - \left(1 - \frac{2}{RK}\right)^{RK/2}\right)\frac{\eta\sigma}{4} \tag{181}$$

$$\leq \frac{\eta\sigma}{48} \tag{182}$$

These inequalities hold only as long as $\mathbb{E}\bar{x}_r > \frac{-\eta\sigma}{48}$. But, if for some r, $\mathbb{E}\bar{x}_r \leq \frac{-\eta\sigma}{48}$ then $\mathbb{E}\bar{x}_R \leq \frac{-\eta\sigma}{48}$ by the same argument as above. We conclude that

$$\mathbb{E}\bar{x}_R \leq \frac{-\eta\sigma}{48}\mathbb{1}_{\left\{\eta \leq \frac{1}{2L}\right\}}\mathbb{1}_{\left\{c \geq \frac{\eta\sigma}{48} \vee \eta \geq \frac{2}{LRK}\right\}} \tag{183}$$

Since $\mathbb{E}\hat{x}_3 - c = \mathbb{E}\bar{x}_R$, by Jensen's inequality

$$\mathbb{E}\left[\frac{L}{2}\left((\hat{x}_3 - c)^2 + [\hat{x}_3 - c]_+^2\right)\right] \geq \frac{L}{2}\left((\mathbb{E}\bar{x}_R)^2 + [\mathbb{E}\bar{x}_R]_+^2\right) \tag{184}$$

$$\geq \frac{L\eta^2\sigma^2}{4608}\mathbb{1}_{\left\{\eta \leq \frac{1}{2L}\right\}}\mathbb{1}_{\left\{c \geq \frac{\eta\sigma}{48} \vee \eta \geq \frac{2}{LRK}\right\}} \tag{185}$$

□

We now analyze the progress of SGD on the first two coordinates of F in the following lemma:

Lemma 11. *Let $\hat{x}$ be the output of local-SGD(K, R, M) on F using a fixed stepsize η and initialized at zero. Then with probability 1,*

$$\frac{\mu}{2}(\hat{x}_1 - b)^2 \geq \frac{\mu b^2}{8}\mathbb{1}_{\left\{\eta < \frac{1}{2\mu KR}\right\}}$$

and

$$\frac{H}{2}(\hat{x}_2 - b)^2 \geq \frac{Hb^2}{2}\mathbb{1}_{\left\{\eta > \frac{2}{H}\right\}}.$$

Proof. Since the stochastic gradient estimator has no noise along the first and second coordinates, and since the separate coordinates evolve independently, $\hat{x}_1$ is exactly the output of KR steps of deterministic gradient descent with fixed stepsize η on the univariate function $x \mapsto \frac{\mu}{2}(x-b)^2$. Similarly, $\hat{x}_2$ is the output of KR steps of deterministic gradient descent with fixed stepsize η on $x \mapsto \frac{H}{2}(x-b)^2$. Thus,

$$x_1^{(t+1)} - b \;=\; x_1^{(t)} - b - \eta\mu\Big(x_1^{(t)} - b\Big) \quad\Longrightarrow\quad \hat{x}_1 \;=\; b + (1-\eta\mu)^{KR}\Big(x_1^{(0)} - b\Big) \;=\; b\Big(1 - (1-\eta\mu)^{KR}\Big) \tag{186}$$

Thus, if $\eta < \frac{1}{2\mu KR}$, then

$$\hat{x}_1 \leq b\eta\mu KR < \frac{b}{2} \implies \frac{\mu}{2}(\hat{x}_1 - b)^2 \geq \frac{\mu b^2}{8}\mathbb{1}_{\{\eta < \frac{1}{2\mu KR}\}} \tag{187}$$

Similarly,

$$x_2^{(t+1)} - b \;=\; x_2^{(t)} - b - \eta H\Big(x_2^{(t)} - b\Big) \quad\Longrightarrow\quad \hat{x}_2 - b \;=\; (1-\eta H)^{KR}\Big(x_2^{(0)} - b\Big) \;=\; -b(1-\eta H)^{KR} \tag{188}$$

Thus, if $\eta > \frac{2}{H}$, then

$$|\hat{x}_2 - b| \geq b \implies \frac{H}{2}(\hat{x}_2 - b)^2 \geq \frac{Hb^2}{2}\mathbb{1}_{\{\eta > \frac{2}{H}\}} \tag{189}$$

□

Combining Lemmas 10 and 11, we are ready to prove the theorem:

Theorem 3. *For $0 \leq \lambda \leq \frac{H}{16}$, there exists $(f,\mathcal{D}) \in \mathcal{F}(H,\lambda,B,\sigma^2)$ such that for any $K \geq 2$ and $M, R \geq 1$, local SGD initialized at 0 with any fixed stepsize, will output a point $\hat{x}$ such that for a universal constant c*

$$\begin{aligned}
&\mathbb{E}F(\hat{x}) - \min_x F(x) \\
&\geq c \cdot \min\left\{\frac{H^{1/3}\sigma^{2/3}B^{4/3}}{K^{2/3}R^{2/3}}, \frac{H\sigma^2}{\lambda^2K^2R^2}, HB^2\right\} \\
&\quad + c \cdot \min\left\{\frac{\sigma B}{\sqrt{MKR}}, \frac{\sigma^2}{\lambda MKR}\right\}.
\end{aligned}$$

Proof. Consider optimizing the objective F defined in (119) using the stochastic gradient oracle (120) initialized at zero and using a fixed stepsize η. The variance of the stochastic gradient oracle is equal to σ^2. This function is $\max\{\mu, H, 2L\}$-smooth, and $\min\{\mu, H, L\}$-strongly convex. We will be choosing $L = \frac{H}{4}$ and $\mu \in \big[\lambda, \frac{H}{16}\big]$ so that F is H-smooth and λ-strongly convex. Finally, the objective F is minimized at the point $x^* = [b, b, c]^\top$ and $F(x^*) = 0$. This point has norm $\|x^*\| = \sqrt{2b^2 + c^2}$ we will choose $b = c = \frac{B}{\sqrt{3}}$ so that $\|x^*\| = B$.

By Lemma 10, the output of local-SGD(K,R,M), $\hat{x}$ satisfies

$$\mathbb{E}\left[\frac{L}{2}\Big((\hat{x}_3 - c)^2 + [\hat{x}_3 - c]_+^2\Big)\right] \geq \frac{L\eta^2\sigma^2}{4608}\mathbb{1}_{\{\eta \leq \frac{1}{2L}\}}\mathbb{1}_{\{c \geq \frac{\eta\sigma}{48} \vee \eta \geq \frac{2}{LRK}\}} \tag{190}$$

By Lemma 11, the output of local-SGD(K,R,M), $\hat{x}$ satisfies

$$\frac{\mu}{2}(\hat{x}_1 - b)^2 + \frac{H}{2}(\hat{x}_2 - b)^2 \geq \frac{\mu b^2}{8}\mathbb{1}_{\{\eta < \frac{1}{2\mu KR}\}} + \frac{Hb^2}{2}\mathbb{1}_{\{\eta > \frac{2}{H}\}} \tag{191}$$

Combining these, we have

$$\mathbb{E}F(\hat{x}) \quad-\quad \min_x F(x) \quad\geq\quad \frac{\mu b^2}{8}\mathbb{1}_{\{\eta < \frac{1}{2\mu KR}\}} \quad+\quad \frac{Hb^2}{2}\mathbb{1}_{\{\eta > \frac{2}{H}\}} \quad+\quad \frac{L\eta^2\sigma^2}{4608}\mathbb{1}_{\{\eta \leq \frac{1}{2L}\}}\mathbb{1}_{\{\eta \leq \frac{48c}{\sigma} \vee \eta \geq \frac{2}{LRK}\}} \tag{192}$$

Consider two cases: first, suppose that $\eta \notin \Big[\frac{1}{2\mu KR}, \frac{2}{H}\Big]$. Then,

$$\mathbb{E}F(\hat{x}) - \min_x F(x) \geq \min\left\{\frac{\mu b^2}{8}, \frac{Hb^2}{2}\right\} = \frac{\mu b^2}{8} \tag{193}$$

Suppose instead that $\eta \in \left[\frac{1}{2\mu KR}, \frac{2}{H}\right]$. Since $L = \frac{H}{4}$, $\eta \leq \frac{2}{H} \leq \frac{1}{2L}$. Similarly, since $\mu \leq \frac{H}{16} = \frac{L}{4}$, $\eta \geq \frac{1}{2\mu KR} \geq \frac{2}{LRK}$. Therefore, $\eta \in \left[\frac{1}{2\mu KR}, \frac{2}{H}\right]$ implies

$$\mathbb{E}F(\hat{x}) - \min_x F(x) \geq \min_{\eta \in \left[\frac{1}{2\mu KR}, \frac{2}{H}\right]} \frac{L\eta^2\sigma^2}{4608} \mathbb{1}_{\left\{\eta \leq \frac{1}{2L}\right\}} \mathbb{1}_{\left\{\eta \leq \frac{48c}{\sigma} \vee \eta \geq \frac{2}{LRK}\right\}} \tag{194}$$

$$= \min_{\eta \in \left[\frac{1}{2\mu KR}, \frac{2}{H}\right]} \frac{L\eta^2\sigma^2}{4608} \tag{195}$$

$$= \frac{L\sigma^2}{18432\mu^2K^2R^2} \tag{196}$$

Combining (193) and (196) yields

$$\mathbb{E}F(\hat{x}) - \min_x F(x) \geq \min\left\{\frac{\mu B^2}{24}, \frac{H\sigma^2}{73728\mu^2K^2R^2}\right\} \tag{197}$$

This statement holds for any $\mu \in \left[\lambda, \frac{H}{16}\right]$. Consider three cases: first, suppose $\mu = \left(\frac{H\sigma^2}{3072B^2K^2R^2}\right)^{1/3} \in \left[\lambda, \frac{H}{16}\right]$. Then

$$\mathbb{E}F(\hat{x}) - \min_x F(x) \geq \frac{H^{1/3}\sigma^{2/3}B^{4/3}}{350K^{2/3}R^{2/3}} \tag{198}$$

Consider next the case that $\left(\frac{H\sigma^2}{3072B^2K^2R^2}\right)^{1/3} > \frac{H}{16} \implies \frac{\sigma^2}{192B^2K^2R^2} > \frac{H^2}{256}$ and choose $\mu = \frac{H}{16}$. Then

$$\mathbb{E}F(\hat{x}) - \min_x F(x) \geq \min\left\{\frac{HB^2}{384}, \frac{H\sigma^2}{73728K^2R^2 \cdot \frac{H^2}{256}}\right\} = \frac{HB^2}{384} \tag{199}$$

Finally, consider the case that $\left(\frac{H\sigma^2}{3072B^2K^2R^2}\right)^{1/3} < \lambda$ and choose $\mu = \lambda$. Then,

$$\mathbb{E}F(\hat{x}) - \min_x F(x) \geq \min\left\{\frac{\lambda B^2}{24}, \frac{H\sigma^2}{73728\lambda^2K^2R^2}\right\} = \frac{H\sigma^2}{73728\lambda^2K^2R^2} \tag{200}$$

Combining these cases completes the proof. □