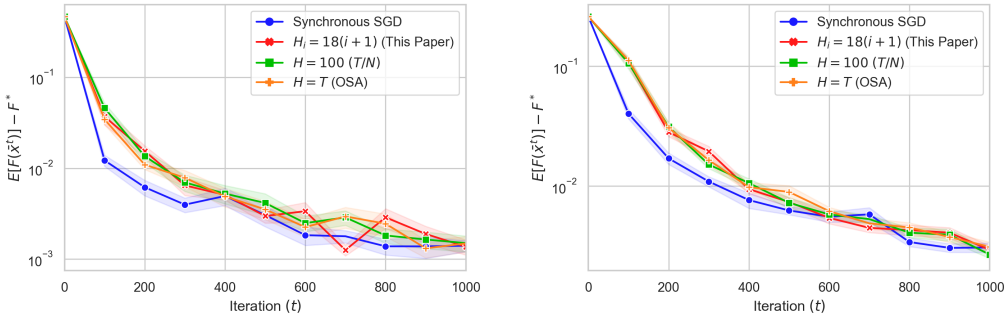


(a) NSQIP data set.

(b) a9a data set.

Figure 3: Minimizing (5) using Local SGD with different communication strategies. Figures (a) and (b) show the error over iteration for NSQIP and a9a datasets, respectively. The shaded areas show the 1-standard deviation error bar.

# A More numerical experiments

In this section we present additional numerical experiments. We consider binary classification and select $l_2$-regularized logistic regression with its corresponding loss function as the objective function $F$ to be minimized, i.e.,

$$F(\mathbf{x}) = \frac{1}{M}\sum_{j=1}^{M}\left(\ln(1+\exp(\mathbf{x}^\top \mathbf{A}_j)) - 1_{(b_j=1)}\mathbf{x}^\top \mathbf{A}_j\right) + \frac{\lambda}{2}\|\mathbf{x}\|_2^2, \tag{5}$$

where $\lambda$ is the regularization parameter, $\mathbf{A}_j \in \mathbb{R}^d$ and $b_j \in \{0,1\}$, $j = 1,\ldots,M$ are features (data points) and their corresponding class labels, respectively.

## A.1 Fixed number of workers

Here we use two large datasets. One, a real dataset from the American College of Surgeons National Surgical Quality Improvement Program (NSQIP) to predict whether a specific patient will be re-admitted within 30 days from discharge after general surgery. This dataset consists of $M = 722{,}101$ data points for training with $d = 231$ features including (i) baseline demographic and healthcare status characteristics, (ii) procedure information and (iii) pre-operative, intra-operative, and post-operative variables. Second, the a9a dataset from LIBSVM (Chang, Lin, 2011). This dataset consists of $M = 32{,}561$ data points for training with $d = 124$ features.

We perform Local SGD with $N = 10$ workers, $\lambda = 0.05$, step-size sequence $\eta_t = 3/(\mu(t+1))$ ($\beta = 1$), $T = 1000$ iterations and batch size of $b = 1$ with different communication strategies: (i) synchronized SGD with $H = 1$, (ii) a strategy with the time varying communication intervals with $H_i = a(i+1), a \approx 18$ and $R = 10$ communication rounds proposed in this paper, (iii) a strategy with the same number of communications however with a fixed $H = T/N = 100$, and finally, (iv) one-shot averaging with $H = T$. Each simulation has been repeated 10 times and the average of their performance is reported in Figure 3.

It can be seen from Figure 3 that all of the communication methods, including OSA, have similar terminal error as synchronized SGD. This further validated our results, especially Theorem 2, since the logistic loss is both twice differentiable and satisfies the PL condition, due to strong convexity of the $l_2$-regularization. Moreover, we do not notice any significant difference between the performance of the varying and constant local steps, mainly because even a method with only one communication round (OSA) performs just as well.

### A.2 Comparison with FedAC

Here, we perform an extensive comparison between different methods using different number of workers $N$ and communication rounds $R$. We adopt a setting similar to that of Figure 4 in Yuan, Ma (2020). More specifically, we compare our communication strategy with other baselines and FedAC, using logistic regression (5) on the a9a dataset with $\lambda = 0.01$ and $T = 8192$.

The results in Figures 4 and 5 are obtained by tuning the fixed learning rate $\eta$ over the set $\{1e^{-3}, 2e^{-3}, 5e^{-3}, 1e^{-2}, \ldots, 2, 5, 10\}$ for all the methods except for Local SGD with growing intervals, where we used $\eta_t = 3/(\mu(t+1))$ without any tuning.

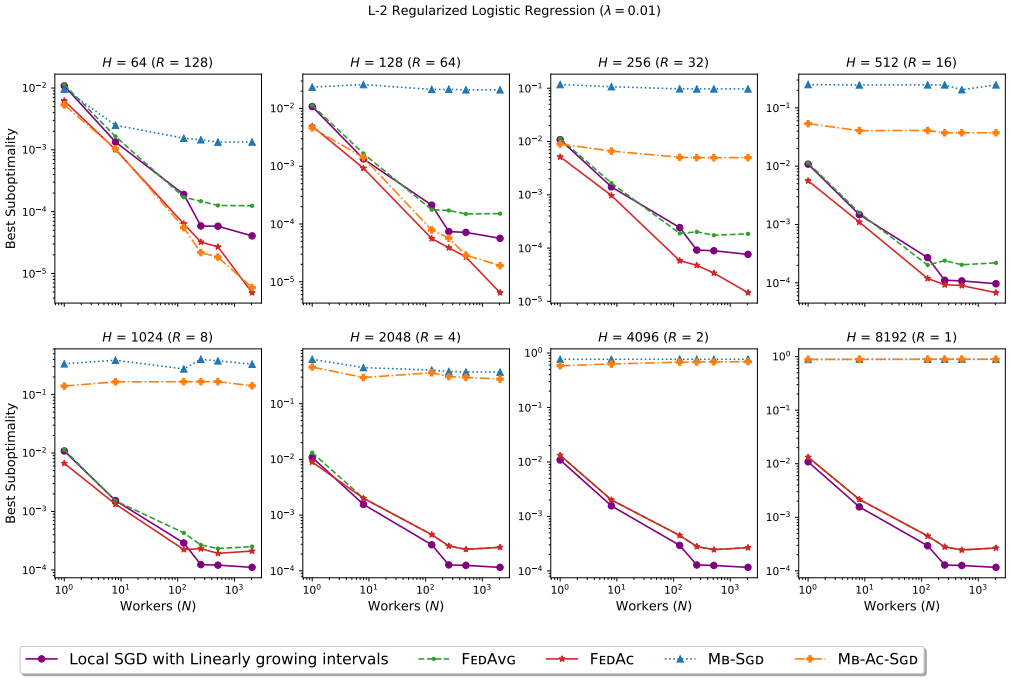


Figure 4: Comparison of Local SGD with (linearly) growing communication intervals introduced in this paper with other baseline methods on the observed linear speed-up w.r.t. $N$ workers ($\lambda = 0.01$).

We observe from Figure 4 that when the number of communications $R$ is large ($R \geq 16$), FedAC has better performance across different values of $N$. However, as the communication becomes sparse, Local SGD with growing communication intervals outperforms all the other methods, specifically as the number of workers increases. We also notice that both Mini-Batch SGD and its accelerated version have a relatively poor performance as $N$ or $H$ increase. Similar observations can be made from Figure 5.

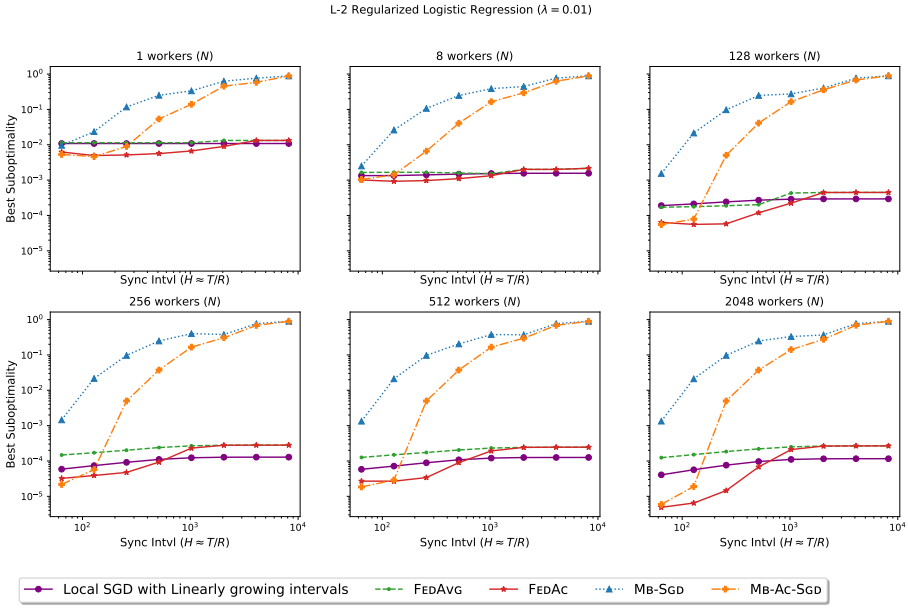


Figure 5: Comparison of Local SGD with (linearly) growing communication intervals introduced in this paper with other baseline methods on the dependency on number of communications ($\lambda = 0.01$).

We notice that increasing strong convexity to $\lambda = 1.0$, results in our communication strategy to uniformly outperform all the other methods, across all values of $N$ and $R$ (see Figure 6).

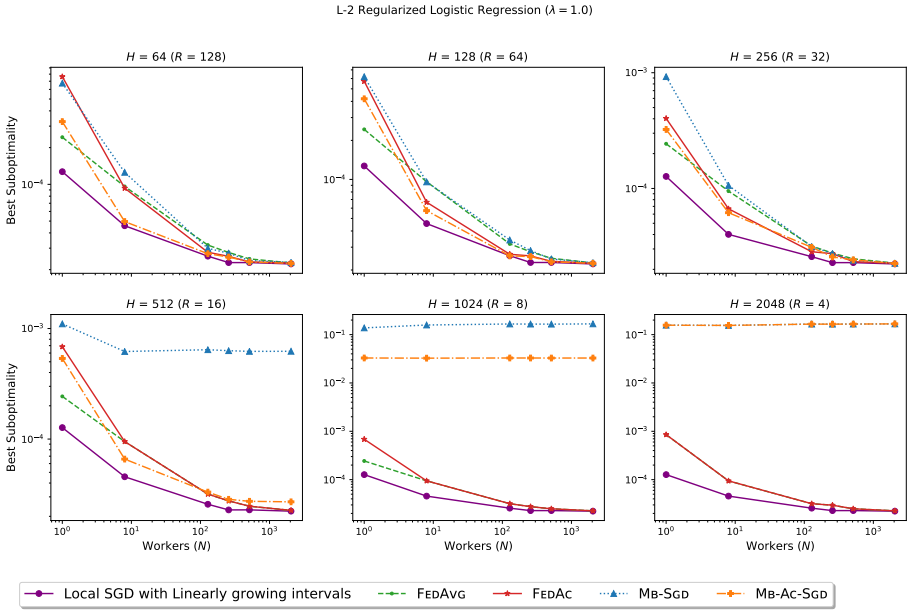


Figure 6: Comparison of Local SGD with (linearly) growing communication intervals introduced in this paper with other baseline methods on the observed linear speed-up w.r.t. $N$ workers ($\lambda = 1.0$).

## B Local SGD

Here we present a few results which will be used later to prove Theorem 1 as well as to better understand the choice of varying number of local steps. In the following theorem, we show an upper bound for the sub-optimality error, in the sense of function value, for any choice of communication times $\mathcal{I}$. Theorem 1 will be obtained by specializing the following bound.

First, let us introduce some notation. Let $0 = \tau_0 < \tau_1 < \ldots < \tau_R = T$ be the communication times and denote the most recent communication time by $\tau(t) := \max\{t' \in \mathcal{I} | t' \leq t\}$. Define $H_i := \tau_{i+1} - \tau_i$, as the length of the $(i+1)$-st inter-communication interval, for $i = 0, \ldots, R-1$.

**Theorem 3.** *Suppose Assumptions 1, 2 and 5 hold. Choose $\beta \geq 9\kappa$ and communication times $\mathcal{I} = \{\tau_i | i = 1, \ldots, R\}$ such that it holds for $i = 0, \ldots, R-1$,*

$$12\kappa^2 c \ln(1 + \frac{H_i - 1}{\tau_i + \beta}) + 3\kappa(1 + \frac{c}{N}) - (\tau_i + \beta) \leq 0. \tag{6}$$

*Set step-sizes $\eta_t = 3/(\mu(t+\beta))$, $t = 0, 1, \ldots, T-1$. Then, using Algorithm 1, we have*

$$\mathbb{E}[f(\bar{\mathbf{x}}^T)] - f^* \leq \frac{\beta^2 (f(\bar{\mathbf{x}}^0) - f^*)}{T^2} + \frac{9L\sigma^2}{2\mu^2 NT} + \frac{18L^2\sigma^2}{\mu^3 T^2} \sum_{t=0}^{T-1} \frac{t - \tau(t)}{t+\beta}, \tag{7}$$

The last term in Equation (7) is due the to disagreement between workers (consensus error), introduced by local computations without any communication. As the inter-communication intervals become larger, $t - \tau(t)$ becomes larger as well and increases the overall optimization error. This term explains the trade-off between communication efficiency and the optimization error.

Note that condition (6) is mild. For instance, it suffices to set $\beta \geq \max\{12\kappa^2 c \ln(1 + T/(9\kappa)) + 3\kappa(1 + c/N), 9\kappa\}$. Moreover, the bound in (7) is for the last iterate $T$, and does not require keeping track of a weighted average of all the iterates.

Theorem 3 not only bounds the optimization error, but introduces a methodological approach to select the communication times to achieve smaller errors. For the scenarios when the user can afford to have a certain number of a communications, they can select $\tau_i$ to minimize the last term in (7).

**One-shot averaging.** Plugging $H = T$ in Theorem 3, we obtain a convergence rate of $\mathcal{O}(\kappa^2\sigma^2/(\mu T))$ without any linear speed-up. Among previous works, only Khaled et al. (2020) show a similar result.

### B.1 Fixed-length intervals

A simple way to select the communication times $\mathcal{I}$, is to split the whole training time $T$ to $R$ intervals of length at most $H$. Then we can use the following bound in Equation (7),

$$\sum_{t=0}^{T-1} \frac{t - \tau(t)}{t+\beta} \leq (H-1) \sum_{t=0}^{T-1} \frac{1}{t+\beta} \leq (H-1) \ln(1 + \frac{T}{\beta - 1}).$$

We state this result formally in the following corollary.

**Corollary 2.** *Suppose assumptions of Theorem 3 hold and in addition, workers communicate at least once every $H$ iterations. Then,*

$$\mathbb{E}[f(\bar{\mathbf{x}}^T)] - f^* \leq \frac{\beta^2 (f(\bar{\mathbf{x}}^0) - F^*)}{T^2} + \frac{9L\sigma^2}{2\mu^2 NT} + \frac{18L^2\sigma^2(H-1)}{\mu^3 T^2} \ln(1 + \frac{T}{\beta - 1}). \tag{8}$$

**Linear speed-up.** Setting $H = \mathcal{O}(T/(N \ln(T)))$ we achieve linear speed-up in the number of workers, which is equivalent to a communication complexity of $R = \Omega(N \ln(T))$. To the best of the authors' knowledge, this is the tightest communication complexity that is shown to achieve linear speed-up. Khaled et al. (2020) and Stich, Karimireddy (2019) have shown a similar communication complexity.

**Recovering synchronized SGD.** When $H = 1$, the last term in (8) disappears and we recover the convergence rate of parallel SGD, albeit, with a worse dependence on $\kappa$.

### B.2 Sketch of proof

Here we give an outline of the proofs for the Local SGD results presented in this paper. The proof of the following lemmas are provided in the next section.

**Perturbed iterates.** A common approach in analyzing parallel algorithms such as Local SGD is to study the evolution of the sequence $\{\bar{\mathbf{x}}^t\}_{t\geq 0}$. We have,

$$\bar{\mathbf{x}}^{t+1} = \bar{\mathbf{x}}^t - \frac{\eta_t}{N}\sum_{i=1}^{N}\hat{\mathbf{g}}_i^t = \bar{\mathbf{x}}^t - \eta_t\tilde{\mathbf{g}}^t, \tag{9}$$

where $\tilde{\mathbf{g}}^t := (\sum_{i=1}^{N}\hat{\mathbf{g}}_i^t)/N$ is the average of the stochastic gradient estimates of all workers.

Let us define $\xi^t := \mathbb{E}[f(\bar{\mathbf{x}}^t)] - f^*$ to be the optimality error. The following lemma, which is similar to a part of the proof found in Haddadpour et al. (2019), bounds the optimality error at each iteration recursively.

**Lemma 1.** *Let Assumptions 1, 2 and 5 hold. Then,*

$$\xi^{t+1} \leq \xi^t(1-\mu\eta_t) + \frac{L^2\eta_t}{2N}\mathbb{E}\left[\sum_{i=1}^{N}\|\bar{\mathbf{x}}^t - \mathbf{x}_i^t\|^2\right] + \frac{\eta_t^2 L}{2}\mathbb{E}[\|\tilde{\mathbf{g}}^t\|_2^2] - \frac{\eta_t}{2N}\mathbb{E}\left[\sum_{i=1}^{N}\|\nabla f(\mathbf{x}_i^t)\|^2\right].$$

Equipped with Lemma 1, we can bound the consensus error ($\mathbb{E}[\sum_{i=1}^{N}\|\bar{\mathbf{x}}^t - \mathbf{x}_i^t\|^2]$) as well as the term $\mathbb{E}[\|\tilde{\mathbf{g}}^t\|^2]$ in the following lemmas.

**Consensus error.** In the following lemmas, we utilize the structure of the problem to bound the consensus error recursively. Let us define $\mathbf{g}_i^t = \nabla f(\mathbf{x}_i^t)$ as the true gradient at worker $i$'s iterate at time $t$.

**Lemma 2.** *Let Assumptions 1, 2 and 5 hold. Then,*

$$\mathbb{E}\left[\sum_{i=1}^{N}\|\mathbf{x}_i^{t+1} - \bar{\mathbf{x}}^{t+1}\|^2\right] \leq \mathbb{E}\left[\sum_{i=1}^{N}\|\mathbf{x}_i^t - \bar{\mathbf{x}}^t\|^2\right](1-\eta_t\mu+\eta_t^2\mu L) + (N-1)\eta_t^2\sigma^2 + \left(1-\frac{1}{N}\right)\eta_t^2 c\mathbb{E}\left[\sum_{i=1}^{N}\|\mathbf{g}_i^t\|^2\right]. \tag{10}$$

This lemma, bounds how much the consensus error grows at each iteration. Of course, when workers communicate, this error resets to zero and thus, we can calculate an upper bound for the consensus error, knowing the last iteration communication occurred and the step-size sequence. The following lemma takes care of that. Before stating the following lemma, let us define $G^t := \frac{1}{n}\sum_{i=1}^{N}\|\mathbf{g}_i^t\|^2$.

**Lemma 3.** *Let assumptions of Theorem 3 hold. Then,*

$$\mathbb{E}\left[\sum_{i=1}^{N}\|\mathbf{x}_i^t - \bar{\mathbf{x}}^t\|^2\right] \leq 12(N-1)\sum_{k=\tau(t)}^{t-1}\frac{c\mathbb{E}\left[G^k\right]+\sigma^2}{\mu^2(t+\beta)^2}. \tag{11}$$

**Variance.** Our next lemma bounds $E[\|\tilde{\mathbf{g}}^t\|^2]$.

**Lemma 4.** *Under Assumption 2 we have,*

$$\mathbb{E}\left[\left\|\tilde{\mathbf{g}}^t\right\|^2\right] \leq \left(1+\frac{c}{N}\right)\mathbb{E}\left[G^t\right] + \frac{\sigma^2}{N}.$$

### B.3 Proofs

Let us define the following notations used in the proofs presented here.

$$\bar{\mathbf{g}}^t := \frac{1}{N}\sum_{i=1}^{n}\mathbf{g}_i^t, \qquad G^t := \frac{1}{N}\sum_{i=1}^{N}\|\mathbf{g}_i^t\|^2, \qquad \mathbf{w}_i^t := \hat{\mathbf{g}}_i^t - \mathbf{g}_i^t.$$

Moreover, define $\mathcal{F}^t := \{\mathbf{x}_i^k, \hat{\mathbf{g}}_i^k | 1\leq i\leq N, 0\leq k\leq t-1\} \cup \{\mathbf{x}_i^t | 1\leq i\leq N\}$.

*Proof of Lemma 1.* By Assumptions 1 and 2 and (9) we have,

$$\mathbb{E}[f(\bar{\mathbf{x}}^{t+1}) - f(\bar{\mathbf{x}}^t)] \leq -\eta_t \mathbb{E}[\langle \nabla f(\bar{\mathbf{x}}^t), \tilde{\mathbf{g}}^t \rangle] + \frac{\eta_t^2 L}{2} \mathbb{E}[\|\tilde{\mathbf{g}}^t\|_2^2]. \tag{12}$$

We bound the first term on the R.H.S of (12) by conditioning on $\mathcal{F}^t$ as follows:

$$\begin{aligned}
\mathbb{E}[\langle \nabla f(\bar{\mathbf{x}}^t), \tilde{\mathbf{g}}^t \rangle | \mathcal{F}^t] &= \frac{1}{N} \sum_{i=1}^N \langle \nabla f(\bar{\mathbf{x}}^t), \mathbb{E}[\hat{\mathbf{g}}_i^t | \mathbf{x}_i^t] \rangle \\
&= \frac{1}{2} \|\nabla f(\bar{\mathbf{x}}^t)\|^2 + \frac{1}{2N} \sum_{i=1}^N \|\nabla f(\mathbf{x}_i^t)\|^2 - \frac{1}{2N} \sum_{i=1}^N \|\nabla f(\bar{\mathbf{x}}^t) - \nabla f(\mathbf{x}_i^t)\|^2 \\
&\geq \mu(f(\bar{\mathbf{x}}^t) - f^*) + \frac{1}{2N} \sum_{i=1}^N \|\nabla f(\mathbf{x}_i^t)\|^2 - \frac{L^2}{2N} \sum_{i=1}^N \|\bar{\mathbf{x}}^t - \mathbf{x}_i^t\|^2,
\end{aligned} \tag{13}$$

where we used $\langle a, b \rangle = \frac{1}{2}\|a\|^2 + \frac{1}{2}\|b\|^2 - \frac{1}{2}\|a - b\|^2$ in the second equation and $(1/2)\|\nabla f(\mathbf{x})\|^2 \geq \mu(f(\mathbf{x}) - f^*)$ as well as smoothness of $f$ in the last inequality. Taking full expectation of (13) and combining it with (12) concludes the lemma. □

We state an important identity in the following lemma.

**Lemma 5.** *Let $\mathbf{u}_1, \dots \mathbf{u}_n \in \mathbb{R}^d$ be $n$ arbitrary vectors. Define $\bar{\mathbf{u}} = (\sum_{i=1}^n \mathbf{u}_i)/n$. Then,*

$$\sum_{i=1}^n \|\mathbf{u}_i - \bar{\mathbf{u}}\|^2 = \sum_{i=1}^n \|\mathbf{u}_i\|^2 - n\|\bar{\mathbf{u}}\|^2.$$

*Proof.* We have

$$\begin{aligned}
\sum_{i=1}^n \|\mathbf{u}_i - \bar{\mathbf{u}}\|^2 &= \sum_{i=1}^n \|\mathbf{u}_i\|^2 + n\|\bar{\mathbf{u}}\|^2 - 2\sum_{i=1}^n \langle \mathbf{u}_i, \bar{\mathbf{u}} \rangle \\
&= \sum_{i=1}^n \|\mathbf{u}_i\|^2 + n\|\bar{\mathbf{u}}\|^2 - 2n\langle \bar{\mathbf{u}}, \bar{\mathbf{u}} \rangle \\
&= \sum_{i=1}^n \|\mathbf{u}_i\|^2 - n\|\bar{\mathbf{u}}\|^2.
\end{aligned}$$

□

*Proof of Lemma 2.* We have,

$$\sum_{i=1}^N \mathbb{E}\left[\left\|\mathbf{x}_i^{t+1} - \bar{\mathbf{x}}^{t+1}\right\|^2\right] = \sum_{i=1}^N \left\|\mathbb{E}\left[\mathbf{x}_i^{t+1} - \bar{\mathbf{x}}^{t+1}\right]\right\|^2 + \sum_{i=1}^N \mathbb{E}\left[\left\|\mathbf{x}_i^{t+1} - \bar{\mathbf{x}}^{t+1} - \mathbb{E}\left[\mathbf{x}_i^{t+1} - \bar{\mathbf{x}}^{t+1}\right]\right\|^2\right]. \tag{14}$$

Let us consider the first term on the right hand side of (14). Taking conditional expectation of both sides of (9) implies,

$$\begin{aligned}
\sum_{i=1}^N \left\|\mathbb{E}\left[\mathbf{x}_i^{t+1} - \bar{\mathbf{x}}^{t+1} | \mathcal{F}^t\right]\right\|^2 &= \sum_{i=1}^N \|\mathbf{x}_i^t - \bar{\mathbf{x}}^t - \eta_t(\mathbf{g}_i^t - \bar{\mathbf{g}}^t)\|^2 \\
&= \sum_{i=1}^N \left(\|\mathbf{x}_i^t - \bar{\mathbf{x}}^t\|^2 + \eta_t^2 \|\mathbf{g}_i^t - \bar{\mathbf{g}}^t\|^2 - 2\eta_t \langle \mathbf{g}_i^t, \mathbf{x}_i^t - \bar{\mathbf{x}}^t \rangle\right).
\end{aligned} \tag{15}$$

By $L$-smoothness of $F$, $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|^2 \leq 2L(f(\mathbf{x}) - f(\mathbf{y}) - \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle)$. Thus,

$$\begin{aligned}
\sum_{i=1}^N \|\mathbf{g}_i^t - \bar{\mathbf{g}}^t\|^2 \leq \sum_{i=1}^N \|\mathbf{g}_i^t - \nabla f(\bar{\mathbf{x}}^t)\|^2 \leq \\
\sum_{i=1}^N 2L\left(f(\bar{\mathbf{x}}^t) - f(\mathbf{x}_i^t) - \langle \mathbf{g}_i^t, \bar{\mathbf{x}}^t - \mathbf{x}_i^t \rangle\right) \leq 2L \sum_{i=1}^N \langle \mathbf{g}_i^t, \mathbf{x}_i^t - \bar{\mathbf{x}}^t \rangle.
\end{aligned} \tag{16}$$

Moreover, by $\mu$-strong convexity of $F$,

$$\sum_{i=1}^{N}\langle \mathbf{g}_i^t, \mathbf{x}_i^t - \bar{\mathbf{x}}^t\rangle \geq \sum_{i=1}^{N}\left(f(\mathbf{x}_i^t) - f(\bar{\mathbf{x}}_i^t) + \frac{\mu}{2}\|\mathbf{x}_i^t - \bar{\mathbf{x}}^t\|^2\right) \geq \frac{\mu}{2}\sum_{i=1}^{N}\|\mathbf{x}_i^t - \bar{\mathbf{x}}^t\|^2. \tag{17}$$

We used the Jensen's inequality $\sum_{i=1}^{N} f(\mathbf{x}_i^t) - f(\bar{\mathbf{x}}^t) \leq 0$ in both equations above. Combining (15)-(17) and having $\eta_t < 1/L$ we obtain,

$$\begin{aligned}\sum_{i=1}^{N}\|\mathbb{E}[\mathbf{x}_i^{t+1} - \bar{\mathbf{x}}^{t+1}|\mathcal{F}^t]\|^2 &\leq \sum_{i=1}^{N}\|\mathbf{x}_i^t - \bar{\mathbf{x}}^t\|^2 - (2\eta_t - 2\eta_t^2 L)\sum_{i=1}^{N}\langle \mathbf{g}_i^t, \mathbf{x}_i^t - \bar{\mathbf{x}}^t\rangle \\ &\leq \sum_{i=1}^{N}\|\mathbf{x}_i^t - \bar{\mathbf{x}}^t\|^2\left(1 - \eta_t\mu + \eta_t^2\mu L\right).\end{aligned}$$

Now, consider the second term on the right hand side of (14). We have,

$$\begin{aligned}\sum_{i=1}^{N}\mathbb{E}\left[\left\|\mathbf{x}_i^{t+1} - \bar{\mathbf{x}}^{t+1} - \mathbb{E}[\mathbf{x}_i^{t+1} - \bar{\mathbf{x}}^{t+1}]\right\|^2|\mathcal{F}^t\right] &= \sum_{i=1}^{N}\mathbb{E}\left[\left\|\mathbf{x}_i^{t+1} - \mathbb{E}[\mathbf{x}_i^{t+1}] - (\bar{\mathbf{x}}^{t+1} - \mathbb{E}[\bar{\mathbf{x}}^{t+1}])\right\|^2|\mathcal{F}^t\right] \\ &= \eta_t^2\sum_{i=1}^{N}\mathbb{E}\left[\left\|\mathbf{w}_i^t - \bar{\mathbf{w}}^t\right\|^2|\mathcal{F}^t\right] \\ &= \eta_t^2\left(\sum_{i=1}^{N}\mathbb{E}\left[\left\|\mathbf{w}_i^t\right\|^2|\mathcal{F}^t\right] - N\mathbb{E}\left[\left\|\bar{\mathbf{w}}^t\right\|^2|\mathcal{F}^t\right]\right) \\ &= \eta_t^2\sum_{i=1}^{N}\mathbb{E}\left[\left\|\mathbf{w}_i^t\right\|^2|\mathcal{F}^t\right](1 - \frac{1}{N}) \\ &\leq (N-1)\eta_t^2\sigma^2 + (1 - \frac{1}{N})\eta_t^2 c\sum_{i=1}^{N}\|\mathbf{g}_i^t\|^2,\end{aligned}$$

where $\mathbf{w}_i^t$ are defined at the beginning of this section and $\bar{\mathbf{w}}^t := (\sum_{i=1}^{N}\mathbf{w}_i^t)/n$ and we used Lemma 5 in the third equation and the conditional independence of $\mathbf{w}_i^t$ to use $\mathbb{E}[\|\bar{\mathbf{w}}^t\|^2|\mathcal{F}^t] = (1/N^2)\sum_{i=1}^{N}\mathbb{E}[\|\mathbf{w}_i^t\|^2|\mathcal{F}^t]$ in the last equality. Taking full expectation of the two relations above with respect to $\mathcal{F}^t$ and combining them with (14) completes the proof. □

Before proving Lemma 3, let us state and prove the following lemma.

**Lemma 6.** *Let $b \geq a > 2$ be integers. Define $\Phi(a,b) = \prod_{i=a}^{b}\left(1 - \frac{2}{i}\right)$. We then have $\Phi(a,b) \leq \left(\frac{a}{b+1}\right)^2$.*

*Proof.* Indeed,

$$\ln(\Phi(a,b)) = \sum_{i=a}^{b}\ln\left(1 - \frac{2}{i}\right) \leq \sum_{i=a}^{b} -\frac{2}{i} \leq -2\left[\ln(b+1) - \ln(a)\right].$$

where we used the inequality $\ln(1-x) \leq -x$ as well as the standard technique of viewing $\sum_{i=a}^{b} 1/i$ as a Riemann sum for $\int_a^{b+1} 1/x\ dx$ and observing that the Riemann sum overstates the integral. Exponentiating both sides now implies the lemma. □

*Proof of Lemma 3.* Define $a^k = \mathbb{E}\left[\sum_{i=1}^N \|\mathbf{x}_i^k - \bar{\mathbf{x}}^k\|^2\right]$ and $\Delta_k = (1 - \eta_k\mu + \eta_k^2\mu L)$ for $k \geq 0$ . By Lemma 2,

$$\begin{aligned}
a^t &\leq \Delta_{t-1}a^{t-1} + \eta_{t-1}^2(N-1)(\sigma^2 + c\mathbb{E}[G^{t-1}]) \\
&\leq \Delta_{t-1}(\Delta_{t-2}a^{t-2} + \eta_{t-2}^2(N-1)(\sigma^2 + c\mathbb{E}[G^{t-2}])) + \eta_{t-1}^2(N-1)(\sigma^2 + c\mathbb{E}[G^{t-1}]) \\
&\leq \ldots \leq \prod_{k=\tau(t)}^{t-1}\Delta_k a^{\tau(t)} + (N-1)\sum_{k=\tau(t)}^{t-1}\eta_k^2(\sigma^2 + c\mathbb{E}[G^k])\prod_{i=k+1}^{t-1}\Delta_i \\
&= (N-1)\sum_{k=\tau(t)}^{t-1}\eta_k^2(\sigma^2 + c\mathbb{E}[G^k])\prod_{i=k+1}^{t-1}\Delta_i,
\end{aligned}$$

where we used $a^{\tau(t)} = 0$ in the last equation. By the choice of stepsize and $\beta \geq 9\kappa$, we have

$$\Delta_k = 1 - \frac{3}{k+\beta} + \frac{9L}{\mu(k+\beta)^2} \leq 1 - \frac{3}{k+\beta} + \frac{9\kappa}{(k+\beta)\beta} \leq 1 - \frac{3}{k+\beta} + \frac{1}{(k+\beta)} = 1 - \frac{2}{k+\beta}.$$

Therefore, by Lemma 6,

$$a^t \leq (N-1)\sum_{k=\tau(t)}^{t-1}\frac{9(\sigma^2 + c\mathbb{E}[G^k])}{\mu^2(k+\beta)^2}\frac{(k+\beta+1)^2}{(t+\beta)^2} \leq (N-1)\sum_{k=\tau(t)}^{t-1}\frac{12(\sigma^2 + c\mathbb{E}[G^k])}{\mu^2(t+\beta)^2},$$

where we used $9(k+\beta+1)^2/(k+\beta)^2 \leq 9(\beta+1)^2/\beta^2 \leq 9(10/9)^2 \leq 12$ since $\beta \geq 9\kappa \geq 9$. □

*Proof of Lemma 4.* We have,

$$\mathbb{E}[\|\tilde{\mathbf{g}}^t\|^2|\mathcal{F}^t] = \mathbb{E}[\|\bar{\mathbf{g}}^t + \bar{\epsilon}^t\|^2|\mathcal{F}_t] = \|\bar{\mathbf{g}}^t\|^2 + \mathbb{E}[\|\bar{\mathbf{w}}^t\|^2|\mathcal{F}^t] \leq \frac{1}{N}\sum_{i=1}^N\|\mathbf{g}_i^t\|^2 + \frac{1}{N^2}\sum_{i=1}^N(\sigma^2 + c\|\mathbf{g}_i^t\|^2),$$

where in the last inequality we used Lemma 5 and the conditional independency of $\mathbf{w}_i^t$ to decouple the noise terms. □

*Proof of Theorem 3.* Combining Equations Lemmas 1-4 and plugging $\eta_t = 3/(\mu(t+\beta))$ we obtain

$$\begin{aligned}
\xi^{t+1} \leq \xi^t(1 - \mu\eta_t) &+ \frac{18L^2}{\mu^3(t+\beta)^3}\sum_{k=\tau(t)}^{t-1}\left(c\mathbb{E}[G^k] + \sigma^2\right) \\
&+ \frac{9L}{2\mu^2(t+\beta)^2}\left(\left(1 + \frac{c}{N}\right)\mathbb{E}[G^t] + \frac{\sigma^2}{N}\right) - \frac{3}{2\mu(t+\beta)}\mathbb{E}[G^t].
\end{aligned}$$

Let us multiply both sides of relation above by $(t+\beta)^2$ and use the following inequality

$$(1 - \mu\eta_t)(t+\beta)^2 = \left(1 - \frac{2}{t+\beta}\right)(t+\beta)^2 = (t+\beta)^2 - 2(t+\beta) < (t+\beta-1)^2,$$

to obtain,

$$\begin{aligned}
\xi^{t+1}(t+\beta)^2 \leq \xi^t(t+\beta-1)^2 &+ \frac{9L\sigma^2}{2\mu^2 N} + \\
&\frac{18L^2}{\mu^3(t+\beta)}\sum_{k=\tau(t)}^{t-1}\left(c\mathbb{E}[G^k] + \sigma^2\right) + \left(\frac{9L}{2\mu^2}\left(1 + \frac{c}{N}\right) - \frac{3(t+\beta)}{2\mu}\right)\mathbb{E}[G^t].
\end{aligned}$$

Summing relation above for $t = \tau_i, \ldots, \tau_{i+1} - 1$, where $\tau_i, \tau_{i+1} \in \mathcal{I}$ are two consecutive communication times, implies,

$$\begin{aligned}
\xi^{\tau_{i+1}}(\tau_{i+1}+\beta-1)^2 \leq \xi^{\tau_i}(\tau_i+\beta-1)^2 &+ \frac{9L\sigma^2}{2\mu^2 N}(\tau_{i+1} - \tau_i) + \frac{18L^2\sigma^2}{\mu^3}\sum_{t=\tau_i}^{\tau_{i+1}-1}\frac{t-\tau_i}{t+\beta} \\
&+ \sum_{t=\tau_i}^{\tau_{i+1}-1}\mathbb{E}[G^t]\left(\sum_{k=t+1}^{\tau_{i+1}-1}\frac{18L^2c}{\mu^3(k+\beta)} + \frac{9L}{2\mu^2}\left(1 + \frac{c}{N}\right) - \frac{3(t+\beta)}{2\mu}\right).
\end{aligned}$$

Each of the coefficients of $\mathbb{E}[G^t]$ in above can be bounded by,

$$\begin{aligned}\sum_{k=t+1}^{\tau_{i+1}-1} \frac{18L^2c}{\mu^3(k+\beta)} + \frac{9L}{2\mu^2}\left(1+\frac{c}{N}\right) - \frac{3(t+\beta)}{2\mu} &\leq \frac{18L^2c}{\mu^3}\ln\left(\frac{\tau_{i+1}+\beta-1}{\tau_i+\beta}\right) + \frac{9L}{2\mu^2}\left(1+\frac{c}{N}\right) - \frac{3\tau_i+\beta}{2\mu} \\ &= \frac{3}{2\mu}\left(12\kappa^2 c\ln\left(1+\frac{H_i-1}{\tau_i+\beta}\right) + 3\kappa\left(1+\frac{c}{N}\right) - (\tau_i+\beta)\right) \\ &\leq 0,\end{aligned}$$

where we used $\sum_{k=t_1+1}^{t_2} 1/k \leq \int_{t_1}^{t_2} dx/x = \ln(t_2/t_1)$ in the first inequality and the last inequality comes from the assumption of the theorem. Now that the coefficients of $\mathbb{E}[G^k]$ are non-positive, we can simply ignore them and obtain,

$$\xi^{\tau_{i+1}}(\tau_{i+1}+\beta-1)^2 \leq \xi^{\tau_i}(\tau_i+\beta-1)^2 + \frac{9L\sigma^2}{2\mu^2 N}(\tau_{i+1}-\tau_i) + \frac{18L^2\sigma^2}{\mu^3}\sum_{t=\tau_i}^{\tau_{i+1}-1}\frac{t-\tau_i}{t+\beta}.$$

Recursing relation above for $i = 0, \ldots, R-1$ implies,

$$\xi^T(T+\beta-1)^2 \leq \xi^0(\beta-1)^2 + \frac{9L\sigma^2}{2\mu^2 N}T + \frac{18L^2\sigma^2}{\mu^3}\sum_{t=0}^{T-1}\frac{t-\tau(t)}{t+\beta}.$$

Dividing both sides by $(T+\beta-1)^2$ concludes the proof. □

*Proof of Theorem 1.* We have,

$$\tau_j = \tau_0 + \sum_{i=0}^{j-1} H_i = a\frac{j(j+1)}{2}, \qquad j = 0, \ldots, k-1.$$

Hence,

$$\begin{aligned}1 + \frac{H_0-1}{\tau_0+\beta} &= 1 + \frac{a-1}{\beta} \leq 1 + \frac{2T}{9\kappa R^2} \leq 1 + \frac{T}{4\kappa R^2}, \\ 1 + \frac{H_i-1}{\tau_i+\beta} &\leq 1 + \frac{a(i+1)}{\frac{ai(i+1)}{2}} \leq 3, \qquad i \geq 1.\end{aligned}$$

Thus, $12\kappa^2 c\ln(1+\frac{H_i-1}{\tau_i+\beta}) + 3\kappa(1+\frac{c}{N}) - (\tau_i+\beta) \leq 0, i = 0, \ldots, R-1$ and we can use Theorem 3. Moreover,

$$\begin{aligned}\sum_{t=0}^{T-1}\frac{t-\tau(t)}{t+\beta} &\leq \sum_{j=0}^{R-1}\sum_{i=1}^{H_j-1}\frac{i}{\tau_j+i+\beta} \leq H_0 + \sum_{j=1}^{R-1}\sum_{i=1}^{H_j-1}\frac{i}{\tau_j+1+\beta} \\ &= a + \sum_{j=1}^{R-1}\frac{H_j(H_j-1)}{2(\tau_j+1+\beta)} = a + \sum_{j=1}^{R-1}\frac{a(j+1)(a(j+1)-1)}{aj(j+1)+2(1+\beta)} \\ &\leq a + \sum_{j=1}^{R-1}\frac{a^2(j+1)^2}{aj(j+1)} \leq 2aR.\end{aligned}$$

Plugging the values of $R$ and $a$ implies,

$$\sum_{t=0}^{T-1}\frac{t-\tau(t)}{t+\beta} \leq 2aR \leq 2(\frac{2T}{R^2}+1)R = \frac{4T}{R} + 2R \leq \frac{4T}{R} + \frac{4T}{R} = \frac{8T}{R},$$

where we used $R \leq \sqrt{2T}$ in the last inequality. Using the relation above together with Theorem 3 concludes the proof. □

## C One-shot averaging

In this section we prove Theorem 2 for one-shot averaging. The main idea is to use second order approximation for gradients at any point with respect to the minimizer and show that the residual errors are *insignificant*, using concentration results from Karimi et al. (2016).

### C.1 Preliminaries

Define $\mathbf{v}(\mathbf{y}, \mathbf{x}) := \nabla f(\mathbf{y}) - \big(\nabla f(\mathbf{x}) + \nabla^2 f(\mathbf{x})(\mathbf{y} - \mathbf{x})\big)$ and $\mathbf{v}_i^t = \mathbf{v}(\mathbf{x}_i^t, \mathbf{x}^*)$.

**Lemma 7.** *Let Assumption 4 hold. Then* $|[\mathbf{v}(\mathbf{x}, \mathbf{x}^*)]_i| = o(\|\mathbf{x} - \mathbf{x}^*\|)$ *for* $i = 1, \ldots, d$.

*Proof.* Denote $h_i(\mathbf{x}) = [\nabla f(\mathbf{x})]_i$. Then by Assumption 4, $h_i$ is continuously differentiable over an open set containing $\mathbf{x}^*$. Thus,

$$h_i(\mathbf{x}) = h_i(\mathbf{x}^*) + \nabla h_i(\mathbf{x}^*)^\top (\mathbf{x} - \mathbf{x}^*) + o(\|\mathbf{x} - \mathbf{x}^*\|) = \nabla h_i(\mathbf{x}^*)^\top (\mathbf{x} - \mathbf{x}^*) + o(\|\mathbf{x} - \mathbf{x}^*\|).$$

Therefore,

$$[\mathbf{v}(\mathbf{x}, \mathbf{x}^*)]_i = h_i(\mathbf{x}) - \sum_{j=1}^{d} \frac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{x}^*)[\mathbf{x} - \mathbf{x}^*]_j = h_i(\mathbf{x}) - \nabla h_i(\mathbf{x}^*)^\top (\mathbf{x} - \mathbf{x}^*) = o(\|\mathbf{x} - \mathbf{x}^*\|).$$

□

Let us define $u(r) := \max_{\|\mathbf{x} - \mathbf{x}^*\| \le r} \|\mathbf{v}(\mathbf{x}, \mathbf{x}^*)\|$. We have $u(r) = o(r)$.

**Theorem 4** (Karimi et al. (2016), Theorem 1)**.** *Under Assumptions 1 and 3, the following inequality, known as the quadratic growth (QG) condition holds:*

$$\|\mathbf{x} - \mathbf{x}^*\|^2 \le \frac{2}{\mu}(f(\mathbf{x}) - f^*).$$

**Lemma 8.** *Under Assumptions 1, 3 and 4 we have,*

$$\nabla^2 f(\mathbf{x}^*) \succeq \mu.$$

*Proof.* The result is established by using the *linear approximation theorem* on a sequence of points converging to $x^*$ on a line, continuity of Hessian as well as the quadratic growth from Theorem 4. Similar approach can be found in the proof of Theorem 2.26 Beck (2014). □

Next, we state a Theorem from Madden et al. (2020) which we will use frequently in the rest of our results.

**Theorem 5** (Madden et al. (2020), Theorem 4 and 13)**.** *Under Assumptions 1, 3 and 6, SGD with step-size sequence* $\{\eta_t\} = \{\theta_t\}$ *defined in* (2)*, constructs a sequence of* $\{\mathbf{x}^t\}$ *such that there exist* $C_1, C_2 > 0$ *such that for* $t \ge t_0$*,*

$$\mathbb{E}[f(\mathbf{x}^t)] - f^* = C_1 \frac{L\sigma^2}{\mu^2 t},$$

*and w.p.* $\ge 1 - \delta$ *for all* $\delta \in (0, 1/e)$*,*

$$f(\mathbf{x}^t) - f^* \le C_2 \frac{L\sigma^2 \log(e/\delta)}{\mu^2 t}.$$

**Lemma 9.** *Under Assumptions 1 and 4 we have,*

$$\|\mathbf{v}(\mathbf{x}, \mathbf{x}^*)\| \le 2L\|\mathbf{x} - \mathbf{x}^*\|. \tag{18}$$

*Proof.* We have,

$$\begin{aligned}
\|\mathbf{v}(\mathbf{x}, \mathbf{x}^*)\| &= \|\nabla f(\mathbf{x}) - \nabla^2 f(\mathbf{x}^*)(\mathbf{x} - \mathbf{x}^*)\| \\
&\le \|\nabla f(\mathbf{x})\| + \|\nabla^2 f(\mathbf{x}^*)(\mathbf{x} - \mathbf{x}^*)\| \\
&\le L\|\mathbf{x} - \mathbf{x}^*\| + \|\nabla^2 f(\mathbf{x}^*)\|_2 \|\mathbf{x} - \mathbf{x}^*\| \\
&\le 2L\|\mathbf{x} - \mathbf{x}^*\|,
\end{aligned}$$

where we used $\|\nabla^2 f(\mathbf{x}^*)\| \le L$ in the last inequality. □

The following lemma is the key result we need to show the asymptotic performance of OSA.

**Lemma 10.** *Under Assumptions 1, 3, 4 and 6 and steps-size sequence $\{\eta_t\} = \{\theta_t\}$ defined in (2), we have*

1. $\mathbb{E}[\|\mathbf{v}_i^t\|^2] = o(\frac{1}{t})$,
2. $\mathbb{E}[\|\mathbf{v}_i^t\|\|\mathbf{x}_i^t - \mathbf{x}^*\|] = o(\frac{1}{t})$.

*Proof.* Let us define $u(r) := \max_{\|\mathbf{x}-\mathbf{x}^*\|\leq r} \|\mathbf{v}(\mathbf{x}, \mathbf{x}^*)\|$. By Lemma 7 we have $u(r) = o(r)$. Also define random variable $r_i^t = \|\mathbf{x}_i^t - \mathbf{x}^*\|$.

Since $u(r) = o(r)$, for any $\epsilon > 0$ there exists $s > 0$ such that for $r \leq s$, $u(r) \leq \sqrt{\epsilon} r$ or $u(r)^2 \leq \epsilon r^2$. We have,

$$\begin{aligned}
\mathbb{E}[\|\mathbf{v}_i^t\|^2] &= \mathbb{E}_{\mathbf{x}_i^t}[\|\mathbf{v}(\mathbf{x}_i^t, \mathbf{x}^*)\|^2] \leq \mathbb{E}_{r_i^t}[u(r_i^t)^2] \\
&= \int_0^\infty u(r)^2 p_{r_i^t}(r) dr \\
&= \int_0^s u(r)^2 p_{r_i^t}(r) dr + \int_s^\infty u(r)^2 p_{r_i^t}(r) dr \\
&\leq \epsilon \int_0^s r^2 p_{r_i^t}(r) dr + 4L^2 \int_s^\infty r^2 p_{r_i^t}(r) dr \\
&= \epsilon \mathbb{E}[(r_i^t)^2] + (4L^2 - \epsilon) \int_s^\infty r^2 p_{r_i^t}(r) dr, \qquad (19)
\end{aligned}$$

where $p_X$ denotes the Probability Density Function (PDF) for random variable $X$ and we used $u(r) \leq 2Lr$ from (18).

Without loss of generality, we assume $t \geq t_0$ for the rest of the proof. By Theorems 4 and 5 we have,

$$\mathbb{E}\left[(r_i^t)^2\right] = \mathbb{E}\left[\left\|\mathbf{x}_i^t - \mathbf{x}^*\right\|^2\right] \leq \frac{2}{\mu}\mathbb{E}[f(\mathbf{x}_i^t) - f^*] \leq \frac{2C_1 L\sigma^2}{\mu^2 t} = \mathcal{O}\left(\frac{1}{t}\right).$$

Moreover, define $J_t(\delta) := C_2 L\sigma^2 \log(e/\delta)/(\mu^2 t)$. Then,

$$\Pr\left((r_i^t)^2 \leq \frac{2J_t(\delta)}{\mu}\right) \geq \Pr\left(f(\mathbf{x}_i^t) - f^* \leq J_t(\delta)\right) \geq 1 - \delta, \qquad \text{for } \delta \in (0, 1/e), \qquad (20)$$

or,

$$F^{-1}_{(r_i^t)^2}(1-\delta) \leq \frac{2J_t(\delta)}{\mu}, \qquad \text{for } \delta \in (0, 1/e), \qquad (21)$$

where $F_X$ denotes the Cumulative Distribution Function (CDF) for random variable $X$. Since $\lim_{t\to\infty} J_t(\delta) = 0$, $\exists t_1 \geq t_0$ such that for $t \geq t_1$, $J_t^{-1}(\mu s^2/2) \in (0, 1/e)$. It follows that,

$$\begin{aligned}
\int_s^\infty r^2 p_{r_i^t}(r) dr &= \int_{s^2}^\infty r^2 p_{(r_i^t)^2}(r^2) dr^2 = \int_{s^2}^\infty r^2 dF_{(r_i^t)^2}(r^2) \\
&= \int_{F_{(r_i^t)^2}(s^2)}^1 F^{-1}_{(r_i^t)^2}(x) dx = \int_{1-F_{(r_i^t)^2}(s^2)}^0 -F^{-1}_{(r_i^t)^2}(1-\delta) d\delta \\
&= \int_0^{1-F_{(r_i^t)^2}(s^2)} F^{-1}_{(r_i^t)^2}(1-\delta) d\delta \\
&\leq \frac{2}{\mu} \int_0^{1-F_{(r_i^t)^2}(s^2)} J_t(\delta) d\delta \leq \frac{2}{\mu} \int_0^{J_t^{-1}(\frac{\mu s^2}{2})} J_t(\delta) d\delta. \qquad (22)
\end{aligned}$$

In the equation above, we switched from Probability Density Function (PDF) $p_{r_i^t}$ to $p_{(r_i^t)^2}$ in the first equality. In the next equality we used $p_X = dF_X/dX$ that holds for any continuous random variable $X$. In third equality, we simply changed variable to $x = F_{(r_i^t)^2}(r^2)$ and without loss of generality we define $F_X^{-1}(y) := \inf\{x | F_X(x) \geq y\}$. In the next equation, again, we simply changed

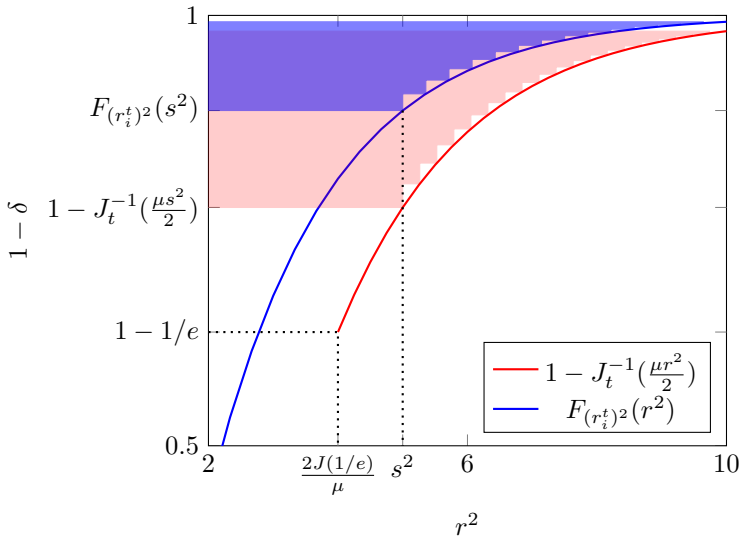


Figure 7: Illustration of integrals in (22)

variable to $\delta = 1 - x$. Finally, in the last two inequalities, we used (21) and a direct result of (20), $1 - F_{(r_i^t)^2}(s^2) \leq J_t^{-1}(\frac{\mu s^2}{2})$ (see Figure 7).

By Lemma 11, $\exists t_2 \geq t_1$ such that for $t \geq t_2$, $\int_0^{J_t^{-1}(\mu s^2/2)} J_t(\delta)d\delta \leq \epsilon B_1/t$, where $B_1 := 2C_2L\sigma^2 e/\mu^2$. Combining with (19) we obtain,

$$\mathbb{E}[\|\mathbf{v}_i^t\|^2] \leq \frac{\epsilon}{t}\left(\frac{2C_1L\sigma^2}{\mu^2} + \frac{16C_2L^3\sigma^2 e}{\mu^3}\right), \qquad t \geq t_2.$$

Next, we show $\mathbb{E}[\|\mathbf{v}_i^t\|\|\mathbf{x}_i^t - \mathbf{x}^*\|] = o(1/t)$. Since $u(r) = o(r)$, for any $\epsilon > 0$, there exists $s' > 0$ such that for $r \leq s'$, $u(r) \leq \epsilon r$. Then,

$$\begin{aligned}
\mathbb{E}[\|\mathbf{v}_i^t\|\|\mathbf{x}_i^t - \mathbf{x}^*\|] &\leq \mathbb{E}[u(r_i^t)r_i^t] \\
&= \int_0^{s'} u(r)rp_{r_i^t}(r)dr + \int_{s'}^{\infty} u(r)rp_{r_i^t}(r)dr \\
&\leq \epsilon\int_0^{s'} r^2 p_{r_i^t}(r) + 4L^2\int_{s'}^{\infty} r^2 p_{r_i^t}(r)dr.
\end{aligned}$$

Following the same steps from the first part of this proof, we obtain $\exists t_3 > 0$ such that,

$$\mathbb{E}[\|\mathbf{v}_i^t\|\|\mathbf{x}_i^t - \mathbf{x}^*\|] \leq \frac{\epsilon}{t}\left(\frac{2C_1L\sigma^2}{\mu^2} + \frac{16C_2L^3\sigma^2 e}{\mu^3}\right), \qquad t \geq t_3.$$

Since we could pick $\epsilon$ arbitrarily small, we showed that $\mathbb{E}[\|\mathbf{v}_i^t\|^2] = o(1/t)$ and $\mathbb{E}[\|\mathbf{v}_i^t\|\|\mathbf{x}_i^t - \mathbf{x}^*\|] = o(1/t)$. □

**Lemma 11.** *Let $q_t : (0, 1/e) \to \mathbb{R}_+$ be defined as $q_t(\delta) = a_1 \log(e/\delta)/t$ for some $a_1 > 0$ and $\forall t \geq 1$. Suppose $y \in range(q_t)$ for $t \geq t_1$, then for any $\epsilon > 0$, there exists $t_2 \geq t_1$ such that for any $t \geq t_2$,*

$$\int_0^{q_t^{-1}(y)} q_t(\delta)d\delta \leq \frac{B\epsilon}{t},$$

*where $B = 2a_1 e$.*

*Proof.* Define $x_t$ such that $q_t(x_t) = y$. Then,

$$\frac{a_1\log(e/x_t)}{t} = y \iff \log(\frac{e}{x_t}) = \frac{yt}{a_1} \iff x_t = \exp(1 - \frac{yt}{a_1}). \tag{23}$$

Moreover,

$$\begin{aligned}\int_0^{x_t} q_t(\delta)d\delta &= \frac{a_1}{t}\int_0^{x_t}\log(\frac{e}{\delta})d\delta \\ &= \frac{a_1}{t}\left(x_t - x_t(\log(x_t)-1)\right) \\ &= \frac{a_1}{t}\left(x_t + x_t(\frac{yt}{a_1})\right) = x_t(y+\frac{a_1}{t}).,\end{aligned}$$

where we used (23) in third equality. First, we note that for $t \geq a_1/y$, we have $y + a_1/t \leq 2y$. Next, we show that for $t$ large enough, $x_t \leq B\epsilon/(2yt)$ for some $B > 0$. We have $\lim_{s\to\infty}\exp(s)/s = \infty$. Therefore $\exists s_0 \geq 1$ such that for $s \geq s_0$, $\exp(s)/s \geq 1/\epsilon$. Thus for $t \geq s_0 a_1/y$ we have,

$$\begin{aligned}&\exp(\frac{ty}{a_1}) \geq \frac{ty}{a_1\epsilon}, \\ \Rightarrow x_t = \exp(1-\frac{yt}{a_1}) &\leq e(\frac{a_1\epsilon}{ty}) = \frac{B\epsilon}{2yt},\end{aligned}$$

where $B := 2a_1 e$. Therefore, for $t \geq t_2 := \max\{s_0 a_1/y, t_1\}$ we have $\int_0^{x_t} q_t(\delta)d\delta \leq 2x_t y \leq B\epsilon/t$. □

### C.2 One-step progress

**Lemma 12.** *Under Assumptions 1, 3, 4 and 6 and steps-size sequence $\{\eta_t\} = \{\theta_t\}$ defined in (2), we have*

$$\mathbb{E}[\|\bar{\mathbf{x}}^{t+1} - \mathbf{x}^*\|^2] \leq (1-\eta_t\mu)^2\mathbb{E}[\|\bar{\mathbf{x}}^t - \mathbf{x}^*\|^2] + \frac{\eta_t^2\sigma^2}{N} + o\left(\frac{1}{t^2}\right). \tag{24}$$

*Proof.* Let us define $\mathbf{A} = \nabla^2 f(\mathbf{x}^*)$. By definition,

$$\nabla f(\mathbf{x}_i^t) = \mathbf{A}(\mathbf{x}_i^t - \mathbf{x}^*) + \mathbf{v}_i^t. \tag{25}$$

Plugging (25) in SGD process and averaging over all $i$ we obtain,

$$\begin{aligned}\bar{\mathbf{x}}^{t+1} &= \bar{\mathbf{x}}^t - \frac{\eta_t}{N}\sum_{i=1}^N \hat{\mathbf{g}}_i^t = \bar{\mathbf{x}}^t - \frac{\eta_t}{N}\sum_{i=1}^N\left(\nabla f(\mathbf{x}_i^t) + \mathbf{w}_i^t\right) \\ &= \bar{\mathbf{x}}^t - \frac{\eta_t}{N}\sum_{i=1}^N\left(\mathbf{A}(\mathbf{x}_i^t - \mathbf{x}^*) + \mathbf{v}_i^t + \mathbf{w}_i^t\right) = \bar{\mathbf{x}}^t - \eta_t\mathbf{A}(\bar{\mathbf{x}}^t - \mathbf{x}^*) + \frac{\eta_t}{N}\sum_{i=1}^N(\mathbf{v}_i^t + \mathbf{w}_i^t).\end{aligned}$$

Thus,

$$\begin{aligned}\mathbb{E}[\|\bar{\mathbf{x}}^{t+1} - \mathbf{x}^*\|^2|\mathcal{F}_t] &= \mathbb{E}[\|(I-\eta_t\mathbf{A})(\bar{\mathbf{x}}^t - \mathbf{x}^*) + \frac{\eta_t}{N}\sum_{i=1}^N(\mathbf{v}_i^t + \mathbf{w}_i^t)\|^2|\mathcal{F}_t] \\ &= \mathbb{E}[\|(I-\eta_t\mathbf{A})(\bar{\mathbf{x}}^t - \mathbf{x}^*) + \frac{\eta_t}{N}\sum_{i=1}^N\mathbf{v}_i^t\|^2|\mathcal{F}_t] + \mathbb{E}[\|\frac{\eta_t}{N}\sum_{i=1}^N\mathbf{w}_i^t\|^2|\mathcal{F}_t] \\ &\leq \mathbb{E}[\|(I-\eta_t\mathbf{A})(\bar{\mathbf{x}}^t - \mathbf{x}^*) + \frac{\eta_t}{N}\sum_{i=1}^N\mathbf{v}_i^t\|^2|\mathcal{F}_t] + \frac{\eta_t^2\sigma^2}{N}.\end{aligned}$$

Taking full expectation with respect to $\mathcal{F}_t$ yields,

$$\begin{aligned}\mathbb{E}[\|\bar{\mathbf{x}}^{t+1} - \mathbf{x}^*\|^2] &\leq \mathbb{E}[\|(I-\eta_t\mathbf{A})(\bar{\mathbf{x}}^t - \mathbf{x}^*) + \frac{\eta_t}{N}\sum_{i=1}^N\mathbf{v}_i^t\|^2] + \frac{\eta_t^2\sigma^2}{N} \\ &= \mathbb{E}[\|(I-\eta_t\mathbf{A})(\bar{\mathbf{x}}^t - \mathbf{x}^*)\|^2] + \frac{\eta_t^2\sigma^2}{N} \\ &+ \underbrace{\mathbb{E}[\|\frac{\eta_t}{N}\sum_{i=1}^N\mathbf{v}_i^t\|^2]}_{T_2} + \underbrace{\mathbb{E}[\|(I-\eta_t\mathbf{A})(\bar{\mathbf{x}}^t - \mathbf{x}^*)\|\|\frac{\eta_t}{N}\sum_{i=1}^N\mathbf{v}_i^t\|]}_{T_3}.\end{aligned} \tag{26}$$

Next we bound $T_2$ and $T_3$. Using Lemma 10 we have,

$$\mathbb{E}[\|\frac{\eta_t}{N}\sum_{i=1}^{N}\mathbf{v}_i^t\|^2] \leq \frac{\eta_t^2}{N}\sum_{i=1}^{N}\mathbb{E}[\|\mathbf{v}_i^t\|^2] \leq \frac{4}{\mu^2 t^2}o\left(\frac{1}{t}\right) = o\left(\frac{1}{t^3}\right). \tag{27}$$

Moreover, by Lemma 8 and $f$ being $L$-smooth we have $\mu \preceq \mathbf{A} \preceq L$. It follows

$$1-\eta_t L \preceq I - \eta_t \mathbf{A} \preceq 1 - \eta_t \mu.$$

Since $\eta_t \leq 1/L$ and $I-\eta_t\mathbf{A}$ is symmetric, we have $\|I-\eta_t\mathbf{A}\| \leq 1-\eta_t\mu \leq 1$. Then,

$$\|(I-\eta_t\mathbf{A})(\bar{\mathbf{x}}^t - \mathbf{x}^*)\| \leq \|I-\eta_t\mathbf{A}\|\|\bar{\mathbf{x}}^t-\mathbf{x}^*\| \leq \|\bar{\mathbf{x}}^t-\mathbf{x}^*\|.$$

Thus,

$$\begin{aligned}
\mathbb{E}[\|(I-\eta_t\mathbf{A})(\bar{\mathbf{x}}^t-\mathbf{x}^*)\|\|\frac{\eta_t}{N}\sum_{i=1}^{N}\mathbf{v}_i^t\|] &\leq \mathbb{E}[\|\bar{\mathbf{x}}^t-\mathbf{x}^*\|\|\frac{\eta_t}{N}\sum_{i=1}^{N}\mathbf{v}_i^t\|] \\
&\leq \mathbb{E}\left[\left(\frac{1}{N}\sum_{i=1}^{N}\|\mathbf{x}_i^t-\mathbf{x}^*\|\right)\left(\frac{\eta_t}{N}\sum_{i=1}^{N}\|\mathbf{v}_i^t\|\right)\right] \\
&\leq \frac{\eta_t}{N^2}\sum_{i=1}^{N}\mathbb{E}[\|\mathbf{x}_i^t-\mathbf{x}^*\|\|\mathbf{v}_i^t\|] + \frac{\eta_t}{N^2}\sum_{i\neq j}\mathbb{E}[\|\mathbf{x}_j^t-\mathbf{x}^*\|\|\mathbf{v}_i^t\|] \\
&= \frac{\eta_t}{N^2}\sum_{i=1}^{N}\mathbb{E}[\|\mathbf{x}_i^t-\mathbf{x}^*\|\|\mathbf{v}_i^t\|] + \frac{\eta_t}{N^2}\sum_{i\neq j}\mathbb{E}[\|\mathbf{x}_j^t-\mathbf{x}^*\|]\mathbb{E}[\|\mathbf{v}_i^t\|] \\
&\leq \frac{2}{\mu t}o\left(\frac{1}{t}\right) + \frac{2}{\mu t}o\left(\frac{1}{\sqrt{t}}\right)o\left(\frac{1}{\sqrt{t}}\right) = o\left(\frac{1}{t^2}\right),
\end{aligned} \tag{28}$$

where we used $|E[X]| \leq \sqrt{E[X^2]}$ for random variables $\|\mathbf{x}_j^t-\mathbf{x}^*\|$ and $\mathbf{v}_i^t$ and Lemma 10 in last equation above. plugging (27) and (28) in (26) we obtain the desired result. □

Now we are ready to present the proof of Theorem 2.

*Proof of Theorem 2.* Denote $\psi^t := \mathbb{E}[\|\bar{\mathbf{x}}^t-\mathbf{x}^*\|]$ for $t \geq 0$. By Lemma 12 we can write,

$$\psi^{t+1} \leq \psi^t(1-\eta_t\mu)^2 + \frac{\eta_t^2\sigma^2}{N} + \nu^t,$$

where $\nu^t \geq 0$ and $\nu^t = o(1/t^2)$. It follows,

$$\psi^k \leq \underbrace{\psi^0\prod_{t=0}^{k-1}(1-\eta_t\mu)^2}_{S_1} + \underbrace{\sum_{t=0}^{k-1}\frac{\eta_t^2\sigma^2}{N}\prod_{l=t+1}^{k-1}(1-\eta_l\mu)^2}_{S_2} + \underbrace{\sum_{t=0}^{k-1}\nu^t\prod_{l=t+1}^{k-1}(1-\eta_l\mu)^2}_{S_3}. \qquad \forall k \geq t_0. \tag{29}$$

Next, we will bound each of the terms $S_1$, $S_2$, and $S_3$. Before that, we note that for $t \geq t_0$,

$$1-\eta_t\mu = 1 - \frac{2t}{(t+1)^2} \leq 1 - \frac{2}{t}.$$

Therefore, for $t_2 > t_1 \geq t_0$ we have,

$$\begin{aligned}
\prod_{l=t_1}^{t_2-1}(1-\eta_l\mu) \leq \prod_{l=t_1}^{t_2-1}(1-\frac{2}{l}) &= \exp\left(\sum_{l=t_1}^{t_2-1}\log(1-\frac{2}{l})\right) \\
&\leq \exp\left(\sum_{l=t_1}^{t_2-1}\frac{-2}{l}\right) \leq \exp\left(2\log(t_1) - 2\log(t_2)\right) = \left(\frac{t_1}{t_2}\right)^2.
\end{aligned} \tag{30}$$

Now we have the tools we need to bound $S_1$, $S_2$, and $S_3$. we have,

$$S_1 = \|\bar{\mathbf{x}}^0 - \mathbf{x}^*\|^2 \prod_{t=0}^{t_0-1}(1-\eta_t\mu)^2 \prod_{t=t_0}^{k-1}(1-\eta_t\mu)^2 \leq (1-\frac{\mu}{L})^{2t_0}\left(\frac{t_0}{k}\right)^4 \|\bar{\mathbf{x}}^0 - \mathbf{x}^*\|^2. \tag{31}$$

$$\begin{aligned}
S_2 &= \sum_{t=0}^{t_0-1}\frac{\eta_t^2\sigma^2}{N}\prod_{l=t+1}^{t_0-1}(1-\eta_l\mu)^2\prod_{l=t_0}^{k-1}(1-\eta_l\mu)^2 + \sum_{t=t_0}^{k-1}\frac{\eta_t^2\sigma^2}{N}\prod_{l=t+1}^{k-1}(1-\eta_l\mu)^2 \\
&\leq \frac{\sigma^2}{N}\left[\sum_{t=0}^{t_0-1}\frac{1}{L^2}(1-\frac{\mu}{L})^{2(t_0-1-t)}\left(\frac{t_0}{k}\right)^4 + \sum_{t=t_0}^{k-1}\left(\frac{2t}{\mu(t+1)^2}\right)^2\left(\frac{t+1}{k}\right)^4\right] \\
&= \frac{\sigma^2}{N}\left[\frac{t_0^4}{L^2k^4}\sum_{t=0}^{t_0-1}(1-\frac{\mu}{L})^{2t} + \frac{4}{\mu^2k^4}\sum_{t=t_0}^{k-1}t^2\right] \\
&\leq \frac{\sigma^2}{N}\left[\frac{t_0^4}{L^2k^4}\sum_{t=0}^{\infty}(1-\frac{\mu}{L})^{2t} + \frac{4}{\mu^2k^4}\sum_{t=1}^{k-1}t^2\right] \\
&= \frac{\sigma^2}{N}\left[\frac{t_0^4}{L^2k^4(1-(1-\frac{\mu}{L})^2)} + \frac{2k(k-1)(2k-1)}{3\mu^2k^4}\right] \\
&\leq \frac{\sigma^2}{N}\left[\frac{t_0^4}{L\mu k^4} + \frac{4}{3\mu^2k}\right] = \frac{4\sigma^2}{3N\mu^2k}\left[1+\frac{3\mu t_0^4}{4Lk^3}\right].
\end{aligned} \tag{32}$$

Next, we show $S_3 = o(1/k)$. Since $\nu^t = o(1/t^2)$, without loss of generality, we can assume there exists $B_1, B_2 > 0$ such that for any $\epsilon > 0$, there exists $k_1 \geq t_0$ such that,

$$\nu^t \leq \begin{cases} \frac{B_1}{(t+1)^2}, & t \geq 0, \\ \frac{\epsilon B_2}{(t+1)^2}, & t \geq k_1. \end{cases}$$

It follows,

$$\begin{aligned}
S_3 &\leq \sum_{t=0}^{t_0-1}(1-\frac{\mu}{L})^{2(t_0-1-t)}\left(\frac{t_0}{k}\right)^4\frac{B_1}{(t+1)^2} + \sum_{t=t_0}^{k_1-1}\left(\frac{t+1}{k}\right)^4\frac{B_1}{(t+1)^2} + \sum_{t=k_1}^{k-1}\left(\frac{t+1}{k}\right)^4\frac{\epsilon B_2}{(t+1)^2} \\
&\leq \sum_{t=0}^{\infty}\left(\frac{t_0}{k}\right)^4\frac{B_1}{(t+1)^2} + \sum_{t=t_0}^{k_1-1}\frac{B_1}{k^2} + \sum_{t=k_1}^{k-1}\frac{\epsilon B_2}{k^2} \\
&\leq \frac{2t_0^4B_1}{k^4} + \frac{B_1(k_1-t_0)}{k^2} + \frac{\epsilon B_2(k-k_1)}{k^2} \\
&\leq \frac{2\epsilon t_0B_1}{k} + \frac{\epsilon B_1}{k} + \frac{\epsilon B_2}{k} = \frac{\epsilon(B_1(2t_0+1)+B_2)}{k}, \qquad \text{for } k \geq \left\lceil\frac{k_1}{\epsilon}\right\rceil.
\end{aligned}$$

Thus,

$$S_3 = o\left(\frac{1}{k}\right). \tag{33}$$

Plugging (31)-(33) in (29) results,

$$\begin{aligned}
\mathbb{E}[\|\bar{\mathbf{x}}^k - \mathbf{x}^*\|^2] &\leq \frac{(1-\frac{\mu}{L})^{2t_0}t_0^4}{k^4}\|\bar{\mathbf{x}}^0 - \mathbf{x}^*\|^2 + \frac{4\sigma^2}{3N\mu^2k} + \frac{\sigma^2t_0^4}{NL\mu k^4} + o\left(\frac{1}{k}\right) \\
&= \frac{4\sigma^2}{3N\mu^2k} + o\left(\frac{1}{k}\right).
\end{aligned}$$

□