A Proof of Lemma 1

Proof For any given vector $g \in \mathbb{R}^d$, the ratio $|g_i|/\|g\|_\infty$ lies in an interval of the form $[l_i/s, (l_i+1)/s]$ where $l_i \in \{0, 1, \ldots, s-1\}$. Hence, for that specific l_i, the following inequalities

$$\frac{l_i}{s} \leq \frac{|g_i|}{\|g\|_\infty} \leq \frac{l_i+1}{s} \tag{14}$$

are satisfied. Moreover, based on the probability distribution of b_i we know that

$$\frac{l_i}{s} \leq b_i \leq \frac{l_i+1}{s}. \tag{15}$$

Therefore, based on the inequalities in (14) and (15) we can write

$$-\frac{1}{s} \leq \frac{|g_i|}{\|g\|_\infty} - b_i \leq \frac{1}{s} \tag{16}$$

Hence, we can show that the variance of s-`Partition Encoding Scheme` is upper bounded by

$$\begin{aligned}
\mathrm{Var}[\phi'(g)|g] &= \mathbb{E}[\|\phi'(g) - g\|^2|g] \\
&= \sum_{i=1}^{d} \mathbb{E}[(g_i - \mathrm{sgn}(g_i) b_i \|g\|_\infty)^2|g] \\
&= \sum_{i=1}^{d} \mathbb{E}[(|g_i| - b_i\|g\|_\infty)^2|g] \\
&= \sum_{i=1}^{d} \|g\|_\infty^2 \mathbb{E}\left[\left(\frac{|g_i|}{\|g\|_\infty} - b_i\right)^2 \mid g\right] \\
&\leq \frac{d}{s^2}\|g\|_\infty^2,
\end{aligned} \tag{17}$$

where the inequality holds due to (16). ∎

B Proof of Theorem 1 and Corollary 1

The key to the proofs of Theorem 1 is to upper bound the difference between the true gradient $\nabla f(x_t) = \nabla f(x_{i,k})$ and the estimated gradient $\bar{g}_{i,k}$. Intuitively, if the error is small enough, then we can approximate $\nabla f(x_{i,k})$ by $\bar{g}_{i,k}$. Thus the algorithm fed with the estimated gradient $\bar{g}_{i,k}$ will still converge.

So we first address the bound of $\|\nabla f(x_{i,k}) - \bar{g}_{i,k}\|$, which is resolved in the following lemma.

Lemma 4 *Under the condition of Theorem 1, we have*

$$\mathbb{E}[\|\nabla f(x_{i,k}) - \bar{g}_{i,k}\|^2] \leq \frac{2(G_\infty^2 + 2L^2D^2 + 4L_\infty^2 D^2)}{p_i^2}. \tag{18}$$

Proof We first define a few auxiliary variables. On each worker m, we define the average function of its n component functions as $f^{(m)}(x) = \frac{\sum_{j=1}^{n} f_{m,j}(x)}{n}$, so $f(x) = \frac{\sum_{m=1}^{M} f^{(m)}(x)}{M}$. We also define

$$g_{i,k}^{(m)} = \begin{cases} g_{i,k}^m & k = 1 \\ g_{i,k-1}^{(m)} + g_{i,k}^m = \sum_{j=1}^{k} g_{i,j}^m & k \geq 2, \end{cases}$$

where $g_{i,k}^m$ is defined in Algorithm 1. Then $g_{i,k}^{(m)}$ is an unbiased estimator of $\nabla f^{(m)}(x_{i,k})$. We define the average of $g_{i,k}^{(m)}$ as

$$g_{i,k} = \frac{\sum_{m=1}^{M} g_{i,k}^{(m)}}{M}.$$

We also define $\mathcal{F}_{i,k}$ to be the σ-field generated by all the randomness before round (i,k), *i.e,* round $t = \sum_{j=1}^{i-1} p_j + k$. We note that given $\mathcal{F}_{i,k}$, $x_{i,k}$ is actually determined, and we can verify that $\mathbb{E}[g_{i,k}|\mathcal{F}_{i,k}] = \nabla f(x_{i,k})$, and $\mathbb{E}[\bar{g}_{i,k}|\mathcal{F}_{i,k}, g_{i,k}] = g_{i,k}$, for all (i,k). Here, with abuse of notation, $\mathbb{E}[\cdot|g_{i,k}]$ is the conditional expectation given not only the value of $g_{i,k}$, but also the sampled gradients $\nabla f_{m,j}(x_{i,k}), \nabla f_{m,j}(x_{i,k-1})$(if defined) for all $j \in \mathcal{S}_{i,k}^m, m \in [M]$.

Then by law of total expectation, we have

$$\begin{aligned}
\mathbb{E}[\|\nabla f(x_{i,k}) - \bar{g}_{i,k}\|^2] &= \mathbb{E}[\mathbb{E}[\|\nabla f(x_{i,k}) - \bar{g}_{i,k}\|^2|\mathcal{F}_{i,k}]] \\
&= \mathbb{E}[\mathbb{E}[\|\nabla f(x_{i,k}) - g_{i,k} + g_{i,k} - \bar{g}_{i,k}\|^2|\mathcal{F}_{i,k}]] \\
&= \mathbb{E}[\mathbb{E}[\|\nabla f(x_{i,k}) - g_{i,k}\|^2|\mathcal{F}_{t-1}]] + \mathbb{E}[\mathbb{E}[\|g_{i,k} - \bar{g}_{i,k}\|^2|\mathcal{F}_{i,k}]] \\
&\quad + 2\mathbb{E}[\mathbb{E}[\langle \nabla f(x_{i,k}) - g_{i,k}, g_{i,k} - \bar{g}_{i,k}\rangle|\mathcal{F}_{i,k}]] \\
&= \mathbb{E}[\|\nabla f(x_{i,k}) - g_{i,k}\|^2] + \mathbb{E}[\|g_{i,k} - \bar{g}_{i,k}\|^2],
\end{aligned} \tag{19}$$

where the last equation holds since

$$\begin{aligned}
\mathbb{E}[\langle \nabla f(x_{i,k}) - g_{i,k}, g_{i,k} - \bar{g}_{i,k}\rangle|\mathcal{F}_{i,k}] &= \mathbb{E}[\mathbb{E}[\langle \nabla f(x_{i,k}) - g_{i,k}, g_{i,k} - \bar{g}_{i,k}\rangle|\mathcal{F}_{i,k}, g_{i,k}]|\mathcal{F}_{i,k}] \\
&= \mathbb{E}[\langle \nabla f(x_{i,k}) - g_{i,k}, \mathbb{E}[g_{i,k} - \bar{g}_{i,k}|\mathcal{F}_{i,k}, g_{i,k}]\rangle|\mathcal{F}_{i,k}] \\
&= 0.
\end{aligned}$$

Now we turn to bound $\mathbb{E}[\|\nabla f(x_{i,k}) - g_{i,k}\|^2]$. In fact, we have

$$\begin{aligned}
\mathbb{E}[\|\nabla f(x_{i,k}) - g_{i,k}\|^2] &= \mathbb{E}[\|\frac{\sum_{m=1}^M \nabla f^{(m)}(x_{i,k})}{M} - \frac{\sum_{m=1}^M g_{i,k}^{(m)}}{M}\|^2] \\
&= \frac{\sum_{m=1}^M \mathbb{E}[\|\nabla f^{(m)}(x_{i,k}) - g_{i,k}^{(m)}\|^2]}{M^2}.
\end{aligned} \tag{20}$$

For $k \geq 2$, we have

$$\begin{aligned}
&\mathbb{E}[\|\nabla f^{(m)}(x_{i,k}) - g_{i,k}^{(m)}\|^2] \\
=& \mathbb{E}[\mathbb{E}[\|[\nabla f^{(m)}(x_{i,k}) - \nabla f^{(m)}(x_{i,k-1})] - g_{i,k}^m\|^2|\mathcal{F}_{i,k}]] + \mathbb{E}[\mathbb{E}[\|\nabla f^{(m)}(x_{i,k-1}) - g_{i,k-1}^{(m)}\|^2|\mathcal{F}_{i,k}]] \\
=& \mathbb{E}[\mathrm{Var}[g_{i,k}^m]|\mathcal{F}_{i,k}]] + \mathbb{E}[\|\nabla f^{(m)}(x_{i,k-1}) - g_{i,k-1}^{(m)}\|^2] \\
=& \mathbb{E}[\mathrm{Var}[\frac{\sum_{j\in\mathcal{S}_{i,k}^m} \nabla f_j(x_{i,k}) - \nabla f_j(x_{i,k-1})}{S_{i,k}}|\mathcal{F}_{i,k}]] + \mathbb{E}[\|\nabla f^{(m)}(x_{i,k-1}) - g_{i,k-1}^{(m)}\|^2] \\
=& \mathbb{E}[\frac{\sum_{j\in\mathcal{S}_{i,k}^m} \mathrm{Var}[\nabla f_j(x_{i,k}) - \nabla f_j(x_{i,k-1})|\mathcal{F}_{i,k}]}{[S_{i,k}]^2}] + \mathbb{E}[\|\nabla f^{(m)}(x_{i,k-1}) - g_{i,k-1}^{(m)}\|^2] \\
\leq& \mathbb{E}[\frac{\sum_{j\in\mathcal{S}_{i,k}^m} \mathbb{E}[\|\nabla f_j(x_{i,k}) - \nabla f_j(x_{i,k-1})\|^2|\mathcal{F}_{i,k}]}{[S_{i,k}]^2})] + \mathbb{E}[\|\nabla f^{(m)}(x_{i,k-1}) - g_{i,k-1}^{(m)}\|^2] \\
\leq& \frac{1}{S_{i,k}}(LD\eta_{i,k-1})^2 + \mathbb{E}[\|\nabla f(x_{i,k-1}) - g_{i,k-1}\|^2] \\
=& \frac{L^2 D^2 \eta_{i,k-1}^2}{S_{i,k}} + \mathbb{E}[\|\nabla f^{(m)}(x_{i,k-1}) - g_{i,k-1}^{(m)}\|^2].
\end{aligned}$$

For $k = 1$, we have $g_{i,1}^{(m)} = \nabla f^{(m)}(x_{i,1})$. So

$$\mathbb{E}[\|\nabla f^{(m)}(x_{i,k}) - g_{i,k}^{(m)}\|^2] \leq L^2 D^2 \sum_{j=2}^k \frac{\eta_{i,j-1}^2}{S_{i,j}} = \frac{L^2 D^2 M}{p_i} \sum_{j=2}^k \eta_{i,j-1}^2.$$

Since

$$\sum_{j=2}^k \eta_{i,j-1}^2 = \sum_{j=2}^k \frac{4}{(p_i + j - 1)^2} \leq \sum_{j=2}^k \frac{4}{p_i^2} \leq \frac{4}{p_i},$$

we have

$$\mathbb{E}[\|\nabla f^{(m)}(x_{i,k}) - g_{i,k}^{(m)}\|^2] \leq \frac{4ML^2D^2}{p_i^2}. \tag{21}$$

Combine with Equation (20), we have

$$\mathbb{E}[\|\nabla f(x_{i,k}) - g_{i,k}\|^2] \leq \frac{M \cdot 4ML^2D^2}{M^2 \cdot p_i^2} = \frac{4L^2D^2}{p_i^2}. \tag{22}$$

Now we only need to bound $\mathbb{E}[\|g_{i,k} - \bar{g}_{i,k}\|^2]$. For $k \geq 2$, we have

$$\begin{aligned}
&\mathbb{E}[\|g_{i,k} - \bar{g}_{i,k}\|^2] \\
=& \mathbb{E}[\mathbb{E}[\|\frac{\sum_{m=1}^M g_{i,k}^m}{M} + g_{i,k-1} - \phi'_{2,i,k}(\tilde{g}_{i,k}) - \bar{g}_{i,k-1}\|^2|\mathcal{F}_{i,k}, g_{i,k}]] \\
=& \mathbb{E}[\mathbb{E}[\|\frac{\sum_{m=1}^M g_{i,k}^m}{M} - \phi'_{2,i,k}(\tilde{g}_{i,k})\|^2|\mathcal{F}_{i,k}, g_{i,k}]] + \mathbb{E}[\|g_{i,k-1} - \bar{g}_{i,k-1}\|^2] \\
&+ 2\mathbb{E}[\mathbb{E}[\frac{\sum_{m=1}^M g_{i,k}^m}{M} - \phi'_{2,i,k}(\tilde{g}_{i,k}), g_{i,k-1} - \bar{g}_{i,k-1}\rangle|\mathcal{F}_{i,k}, g_{i,k-1}]].
\end{aligned}$$

Moreover

$$\begin{aligned}
\mathbb{E}[\phi'_{2,i,k}(\tilde{g}_{i,k})|\mathcal{F}_{i,k}, g_{i,k}] &= \mathbb{E}[\tilde{g}_{i,k}|\mathcal{F}_{i,k}, g_{i,k}] \\
&= \mathbb{E}[\sum_{m=1}^M \phi'_{1,i,k}(g_{i,k}^m)/M|\mathcal{F}_{i,k}, g_{i,k}] \\
&= \frac{\sum_{m=1}^M g_{i,k}^m}{M},
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{E}[\mathbb{E}[\|\frac{\sum_{m=1}^M g_{i,k}^m}{M} - \phi'_{2,i,k}(\tilde{g}_{i,k})\|^2|\mathcal{F}_{i,k}, g_{i,k}]] \\
=& \mathbb{E}[\mathbb{E}[\|\frac{\sum_{m=1}^M g_{i,k}^m}{M} - \tilde{g}_{i,k} + \tilde{g}_{i,k} - \phi'_{2,i,k}(\tilde{g}_{i,k})\|^2|\mathcal{F}_{i,k}, g_{i,k}]] \\
=& \mathbb{E}[\mathbb{E}[\|\frac{\sum_{m=1}^M g_{i,k}^m}{M} - \sum_{m=1}^M \phi'_{1,i,k}(g_{i,k}^m)/M\|^2|\mathcal{F}_{i,k}, g_{i,k}]] + \mathbb{E}[\mathbb{E}[\|\tilde{g}_{i,k} - \phi'_{2,i,k}(\tilde{g}_{i,k})\|^2|\mathcal{F}_{i,k}, g_{i,k}, \tilde{g}_{i,k}]] \\
\leq& \frac{1}{M}\frac{d}{s_{1,i,k}^2}(\eta_{i,k-1}LD)^2 + \frac{d}{s_{2,i,k}^2}(\eta_{i,k-1}LD)^2 \\
=& \frac{\eta_{i,k-1}^2 dL^2D^2}{Ms_{1,i,k}^2} + \frac{\eta_{i,k-1}^2 dL^2D^2}{s_{2,i,k}^2},
\end{aligned}$$

where in the inequality, we apply Lemma 1 with $\|g_{i,k}^m\|_\infty = \|\nabla f_{\mathcal{S}_{i,k}^m}(x_{i,k}) - \nabla f_{\mathcal{S}_{i,k}^m}(x_{i,k-1})\|_\infty \leq \|\nabla f_{\mathcal{S}_{i,k}^m}(x_{i,k}) - \nabla f_{\mathcal{S}_{i,k}^m}(x_{i,k-1})\|_2 \leq \eta_{i,k-1}LD$ and $\|\tilde{g}_{i,k}\|_\infty = \|\sum_{m=1}^M \phi'_{1,i,k}(g_{i,k}^m)/M\|_\infty \leq \eta_{i,k-1}LD$. Now for $k \geq 2$ we have,

$$\mathbb{E}[\|g_{i,k} - \bar{g}_{i,k}\|^2] \leq \frac{\eta_{i,k-1}^2 dL^2D^2}{Ms_{1,i,k}^2} + \frac{\eta_{i,k-1}^2 dL^2D^2}{s_{2,i,k}^2} + \mathbb{E}[\|g_{i,k-1} - \bar{g}_{i,k-1}\|^2].$$

If $k=1$, we have

$$\begin{aligned}
\mathbb{E}[\|g_{i,k}-\bar{g}_{i,k}\|^2] &= \mathbb{E}[\|\nabla f(x_{i,k})-\tilde{g}_{i,k}+\tilde{g}_{i,k}-\phi'_{2,i,k}(\tilde{g}_{i,k})\|^2] \\
&= \mathbb{E}[\mathbb{E}[\|\nabla f(x_{i,k})-\frac{\sum_{m=1}^{M}\phi'_{1,i,k}(\nabla f^{(m)}(x_{i,k}))}{M}\|^2|\mathcal{F}_{i,k},g_{i,k}]] \\
&\quad +\mathbb{E}[\mathbb{E}[\|\tilde{g}_{i,k}-\phi'_{2,i,k}(\tilde{g}_{i,k})\|^2|\mathcal{F}_{i,k},g_{i,k},\tilde{g}_{i,k}]] \\
&\leq \frac{1}{M^2}\mathbb{E}[\sum_{m=1}^{M}\mathbb{E}[\|\nabla f^{(m)}(x_{i,k})-\phi'_{1,t}(\nabla f^{(m)}(x_{i,k}))\|^2|\mathcal{F}_{i,k},g_{i,k}]]+\frac{d}{s_{2,i,k}^2}G_\infty^2 \\
&\leq \frac{dG_\infty^2}{Ms_{1,i,k}^2}+\frac{dG_\infty^2}{s_{2,i,k}^2},
\end{aligned}$$

where in the inequality, we apply Lemma 1 with $\|\nabla f^{(m)}(x_k)\|_\infty \leq G_\infty$ and $\|\tilde{g}_{i,k}\|_\infty = \|\frac{\sum_{m=1}^{M}\phi'_{1,i,k}(\nabla f^{(m)}(x_k))}{M}\|_\infty \leq G_\infty$. Then we have

$$\begin{aligned}
\mathbb{E}[\|g_{i,k}-\bar{g}_{i,k}\|^2] &\leq \sum_{j=2}^{k}\frac{\eta_{i,j-1}^2 dL^2D^2}{Ms_{1,i,j}^2}+\sum_{j=2}^{k}\frac{\eta_{i,j-1}^2 dL^2D^2}{s_{2,i,j}^2}+\frac{dG_\infty^2}{Ms_{1,i,1}^2}+\frac{dG_\infty^2}{s_{2,i,1}^2} \\
&\leq \frac{dL^2D^2}{Ms_{1,i}^2}\sum_{j=2}^{k}\eta_{i,j-1}^2+\frac{dL^2D^2}{s_{2,i}^2}\sum_{j=2}^{k}\eta_{i,j-1}^2+\frac{dG_\infty^2}{Ms_{1,i,1}^2}+\frac{dG_\infty^2}{s_{2,i,1}^2} \\
&\leq \frac{dL^2D^2}{M\frac{p_i d}{M}}\frac{4}{p_i}+\frac{dL^2D^2}{p_i d}\frac{4}{p_i}+\frac{dG_\infty^2}{M\frac{dp_i^2}{M}}+\frac{dG_\infty^2}{dp_i^2} \\
&= \frac{2G_\infty^2+8L^2D^2}{p_i^2}.
\end{aligned} \tag{23}$$

Now combine Equations (19), (22) and (23), we have

$$\mathbb{E}[\|\nabla f(x_{i,k})-\bar{g}_{i,k}\|^2] \leq \frac{2(G_\infty^2+6L^2D^2)}{p_i^2} \triangleq \frac{C_1^2}{p_i^2}.$$

■

Now we turn to prove Theorem 1. First, since $x_{t+1}=(1-\eta_{i,k})x_t+\eta_{i,k}v_{i,k}$ is a convex combination of $x_t, v_{i,k}$, and $x_1\in\mathcal{K}, v_{i,k}\in\mathcal{K}$, for all t, we can prove $x_t\in\mathcal{K}$, for all t by induction. So $x_{T+1}\in\mathcal{K}$. Then we need the following lemma.

Lemma 5 (Proof of Theorem 1 in Yurtsever et al. [2019]) *Consider Algorithm 1, under the conditions of Theorem 1, we have*

$$\mathbb{E}[f_{i,k+1}]-f(x^*) \leq (1-\eta_{i,k})(\mathbb{E}[f(x_{i,k})-f(x^*)])+\eta_{i,k}D\mathbb{E}[\|\nabla f(x_i,k)-\bar{g}_{i,k}\|]+\eta_{i,k}^2\frac{LD^2}{2}.$$

Moreover, by analyzing the telescopic sum of the inequality over (i,k), we have

$$\mathbb{E}[f(x_{i,k+1})]-f(x^*) \leq \sum_{(\tau,j)}\left(\eta_{\tau,j}D\mathbb{E}[\|\nabla f(x_{\tau,j})-\bar{g}_{\tau,j}\|]+\eta_{\tau,j}^2\frac{LD^2}{2}\right)\frac{(p_\tau+j-2)(p_\tau+j-1)}{(p_i+k-1)(p_i+k)}.$$

By Lemma 4 and Jensen' inequality, we have

$$\mathbb{E}[\|\nabla f(x_i,k)-\bar{g}_{i,k}\|] \leq \sqrt{\mathbb{E}[\|\nabla f(x_i,k)-\bar{g}_{i,k}\|^2]} \leq \frac{C_1}{p_i}.$$

So

$$\begin{aligned}
&\sum_{(\tau,j)} \eta_{\tau,j} D\mathbb{E}[\|\nabla f(x_i,k) - \bar{g}_{i,k}\|] \frac{(p_\tau+j-2)(p_\tau+j-1)}{(p_i+k-1)(p_i+k)} \\
\leq &\sum_{(\tau,j)} \frac{2}{p_\tau+j} D \frac{C_1}{p_\tau} \frac{(p_\tau+j-2)(p_\tau+j-1)}{(p_i+k-1)(p_i+k)} \\
\leq &\frac{4C_1 D}{(p_i+k-1)(p_i+k)} \sum_{(\tau,j)} 1 \\
\leq &\frac{4C_1 D}{p_i+k}.
\end{aligned}$$

We also have

$$\begin{aligned}
\sum_{(\tau,j)} \eta_{\tau,j}^2 \frac{LD^2}{2} \frac{(p_\tau+j-2)(p_\tau+j-1)}{(p_i+k-1)(p_i+k)} &= \sum_{(\tau,j)} \frac{4}{(p_\tau+j)^2} \frac{LD^2}{2} \frac{(p_\tau+j-2)(p_\tau+j-1)}{(p_i+k-1)(p_i+k)} \\
&\leq \frac{2LD^2}{(p_i+k-1)(p_i+k)} \sum_{(\tau,j)} 1 \\
&\leq \frac{2LD^2}{p_i+k}.
\end{aligned}$$

Thus by Lemma 5, we have

$$\mathbb{E}[f(x_{i,k+1})] - f(x^*) \leq \frac{4C_1 D + 2LD^2}{p_i+k}.$$

By definition, $x_{i,k+1} = x_t$, where $t = \sum_{j=1}^{i-1} p_j + k + 1 = p_i + k$. When $t = T$, we have

$$\mathbb{E}[f(x_T)] - f(x^*) \leq \frac{4C_1 D + 2LD^2}{T}.$$

Therefore, to obtain an ϵ-suboptimal solution, we need $\mathcal{O}(1/\epsilon)$ iterations. Let $T = \sum_{j=1}^{I-1} p_i + K = p_I + K - 1$, then $I \leq \log_2(T) + 1$, and thus IFO complexity per worker is

$$\begin{aligned}
IFO &\leq \sum_{i=1}^{I} (n + \sum_{j=2}^{p_i} S_{i,k}) \leq \sum_{i=1}^{I} (n + \frac{2^{2(i-1)}}{M}) \leq nI + 2^{2I}/M \leq [\log_2(T)+1]N/M + 4T^2/M \\
&= \mathcal{O}(\frac{N\ln(1/\epsilon) + 1/\epsilon^2}{M}).
\end{aligned}$$

C Proof of Theorem 2 and Corollary 2

The proof is quite similar to that of Theorem 1.

We first need to upper bound $\mathbb{E}[\|\nabla f(x_{i,k}) - \bar{g}_{i,k}\|^2]$. Equations (19) and (23) still hold. Similarly, we also have for $k \geq 2$,

$$\begin{aligned}
\mathbb{E}[\|f(x_{i,k}) - g_{i,k}\|^2] &\leq \frac{L^2 D^2 \eta_{i,k-1}^2}{MS_{i,k}} + \mathbb{E}[\|f(x_{i,k-1}) - g_{i,k-1}\|^2 \\
&= \frac{L^2 D^2 \eta_{i,k-1}^2}{p_i} + \mathbb{E}[\|f(x_{i,k-1}) - g_{i,k-1}\|^2
\end{aligned}$$

For $k = 1$,

$$\mathbb{E}[\|f(x_{i,k}) - g_{i,k}\|^2] \leq \frac{\sigma^2}{MS_{i,1}} = \frac{\sigma^2}{M\frac{\sigma^2 p_i^2}{ML^2D^2}} = \frac{L^2D^2}{p_i^2}$$

So

$$\mathbb{E}[\|f(x_{i,k}) - g_{i,k}\|^2] \le \frac{L^2D^2}{p_i^2} + \frac{L^2D^2}{p_i}\sum_{j=2}^{k}\eta_{i,j-1}^2 \le \frac{L^2D^2}{p_i^2} + \frac{4L^2D^2}{p_i^2} = \frac{5L^2D^2}{p_i^2}. \tag{24}$$

Combine Equations (19), (23) and (24), we have

$$\mathbb{E}[\|f(x_{i,k}) - \bar{g}_{i,k}\|^2] \le \frac{13L^2D^2 + 2G_\infty^2}{p_i^2} \triangleq \frac{C_2^2}{p_i^2}.$$

Applying Lemma 5, we have

$$\mathbb{E}[f(x_{i,k+1})] - f(x^*) \le \frac{4C_2D + 2LD^2}{p_i + k}.$$

By definition, $x_{i,k+1} = x_t$, where $t = \sum_{j=1}^{i-1} p_j + k + 1 = p_i + k$. When $t = T$, we have

$$\mathbb{E}[f(x_T)] - f(x^*) \le \frac{4C_2D + 2LD^2}{T}.$$

Therefore, to obtain an ϵ-suboptimal solution, we need $\mathcal{O}(1/\epsilon)$ iterations. Let $T = \sum_{j=1}^{I-1} p_i + K = p_I + K - 1$, then $I \le \log_2(T) + 1$, and thus SFO complexity per worker is

$$\begin{aligned}
SFO &\le \sum_{i=1}^{I}\left(\frac{\sigma^2 p_i^2}{ML^2D^2} + \sum_{j=2}^{p_i} S_{i,k}\right) \le \sum_{i=1}^{I}\left(\frac{\sigma^2 2^{2(i-1)}}{ML^2D^2} + \frac{2^{2(i-1)}}{M}\right)\\
&\le \frac{2^{2I}}{M}\left(\frac{\sigma^2}{L^2D^2} + 1\right) \le \frac{4T^2}{M}\left(\frac{\sigma^2}{L^2D^2} + 1\right)\\
&= \mathcal{O}(1/(M\epsilon^2)).
\end{aligned}$$

D Proof of Theorem 3 and Corollary 3

First, since $x_{t+1} = (1-\eta_t)x_t + \eta_t v_t$ is a convex combination of x_t, v_t, and $x_1 \in \mathcal{K}, v_t \in \mathcal{K}$, for all t, we can prove $x_t \in \mathcal{K}$, for all t by induction. So $x_o \in \mathcal{K}$.

Then we turn to upper bound $\mathbb{E}[\|\nabla f(x_{i,k}) - \bar{g}_{i,k}\|^2]$. Equation (19) still holds. Similarly, we also have for $k \ge 2$,

$$\begin{aligned}
\mathbb{E}[\|f^{(m)}(x_{i,k}) - g_{i,k}^{(m)}\|^2] &\le \frac{L^2D^2\eta_{i,k-1}^2}{S_{i,k}} + \mathbb{E}[\|f^{(m)}(x_{i,k-1}) - g_{i,k-1}^{(m)}\|^2\\
&= \frac{L^2D^2T^{-1}}{\frac{\sqrt{n}}{M}} + \mathbb{E}[\|f^{(m)}(x_{i,k-1}) - g_{i,k-1}^{(m)}\|^2\\
&= \frac{ML^2D^2}{\sqrt{n}T} + \mathbb{E}[\|f^{(m)}(x_{i,k-1}) - g_{i,k-1}^{(m)}\|^2
\end{aligned}$$

For $k = 1$, we have $g_{i,1}^{(m)} = \nabla f^{(m)}(x_{i,1})$. So

$$\mathbb{E}[\|\nabla f^{(m)}(x_{i,k}) - g_{i,k}^{(m)}\|^2] \le \frac{ML^2D^2}{\sqrt{n}T}(k-1) \le \frac{ML^2D^2}{\sqrt{n}T}p_i = \frac{ML^2D^2}{T}.$$

By Equation (20),

$$\mathbb{E}[\nabla f(x_{i,k}) - g_{i,k}\|^2] \le \frac{M\frac{ML^2D^2}{T}}{M^2} = \frac{L^2D^2}{T}. \tag{25}$$

we also have

$$\begin{aligned}
\mathbb{E}[\|g_{i,k} - \bar{g}_{i,k}\|^2] &\le \sum_{j=2}^{k}\frac{\eta_{i,j-1}^2 dL^2D^2}{Ms_{1,i,j}^2} + \sum_{j=2}^{k}\frac{\eta_{i,j-1}^2 dL^2D^2}{s_{2,i,j}^2} + \frac{dG_\infty^2}{Ms_{1,i,1}^2} + \frac{dG_\infty^2}{s_{2,i,1}^2}\\
&\le \frac{p_i dL^2D^2}{TM\frac{d\sqrt{n}}{M}} + \frac{p_i dL^2D^2}{Td\sqrt{n}} + \frac{dG_\infty^2}{M\frac{Td}{M}} + \frac{dG_\infty^2}{dT}\\
&= \frac{2(L^2D^2 + G_\infty^2)}{T}.
\end{aligned} \tag{26}$$

Combine Equations (19), (25) and (26)

$$\mathbb{E}[\|\nabla f(x_{i,k}) - \bar{g}_{i,k}\|^2] \leq \frac{3L^2D^2 + 2G_\infty^2}{T}. \tag{27}$$

By Assumption 4, f is also a bounded (potentially) non-convex function on $\mathcal{K}$ with L-Lipschitz continuous gradient. Specifically, we have $\sup_{x\in\mathcal{K}}|f(x)|\leq M_0$. Note that if we define $v_t' = \operatorname{argmin}_{v\in\mathcal{K}}\langle v, \nabla f(x_t)\rangle$, then $\mathcal{G}(x_t) = \langle v_t' - x_t, -\nabla f(x_t)\rangle = -\langle v_t' - x_t, \nabla f(x_t)\rangle$. So we have

$$\begin{aligned}
f(x_{t+1}) &\overset{(a)}{\leq} f(x_t) + \langle \nabla f(x_t), x_{t+1} - x_t\rangle + \frac{L}{2}\|x_{t+1} - x_t\|^2 \\
&= f(x_t) + \langle \nabla f(x_t), \eta_t(v_t - x_t)\rangle + \frac{L}{2}\|\eta_t(v_t - x_t)\|^2 \\
&\overset{(b)}{\leq} f(x_t) + \eta_t\langle \nabla f(x_t), v_t - x_t\rangle + \frac{L\eta_t^2D^2}{2} \\
&= f(x_t) + \eta_t\langle \bar{g}_t, v_t - x_t\rangle + \eta_t\langle \nabla f(x_t) - \bar{g}_t, v_t - x_t\rangle + \frac{L\eta_t^2D^2}{2} \\
&\overset{(c)}{\leq} f(x_t) + \eta_t\langle \bar{g}_t, v_t' - x_t\rangle + \eta_t\langle \nabla f(x_t) - \bar{g}_t, v_t - x_t\rangle + \frac{L\eta_t^2D^2}{2} \\
&= f(x_t) + \eta_t\langle \nabla f(x_t), v_t' - x_t\rangle + \eta_t\langle \bar{g}_t - \nabla f(x_t), v_t' - x_t\rangle \\
&\qquad + \eta_t\langle \nabla f(x_t) - \bar{g}_t, v_t - x_t\rangle + \frac{L\eta_t^2D^2}{2} \\
&= f(x_t) - \eta_t\mathcal{G}(x_t) + \eta_t\langle \nabla f(x_t) - \bar{g}_t, v_t - v_t'\rangle + \frac{L\eta_t^2D^2}{2} \\
&\overset{(d)}{\leq} f(x_t) - \eta_t\mathcal{G}(x_t) + \eta_t\|\nabla f(x_t) - \bar{g}_t\|\|v_t - v_t'\| + \frac{L\eta_t^2D^2}{2} \\
&\overset{(e)}{\leq} f(x_t) - \eta_t\mathcal{G}(x_t) + \eta_t D\|\nabla f(x_t) - \bar{g}_t\| + \frac{L\eta_t^2D^2}{2},
\end{aligned}$$

where we used the assumption that f has L-Lipschitz continuous gradient in inequality (a). Inequalities (b), (e) hold because of Assumption 1. Inequality (c) is due to the optimality of v_t, and in (d), we applied the Cauchy-Schwarz inequality.

Rearrange the inequality above, we have

$$\eta_t\mathcal{G}(x_t) \leq f(x_t) - f(x_{t+1}) + \eta_t D\|\nabla f(x_t) - \bar{g}_t\| + \frac{L\eta_t^2D^2}{2}. \tag{28}$$

Apply Equation (28) recursively for $t = 1, 2, \cdots, T$, and take expectations, we attain the following inequality:

$$\sum_{t=1}^{T}\eta_t\mathbb{E}[\mathcal{G}(x_t)] \leq f(x_1) - f(x_{T+1}) + D\sum_{t=1}^{T}\eta_t\mathbb{E}[\|\nabla f(x_t) - \bar{g}_t\|] + \frac{LD^2}{2}\sum_{t=1}^{T}\eta_t^2. \tag{29}$$

Since we have $\mathbb{E}[\|\nabla f(x_{i,k}) - \bar{g}_{i,k}\|^2] \leq \frac{3L^2D^2+2G_\infty^2}{T} \triangleq \frac{c^2}{T}$, we have

$$\mathbb{E}[\|\nabla f(x_t) - \bar{g}_t\|] \leq \sqrt{\mathbb{E}[\|\nabla f(x_t) - \bar{g}_t\|^2]} \leq \frac{c}{\sqrt{T}}.$$

With $\eta_t = T^{-1/2}$, we then have

$$\begin{aligned}
\sum_{t=1}^{T}\mathbb{E}[\mathcal{G}(x_t)] &\leq \sqrt{T}[f(x_1) - f(x_{T+1})] + D\sum_{t=1}^{T}\mathbb{E}[\|\nabla f(x_t) - \bar{g}_t\|] + \sqrt{T}\frac{LD^2}{2}T(T^{-1/2})^2 \\
&\leq 2M_0\sqrt{T} + DT\frac{c}{\sqrt{T}} + \frac{LD^2}{2}\sqrt{T} = (2M_0 + cD + \frac{LD^2}{2})\sqrt{T}.
\end{aligned}$$

So

$$\mathbb{E}[\mathcal{G}(x_o)] = \frac{\sum_{t=1}^{T} \mathbb{E}[\mathcal{G}(x_t)]}{T} \le \frac{2M_0 + cD + \frac{LD^2}{2}}{\sqrt{T}}.$$

Therefore, in order to find an ϵ-first order stationary points, we need at most $\mathcal{O}(1/\epsilon^2)$ iterations. The IFO complexity per worker is $[n + 2(p-1)S_{i,k}] \cdot \frac{T}{p} = \mathcal{O}(\sqrt{n}/\epsilon^2) = \mathcal{O}(\sqrt{N}/(\epsilon^2\sqrt{M}))$. The average communication bits per round is $\frac{1}{p}\{M[32 + d(z_{1,i,1}+1) + (p-1)(32 + d(z_{1,i,k}+1))] + [32 + d(z_{2,i,1}+1) + (p-1)(32 + d(z_{2,i,k}+1))]\} = (32+d)(M+1) + \frac{Md}{\sqrt{n}}\log_2(\sqrt{\frac{Td}{M}}+1) + Md\log_2(\frac{d^{1/2}n^{1/4}}{\sqrt{M}}+1) + \frac{d}{\sqrt{n}}\log_2(\sqrt{TD}+1) + d\log_2(d^{1/2}n^{1/4}+1)$.