Supplementary Material to Decentralized Stochastic Projection-free Learning with Compressed Push-sum

APPENDIX

This document contains proofs of Lemmas 1 to 5 and Theorems 1 and 2 in the paper that establish the convergence rate of CE-DSFW summarized in Algorithm 1.

Algorithm 1 The CE-DSFW Algorithm

Initialize $\boldsymbol{x}_0^{(i)} \in \mathcal{C}$, $\boldsymbol{d}_{-1}^{(i)} = \mathbf{0}$, $\boldsymbol{y}_{-1}^{(i)} = \mathbf{1}$, $\forall i \in [n]$.
for each node $i \in [n]$ and $k = 0, 1, \cdots$ **do**
  Sample $\zeta_k^{(i)}$ at random according to $\mathcal{D}^{(i)}$
  $\boldsymbol{a}_k^{(i)} = \boldsymbol{a}_{k-1}^{(i)} + \eta_k \operatorname{Comp}\left(\nabla f(\boldsymbol{x}_k^{(i)}, \zeta_k^{(i)}) - \boldsymbol{a}_{k-1}^{(i)}\right)$
  $\boldsymbol{z}_k^{(i)} = \sum_{j=1}^{n} W_{ij} \boldsymbol{a}_k^{(i)}$
  $\boldsymbol{y}_k^{(i)} = \sum_{j=1}^{n} W_{ij} \boldsymbol{y}_{k-1}^{(j)}$
  $\boldsymbol{c}_k^{(i)} = \boldsymbol{z}_k^{(i)} \oslash \boldsymbol{y}_k^{(i)}$
  $\boldsymbol{d}_k^{(i)} = (1 - \eta_{k-1}) \boldsymbol{d}_{k-1}^{(i)} + \eta_{k-1} \boldsymbol{c}_k^{(i)}$
  $\boldsymbol{s}_k^{(i)} = \operatorname*{argmin}_{\boldsymbol{s} \in \mathcal{C}} \ \langle \boldsymbol{d}_k^{(i)}, \boldsymbol{s} \rangle$
  $\boldsymbol{x}_{k+1}^{(i)} = \boldsymbol{x}_k^{(i)} + \gamma_k (\boldsymbol{s}_k^{(i)} - \boldsymbol{x}_k^{(i)})$
  $\gamma_k \leftarrow \frac{2}{(k+2)^\beta}$ and η_k.
end for

Before proceeding with the derivations, we present the assumptions and some backgrounds.

Assumption 1. **L-smooth local objectives.** There exists a constant $L > 0$ such that

$$\|\nabla g(\boldsymbol{y}) - \nabla g(\boldsymbol{x})\| \leq L \|\boldsymbol{y} - \boldsymbol{x}\| \tag{1}$$

or equivalently

$$g(\boldsymbol{y}) \leq g(\boldsymbol{x}) + \langle \nabla g(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle + \frac{L}{2} \|\boldsymbol{y} - \boldsymbol{x}\|^2, \ \ \forall \boldsymbol{y}, \boldsymbol{x} \in \mathcal{R}^d. \tag{2}$$

We assume that functions $f(\cdot)$ and $f(\cdot, \cdot)$ satisfy the smoothness assumption.

Assumption 2. Diameter of $\mathcal{C}$. The constraint set $\mathcal{C}$ is convex and compact with diameter D, i.e.,

$$\|\boldsymbol{y}-\boldsymbol{x}\| \leq D, \quad \forall \boldsymbol{y}, \boldsymbol{x} \in \mathcal{C} \subset \mathcal{R}^d.$$

Assumption 3. Bounded local stochastic gradients. Local stochastic gradients $\nabla f(\boldsymbol{x}_k^{(i)}, \zeta_k^{(i)})$ have bounded variance

$$\mathbb{E}\left[\|\nabla f(\boldsymbol{x}_k^{(i)}, \zeta_k^{(i)}) - \nabla f(\boldsymbol{x}_k^{(i)})\|^2\right] \leq \nu^2 \quad \forall i \in [n], \tag{3}$$

where $\nabla f(\boldsymbol{x}_k^{(i)})$ is the full (non-stochastic) gradient evaluated at $\boldsymbol{x}_k^{(i)}$.

Assumption 4. Unbiasedness and variance of compression operator. For any vector $\boldsymbol{x} \in \mathcal{R}^d$, the compression operator $\mathrm{Comp}(\cdot)$ satisfies

$$\mathbb{E}\left[\mathrm{Comp}(\boldsymbol{x})|\boldsymbol{x}\right] = \boldsymbol{x}, \quad \mathbb{E}\|\boldsymbol{x} - \mathrm{Comp}(\boldsymbol{x})\|^2 \leq \delta^2. \tag{4}$$

In the following lemma, we give the convergence guarantees for the push-sum method employed to reach a consensus. Consider applying push-sum to the parameter $vp_{k-1}^{(i)}$ to obtain the $vp_k^{(i)}$, then the parameter is guaranteed to converge to the consensus $\bar{\boldsymbol{p}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{p}_k^{(i)}$ and is formally stated below.

Lemma A1. If the update from the push-sum algorithm $\boldsymbol{p}_k^{(i)} = \boldsymbol{a}_t^{(i)} \oslash \boldsymbol{y}_t^{(i)}$ converges to a consensus, then the consensus is the average vector $\bar{\boldsymbol{p}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{p}_k^{(i)}$ [1][Theorem 3.1].

Given the push-sum converges to the consensus we bound the error at each iterate inn the folloeing lemma.

Lemma A2. For a directed network characterized by a row stochastic connectivity matrix $\boldsymbol{W} \in \mathbb{R}^{n\times n}$ the distance of the parameter to the consensus can be bounded as

$$\|\boldsymbol{p}_k^{(i)} - \bar{\boldsymbol{p}}\|_2 \leq \sqrt{n}\|\boldsymbol{p}_0^{(i)}\|_\infty u(k) \tag{5}$$

where $u(k)$ is a non-increasing sequence.

Proof. Using [1][Lemma 4.1], we have

$$\|\boldsymbol{p}_k^{(i)} - \bar{\boldsymbol{p}}\|_\infty \leq \|\boldsymbol{p}_0^{(i)}\|_\infty u(k).$$

The bound in Equation (5) follows from the norm inequality $\|\boldsymbol{p}\|_2 \leq \sqrt{n}\|\boldsymbol{p}\|_\infty$ for $\boldsymbol{p} \in \mathbb{R}^n$. $\square$

Equation (5) bounds the error at each iterate obtained from the push-sum algorithm. Now we want to bound the improvement at each round given the worst-case performance until the previous iterate, which is formally stated below.

Lemma A3. Consider the parameter $\boldsymbol{p}_k^{(i)}$ at each node $i \in [n]$, then we have

$$\|\boldsymbol{p}_{k+1}^{(i)} - \bar{\boldsymbol{p}}\|_2 \leq \rho\|\boldsymbol{p}_k^{(i)} - \bar{\boldsymbol{p}}\|_2,$$

where $\rho \leq 1$ is the improvement factor at each gossip step.

Proof. It immediately follows from Lemma A2 that

$$\frac{\|\boldsymbol{p}_{k+1}^{(i)} - \bar{\boldsymbol{p}}\|_2}{\|\boldsymbol{p}_k^{(i)} - \bar{\boldsymbol{p}}\|_2} \leq \frac{u(k+1)}{u(k)} =: \rho_k$$

where $\rho_k = \frac{u(k+1)}{u(k)} \leq 1$ as $u(\cdot)$ is a non-increasing function and $\rho = \max_k \rho_k$. $\square$

The constant ρ depends on the network topology and is an indicator for improvement at each iterate. For undirected networks, the constant reduces to the second largest eigenvalue of the adjacency matrix $\boldsymbol{W}$, i.e., $\rho = \lambda_2(\boldsymbol{W})$ [2].

Using Assumption 1 and 3, we introduce a non-standard smoothness assumption for the objective function as in the next lemma.

Lemma A4. The exists a constant $\bar{L} > 0$, such that

$$\mathbb{E}\|\nabla f(\boldsymbol{x}_k^{(i)}, \zeta_k^{(i)}) - \nabla f(\boldsymbol{x}_k^{(j)})\| \leq \bar{L}D, \tag{6}$$

where $\nabla f(\boldsymbol{x}_k^{(i)}, \zeta_k^{(i)})$, is the stochastic gradient evaluated at $\boldsymbol{x}_k^{(i)}$ and $\nabla f(\boldsymbol{x}_k^{(j)})$, the deterministic gradient evaluated at $\boldsymbol{x}_k^{(j)}$.

Proof. Consider $\nabla f(\boldsymbol{x}_k^{(i)})$, the deterministic gradient at the ith node in the kth iterate, then we have

$$\begin{aligned}
\mathbb{E}\|\nabla f(\boldsymbol{x}_k^{(i)}, \zeta_k^{(i)}) - \nabla f(\boldsymbol{x}_k^{(j)})\| &= \mathbb{E}\|\nabla f(\boldsymbol{x}_k^{(i)}, \zeta_k^{(i)}) - \nabla f(\boldsymbol{x}_k^{(i)}) + \nabla f(\boldsymbol{x}_k^{(i)}) - \nabla f(\boldsymbol{x}_k^{(j)})\| \\
&\overset{(a)}{\leq} \mathbb{E}\|\nabla f(\boldsymbol{x}_k^{(i)}, \zeta_k^{(i)}) - \nabla f(\boldsymbol{x}_k^{(i)})\| + \mathbb{E}\|\nabla f(\boldsymbol{x}_k^{(i)}) - \nabla f(\boldsymbol{x}_k^{(j)})\| \\
&\overset{(b)}{\leq} \nu + \mathbb{E}\|\nabla f(\boldsymbol{x}_k^{(i)}) - \nabla f(\boldsymbol{x}_k^{(j)})\| \\
&\overset{(c)}{\leq} \nu + L\,\mathbb{E}\|\boldsymbol{x}_k^{(i)} - \boldsymbol{x}_k^{(j)}\| \\
&\leq \bar{L}\,D,
\end{aligned}$$

where (a) follows from triangle inequality, (b) follows as

$$\mathbb{E}\|\nabla f(\boldsymbol{x}_k^{(i)}, \zeta_k^{(i)}) - \nabla f(\boldsymbol{x}_k^{(i)})\| = \mathbb{E}\sqrt{\|\nabla f(\boldsymbol{x}_k^{(i)}, \zeta_k^{(i)}) - \nabla f(\boldsymbol{x}_k^{(i)})\|^2} \leq \sqrt{\mathbb{E}\|\nabla f(\boldsymbol{x}_k^{(i)}, \zeta_k^{(i)}) - \nabla f(\boldsymbol{x}_k^{(i)})\|^2} \leq \nu,$$

(c) follows from Assumption 1, and final expression follows from Assumption 2. $\square$

Now using Assumption 3, we can bound the variance of the local stochastic gradient at each node after a gossip step as in the next lemma.

Lemma 1. Uncompressed local stochastic gradients after one gossip step, i.e., $\boldsymbol{z}_k^{(i)} = \sum_{j\in[n]} w_{ij} \nabla f(\boldsymbol{x}_k^{(j)}, \zeta_k^{(j)})$, has bounded variance

$$\mathbb{E}\left[\|\boldsymbol{z}_k^{(i)} - \nabla f(\boldsymbol{x}_k^{(i)})\|^2\right] \leq \sigma^2 \quad \forall i \in [n], \tag{7}$$

where σ is related to ν and the network topology.

Proof. The average stochastic gradient (averaged over n nodes) at iteration k is $\frac{1}{n}\sum_{j=1}^{n}\nabla f(\boldsymbol{x}_k^{(j)},\zeta_k^{(j)})$. Then, we have the following

$$\begin{aligned}
\mathbb{E}\left[\|\boldsymbol{z}_k^{(i)}-\nabla f(\boldsymbol{x}_k^{(i)})\|^2\right] &= \mathbb{E}\left[\|\boldsymbol{z}_k^{(i)}-\frac{1}{n}\sum_{j=1}^{n}\nabla f(\boldsymbol{x}_k^{(j)},\zeta_k^{(j)})+\frac{1}{n}\sum_{j=1}^{n}\nabla f(\boldsymbol{x}_k^{(j)},\zeta_k^{(j)})-\nabla f(\boldsymbol{x}_k^{(i)})\|^2\right]\\
&= \mathbb{E}\left[\|\boldsymbol{z}_k^{(i)}-\frac{1}{n}\sum_{j=1}^{n}\nabla f(\boldsymbol{x}_k^{(j)},\zeta_k^{(j)})\|^2\right]+\mathbb{E}\left[\|\frac{1}{n}\sum_{j=1}^{n}\nabla f(\boldsymbol{x}_k^{(j)},\zeta_k^{(j)})-\nabla f(\boldsymbol{x}_k^{(i)})\|^2\right]\\
&\qquad + 2\,\mathbb{E}\left[\langle\boldsymbol{z}_k^{(i)}-\frac{1}{n}\sum_{j=1}^{n}\nabla f(\boldsymbol{x}_k^{(j)},\zeta_k^{(j)}),\frac{1}{n}\sum_{j=1}^{n}\nabla f(\boldsymbol{x}_k^{(j)},\zeta_k^{(j)})-\nabla f(\boldsymbol{x}_k^{(i)})\rangle\right]\\
&\overset{(a)}{\leq} 2\mathbb{E}\left[\|\boldsymbol{z}_k^{(i)}-\frac{1}{n}\sum_{j=1}^{n}\nabla f(\boldsymbol{x}_k^{(j)},\zeta_k^{(j)})\|^2\right]+2\mathbb{E}\left[\|\frac{1}{n}\sum_{j=1}^{n}\nabla f(\boldsymbol{x}_k^{(j)},\zeta_k^{(j)})-\nabla f(\boldsymbol{x}_k^{(i)})\|^2\right]\\
&\overset{(b)}{\leq} \underbrace{2\,\mathbb{E}\left[\|\boldsymbol{z}_k^{(i)}-\frac{1}{n}\sum_{j=1}^{n}\nabla f(\boldsymbol{x}_k^{(j)},\zeta_k^{(j)})\|^2\right]}_{:=T_1}\\
&\qquad + \underbrace{2\,\frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\left[\|\nabla f(\boldsymbol{x}_k^{(j)},\zeta_k^{(j)})-\nabla f(\boldsymbol{x}_k^{(i)})\|^2\right]}_{:=T_2},
\end{aligned}$$

where (a) follows from the Young's inequality $2\langle\boldsymbol{k},\boldsymbol{l}\rangle\leq\eta\|\boldsymbol{k}\|^2+\frac{1}{\eta}\|\boldsymbol{l}\|^2$ for all $\eta>0$, (we use $\boldsymbol{k}=\boldsymbol{z}_k^{(i)}-\frac{1}{n}\sum_{j=1}^{n}\nabla f(\boldsymbol{x}_k^{(j)},\zeta_k^{(j)})$, $\boldsymbol{l}=\frac{1}{n}\sum_{j=1}^{n}\nabla f(\boldsymbol{x}_k^{(j)},\zeta_k^{(j)})-\nabla f(\boldsymbol{x}_k^{(i)})$, and $\eta=1$) and (b) follows from the Jensen's inequality $l(\sum_i\alpha_i\boldsymbol{x}_i)\leq\sum_i\alpha_i l(\boldsymbol{x}_i)$ with $l(\cdot)$ being a convex function and $\sum\limits_{i=1}^{n}\alpha_i=1$.

Using Lemma A4, we have

$$T_2\leq\bar{L}^2D^2.$$

Recall that $\boldsymbol{z}_k^{(i)}$ is the uncompressed stochastic gradient after one gossip step and $\nabla f(\boldsymbol{x}_k^{(i)},\zeta_k^{(i)})$ is the local stochastic gradient before gossip at node i. From Lemma A3, the first term satisfies

$$\begin{aligned}
T_1 &\leq \rho^2\mathbb{E}\left[\|\nabla f(\boldsymbol{x}_k^{(i)},\zeta_k^{(i)})-\frac{1}{n}\sum_{j=1}^{n}\nabla f(\boldsymbol{x}_k^{(j)},\zeta_k^{(j)})\|^2\right]\\
&\overset{(a)}{=} \rho^2\mathbb{E}\left[\|\frac{1}{n}\sum_{j=1}^{n}\nabla f(\boldsymbol{x}_k^{(i)},\zeta_k^{(i)})-\frac{1}{n}\sum_{j=1}^{n}\nabla f(\boldsymbol{x}_k^{(j)},\zeta_k^{(j)})\|^2\right]\\
&\overset{(b)}{\leq} \frac{\rho^2}{n}\sum_{j=1}^{n}\mathbb{E}\left[\|\nabla f(\boldsymbol{x}_k^{(i)},\zeta_k^{(i)})-\nabla f(\boldsymbol{x}_k^{(j)},\zeta_k^{(j)})\|^2\right]\\
&\overset{(c)}{\leq} \frac{L^2\rho^2}{n}\sum_{j=1}^{n}\mathbb{E}\left[\|\boldsymbol{x}_k^{(i)}-\boldsymbol{x}_k^{(j)}\|^2\right]\\
&\leq \rho^2L^2D^2,
\end{aligned}$$

where (a) is obtained by substituting $\nabla f(\boldsymbol{x}_k^{(i)}) = \frac{1}{n}\sum_{j=1}^{n} \nabla f(\boldsymbol{x}_k^{(i)})$, (b) follows from Jensen's inequality, and (c) follows from the Assumption 1. Now combining T_1 and T_2, we have

$$\mathbb{E}\left[\|\boldsymbol{z}_k^{(i)} - \nabla f(\boldsymbol{x}_k^{(i)})\|^2\right] \leq \bar{L}^2 D^2 + \rho^2 L^2 D^2 := \sigma^2.$$

□

A. *Proof of Lemma 2: Inexact Iterates due to Decentralization*

To prove the Lemma 2 (stated below as well), we use the following result from [3].

Lemma A6. [3][Lemma 17] Suppose a sequence of real numbers ϕ_k satisfies

$$\phi_k = \left(1 - \frac{a_1}{(k+k_0)^{a_3}}\right)\phi_{k-1} + \frac{a_2}{(k+k_0)^{2a_3}}, \tag{8}$$

for some scalars $k_0 \geq 0$, $a_1 > 1$, a_2, and $a_3 \leq 1$. Then the sequence ϕ_k converges at the rate

$$\phi_k \leq \frac{\max\{k_0^{a_3}\phi_0, a_2/(a_1-1)\}}{(k+k_0+1)^{a_3}}. \tag{9}$$

Lemma 2. The error associated with the iterate $\boldsymbol{x}_k^{(i)}$ to the global consensus $\bar{\boldsymbol{x}}_k$ in expectation satisfies

$$q_k^2 = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\|\boldsymbol{x}_k^{(i)} - \bar{\boldsymbol{x}}_k\|^2\right] \leq \frac{4D^2}{(k+2)^{\beta}}.$$

Proof. Consider the distance of the iterate $\boldsymbol{x}_{k+1}^{(i)}$ and the global consensus $\bar{\boldsymbol{x}}_{k+1}$, we have

$$\begin{aligned}
q_{k+1}^2 &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\|\boldsymbol{x}_{k+1}^{(i)} - \bar{\boldsymbol{x}}_{k+1}\|^2\right] \\
&\stackrel{(a)}{=} \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\|(1-\gamma_k)(\boldsymbol{x}_k^{(i)} - \bar{\boldsymbol{x}}_k) + \gamma_k(\boldsymbol{s}_k^{(i)} - \bar{\boldsymbol{s}}_k)\|^2\right] \\
&\stackrel{(b)}{\leq} \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\|(1-\gamma_k)(\boldsymbol{x}_k^{(i)} - \bar{\boldsymbol{x}}_k) + \gamma_k(\boldsymbol{s} - \bar{\boldsymbol{s}}_k)\|^2\right]
\end{aligned}$$

where (a) is from the update [cf. Algorithm 1, Line 13] and (b) follows from the definition

$$\boldsymbol{s} = \operatorname*{argmax}_{s\in\mathcal{C}} \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\|(1-\gamma_k)(\boldsymbol{x}_k^{(i)} - \bar{\boldsymbol{x}}_k) + \gamma_k(\boldsymbol{s} - \bar{\boldsymbol{s}}_k)\|^2\right].$$

On expanding the above expression we get

$$\begin{aligned}
=&(1-\gamma_k)^2\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\|\boldsymbol{x}_k^{(i)} - \bar{\boldsymbol{x}}_k\|^2\right] + \gamma_k^2\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\|\boldsymbol{s} - \bar{\boldsymbol{s}}_k\|^2\right] \\
&\qquad + 2(1-\gamma_k)\gamma_k\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\langle \boldsymbol{x}_k^{(i)} - \bar{\boldsymbol{x}}_k, \boldsymbol{s} - \bar{\boldsymbol{s}}_k\rangle\right] \\
\stackrel{(a)}{=}& (1-\gamma_k)^2\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\|\boldsymbol{x}_k^{(i)} - \bar{\boldsymbol{x}}_k\|^2\right] + \gamma_k^2\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\|\boldsymbol{s}_k^{(i)} - \bar{\boldsymbol{s}}_k\|^2\right]
\end{aligned}$$

where (a) is from the definition $\bar{\boldsymbol{x}}_k = \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_k^{(i)}$ due to which the cross term vanishes. From Assumption 2, we have

$$\begin{aligned}
q_{k+1}^2 &\leq (1-\gamma_k)q_k^2 + \gamma_k^2 D^2 \\
&\overset{(a)}{=} \left(1 - \frac{2}{(k+2)^\beta}\right) q_k^2 + \frac{4D^2}{(k+2)^{2\beta}} \\
&\overset{(b)}{\leq} \frac{\max\{2^\beta q_0^2, 4D^2\}}{(k+2)^\beta} \\
&\overset{(c)}{\leq} \frac{4D^2}{(k+3)^\beta}
\end{aligned}$$

where we use the step-size $\gamma_k = \frac{2}{(k+2)^\beta}$ in (a), expression (b) follows from Lemma A6 and (c) follows as $2^\beta q_0^2 = 2^\beta \frac{2D^2}{2^\beta} = 2D^2$, hence $\max\{2^\beta q_0^2, 4D^2\} = 4D^2$. □

B. Proof of Lemma 3: Inexact Iterates due to Compression

Lemma 3. The error associated with the estimate of the compressed gradient with memory $\boldsymbol{a}_k^{(i)}$ to the stochastic gradient $\nabla f(\boldsymbol{x}_k^{(i)}, \zeta_k^{(i)}),\ \forall i \in [n]$ satisfies

$$\begin{aligned}
u_k^2 &= \mathbb{E}\|\nabla f(\boldsymbol{x}_k^{(i)}, \zeta_k^{(i)}) - \boldsymbol{a}_k^{(i)}\|^2 \\
&\leq \frac{C_u}{n(k+1)^\alpha},
\end{aligned}$$

where $C_u = 2L^2D^2 + 4n^{3/2}\delta^2$.

Proof. Consider the case $k = 0$,

$$\begin{aligned}
u_0^2 &= \mathbb{E}\|\nabla f(\boldsymbol{x}_0^{(i)}, \zeta_0^{(i)}) - \boldsymbol{a}_0^{(i)}\|^2 \\
&= \mathbb{E}\|\nabla f(\boldsymbol{x}_0^{(i)}, \zeta_0^{(i)}) - \text{Comp}\left(\nabla f(\boldsymbol{x}_0^{(i)}, \zeta_0^{(i)})\right)\|^2 \\
&\overset{(a)}{\leq} \delta^2, \qquad (10)
\end{aligned}$$

where a, follows from Assumption 4. Consider the conditional error (conditioned on all randomness until iterate $k+1$)

$$
\begin{aligned}
&\mathbb{E}_{\cdot|\mathcal{F}_{k+1}}\left[\|\nabla f(\boldsymbol{x}_k^{(i)},\zeta_k^{(i)})-\boldsymbol{a}_k^{(i)}\|^2\right]\\
&\quad\overset{(a)}{=}\mathbb{E}_{\cdot|\mathcal{F}_{k+1}}\Big[\|\nabla f(\boldsymbol{x}_k^{(i)},\zeta_k^{(i)})-\boldsymbol{a}_{k-1}^{(i)}-\eta_k\operatorname{Comp}(\nabla f(\boldsymbol{x}_k^{(i)},\zeta_k^{(i)})-\boldsymbol{a}_{k-1}^{(i)})\|^2\Big]\\
&\quad\overset{(b)}{=}\mathbb{E}_{\cdot|\mathcal{F}_{k+1}}\Big[\|\nabla f(\boldsymbol{x}_k^{(i)},\zeta_k^{(i)})-(1-\eta_k)\nabla f(\boldsymbol{x}_{k-1}^{(i)},\zeta_{k-1}^{(i)})+(1-\eta_k)\nabla f(\boldsymbol{x}_{k-1}^{(i)},\zeta_{k-1}^{(i)})-\boldsymbol{a}_{k-1}^{(i)}\\
&\qquad\quad-\eta_k\operatorname{Comp}(\nabla f(\boldsymbol{x}_k^{(i)},\zeta_k^{(i)})-\boldsymbol{a}_{k-1}^{(i)})\|^2\Big]\\
&\quad=\mathbb{E}_{\cdot|\mathcal{F}_{k+1}}\Big[\|\underbrace{(1-\eta_k)\Big(\nabla f(\boldsymbol{x}_k^{(i)},\zeta_k^{(i)})-\nabla f(\boldsymbol{x}_{k-1}^{(i)},\zeta_{k-1}^{(i)})\Big)}_{:=\boldsymbol{\alpha_i}}+\underbrace{(1-\eta_k)\Big(\nabla f(\boldsymbol{x}_{k-1}^{(i)},\zeta_{k-1}^{(i)})-\boldsymbol{a}_{k-1}^{(i)}\Big)}_{:=\boldsymbol{\beta_i}}\\
&\qquad\quad+\underbrace{\eta_k\Big(\nabla f(\boldsymbol{x}_k^{(i)},\zeta_k^{(i)})-\boldsymbol{a}_{k-1}^{(i)}-\operatorname{Comp}(\nabla f(\boldsymbol{x}_k^{(i)},\zeta_k^{(i)})-\boldsymbol{a}_{k-1}^{(i)})\Big)}_{:=\boldsymbol{\gamma_i}}\|^2\Big]\\
&\quad=\mathbb{E}_{\cdot|\mathcal{F}_{k+1}}\Big[\|\boldsymbol{\alpha_i}\|^2+\|\boldsymbol{\beta_i}\|^2+\|\boldsymbol{\gamma_i}\|^2\\
&\qquad\quad+2(1-\eta_k)^2\langle\nabla f(\boldsymbol{x}_k^{(i)},\zeta_k^{(i)})-\nabla f(\boldsymbol{x}_{k-1}^{(i)},\zeta_{k-1}^{(i)}),\nabla f(\boldsymbol{x}_{k-1}^{(i)},\zeta_{k-1}^{(i)})-\boldsymbol{a}_{k-1}^{(i)}\rangle\\
&\qquad\quad+2(1-\eta_k)\eta_k\langle\nabla f(\boldsymbol{x}_k^{(i)},\zeta_k^{(i)})-\nabla f(\boldsymbol{x}_{k-1}^{(i)},\zeta_{k-1}^{(i)}),\nabla f(\boldsymbol{x}_k^{(i)},\zeta_k^{(i)})-\boldsymbol{a}_{k-1}^{(i)}-\operatorname{Comp}(\nabla f(\boldsymbol{x}_k^{(i)},\zeta_k^{(i)})-\boldsymbol{a}_{k-1}^{(i)})\rangle\\
&\qquad\quad+2(1-\eta_k)\eta_k\langle\nabla f(\boldsymbol{x}_{k-1}^{(i)},\zeta_{k-1}^{(i)})-\boldsymbol{a}_{k-1}^{(i)},\nabla f(\boldsymbol{x}_k^{(i)},\zeta_k^{(i)})-\boldsymbol{a}_{k-1}^{(i)}-\operatorname{Comp}(\nabla f(\boldsymbol{x}_k^{(i)},\zeta_k^{(i)})-\boldsymbol{a}_{k-1}^{(i)})\rangle
\end{aligned}
$$

where (a) follows from the update of the quantized gradient vector with memory $\boldsymbol{a}_k^{(i)}=\boldsymbol{a}_{k-1}^{(i)}+\eta_k\operatorname{Comp}(\nabla f(\boldsymbol{x}_k^{(i)},\zeta_k^{(i)})-\boldsymbol{a}_{k-1}^{(i)})$ and we introduce the term $(1-\eta_k)\nabla f(\boldsymbol{x}_{k-1}^{(i)},\zeta_{k-1}^{(i)})$ to obtain (b). The cross terms ($\langle\boldsymbol{\alpha}_i,\boldsymbol{\gamma}_i\rangle$ and $\langle\boldsymbol{\beta}_i,\boldsymbol{\gamma}_i\rangle$) vanish due to the unbiasedness of the compression operator $\operatorname{Comp}(\cdot)$, and we have

$$
\begin{aligned}
\mathbb{E}_{\cdot|\mathcal{F}_{k+1}}\left[\|\nabla f(\boldsymbol{x}_k^{(i)},\zeta_k^{(i)})-\boldsymbol{a}_k^{(i)}\|^2\right]&=\|\boldsymbol{\alpha_i}\|^2+\|\boldsymbol{\beta_i}\|^2+\|\boldsymbol{\gamma_i}\|^2\\
&\quad+2(1-\eta_k)^2\langle\nabla f(\boldsymbol{x}_k^{(i)},\zeta_k^{(i)})-\nabla f(\boldsymbol{x}_{k-1}^{(i)},\zeta_{k-1}^{(i)}),\nabla f(\boldsymbol{x}_{k-1}^{(i)},\zeta_{k-1}^{(i)})-\boldsymbol{a}_{k-1}^{(i)}\rangle.
\end{aligned}
$$

Now using the iterated property of conditional expectation, i.e., $\mathbb{E}[\cdot]=\mathbb{E}\left[\mathbb{E}_{\cdot|\mathcal{F}_{k+1}}\right]$

$$
\begin{aligned}
u_k^2&=\mathbb{E}\left[\mathbb{E}_{\cdot|\mathcal{F}_{k+1}}\left[\|\nabla f(\boldsymbol{x}_k^{(i)},\zeta_k^{(i)})-\boldsymbol{a}_k^{(i)}\|^2\right]\right]\\
&=\mathbb{E}\Big[\|\boldsymbol{\alpha_i}\|^2+\|\boldsymbol{\beta_i}\|^2+\|\boldsymbol{\gamma_i}\|^2\\
&\qquad+2(1-\eta_k)^2\langle\nabla f(\boldsymbol{x}_k^{(i)},\zeta_k^{(i)})-\nabla f(\boldsymbol{x}_{k-1}^{(i)},\zeta_{k-1}^{(i)}),\nabla f(\boldsymbol{x}_{k-1}^{(i)},\zeta_{k-1}^{(i)})-\boldsymbol{a}_{k-1}^{(i)}\rangle\Big]\\
&=\mathbb{E}\Big[(1-\eta_k)^2\|\nabla f(\boldsymbol{x}_k^{(i)},\zeta_k^{(i)})-\nabla f(\boldsymbol{x}_{k-1}^{(i)},\zeta_{k-1}^{(i)})\|^2+(1-\eta_k)^2\|\nabla f(\boldsymbol{x}_{k-1}^{(i)},\zeta_{k-1}^{(i)})-\boldsymbol{a}_{k-1}^{(i)}\|^2\\
&\qquad+\eta_k^2\|\nabla f(\boldsymbol{x}_k^{(i)},\zeta_k^{(i)})-\boldsymbol{a}_{k-1}^{(i)}-\operatorname{Comp}(\nabla f(\boldsymbol{x}_k^{(i)},\zeta_k^{(i)})-\boldsymbol{a}_{k-1}^{(i)})\|^2\\
&\qquad+2(1-\eta_k)^2\langle\nabla f(\boldsymbol{x}_k^{(i)},\zeta_k^{(i)})-\nabla f(\boldsymbol{x}_{k-1}^{(i)},\zeta_{k-1}^{(i)}),\nabla f(\boldsymbol{x}_{k-1}^{(i)},\zeta_{k-1}^{(i)})-\boldsymbol{a}_{k-1}^{(i)}\rangle\Big]\\
&\overset{(a)}{\le}\mathbb{E}\Big[(1-\eta_k)^2(1+\frac{1}{\eta_k})\|\nabla f(\boldsymbol{x}_k^{(i)},\zeta_k^{(i)})-\nabla f(\boldsymbol{x}_{k-1}^{(i)},\zeta_{k-1}^{(i)})\|^2+(1-\eta_k)^2(1+\eta_k)\|\nabla f(\boldsymbol{x}_{k-1}^{(i)},\zeta_{k-1}^{(i)})-\boldsymbol{a}_{k-1}^{(i)}\|^2\\
&\qquad+\eta_k^2\|\nabla f(\boldsymbol{x}_k^{(i)},\zeta_k^{(i)})-\boldsymbol{a}_{k-1}^{(i)}-\operatorname{Comp}(\nabla f(\boldsymbol{x}_k^{(i)},\zeta_k^{(i)})-\boldsymbol{a}_{k-1}^{(i)})\|^2\Big],
\end{aligned}
$$

where we use the Young's inequality $2\langle \boldsymbol{\alpha}_i, \boldsymbol{\beta}_i \rangle \leq \frac{1}{\eta}\|\boldsymbol{\alpha}_i\|^2 + \eta\|\boldsymbol{\beta}_i\|^2$ with $\eta = \eta_k$, to obtain the expression in (a). Using the L-lipschitz continuity of gradient and from the definition u_{k-1}

$$\begin{aligned}
u_k^2 &= (1-\eta_k)(1-\eta_k^2)u_{k-1}^2 + (1-\eta_k)(-\eta_k + \frac{1}{\eta_k})\frac{1}{n}\sum_{i=1}^{n} L^2\mathbb{E}\left[\|\boldsymbol{x}_k^{(i)} - \boldsymbol{x}_{k-1}^{(i)}\|^2\right] \\
&\quad + \eta_k^2 \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\|\nabla f(\boldsymbol{x}_k^{(i)}, \zeta_k^{(i)}) - \boldsymbol{a}_{k-1}^{(i)} - \mathrm{Comp}(\nabla f(\boldsymbol{x}_k^{(i)}, \zeta_k^{(i)}) - \boldsymbol{a}_{k-1}^{(i)})\|^2\right] \\
&\overset{(a)}{\leq} (1-\eta_k)u_{k-1}^2 + \frac{1}{\eta_k}\gamma_k^2 L^2 D^2 + \eta_k^2\delta^2,
\end{aligned} \tag{11}$$

where we substitute $\boldsymbol{x}_k^{(i)} - \boldsymbol{x}_{k-1}^{(i)} = \gamma_k\left(\boldsymbol{s}_k^{(i)} - \boldsymbol{x}_{k-1}^{(i)}\right)$ from Step 9 of Algorithm 1 and then using Assumption 2 to bound the 2nd term, and from the unbaisedness of the compression operator Assumption 4 to obtain (a). Consider the case when $k > T$, with step-size $\gamma_k = \frac{2}{(k+2)^\beta}$ and the weight $\eta_k = \frac{2\sqrt{n}}{(k+2)^\alpha}$

$$\begin{aligned}
u_k^2 &\leq \left(1 - \frac{2\sqrt{n}}{(k+2)^\alpha}\right)u_{k-1}^2 + \frac{1}{\frac{2\sqrt{n}}{(k+2)^\alpha}}\frac{4}{(k+2)^{2\beta}}L^2D^2 + \frac{4n}{(k+2)^{2\alpha}}\delta^2 \\
&\overset{(b)}{=} \left(1 - \frac{2\sqrt{n}}{(k+2)^\alpha}\right)u_{k-1}^2 + \frac{2L^2D^2}{\sqrt{n}(k+2)^{2\alpha}} + \frac{4n\delta^2}{(k+2)^{2\alpha}} \\
&= \left(1 - \frac{2\sqrt{n}}{(k+2)^\alpha}\right)u_{k-1}^2 + \frac{2L^2D^2 + 4n^{3/2}\delta^2}{\sqrt{n}(k+2)^{2\alpha}}
\end{aligned}$$

where in (b) we use the fact that $\alpha \leq \frac{2}{3}\beta$. Finally, from Lemma A6 with $a_1 = 2\sqrt{n}, a_2 = \frac{2L^2D^2+4n^{3/2}\delta^2}{\sqrt{n}}$ and $a_3 = \alpha$, we obtain

$$\begin{aligned}
u_k^2 &\leq \frac{\max\{2u_0^2, \frac{(2L^2D^2+4n^{3/2}\delta^2)/\sqrt{n}}{2\sqrt{n}-1}\}}{(k+2)^\alpha} \\
&\overset{(a)}{\leq} \frac{\max\{2u_0^2, \frac{2L^2D^2+4n^{3/2}\delta^2}{n}\}}{(k+2)^\alpha} \\
&\leq \frac{\max\{2nu_0^2, 2L^2D^2 + 4n^{3/2}\delta^2\}}{n(k+2)^\alpha} \\
&= \frac{C_u}{n(k+2)^\alpha},
\end{aligned}$$

where, (a) follows from the fact that $2\sqrt{n} - 1 \geq \sqrt{n}$, for $n \geq 2$. Using Equation (10), the expression simplifies to $\max\{2nu_0^2, 2L^2D^2 + 4n^{3/2}\delta^2\} = 2L^2D^2 + 4n^{3/2}\delta^2$, that is, $C_u = 2L^2D^2 + 4n^{3/2}\delta^2$.

Now consider the case $k \leq T$,

$$\begin{aligned}
u_k^2 &= \mathbb{E}\|\nabla f(\boldsymbol{x}_k^{(i)}, \zeta_k^{(i)}) - \boldsymbol{a}_k^{(i)}\|^2 \\
&= \mathbb{E}\|\nabla f(\boldsymbol{x}_k^{(i)}, \zeta_k^{(i)}) - \boldsymbol{a}_{k-1}^{(i)} - \mathrm{Comp}\left(\nabla f(\boldsymbol{x}_k^{(i)}, \zeta_k^{(i)}) - \boldsymbol{a}_{k-1}^{(i)}\right)\|^2 \\
&\overset{(a)}{\leq} \delta^2,
\end{aligned}$$

where (a) follows from Assumption 4. Consider the term

$$\begin{aligned}
\frac{4n^{3/2}\delta^2}{n(T+2)^\alpha} &= \frac{4\sqrt{n}\delta^2}{\left((2\sqrt{n})^{1/\alpha} - 2 + 2\right)^\alpha} \\
&= \frac{4\sqrt{n}\delta^2}{2\sqrt{n}} = 2\delta^2.
\end{aligned}$$

From the above expression we see that $\delta^2 < \frac{4n^{3/2}\delta^2}{n(k+2)^\alpha}$ and $u_k^2 < \frac{C_u}{n(k+2)^\alpha}$ $\forall k \le T$, therefore we have $u_k^2 \le \frac{C_u}{n(k+2)^\alpha}$ $\forall k$. The bound follows for every node $i \in [n]$, as we are using same compression operation and variance reduction technique. So we neglected the depends of nodes on u_k. □

C. Proof of Lemma 4: Inexact Iterates due to Stochasticity and Compression

Lemma 4. The error associated with the stochastic estimate of the gradient $\boldsymbol{d}_k^{(i)}$ to the deterministic gradient $\nabla f^{(i)}(\boldsymbol{x}_k^{(i)})$ satisfies

$$r_k^2 = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\|\boldsymbol{d}_k^{(i)} - \nabla f(\boldsymbol{x}_k^{(i)})\|^2\right] \le \underbrace{\frac{2L^2D^2 + 4n^{(3/2)}\sigma^2}{n(k+1)^\alpha}}_{\text{stochasticity}} + \underbrace{\frac{4\sqrt{n}C_u}{n(k+1)^\alpha}}_{\text{quantization}} := \frac{C_r^2}{n(k+1)^\alpha},$$

where $C_r^2 = 2L^2D^2 + 4n^{(3/2)}\sigma^2 + 4\sqrt{n}C_u$.

Proof. The conditional error (conditioned on all randomness up to iterate k) is given by

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{.|\mathcal{F}_k}\left[\|\nabla f(\boldsymbol{x}_k^{(i)}) - \boldsymbol{d}_k^{(i)}\|^2\right] \\
&\quad \overset{(a)}{=} \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{.|\mathcal{F}_k}\left[\|\nabla f(\boldsymbol{x}_k^{(i)}) - (1-\eta_{k-1})\boldsymbol{d}_{k-1}^{(i)} + \eta_{k-1}\boldsymbol{c}_k^{(i)})\|^2\right] \\
&\quad \overset{(b)}{=} \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{.|\mathcal{F}_k}\Bigg[\|(1-\eta_{k-1})\underbrace{\left(\nabla f(\boldsymbol{x}_k^{(i)}) - \nabla f(\boldsymbol{x}_{k-1}^{(i)})\right)}_{:=\boldsymbol{\alpha}_i} \\
&\qquad + (1-\eta_{k-1})\underbrace{\left(\nabla f(\boldsymbol{x}_{k-1}^{(i)}) - \boldsymbol{d}_{k-1}^{(i)}\right)}_{:=\boldsymbol{\beta}_i} + \eta_{k-1}\underbrace{\left(\nabla f(\boldsymbol{x}_k^{(i)}) - \boldsymbol{z}_k^{(i)}\right)}_{:=\boldsymbol{\gamma}_i} + \eta_{k-1}\underbrace{\left(\boldsymbol{z}_k^{(i)} - \boldsymbol{c}_k^{(i)}\right)}_{:=\boldsymbol{\omega}_i}\|^2\Bigg],
\end{aligned}$$

where (a) follows from Step 8 and we introduce $(1-\eta_{k-1})\nabla f(\boldsymbol{x}_{k-1}^{(i)}), \eta_{k-1}\boldsymbol{z}_k^{(i)}$ to obtain (b). Here the vector $\boldsymbol{z}_k^{(i)} = \sum_{j\in[n]} w_{ij}\nabla f(\boldsymbol{x}_k^{(j)}, \zeta_k^{(j)})$, the uncompressed stochastic gradient after one round of gossip. On expanding the conditional error is given as

$$\begin{aligned}
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{.|\mathcal{F}_k}\Bigg[(1-\eta_{k-1})^2\|\boldsymbol{\alpha}_i\|^2 + (1-\eta_{k-1})^2\|\boldsymbol{\beta}_i\|^2 + \eta_{k-1}^2\|\boldsymbol{\gamma}_i\|^2 + \eta_{k-1}^2\|\boldsymbol{\omega}_i\|^2 \\
&\qquad + 2(1-\eta_{k-1})^2\langle\boldsymbol{\alpha}_i, \boldsymbol{\beta}_i\rangle + 2(1-\eta_{k-1})\eta_{k-1}\langle\boldsymbol{\alpha}_i, \boldsymbol{\gamma}_i\rangle + 2(1-\eta_{k-1})\eta_{k-1}\langle\boldsymbol{\beta}_i, \boldsymbol{\gamma}_i\rangle \\
&\qquad + 2(1-\eta_{k-1})\eta_{k-1}\langle\boldsymbol{\beta}_i, \boldsymbol{\omega}_i\rangle + 2(1-\eta_{k-1})\eta_{k-1}\langle\boldsymbol{\alpha}_i, \boldsymbol{\omega}_i\rangle + 2\eta_{k-1}^2\langle\boldsymbol{\omega}_i, \boldsymbol{\gamma}_i\rangle\Bigg] \\
&\overset{(a)}{=} \frac{1}{n}\sum_{i=1}^{n}(1-\eta_{k-1})^2\|\boldsymbol{\alpha}_i\|^2 + (1-\eta_{k-1})^2\|\boldsymbol{\beta}_i\|^2 + \eta_{k-1}^2\|\boldsymbol{\gamma}_i\|^2 + \eta_{k-1}^2\mathbb{E}_{.|\mathcal{F}_k}\|\boldsymbol{\omega}_i\|^2 \\
&\qquad + 2(1-\eta_{k-1})^2\langle\boldsymbol{\alpha}_i, \boldsymbol{\beta}_i\rangle + 2(1-\eta_{k-1})\eta_{k-1}\langle\boldsymbol{\alpha}_i, \boldsymbol{\gamma}_i\rangle + 2(1-\eta_{k-1})\eta_{k-1}\langle\boldsymbol{\beta}_i, \boldsymbol{\gamma}_i\rangle,
\end{aligned}$$

where the three cross terms ($\langle \boldsymbol{\alpha}_i, \boldsymbol{\omega}_i \rangle$, $\langle \boldsymbol{\beta}_i, \boldsymbol{\omega}_i \rangle$, and $\langle \boldsymbol{\omega}_i, \boldsymbol{\gamma}_i \rangle$) vanishes due to the unbiasedness of the compression operator $\mathrm{Comp}(\cdot)$, i.e., $\mathbb{E}_{.|\mathcal{F}_k}[\boldsymbol{\omega}_i] = 0 \ \forall i$. Now considering all the randomness until $k-1$th iterate

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{.|\mathcal{F}_{k-1}}\left[\mathbb{E}_{.|\mathcal{F}_k}\left[\|\nabla f(\boldsymbol{x}_k^{(i)}) - \boldsymbol{d}_k^{(i)}\|^2\right]\right] \\
&\qquad\qquad \leq \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{.|\mathcal{F}_{k-1}}\Big[(1-\eta_{k-1})^2\|\boldsymbol{\alpha}_i\|^2 + (1-\eta_{k-1})^2\|\boldsymbol{\beta}_i\|^2 + \eta_{k-1}^2\|\boldsymbol{\gamma}_i\|^2 + \eta_{k-1}^2\mathbb{E}_{.|\mathcal{F}_k}\|\boldsymbol{\omega}_i\|^2 \\
&\qquad\qquad\qquad + 2(1-\eta_{k-1})^2\langle \boldsymbol{\alpha}_i, \boldsymbol{\beta}_i \rangle + 2(1-\eta_{k-1})\eta_{k-1}\langle \boldsymbol{\alpha}_i, \boldsymbol{\gamma}_i \rangle + 2(1-\eta_{k-1})\eta_{k-1}\langle \boldsymbol{\beta}_i, \boldsymbol{\gamma}_i \rangle\Big] \\
&\qquad\qquad \overset{(a)}{\leq} \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{.|\mathcal{F}_{k-1}}\Big[(1-\eta_{k-1})^2\|\boldsymbol{\alpha}_i\|^2 + (1-\eta_{k-1})^2\|\boldsymbol{\beta}_i\|^2 + \eta_{k-1}^2\|\boldsymbol{\gamma}_i\|^2 + \eta_{k-1}^2\mathbb{E}_{.|\mathcal{F}_k}\|\boldsymbol{\omega}_i\|^2 \\
&\qquad\qquad\qquad + 2(1-\eta_{k-1})^2\langle \boldsymbol{\alpha}_i, \boldsymbol{\beta}_i \rangle\Big].
\end{aligned}
$$

In a deterministic setting, since each node has access to the entire data, the sequence of iterates and the full gradients are deterministic and the same (i.e., $\boldsymbol{x}_k^{(i)} = \boldsymbol{x}_k^{(j)}$ and $\nabla f(\boldsymbol{x}_k^{(i)}) = \nabla f(\boldsymbol{x}_k^{(j)})$ for $i, j \in [n]$) when each node has the same initial point. Due to unbiasedness of the local stochastic gradient $\nabla f(\boldsymbol{x}_k^{(i)}, \zeta_k^{(i)})$ and that $\boldsymbol{c}_k^{(i)}$ is obtained by a convex combination of full (deterministic) gradients under $\mathbb{E}_{.|\mathcal{F}_k}$, we have $\mathbb{E}_{.|\mathcal{F}_{k-1}}[\boldsymbol{c}_k^{(i)}] = \nabla f(\boldsymbol{x}_k^{(i)})$. Consequently, the two cross terms ($\langle \boldsymbol{\alpha}_i, \boldsymbol{\gamma}_i \rangle$ and $\langle \boldsymbol{\beta}_i, \boldsymbol{\gamma}_i \rangle$) vanish under $\mathbb{E}_{.|\mathcal{F}_{k-1}}$ in (a).

Due to the fact that $\mathbb{E}\left[\mathbb{E}_{.|\mathcal{F}_{k-1}}\left[\mathbb{E}_{.|\mathcal{F}_k}[\cdot]\right]\right] = \mathbb{E}[\cdot]$, we have

$$
\begin{aligned}
r_k^2 &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\|\nabla f(\boldsymbol{x}_k^{(i)}) - \boldsymbol{d}_k^{(i)}\|^2\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\mathbb{E}_{.|\mathcal{F}_{k-1}}\left[\mathbb{E}_{.|\mathcal{F}_k}\left[\|\nabla f(\boldsymbol{x}_k^{(i)}) - \boldsymbol{d}_k^{(i)}\|^2\right]\right]\right] \\
&\leq \frac{1}{n}\sum_{i=1}^{n}(1-\eta_{k-1})^2\mathbb{E}\|\boldsymbol{\alpha}_i\|^2 + (1-\eta_{k-1})^2\mathbb{E}\|\boldsymbol{\beta}_i\|^2 + \eta_{k-1}^2\mathbb{E}\|\boldsymbol{\gamma}_i\|^2 + \eta_{k-1}^2\mathbb{E}\|\boldsymbol{\omega}_i\|^2 \\
&\qquad + 2(1-\eta_{k-1})^2\mathbb{E}\langle \boldsymbol{\alpha}_i, \boldsymbol{\beta}_i \rangle \\
&\overset{(a)}{\leq} \frac{1}{n}\sum_{i=1}^{n}(1-\eta_{k-1})^2\mathbb{E}\|\boldsymbol{\alpha}_i\|^2 + (1-\eta_{k-1})^2\mathbb{E}\|\boldsymbol{\beta}_i\|^2 + \eta_{k-1}^2\mathbb{E}\|\boldsymbol{\gamma}_i\|^2 + \eta_{k-1}^2\mathbb{E}\|\boldsymbol{\omega}_i\|^2 \\
&\qquad + (1-\eta_{k-1})^2\mathbb{E}\left[\frac{1}{\eta_{k-1}}\|\boldsymbol{\alpha}_i\|^2 + \eta_{k-1}\|\boldsymbol{\beta}_i\|^2\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}(1-\eta_{k-1})^2(1+\frac{1}{\eta_{k-1}})\mathbb{E}\|\boldsymbol{\alpha}_i\|^2 + (1-\eta_{k-1})^2(1+\eta_{k-1})\mathbb{E}\|\boldsymbol{\beta}_i\|^2 + \eta_{k-1}^2\mathbb{E}\|\boldsymbol{\gamma}_i\|^2 + \eta_{k-1}^2\mathbb{E}\|\boldsymbol{\omega}_i\|^2,
\end{aligned}
$$

where (a) obtained using Young's inequality with $\eta = \eta_{k-1}$. For the case $k \geq T$, using the inequalities $(1-\eta_{k-1})(1+\eta_{k-1}) = (1-\eta_{k-1}^2) \leq 1$ and $(1-\eta_{k-1})^2(1+\frac{1}{\eta_{k-1}}) \leq \frac{1}{\eta_{k-1}}$, we have

$$
r_k^2 \leq \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\eta_{k-1}}\mathbb{E}\|\boldsymbol{\alpha}_i\|^2 + (1-\eta_{k-1})\mathbb{E}\|\boldsymbol{\beta}_i\|^2 + \eta_{k-1}^2\mathbb{E}\|\boldsymbol{\gamma}_i\|^2 + \eta_{k-1}^2\mathbb{E}\|\boldsymbol{\omega}_i\|^2. \tag{12}
$$

Consider the term,

$$\begin{aligned}
\mathbb{E}\|\boldsymbol{\omega}_i\|^2 &= \mathbb{E}\|\boldsymbol{z}_k^{(i)} - \boldsymbol{c}_k^{(i)}\|^2 \\
&= \mathbb{E}\left\|\sum_{j\in[n]} w_{ij}\nabla f(\boldsymbol{x}_k^{(j)},\zeta_k^{(j)}) - \sum_{j\in[n]} w_{ij}\boldsymbol{a}_k^{(j)}\right\|^2 \\
&\overset{(a)}{\leq} \sum_{j\in[n]} w_{ij}\mathbb{E}\left\|\nabla f(\boldsymbol{x}_k^{(j)},\zeta_k^{(j)}) - \boldsymbol{a}_k^{(j)}\right\|^2 \\
&\overset{(b)}{\leq} u_k^2,
\end{aligned} \tag{13}$$

where (a), follows from Jensen's inequality for convex functions and (b) follows from the definition of u_k^2. Substituting $\boldsymbol{\alpha}_i, \boldsymbol{\beta}_i$, Equation (13), and Lemma 1 in Equation (12), we have the following bound

$$\begin{aligned}
r_k^2 &\leq \frac{1}{\eta_{k-1}}\frac{1}{n}\sum_{i=1}^n \mathbb{E}\|\nabla f(\boldsymbol{x}_k^{(i)}) - \nabla f(\boldsymbol{x}_{k-1}^{(i)})\|^2 + \frac{(1-\eta_{k-1})}{n}\sum_{i=1}^n \mathbb{E}\|\nabla f(\boldsymbol{x}_{k-1}^{(i)}) - \boldsymbol{d}_{k-1}^{(i)}\|^2 + \eta_{k-1}^2\sigma^2 + \eta_{k-1}^2 u_k^2 \\
&\overset{(a)}{\leq} (1-\eta_{k-1})r_{k-1}^2 + \frac{1}{\eta_{k-1}}\frac{L^2}{n}\sum_{i=1}^n \mathbb{E}\left[\|\boldsymbol{x}_k^{(i)} - \boldsymbol{x}_{k-1}^{(i)}\|^2\right] + \eta_{k-1}^2\sigma^2 + \eta_{k-1}^2 u_k^2 \\
&\overset{(b)}{\leq} (1-\eta_{k-1})r_{k-1}^2 + \frac{1}{\eta_{k-1}}\gamma_{k-1}^2\frac{L^2D^2}{n} + \eta_{k-1}^2\sigma^2 + \eta_{k-1}^2 u_k^2,
\end{aligned}$$

where, we substitute $\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\|\nabla f(\boldsymbol{x}_{k-1}^{(i)}) - \boldsymbol{d}_{k-1}^{(i)}\|^2\right]$ with r_{k-1}^2 and use Assumption 1 in (a) and (b) follows from the Step 9 and Assumption 2. Using Lemma 3,

$$\begin{aligned}
r_k^2 &\leq (1-\eta_{k-1})r_{k-1}^2 + \frac{1}{\eta_{k-1}}\gamma_{k-1}^2 L^2D^2 + \eta_{k-1}^2\sigma^2 + \eta_{k-1}^2\frac{C_u}{n(k+2)^\alpha} \\
&\overset{(a)}{=} \left(1 - \frac{2\sqrt{n}}{(k+1)^\alpha}\right)r_{k-1}^2 + \frac{1}{\frac{2\sqrt{n}}{(k+1)^\alpha}}\frac{4}{(k+1)^{2\beta}}L^2D^2 + \frac{4n}{(k+1)^{2\alpha}}\sigma^2 + \frac{4n}{(k+1)^{2\alpha}}\frac{C_u}{n(k+2)^\alpha} \\
&\overset{(b)}{\leq} \left(1 - \frac{2\sqrt{n}}{(k+1)^\alpha}\right)r_{k-1}^2 + \frac{2L^2D^2 + 4n^{(3/2)}\sigma^2 + 4\sqrt{n}C_u}{\sqrt{n}(k+1)^{2\alpha}},
\end{aligned}$$

where we use the step-size $\gamma_{k-1} = \frac{2}{(k+1)^\beta}$ and the weight $\eta_{k-1} = \frac{2\sqrt{n}}{(k+1)^\alpha}$ in (a) and in (b), we use the fact that $\alpha \leq \frac{2}{3}\beta$. Using Lemma A6

$$\begin{aligned}
r_k^2 &\leq \frac{\max\{r_0^2, \frac{(2L^2D^2+4n^{(3/2)}\sigma^2+4\sqrt{n}C_u)/\sqrt{n}}{2\sqrt{n}-1}\}}{(k+2)^\alpha} \\
&\overset{(a)}{\leq} \frac{\max\{r_0^2, \frac{2L^2D^2+4n^{(3/2)}\sigma^2+4\sqrt{n}C_u}{n}\}}{(k+2)^\alpha} \\
&= \frac{\max\{nr_0^2, 2L^2D^2 + 4n^{(3/2)}\sigma^2 + 4\sqrt{n}C_u\}}{n(k+2)^\alpha} \\
&= \frac{C_r^2}{n(k+2)^\alpha},
\end{aligned}$$

where, (a) follows from the fact that $2\sqrt{n}-1 \geq \sqrt{n}$, with $n \geq 2$. Consider the term

$$\begin{aligned}
r_0^2 &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\|\nabla f(\boldsymbol{x}_0^{(i)}) - \boldsymbol{d}_0^{(i)}\|^2\right] \\
&\overset{(a)}{=} \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\|\nabla f(\boldsymbol{x}_0^{(i)}) - \boldsymbol{c}_0^{(i)}\|^2\right] \\
&\overset{(b)}{\leq} \sigma^2,
\end{aligned} \tag{14}$$

where, (a) follows from Step 8 with $\eta_0 = 1$ and (b) follows from Lemma 1. Using Equation (14), $\max\{nr_0^2, 2L^2D^2 + 4n^{(3/2)}\sigma^2 + 4\sqrt{n}C_u\} = 2L^2D^2 + 4n^{(3/2)}\sigma^2 + 4\sqrt{n}C_u := C_r^2$.

Now consider the case $k \leq T$,

$$r_k^2 = \mathbb{E}\|\nabla f(\boldsymbol{x}_k^{(i)}) - \boldsymbol{d}_k^{(i)}\|^2 = \mathbb{E}\|\nabla f(\boldsymbol{x}_k^{(i)}) - \boldsymbol{c}_k^{(i)}\|^2 \leq \ \sigma^2,$$

where (a) follows from Lemma 1. Consider the term

$$\begin{aligned}
\frac{4n^{3/2}\sigma^2}{n(T+2)^\alpha} &= \frac{4\sqrt{n}\sigma^2}{\left((2\sqrt{n})^{1/\alpha} - 2 + 2\right)^\alpha} \\
&= \frac{4\sqrt{n}\sigma^2}{2\sqrt{n}} = 2\sigma^2.
\end{aligned}$$

From the above expression we see that $\sigma^2 < \frac{4n^{3/2}\sigma^2}{n(k+2)^\alpha}$ and $r_k^2 < \frac{C_r^2}{n(k+2)^\alpha}$ $\forall k \leq T$, therefore we bound $r_k^2 \leq \frac{C_r^2}{n(k+2)^\alpha}$ $\forall k$.

□

D. Proof of Lemma 5: Generalized Suboptimality Gap

Lemma 5. For an L-smooth convex function $f : \mathcal{R}^d \to \mathcal{R}$, convex and compact constraint set $\mathcal{C}$ with diameter D, the suboptimality gap $h_k = \mathbb{E}[f(\bar{\boldsymbol{x}}_k) - f_{\text{opt}}]$ satisfies

$$h_{k+1} \leq (1-\gamma_k)h_k + \gamma_k^2\frac{LD^2}{2} + \underbrace{\gamma_k 2LDq_k}_{\text{decentralization}} + \underbrace{\gamma_k 2Dr_k}_{\text{stochasticity}}.$$

Proof. We derive an upper bound on the suboptimality gap due to inexact updates using L-smoothness of f [cf. Assumption 1] as

$$\begin{aligned}
f(\bar{\boldsymbol{x}}_{k+1}) &\leq f(\bar{\boldsymbol{x}}_k) + \langle\nabla f(\bar{\boldsymbol{x}}_k), \bar{\boldsymbol{x}}_{k+1} - \bar{\boldsymbol{x}}_k\rangle + \frac{L}{2}\|\bar{\boldsymbol{x}}_{k+1} - \bar{\boldsymbol{x}}_k\|_2^2 \\
&\overset{(a)}{=} f(\bar{\boldsymbol{x}}_k) + \gamma_k\langle\nabla f(\bar{\boldsymbol{x}}_k), \bar{\boldsymbol{s}}_k - \bar{\boldsymbol{x}}_k\rangle + \gamma_k^2\frac{L}{2}\|\bar{\boldsymbol{s}}_k - \bar{\boldsymbol{x}}_k\|_2^2 \\
&\overset{(b)}{=} f(\bar{\boldsymbol{x}}_k) + \gamma_k\langle\nabla f(\bar{\boldsymbol{x}}_k), \bar{\boldsymbol{s}}_k - \bar{\boldsymbol{x}}_k\rangle + \gamma_k^2\frac{L}{2}D^2,
\end{aligned} \tag{15}$$

where (a) follows from the FW update Step 9 and (b) follows from Assumption 2.

Let us introduce the average variance-reduced gradient $\bar{\boldsymbol{d}}_k = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{d}_k^{(i)}$ and the average deterministic gradient $\overline{\nabla f}_k = \frac{1}{n}\sum_{i=1}^{n} \nabla f(\boldsymbol{x}_k^{(i)})$ in $\langle \nabla f(\bar{\boldsymbol{x}}_k), \bar{\boldsymbol{s}}_k - \bar{\boldsymbol{x}}_k\rangle$ as

$$\begin{aligned}
\langle \nabla f(\bar{\boldsymbol{x}}_k), \bar{\boldsymbol{s}}_k - \bar{\boldsymbol{x}}_k\rangle &= \langle \bar{\boldsymbol{d}}_k, \bar{\boldsymbol{s}}_k - \bar{\boldsymbol{x}}_k\rangle + \langle \nabla f(\bar{\boldsymbol{x}}_k) - \bar{\boldsymbol{d}}_k, \bar{\boldsymbol{s}}_k - \bar{\boldsymbol{x}}_k\rangle \\
&= \langle \bar{\boldsymbol{d}}_k, \bar{\boldsymbol{s}}_k - \bar{\boldsymbol{x}}_k\rangle + \langle \nabla f(\bar{\boldsymbol{x}}_k) - \overline{\nabla f}_k, \bar{\boldsymbol{s}}_k - \bar{\boldsymbol{x}}_k\rangle \\
&\quad + \langle \overline{\nabla f}_k - \bar{\boldsymbol{d}}_k, \bar{\boldsymbol{s}}_k - \bar{\boldsymbol{x}}_k\rangle \\
&= \langle \bar{\boldsymbol{d}}_k, \bar{\boldsymbol{s}}_k - \bar{\boldsymbol{x}}_k\rangle + \frac{1}{n}\sum_{i=1}^{n} \langle \nabla f(\bar{\boldsymbol{x}}_k) - \nabla f(\boldsymbol{x}_k^{(i)}), \bar{\boldsymbol{s}}_k - \bar{\boldsymbol{x}}_k\rangle \\
&\quad + \frac{1}{n}\sum_{i=1}^{n} \langle \nabla f(\boldsymbol{x}_k^{(i)}) - \boldsymbol{d}_k^{(i)}, \bar{\boldsymbol{s}}_k - \bar{\boldsymbol{x}}_k\rangle.
\end{aligned}$$

Using Cauchy-Schwarz inequality, we get

$$\begin{aligned}
\langle \nabla f(\bar{\boldsymbol{x}}_k), \bar{\boldsymbol{s}}_k - \bar{\boldsymbol{x}}_k\rangle &\le \langle \bar{\boldsymbol{d}}_k, \bar{\boldsymbol{s}}_k - \bar{\boldsymbol{x}}_k\rangle + \frac{1}{n}\sum_{i=1}^{n} \|\nabla f(\bar{\boldsymbol{x}}_k) - \nabla f(\boldsymbol{x}_k^{(i)})\|_2 \|\bar{\boldsymbol{s}}_k - \bar{\boldsymbol{x}}_k\|_2 \\
&\quad + \frac{1}{n}\sum_{i=1}^{n} \|\nabla f(\boldsymbol{x}_k^{(i)}) - \boldsymbol{d}_k^{(i)}\|_2 \|\bar{\boldsymbol{s}}_k - \bar{\boldsymbol{x}}_k\|_2 \\
&\overset{(a)}{\le} \langle \bar{\boldsymbol{d}}_k, \bar{\boldsymbol{s}}_k - \bar{\boldsymbol{x}}_k\rangle + \frac{1}{n}\sum_{i=1}^{n} \|\nabla f(\bar{\boldsymbol{x}}_k) - \nabla f(\boldsymbol{x}_k^{(i)})\|_2 D \\
&\quad + \frac{1}{n}\sum_{i=1}^{n} \|\nabla f(\boldsymbol{x}_k^{(i)}) - \boldsymbol{d}_k^{(i)}\|_2 D \\
&\overset{(b)}{\le} \langle \bar{\boldsymbol{d}}_k, \bar{\boldsymbol{s}}_k - \bar{\boldsymbol{x}}_k\rangle + \frac{1}{n}\sum_{i=1}^{n} LD\|\bar{\boldsymbol{x}}_k - \boldsymbol{x}_k^{(i)}\|_2 + \frac{1}{n}\sum_{i=1}^{n} D\|\nabla f(\boldsymbol{x}_k^{(i)}) - \boldsymbol{d}_k^{(i)}\|_2, \qquad (16)
\end{aligned}$$

where (a) follows from Assumption 2 and (b) follows from Assumption 1.

Next we bound $\langle \bar{\boldsymbol{d}}_k, \bar{\boldsymbol{s}}_k - \bar{\boldsymbol{x}}_k \rangle$ by introducing the terms $\nabla f(\bar{\boldsymbol{x}}_k)$ and $\overline{\nabla f}_k$ as

$$\begin{aligned}
\langle \bar{\boldsymbol{d}}_k, \bar{\boldsymbol{s}}_k - \bar{\boldsymbol{x}}_k \rangle &= \langle \nabla f(\bar{\boldsymbol{x}}_k), \bar{\boldsymbol{s}}_k - \bar{\boldsymbol{x}}_k \rangle + \langle \overline{\nabla f}_k - \nabla f(\bar{\boldsymbol{x}}_k), \bar{\boldsymbol{s}}_k - \bar{\boldsymbol{x}}_k \rangle \\
&\qquad + \langle \bar{\boldsymbol{d}}_k - \overline{\nabla f}_k, \bar{\boldsymbol{s}}_k - \bar{\boldsymbol{x}}_k \rangle \\
&= \langle \nabla f(\bar{\boldsymbol{x}}_k), \bar{\boldsymbol{s}}_k - \bar{\boldsymbol{x}}_k \rangle + \frac{1}{n}\sum_{i=1}^{n} \langle \nabla f(\boldsymbol{x}_k^{(i)}) - \nabla f(\bar{\boldsymbol{x}}_k), \bar{\boldsymbol{s}}_k - \bar{\boldsymbol{x}}_k \rangle \\
&\qquad + \frac{1}{n}\sum_{i=1}^{n} \langle \boldsymbol{d}_k^{(i)} - \nabla f(\boldsymbol{x}_k^{(i)}), \bar{\boldsymbol{s}}_k - \bar{\boldsymbol{x}}_k \rangle \\
&\overset{(a)}{\leq} \langle \nabla f(\bar{\boldsymbol{x}}_k), \bar{\boldsymbol{s}}_k - \bar{\boldsymbol{x}}_k \rangle + \frac{1}{n}\sum_{i=1}^{n} \|\nabla f(\boldsymbol{x}_k^{(i)}) - \nabla f(\bar{\boldsymbol{x}}_k)\|_2 \|\bar{\boldsymbol{s}}_k - \bar{\boldsymbol{x}}_k\|_2 \\
&\qquad + \frac{1}{n}\sum_{i=1}^{n} \|\boldsymbol{d}_k^{(i)} - \nabla f(\boldsymbol{x}_k^{(i)})\|_2 \|\bar{\boldsymbol{s}}_k - \bar{\boldsymbol{x}}_k\|_2 \\
&\overset{(b)}{\leq} \langle \nabla f(\bar{\boldsymbol{x}}_k), \bar{\boldsymbol{s}}_k - \bar{\boldsymbol{x}}_k \rangle + \frac{1}{n}\sum_{i=1}^{n} \|\nabla f(\boldsymbol{x}_k^{(i)}) - \nabla f(\bar{\boldsymbol{x}}_k)\|_2 D \\
&\qquad + \frac{1}{n}\sum_{i=1}^{n} \|\boldsymbol{d}_k^{(i)} - \nabla f(\boldsymbol{x}_k^{(i)})\|_2 D \\
&\overset{(c)}{\leq} \langle \nabla f(\bar{\boldsymbol{x}}_k), \bar{\boldsymbol{s}}_k - \bar{\boldsymbol{x}}_k \rangle + \frac{1}{n}\sum_{i=1}^{n} LD \|\boldsymbol{x}_k^{(i)} - \bar{\boldsymbol{x}}_k\|_2 \\
&\qquad + \frac{1}{n}\sum_{i=1}^{n} D \|\boldsymbol{d}_k^{(i)} - \nabla f(\boldsymbol{x}_k^{(i)})\|_2,
\end{aligned} \tag{17}$$

where we employ the Cauchy-Schwarz inequality in (a), Assumption 1 in (b), and Assumption 2 in (c).

Using Equation (16) and Equation (17) in Equation (15), we get

$$\begin{aligned}
f(\bar{\boldsymbol{x}}_{k+1}) &\leq f(\bar{\boldsymbol{x}}_k) + \gamma_k \langle \nabla f(\bar{\boldsymbol{x}}_k), \bar{\boldsymbol{s}}_k - \bar{\boldsymbol{x}}_k \rangle + \gamma_k^2 \frac{L}{2} D^2 + \gamma_k \frac{1}{n}\sum_{i=1}^{n} 2LD \|\boldsymbol{x}_k^{(i)} - \bar{\boldsymbol{x}}_k\|_2 \\
&\qquad + \gamma_k \frac{1}{n}\sum_{i=1}^{n} 2D \|\boldsymbol{d}_k^{(i)} - \nabla f(\boldsymbol{x}_k^{(i)})\|_2 \\
&\overset{(a)}{\leq} f(\bar{\boldsymbol{x}}_k) + \gamma_k \langle \nabla f(\bar{\boldsymbol{x}}_k), \bar{\boldsymbol{s}}_k - \bar{\boldsymbol{x}}_k \rangle + \gamma_k^2 \frac{L}{2} D^2 + \gamma_k\, 2LD \sqrt{\frac{1}{n}\sum_{i=1}^{n} \|\boldsymbol{x}_k^{(i)} - \bar{\boldsymbol{x}}_k\|_2^2} \\
&\qquad + \gamma_k\, 2D \sqrt{\frac{1}{n}\sum_{i=1}^{n} \|\boldsymbol{d}_k^{(i)} - \nabla f(\boldsymbol{x}_k^{(i)})\|_2^2},
\end{aligned}$$

where the above expression follows from the concavity of $\sqrt{(\cdot)}$. Finally, we have

$$
\begin{aligned}
\mathbb{E}\Big[f(\bar{\boldsymbol{x}}_{k+1}) - f_{\text{opt}}\Big] \leq \mathbb{E}\Big[& f(\bar{\boldsymbol{x}}_k) - f_{\text{opt}} + \gamma_k \langle \nabla f(\bar{\boldsymbol{x}}_k), \bar{\boldsymbol{s}}_k - \bar{\boldsymbol{x}}_k \rangle + \gamma_k^2 \frac{L}{2} D^2 \\
& + \gamma_k 2LD \sqrt{\frac{1}{n} \sum_{i=1}^{n} \|\boldsymbol{x}_k^{(i)} - \bar{\boldsymbol{x}}_k\|_2^2} \\
& + \gamma_k 2D \sqrt{\frac{1}{n} \sum_{i=1}^{n} \|\boldsymbol{d}_k^{(i)} - \nabla f(\boldsymbol{x}_k^{(i)})\|_2^2}\Big] \\
\overset{(a)}{\leq} \mathbb{E}\Big[& f(\bar{\boldsymbol{x}}_k) - f_{\text{opt}}\Big] + \gamma_k \mathbb{E}\Big[\langle \nabla f(\bar{\boldsymbol{x}}_k), \bar{\boldsymbol{s}}_k - \bar{\boldsymbol{x}}_k \rangle\Big] + \gamma_k^2 \frac{L}{2} D^2 \\
& + \gamma_k 2LD \sqrt{\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[\|\boldsymbol{x}_k^{(i)} - \bar{\boldsymbol{x}}_k\|_2^2\right]} \\
& + \gamma_k 2D \sqrt{\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\Big[\|\boldsymbol{d}_k^{(i)} - \nabla f(\boldsymbol{x}_k^{(i)})\|_2^2\Big]},
\end{aligned}
$$

where we use Jensen's inequality for concave functions in (a). Thus we have

$$
\begin{aligned}
h_{k+1} &\leq h_k + \gamma_k \mathbb{E}[\langle \nabla f(\bar{\boldsymbol{x}}_k), \bar{\boldsymbol{s}}_k - \bar{\boldsymbol{x}}_k \rangle] + \gamma_k^2 \frac{L}{2} D^2 + \gamma_k 2LD q_k + \gamma_k 2D r_k \\
&\overset{(a)}{\leq} h_k - \gamma_k \mathbb{E}[g_k] + \gamma_k^2 \frac{L}{2} D^2 + \gamma_k 2LD q_k + \gamma_k 2D r_k \\
&\overset{(b)}{\leq} (1 - \gamma_k) h_k + \gamma_k^2 \frac{L}{2} D^2 + \gamma_k 2LD q_k + \gamma_k 2D r_k,
\end{aligned} \tag{18}
$$

where (a) follows from the definition of the Frank-Wolfe duality gap

$$
g_k = \max_{\boldsymbol{s} \in \mathcal{C}} \langle \nabla f(\overline{\boldsymbol{x}}_k), \overline{\boldsymbol{x}}_k - \boldsymbol{s} \rangle
$$

and (b) is because of the fact that the duality gap upper bounds the primal suboptimality gap for convex functions. □

E. Proof of Theorem 1: Convergence Rate for Convex Objectives

Theorem 1. For an L-smooth convex function $f : \mathcal{R}^d \to \mathcal{R}$, convex and compact constraint set $\mathcal{C}$ with diameter D, the average suboptimality gap at the kth iterate with $\gamma_k = \frac{2}{k+2}$ and $\alpha = 2/3$ satisfies

$$
\mathbb{E}[f(\bar{\boldsymbol{x}}_k) - f_{\text{opt}}] \leq \frac{2LD^2}{(k+1)} \left(-1 + \log\left(k+1\right)\right) + \frac{16LD^2}{\sqrt{k+1}} + \frac{6C_r D}{\sqrt{n}(k+1)^{1/3}}, \tag{19}
$$

where $C_r^2 = 2L^2D^2 + 4n^{(3/2)}\sigma^2 + 4\sqrt{n}C_u$.

Proof. Consider the bound on the suboptimality gap in Lemma 5

$$
h_{k+1} \leq (1 - \gamma_k) h_k + \gamma_k^2 \frac{L}{2} D^2 + \gamma_k 2LD q_k + \gamma_k 2D r_k.
$$

Unrolling the bound up to K iterations

$$
\begin{aligned}
h_K \leq & \prod_{j=0}^{K-1}(1-\gamma_j)h_0 + \frac{LD^2}{2}\sum_{i=0}^{K-1}\gamma_i^2\prod_{j=i+1}^{K-1}(1-\gamma_j) + 2LD\sum_{i=0}^{K-1}\gamma_i q_i\prod_{j=i+1}^{K-1}(1-\gamma_j) \\
& + 2D\sum_{i=0}^{K-1}\gamma_i r_i\prod_{j=i+1}^{K-1}(1-\gamma_j) \\
\overset{(a)}{=} & \, 0 + \frac{LD^2}{2}\sum_{i=0}^{K-1}\gamma_i^2\prod_{j=i+1}^{K-1}(1-\gamma_j) + 2LD\sum_{i=0}^{K-1}\gamma_i q_i\prod_{j=i+1}^{K-1}(1-\gamma_j) + 2D\sum_{i=0}^{K-1}\gamma_i r_i\prod_{j=i+1}^{K-1}(1-\gamma_j) \\
\overset{(b)}{\leq} & \frac{LD^2}{2}\sum_{i=0}^{K-1}\frac{4}{(i+2)^2}\prod_{j=i+1}^{K-1}(1-\frac{2}{j+2}) + 2LD\sum_{i=0}^{K-1}\frac{2}{i+2}\frac{2D}{\sqrt{i+1}}\prod_{j=i+1}^{K-1}(1-\frac{2}{j+2}) \\
& + 2D\sum_{i=0}^{K-1}\frac{2}{i+2}\frac{C_r}{\sqrt{n}(i+2)^{1/3}}\prod_{j=i+1}^{K-1}(1-\frac{2}{j+2})
\end{aligned}
$$

where we invoke the fact that $\gamma_0 = 1$ and $1-\gamma_j \leq 1, \ \forall j$ to obtain (a). We substitute for q_k from Lemma 2 and r_k from Lemma 3, with $\alpha = 2/3$ and use the fact that step-size of $\gamma_j = \frac{2}{j+2}$ in (b). Consider the term

$$
\begin{aligned}
\sum_{i=0}^{K-1}\frac{4}{(i+2)^2}\prod_{j=i+1}^{K-1}(1-\frac{2}{j+2}) &= \frac{4}{(K+1)^2} + \frac{4}{K^2}\frac{K-1}{K+1} + \frac{4}{(K-1)^2}\frac{K-2}{K}\frac{K-1}{K+1} + \dots \\
&= \frac{1}{K+1}\left(\frac{4}{K+1} + \frac{4(K-1)}{K^2} + \frac{4(K-2)}{(K-1)K} + \dots\right) \\
&\leq \frac{1}{K+1}\left(\frac{4}{K+1} + \frac{4}{K} + \frac{4}{(K-1)} + \dots\right) \\
&= \frac{1}{K+1}\sum_{i=0}^{K-1}\frac{4}{i+2}.
\end{aligned}
$$

Now consider the term

$$
\begin{aligned}
\sum_{i=0}^{K-1}\frac{2}{i+2}\frac{2D}{\sqrt{i+1}}\prod_{j=i+1}^{K-1}(1-\frac{2}{j+2}) &= \frac{4D}{(K+1)\sqrt{K}} + \frac{4D}{K\sqrt{K-1}}\frac{K-1}{K+1} + \frac{4D}{(K-1)\sqrt{K-2}}\frac{K-2}{K}\frac{K-1}{K+1} + \dots \\
&= \frac{1}{K+1}\left(\frac{4D}{\sqrt{K}} + \frac{4D(K-1)}{K\sqrt{K-1}} + \frac{4D(K-2)}{K\sqrt{K-2}} + \dots\right) \\
&\leq \frac{1}{K+1}\left(\frac{4D}{\sqrt{K}} + \frac{4D}{\sqrt{K-1}} + \frac{4D}{\sqrt{K-2}} + \dots\right) \\
&= \frac{1}{K+1}\sum_{i=0}^{K-1}\frac{4D}{\sqrt{i+1}}.
\end{aligned}
$$

Similarly, we have

$$
\sum_{i=0}^{K-1}\frac{2}{i+2}\frac{C_r}{\sqrt{n}(i+2)^{1/3}}\prod_{j=i+1}^{K-1}(1-\frac{2}{j+2}) = \frac{1}{K+1}\sum_{i=0}^{K-1}\frac{2C_r}{\sqrt{n}(i+2)^{1/3}}.
$$

Using the above inequalities we have

$$
h_K \leq \frac{LD^2}{2(K+1)}\sum_{i=0}^{K-1}\frac{4}{(i+2)} + \frac{8LD^2}{K+1}\sum_{i=0}^{K-1}\frac{1}{\sqrt{i+1}} + \frac{4C_rD}{\sqrt{n}(K+1)}\sum_{i=0}^{K-1}\frac{1}{(i+2)^{1/3}}.
$$

Now let us use the following inequalities

$$\sum_{j=0}^{K-1} \frac{1}{j+2} \leq -1 + \int_{j=1}^{K+1} \frac{1}{j} dj$$
$$= -1 + \log(K+1),$$
$$\sum_{j=0}^{K-1} \frac{1}{\sqrt{j+1}} \leq \int_{j=0}^{K} \frac{1}{\sqrt{j}} dj$$
$$= 2\sqrt{K},$$

and

$$\sum_{j=0}^{K-1} \frac{1}{(j+2)^{1/3}} \leq \int_{j=0}^{K+1} \frac{1}{j^{1/3}} dj$$
$$= \frac{(K+1)^{1-1/3}}{1-1/3}$$
$$= \frac{3}{2}(K+1)^{2/3}.$$

Finally, we have

$$h_K \leq \frac{2LD^2}{(K+1)}\left(-1 + \log(K+1)\right) + \frac{8LD^2}{K+1}2\sqrt{K} + \frac{4C_rD}{\sqrt{n}(K+1)}\frac{3}{2}(K+1)^{2/3}$$
$$\leq \frac{2LD^2}{(K+1)}\left(-1 + \log(K+1)\right) + \frac{16LD^2}{\sqrt{K+1}} + \frac{6C_rD}{\sqrt{n}(K+1)^{1/3}}.$$

□

F. Proof of Lemma 6: Convergence of Average Duality Gap

Lemma 6. For an L-smooth but non-convex $f : \mathcal{R}^d \to \mathcal{R}$, convex and compact constraint set $\mathcal{C}$, the distance between the duality gaps g_k and $\hat{\bar{g}}_k$, in expectation satisfies

$$\mathbb{E}\left[|g_k - \hat{\bar{g}}_k|\right] \leq LDq_k + Dr_k. \tag{20}$$

Proof. To obtain a bound on the distance between the true duality gap and the average duality gap at each node, we use the definition of the duality gap at node i

$$
\begin{aligned}
g_k &= \max_{\boldsymbol{s}\in\mathcal{C}} \langle \nabla f(\bar{\boldsymbol{x}}_k), \bar{\boldsymbol{x}}_k - \boldsymbol{s}\rangle = \max_{\boldsymbol{s}\in\mathcal{C}} \frac{1}{n}\sum_{i=1}^{n} \langle \nabla f(\bar{\boldsymbol{x}}_k), \boldsymbol{x}_k^{(i)} - \boldsymbol{s}\rangle \\
&\le \frac{1}{n}\sum_{i=1}^{n} \max_{\boldsymbol{s}\in\mathcal{C}} \langle \nabla f(\bar{\boldsymbol{x}}_k), \boldsymbol{x}_k^{(i)} - \boldsymbol{s}\rangle \\
&\overset{(a)}{\le} \frac{1}{n}\sum_{i=1}^{n} \langle \nabla f(\bar{\boldsymbol{x}}_k), \boldsymbol{x}_k^{(i)} - \boldsymbol{s}_k^{(i)}\rangle \\
&\overset{(b)}{=} \frac{1}{n}\sum_{i=1}^{n} \langle \nabla f(\boldsymbol{x}_k^{(i)}), \boldsymbol{x}_k^{(i)} - \boldsymbol{s}_k^{(i)}\rangle + \frac{1}{n}\sum_{i=1}^{n} \langle \nabla f(\bar{\boldsymbol{x}}_k) - \nabla f(\boldsymbol{x}_k^{(i)}), \boldsymbol{x}_k^{(i)} - \boldsymbol{s}_k^{(i)}\rangle \\
&\overset{(c)}{=} \frac{1}{n}\sum_{i=1}^{n} \langle \boldsymbol{d}_k^{(i)}, \boldsymbol{x}_k^{(i)} - \boldsymbol{s}_k^{(i)}\rangle + \frac{1}{n}\sum_{i=1}^{n} \langle \nabla f(\boldsymbol{x}_k^{(i)}) - \boldsymbol{d}_k^{(i)}, \boldsymbol{x}_k^{(i)} - \boldsymbol{s}_k^{(i)}\rangle \\
&\qquad + \frac{1}{n}\sum_{i=1}^{n} \langle \nabla f(\bar{\boldsymbol{x}}_k) - \nabla f(\boldsymbol{x}_k^{(i)}), \boldsymbol{x}_k^{(i)} - \boldsymbol{s}_k^{(i)}\rangle,
\end{aligned}
$$

where we use the solution to the linear program $\boldsymbol{s}_k^{(i)}$ in (a). We introduce the full gradient $\nabla f(\boldsymbol{x}_k^{(i)})$ in (b) and the variance-reduced stochastic gradient direction $\boldsymbol{d}_k^{(i)}$ in (c). From the definition of $\hat{\bar{g}}_k$ and the Cauchy-Scharwz inequality, we have

$$
\begin{aligned}
g_k &\le \hat{\bar{g}}_k + \frac{1}{n}\sum_{i=1}^{n} \|\nabla f(\boldsymbol{x}_k^{(i)}) - \boldsymbol{d}_k^{(i)}\|_2 \|\boldsymbol{x}_k^{(i)} - \boldsymbol{s}_k^{(i)}\|_2 \\
&\qquad + \frac{1}{n}\sum_{i=1}^{n} \|\nabla f(\bar{\boldsymbol{x}}_k) - \nabla f(\boldsymbol{x}_k^{(i)})\|_2 \|\boldsymbol{x}_k^{(i)} - \boldsymbol{s}_k^{(i)}\|_2 \\
&\overset{(a)}{\le} \hat{\bar{g}}_k + \frac{1}{n}\sum_{i=1}^{n} \|\nabla f(\boldsymbol{x}_k^{(i)}) - \boldsymbol{d}_k^{(i)}\|_2 D + \frac{1}{n}\sum_{i=1}^{n} \|\nabla f(\bar{\boldsymbol{x}}_k) - \nabla f(\boldsymbol{x}_k^{(i)})\|_2 D \\
&\overset{(b)}{\le} \hat{\bar{g}}_k + \frac{D}{n}\sum_{i=1}^{n} \|\nabla f(\boldsymbol{x}_k^{(i)}) - \boldsymbol{d}_k^{(i)}\|_2 + \frac{LD}{n}\sum_{i=1}^{n} \|\bar{\boldsymbol{x}}_k - \boldsymbol{x}_k^{(i)}\|_2 \\
\mathbb{E}[g_k] &\le \mathbb{E}[\hat{\bar{g}}_k] + Dr_k + LDq_k
\end{aligned}
$$

where we use Assumption 2 in (a) and Assumption 1 in (b). □

G. Proof of Theorem 2: Convergence Rate for Non-convex Objectives

Theorem 2. For an L-smooth but non-convex function $f : \mathcal{R}^d \to \mathcal{R}$, over a convex and compact constraint set $\mathcal{C}$ with diameter D, the average FW duality gap $\mathbb{E}\left[\bar{g}_K\right]$, with $\beta = 2/3$ and $\alpha = 4/9$ satisfies

$$
\begin{aligned}
\mathbb{E}\left[\bar{g}_K\right] &= \frac{1}{K+1}\sum_{k=0}^{K} \mathbb{E}\left[g_k\right] \\
&\le \frac{\mathbb{E}\left[f(\bar{\boldsymbol{x}}_0) - f(\bar{\boldsymbol{x}}_{K+1})\right] + 6LD^2}{(K+1)^{-1/3}} + \frac{LD^2}{2(K+1)^{-2/3}} + \frac{3C_r D}{\sqrt{n}(K+2)^{-2/9}}.
\end{aligned}
$$

Proof. From the bound on suboptimality gap in Equation (18), we have

$$h_{k+1} \leq h_k - \gamma_k \mathbb{E}[g_k] + \gamma_k^2 \frac{LD^2}{2} + \gamma_k 2LDq_k + \gamma_k 2Dr_k$$
$$\gamma_k \mathbb{E}[g_k] \leq \mathbb{E}\left[f(\bar{\boldsymbol{x}}_k) - f(\bar{\boldsymbol{x}}_{k+1})\right] + \gamma_k^2 \frac{LD^2}{2} + \gamma_k 2LDq_k + \gamma_k 2Dr_k.$$

Summing from $k = 0, \ldots, K$, $f(\bar{\boldsymbol{x}}_k) - f(\bar{\boldsymbol{x}}_{k+1})$ telescopes to

$$\sum_{k=0}^{K} \gamma_k \mathbb{E}[g_k] \leq \mathbb{E}\left[f(\bar{\boldsymbol{x}}_0) - f(\bar{\boldsymbol{x}}_{K+1})\right] + \sum_{k=0}^{K} \gamma_k^2 \frac{LD^2}{2} + \sum_{k=0}^{K} \gamma_k 2LDq_k + \sum_{k=0}^{K} \gamma_k 2Dr_k.$$

Consider

$$\mathbb{E}\left[\bar{g}_K\right] = \frac{1}{K+1} \sum_{k=0}^{K} \mathbb{E}\left[g_k\right] \tag{21}$$
$$\overset{(a)}{\leq} \frac{1}{(K+1)\cdot(K+2)^{-2/3}} \Bigg[\mathbb{E}\left[f(\bar{\boldsymbol{x}}_0) - f(\bar{\boldsymbol{x}}_{K+1})\right] + \frac{LD^2}{2} \sum_{k=0}^{K} (K+2)^{-4/3} + 2LD \sum_{k=0}^{K} (K+2)^{-2/3} q_k \tag{22}$$
$$+ 2D \sum_{k=0}^{K} (K+2)^{-2/3} r_k \Bigg]$$
$$\overset{(b)}{\leq} \frac{1}{(K+1)^{1/3}} \Bigg[\mathbb{E}\left[f(\bar{\boldsymbol{x}}_0) - f(\bar{\boldsymbol{x}}_{K+1})\right] + \frac{LD^2}{2} \sum_{k=0}^{K} (K+2)^{-4/3} + 2LD \sum_{k=0}^{K} (K+2)^{-2/3} \frac{2D}{(k+2)^{1/3}} \tag{23}$$
$$+ 2D \sum_{k=0}^{K} (K+2)^{-2/3} \frac{C_r}{\sqrt{n}(k+2)^{2/9}} \Bigg], \tag{24}$$

where (a) follows by substituting $\gamma_k = (K+2)^{-2/3}$ and (b) follows by substituting for r_k and q_k with $\alpha = 4/9$. Now consider

$$\sum_{j=0}^{K} \frac{1}{(j+2)^{2/9}} \leq \int_{j=1}^{K+2} \frac{1}{j^{2/9}} dj$$
$$= \frac{(K+2)^{1-2/9}}{1-2/9}$$
$$= \frac{9}{7}(K+2)^{7/9}.$$

Using the above inequlaity in Equation (24)

$$\mathbb{E}\left[\bar{g}_K\right] \leq \frac{1}{(K+1)^{1/3}} \Bigg[\mathbb{E}\left[f(\bar{\boldsymbol{x}}_0) - f(\bar{\boldsymbol{x}}_{K+1})\right] + \frac{LD^2}{2}(K+1)\cdot(K+2)^{-4/3} + 4LD^2(K+2)^{-2/3}\frac{3}{2}(K+2)^{2/3}$$
$$+ \frac{2C_rD}{\sqrt{n}}(K+2)^{-2/3}\frac{9}{7}(K+2)^{7/9}\Bigg]$$
$$\leq \frac{1}{(K+1)^{1/3}} \Bigg[\mathbb{E}\left[f(\bar{\boldsymbol{x}}_0) - f(\bar{\boldsymbol{x}}_{K+1})\right] + \frac{LD^2}{2}(K+1)^{-1/3} + 6LD^2 + \frac{2C_rD}{\sqrt{n}}\frac{9}{7}(K+2)^{1/9}\Bigg]$$
$$= \frac{\mathbb{E}\left[f(\bar{\boldsymbol{x}}_0) - f(\bar{\boldsymbol{x}}_{K+1})\right]}{(K+1)^{1/3}} + \frac{LD^2}{2(K+1)^{2/3}} + \frac{6LD^2}{(K+1)^{1/3}} + \frac{3C_rD}{\sqrt{n}(K+2)^{2/9}}$$
$$= \frac{\mathbb{E}\left[f(\bar{\boldsymbol{x}}_0) - f(\bar{\boldsymbol{x}}_{K+1})\right] + 6LD^2}{(K+1)^{1/3}} + \frac{LD^2}{2(K+1)^{2/3}} + \frac{3C_rD}{\sqrt{n}(K+2)^{2/9}}.$$

∎

REFERENCES

[1] F. Bénézit, V. Blondel, P. Thiran, J. Tsitsiklis, and M. Vetterli, "Weighted gossip: Distributed averaging using non-doubly stochastic matrices," in *Proceedings of the IEEE International Symposium on Information Theory*, 2010, pp. 1753–1757.

[2] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah, "Gossip algorithms: design, analysis and applications," in *Proceedings IEEE 24th Annual Joint Conference of the IEEE Computer and Communications Societies.*, vol. 3, 2005, pp. 1653–1664 vol. 3.

[3] A. Mokhtari, H. Hassani, and A. Karbasi, "Stochastic conditional gradient methods: From convex minimization to submodular maximization," 2018.