A One-Sample Decentralized Proximal Algorithm for Non-Convex Stochastic Composite Optimization (Supplementary Material)

Tesi Xiao[1] **Xuxing Chen**[2] **Krishnakumar Balasubramanian**[1] **Saeed Ghadimi**[3]

[1]Department of Statistics, University of California, Davis
[2]Department of Mathematics, University of California, Davis
[3]Department of Management Sciences, University of Waterloo

Notations. $\|\cdot\|$ denotes the ℓ_2-norm for vectors and Frobenius norm for matrices. $\|\cdot\|_2$ denotes the spectral norm for matrices. $\mathbf{1}$ represents the all-one vector, and $\mathbf{I}$ is the identity matrix as a standard practice. We identify vectors at agent i in the subscript and use the superscript for the algorithm step. For example, the optimization variable of agent i at step k is denoted as x_i^k, and z_i^k is the corresponding dual variable. We use uppercase bold letters to represent the matrix that collects all the variables from nodes (corresponding lowercase) as columns. We add an overbar to a letter to denote the average over all nodes. For example, we denote the optimization variables over all nodes at step k as

$$\mathbf{X}_k = \left[x_1^k, \ldots, x_n^k\right].$$

The corresponding average over all notes can be thereby defined as

$$\bar{x}^k = \frac{1}{n}\sum_{i=1}^{n} x_i^k = \frac{1}{n}\mathbf{X}_k\mathbf{1},$$
$$\bar{\mathbf{X}}_k = [\bar{x}^k, \ldots, \bar{x}^k] = \bar{x}^k\mathbf{1}^\top = \frac{1}{n}\mathbf{X}_k\mathbf{1}\mathbf{1}^\top.$$

For an extended valued function $\Psi : \mathbb{R}^d \to \mathbb{R}\cup\{+\infty\}$, its effective domain is written as $\mathrm{dom}(\Psi) = \{x \mid \Psi(x) < +\infty\}$. A function Ψ is said to be proper if $\mathrm{dom}(\Psi)$ is non-empty. For any proper closed convex function Ψ, $x \in \mathbb{R}^d$, and scalar $\gamma > 0$, the proximal operator is defined as

$$\mathbf{prox}_\Psi^\gamma(x) = \operatorname*{arg\,min}_{y\in\mathbb{R}^d}\left\{\frac{1}{2\gamma}\|y-x\|^2 + \Psi(y)\right\}.$$

All random objects are properly defined in a probability space $(\Omega, \mathscr{F}, \mathbb{P})$ and write $x \in \mathcal{H}$ if x is $\mathcal{H}$-measurable given a sub-σ-algebra $\mathcal{H} \subseteq \mathscr{F}$ and a random vector x. We use $\sigma(\cdot)$ to denote the σ-algebra generated by all the argument random vectors. Without loss of generality, we assume $n \geq 2$.

Assumption 1. *The adjacency matrix $\mathbf{W} = (w_{ij}) \in \mathbb{R}^{n\times n}$ is symmetric and doubly stochastic, i.e.,*

$$\mathbf{W} = \mathbf{W}^\top, \quad \mathbf{W}\mathbf{1}_n = \mathbf{1}_n, \quad w_{ij} \geq 0, \forall i, j,$$

and its eigenvalues satisfy $1 = \lambda_1 > \lambda_2 \geq \cdots \geq \lambda_n$ and $\rho \coloneqq \max\{|\lambda_2|, |\lambda_n|\} < 1$.

Assumption 2. *All functions $\{F_i\}_{1\leq i\leq n}$ have Lipschitz continuous gradients with Lipschitz constants $L_{\nabla F_i}$, respectively. Therefore, ∇F is $L_{\nabla F}$-Lipchitz continous with $L_{\nabla F} = \max_{1\leq i\leq n}\{L_{\nabla F_i}\}$.*

Assumption 3. *The function $\Psi : \mathbb{R}^d \to \mathbb{R}\cup\{+\infty\}$ is a closed proper convex function.*

For stochastic oracles, we assume that each node i at every iteration k is able to obtain a local random data vector ξ_i^k. The induced natural filtration is given by $\mathscr{F}_0 = \{\emptyset, \Omega\}$ and

$$\mathscr{F}_k \coloneqq \sigma\left(\xi_i^t \mid i = 1, \ldots, n,\ t = 1, \ldots, k\right), \forall k \geq 1.$$

We require that the stochastic gradient $\nabla G_i(\cdot, \xi_i^{k+1})$ is unbiased conditioned on the filteration $\mathscr{F}_k$.

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).

Algorithm 1: `Prox-DASA`

Input: $x_i^0 = z_i^0 = \mathbf{0}, \gamma, \{\alpha_k\}_{\geq 0}, m$
for $k = 0, 1, \ldots, K-1$ **do**
  `# Local Update`
  for $i = 1, 2, \ldots, n$ *(in parallel)* **do**
    $y_i^k = \mathbf{prox}_{\Psi}^{\gamma}\left(x_i^k - \gamma z_i^k\right)$
    $\tilde{x}_i^{k+1} = (1-\alpha_k)x_i^k + \alpha_k y_i^k$
    `# Compute stochastic gradient`
    $v_i^{k+1} = \nabla G_i(x_i^k, \xi_i^{k+1})$
    $\tilde{z}_i^{k+1} = (1-\alpha_k)z_i^k + \alpha_k v_i^{k+1}$
  end
  `# Communication`
  $[x_1^{k+1}, \ldots, x_n^{k+1}] = [\tilde{x}_1^{k+1}, \ldots, \tilde{x}_n^{k+1}]\mathbf{W}^m$
  $[z_1^{k+1}, \ldots, z_n^{k+1}] = [\tilde{z}_1^{k+1}, \ldots, \tilde{z}_n^{k+1}]\mathbf{W}^m$
end

Algorithm 2: `Prox-DASA-GT`

Input: $x_i^0 = z_i^0 = \mathbf{0}, \gamma, \{\alpha_k\}_{\geq 0}, m$
for $k = 0, 1, \ldots, K$ **do**
  `# Local Update`
  for $i = 1, 2, \ldots, n$ *(in parallel)* **do**
    $y_i^k = \mathbf{prox}_{\Psi}^{\gamma}\left(x_i^k - \gamma z_i^k\right)$
    $\tilde{x}_i^{k+1} = (1-\alpha_k)x_i^k + \alpha_k y_i^k$
    `# Compute stochastic gradient`
    $v_i^{k+1} = \nabla G_i(x_i^k, \xi_i^{k+1})$
    $\tilde{u}_i^{k+1} = u_i^k + v_i^{k+1} - v_i^k$
    $\tilde{z}_i^{k+1} = (1-\alpha_k)z_i^k + \alpha_k u_i^k$
  end
  `# Communication`
  $[x_1^{k+1}, \ldots, x_n^{k+1}] = [\tilde{x}_1^{k+1}, \ldots, \tilde{x}_n^{k+1}]\mathbf{W}^m$
  $[u_1^{k+1}, \ldots, u_n^{k+1}] = [\tilde{u}_1^{k+1}, \ldots, \tilde{u}_n^{k+1}]\mathbf{W}^m$
  $[z_1^{k+1}, \ldots, z_n^{k+1}] = [\tilde{z}_1^{k+1}, \ldots, \tilde{z}_n^{k+1}]\mathbf{W}^m$
end

Assumption 4 (Unbiasness). *For any $k \geq 0, x \in \mathscr{F}_k$, and $1 \leq i \leq n$, $\mathbb{E}\left[\nabla G_i(x, \xi_i^{k+1}) \mid \mathscr{F}_k\right] = \nabla F_i(x)$.*

Assumption 5 (Independence). *For any $k \geq 0, 1 \leq i, j \leq n, i \neq j$, ξ_i^{k+1} is independent of $\mathscr{F}_k$, and ξ_i^{k+1} is independent of ξ_j^{k+1}.*

In addition, we consider two standard assumptions on the variance and heterogeneity of stochastic gradients.

Assumption 6 (Bounded variance). *For any $k \geq 0, x \in \mathscr{F}_k$, and $1 \leq i \leq n$,*

$$\mathbb{E}\left[\left\|\nabla G_i(x, \xi_i^{k+1}) - \nabla F_i(x)\right\|^2 \middle| \mathscr{F}_k\right] \leq \sigma_i^2.$$

Assumption 7 (Gradient heterogeneity). *There exists a constant $\nu \geq 0$ such that for all $1 \leq i \leq n, x \in \mathbb{R}^d$,*

$$\|\nabla F_i(x) - \nabla F(x)\| \leq \nu.$$

1 EXPERIMENTAL DETAILS

All experiments in Section 5.2 are conducted on a laptop with Intel Core i7-11370H Processor and Windows 11 operating system. The total iteration numbers for a9a and MNIST are 10000 and 3000 respectively. The graph that represents the network topology is set to be ring (or cycle in graph theory) for a9a and random graph (given by Mancino-ball et al. [2023]) for MNIST (See Figure 1).

We summarize the outputs of all experiments in Table 1, from which we can tell `Prox-DASA` and `Prox-DASA-GT` achieve good performance in a relatively short amount of time. The stationarity is defined as $\|\mathcal{G}(\bar{x}^k, \nabla F(\bar{x}^k), 1)\|^2 + \|\mathbf{X}_k - \bar{\mathbf{X}}_k\|^2$, which is the same as that in Mancino-ball et al. [2023]. As mentioned in the caption of Figure 2 in the main paper, there is an extra hyperparameter q in `ProxGT-SR-E`, and we found that large q already works well for a9a experiment, but q has to be small in the MNIST experiment otherwise the final accuracy will be much smaller than the one presented in Table 1. Hence in `ProxGT-SR-E` we choose $q = 1000$ for a9a and $q = 32$ for MNIST, and the plots that take this amount of epochs into account are in Figure 2.

Table 1: Comparisons between all algorithms

Algorithm	**Accuracy**	**Training Loss**	**Stationarity**	**Communication time per iteration (s)**	**Computation time per iteration (s)**	**Total time per iteration (s)**
			a9a			
`SPPDM`	84.64%	0.3340	0.0174	0.0260	0.0305	0.0565
`ProxGT-SR-E`	76.38%	0.6528	0.0797	0.0521	0.0394	0.0915
`DEEPSTORM v2`	**84.90%**	**0.3274**	0.0029	0.0525	0.0398	0.0923
`Prox-DASA`	84.71%	0.3338	**0.0017**	0.0360	0.0298	0.0658
`Prox-DASA-GT`	84.69%	0.3342	**0.0017**	0.0390	0.0301	0.0691
			MNIST			
`SPPDM`	76.54%	0.7854	0.0436	0.1587	0.1246	0.2833
`ProxGT-SR-E`	92.26%	0.3042	0.0250	0.1771	0.3368	0.5139
`DEEPSTORM v2`	94.52%	0.1759	**0.0016**	0.1758	0.2030	0.3788
`Prox-DASA`	96.74%	0.1469	0.0081	0.1912	0.1299	0.3211
`Prox-DASA-GT`	**96.84%**	**0.1460**	0.0058	0.1935	0.1317	0.3252

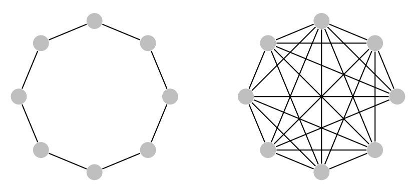

Figure 1: Network topology. The left represents the ring topology and the right represents the random graph.

2 ACCELERATED CONSENSUS

When the number of communication round $m > 1$, we can replace $\mathbf{W}^m$ with the Chebyshev mixing protocol described in Algorithm 3.

Algorithm 3: Chebyshev Mixing Protocol

Input: Matrix $\mathbf{X}$, mixing matrix $\mathbf{W}$, rounds m

Set $\mathbf{A}_0 = \mathbf{X}, \mathbf{A}_1 = \mathbf{X}\mathbf{W}, \rho = \max\{|\lambda_2(\mathbf{W})|, |\lambda_n(\mathbf{W})|\} < 1, \mu_0 = 1, \mu_1 = \frac{1}{\rho}$

for $t = 1, \ldots, m-1$ **do**

  $\mu_{t+1} = \frac{2}{\rho}\mu_t - \mu_{t-1}$

  $\mathbf{A}_{t+1} = \frac{2\mu_t}{\rho\mu_{t+1}}\mathbf{A}_t\mathbf{W} - \frac{\mu_{t-1}}{\mu_{t+1}}\mathbf{A}_{t-1}$

end

Output: $\mathbf{A}_m$

Then, we have the following lemma.

Lemma 2.1. *Suppose* $\mathbf{W}$ *satisfies Assumption 1. Let* $\mathbf{A}_0, \mathbf{A}_m$ *be the input and output matrix of Algorithm 3 respectively. Then, we have*

$$\left\|\mathbf{A}_m - \bar{\mathbf{A}}_m\right\| \leq 2\left(1 - \sqrt{1-\rho}\right)^m \left\|\mathbf{A}_0 - \bar{\mathbf{A}}_0\right\|.$$

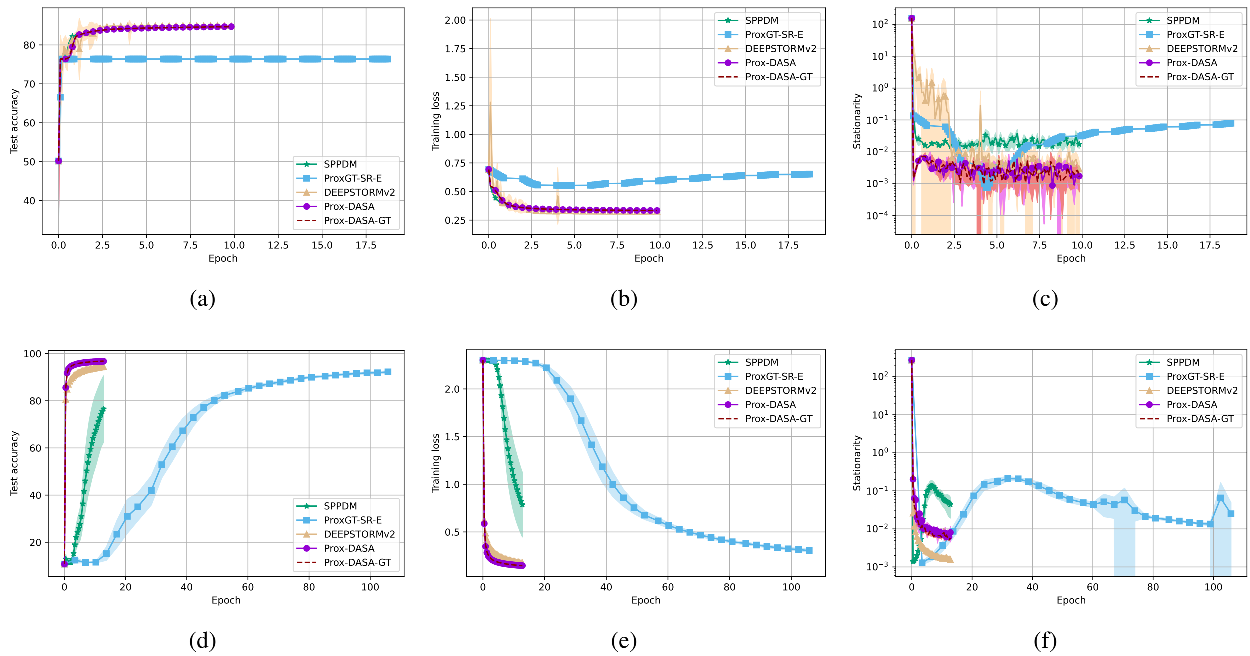

Figure 2: Comparisons between `SPPDM` [Wang et al., 2021], `ProxGT-SR-E` [Xin et al., 2021], `DEEPSTORM` [Mancino-ball et al., 2023], `Prox-DASA` 1, and `Prox-DASA-GT` 2. In each experiments `ProxGT-SR-E` computes 1 more epoch than other algorithms every q iterations. q is chosen to be 1000 for a9a and 32 for MNIST.

Hence, we obtain a linear convergence rate of $\left(1-\sqrt{1-\rho}\right)$ instead of ρ. By virtue of that, we can set $m = \lceil \frac{1}{\sqrt{1-\rho}} \rceil$ to obtain a topology-independent iteration complexity.

3 CONVERGENCE ANALYSIS

We present the complete proof in this section. In the sequel, $\|\cdot\|$ denotes the ℓ_2-norm for vectors and Frobenius norm for matrices. $\|\cdot\|_2$ denotes the spectral norm for matrices. $\mathbf{1}$ represents the all-one vector. We identify vectors at agent i in the subscript and use the superscript for the algorithm step. For example, the optimization variable of agent i at step k is denoted as x_i^k, and z_i^k is the corresponding dual variable. We use uppercase bold letters to represent the matrix that collects all the variables from agents (corresponding lowercase) as columns. To be specific,

$$\mathbf{X}_k = \left[x_1^k, \dots, x_n^k\right], \quad \mathbf{Z}_k = \left[z_1^k, \dots, z_n^k\right], \quad \mathbf{Y}_k = \left[y_1^k, \dots, y_n^k\right], \quad \mathbf{V}_{k+1} = \left[v_1^{k+1}, \dots, v_n^{k+1}\right].$$

We add an overbar to a letter to denote the average over all agents. For example,

$$\bar{x}^k = \frac{1}{n}\sum_{i=1}^n x_i^k = \frac{1}{n}\mathbf{X}_k\mathbf{1}, \quad \bar{\mathbf{X}}_k = [\bar{x}^k, \dots, \bar{x}^k] = \bar{x}^k\mathbf{1}^\top = \frac{1}{n}\mathbf{X}_k\mathbf{1}\mathbf{1}^\top$$

Hence, the consensus errors for iterates $\{x_i^k\}$ and dual variables $\{z_i^k\}$ can be written as

$$\frac{1}{n}\sum_{i=1}^n \left\|x_i^k - \bar{x}^k\right\|^2 = \frac{1}{n}\left\|\mathbf{X}_k - \bar{\mathbf{X}}_k\right\|^2, \quad \frac{1}{n}\sum_{i=1}^n \left\|z_i^k - \bar{z}^k\right\|^2 = \frac{1}{n}\left\|\mathbf{Z}_k - \bar{\mathbf{Z}}_k\right\|^2.$$

We denote $L_{\nabla F} = \max_{1\le i\le n}\{L_{\nabla F_i}\}$ for ease of presentation. Our proof heavily relies on the merit function below:

$$W(\bar{x}^k, \bar{z}^k) = \underbrace{\Phi(\bar{x}^k) - \Phi_*}_{\text{function value gap}} + \underbrace{\Psi(\bar{x}^k) - \eta(\bar{x}^k, \bar{z}^k)}_{\text{primal convergence}} + \lambda \underbrace{\left\|\nabla F(\bar{x}^k) - \bar{z}^k\right\|^2}_{\text{dual convergence}}, \tag{1}$$

where

$$\eta(x, z) = \min_{y\in\mathbb{R}^d}\left\{\langle z, y - x\rangle + \frac{1}{2\gamma}\|y - x\|^2 + \Psi(y)\right\}. \tag{2}$$

3.1 TECHNICAL LEMMAS

Lemma 3.1. *For any $p, q, r \in \mathbb{N}_+$ and matrix $\mathbf{A} \in \mathbb{R}^{p\times q}, \mathbf{B} \in \mathbb{R}^{q\times r}$, we have:*

$$\|\mathbf{AB}\| \le \min\left(\|\mathbf{A}\|_2 \cdot \|\mathbf{B}\|, \|\mathbf{A}\| \cdot \|\mathbf{B}^\top\|_2\right).$$

Lemma 3.2. *Suppose $\mathbf{W}$ satisfies Assumption 1. For any $m \in \mathbb{N}_+$, we have*

$$\left\|\mathbf{W}^m - \frac{\mathbf{1}_n\mathbf{1}_n^\top}{n}\right\|_2 \le \rho^m$$

Lemma 3.3. *Suppose we are given three sequences $\{a_n\}_{n=0}^\infty, \{b_n\}_{n=0}^\infty, \{\tau_n\}_{n=-1}^\infty$, and a constant r satisfying*

$$a_{k+1} \le ra_k + b_k,\ a_k \ge 0,\ b_k \ge 0,\ 0 = \tau_{-1} \le \tau_{k+1} \le \tau_k \le 1, \tag{3}$$

for all $k \ge 0$. Then for any $K > 0$, we have

$$\sum_{k=0}^K \tau_k a_k \le \frac{1}{1-r}\left(\tau_0 a_0 + \sum_{k=0}^K \tau_k b_k\right)$$

Proof. Note that we have

$$(1-r)\sum_{k=0}^K \tau_k a_k \le \sum_{k=0}^K \tau_k(a_k - a_{k+1} + b_k) = \sum_{k=0}^K (\tau_k - \tau_{k-1})a_k - \tau_K a_{K+1} + \sum_{k=0}^K \tau_k b_k \le \tau_0 a_0 + \sum_{k=0}^K \tau_k b_k,$$

where the inequalities use (3), and the equality uses summation by parts. □

Lemma 3.4. *Let $\Psi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a closed proper convex function.*

(a) Let $\eta(x, z)$ be the function defined in (2). Then, $\nabla\eta$ is C_γ-Lipschitz continuous where

$$C_\gamma = 2\sqrt{(1+\frac{1}{\gamma})^2 + (1+\frac{\gamma}{2})^2}. \tag{4}$$

(b) For $x, z \in \mathbb{R}^d$ and $\gamma \in \mathbb{R}$, let $y_+ = \mathbf{prox}_\Psi^\gamma(x - \gamma z) = \arg\min_{y\in\mathbb{R}^d} \left\{\langle z, y - x\rangle + \frac{1}{2\gamma}\|y - x\|^2 + \Psi(y)\right\}$, then for any $y \in \mathbb{R}^d$, we have

$$\Psi(y_+) - \Psi(y) \le \left\langle z + \gamma^{-1}(y_+ - x), y - y_+\right\rangle$$

Proof. We prove (a) at first. Recall that the Moreau envelope of a convex and closed function Ψ multiplied by a scalar γ is defined by

$$\mathrm{env}_{\gamma\Psi}(x) = \min_{y\in\mathbb{R}^d}\left\{\frac{1}{2\gamma}\|y - x\|^2 + \Psi(y)\right\},$$

and its gradient is given by $\nabla\mathrm{env}_{\gamma\Psi}(x) = \frac{1}{\gamma}(x - \mathbf{prox}_\Psi^\gamma(x))$ where $\mathbf{prox}_\Psi^\gamma(x) = \arg\min_{y\in\mathbb{R}^d}\left\{\frac{1}{2\gamma}\|y - x\|^2 + \Psi(y)\right\}$. Note that $\eta(x, z) = \mathrm{env}_{\gamma\Psi}(x - \gamma z) - \frac{\gamma}{2}\|z\|^2$. Therefore, the partial gradients of η are given by

$$\nabla_x\eta(x, z) = -z - \gamma^{-1}\left(\mathbf{prox}_\Psi^\gamma(x - \gamma z) - x\right), \quad \nabla_z\eta(x, z) = \mathbf{prox}_\Psi^\gamma(x - \gamma z) - x. \tag{5}$$

Hence, for any (x, z) and (x', z'),

$$\begin{aligned}\|\nabla\eta(x, z) - \nabla\eta(x', z')\| &\le \|\nabla_x\eta(x, z) - \nabla_x\eta(x', z')\| + \|\nabla_z\eta(x, z) - \nabla_z\eta(x', z')\| \\ &\le 2(1 + 1/\gamma)\|x - x'\| + (2+\gamma)\|z - z'\| \le C_\gamma\|(x, z) - (x', z')\|.\end{aligned}$$

To prove (b), denote the subdifferential of $\Psi(x)$ as $\partial\Psi(x)$. By the optimality condition, we have $\mathbf{0}$ is a subgradient of $H(y) = \langle z, y - x\rangle + \frac{1}{2\gamma}\|y - x\|^2 + \Psi(y)$ at y_+, i.e.,

$$\mathbf{0} \in z + \gamma^{-1}(y_+ - x) + \partial\Psi(y_+).$$

Hence, there exists a subgradient of $\Psi(y)$ at y_+, denoted by $\tilde{\nabla}\Psi(y_+)$, such that

$$\tilde{\nabla}\Psi(y_+) = -z - \gamma^{-1}(y_+ - x).$$

Finally, by the convexity of Ψ, we have for any $y \in \mathbb{R}^d$,

$$\Psi(y) - \Psi(y_+) \geq \left\langle \tilde{\nabla}\Psi(y_+), y - y_+ \right\rangle = \left\langle -z - \gamma^{-1}(y_+ - x), y - y_+ \right\rangle,$$

which completes the proof. □

3.2 BUILDING BLOCKS OF MAIN PROOF

The following lemma connects the consensus error of $\mathbf{Y}$ to the consensus errors of $\mathbf{X}$ and $\mathbf{Z}$.

Lemma 3.5. *Let $y_+^k = \mathbf{prox}(\bar{x}^k - \gamma\bar{z}^k)$. Then for any $k \geq 0$ and $\gamma > 0$, we have*

$$\left\|y_+^k - \bar{y}^k\right\|^2 + \frac{1}{n}\left\|\mathbf{Y}_k - \bar{\mathbf{Y}}_k\right\|^2 = \frac{1}{n}\sum_{i=1}^n \left\|y_i^k - y_+^k\right\|^2 \leq \frac{2}{n}\left\{\|\mathbf{X}_k - \bar{\mathbf{X}}_k\|^2 + \gamma^2\|\mathbf{Z}_k - \bar{\mathbf{Z}}_k\|^2\right\}.$$

Proof. By the non-expansiveness of proximal operator, we have

$$\|y_i^k - y_+^k\| \leq \|x_i^k - \bar{x}^k - \gamma\left(z_i^k - \bar{z}^k\right)\| \leq \|x_i^k - \bar{x}^k\| + \gamma\|z_i^k - \bar{z}^k\|. \tag{6}$$

Hence we know the consensus error of y can be bounded

$$\begin{aligned}
&\frac{1}{n}\|\mathbf{Y}_k - \bar{\mathbf{Y}}^k\|^2 = \frac{1}{n}\sum_{i=1}^n \|y_i^k - \bar{y}^k\|^2 = \frac{1}{n}\sum_{i=1}^n \|y_i^k - y_+^k + \frac{1}{n}\sum_{j=1}^n (y_+^k - y_j^k)\|^2 \\
=&\frac{1}{n}\sum_{i=1}^n \|y_i^k - y_+^k\|^2 - \|\frac{1}{n}\sum_{j=1}^n \left(y_j^k - y_+^k\right)\|^2 \leq \frac{1}{n}\sum_{i=1}^n \|y_i^k - y_+^k\|^2 \\
\leq&\frac{2}{n}\left\{\|\mathbf{X}_k - \bar{\mathbf{X}}_k\|^2 + \gamma^2\|\mathbf{Z}_k - \bar{\mathbf{Z}}_k\|^2\right\}
\end{aligned} \tag{7}$$

where the third equality uses the fact that

$$\frac{1}{n}\sum_{i=1}^n \left\| v_i - \left(\frac{1}{n}\sum_{j=1}^n v_j\right)\right\|^2 = \frac{1}{n}\sum_{i=1}^n \|v_i\|^2 - \left\|\frac{1}{n}\sum_{j=1}^n v_j\right\|^2$$

for any vectors v_i $(1 \leq i \leq n)$. □

The following technical lemma explicitly characterizes the consensus error.

Lemma 3.6 (Conensus error of Algorithm 1: `Prox-DASA`)**.** *Suppose Assumptions 1, 4, 5, 6, and 7 hold. Let $\varrho(m) = \frac{(1+\rho^{2m})\rho^{2m}}{(1-\rho^{2m})^2}$, and ρ, m and α_k satisfy*

$$\varrho(m)\alpha_k^2 \leq \min\left\{\frac{1}{8}, \frac{1}{24L_{\nabla F}^2\gamma^2}\right\},\ 0 = \alpha_{-1} \leq \alpha_{k+1} \leq \alpha_k \leq 1 \tag{8}$$

for any $k \geq 0$. Then in Algorithm 1 for any $p \geq 0$, we have

$$\begin{aligned}
\sum_{k=0}^K \frac{\alpha_k^p}{n}\mathbb{E}\left[\|\mathbf{X}_k - \bar{\mathbf{X}}_k\|^2\right] &\leq 4\gamma^2(\sigma^2 + 3L_{\nabla F}^2\nu^2)\varrho(m)\sum_{k=0}^K \alpha_k^{p+2}, \\
\sum_{k=0}^K \frac{\alpha_k^p}{n}\mathbb{E}\left[\|\mathbf{Z}_k - \bar{\mathbf{Z}}_k\|^2\right] &\leq 4(\sigma^2 + 3L_{\nabla F}^2\nu^2)\varrho(m)\sum_{k=0}^K \alpha_k^{p+2}, .
\end{aligned}$$

Proof. By Assumption 1, the iterates in Algorithm 1 satisfy

$$\begin{aligned}&\mathbf{X}_{k+1}=(1-\alpha_k)\mathbf{X}_k\mathbf{W}^m+\alpha_k\mathbf{Y}_k\mathbf{W}^m,\ \bar{x}^{k+1}=(1-\alpha_k)\bar{x}^k+\alpha_k\bar{y}^k,\\&\mathbf{Z}_{k+1}=(1-\alpha_k)\mathbf{Z}_k\mathbf{W}^m+\alpha_k\mathbf{V}_{k+1}\mathbf{W}^m,\ \bar{z}^{k+1}=(1-\alpha_k)\bar{z}^k+\alpha_k\bar{v}^{k+1}.\end{aligned}\tag{9}$$

Hence, for the consensus error of iterates $\{x_i^k\}$, we have

$$\begin{aligned}&\left\|\mathbf{X}_{k+1}-\bar{\mathbf{X}}_{k+1}\right\|^2\\=&\left\|\Big((1-\alpha_k)\left(\mathbf{X}_k-\bar{\mathbf{X}}_k\right)+\alpha_k\left(\mathbf{Y}_k-\bar{\mathbf{Y}}_k\right)\Big)\left(\mathbf{W}^m-\frac{\mathbf{1}\mathbf{1}^\top}{n}\right)\right\|^2\\\leq&\left\{\left(1+\frac{1-\rho^{2m}}{2\rho^{2m}}\right)(1-\alpha_k)^2\left\|\mathbf{X}_k-\bar{\mathbf{X}}_k\right\|^2+\left(1+\frac{2\rho^{2m}}{1-\rho^{2m}}\right)\alpha_k^2\left\|\mathbf{Y}_k-\bar{\mathbf{Y}}_k\right\|^2\right\}\rho^{2m}\\\leq&\frac{(1+\rho^{2m})}{2}\left\|\mathbf{X}_k-\bar{\mathbf{X}}_k\right\|^2+\frac{(1+\rho^{2m})\rho^{2m}}{1-\rho^{2m}}\alpha_k^2\left\|\mathbf{Y}_k-\bar{\mathbf{Y}}_k\right\|^2,\end{aligned}\tag{10}$$

where the first inequality uses Lemma 3.1 and 3.2. Combining (8), (10), and Lemma 3.5, we have

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{X}_{k+1}-\bar{\mathbf{X}}_{k+1}\|^2\right]&\leq\frac{(1+\rho^{2m})}{2}\mathbb{E}\left[\|\mathbf{X}_k-\bar{\mathbf{X}}_k\|^2\right]+\frac{(1-\rho^{2m})}{4}\mathbb{E}\left[\|\mathbf{X}_k-\bar{\mathbf{X}}_k\|^2+\gamma^2\|\mathbf{Z}_k-\bar{\mathbf{Z}}_k\|^2\right]\\&=\frac{(3+\rho^{2m})}{4}\mathbb{E}\left[\|\mathbf{X}_k-\bar{\mathbf{X}}_k\|^2\right]+\frac{(1-\rho^{2m})\gamma^2}{4}\mathbb{E}\left[\|\mathbf{Z}_k-\bar{\mathbf{Z}}_k\|^2\right]\end{aligned}$$

Using Lemma 3.3 in the above inequality with $\tau_k=\frac{\alpha_k^p}{n}$ for any fixed $p\geq 0$ we know

$$\sum_{k=0}^K\frac{\alpha_k^p}{n}\mathbb{E}\left[\|\mathbf{X}_k-\bar{\mathbf{X}}_k\|^2\right]\leq\sum_{k=0}^K\frac{\gamma^2\alpha_k^p}{n}\mathbb{E}\left[\|\mathbf{Z}_k-\bar{\mathbf{Z}}_k\|^2\right].\tag{11}$$

Similarly to (10), we can obtain the following results on the consensus error of dual variables $\{z_i^k\}$:

$$\left\|\mathbf{Z}_{k+1}-\bar{\mathbf{Z}}_{k+1}\right\|^2\leq\frac{(1+\rho^{2m})}{2}\left\|\mathbf{Z}_k-\bar{\mathbf{Z}}_k\right\|^2+\frac{(1+\rho^{2m})\rho^{2m}}{1-\rho^{2m}}\alpha_k^2\left\|\mathbf{V}_{k+1}-\bar{\mathbf{V}}_{k+1}\right\|^2,\tag{12}$$

Using (8) and Lemma 3.3 in (12) with $\tau_k=\frac{\alpha_k^p}{n}$, we have

$$\sum_{k=0}^K\frac{\alpha_k^p}{n}\mathbb{E}\left[\|\mathbf{Z}_k-\bar{\mathbf{Z}}_k\|^2\right]\leq 2\varrho(m)\sum_{k=0}^K\frac{\alpha_k^{p+2}}{n}\mathbb{E}\left[\|\mathbf{V}_{k+1}-\bar{\mathbf{V}}_{k+1}\|^2\right].\tag{13}$$

To bound $\|\mathbf{V}_{k+1}-\bar{\mathbf{V}}_{k+1}\|$ we first notice that

$$\begin{aligned}v_i^{k+1}-\bar{v}^{k+1}=&v_i^{k+1}-\mathbb{E}\left[v_i^{k+1}|\mathscr{F}_k\right]-\frac{1}{n}\sum_{j=1}^n(v_j^{k+1}-\mathbb{E}\left[v_j^{k+1}|\mathscr{F}_k\right])\\&+\mathbb{E}\left[v_i^{k+1}|\mathscr{F}_k\right]-\nabla F_i(\bar{x}^k)+\nabla F_i(\bar{x}^k)-\nabla F(\bar{x}^k)+\nabla F(\bar{x}^k)-\frac{1}{n}\sum_{j=1}^n\mathbb{E}\left[v_j^{k+1}|\mathscr{F}_k\right]\\&=\left(1-\frac{1}{n}\right)(v_i^{k+1}-\mathbb{E}\left[v_i^{k+1}|\mathscr{F}_k\right])-\frac{1}{n}\sum_{j\neq i}(v_j^{k+1}-\mathbb{E}\left[v_j^{k+1}|\mathscr{F}_k\right])\\&+\left(1-\frac{1}{n}\right)\left(\nabla F_i(x_i^k)-\nabla F_i(\bar{x}^k)\right)+\nabla F_i(\bar{x}^k)-\nabla F(\bar{x}^k)+\frac{1}{n}\sum_{j\neq i}\left(\nabla F_j(\bar{x}^k)-\nabla F_i(x_j^k)\right)\end{aligned}$$

which gives

$$\mathbb{E}\left[\|v_i^{k+1}-\bar{v}^{k+1}\|^2\right]$$

$$
\begin{aligned}
&= \left(1-\frac{1}{n}\right)^2 \mathbb{E}\left[\|v_i^{k+1} - \mathbb{E}\left[v_i^{k+1}|\mathscr{F}_k\right]\|^2\right] + \frac{1}{n^2}\sum_{j\neq i}^{n} \mathbb{E}\left[\|v_j^{k+1} - \mathbb{E}\left[v_j^{k+1}|\mathscr{F}_k\right]\|^2\right] \\
&+ \left\|\left(1-\frac{1}{n}\right)\left(\nabla F_i(x_i^k) - \nabla F_i(\bar{x}^k)\right) + \nabla F_i(\bar{x}^k) - \nabla F(\bar{x}^k) + \frac{1}{n}\sum_{j\neq i}\left(\nabla F_j(\bar{x}^k) - \nabla F_i(x_j^k)\right)\right\|^2 \\
&\leq \sigma^2 + 3L_{\nabla F}^2\left(\left(1-\frac{1}{n}\right)^2\|x_i^k - \bar{x}^k\|^2 + \nu^2 + \frac{1}{n}\sum_{j\neq i}\|x_j^k - \bar{x}^k\|^2\right),
\end{aligned}
$$

where the first equality uses Assumption 5, and the second inequality uses Cauchy-Schwarz inequality, Assumptions 2, 6, and 7. Hence we have

$$
\mathbb{E}\left[\|\mathbf{V}_{k+1} - \bar{\mathbf{V}}_{k+1}\|^2\right] \leq 6L_{\nabla F}^2\mathbb{E}\left[\|\mathbf{X}_k - \bar{\mathbf{X}}_k\|^2\right] + n\sigma^2 + 3nL_{\nabla F}^2\nu^2. \tag{14}
$$

Combining (13) and (14), we have

$$
\begin{aligned}
\sum_{k=0}^{K}\frac{\alpha_k^p}{n}\mathbb{E}\left[\|\mathbf{Z}_k - \bar{\mathbf{Z}}_k\|^2\right] &\leq 2\varrho(m)\sum_{k=0}^{K}\left\{\frac{6L_{\nabla F}^2\alpha_k^{p+2}}{n}\mathbb{E}\left[\|\mathbf{X}_k - \bar{\mathbf{X}}_k\|^2\right] + (\sigma^2 + 3L_{\nabla F}^2\nu^2)\sum_{k=0}^{K}\alpha_k^{p+2}\right\} \\
&\leq \sum_{k=0}^{K}\left\{12\varrho(m)\alpha_k^2 L_{\nabla F}^2\gamma^2\right\}\frac{\alpha_k^p}{n\gamma^2}\mathbb{E}\left[\|\mathbf{X}_k - \bar{\mathbf{X}}_k\|^2\right] + 2(\sigma^2 + 3L_{\nabla F}^2\nu^2)\varrho(m)\sum_{k=0}^{K}\alpha_k^{p+2} \\
&\leq \sum_{k=0}^{K}\frac{\alpha_k^p}{2n}\mathbb{E}\left[\|\mathbf{Z}_k - \bar{\mathbf{Z}}_k\|^2\right] + 2(\sigma^2 + 3L_{\nabla F}^2\nu^2)\varrho(m)\sum_{k=0}^{K}\alpha_k^{p+2},
\end{aligned} \tag{15}
$$

where the second inequality uses (8). By (11) and (15) we can finally obtain that

$$
\sum_{k=0}^{K}\frac{\alpha_k^p}{n}\mathbb{E}\left[\|\mathbf{X}_k - \bar{\mathbf{X}}_k\|^2\right] \leq 4\gamma^2(\sigma^2 + 3L_{\nabla F}^2\nu^2)\varrho(m)\sum_{k=0}^{K}\alpha_k^{p+2},, \tag{16}
$$

$$
\sum_{k=0}^{K}\frac{\alpha_k^p}{n}\mathbb{E}\left[\|\mathbf{Z}_k - \bar{\mathbf{Z}}_k\|^2\right] \leq 4(\sigma^2 + 3L_{\nabla F}^2\nu^2)\varrho(m)\sum_{k=0}^{K}\alpha_k^{p+2},. \tag{17}
$$

□

Lemma 3.7 (Conensus error of Algorithm 2: `Prox-DASA-GT`)**.** *Suppose Assumptions 1, 4, 6 and 5 hold. Let* $\varrho(m) = \frac{(1+\rho^{2m})\rho^{2m}}{(1-\rho^{2m})^2}$*, and* ρ, m *and* α_k *satisfy*

$$
\varrho(m)\alpha_k^2 \leq \frac{1}{8},\ \varrho(m)\alpha_k \leq \frac{1}{9L_{\nabla F}\gamma},\ 0 = \alpha_{-1} \leq \alpha_{k+1} \leq \alpha_k \leq 1 \tag{18}
$$

for any $k \geq 0$*, and the initialization satisfies* $u_i^0 = v_i^0 = 0$ *for all* i*. Then in Algorithm 2 for any* $p \geq 0$ *we have*

$$
\begin{aligned}
\sum_{k=0}^{K}\frac{\alpha_k^p}{n}\mathbb{E}\left[\|\mathbf{X}_k - \bar{\mathbf{X}}_k\|^2\right] &\leq 40\gamma^2\varrho(m)^2\sum_{k=0}^{K}\alpha_k^{p+2}\left\{L_{\nabla F}^2\alpha_k^2\mathbb{E}\left[\|\bar{x}^k - \bar{y}^k\|^2\right] + 2\sigma^2\right\}, \\
\sum_{k=0}^{K}\frac{\alpha_k^p}{n}\mathbb{E}\left[\|\mathbf{Z}_k - \bar{\mathbf{Z}}_k\|^2\right] &\leq 40\varrho(m)^2\sum_{k=0}^{K}\alpha_k^{p+2}\left\{L_{\nabla F}^2\alpha_k^2\mathbb{E}\left[\|\bar{x}^k - \bar{y}^k\|^2\right] + 2\sigma^2\right\}.
\end{aligned}
$$

Proof. The updates in Algorithm 2 take the form:

$$
\begin{aligned}
&\mathbf{X}_{k+1} = (1-\alpha_k)\mathbf{X}_k\mathbf{W}^m + \alpha_k\mathbf{Y}_k\mathbf{W}^m,\ \bar{x}^{k+1} = (1-\alpha_k)\bar{x}^k + \alpha_k\bar{y}^k, \\
&\mathbf{U}_{k+1} = \mathbf{U}_k\mathbf{W}^m + (\mathbf{V}_{k+1} - \mathbf{V}_k)\mathbf{W}^m,\ \bar{u}^{k+1} = \bar{u}^k + \bar{v}^{k+1} - \bar{v}^k, \\
&\mathbf{Z}_{k+1} = (1-\alpha_k)\mathbf{Z}_k\mathbf{W}^m + \alpha_k\mathbf{U}_k\mathbf{W}^m,\ \bar{z}^{k+1} = (1-\alpha_k)\bar{z}^k + \alpha_k\bar{u}^k.
\end{aligned} \tag{19}
$$

Setting $u_i^0 = v_i^0$, we can prove by induction that $\bar{u}^k = \bar{v}^k$. To analyze the consensus error of $\mathbf{U}_k$, we first notice:

$$\begin{aligned}
&\mathbf{U}_{k+1} - \bar{\mathbf{U}}_{k+1} \\
&= \left(\mathbf{U}_k - \bar{\mathbf{U}}_k + \mathbf{V}_{k+1} - \mathbf{V}_k - \bar{\mathbf{V}}^{k+1} + \bar{\mathbf{V}}^k\right)\left(\mathbf{W}^m - \frac{\mathbf{1}\mathbf{1}^\top}{n}\right) \\
&= \left(\mathbf{U}_k - \bar{\mathbf{U}}_k + (\mathbf{V}_{k+1} - \mathbf{V}_k)\left(\mathbf{I} - \frac{\mathbf{1}\mathbf{1}^\top}{n}\right)\right)\left(\mathbf{W}^m - \frac{\mathbf{1}\mathbf{1}^\top}{n}\right)
\end{aligned}$$

which gives

$$\begin{aligned}
&\|\mathbf{U}_{k+1} - \bar{\mathbf{U}}_{k+1}\|^2 \\
\leq &\left\{\left(1 + \frac{1-\rho^{2m}}{2\rho^{2m}}\right)\left\|\mathbf{U}_k - \bar{\mathbf{U}}_k\right\|^2 + \left(1 + \frac{2\rho^{2m}}{1-\rho^{2m}}\right)\|\mathbf{V}_{k+1} - \mathbf{V}_k\|^2\right\}\rho^{2m} \\
= &\frac{(1+\rho^{2m})}{2}\left\|\mathbf{U}_k - \bar{\mathbf{U}}_k\right\|^2 + \frac{(1+\rho^{2m})\rho^{2m}}{1-\rho^{2m}}\|\mathbf{V}_{k+1} - \mathbf{V}_k\|^2.
\end{aligned}$$

Using Lemma 3.3, we know for any $k \geq 0$ and $p \geq 0$,

$$\sum_{k=0}^{K} \alpha_k^p \|\mathbf{U}_k - \bar{\mathbf{U}}_k\|^2 \leq 2\varrho(m) \sum_{k=0}^{K} \alpha_k^p \|\mathbf{V}_{k+1} - \mathbf{V}_k\|^2. \tag{20}$$

Note that we also have

$$\begin{aligned}
\mathbf{V}_{k+1} - \mathbf{V}_k = &\mathbf{V}_{k+1} - \mathbb{E}[\mathbf{V}_{k+1}|\mathscr{F}_k] - (\mathbf{V}_k - \mathbb{E}[\mathbf{V}_k|\mathscr{F}_{k-1}]) \\
&+ \mathbb{E}[\mathbf{V}_{k+1}|\mathscr{F}_k] - \nabla\mathbf{F}(\bar{x}^k) + \nabla\mathbf{F}(\bar{x}^k) - \nabla\mathbf{F}(\bar{x}^{k-1}) + \nabla\mathbf{F}(\bar{x}^{k-1}) - \mathbb{E}[\mathbf{V}_k|\mathscr{F}_{k-1}]
\end{aligned}$$

where we overload the notation and define $\nabla\mathbf{F}(x) = [\nabla F_1(x), ..., \nabla F_n(x)]$. Hence we know

$$\begin{aligned}
&\mathbb{E}\left[\|\mathbf{V}_{k+1} - \mathbf{V}_k\|^2\right] \\
\leq &5\left\{\mathbb{E}\left[\|\mathbf{V}_{k+1} - \mathbb{E}[\mathbf{V}_{k+1}|\mathscr{F}_k]\|^2\right] + \mathbb{E}\left[\|\mathbf{V}_k - \mathbb{E}[\mathbf{V}_k|\mathscr{F}_{k-1}]\|^2\right] + \mathbb{E}\left[\sum_{i=1}^n \|\nabla F_i(x_i^k) - \nabla F_i(\bar{x}^k)\|^2\right]\right. \\
&\left. + \mathbb{E}\left[\sum_{i=1}^n \|\nabla F_i(\bar{x}^k) - \nabla F_i(\bar{x}^{k-1})\|^2\right] + \mathbb{E}\left[\sum_{i=1}^n \|\nabla F_i(x_i^{k-1}) - \nabla F_i(\bar{x}^{k-1})\|^2\right]\right\} \\
\leq &5\left(2n\sigma^2 + L_{\nabla F}^2 \mathbb{E}\left[\|\mathbf{X}_k - \bar{\mathbf{X}}_k\|^2 + \|\mathbf{X}^{k-1} - \bar{\mathbf{X}}^{k-1}\|^2 + n\alpha_{k-1}^2\|\bar{x}^{k-1} - \bar{y}^{k-1}\|^2\right]\right)
\end{aligned} \tag{21}$$

where the first inequality uses Cauchy-Schwarz inequality, and the second inequality uses Lipschitz continuity of ∇f_i and (19). For simplicity we set $x_i^{-1} = y_i^{-1} = 0$ for all i so that it is easy to check the above inequality holds for all $k \geq 0$. Using (20) and (21) we know:

$$\sum_{k=0}^{K} \frac{\alpha_k^p}{n}\|\mathbf{U}_k - \bar{\mathbf{U}}_k\|^2 \tag{22}$$

$$\leq \frac{10\varrho(m)}{n}\sum_{k=0}^{K}\alpha_k^p\left(2n\sigma^2 + L_{\nabla F}^2\mathbb{E}\left[\|\mathbf{X}_k - \bar{\mathbf{X}}_k\|^2 + \|\mathbf{X}^{k-1} - \bar{\mathbf{X}}^{k-1}\|^2 + n\alpha_{k-1}^2\|\bar{x}^{k-1} - \bar{y}^{k-1}\|^2\right]\right).$$

$$\leq \frac{20L_{\nabla F}^2\varrho(m)}{n}\sum_{k=0}^{K}\alpha_k^p\mathbb{E}\left[\|\mathbf{X}_k - \bar{\mathbf{X}}_k\|^2\right] + 10L_{\nabla F}^2\varrho(m)\sum_{k=0}^{K}\alpha_k^{p+2}\mathbb{E}\left[\|\bar{x}^k - \bar{y}^k\|^2\right] + 20\sigma^2\varrho(m)\sum_{k=0}^{K}\alpha_k^p, \tag{23}$$

where the third inequality uses (18). For other consensus error terms we follow the same proof in Lemma 3.6 to get

$$\left\|\mathbf{X}_{k+1} - \bar{\mathbf{X}}_{k+1}\right\|^2 \leq \frac{(1+\rho^{2m})}{2}\left\|\mathbf{X}_k - \bar{\mathbf{X}}_k\right\|^2 + \frac{(1+\rho^{2m})\rho^{2m}}{1-\rho^{2m}}\alpha_k^2\left\|\mathbf{Y}_k - \bar{\mathbf{Y}}_k\right\|^2, \tag{24}$$

$$\left\|\mathbf{Y}_k - \bar{\mathbf{Y}}_k\right\|^2 \leq 2(\left\|\mathbf{X}_k - \bar{\mathbf{X}}_k\right\|^2 + \gamma^2\left\|\mathbf{Z}_k - \bar{\mathbf{Z}}_k\right\|^2), \tag{25}$$

$$\left\|\mathbf{Z}_{k+1}-\bar{\mathbf{Z}}_{k+1}\right\|^2 \leq \frac{(1+\rho^{2m})}{2}\left\|\mathbf{Z}_k-\bar{\mathbf{Z}}_k\right\|^2+\frac{(1+\rho^{2m})\rho^{2m}}{1-\rho^{2m}}\alpha_k^2\left\|\mathbf{U}_k-\bar{\mathbf{U}}_k\right\|^2. \tag{26}$$

Hence we know (11) still holds:

$$\sum_{k=0}^{K}\frac{\alpha_k^p}{n}\mathbb{E}\left[\|\mathbf{X}_k-\bar{\mathbf{X}}_k\|^2\right] \leq \sum_{k=0}^{K}\frac{\gamma^2\alpha_k^p}{n}\mathbb{E}\left[\|\mathbf{Z}_k-\bar{\mathbf{Z}}_k\|^2\right]. \tag{27}$$

Applying Lemma (3.3) in (26) with $\tau_k=\frac{\alpha_k^p}{n}$, we have

$$\sum_{k=0}^{K}\frac{\alpha_k^p}{n}\mathbb{E}\left[\|\mathbf{Z}_k-\bar{\mathbf{Z}}_k\|^2\right] \leq 2\varrho(m)\sum_{k=0}^{K}\frac{\alpha_k^{p+2}}{n}\mathbb{E}\left[\|\mathbf{U}_k-\bar{\mathbf{U}}_k\|^2\right]. \tag{28}$$

The above two inequalities together with (23) and (18) imply

$$\begin{aligned}
&\sum_{k=0}^{K}\frac{\alpha_k^p}{n}\mathbb{E}\left[\|\mathbf{X}_k-\bar{\mathbf{X}}_k\|^2\right] \leq 2\varrho(m)\gamma^2\sum_{k=0}^{K}\frac{\alpha_k^{p+2}}{n}\mathbb{E}\left[\|\mathbf{U}_k-\bar{\mathbf{U}}_k\|^2\right]\\
\leq&\sum_{k=0}^{K}\left\{40L_{\nabla F}^2\gamma^2\varrho(m)^2\alpha_k^2\right\}\frac{\alpha_k^p}{n}\mathbb{E}\left[\|\mathbf{X}_k-\bar{\mathbf{X}}_k\|^2\right]+20\gamma^2\varrho(m)^2\sum_{k=0}^{K}\alpha_k^{p+2}\left\{L_{\nabla F}^2\alpha_k^2\mathbb{E}\left[\|\bar{x}^k-\bar{y}^k\|^2\right]+2\sigma^2\right\}\\
\leq&\frac{1}{2}\sum_{k=0}^{K}\frac{\alpha_k^p}{n}\mathbb{E}\left[\|\mathbf{X}_k-\bar{\mathbf{X}}_k\|^2\right]+20\gamma^2\varrho(m)^2\sum_{k=0}^{K}\alpha_k^{p+2}\left\{L_{\nabla F}^2\alpha_k^2\mathbb{E}\left[\|\bar{x}^k-\bar{y}^k\|^2\right]+2\sigma^2\right\},
\end{aligned}$$

which gives

$$\sum_{k=0}^{K}\frac{\alpha_k^p}{n}\mathbb{E}\left[\|\mathbf{X}_k-\bar{\mathbf{X}}_k\|^2\right] \leq 40\gamma^2\varrho(m)^2\sum_{k=0}^{K}\alpha_k^{p+2}\left\{L_{\nabla F}^2\alpha_k^2\mathbb{E}\left[\|\bar{x}^k-\bar{y}^k\|^2\right]+2\sigma^2\right\}. \tag{29}$$

Combining (18), (23), (28), and (29), we obtain that

$$\begin{aligned}
&\sum_{k=0}^{K}\frac{\alpha_k^p}{n}\mathbb{E}\left[\|\mathbf{Z}_k-\bar{\mathbf{Z}}_k\|^2\right] \leq 2\varrho(m)\sum_{k=0}^{K}\frac{\alpha_k^{p+2}}{n}\mathbb{E}\left[\|\mathbf{U}_k-\bar{\mathbf{U}}_k\|^2\right]\\
\leq&\frac{1}{2\gamma^2}\sum_{k=0}^{K}\frac{\alpha_k^p}{n}\mathbb{E}\left[\|\mathbf{X}_k-\bar{\mathbf{X}}_k\|^2\right]+20\varrho(m)^2\sum_{k=0}^{K}\alpha_k^{p+2}\left\{L_{\nabla F}^2\alpha_k^2\mathbb{E}\left[\|\bar{x}^k-\bar{y}^k\|^2\right]+2\sigma^2\right\},\\
\leq&40\varrho(m)^2\sum_{k=0}^{K}\alpha_k^{p+2}\left\{L_{\nabla F}^2\alpha_k^2\mathbb{E}\left[\|\bar{x}^k-\bar{y}^k\|^2\right]+2\sigma^2\right\}.
\end{aligned}$$

□

Lemma 3.8 (Basic inequalities of dual convergence)**.**

$$\delta^k=\frac{\nabla F(\bar{x}^k)-\nabla F(\bar{x}^{k+1})}{\alpha_k}+\frac{1}{n}\sum_{i=1}^{n}\nabla F_i(x_i^k)-\nabla F(\bar{x}^k),\quad \bar{\Delta}^{k+1}=\bar{v}^{k+1}-\frac{1}{n}\sum_{i=1}^{n}\nabla F_i(x_i^k). \tag{30}$$

Under Assumption 2, we have

$$\begin{aligned}
\|\bar{z}^{k+1}-\nabla F(\bar{x}^{k+1})\|^2 \leq& (1-\alpha_k)\left\|\bar{z}^k-\nabla F(\bar{x}^k)\right\|^2+2L_{\nabla F}^2\alpha_k\left\|\bar{x}^k-\bar{y}^k\right\|^2+\alpha_k^2\left\|\bar{\Delta}^{k+1}\right\|^2\\
&+\frac{2L_{\nabla F}^2\alpha_k}{n}\left\|\mathbf{X}_k-\bar{\mathbf{X}}_k\right\|^2+2\left\langle\alpha_k\bar{\Delta}^{k+1},(1-\alpha_k)\left(\bar{z}^k-\nabla F(\bar{x}^k)\right)+\alpha_k\delta^k\right\rangle,
\end{aligned} \tag{31}$$

and

$$\begin{aligned}
\left\|\bar{z}^{k+1}-\bar{z}^k\right\|^2 \leq \alpha_k^2\Bigg\{&2\left\|\nabla F(\bar{x}^k)-\bar{z}^k\right\|^2+\frac{2L_{\nabla F}^2}{n}\left\|\mathbf{X}_k-\bar{\mathbf{X}}_k\right\|^2+\left\|\bar{\Delta}^{k+1}\right\|^2\\
&+2\left\langle\bar{\Delta}^{k+1},\frac{1}{n}\sum_{i=1}^{n}\nabla F_i(x_i^k)-\bar{z}^k\right\rangle\Bigg\}.
\end{aligned} \tag{32}$$

Proof. By definitions in (30), we have

$$\bar{z}^{k+1} - \nabla F(\bar{x}^{k+1}) = (1-\alpha_k)\left(\bar{z}^k - \nabla F(\bar{x}^k)\right) + \alpha_k \delta^k + \alpha_k \bar{\Delta}^{k+1},$$

Hence, we can get

$$\begin{aligned}
&\left\|\bar{z}^{k+1} - \nabla F(\bar{x}^{k+1})\right\|^2 \\
&= \left\|(1-\alpha_k)\left(\bar{z}^k - \nabla F(\bar{x}^k)\right) + \alpha_k\delta^k\right\|^2 + \alpha_k^2\left\|\bar{\Delta}^{k+1}\right\|^2 + 2\left\langle \alpha_k\bar{\Delta}^{k+1}, (1-\alpha_k)\left(\bar{z}^k - \nabla F(\bar{x}^k)\right) + \alpha_k\delta^k\right\rangle \\
&\leq (1-\alpha_k)\left\|\bar{z}^k - \nabla F(\bar{x}^k)\right\|^2 + \alpha_k\left\|\delta^k\right\|^2 + \alpha_k^2\left\|\bar{\Delta}^{k+1}\right\|^2 + 2\left\langle \alpha_k\bar{\Delta}^{k+1}, (1-\alpha_k)\left(\bar{z}^k - \nabla F(\bar{x}^k)\right) + \alpha_k\delta^k\right\rangle
\end{aligned}$$

where the inequality uses the convexity of $\|\cdot\|^2$. In addition, we have

$$\begin{aligned}
\left\|\delta^k\right\|^2 &\leq 2\left\|\frac{\nabla F(\bar{x}^k) - \nabla F(\bar{x}^{k+1})}{\alpha_k}\right\|^2 + 2\left\|\frac{1}{n}\sum_{i=1}^n\left(\nabla F_i(x_i^k) - \nabla F_i(\bar{x}^k)\right)\right\|^2 \\
&\leq 2L_{\nabla F}^2\left\|\bar{x}^k - \bar{y}^k\right\|^2 + \frac{2L_{\nabla F}^2}{n}\left\|\mathbf{X}_k - \bar{\mathbf{X}}_k\right\|^2,
\end{aligned}$$

which completes the proof of (31). The inequality (32) can be proved similarly by noting that

$$\begin{aligned}
&\left\|\bar{z}^{k+1} - \bar{z}^k\right\|^2 = \alpha_k^2\left\|-\bar{z}^k + \bar{v}^{k+1}\right\|^2 \\
&= \alpha_k^2\left\|(\nabla F(\bar{x}^k) - \bar{z}^k) + \left(\frac{1}{n}\sum_{i=1}^n\left(\nabla F_i(x_i^k) - \nabla F_i(\bar{x}^k)\right)\right) + \alpha_k\bar{\Delta}^{k+1}\right\|^2 \\
&= \alpha_k^2\left\{\left\|(\nabla F(\bar{x}^k) - \bar{z}^k) + \left(\frac{1}{n}\sum_{i=1}^n\left(\nabla F_i(x_i^k) - \nabla F_i(\bar{x}^k)\right)\right)\right\|^2 + \left\|\bar{\Delta}^{k+1}\right\|^2 + 2\left\langle \bar{\Delta}^{k+1}, \frac{1}{n}\sum_{i=1}^n \nabla F_i(x_i^k) - \bar{z}^k\right\rangle\right\}.
\end{aligned}$$

□

Lemma 3.9. *Under Assumption 3,*

$$\Psi(\bar{y}^k) - \Psi(y_+^k) \leq \left\langle \bar{z}^k + \gamma^{-1}(\bar{y}^k - \bar{x}^k), y_+^k - \bar{y}^k\right\rangle + \frac{\gamma}{2n}\left\|\mathbf{Z}_k - \bar{\mathbf{Z}}_k\right\|^2 + \frac{\gamma^{-1}}{2n}\left\|\mathbf{X}_k - \bar{\mathbf{X}}_k\right\|^2. \tag{33}$$

Proof. By the convexity of Ψ and part (b) of Lemma 3.4, we have

$$\begin{aligned}
\Psi(\bar{y}^k) - \Psi(y_+^k) &\overset{\text{cvx}}{\leq} \frac{1}{n}\sum_{i=1}^n\left(\Psi(y_i^k) - \Psi(y_+^k)\right) \overset{\text{Lemma 3.4 (b)}}{\leq} \frac{1}{n}\sum_{i=1}^n\left\langle z_i^k + \gamma^{-1}(y_i^k - x_i^k), y_+^k - y_i^k\right\rangle \\
&= \left\langle \bar{z}^k + \gamma^{-1}(\bar{y}^k - \bar{x}^k), y_+^k - \bar{y}^k\right\rangle + \frac{1}{n}\sum_{i=1}^n\left\langle z_i^k - \bar{z}^k + \gamma^{-1}(y_i^k - \bar{y}^k + \bar{x}^k - x_i^k), \bar{y}^k - y_i^k\right\rangle \\
&\leq \left\langle \bar{z}^k + \gamma^{-1}(\bar{y}^k - \bar{x}^k), y_+^k - \bar{y}^k\right\rangle + \frac{\gamma}{2n}\left\|\mathbf{Z}_k - \bar{\mathbf{Z}}_k\right\|^2 + \frac{1}{2n\gamma}\left\|\mathbf{X}_k - \bar{\mathbf{X}}_k\right\|^2.
\end{aligned}$$

The equality above comes from the fact that for sequences $\{a_i\}_{1\leq i\leq n}, \{b_i\}_{1\leq i\leq n} \in \mathbb{R}^d$, we have

$$\sum_{i=1}^n\left\langle a_i - \frac{1}{n}\sum_{i=1}^n a_i, b_i - \frac{1}{n}\sum_{i=1}^n b_i\right\rangle = \sum_{i=1}^n\langle a_i, b_i\rangle - \left(\frac{1}{n}\sum_{i=1}^n a_i\right)\left(\frac{1}{n}\sum_{i=1}^n b_i\right).$$

The last inequality above is obtained by Young's inequalities:

$$\begin{aligned}
\left\langle z_i^k - \bar{z}^k, \bar{y}^k - y_i^k\right\rangle &\leq \frac{\gamma}{2}\left\|z_i^k - \bar{z}^k\right\|^2 + \frac{1}{2\gamma}\left\|y_i^k - \bar{y}^k\right\|^2, \\
\gamma^{-1}\left\langle \bar{x}^k - x_i^k, \bar{y}^k - y_i^k\right\rangle &\leq \frac{1}{2\gamma}\left\|x_i^k - \bar{x}^k\right\|^2 + \frac{1}{2\gamma}\left\|y_i^k - \bar{y}^k\right\|^2.
\end{aligned}$$

□

Lemma 3.10 (Basic lemma of merit function difference). *Let $W(\bar{x}^k, \bar{z}^k)$ be the merit function defined in* (1) *with $\lambda = \frac{\gamma^{-1}}{8L_{\nabla F}^2}$. Under Assumption 2, 3, for any $k \geq 0$, setting $\alpha_k \leq \min\{\frac{\gamma^{-1}}{8L_{\nabla F}}, \frac{\gamma^{-1}}{8C_\gamma}, \frac{\gamma^{-1}}{32C_\gamma L_{\nabla F}^2}\}$, we have*

$$W(\bar{x}^{k+1}, \bar{z}^{k+1}) - W(\bar{x}^k, \bar{z}^k) \leq -\alpha_k \left\{\Theta^k + \Upsilon^k + \alpha_k \Lambda^k + r^{k+1}\right\},$$

where

$$\begin{aligned}
\Theta^k &= \left\{\frac{\gamma^{-1}}{4}\|\bar{x}^k - \bar{y}^k\|^2 + \frac{\lambda}{4}\left\|\nabla F(\bar{x}^k) - \bar{z}^k\right\|^2\right\}, \quad \Lambda^k = \left\{\frac{C_\gamma + 2\lambda}{2}\left\|\bar{\Delta}^{k+1}\right\|^2\right\},\\
\Upsilon^k &= \left\{\frac{2\gamma(1+4\gamma^2 L_{\nabla F}^2)}{n}\left\|\mathbf{Z}_k - \bar{\mathbf{Z}}_k\right\|^2 + \frac{2\left(\gamma^{-1} + 3\gamma L_{\nabla F}^2\right)}{n}\left\|\mathbf{X}_k - \bar{\mathbf{X}}_k\right\|^2\right\},\\
r^{k+1} &= \left\langle \bar{\Delta}^{k+1}, \bar{x}^k - y_+^k + C_\gamma \alpha_k \left(\frac{1}{n}\sum_{i=1}^n \nabla F_i(x_i^k) - \bar{z}^k\right) + 2\lambda\left((1-\alpha_k)\left(\bar{z}^k - \nabla F(\bar{x}^k)\right) + \alpha_k \delta^k\right)\right\rangle.
\end{aligned} \tag{34}$$

Proof. By the smoothness of F and η, we have

$$\begin{aligned}
&F(\bar{x}^{k+1}) - F(\bar{x}^k)\\
\leq&\left\langle \nabla F(\bar{x}^k), \bar{x}^{k+1} - \bar{x}^k\right\rangle + \frac{L_{\nabla F}}{2}\|\bar{x}^{k+1} - \bar{x}^k\|^2 = -\alpha_k\left\langle \nabla F(\bar{x}^k), \bar{x}^k - \bar{y}^k\right\rangle + \frac{L_{\nabla F}\alpha_k^2}{2}\|\bar{x}^k - \bar{y}^k\|^2
\end{aligned} \tag{35}$$

$$\begin{aligned}
&\eta(\bar{x}^k, \bar{z}^k) - \eta(\bar{x}^{k+1}, \bar{z}^{k+1})\\
\leq&\left\langle -\bar{z}^k - \gamma^{-1}(y_+^k - \bar{x}^k), \bar{x}^k - \bar{x}^{k+1}\right\rangle + \left\langle y_+^k - \bar{x}^k, \bar{z}^k - \bar{z}^{k+1}\right\rangle + \frac{C_\gamma}{2}\left(\|\bar{x}^{k+1} - \bar{x}^k\|^2 + \|\bar{z}^{k+1} - \bar{z}^k\|^2\right)\\
=&2\alpha_k\left\langle \bar{z}^k, y_+^k - \bar{x}^k\right\rangle + \gamma^{-1}\alpha_k\|\bar{x}^k - y_+^k\|^2 + \alpha_k\left\langle \bar{v}^{k+1}, \bar{x}^k - \bar{y}^k\right\rangle\\
&+\alpha_k\left\langle \bar{z}^k + \gamma^{-1}(y_+^k - \bar{x}^k) + \bar{v}^{k+1}, \bar{y}^k - y_+^k\right\rangle + \frac{C_\gamma}{2}\left(\alpha_k^2\|\bar{x}^k - \bar{y}^k\|^2 + \|\bar{z}^{k+1} - \bar{z}^k\|^2\right).
\end{aligned} \tag{36}$$

Since y_+^k is the minimizer of a $1/\gamma$-strongly convex function, i.e.,

$$\left\langle \bar{z}^k, y_+^k - \bar{x}^k\right\rangle + \frac{1}{2\gamma}\|y_+^k - \bar{x}^k\|^2 + \Psi(y_+^k) \leq \Psi(\bar{x}^k) - \frac{1}{2\gamma}\|y_+^k - \bar{x}^k\|^2,$$

which together with (36) gives

$$\begin{aligned}
&\eta(\bar{x}^k, \bar{z}^k) - \eta(\bar{x}^{k+1}, \bar{z}^{k+1})\\
\leq& -\gamma^{-1}\alpha_k\|\bar{x}^k - y_+^k\|^2 + \alpha_k\left\langle \bar{v}^{k+1}, \bar{x}^k - \bar{y}^k\right\rangle + \alpha_k\left\langle \bar{z}_k + \gamma^{-1}(y_+^k - \bar{x}^k) + \bar{v}^{k+1}, \bar{y}^k - y_+^k\right\rangle\\
&+2\alpha_k\left(\Psi(\bar{x}^k) - \Psi(y_+^k)\right) + \frac{C_\gamma}{2}\left(\|\bar{x}^{k+1} - \bar{x}^k\|^2 + \|\bar{z}^{k+1} - \bar{z}^k\|^2\right).
\end{aligned} \tag{37}$$

By the convexity of Ψ, we have

$$\Psi(\bar{x}^{k+1}) - \Psi(\bar{x}^k) \leq (1-\alpha_k)\Psi(\bar{x}^k) + \alpha_k\Psi(\bar{y}^k) - \Psi(\bar{x}^k) = \alpha_k\left(\Psi(\bar{y}^k) - \Psi(\bar{x}_i^k)\right). \tag{38}$$

Combining (35), (37), and (38), we have

$$\begin{aligned}
&\left[\Phi(\bar{x}^{k+1}) + \Psi(\bar{x}^{k+1}) - \eta(\bar{x}^{k+1}, \bar{z}^{k+1})\right] - \left[\Phi(\bar{x}^k) + \Psi(\bar{x}^k) - \eta(\bar{x}^k, \bar{z}^k)\right]\\
\leq& -\gamma^{-1}\alpha_k\|\bar{x}^k - y_+^k\|^2 + \alpha_k\left\langle \bar{v}^{k+1} - \nabla F(\bar{x}^k), \bar{x}^k - \bar{y}^k\right\rangle + 2\alpha_k(\Psi(\bar{y}^k) - \Psi(y_+^k))\\
&+\alpha_k\left\langle \bar{z}^k + \gamma^{-1}(y_+^k - \bar{x}^k) + \bar{v}^{k+1}, \bar{y}^k - y_+^k\right\rangle + \frac{(L_{\nabla F} + C_\gamma)\alpha_k^2}{2}\|\bar{x}^k - \bar{y}^k\|^2 + \frac{C_\gamma}{2}\|\bar{z}^{k+1} - \bar{z}^k\|^2.
\end{aligned} \tag{39}$$

Removing non-smooth terms in (39) using (33) in Lemma 3.9, and re-organizing (39) using the decomposition that $\bar{z}^{k+1} - \bar{z}^k = \alpha_k(-\bar{z}^k + \bar{v}^{k+1}) = \alpha_k(\nabla F(\bar{x}^k) - \bar{z}^k) + \alpha_k(\frac{1}{n}\sum_{i=1}^n(\nabla F_i(x_i^k) - \nabla F_i(\bar{x}^k))) + \alpha_k\bar{\Delta}^{k+1}$, we can get

$$\left[\Phi(\bar{x}^{k+1}) + \Psi(\bar{x}^{k+1}) - \eta(\bar{x}^{k+1}, \bar{z}^{k+1})\right] - \left[\Phi(\bar{x}^k) + \Psi(\bar{x}^k) - \eta(\bar{x}^k, \bar{z}^k)\right]$$

$$\le \underbrace{\gamma^{-1}\alpha_k\Big\{ -\|\bar{x}^k - y_+^k\|^2 + \big\langle (y_+^k - \bar{y}^k) + (\bar{x}^k - \bar{y}^k), \bar{y}^k - y_+^k \big\rangle \Big\}}_{\varkappa_1}$$

$$+ \underbrace{\alpha_k \left\langle \frac{1}{n}\sum_{i=1}^n \left(\nabla F_i(x_i^k) - \nabla F_i(\bar{x}^k)\right), \bar{x}^k - y_+^k \right\rangle}_{\varkappa_2} + \underbrace{\alpha_k \left\langle \nabla F(\bar{x}^k) - \bar{z}^k, \bar{y}^k - y_+^k \right\rangle}_{\varkappa_3} + \alpha_k \left\langle \bar{\Delta}^{k+1}, \bar{x}^k - y_+^k \right\rangle$$

$$\frac{(L_{\nabla F} + C_\gamma)\alpha_k^2}{2}\|\bar{x}^k - \bar{y}^k\|^2 + \underbrace{\frac{C_\gamma}{2}\|\bar{z}^{k+1} - \bar{z}^k\|^2}_{\varkappa_4} + \frac{\gamma\alpha_k}{n}\left\|\mathbf{Z}_k - \bar{\mathbf{Z}}_k\right\|^2 + \frac{\gamma^{-1}\alpha_k}{n}\left\|\mathbf{X}_k - \bar{\mathbf{X}}_k\right\|^2.$$

To further simplify the above inequalities, we analyze the terms $\varkappa_1, \varkappa_2, \varkappa_3, \varkappa_4$ separately as follows:

$$\begin{aligned}
\varkappa_1 =& \gamma^{-1}\alpha_k\left\{ -\left\|\bar{x}^k - \bar{y}^k\right\|^2 - \left\langle \bar{x}^k - \bar{y}^k, \bar{y}^k - y_+^k \right\rangle - 2\left\|\bar{y}^k - y_+^k\right\|^2 \right\} \le -\frac{7\gamma^{-1}\alpha_k}{8}\left\|\bar{x}^k - \bar{y}^k\right\|^2, \\
\varkappa_2 \le& 2\gamma\alpha_k \left\| \frac{1}{n}\sum_{i=1}^n \left(\nabla F_i(x_i^k) - \nabla F_i(\bar{x}^k)\right) \right\|^2 + \frac{\gamma^{-1}\alpha_k}{8}\left\|\bar{x}^k - y_+^k\right\|^2 \\
\le& \frac{2\gamma\alpha_k L_{\nabla F}^2}{n}\left\|\mathbf{X}_k - \bar{\mathbf{X}}_k\right\|^2 + \frac{\gamma^{-1}\alpha_k}{4}\left\|\bar{x}^k - \bar{y}^k\right\|^2 + \frac{\gamma^{-1}\alpha_k}{4}\left\|\bar{y}^k - y_+^k\right\|^2, \\
\varkappa_3 \le& \frac{\lambda\alpha_k}{2}\left\|\nabla F(\bar{x}^k) - \bar{z}^k\right\|^2 + \frac{\lambda^{-1}\alpha_k}{2}\left\|\bar{y}^k - y_+^k\right\|^2, \\
\varkappa_4 \le& \frac{C_\gamma\alpha_k^2}{2}\left\{ 2\left\|\nabla F(\bar{x}^k) - \bar{z}^k\right\|^2 + \frac{2L_{\nabla F}^2}{n}\left\|\mathbf{X}_k - \bar{\mathbf{X}}_k\right\|^2 + \left\|\bar{\Delta}^{k+1}\right\|^2 + 2\left\langle \bar{\Delta}^{k+1}, \frac{1}{n}\sum_{i=1}^n \nabla F_i(x_i^k) - \bar{z}^k \right\rangle \right\}.
\end{aligned}$$

Combining the above results with (31) in Lemma 3.8 and the definition of $W(\bar{x}^k, \bar{z}^k)$ in (1), we have

$$\begin{aligned}
&W(\bar{x}^{k+1}, \bar{z}^{k+1}) - W(\bar{x}^k, \bar{z}^k) \le \alpha_k\left\{ -\frac{5}{8}\gamma^{-1} + \frac{(L_{\nabla F} + C_\gamma)\alpha_k}{2} + 2\lambda L_{\nabla F}^2 \right\}\|\bar{x}^k - \bar{y}^k\|^2 \\
&+ \alpha_k\left\{ -\frac{\lambda}{2} + C_\gamma\alpha_k \right\}\left\|\nabla F(\bar{x}^k) - \bar{z}^k\right\|^2 + \frac{C_\gamma\alpha_k^2}{2}\left\|\bar{\Delta}^{k+1}\right\|^2 + \frac{(\gamma^{-1} + 2\lambda^{-1})\alpha_k}{4}\left\|y_+^k - \bar{y}^k\right\|^2 \\
&+ \frac{\gamma\alpha_k}{n}\left\|\mathbf{Z}_k - \bar{\mathbf{Z}}_k\right\|^2 + \frac{\left(\gamma^{-1} + 2\gamma L_{\nabla F}^2 + 2\lambda L_{\nabla F}^2 + C_\gamma L_{\nabla F}^2\alpha_k\right)\alpha_k}{n}\left\|\mathbf{X}_k - \bar{\mathbf{X}}_k\right\|^2 \\
&+ \alpha_k \underbrace{\left\langle \bar{\Delta}^{k+1}, \bar{x}^k - y_+^k + C_\gamma\alpha_k\left(\frac{1}{n}\sum_{i=1}^n \nabla F_i(x_i^k) - \bar{z}^k\right) + 2\lambda\left((1-\alpha_k)\left(\bar{z}^k - \nabla F(\bar{x}^k)\right) + \alpha_k\delta^k\right) \right\rangle}_{r^{k+1}}.
\end{aligned} \tag{40}$$

In addition, from Lemma 3.5, we already know

$$\left\|y_+^k - \bar{y}^k\right\|^2 \le \frac{2}{n}\left\{ \|\mathbf{X}_k - \bar{\mathbf{X}}_k\|^2 + \gamma^2\|\mathbf{Z}_k - \bar{\mathbf{Z}}_k\|^2 \right\}.$$

Finally, choosing α_k such that $\alpha_k \le \min\{\frac{\gamma^{-1}}{8L_{\nabla F}}, \frac{\gamma^{-1}}{8C_\gamma}, \frac{\gamma^{-1}}{32C_\gamma L_{\nabla F}^2}\}$ and $\lambda = \frac{\gamma^{-1}}{8L_{\nabla F}^2}$, we can re-organize the terms in (40) as follows and complete the proof.

$$\begin{aligned}
&W(\bar{x}^{k+1}, \bar{z}^{k+1}) - W(\bar{x}^k, \bar{z}^k) \\
&\le -\alpha_k \underbrace{\left\{ \frac{\gamma^{-1}}{4}\|\bar{x}^k - \bar{y}^k\|^2 + \frac{\lambda}{4}\left\|\nabla F(\bar{x}^k) - \bar{z}^k\right\|^2 \right\}}_{\Theta^k} + \alpha_k^2 \underbrace{\left\{ \frac{C_\gamma + 2\lambda}{2}\left\|\bar{\Delta}^{k+1}\right\|^2 \right\}}_{\Lambda^k} + \alpha_k r^k \\
&+ \alpha_k \underbrace{\left\{ \frac{2\gamma(1 + 4\gamma^2 L_{\nabla F}^2)}{n}\left\|\mathbf{Z}_k - \bar{\mathbf{Z}}_k\right\|^2 + \frac{2\left(\gamma^{-1} + 3\gamma L_{\nabla F}^2\right)}{n}\left\|\mathbf{X}_k - \bar{\mathbf{X}}_k\right\|^2 \right\}}_{\Upsilon^k}.
\end{aligned} \tag{41}$$

□

4 DISCUSSION ON THE DEFINTION OF CONSENSUS ERROR

In this section, we briefly discuss two different functions that measure the consensus violation of vectors among agents. Suppose agent i has $x_i \in \mathbb{R}^d$, our consensus error can be viewed as

$$f(x_1, ..., x_n) = \frac{1}{n} \sum_{i=1}^{n} \|x_i - \bar{x}\|^2,$$

where $\bar{x} := \frac{1}{n} \sum_{i=1}^{n} x_i$, while SPPDM in Wang et al. [2021] defines (see Eq. (4a), (4b), (5a), (5b), and (41) in Wang et al. [2021])

$$\begin{aligned} g_W(x_1, ..., x_n) &= \sum_{i \sim j, 1 \le i < j \le n} \|x_i - x_j\|^2 \\ &= \frac{1}{2} \sum_{i=j \text{ or } i \sim j} \left(\|x_i - \bar{x}\|^2 + \|x_j - \bar{x}\|^2 - 2 \langle x_i - \bar{x}, x_j - \bar{x} \rangle \right) \end{aligned} \tag{42}$$

over a connected network whose weighted adjacency matrix (i.e., mixing matrix) is W, and the stationarity therein is defined by using g_W. $i \sim j$ means agents i and j are neighbors. Note that in general the relationship between f and g_W largely depends on W. We consider several special cases:

- W is a complete graph. By (42) we have

$$g_W(x_1, ..., x_n) = n \sum_{i=1}^{n} \|x_i - \bar{x}\|^2 - \left\langle \sum_{i=1}^{n} (x_i - \bar{x}), \sum_{j=1}^{n} (x_j - \bar{x}) \right\rangle = n^2 f(x_1, ..., x_n).$$

- W is a cycle. By (42) we have

$$g_W(x_1, ..., x_n) \le \sum_{i \sim j, 1 \le i < j \le n} 2 \left(\|x_i - \bar{x}\|^2 + \|x_j - \bar{x}\|^2 \right) = 4n f(x_1, ..., x_n).$$

- W is a simple path such that i and $i+1$ are adjacent for all $1 \le i \le n-1$, and $x_i = i \in \mathbb{R}$. Note that in this case, we can directly obtain $g_W(x_1, ..., x_n) = n - 1$. For f we have

$$f(x_1, ..., x_n) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{n+1}{2} - i \right)^2 = \Theta(n^2),$$

  which implies $g_W = \Theta(\frac{f}{n})$.

We know from the above examples that the order (in terms of n) of g_W/f can range from $\frac{1}{n}$ to n^2. Hence these two types of consensus error are not comparable if no additional assumptions are given, and thus we only include SPPDM in the experiments and do not compare their complexity results to ours.

References

Gabriel Mancino-ball, Shengnan Miao, Yangyang Xu, and Jie Chen. Proximal stochastic recursive momentum methods for nonconvex composite decentralized optimization. In *AAAI Conference on Artificial Intelligence*, 2023.

Zhiguo Wang, Jiawei Zhang, Tsung-Hui Chang, Jian Li, and Zhi-Quan Luo. Distributed stochastic consensus optimization with momentum for nonconvex nonsmooth problems. *IEEE Transactions on Signal Processing*, 69:4486–4501, 2021.

Ran Xin, Subhro Das, Usman A Khan, and Soummya Kar. A stochastic proximal gradient framework for decentralized non-convex composite optimization: Topology-independent sample complexity and communication efficiency. *arXiv preprint arXiv:2110.01594*, 2021.