Supporting Material – Appendix

This document serves as supporting material of the paper entitled "Accelerated Primal-Dual Algorithms for Distributed Smooth Convex Optimization over Networks" and contains all the proofs of the main results in the paper.

A Review of existing distributed algorithms and their connections

This section shows the generality of the first-order oracle $\mathcal{A}$ in (6) and the proposed distributed primal-dual algorithmic framework (12) by casting several existing distributed algorithms in the oracle form (6) and algorithmic form (12).

A.1 Some distributed optimization methods

Distributed gradient methods One of the first distributed algorithms for Problem (1) was proposed in the seminal work Nedic and Ozdaglar (2009) and called Distributed Gradient Algorithm (DGD). DGD employing constant step-size can be written in compact form as:

$$\mathbf{x}^{k+1} = \mathbf{W}\mathbf{x}^k - \gamma\nabla f(\mathbf{x}^k), \tag{20}$$

where $\mathbf{W} \in \mathcal{W}_{\mathcal{G}}$. Defining $\mathbf{x}^{(t_k)} = \mathbf{x}^k$, DGD can be rewritten in a piece-wise continuous form as

$$\begin{aligned} \mathbf{x}^{(t_{k+1})} &= \mathbf{W}\mathbf{x}^{(t_k)} - \gamma\nabla f(\mathbf{x}^{(t_k)}), \\ \mathbf{x}^{(t)} &= \mathbf{x}^{(t_k)},\ t_k \le t < t_{k+1}, \end{aligned} \tag{21}$$

which is an instance of the oracle $\mathcal{A}$.

Distributed gradient tracking methods The distributed gradient tracking algorithm, first proposed in Di Lorenzo and Scutari (2016); Xu et al. (2015) and further analyzed in Nedich et al. (2017); Qu and Li (2017b), reads

$$\mathbf{x}^{k+1} = \mathbf{W}\mathbf{x}^k - \gamma\mathbf{y}^k \tag{22a}$$

$$\mathbf{y}^{k+1} = \mathbf{W}\mathbf{y}^k + \nabla f(\mathbf{x}^{k+1}) - \nabla f(\mathbf{x}^k) \tag{22b}$$

where $\mathbf{y}_k$ is an auxiliary variable aiming at tracking the gradient of the sum-cost function. The above algorithm is proved to converge at linear rate to a solution of Problem (2), under proper conditions on the stepsize γ. To show its relationship to the oracle, we first rewrite (22) absorbing the tracking variable $\mathbf{y}$, which yields

$$\mathbf{x}^{k+2} = 2\mathbf{W}\mathbf{x}^{k+1} - \mathbf{W}^2\mathbf{x}^k - \gamma(\nabla f(\mathbf{x}^{k+1}) - \nabla f(\mathbf{x}^k)),$$

with $\mathbf{x}^1 = \mathbf{W}\mathbf{x}^0 - \gamma\nabla f(\mathbf{x}^0)$. It is clear that the gradient tracking algorithm belongs to the oracle $\mathcal{S}$, as each iteration k only involves the historical neighboring information and local gradients at $k-1$ and $k-2$.

Distributed primal-dual methods Distributed primal-dual algorithms can be generally written in the following form Shi et al. (2015)

$$\mathbf{x}^{k+1} = \mathbf{W}\mathbf{x}^k - \gamma\nabla f(\mathbf{x}^k) - \mathbf{y}^k \tag{23a}$$

$$\mathbf{y}^{k+1} = \mathbf{y}^k + (\mathbf{I} - \mathbf{W})\mathbf{x}^{k+1} \tag{23b}$$

where $\mathbf{y}_k$ is the dual variable. When $\mathbf{y}^0 = \mathbf{0}$, the algorithm 23 can solve problem (2). Evaluating (23a) at $k+1$ and substituting it into (23b) yields

$$\mathbf{x}^{k+2} = 2\mathbf{W}\mathbf{x}^{k+1} - \mathbf{W}\mathbf{x}^k - \gamma(\nabla f(\mathbf{x}^{k+1}) - \nabla f(\mathbf{x}^k)), \tag{24}$$

with $\mathbf{x}^1 = \mathbf{W}\mathbf{x}^0 - \gamma\mathbf{W}\nabla f(\mathbf{x}^0)$. It is easy to check that (24) belongs to the oracle $\mathcal{A}$.

Remark 4. There are some other distributed algorithms that do not belong to the categories above such as Chen and Sayed (2012). However, using similar arguments as above, one can show that they are instances of the oracle $\mathcal{A}$.

A.2 Connections between gradient tracking and primal-dual methods

We reveal here an unknown interesting connection between primal-dual methods and gradient tracking based methods. More specifically, setting in (12a) $\mathbf{A} = \mathbf{W}^2$ and $\mathbf{B} = (\mathbf{I} - \mathbf{W})^2$, one can easily recover gradient tracking methods from the primal-dual ones. To simplify the presentation, we consider a slightly different form of (12a), that is

$$\mathbf{x}^{k+1} = \mathbf{W}^2(\mathbf{x}^k - \gamma(\nabla f(\mathbf{x}^k)) + (\mathbf{I} - \mathbf{W})\mathbf{y}^k, \tag{25a}$$
$$\mathbf{y}^{k+1} = \mathbf{y}^k + (\mathbf{I} - \mathbf{W})\mathbf{x}^{k+1}. \tag{25b}$$

Then, from (25a), we have at iteration $k+1$

$$\mathbf{x}^{k+2} = \mathbf{W}^2\mathbf{x}^{k+1} - \gamma\mathbf{W}^2\nabla f(\mathbf{x}^{k+1}) - (\mathbf{I} - \mathbf{W})\mathbf{y}^{k+1}$$

Subtracting (25a) from the above equation we have

$$\begin{aligned}\mathbf{x}^{k+2} - \mathbf{x}^{k+1} &= \mathbf{W}^2\mathbf{x}^{k+1} - \mathbf{W}^2\mathbf{x}^k - \gamma\mathbf{W}^2(\nabla f(\mathbf{x}^{k+1}) - \nabla f(\mathbf{x}^k)) - (\mathbf{I} - \mathbf{W})(\mathbf{y}^{k+1} - \mathbf{y}^k)\\ &= \mathbf{W}^2\mathbf{x}^{k+1} - \mathbf{W}^2\mathbf{x}^k - \gamma\mathbf{W}^2(\nabla f(\mathbf{x}^{k+1}) - \nabla f(\mathbf{x}^k)) - (\mathbf{I} - \mathbf{W})^2\mathbf{x}^{k+1}\\ &= 2\mathbf{W}\mathbf{x}^{k+1} - \mathbf{x}^{k+1} - \mathbf{W}^2\mathbf{x}^k - \gamma\mathbf{W}^2(\nabla f(\mathbf{x}^{k+1}) - \nabla f(\mathbf{x}^k)).\end{aligned}$$

Rearranging terms leads to

$$\mathbf{x}^{k+2} - \mathbf{W}\mathbf{x}^{k+1} = \mathbf{W}(\mathbf{x}^{k+1} - \mathbf{W}\mathbf{x}^k) - \gamma\mathbf{W}^2(\nabla f(\mathbf{x}^{k+1}) - \nabla f(\mathbf{x}^k)).$$

Let $-\gamma\mathbf{W}\mathbf{y}^k = \mathbf{x}^{k+1} - \mathbf{W}\mathbf{x}^k$ and suppose $\mathbf{W}$ is invertible. Then, we have

$$\mathbf{x}^{k+1} = \mathbf{W}(\mathbf{x}^k - \gamma\mathbf{y}^k) \tag{26a}$$
$$\mathbf{y}^{k+1} = \mathbf{W}(\mathbf{y}^k + \nabla f(\mathbf{x}^{k+1}) - \nabla f(\mathbf{x}^k)) \tag{26b}$$

which is exactly the standard gradient tracking method in the ATC form (Di Lorenzo and Scutari, 2016; Xu et al., 2015).

B Proof of Proposition 1

Statement (a) is a direct result of (Bertsekas et al., 2003, Prop. 6.1.1). We prove next statement (b). Suppose that there are two optimal solutions $\mathbf{x}^\star$ and $\widetilde{\mathbf{x}}^\star$ such that

$$\nabla f(\mathbf{x}^\star),\, \nabla f(\widetilde{\mathbf{x}}^\star) \in \mathcal{C}^\perp, \qquad \mathbf{x}^\star,\, \widetilde{\mathbf{x}}^\star \in \mathcal{C} \qquad \text{and} \qquad f(\mathbf{x}^\star) = f(\widetilde{\mathbf{x}}^\star).$$

Since $G(\mathbf{x}, \mathbf{x}^\star) = f(\mathbf{x}) - f(\mathbf{x}^\star) - \langle \nabla f(\mathbf{x}^\star), \mathbf{x} - \mathbf{x}^\star\rangle \geq 0$ for all $\mathbf{x} \in \mathbb{R}^{m\times d}$, and $G(\widetilde{\mathbf{x}}^\star, \mathbf{x}^\star) = 0$, $\widetilde{\mathbf{x}}^\star$ is the global minimizer of G. Hence, it must be $\nabla f(\mathbf{x}^\star) = \nabla f(\widetilde{\mathbf{x}}^\star)$, implying

$$\begin{aligned}G(\mathbf{x}, \mathbf{x}^\star) &= f(\mathbf{x}) - f(\mathbf{x}^\star) - \langle \nabla f(\mathbf{x}^\star), \mathbf{x} - \mathbf{x}^\star\rangle\\ &= f(\mathbf{x}) - f(\widetilde{\mathbf{x}}^\star) - \langle \nabla f(\widetilde{\mathbf{x}}^\star), \mathbf{x} - \widetilde{\mathbf{x}}^\star\rangle = G(\mathbf{x}, \widetilde{\mathbf{x}}^\star), \qquad \forall \mathbf{x} \in \mathbb{R}^{m\times d},\end{aligned}$$

where we have used the fact that $\langle \nabla f(\mathbf{z}), \mathbf{z}\rangle = 0$ for any optimal solution $\mathbf{z}$.

C Proof of Theorem 2

As elaborated in Section 3.1, to study the lower complexity bound of the first order distributed oracle $\mathcal{A}$ solving Problem (2) [and thus (1)], one can consider ϵ-solutions (i.e., $\bar{\mathbf{x}} \in \mathbb{R}^{m\times d}$ such that $G(\bar{\mathbf{x}}) \leq \epsilon$) of the following convex optimization problem:

$$\min_{\mathbf{x}\in\mathbb{R}^{m\times d}} G(\mathbf{x}) = f(\mathbf{x}) - \langle \nabla f(\mathbf{x}^\star), \mathbf{x} - \mathbf{x}^\star\rangle - f(\mathbf{x}^\star). \tag{27}$$

The proof is based on building a worst-case objective function in (27) and network graph for which the lower bound is achieved by the best available gossip, distributed algorithm in the oracle $\mathcal{A}$. To do so we build on the

cost function first introduced in Arjevani and Shamir (2015) for a fully connected network and later used for a peer-to-peer network in Scaman et al. (2017), both for smooth strongly convex problems. Since we use a different metric (the Bregman distance) to define the lower bound and consider smooth convex problems (not necessarily strongly-convex), the analysis in Scaman et al. (2017) cannot be readily applied to our setting and an ad-hoc proof of the theorem is needed.

The path of our proof is the following: i) We start with a simple network consisting of two agents such that the diameter of the network will not come into play–see Sec. C.1; and ii) then we extend our results to a general network composed by an arbitrary number of agents–see Sec. C.2.

C.1 A simple two-agent network

We state the result on the simple two-agent network as the following.

Theorem 8. *Consider a two-agent network with cost functions given in* (28). *Let* $\{\mathbf{x}^k\}_{k=0}^{\infty}$ *be the sequence generated by any first-order algorithm* $\mathcal{A}$. *Suppose* $0 \le k \le \frac{d-1}{2}$. *Then, we have*

$$G(\mathbf{x}^k) = \Omega\left(\frac{L_f \left\|\mathbf{x}^0 - \mathbf{x}^\star\right\|^2}{(k+1)^2} + \frac{\left\|\mathbf{x}^0 - \mathbf{x}^\star\right\| \|\nabla f(\mathbf{x}^\star)\|}{k+1}\right).$$

We prove the above result in three steps: i) we construct the hard function in Sec. C.1.1, which is the worst-case function for all methods belonging to the oracle $\mathcal{A}$; ii) we introduce some intermediate result in Sec. C.1.2, which is related to our specific metric–the Bregman distance G, and iii) building on step i-ii, we derive the lower bound in Sec. C.1.3.

C.1.1 Construction of the hard function

Consider a network composed of two agents. The idea of the proof of the lower complexity bound relies on splitting the "hard" function used by Nesterov to prove the iteration complexity of first-order gradient methods for (centralized) smooth convex problems across the agents(Nesterov, 2013, Chapter 2). More specifically, consider the following cost functions for the two agents:

$$\begin{cases} f_{1,[k]}(x_1) = \frac{L_f}{8} x_1^\top \mathbf{A}_{1,[k]} x_1 - \frac{L_f}{4} e_1^\top x_1, \\ f_{2,[k]}(x_2) = \frac{L_f}{8} x_2^\top \mathbf{A}_{2,[k]} x_2, \end{cases} \tag{28}$$

where

$$\mathbf{A}_{1,[k]} := \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & \cdots \\ 0 & 1 & -1 & 0 & 0 & \cdots \\ 0 & -1 & 1 & 0 & 0 & \cdots \\ 0 & 0 & 0 & 1 & -1 & \cdots \\ 0 & 0 & 0 & -1 & 1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}, \qquad \mathbf{A}_{2,[k]} := \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & \cdots \\ -1 & 1 & 0 & 0 & 0 & \cdots \\ 0 & 0 & 1 & -1 & 0 & \cdots \\ 0 & 0 & -1 & 1 & 0 & \cdots \\ 0 & 0 & 0 & 0 & 1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \tag{29}$$

are two $d \times d$ matrices with their leading principal minors of order $k \in [1, d]$ having non-zero block diagonals while the rest being zero.

The key idea of Nesterov proof for the lower complexity bound of centralized first-order gradient methods consists in designing the "hardest" function to be minimized by any method belonging to the oracle. This function was shown to be such that, at iteration k, all these methods produce a new iterate whereby only the kth component is updated. The choice of the two agents' cost functions in (28) follows the same rationale: the structure of $\mathbf{A}_{1,[k]}$ and $\mathbf{A}_{2,[k]}$ is such that none of the two agents is able to make progresses towards optimality, i.e., updating the next component in their local optimization vector (with odd index for agent 1 and even index for agent 2) just performing local gradient updates and without communication with each other. This means that at certain stages a communication between the two agents is necessary for the algorithm to make progresses towards optimality. Building on the above idea, we begin establishing the lower complexity bound for the two-agent network problem in terms of gradient evaluations.

C.1.2 Intermediate results

Now substituting $f(\mathbf{x}) = f_{1,[k]}(x_1) + f_{2,[k]}(x_2)$ in (27) and ignoring constants, we obtain

$$\min_{\mathbf{x}\in\mathbb{R}^{2\times d}} f_{[k]}(\mathbf{x}) := f_{1,[k]}(x_1) + f_{2,[k]}(x_2) - \left\langle [\nabla f_{1,[k]}(x_1^\star), \nabla f_{2,[k]}(x_2^\star)]^\top, \mathbf{x} \right\rangle. \tag{30}$$

We denote the optimal function value of the above problem as $f_{[k]}^\star$. It is obvious that, when agents reach consensus, i.e., $x_1 = x_2$, the function $f_{[k]}(\mathbf{x})$ will reduce to the Nesterov's "hard" function (Nesterov, 2013, Section 2.1.2), for which we have the optimal solution

$$x_1^\star = x_2^\star = [\underbrace{\frac{k}{k+1}, \frac{k-1}{k+1}, ..., \frac{1}{k+1}}_{\text{the first } k \text{ components}}, 0, \dots, 0]^\top \in \mathbf{span}(e_1, e_2, ..., e_k),$$

and it yields

$$\|\mathbf{x}^\star\|^2 = \|x_1^\star\|^2 + \|x_2^\star\|^2 \le \frac{2}{3}(k+1) \tag{31}$$

and $f_{[k]}^\star = \frac{L_f}{8}(-1 + \frac{1}{k+1})$. Also, we have

$$\begin{cases} \nabla f_{1,[k]}(x_1^\star) = \frac{L_f}{4}(\mathbf{A}_{1,[k]} x_1^\star - e_1) = -\frac{L_f}{4}\frac{1}{k+1} a_{[k]} \\ \nabla f_{2,[k]}(x_2^\star) = \frac{L_f}{4}\mathbf{A}_{2,[k]} x_2^\star = \frac{L_f}{4}\frac{1}{k+1} a_{[k]}, \end{cases} \tag{32}$$

where

$$a_{[k]} = [\underbrace{1, -1, 1, -1, 1, -1, ...}_{1/-1 \text{ alternates } k \text{ times}}, 0, \dots, 0]^\top.$$

Thus, we further have

$$\|\nabla f(\mathbf{x}^\star)\| = \sqrt{\left\|\nabla f_{1,[k]}(x_1^\star)\right\|^2 + \left\|\nabla f_{2,[k]}(x_2^\star)\right\|^2} = \sqrt{\frac{2L_f^2 a_{[k]}^\top a_{[k]}}{16(k+1)^2}} = \frac{\sqrt{2k} L_f}{4(k+1)}. \tag{33}$$

Note that quantities (31) and (33) will be useful later to relate the complexities with $\left\|\mathbf{x}^0 - \mathbf{x}^\star\right\|$ and $\|\nabla f(\mathbf{x}^\star)\|$. According to (32), Problem (30) further becomes

$$\min_{\mathbf{x}\in\mathbb{R}^{2\times d}} f_{[k]}(\mathbf{x}) = f_{1,[k]}(x_1) + f_{2,[k]}(x_2) + \frac{L_f}{4(k+1)} \left\langle a_{[k]}, x_1 - x_2 \right\rangle. \tag{34}$$

In the following, we study the above problem when the local variables x_1 and x_2 are restricted to the truncating subspace of $\mathbb{R}^d$, as a stepping stone to prove Theorem 8.

Let $\mathbb{R}^{k,d} := \mathbf{span}(e_i \in \mathbb{R}^d \,|\, 1 \le i \le k)$ denote the subspace composed of vectors whose only first k components are possibly non-zeros and $\mathcal{L}^k := \mathbf{span}(\nabla f_i(x_i^l) \,|\, 0 \le l \le k-1, i \in \mathcal{V})$. It should be noted that the local cost functions constructed in (28) are dependent on k, but hereafter subscripts indicating this dependence are omitted for simplicity.

Lemma 9 (Linear Span)**.** *Let $\{\mathbf{x}^k\}_{k=0}^\infty$ be the sequence generated by any distributed first-order algorithm $\mathcal{A}$ with $\mathbf{x}^0 = \mathbf{0}$. Then, $x_i^k \in \mathcal{L}^k$ for all $k \ge 0$ and all $i \in \mathcal{V}$.*

The proof of the above lemma is straightforward, since local communication steps do not change the space spanned by the historical gradient vectors generated over the network.

Lemma 10. *Let $\mathbf{x}^0 = \mathbf{0}$. For the two-agent problem* (28)*, we have $\mathcal{L}^k \subseteq \mathbb{R}^{k,d}$.*

Proof. Since $\mathbf{x}^0 = \mathbf{0}$, we have $\nabla f_1(x_1^0) = -\frac{L_f}{4} e_1 \in \mathbb{R}^{1,d}, \nabla f_2(x_2^0) = \mathbf{0} \in \mathbb{R}^{1,d}$ and thus $\mathcal{L}^1 = \mathbf{span}(\nabla f_1(x_1^0), \nabla f_2(x_2^0)) \subseteq \mathbb{R}^{1,d}$. Now, let $x_i^j \in \mathcal{L}^j \subseteq \mathbb{R}^{j,d}$. Without loss of generality, let us assume j is odd. Then, according to the structure of ∇f_1, we have $\nabla f_1(x_1^j) = \frac{L_f}{4}(\mathbf{A}_{1,[k]} x_1^j - e_1) \in \mathbb{R}^{j,d}$, but multiplying $\mathbf{A}_{1,[k]}$ from the left of $x_1^j \in \mathbb{R}^{j,d}$ will not increase the number of nonzeros to $j+1$. By contrast, for ∇f_2, we have $\nabla f_2(x_2^j) = \frac{L_f}{4}\mathbf{A}_{2,[k]} x_2^j \in \mathbb{R}^{j+1,d}$ and $\mathbf{A}_{2,[k]}$ is now able to increase the number of non-zeros. Therefore, we have $\mathcal{L}^{j+1} = \mathcal{L}^j + \mathbf{span}(\nabla f_1(x_1^j), \nabla f_2(x_2^j)) \subseteq \mathbb{R}^{j+1,d}$ and we can complete the proof by induction. □

Lemma 11. *Consider Problem* (34). *Let* $f^\star_{[k,j]} := \min_{\mathbf{x}_i \in \mathbb{R}^{j,d}, \forall i \in \mathcal{V}} f_{[k]}(\mathbf{x})$*; we have*

$$f^\star_{[k,j]} = -\frac{L_f}{8}\left(\frac{k^2}{(k+1)^2} + \frac{j}{(k+1)^2}\right).$$

Proof. Let $x_i \in \mathbb{R}^{1,d}$, $i \in \mathcal{V}$. Then, the cost function in (34) becomes

$$f_{[k,1]}(\mathbf{x}) := \frac{L_f}{4}[0.5x_{11}^2 - x_{11} + \frac{1}{(k+1)}(x_{11} - x_{21}) + 0.5x_{21}^2]$$

which attains the optimum $f^\star_{[k,1]} = \frac{L_f}{8}(-\frac{k^2}{(k+1)^2} - \frac{1}{(k+1)^2})$.

Likewise, letting $x_i \in \mathbb{R}^{2,d}$, $i \in \mathcal{V}$, we have

$$f_{[k,2]}(\mathbf{x}) := \frac{L_f}{4}[0.5x_{11}^2 + 0.5x_{12}^2 - x_{11} - \frac{1}{k+1}(x_{21} - x_{11}) + \frac{1}{k+1}(x_{22} - x_{12}) + 0.5(x_{21} - x_{22})^2]$$

which yields $f^\star_{[k,2]} = \frac{L_f}{8}(-\frac{k^2}{(k+1)^2} - \frac{2}{(k+1)^2})$. Also, for $\mathbf{x}_i \in \mathbb{R}^{3,d}$, $i \in \mathcal{V}$, we have

$$\begin{aligned} f_{[k,3]}(\mathbf{x}) := \frac{L_f}{4}[&0.5x_{11}^2 + 0.5(x_{12} - x_{13})^2 - x_{11} - \frac{1}{k+1}(x_{21} - x_{11}) + \frac{1}{k+1}(x_{22} - x_{12}) \\ &- \frac{1}{k+1}(x_{23} - x_{13}) + 0.5(x_{21} - x_{22})^2 + 0.5x_{23}^2], \end{aligned}$$

which gives $f^\star_{[k,3]} = \frac{L_f}{8}(-\frac{k^2}{(k+1)^2} - \frac{3}{(k+1)^2})$.

In fact, by induction, it is not difficult to show that, when j is odd, for $\mathbf{x}_i \in \mathbb{R}^{j,d}$, $i \in \mathcal{V}$, we have

$$\begin{aligned} f_{[k,j]}(\mathbf{x}) := \frac{L_f}{4}\Bigg(&0.5x_{11}^2 - \frac{k}{k+1}x_{11} + \sum_{i=1}^{\frac{j-1}{2}}\left(0.5(x_{2(2i)} - x_{2(2i-1)})^2 - \frac{1}{k+1}(x_{2(2i)} - x_{2(2i-1)})\right) \\ &+0.5x_{2j}^2 - \frac{1}{k+1}x_{2j} + \sum_{i=1}^{\frac{j-1}{2}}\left(0.5(x_{1(2i)} - x_{1(2i+1)})^2 - \frac{1}{k+1}(x_{2i} - x_{1(2i+1)})\right)\Bigg), \end{aligned}$$

which yields

$$f^\star_{[k,j]} = -\frac{L_f}{8}\left(\frac{k^2}{(k+1)^2} + \frac{j}{(k+1)^2}\right).$$

When j is even, for $\mathbf{x}_i \in \mathbb{R}^{j,d}$, $i \in \mathcal{V}$, we have

$$\begin{aligned} f_{[k,j]}(\mathbf{x}) = \frac{L_f}{4}\Bigg(&0.5x_{11}^2 - \frac{k}{k+1}x_{11} + \sum_{i=1}^{\frac{j}{2}}\left(0.5(x_{2(2i)} - x_{2(2i-1)})^2 - \frac{1}{k+1}(x_{2(2i)} - x_{2(2i-1)})\right) \\ &+0.5x_{1j}^2 - \frac{1}{k+1}x_{1j} + \sum_{i=1}^{\frac{j}{2}-1}\left(0.5(x_{1(2i)} - x_{1(2i+1)})^2 - \frac{1}{k+1}(x_{2i} - x_{1(2i+1)})\right)\Bigg) \end{aligned}$$

which also yields

$$f^\star_{[k,j]} = -\frac{L_f}{8}\left(\frac{k^2}{(k+1)^2} + \frac{j}{(k+1)^2}\right).$$

The proof is completed by combining the two cases above. □

C.1.3 Proof of Theorem 8

We can now prove the theorem. Let us fix k and apply the first-order gossip algorithm $\mathcal{A}$ to minimize $f_{[2k+1]}$. Since $\mathbf{x}^0 = \mathbf{0}$, invoking Lemma 11, we have

$$\begin{aligned} G(\mathbf{x}^k) &= f_{[2k+1]}(\mathbf{x}^k) - f^\star_{[2k+1]} \geq \min_{\mathbf{x}\in\mathbb{R}^{k,d}} f_{[2k+1]}(\mathbf{x}) - f^\star_{[2k+1]} = f^\star_{[2k+1,k]} - f^\star_{[2k+1]} \\ &\geq \frac{L_f}{8}(1 - \frac{1}{2(k+1)} - \frac{(2k+1)^2}{4(k+1)^2} - \frac{k}{4(k+1)^2}) \\ &= \frac{L_f}{32(k+1)} = \Theta\left(\frac{L_f \left\|\mathbf{x}^0 - \mathbf{x}^\star\right\|^2}{(k+1)^2} + \frac{\left\|\mathbf{x}^0 - \mathbf{x}^\star\right\| \|\nabla f(\mathbf{x}^\star)\|}{(k+1)}\right), \end{aligned}$$

where the last inequality comes from the previously developed facts $\|\mathbf{x}^\star\|^2 = \Theta\left(k+1\right)$, $\|\nabla f(\mathbf{x}^\star)\| = \Theta\left(\frac{L_f}{\sqrt{k+1}}\right)$ and thus $\left\|\mathbf{x}^0 - \mathbf{x}^\star\right\|^2 = \Theta\big(\frac{k+1}{L_f}\left\|\mathbf{x}^0 - \mathbf{x}^\star\right\| \|\nabla f(\mathbf{x}^\star)\|\big)$. This completes the proof for the two-agent network. □

Remark 5. The lower bound we develop in Theorem 8 for distributed scenarios has similar structure of that of the recent paper Ouyang and Xu (2018), where the lower bound is derived for general equality-constrained problems in centralized scenarios (i.e., $\mathbf{A}\mathbf{x} = \mathbf{b}$). Notice that the results and techniques therein can not apply to our distributed setting, as we require $\mathbf{b} = \mathbf{0}$ and $\mathbf{A} \in \mathcal{W}_\mathcal{G}$ while the lower bound in Ouyang and Xu (2018) is determined by a choice of $\mathbf{b}$ and $\mathbf{A}$ that does not meet our requirement.

C.2 Proof of Theorem 2

Following the same path of Scaman et al. (2017), we now extend the above analysis to the general network setting (arbitrary number of agents) by employing a line graph and constructing certain number of pairwise two-agent networks as in (28) from the left and the right of the line graph, respectively, yielding two subgroups. Between these two subgroups, we place a number (proportional to the diameter of the network) of agents with zero cost functions to ensure the necessity of communications between the agents in the two subgroups. To prove the time complexity lower bound, we then leverage the effect of the network by establishing the connection between the diameter of the network and the eigengap of the gossip matrix.

Let $\eta_n = \frac{1-\cos\left(\frac{\pi}{T}\right)}{1+\cos\left(\frac{\pi}{T}\right)}$. For a given $\eta \in (0,1]$, there exists $n \geq 2$ such that $\eta_n \geq \eta > \eta_{n+1}$. We treat the cases $n = 2$ and $n \geq 3$ separately. Let us first consider the case $n \geq 3$. There exists a line graph of $m = n$ agents and associated Laplacian weight matrix with eigengap η. Now, let us define two subsets of agents as $\mathcal{A}_l = \left\{i \middle| 1 \leq i \leq \lceil \zeta m \rceil\right\}$ and $\mathcal{A}_r = \left\{i \middle| \lfloor (1-\zeta)m \rfloor + 1 \leq i \leq m\right\}$, which lie on the left and the right of the line graph, respectively; the parameter $\zeta \in (0, \frac{1}{2})$ is to be determined. The distance between the two subsets is thus $d_c \triangleq \lfloor (1-\zeta)m \rfloor + 1 - \lceil \zeta m \rceil$. The class of local functions is defined as follows

$$f_i = \begin{cases} \frac{L_f}{8} x_i^\top \mathbf{A}_{1,[k]} x_i - \frac{L_f}{4} e_1^\top x_i & \forall i \in \mathcal{A}_l \\ \frac{L_f}{8} x_i^\top \mathbf{A}_{2,[k]} x_i & \forall i \in \mathcal{A}_r \\ 0 & \text{otherwise} \end{cases} \tag{35}$$

where $\mathbf{A}_{1,[k]}, \mathbf{A}_{2,[k]}$ are the two matrices defined in (29). Similarly to the two-agent network case (cf. Sec. C.1), we have

$$\|\mathbf{x}^\star\|^2 \leq \frac{m}{3}(k+1), \;\; \|\nabla f(\mathbf{x}^\star)\| \leq \sqrt{2(\zeta m + 1)}\frac{\sqrt{k} L_f}{4(k+1)},$$

and Problem (27) becomes

$$\min_{\mathbf{x}\in\mathbb{R}^{m\times d}} f_{[k]}(\mathbf{x}) = \sum_{i=1}^{\lceil \zeta m \rceil} f_i(x_i) + f_{m+1-i}(x_{m+1-i}) + \frac{L_f}{4(k+1)} \left\langle a_{[k]}, x_i - x_{m+1-i} \right\rangle \tag{36}$$

which further yields

$$f^\star_{[k]} = \lceil \zeta m \rceil \frac{L_f}{8}(-1 + \frac{1}{k+1}) \quad \text{and} \quad f^\star_{[k,i]} = -\lceil \zeta m \rceil \frac{L_f}{8}\left(\frac{k^2}{(k+1)^2} + \frac{i}{(k+1)^2}\right).$$

Let each row of $\mathbf{x}^k$ belongs to $\mathbb{R}^{k,d}$. Then, since $\mathbf{x}^0 = \mathbf{0}$, we have

$$\begin{aligned} G(\mathbf{x}^k) &= f_{[2k+1]}(\mathbf{x}^k) - f^\star_{[2k+1]} \geq \min_{x_i \in \mathbb{R}^{k,d}} f_{[2k+1]}(\mathbf{x}) - f^\star_{[2k+1]} = f^\star_{[2k+1,k]} - f^\star_{[2k+1]} \\ &\geq \frac{\zeta m L_f}{8}\left(1 - \frac{1}{2(k+1)} - \frac{(2k+1)^2}{4(k+1)^2} - \frac{k}{4(k+1)^2}\right) \\ &= \frac{\zeta m L_f}{32(k+1)} = \Theta\left(\frac{L_f \left\|\mathbf{x}^0 - \mathbf{x}^\star\right\|^2}{(k+1)^2} + \frac{\left\|\mathbf{x}^0 - \mathbf{x}^\star\right\| \left\|\nabla f(\mathbf{x}^\star)\right\|}{k+1}\right). \end{aligned} \tag{37}$$

Similarly as the two-agent case, one can verify that $\left\|\mathbf{x}^0 - \mathbf{x}^\star\right\|^2 = \Theta\big(\frac{k+1}{L_f}\left\|\mathbf{x}^0 - \mathbf{x}^\star\right\| \left\|\nabla f(\mathbf{x}^\star)\right\|\big)$.

To have at least one non-zero element at the kth component among the local copies of agents in both of the above two subsets, one must perform at least k local computation steps and $(k-1)d_c$ communication steps. Thus, we have

$$k \leq \left\lfloor \frac{t-1}{1+d_c\tau_c} \right\rfloor + 1 \leq \frac{t}{1+d_c\tau_c} + 1. \tag{38}$$

Choosing $\zeta = \frac{1}{32}$, we have

$$\begin{aligned} d_c &= \lfloor (1-\zeta)m \rfloor + 1 - \lceil \zeta m \rceil \geq (1-2\zeta)m - 1 \\ &= \frac{15}{16}m - 1 \overset{(a)}{\geq} \frac{15}{16}\left(\sqrt{\frac{2}{\eta}} - 1\right) - 1 \overset{(b)}{\geq} \frac{1}{5\sqrt{\eta}}, \end{aligned}$$

where (a) is due to $\eta > \eta_{m+1} > \frac{2}{(m+1)^2}$ and (b) is due to $\eta \leq \eta_3 = \frac{1}{3}$. Further, since d_c is an integer, we have $d_c \geq \left\lceil \frac{1}{5\sqrt{\eta}} \right\rceil$. Combining (37) and (38) leads to

$$G(\mathbf{x}^{(t)}) \geq \Omega\left(\frac{L_f \left\|\mathbf{x}^0 - \mathbf{x}^\star\right\|^2}{(\frac{t}{1+\lceil \frac{1}{5\sqrt{\eta}} \rceil \tau_c} + 2)^2} + \frac{\left\|\mathbf{x}^0 - \mathbf{x}^\star\right\| \left\|\nabla f(\mathbf{x}^\star)\right\|}{\frac{t}{1+\lceil \frac{1}{5\sqrt{\eta}} \rceil \tau_c} + 2}\right). \tag{39}$$

We focus now on the case $n = 2$. Consider a complete graph of 3 agents with associated Laplacian matrix having eigengap equal to η. The agents' cost functions are

$$f_i = \begin{cases} \frac{L_f}{8} x_i^\top \mathbf{A}_{1,[k]} x_i - \frac{L_f}{4} e_1^\top \mathbf{x}_i & i = 1 \\ \frac{L_f}{8} x_i^\top \mathbf{A}_{2,[k]} x_i & i = 2 \\ 0 & i = 3 \end{cases}$$

Following similar steps as above, one can show that

$$G(\mathbf{x}^k) \geq \Omega\left(\frac{L_f \left\|\mathbf{x}^0 - \mathbf{x}^\star\right\|^2}{(k+1)^2} + \frac{\left\|\mathbf{x}^0 - \mathbf{x}^\star\right\| \left\|\nabla f(\mathbf{x}^\star)\right\|}{(k+1)}\right) \text{ with } k \leq \frac{t}{1+\tau_c} + 1 \text{ and } 1 \geq \left\lceil \frac{1}{5\sqrt{\eta}} \right\rceil,$$

which leads to the same expression of the lower bound as in (39). This concludes the proofs.

D Proof of Theorem 4

The proof follows the similar line of (Nesterov, 2013, Section 2.1.2). We consider the same set of local cost functions as depicted in (35), with the subscript $[k]$ of the $\mathbf{A}$ matrices replaced by $[2k+1]$. Then, it is not difficult to see that $\min_{x \in \mathbb{R}^d} F(x) = f^\star_{[2k+1]}$ and, for any $x \in \mathbb{R}^{k,d}$, we have

$$\begin{aligned} F(x) - f^\star_{[2k+1]} &= \min_{y \in \mathbb{R}^{k,d}} F(y) - f^\star_{[2k+1]} = f^\star_{[k]} - f^\star_{[2k+1]} \\ &= \lceil \zeta m \rceil \frac{L_f}{8}\left(-1 + \frac{1}{k+1} + 1 - \frac{1}{2k+1+1}\right) = \lceil \zeta m \rceil \frac{L_f}{16}\frac{1}{k+1} = \Theta\left(\frac{L_f\, m}{k+1}\right). \end{aligned} \tag{40}$$

For the cost functions as mentioned above, one can also verify that (cf. Appendix C.1.3)

$$\left\|\mathbf{x}^\star - \mathbf{x}^0\right\|^2 = \Theta\left(m(k+1)\right), \quad \|\nabla f(\mathbf{x}^\star)\| = \Theta\left(\frac{\sqrt{m}L_f}{\sqrt{k+1}}\right),$$

and thus $\frac{L_f\left\|\mathbf{x}^\star-\mathbf{x}^0\right\|^2}{k+1} = \Theta\left(\left\|\mathbf{x}^\star - \mathbf{x}^0\right\| \|\nabla f(\mathbf{x}^\star)\|\right)$. As a result, the RHS of (40) can be rewritten as:

$$\Theta\left(\frac{L_f\left\|\mathbf{x}^\star - \mathbf{x}^0\right\|^2}{(k+1)^2} + \frac{\left\|\mathbf{x}^\star - \mathbf{x}^0\right\| \|\nabla f(\mathbf{x}^\star)\|}{k+1}\right), \text{ or equivalently, } \quad \Theta\left(\frac{L_f\left\|\mathbf{x}^\star - \mathbf{x}^0\right\|^2}{(k+1)^2}\right),$$

which translate to the following lower bounds in terms of number of iterations, respectively:

$$\Omega\left(\sqrt{\frac{L_f\left\|\mathbf{x}^0 - \mathbf{x}^\star\right\|^2}{\epsilon}} + \frac{\left\|\mathbf{x}^0 - \mathbf{x}^\star\right\| \|\nabla f(\mathbf{x}^\star)\|}{\epsilon}\right) \quad \text{and} \quad \Omega\left(\sqrt{\frac{L_f\left\|\mathbf{x}^0 - \mathbf{x}^\star\right\|^2}{\epsilon}}\right).$$

The rest of proof follows by the same argument as in Section C.2 to relate k to the absolute time t as well as the eigengap η of the network.

E Proofs for the Upper Complexity Bounds

This section is devoted to the proofs of the upper complexity bounds of the proposed algorithms. We begin in Sec. E.1 establishing two fundamental inequalities that are valid for all feasible primal-dual solutions of Problem (3); see Lemma 12 and Lemma 13. Then, applying these inequalities to a saddle point solution of Problem (3), we obtain the convergence rate of Algorithms 1 and 2 in terms of the Bregman distance, see Sec. E.2and Sec. E.3 respectively. Finally in Sec. E.4, we apply the analysis of Chebyshev polynomials to show that the eigengap of the communication matrix $\mathbf{B}$, as a polynomial of the gossip matrix, can be upper bounded by a constant, leading to the upper complexity bound that matches the established lower bound.

E.1 Intermediate results

Lemma 12 (Fundamental Inequality I)**.** *Consider Algorithm* (15)*. We define* $\tau = \frac{1}{\nu T \lambda_m(\mathbf{B})}$*. Then we have*

$$\sigma_k = \frac{1}{\theta_{k+1}}, \quad \alpha_k = \frac{\theta_{k+1}}{\theta_k} - \theta_{k+1}, \quad \beta_k = \frac{\tau_{k+1}}{\tau_k}, \quad \tau_k = \frac{\tau}{\theta_k}.$$

Suppose Assumptions 1 and 3 hold. Then, for any $\mathbf{x} \in \mathbb{R}^{m\times d}$ *and* $\mathbf{y} \in \mathcal{C}^\perp$*, we have*

$$\begin{aligned}
&\Phi(\mathbf{u}^{k+1}, \mathbf{y}) - \Phi(\mathbf{A}\mathbf{x}, \mathbf{y}) + h(\mathbf{u}^{k+\frac{1}{2}}) - h(\mathbf{x}) \\
&\le -\left\langle \mathbf{y}^{k+1} - \mathbf{y}, \mathbf{u}^{k+1} - \mathbf{x}\right\rangle - \frac{1}{\gamma}\left\langle \theta_k(\mathbf{I} - \frac{\gamma\tau}{\theta_k^2}\mathbf{B})(\hat{\mathbf{x}}^{k+1} - \hat{\mathbf{x}}^k), \mathbf{u}^{k+1} - \mathbf{A}\mathbf{x}\right\rangle + \frac{L_f}{2}\left\|\mathbf{u}^{k+1} - \mathbf{x}^k\right\|^2,
\end{aligned}$$

where $h(\cdot) = \frac{1}{2\gamma}\|\cdot\|^2_{\mathbf{A}-\mathbf{A}^2}$*,* $\mathbf{u}^{k+\frac{1}{2}} = \mathbf{x}^k - \gamma(\nabla f(\mathbf{x}^k) + \hat{\mathbf{y}}^k)$ *and* $L_f = \max_i\{L_{f_i}\}$*.*

Proof. Since f is L_f-smooth by Assumption 1, we have

$$f(\mathbf{u}^{k+1}) \le f(\mathbf{x}^k) + \left\langle \nabla f(\mathbf{x}^k), \mathbf{u}^{k+1} - \mathbf{x}^k\right\rangle + \frac{L_f}{2}\left\|\mathbf{u}^{k+1} - \mathbf{x}^k\right\|^2,$$

and using $f(\mathbf{A}\mathbf{x}) \ge f(\mathbf{x}^k) + \left\langle \nabla f(\mathbf{x}^k), \mathbf{A}\mathbf{x} - \mathbf{x}^k\right\rangle$, further gives

$$f(\mathbf{u}^{k+1}) \le f(\mathbf{A}\mathbf{x}) + \left\langle \nabla f(\mathbf{x}^k), \mathbf{u}^{k+1} - \mathbf{A}\mathbf{x}\right\rangle + \frac{L_f}{2}\left\|\mathbf{u}^{k+1} - \mathbf{x}^k\right\|^2. \tag{41}$$

Also, subtracting $\mathbf{A}\mathbf{u}^{k+1}$ from both sides of (15a), multiplying (15d) by $\gamma\mathbf{A}$, and adding the obtained two equations while using (15e) lead to

$$\begin{aligned}
(\mathbf{I} - \mathbf{A})\mathbf{u}^{k+1} &= -\mathbf{A}\left(\mathbf{u}^{k+1} - \mathbf{x}^k + \gamma\left(\nabla f(\mathbf{x}^k) + \mathbf{y}^{k+1}\right)\right) - \gamma\mathbf{A}\mathbf{B}(\tau_{k-1}\beta_{k-1}\hat{\mathbf{x}}^k - \tau_k\hat{\mathbf{x}}^{k+1}) \\
&\overset{(*)}{=} -\mathbf{A}\left(\mathbf{u}^{k+1} - \mathbf{x}^k + \gamma\left(\nabla f(\mathbf{x}^k) + \mathbf{y}^{k+1}\right)\right) - \frac{\gamma\tau}{\theta_k}\mathbf{A}\mathbf{B}(\hat{\mathbf{x}}^k - \hat{\mathbf{x}}^{k+1}),
\end{aligned}$$

where in $(*)$ we used $\beta_{k-1}=\frac{\tau_k}{\tau_{k-1}}, \tau_k=\frac{\tau}{\theta_k}$. Notice that, for the above derivation, we implicitly assume that $k\geq 2$. However, with the definition of $\hat{\mathbf{x}}^1:=\mathbf{x}^1$ and the fact that $\hat{\mathbf{y}}^1=\tau_1\mathbf{B}\mathbf{x}^1$, we still have $(\mathbf{I}-\mathbf{A})\mathbf{u}^2=-\mathbf{A}\left(\mathbf{u}^2-\mathbf{x}^1+\gamma\left(\nabla f(\mathbf{x}^1)+\mathbf{y}^2\right)\right)-\frac{\gamma\tau}{\theta_1}\mathbf{A}\mathbf{B}(\hat{\mathbf{x}}^1-\hat{\mathbf{x}}^2)$.

Multiplying $\mathbf{u}^{k+\frac{1}{2}}-\mathbf{x}$ from both sides of the above equation and using the convexity of $h(\cdot)$ and the fact that $\mathbf{u}^{k+1}=\mathbf{A}\mathbf{u}^{k+\frac{1}{2}}$ we obtain

$$h(\mathbf{u}^{k+\frac{1}{2}})\leq h(\mathbf{x})-\frac{1}{\gamma}\left\langle \mathbf{u}^{k+1}-\mathbf{x}^k+\gamma\left(\nabla f(\mathbf{x}^k)+\mathbf{y}^{k+1}\right)-\frac{\gamma\tau}{\theta_k}\mathbf{B}(\hat{\mathbf{x}}^{k+1}-\hat{\mathbf{x}}^k),\mathbf{u}^{k+1}-\mathbf{A}\mathbf{x}\right\rangle. \tag{42}$$

Since $\sigma_k=\frac{1}{\theta_{k+1}}$ and $\alpha_k=\frac{\theta_{k+1}}{\theta_k}-\theta_{k+1}$, using (15b) and (15c) leads to

$$\theta_k\hat{\mathbf{x}}^{k+1}=\mathbf{u}^{k+1}-(1-\theta_k)\mathbf{u}^k \tag{43}$$

and

$$\begin{aligned}\hat{\mathbf{x}}^{k+1}-\hat{\mathbf{x}}^k&=\frac{1}{\theta_k}(\mathbf{u}^{k+1}-(1-\theta_k)\mathbf{u}^k)-\frac{1}{\theta_{k-1}}(\mathbf{u}^k-(1-\theta_{k-1})\mathbf{u}^{k-1})\\&=\frac{1}{\theta_k}\mathbf{u}^{k+1}-\frac{1}{\theta_k}\left((1-\theta_k)\mathbf{u}^k+\frac{\theta_k}{\theta_{k-1}}\mathbf{u}^k-\frac{\theta_k}{\theta_{k-1}}(1-\theta_{k-1})\mathbf{u}^{k-1}\right)\\&=\frac{1}{\theta_k}\mathbf{u}^{k+1}-\frac{1}{\theta_k}\left(\mathbf{u}^k+(\frac{\theta_k}{\theta_{k-1}}-\theta_k)(\mathbf{u}^k-\mathbf{u}^{k-1})\right)\\&\overset{(15b)}{=}\frac{1}{\theta_k}(\mathbf{u}^{k+1}-\mathbf{x}^k).\end{aligned}$$

We implicitly assumed $k\geq 2$; still we have $\hat{\mathbf{x}}^2-\hat{\mathbf{x}}^1=\frac{1}{\theta_k}\left(\mathbf{u}^2-\mathbf{x}^1\right)$, recalling that $\hat{\mathbf{x}}^1=\mathbf{x}^1$. Thus, (42) becomes

$$h(\mathbf{u}^{k+\frac{1}{2}})\leq h(\mathbf{x})-\frac{1}{\gamma}\left\langle \theta_k(\mathbf{I}-\frac{\gamma\tau}{\theta_k^2}\mathbf{B})(\hat{\mathbf{x}}^{k+1}-\hat{\mathbf{x}}^k)+\gamma\left(\nabla f(\mathbf{x}^k)+\mathbf{y}^{k+1}\right),\mathbf{u}^{k+1}-\mathbf{A}\mathbf{x}\right\rangle. \tag{44}$$

Combining (41) and (44) yields: for any $\mathbf{x}\in\mathbb{R}^{m\times d}$ and $\mathbf{y}\in\mathcal{C}^\perp$,

$$\begin{aligned}&f(\mathbf{u}^{k+1})+h(\mathbf{u}^{k+\frac{1}{2}})+\left\langle\mathbf{y},\mathbf{u}^{k+1}-\mathbf{A}\mathbf{x}\right\rangle-f(\mathbf{A}\mathbf{x})-h(\mathbf{x})\\&\leq\left\langle-(\mathbf{y}^{k+1}-\mathbf{y})-\frac{1}{\gamma}\theta_k(\mathbf{I}-\frac{\gamma\tau}{\theta_k^2}\mathbf{B})(\hat{\mathbf{x}}^{k+1}-\hat{\mathbf{x}}^k),\mathbf{u}^{k+1}-\mathbf{A}\mathbf{x}\right\rangle+\frac{L_f}{2}\left\|\mathbf{u}^{k+1}-\mathbf{x}^k\right\|^2,\end{aligned}$$

which, recalling that $\Phi(\mathbf{x},\mathbf{y})=f(\mathbf{x})+\langle\mathbf{y},\mathbf{x}\rangle$, completes the proof. □

Lemma 13 (Fundamental Inequality II)**.** *In the setting of Lemma 12, let* $\frac{1}{\theta_{k-1}^2}-\frac{1-\theta_k}{\theta_k^2}=0$*, that is,* $\frac{1}{\theta_k}=\frac{1+\sqrt{1+4(\frac{1}{\theta_{k-1}})^2}}{2}$*, with* $\theta_1=1$ *and* $(1-\gamma L_f)\mathbf{I}-\frac{\gamma\tau}{\theta_k^2}\mathbf{B}\succeq\mathbf{0}$*, for all* $1\leq k\leq T-1$*. Suppose Assumptions 1 and 3 hold. Then, for any* $\mathbf{x}\in\mathcal{C},\mathbf{y}\in\mathcal{C}^\perp$*, we have*

$$\Phi(\mathbf{u}^T,\mathbf{y})-\Phi(\mathbf{x},\mathbf{y})\leq\frac{1}{T^2}\left(\frac{2}{\gamma}\left\|\mathbf{u}^1-\mathbf{x}\right\|^2+\frac{2}{\tau\lambda_2(\mathbf{B})}\left\|\mathbf{y}\right\|^2\right).$$

where N *is the overall number of iterations.*

Proof. Applying Lemma 12 with $\mathbf{x}\in\mathcal{C}$, we have (note that $\mathbf{A}\mathbf{x}=\mathbf{x}$ by Assumption 3)

$$\begin{aligned}&\Phi(\mathbf{u}^{k+1},\mathbf{y})-\Phi(\mathbf{x},\mathbf{y})+h(\mathbf{u}^{k+\frac{1}{2}})-h(\mathbf{x})\\&\leq-\left\langle\mathbf{y}^{k+1}-\mathbf{y},\mathbf{u}^{k+1}-\mathbf{x}\right\rangle-\left\langle\frac{1}{\gamma}\theta_k(\mathbf{I}-\frac{\gamma\tau}{\theta_k^2}\mathbf{B})(\hat{\mathbf{x}}^{k+1}-\hat{\mathbf{x}}^k),\mathbf{u}^{k+1}-\mathbf{x}\right\rangle+\frac{L_f}{2}\left\|\mathbf{u}^{k+1}-\mathbf{x}^k\right\|^2.\end{aligned} \tag{45}$$

Likewise, with $\mathbf{x}=\mathbf{u}^{k-\frac{1}{2}}$ we have

$$\begin{aligned}&\Phi(\mathbf{u}^{k+1},\mathbf{y})-\Phi(\mathbf{u}^k,\mathbf{y})+h(\mathbf{u}^{k+\frac{1}{2}})-h(\mathbf{u}^{k-\frac{1}{2}})\\&\leq-\left\langle\mathbf{y}^{k+1}-\mathbf{y},\mathbf{u}^{k+1}-\mathbf{u}^k\right\rangle-\left\langle\frac{1}{\gamma}\theta_k(\mathbf{I}-\frac{\gamma\tau}{\theta_k^2}\mathbf{B})(\hat{\mathbf{x}}^{k+1}-\hat{\mathbf{x}}^k),\mathbf{u}^{k+1}-\mathbf{u}^k\right\rangle+\frac{L_f}{2}\left\|\mathbf{u}^{k+1}-\mathbf{x}^k\right\|^2.\end{aligned} \tag{46}$$

Let $V_k = \Phi(\mathbf{u}^k, \mathbf{y}) - \Phi(\mathbf{x}, \mathbf{y}) + h(\mathbf{u}^{k+\frac{1}{2}}) - h(\mathbf{x})$. Then, multiplying (46) by $1-\theta_k$ and (45) by θ_k, and combing the obtained equations yield

$$\begin{aligned}
&V_{k+1} - (1-\theta_k)V_k \\
&\le -\left\langle \mathbf{y}^{k+1} - \mathbf{y}, \mathbf{u}^{k+1} - \theta_k\mathbf{x} - (1-\theta_k)\mathbf{u}^k \right\rangle \\
&- \frac{1}{\gamma}\left\langle \theta_k(\mathbf{I} - \frac{\gamma\tau}{\theta_k^2}\mathbf{B})(\hat{\mathbf{x}}^{k+1} - \hat{\mathbf{x}}^k), \mathbf{u}^{k+1} - \theta_k\mathbf{x} - (1-\theta_k)\mathbf{u}^k \right\rangle + \frac{L_f}{2}\left\|\mathbf{u}^{k+1} - \mathbf{x}^k\right\|^2 \\
&\overset{(43)}{=} -\theta_k\left\langle \mathbf{y}^{k+1} - \mathbf{y}, \hat{\mathbf{x}}^{k+1} - \mathbf{x} \right\rangle - \frac{\theta_k^2}{\gamma}\left\langle \hat{\mathbf{x}}^{k+1} - \hat{\mathbf{x}}^k, \hat{\mathbf{x}}^{k+1} - \mathbf{x} \right\rangle_{\mathbf{I}-\frac{\gamma\tau}{\theta_k^2}\mathbf{B}} + \frac{\theta_k^2 L_f}{2}\left\|\hat{\mathbf{x}}^{k+1} - \hat{\mathbf{x}}^k\right\|^2 \\
&= -\frac{\theta_k^2}{\tau}\left\langle \mathbf{y}^{k+1} - \mathbf{y}, \mathbf{y}^{k+1} - \mathbf{y}^k \right\rangle_{(\mathbf{B}+\mathbf{J})^{-1}} - \frac{\theta_k^2}{\gamma}\left\langle \hat{\mathbf{x}}^{k+1} - \hat{\mathbf{x}}^k, \hat{\mathbf{x}}^{k+1} - \mathbf{x} \right\rangle_{\mathbf{I}-\frac{\gamma\tau}{\theta_k^2}\mathbf{B}} + \frac{\theta_k^2 L_f}{2}\left\|\hat{\mathbf{x}}^{k+1} - \hat{\mathbf{x}}^k\right\|^2,
\end{aligned} \tag{47}$$

where in the last equality we used $\mathbf{1}^\top\mathbf{y}^k = \mathbf{0}, \forall k \ge 1$ and the following result (recall $\mathbf{BJ} = \mathbf{JB} = \mathbf{0}$ and $\mathbf{y} \in \mathcal{C}^\perp$):

$$\begin{aligned}
&\left\langle \mathbf{y}^{k+1} - \mathbf{y}, \hat{\mathbf{x}}^{k+1} - \mathbf{x} \right\rangle \\
&\quad= \left\langle (\mathbf{B}+\mathbf{J})^{-1}(\mathbf{B}+\mathbf{J})(\mathbf{y}^{k+1} - \mathbf{y}), \hat{\mathbf{x}}^{k+1} - \mathbf{x} \right\rangle \\
&\quad= \left\langle \mathbf{B}(\mathbf{B}+\mathbf{J})^{-1}(\mathbf{y}^{k+1} - \mathbf{y}), \hat{\mathbf{x}}^{k+1} - \mathbf{x} \right\rangle \\
&\quad= \left\langle \mathbf{y}^{k+1} - \mathbf{y}, \mathbf{B}(\hat{\mathbf{x}}^{k+1} - \mathbf{x}) \right\rangle_{(\mathbf{B}+\mathbf{J})^{-1}} \\
&\quad\overset{(15d)}{=} \frac{\theta_k}{\tau}\left\langle \mathbf{y}^{k+1} - \mathbf{y}, \mathbf{y}^{k+1} - \mathbf{y}^k \right\rangle_{(\mathbf{B}+\mathbf{J})^{-1}}.
\end{aligned}$$

Dividing θ_k^2 from both sides of (47) leads to

$$\begin{aligned}
&\frac{V_{k+1}}{\theta_k^2} - \frac{1-\theta_k}{\theta_k^2}V_k \\
&\le -\frac{1}{\gamma}\left\langle \hat{\mathbf{x}}^{k+1} - \hat{\mathbf{x}}^k, \hat{\mathbf{x}}^{k+1} - \mathbf{x} \right\rangle_{\mathbf{I}-\frac{\gamma\tau}{\theta_k^2}\mathbf{B}} - \frac{1}{\tau}\left\langle \mathbf{y}^{k+1} - \mathbf{y}, \mathbf{y}^{k+1} - \mathbf{y}^k \right\rangle_{(\mathbf{B}+\mathbf{J})^{-1}} + \frac{L_f}{2}\left\|\hat{\mathbf{x}}^{k+1} - \hat{\mathbf{x}}^k\right\|^2 \\
&= -\frac{1}{2\gamma}\left(\left\|\hat{\mathbf{x}}^{k+1} - \hat{\mathbf{x}}^k\right\|^2_{(1-\gamma L_f)\mathbf{I}-\frac{\gamma\tau}{\theta_k^2}\mathbf{B}} + \left\|\hat{\mathbf{x}}^{k+1} - \mathbf{x}\right\|^2_{\mathbf{I}-\frac{\gamma\tau}{\theta_k^2}\mathbf{B}} - \left\|\hat{\mathbf{x}}^k - \mathbf{x}\right\|^2_{\mathbf{I}-\frac{\gamma\tau}{\theta_k^2}\mathbf{B}}\right) \\
&\quad - \frac{1}{2\tau}\left(\left\|\mathbf{y}^{k+1} - \mathbf{y}^k\right\|^2_{(\mathbf{B}+\mathbf{J})^{-1}} + \left\|\mathbf{y}^{k+1} - \mathbf{y}\right\|^2_{(\mathbf{B}+\mathbf{J})^{-1}} - \left\|\mathbf{y}^k - \mathbf{y}\right\|^2_{(\mathbf{B}+\mathbf{J})^{-1}}\right),
\end{aligned} \tag{48}$$

where in the last equality we used

$$\mathbf{2}\left\langle \mathbf{a}-\mathbf{c}, \mathbf{b}-\mathbf{c} \right\rangle_{\mathbf{G}} = \|\mathbf{a}-\mathbf{c}\|^{\mathbf{2}}_{\mathbf{G}} + \|\mathbf{b}-\mathbf{c}\|^{\mathbf{2}}_{\mathbf{G}} - \|\mathbf{a}-\mathbf{b}\|^{\mathbf{2}}_{\mathbf{G}},\ \forall \mathbf{a}, \mathbf{b} \in \mathbb{R}^{m\times d}.$$

Summing (48) over k from 1 to $T-1$ yields

$$\begin{aligned}
&\frac{V_T}{\theta_{T-1}^2} - \frac{1-\theta_1}{\theta_1^2}V_1 + \sum_{k=2}^{T-1}\left(\frac{1}{\theta_{k-1}^2} - \frac{1-\theta_k}{\theta_k^2}\right)V_k \\
&\le -\frac{1}{2\gamma}\sum_{k=1}^{T-1}\left\|\hat{\mathbf{x}}^{k+1} - \hat{\mathbf{x}}^k\right\|^2_{(1-\gamma L_f)\mathbf{I}-\frac{\gamma\tau}{\theta_k^2}\mathbf{B}} - \frac{1}{2\gamma}\sum_{k=2}^{T-1}\gamma\tau\left(\frac{1}{\theta_k^2} - \frac{1}{\theta_{k-1}^2}\right)\left\|\hat{\mathbf{x}}^k - \mathbf{x}\right\|^2_{\mathbf{B}} \\
&\quad - \frac{1}{2\gamma}\left(\left\|\hat{\mathbf{x}}^T - \mathbf{x}\right\|^2_{\mathbf{I}-\frac{\gamma\tau}{\theta_{T-1}^2}\mathbf{B}} - \left\|\hat{\mathbf{x}}^1 - \mathbf{x}\right\|^2_{\mathbf{I}-\frac{\gamma\tau}{\theta_1^2}\mathbf{B}}\right) \\
&\quad - \frac{1}{2\tau}\left(\sum_{k=1}^{T-1}\left\|\mathbf{y}^{k+1} - \mathbf{y}^k\right\|^2_{(\mathbf{B}+\mathbf{J})^{-1}} + \left\|\mathbf{y}^T - \mathbf{y}\right\|^2_{(\mathbf{B}+\mathbf{J})^{-1}} - \left\|\mathbf{y}^1 - \mathbf{y}\right\|^2_{(\mathbf{B}+\mathbf{J})^{-1}}\right).
\end{aligned} \tag{49}$$

Recalling that $\frac{1}{\theta_{k-1}^2} - \frac{1-\theta_k}{\theta_k^2} = 0$ and $\theta_1 = 1$, by induction it is easy to see that $k+1 > \frac{1}{\theta_k} \ge \frac{k+1}{2}$ and thus

$\frac{1}{\theta_k^2} - \frac{1}{\theta_{k-1}^2} = \frac{1}{\theta_k} > 0$. Then, with $\hat{\mathbf{x}}^1 = \mathbf{x}^1 := \mathbf{u}^1, \mathbf{y}^1 := \mathbf{0}$, (49) can be simplified as

$$\begin{aligned}\frac{T^2}{4}V_T + \sum_{i=1}^{T}\frac{1}{2\tau}\left\|\mathbf{y}^{k+1}-\mathbf{y}^k\right\|^2_{(\mathbf{B}+\mathbf{J})^{-1}} + \frac{1}{2\gamma}\sum_{k=1}^{T-1}\left\|\hat{\mathbf{x}}^{k+1}-\hat{\mathbf{x}}^k\right\|^2_{(1-\gamma L_f)\mathbf{I}-\frac{\gamma\tau}{\theta_k^2}\mathbf{B}} \\ \leq \frac{1}{2\gamma}\left\|\mathbf{u}^1-\mathbf{x}\right\|^2_{\mathbf{I}-\frac{\gamma\tau}{\theta_1^2}\mathbf{B}} + \frac{1}{2\tau}\left\|\mathbf{y}\right\|^2_{(\mathbf{B}+\mathbf{J})^{-1}}.\end{aligned} \tag{50}$$

Since $\rho\left((\mathbf{B}+\mathbf{J})^{-1}\right) = \frac{1}{\lambda_{\min}(\mathbf{B}+\mathbf{J})} = \frac{1}{\lambda_2(\mathbf{B})}$, $\mathbf{B}\succeq\mathbf{0}$ and $(1-\gamma L_f)\mathbf{I} - \frac{\gamma\tau}{\theta_k^2}\mathbf{B}\succeq\mathbf{0}$, we further have

$$\frac{T^2}{4}V_T \leq \frac{1}{2\gamma}\left\|\mathbf{u}^1-\mathbf{x}\right\|^2 + \frac{1}{2\tau}\frac{\|\mathbf{y}\|^2}{\lambda_2(\mathbf{B})},$$

which, together with the fact that $V_k \geq \Phi(\mathbf{u}^k,\mathbf{y}) - \Phi(\mathbf{x},\mathbf{y})$, completes the proof. □

E.2 Proof of Theorem 5

Note that the primal-dual method (12) is a special case of the update (15) with the setting $\theta_k \equiv 1, \alpha_k \equiv 0, \sigma_k \equiv 1, \tau_k \equiv \tau, \beta_k \equiv 1$. Furthermore, γ and τ defined in (13) satisfy $(1-\gamma L_f)\mathbf{I} - \gamma\tau\mathbf{B} \geq 0$ and $\mathbf{x}^k \equiv \mathbf{u}^k$. Invoking (49) with these parameter settings and $\mathbf{x} := \mathbf{x}^\star$, $\mathbf{y} := \mathbf{y}^\star = -\nabla f(\mathbf{x}^\star)$, $\hat{\mathbf{x}}^1 := \mathbf{x}^1, \mathbf{y}^1 := \mathbf{0}$, we have

$$\sum_{k=2}^{T}(\Phi(\mathbf{x}^k,\mathbf{y}^\star) - \Phi(\mathbf{x}^\star,\mathbf{y}^\star)) \leq \frac{1}{2\gamma}\left\|\mathbf{x}^1-\mathbf{x}^\star\right\|^2 + \frac{1}{2\tau}\frac{\|\mathbf{y}^\star\|^2}{\lambda_2(\mathbf{B})},$$

Let $\bar{\mathbf{x}}^T := \frac{1}{T-1}\sum_{k=2}^{T}\mathbf{x}_k$. Using the convexity of Φ, we furhter have

$$\begin{aligned}\Phi(\bar{\mathbf{x}}^T,\mathbf{y}^\star) - \Phi(\mathbf{x}^\star,\mathbf{y}^\star)) &\leq \frac{1}{T-1}\left(\frac{1}{2\gamma}\left\|\mathbf{x}^1-\mathbf{x}^\star\right\|^2 + \frac{1}{2\tau}\frac{\|\mathbf{y}^\star\|^2}{\lambda_2(\mathbf{B})}\right) \\ &\leq \frac{1}{T-1}\left(\frac{L_f}{2}\left\|\mathbf{x}^1-\mathbf{x}^\star\right\|^2 + \frac{1}{2\nu}\left\|\mathbf{x}^1-\mathbf{x}^\star\right\|^2 + \frac{\nu}{2\eta(\mathbf{B})}\left\|\mathbf{y}^\star\right\|^2\right).\end{aligned}$$

Setting $\nu = \frac{\sqrt{\eta(\mathbf{B})}\|\mathbf{x}^1-\mathbf{x}^\star\|}{\nabla f(\mathbf{x}^\star)}$ yields (14). Finally, it follows from (49) that $\hat{\mathbf{x}}^T = \mathbf{x}^T$ is bounded, for every $T \in \mathbb{N}_+$. The rest of proof is to show that $\mathbf{x}^T \to \mathbf{x}^\star$, which follows the standard cluster point analysis, as in the proof of (Chambolle and Pock, 2011, Th. 1) (refer also to (Chambolle and Pock, 2016, Remark 3)).

E.3 Proof of Theorem 6

Since $\gamma = \frac{\nu}{\nu L_f + T}, \tau = \frac{1}{\nu T\lambda_m(\mathbf{B})}$ and $\frac{1}{k+1} < \theta_k < \frac{2}{k+1}$, we have

$$(1-\gamma L_f)\mathbf{I} - \frac{\gamma\tau}{\theta_k^2}\mathbf{B} \succeq \mathbf{0}, \forall 1 \leq k \leq T-1. \tag{51}$$

Then, invoking Lemma 13 with $\mathbf{x} = \mathbf{x}^\star, \mathbf{y} = \mathbf{y}^\star = -\nabla f(\mathbf{x}^\star)$ and knowing that $\Phi(\mathbf{u}^k,\mathbf{y}^\star) - \Phi(\mathbf{x}^\star,\mathbf{y}^\star) = G(\mathbf{u}^k) \geq 0$ (cf., the relation (5) in the main text), we obtain

$$\begin{aligned}G(\mathbf{u}^T) &\leq \frac{\frac{2}{\gamma}R_x + \frac{2}{\tau}\frac{R_y}{\lambda_2(\mathbf{B})}}{T^2} = \frac{2(L_f + T/\nu)R_x + 2\nu T\lambda_m(\mathbf{B})\frac{R_y}{\lambda_2(\mathbf{B})}}{T^2} \\ &= \frac{2L_fR_x}{T^2} + \frac{\frac{2}{\nu}R_x + 2\nu\frac{R_y}{\eta(\mathbf{B})}}{T},\end{aligned}$$

where $R_x = \left\|\mathbf{u}^1-\mathbf{x}^\star\right\|^2, R_y = \|\nabla f(\mathbf{x}^\star)\|^2$. Setting $\nu = \sqrt{\eta(\mathbf{B})}$ we have

$$G(\mathbf{u}^T) \leq \frac{2L_f}{T^2}R_x + \frac{2}{\sqrt{\eta(\mathbf{B})}T}(R_x + R_y),$$

which, together with the time $(1+t_c)$ needed at each iteration, gives the overall time complexity.

If we set[3] $\nu = \sqrt{\frac{\eta(\mathbf{B})R_x}{R_y}}$, we have

$$G(\mathbf{u}^T, \mathbf{x}^\star) \leq \frac{2L_f R_x}{T^2} + \frac{4\sqrt{R_x R_y}}{\sqrt{\eta(\mathbf{B})}T},$$

which matches the lower bound also with respect to R_x, R_y.

In the following, we show that both consensus error and the absolute value of the objective error will converge at the same rate as the Bregman distance. Invoking Lemma 13 with $\mathbf{x} = \mathbf{x}^\star$, $\gamma = \frac{\nu}{\nu L_f + T}$, $\tau = \frac{1}{\nu T \lambda_m(\mathbf{B})}$ and $\nu = \sqrt{\eta(\mathbf{B})}$, we have

$$f(\mathbf{u}^T) - f(\mathbf{x}^\star) + \left\langle \mathbf{u}^T, \mathbf{y} \right\rangle = \Phi(\mathbf{u}^T, \mathbf{y}) - \Phi(\mathbf{x}^\star, \mathbf{y}) \leq \phi(\|\mathbf{y}\|)$$

where $\phi(\cdot) := \frac{2L_f R_x}{T^2} + \frac{\frac{2}{\sqrt{\eta}}(R_x + (\cdot)^2)}{T}$.

Now, setting $\mathbf{y} = 2\frac{\tilde{\mathbf{u}}^T}{\|\tilde{\mathbf{u}}^T\|}\|\mathbf{y}^\star\|$ where $\tilde{\mathbf{u}}^T = (\mathbf{I} - \frac{\mathbf{1}\mathbf{1}^T}{m})\mathbf{u}^T$, we have

$$f(\mathbf{u}^T) - f(\mathbf{x}^\star) + 2\|\mathbf{y}^\star\| \left\|\tilde{\mathbf{u}}^T\right\| \leq \phi(2\|\mathbf{y}^\star\|)$$

Also, since $f(\mathbf{u}^T) - f(\mathbf{x}^\star) + \left\langle \mathbf{u}^T, \mathbf{y}^\star \right\rangle \geq 0$, we have $f(\mathbf{u}^T) - f(\mathbf{x}^\star) \geq -\|\mathbf{y}^\star\| \left\|\tilde{\mathbf{u}}^T\right\|$. Thus, combining the above two inequalities yields

$$\left\|\tilde{\mathbf{u}}^T\right\| \leq \frac{\phi(2\|\mathbf{y}^\star\|)}{\|\mathbf{y}^\star\|} \text{ and } \left|f(\mathbf{u}^T) - f(\mathbf{x}^\star)\right| \leq \phi(2\|\mathbf{y}^\star\|).$$

□

E.4 Proof of Theorem 7

Following the similar lines in (Scaman et al., 2017, Theorem 4), we first consider the normalized Laplacian $\mathbf{L}$ has a spectrum in $[1 - c_1^{-1}, 1 + c_1^{-1}]$. According to Scaman et al. (2017); Wien (2011), the Chebyshev polynomail $P_K(x) = 1 - \frac{T_K(c_1(1-x))}{T_K(c_1)}$ is the solution of the following problem

$$\min_{p \in \mathbb{P}_K, p(0)=0} \max_{x \in [1-c_1^{-1}, 1+c_1^{-1}]} |p(x) - 1|.$$

As a result, we have

$$\max_{x \in [1-c_1^{-1}, 1+c_1^{-1}]} |P_K(x) - 1| \leq 2\frac{c_0^K}{1 + c_0^{2K}}. \tag{52}$$

Define $\delta = 2\frac{c_0^K}{1+c_0^{2K}}$. Since Algorithm 2 amounts to an instance of Procedure (15) with $\mathbf{A} = \mathbf{I} - c_2 \cdot P_K(\mathbf{L})$ and $\mathbf{B} = P_K(\mathbf{L})$, its convergence proof follows the same lines as that of Theorem 6 with the following properties of $P_K(\mathbf{L})$: i) $P_K(\mathbf{L})$ is symmetric; ii) according to (52), $\mathbf{0} \preceq \mathbf{I} - c_2 \cdot P_K(\mathbf{L}) \preceq \mathbf{I}$ and $P_K(\mathbf{L}) \succeq \mathbf{0}$, and $\mathbf{null}(P_K(\mathbf{L})) = \mathcal{C}$; iii) The values given for γ and τ in Algorithm 2 ensures that $(1 - \gamma L_f)\mathbf{I} - \frac{\gamma\tau}{\theta_k^2} P_K(\mathbf{L}) \succeq \mathbf{0}$, analogus to (51). Therefore we have

$$\begin{aligned} G(\mathbf{u}^T) &\leq \frac{\frac{2}{\gamma}R_x + \frac{2}{\tau}\frac{R_y}{\lambda_{\min}(P_K(\mathbf{L}) + \mathbf{J})}}{T^2} \leq \frac{2(L_f + T/\nu)R_x + 2\nu T(1+\delta)\frac{R_y}{1-\delta}}{T^2} \\ &= \frac{2L_f R_x}{T^2} + \frac{\frac{2}{\nu}R_x + 2\,\nu\, R_y \frac{1+\delta}{1-\delta}}{N} \overset{(*)}{=} \frac{2L_f R_x}{T^2} + \frac{4\sqrt{R_x R_y}}{N}\sqrt{\frac{1+\delta}{1-\delta}}, \end{aligned}$$

where (*) requires a specified ν. Finally, we have

$$\sqrt{\frac{1+\delta}{1-\delta}} = \left(1 + \left(\frac{1-\sqrt{\eta}}{1+\sqrt{\eta}}\right)^K\right) \Big/ \left(1 - \left(\frac{1-\sqrt{\eta}}{1+\sqrt{\eta}}\right)^K\right).$$

[3] Note that this requires accurate estimates on the ratio of R_x/R_y, which, indeed, plays a key role of trade-off parameter balancing gradient computation steps and communication steps.

Taking $K = \left\lceil \frac{1}{\sqrt{\eta}} \right\rceil$, we have

$$\left(\frac{1-\sqrt{\eta}}{1+\sqrt{\eta}}\right)^{\left\lceil \frac{1}{\sqrt{\eta}} \right\rceil} = \left(1 - \frac{2\sqrt{\eta}}{1+\sqrt{\eta}}\right)^{\left\lceil \frac{1}{\sqrt{\eta}} \right\rceil} \leq \left(1 - \frac{2}{1+\left\lceil \frac{1}{\sqrt{\eta}} \right\rceil}\right)^{\left\lceil \frac{1}{\sqrt{\eta}} \right\rceil}$$
$$\overset{(*)}{\leq} \left(1 - \frac{1}{\left\lceil \frac{1}{\sqrt{\eta}} \right\rceil}\right)^{\left\lceil \frac{1}{\sqrt{\eta}} \right\rceil} < e^{-1},$$

where (*) is due to the fact that $\left\lceil \frac{1}{\sqrt{\eta}} \right\rceil \geq 1$. Thus, we have $\sqrt{\frac{1+\delta}{1-\delta}} \leq \frac{1+e^{-1}}{1-e^{-1}} \leq 2.5$, which, together with the time $(1 + K\tau_c)$ needed at each iteration, gives the time complexity as announced. □

F Additional numerical results

F.1 Supplementary plots for the experiments in Section 5

This section provides additional numerical results, complementing those reported in the paper (cf. Section 5). In Figure 2 we plot the FEM-metric (11) versus the overall number of communications and computations performed by each agent (left panel), the number of communications (middle panel), and the number of computations (right panel). The comparison of the different schemes suggests to the same conclusions as in Sec.5; the only exception is that in the FEM-metric, the stochastic algorithm–DPSGD–does not present a significant advantage with respect to the non-accelerated algorithms–NEXT, DIGing and EXTRA.

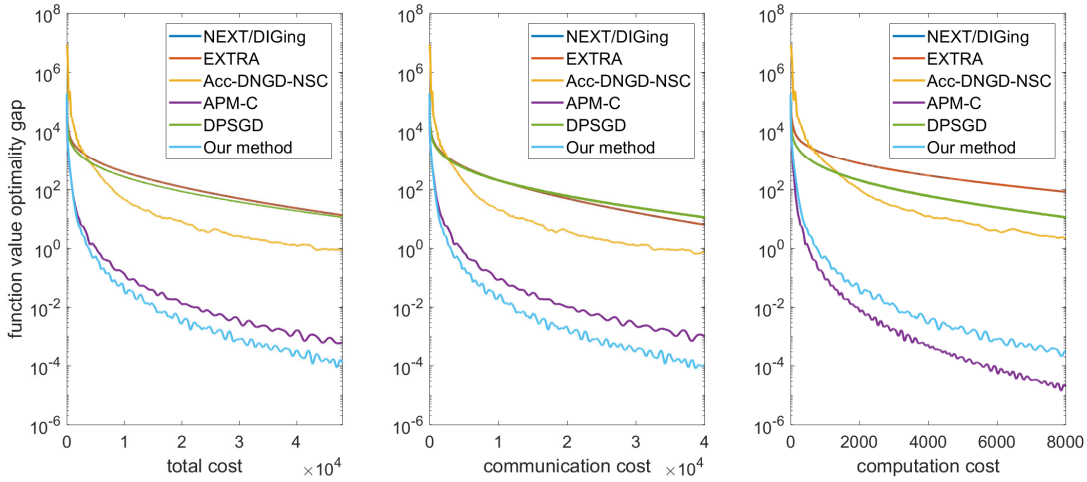

Figure 2: Comparison of distributed first-order gradient algorithms for solving the decentralized linear regression problem in terms of the traditional FEM-metric.

F.2 Decentralized logistic regression

To further verify the effectiveness of our proposed scheme, we also include a decentralized logistic regression task on the Parkinson's Disease Classification Data Set[4]. We preprocess the data by deleting the first column-id number, rescaling feature values to the range $(0, 1)$, and changing the label notation from $\{1, 0\}$ to $\{1, -1\}$. We denote the processed data set as $\{(u_i, y_i)\}_{i\in\mathcal{D}}$, where $u_i \in \mathbb{R}^d$ is the feature vector and $y_i \in \{1, -1\}$ is the label of the i-th observation. We simulated a network of 60 agents, generated by the Erdös-RéTyi model with the parameter of connection probability as 0.1. Then, we distributed the data set to all agents evenly, corresponding

[4]The data set is available at https://archive.ics.uci.edu/ml/datasets/Parkinson%27s+Disease+Classification

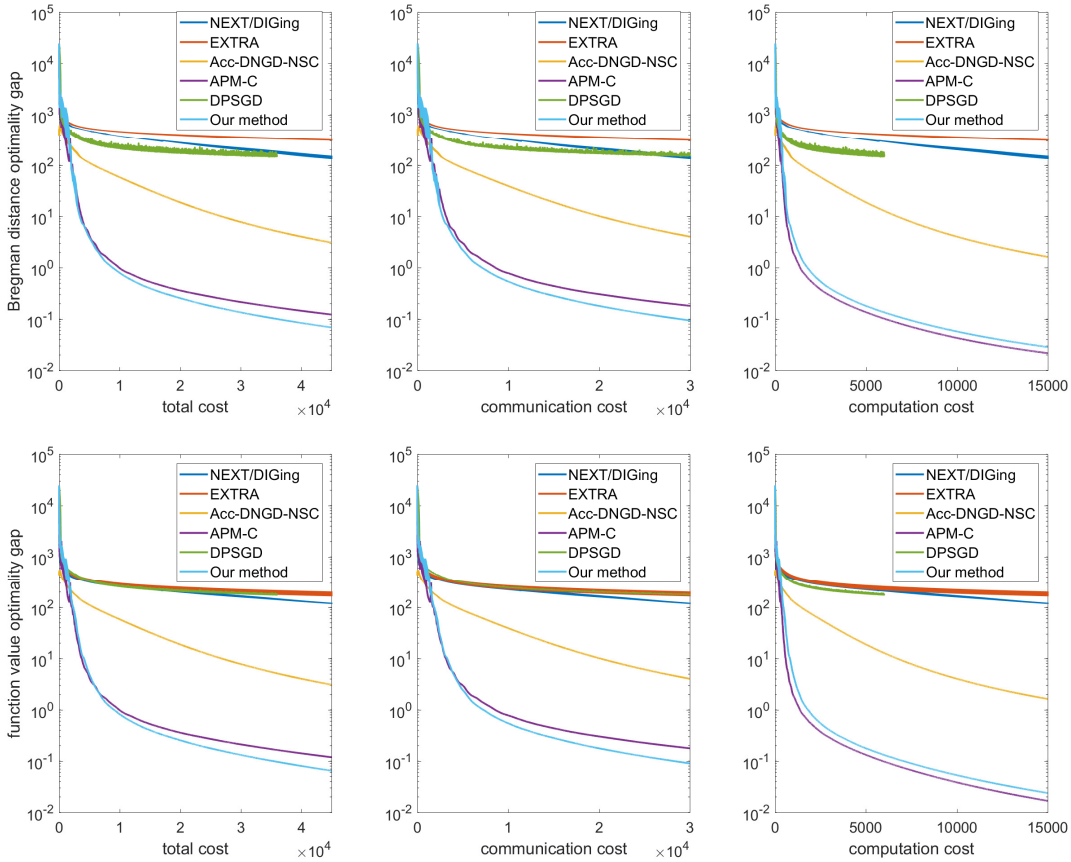

Figure 3: Comparison of distributed first-order gradient algorithms for solving the decentralized logistic regression problem in terms of both the Bregman distance and the traditional FEM-metric.

to a partition of the index set $\mathcal{D}$ across agents as $\mathcal{D} = \cup_{i=1}^{60} \mathcal{D}_i$. The decentralized logistic regression problem reads

$$\min_{x\in\mathbb{R}} \sum_{i=1}^{m} \sum_{j\in\mathcal{D}_i} \log\left(1+\exp(-y_j u_j^\top x)\right).$$

We estimated L_f for the problem as $L_f = 24$ and tuned the free parameters of the simulated algorithms manually to achieve the best practical performance for each algorithm. This leads to the following choices: **i)** the step size of NEXT/DIGing is set to 0.01; **ii)** the step size of EXTRA is set to 0.005; **iii)** for Acc-DNGD-NSC, we used the fixed step-size rule, with $\eta = 0.01/L_f$; **iv)** for APM-C, we set (see notation therein) $T_k = \lceil c \cdot (\log k/\sqrt{1-\sigma_2(\mathbf{W})})\rceil$, with $c = 0.2$ and $\beta_0 = 10^4$; **v)** for DPSGD, we set its step size as 0.001 and the portion of batch size to the full local data set as 20% and for **vi)** for our algorithm, we set $\nu = 1500$ and $K = 2$.

The experiment result is reported in Figure 3. The first row of panels shows the Bregman distance versus the total cost (left panel), the communication cost (middle panel), and the gradient computation cost (right panel). The second row plots the FEM-metric versus the same quantities as in the first row. Both the communication time unit and the computation time unit for a full epoch of local data is set as 1. For DPSGD, the computation time unit is scaled in proportion to the local batch size. The only existing algorithm that has a comparable performance with the proposed OPTRA is APM-C. As discussed in the task of decentralized linear regression, APM-C performs better than OPTRA in terms of the number of gradient computations, while suffers from high communication cost. In terms of the overall number of communications and computations, OPTRA outperforms all the other simulated schemes under the above setting.

F.3 Different ratio of communication time versus computation time

In all the previous experiments, we set both the communication time unit and the computation time unit for a full epoch of local data as 1. To incorporate scenarios where a full epoch computation of local gradient is much more expensive than one communication process, we re-conducted the previous experiments in the setting where the communication time unit is 1 while the computation time unit for a full epoch of local data is 5. Note that all the process of data generation and parameters tunings are the same as in the Sec. F.2. The results are reported in Figure 4 and Figure 5 respectively for decentralized linear regression problem and the decentralized logistic regression problem. It can be seen that OPTRA outperforms all the other simulated schemes in terms of the overall number of communications and computations, especially when the communication cost is not negligible.

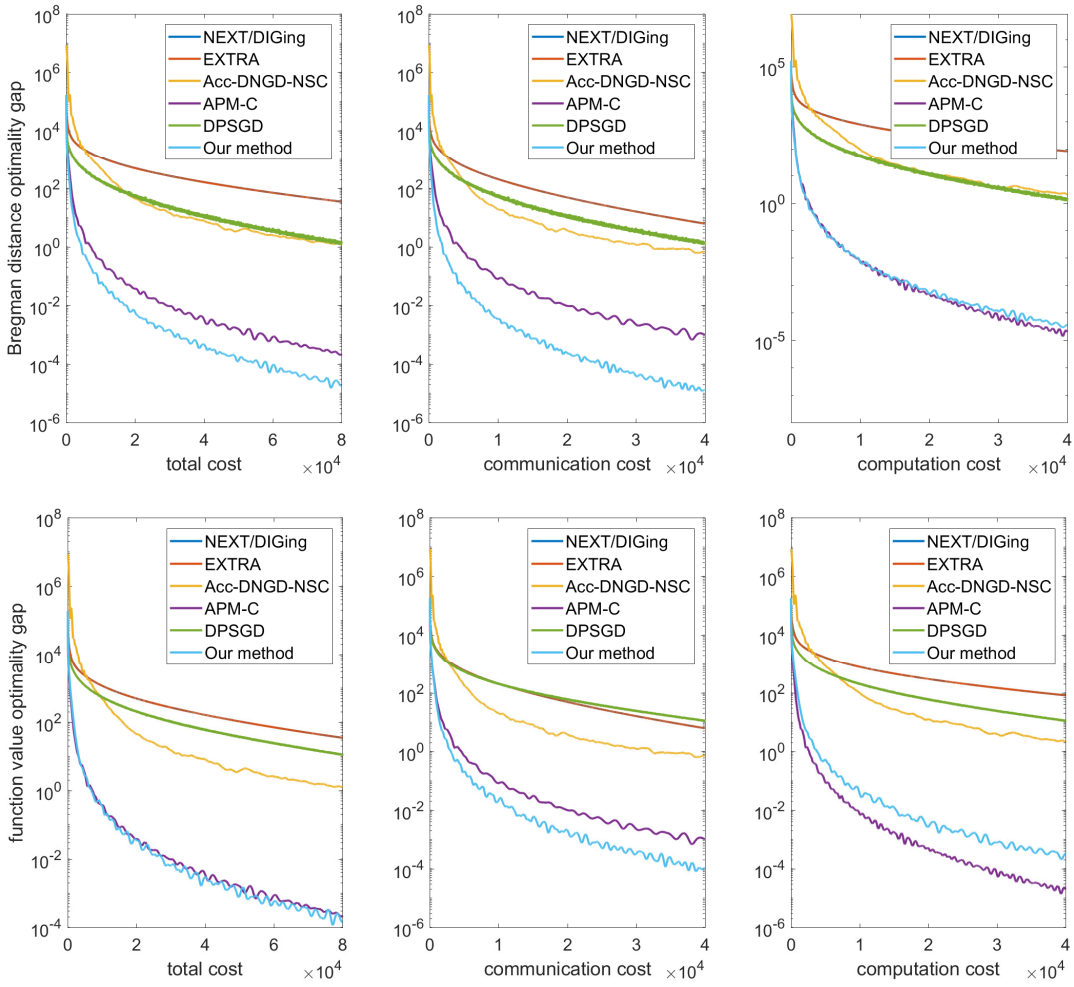

Figure 4: Comparison for distributed algorithms for solving the decentralized linear regression problem with the communication time unit being "1" and the computation time unit "5" for a full epoch of local data.

G Additional Comments

Scaman et al. (2017) presented optimal algorithms for decentralized optimization of strongly convex smooth functions. The functions considered in our paper are just convex (smooth). However, adding a small regularization (of the order of $\epsilon \|\mathbf{x}\|^2 / R^2$), the problem becomes strongly convex and the results of (Scaman et al., 2017) apply. In so doing, one can show that the method in (Scaman et al., 2017) achieves an ϵ-solution in $O\left(\left(1+\frac{1}{\sqrt{\eta}}\tau_c\right)\sqrt{\frac{L_f R^2}{\epsilon}}\log\left(\frac{1}{\epsilon}\right)\right)$. However, this requires an accurate estimate of R and the resulting rate differs from the lower bound for smooth convex functions by an extra log-factor "$\log 1/\epsilon$", meaning that this is not optimal as our scheme. Also, more importantly, i) Scaman et al. (2017) requires the computation in closed form

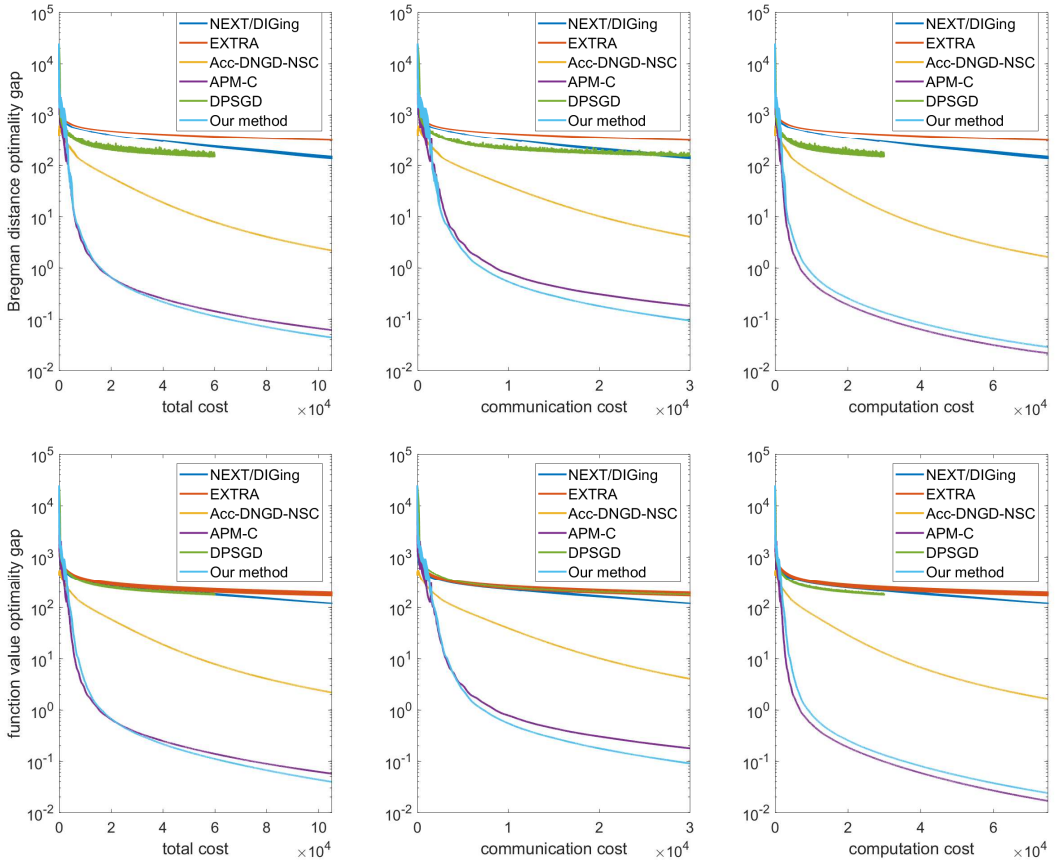

Figure 5: Comparison for distributed algorithms for solving the decentralized logistic regression problem, in the setting where the communication time unit is 1 while the computation time unit for a full epoch of local data is 5.

of the gradient of the Fenchel conjugate while our scheme does not have this limitation; ii) the Lipschitz constant of the gradient of the Fenchel conjugate in the setting above will scale as $1/\epsilon$ and thus the condition number of the dual problem is $O(1/\epsilon)$, which becomes arbitrarily large as ϵ decreases. As a consequence, (Scaman et al., 2017) significantly slow-downs in practice. This motivates our design of distributed algorithms specifically for convex (but not strongly convex) functions.