

# Accelerated Distributed Projected Gradient Descent for Convex Optimization with Clique-wise Coupled Constraints

Yuto Watanabe\* Kazunori Sakurama\*\*

\* Department of System Science, Graduate school of Informatics,  
Kyoto University (email: y-watanabe@sys.i.kyoto-u.ac.jp)

\*\* Department of System Science, Graduate school of Informatics,  
Kyoto University(email: sakurama@sys.i.kyoto-u.ac.jp)

---

**Abstract:** This paper addresses a distributed convex optimization problem with a class of coupled constraints, which arise in a multi-agent system composed of multiple communities modeled by *cliques*. First, we propose a fully distributed gradient-based algorithm with a novel operator inspired by the convex projection, called the clique-based projection. Next, we scrutinize the convergence properties for both diminishing and fixed step sizes. For diminishing ones, we show the convergence to an optimal solution under the assumptions of the smoothness of an objective function and the compactness of the constraint set. Additionally, when the objective function is strongly monotone, the strict convergence to the unique solution is proved without the assumption of compactness. For fixed step sizes, we prove the non-ergodic convergence rate of  $O(1/k)$  concerning the objective residual under the assumption of the smoothness of the objective function. Furthermore, we apply Nesterov’s acceleration method to the proposed algorithm and establish the convergence rate of  $O(1/k^2)$ . Numerical experiments illustrate the effectiveness of the proposed method.

*Keywords:* Distributed optimization, multi-agent system, convex optimization, coupled constraints, graph theory

---

## 1. INTRODUCTION

Due to the rapid improvement of information technology, devices with built-in computers are becoming more prevalent in various application domains (e.g., sensor networks, traffic networks, and power systems). As a result, systems are getting larger and larger by connecting these devices. In such large-scale systems, many challenges arise. For example, the central management of these systems leads to heavy computational burdens, and low-spec devices are incorporated into the network. To deal with these challenges, researchers in the control and machine learning communities have vigorously investigated distributed optimization, which aims to solve optimization problems by local communication without central management. Various papers have addressed the issue of the fast convergence and efficient computation of algorithms to reduce computational costs, communication frequency, and so on.

In recent years, distributed optimization problems with constraint-coupling have been studied in several papers (e.g., Falsone et al. (2020); Notarnicola and Notarstefano (2020); Chang (2016); Su et al. (2022); Wu et al. (2022)). As opposed to conventional methods like the dual decomposition in Terelius et al. (2011) and the ADMM in Boyd et al. (2011), the methods proposed in those papers do not require central management. Such a constraint-

coupled problem is a practical and general framework, which contains the widely studied problem in which the agents have to seek a common solution under agent-wise constraints, e.g., Nedic et al. (2010); Zhu and Martinez (2012); Yang et al. (2019), etc. However, the existing methods for constraint-coupling setups have slow convergence rates or do not explicitly show their convergence rate. Falsone et al. (2020) proposed an ADMM-based algorithm with the dynamic average consensus strategy but did not explicitly show the convergence rate. Similarly, Notarnicola and Notarstefano (2020) proposed a distributed algorithm based on a relaxed primal problem and the duality theory without offering the convergence rate. In Chang (2016); Wu et al. (2022), the convergence rate of  $O(1/k)$  was established in an ergodic sense. Just recently, in Su et al. (2022), the non-ergodic convergence rate of  $O(1/k)$  has been achieved for constraint-coupled optimization problems. To the authors’ knowledge, the convergence rate of existing methods for constraint-coupling setups has been at most  $O(1/k)$  so far. Given implementation to low-spec devices, it is meaningful to construct a faster and more efficient method.

This paper addresses a distributed convex optimization problem over a multi-agent network with a class of coupling in constraints called clique-wise coupled constraints. This class corresponds to a system consisting of multiple communities (with some overlap) modeled by cliques, i.e., complete subgraphs (see Bollobas (1998)), and each community has some constraints. This formulation con-

---

\* This work was partially supported by the joint project of Kyoto University and Toyota Motor Corporation, titled “Advanced Mathematical Science for Mobility Society”.

tains many conventional distributed optimization problems, e.g., the consensus and agent-wise constraints studied in Nedic et al. (2010), etc. First, we develop a clique-based projection operator, which is a novel extension of the convex projection. Second, by using this developed operator, we propose a distributed optimization algorithm, the *clique-based projected gradient descent* (CPGD), which is a generalization of the well-known projected gradient descent (PGD, see Calamai and Moré (1987)). Next, we prove its convergence to an optimal solution for diminishing and fixed step sizes. Moreover, using the Nesterov acceleration scheme (Nesterov (1983); Beck and Teboulle (2009)), we show that the proposed CPGD achieves the outstanding convergence rate of  $O(1/k^2)$ . Finally, we demonstrate the effectiveness of the proposed method through numerical experiments.

The major contributions of our proposed method, CPGD are threefold as follows. (i) The proposed CPGD does not require central management and is implementable only with peer-to-peer communication between agents. (ii) In the CPGD, by repeatedly operating the clique-based projection, we can generate a sequence arbitrarily close to a constraint set, which is challenging for the existing methods such as Falsone et al. (2020); Notarnicola and Notarstefano (2020), etc. Although there is a trade-off between the number of its operation and communication costs, the number can be tuned depending on the situation. (iii) The CPGD is fast and efficient for some practical constraints. The CPGD with the Nesterov acceleration achieves the non-ergodic convergence rate  $O(1/k^2)$ , which is faster than the existing distributed algorithms for constraint-coupled setups. Moreover, the computational cost in each iteration is very small for some typical constraints (e.g., linear and norm constraints).

The rest of this paper is organized as follows. Section 2 provides preliminaries. Section 3 presents the problem setting and show several examples that this paper considers. In Section 4, we extend the convex projection and propose a new distributed algorithm, the clique-based projected gradient descent (CPGD). In Section 5, we show the convergence properties of the CPGD. Section 6 presents the accelerated CPGD, which accomplishes the convergence rate of  $O(1/k^2)$ . In Section 7, numerical examples illustrate the effectiveness of our CPGD. Finally, Section 8 concludes this paper.

## 2. PRELIMINARIES

### 2.1 Notation

Let  $\mathbb{R}$  and  $\mathbb{N}$  be the set of real numbers and that of positive integers, respectively. Let  $|\cdot|$  be the number of elements in a countable finite set. The closure of a set is denoted by  $\text{cl}(\cdot)$ . For a mapping  $T : \mathbb{R}^m \rightarrow \mathbb{R}^m$ , define the fixed points set of  $T$  as  $\text{Fix}(T) = \{x \in \mathbb{R}^m : T(x) = x\}$ . Let  $I_d \in \mathbb{R}^{d \times d}$  denote the  $d \times d$  identity matrix. Let  $\mathbf{1}_d = [1, \dots, 1]^T \in \mathbb{R}^d$  denote the vector of  $d$  ones. With a positive definite and symmetric matrix  $Q \in \mathbb{R}^{m \times m}$ , we define the norm  $\|\cdot\|_Q$  as  $\|v\|_Q = \sqrt{v^T Q v}$  for a vector  $v \in \mathbb{R}^m$ . When  $Q = I_d$ , we simply write  $\|\cdot\|_{I_d}$  as  $\|\cdot\|$ .

For a vector  $v = [v_1^T, \dots, v_j^T, \dots, v_N^T]^T \in \mathbb{R}^{Nd}$  with vectors  $v_1, \dots, v_N \in \mathbb{R}^d$ ,  $[v]_j$  represents the operation to extract

the  $j$ th vector  $v_j$  from  $v$ , that is,

$$[v]_j = v_j \in \mathbb{R}^d.$$

For a vector  $x = [x_1^T, \dots, x_n^T]^T \in \mathbb{R}^{nd}$  with  $x_1, \dots, x_n \in \mathbb{R}^d$  and a subset  $\mathcal{C} = \{j_1, \dots, j_{|\mathcal{C}|}\} \subset \{1, \dots, n\}$ , let  $x_{\mathcal{C}}$  be

$$x_{\mathcal{C}} = [x_{j_1}^T, \dots, x_{j_{|\mathcal{C}|}}^T]^T \in \mathbb{R}^{|\mathcal{C}|d},$$

where  $\{j_1, \dots, j_{|\mathcal{C}|}\}$  is a strictly monotonically increasing sequence.

For  $x = [x_1^T, \dots, x_n^T]^T \in \mathbb{R}^{nd}$  and a differentiable function  $f : \mathbb{R}^{nd} \rightarrow \mathbb{R}$ , we write  $\nabla f(x) = [\nabla_1 f(x)^T, \dots, \nabla_n f(x)^T]^T \in \mathbb{R}^{nd}$  with  $\nabla = \partial/\partial x$  and  $\nabla_i = \partial/\partial x_i$ .

### 2.2 Graph Theory

In this subsection, we provide graph theoretic concepts. Consider a graph  $G = (\mathcal{N}, \mathcal{E})$  with a node set  $\mathcal{N} = \{1, \dots, n\}$  and an edge set  $\mathcal{E}$  consisting of pairs  $(i, j)$  of nodes  $i, j \in \mathcal{N}$ . If  $(i, j) \in \mathcal{E} \Leftrightarrow (j, i) \in \mathcal{E}$  holds for all  $(i, j) \in \mathcal{E}$ , the graph  $G$  is said to be *undirected*. In the following, we consider a time-invariant undirected graph  $G$ . For  $i \in \mathcal{N}$  and  $G$ , let  $\mathcal{N}_i \subset \mathcal{N}$  be the *neighbor set* of node  $i$  over  $G$ , defined as  $\mathcal{N}_i = \{j \in \mathcal{N} : (i, j) \in \mathcal{E}\} \cup \{i\}$ .

For an undirected graph  $G$ , we consider a set  $\mathcal{C} \subset \mathcal{N}$ . For  $\mathcal{C}$  and  $\mathcal{E}$ , let  $\mathcal{E}|_{\mathcal{C}}$  denote the subset of  $\mathcal{E}$  defined as  $\mathcal{E}|_{\mathcal{C}} = \{(i, j) \in \mathcal{E} : i, j \in \mathcal{C}\}$ . We call  $G|_{\mathcal{C}} = (\mathcal{C}, \mathcal{E}|_{\mathcal{C}})$  a subgraph induced by  $\mathcal{C}$ . If  $G|_{\mathcal{C}}$  is complete,  $\mathcal{C}$  is called a *clique* in  $G$ . If a clique  $\mathcal{C}$  is not contained by any other cliques,  $\mathcal{C}$  is said to be *maximal*. Let  $\text{clq}(G) = \{1, 2, \dots, q\}$  be a set of indices of maximal cliques in  $G$ . For  $i \in \mathcal{N}$ , we define  $\text{clq}_i(G)$  as an index set of the maximal cliques containing  $i$ , that is,  $\text{clq}_i(G) = \{k \in \text{clq}(G) : i \in \mathcal{C}_k\}$ . Note that, for each  $i \in \mathcal{N}$ ,  $\mathcal{N}_i$ , and  $\mathcal{C}_l$ ,  $l \in \text{clq}_i(G)$ , the following relationship holds (Sakurama and Sugie (2021)):

$$\mathcal{N}_i = \bigcup_{l \in \text{clq}_i(G)} \mathcal{C}_l. \quad (1)$$

## 3. PROBLEM STATEMENT

Consider a multi-agent system with  $n$  agents. Let  $\mathcal{N} = \{1, \dots, n\}$  be the set of agent indices. The communication network is expressed by a time-invariant undirected graph  $G = (\mathcal{N}, \mathcal{E})$  with an edge set  $\mathcal{E}$ , representing communication paths. For the graph  $G$  and  $l \in \text{clq}(G) = \{1, \dots, q\}$ ,  $\mathcal{C}_l$  denotes the  $l$ th maximal clique of  $G$ .

Our aim is to design a distributed algorithm that can solve the following distributed optimization problem only using local data and peer-to-peer communication:

$$\min_{x \in \mathbb{R}^{nd}} f(x) = \sum_{i=1}^n f_i(x_i) \quad (2a)$$

$$\text{s.t. } x \in \mathcal{D} = \bigcap_{l \in \text{clq}(G)} \{x \in \mathbb{R}^{nd} : x_{\mathcal{C}_l} \in \mathcal{D}_l\} \quad (2b)$$

with a convex objective function  $f$  in (2a) and a non-empty closed convex constraint set  $\mathcal{D}$  in (2b), where  $x = [x_1^T, \dots, x_n^T]^T \in \mathbb{R}^{nd}$ . Here, each  $\mathcal{D}_l \subset \mathbb{R}^{|\mathcal{C}_l|d}$  is non-empty and closed convex.

The set  $\mathcal{D}$  in (2b) can describe various coupled-constraints in accordance with  $G$ . The following example represents

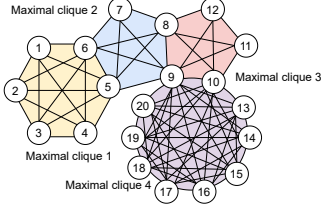


Fig. 1. Example of a network and its maximal cliques.

multiple communities forming a communication network with some constraints.

*Example 1.* Consider graph  $G$  in Fig. 1. This network  $G$  consists of four communities expressed by maximal cliques, and each of them has some overlapping nodes. Consider convex constraints imposed on each communities such as  $A_l x_{\mathcal{C}_l} = b_l$  and/or  $\|x_{\mathcal{C}_l}\| \leq r_l$  for  $l \in \text{clq}(G) = \{1, 2, 3, 4\}$  with some  $A_l \in \mathbb{R}^{m \times |\mathcal{C}_l|^d}$ ,  $b_l \in \mathbb{R}^m$ , and  $r_l > 0$ . Then, the intersection of these constraints can be written as (2b).

Note that conventional agent-wise and pair-wise constraints are included in the class of  $\mathcal{D}$  in (2b) as follows.

*Example 2.* Consider a connected graph  $G = (\mathcal{N}, \mathcal{E})$  and non-empty closed convex sets  $\mathcal{X}_{ij} \subset \mathbb{R}^{2d}$ ,  $(i, j) \in \mathcal{E}$ . Then,  $\mathcal{D} = \bigcap_{(i,j) \in \mathcal{E}} \{x \in \mathbb{R}^{nd} : [x_i^\top, x_j^\top]^\top \in \mathcal{X}_{ij}\}$  can be rewritten as (2b) because each edge belongs to some maximal clique of  $G$ . Also, the conventional constraints, studied in Nedic et al. (2010), etc, can be expressed by (2b) for  $\mathcal{D} = \bigcap_{(i,j) \in \mathcal{E}} \{x \in \mathbb{R}^{nd} : x_i = x_j\} \cap \bigcap_{i=1}^n \{x \in \mathbb{R}^{nd} : x_i \in \mathcal{X}_i\}$ .

Finally, we impose the following assumption on  $f$ .

*Assumption 1.* The function  $f$  is  $L$ -smooth, i.e.,  $\nabla f$  satisfies  $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$  with  $L > 0$  for any  $x, y \in \mathbb{R}^{nd}$ .

## 4. CLIQUE-BASED PROJECTED GRADIENT DESCENT

### 4.1 Algorithm description

We develop the clique-based projection and propose a novel distributed algorithm to solve the problem (2) using the projection. Let  $x_i(k) \in \mathbb{R}^d$  be the estimate  $x_i(k) \in \mathbb{R}^d$  of an solution at each iteration  $k$ , updated by agent  $i$ .

The clique-based projection is defined as follows.

*Definition 1.* For a non-empty closed convex set  $\mathcal{D} \subset \mathbb{R}^{nd}$  in (2b), a graph  $G$ , and its maximal cliques  $\mathcal{C}_l$ ,  $l \in \text{clq}(G)$ , the *clique-based projection*  $T : \mathbb{R}^{nd} \rightarrow \mathbb{R}^{nd}$  of  $x \in \mathbb{R}^{nd}$  onto  $\mathcal{D}$  is defined as

$$T(x) = [T_1(x_{\mathcal{N}_1})^\top, \dots, T_n(x_{\mathcal{N}_n})^\top]^\top \quad (3)$$

with

$$T_i(x_{\mathcal{N}_i}) = \frac{1}{|\text{clq}_i(G)|} \sum_{l \in \text{clq}_i(G)} [P_{\mathcal{D}_l}(x_{\mathcal{C}_l})]_{m_{l,i}} \quad (4)$$

for each  $i \in \mathcal{N}$ , where  $P_{\mathcal{D}_l} : \mathbb{R}^{|\mathcal{C}_l|^d} \rightarrow \mathcal{D}_l \subset \mathbb{R}^{|\mathcal{C}_l|^d}$  is the convex projection of  $x_{\mathcal{C}_l}$  onto  $\mathcal{D}_l$  with respect to the norm  $\|\cdot\|_{\text{diag}(\gamma_{\mathcal{C}_l})}$ , i.e.,

$$P_{\mathcal{D}_l}(x_{\mathcal{C}_l}) = \arg \min_{z \in \mathcal{D}_l} \|z - x_{\mathcal{C}_l}\|_{\text{diag}(\gamma_{\mathcal{C}_l})}. \quad (5)$$

---

### Algorithm 1 Clique-based projected gradient descent (CPGD) for agent $i$

---

**Require:**  $x_i(0) \in \mathbb{R}^d$ ,  $\{\lambda_k\}_{k \geq 1} \subset \mathbb{R}$ ,  $p \in \mathbb{N}$

- 1: **for**  $k = 0, 1, \dots$  **do**
- 2:    $x_i^{[1]}(k) \leftarrow x_i(k) - \lambda_{k+1} \nabla_i f_i(x_i(k))$ .
- 3:   **for**  $s = 1, 2, \dots, p$  **do**
- 4:     Gather  $x_j^{[s]}(k)$  from each  $j \in \mathcal{N}_i \setminus \{i\}$ .
- 5:      $x_i^{[s+1]}(k) \leftarrow T_i(x_{\mathcal{N}_i}^{[s]}(k))$  with  $T_i$  in (4).
- 6:   **end for**
- 7:    $x_i(k+1) \leftarrow x_i^{[p+1]}(k)$ .
- 8: **end for**

---

Here,  $\gamma = [1/|\text{clq}_1(G)|, \dots, 1/|\text{clq}_n(G)|]^\top \otimes \mathbf{1}_d \in \mathbb{R}^{nd}$ , and  $m_{l,i} \in \{1, \dots, |\mathcal{C}_l|\}$  denotes an order of  $i$  in  $\mathcal{C}_l$  for  $i \in \mathcal{N}$  and a clique  $\mathcal{C}_l$ ,  $l \in \text{clq}_i(G)$ , i.e.,  $\mathcal{C}_l = \{\dots, i, \dots\}_{m_{l,i}}$ .

Using the clique-based projection, we present a novel distributed algorithm for the problem (2), the *clique-based projected gradient descent* (CPGD), in Algorithm 1. Here,  $x_{\mathcal{N}_i}^{[s]}(k)$  is the aggregated vector of  $x_j^{[s]}(k)$  according to  $j \in \mathcal{N}_i$ . The step size  $\{\lambda_k\}$  is a sequence of positive numbers and  $p$  is a positive integer. To guarantee the convergence of the CPGD, we must impose some conditions on  $\{\lambda_k\}$ . Section 5 shows convergence properties of Algorithm 1.

Note that from (1), each agent  $i$  can implement the CPGD in Algorithm 1 only with local communication with neighbors  $\mathcal{N}_i$ .

By aggregating Algorithm 1 for all  $i \in \mathcal{N}$ , we obtain

$$x^{[1]}(k) = x(k) - \lambda_{k+1} \nabla f(x(k)) \quad (6a)$$

$$x(k+1) = T^p(x^{[1]}(k)), \quad (6b)$$

where  $T^p = \overbrace{T \circ T \circ \dots \circ T}^p$ .

### 4.2 Discussion on the clique-based projection

Here, we view the clique-based projection in Definition 1 from the perspective of the convex projection and the proximal operator.

First, we prove the following proposition, which indicates that the clique-based projection  $T$  can be obtained by taking the gradient of a function  $V$ .

*Proposition 1.* Let  $V : \mathbb{R}^{nd} \rightarrow \mathbb{R}$  be

$$V(x) = \frac{1}{2} \sum_{l \in \text{clq}(G)} \|x_{\mathcal{C}_l} - P_{\mathcal{D}_l}(x_{\mathcal{C}_l})\|_{\text{diag}(\gamma_{\mathcal{C}_l})}^2 \quad (7)$$

with  $P_{\mathcal{D}_l}$  in (5) and  $\gamma = [1/|\text{clq}_1(G)|, \dots, 1/|\text{clq}_n(G)|]^\top \otimes \mathbf{1}_d \in \mathbb{R}^{nd}$ . Then, the following holds for any  $x \in \mathbb{R}^{nd}$ :

$$T(x) = x - \nabla V(x). \quad (8)$$

**Proof.** Since each  $\mathcal{D}_l$  is closed and convex,  $1/2 \|x_{\mathcal{C}_l} - P_{\mathcal{D}_l}(x_{\mathcal{C}_l})\|_{\text{diag}(\gamma_{\mathcal{C}_l})}^2$  is differentiable and thus  $V(x)$  in (7) is also differentiable. Then, for all  $i \in \mathcal{N}$ , we have  $\nabla_i V(x) = \sum_{l \in \text{clq}_i(G)} \frac{1}{|\text{clq}_i(G)|} (x_i - [P_{\mathcal{D}_l}(x_{\mathcal{C}_l})]_{m_{l,i}}) = x_i - \sum_{l \in \text{clq}_i(G)} \frac{1}{|\text{clq}_i(G)|} [P_{\mathcal{D}_l}(x_{\mathcal{C}_l})]_{m_{l,i}} = x_i - T_i(x)$  from (1) and (4). Hence, we obtain (8).  $\square$

Through Proposition 1, we can interpret the proposed method in (3) as a variant of the proximal gradient method (Beck and Teboulle (2009)). From (8), the clique-based projection  $T$  satisfies

$$T(x) = \arg \min_{y \in \mathbb{R}^{nd}} \frac{1}{2} \|x - y\|^2 + V(x) + \nabla V(x)^\top (y - x).$$

Hence, the clique-based projection  $T$  can be regarded as the proximal operator  $\text{prox}_\psi(x) = \arg \min_{y \in \mathbb{R}^{nd}} 1/2 \|x - y\|^2 + \psi(y)$  for  $\psi(y) = V(x) + \nabla V(x)^\top (y - x)$ , which is the first order approximation of  $V(y)$  at  $x$ .

Moreover, if  $G$  is complete, the CPGD in Algorithm 1 equals to the well-known *projected gradient descent* (PGD) (see Calamai and Moré (1987)):

$$x(k+1) = P_{\mathcal{D}}(x(k) - \lambda_{k+1} \nabla f(x(k))) \quad (9)$$

with  $P_{\mathcal{D}}(x) = \arg \min_{z \in \mathcal{D}} \|x - z\|$  because  $\text{clq}(G) = \{1\}$  and  $\mathcal{C}_1 = \mathcal{N}$  hold for complete  $G$ . Note that the PGD in (9) is centralized due to the operation  $P_{\mathcal{D}}(\cdot)$ .

## 5. CONVERGENCE ANALYSIS

For the CPGD in Algorithm 1, we present convergence theorems for both diminishing and fixed step sizes.

Before proceeding to analyze convergence properties, we provide the following lemma, which shows key features of the clique-based projection  $T$  in Definition 1.

*Lemma 1.* For the clique-based projection  $T$  in Definition 1 and the closed convex set  $\mathcal{D}$  in (2b), the following statements hold:

- The mapping  $T$  is nonexpansive, i.e.,  $\|T(x) - T(y)\| \leq \|x - y\|$  holds for any  $x, y \in \mathbb{R}^{nd}$ .
- The fixed points set of  $T$  satisfies  $\text{Fix}(T) = \mathcal{D}$ .
- For any  $x \in \mathbb{R}^{nd} \setminus \mathcal{D}$  and any  $z \in \mathcal{D}$ ,  $\|T(x) - z\| < \|x - z\|$  holds.
- For any  $x \in \mathbb{R}^{nd}$ ,  $T^\infty(x) = \lim_{p \rightarrow \infty} T^p(x) \in \mathcal{D}$  holds.

**Proof.** See Appendix A.  $\square$

By Lemma 1a-b, the clique-based projection  $T$  preserves important features of the convex projection for  $\mathcal{D}$ , that is, the nonexpansiveness and fixed point set. Moreover, from Lemma 1c-d, by repeatedly operating  $T$ ,  $x$  gradually converges to  $\mathcal{D}$ .

With this in mind, we show the first main result as follows.

*Theorem 1.* Assume that a convex objective function  $f$  in (2a) satisfies Assumption 1, and that  $\mathcal{D} \subset \mathbb{R}^{nd}$  in (2b) is a non-empty closed convex set. Consider the sequence  $\{x(k)\}$  generated by CPGD in Algorithm 1 (or (6)).

- Let a sequence  $\{\lambda_k\}$  of positive integers satisfy  $\lim_{k \rightarrow \infty} \lambda_k = 0$ ,  $\sum_{k=1}^{\infty} \lambda_k = \infty$ , and  $\sum_{k=1}^{\infty} \lambda_k^2 < \infty$ .<sup>1</sup> Assume that  $\mathcal{D}$  in (2b) is bounded. Then, for any initial point  $x(0) = x_0 \in \mathbb{R}^{nd}$  and any  $p \in \mathbb{N}$ ,  $\{x(k)\}$  converges to an optimal solution  $x_* \in \arg \min_{x \in \mathcal{D}} f(x)$ .
- Let a sequence  $\{\lambda_k\}$  of positive integers satisfy  $\lim_{k \rightarrow \infty} \lambda_k = 0$ ,  $\sum_{k=1}^{\infty} \lambda_k = \infty$ , and  $\sum_{k=1}^{\infty} |\lambda_k - \lambda_{k+1}| < \infty$ .<sup>2</sup> Additionally assume that the gradient  $\nabla f : \mathbb{R}^{nd} \rightarrow \mathbb{R}^{nd}$  of  $f$  is strongly monotone, i.e., there

<sup>1</sup> For example,  $\lambda_k = 1/k$  satisfies the conditions.

<sup>2</sup> For example,  $\lambda_k = 1/k$  and  $\lambda_k = 1/\sqrt{k}$  satisfy the conditions.

exists some  $\mu > 0$  such that  $(\nabla f(x) - \nabla f(y))^\top (x - y) \geq \mu \|x - y\|^2$  is satisfied for any  $x, y \in \mathbb{R}^{nd}$ . Then  $\{x(k)\}$  converges to the unique optimal solution  $x_* = \arg \min_{z \in \mathcal{D}} f(z)$  for any initial point  $x(0) = x_0 \in \mathbb{R}^{nd}$  and any  $p \in \mathbb{N}$ .

- Let  $\lambda_k = t \in (0, 1/L]$  for any  $k \in \mathbb{N}$ . Let  $J : \mathbb{R}^{nd} \rightarrow \mathbb{R}$  be

$$J(x) = f(x) + V(x)/t \quad (10)$$

with  $V$  in (7). Then, for any initial point  $x(0) = x_0 \in \mathbb{R}^{nd}$  and  $p = 1$ ,

$$J(x(k)) - J(x_*) \leq \frac{\|x_0 - x_*\|^2}{2tk} \quad (11)$$

holds with  $x_* \in \arg \min_{x \in \mathcal{D}} f(x)$ .

**Proof.** a) From Lemma 1a-b, the CPGD in (6) can be regarded as the hybrid steepest descent in Yamada (2001); Yamada et al. (2002) for any  $p \in \mathbb{N}$ . Hence, Theorem 1a follows from Theorem 2.18, Remark 2.17 in Yamada et al. (2002), and Lemma 1c. b) The statement follows from Theorem 2.15 in Yamada et al. (2002) and Lemma 1a-b. c) See Appendix B.1 and B.2.  $\square$

Theorem 1a-b imply that the CPGD with a diminishing step size provides an optimal solution regardless of the number  $p$  of operations of  $T$  in each iteration. Thus, with a sufficiently large  $p$ , the CPGD can generate points arbitrarily close to  $\mathcal{D}$  by Lemma 1c-d and thus can work just like the conventional PGD in (9), which is centralized. Note that the assumption of the boundedness of  $\mathcal{D}$  in Theorem 1a is not restrictive in practice, and the additional assumption of the strong monotonicity of  $\nabla f$  in Theorem 1b is satisfied for strongly convex  $f$ .

Theorem 1c guarantees the non-ergodic convergence rate of  $O(1/k)$  for smooth  $f$  when  $\lambda_k$  is fixed and  $p = 1$ . Practically, the CPGD is expected to perform well even if  $p > 1$ , which is shown in the numerical results in Section 7.

## 6. NESTEROV'S ACCELERATED CPGD

In this section, to enhance the convergence rate, we modify the proposed CPGD with Nesterov's acceleration scheme (Nesterov (1983); Beck and Teboulle (2009)). The modified method is called the accelerated clique-based projected gradient descent (ACPGD), given as follows:

$$x(k+1) = T^p(\hat{x}(k) - \lambda_{k+1} \nabla f(\hat{x}(k))) \quad (12a)$$

$$\hat{x}(k+1) = x(k+1) + \frac{\sigma_k - 1}{\sigma_{k+1}} (x(k+1) - x(k)), \quad (12b)$$

where  $\hat{x}(0) = x(0)$  and  $\sigma_{k+1} = (1 + \sqrt{1 + 4\sigma_k^2})/2$ ,  $\sigma_0 = 1$ . Agent  $i \in \mathcal{N}$  can run the ACPGD in a distributed fashion by updating  $x_i(k+1)$  as  $\hat{x}_i(k+1) = x_i(k+1) + \frac{\sigma_k - 1}{\sigma_{k+1}} (x_i(k+1) - x_i(k))$  with  $x_i(k)$  in addition to Algorithm 1.

The ACPGD achieves the convergence rate of  $O(1/k^2)$  as follows.

*Theorem 2.* Assume that a convex function  $f$  in (2a) satisfies Assumption 1, and that  $\mathcal{D} \subset \mathbb{R}^{nd}$  in (2b) is a non-empty closed convex set. Let  $p = 1$  and  $\lambda_k = t \in (0, 1/L]$  for all  $k \in \mathbb{N}$ . Consider the sequence  $\{x(k)\}$  generated by the ACPGD in (12). Then, for any initial state  $x(0) = \hat{x}(0) = x_0 \in \mathbb{R}^{nd}$ , the following inequality holds:

$$J(x(k)) - J(x_*) \leq \frac{2\|x_0 - x_*\|^2}{tk^2}, \quad (13)$$

where  $x_* \in \arg \min_{x \in \mathcal{D}} f(x)$  and  $J(x)$  is given as (10).

**Proof.** See Appendix B.1 and B.3.  $\square$

Although the case of  $p = 1$  is only proved in Theorem 2 like Theorem 1c, the ACPGD for  $p > 1$  can also perform well as shown in numerical results.

## 7. NUMERICAL EXPERIMENT

We demonstrate the effectiveness of the proposed method through numerical experiments.

Consider a multi-agent system with  $n = 20$  agents. The communication network  $G$  is given as Fig. 1. Then we have  $\text{clq}(G) = \{1, 2, 3, 4\}$  and  $\mathcal{C}_1 = \{1, 2, \dots, 6\}$ ,  $\mathcal{C}_2 = \{5, 6, \dots, 9\}$ ,  $\mathcal{C}_3 = \{8, 9, \dots, 12\}$ ,  $\mathcal{C}_4 = \{9, 10, 13, 14, \dots, 20\}$ . We consider the following allocation problem:

$$\min_{x \in \mathbb{R}^{20}} \frac{1}{2} \sum_{i=1}^{20} (x_i - a_i)^2 \quad (14a)$$

$$\text{s.t.} \quad \sum_{j \in \mathcal{C}_l} x_j = N_l, \quad \forall l \in \text{clq}(G) = \{1, \dots, 4\}, \quad (14b)$$

where  $a_1, \dots, a_{20} \in \mathbb{R}$  are randomly generated by the uniform distribution for the interval  $[0, 10]$  and  $N_l, l \in \text{clq}(G)$  are given as  $[N_1, \dots, N_4] = [7, 3, 5, 10]$ . Letting  $\mathcal{D}_l = \{y = [y_1, \dots, y_{|\mathcal{C}_l|}]^\top \in \mathbb{R}^{|\mathcal{C}_l|} : \sum_{j=1}^{|\mathcal{C}_l|} y_j = N_l\}$  for  $l \in \text{clq}(G)$ , we can apply Algorithm 1 to the problem (14).

We conduct simulations of the CPGD in Algorithm 1 with  $\lambda_k = 1/k$  and  $\lambda_k = 0.001$  for  $p = 1, 10, 50$ . Additionally, we conduct simulations of the ACPGD in (12) with  $\lambda_k = 0.001$  for  $p = 1, 10, 50$ . For comparison, we run the conventional PGD in (9), which is centralized.

Figs. 2a-2c plot the evolution of the relative optimality gap  $|f(x(k)) - f^*|/f^*$  between the value of the objective function  $f(x(k))$  and its optimal value  $f^*$  under  $p = 1, 10, 50$ , respectively. Besides, Fig. 2d shows the result of the conventional PGD. From Figs. 2a-2c, the CPGD with both types of step sizes and the ACPGD succeed in the swift convergence of  $x(k)$  to a point close to the optimal solution. For a larger  $p$ , a more accurate solution is obtained. In addition, in the early stage of the simulation, the ACPGD achieves the best performance for all  $p$ , especially the lower  $p$ . After about 1000 iterations, the diminishing step size  $\lambda_k = 1/k$  gives the closest solution to the optimal one. Furthermore, from Fig. 2c and Fig. 2d, the CPGD and the ACPGD run as fast as the PGD when  $p = 50$ . These results illustrate the effectiveness of the proposed method.

## 8. CONCLUSION

This paper addressed a distributed convex optimization problem with clique-wise couplings in constraints. First, we developed the clique-based projection operator and proposed a new distributed algorithm with the operator, the clique-based projected gradient descent (CPGD). Next, we proved its convergence properties for diminishing and fixed step sizes. Moreover, we presented the accelerated version of the CPGD, which achieved the convergence

rate of  $O(1/k^2)$ . Finally, numerical experiments illustrated the effectiveness of the proposed method. Our future directions are to consider more general coupled constraints and to apply the CPGD to some applications, such as traffic networks.

## Appendix A. PROOF OF LEMMA 1

### A.1 Preliminaries

As a preliminary, we provide several useful inequalities and some propositions.

First, for a convex function  $h : \mathbb{R}^m \rightarrow \mathbb{R}$ , any positive integer  $k \in \mathbb{N}$ , any  $x_1, \dots, x_k \in \mathbb{R}^m$ , and any  $\alpha_j \geq 0$  ( $j = 1, \dots, k$ ) satisfying  $\sum_{j=1}^k \alpha_j = 1$ , the following inequality holds, which is called *Jensen's inequality*:

$$h(\sum_{j=1}^k \alpha_j x_j) \leq \sum_{j=1}^k \alpha_j h(x_j). \quad (A.1)$$

Note that  $h(\sum_{j=1}^k \alpha_j x_j) = \sum_{j=1}^k \alpha_j h(x_j)$  holds if and only if  $x_1 = \dots = x_k$  holds or  $h$  is affine (see Peressini et al. (1988)).

Second, we give some properties of the convex projection without proof. For details, see textbooks of function analysis, e.g., Kreyszig (1991). For any  $x, y \in \mathbb{R}^m$ , a norm  $\|\cdot\|_Q$ , a non-empty closed convex set  $\mathcal{M} \subset \mathbb{R}^m$ , and any  $z \in \mathcal{M}$ , the convex projection  $P_{\mathcal{M}} : \mathbb{R}^m \rightarrow \mathcal{M}$  with respect to  $\|\cdot\|_Q$ , i.e.,  $P_{\mathcal{M}}(x) = \arg \min_{z \in \mathcal{M}} \|x - z\|_Q$ , satisfies the following inequalities:

$$\|P_{\mathcal{M}}(x) - P_{\mathcal{M}}(y)\|_Q \leq \|x - y\|_Q \quad (A.2)$$

$$\|P_{\mathcal{M}}(x) - z\|_Q \leq \|x - z\|_Q, \quad (A.3)$$

implying the nonexpansiveness and quasi-nonexpansiveness of  $P_{\mathcal{M}}(\cdot)$ , respectively. Moreover, for any  $x, y \in \mathbb{R}^m$ , the following inequalities hold:

$$\|P_{\mathcal{M}}(x) - P_{\mathcal{M}}(y)\|_Q^2 \leq (x - y)^\top Q (P_{\mathcal{M}}(x) - P_{\mathcal{M}}(y)) \quad (A.4)$$

$$(x - P_{\mathcal{M}}(x))^\top Q (z - P_{\mathcal{M}}(x)) \leq 0 \quad (\forall z \in \mathcal{M}). \quad (A.5)$$

Next, we present important properties of the function  $V(x)$  in (7) for  $\mathcal{D}$  in (2b) as follows. Note that the function  $V$  in (7) is convex because of the convexity of each  $\mathcal{D}_l$ .

*Proposition 2.* For  $V(x)$  in (7) and a non-empty closed convex set  $\mathcal{D}$  in (2b),  $V(x) = 0 \Leftrightarrow x \in \mathcal{D}$  holds.

**Proof.** If  $V(x) = 0$  for  $x \in \mathbb{R}^{nd}$ , we obtain  $x_{\mathcal{C}_l} = P_{\mathcal{D}_l}(x_{\mathcal{C}_l}) \in \mathcal{D}_l$  for all  $l \in \text{clq}(G)$ , which yields  $x \in \mathcal{D}$  because of (2b). Conversely, if  $x \in \mathcal{D}$ , then we have  $x_{\mathcal{C}_l} \in \mathcal{D}_l$  for all  $l \in \text{clq}(G)$ . Thus,  $V(x) = 0$  holds.  $\square$

*Proposition 3.* The function  $V(x)$  in (7) is a 1-smooth function, i.e., its gradient  $\nabla V(x)$  is 1-Lipschitzian.

**Proof.** From Definition 1, the inequality (A.4), and Proposition 1, we obtain the following for any  $x, y \in \mathbb{R}^{nd}$ :

$$\begin{aligned} \|\nabla V(x) - \nabla V(y)\|^2 &= \|(x - y) - (T(x) - T(y))\|^2 \\ &= \|x - y\|^2 + \|T(x) - T(y)\|^2 - 2(x - y)^\top (T(x) - T(y)) \\ &= \|x - y\|^2 + \|T(x) - T(y)\|^2 \\ &\quad - 2 \sum_{l \in \text{clq}(G)} (x_{\mathcal{C}_l} - y_{\mathcal{C}_l})^\top \text{diag}(\gamma_{\mathcal{C}_l}) (P_{\mathcal{D}_l}(x_{\mathcal{C}_l}) - P_{\mathcal{D}_l}(y_{\mathcal{C}_l})) \\ &\leq \|x - y\|^2 + \|T(x) - T(y)\|^2 \\ &\quad - 2 \sum_{l \in \text{clq}(G)} \|P_{\mathcal{D}_l}(x_{\mathcal{C}_l}) - P_{\mathcal{D}_l}(y_{\mathcal{C}_l})\|_{\text{diag}(\gamma_{\mathcal{C}_l})}^2 \\ &\leq \|x - y\|^2 - \|T(x) - T(y)\|^2 \leq \|x - y\|^2. \end{aligned}$$

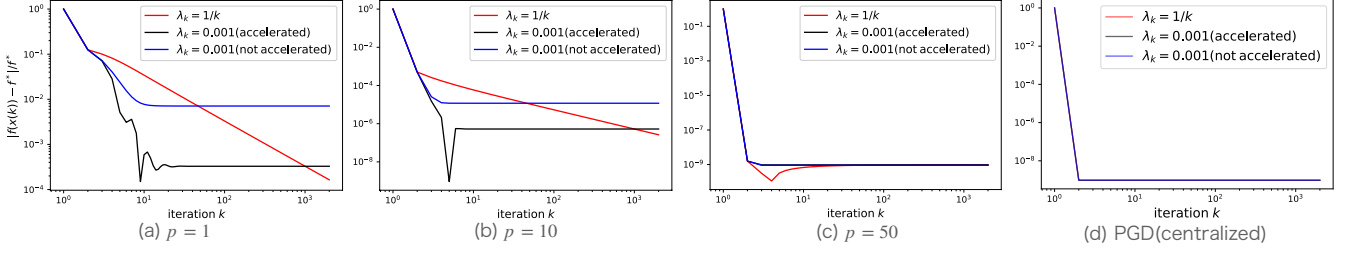


Fig. 2. Plots of relative residuals of  $f$  versus the number  $k$  of iterations under the CPGD with  $\lambda_k = 1/k$  (red line) and  $\lambda_k = 0.001$  (blue line), and the ACPGD with  $\lambda_k = 0.001$  (black line) for (a)  $p = 1$ , (b)  $p = 10$ , and (c)  $p = 50$ . For comparison purposes, the plot (d) shows the results of the PGD in (9) with  $\lambda_k = 1/k$ , 0.001 and the accelerated PGD with  $\lambda_k = 0.001$ .

The last line follows from (A.6) in the proof of Lemma 1a. It completes the proof.  $\square$

## A.2 Proof

Here, we show the proof of Lemma 1.

a) From (A.1) and (A.2), the following inequality holds:

$$\begin{aligned}
\|x - y\|^2 &= \sum_{l \in \text{clq}(G)} \|x_{C_l} - y_{C_l}\|_{\text{diag}(\gamma_{C_l})}^2 \\
&\geq \sum_{l \in \text{clq}(G)} \|P_{\mathcal{D}_l}(x_{C_l}) - P_{\mathcal{D}_l}(y_{C_l})\|_{\text{diag}(\gamma_{C_l})}^2 \\
&= \sum_{i=1}^n \sum_{l \in \text{clq}_i(G)} \frac{1}{|\text{clq}_i(G)|} \left\| [P_{\mathcal{D}_l}(x_{C_l})]_{m_{l,i}} \right. \\
&\quad \left. - [P_{\mathcal{D}_l}(y_{C_l})]_{m_{l,i}} \right\|^2 \\
&\geq \sum_{i=1}^n \left\| \sum_{l \in \text{clq}_i(G)} \frac{1}{|\text{clq}_i(G)|} \left( [P_{\mathcal{D}_l}(x_{C_l})]_{m_{l,i}} \right. \right. \\
&\quad \left. \left. - [P_{\mathcal{D}_l}(y_{C_l})]_{m_{l,i}} \right) \right\|^2 \\
&= \sum_{i=1}^n \|T_i(x_{\mathcal{N}_i}) - T_i(y_{\mathcal{N}_i})\|^2 = \|T(x) - T(y)\|^2. \quad (\text{A.6})
\end{aligned}$$

Thus, we obtain  $\|T(x) - T(y)\| \leq \|x - y\|$ .  $\square$

b)  $\mathcal{D} \subset \text{Fix}(T)$  holds because  $x_{C_l} = P_{\mathcal{D}_l}(x_{C_l})$  holds for any  $x \in \mathcal{D}$  and all  $l \in \text{clq}(G)$ . In the following, we prove the converse inclusion  $\text{Fix}(T) \subset \mathcal{D}$ . Let  $y \in \mathcal{D}$ . We show  $\hat{y} \in \text{Fix}(T) \setminus \{y\} \Rightarrow \hat{y} \in \mathcal{D}$ . From  $\hat{y} \in \text{Fix}(T)$ , we obtain  $\hat{y}_i = T_i(\hat{y}_{\mathcal{N}_i})$  for all  $i \in \mathcal{N}$ . In addition, from (A.1) and (A.3), we have

$$\begin{aligned}
\|y - \hat{y}\|^2 &\geq \sum_{l \in \text{clq}(G)} \|y_{C_l} - P_{\mathcal{D}_{C_l}}(\hat{y}_{C_l})\|_{\text{diag}(\gamma_{C_l})}^2 \\
&= \sum_{i=1}^n \sum_{l \in \text{clq}_i(G)} \frac{1}{|\text{clq}_i(G)|} \|y_i - [P_{\mathcal{D}_l}(\hat{y}_{C_l})]_{m_{l,i}}\|^2 \\
&\geq \sum_{i=1}^n \|y_i - \sum_{l \in \text{clq}_i(G)} \frac{1}{|\text{clq}_i(G)|} [P_{\mathcal{D}_l}(\hat{y}_{C_l})]_{m_{l,i}}\|^2 \\
&= \sum_{i=1}^n \|y_i - T_i(\hat{y}_{\mathcal{N}_i})\|^2 = \|y - \hat{y}\|^2.
\end{aligned}$$

Thus, from the equality condition of (A.1), we obtain  $y_i - [P_{\mathcal{D}_k}(\hat{y}_{C_k})]_{m_{k,i}} = y_i - [P_{\mathcal{D}_l}(\hat{y}_{C_l})]_{m_{l,i}}$  for all  $C_k, C_l (k, l \in \text{clq}_i(G))$  for all  $i \in \mathcal{N}$ . Then, we have  $[P_{\mathcal{D}_k}(\hat{y}_{C_k})]_{m_{k,i}} = [P_{\mathcal{D}_l}(\hat{y}_{C_l})]_{m_{l,i}}$  for all  $C_k, C_l (k, l \in \text{clq}_i(G))$ . Therefore, because  $\hat{y} \in \text{Fix}(T)$ , we have  $2V(\hat{y}) = \sum_{i=1}^n \sum_{l \in \text{clq}_i(G)} \frac{1}{|\text{clq}_i(G)|} \|\hat{y}_i - [P_{\mathcal{D}_l}(\hat{y}_{C_l})]_{m_{l,i}}\|^2 = \sum_{i=1}^n \|\hat{y}_i - T_i(\hat{y}_{\mathcal{N}_i})\|^2 = 0$ . Then,  $\hat{y} \in \mathcal{D}$  holds from Proposition 2. Hence, we obtain  $\text{Fix}(T) \subset \mathcal{D}$ .  $\square$

c) For a non-empty closed convex set  $\mathcal{D}$  in (2b) and  $x \in \mathbb{R}^{nd} \setminus \mathcal{D}$ , there exists  $\hat{l} \in \text{clq}(G)$  such that  $\|x_{C_{\hat{l}}} - P_{\mathcal{D}_{C_{\hat{l}}}}(x_{C_{\hat{l}}})\|_{\text{diag}(\gamma_{C_{\hat{l}}})} > 0$ . Hence, for  $\hat{l} \in \text{clq}(G)$ ,  $x \in \mathbb{R}^{nd} \setminus \mathcal{D}$ , and  $z \in \mathcal{D}$ , we have  $\|x_{C_{\hat{l}}} - z_{C_{\hat{l}}}\|_{\text{diag}(\gamma_{C_{\hat{l}}})}^2 > \|P_{\mathcal{D}_{C_{\hat{l}}}}(x_{C_{\hat{l}}}) - z_{C_{\hat{l}}}\|_{\text{diag}(\gamma_{C_{\hat{l}}})}^2$  because  $\|x_{C_{\hat{l}}} - z_{C_{\hat{l}}}\|_{\text{diag}(\gamma_{C_{\hat{l}}})}^2 =$

$\|x_{C_{\hat{l}}} - P_{\mathcal{D}_{C_{\hat{l}}}}(x_{C_{\hat{l}}})\|_{\text{diag}(\gamma_{C_{\hat{l}}})}^2 + \|P_{\mathcal{D}_{C_{\hat{l}}}}(x_{C_{\hat{l}}}) - z_{C_{\hat{l}}}\|_{\text{diag}(\gamma_{C_{\hat{l}}})}^2 - 2(x_{C_{\hat{l}}} - P_{\mathcal{D}_{C_{\hat{l}}}}(x_{C_{\hat{l}}}))^\top \text{diag}(\gamma_{C_{\hat{l}}})(z_{C_{\hat{l}}} - P_{\mathcal{D}_{C_{\hat{l}}}}(x_{C_{\hat{l}}})) > \|P_{\mathcal{D}_{C_{\hat{l}}}}(x_{C_{\hat{l}}}) - z_{C_{\hat{l}}}\|_{\text{diag}(\gamma_{C_{\hat{l}}})}^2$  holds, where the last line follows from (A.5).

Thus, by (A.1) and (A.2), for any  $x \in \mathbb{R}^{nd} \setminus \mathcal{D}$  and  $z \in \mathcal{D}$ , we obtain

$$\begin{aligned}
\|x - z\|^2 &= \sum_{l \in \text{clq}(G)} \|x_{C_l} - z_{C_l}\|_{\text{diag}(\gamma_{C_l})}^2 \\
&> \sum_{l \in \text{clq}(G)} \|P_{\mathcal{D}_l}(x_{C_l}) - z_{C_l}\|_{\text{diag}(\gamma_{C_l})}^2 \\
&\geq \sum_{i=1}^n \left\| \sum_{l \in \text{clq}_i(G)} \frac{1}{|\text{clq}_i(G)|} [P_{\mathcal{D}_l}(x_{C_l})]_{m_{l,i}} - z_i \right\|^2 \\
&= \sum_{i=1}^n \|T_i(x_{\mathcal{N}_i}) - z_i\|^2 = \|T(x) - z\|^2.
\end{aligned}$$

Therefore,  $\|T(x) - z\| < \|x - z\|$  holds for any  $x \in \mathbb{R}^{nd} \setminus \mathcal{D}$  and any  $z \in \mathcal{D}$ .  $\square$

d) For  $x \in \mathbb{R}^{nd}$ , we define  $\{a_k\}$  as  $a_{k+1} = T(a_k)$  with  $a_0 = x$ . Then, we obtain  $\lim_{k \rightarrow \infty} a_{k+1} = \lim_{k \rightarrow \infty} T(a_k)$ . Thus, from the continuity of  $T$  shown in Lemma 1a, we have  $T^\infty(x) = \lim_{k \rightarrow \infty} a_{k+1} = T(\lim_{k \rightarrow \infty} a_k) = T(T^\infty(x))$ . Hence, Lemma 1b yields  $T^\infty(x) \in \text{Fix}(T) = \mathcal{D}$ .  $\square$

## Appendix B. PROOF OF THEOREM 1C AND 2

Here, we show the proofs of Theorem 1c and 2. These proofs are based on the convergence theorems for ISTA and FISTA (Theorem 3.1 and 4.4 in Beck and Teboulle (2009)), respectively.

### B.1 Supporting lemmas

Before proceeding to prove the theorems, we show some inequalities corresponding to those obtained from Lemma 2.3 in Beck and Teboulle (2009), which is a key to prove the convergence theorems. Note that a differentiable function  $h : \mathbb{R}^m \rightarrow \mathbb{R}$  is convex if and only if

$$h(y) \geq h(x) + \nabla h(x)^\top (y - x) \quad (\text{B.1})$$

holds for any  $x, y \in \mathbb{R}^{nd}$ . Besides, if  $h$  is  $\beta$ -smooth and convex, then

$$h(y) \leq h(x) + \nabla h(x)^\top (y - x) + \frac{\beta}{2} \|y - x\|^2 \quad (\text{B.2})$$

$$h(y) \geq h(x) + \nabla h(x)^\top (y - x) + \frac{1}{2\beta} \|\nabla h(x) - \nabla h(y)\|^2 \quad (\text{B.3})$$

holds for any  $x, y \in \mathbb{R}^{nd}$ . For details, see textbooks on nonlinear programming, e.g., Bertsekas (1999).

In preparation for showing lemmas, let  $t \in (0, 1/L]$  and

$$V_t(x) = V(x)/t$$

with  $V(x)$  in (7). Additionally, for  $s \in \mathbb{R}^{nd}$ , we define  $F_y : \mathbb{R}^{nd} \rightarrow \mathbb{R}$  with some  $y \in \mathbb{R}^{nd}$  as

$$F_y(s) = f(s) + V_t(y) + \nabla V_t(y)^\top (s - y). \quad (\text{B.4})$$

For  $F_y(s)$  in (B.4), the following inequalities hold.

*Proposition 4.* Assume that the convex function  $f$  satisfies Assumption 1. Let  $y = x - t\nabla f(x)$  for  $x \in \mathbb{R}^{nd}$ . Then,

$$F_y(T(y)) \leq F_y(z) + \frac{1}{t}(x - T(y))^\top (x - z) - \frac{1}{2t}\|x - T(y)\|^2 \quad (\text{B.5})$$

holds for any  $z \in \mathbb{R}^{nd}$ .

**Proof.** Let  $G_y(s) = f(s) + \nabla V_t(y)^\top (s - y)$  and  $z \in \mathbb{R}^{nd}$ . Then, by using  $L$ -smoothness of  $f$  in Assumption 1,  $\nabla f(x) = (x - y)/t$ , and  $\nabla V_t(y) = (y - T(y))/t$  (see Proposition 1),

$$\begin{aligned} & G_y(T(y)) = f(T(y)) + \nabla V_t(y)^\top (T(y) - y) \\ & \leq f(x) - \nabla f(x)^\top (x - T(y)) + \frac{1}{2t}\|x - T(y)\|^2 \\ & \quad + \nabla V_t(y)^\top (T(y) - y) \\ & \leq f(z) + \nabla f(x)^\top (x - z) - \nabla f(x)^\top (x - T(y)) \\ & \quad + \frac{1}{2t}\|x - T(y)\|^2 + \nabla V_t(y)^\top (z - y) + \nabla V_t(y)^\top (T(y) - z) \\ & = G_y(z) + \frac{1}{t}(x - T(y))(T(y) - z) + \frac{1}{2t}\|x - T(y)\|^2 \\ & = G_y(z) + \frac{1}{t}(x - T(y))^\top (x - z) - \frac{1}{2t}\|x - T(y)\|^2 \end{aligned}$$

is obtained from (B.1) and (B.2). Thus, adding  $V_t(y)$  to the both sides, we obtain (B.5).  $\square$

*Proposition 5.* Let  $x(k+1) = T(y(k))$  with some  $\{y(k)\} \subset \mathbb{R}^{nd}$ . Then, it holds that

$$\begin{aligned} & F_{y(k)}(x(k)) + \frac{t}{2}\|\nabla V_t(y(k))\|^2 \\ & \leq F_{y(k-1)}(x(k)) + \frac{t}{2}\|\nabla V_t(y(k-1))\|^2. \end{aligned} \quad (\text{B.6})$$

**Proof.** By  $1/t$ -smoothness of  $V_t(x)$  (see Proposition 3) and Proposition 1,

$$\begin{aligned} & F_{y(k-1)}(x(k)) = f(x(k)) + V_t(y(k-1)) \\ & \quad + \nabla V_t(y(k-1))^\top (x(k) - y(k-1)) \\ & = f(x(k)) + V_t(y(k-1)) - t\|\nabla V_t(y(k-1))\|^2 \\ & \geq f(x(k)) + V_t(y(k)) + \nabla V_t(y(k))^\top (y(k-1) - y(k)) \\ & \quad + \frac{t}{2}\|\nabla V_t(y(k-1)) - \nabla V_t(y(k))\|^2 - t\|\nabla V_t(y(k-1))\|^2 \\ & = \underbrace{f(x(k)) + V_t(y(k)) + \nabla V_t(y(k))^\top (x(k) - y(k))}_{=F_{y(k)}(x(k))} \\ & \quad + \nabla V_t(y(k))^\top \underbrace{(y(k-1) - x(k))}_{=t\nabla V_t(y(k-1))} \\ & \quad + \frac{t}{2}\|\nabla V_t(y(k-1)) - \nabla V_t(y(k))\|^2 - t\|\nabla V_t(y(k-1))\|^2 \\ & = F_{y(k)}(x(k)) + \frac{t}{2}\|\nabla V_t(y(k))\|^2 - \frac{t}{2}\|\nabla V_t(y(k-1))\|^2 \end{aligned}$$

is obtained from (B.3). Hence, (B.6) holds.  $\square$

With this in mind, we consider the following update rule with  $\hat{x}(0) = x(0)$  and some  $\{\theta_k\} \subset \mathbb{R}$ :

$$\begin{aligned} & y(k) = \hat{x}(k) - t\nabla f(\hat{x}(k)) \\ & x(k+1) = T(y(k)) \\ & \hat{x}(k+1) = x(k+1) + \theta_k(x(k+1) - x(k)). \end{aligned} \quad (\text{B.7})$$

In addition, we define  $H_k : \mathbb{R}^{nd} \rightarrow \mathbb{R}$  as

$$H_k = F_{y(k-1)}(x(k)) + \frac{t}{2}\|\nabla V_t(y(k-1))\|^2. \quad (\text{B.8})$$

with  $F_y$  in (B.4). By  $x(k) - y(k-1) = -t\nabla V_t(y(k-1))$ ,  $H_k$  can be rewritten as  $H_k = f(x(k)) + V_t(y(k-1)) - \frac{1}{2t}\|y(k-1) - T(y(k-1))\|^2 = f(x(k)) + V_t(y(k-1)) - \frac{1}{2t}\|y(k) - x(k+1)\|^2$ .

Remarkably,  $H_k$  in (B.8) satisfies the following lemma.

*Lemma 2.* Consider the sequence generated by (B.7). Then, it holds that

$$f(x(k)) + V_t(x(k)) \leq H_k. \quad (\text{B.9})$$

**Proof.** In light of  $1/t$ -smoothness of  $V_t$  and  $\nabla V_t(y(k-1)) = -(y(k-1) - x(k))/t$ , we obtain  $V_t(x(k)) \leq V_t(y(k-1)) + \nabla V_t(y(k-1))^\top (y(k-1) - x(k)) + \frac{1}{2t}\|y(k) - x(k+1)\|^2 = V_t(y(k-1)) - \frac{1}{2t}\|y(k) - x(k+1)\|^2$ . Hence, adding  $f(x(k))$  to both sides yields (B.9).  $\square$

Furthermore, for  $H_k$  in (B.8), we prove the following inequality, which is essential to the proof of Theorem 1c and 2.

*Lemma 3.* For the sequence generated by (B.7) and  $H_k$  defined in (B.8), it holds that

$$\begin{aligned} & H_k - H_{k+1} \geq \frac{1}{2t}\|\hat{x}(k) - x(k+1)\|^2 \\ & \quad + \frac{1}{t}(x(k+1) - \hat{x}(k))^\top (\hat{x}(k) - x(k)). \end{aligned} \quad (\text{B.10})$$

**Proof.** Substituting  $x = x(k+1)$ ,  $y = y(k)$ , and  $z = x(k)$  into (B.5), we obtain

$$\begin{aligned} & H_{k+1} = f(x(k+1)) + V_t(y(k)) \\ & \quad + \nabla V_t(y(k))^\top (x(k+1) - y(k)) + \frac{t}{2}\|\nabla V_t(y(k))\|^2 \\ & \leq f(x(k)) + V_t(y(k)) \\ & \quad + \nabla V_t(y(k))^\top (x(k) - y(k)) + \frac{t}{2}\|\nabla V_t(y(k))\|^2 \\ & \quad + \frac{1}{t}(\hat{x}(k) - x(k+1))^\top (\hat{x}(k) - x(k)) - \frac{1}{2t}\|\hat{x}(k) - x(k+1)\|^2 \\ & = F_{y(k)}(x(k)) + \frac{t}{2}\|\nabla V_t(y(k))\|^2 \\ & \quad + \frac{1}{t}(\hat{x}(k) - x(k+1))^\top (\hat{x}(k) - x(k)) - \frac{1}{2t}\|\hat{x}(k) - x(k+1)\|^2 \\ & \leq F_{y(k-1)}(x(k)) + \frac{t}{2}\|\nabla V_t(y(k-1))\|^2 \\ & \quad + \frac{1}{t}(\hat{x}(k) - x(k+1))^\top (\hat{x}(k) - x(k)) - \frac{1}{2t}\|\hat{x}(k) - x(k+1)\|^2 \\ & = H_k + \frac{1}{t}(\hat{x}(k) - x(k+1))^\top (\hat{x}(k) - x(k)) \\ & \quad - \frac{1}{2t}\|\hat{x}(k) - x(k+1)\|^2. \end{aligned}$$

from (B.1), (B.2), and (B.6). Thus, (B.10) holds.  $\square$

Moreover, for the relationship between  $x(k)$  and an optimal solution  $x_*$ , we present the following lemma.

*Lemma 4.* For  $x_* \in \arg \min_{x \in \mathcal{D}} f(x)$ , it holds that

$$f(x_*) + V_t(x_*) - H_{k+1} \geq \frac{1}{2t} \|\hat{x}(k) - x(k+1)\|^2 + \frac{1}{t} (x(k+1) - \hat{x}(k))^\top (\hat{x}(k) - x_*). \quad (\text{B.11})$$

**Proof.** Recalling (B.7),  $L$ -smoothness of  $f$ , and  $1/t$ -smoothness of  $V_t$  for  $t \in (0, 1/L]$ , we obtain

$$\begin{aligned} H_{k+1} &\leq f(\hat{x}(k)) - \nabla f(\hat{x}(k))^\top (\hat{x}(k) - x(k+1)) \\ &+ \frac{1}{2t} \|\hat{x}_k - x(k+1)\|^2 + V_t(y(k)) - \frac{1}{2t} \|y(k) - T(y(k))\|^2 \\ &\leq f(x_*) + \nabla f(\hat{x}(k))^\top (\hat{x} - z) - \nabla f(\hat{x}(k))^\top (\hat{x} - T(y(k))) \\ &+ \frac{1}{2t} \|\hat{x}(k) - T(y(k))\|^2 + V_t(x_*) - \frac{1}{2t} \|y(k) - T(y(k))\|^2 \\ &+ \frac{1}{t} (y(k) - T(y(k)))^\top (T(y(k)) - x_* + y(k) - T(y(k))) \\ &- \frac{1}{2t} \|y(k) - T(y(k)) - (x_* - T(x_*))\|^2 \\ &= f(x_*) + V_t(x_*) + \frac{1}{t} (\hat{x}(k) - x(k+1))^\top (\hat{x}(k) - x_*) \\ &- \frac{1}{2t} \|\hat{x}(k) - x(k+1)\|^2, \end{aligned}$$

from (B.1), (B.2), and (B.3), where the last line is obtained because  $x_* = T(x_*)$  holds for  $x_* \in \mathcal{D}$ . Therefore, (B.11) is obtained.  $\square$

### B.2 Proof of Theorem 1c

In this proof, assume that  $\theta_k = 0$  for all  $k$ . Then,  $\hat{x}(k) = x(k)$  holds and the algorithm in (B.7) equals to the CPGD with  $\lambda_k = t \in (0, 1/L]$  for all  $k \in \mathbb{N}$ .

In light of (B.11) and  $\hat{x}(k) = x(k)$ , we obtain  $2t(H_{k+1} - f^* + V_t(x_*)) \leq \|x_* - x(k)\|^2$  because  $2t(H_{k+1} - f^* + V(x_*)) \leq 2(x(k) - x(k+1))^\top (x(k) - x_*) - \frac{1}{2t} \|x(k) - x(k+1)\|^2 = \|x_* - x(k)\|^2 - \|x_* - x_{k+1}\|^2 \leq \|x_* - x(k)\|^2$ . Besides, invoking (B.10), we have

$$2t(H_{k+1} - H_k) \leq \|x(k) - x(k+1)\|^2 \leq 0.$$

Then, following the same procedure as Theorem 3.1 in Beck and Teboulle (2009) (with  $F(\mathbf{x}_k) \equiv H_k$  and  $F(\mathbf{x}^*) \equiv f^* + V_t(x_*)$ ) and using (B.9), we obtain (11).  $\square$

### B.3 Proof of Theorem 2

Substituting  $\theta_k = (\sigma_k - 1)/\sigma_{k+1}$  in (12) into (B.7) yields the ACPGD in (12).

Now, by (B.10), (B.11), and  $\sigma_{k-1}^2 = \sigma_k(\sigma_k - 1)$ , following the procedure of the proof for Theorem 4.4 in Beck and Teboulle (2009) gives

$$\begin{aligned} &\sigma_{k-1}^2 (H_k - (f^* + V_t(x_*))) - \sigma_k^2 (H_{k+1} - (f^* + V_t(x_*))) \\ &\leq \frac{1}{2t} (\|w_{k+1}\|^2 - \|w_k\|^2), \end{aligned}$$

where  $w_k = \sigma_k(\hat{x}(k) - x_*) - (\sigma_k - 1)(x(k) - x_*)$ . Thus, summing both sides over  $k = 1, 2, \dots$  yields

$$\sigma_k^2 (H_{k+1} - (f(x_*) + V_t(x_*))) \leq \frac{1}{2t} \|w_0\|^2 = \frac{1}{2t} \|x_0 - x_*\|^2.$$

By  $\sigma_k \geq (k+1)/2$ , which can be shown by mathematical induction, we obtain

$$H_{k+1} - (f(x_*) + V_t(x_*)) \leq \frac{2\|x_0 - x_*\|^2}{t(k+1)^2}.$$

Therefore, the inequality (13) follows from (B.9).  $\square$

- Beck, A. and Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.*, 2(1), 183–202.
- Bertsekas, D.P. (1999). *Nonlinear Programming*. Athena Scientific, Belmont, MA, USA.
- Bollobas, B. (1998). *Modern Graph Theory*. Springer Science & Business Media.
- Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.*, 3(1), 1–122.
- Calamai, P.H. and Moré, J.J. (1987). Projected gradient methods for linearly constrained problems. *Math. Program.*, 39(1), 93–116.
- Chang, T.H. (2016). A proximal dual consensus ADMM method for multi-agent constrained optimization. *IEEE Trans. Signal Process.*, 64(14), 3719–3734.
- Falson, A., Notarnicola, I., Notarstefano, G., and Prandini, M. (2020). Tracking-ADMM for distributed constraint-coupled optimization. *Automatica*, 117(108962), 108962.
- Kreyszig, E. (1991). *Introductory Functional Analysis with Applications*. John Wiley & Sons.
- Nedic, A., Ozdaglar, A., and Parrilo, P.A. (2010). Constrained consensus and optimization in multi-agent networks. *IEEE Trans. Automat. Contr.*, 55(4), 922–938.
- Nesterov, Y. (1983). A method of solving a convex programming problem with convergence rate. *Dokl. Akad. Nauk*, 269, 543–547.
- Notarnicola, I. and Notarstefano, G. (2020). Constraint-coupled distributed optimization: A relaxation and duality approach. *IEEE Trans. Control Netw. Syst.*, 7(1), 483–492.
- Peressini, A.L., Sullivan, F.E., Jr, U., and Jerry, J. (1988). *The Mathematics of Nonlinear Programming*. Springer-Verlag, Berlin, Heidelberg.
- Sakurama, K. and Sugie, T. (2021). Generalized coordination of multi-robot systems. *Foundations and Trends in Systems and Control*, 9(1), 1–170.
- Su, Y., Wang, Q., and Sun, C. (2022). Distributed primal-dual method for convex optimization with coupled constraints. *IEEE Trans. Signal Process.*, 70, 523–535.
- Terelius, H., Topcu, U., and Murray, R.M. (2011). Decentralized multi-agent optimization via dual decomposition. *IFAC Proceedings Volumes*, 44(1), 11245–11251.
- Wu, X., Wang, H., and Lu, J. (2022). Distributed optimization with coupling constraints. *IEEE Trans. Automat. Contr.*, 1–1.
- Yamada, I. (2001). The hybrid steepest descent method for the variational inequality problem over the intersection of fixed point sets of nonexpansive mappings. *Stud. Comput. Math*, 8, 473–504.
- Yamada, I., Ogura, N., and Shirakawa, N. (2002). A numerically robust hybrid steepest descent method for the convexly constrained generalized inverse problems. *Contemp. Math.*, 313, 269–305.
- Yang, T., Yi, X., Wu, J., Yuan, Y., Wu, D., Meng, Z., Hong, Y., Wang, H., Lin, Z., and Johansson, K.H. (2019). A survey of distributed optimization. *Annu. Rev. Control*, 47, 278–305.
- Zhu, M. and Martinez, S. (2012). On distributed convex optimization under inequality and equality constraints.



*IEEE Trans. Automat. Contr.*, 57(1), 151–164.