A Cutting Plane-based Distributed Algorithm for Non-smooth Optimisation with Coupling Constraints

Tianyi Zhong and David Angeli *Fellow, IEEE*

***Abstract*—In this paper, we study a general setup for constrained convex optimisation over time-varying networks. We propose a distributed algorithm, based on the cutting plane method, to address non-smooth optimisation challenges. Cutting plane-based approaches require constraint consensus which is structurally different from established consensus schemes. We bridge this gap by linking the cutting plane-based algorithm with a dynamic average tracking scheme. The distributed cutting plane algorithm is presented and its convergence is analysed. Its performance is investigated through a numerical example.**

***Index Terms*—Cutting Plane, Non-Smooth Optimisation, dynamic average tracking, Distributed Algorithm**

I. Introduction

DISTRIBUTED non-smooth convex optimisation, as a crucial class of convex optimisation problems, plays an important role in many engineering applications, such as image decomposition and reduction (e.g., L1-norm minimisation problem), compressed perception (e.g., Lasso problem) and artificial intelligence (e.g., L1-regularised empirical loss models). The growing attention towards non-smooth convex optimisation problems can be attributed to their versatility and efficacy in addressing real-world challenges.

Within this area, distributed cutting plane methods serve as an alternative direction to distributed subgradient methods for non-smooth minimisation problems. Typically, cutting plane methods do not require evaluation of the objective and all the constraint functions at each iteration. Hence, the local problem is much easier to solve. Moreover, cutting plane-based approaches alleviate the necessity for a global tuning step-size parameter, enhancing their practicality and efficiency in distributed optimisation frameworks.

This class of methods have been investigated by recent research. Existing fully distributed cutting plane-based algorithms aim at achieving constraints consensus and are structurally different from those well-known consensus-based algorithms. Most approaches (i.e. [1]) focus on solving convex optimisation problems with a common cost, where constraints are distributed through the network processors. The consensus and convergence are achieved by exchanging sets of constraints to form an outer approximation of the global feasible set. A decentralised cutting plane method was proposed in [2] where the lower bounds are constructed by querying agents in parallel.

The proposed distributed cutting plane tracking algorithm is different from the existing cutting plane consensus literature in the sense that constraints consensus is achieved along with sample points consensus. Therefore, existing consensus techniques can be deployed. Specifically, the same constraints are cooperatively constructed by individual agents via sample points consensus. In contrast with [2], our algorithm is fully distributed in the sense that individual agents decide the next sample point locally. Unlike [1], which adopts constraints consensus by directly exchanging the constraint set, most established consensus methods are compatible with our setup. The proposed algorithm and the algorithm in [1] address a local linear program with k and n (n represents the dimension of the common decision variable) inequality constraints at iteration step $k \geq n$. Compared with [1] that construct the constraint basis, the constraints used in our algorithm are built by computing $N-1$ (N denotes the number of agents) approximation functions. The approximation computation also exists in [2] where the central unit solves a dynamic unconstrained quadratic problem. The algorithm in [1] transmits at most $n(n+1)$ numbers per step whereas our proposed algorithm transmits $2n+1$ numbers at each step. Since the computation complexity and the communication burden depend on the number of iteration steps and the number of neighbours, the algorithm scales nicely with the network size.

II. Problem Setup and preliminaries

Consider a finite group of agents $\mathcal{N} = \{1, 2, \cdots, N\}$ in a time-varying network to cooperatively solve:

$$\begin{aligned} \min_{x_1,\dots,x_N} \; & \sum_{i\in\mathcal{N}} f_i(x_i) \\ \text{s.t.} \; & \sum_{i\in\mathcal{N}} g_i(x_i) \leq \mathbf{0} \\ & x_i \in \mathcal{X}_i \quad \forall i \in \mathcal{N} \end{aligned} \tag{1}$$

where $x_i \in \mathcal{X}_i \subseteq \mathbb{R}^{n_i}$ is the local decision variable, $\mathbf{0}$ is a vector of all zeros of dimension s, and the objective function $f_i : \mathbb{R}^{n_i} \to \mathbb{R}$ is private and only known to agent i. The

This paragraph of the first footnote will contain the date on which you submitted your paper for review.

The work was partially supported by project PRIN 20225MESZB.

Tianyi Zhong and David Angeli are with the Department of Electrical and Electronic Engineering, Imperial College London, United Kingdom (e-mail: tianyi.zhong20@imperial.ac.uk; dangeli@imperial.ac.uk).

David Angeli is also with the Department of Information Engineering of University of Florence, Italy.

decision variables are coupled by means of separable private and coupling constraints: $g_i : \mathbb{R}^{n_i} \to \mathbb{R}^s$.

We impose the following assumptions on problem (1).

Assumption 1 (Convexity and Compactness): For each $i \in \mathcal{N}$, the objective function $f_i : \mathbb{R}^n \to \mathbb{R}$ is strongly convex and the constraint function $g_i : \mathbb{R}^{n_i} \to \mathbb{R}^s$ is convex. Also, the constraint function $g_i : \mathbb{R}^{n_i} \to \mathbb{R}^s$ is bounded on $\mathcal{X}_i$. The set $\mathcal{X}_i$ is convex and compact.

Assumption 2 (Existence of Optimal Solutions and Interior Point): The solution of Problem (1) exists and the feasibility region has a nonempty interior. There exists $\tilde{x} = [\tilde{x}_1^T, \ldots, \tilde{x}_N^T]^T \in \text{ri}(\Pi_{i\in\mathcal{N}}\mathcal{X}_i)$ such that $\sum_{i\in\mathcal{N}} g_i(\tilde{x}_i) < \mathbf{0}$, where ri$(\cdot)$ denotes the relative interior of the set.

Assumption 2 states the satisfaction of Slater's condition and hence strong duality holds. Denote the multiplier associated with the coupling constraint $\sum_{i\in\mathcal{N}} g_i(x_i) \leq \mathbf{0}$ as $\mu \in \mathbb{R}^s$ and derive the dual problem of (1):

$$\max_{\mu\geq\mathbf{0}} \sum_{i\in\mathcal{N}} q_i(\mu) \tag{2}$$

where q_i is defined as:

$$q_i(\mu) = \min_{x_i\in\mathcal{X}_i} f_i(x_i) + \mu^T g_i(x_i) \tag{3}$$

It is known that each $q_i : \mathbb{R}^s \to \mathbb{R}$ is concave. Moreover, let $\mu_s \in \mathbb{R}^s$ and suppose that x_{μ_s} minimises the Lagrangian over $\mathcal{X}_i$, $x_{\mu_s} = \arg\min_{x_i\in\mathcal{X}_i}\{f_i(x_i) + \mu_s^T g_i(x_i)\}$. Under Assumption 1, $g_i(x_{\mu_s})$ is the gradient of q_i at μ_s (Theorem 6.3.3 in [3]) and the dual function is differentiable. The concavity of q_i and its bounded gradient indicates Lipschitz continuity of $q_i, i \in \mathcal{N}$.

A. Linear Approximation and Cutting Plane Oracle

The proposed distributed algorithm relies on the cutting plane method. The basic idea of the cutting plane method is interpreted as solving and iteratively refining a linear approximation problem of the dual problem. The sequence of linear approximation problems is constructed through oracles. Formally, consider the following maximisation problem:

$$\max_{\mu\geq\mathbf{0}, Q\in\mathbb{R}} Q \tag{4a}$$

$$\text{s.t.} \quad Q \leq \sum_{i\in\mathcal{N}} (g_s^i)^T(\mu - \mu_s) + q_s^i \quad \forall \mu_s \in \mathbb{R}^s \tag{4b}$$

where $q_s^i = q_i(\mu_s)$ and g_s^i is the gradient of $q_i(\mu)$ at μ_s. In principle, the right-hand side of (4b) is an upper bound of $\sum_{i\in\mathcal{N}} q_i(\mu)$ in (2) and they touch each other at point μ_s. We say that problem (4) is a linear approximation of problem (2).

Taking $\mathcal{Z} \subset \mathbb{R}^s$ as the target set that contains maxima of problem (2), $\mathcal{Z} = \arg\max_{\mu\geq\mathbf{0}} \sum_{i\in\mathcal{N}} q_i(\mu)$, we have that if $\mu_s \notin \mathcal{Z}$, then $\sum_{i\in\mathcal{N}} q_i(\mu_s) < \sum_{i\in\mathcal{N}} q_i(\mu), \forall \mu \in \mathcal{Z}$. Together with the concavity of $\sum_{i\in\mathcal{N}} q_i(\mu)$, we see:

$$\sum_{i\in\mathcal{N}} (g_s^i)^T(\mu - \mu_s) + q_s^i \geq \sum_{i\in\mathcal{N}} q_i(\mu) > \sum_{i\in\mathcal{N}} q_i(\mu_s) \tag{5}$$

and ultimately,

$$\sum_{i\in\mathcal{N}} (g_s^i)^T(\mu - \mu_s) > 0, \quad \forall \mu \in \mathcal{Z} \tag{6}$$

A cutting plane with respect to a query point $\mu_s \in \mathbb{R}^s$ is defined as a hyperplane $h(\mu_s) := \{\mu : a^T(\mu_s)\mu - b(\mu_s) = 0\}, a(\mu_s) \in \mathbb{R}^s$ and $b(\mu_s) \in \mathbb{R}$ that satisfies:

$$a^T(\mu_s)\mu - b(\mu_s) > 0, \quad \forall \mu \in \mathcal{Z} \tag{7a}$$

$$a^T(\mu_s)\mu_s - b(\mu_s) \leq 0 \tag{7b}$$

The cutting plane $h(\mu_s)$ separates μ_s and $\mathcal{Z}$. Given a query point μ_s and a set $\mathcal{Z}$, the cutting plane oracle either returns a cutting plane $h(\mu_s)$ that separates the point μ_s and the set $\mathcal{Z}$ or confirms that the point is contained in the set. From (6) and (7), we know $\{a^T(\mu_s)\mu - b(\mu_s) = 0\}$ is a cutting plane with respect to the query point μ_s by letting $a(\mu_s) = \sum_{i\in\mathcal{N}} g_s^i$ and $b(\mu_s) = \sum_{i\in\mathcal{N}} (g_s^i)^T\mu_s$.

Cutting-plane oracle:

Given a query point μ_s and a set $\mathcal{Z}$.

If $\mu_s \notin \mathcal{Z}$, generates a cutting plane $h(\mu_s)$ separating μ_s and $\mathcal{Z}$.

Otherwise asserts $\mu_s \in \mathcal{Z}$ and returns an empty h

To numerically solve problem (4), the query point is selected iteratively to generate cutting planes. The obtained optimum could gradually achieve the global optimum of the dual problem (2). Even though evaluating g_s^i and q_s^i in (4b) can be done in parallel across all agents, a central node is required to form the corresponding cutting plane.

III. Tracking Cutting Plane

To fully distribute the algorithm, the local dual updating process of agent i at step k is designed as:

$$\begin{aligned} [y_k^i, Q_k^i] \in &\arg\max_{y\geq\mathbf{0}, Q\in\mathbb{R}} Q \\ &\text{s.t.}\ Q \leq (g_s^i)^T(y - \mu_s^i) + q_s^i + \sum_{j\neq i}(g_s^{i:j})^T(y - \mu_s^i) \\ &\qquad + q_s^{i:j}, \quad s \in \{1, \ldots, k-1\} \end{aligned} \tag{8}$$

where $g_s^i = g_i(x_s^i)$, x_s^i is the primal minimiser related to μ_s^i and $q_s^i = q_i(\mu_s^i)$. The auxiliary variable $g_s^{i:j}$ estimates $g_j(x_s^i), j \neq i$. In this setup, we denote query points $\{\mu_s^i\}_{s\leq k-1}$ as sample points in the sense that each agent decides locally at which points to generate cutting planes.

Unlike the centralised algorithm that set μ_k^i as the next sample point, the distributed sample point is selected based on the following dynamic average tracking scheme:

$$\mu_k^i = \sum_{j\in\mathcal{N}} a_k^{ij}\mu_{k-1}^j + y_k^i - y_{k-1}^i \tag{9}$$

which is initialised identically for all agent $i \in \mathcal{N}$ as $\mu_0^i = \frac{1}{N}\sum_{j\in\mathcal{N}} y_0^j$. Under this scheme and the following assumption, μ_k^i acts as a distributed tracker of the time-varying signal $\frac{1}{N}\sum_{i\in\mathcal{N}} y_k^i$ (see proof of Lemma 2 (i)).

Assumption 3 (Weight Coefficients): If agent j transmits information to agent i, then $a_k^{ij} > 0$, otherwise $a_k^{ij} = 0$. There exists $\eta \in (0, 1)$ such that for all $i, j \in \mathcal{N}$ and all $k \geq 0$, if $a_k^{ij} > 0$ then $a_k^{ij} \geq \eta$. Moreover,

1) $\sum_{j=1}^N a_k^{ij} = 1$, for all $i = 1, \ldots, N$
2) $\sum_{i=1}^N a_k^{ij} = 1$, for all $j = 1, \ldots, N$

Each agent i gives weights to the received information and assigns zero weights to j when there is no communication link. More generally, when agent j's transmitted information is not available to agent i at the current step, the communication link is modelled as unconnected. This unavailability could arise when there exists temporary communication failures or bounded time delays.

The dual problem (2) may have multiple optimal solutions, while any optimal dual solution can lead to the primal minimum under the strong duality condition. In a distributed setup, reconstruction of the optimal global dual solution depends on consensus. In order to guarantee consensus on local sample points, it is necessary for agents to employ a common rule to select one optimiser. Hence, we adopt the lexicographic criterion (see Definition 42 of Appendix C in [4] for the definition) to obtain a unique solution in (8), termed lex-optimal solution.

A. Communication and Local Estimation

Problem (8) is a local estimation of problem (4). To update this local estimation, agents share some of the information with others. The shared information is processed locally by each individual agent and cannot be requested by other agents. We assume the agents update and transmit the shared information at discrete time steps. The communication between agents is modelled by a time-varying undirected graph sequence $\{\mathcal{G}_a(0), \mathcal{G}_a(1), \dots\}$. Every graph instance $\mathcal{G}_a(k)$ includes a time-invariant vertex set $\mathcal{N} = \{1, 2, \cdots, N\}$ that corresponds to agents and a time-varying edge set $E(k)$ represents their communication links. The agent i and j can transmit information with each other at step k through the undirected edge $(i, j) \in E(k)$. Also, $(i, i) \in E(k), \forall i \in \mathcal{N}, \forall k \geq 0$. The assumption on the communication is the following.

Assumption 4 (Connectivity and Communication): The undirected graph $\mathcal{G}_a^\infty := (\mathcal{N}, \bigcup_{\tau=1}^{\infty} E(\tau))$ is fully connected. Moreover, there exists $T \geq 1$ such that $\bigcup_{\tau=k}^{k+T} E(\tau) = \mathcal{N}^2$ for all $k \geq 0$, i.e. agent i receives information from a neighbouring agent j at least once every consecutive T iterations.

Assumption 4 ensures that the information transmitted by one agent can influence any other agents infinitely often. It is a crucial assumption for achieving optimality.

At step k, agent j solves the local optimisation problem, samples his local objective function at the obtained point μ_k^j, broadcasts $\mu_k^j, q_k^j = q_j(\mu_k^j), g_k^j = \nabla q_j(\mu_k^j)$ to and receives $\mu_k^i, q_k^i = q_i(\mu_k^i), g_k^i = \nabla q_i(\mu_k^i)$ from his current neighbours $i \in \{i \in \mathcal{N} : (i, j) \in E(k)\}$.

Suppose agent i is agent j's neighbour at step $m \in M_{ji}^{k-1}$ where $M_{ji}^{k-1} = \{m | (i, j) \in \mathcal{G}_a(m) \, \text{and} \, m \leq k-1\}$. Storing all received information of agent j including $\{\mu_m^j\}$ and corresponding $\{q_m^j = q_j(\mu_m^j)\}, \{g_m^j = \nabla q_j(\mu_m^j)\}$, a piecewise linear function $q_k^{i:j}(\mu)$ can be constructed by agent i:

$$q_k^{i:j}(\mu) := \min_{m \in M_{ji}^{k-1}} q_m^j + (g_m^j)^T(\mu - \mu_m^j) \tag{10}$$

Notice that, for all $\mu_s^i \in \{\mu_m^j, m \in M_{ji}^k\}$, $q_k^{i:j}(\mu_s^i) = q_j(\mu_s^i)$, otherwise $q_k^{i:j}(\mu_s^i) > q_j(\mu_s^i)$. Therefore, $q_k^{i:j}(\mu)$ bounds the objective function $q_j(\mu)$ from above and this upper bound becomes tighter as the cardinality of the set M_{ji}^{k-1} goes to infinity.

The global cutting plane can be reconstructed locally as long as the local sample points coincide. Then, solving problem (8) with approximated information can lead to a reliable optimal solution.

B. Distributed Cutting Plane Algorithm

The detailed iterative approach of the proposed distributed cutting plane tracking algorithm is presented in Algorithm 1. The network topology is initialised as a fully connected graph. The initial points y_0^i can be arbitrarily chosen in the set $\mathbb{R}_+^s$ and μ_0^i should be initialised as the average of all y_0^i, $\mu_0^i = 1/N \sum_{i \in \mathcal{N}} y_0^i$. This initialisation could also be achieved through average consensus ahead of optimisation.

Algorithm 1 Distributed Cutting Plane Tracking Algorithm

1: **Initialization** :$\{a_0^{ij}\}$, $\mathcal{G}_a(0)$, $y_0^i \in \mathbb{R}_+^s$, $\mu_0^i = \frac{1}{N} \sum_{i=1}^{N} y_0^i$
2: **Evolution** : for $k = 1, 2, \dots$
3: Update $x_{k-1}^i = \arg\min_{x_i \in \mathcal{X}_i} f_i(x_i) + (\mu_{k-1}^i)^T g_i(x_i)$
4: Gather $\mu_{k-1}^j, g_{k-1}^j = g_j(x_{k-1}^j), q_{k-1}^j = q_j(\mu_{k-1}^j)$ from all neighbours j.
5: Update for all neighbours j
$M_{ji}^{k-1} = \{m | (i, j) \in \mathcal{G}_a(m) \, \text{and} \, m \leq k-1\}$.
$q_{k-1}^{i:j} \leftarrow \min_{m \in M_{ji}^{k-1}} q_m^j + (g_m^j)^T(\mu_{k-1}^i - \mu_m^j)$
$m^* = \max\{\arg\max_{m \in M_{ji}^{k-1}} q_m^j + (g_m^j)^T(\mu_{k-1}^i - \mu_m^j)\}$
$g_{k-1}^{i:j} \leftarrow g_{m^*}^j$
6: Compute the lex-optimal solution of the following optimisation problem
$[y_k^i, Q_k^i] = \arg\max_{y \geq \mathbf{0}, Q \in \mathbb{R}} Q$
s.t. $Q \leq (g_s^i)^T(y - \mu_s^i) + q_s^i + \sum_{j \neq i} (g_s^{i:j})^T(y - \mu_s^i) + q_s^{i:j}, \quad s \in \{1, \dots, k-1\}$
$\mu_k^i = \sum_{j \in \mathcal{N}} a_k^{ij} \mu_{k-1}^j + y_k^i - y_{k-1}^i$
7: Update $\mathcal{G}_a(k)$ and $\{a_k^{ij}\}$

Updating coefficients $\{a_k^{ij}\}_{k \geq 0}$ that fulfil Assumption 3 requires agents to agree on an infinite sequence of doubly stochastic matrices. This agreement is feasible, for example, by using Metropolis weights.

In practice, agent $i \in \mathcal{N}$ will terminate its update process when the local estimation for the gradient of the global dual function at the latest sample point approaches zero. More specifically, $\|g_i(x_k^i) + \sum_{j \neq i} g_k^{i:j}\|$ (where $g_k^{i:j}$ is defined in Algorithm 1 step 5) keeps below the user-defined tolerance for a number of iterations greater than T (T is defined in Assumption 4).

Some appealing features of the proposed algorithm are worth highlighting. The algorithm does not tune a global step-size sequence, making the convergence rate insensitive to parameters. Meanwhile, the algorithm enhances privacy protection of all agents by transmitting aggregated information derived from the local dual functions so that agents do not disclose sensitive primal information (such as local primal state x_i, local cost f_i or local (sub)gradient ∇f_i).

C. Technical Analysis

The process of generating local cutting planes could be interpreted as separately approximating every other agent's dual function $q_i(\mu)$. The following analysis is based on this interpretation. The approximated dual function of agent j is updated by adding a new hyperplane which is defined as: $h_{j,k-1} := \{(\mu, q) \in \mathbb{R}^s \times \mathbb{R} : (g^j_{k-1})^T(\mu - \mu^j_{k-1}) + q^j_{k-1} - q = 0\}$. According to the algorithm, the next sample point μ^i_k is chosen as:

$$y^i_k = \arg\max_{y \geq \mathbf{0}} \sum_{j \in \mathcal{N}} q^{i:j}_k(y) \tag{11}$$

$$\mu^i_k = \sum_{j \in \mathcal{N}} a^{ij}_k \mu^j_{k-1} + y^i_k - y^i_{k-1} \tag{12}$$

where y^i_k is the lex-optimal solution. Let us denote $\sum_{j \in \mathcal{N}} q^{i:j}_k(y)$ as $C^i_k(y)$. By construction, $C^i_k(y)$ is concave and non-increasing with respect to k and $g^i_k + \sum_{j \neq i} g^{i:j}_k$ is the graident of $C^i_k(y)$ at point μ^i_k.

We start the analysis of Algorithm 1 with the property of sequence $\{C^i_k(y)\}_{k \geq 0}$ and results derived directly from previous research.

Lemma 1: The sequence of functions $C^i_k(y)$ uniformly converges to $\bar{C}_i(y)$ and the sequence of sample points $\{y^i_k\}$ converges to a limit point $\hat{y}_i$ as k goes to infinity, for all $i \in \mathcal{N}$.

Lemma 2: Let *Assumptions* 1-4 hold. Consider sequences $\{y^i_k\}_{k \geq 0}$ and $\{\mu^i_k\}_{k \geq 0}, i \in \mathcal{N}$ generated by the proposed algorithm and let $\bar{\mu}_k = \frac{1}{N}\sum_{i \in \mathcal{N}} \mu^i_k$, we have:

(i) [5] The network average of the local variables μ^i_k defined by $\bar{\mu}_k = \frac{1}{N}\sum_{i \in \mathcal{N}} \mu^i_k$ coincides with the average of y^i_k:

$$\bar{\mu}_k = \frac{1}{N}\sum_{i \in \mathcal{N}} y^i_k \tag{13}$$

(ii) (Following similar lines as Lemma 2 in [6]) For all k, s with $s \geq 0, k > s$, and for all $i \in \mathcal{N}$, the following holds:

$$\begin{aligned}\|\mu^i_k - \bar{\mu}_k\| \leq & \gamma q^{k-s-1} \sum_{j=1}^{N} \|\mu^j_s\| + \sum_{r=s+1}^{k-1} \gamma q^{k-r-1} \sum_{j=1}^{N} \|d^j_r\| \\ & + \|d^i_k\| + \frac{1}{N}\sum_{j=1}^{N} \|d^j_k\|\end{aligned} \tag{14}$$

where η is a lower bound defined in *Assumption 3*, $\gamma = 2(1 + \eta^{-(N-1)T})/(1 - \eta^{(N-1)T}) \in \mathbb{R}_+$, $q = (1 - \eta^{(N-1)T})^{\frac{1}{(N-1)T}} \in (0, 1)$ and $d^i_k = y^i_k - y^i_{k-1}$.

We are now ready to state our main result.

Theorem 3: Let *Assumptions* 1-4 hold. Consider sequences $\{y^i_k\}_{k \geq 0}$ and $\{\mu^i_k\}_{k \geq 0}$ generated by Algorithm 1 with $y^i_0 \in \mathbb{R}^s_+$ and $\mu^i_0 = \frac{1}{N}\sum_{i \in \mathcal{N}} y^i_0, \forall i \in \mathcal{N}$. Let $\bar{y}_k$ and $\bar{\mu}_k$ be their average: $\bar{y}_k = \frac{1}{N}\sum_{i \in \mathcal{N}} y^i_k$ and $\bar{\mu}_k = \frac{1}{N}\sum_{i \in \mathcal{N}} \mu^i_k$. We have:

(i) The tracking error defined as $e^i_k = \mu^i_k - \bar{\mu}_k$ shrinks to 0, $\lim_{k \to \infty} \|e^i_k\| = 0, \forall i \in \mathcal{N}$.

(ii) All local quantities μ^i_k achieve consensus: $\lim_{k \to \infty} \|\mu^i_k - \mu^j_k\|, i, j \in \mathcal{N}$.

(iii) The sequence $\{\mu^i_k\}_{k \geq 0}$ converges to a limit point $\hat{\mu}$, $\lim_{k \to \infty} \|\mu^i_k - \hat{\mu}\| = 0$ holds for all $i \in \mathcal{N}$.

(iv) The limit point $\hat{\mu}$ maximises the global dual problem (2), $\hat{\mu} \in \arg\max_{\mu \geq \mathbf{0}} \sum_{i \in \mathcal{N}} q_i(\mu)$.

Proof: (i) According to *Lemma 1*, the convergence of sequence $\{y^i_k\}_{k \geq 0}$ implies $\lim_{k \to \infty} \|d^i_k\| = \lim_{k \to \infty} \|y^i_k - y^i_{k-1}\| = 0$ for all $i \in \mathcal{N}$. Choose any $\varepsilon \geq 0$, there exists some sufficiently large s such that $\|d^i_k\| \leq \varepsilon$ for all $k > s$. By *Lemma 2* (ii), we have:

$$\begin{aligned}\|\mu^i_k - \bar{\mu}_k\| &\leq \gamma q^{k-s-1} \sum_{j=1}^{N} \|\mu^j_s\| + N\gamma\varepsilon \sum_{r=s+1}^{k-1} q^{k-r-1} + 2\varepsilon \\ &= \gamma q^{k-s-1} \sum_{j=1}^{N} \|\mu^j_s\| + N\gamma\varepsilon \sum_{t=0}^{k-s-2} q^t + 2\varepsilon \\ &< \gamma q^{k-s-1} \sum_{j=1}^{N} \|\mu^j_s\| + N\gamma\varepsilon \sum_{t=0}^{\infty} q^t + 2\varepsilon \\ &\leq N\gamma D q^{k-s-1} + N\gamma \frac{\varepsilon}{1-q} + 2\varepsilon\end{aligned} \tag{15}$$

Taking limsup on both sides of (15), because $q \in (0, 1)$ we have for all $i \in \mathcal{N}$:

$$\limsup_{k \to \infty} \|\mu^i_k - \bar{\mu}_k\| \leq N\gamma \frac{\varepsilon}{1-q} + 2\varepsilon \tag{16}$$

Hence, by arbitrariness of $\varepsilon \geq 0$, the following holds:

$$\lim_{k \to \infty} \|e^i_k\| = \lim_{k \to \infty} \|\mu^i_k - \bar{\mu}_k\| = 0 \tag{17}$$

(ii) Directly from (i) , we know all local quantities μ^i_k achieve consensus:

$$\lim_{k \to \infty} \|\mu^i_k - \mu^j_k\| \leq \lim_{k \to \infty} \|\mu^i_k - \bar{\mu}_k\| + \lim_{k \to \infty} \|\mu^j_k - \bar{\mu}_k\| = 0 \tag{18}$$

(iii) According to *Lemma 1*, we know the limit of sequences $\{y^i_k\}_{k \geq 0}$ exist, denoted as $\hat{y}_i, \forall i \in \mathcal{N}$. Meanwhile, as discussed in *Lemma 2* (i), $\bar{\mu}_k = \frac{1}{N}\sum_{i \in \mathcal{N}} y^i_k$ holds. Therefore, the limit of sequence $\{\bar{\mu}_k\}_{k \geq 0}$ exists. Let $\hat{\mu} = \frac{1}{N}\sum_{i \in \mathcal{N}} \hat{y}_i$, we obtain:

$$\begin{aligned}\lim_{k \to \infty} \|\mu^i_k - \hat{\mu}\| &= \lim_{k \to \infty} \|\mu^i_k - \frac{1}{N}\sum_{i \in \mathcal{N}} \hat{y}_i\| \\ &\leq \lim_{k \to \infty} (\|\mu^i_k - \bar{\mu}_k\| + \|\bar{\mu}_k - \frac{1}{N}\sum_{i \in \mathcal{N}} y^i_k\|) \\ &+ \lim_{k \to \infty} \frac{1}{N} \|\sum_{i \in \mathcal{N}} y^i_k - \sum_{i \in \mathcal{N}} \hat{y}_i\| \\ &= 0\end{aligned} \tag{19}$$

(iv) By convergence of the Lagrangian multiplier μ^i_k, the primal solution $x^i_k = \arg\min_{x_i \in \mathcal{X}_i} f_i(x_i) + (\mu^i_k)^T g_i(x_i)$ also admits a limit point $\lim_{k \to \infty} \|x^i_k - \hat{x}_i\| = 0$ for each agent $i \in \mathcal{N}$. Based on properties of the dual function, we know $\lim_{k \to \infty} g^i_k = g_i(\hat{x}_i), \forall i \in \mathcal{N}$. We first show $\lim_{k \to \infty} g^{i:j}_k = \lim_{k \to \infty} g^j_k = g_j(\hat{x}_j)$ holds for all $i, j \in \mathcal{N}$. By construction (shown in Algorithm 1 step 5), we have:

$$q^{i:j}_k(\mu^i_k) = \min_{m \in M^{k-1}_{ji}} q_j(\mu^j_m) + (g^j_m)^T(\mu^i_k - \mu^j_m) \tag{20}$$

$$= q_j(\mu^j_{m_{j,k}}) + (g^j_{m_{j,k}})^T(\mu^i_k - \mu^j_{m_{j,k}}) \tag{21}$$

$$\leq q_j(\mu^j_k) + (g^j_k)^T(\mu^i_k - \mu^j_k) \tag{22}$$

where $m_{j,k}$ is the largest element in set $\mathcal{M}_j^k$ defined as:

$$\mathcal{M}_j^k := \arg \min_{m \in M_{ji}^{k-1}} q_j(\mu_m^j) + (g_m^j)^T(\mu_k^i - \mu_m^j) \tag{23}$$

$$m_{j,k} = \max \mathcal{M}_j^k \tag{24}$$

Assumption 4 implies $|M_{ji}^{k-1}| \to \infty$ as $k \to \infty$ for all $i, j \in \mathcal{N}$. In general, this gives us two possibilities: $m_{j,k}$ in (21) is finite or $m_{j,k} \to \infty$ as $k \to \infty$.

First, we deal with the case where $m_{j,k} \to \infty$ as $k \to \infty$ for some $j \neq i$. By construction $g_k^{i:j} = g_{m_{j,k}}^j$. Hence, $\lim_{k\to\infty} g_k^{i:j} = \lim_{k\to\infty} g_k^j = g_j(\hat{x}_j)$.

Next, we discuss the situation where there exists a finite step $\hat{m}$ such that $m_{j,k} = \hat{m}$ for all $k \geq \hat{m}$ and for some $j \neq i$ and we will show that $\lim_{k\to\infty} g_k^{i:j} = g_j(\hat{x}_j)$ also holds. From relationship (21)-(22), we have:

$$q_k^{i:j}(\mu_k^i) = q_j(\mu_{\hat{m}}^j) + (g_{\hat{m}}^j)^T(\mu_k^i - \mu_{\hat{m}}^j) \tag{25}$$

$$\leq q_j(\mu_k^j) + (g_k^j)^T(\mu_k^i - \mu_k^j) \tag{26}$$

$$\leq q_j(\mu_k^j) + \|g_k^j\|\|\mu_k^i - \mu_k^j\| \tag{27}$$

holds for all $k \geq \hat{m}$. According to consensus, we see $\lim_{k\to\infty} \|\mu_k^i - \mu_k^j\| = 0$, $\forall i \in \mathcal{N}$. Choose any $\epsilon \geq 0$, since $|M_{ji}^k| \to \infty$ as $k \to \infty$ there exists a step $\bar{k}$ such that $\|\mu_k^i - \mu_k^j\| \leq \epsilon/2$ holds for all $k \geq \bar{k}$ and all $i \in \mathcal{N}$. Hence from (25) and (27):

$$q_j(\mu_{\hat{m}}^j) + (g_{\hat{m}}^j)^T(\mu_{\bar{k}}^i - \mu_{\hat{m}}^j) \leq q_j(\mu_{\bar{k}}^j) + \epsilon\|g_{\bar{k}}^j\| \tag{28}$$

By concavity of $q_j(\mu)$, we see:

$$q_j(\mu_{\bar{k}}^i) \leq q_j(\mu_{\hat{m}}^j) + (g_{\hat{m}}^j)^T(\mu_{\bar{k}}^i - \mu_{\hat{m}}^j) \leq q_j(\mu_{\bar{k}}^j) + \epsilon\|g_{\bar{k}}^j\| \tag{29}$$

Together with Lipschitz continuity of $q_j(\mu)$, we have:

$$q_j(\mu_{\bar{k}}^i) \leq q_j(\mu_{\hat{m}}^j) + (g_{\hat{m}}^j)^T(\mu_{\bar{k}}^i - \mu_{\hat{m}}^j) \tag{30}$$

$$\leq q_j(\mu_{\bar{k}}^i) + \epsilon(\|g_{\bar{k}}^j\| + L) \tag{31}$$

We emphasise that $\bar{k}$ depends on ϵ. Choosing $\epsilon \to 0$, we have either (1): $\bar{k} \to \infty$ or (2): finite step convergence holds.

1) Notice that $\lim_{\bar{k}\to\infty} q_j(\mu_{\bar{k}}^i) = q_j(\hat{\mu})$ and $\lim_{\epsilon\to 0} q_j(\mu_{\bar{k}}^i) + \epsilon(\|g_{\bar{k}}^j\| + L) = q_j(\hat{\mu})$, therefore the following holds according to (30) and (31):

$$\lim_{\bar{k}\to\infty} q_j(\mu_{\hat{m}}^j) + (g_{\hat{m}}^j)^T(\mu_{\bar{k}}^i - \mu_{\hat{m}}^j) \tag{32}$$

$$= q_j(\mu_{\hat{m}}^j) + (g_{\hat{m}}^j)^T(\hat{\mu} - \mu_{\hat{m}}^j) \tag{33}$$

$$= q_j(\hat{\mu}) \tag{34}$$

where the latter equality implies $g_{\hat{m}}^j$ is the gradient of $q_j(\mu)$ at $\mu_{\hat{m}}^j$ and at $\hat{\mu}$.

2) Finite convergence implies there exists a step $\bar{k} < \infty$ independent of ϵ such that $\mu_{\bar{k}}^i = \hat{\mu}$ for all $k \geq \bar{k}$. According to (30)-(31) and take the limit of $\epsilon \to 0$:

$$q_j(\hat{\mu}) \leq q_j(\mu_{\hat{m}}^j) + (g_{\hat{m}}^j)^T(\hat{\mu} - \mu_{\hat{m}}^j) \leq q_j(\hat{\mu}) \tag{35}$$

which also implies $g_{\hat{m}}^j$ is the gradient of $q_j(\mu)$ at $\mu_{\hat{m}}^j$ and at $\hat{\mu}$.

For both cases, by definition, $g_{\hat{m}}^j = \nabla q_j(\hat{\mu}) = g_j(\hat{x}_j)$. Together with the construction $g_k^{i:j} = g_{m_{j,k}}^j$, we see $g_k^{i:j} = g_{\hat{m}}^j$ holds for all $k \geq \hat{m}$ and $\lim_{k\to\infty} g_k^{i:j} = g_j(\hat{x}_j)$ as claimed.

Let $g_i = \lim_{k\to\infty}(g_k^i + \sum_{j\neq i} g_k^{i:j})$, we see $g_i = \sum_{j\in\mathcal{N}} g_j(\hat{x}_j)$. By construction, g_i is a subgradient of $\bar{C}_i$ at $\hat{\mu}$. Notice that $\sum_{j\in\mathcal{N}} g_j(\hat{x}_j)$ is independent of i and therefore $g_i = g_j$, $\forall i, j \in \mathcal{N}$. This result implies that the local cutting planes achieve consensus. By defining $G = g_i$ and since $\hat{\mu} = \frac{1}{N}\sum_{i\in\mathcal{N}} \hat{y}_i$, we have:

$$\sum_{i\in\mathcal{N}} g_i^T(\hat{y}_i - \hat{\mu}) = \sum_{i\in\mathcal{N}} G^T(\hat{y}_i - \hat{\mu})$$

$$= G^T \sum_{i\in\mathcal{N}} (\hat{y}_i - \hat{\mu}) = 0 \tag{36}$$

Meanwhile, $\hat{y}_i$ maximises $\bar{C}_i(y)$, the following holds for all $i \in \mathcal{N}$:

$$g_i^T(\hat{y}_i - \hat{\mu}) \geq 0 \tag{37}$$

Relationship (36) and (37) implies $g_i^T(\hat{y}_i - \hat{\mu}) = 0$ holds for all $i \in \mathcal{N}$. According to the concavity property:

$$\bar{C}_i(\hat{y}_i) \leq \bar{C}_i(\hat{\mu}) + g_i^T(\hat{y}_i - \hat{\mu}) = \bar{C}_i(\hat{\mu}) \tag{38}$$

Hence, $\hat{\mu}$ coincides with the maximiser of all $\bar{C}_i(y)$, i.e. $\hat{\mu} = \hat{y}_i$, $\forall i \in \mathcal{N}$. Denote $Q(\mu) = \sum_{i\in\mathcal{N}} q_i(\mu)$. We conclude that:

$$|\bar{C}_i(\hat{\mu}) - Q(\hat{\mu})| \leq \lim_{k\to\infty} \sum_{j\neq i} |q_k^{i:j}(\hat{\mu}) - q_j(\hat{\mu})| = 0 \tag{39}$$

which directly shows the following relationship for all $i \in \mathcal{N}$:

$$\bar{C}_i(\hat{\mu}) = Q(\hat{\mu}) \leq Q(\mu^*) \leq \bar{C}_i(\mu^*) \leq \bar{C}_i(\hat{y}_i) = \bar{C}_i(\hat{\mu}) \tag{40}$$

where μ^* is a maximiser of $Q(\mu)$. Hence, we obtain $\hat{\mu}$ is also a maximiser of $Q(\mu)$. ■

By Assumption 2, strong duality holds. Hence, local primal solutions also converge to the primal optimum.

IV. Numerical Example

To demonstrate the performance of the algorithm, we test the proposed algorithm on problem (1) with $N = 10$ agents. The dimension of each agent's decision variable is set as $[n_1, \ldots, n_{10}] = [2, 2, 2, 2, 3, 2, 4, 3, 2, 2]$. The dimension of coupled constraints is $s = 2$. Local objective functions are:

$$f_i(x_i) = x_i^T Q_i x_i + r_i\|x_i\|_1 \tag{41}$$

where each $Q_i \in \mathbb{R}^{n_i \times n_i}$ is randomly generated such that its eigenvalues are drawn uniformly from $[1, 5]$ and each r_i is selected uniformly from $[0, 1]$. Communication graphs are generated randomly with T in Assumption 4 be set as $T = 30$. The local feasible set is $\mathcal{X}_i := \{x_i \in \mathbb{R}^{n_i} | \underline{x_i} \leq x_i \leq \overline{x_i}\}$ where $\underline{x_i}$ and $\overline{x_i}$ are drawn uniformly from $[-15, -5]$ and $[5, 15]$, respectively. The coupled constraint functions $g_i(x_i)$ are:

$$g_i(x_i) = A_i^T x_i - b_i \tag{42}$$

where entries of $A_i \in \mathbb{R}^{s\times n_i}$ are randomly selected from $[-1, 6]$ and $b_i \in \mathbb{R}^s \sim Normal(0, 5)$. The behaviour of Algorithm 1 and the comparison among other algorithms are shown in Fig.1 and Fig.2.

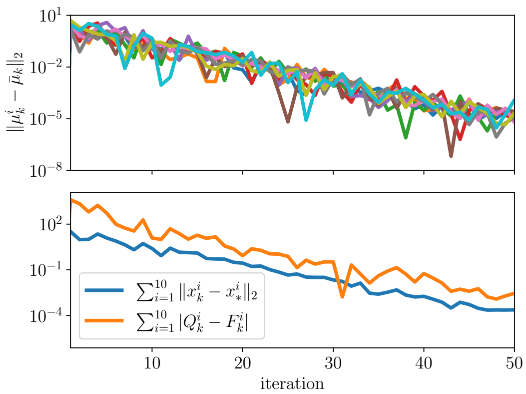

Fig. 1. The plot at the top is the evolution of normalised dual consensus residual $\|\boldsymbol{\mu}_k^i - \bar{\boldsymbol{\mu}}_k\|_2$ from $\boldsymbol{k} = \mathbf{1}$ to $\boldsymbol{k} = \mathbf{50}$. Different colour represents different agent. The plot at the bottom shows the evolution of normalised network primal residual $\sum_{i=1}^{10} \|\boldsymbol{x}_k^i - \boldsymbol{x}_*^i\|_2$ and the evolution of normalised network duality gap $\sum_{i=1}^{10} |\boldsymbol{Q}_k^i - \boldsymbol{F}_k^i|$.

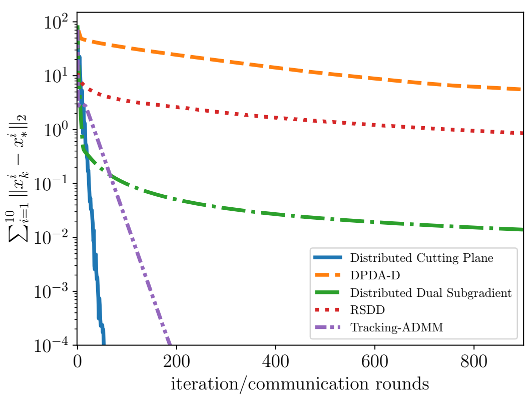

Fig. 2. Evolution of $\sum_{i=1}^{10} \|\boldsymbol{x}_k^i - \boldsymbol{x}_i^*\|_2$ among different algorithms. The other four algorithms need tuning parameters and we choose them based on the recommended parametric structure.

V. Conclusion

We have presented a distributed optimisation algorithm that combines the cutting plane method and dynamic average tracking protocol. We showed that the method links the constraint consensus scheme with the well-established decision variable consensus techniques. The convergence behaviour has been verified theoretically and numerically. Future investigations include relaxing topology assumptions and extending the algorithm to asynchronous implementation.

References

[1] M. Bürger, G. Notarstefano, and F. Allgöwer, "A polyhedral approximation framework for convex and robust distributed optimization," *IEEE Transactions on Automatic Control*, vol. 59, no. 2, pp. 384–395, 2013.

[2] T. Parshakova, F. Zhang, and S. Boyd, "Implementation of an oracle-structured bundle method for distributed optimization," *Optimization and Engineering*, Nov. 2023, ISSN: 1573-2924.

[3] M. S. Bazaraa, H. D. Sherali, and C. M. Shetty, *Nonlinear programming: theory and algorithms*. John wiley & sons, 2013.

[4] G. Notarstefano, I. Notarnicola, A. Camisa, *et al.*, "Distributed optimization for smart cyber-physical networks," *Foundations and Trends® in Systems and Control*, vol. 7, no. 3, pp. 253–383, 2019.

[5] A. Falsone, I. Notarnicola, G. Notarstefano, and M. Prandini, "Tracking-admm for distributed constraint-coupled optimization," *Automatica*, vol. 117, p. 108 962, 2020.

[6] K. Margellos, A. Falsone, S. Garatti, and M. Prandini, "Distributed constrained optimization and consensus in uncertain networks via proximal minimization," *IEEE Transactions on Automatic Control*, vol. 63, no. 5, pp. 1372–1387, 2017.

[7] D. Gale, V. Klee, and R. Rockafellar, "Convex functions on convex polytopes," *Proceedings of the American Mathematical Society*, vol. 19, no. 4, pp. 867–873, 1968.

Appendix I
Proof of preparatory results

A. Lemma 1

By definition, the sequence of functions $C_k^i(y)$ is monotonically non-increasing in k. Moreover, denote $Q(y) = \sum_{i\in\mathcal{N}} q_i(y)$, $C_k^i(y) \geq Q(y)$ holds for all $y \geq \mathbf{0}$ and $k \geq 0$. This shows existence of the limit $\bar{C}_i(y) = \lim_{k\to\infty} C_k^i(y)$. Notice that $\bar{C}_i(y)$ is a point-wise infimum of continuous functions and therefore it is upper semi-continuous. Moreover, $\bar{C}_i(y) = q_i(x) + \lim_{k\to\infty} q_k^{i:j}(y)$ and q_i is a concave function. Hence, $\bar{C}_i(y)$ is also concave. By the Gale-Klee-Rockafellar Theorem [7], a convex function on a polytopic set is upper semi-continuous. Apply this theorem on $-\bar{C}_i(y)$, we see $\bar{C}_i(y)$ is lower semi-continuous. Hence, $\bar{C}_i(x)$ is continuous. By Dini's Theorem, convergence of $C_k^i(x)$ to $\bar{C}_i(x)$ is uniform on the convex set $y \geq \mathbf{0}$.

Let $\hat{y}_k^i$ be one of the maxima of $C_k^i(y)$ over set $y \geq \mathbf{0}$, $\hat{y}_k^i \in \operatorname{argmin}_{y\geq\mathbf{0}} C_k^i(y)$, we know $C_k^i(\hat{y}_k^i) \geq C_k^i(y) \geq \bar{C}_i(y)$ for all $k \geq 0$ and $\forall y \geq \mathbf{0}$. Let Ω be the non-empty set of limit points of the sequence $\{\hat{y}_k^i\}$ and arbitrarily take $\omega \in \Omega$ and let $\{\hat{y}_{k_q}^i\}$ be a subsequence converging to ω as $q \to +\infty$. The following holds:

$$
\begin{aligned}
&\lim_{q\to+\infty} |C_{k_q}^i(\hat{y}_{k_q}^i) - \bar{C}_i(\omega)| \qquad (43)\\
&\leq \lim_{q\to+\infty} (|C_{k_q}^i(\hat{y}_{k_q}^i) - \bar{C}_i(\hat{y}_{k_q}^i)| + |\bar{C}_i(\hat{y}_{k_q}^i) - \bar{C}_i(\omega)|) = 0
\end{aligned}
$$

Then, we obtain that for any $y \geq \mathbf{0}$,

$$
\bar{C}_i(\omega) = \lim_{q\to+\infty} C_{k_q}^i(\hat{y}_{k_q}^i) \geq \lim_{q\to+\infty} C_{k_q}^i(y) \geq \bar{C}_i(y) \qquad (44)
$$

Hence, any $\omega \in \Omega$ is a maximiser of $\bar{C}_i(y)$. Under the lexicographic criterion, the obtained sequence $\{y_k^i\}_{k\geq 0}$ converges to a unique limit point $\hat{y}_i \in \Omega$.

B. Lemma 2 (i)

We prove this lemma by induction. Relationship (13) holds for $k = 0$ by definition of $\bar{\mu}_k$ and initialisation: $\bar{\mu}_0 = \frac{1}{N}\sum_{i\in\mathcal{N}} y_0^i$. Assume now that (13) holds up to k, then:

$$
\begin{aligned}
\bar{\mu}_{k+1} &= \frac{1}{N}\sum_{i\in\mathcal{N}}(\sum_{j\in\mathcal{N}} a_{k+1}^{ij}\mu_k^j + y_{k+1}^i - y_k^i)\\
&\overset{(a)}{=} \frac{1}{N}\sum_{j\in\mathcal{N}}(\sum_{i\in\mathcal{N}} a_{k+1}^{ij})\mu_k^j + \frac{1}{N}\sum_{i\in\mathcal{N}}(y_{k+1}^i - y_k^i)\\
&\overset{(b)}{=} \bar{\mu}_k - \frac{1}{N}\sum_{i\in\mathcal{N}} y_k^i + \frac{1}{N}\sum_{i\in\mathcal{N}} y_{k+1}^i\\
&= \frac{1}{N}\sum_{i\in\mathcal{N}} y_{k+1}^i \qquad (45)
\end{aligned}
$$

where in (a) we exchanged the summations, and in (b) we used column stochasticity of the weights and the definition of $\bar{\mu}_k$.