On linear convergence of a distributed dual gradient algorithm for linearly constrained separable convex problems

Ion Necoara and Valentin Nedelcu

Abstract

In this paper we propose a distributed dual gradient algorithm for minimizing linearly constrained separable convex problems and analyze its rate of convergence. In particular, we prove that under the assumption of strong convexity and Lipshitz continuity of the gradient of the primal objective function we have a global error bound type property for the dual problem. Using this error bound property we devise a new fully distributed dual gradient scheme, i.e. a gradient scheme based on a weighted step size, for which we derive global linear rate of convergence for both dual and primal suboptimality and for primal feasibility violation. Many real applications, e.g. distributed model predictive control, network utility maximization or optimal power flow, can be posed as linearly constrained separable convex problems for which dual gradient type methods from literature have sublinear convergence rate. In the present paper we prove for the first time that in fact we can achieve linear convergence rate for such algorithms when they are used for solving these applications. Numerical simulations are also provided to confirm our theory.

1 Introduction

Nowadays, many engineering applications which appear in the context of communications networks or networked systems can be posed as large scale linearly constrained separable convex problems. Several important applications that can be modeled in this framework, distributed model predictive control (DMPC) problem for networked systems [16], the network utility maximization (NUM) problem [2], and the direct current optimal power flow (DC-OPF) problem for a power system [1], have attracted great attention lately. Due to the large dimension and the separable structure of these problems, distributed optimization methods have become an appropriate tool for solving them.

Distributed optimization methods are based on decomposition [16]. Decomposition methods represent a powerful tool for solving these types of problems due to their ability of dividing the original large scale problem into smaller subproblems which are coordinated by a master problem. Decomposition methods can be divided into two main classes: primal and dual decomposition. While in the primal decomposition methods the optimization problem is solved using the original formulation and variables [5, 16, 27], in dual decomposition the constraints are moved into the cost using the Lagrange multipliers and then the dual problem is solved [3, 17, 26]. In many applications, such as (DMPC), (NUM) and (DC-OPF) problems, when the constraints set is complicated (i.e. the projection on this set is hard to compute), dual decomposition is more effective since a primal approach would require at each iteration a projection onto the feasible set, operation that is numerically expensive.

First order decomposition methods for solving dual problems have been extensively studied in the literature. Dual subgradient methods based on averaging, that produce primal solutions in the limit, can be found e.g. in [26]. Convergence rate analysis for the dual subgradient method is given e.g. in [20], where estimates of order $O(1/\sqrt{k})$ for suboptimality and feasibility violation of an average primal sequence are provided, with $k$ denoting the iteration counter. In [17] the authors derive a dual decomposition method based on a fast gradient algorithm and a smoothing technique and prove rate of convergence of order $O\left(\frac{1}{k}\right)$ for primal suboptimality and feasibility violation for an average primal sequence. Also, in [15, 19] ( [22]) the authors propose inexact (exact) dual fast gradient algorithms for which estimates of order $O\left(\frac{1}{k}\right)$ in an average primal sequence are provided for primal suboptimality and feasibility violation. For the special case of QP problems, dual fast gradient algorithms were also analyzed in [6,9,23]. To our knowledge, the first result on the linear convergence of dual gradi-
ent methods was provided in [11]. However, the authors in [11] were able to show linear convergence only \textit{locally} using a local error bound condition that estimates the distance from the dual optimal solution set in terms of norm of a proximal residual. Another strand of this literature uses alternating direction method of multipliers (ADMM) [8, 29] or Newton methods [18, 30]. For example, [8] established a linear convergence rate of (ADMM) using an error bound condition that holds under specific assumptions on the primal problem, while in [29] sublinear rate of convergence is proved for (ADMM), but for more general assumptions on the primal objective function. In [18, 30] distributed Newton algorithms are derived with fast convergence under the assumption that the primal objective function is self-concordant. Finally, very few results were known in the literature on distributed implementations of dual gradient type methods since most of the papers enumerated above require a centralized step size. Recently, in [2] (and [14]), distributed dual fast gradient algorithms are given, where the step size is chosen distributively, and estimates of order $O\left(\frac{1}{k}\right)$ for primal suboptimality and infeasibility in the last primal iterate (linear) are given.

Despite widespread use of gradient methods for solving dual problems, there are some aspects that have not been fully studied. In particular, in applications the main interest is in finding primal vectors that are near-feasible and near-optimal. We also need to devise algorithms with fast convergence rate, e.g. linear convergence. Finally, we are interested in providing distributed schemes, i.e. methods based on distributed computations. These represent the main issues that we pursue in this paper.

Contributions: In this paper we propose a novel distributed dual gradient method generating approximate primal feasible and optimal solutions but with great improvement on the convergence rate w.r.t. existing results. Under the assumptions of strong convexity and Lipschitz continuity of the gradient of the primal objective function, which are often satisfied in practical applications (e.g. (DMPC), (NUM) or (DC-OPF)), we prove that the corresponding dual problem satisfies a certain global error bound property that estimates the distance from the dual optimal solution set in terms of the norm of a proximal residual. In order to prove such a result we extend the approach developed in [11, 28], where the authors show such a property for objective functions having a certain structure and constraints set defined in terms of bounded polyhedra, to the case where the objective function is more general and the constraints set is an unbounded polyhedron. This nontrivial extension also allows us to tackle dual problems, where e.g. the constraints are defined in terms of the nonnegative orthant. In these settings we analyze the convergence behavior of a new distributed dual gradient algorithm, for which we are able to provide for the first time \textit{global} linear convergence rate on primal suboptimality and feasibility violation for the last primal iterate, as opposed to the results in [11] where only \textit{local} linear convergence was derived for such an algorithm. Moreover, our algorithm is fully distributed since is based on a weighted step size, as opposed to typical dual distributed schemes existing in literature, where a centralized step size is used and sublinear convergence is proved [4, 6, 13, 15]. Note that our results are also related to those in [8]: in particular, paper [8] established an error bound property for the \textit{augmented} dual function and then proved linear convergence for the (ADMM) method. However, the main drawbacks with (ADMM) consist of the difficulty in tuning the penalty parameter in the augmented Lagrangian and the centralized choice of it.

Paper Outline: In Section 2 we introduce our optimization model and discuss the (DMPC) problem for a networked system. In Section 3 we prove a certain error bound property of the dual function which allows us to derive global linear converge for a fully distributed dual gradient method in Section 4. In Sections 5 we discuss implementation issues, in particular in the context of (DMPC), and finally in Section 6 we provide some numerical simulations that confirm our theory.

Notations: For $z, y \in \mathbb{R}^n$ we denote the Euclidean inner product $\langle z, y \rangle = z^T y = \sum_{i=1}^n z_i y_i$, the Euclidean norm $\|z\| = \sqrt{\langle z, z \rangle}$ and the infinity norm $\|z\|_\infty = \sup_{i} |z_i|$. For a matrix $G \in \mathbb{R}^{n \times n}$, $\|G\|$ denotes its spectral norm. Also, we denote the orthogonal projection onto the nonnegative orthant $\mathbb{R}^+_n$ by $[z]_+$ and the orthogonal projection onto the convex set $D$ by $[z]_D$. For a positive definite matrix $W$ we define norm $\|z\|_W = \sqrt{z^T W z}$ and the projection of vector $z$ onto a convex set $D$ w.r.t. norm $\|\cdot\|_W$ by $[z]_D^W$. For a matrix $A$, $A_i$ is its $i$th (block) column.

2 Problem formulation

We consider the following large scale linearly constrained separable convex optimization problem:

$$
\begin{align*}
    f^* &= \min_{z \in \mathbb{R}^n} f(z) \\
    &\quad \text{s.t.: } Az = b, \quad Cz \leq c,
\end{align*}
$$

where $f_i : \mathbb{R}^n \to \mathbb{R}$ are convex functions, $z = [z_1^T \ldots z_M^T]^T \in \mathbb{R}^n$, $A \in \mathbb{R}^{p \times n}$, $C \in \mathbb{R}^{q \times n}$, $b \in \mathbb{R}^p$ and $c \in \mathbb{R}^q$. To our problem (1) we associate a bipartite communication graph $G = (V_1, V_2, E)$, where $V_1 = \{1, \ldots, M\}$, $V_2 = \{1, \ldots, M\}$ and $E \in \{0, 1\}^{M \times M}$ represents the incidence matrix. E.g., in the context of (NUM) and (DC-OPF), $V_1$ denotes the set of sources, $V_2$ the set of links between sources and the incidence matrix $E$ models the way sources interact. In (DMPC), $V_1 = V_2$ represents the set of interacting subsystems, while the incidence matrix $E$ indicates the dynamic couplings between these subsystems. We assume that $A$ and $C$ are block matrices with the blocks $A_{ji} \in \mathbb{R}^{p_j \times n_i}$ and $C_{ji} \in \mathbb{R}^{q_j \times n_i}$, where $\sum_{i=1}^M n_i = n$, $\sum_{j=1}^M p_j = p$. 

and $\sum_{j=1}^{M} g_j = q$. We also assume that if $E_{ji} = 0$, then both blocks $A_{ji}$ and $C_{ji}$ are zero. In these settings we allow a block $A_{ji}$ or $C_{ji}$ to be zero even if $E_{ji} = 1$. We also introduce the index sets:

$$\tilde{N}_i = \{ j \in V_2 : E_{ji} \neq 0 \} \text{ and } \tilde{N}_j = \{ i \in V_1 : E_{ji} \neq 0 \}$$

for all $i \in V_1, j \in V_2$, which describe the local information flow in the graph. Note that the cardinality of the sets $\tilde{N}_i$ and $\tilde{N}_j$ can be viewed as a measure for the degree of separability of problem (1). Therefore, the local information structure imposed by the graph $\mathcal{G}$ should be considered as part of the problem formulation. Further, we make the following assumptions regarding the optimization problem (1):

**Assumption 2.1**

(a) The functions $f_i$ have Lipschitz continuous gradient with constants $L_i$ and are $\sigma_i$-strongly convex w.r.t. the Euclidean norm $\| \cdot \|$ on $\mathbb{R}^n$ [21].

(b) Matrix $A$ has full row rank and there exists a feasible point $\tilde{z}$ for problem (1) such that $A\tilde{z} = b$ and $C\tilde{z} < c$.

Note that if Assumption 2.1 (a) does not hold, we can apply smoothing techniques by adding a regularization term to the function $f_i$ in order to obtain a strongly convex approximation of it (see e.g. [17] for more details). Assumption 2.1 (b) implies that strong duality holds for optimization problem (1) and the set of optimal Lagrange multipliers is bounded [7, Theorem 2.3.2]. Note that Assumption 2.1 (b) is not restrictive; we can always remove the redundant equalities so that matrix $A$ has full row rank and strict feasibility for the inequality constraints is usually satisfied in applications (e.g. (DMPC), (NUM) or (DC-OPF)). In particular, we have:

$$f^* = \max_{\nu \in \mathbb{R}^n, \mu \in \mathbb{R}_+} d(\nu, \mu),$$

where $d(\nu, \mu)$ denotes the dual function of (1):

$$d(\nu, \mu) = \min_{z \in \mathbb{R}^n} \mathcal{L}(z, \nu, \mu),$$

with the Lagrangian function

$$\mathcal{L}(z, \nu, \mu) = f(z) + \langle \nu, A z - b \rangle + \langle \mu, C z - c \rangle.$$

For simplicity of the exposition we introduce further the following notations:

$$G = \begin{bmatrix} A \\ C \end{bmatrix} \text{ and } g = \begin{bmatrix} b \\ c \end{bmatrix}.$$  

Since $f_i$ are strongly convex functions, then $f$ is also strongly convex w.r.t. the Euclidean norm $\| \cdot \|$ on $\mathbb{R}^n$, with convexity parameter e.g. $\sigma_i = \min_{i=1,\ldots,M} \sigma_i$. Further, the dual function $d$ is differentiable and its gradient is given by the following expression [15]:

$$\nabla d(\nu, \mu) = G z(\nu, \mu) - g,$$

where $z(\nu, \mu)$ denotes the unique optimal solution of the inner problem (3), i.e.:

$$z(\nu, \mu) = \arg \min_{z \in \mathbb{R}^n} \mathcal{L}(z, \nu, \mu).$$

Moreover, the gradient $\nabla d$ of the dual function is Lipschitz continuous w.r.t. Euclidean norm $\| \cdot \|$, with constant [15]:

$$L_d = \frac{\|G\|^2}{\sigma_i}.$$

If we denote by $\nu_{\tilde{N}_i} = [\nu_{ji}]_{j \in \tilde{N}_i}$ and by $\mu_{\tilde{N}_i} = [\mu_{ji}]_{j \in \tilde{N}_i}$, we can observe that the dual function can be written in the following separable form:

$$d(\nu, \mu) = \sum_{i=1}^{M} d_i(\nu_{\tilde{N}_i}, \mu_{\tilde{N}_i}) = \langle \nu, b \rangle + \langle \mu, c \rangle,$$

with

$$d_i(\nu_{\tilde{N}_i}, \mu_{\tilde{N}_i}) = \min_{z_i \in \mathbb{R}^n} f_i(z_i) + \langle \nu, A_i z_i \rangle + \langle \mu, C_i z_i \rangle$$

$$= \min_{z_i \in \mathbb{R}^n} f_i(z_i) + \sum_{j \in \tilde{N}_i} \langle A_{ji} \nu_j + C_{ji} \mu_j, z_i \rangle.$$  

In these settings, we have that the gradient $\nabla d_i$ is:

$$\nabla d_i(\nu_{\tilde{N}_i}, \mu_{\tilde{N}_i}) = \left[ [A_{ji}]_{j \in \tilde{N}_i} \right] z_i(\nu_{\tilde{N}_i}, \mu_{\tilde{N}_i}),$$

where $z_i(\nu_{\tilde{N}_i}, \mu_{\tilde{N}_i})$ denotes the unique optimal solution in (6). Note that $\nabla d_i$ is Lipschitz continuous w.r.t. Euclidean norm $\| \cdot \|$, with constant [15]:

$$L_{d_i} = \left\| \left[ [A_{ji}]_{j \in \tilde{N}_i} \right] \right\|_{\sigma_i}.$$  

For simplicity of the exposition we will consider the notation $\lambda = [\nu^T \mu^T]^T$ and we will also denote the effective domain of the dual function by $\mathcal{D} = \mathbb{R}^p \times \mathbb{R}_+^q$. The following result, which is a distributed version of descent lemma is central in our derivations of a distributed dual algorithm and in the proofs of its convergence rate.

**Lemma 2.2** Let Assumption 2.1 (a) hold. Then, the following inequality is valid:

$$d(\lambda) \geq d(\tilde{\lambda}) + \langle \nabla d(\tilde{\lambda}), \lambda - \tilde{\lambda} \rangle - \frac{1}{2} \| \tilde{\lambda} - \lambda \|^2_W \quad \forall \lambda, \tilde{\lambda} \in \mathcal{D},$$

where $z(\nu, \mu)$ denotes the unique optimal solution of the inner problem (3), i.e.:

$$z(\nu, \mu) = \arg \min_{z \in \mathbb{R}^n} \mathcal{L}(z, \nu, \mu).$$

Moreover, the gradient $\nabla d$ of the dual function is Lipschitz continuous w.r.t. Euclidean norm $\| \cdot \|$, with constant [15]:

$$L_d = \frac{\|G\|^2}{\sigma_i}.$$
where the matrix \( W = \text{diag}(W_\nu, W_\mu) \) with the matrices \( W_\nu = \text{diag} \left( \sum_{i \in \mathcal{N}_j} L_d I_{p_j} ; j \in V_2 \right) \) and \( W_\mu = \text{diag} \left( \sum_{i \in \mathcal{N}_j} L_d I_{q_j} ; j \in V_2 \right) \).

**Proof.** Note that a similar result for the case of inequality constraints was given in [2,14]. The proof, when we have both equality and inequality constraints, follows similar lines. □

**Tightness of the descent lemma.** Our descent lemma (Lemma 2.2) is "tight" in the following sense: there are functions for which \( W \) in (8) cannot be replaced by smaller diagonal matrices in positive definite sense. We show this on a simple example. Let \( f_3(z_i) = \frac{1}{2} z_i^T z_i \) and \( n_i \) = 1. In this case \( d(\lambda) = -\frac{1}{2} \lambda^T (GG^{-1}G^T) \lambda \) and \( Q = \text{diag}(\sigma_i; i \in V_1) \). We note that we can write \( d(\lambda) = d(\lambda) + \nabla d(\lambda) \lambda - \frac{1}{2} \| \lambda - \bar{\lambda} \|_G^{-1}G \| \_2^2 \). Let us define \( \bar{\lambda} = \lambda = h \). We need to show that there exists a matrix \( G \) for which \( \max_{h \neq 0} \frac{\| h \|_G^{-1}G \|_2^2}{\| h \|_2^2} = 1 \). Since we know that the maximal value in the previous optimization problem is always smaller than 1 (otherwise (8) would not hold), we have to show that there exists matrix \( G \) and vector \( h \) for which:

\[
\| h \|_G^{-1}G \|_2^2 = \| h \|_2^2. \tag{9}
\]

Let us consider matrices \( G \) with \( 0/\sqrt{\sigma_i} \) entries and that have exactly \( \omega \) nonzeros on each row and on each column. If we let \( h \) be the vector with all entries equal to 1, then (9) holds.

We denote by \( \Lambda^* \) the set of optimal solutions of dual problem (2). According to [7, Theorem 2.3.2], if Assumption 2.1 holds for our original problem (1), then \( \Lambda^* \) is nonempty, convex and bounded. For any \( \lambda \in \mathbb{R}^{p+q} \), we can define the following finite quantity:

\[
\mathcal{R}(\lambda) = \min_{\lambda^* \in \Lambda^*} \| \lambda^* - \lambda \|_W. \tag{10}
\]

In this paper we propose a novel distributed dual gradient method for which we are interested in deriving estimates for both dual and primal suboptimality and also for primal feasibility violation, i.e. for a given accuracy \( \epsilon \) find a primal-dual pair \( (\hat{z}, \hat{\lambda}) \) such that:

\[
\| [G \hat{z} - g] \|_W^{-1} \leq \mathcal{O}(\epsilon), \quad \| \hat{z} - z^* \|_2^2 \leq \mathcal{O}(\epsilon), \quad -\mathcal{O}(\epsilon) \leq f(\hat{z}) - f^* \leq \mathcal{O}(\epsilon) \quad \text{and} \quad f^* - d(\hat{\lambda}) \leq \mathcal{O}(\epsilon). \tag{11}
\]

2.1 Motivation: Distributed MPC (DMPC) for networked systems

We consider a discrete time networked system, modeled by a directed graph \( \mathcal{G} = (V, E) \), for which the set \( V = \{1, \ldots, M\} \) represents the subsystems and the adjacency matrix \( E \) indicates the dynamic couplings between these subsystems. Note that in these settings, the graph \( \mathcal{G} \) is a particular case of the bipartite graph previously presented for which we have \( V_1 = V_2 = V \). The dynamics of the subsystems can be defined by the following linear state equations [3,16]:

\[
x_i(t + 1) = \sum_{j \in \mathcal{N}_i} A_{ij} x_j(t) + B_{ij} u_j(t), \quad \forall i \in V. \tag{12}
\]

where \( x_i(t) \in \mathbb{R}^{n_i} \) and \( u_i(t) \in \mathbb{R}^{n_u} \) represent the state and the input of the \( i \)th subsystem at time \( t \), \( A_{ij} \in \mathbb{R}^{n_i \times n_j} \) and \( B_{ij} \in \mathbb{R}^{n_u \times n_j} \). Note that in this case \( \mathcal{N}_j \) denotes the set of subsystems, including \( j \), whose dynamics directly affect the dynamics of subsystem \( j \) and \( \mathcal{N}_i \) represents the set of subsystems, including \( i \), whose dynamics are affected by the dynamics of subsystem \( i \). We also impose coupled state and input constraints:

\[
\sum_{j \in \mathcal{N}_i} \bar{C}_{ij} x_j(t) + \bar{C}_{ij} u_j(t) \leq c_i, \quad \forall i \in V, \quad t \geq 0. \tag{13}
\]

For a prediction horizon of length \( N \), we consider strongly convex stage and final costs for each subsystem:

\[
\sum_{t=0}^{N-1} \ell_i(x_i(t), u_i(t)) + \ell_f(x_i(N)), \quad \text{where the final costs} \quad \ell_f \quad \text{and the terminal set} \quad X_i^f \quad \text{are chosen such that} \quad \text{the control scheme ensures closed-loop stability} \quad [12].
\]

For the state and input trajectory of subsystem \( i \) and the overall state and input trajectory we use the notations:

\[
z_i = [u_i(0)^T x_i(1)^T \cdots u_i(N-1)^T x_i(N)^T]^T, \quad z = [z_1^T \cdots z_M^T]^T
\]

and for the total local cost over the prediction horizon

\[
f_i(z_i) = \sum_{t=0}^{N-1} \ell_i(x_i(t), u_i(t)) + \ell_f(x_i(N)).
\]

In these settings, the centralized MPC problem for the networked system (12), for a given initial state \( x = [x_1^T \cdots x_M^T]^T \), can be posed as the separable convex optimization problem (1), where \( n_i = N(n_{u_i} + n_{x_i}) \), the equality constraints \( A z = b \) are obtained by stacking all the dynamics (12) together, while the inequality constraints \( C z \leq c \) are obtained by writing the state and input constraints (13) in compact form, over the prediction horizon (see e.g. [18]). Note also that for the matrices \( A \) and \( C \), each block \( A_{ji} \) and \( C_{ji} \) is zero when \( E_{ji} = 0 \).
In the following sections, we analyze the structural properties of the dual problem (2) and then we propose a fully distributed dual gradient method for solving this problem which exploits the separability of the dual function and allow us to recover a suboptimal and nearly feasible solution for our original problem (1) in linear time.

3 Error bound property of the dual problem

In this section, under Assumption 2.1, we prove an error bound type property on the corresponding dual problem (2). For completeness, first we briefly review the existing results on error bound properties for a convex problem in the form:

\[
\min_{y \in D} \psi(y),
\]

where \(\psi(\cdot)\) is convex function, with Lipschitz continuous gradient, and \(D\) is a polyhedral set. We are interested in finding optimal points for this problem, i.e. points \(y\) satisfying \(y = [y - \nabla \psi(y)]_D\). Typically, in order to show linear convergence for gradient based methods used for solving the above problem, we need to require some nondegeneracy assumption on the problem (e.g. strong convexity) which does not hold for many practical applications (e.g. (DMPC), (NUM) or (DCOPF) problems). A new line of analysis, that circumvents these difficulties, was developed using the notion of error bound, which estimates the distance to the solution set from an \(y \in D\) by the norm of the proximal residual \(\nabla^+ \psi(y) = [y - \nabla \psi(y)]_D - y\). For objective functions of the form \(\psi(y) = \psi(G^T y)\), with \(\psi(\cdot)\) strongly convex function and \(G\) a general matrix, the authors in [10] show a local error bound property that holds in a neighborhood of the solution set, while in [28] the authors show a global error bound property provided that the set \(D\) is a bounded polyhedron or the entire space.

Our approach for proving a global error bound property for the dual problem (2) is in a way similar to the one in [10, 11, 28]. However, our results are more general in the sense that: we derive a global error bound property as opposed to the results in [10, 11] where the authors show this property only locally in a neighborhood of the solution set, and we allow the constraints set to be an unbounded polyhedron, as opposed to the results in [28] where the authors show an error bound property only for constraints defined in terms of bounded polyhedra. Also, our proximal residual introduced below is more general than the one used in the standard analysis of the error bound property (see e.g. [10, 11, 28]). Last but not least important is that our approach also works for dual problems, which allows us to prove for the first time a global error bound property for such problems.

For the convex function \(f\), we denote its conjugate [25]:

\[
\hat{f}(y) = \sum_{i=1}^{M} \hat{f}_i(y),
\]

where \(\hat{f}_i(y) = \max_{z_i \in \mathbb{R}^{n_i}} \langle y, z_i \rangle - f_i(z_i)\). According to Proposition 12.60 in [25], under the Assumption 2.1 (a) (in particular, under the assumption that \(f_i\) has Lipschitz gradient), each function \(\hat{f}_i(y)\) is strongly convex w.r.t. Euclidean norm, with constant \(\frac{1}{\theta_1}\), which implies that function \(\hat{f}\) is strongly convex w.r.t. the same norm, with constant:

\[
\sigma^2 = \sum_{i=1}^{M} \frac{1}{\theta_1}.
\]

Note that in these settings our dual function can be written as:

\[
d(\lambda) = -\hat{f}(\lambda^T G) - g^T \lambda. \quad (14)
\]

Recall that for the projection of \(z\) onto the set \(\mathbb{D}\) w.r.t. the norm \(\|\cdot\|_W\) we use \([z]_W\). We denote further the proximal residual:

\[
\nabla^+ d(\lambda) = [\lambda + W^{-1} \nabla d(\lambda)]_W - \lambda \quad \forall \lambda \in \mathbb{D}. \quad (15)
\]

The following lemma whose proof can be found e.g. in [14, Lemma 6.4] (see also [10, 28]) will help us prove the desired error bound property for our dual problem (2).

Lemma 3.1 Let Assumption 2.1 hold. Then, there exists a unique \(y^* \in \mathbb{R}^n\) such that:

\[
G^T \lambda^* = y^* \quad \forall \lambda^* \in \Lambda^*. \quad (16)
\]

Moreover, \(\nabla d(\lambda) = G \nabla \hat{f}(-y^*) - g\) is constant for all \(\lambda \in \Lambda\), where the set \(\Lambda = \{\lambda \in \mathbb{D} : G^T \lambda = y^*\}\).

The following theorem, which is one of the main results of the paper, establishes a global error bound like property for our dual problem (2):

Theorem 3.2 Let Assumption 2.1 hold. Then, there exists a constant \(\kappa\), depending on the data of problem (1) and the term \(T(\lambda) = \max_{\lambda^* \in \Lambda^*} \|\lambda - \lambda^*\|_W\), such that the following error bound property holds for dual problem (2):

\[
\|\lambda - \hat{\lambda}\|_W \leq \kappa \left(\frac{T(\lambda)}{T(\lambda) + \max_{\lambda^* \in \Lambda^*} \|\lambda - \lambda^*\|_W}\right) \hat{d}(\lambda) \quad \forall \lambda \in \mathbb{D}, \quad (17)
\]

where \(\lambda = [\lambda]_W^1\), and \(\kappa(T(\lambda))\) is given by \(\kappa(T(\lambda)) = \theta_1^2 \frac{4}{\theta_2^2} + 12 \theta_3^2 \frac{4}{\theta_2^2} \left(2 T^2(\lambda) + 2 \nabla d(\lambda)\right) \left(1 + 3 \theta_2^2 \frac{2}{\theta_1^2}\right), \) with \(\theta_1\) and \(\theta_2\) positive constants depending on problem data.

PROOF. Since the proof is involved and makes use of some technical results, for clarity of the exposition we present it in the Appendix. □

Based on Theorem 3.2 we will prove in the following section the linear rate of convergence of a distributed dual
gradient method. To our knowledge this is the first result showing global linear convergence rate on primal suboptimality and infeasibility for the last primal iterate of a dual gradient algorithm, as opposed e.g. to the results in [11] where only local linear convergence was derived for such an algorithm or results in [2, 6, 9, 15, 20, 22, 23] where sublinear convergence is proved.

4 Linear convergence for dual distributed gradient method under an error bound property

The existing convergence results from the literature on dual gradient methods either show sublinear rate of convergence [2, 6, 9, 15, 20, 22, 23] or at most local linear convergence [11]. In this section we show, that under the error bound property for the dual problem as proved in Theorem 3.2, which is valid for quite general assumptions (see Assumption 2.1), we have linear convergence [11]. In this section we show, that under the error bound property for the dual problem as proved in Theorem 3.2, which is valid for quite general assumptions (see Assumption 2.1), we have linear convergence [11]. In this section we show, that under the error bound property for the dual problem as proved in Theorem 3.2, which is valid for quite general assumptions (see Assumption 2.1), we have linear convergence [11]. In this section we show, that under the error bound property for the dual problem as proved in Theorem 3.2, which is valid for quite general assumptions (see Assumption 2.1), we have linear convergence [11]. In this section we show, that under the error bound property for the dual problem as proved in Theorem 3.2, which is valid for quite general assumptions (see Assumption 2.1), we have linear convergence [11]. In this section we show, that under the error bound property for the dual problem as proved in Theorem 3.2, which is valid for quite general assumptions (see Assumption 2.1), we have linear convergence [11].

The main difference between our new Algorithm (DG) and the algorithms proposed in literature [4, 6, 13, 15, 17, 20] consists in the way we update the sequence \(\lambda^k\). Instead of using a classical projected gradient step with a scalar centralized step size as in [4, 6, 13, 15, 17, 20], we update \(\lambda^k\) using a projected weighted gradient step which allows us to obtain a fully distributed scheme. The following relation, which is a generalization of a standard result for gradient methods shows that Algorithm (DG) is an ascent method [21]:

\[
d(\lambda^{k+1}) \geq d(\lambda^k) + \frac{1}{2}\|\lambda^k - \lambda^{k+1}\|_W^2 \quad \forall k \geq 0.
\]  

Using further Lemma 2.2 with \(\lambda = \lambda^1\) and \(\bar{\lambda} = \lambda^0\) we have:

\[
d(\lambda^1) \geq d(\lambda^0) + \langle \nabla d(\lambda^0), \lambda^1 - \lambda^0 \rangle - \frac{1}{2}\|\lambda^1 - \lambda^0\|_W^2
\]

\[
= \max_{\lambda \in \mathbb{D}} d(\lambda^0) + \langle \nabla d(\lambda^0), \lambda - \lambda^0 \rangle - \frac{1}{2}\|\lambda - \lambda^0\|_W^2
\]

\[
\geq \max_{\lambda \in \mathbb{D}} d(\lambda) - \frac{1}{2}\|\lambda - \lambda^0\|_W^2 \geq f^* - \frac{1}{2}R^2(\lambda^0),
\]

where the first equality follows from the definition of \(\lambda^1\) in Algorithm (DG), the second inequality from the concavity of function \(d\) and the last inequality from \(\lambda^* \subseteq \mathbb{D}\) and definition of max. Using now the previous relation we obtain:

\[
f^* - d(\lambda^1) \leq \frac{1}{2}R^2(\lambda^0).
\]

The next lemma will help us analyze the convergence of the Algorithm (DG):

**Lemma 4.1** Let Assumption 2.1 hold and the sequence \(\{\lambda^k\}_{k \geq 0}\) be generated by Algorithm (DG). Then, the following inequalities hold:

\[
\|\lambda^k - \lambda^*\|_W \leq \cdots \leq \|\lambda_0 - \lambda^*\|_W \quad \forall \lambda^* \in \Lambda^*, \ k \geq 0.
\]

**PROOF.** First we notice that the update \(\lambda^{k+1}\) can be also viewed as the unique optimal solution of the maximization of the following quadratic approximation of \(d\):

\[
\max_{\lambda \in \mathbb{D}} d(\lambda^k) + \langle \nabla d(\lambda^k), \lambda - \lambda^k \rangle - \frac{1}{2}\|\lambda - \lambda^k\|_W^2.
\]

Taking now \(\lambda = \lambda^*\) in the optimality condition of (21), we obtain the following inequality:

\[
\langle \nabla d(\lambda^k) - W(\lambda^{k+1} - \lambda^k), \lambda^* - \lambda^{k+1} \rangle \leq 0.
\]

Further, we can write:

\[
\|\lambda^{k+1} - \lambda^*\|_W^2 = \|\lambda^k - \lambda^*\|_W^2 + 2\langle \nabla d(\lambda^k), \lambda^{k+1} - \lambda^* \rangle - \frac{1}{2}\|\lambda^{k+1} - \lambda^k\|_W^2
\]

\[
\leq \|\lambda^k - \lambda^*\|_W^2 + 2\|\nabla d(\lambda^k)\|_W \|\lambda^* - \lambda^k\|_W + \frac{1}{2}\|\lambda^{k+1} - \lambda^k\|_W^2.
\]

where the first inequality follows from (22) and the second one is derived from the concavity of the function \(d\) and Lemma 2.2. □

Using now inequality (20) in (17) we obtain one important relation that estimates the distance from the dual optimal solution set of the sequence \(\lambda^k\) in terms of the norm of a proximal residual:

\[
\|\lambda^k - \bar{\lambda}^k\|_W \leq \pi \|\nabla^+ d(\lambda^k)\|_W \quad \forall k \geq 0,
\]

where from Theorem 3.2 we have:

\[
\pi = \theta^2_1 \frac{4}{\sigma_i} + 12\theta^3_2 (2T^2(\lambda^0) + 2T^2(\Lambda^*)) \left(1 + 3\theta^2_1 \frac{2}{\sigma_i}\right),
\]

where we have defined the positive constant \(T(\lambda^0) = \max_{\lambda^* \in \Lambda^*} \|\lambda^0 - \lambda^*\|_W\), which is finite since \(\lambda^*\) is a bounded
Combining now (24) with (25), we can write:

\[ \|\nabla^+ d(\lambda^k)\|_W = \|\lambda^{k+1} - \lambda^k\|_W. \]  

(25)

The following theorem provides an estimate on the dual suboptimality for Algorithm (DG) and follows similar lines as in [11, 14, 28]:

**Theorem 4.2** Let Assumption 2.1 hold and sequences \((z^k, \lambda^k)_{k \geq 0}\) be generated by Algorithm (DG). Then, an estimate on dual suboptimality for (2) is given by:

\[ f^* - d(\lambda^{k+1}) \leq \frac{1}{2} \left( \frac{4(1 + \pi)}{1 + 4(1 + \pi)} \right)^{k-1} R^2(\lambda^0). \]  

(27)

**Proof.** From the optimality conditions of problem (21) we have:

\[ \langle \nabla d(\lambda^k), \lambda^k - \lambda^{k+1} \rangle \leq \langle W(\lambda^{k+1} - \lambda^k), \lambda^k - \lambda^{k+1} \rangle \leq 0, \]  

(28)

where we recall that \(\lambda^k \in \Lambda^W\). Further, since the optimal value of the dual function is unique we can write:

\[ f^* - d(\lambda^{k+1}) = d(\lambda^k) - d(\lambda^{k+1}) \leq \langle \nabla d(\lambda^{k+1}), \lambda^k - \lambda^{k+1} \rangle = \langle \nabla d(\lambda^{k+1}) - \nabla d(\lambda^k), \lambda^k - \lambda^{k+1} \rangle + \langle \nabla d(\lambda^k), \lambda^k - \lambda^{k+1} \rangle \leq \|\nabla d(\lambda^{k+1}) - \nabla d(\lambda^k)\|_W \|\lambda^k - \lambda^{k+1}\|_W + \langle W(\lambda^{k+1} - \lambda^k), \lambda^k - \lambda^{k+1} \rangle \leq \|\lambda^{k+1} - \lambda^k\|_W \|\lambda^k - \lambda^{k+1}\|_W + \|\lambda^{k+1} - \lambda^k\|_W \|\lambda^k - \lambda^{k+1}\|_W \]

\[ = 2\|\lambda^{k+1} - \lambda^k\|_W \|\lambda^k - \lambda^{k+1}\|_W, \]  

(29)

where the second inequality follows from (28). Using now relation (26) we can write:

\[ \|\lambda^k - \lambda^{k+1}\|_W \leq \|\lambda^k - \lambda^k\|_W + \|\lambda^k - \lambda^{k+1}\|_W \leq (1 + \pi) \|\lambda^k - \lambda^{k+1}\|_W. \]

Introducing now the previous inequality in (29) and using (18) we have:

\[ f^* - d(\lambda^{k+1}) \leq 2 (1 + \pi) \|\lambda^k - \lambda^{k+1}\|_W \leq 4 (1 + \pi) \left( d(\lambda^{k+1}) - d(\lambda^k) \right). \]

Rearranging the terms in the previous inequality we obtain:

\[ f^* - d(\lambda^{k+1}) \leq \frac{4(1 + \pi)}{1 + 4(1 + \pi)} \left( f^* - d(\lambda^k) \right). \]  

(30)

Applying now (30) recursively and using (19) we obtain (27).

The following theorems give estimates on the primal feasibility violation and suboptimality for Algorithm (DG). Note that usually, for recovering an approximate primal solution from dual gradient based methods, we need to use averaging (see e.g. [4, 9, 15, 17, 20, 22]). In what follows, we do not consider averaging and we prove linear convergence for the last primal iterate, a result which appears to be new.

**Theorem 4.3** Under the assumptions of Theorem 4.2, the following estimate holds for the primal infeasibility:

\[ \|\| G z^k - g \|\|_{W^{-1}} \leq \left( \frac{4(1 + \pi)}{1 + 4(1 + \pi)} \right)^{k-2} R^2(\lambda^0). \]  

(31)

**Proof.** Using the descent property of dual gradient method (18) we have:

\[ \|\lambda^k - \lambda^{k+1}\|_W^2 \leq 2 \|\lambda^{k+1} - \lambda^k\|_W^2 \leq 2 \left( f^* - d(\lambda^k) \right) \leq \left( \frac{4(1 + \pi)}{1 + 4(1 + \pi)} \right)^{k-2} R^2(\lambda^0), \]

where in the last inequality we used Theorem 4.2. In order to prove the statement of the theorem we will first show that \(\|\| G z^k - g \|\|_{W^{-1}} \leq \|\lambda^k - \lambda^{k+1}\|_W^2\). We will prove this inequality componentwise. First, we recall that \(D = R^p \times R^p_i\). Thus, for all \(i = 1, \ldots, p\) we have:

\[ \|\| \nabla_i d(\lambda^k) \|\|_{W_{ii}^{-1}}^2 = \|\nabla_i d(\lambda^k)\|_{W_{ii}}^2 \]

\[ = \|\lambda^k_i - \lambda^{k+1}_i - W_{ii}^{-1} \nabla_i d(\lambda^k)\|_{W_{ii}}^2 = \|\lambda^k_i - \lambda^{k+1}_i\|_{W_{ii}}^2, \]

where in the last equality we used the definition of \(\lambda^{k+1}\). We now introduce the following disjoint sets: \(I_- = \{ i \in [p + 1, p + q] : \nabla_i d(\lambda^k) < 0 \}\) and \(I_+ = \{ i \in [p + 1, p + q] : \nabla_i d(\lambda^k) \geq 0 \}\). Using these notations and the definition of \(D\), we can write:

\[ \|\| \nabla_i d(\lambda^k) \|\|_{W_{ii}^{-1}}^2 = 0 \leq \|\lambda^k_i - \lambda^{k+1}_i\|_{W_{ii}}^2 \quad \forall i \in I_- \]  

(33)

On the other hand, for all \(i \in I_+\) we have:

\[ \|\| \nabla_i d(\lambda^k) \|\|_{W_{ii}^{-1}}^2 = \|\nabla_i d(\lambda^k)\|_{W_{ii}}^2 \]

\[ = \|W_{ii}^{-1} \nabla_i d(\lambda^k)\|_{W_{ii}}^2 = \|\lambda^k_i - \lambda^{k+1}_i\|_{W_{ii}}^2. \]  

(34)
Summing up the relations (32),(33) and (34) for all $i = 1, \ldots, p + q$ we obtain:

$$
\|\nabla d(\lambda^k)\|_W^2 \leq \|\lambda^k - \lambda^{k+1}\|_W^2.
$$

(35)

Combining now (35) and (31) we obtain:

$$
\|\nabla d(\lambda^k)\|_W^{-2} \leq \|\lambda^k - \lambda^{k+1}\|_W^2 \\
\leq \left(\frac{4(1 + \pi)}{1 + 4(1 + \pi)}\right)^{k-2} R^2(\lambda^0).
$$

Taking into account the definition of $\nabla d$ we conclude the statement. \qed

We now characterize the primal suboptimality and the distance from the last iterate $z^k$, generated by Algorithm (DG), to the optimal solution $z^*$. 

**Theorem 4.4** Under the assumptions of Theorem 4.3, the following estimate on primal suboptimality for problem (1) can be derived:

$$
\|z^k - z^*\| \leq \sqrt{\frac{1}{\sigma_f}} \left(\frac{4(1 + \pi)}{1 + 4(1 + \pi)}\right)^{k-2} R(\lambda^0).
$$

(37)

**Proof.** In order to prove the left-hand side inequality of (36) we can write:

$$
f^* = d(\lambda^*) = \min_{z \in \mathbb{R}^n} f(z) + \langle \lambda^*, Gz - g \rangle
$$

(38)

\leq f(z^k) + \langle \lambda^*, Gz^k - g \rangle

\leq f(z^k) + \langle \lambda^*, [Gz^k - g]_D \rangle

\leq f(z^k) + \|\lambda^*\|_W \|[Gz^k - g]_D\|_{W^{-1}}

\leq f(z^k) + (R(\lambda^0) + \|\lambda^0\|_W) \|[Gz^k - g]_D\|_{W^{-1}},

where the second inequality follows from the fact that $\lambda^* \in \mathbb{D}$ and the third one from the Cauchy-Schwartz inequality. Using now Theorem 4.3 we obtain the result.

For proving the right-hand side inequality of (36) we first show (37). First, let us note that since $f$ is $\sigma_f$-strongly convex w.r.t. Euclidean norm, it follows that $\mathcal{L}(z, \lambda^k)$ is also $\sigma_f$-strongly convex in the variable $z$ in the same norm. We recall that $d(\lambda^k) = f(z^k) + \langle \lambda^k, Gz^k - g \rangle$ and $\nabla d(\lambda^k) = Gz^k - g$. Taking now into account that $z^k = \arg\min_{z \in \mathbb{R}^n} \mathcal{L}(z, \lambda^k)$ and the fact that $\langle \lambda^k, \nabla d(\lambda^k) \rangle \leq 0$, from the strong convexity of $\mathcal{L}$ we have:

$$
\frac{\sigma_f}{2} \|z^k - z^*\|^2 \leq \mathcal{L}(z^*, \lambda^k) - \mathcal{L}(z^k, \lambda^k)
$$

$$
= f(z^*) + \langle \lambda^*, \nabla d(\lambda^*) \rangle - f(z^k) - \langle \lambda^k, \nabla d(\lambda^k) \rangle
$$

$$
= f^* + \langle \lambda^*, \nabla d(\lambda^*) \rangle - d(\lambda^k)
$$

$$
\leq f^* - d(\lambda^k).
$$

Using further Theorem 4.2 in the previous inequality we obtain (37). From the Lipschitz continuity property of $\nabla f$ we obtain:

$$
f(z^k) - f^* \leq \langle \nabla f(z^k), z^k - z^* \rangle + \max_i L_i \|z^k - z^*\|^2
$$

$$
= \langle -G^T \lambda^*, z^k - z^* \rangle + \max_i L_i \|z^k - z^*\|^2
$$

$$
\leq \|\lambda^*\|_W \|G\| \|z^k - z^*\| + \max_i L_i \|z^k - z^*\|^2
$$

where the first equality is deduced from the optimality conditions of problem $z^* = \arg\min_{z \in \mathbb{R}^n} f(z) + \langle \lambda^*, Gz - g \rangle$ and in the last inequality we used the Cauchy-Schwartz inequality and the fact that $\| \cdot \| \leq \frac{1}{\sigma_f} \| \cdot \|_W$ and $\|G^T \lambda^*\| \leq \|G\| \|\lambda^*\|$. Using now the definition of $\mathcal{R}(\lambda^0)$ and (37) we obtain the result. \qed

5 Distributed implementation

In this section we analyze the distributed implementation of Algorithm (DG). We look first at step 1 of the algorithm. According to (6), for all $i \in V_1$ we have:

$$
z_i^k = \arg\min_{z_i \in \mathbb{R}^n} f_i(z_i) + \langle \lambda_i^k, [A_i^T C_i^T]^T z_i \rangle
$$

$$
= \arg\min_{z_i \in \mathbb{R}^n} f_i(z_i) + \sum_{j \in N_i} ([A_{ij}^T C_{ij}^T]^T \lambda_j^k)^T z_i.
$$

(39)

Thus, in order to compute $z_i^k$ the algorithm requires only local information, namely $\{A_{ij}, C_{ij}, \lambda_j^k\}_{j \in N_i}$. Using now the definitions of $W$ and $\nabla d$, step 2 in Algorithm (DG)
can be written in the following form for all \( j \in V_2 \):

\[
\lambda_j^{k+1} = \left[ \lambda_j^k + \left[ W_{\nu,j}^{-1} \sum_{i \in N_j} A_{ji} \bar{z}_i^k + W_{\mu,j}^{-1} \sum_{i \in N_j} C_{ji} \bar{z}_i^k \right] \right],\]

where \( W_{\nu,j} \) and \( W_{\mu,j} \) denote the \( j \)th block-diagonal element of matrix \( W_\nu \) and \( W_\mu \), respectively. Taking into account the definitions of \( W_{\nu,j} \) and \( W_{\mu,j} \) we can conclude that in order to update the dual variable \( \lambda_j^{k+1} \) in step 2 of Algorithm (DG) we require only local information \( \left\{ L_{di}, A_{ji}, C_{ji}, \bar{z}_i^k \right\}_{i \in N_j} \).

We discuss further some implementation issues for the case of (DMPC) problems. A standard approach for such problems is to consider quadratic stage and final costs for each subsystem, i.e.:

\[
\ell_i(x_i(t), u_i(t)) = 0.5\|x_i(t)\|_Q^2 + 0.5\|u_i(t)\|_{\bar{R}_i}^2 \quad \text{and} \quad \ell_i^2(x_i(N)) = 0.5\|x_i(N)\|_{\bar{P}_i}^2,
\]

where \( Q_i, \bar{R}_i \) and \( \bar{P}_i \) are positive definite matrices of appropriate dimensions. In this case, each objective function is quadratic, i.e. \( f_i(z_i) = 0.5\|z_i\|_Q^2 \), where \( Q_i \) is given by:

\[
Q_i = \text{diag} \left( I_{N_i-1} \otimes \begin{bmatrix} \bar{R}_i & 0 \\ 0 & \bar{Q}_i \end{bmatrix} \right).
\]

Further, note that the matrices \( A \) and \( C \) are block-sparse having each block \((i, j)\) zero for every subsystem \( j \) which is not influenced by subsystem \( i \). According to (39), the update \( \bar{z}_i^k \) of subsystem \( i \) can be done in closed form:

\[
\bar{z}_i^k = Q_i^{-1} \sum_{j \in N_i} \left( [A_{ji}^T C_{ji}^T] \lambda_j^k \right),
\]

and requires only information from those subsystems that are influenced by subsystem \( i \), i.e. \( N_i \). Similarly, for updating \( \lambda_j^{k+1} \) corresponding to subsystem \( j \), we need information from its neighbors, i.e. \( N_j \) (those subsystems that affect subsystem \( j \)). Moreover, if we define the measure of sparsity for the incidence matrix \( E \) as:

\[
\omega = \max_{i,j \in V} \{ |N_i|, |N_j| \},
\]

then due to the structure of block matrices \( Q_i, A_{ji}, C_{ji} \), the computational complexity of one step of Algorithm (DG) for (DMPC) is linear in the number of subsystems \( M \) and the horizon length \( N \), i.e. \( O(MN) \), provided that \( \omega \ll M \). Finally, our algorithm is scalable in the sense that removing or adding a new node (subsystem) can be done immediately using only local information.

\section{Numerical simulations}

In this section we consider problems of the form (1), where the objective functions are given by:

\[
f_i(z_i) = 0.5\|z_i\|_Q^2 + q_i^T z_i + \gamma_i \log \left( 1 + e^{a_i z_i} \right).
\]

Note that this type of function satisfies Assumption 2.1 (a), provided that \( Q_i \) are positive definite matrices, and are intensively used in (DMPC) applications for \( \gamma_i = 0 \) or (NUM) applications for \( \gamma_i > 0 \). We first generate a sparse communication graph \( G = (V, E) \) characterized by an incidence matrix \( E \in \mathbb{R}^{M \times M} \) generated randomly with different degrees of sparsity given by \( N_i \) and \( N_j \).

Recall that we have defined the measure of sparsity of the incidence matrix \( E \) as:

\[
\omega = \max_{i,j \in V} \{ |N_i|, |N_j| \}.
\]

We take \( n_i = n_j \) for all \( i, j \). Matrices \( Q_i \in \mathbb{R}^{n_i \times n_i}, A_{ji} \in \mathbb{R}^{\frac{3n_i}{4} \times n_i} \) and \( C_{ji} \in \mathbb{R}^{\frac{2n_i}{4} \times n_i} \) are taken from a normal distribution with zero mean and unit variance. Matrix \( Q_i \) is then made positive definite by transformations \( Q_i \leftarrow Q_i^T Q_i + \sigma_i I_{n_i} \), where \( \sigma_i \) are randomly generated from the interval \([1, 10] \).

In order to analyze the behavior of Algorithm (DG) we first consider a problem with the number of subsystems \( M = 100 \), the dimension of local variables \( n_i = 10 \) for all \( i \in \{1, \ldots, 100\} \) and \( \omega = 15 \). We are interested in analyzing the evolution of both, primal and dual suboptimality, w.r.t. the number of iterations. We consider an accuracy \( \epsilon = 10^{-4} \) and impose a stopping criterion of the form:

\[
\frac{|f(z^k) - f^*|}{|f^*|} \leq \epsilon,
\]

where \( f^* \) is computed using CVX. We plot the results in logarithmic scale in Figure 1a. We can observe that both dual and primal suboptimality converge linearly, which confirms the theoretical results derived in Section 4.

We are also interested in comparing the performances of Algorithm (DG) with the ones of the centralized dual gradient method, called Algorithm (CG), where for updating the dual variable \( \lambda \) we use the centralized step (see also [15]):

\[
(CG) : \quad \lambda^{k+1} = [\lambda^k + L_{da}^{-1} \nabla d(\lambda^k)]_D.
\]

For this purpose, we consider a set of 10 problems with fixed dimension, \( M = 100 \) and \( n = 10 \), accuracy \( \epsilon = 10^{-2} \) and the measure of sparsity \( \omega \) ranging from 5 to 50. We plot in Figure 1b the ratio between the number of iterations performed by Algorithm (DG) (\( k_{DG} \)) and by Algorithm (CG) (\( k_{CG} \)), respectively. On the one hand, we observe that for small values of the sparsity...
measure $\omega$, Algorithm (DG) clearly outperforms Algorithm (CG). On the other hand, by increasing the sparsity measure $\omega$ we observe a reduction in the ratio between the number of iterations performed by the two algorithms.

Further, we consider problems of different dimensions, having the measure of sparsity $\omega = 0.15M$, and compare the Algorithms (DG) and (CG) in terms of number of iterations performed for obtaining a suboptimal solution with accuracy $\epsilon = 10^{-2}$. In Table 1 we present the average results obtained for 10 randomly generated problems for each dimension. We use the notation $\overline{w} = \lambda_{\max}(W)$ and we recall that $\overline{w} = \lambda_{\min}(W)$. We can observe that for all dimensions, Algorithm (DG) outperforms Algorithm (CG) due to the reduced level of sparsity $\omega$ and smaller values of the entries of weighted step size $W$ compared to the centralized one $L_d$.

7 Conclusions

In this paper we have proposed and analyzed a fully distributed dual gradient method for solving the Lagrangian dual of a primal separable convex optimization problem with linear constraints. Under the strong convexity and Lipschitz continuous gradient property assumptions of the primal objective function we have provided a global error bound for the dual problem. Using this property, we have proved global linear convergence rate for both primal and dual suboptimality and for primal feasibility violation for our distributed dual gradient algorithm. We have also discussed distributed implementation aspects for our method and provided several numerical simulations which confirm the theoretical results and the efficiency of our approach.
Appendix

In order to prove Theorem 3.2 we first need some technical results. First we recall from Lemma 3.1 that there exists a unique \( y^* \in \mathbb{R}^n \) such that:

\[
G^T \lambda^* = y^* \quad \forall \lambda^* \in \Lambda^*.
\]

Moreover, \( \nabla d(\lambda) = G \nabla \tilde{f}(\cdot - y^*) - g \) is constant for all \( \lambda \in \Lambda \), where we have defined the set:

\[
\Lambda = \{ \lambda \in \mathbb{D} : G^T \lambda = y^* \}.
\]

We introduce further the following notations:

\[
r = [\lambda]_\Lambda^W \text{ and } \bar{r} = [\bar{r}]_\Lambda^W, \quad \forall \lambda \in \mathbb{D}. \quad (42)
\]

Using now the notations (42) and \( \bar{\lambda} = [\lambda]_\Lambda^W \), we can write:

\[
\| \lambda - \bar{\lambda} \|^2_W \leq \| \lambda - \bar{r} \|^2_W \leq (\| \lambda - r \|_W + \| r - \bar{r} \|_W)^2 \\
\leq 2\| \lambda - r \|^2_W + 2\| r - \bar{r} \|^2_W, \quad \forall \lambda \in \mathbb{D}. \quad (43)
\]

In what follows we show how we can find upper bounds on \( \| \lambda - r \|_W \) and \( \| r - \bar{r} \|_W \) such that we will be able to establish the error bound property on the dual problem (2) given in Theorem 3.2.

Lemma 7.1 Let Assumption 2.1 hold and \( \nabla^+ d \) given in (15). Then, the following inequality holds for all \( \lambda, \omega \in \mathbb{D} \):

\[
\langle \nabla d(\omega) - \nabla d(\lambda), \lambda - \omega \rangle \leq 2\| \nabla^+ d(\lambda) - \nabla^+ d(\omega) \|_W \| \lambda - \omega \|_W.
\]

PROOF. First, let us recall that \([\lambda + W^{-1} \nabla d(\lambda)]_\mathbb{D}^W \) is the unique solution of the optimization problem:

\[
\min_{\xi \in \mathbb{D}} \| \xi - \lambda - W^{-1} \nabla d(\lambda) \|_W^2,
\]

for which the optimality conditions reads:

\[
\left\langle W \left[ (\xi + W^{-1} \nabla d(\lambda))_\mathbb{D}^W - (\lambda + W^{-1} \nabla d(\lambda))_\mathbb{D}^W \right], \xi - [\lambda + W^{-1} \nabla d(\lambda)]_\mathbb{D}^W \right\rangle \geq 0 \quad \forall \xi \in \mathbb{D}.
\]

Taking now \( \xi = [\omega + W^{-1} \nabla d(\omega)]_\mathbb{D}^W \) in the previous inequality, adding and subtracting both, \( \lambda \) and \( \omega \), in the right term of the scalar product and using the definition of \( \nabla^+ d \) we obtain:

\[
\langle W \left( \nabla^+ d(\lambda) - W^{-1} \nabla d(\lambda) \right), \nabla d(\lambda) + \lambda - \omega - \nabla^+ d(\omega) \rangle \leq 0,
\]

and the symmetry of matrix \( W \) leads to:

\[
\langle \nabla^+ d(\lambda) - W^{-1} \nabla d(\lambda), W (\lambda - \omega) + W \left( \nabla^+ d(\lambda) - \nabla^+ d(\omega) \right) \rangle \leq 0.
\]

Rearranging the terms in the previous inequality we get:

\[
-\langle \nabla d(\lambda), \lambda - \omega \rangle \\
\leq -\langle \nabla^+ d(\lambda), W (\lambda - \omega) \rangle + \langle \nabla d(\lambda), \nabla^+ d(\lambda) - \nabla^+ d(\omega) \rangle \\
- \langle \nabla^+ d(\lambda), W (\nabla^+ d(\lambda) - \nabla^+ d(\omega)) \rangle.
\]

Writing now the previous inequality with \( \lambda \) and \( \xi \) interchanged and summing them up we can write:

\[
\langle \nabla d(\xi) - \nabla d(\lambda), \lambda - \xi \rangle \\
\leq \langle \nabla^+ d(\xi) - \nabla^+ d(\lambda), W (\lambda - \xi) \rangle \\
+ \langle \nabla d(\lambda) - \nabla d(\xi), \nabla^+ d(\lambda) - \nabla^+ d(\xi) \rangle \\
- \| \nabla^+ d(\lambda) - \nabla^+ d(\xi) \|^2_W \\
\leq \| \nabla^+ d(\lambda) - \nabla^+ d(\xi) \|^2_W \| W (\lambda - \xi) \| W^{-1} \\
+ \| \nabla^+ d(\lambda) - \nabla^+ d(\xi) \|^2_W \| \nabla d(\lambda) - \nabla d(\xi) \| W^{-1} \\
\leq 2\| \nabla^+ d(\lambda) - \nabla^+ d(\xi) \|^2_W \| \lambda - \xi \|^2_W,
\]

which concludes the statement. \( \square \)

The next lemma gives us an upper bound on \( \| \lambda - r \|_W \):

Lemma 7.2 Under Assumption 2.1 there exists a constant \( \kappa_1 \) such that the following inequality holds:

\[
\| \lambda - r \|^2_W \leq \kappa_1 \| \nabla^+ d(\lambda) \|^2_W \| \lambda - \bar{\lambda} \|^2_W \quad \forall \lambda \in \mathbb{D}, \quad (45)
\]

where \( \kappa_1 = \frac{2}{\theta_1^2} \theta_1 \) and \( \theta_1 \) depends on the matrix \( G \).

PROOF. First, let us notice that we can write the set \( \Lambda \) explicitly as:

\[
\Lambda = \{ \lambda \in \mathbb{R}^{p+q} : F \lambda \leq 0, \quad G^T \lambda = y^* \}, \quad (46)
\]

where \( F = [I_q, I_q - I_p] \). Since \( \lambda \in \mathbb{D} \) this implies \( F \lambda \leq 0 \) and therefore, according to Theorem 2 in [24], we can write:

\[
\| \lambda - r \|_W \leq \theta_1 \| G^T \lambda - y^* \| \infty \leq \theta_1 \| G^T \lambda - y^* \|,
\]

where \( \theta_1 \) is finite and depends only on the matrix \( G \) and on the norms \( \| \cdot \|_W \) and \( \| \cdot \| \infty \). From the strong convexity property of \( \tilde{f} \) combined with the fact that \( G^T \bar{\lambda} = y^* \) we have:

\[
\sigma_1 \| G^T \lambda - y^* \|^2 \\
\leq \langle \nabla \tilde{f}(G^T \lambda) - \nabla \tilde{f}(G^T \bar{\lambda}), G^T \lambda - G^T \bar{\lambda} \rangle \\
= -\langle G \nabla \tilde{f}(G^T \lambda) + g, G \nabla \tilde{f}(G^T \bar{\lambda}) - g, \bar{\lambda} - \bar{\lambda} \rangle \\
= \langle \nabla d(\lambda) - \nabla d(\bar{\lambda}), \lambda - \bar{\lambda} \rangle \\
\leq 2\| \nabla^+ d(\lambda) - \nabla^+ d(\bar{\lambda}) \|^2_W \| \lambda - \bar{\lambda} \|^2_W \\
= 2\| \nabla^+ d(\lambda) \|^2_W \| \lambda - \bar{\lambda} \|^2_W,
\]
where the last inequality follows from Lemma 7.1 and the last equality follows from the fact that $\nabla^+ d(\bar{\lambda}) = 0$ for $\lambda \in \Lambda^*$. Combining now (47) with (48) we obtain the result. □

The next result establishes an upper bound on $\|r - \tilde{r}\|_W$:

**Lemma 7.3** Let Assumption 2.1 be satisfied. Then, the following inequality is valid:

$$\|r - \tilde{r}\|_W^2 \leq \kappa_2(T(\lambda))\|\nabla^+ d(\lambda)\|_W\|\lambda - \bar{\lambda}\|_W \quad \forall \lambda \in \mathbb{D}, \quad (49)$$

where $r, \tilde{r}$ are given in (42), $\bar{\lambda} = [\lambda]^W$, and

$$\kappa_2(T(\lambda)) = \max_{\lambda^* \in \Lambda^*} \|\lambda - \lambda^*\|_W \text{ and } \theta_2 \text{ being a constant depending on } C, \nabla d(\lambda) \text{ and } y^*.$$  

**PROOF.** Since $\Lambda^* \subseteq \Lambda \subseteq \mathbb{D}$ and $G^T y = y^*$ for all $\xi \in \Lambda$, the dual problem (2) has the same optimal solutions as the following linear program:

$$\arg \max_{\lambda \in \mathbb{D}} d(\lambda) = \arg \max_{\lambda \in \mathbb{D}} d(\lambda) = \arg \max_{\lambda \in \mathbb{D}} -\bar{f}(y^*) - \langle g, \lambda \rangle = \max_{\lambda \in \mathbb{D}} -\langle g, \lambda \rangle. \quad (50)$$

Further, let us recall that $\nabla d(\zeta) = G(\bar{\lambda} - y^*) - g$ for any $\zeta \in \Lambda$ and thus we have that $\langle \nabla d(\zeta), \xi \rangle = \langle \nabla \bar{f}(y^*), y^* \rangle - \langle g, \xi \rangle$ for all $\zeta, \xi \in \Lambda$. Therefore, we can write further for any $\zeta \in \Lambda$:

$$\arg \max_{\xi \in \Lambda} \langle \nabla d(\zeta), \xi \rangle = \arg \max_{\xi \in \Lambda} \langle \nabla \bar{f}(y^*) + y^*, y^* \rangle - \langle g, \xi \rangle = \arg \max_{\xi \in \Lambda} -\langle g, \xi \rangle. \quad (51)$$

Combining now (50) with (51), we can conclude that any solution of the dual problem (2) $\xi = [\xi]^W$, with $\xi \in \mathbb{E}$, is also a solution of linear program (51). Since $\Lambda^* \subseteq \Lambda$, then for any $[\lambda]^W \in \Lambda^*$ we have that $\nabla d(\bar{\lambda}) = A\bar{\lambda} - b = 0$ and $\nabla d(\lambda) = C\bar{\lambda} - c \leq 0$, and thus we also have that the maximum in (51) is finite for any $\zeta = \bar{\lambda} = [\lambda]^W \in \Lambda^*$. Thus problem (51) is solvable for any $\zeta = \bar{\lambda} = [\lambda]^W \in \Lambda^*$. Applying now Theorem 2 in [24] to the optimality conditions of problem (51) and its dual we obtain:

$$\|\xi - \bar{\xi}\|_W \leq \theta_2(\|\nabla d(\bar{\lambda})\|_W, \bar{\lambda} - \langle \nabla d(\bar{\lambda}), \xi \rangle) \quad \forall \xi \in \Lambda, \quad (52)$$

where $\theta_2$ is a constant depending only on the matrix $C$ and vectors $\nabla d(\bar{\lambda})$ and $y^*$ (see eq. (6) in [24] for details).

Using the previous relation we have:

$$\|\xi - \bar{\xi}\|_W \leq \theta_2(\|\nabla d(\bar{\lambda})\|_W, \bar{\lambda} - \langle \nabla d(\bar{\lambda}), \xi \rangle) = \theta_2(\|\nabla d(\bar{\lambda}), \xi \rangle, \xi - \bar{\xi}). \quad (53)$$

For any $\xi \in \Lambda$ the optimality conditions of the following projection problem $\min_{\omega \in \Lambda} \|\omega - \xi - W^{-1}\nabla d(\bar{\lambda})\|_W^2$ become:

$$\langle W \left(\xi + W^{-1}\nabla d(\bar{\lambda})\right) - \xi - W^{-1}\nabla d(\bar{\lambda}), \omega - \xi \rangle \leq 0, \quad \forall \omega \in \Lambda.$$  

Taking now $\omega = \xi = [\xi]^W$, and since $W$ is a symmetric matrix we obtain:

$$\langle \nabla d(\bar{\lambda}), \xi - \bar{\xi} \rangle \leq \langle \xi + W^{-1}\nabla d(\bar{\lambda})\rangle - \xi - W^{-1}\nabla d(\bar{\lambda}) \leq \langle \xi + W^{-1}\nabla d(\bar{\lambda})\rangle - \xi \leq \| W(\xi - \xi) \|_W^2 \leq \| W(\xi - \xi) \|_W^2.$$  

where in the last equality we used the definition of $\nabla^+ d$ and the fact that $\nabla d(\bar{\lambda}) = \nabla d(\bar{\lambda})$ for all $\xi \in \Lambda$ (see Lemma 3.1). Combining now the previous inequality with (53) and taking $\xi = r \in \Lambda$ we obtain:

$$\|r - \tilde{r}\|_W^2 \leq \bar{\theta}_2(\|r - \tilde{r}\|_W + \|\nabla d(\bar{\lambda})\|_W - 1)^2 + \|\nabla^+ d(r)\|_W^2. \quad (54)$$

Since $\tilde{r} = [r]^W$, and $\Lambda^* \subseteq \Lambda$ we also have $\tilde{r} = [\tilde{r}]_\Lambda$. Thus, using the nonexpansive property of the projection we can bound the term $\|r - \tilde{r}\|_W$ from above with a finite positive constant $T(\lambda) = \max_{\lambda^* \in \Lambda^*} \|\lambda - \lambda^*\|_W$:

$$\|r - \tilde{r}\|_W \leq \|\lambda - \tilde{r}\|_W \leq \max_{\lambda^* \in \Lambda^*} \|\lambda - \lambda^*\|_W = T(\lambda). \quad (55)$$

Note that $T(\lambda)$ is finite for any $\lambda \in \mathbb{D}$, provided that $\Lambda^*$ is a bounded set. Further, our goal is to find an upper bound for $\|\nabla^+ d(r)\|_W$ in terms of $\|\lambda - \bar{\lambda}\|_W$ and $\|\nabla^+ d(\bar{\lambda})\|_W$. To this purpose, let us first prove that $\nabla^+ d$ is Lipschitz continuous with constant $3$ w.r.t. the norm
\[ \| \cdot \|_W \cdot \|W \cdot \] For any \( \lambda, \tilde{\lambda} \in \mathcal{D} \) we can write:

\[
\| \nabla^+ d(\lambda) - \nabla^+ d(\tilde{\lambda}) \|_W \leq \| \lambda - \tilde{\lambda} \|_W \tag{56}
\]

\[
+ \left\| \left[ \lambda + W^{-1} \nabla d(\lambda) \right]_\mathcal{D} - \left[ \tilde{\lambda} + W^{-1} \nabla d(\tilde{\lambda}) \right]_\mathcal{D} \right\|_W \\
\leq \| \lambda - \tilde{\lambda} \|_W + \| \lambda + W^{-1} \nabla d(\lambda) - \tilde{\lambda} - W^{-1} \nabla d(\tilde{\lambda}) \|_W \\
\leq 2\| \lambda - \tilde{\lambda} \|_W + \| \nabla d(\lambda) - \nabla d(\tilde{\lambda}) \|_W \leq 3\| \lambda - \tilde{\lambda} \|_W.
\]

Using now (56) with \( \tilde{\lambda} = \lambda \) and taking into account that \( \nabla^+ d(\lambda) = 0 \), we have:

\[
\| \nabla^+ d(\lambda) \|_W = \| \nabla^+ d(\lambda) - \nabla^+ d(\tilde{\lambda}) \|_W \leq 3\| \lambda - \tilde{\lambda} \|_W.
\]

Using now again (56) and the previous inequality we get:

\[
\| \nabla^+ d(r) \|_W^2 \leq \left( \| \nabla^+ d(\lambda) \|_W + \| \nabla^+ d(r) - \nabla^+ d(\lambda) \|_W \right)^2 \\
\leq 2\| \nabla^+ d(\lambda) \|_W^2 + 2\| \nabla^+ d(r) \|_W^2 \\
\leq 6\| \nabla^+ d(\lambda) \|_W^2 + 18\| \lambda - \tilde{\lambda} \|_W^2 \\
\leq 6 \left[ 1 + 30\alpha^2 + \frac{2}{\alpha^2} \right] \| \nabla^+ d(\lambda) \|_W \| \lambda - \tilde{\lambda} \|_W, \tag{57}
\]

where in the last inequality we used (45). Introducing now (55) and (57) in (54) and using the inequality \( (\alpha + \beta)^2 \leq 2\alpha^2 + 2\beta^2 \) we obtain the result. \( \Box \)

**PROOF.** (Proof of Theorem 3.2) The result given in Theorem 3.2 follows immediately by using (45) from Lemma 7.2 and (49) from Lemma 7.3 in (43) and dividing both sides by \( \| \lambda - \tilde{\lambda} \|_W \). \( \Box \)

---

**References**

[1] A.G. Bakirtzis and P.N. Biskas. A decentralized solution to the dc-opf of interconnected power systems. *IEEE Transactions on Power Systems*, 18(3):1007–1013, 2003.

[2] A. Beck, A. Nedic, A. Ozdaglar, and M. Teboulle. Optimal distributed gradient methods for network resource allocation problems. Technical report, Technion, to appear in IEEE Transactions on Control of Network Systems, 2014.

[3] E. Camponogara and L.B. de Oliveira. Distributed optimization for model predictive control of linear-dynamic networks. *IEEE Transactions on Systems, Man and Cybernetics: Part A*, 39(6):1321–1338, 2009.

[4] M.D. Doan, T. Keviczky, and B. De Schutter. A distributed optimization-based approach for hierarchical mpc of large-scale systems with coupled dynamics and constraints. In *Proceedings of 50th Conference on Decision and Control*, pages 5236–5241, 2011.

[5] M. Farina and R. Scattolini. Distributed predictive control: a non-cooperative algorithm with neighbor-to-neighbor communication for linear systems. *Automatica*, 48(6):1088–1096, 2012.

[6] P. Giselsson, M. D. Doan, T. Keviczky, B. De Schutter, and A. Rantzer. Accelerated gradient methods and dual decomposition in distributed model predictive control. *Automatica*, 49:829–833, 2013.

[7] J.B. Hiriart-Urruty and C. Lemarechal. *Convex analysis and minimization algorithms: vol. I*. Springer-Verlag, 1996.

[8] M. Hong and Z.Q. Luo. On the linear convergence of the alternating direction method of multipliers. Technical Report arxiv preprint: 1208.3922, 2013.

[9] M. Kogel and R. Findeisen. Fast predictive control of linear systems combining nesterov’s gradient method and the method of multipliers. In *Proceedings of 50th Conference on Decision and Control*, pages 501–506, 2011.

[10] Z.Q. Luo and P. Tseng. Error bounds and convergence analysis of feasible descent methods: A general approach. *Annals of Operations Research*, 46–47(1):157–178, 1993.

[11] Z.Q. Luo and P. Tseng. On the convergence rate of dual ascent methods for linearly constrained convex minimization. *Mathematics of Operations Research*, 18(4):846–867, 1993.

[12] D.Q. Mayne, J.B. Rawlings, C.V. Rao, and P.O.M. Scokaert. Constrained model predictive control: Stability and optimality. *Automatica*, 36(6):789–814, 2000.

[13] M.C. Meinel, M. Ulbrich, and S. Albrecht. A class of distributed optimization methods with event-triggered communication. *Computational Optimization and Applications*, 57:517–553, 2014.

[14] I. Necoara and D. Clipici. Efficient distributed coordinate descent methods on smooth and error bound convex minimization. *SIAM Journal of Optimization*, submitted:1–40, 2013.

[15] I. Necoara and V. Nedelcu. Rate analysis of inexact dual first order methods: application to dual decomposition. *IEEE Transactions on Automatic Control*, 59(5):1232–1243, 2014.

[16] I. Necoara, V. Nedelcu, and I. Dumitrasche. Parallel and distributed optimization methods for estimation and control in networks. *Journal of Process Control*, 21(5):756–766, 2011.

[17] I. Necoara and J.A.K. Suykens. Application of a smoothing technique to decomposition in convex optimization. *IEEE Transactions on Automatic Control*, 53(11):2674–2679, 2008.

[18] I. Necoara and J.A.K. Suykens. An interior-point lagrangian decomposition method for separable convex optimization. *Journal of Optimization Theory and Applications*, 143(3):567–588, 2009.

[19] V. Nedelcu, I. Necoara, and D. Q. Quoc. Computational complexity of inexact gradient augmented lagrangian methods: application to constrained mpc. *SIAM Journal on Control and Optimization*, partially accepted:1–25, November 2012.

[20] A. Nedic and A. Ozdaglar. Approximate primal solutions and rate analysis for dual subgradient methods. *SIAM Journal on Optimization*, 19(4):1757–1780, 2009.

[21] Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Kluer, Boston, USA, 2004.

[22] P. Patrinos and A. Bemporad. An accelerated dual gradient-projection algorithm for embedded linear model predictive control. *IEEE Transactions on Automatic Control*, 59(1):18–33, 2014.

[23] S. Richter, M. Morari, and C.N. Jones. Towards computational complexity certification for constrained mpc. *Automatica*, 40, 2013.

[24] S. Richter, M. Morari, and C.N. Jones. Towards computational complexity certification for constrained mpc. *Automatica*, 40, 2013.

[25] S. Robinson. Bounds for error in the solution set of a perturbed linear program. *Linear Algebra and its Applications*, 6:69–81, 1973.

[26] R.T. Rockafellar and R.J. Wets. *Variational Analysis*. Springer-Verlag, New York, 1998.
[26] S. Sen and H.D. Sherali. A class of convergent primal-dual subgradient algorithms for decomposable convex programs. *Mathematical Programming*, 35(3):279–297, 1986.

[27] B. T. Stewart, A.N. Venkat, J.B. Rawlings, S.J. Wright, and G. Pannocchia. Cooperative distributed model predictive control. *Systems and Control Letters*, 59(8):460–469, 2010.

[28] P.W. Wang and C.J. Lin. Iteration complexity of feasible descent methods for convex optimization. *Journal of Machine Learning Research*, 15:1523–1548, 2014.

[29] E. Wei and A. Ozdaglar. On the o(1/k) convergence of asynchronous distributed alternating direction method of multipliers. Technical report, MIT, 2013.

[30] E. Wei, A. Ozdaglar, and A. Jadabaie. A distributed newton method for network utility maximization – part I and II. *IEEE Transactions on Automatic Control*, 58(9), 2013.