Application of a smoothing technique to decomposition in convex optimization

Ion Necoara and Johan A.K. Suykens

Abstract

Dual decomposition is a powerful technique for deriving decomposition schemes for convex optimization problems with specific structure. Although the Augmented Lagrangian is computationally more stable than the ordinary Lagrangian, the *prox-term* destroys the separability of the given problem. In this paper we use another approach to obtain a smooth Lagrangian, based on a smoothing technique developed by Nesterov, which preserves separability of the problem. With this approach we derive a new decomposition method, which from the viewpoint of efficiency estimates improves the bounds on the number of iterations of the classical dual gradient scheme by an order of magnitude. This can be achieved with the new decomposition algorithm since the resulting dual function has good smoothness properties and since we make use of the particular structure of the corresponding problem.

Index Terms

Smooth convex minimization, dual decomposition, parallel algorithms, distributed agents, coordination.

I. INTRODUCTION

There has been considerable recent interest in parallel and distributed computation methods for solving large-scale optimization problems (e.g. [1]). For separable convex problems, i.e. separable objective function but with coupling constraints (this type of problems arise in many fields of engineering: e.g. networks [2], [3], distributed model predictive control (MPC) [4], stochastic programming [5], etc), many researchers have proposed dual decomposition algorithms such as the dual subgradient method [1], [6], alternating direction method [1], [7]–[9], proximal method

The authors are with the Katholieke Universiteit Leuven, Department of Electrical Engineering, ESAT, Kasteelpark Arenberg 10, B–3001 Leuven (Heverlee), Belgium, tel/fax:+321632 0362/1970 (email:{ion.necoara,johan.suykens}@esat.kuleuven.be).

of multipliers [10], partial inverse method [11], [12], etc. In general, these methods are based on alternating minimization in a Gauss-Seidel fashion of an (Augmented) Lagrangian followed by a steepest ascent update for the multipliers. However, the prox-term used to obtain the smooth Lagrangian destroys the separability of the given problem. The step-size parameter which has a very strong influence on the convergence rate of these methods is also very difficult to tune and they do not provide any complexity estimates for the general case (linear convergence is obtained e.g. under strong convexity assumptions). Moreover, these methods use the steepest ascent update for the multipliers, while we know from [13] that this update is inferior with one order of magnitude compared to Nesterov's accelerated scheme.

The purpose of this paper is to propose a new decomposition method for separable convex optimization problems that overcomes the disadvantages mentioned above. Based on a smoothing technique recently developed by Nesterov in [14], we obtain a smooth Lagrangian that preserves separability of the problem. Using this smooth Lagrangian, we derive a new dual decomposition method in which the corresponding parameters are selected optimally and thus straightforward to tune. In contrast to the dual gradient update for the multipliers used by most of the decomposition methods from the literature, our method uses an optimal gradient-based scheme (see e.g. [13], [14]) for updating the multipliers. Therefore, we derive for the new method an efficiency estimate for the general case which improves with one order of magnitude the complexity of the classical dual gradient method (i.e. the steepest ascent update). Up to our knowledge these are the first efficiency estimate results of a dual decomposition method for separable *non-strongly* convex programs. The new algorithm is suitable for decomposition since it is highly parallelizable and thus it can be effectively implemented on parallel processors. This is a distinct feature of our method compared to alternating direction methods based on Gauss-Seidel iterations that obviously do not share this advantage.

This paper is organized as follows. Section II contains the problem formulation, followed by a brief introduction of some of the dual decomposition methods derived in the literature and the description of an accelerated scheme for smooth minimization developed by Nesterov in [13], [14]. The main results of the paper are presented in Section III, where we describe our new decomposition method and its main properties. We conclude with some applications and preliminary computational results on some test problems in Section IV.

II. PRELIMINARIES

A. Splitting methods for separable convex programs

An important application of convex duality theory is in decomposition algorithms for solving large-scale problems but with special structure. One such example, that we also consider in this paper, is the following separable convex program:

$$f^* = \min_{x\in X, z\in Z} \big\{\phi_1(x) + \phi_2(z) : \ Ax + Bz = b\big\}, \tag{1}$$

where $\phi_1 : \mathbb{R}^m \to \mathbb{R}$ and $\phi_2 : \mathbb{R}^p \to \mathbb{R}$ are continuous convex functions on X and Z, respectively, A is a given $n \times m$ matrix, B is a given $n \times p$ matrix, and b is a given vector in $\mathbb{R}^n$. In this paper we do *not* assume ϕ_1 and/or ϕ_2 to be strongly convex or smooth. Moreover, we assume that $X \subseteq \mathbb{R}^m$ and $Z \subseteq \mathbb{R}^p$ are given compact convex sets. We also use different norms on $\mathbb{R}^n, \mathbb{R}^m$ and $\mathbb{R}^p$, not necessarily the corresponding Euclidian norms. However, for simplicity in notation we do not use indices to specify the norms on $\mathbb{R}^n, \mathbb{R}^m$ and $\mathbb{R}^p$, since from the context it will be clear in which Euclidian space we are and which norm we use.

***Remark** 2.1:* This type of problems (1) arises e.g. in the context of large-scale networks consisting of multiple agents with different objectives or in control (see Section IV). We can also have any number of agents with different ϕ_i's, not necessarily two agents. Moreover, the method developed in this paper can handle both coupling equalities ($Ax + Bz{=}b$) and/or inequalities ($Ax + Bz \leq b$). However, for simplicity of the exposition we restrict ourselves to (1).

Let $\langle \cdot, \cdot \rangle$ denote the scalar product on the Euclidian space $\mathbb{R}^n$. By forming the Lagrangian corresponding to the linear constraints (with the Lagrange multipliers $\lambda \in \mathbb{R}^n$), i.e. $\mathcal{L}_0(x, z, \lambda) = \phi_1(x) + \phi_2(z) + \langle \lambda, Ax + Bz - b\rangle$, and using the dual decomposition method, one arrives at the following decomposition algorithm:

***Algorithm** 2.2:* ([1], [6]) **for** $k \geq 0$ **do**

1. given λ^k, minimize the Lagrangian $(x^{k+1}, z^{k+1}) = \arg\min_{x\in X, z\in Z} \mathcal{L}_0(x, z, \lambda^k)$, or equivalently minimize over x and z independently

$$x^{k+1} = \arg\min_{x\in X}[\phi_1(x) + \langle \lambda^k, Ax\rangle], \quad z^{k+1} = \arg\min_{z\in Z}[\phi_2(z) + \langle \lambda^k, Bz\rangle]$$

2. update the multipliers: $\lambda^{k+1} = \lambda^k + c_k(Ax^{k+1} + Bz^{k+1} - b)$,

where c_k is a positive step-size. The following assumption is valid throughout the paper:

***Assumption** 2.3:* The set of optimal Lagrange multipliers Λ^* is nonempty for problem (1).

It is known that Algorithm 2.2 is convergent under Assumption 2.3 and the assumption that both ϕ_1 and ϕ_2 are strongly convex functions (the later guarantees that the minimizer (x^{k+1}, z^{k+1}) is unique). In fact, under the assumption of strong convexity, the dual function

$$f_0(\lambda) = \min_{x \in X, z \in Z} \phi_1(x) + \phi_2(z) + \langle \lambda, Ax + Bz - b \rangle$$

is differentiable [1], [15], and thus Algorithm 2.2 can be seen as the gradient method with step-size c_k for maximizing the dual function.

However, for many interesting problems, especially arising from transformations that leads to decomposition (see Section IV), the functions ϕ_1 and ϕ_2 are not strongly convex. There are some methods (e.g. alternating direction method [1], [7], [8], proximal point method [10], partial inverse method [11]) that overcome this difficulty based on alternating minimization in a Gauss-Seidel fashion of the *Augmented Lagrangian*, followed by a steepest ascent update of the multipliers. A computational drawback of these schemes is that the *prox-term* $\frac{c}{2}\|Ax + Bz - b\|^2$, using the Euclidian norm framework, present in the Augmented Lagrangian is not separable in x and z. Another disadvantage is that they cannot deal with coupling *inequalities* in general. Moreover, these schemes were shown to be very sensitive to the value of the parameter c, with difficulties in practice to obtain the best convergence rate. Some heuristics for choosing c can be found in the literature [7], [8], [10]. But, these heuristics have not been formally analyzed from the viewpoint of efficiency estimates for the general non-smooth case (linear convergence results were obtained e.g. for strongly convex functions). Note that alternating direction method variants which allow for inexact minimization were proposed in [9], [10]. A closely related method is the partial inverse of a monotone operator developed in [11], [12].

B. Accelerated scheme for smooth maximization

In this section we briefly describe an accelerated scheme that also uses only first-order information for smooth convex functions developed by Nesterov in [13], [14]. Let f be a concave and differentiable function on a closed convex set $\hat{Q} \subseteq \mathbb{R}^n$. We further assume that the gradient of this function is Lipschitz continuous: $\|\nabla f(x) - \nabla f(y)\|_* \leq L\|x - y\| \quad \forall x, y \in \hat{Q}$, where $\|s\|_* = \max_{\|x\| \leq 1} \langle s, x \rangle$ is the corresponding dual norm of the norm used on $\mathbb{R}^n$ [15], [16].

***Definition** 2.4:* [14] We define a *prox-function* d of the set $\hat{Q}$ as a function with the following properties: d is continuous, strongly convex on $\hat{Q}$ with convexity parameter σ, and u^0 is the center of the set $\hat{Q}$, i.e. $u^0 = \arg\min_{x \in \hat{Q}} d(x)$ such that $d(u^0) = 0$.

The goal is to find an approximate solution to the smooth convex problem $x^* = \arg\max_{x\in\hat{Q}} f(x)$. In Nesterov's scheme three sequences of points from $\hat{Q}$ are updated recursively: $\{u^k\}_{k\geq 0}, \{x^k\}_{k\geq 0}$, and $\{v^k\}_{k\geq 0}$. The algorithm can be described as follows:

***Algorithm** 2.5:* ([14]) **for** $k \geq 0$ **do**

1. compute $f(u^k)$ and $\nabla f(u^k)$

2. find $\bar{x}^k = \arg\max_{x\in\hat{Q}}[f(u^k) + \langle \nabla f(u^k), x - u^k\rangle - \frac{L}{2}\|x - u^k\|^2]$ and define $x^k = \arg\max_{x\in\{\bar{x}^k, x^{k-1}, u^k\}} f(x)$

3. find $v^k = \arg\max_{x\in\hat{Q}} -\frac{L}{\sigma}d(x) + \sum_{l=0}^{k} \frac{l+1}{2}[f(u^l) + \langle \nabla f(u^l), x - u^l\rangle]$

4. set $u^{k+1} = \frac{k+1}{k+3}x^k + \frac{2}{k+3}v^k$.

The derivation of Algorithm 2.5 is based on the notion of *estimate sequence*. The main property of the estimate sequence corresponding to Algorithm 2.5 is the following relation [14]:

$$\frac{(k+1)(k+2)}{4} f(x^k) \geq \max_{x\in\hat{Q}} -\frac{L}{\sigma}d(x) + \sum_{l=0}^{k} \frac{l+1}{2}[f(u^l) + \langle \nabla f(u^l), x - u^l\rangle]. \tag{2}$$

The convergence properties of Algorithm 2.5 are summarized in the following theorem:

***Theorem** 2.6:* [14] Let sequence $\{x^k\}_{k\geq 0}$ be generated by Algorithm 2.5. Then, $\{f(x^k)\}_{k\geq 0}$ is nondecreasing and we have the following efficiency estimate: $f(x^*) - f(x^k) \leq \frac{4Ld(x^*)}{\sigma(k+1)(k+2)}$.

Theorem 2.6 tells us that from the viewpoint of efficiency estimates Nesterov's method applied to maximization of a concave function with Lipschitz continuous gradient has the order $\mathcal{O}\big(\sqrt{\frac{L}{\epsilon}}\big)$. Therefore, the efficiency of the method is higher by an order of magnitude than the corresponding pure gradient method (steepest ascent update with complexity $\mathcal{O}\big(\frac{1}{\epsilon}\big)$) for the same smooth problem (see [13]). Note that we can define directly $x^k = \bar{x}^k$ in step 2. The conclusions of Theorem 2.6 remain the same except that the sequence $\{f(x^k)\}_{k\geq 0}$ is not necessarily monotone.

III. A NEW DECOMPOSITION METHOD BASED ON SMOOTHING THE LAGRANGIAN

In this section we propose a new method to smoothen the Lagrangian of (1), inspired from [14]. This smoothing technique preserves the separability of the problem and moreover the corresponding parameters are easy to tune. Since separability is preserved under this smoothing technique, we derive a new dual decomposition method in which the multipliers are updated according to Algorithm 2.5. Moreover, we obtain an efficiency estimate of the new method for the general case. Note that with our method we can treat both coupling equalities $Ax + Bz = b$ and/or inequalities $Ax + Bz \leq b$ (see also Remark 3.8).

A. Smoothing the Lagrangian

Let d_X and d_Z be two prox-functions for the compact convex sets X and Z, with convexity parameter σ_X and σ_Z, respectively. Denote $x^0 = \arg\min_{x\in X} d_X(x), \quad z^0 = \arg\min_{z\in Z} d_Z(z)$. Since X and Z are compact and d_X and d_Z are continuous, we can choose finite and positive constants $D_X \geq \max_{x\in X} d_X(x)$ and $D_Z \geq \max_{z\in Z} d_Z(z)$. We also introduce the following notation $\|A\| = \max_{\|\lambda\|=1,\|x\|=1}\langle\lambda, Ax\rangle$. Since the linear operator A is defined as $A : \mathbb{R}^m \to \mathbb{R}^n_*$, where $\mathbb{R}^n_*$ is the dual of $\mathbb{R}^n$ (in fact $\mathbb{R}^n_* = \mathbb{R}^n$), we have

$$\|A\| = \max_{\|x\|=1} \|Ax\|_* \text{ and } \|Ax\|_* \leq \|A\|\|x\| \ \forall x.$$

Similarly for B. Let us introduce the following family of functions:

$$f_c(\lambda) = \min_{x\in X, z\in Z} \phi_1(x) + \phi_2(z) + \langle\lambda, Ax + Bz - b\rangle + c\big(d_X(x) + d_Z(z)\big), \tag{3}$$

where c is a positive *smoothness parameter* that will be defined later in this section. Note that we could also choose different parameters c_1 and c_2 for each prox-term. The generalization is straightforward. It is clear that the objective function in (3) is separable in x and z, i.e.

$$f_c(\lambda) = -\langle\lambda, b\rangle + \min_{x\in X}[\phi_1(x) + \langle\lambda, Ax\rangle + c\ d_X(x)] + \min_{z\in Z}[\phi_2(z) + \langle\lambda, Bz\rangle + c\ d_Z(z)]. \tag{4}$$

Denote by $x(\lambda)$ and $z(\lambda)$ the optimal solution of the minimization problem in x and z, respectively. Function f_c has the following smoothness properties:

***Theorem** 3.1:* The function f_c is concave and continuously differentiable at any $\lambda \in \mathbb{R}^n$. Moreover, its gradient $\nabla f_c(\lambda) = Ax(\lambda) + Bz(\lambda) - b$ is Lipschitz continuous with Lipschitz constant $L_c = \frac{\|A\|^2}{c\sigma_X} + \frac{\|B\|^2}{c\sigma_Z}$. The following inequalities also hold:

$$f_c(\lambda) \geq f_0(\lambda) \geq f_c(\lambda) - c(D_X + D_Z) \ \ \forall\lambda \in \mathbb{R}^n. \tag{5}$$

Proof: Since the functions d_X and d_Z are strongly convex, it follows that the optimal solution $(x(\lambda), z(\lambda))$ of (3) is unique for any λ and thus the function f_c is well defined at any λ. Concavity and continuous differentiability of f_c follows from standard duality theory [1], [15]. It remains to show that its gradient $\nabla f_c(\lambda) = Ax(\lambda) + Bz(\lambda) - b$ is Lipschitz continuous. For simplicity of notation we assume that all the functions involved in the minimization problem (3) are differentiable. Let λ and η be two Lagrange multipliers. Using first-order optimality

conditions for the minimization problem in x (see e.g. [13]) we obtain:

$$\langle \nabla\phi_1(x(\lambda)) + A^T\lambda + c\nabla d_X(x(\lambda), x(\eta) - x(\lambda)\rangle \geq 0$$
$$\langle \nabla\phi_1(x(\eta)) + A^T\eta + c\nabla d_U(x(\eta)), x(\lambda) - x(\eta)\rangle \geq 0.$$

Adding these two inequalities and since ϕ_1 is convex and d_X is strongly convex, we obtain

$$\langle A^T(\eta - \lambda), x(\lambda) - x(\eta)\rangle \geq \langle \nabla\phi_1(x(\lambda)) - \nabla\phi_1(x(\eta)), x(\lambda) - x(\eta)\rangle +$$
$$c\langle \nabla d_X(x(\lambda)) - \nabla d_X(x(\eta)), x(\lambda) - x(\eta)\rangle \geq c\sigma_X\|x(\lambda) - x(\eta)\|^2.$$

In view of the last relation and the Cauchy-Schwartz inequality we have:

$$\|Ax(\lambda) - Ax(\eta)\|_*^2 \leq \|A\|^2\|x(\lambda) - x(\eta)\|^2 \leq \frac{\|A\|^2}{c\sigma_X}\langle A^T(\eta - \lambda), x(\lambda) - x(\eta)\rangle \leq$$
$$\frac{\|A\|^2}{c\sigma_X}\|\lambda - \eta\|\|Ax(\lambda) - Ax(\eta)\|_*,$$

and thus $\|Ax(\lambda) - Ax(\eta)\|_* \leq \frac{\|A\|^2}{c\sigma_X}\|\lambda - \eta\|$. Similarly, for the minimization problem in z we obtain: $\|Bz(\lambda) - Bz(\eta)\|_* \leq \frac{\|B\|^2}{c\sigma_Z}\|\lambda - \eta\|$. In conclusion, the gradient of f_c satisfies

$$\|\nabla f_c(\lambda) - \nabla f_c(\eta)\|_* \leq \|Ax(\lambda) - Ax(\eta)\|_* + \|Bz(\lambda) - Bz(\eta)\|_* \leq \big(\frac{\|A\|^2}{c\sigma_X} + \frac{\|B\|^2}{c\sigma_Z}\big)\|\lambda - \eta\|.$$

Furthermore, the first inequality in (5) is a consequence of the fact that $d_X(x) \geq 0$ for all x, and $d_Z(z) \geq 0$ for all z. The second inequality in (5) follows from: $f_c(\lambda) \leq \min_{x\in X, z\in Z}[\phi_1(x) + \phi_2(z) + \langle\lambda, Ax + Bz - b\rangle] + c\ \max_{x\in X, z\in Z}[d_X(x) + d_Z(z)]$. □

B. A proximal center–based decomposition method

In this section we derive a new decomposition method based on the smoothing technique described in Section III-A. The new algorithm, called here the *proximal center algorithm*, has the nice feature that the *coordination* between the agents involves the maximization of a smooth convex objective function (i.e. with Lipschitz continuous gradient). Moreover, the *resource allocation* stage consists in solving in *parallel* two independent minimization problems with strongly convex objectives. The new method belongs to the class of two-level algorithms [17] and is particularly suitable for separable convex problems where the minimizations over x and z in (4) are easily carried out.

Let us apply Nesterov's accelerated method described in Algorithm 2.5 to the concave function f_c that has a Lipschitz continuous gradient:

$$\max_{\lambda \in Q} f_c(\lambda), \tag{6}$$

where Q is a given closed convex set in $\mathbb{R}^n$ that contains at least one optimal multiplier $\lambda^* \in \Lambda^*$. Notice that $Q \subseteq \mathbb{R}^n$ for linear equalities (i.e. $Ax + Bz - b = 0$), $Q \subseteq \mathbb{R}^n_+$, where $\mathbb{R}_+$ denotes the set of nonnegative real numbers, for linear inequalities (i.e. $Ax+Bz-b \leq 0$), or $Q \subseteq (\mathbb{R}^{n_1} \times \mathbb{R}^{n_2}_+)$, where $n_1 + n_2 = n$, when both, equalities and inequalities are present. Note that according to Algorithm 2.5 we also need to choose a prox-function d_Q for the set Q with the convexity parameter σ_Q and center u^0. The *proximal center algorithm* can be described as follows:

***Algorithm** 3.2:* **for** $k \geq 0$ **do**

1. given u^k compute in *parallel*

$$x^{k+1} = \arg\min_{x \in X}[\phi_1(x) + \langle u^k, Ax\rangle + c\, d_X(x)],\ z^{k+1} = \arg\min_{z \in Z}[\phi_2(z) + \langle u^k, Bz\rangle + c\, d_Z(z)].$$

2. compute $f_c(u^k) = \mathcal{L}_0(x^{k+1}, z^{k+1}, u^k) + c\big(d_X(x^{k+1}) + d_Z(z^{k+1})\big)$, $\nabla f_c(u^k) = Ax^{k+1} + Bz^{k+1} - b$

3. find $\bar{\lambda}^k = \arg\max_{\lambda \in Q} f_c(u^k) + \langle \nabla f_c(u^k), \lambda - u^k\rangle - \frac{L_c}{2}\|\lambda - u^k\|^2$ and define $\lambda^k = \arg\max_{\lambda \in \{\bar{\lambda}^k, \lambda^{k-1}, u^k\}} f_c(\lambda)$

4. find $v^k = \arg\max_{\lambda \in Q} -\frac{L_c}{\sigma_Q} d_Q(\lambda) + \sum_{l=0}^{k} \frac{l+1}{2}[f_c(u^l) + \langle \nabla f_c(u^l), \lambda - u^l\rangle]$

5. set $u^{k+1} = \frac{k+1}{k+3}\lambda^k + \frac{2}{k+3} v^k$.

The proximal center algorithm is suitable for decomposition since it is highly parallelizable: the agents can solve in parallel their corresponding minimization problems. We now derive a lower bound for the value of the objective function which will be used frequently in the sequel:

***Lemma** 3.3:* For any $\lambda^* \in \Lambda^*$ and $\hat{x} \in X, \hat{z} \in Z$, the following lower bound on primal gap holds: $[\phi_1(\hat{x}) + \phi_2(\hat{z})] - f^* \geq \langle A\hat{x} + B\hat{z} - b, \lambda^*\rangle \geq -\|\lambda^*\|_* \|A\hat{x} + B\hat{z} - b\|_*$.

Proof: From our assumptions we have that

$$f^* = \min_{x \in X, z \in Z}[\phi_1(x) + \phi_2(z) - \langle Ax + Bz - b, \lambda^*\rangle] \leq \phi_1(\hat{x}) + \phi_2(\hat{z}) - \langle A\hat{x} + B\hat{z} - b, \lambda^*\rangle,$$

thus the first inequality is proved. The second one follows from Cauchy-Schwarz inequality. □

The previous lemma shows that if $\|A\hat{x} + B\hat{z} - b\|_* \leq \epsilon_c$, then the primal gap is bounded by:

$$-\epsilon_c \|\lambda^*\|_* \leq \phi_1(\hat{x}) + \phi_2(\hat{z})] - f^* \leq \phi_1(\hat{x}) + \phi_2(\hat{z}) - f_0(\hat{\lambda}) \quad \forall \hat{\lambda} \in Q. \tag{7}$$

Therefore, if we are able to derive an upper bound ϵ for the duality gap and ϵ_c for the coupling constraints for some given $\hat{\lambda}$ and $\hat{x} \in X, \hat{z} \in Z$, then we conclude that $(\hat{x}, \hat{z})$ is an (ϵ, ϵ_c)-solution for problem (1) (since in this case $-\epsilon_c \|\lambda^*\|_* \leq \phi_1(\hat{x}) + \phi_2(\hat{z}) - f^* \leq \epsilon$ for all $\lambda^* \in \Lambda^*$). The next theorem derives an upper bound on the duality gap for our method.

***Theorem** 3.4:* Assume that there exists a closed convex set Q that contains a $\lambda^* \in \Lambda^*$. Then, after k iterations we obtain an approximate solution to the problem (1) $(\hat{x}, \hat{z}) = \sum_{l=0}^{k} \frac{2(l+1)}{(k+1)(k+2)}(x^{l+1}, z^{l+1})$ and $\hat{\lambda} = \lambda^k$ which satisfy the following duality gap:

$$[\phi_1(\hat{x}) + \phi_2(\hat{z})] - f_0(\hat{\lambda}) \leq c(D_X + D_Z) - \max_{\lambda \in Q} \big[-\frac{4L_c}{\sigma_Q (k+1)^2} d_Q(\lambda) + \langle A\hat{x} + B\hat{z} - b, \lambda \rangle \big]. \quad (8)$$

Proof: For an arbitrary c, we have from the inequality (2) that after k iterations the Lagrange multiplier $\hat{\lambda}$ satisfies the following relation:

$$\frac{(k+1)(k+2)}{4} f_c(\hat{\lambda}) \geq \max_{\lambda \in Q} -\frac{L_c}{\sigma_Q} d_Q(\lambda) + \sum_{l=0}^{k} \frac{l+1}{2} [f_c(u^l) + \langle \nabla f_c(u^l), \lambda - u^l \rangle].$$

In view of the previous inequality we have:

$$f_c(\hat{\lambda}) \geq \max_{\lambda \in Q} -\frac{4L_c}{\sigma_Q (k+1)^2} d_Q(\lambda) + \sum_{l=0}^{k} \frac{2(l+1)}{(k+1)(k+2)} [f_c(u^l) + \langle \nabla f_c(u^l), \lambda - u^l \rangle].$$

Now, we replace $f_c(u^l)$ and $\nabla f_c(u^l)$ with the expressions given in step 2 of Algorithm 3.2:

$$\sum_{l=0}^{k} \frac{2(l+1)}{(k+1)(k+2)} [f_c(u^l) + \langle \nabla f_c(u^l), \lambda - u^l \rangle] =$$

$$\sum_{l=0}^{k} \frac{2(l+1)}{(k+1)(k+2)} [\phi_1(x^{l+1}) + \phi_2(z^{l+1}) + c\big(d_X(x^{l+1}) + d_Z(z^{l+1})\big) + \langle Ax^{l+1} + Bz^{l+1} - b, \lambda \rangle] \geq$$

$$\sum_{l=0}^{k} \frac{2(l+1)}{(k+1)(k+2)} [\phi_1(x^{l+1}) + \phi_2(z^{l+1}) + \langle Ax^{l+1} + Bz^{l+1} - b, \lambda \rangle] \geq$$

$$\phi_1(\hat{x}) + \phi_2(\hat{z}) + \langle A\hat{x} + B\hat{z} - b, \lambda \rangle, \quad \forall \lambda \in Q.$$

The first inequality follows from the definition of a prox-function (see Definition 2.4) and the second inequality follows from the fact that ϕ_1 and ϕ_2 are convex functions. Using the last relation and (5) we derive the bound (8) on the duality gap. □

We now show how to construct the set Q and how to choose optimally the smoothness parameter c. In the next two sections we discuss two cases depending on the choices for Q and d_Q.

C. Efficiency estimates for compact Q

Let D_Q be a positive constant satisfying

$$\max_{\lambda \in Q} d_Q(\lambda) \leq D_Q. \tag{9}$$

Let us note that we can choose D_Q finite whenever Q is compact. In this section we specialize the result of Theorem 3.4 for the case when Q has the following form: $Q = \{\lambda \in \mathbb{R}^n : \|\lambda\| \leq R\}$.

***Theorem** 3.5:* Assume that Λ^* is bounded. Then, the sequence $\{\lambda^k\}_{k\geq 0}$ generated by Algorithm 3.2 is also bounded.

Proof: Note that $\Lambda^* = \{\lambda : f_0(\lambda) \geq f^*\}$. Let us introduce the sets

$$\Lambda^0 = \{\lambda : f_0(\lambda) \geq f^* - c(D_X + D_Z)\} \text{ and } \Lambda^c = \{\lambda : f_c(\lambda) \geq f^*\}.$$

From the inequalities in (5) it follows immediately that $\Lambda^* \subseteq \Lambda^0$ and also $\Lambda^* \subseteq \Lambda^c$. Therefore, the sets Λ^c and Λ^0 are nonempty. Since Λ^* is bounded, from Corollary 8.7.1 in [15] it follows that the set Λ^0 is also bounded. We can also show that Λ^c is a bounded set. Indeed let $\lambda \in \Lambda^c$, then using once more the second inequality in (5) we obtain: $f_0(\lambda) + c(D_X + D_Z) \geq f_c(\lambda) \geq f^*$, i.e. $\lambda \in \Lambda^0$. In conclusion, $\Lambda^c \subseteq \Lambda^0$ and thus Λ^c is also bounded.

Let us now show that the sequence $\{\lambda^k\}_{k\geq 0}$ is bounded. From Theorem 2.6 it follows that the sequence $\{f_c(\lambda^k)\}_{k\geq 0}$ is nondecreasing and thus $\{\lambda^k : k \geq 0\} \subseteq \{\lambda : f_c(\lambda) \geq f_c(\lambda^0)\}$. But since Λ^c is bounded, using once again Corollary 8.7.1 in [15] it follows that the set $\{\lambda : f_c(\lambda) \geq f_c(\lambda^0)\}$ is bounded. In conclusion, the sequence $\{\lambda^k\}_{k\geq 0}$ is bounded. □

Since Assumption 2.3 holds, then Λ^* is nonempty. Conditions under which Λ^* is bounded can be found in [15] (e.g. when the matrix $[A\ B]$ has full row rank). Under the assumptions of Theorem 3.5, it follows that there exists $R > 0$ sufficiently large such that the set $Q = \{\lambda \in \mathbb{R}^n : \|\lambda\| \leq R\}$ contains[1] a $\lambda^* \in \Lambda^*$, and thus we can assume D_Q to be finite. Notice that similar arguments were used in order to prove convergence of two-level algorithms for convex optimization problems (see e.g. [17]).

The next theorem shows how to choose optimally the smoothness parameter c and provides the complexity estimates of our method for the case when Q is a ball.

[1]In a practical algorithm R is increased adaptively: if λ^k approaches the boundary of Q we take $R_{k+1} = \alpha R_k$ for some $\alpha > 1$. An upper bound on $\|\lambda^*\|_*$ can also be estimated using R.

***Theorem** 3.6:* Assume that there exists $R > 0$ such that the set $Q = \{\lambda \in \mathbb{R}^n : \|\lambda\| \leq R\}$ contains a $\lambda^* \in \Lambda^*$. Taking $c = \frac{2}{k+1}\sqrt{\frac{D_Q}{D_X+D_Z}\big(\frac{\|A\|^2}{\sigma_Q\sigma_X} + \frac{\|B\|^2}{\sigma_Q\sigma_Z}\big)}$, then after k iterations we obtain an approximate solution to the problem (1) $(\hat{x}, \hat{z}) = \sum_{l=0}^{k} \frac{2(l+1)}{(k+1)(k+2)}(x^{l+1}, z^{l+1})$ and $\hat{\lambda} = \lambda^k$ which satisfy the following bounds on the duality gap and constraints:

$$[\phi_1(\hat{x}) + \phi_2(\hat{z})] - f_0(\hat{\lambda}) \leq \frac{4}{k+1}\sqrt{D_Q(D_X + D_Z)\big(\frac{\|A\|^2}{\sigma_Q\sigma_X} + \frac{\|B\|^2}{\sigma_Q\sigma_Z}\big)}, \tag{10}$$

$$\|A\hat{x} + B\hat{z} - b\|_* \leq \frac{4}{(R - \|\lambda^*\|_*)(k+1)}\sqrt{D_Q(D_X + D_Z)\big(\frac{\|A\|^2}{\sigma_Q\sigma_X} + \frac{\|B\|^2}{\sigma_Q\sigma_Z}\big)}. \tag{11}$$

Proof: Using (9) and the form of Q we obtain that

$$\max_{\lambda\in Q} -\frac{4L_c}{\sigma_Q(k+1)^2}d_Q(\lambda) + \langle A\hat{x} + B\hat{z} - b, \lambda\rangle \geq -\frac{4L_cD_Q}{\sigma_Q(k+1)^2} + \max_{\|\lambda\|\leq R}\langle A\hat{x} + B\hat{z} - b, \lambda\rangle =$$
$$-\frac{4L_cD_Q}{\sigma_Q(k+1)^2} + R\|A\hat{x} + B\hat{z} - b\|_*.$$

In view of the previous relation and Theorem 3.4 we obtain the following duality gap:

$$[\phi_1(\hat{x}) + \phi_2(\hat{z})] - f_0(\hat{\lambda}) \leq c(D_X + D_Z) + \frac{4L_cD_Q}{\sigma_Q(k+1)^2} - R\|A\hat{x} + B\hat{z} - b\|_* \leq$$
$$c(D_X + D_Z) + \frac{4L_cD_Q}{\sigma_Q(k+1)^2}$$

Minimizing the right-hand side of this inequality over c we get the above expressions for c and for the upper bound on the duality gap (10). Moreover, for the constraints using Lemma 3.3 and inequality (7) we have that

$$(R - \|\lambda^*\|_*)\|A\hat{x} + B\hat{z} - b\|_* \leq c(D_X + D_Z) + \frac{4L_cD_Q}{\sigma_Q(k+1)^2}$$

and replacing c derived above we also get the bound on the constraints violation (11). □

From Theorem 3.6 and inequality (7) we obtain that the complexity for finding an (ϵ, ϵ_c)-approximation of the optimal value function f^*, when the set Q is a ball, is $k + 1 = 4\sqrt{D_Q(D_X + D_Z)\big(\frac{\|A\|^2}{\sigma_Q\sigma_X} + \frac{\|B\|^2}{\sigma_Q\sigma_Z}\big)}\ \frac{1}{\epsilon}$, i.e. the efficiency estimates of our scheme is of the order $\mathcal{O}(\frac{1}{\epsilon})$, better than most non-smooth optimization schemes such as the subgradient method that have an efficiency estimate of the order $\mathcal{O}(\frac{1}{\epsilon^2})$ (see e.g. [13]). Moreover the dependence of the parameters c and L_c on ϵ is as follows: $c = \frac{\epsilon}{2(D_X+D_Z)}$ and $L_c = \big(\frac{\|A\|^2}{\sigma_X} + \frac{\|B\|^2}{\sigma_Z}\big)\frac{D_X+D_Z}{2\epsilon}$. Another advantage of the proximal center method is that we are free in the choice of the norms in the spaces $\mathbb{R}^n, \mathbb{R}^m$ and $\mathbb{R}^p$, while most of the decomposition schemes are based on the Euclidian norm. Thus, we can choose the norms which make the ratio $\frac{\|A\|^2}{\sigma_Q\sigma_X}$ as small as possible.

D. Efficiency estimates for the Euclidian norm

In this section we assume that $\mathbb{R}^n$ is endowed with the Euclidian norm.

***Theorem** 3.7:* Assume that $Q = \mathbb{R}^n$ and $d_Q(\lambda) = \frac{1}{2}\|\lambda\|^2$, with the Euclidian norm on $\mathbb{R}^n$. Taking $c = \frac{\epsilon}{D_X+D_Z}$ and $k+1 = 2\sqrt{\big(\frac{\|A\|^2}{\sigma_X} + \frac{\|B\|^2}{\sigma_Z}\big)(D_X+D_Z)}\ \frac{1}{\epsilon}$, then after k iterations the duality gap is less than ϵ and the constraints satisfy $\|A\hat{x} + B\hat{z} - b\| \leq \epsilon\big(\|\lambda^*\| + \sqrt{\|\lambda^*\|^2+2}\big)$.

Proof: Let us note that for these choices of Q and d_Q we have $\sigma_Q = 1$ and thus

$$\max_{\lambda\in\mathbb{R}^n} -\frac{4L_c}{\sigma_Q(k+1)^2}d_Q(\lambda) + \langle A\hat{x} + B\hat{z} - b, \lambda\rangle = \frac{(k+1)^2}{8L_c}\|A\hat{x} + B\hat{z} - b\|^2.$$

In this case we obtain the following bound on the duality gap (see Theorem 3.4):

$$[\phi_1(\hat{x}) + \phi_2(\hat{z})] - f_0(\hat{\lambda}) \leq c(D_X + D_Z) - \frac{(k+1)^2}{8L_c}\|A\hat{x} + B\hat{z} - b\|^2 \leq c(D_X + D_Z).$$

It follows that taking $c = \frac{\epsilon}{D_X+D_Z}$, the duality gap is less than ϵ. For the constraints using Lemma 3.3 and inequality (7) we get that $\|A\hat{x} + B\hat{z} - b\|$ satisfies the second order inequality in y: $\frac{(k+1)^2}{8L_c}y^2 - \|\lambda^*\|y - \epsilon \leq 0$. Therefore, $\|A\hat{x} + B\hat{z} - b\|$ must be less than the largest root of the corresponding second-order equation, i.e.

$$\|A\hat{x} + B\hat{z} - b\| \leq \big(\|\lambda^*\| + \sqrt{\|\lambda^*\|^2 + \frac{\epsilon(k+1)^2}{2L_c}}\big)\frac{4L_c}{(k+1)^2}.$$

After some long but straightforward computations we get that after k iterations, where k defined as in the theorem, we also get $\|A\hat{x} + B\hat{z} - b\| \leq \epsilon\big(\|\lambda^*\| + \sqrt{\|\lambda^*\|^2+2}\big)$. □

***Remark** 3.8:* When coupling inequalities $Ax + Bz - b \leq 0$ are present, then we choose $Q = \mathbb{R}^n_+$. Using the same reasoning as before we get that $\max_{\lambda\geq 0} -\frac{4L_c}{\sigma_Q(k+1)^2}d_Q(\lambda) + \langle A\hat{x} + B\hat{z} - b, \lambda\rangle = \frac{(k+1)^2}{8L_c}\|[A\hat{x} + B\hat{z} - b]^+\|^2$, where $[y]^+$ denotes the projection of $y \in \mathbb{R}^n$ onto $\mathbb{R}^n_+$. Taking for c and k the same values as in Theorem 3.7, we can conclude that after k iterations the duality gap is less than ϵ and using a modified version of Lemma 3.3 (i.e. a generalized version of Cauchy-Schwarz inequality: $-\langle y, \lambda\rangle \geq -\|[y]^+\|\ \|\lambda\|$ for any $y \in \mathbb{R}^n, \lambda \geq 0$) the constraints violation satisfy $\|[A\hat{x} + B\hat{z} - b]^+\| \leq \epsilon\big(\|\lambda^*\| + \sqrt{\|\lambda^*\|^2+2}\big)$. □

Finally, let us mention that our decomposition method described in Algorithm 3.2 bears similarity with proximal type methods [1], [7]–[10], [12], but is different both in the computational steps and in the choice of the parameter c. More precisely, our method uses a fixed center in the

prox-terms that allows the inner iterations at each k to move freely, contrary to most proximal type methods that force the next iterates to be close to the previous ones. The main advantage of our scheme is that it is fully automatic, the parameter c is chosen unambiguously, which is crucial for justifying the convergence properties of Algorithm 3.2.

IV. APPLICATIONS

A. Applications with separable structure

In this section we briefly discuss some of the applications to which our method can be applied.

First application that we will discuss here is the control of large-scale systems with interacting subsystem dynamics. A distributed MPC framework is appealing in this context since this framework allows us to design local subsystem-base controllers that take care of the interactions between different subsystems and physical constraints. We assume that the overall system model can be decomposed into M appropriate subsystem models:

$$x^i(k+1) = \sum_{j\in\mathcal{N}(i)} A_{ij}x^j(k) + B_{ij}u^j(k) \;\; \forall i = 1\cdots M,$$

where $\mathcal{N}(i)$ denotes the set of subsystems that interact with the ith subsystem, including itself. The control and state sequence must satisfy local constraints: $x^i(k) \in \Omega_i$ and $u^i(k) \in U_i$ for all i and $k \geq 0$, where the sets Ω_i and U_i are usually convex compact sets with the origin in their interior. In general the control objective is to steer the state of the system to origin or any other set point in a "best" way. Performance is expressed via a stage cost, which in this paper we assume to have the following form [4]: $\sum_{i=1}^{M} \ell_i(x^i, u^i)$, where usually ℓ_i is a convex quadratic function, not necessarily strictly convex. Let N denote the prediction horizon. In MPC we must solve at each step k, given $x^i(k) = x^i$, an optimal control problem of the following form:

$$\min_{x_l^i, u_l^i} \Big\{ \sum_{l=0}^{N-1}\sum_{i=1}^{M} \ell_i(x_l^i, u_l^i) : x_0^i = x^i,\; x_{l+1}^i = \sum_{j\in\mathcal{N}(i)} A_{ij}x_l^j + B_{ij}u_l^j,\; x_l^i \in \Omega_i,\; u_l^i \in U_i \;\forall\, l, i\Big\} \quad (12)$$

A similar formulation of distributed MPC for coupled linear subsystems with decoupled costs was given in [4], but without state constraints. Let us introduce $\mathbf{x}_i = (x_0^i \cdots x_N^i \;\; u_0^i \cdots u_{N-1}^i), X_i = \Omega_i^{N+1} \times U_i^N$ and the non-strictly convex quadratic functions $\psi_i(\mathbf{x}^i) = \sum_{l=0}^{N-1} \ell_i(x_l^i, u_l^i)$. Then, the control problem (12) can be recast as a separable convex program:

$$\min_{\mathbf{x}_i \in X_i} \Big\{ \sum_{i=1}^{M} \psi_i(\mathbf{x}_i) : \sum_{i=1}^{M} C_i\mathbf{x}_i - \gamma = 0\Big\}, \quad (13)$$

where the matrices C_i's and the vector γ are defined appropriately.

Network optimization furnishes another areas in which our Algorithm 3.2 leads to a new method of solution. In this application the convex optimization problem has the form [2], [3]:

$$\min_{x_i \in X_i} \Big\{ \sum_{i=1}^{M} \psi_i(x_i) : \sum_{i=1}^{M} C_i x_i - \gamma = 0,\ \sum_{i=1}^{M} D_i x_i - \beta \leq 0 \Big\}, \tag{14}$$

where X_i are compact convex sets (in general balls) in the Euclidian space $\mathbb{R}^m$, ψ_i's are non-strictly convex functions and M denotes the number of agents in the network.

In [4] the optimization problem (12) (or equivalently (13)) was solved in a decentralized fashion, iterating the Jacobi algorithm [1] $p_{\max}$ times. But, there is no theoretical guarantee of the Jacobi algorithm about how good is the approximation of the optimum after $p_{\max}$ iterations and moreover we need strictly convex functions ψ_i to prove asymptotic convergence to the optimum. However, if we solve (12) using Algorithm 3.2, we have a guaranteed upper bound (see Theorem 3.6 or 3.7) on the approximation of the optimum after $p_{\max}$ iterations. In [2], [3] the optimization problem (14) is solved using the dual subgradient method described in the Algorithm 2.2. Some preliminary simulation tests from Section IV-B show that Algorithm 3.2 is superior to Algorithm 2.2.

Let us describe briefly the main ingredients of Algorithm 3.2 for problem (14). Let d_{X_i} be properly chosen prox-functions, according to the structure of the sets X_i's and the norm[2] used on $\mathbb{R}^m$. Then, the smooth dual function f_c has the following form:

$$f_c(\lambda) = \min_{x_i \in X_i} \sum_{i=1}^{M} \psi_i(x_i) + \langle \lambda_1, \sum_{i=1}^{M} C_i x_i - \alpha \rangle + \langle \lambda_2, \sum_{i=1}^{M} D_i x_i - \beta \rangle + c \sum_{i=1}^{M} d_{X_i}(x_i).$$

Moreover, $Q = \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}_+$, where n_1 and n_2 denote the number of equality and inequality constraints. In this case all the minimization problems of Algorithm 3.2 are decomposable in x_i and thus the agents can solve the optimization problem (14) in a decentralized fashion.

B. Computational results

We conclude this paper with the results of computational experiments on a random set of problems of the form (14), where ψ_i's are convex quadratic functions (but not strictly convex)

[2]For example if $X_i = \{x \in \mathbb{R}^m : \|x - x_0\|_2 \leq R\}$, where $\|\cdot\|$ denotes here the Euclidian norm, then it is natural to take $d_{X_i}(x) = \frac{\|x-x_0\|_2^2}{2}$. If $X_i = \{x \in \mathbb{R}^m : x \geq 0,\ \sum_{i=1}^m x_{(i)} = 1\}$, then the norm $\|x\| = \sum_{i=1}^m |x_{(i)}|$ and $d_{X_i}(x) = \ln m + \sum_{i=1}^m x_{(i)} \ln x_{(i)}$ is more suitable (see [14] for more details).

and X_i's are balls in $\mathbb{R}^m$ defined with the Euclidian norm. Similarly, $\mathbb{R}^n$ (corresponding to the Lagrange multipliers) is endowed with the Euclidian norm (see Section III-D).

		$M = 2$		$M = 10$	
m	ϵ	nr iter Alg 3.2	nr iter Alg 2.2	nr iter Alg 3.2	nr iter Alg 2.2
50	0.01	202	5000(0.05)	1211	5000(0.29)
200	0.01	625	5000(0.19)	3128	5000(0.51)
1000	0.01	890	10000(0.34)	7240	10000(0.62)
50	0.001	688	5000(0.05)	4898	5000(0.29)
200	0.001	2198	5000(0.19)	5000(0.005)	5000(0.51)
1000	0.001	3926	10000(0.34)	10000(0.008)	10000(0.62)

In the table we display the number of iterations of the Algorithm 2.2 and 3.2 for different values of m, M and of the accuracy ϵ. For $m = 50$ and $m = 200$ the maximum number of iterations that we allow is 5000. For $m = 1000$ we iterate at most 10000 times. When maximum number of iterations is reached we also display between brackets the corresponding accuracy. We see that the duality gap is much better with our Algorithm 3.2 than with Algorithm 2.2.

V. CONCLUSIONS

A new decomposition method in convex programming is developed in this paper using the framework of dual decomposition. Our method combines the computationally non-expensive gradient method with the efficiency of structural optimization for solving separable convex programs. Although our method resembles proximal-based methods, it differs both in the computational steps and in the choice of the parameters. Contrary to most proximal-based methods that enforce the next iterates to be close to the previous ones, our method uses fixed centers in the prox-terms which leads to more freedom in the next iterates. Another advantage of our proximal center method is that it is fully automatic, i.e. the parameters of the scheme are chosen optimally, which are crucial for justifying its convergence properties. We also presented efficiency estimate results of the new method for general separable convex problems. The computations on some test problems confirm that the proposed method works well in practice.

Acknowledgment. The authors thank Michel Baes for helpful discussions. We also acknowledge financial support by Research Council K.U. Leuven: GOA AMBioRICS, CoE EF/05/006, OT/03/12, PhD/postdoc & fellow grants;

Flemish Government: FWO PhD/postdoc grants, FWO projects G.0499.04, G.0211.05, G.0226.06, G.0302.07; Research communities (ICCoS, ANMMM, MLDM); AWI: BIL/05/43, IWT: PhD Grants; Belgian Federal Science Policy Office: IUAP DYSCO.

REFERENCES

[1] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and distributed computation: Numerical Methods*. Englewood Cliffs, NJ: Prentice-Hall, 1989.

[2] L. Xiao, M. Johansson, and S. Boyd, "Simultaneous routing and resource allocation via dual decomposition," *IEEE Transactions on Communications*, vol. 52, no. 7, pp. 1136–1144, 2004.

[3] D. Palomar and M. Chiang, "Alternative distributed algorithms for network utility maximization: Framework and applications," *IEEE Transactions on Automatic Control*, vol. 52, no. 12, pp. 2254–2268, 2007.

[4] A. Venkat, I. Hiskens, J. Rawlings, and S. Wright, "Distributed MPC strategies with application to power system automatic generation control," *IEEE Transactions on Control Systems Technology*, vol. to appear, 2007.

[5] R. Rockafellar and R.-B. Wets, "Scenarios and policy aggregation in optimization under uncertainty," *Mathematics of Operations Research*, vol. 16, no. 1, pp. 119–147, 1991.

[6] H. Uzawa, "Iterative methods for concave programming," in *Studies in Linear and Nonlinear Programming*, K. Arrow, L. Hurwicz, and H. Uzawa, Eds., 1958, pp. 154–165.

[7] P. Tseng, "Applications of a splitting algorithm to decomposition in convex programming and variational inequalities," *SIAM Journal of Control and Optimization*, vol. 29, no. 1, pp. 119–138, 1991.

[8] S. Kontogiorgis, R. D. Leone, and R. Meyer, "Alternating direction splitings for block angular parallel optimization," *Journal of Optimization Theory and Applications*, vol. 90, no. 1, pp. 1–29, 1996.

[9] J. Eckstein and D. Bertsekas, "On the Douglas–Rachford splitting method and the proximal point method for maximal monotone operators," *Mathematical Programming (A)*, vol. 55, no. 3, pp. 293–318, 1992.

[10] G. Chen and M. Teboulle, "A proximal-based decomposition method for convex minimization problems," *Mathematical Programming (A)*, vol. 64, pp. 81–101, 1994.

[11] J. E. Spingarn, "Applications of the method of partial inverses to convex programming: decomposition," *Mathematical Programming (A)*, vol. 32, pp. 199–223, 1985.

[12] P. Mahey, S. Oualibouch, and P. Tao, "Proximal decomposition on the graph of a maximal monotone operator," *SIAM Journal of Optimization*, vol. 5, pp. 454–466, 1995.

[13] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*. Boston: Kluwer, 2004.

[14] ——, "Smooth minimization of non-smooth functions," *Mathematical Programming (A)*, vol. 103, no. 1, pp. 127–152, 2005.

[15] R. T. Rockafellar, *Convex Analysis*, ser. Princeton Mathematics Series. Princeton University Press, 1970, vol. 28.

[16] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[17] G. Cohen, "Optimization by decomposition and coordination: A unified approach," *IEEE Transactions on Automatic Control*, vol. AC–23, no. 2, pp. 222–232, 1978.