Decomposition of Large Scale Linear Programming Problems Containing both Linking Variables and Constraints via Accuracy Certificates

Bruce A. Cox

3 March 2020

Abstract Several well known large scale linear programming decomposition methodologies exist. Benders Decomposition, which covers the case where some small subset of variables link the otherwise separable subproblems. Dantzig-Wolfe decomposition and Lagrangian decompositions, which cover the case where some few constraints link the otherwise separable subproblems, and finally the ”Cross-Decomposition” originating from TJ Van Roy which enables one to deal with both linking constraints and linking variables by essentially alternating iteratively between the Benders and the Lagrangian Decomposition. In this paper we present a novel alternative to Cross-decomposition that deals with both linking constraints and linking variables through the application of accuracy certificates for black-box, sub-gradient based algorithms such as NERML.

1 Introduction

Our paper is motivated by the consideration of a large scale Linear Programming problem such as:

$$\text{Opt} = \min_{x \in [x_1,x_2]} \left\{ c_1^T x_1 + c_2^T x_2 : A^{11} x_1 + A^{12} x_2 \leq b_1, A^{21} x_1 + A^{22} x_2 \leq b_2, \|x\|_\infty \leq R \right\}$$

(1)

With $n = n_1 + n_2$ variables and $m = m_1 + m_2$ linear inequality constraints, and the sizes of $A^{11}$, $A^{12}$, $A^{21}$, and $A^{22}$ are $m_1 \times n_1$, $m_1 \times n_2$, $m_2 \times n_1$, and $m_2 \times n_2$ respectively. Note that the bounds on variables $\|x\|_\infty \leq R$ are added for technical reasons and have no real practical significance since $R$ can be arbitrarily large.

B. Cox
Air Force Institute of Technology
E-mail: BruceACox1@gmail.com
In addition we make the following critical assumptions regarding the nature of the underlying problem.

(1) It is relatively easy to solve Linear Programming problems of the form

\[ \min_{x_1} \{ c^T x_1 : A^{11} x_1 \leq b, \|x_1\|_\infty \leq R \} \]  

(2) The number of linking constraints \( m_2 \) is \( \ll \) the total number of constraints \( m \) in (1), and the number of linking variables \( n_2 \) is \( \ll \) the total number of variables \( n \) in (1).

For example the following graphic outlines one such case that meets both assumptions.

The fundamental question is: How can we exploit the specific structure of (1) in order to accelerate its solution? This is the question we address in this paper. The organization of the paper is as follows: In section 2 we address two well known cases of the outlined situation where we already know how to act — one where there are no linking constraints, and one where there are no linking variables. In section 3 we address the more interesting general case where both linking variables and linking constraints are present, via a novel approach utilizing accuracy certificates.
2 Two Well Known Cases

2.1 Case A: No Linking Constraints

Consider the case where there are no linking constraints, that is, \( m_2 = 0 \). In this case, we can define the convex function

\[
f(x_2) = \min_{x_1} \left\{ c_1^T x_1 + c_2^T x_2 : A^{11} x_1 + A^{12} x_2 \leq b_1, \|x_1\|_\infty \leq R \right\}
\]

and replace the problem of interest with an equivalent problem

\[
\min_{x_2 : \|x_2\|_\infty \leq R} f(x_2) = \min_{x_2 : \|x_2\|_\infty \leq R} \left( \min_{x_1} \left\{ c_1^T x_1 + c_2^T x_2 : A^{11} x_1 + A^{12} x_2 \leq b_1, \|x_1\|_\infty \leq R \right\} \right).
\]

(3)

According to assumption (1), it is relatively easy to check, given \( x_2 \), whether \( x_2 \in \text{Dom } f \), and if so, to compute \( f(x_2) \) and a subgradient \( f'(x_2) \) of \( f_2 \) at \( x_2 \). Indeed, to this end one should solve the LP program

\[
\min_{x_1} \left\{ c_1^T x_1 : A^{11} x_1 \leq b_1 - A^{12} x_2, \|x_1\|_\infty \leq R \right\},
\]

(4)

which by our assumption is easy. If this problem is solvable, then \( x_2 \in \text{Dom } f \), and the optimal solution \( x_1(x_2) \) to the LP (4), along with an optimal solution \( y_1 = y_1(x_2) \) to its associated dual problem

\[
\max_{y_1} \left\{ [A^{12} x_2 - b_1]^T y_1 : [A^{11}]^T y_1 + c_1 = 0, y_1 \geq 0 \right\}
\]

allows us to compute \( f(x_2) \), and \( f'(x_2) \) according to

\[
f(x_2) = c_1^T x_1(x_2) + c_2^T x_2 = y_1^T [A^{12} x_2 - b_1] + c_2^T x_2, \quad f'(x_2) = [A^{12}]^T y_1 + c_2.
\]

Conversely if the problem (4) is unsolvable (which, since the feasible set of the problem clearly is bounded, can happen only if the problem (4) is infeasible), we will receive, via the General Theorem on Alternative, an infeasibility certificate \( y_1 = y_1(x_2) \) such that

\[
y_1 \geq 0 \& y_1^T [A^{12} x_2 - b_1] > \| [A^{11}]^T y_1 \|_\infty;
\]

in this case, the linear form \( \langle [A^{12}]^T y_1, \cdot \rangle \) is nonzero and separates \( x_2 \) from the feasible domain \( X_2 \) of (3). Thus, (3) is naturally equipped with both a Separation oracle, and a First Order oracle, and as such can be solved by a black box oriented method for convex minimization, see, e.g., [3]. By assumption, the dimension of this problem \( n_2 \) is \( \ll \) the dimension of (1), which can make this outlined approach highly preferable to a direct attack on (1) with an LP solver.

Note that after a feasible \( \epsilon \)-optimal solution \( x_2 \) to (3) is found, we automatically have at our disposal an \( \epsilon \)-optimal feasible solution \( [x_1(x_2) ; x_2] \) to the problem of interest.

Note that one of the most popular implementations of this scheme is the classical Benders decomposition, see, e.g., [3].
2.2 Case B: No Linking Variables

Now consider the case where there are no linking variables, that is, $n_2 = 0$. One well-known way to utilize the resulting structure is to use the well-known Dantzig-Wolfe decomposition which is intrinsically linked to the revised Primal Simplex Method, see, e.g., [6]. Another standard way to handle (1) with no linking variables is to use Lagrangian decomposition, that is, to associate with (1) its partial dual problem (where we dualize the linking constraints) – which we write down in the minimization form

$$
\min_{y_2 \geq 0} f(y_2) = \max_{x_1} \left\{ -[c_1 + [A^{21}]^T y_2]^T x_1 : A^{11} x_1 \leq b_1, \|x_1\|_\infty \leq R \right\} \cdot b_2^T y_2. 
$$

Observe that $f(y_2)$ is an everywhere finite convex function which can be easily equipped with a First Order oracle, since by our assumption the parametric LP problem specifying $f$ is relatively easy to solve. Indeed the LP associated with (5) can be rewritten as

$$
\max_{x_1} \left\{ -[c_1 + [A^{21}]^T y_2]^T x_1 : A^{11} x_1 \leq b_1, \|x_1\|_1 \leq R \right\}.
$$

In many cases we can find a finite upper bound $L$ on, say, the $\|\cdot\|_1$-norm of an optimal solution to (5) and thus reduce this problem to a convex minimization on a simple solid, which we can solve by a black box oriented method (see e.g. [5]). Since the dimension $n_2$ of (5) is small when compared to the sizes of the problem of interest (1), this approach can be more attractive than a direct attack on (1) with an LP solver.

Note that the Lagrange decomposition approach requires us to solve the nontrivial problem of recovering a good approximate solution to the problem of interest (1) from a good approximate solution to the problem

$$
- \text{Opt}(L) := \min_{y_2 \in \mathcal{B}} f(y_2), \quad \mathcal{B} = \mathcal{B}_L = \left\{ y_2 : y_2 \geq 0, \sum_{i=1}^{m_2} [y_2]_i \leq L \right\}
$$

(which is nothing but the problem (5) with an added bound on variables). Observe that by Weak Duality $-\text{Opt}(L) \leq \text{Opt}$ for all $L$. One of the ways to resolve this problem is to use accuracy certificates; to this end, we have proved the following result:

**Proposition 1** Assume that $n_2 = 0$ and that (6) is solved by an algorithm with accuracy certificates. By construction of the First Order oracle for the
objective of this problem, at every productive step $t$ of this algorithm, the search point being $y^*_t \in B_L$, we have at our disposal a point

$$x^*_t = x_1(y^*_t) \in \text{Argmax} \left\{ -[c_1 + [A^{21}]^T y^*_t]^T x_1 : A^{11} x_1 \leq b_1, \|x_1\|_\infty \leq R \right\}.$$  

Now let $\tau$ be a step such that the accuracy certificate $\xi^\tau$ associated with the corresponding execution protocol $P_\tau = \{(y^*_t, \epsilon_t)_{t=1}^{\tau}; I_\tau, J_\tau\}$ is well defined. Setting

$$x^*_\tau = \sum_{t \in I_\tau} \xi^\tau_t x^*_t,$$

we get an approximate solution to the problem of interest (1) such that

$$A^{11} x^*_\tau \leq b_1 \& \|x^*_\tau\|_\infty \leq R,$$

$$A^{21} x^*_\tau \leq b_2 + L^{-1} [\epsilon_{\text{cert}}(\xi^\tau \mid P_\tau, B_L) + R\|c_1\|_1] \tag{9}$$

where $1$ is the all-ones vector of dimension $m_2$.

In addition, let $\tilde{y} = [\tilde{y}_1; \tilde{y}_2; \tilde{y}_1; \tilde{y}_2] \geq 0$ be the vector of optimal Lagrange multipliers for (1), so that

$$c^T_1 x_1 + \tilde{y}^T_1 [A^{11} x_1 - b_1] + \tilde{y}^T_2 [A^{21} x_1 - b_2] + \tilde{y}^T_1 [x_1 - R1] + \tilde{y}^T_2 [-x_1 - R1] \equiv \text{Opt} \forall x_1. \tag{10}$$

When $t := L - \sum_{i=2}^{m_2} |\tilde{y}_i| > 0$, then, in addition to the second relation in (9), we have

$$A^{21} x^*_\tau \leq b_2 + t^{-1} \epsilon_{\text{cert}}(\xi^\tau \mid P_\tau, B_L) 1 \tag{11}$$

**Proof.** Let $x^*_\tau$ be as stated in Proposition. Then

i. Clearly $A^{11} x^*_\tau \leq b_1 \& \|x^*_\tau\|_\infty \leq R$ holds since by construction $x^*_\tau$ is a convex combination of feasible solutions to (1).

ii. We have for every $y_2 \in B_L$:

$$\epsilon_{\text{cert}}(\xi^\tau \mid P_\tau, B_L) \geq \sum_{t=1}^{\tau} \xi_t (f(y^*_t, y^*_2) - y_2) \quad \text{[definition of } \epsilon_{\text{cert}}]\tag{9}\]

$$= \sum_{t \in I_\tau} \xi_t (f(y^*_t, y^*_2) - y_2) \quad \text{[since } \langle \epsilon_t, y^*_2 - y_2 \rangle \geq 0 \text{ for } t \notin I_\tau]\tag{9}\]

$$= \sum_{t \in I_\tau} \xi_t [(b - A^{21} x^*_t, y^*_2) - (b - A^{21} x^*_t, y_2)]\tag{9}\]

$$= \sum_{t \in I_\tau} \xi_t [f(y^*_t, y^*_2) + c^T_1 x^*_t - (b - A^{21} x^*_t, y_2)] \quad \text{[see (7)]}\tag{9}\]

whence, taking into account that $f(y^*_2) \geq -\text{Opt}(L)$ for $t \in I_\tau$,

$$\langle A^{21} x^*_\tau - b_2, y_2 \rangle \leq \text{Opt}(L) + \epsilon_{\text{cert}}(\xi^\tau \mid P_\tau, B_L) - c^T_1 x^*_\tau. \tag{11}$$

Relation (11) holds true for all $y_2 \in B_L$; recalling what $B_L$ is and maximizing the left hand side in (11) over $y_2 \in B_L$ with $\sum_i y_{2i} = L$, we get

$$L \max_i |A^{21} x^*_\tau - b_2|_i \leq \text{Opt}(L) + \epsilon_{\text{cert}}(\xi^\tau \mid P_\tau, B_L) - c^T_1 x^*_\tau \leq \text{Opt}(L) + \epsilon_{\text{cert}}(\xi^\tau \mid P_\tau, B_L) - \|c_1\|_1 R$$
which is nothing but the second relation in (9). Setting in (!) \( y_2 = 0 \), we get
\[ c^T_1 \hat{x}^* _1 \leq \text{Opt}(L) + \varepsilon_{\text{cert}}(\xi^\tau | P_\tau, B_L). \]
Taking into account that by weak duality \( \text{Opt} \geq \sup_{y_2 \geq 0} [-f(y_2)] \geq \sup_{y_2 \in B_L} [-f(y_2)] = \text{Opt}(L) \), we arrive at the third relation in (9).

iii. To prove (11), we can assume w.l.o.g. that the set
\[ K = \{ i : [A^{21} \hat{x}^* _1 - b_2]_i > 0 \} \]
is nonempty, since otherwise (11) is evident. Setting in (10) \( x_1 = \hat{x}^* _1 \) and taking into account the first relation in (9), we get
\[ c^T_1 \hat{x}^* _1 + \tilde{y}^T_2 [A^{21} \hat{x}^* _1 - b_2] \geq \text{Opt} \geq \text{Opt}(L), \]
so that
\[ \text{Opt}(L) - c^T_1 \hat{x}^* _1 \leq \tilde{y}^T_2 [A^{21} \hat{x}^* _1 - b_2], \]
which combines with (!) to imply that for all \( y_2 \in B_L \) one has
\[ \langle y_2, A^{21} \hat{x}^* _1 - b_2 \rangle \leq \varepsilon_{\text{cert}}(\xi^\tau | P_\tau, B_L) + \tilde{y}^T_2 [A^{21} \hat{x}^* _1 - b_2] \]
\[ \leq \varepsilon_{\text{cert}}(\xi^\tau | P_\tau, B_L) + \sum_{i=1}^{m_2} [\tilde{y}^*_2]_i \max_i [A^{12} \hat{x}^* _1 - b_2]_i. \]
Let \( \mu := \max_i [A^{12} \hat{x}^* _1 - b_2]_i = [A^{12} \hat{x}^* _1 - b_2]_{i*} \), so that the above relation reads
\[ \langle y_2, b_2 - A^{21} \hat{x}^* _1 \rangle \leq \varepsilon_{\text{cert}}(\xi^\tau | P_\tau, B_L) + \mu \sum_{i=1}^{m_2} [\tilde{y}^*_2]_i. \]
Setting the \( i_* \)-th coordinate of \( y_2 \) to the value \( L \), and the remaining coordinates – to the value 0, we get from the latter inequality
\[ \mu L \leq \varepsilon_{\text{cert}}(\xi^\tau | P_\tau, B_L) + \mu \sum_{i=1}^{m_2} [\tilde{y}^*_2]_i, \]
and (11) follows. \( \blacksquare \)

3 General Case

Now consider the general case, when there are both linking variables and linking constraints. To the best of our knowledge, the only decomposition scheme proposed for this case is “cross decomposition” originating from T.J. Van Roy \cite{43,44}, see also \cite{15,10,17,37} and references therein. In this scheme, essentially, one alternates iteratively between the Benders and the Lagrange decompositions. To the best of our understanding, no complexity results for this scheme are known. \footnote{This being said, note that the primary motivation and application of cross decomposition is the Mixed Integer version of (1), where the linking variables \( x_2 \) are subject to additional constraints of integrality.}
3.1 Assumptions and Approach

Consider the Lagrange function of problem (1), assuming that both linking constraints and linking variables exist. This bilinear function is:

\[
\Psi(x_1, x_2, y_1, y_2) = (c_1, x_1) + (c_2, x_2) + \langle y_1, A^{11}x_1 + A^{12}x_2 - b_1 \rangle + \langle y_2, A^{21}x_1 + A^{22}x_2 - b_2 \rangle
\]

This function leads us naturally to consider the bilinear convex-concave saddle point problem equivalent to problem (1).

\[
\inf_{x=[x_1, x_2] \in X_1 \times X_2} \sup_{y=[y_1, y_2] \geq 0} \Psi(x_1, x_2, y_1, y_2), \quad (13)
\]

where \(X_1 = \{x_1 : \|x_1\|_\infty \leq R\}\) and \(X_2 = \{x_2 : \|x_2\|_\infty \leq R\}\) are simple solids. Assume, as is usually the case, that we can shrink the \(y\)-domain of this saddle point problem, which is a nonnegative orthant, to a direct product \(Y = Y_1 \times Y_2\) of two simple solids which are “large enough” to ensure that the saddle point problem

\[
\min_{x=[x_1, x_2] \in X_1 \times X_2} \max_{y=[y_1, y_2] \in Y_1 \times Y_2} \Psi(x_1, x_2, y_1, y_2)
\]

is equivalent to (13) and thus – to the LP of interest (or approximates (13) “well enough”).

**Assumption.** At the beginning of this chapter we have assumed that it is easy to solve the LPs of the form \(\min_{x} c^T x_1 : A^{11}x_1 \leq b, \|x\|_\infty \leq R\). From now on we modify this assumption, specifically, assume that

**A:** Given \(x_2 \in X_2, y_2 \in Y_2\), it is easy to solve the saddle point problem

\[
\min_{x_1 \in X_1} \max_{y_1 \in Y_1} \{\Psi(x_1, x_2, y_1, y_2) = (c + [A^{11}]^T y_1, x_1) - \langle b, y_1 \rangle\},
\]

\[
c = c_1 + [A^{12}]^T y_2, \quad b = b_1 - A^{12}x_2 \]

(15)

Note that the original assumption is nothing but the modified one with \(Y_1 = \mathbb{R}_{+}^m\). For our machinery to work, we need \(Y_1\) to be bounded, this is where the modification comes from. Note also that in typical situations where the unmodified assumption holds true, so is the modified one, provided that \(Y_1\) is chosen properly. For example, this is so in the case considered in section ??, where the problem \(\min_{x_1 \in X_1} \{c^T x_1 : A^{11}x_1 \leq b, \|x_1\|_\infty \leq R\}\) is just a collection of a large number of independent of each other LPs, or, which is the same, \(x_1 = [x_{11}; \ldots; x_{1K}]\) is a collection of a large number \(K\) of components \(x_{1k}\) with dimensions low as compared to \(n_1\), and \(A^{11}\) is block-diagonal: \(A^{11}x_1 = [A_{11}^{11}x_{11}; A_{12}^{11}x_{12}; \ldots; A_{1K}^{11}x_{1K}]\), with the dimensions of the blocks \(A_{ik}^{11}\) low as compared to \(m_1\). In this case, choosing \(Y_1\) as a large box:

\[
Y_1 = \{y = [y_{11}; \ldots; y_{1K}] : 0 \leq y_{1k} \leq R_k[1; \ldots; 1]\}
\]
with the blocks \( y_{1k} \) in \( y_1 \) corresponding to the blocks \( A_{1k}^{11} x_{1k} \) in \( A^{11} x_1 \), the saddle point problem in \( A \) decomposes into \( K \) independent low dimensional bilinear saddle point problems of the form

\[
\min_{x_{1k}, \|x_{1k}\| \leq R} \max_{y_{1k}, 0 \leq y_{1k} \leq R_k[1;\ldots;1]} \left[ \langle p_k, x_{1k} \rangle + \langle q_k, y_{1k} \rangle + \langle y_{1k}, A_{1k}^{11} x_{1k} \rangle \right],
\]

thus making assumption \( A \) quite realistic.

**Approach.** On a closer inspection, Assumption \( A \) says that it is easy to build the First Order oracle for the (clearly convex-concave) function

\[
\tilde{\Psi}(x_2, y_2) = \min_{x_1 \in X_1, y_1 \in Y_1} \max_{x_2, y_2} \psi(x_1, x_2, y_1, y_2),
\]

so that the saddle point problem

\[
\min_{x_2 \in X_2} \max_{y_2 \in Y_2} \tilde{\Psi}(x_2, y_2)
\]

(which, by assumption, is of dimensions much smaller than those of the problem of interest) is well suited for solving by a black-box-oriented method, like NERML or perhaps even the Ellipsoid method\(^2\).

When the sizes of (16) are small as compared to those of (14) and elimination of \( x_1, y_1 \), that is, computing the first order information for \( \tilde{\Psi} \), is cheap, the outlined approach can be computationally much more attractive than a direct attack on (13) (or, equivalently, on (1)). Note that with this approach, we in fact simultaneously eliminate both linking variables and linking constraints (and do it just once), in a sharp contrast to cross decomposition, where we iteratively eliminate, in an alternating fashion, either linking variables, or linking constraints.

The crucial (and nontrivial) question underlying the outlined approach is how to recover a good solution to the problem of actual interest (14) (or, which is the same, (1)) from a good solution to the induced saddle point problem (16). Our major goal in the sequel is to demonstrate that when solving (16) by an algorithm with accuracy certificates, these certificates allow to convert \( \epsilon \)-solutions to (16) into \( \epsilon \)-solutions of the saddle point problem of actual interest (14) and thus – to \( \epsilon \)-solutions to the original problem (1).

### 3.2 Induced Pairs of Saddle Point Problems

We intend to consider a situation which is an extension of the one considered in the previous section. Specifically, we intend to investigate saddle point problems *induced* by larger saddle point problems. The general setting is as follows.

\(^2\) The latter is quite realistic, provided that the sizes \( n_2 \) and \( m_2 \) are in the range of tens; note that there are meaningful applications where \( n_2 \) and \( m_2 \) are indeed in the range of tens, while \( m_1 \) and \( n_1 \) are in the range of many thousands.
Proposition 2

One has

$$\text{SadVal}_{X_1 \times Y_1} (\Psi_1) = \text{SadVal}_{X_2 \times Y_2} (\Psi_2) = \text{SadVal}_{X \times Y} (\Psi)$$  \hspace{1cm} (20)
and
\[
\text{SadSet}_{X \times Y}(\Psi) \subset \text{SadSet}_{X_1 \times Y_1}(\Psi_1) \times \text{SadSet}_{X_2 \times Y_2}(\Psi_2),
\]
where \(\text{SadSet}_{U \times V}(\Psi)\) is the set of saddle points of the function \(\Psi(\cdot)\) on \(U \times V\).

Moreover, for every \((x_1, x_2, y_1, y_2) \in Z\) one has
\[
\varepsilon_{\text{sad}}((x_1, y_1)\Psi_1, X_1, Y_1) \leq \varepsilon_{\text{sad}}((x, y)\Psi, X, Y), \quad i = 1, 2.
\]

Proof. Let \((x^*, y^*)\) be a saddle point of \(\Psi\) on \(X \times Y\). We have for \(\Psi_1\), that
\[
\Psi_1(x_1^*, y_1) = \min_{x_2 \in X_2} \max_{y_2 \in Y_2} \Psi(x_1^*, x_2, y_1, y_2) \leq \max_{y_2 \in Y_2} \Psi(x_1^*, x_2^*, y_1, y_2) \leq \Psi(x_1^*, y_1^*),
\]
and
\[
\Psi_2(x_2^*, y_2^*) = \min_{x_1 \in X_1} \max_{y_1 \in Y_1} \Psi(x_1, x_2^*, y_1^*, y_2) \geq \min_{x_1 \in X_1} \Psi(x_1, x_2^*, y_1, y_2) \geq \Psi(x_1^*, y_1^*),
\]
so that
\[
\forall (x_1 \in X_1, y_1 \in Y_1) : \Psi_1(x_1^*, y_1) \leq \Psi(x^*, y^*) \leq \Psi_2(x_2^*, y_2^*).\]

By the standard definition of saddle points as applied to \(\Psi_1\) it follows that \(\Psi_1(x_1^*, y_1^*) = \Psi(x^*, y^*)\), and thus that \(\text{SadVal}(\Psi_1) = \text{SadVal}(\Psi)\) which combines with (23) to imply that \((x_1^*, y_1^*) \in \text{SadSet}(\Psi_1)\). By “symmetric” reasoning, \((x_2^*, y_2^*) \in \text{SadSet}(\Psi_2)\) and \(\text{SadVal}(\Psi_2) = \text{SadVal}(\Psi)\). Thus (20) and (21) are proved.

Now let \((x_1, x_2, y_1, y_2) \in Z\). We have by (??)
\[
\varepsilon_{\text{sad}}((x_1, y_1)\Psi_1, X_1, Y_1) = \max_{\eta_1 \in Y_1} \Psi_1(x_1, \eta_1) - \min_{\xi_1 \in X_1} \Psi_1(\xi_1, y_1)
\]
\[
= \max_{\eta_1 \in Y_1} \min_{\xi_2 \in X_2} \max_{\eta_2 \in Y_2} \Psi(x_1, \xi_2, \eta_1, \eta_2)
\]
\[
- \min_{\xi_1 \in X_1} \min_{\xi_2 \in X_2} \max_{\eta_2 \in Y_2} \Psi(\xi_1, \xi_2, y_1, \eta_2)
\]
\[
\leq \max_{\eta_1 \in Y_1} \max_{\eta_2 \in Y_2} \Psi(\xi_1, \xi_2, \eta_1, \eta_2)
\]
\[
= \varepsilon_{\text{sad}}((x, y)\Psi, X, Y),
\]
and similarly for \(\Psi_2\). (22) is proved. ■

3.3 Recovering Approximate Solutions to the Master Problem: Goal and Assumptions

The goal. Consider a master saddle point problem (19) along with the induced problems (18). By Proposition 2 (specifically by (22)) we can easily extract good approximate solutions to each of the induced problems from a good approximate solution to the master problem. The question is: To what extent is the opposite true?

Specifically, assume we have at our disposal a first-order method capable of solving to within a desired accuracy one of the induced problems, say problem \((P)\), and our goal is to extract from this solution a good approximate solution to the master problem (and thus, by Proposition 2 to problem \((D)\) as well).

When and how could we achieve this goal?

We will demonstrate that this goal is achievable, provided that
the first order information used by the algorithm in question satisfies some
not too restrictive technical assumptions, and that
we have at our disposal, not only the approximate solution to \((P)\) to be
converted into approximate solutions to \((M)\) and \((D)\), but also an accuracy
certificate for this solution.

**Preliminaries.** We start with some technical issues. From now on, we make

**Assumption C.** \(\Psi(x_1, x_2, y_1, y_2)\) is not only continuous convex-concave
on \(Z = X \times Y\), but

- \(\Psi\) is differentiable in \(x_2 \in X_2\) whenever \(x_1 \in \text{int} X_1, y_1 \in \text{int} Y_1\) and \(y_2 \in Y_2\), the derivative being continuous in \(x \in \text{int} X_1 \times X_2\) for every \(y \in \text{int} Y_1 \times Y_2\);
- \(\Psi\) is differentiable in \(y_2 \in Y_2\) whenever \(x_1 \in \text{int} X_1, y_1 \in \text{int} Y_1\) and \(x_2 \in X_2\), the derivative being continuous in \(y \in \text{int} Y_1 \times Y_2\) for every \(x \in \text{int} X_1 \times X_2\).

Under Assumption C, for every point \((x = (x_1, x_2), y = (y_1, y_2)) \in Z\) with \((x_1, y_1) \in \text{int} X_1 \times \text{int} Y_1\), function \(\Psi(\cdot, y)\) at the point \(x\) admits a regular
subgradient \(\Psi'_x(x, y)\) — that is a subgradient whose projection onto the subspace \(\mathbb{R}^{n_2}\) equals the gradient of the continuously differentiable function \(\Psi(x_1, \cdot, y)\) at the point \(x_2\).

Indeed, let \(x_2^t \in \text{int} X_2\) be such that \(x_2^t \to x_2\) as \(t \to \infty\), and let \(g^t\) be a subgradient
of \(\Psi(\cdot, y)\) at the point \(x^t = (x_1, x_2^t) \in \text{int} X\). By evident reasons, such a subgradient is automatically regular. Besides this, the vectors \(g^t\) form a bounded sequence, since
\(\Psi(\xi_1, \xi_2, y)\) is uniformly in \(\xi_1 \in V\) Lipschitz continuous in \(\xi_2 \in X_2\), \(V\) being a
neighbourhood of \(x_1\) with the closure belonging to \(\text{int} X_1\), and is uniformly in \(\xi_2 \in X_2\) Lipschitz continuous in \(\xi_1 \in V\). Passing to a subsequence, we may assume that \(g^t\) has a limit \(g\) as \(t \to \infty\); by construction, \(g\) is a subgradient of \(\Psi(\cdot, y)\) at the point \(x\), and its projection on \(\mathbb{R}^{n_2}\) is \(\lim_{t \to \infty} \nabla s_{x^t} \Psi(x_1, s, y) = \nabla s_{x^t} \Psi(x_1, s, y)\),
as required from a regular subgradient.

We define similarly the notion of a regular supergradient \(\Psi'_y(x, y)\) in \(y\) (a supergradient of the concave function \(\Psi(x, \cdot)\) at the point \(y\) such that the projection of this supergradient onto \(\mathbb{R}^{n_2}\) is the gradient of \(\Psi(x, y_1, \cdot)\) at the point \(y_2\)). Such a supergradient also exists, provided that \((x_1, y_1) \in \text{int} X_1 \times \text{int} Y_1\), and \((x_2, y_2) \in X_2 \times Y_2\).

**Remark 1** Note that Assumption C is automatically satisfied when \(\Psi\) is convex-concave and continuously differentiable (as it is the case, e.g., when \(\Psi\) is bi-
linear in \(x, y\), cf. \([12], [14]\)). In this case, choosing as subgradients of \(\Psi\) w.r.t. \(x\) the corresponding partial gradients, and similarly for supergradients of \(\Psi\) in \(y\), we automatically end up with regular sub- and supergradients.

**Lemma 2** Given \((\bar{x}_1, \bar{y}_1) \in \text{int} X_1 \times \text{int} Y_1\), let \((\bar{x}_2, \bar{y}_2)\) be a saddle point of the
convex-concave continuous function \(\Psi_{x_1, y_1}(x_2, y_2) = \Psi(\bar{x}_1, x_2, \bar{y}_1, y_2)\) on \(X_2 \times Y_2\), and let \(\Psi'_x(\bar{x}), \Psi'_y(\bar{y})\) be regular sub- and supergradients of \(\Psi\) in \(x\) and in \(y\), respectively, computed at the point \(\bar{z} = (\bar{x}_1, \bar{x}_2, \bar{y}_1, \bar{y}_2)\). Let, further, \(\Psi'_{x_1}(\bar{x}_1, \bar{y}_1)\)
be the projection of $\Psi'_s(z)$ onto $\mathbb{R}^{n_1}$, and $\Psi'_{1,y}(\bar{x}_1, \bar{y}_1)$ be the projection of $\Psi'_y(z)$ onto $\mathbb{R}^{n_1}$. Then $\Psi'_{1,y}(\bar{x}_1, \bar{y}_1)$ is a subgradient of the convex function $\Psi_1(x_1, y_1)$ of $x_1 \in X_1$ at the point $x_1 = \bar{x}_1$, and $\Psi'_{1,y}(\bar{x}_1, \bar{y}_1)$ is a supergradient of the concave function $\Psi_1(x_1, y_1)$ of $y_1 \in Y_1$ at the point $y_1 = \bar{y}_1$.

**Proof.** For $x_1 \in X_1$, we have

$$\Psi_1(x_1, \bar{y}_1) = \min_{y_2 \in X_2} \max_{y_2 \in Y_2} \Psi(x_1, x_2, \bar{y}_1, y_2) \geq \min_{y_2 \in X_2} \Psi(x_1, x_2, \bar{y}) = \max_{y_2 \in Y_2} \Psi(x_1, x_2, \bar{y}) \geq \min_{x_2 \in X_2} \Psi(x_1, x_2, \bar{y}) \geq \Psi_1(\bar{x}_1, \bar{y}_1) + \min_{x_2 \in X_2} \left[ \Psi_1(x_1, \bar{y}_1, x_1 - \bar{x}_1) + \langle \nabla \Psi(x_1, \bar{y}_1, x_2 - x_2), y_2 \rangle \right]$$

since $\Psi(x, \bar{y})$ is convex.

The concluding inequality says that $\Psi'_s(z)$ is regular:

$$\Psi_1(x_1, \bar{y}_1) + \min_{x_2 \in X_2} \left[ \Psi_1(x_1, \bar{y}_1, x_1 - \bar{x}_1) \right]$$

The conclusion inequality says that $\Psi'_{1,y}(\bar{x}_1, \bar{y}_1)$ indeed is a subgradient of $\Psi_1(x_1, \bar{y}_1)$ in $x_1 \in X_1$ evaluated at $x_1 = \bar{x}_1$. The “symmetric” reasoning proves the “supergradient” part of the statement.

3.4 Recovering Approximate Solutions to the Master Problem: Construction and Main Result

In the above described situation, let us assume that we have access to Separation oracles for $X_1$ and $Y_1$ (and thus – to a Separation oracle for $Z_1 = X_1 \times Y_1$). We also assume that we have access to a $\Phi$-oracle, where $\Phi: \text{int} Z_1 \rightarrow \mathbb{R}^{n_1} \times \mathbb{R}^{n_1}$ is the monotone mapping associated with the convex-concave saddle point problem

$$\max_{x_1 \in X_1} \min_{y_1 \in Y_1} \Psi(x_1, y_1), \quad (24)$$

specifically, as follows:

Given on input $(x_1, y_1) \in \text{int} X_1 \times \text{int} Y_1$, the $\Phi$-oracle

- solves the saddle point problem

$$\min_{\xi_2 \in X_2} \max_{\eta_2 \in Y_2} \Psi(x_1, \xi_2, y_1, \eta_2) \quad (25)$$

and computes a saddle point $(x_2, y_2)$ of this problem, along with

- the projection $c_{x_2}$ of the regular subgradient of $\Psi(x_1, y_1, \eta_2)$ in $\xi \in X$ computed at the point $\xi = (x_1, x_2)$, onto the space $\mathbb{R}^{n_1}$;

- the projection $-c_{y_2}$ of the regular supergradient of $\Psi(x_1, x_2, \eta)$ in $\eta \in Y$ computed at the point $\eta = (y_1, y_2)$, onto the space $\mathbb{R}^{n_1}$;

- returns the pair $(x_2, y_2)$ and the vector $\Phi(x_1, y_1) = (c_{x_2}, c_{y_2})$, thus, by Lemma 2 reporting the value at $(x_1, y_1)$ of the monotone mapping associated with the saddle point problem (24).
Assume that we have built a \( \tau \)-point execution protocol \( P_{\tau} = \{ (z^t_1, e^t) \}_{t=1}^\tau \), where \( z^t_1 = (x^t_1, y^t_1) \) are the search points, partitioned into those which are strictly feasible \( (z^t_1 \in \text{int} Z_1 \Leftrightarrow t \in I_\tau) \) and all the remaining search points \( (z^t_1 \not\in \text{int} Z_1 \Leftrightarrow t \in J_\tau) \). Also assume that \( e^t \) is either \( \Phi(z^t_1) \) (this is so when \( t \in I_\tau \)), or \( e^t \) is a nonzero separator of \( z^t_1 \) and \( Z_1 \) (this is so when \( t \in J_\tau \)).

According to the construction of the \( \Phi \)-oracle, this protocol can be augmented with pairs \( z^t_2 = (x^t_2, y^t_2), t \in I_\tau, \) reported by the \( \Phi \)-oracle at the productive steps (those from \( I_\tau \)). Our main result is as follows:

**Theorem 1** Let \( B \) be a solid containing \( Z_1 \), and let \( P_{\tau} \) be an execution protocol which admits an accuracy certificate \( \zeta \). Given this certificate, let us set

\[
\bar{x}^\tau_i = \sum_{t \in I_\tau} \zeta_i x^t_1, \quad \bar{y}^\tau_i = \sum_{t \in I_\tau} \zeta_i y^t_1, \quad i = 1, 2.
\]

Then \( (\bar{x}^\tau, \bar{y}^\tau) \in Z \) and

\[
\varepsilon_{\text{sad}}((\bar{x}^\tau, \bar{y}^\tau), \Psi, X, Y) \leq \varepsilon_{\text{cert}}(\zeta | P_{\tau}, B), \quad (26)
\]

whence, by Proposition 2 also

\[
\varepsilon_{\text{sad}}((\bar{x}^\tau, \bar{y}^\tau), \Psi, X_1, Y_1) \leq \varepsilon_{\text{cert}}(\zeta | P_{\tau}, B), \quad i = 1, 2. \quad (27)
\]

**Proof.** For \( t \in I_\tau \), let \( z^t = (x^t_1, x^t_2, y^t_1, y^t_2) = (x^t, y^t) \). Recall that for \( t \in I_\tau \), \( e^t_i \) is the projection onto \( \mathbb{R}^{n+1} \) of a subgradient \( \Psi'_i(x^t, y^t) \) of the function \( \Psi_i(y^t) \) computed at the point \( x^t \), and the projection \( e^t_{x_2} \) of this subgradient onto \( \mathbb{R}^{n+2} \) is the vector \( \nabla x_2 \bigg|_{x_2 = z^t} \Psi(x^t_1, x_2, y^t_1, y^t_2) \), whence

\[
\langle e^t_{x_2}, x_2 - x^t_2 \rangle \geq 0 \quad \forall x_2 \in X_2, \quad (28)
\]

due to the fact that \( \Psi(x^t_1, x_2, y^t_1, y^t_2) \) attains its minimum over \( x_2 \in X_2 \) at the point \( x^t_2 \). Similarly, for \( t \in I_\tau \), \( e^t_y \) is the projection onto \( \mathbb{R}^{n+2} \) of a subgradient \( -\Psi'_y(x^t, y^t) \) of the function \( -\Psi(x^t, \cdot) \) computed at the point \( y^t \), and the projection \( e^t_y \) of this subgradient onto \( \mathbb{R}^{n+2} \) is the vector \( -\nabla y_2 \bigg|_{y_2 = y^t} \Psi(x^t_1, x^t_2, y^t_1, y^t_2) \), whence

\[
\langle e^t_y, y_2 - y^t_2 \rangle \geq 0 \quad \forall y_2 \in Y_2. \quad (29)
\]

Let \( z = (x_1, x_2, y_1, y_2) = (x, y) \in Z \). We have

\[
-\varepsilon_{\text{cert}}(\zeta | P_{\tau}, B) \leq \sum_{t=1}^\tau \zeta_i \langle e^t_i, z - z^t \rangle
\]
\[
\leq \sum_{t \in I_\tau} \zeta_i \langle e^t_i, z - z^t \rangle \quad \text{[since \( e^t_i \) separates \( Z \) and \( z^t \) for \( t \in J_\tau \)]}
\]
\[
= \sum_{t \in I_\tau} \zeta_i \left[ \langle e^t_{x_2}, x_2 - x^t_2 \rangle + \langle e^t_y, y_2 - y^t_2 \rangle \right]
\]
\[
\leq \sum_{t \in I_\tau} \zeta_i \left[ \| \Psi(x^t, y^t) - \Psi(x^t_1, x^t_2, y^t_1, y^t_2) \| + \| \Psi(x^t, y^t) - \Psi(x^t_1, x^t_2, y^t_1, y^t_2) \| \right]
\]
\[
= \sum_{t \in I_\tau} \zeta_i \left[ \| \Psi(x, y) - \Psi(x^t, y^t) \| + \| \Psi(x, y) - \Psi(x^t, y^t) \| \right]
\]
\[
\leq \Psi(x, \bar{y}^\tau) - \Psi(\bar{x}^\tau, y) \quad \text{[since \( \Psi \) is convex-concave]}
\]

Thus,

\[
\forall (x \in X, y \in Y) : \Psi(\bar{x}^\tau, y) - \Psi(\bar{x}^\tau, y) \leq \varepsilon_{\text{cert}}(\zeta | P_{\tau}, B).
\]

Taking the supremum of the left hand side in \( (x, y) \in Z \), we arrive at \( 26 \).
References

1. Cox, Juditsky, Nemirovski, 'Dual subgradient algorithms for large-scale nonsmooth learning problems', Mathematical Programming, Series. B (2014) 148:143180
2. Auslender, A., Optimization Methodes Numeriques, Mason, Paris, 1976.
3. A. Ben-Tal and A. Nemirovski, Lectures on Modern Convex Optimization, SIAM, Philadelphia, 2001.
4. Ben-Tal, A., Margalit, T., and Nemirovski, A., The Ordered Subsets Mirror Descent optimization method with applications to Tomography. SIAM J. Optim. 12 (2001), 79-108.
5. A. Ben-Tal, A. Nemirovski, Non-Euclidean restricted memory level method for large-scale convex optimization. Math. Progr. 102, 407–456, 2005.
6. D. Bertsimas, J.N. Tsitsiklis, Introduction to Linear Optimization. Athena Scientific, 1997.
7. S. Boyd and L. Vandenberghe, Convex Optimization, Cambridge University Press, 2004.
8. V.P. Bulatov and L.O. Shepot’ko, Method of centers of orthogonal simpexes for solving convex programming problems (in Russian), In: Methods of Optimization and Their Application, Nauka, Novosibirsk 1982.
9. D. Gabelev (2003), Polynomial time cutting plane algorithms associated with symmetric cones, M.Sc. Thesis in OR and Systems Analysis, Faculty of Industrial Engineering and Management, Technion – Israel Institute of Technology, Technion City, Haifa 32000, Israel E-print: http://www2.isye.gatech.edu/~nemirovs/Dima.pdf
10. Francisco Facchinei and Jong-Shi Pang, Finite-Dimensional Variational Inequalities and Complementarity Problems I, Springer-Verlag, New York, 2003.
11. Fritz John, Extremum problems with inequalities as subsidiary conditions, in: Studies and Essays Presented to R. Courant on his 60th Birthday, January 8, 1948, Interscience Publishers, Inc., New York, N. Y., 1948, pp. 187–204.
12. E.G. Gol’shtein, Primal-dual vlock method for Linear Programming. Automation and Remote Control No. 11, 1996.
13. M. Gröttschel, L. Lovasz and A. Schrijver, The Ellipsoid Method and Combinatorial Optimization. Springer-Verlag, 1986.
14. K. Holmberg, On the convergence of cross decomposition, Math. Progr. 47:2 (1990), 269-296.
15. K. Holmberg, Linear mean value cross decomposition: A generalization of the Kornai-Liptak method, European Journal of Operational Research 62:1 (1992), 55-73.
16. K. Holmberg, A convergence proof for linear mean value cross decomposition, Zeitschrift für Operations Research 39:2 (1994), 157-186.
17. K. Holmberg, Mean value cross decomposition applied to integer programming problems, European Journal of Operational Research 97:1 (1997), 124-138.
18. C. Lemaréchal, Nonsmooth optimization and descent methods. Research Report 78-4, IIASA, Laxenburg, Austria, 1978.
19. K. Kiwiel, Proximal level bundle method for convex nondifferentable optimization, saddle point problems and variational inequalities. Math. Progr. Series B 69 (1995), 89-109.
20. K.C. Kiwiel, T. Larson, P.O. Lindberg, The efficiency of ballstep subgradient level methods for convex optimization. Math. of Oper. Res. 24 (1999), 237-254.
21. C. Lemaréchal, J.J. Strodiot, and A. Bihain, On a bundle algorithm for nonsmooth optimization. In: O.L. Mangasarian, R.R. Meyer, S.M. Robinson, Eds., Nonlinear Programming 4 (Academic Press, NY, 1981), 245-282.
22. C. Lemaréchal, A. Nemirovski, and Yu. Nesterov, New variants of bundle methods. Math. Progr. 69:1, 111-148, 1995.
23. A.Yu. Levin, On an algorithm for the minimization of convex functions (in Russian), Doklady Akad. Nauk SSSR 160:6 (1965), 1244–1247. (English translation: Soviet Mathematics Doklady, 6:286–290, 1965.)
24. R. Mifflin, A modification and an extension of Lemaréchal’s algorithm for nonsmooth minimization. Math. Progr. Study 17 (1982), 77-90.
25. A. Nemirovskii and D. Yudin, Efficient methods for large-scale convex problems. (in Russian) - Ekonomika i Matematiskie Metody, 15:1 (1979) (the journal is translated into English as Matekon)
26. A. Nemirovskii, Efficient iterative algorithms for variational inequalities with monotone operators. (in Russian) - *Ekonomika i Matematicheskie Metody*, 17:2 (1981), 344-359 (the journal is translated into English as *Matekon*).

27. A. Nemirovskii and D. Yudin, Problem Complexity and Method Efficiency in Optimization, Wiley & Son, 1983.

28. A. Nemirovski, S. Onn and U.G. Rothblum, Accuracy certificates for computational problems with convex structure. *Math. of Oper. Res.* 35 (2010), 52-78.

29. A. Nemirovski, Polynomial time methods in Convex Programming, in: J. Renegar, M. Shub and S. Smale, Eds., *The Mathematics of Numerical Analysis*, AMS-SIAM Summer Seminar on Mathematics in Applied Mathematics, July 17–August 11, 1995, Park City, Utah. Lectures in Applied Mathematics, AMS, Providence, 32:543-589, 1996.

30. A. Nemirovski, Prox-method with rate of convergence $O(1/t)$ for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems, *SIAM J. on Optim.* 15, 229–251, 2004.

31. Y. Nesterov and A. Nemirovskii, *Interior-Point Polynomial Algorithms in Convex Programming*, SIAM, Philadelphia, 1994.

32. Yu. Nesterov, *Introductory Lectures on Convex Optimization. A Basic Course*. Kluwer Academic Press, 2004.

33. Yu. Nesterov, Semidefinite relaxation and nonconvex quadratic minimization – *Optimization and Software* 9 (1998), 141-160.

34. Yu. Nesterov, Smooth minimization of non-smooth functions, *Math. Progr.* 103:1, 127–152, 2005.

35. D.J. Newman, Location of maximum on unimodal surfaces, *Journ. of the Assoc. for Computing Machinery* 12:11-23, 1965.

36. R.T. Rockafellar, *Convex Analysis*, Princeton University press, Princeton, New Jersey, 1970.

37. Han-Suk Sohn and D. Bricker, Cross-Decomposition of the Degree-Constrained Minimum Spanning Tree Problem, *Journal of Systemics, Cybernetics and Informatics* 5:1 (2007).

38. B.T. Polyak, A general method for solving extremal problems. *Soviet Math. Doklady* 174 (1967), 33-36.

39. N.Z. Shor, Generalized gradient descent with application to block programming. *Kibernetika* 1967 No. 3 (in Russian).

40. H. Schramm and J. Zowe, A version of bundle idea for minimizing a non-smooth function: conceptual idea, convergence analysis, numerical results. *SIAM J. Optim.* 2 (1992), 121-152.

41. N.Z. Shor, Cutting plane method with space dilation for the solution of convex programming problems (in Russian), *Kibernetika*, 1977 No. 1, 94-95.

42. S.P. Tarasov, L.G. Khachiyan and I.I. Erlikh, The method of inscribed ellipsoids, *Soviet Mathematics Doklady* 37 (1988), 226–230.

43. T. J. Van Roy, Cross decomposition for mixed integer programs, *Mathematical Programming* 25 (1983), 46-63.

44. T. J. Van Roy, A Cross decomposition algorithm for capacitated facility location, *Operations Research* 34 (1986), 145-163.

45. B. Yamnitsky and L. Levin, An Old Linear Programming Algorithm Runs in Polynomial Time, In: 23rd Annual Symposium on Foundations of Computer Science, IEEE, New York, 327–328, 1982.

46. Chun-Beon Yooa and Dong-Wan Tchaa, A cross decomposition procedure for the facility location problem with a choice of facility type, *Computers & Industrial Engineering* 10:4 (1986), 283-290.

47. D. Yudin and A. Nemirovskii, Informational complexity and effective methods of solution for convex extremal problems (in Russian), *Ekonomika i Matematicheskie Metody* 12:2 (1976), 357–369 (translated into English as *Matekon*: Transl. Russian and East European Math. Economics 13 25–45, Spring ’77).