Noname manuscript No.
(will be inserted by the editor)

Difference of convex algorithms for bilevel programs with applications in hyperparameter selection

Jane J. Ye · Xiaoming Yuan · Shangzhi Zeng · Jin Zhang

Received: date / Accepted: date

Abstract In this paper, we present difference of convex algorithms for solving bilevel programs in which the upper level objective functions are difference of convex functions, and the lower level programs are fully convex. This nontrivial class of bilevel programs provides a powerful modelling framework for dealing with applications arising from hyperparameter selection in machine learning. Thanks to the full convexity of the lower level program, the value function of the lower level program turns out to be convex and hence the bilevel program can be reformulated as a difference of convex bilevel program. We propose two algorithms for solving the reformulated difference of convex program and show their convergence under very mild assumptions. Finally we conduct numerical experiments to a bilevel model of support vector machine classification.

This paper is dedicated to the memory of Olvi L. Mangasarian.

The research of the first author was partially supported by NSERC. The second author was supported by a General Research Fund from Hong Kong Research Grants Council. The last author was supported by NSFC No. 11971220 and Guangdong Basic and Applied Basic Research Foundation 2019A1515011152.

Jane J. Ye
Department of Mathematics and Statistics, University of Victoria, Canada.
E-mail: janeye@uvic.ca

Xiaoming Yuan
Department of Mathematics, The University of Hong Kong, Hong Kong SAR, China.
E-mail: xmyuan@hku.hk

Shangzhi Zeng
Department of Mathematics, The University of Hong Kong, Hong Kong SAR, China.
E-mail: zengsz@connect.hku.hk

Jin Zhang
Department of Mathematics, Southern University of Science and Technology, National Center for Applied Mathematics Shenzhen, Shenzhen, China.
E-mail: zhangj9@sustech.edu.cn

Keywords Bilevel program · difference of convex algorithm · hyperparameter selection

Mathematics Subject Classification (2010) 90C26 · 90C30

1 Introduction

Bilevel programs are a class of hierarchical optimization problems which have constraints containing a lower-level optimization problem parameterized by upper-level variables. Bilevel programs capture a wide range of important applications in various fields including Stackelberg games and moral hazard problems in economics ([37,26]), hyperparameter selection and meta learning in machine learning ([18–20,14,23,24,27,28,31]). More applications can be found in the monographs [3,10,13,36], the survey on bilevel optimization [9,12] and the references within.

In this paper, we develop some numerical algorithms for solving the following difference of convex (DC) bilevel program:

$$\text{(DCBP)} \qquad \begin{array}{ll} \min & F(x,y) := F_1(x,y) - F_2(x,y) \\ s.t. & x \in X, y \in S(x), \end{array}$$

with $S(x)$ being the set of optimal solutions of the lower level problem,

$$(P_x): \quad \min_{y \in Y} \ f(x,y) \quad s.t. \ g(x,y) \leq 0,$$

where $X \subseteq \mathbb{R}^n$ and $Y \subseteq \mathbb{R}^m$ are nonempty closed sets, $g := (g_1, \ldots, g_l)$, all functions $g_i : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$, $i = 1, \ldots, l$ are convex on an open convex set containing the set $X \times Y$, and the functions $F_1, F_2, f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ are convex on an open convex set containing the set

$$C := \{(x,y) \in X \times Y : g(x,y) \leq 0\}.$$

To ensure the bilevel program is well-defined, we assume that $S(x) \neq \emptyset$ for all $x \in X$. Moreover we assume that for all x in an open convex set $\mathcal{O} \supseteq X$, the feasible region for the lower level program $\mathcal{F}(x) := \{y \in Y : g(x,y) \leq 0\}$ is nonempty and the lower level objective function $f(x,y)$ is bounded below on $\mathcal{F}(x)$. Note that all of our results applied to the case where there are also some upper level constraint $G_i(x,y) \leq 0, i = 1, \ldots, k$ with each defining function G_i a difference of convex function. For simplicity of notations, in order to concentrate on the main idea, we choose not to include them here.

Although the objective function in the DC bilevel program we consider must be a DC function, this setting is general enough to capture many cases of practical interests. In particular any lower C^2 function (i.e., a function which can be locally written as a supremum of a family of C^2 functions) and C^{1+} function (i.e., a differentiable function whose gradient is locally Lipschitz continuous) are DC functions and the class of DC functions is closed under many operations encountered frequently in optimization; see, e.g., [17,40]. In the lower level program, we assume all functions are fully convex, i.e., convex in both variables x and y. However as pointed out by [21, Example 1 and Section 5], using some suitable reformulations one may turn a non-fully convex lower level program into a fully convex one. Also as demonstrated in this paper, the bilevel model for hyperparameter selection problem can be reformulated as a bilevel program where the lower level is fully convex.

Solving bilevel programs numerically is extremely challenging. It is known that even when all defining functions are linear, the computational complexity is already NP-hard [4]. If all defining functions are smooth and the lower level program is convex with respect to the lower level variable, the first order approach was popularly used to replace the lower level problem by its first order optimality condition and to solve the resulting problem as the mathematical program with equilibrium constraints (MPEC); see e.g. [3,1,12,25,33]. The first order approach may be problematic since it may not provide an equivalent reformulation to the original bilevel program if only local (not global) optimal solutions are considered; see [11]. Moreover even in the case of a fully convex lower level program, [21, Example 1] shows that it is still possible that a local optimal solution of the corresponding MPEC does not correspond to a local optimal solution of the original bilevel program. Recently some numerical algorithms have been introduced for solving bilevel programs where the lower level problem is not necessarily convex in the lower level variable; see e.g., [22,29,30]. However these approaches have limitations in the numbers of variables in the bilevel program. In most of literature on numerical algorithms for solving bilevel programs, smoothness of all defining functions are assumed. In some special cases, non-smoothness can be dealt with by introducing auxiliary variables and constraints to reformulate a nonsmooth lower level program as a smooth constrained lower level program. But using such an approach the numbers of variables or constraints would increase.

Our research on the DC bilevel program is motivated by a number of important applications in model selection and hyperparameter learning. Recently in the statistical learning, the regularization parameters has been successfully used, e.g., in the least absolute shrinkage and selection operator (lasso) method for regression and support vector machines (SVMs) for classification. However the regularization parameters have to be set *a priori* and the choice of these parameters dramatically affects results on the model selection. The most commonly used method for selecting these parameters is the so-called T-fold cross validation. By T-fold cross validation, a data set Ω is partitioned into T pairwise disjoint subsets called the validation sets Ω_{val}^t, $t = 1, \ldots, T$. For each fold $t = 1, \ldots, T$, a subset of Ω denoted by $\Omega_{trn}^t := \Omega \backslash \Omega_{val}^t$ is used for training and the validation set Ω_{val}^t is used for testing the result. Take the lasso problem for example, suppose the data set $\Omega = \{(\mathbf{a}_j, b_j)\}_{j=1}^{\ell}$ where $\mathbf{a}_j \in \mathbb{R}^n$ and $b_j \in \mathbb{R}$. For each $\lambda > 0$ and each $t = 1, \ldots, T$, the following lasso problem can be solved.

$$(P_\lambda^t) \qquad \min_{\mathbf{w}\in\mathbb{R}^n} \left\{ \sum_{j\in\Omega_{trn}^t} (\mathbf{a}_j^T\mathbf{w} - b_j)^2 + \lambda\|\mathbf{w}\|_1 \right\}.$$

The desirable penalty parameter $\bar{\lambda}$ can be selected by minimizing the cross validation error based on the validation set:

$$\Theta(\mathbf{w}^1,\ldots,\mathbf{w}^T) := \frac{1}{T}\sum_{t=1}^T \frac{1}{|\Omega_{val}^t|} \sum_{j\in\Omega_{val}^t} (\mathbf{a}_j^T\mathbf{w}_\lambda^t - b_j)^2,$$

where $|M|$ denotes the number of elements in set M and $\mathbf{w}_\lambda^t$ denotes a solution to the lasso problem (P_λ^t). In fact, the hyperparameter selection for the lasso problem can be considered as the following bilevel program with T lower level programs:

$$\begin{aligned} \min\ & \Theta(\mathbf{w}^1,\ldots,\mathbf{w}^T) \\ s.t.\ & \lambda > 0, \mathbf{w}^t \in \mathrm{argmin}_{\mathbf{w}} \left\{ \sum_{j\in\Omega_{trn}^t} (\mathbf{a}_j^T\mathbf{w} - b_j)^2 + \lambda\|\mathbf{w}\|_1 \right\}, t = 1,2,\ldots,T. \end{aligned}$$

It is easy to see that we can reformulate the above problem equivalently as the following bilevel program with a single lower level program

$$\begin{aligned} \min\ & \Theta(\mathbf{w}^1,\ldots,\mathbf{w}^T) \\ s.t.\ & \lambda > 0, (\mathbf{w}^1,\ldots,\mathbf{w}^T) \in \mathrm{argmin} \left\{ \sum_{t=1}^T \sum_{j\in\Omega_{trn}^t} \frac{(\mathbf{a}_j^T\mathbf{w}^t - b_j)^2}{\lambda} + \|\mathbf{w}^t\|_1 \right\}. \end{aligned}$$

Moreover using the fact that a function in the form $\phi(\mathbf{x},\lambda) = \|\mathbf{x}\|^2/\lambda$ with $\lambda > 0$ is convex as a perspective function [38, Example 3.18], the above bilevel program has a fully convex lower level program. The classical T-fold cross validation method for selecting hyperparameters usually implements a *grid search*: training T models at each point of a discretized parameter space in order to find an approximate optimal parameter. This method has many drawbacks and limitations. In particular its computational complexity scales exponentially with the number of hyperparameters and the number of grid points for each hyperparameter and hence is not practical for problem requiring several hyperparameters. To deal with limitations of grid search, introduced first in [5] in 2006, the bilevel program has been used to model hyperparameter selection problems in [5,18–20,28,27]. We now consider the bilevel model for support vector classification using cross validation studied in [18,19]. Given a training set Ω containing ℓ labelled data pairs $\{(\mathbf{a}_j, b_j)\}_{j=1}^\ell$ where $\mathbf{a}_j \in \mathbb{R}^n$, and the labels $b_j = \pm 1$ indicate the class membership. Then the bilevel model for support vector (SV) classification is formulated in [18,19] as follows:

$$\begin{aligned}
\min \quad & \Theta(\mathbf{w}^1, \ldots, \mathbf{w}^T, \mathbf{c}) \\
s.t. \quad & \lambda_{lb} \le \lambda \le \lambda_{ub}, \quad \bar{\mathbf{w}}_{lb} \le \bar{\mathbf{w}} \le \bar{\mathbf{w}}_{ub}, \\
& \text{and for } t = 1, \ldots, T: \\
& (\mathbf{w}^t, c_t) \in \operatorname*{argmin}_{\substack{-\bar{\mathbf{w}} \le \mathbf{w} \le \bar{\mathbf{w}} \\ c \in \mathbb{R}}} \left\{ \frac{\lambda}{2} \|\mathbf{w}\|^2 + \sum_{j \in \Omega^t_{trn}} \max(1 - b_j(\mathbf{a}_j^T \mathbf{w} - c), 0) \right\},
\end{aligned} \tag{1}$$

where $\mathbf{c} \in \mathbb{R}^T$ is the vector with c_t as the tth component, and $\Theta(\mathbf{w}^1, \ldots, \mathbf{w}^T, \mathbf{c})$ is some measure of validation accuracy over all folds, typically the average number of misclassifications. Here $\lambda_{lb}, \lambda_{ub}$ are given nonnegative numbers and $\bar{\mathbf{w}}_{lb}, \bar{\mathbf{w}}_{ub}$ are given vectors in $\mathbb{R}^n$. By changing the variable λ to $\mu := \frac{1}{\lambda}$ we can equivalently rewrite the bilevel model for problem (1) so that there is only a single lower level problem and it is fully convex; see the details in (32).

This paper is motivated by an interesting fact that, under our problem setting, the value function of the lower level in (DCBP) defined by

$$v(x) := \inf_{y \in Y} \{ f(x,y) \ s.t. \ g(x,y) \le 0 \}$$

is convex and locally Lipschitz continuous on X. We take full advantage of this convexity and use the value function approach first proposed in [32] for a numerical purpose and further used to study optimality conditions in [44] to reformulate (DCBP) as the following DC program:

$$\text{(VP)} \qquad \begin{aligned} \min_{(x,y) \in C} \quad & F_1(x,y) - F_2(x,y) \\ s.t. \quad & f(x,y) - v(x) \le 0. \end{aligned}$$

Unfortunately, due to the value function constraint, (VP) violates the usual constraint qualification such as the nonsmooth Mangasarian Fromovitz constraint qualification (MFCQ) at each feasible point, see [44, Proposition 3.2] for the smooth case and Proposition 9 for the nonsmooth case. It is well-known that convergence of the difference of convex algorithm (DCA) is only guaranteed under constraint qualifications such as the extended MFCQ, which is MFCQ extended to infeasible points; see, e.g., [39]. To deal with this issue, we consider the following approximate bilevel program

$$\text{(VP)}_\epsilon \qquad \begin{aligned} \min_{(x,y) \in C} \quad & F_1(x,y) - F_2(x,y) \\ s.t. \quad & f(x,y) - v(x) \le \epsilon, \end{aligned}$$

for some $\epsilon > 0$. Such a relaxation strategy has been used for example in [22] based with the reasoning that in numerical algorithms one usually obtain an inexact optimal solution anyway and the solutions of (VP)_ϵ approximate a solution of the original bilevel program (VP) as ϵ approaches zero. In this paper we will show that EMFCQ holds for problem $(\text{VP})_\epsilon$ when $\epsilon > 0$ automatically. Hence we propose to solve problem $(\text{VP})_\epsilon$ with $\epsilon \geq 0$. When $\epsilon > 0$, the convergence of our algorithm is guaranteed and when $\epsilon = 0$, the convergence is not guaranteed but it could still converge if the penalty parameter sequence is bounded.

Using DCA approach, at each iteration point (x^k, y^k), one linearises the concave part of the function, i.e., the functions $F_2(x, y), v(x)$ by using an element of the subdifferentials $\partial F_2(x^k, y^k), \partial v(x^k)$ and solve a resulting convex subproblem. The value function is an implicit function. How do we obtain an element of the subgradient $\partial v(x^k)$? At current iteration x^k, assuming we can solve the lower level problem (P_{x^k}) with a global minimizer $\tilde{y}^k$ and a corresponding Karush-Kuhn-Tucker (KKT) multiplier denoted by γ^k. Suppose that the following partial derivative formula holds:

$$\partial f(x, y) = \partial_x f(x, y) \times \partial_y f(x, y), \;\; \partial g_i(x, y) = \partial_x g_i(x, y) \times \partial_y g_i(x, y) \quad (2)$$

at $(x, y) = (x^k, \tilde{y}^k)$. Then since by convex analysis

$$\partial_x f(x^k, \tilde{y}^k) + \sum_{i=1}^{l} \gamma_i^k \partial_x g_i(x^k, \tilde{y}^k) \subseteq \partial v(x^k),$$

we can select an element of $\partial v(x^k)$ from the set $\partial_x f(x^k, \tilde{y}^k) + \sum_{i=1}^{l} \gamma_i^k \partial_x g_i(x^k, \tilde{y}^k)$ and use it to linearize the value function. We then solve the resulting convex subproblem approximately to obtain a new iterate (x^{k+1}, y^{k+1}). Thanks to recent developments in large-scale convex programming, using this approach we can deal with a large scale DC bilevel program.

Now we summarize our contributions as follows.

- We propose two new algorithms for solving DC program. These algorithms have modified the classical DCA in two ways. First, we add an proximal term in each convex subproblem so the the objective function is strongly convex and at each iteration point, only an approximate solution for the convex subproblem is solved. Second, our penalty parameter update is simplier.
- We have laid down all theoretical foundations from convex analysis that are required for our algorithms to work. In particular we have demonstrated that under the minimal assumptions that we specify for problem (DCBP), the value function is convex and locally Lipschitz on set X automatically.
- Using the two new algorithms for solving DC program, we propose two corresponding algorithms to solve problem (DCBP). Our algorithms hold under very mild and natural assumptions. In particular we allow all defining functions to be nonsmooth and we do not require any constraint qualification to hold for the lower level program. The main assumptions we need are only the partial derivative formula (2) which holds under many practical situations (see Proposition 2 for sufficient conditions) and the existence of an KKT multiplier for the lower level program under each iteration.

– Taking advantage of large scale convex programming, our algorithm can handle high dimensional hyperparameter selection problems. To test effectiveness of our algorithm, we have tested it in the bilevel model for SV classification (1). Our results compare favourably with the MPEC approach [18–20].

This paper is organized as follows. In Section 2 we give preliminaries from convex analysis. We propose two modified DCA and study their convergence for a class of general DC program in Section 3. In Section 4, we derive explicit conditions for the bilevel program under which the algorithms introduced in Section 3 can be applied. Numerical experiments on the bilevel model for SV classification is conducted on Section 5. Section 6 concludes the paper.

2 Preliminaries

Let us recall some notations from convex analysis and variational analysis, which will be needed thereafter. Let $\varphi(x) : \mathbb{R}^n \to [-\infty, \infty]$ be a convex function, and let $\bar{x}$ be a point where φ be finite. Recall that the subdifferential of φ at $\bar{x}$ is a closed convex set defined by

$$\partial\varphi(\bar{x}) := \{\xi \in \mathbb{R}^n |\ \varphi(x) \geq \varphi(\bar{x}) + \langle \xi, x - \bar{x} \rangle,\ \forall x\}.$$

For a function $\varphi : \mathbb{R}^n \times \mathbb{R}^m \to [-\infty, +\infty]$, we denote its partial differential of φ with respect to x and y by $\partial_x \varphi(x, y)$ and $\partial_y \varphi(x, y)$ respectively. Let Σ be a convex subset in $\mathbb{R}^n$ and $\bar{x} \in \Sigma$. The normal cone to Σ at $\bar{x}$ is a closed convex set defined by $\mathcal{N}_\Sigma(\bar{x}) := \{\xi \in \mathbb{R}^n | \langle \xi, x - \bar{x} \rangle \geq 0,\ \forall x \in \Sigma\}$. When $\bar{x} \notin \Sigma$, we let $\mathcal{N}_\Sigma(\bar{x}) = \emptyset$. Let $\delta_\Sigma(x)$ denote the indicator function of set Σ at x, i.e., $\delta_\Sigma(x) := 0$ if $x \in \Sigma$ and $+\infty$ if $x \notin \Sigma$. Then by definition, it is easy to see that $\partial\delta_\Sigma(\bar{x}) = \mathcal{N}_\Sigma(\bar{x})$. It is well-known that when the function $\varphi : \mathbb{R}^n \to [-\infty, \infty]$ is Lipschitz continuous near a point $\bar{x}$ where $\varphi(\bar{x})$ is finite, the Clarke generalized gradient denoted by $\partial^c \varphi(\bar{x})$ coincides with the convex subdifferential; see [7].

Recall that a convex function $\varphi(x) : \mathbb{R}^n \to [-\infty, \infty]$ is proper if $\varphi(x) < +\infty$ for at least one x and $\varphi(x) > -\infty$ for every x. The following calculus rules will be useful.

Proposition 1 (Sum rule) *[34, Theorem 23.8][7, Corollary 1 to Theorem 2.9.8] Let $\varphi_1, \ldots, \varphi_m : \mathbb{R}^n \to [-\infty, \infty]$ be proper convex functions. Suppose that $\bar{x}$ is a point where $\varphi_1(\bar{x}) + \cdots + \varphi_m(\bar{x})$ is finite. Then*

$$\partial\varphi_1(\bar{x}) + \cdots + \partial\varphi_m(\bar{x}) \subseteq \partial(\varphi_1 + \cdots + \varphi_m)(\bar{x}).$$

Moreover if all except one function among $\varphi_i (i = 1, \cdots, m)$ is Lipschitz continuous around point $\bar{x}$, then the inclusion (1) becomes an equality.

Proposition 2 (Partial subdifferentiation) *Let $\varphi : \mathbb{R}^n \times \mathbb{R}^m \to [-\infty, +\infty]$ be a convex function and let $(\bar{x}, \bar{y})$ be a point where φ is finite. Then*

$$\partial\varphi(\bar{x}, \bar{y}) \subseteq \partial_x \varphi(\bar{x}, \bar{y}) \times \partial_y \varphi(\bar{x}, \bar{y}). \tag{3}$$

The inclusion (3) becomes an equality under one of the following conditions.

(a) *For every* $\xi \in \partial_x \varphi(\bar{x}, \bar{y})$*, it holds that* $\varphi(x, y) - \varphi(\bar{x}, y) \geq \langle \xi, x - \bar{x} \rangle$*,* $\forall (x, y) \in \mathbb{R}^n \times \mathbb{R}^m$*.*

(b) $\varphi(x, y) = \varphi_1(x) + \varphi_2(y)$*.*

(c) *For any* $\varepsilon > 0$*, there is* $\delta > 0$ *such that*

$$\text{either} \quad \partial_x \varphi(\bar{x}, \bar{y}) \subseteq \partial_x \varphi(\bar{x}, y) + \varepsilon B_{\mathbb{R}^n} \quad \forall y \in B(\bar{y}; \delta) \tag{4}$$

$$\text{or} \quad \partial_y \varphi(\bar{x}, \bar{y}) \subseteq \partial_y \varphi(x, \bar{y}) + \varepsilon B_{\mathbb{R}^m} \quad \forall x \in B(\bar{x}; \delta), \tag{5}$$

where $B(\bar{x}; \delta)$ *denotes the open ball centered at* $\bar{x}$ *with radius equal to* δ *and* $B_{\mathbb{R}^n}$ *denotes the open unit ball centered at the origin in* $\mathbb{R}^n$*.*

(d) $\varphi(x, y)$ *is continuously differentiable respect to one of the variable* x *or* y *at* $(\bar{x}, \bar{y})$*.*

Moreover $(b) \Longrightarrow (a), (d) \Longrightarrow (c) \Longrightarrow (a)$*.*

Proof The inclusion (3) and its reverse under (a) follow directly from definitions of the convex subdifferential and the partial subdifferential. When $\varphi(x, y) = \varphi_1(x) + \varphi_2(y)$, we have that $\partial \varphi(x, y) = \partial \varphi(x) \times \{0\} + \{0\} \times \partial \varphi(y)$. Hence obviously (b) implies (a). The implication of (d) to (c) is obvious. Now suppose that (4) holds. Let $\xi \in \partial_x \varphi(\bar{x}, \bar{y})$. Then according to (4), for any $\varepsilon > 0$, there is $\delta > 0$ such that $\xi = \eta + \varepsilon e$, where $e \in B_{\mathbb{R}^n}$, and

$$\langle \xi, x - \bar{x} \rangle \leq \varphi(x, y) - \varphi(\bar{x}, y) + \varepsilon \|x - \bar{x}\| \quad \forall y \in B(\bar{y}; \delta).$$

Thanks to the convexity of φ, using the proof technique of [8, Corollary 2.6 (c)], we can easily show that (a) holds. The proof for the case where (5) holds is similar and thus omitted. □

Proposition 3 *[34, Theorem 10.4] Let* $\varphi : \mathbb{R}^n \to [-\infty, \infty]$ *be a proper convex function. Then* φ *is Lipschitz continuous near any point* $\bar{x} \in \mathrm{int}(\mathrm{dom}\varphi)$*, where* $\mathrm{dom}\varphi := \{x | \varphi(x) < +\infty\}$ *is the effective domain of* φ *and* $\mathrm{int}(M)$ *is the interior of set* M*.*

3 Modified inexact proximal DC algorithms

In order to solve the (relaxed) value function reformulation of problem DCBP, in this section we propose numerical algorithms to solve the following difference of convex program:

$$\begin{aligned} (\mathrm{DC}) \qquad & \min_{z \in \Sigma} \; f_0(z) := g_0(z) - h_0(z) \\ & \; s.t. \;\; f_1(z) := g_1(z) - h_1(z) \leq 0, \end{aligned}$$

where Σ is a closed convex subset of $\mathbb{R}^d$ and $g_0(z), h_0(z), g_1(z), h_1(z) : \Sigma \to \mathbb{R}$ are convex functions. Although the results in this section can be generalized to the case where there are more than one inequality in a straight-forwarded manner, to simplify the notation and concentrate on the main idea we assume there is only one inequality constraint in problem (DC). Our algorithms are modifications of the classical DCA (see [39]).

Before we introduce our algorithms and conduct the convergence analyses, we first brief some solution quality characterizations for problem (DC).

Definition 1 Let $\bar{z}$ be a feasible solution of problem (DC). We say that $\bar{z}$ is an KKT point of problem (DC) if there exists a multiplier $\lambda \geq 0$ such that

$$0 \in \partial g_0(\bar{z}) - \partial h_0(\bar{z}) + \lambda(\partial g_1(\bar{z}) - \partial h_1(\bar{z})) + \mathcal{N}_\Sigma(\bar{z}),$$
$$(g_1(\bar{z}) - h_1(\bar{z}))\lambda = 0.$$

Definition 2 Let $\bar{z}$ be a feasible point of problem (DC). We say that the nonzero abnormal multiplier constraint qualification (NNAMCQ) holds at $\bar{z}$ for problem (DC) if either $f_1(\bar{z}) < 0$ or $f_1(\bar{z}) = 0$ but

$$0 \notin \partial g_1(\bar{z}) - \partial h_1(\bar{z}) + \mathcal{N}_\Sigma(\bar{z}). \tag{6}$$

Let $\bar{z} \in \Sigma$, we say that the extended no nonzero abnormal multiplier constraint qualification (ENNAMCQ) holds at $\bar{z}$ for problem (DC) if either $f_1(\bar{z}) < 0$ or $f_1(\bar{z}) \geq 0$ but (6) holds.

Note that NNAMCQ (ENNAMCQ) is equivalent to MFCQ (EMFCQ) respectively; see e.g., [15].

The following optimality condition follows from the nonsmooth multiplier rule in terms of Clarke subgradients (see e.g. [7,15]) and the fact that for two convex functions g, h which are Lipschitz around point $\bar{z}$, we have $\partial^c(g(\bar{z}) - h(\bar{z})) \subseteq \partial^c g(\bar{z}) - \partial^c h(\bar{z}) = \partial g(\bar{z}) - \partial h(\bar{z})$.

Proposition 4 *Let $\bar{z}$ be a local solution of problem (DC). If NNAMCQ holds at $\bar{z}$ and all functions g_0, g_1, h_0, h_1 are Lipschitz around point $\bar{z}$, then $\bar{z}$ is a KKT point of problem (DC).*

3.1 Inexact proximal DCA with simplified penalty parameter update

In this subsection we propose an algorithm called inexact proximal DCA to solve problem (DC) and show its convergence.

By using the main idea of DCA which linearizes the concave part of the DC structure, we propose a sequential convex programming scheme as follows. Given a current iteration $z^k \in \Sigma$ with $k = 0, 1, \ldots$, we select a subdifferential $\xi_i^k \in \partial h_i(z^k)$, for $i = 0, 1$. Then we solve the following subproblem approximately and select z^{k+1} as an approximate minimizer:

$$\begin{aligned} \min_{z \in \Sigma} \ \tilde{\varphi}_k(z) := g_0(z) - h_0(z^k) - \langle \xi_0^k, z - z^k \rangle \\ + \beta_k \max\{g_1(z) - h_1(z^k) - \langle \xi_1^k, z - z^k \rangle, 0\} + \frac{\rho}{2}\|z - z^k\|^2, \end{aligned} \tag{7}$$

where ρ is a given positive constant and β_k represents the adaptive penalty parameter. Our scheme is similar to that of DCA2 in [39] but different in that the subproblem (7) has a strongly convex objective function, the subproblem is only solved approximately, and a simplier penalty parameter update is used. In particular, denote

$$t^{k+1} := \max\{g_1(z^{k+1}) - h_1(z^k) - \langle \xi_1^k, z^{k+1} - z^k \rangle, 0\}. \tag{8}$$

We propose the following two inexact conditions for choosing z^{k+1} as an approximate solution to (7):

$$\operatorname{dist}(0, \partial\tilde{\varphi}_k(z^{k+1}) + \mathcal{N}_\Sigma(z^{k+1})) \leq \zeta_k, \quad \text{for some } \zeta_k \geq 0 \text{ satisfying } \sum_{k=0}^{\infty} \zeta_k^2 < \infty, \tag{9}$$

and

$$\operatorname{dist}(0, \partial\tilde{\varphi}_k(z^{k+1}) + \mathcal{N}_\Sigma(z^{k+1})) \leq \frac{\sqrt{2}}{2}\rho\|z^k - z^{k-1}\|, \tag{10}$$

where $\operatorname{dist}(x, M)$ denotes the distance from a point x to set M.

Using above constructions, we are ready to propose the inexact proximal DCA (iP-DCA) in Algorithm 1.

Algorithm 1 iP-DCA

1: Take an initial point $z^0 \in \Sigma$; $\delta_\beta > 0$; an initial penalty parameter $\beta_0 > 0$, $tol > 0$.
2: **for** $k = 0, 1, \ldots$ **do**
  1. Compute $\xi_i^k \in \partial h_i(z^k)$, $i = 0, 1$.
  2. Obtain an inexact solution z^{k+1} of (7) satisfying (9) or (10).
  3. Stopping test. Compute t^{k+1} in (8). Stop if $\max\{\|z^{k+1} - z^k\|, t^{k+1}\} < tol$.
  4. Penalty parameter update. Set

$$\beta_{k+1} = \begin{cases} \beta_k + \delta_\beta, & \text{if } \max\{\beta_k, 1/t^{k+1}\} < \|z^{k+1} - z^k\|^{-1}, \\ \beta_k, & \text{otherwies.} \end{cases}$$

  5. Set $k := k + 1$.
3: **end for**

In DCA2 of [39], the subproblem (7) was solved as a constrained optimization problem and a Lagrange multiplier is used to update the penalty parameter. Since our penalty parameter update rule does not involve any multipliers, it is easier to implement. In the rest of this section we show that the proposed algorithm converges. Let us start with the following lemma which provides a sufficient decrease of the merit function of (DC) defined by

$$\varphi_k(z) := g_0(z) - h_0(z) + \beta_k \max\{g_1(z) - h_1(z), 0\}.$$

Lemma 1 *Let $\{z^k\}$ be a sequence of iterates generated by iP-DCA as defined in Algorithm 1. If the inexact criterion (9) or (10) is applied, then z^k satisfies*

$$\begin{aligned} &\varphi_k(z^k) \geq \varphi_k(z^{k+1}) + \frac{\rho}{2}\|z^{k+1} - z^k\|^2 - \frac{1}{2\rho}\zeta_k^2, \\ \textit{or} \quad &\varphi_k(z^k) \geq \varphi_k(z^{k+1}) + \frac{\rho}{2}\|z^{k+1} - z^k\|^2 - \frac{\rho}{4}\|z^k - z^{k-1}\|^2, \end{aligned}$$

where $\zeta_k \geq 0$ satisfying $\sum_{k=0}^{\infty} \zeta_k^2 < \infty$ respectively.

Proof Since z^{k+1} is an approximation solution to problem (7) with inexact criterion (9) or (10), there exists a vector e_k such that $e_k \in \partial\tilde{\varphi}_k(z^{k+1}) + \mathcal{N}_\Sigma(z^{k+1}) \subseteq \partial(\tilde{\varphi}_k + \delta_\Sigma)(z^{k+1})$ and

$$\|e_k\| \leq \zeta_k \text{ or } \quad \|e_k\| \leq \frac{\sqrt{2}}{2}\rho\|z^k - z^{k-1}\|, \tag{11}$$

respectively. As $\tilde{\varphi}_k$ is strongly convex with modulus ρ, Σ is a closed convex set and $z^k \in \Sigma$, we have

$$\begin{aligned}\tilde{\varphi}_k(z^k) &\geq \tilde{\varphi}_k(z^{k+1}) + \langle e_k, z^{k+1} - z^k\rangle + \frac{\rho}{2}\|z^{k+1} - z^k\|^2 \\ &\geq \tilde{\varphi}_k(z^{k+1}) - \frac{1}{2\rho}\|e_k\|^2 - \frac{\rho}{2}\|z^{k+1} - z^k\|^2 + \frac{\rho}{2}\|z^{k+1} - z^k\|^2 \\ &= \tilde{\varphi}_k(z^{k+1}) - \frac{1}{2\rho}\|e_k\|^2.\end{aligned} \tag{12}$$

Next, by the convexity of $h_i(z)$ and $\xi_i^k \in \partial h_i(z^k)$, $i = 0, 1$, there holds that

$$h_i(z^{k+1}) \geq h_i(z^k) + \langle \xi_i^k, z^{k+1} - z^k\rangle, \quad i = 0, 1,$$

and thus $\tilde{\varphi}_k(z^{k+1}) \geq \varphi_k(z^{k+1}) + \frac{\rho}{2}\|z^{k+1} - z^k\|^2$. Combined with (12), we have

$$\varphi_k(z^k) = \tilde{\varphi}_k(z^k) \geq \tilde{\varphi}_k(z^{k+1}) - \frac{1}{2\rho}\|e_k\|^2 \geq \varphi_k(z^{k+1}) - \frac{1}{2\rho}\|e_k\|^2 + \frac{\rho}{2}\|z^{k+1} - z^k\|^2.$$

The conclusion follows immediately from (11). □

The following theorem is the main result of this section. It proves that any accumulation point of iP-DCA is an KKT point as long as the penalty parameter sequence $\{\beta_k\}$ is bounded.

Theorem 1 *Suppose f_0 is bounded below on Σ and the sequences $\{z^k\}$ and $\{\beta_k\}$ generated by iP-DCA are bounded. Moreover suppose functions g_0, g_1, h_1, h_0 are locally Lipschitz on set Σ. Then every accumulation point of $\{z^k\}$ is an KKT point of problem (DC).*

Proof Since $\{\beta_k\}$ is bounded, there exists some iteration index k_0 such that $\beta_k = \beta_{k_0}, \quad \forall k \geq k_0$, and thus $\varphi_k(z) = \varphi_{k_0}(z)$ for all $k \geq k_0$. As f_0 is bounded below, $\varphi_k(z)$ is bounded below for all $k \geq k_0$. Then, by the inequality (11) and (11) obtained in Lemma 1, we have

$$\sum_{k=1}^{\infty} \|z^{k+1} - z^k\|^2 < +\infty, \qquad \lim_{k\to\infty} \|z^{k+1} - z^k\| = 0,$$

and thus $\beta_k < \|z^{k+1} - z^k\|^{-1}$ always holds when k is large enough. According to the update strategy of β_k in iP-DCA, there exists some iteration index k_1 such that

$$t^{k+1} := \max\{g_1(z^{k+1}) - h_1(z^k) - \langle \xi_1^k, z^{k+1} - z^k\rangle, 0\} \leq \|z^{k+1} - z^k\| \qquad \forall k \geq k_1,$$

and thus $t^k \to 0$. Since z^{k+1} is an approximate solution to problem (7) and inexact criterion (9) or (10) holds, there exists a vector e_k such that $e_k \in \partial\tilde{\varphi}_k(z^{k+1}) + \mathcal{N}_\Sigma(z^{k+1})$ and (11) holds. According to the sum rule in Proposition 1 and the subdifferential calculus rules for the pointwise maximum (see e.g. [7, Proposition 2.3.12]), there exist $\tilde{\lambda}^{k+1} \in [0,1]$ and $\eta_i^{k+1} \in \partial g_i(z^{k+1})(i=0,1)$ such that

$$e_k \in \eta_0^{k+1} - \xi_0^k + \beta_k\tilde{\lambda}^{k+1}(\eta_1^{k+1} - \xi_1^k) + \rho(z^{k+1} - z^k) + \mathcal{N}_\Sigma(z^{k+1}), \quad (13)$$

$$g_1(z^{k+1}) - h_1(z^k) - \langle \xi_1^k, z^{k+1} - z^k\rangle \le t^{k+1}, \quad (14)$$

$$\tilde{\lambda}^{k+1}(g_1(z^{k+1}) - h_1(z^k) - \langle \xi_1^k, z^{k+1} - z^k\rangle - t^{k+1}) = 0, \quad (15)$$

$$t^{k+1}(1 - \tilde{\lambda}^{k+1}) = 0, \quad t^{k+1} \ge 0. \quad (16)$$

Since $\{\beta_k\tilde{\lambda}^{k+1}\}$ is bounded, we may suppose that $\tilde{z}$ and $\tilde{\lambda}$ are accumulation points of $\{z^k\}$ and $\{\beta_k\tilde{\lambda}^{k+1}\}$ respectively. Taking subsequences if necessary, without loss of generality we may assume that $z^k \to \tilde{z} \in \Sigma$ and $\beta_k\tilde{\lambda}^{k+1} \to \tilde{\lambda}$. Since $g_i(z)$, $h_i(z)$, $i = 0,1$ are locally Lipschitz continuous at $\tilde{z}$, $\partial g_i(z)$, $\partial h_i(z)$, $i = 0,1$ and $\mathcal{N}_\Sigma(z)$ are outer semicontinuous. Now passing onto the limit as $k \to \infty$ in (13)-(15), as $e_k \to 0$ from $\zeta_k \to 0$ in (9) or $\|z^{k+1} - z^k\| \to 0$ in (10) and $t^k \to 0$, we obtain that $\tilde{z}$ is a KKT solution of problem (DC). □

Notice that the boundedness of the penalty parameters is needed for an accumulation point to be an KKT point. The following proposition provides a sufficient condition for the boundedness of the penalty parameters sequence $\{\beta_k\}$.

Proposition 5 *Suppose that the sequence $\{z^k\}$ generated by iP-DCA is bounded. Moreover suppose functions g_0, g_1, h_1, h_0 are Lipschitz around at any accumulation point of $\{z^k\}$. Assume that ENNAMCQ holds at any accumulation points of the sequence $\{z^k\}$. Then the sequence $\{\beta_k\}$ must be bounded.*

Proof The proof is inspired by [39, Theorem 3.1]. To the contrary, suppose that $\beta_k \to +\infty$ as $k \to \infty$. Then there exist infinitely many indices j such that

$$\beta_{k_j} < \|z^{k_j+1} - z^{k_j}\|^{-1} \text{ and } \quad t^{k_j+1} > \|z^{k_j+1} - z^{k_j}\|,$$

and thus

$$\lim_{j\to\infty} \|z^{k_j+1} - z^{k_j}\| = 0, \qquad t^{k_j+1} > 0, \quad \forall j.$$

From (16), since $t^{k_j+1} > 0$ for all j, we have $\tilde{\lambda}^{k_j+1} = 1$ for all j and thus $\lambda^{k_j+1} := \beta_{k_j}\tilde{\lambda}^{k_j+1} \to +\infty$ as $j \to \infty$. Taking a further subsequence, if necessary, we can assume that $z^{k_j} \to \tilde{z} \in \Sigma$ as $j \to \infty$. If $g_1(\tilde{z}) - h_1(\tilde{z}) < 0$, then as g_1, h_1 are continuous at $\tilde{z}$, $\{\xi^{k_j}\}$ is bounded, and $\lim_{j\to\infty} \|z^{k_j+1} - z^{k_j}\| = 0$, when j is sufficiently large, one has $g_1(z^{k_j+1}) - h_1(z^{k_j}) - \langle \xi_1^{k_j}, z^{k_j+1} - z^{k_j}\rangle < 0$, which contradicts to $t^{k_j+1} := \max\{g_1(z^{k_j+1}) - h_1(z^{k_j}) - \langle \xi_1^{k_j}, z^{k_j+1} - z^{k_j}\rangle, 0\} > 0$ for all j. Thus, $g_1(\tilde{z}) - h_1(\tilde{z}) \ge 0$. From (13), we have

$$\begin{aligned} e_{k_j} \in &\partial g_0(z^{k_j+1}) - \partial h_0(z^{k_j}) + \lambda^{k_j+1}\partial g_1(z^{k_j+1}) - \lambda^{k_j+1}\partial h_1(z^{k_j}) \\ &+ \rho(z^{k_j+1} - z^{k_j}) + N_\Sigma(z^{k_j+1}), \end{aligned}$$

where $\lambda^{k+1} := \beta_k \tilde{\lambda}^{k+1}$. Dividing both sides of this equality by λ^{k_j+1}, and passing onto the limit as $j \to \infty$, we have $0 \in \partial g_1(\tilde{z}) - \partial h_1(\tilde{z}) + \mathcal{N}_{\Sigma}(\tilde{z})$, which contradicts ENNAMCQ. □

3.2 Inexact proximal linearized DCA with simplified penalty parameter update

Recall that iP-DCA defined in Algorithm 1 requires minimization of a strongly convex subproblem (7). In this subsection, we assume that g_1 is L-smooth which means that $\nabla g_1(z)$ is Lipschitz continuous with constant L on Σ. This setting motivates a very simple linearization approach in which we not only linearize the concave part but also the convex part of the DC structure. Given a current iteration $z^k \in \Sigma$ with $k = 0, 1, \ldots$, we select a subdifferential $\xi_i^k \in \partial h_i(z^k)$, for $i = 0, 1$. Then we solve the following subproblem approximately.

$$\begin{aligned}\min_{z \in \Sigma} \ \hat{\varphi}_k(z) := g_0(z) - h_0(z^k) - \langle \xi_0^k, z - z^k \rangle + \frac{\rho_k}{2}\|z - z^k\|^2 \\ + \beta_k \max\{g_1(z^k) + \langle \nabla g_1(z^k), z - z^k \rangle - h_1(z^k) - \langle \xi_1^k, z - z^k \rangle, 0\},\end{aligned} \tag{17}$$

where ρ_k and β_k are the adaptive proximal and penalty parameters respectively. Choose z^{k+1} as an *approximate minimizer* of the convex subproblem (17) satisfying one of the following two inexact criteria

$$\operatorname{dist}(0, \partial\hat{\varphi}_k(z^{k+1}) + \mathcal{N}_{\Sigma}(z^{k+1})) \le \zeta_k, \qquad \text{for some } \zeta_k \text{ satisfying } \sum_{k=0}^{\infty} \zeta_k^2 < \infty, \tag{18}$$

and

$$\operatorname{dist}(0, \partial\hat{\varphi}_k(z^{k+1}) + \mathcal{N}_{\Sigma}(z^{k+1})) \le \frac{\sqrt{2}}{2}\sigma\|z^k - z^{k-1}\|. \tag{19}$$

This yields the inexact proximal linearized DCA (iPL-DCA), whose exact description is given in Algorithm 2.

Recall that the merit function of (DC) is defined by $\varphi_k(z) := f_0(z) + \beta_k \max\{f_1(z), 0\}$. Similar to Lemma 1, we first give following sufficiently decrease result of iPL-DCA.

Lemma 2 *Let $\{z^k\}$ be the sequence of iterates generated by iPL-DCA as defined in Algorithm 2. If the inexact criterion (18) or (19) is applied, then z^k satisfies*

$$\begin{aligned}\varphi_k(z^k) &\ge \varphi_k(z^{k+1}) + \frac{\sigma}{2}\|z^{k+1} - z^k\|^2 - \frac{1}{2\sigma}\zeta_k^2, \\ \varphi_k(z^k) &\ge \varphi_k(z^{k+1}) + \frac{\sigma}{2}\|z^{k+1} - z^k\|^2 - \frac{\sigma}{4}\|z^k - z^{k-1}\|^2,\end{aligned}$$

respectively.

Algorithm 2 iPL-DCA

1: Take an initial point $z^0 \in \Sigma$; $\delta_\beta > 0$, $\sigma > 0$, an initial penalty parameter $\beta_0 > 0, \rho_0 = \frac{1}{2}\beta_0 L + \sigma$, $tol > 0$.
2: **for** $k = 0, 1, \dots$ **do**
 1. Compute $\xi_i^k \in \partial h_i(z^k)$, $i = 0, 1$.
 2. Obtain an inexact solution z^{k+1} of (17) satisfying (18) or (19).
 3. Stopping test. Compute $t^{k+1} := \max\{g_1(z^k) + \langle \nabla g_1(z^k), z^{k+1} - z^k\rangle - h_1(z^k) - \langle \xi_1^k, z^{k+1} - z^k\rangle, 0\}$. Stop if $\max\{\|z^{k+1} - z^k\|, t^{k+1}\} < tol$.
 4. Penalty parameter update. Set

$$\beta_{k+1} = \begin{cases} \beta_k + \delta_\beta, & \text{if } \max\{\beta_k, 1/t^{k+1}\} < \|z^{k+1} - z^k\|^{-1}, \\ \beta_k, & \text{otherwies.} \end{cases}$$

$$\rho_{k+1} = \frac{1}{2}\beta_{k+1} L + \sigma.$$

 5. Set $k := k + 1$.
3: **end for**

Proof Since z^{k+1} is an approximation solution to problem (17) with inexact criterion (18) or (19), there exists a vector e_k such that $e_k \in \partial\hat{\varphi}_k(z^{k+1}) + \mathcal{N}_\Sigma(z^{k+1})) \subseteq \partial(\tilde{\varphi}_k + \delta_\Sigma)(z^{k+1})$ and

$$\|e_k\| \le \zeta_k \text{ or } \quad \|e_k\| \le \frac{\sqrt{2}}{2}\sigma\|z^k - z^{k-1}\|, \tag{20}$$

respectively. As $\hat{\varphi}_k$ is strongly convex with modulus ρ_k and Σ is a closed convex set, we have

$$\begin{aligned} \hat{\varphi}_k(z^k) &\ge \hat{\varphi}_k(z^{k+1}) + \langle e_k, z^{k+1} - z^k\rangle + \frac{\rho_k}{2}\|z^{k+1} - z^k\|^2 \\ &\ge \hat{\varphi}_k(z^{k+1}) - \frac{1}{2\sigma}\|e_k\|^2 - \frac{\sigma}{2}\|z^{k+1} - z^k\|^2 + \frac{\rho_k}{2}\|z^{k+1} - z^k\|^2 \\ &= \hat{\varphi}_k(z^{k+1}) - \frac{1}{2\sigma}\|e_k\|^2 + \frac{\rho_k - \sigma}{2}\|z^{k+1} - z^k\|^2. \end{aligned} \tag{21}$$

Next, by the convexity of $h_i(z)$ and $\xi_i^k \in \partial h_i(z^k)$, $i = 0, 1$, we have

$$h_i(z^{k+1}) \ge h_i(z^k) + \langle \xi_i^k, z^{k+1} - z^k\rangle, \quad i = 0, 1.$$

And since g_1 is L-smooth, we have

$$g_1(z^{k+1}) \le g_1(z^k) + \langle \nabla g_1(z^k), z - z^k\rangle + \frac{L}{2}\|z^{k+1} - z^k\|^2.$$

Thus, we have

$$\hat{\varphi}_k(z^{k+1}) \ge \varphi_k(z^{k+1}) + \frac{\rho_k - \beta_k L}{2}\|z^{k+1} - z^k\|^2.$$

Combined with (21), we have

$$\begin{aligned}\varphi_k(z^k) = \hat{\varphi}_k(z^k) &\geq \hat{\varphi}_k(z^{k+1}) - \frac{1}{2\sigma}\|e_k\|^2 + \frac{\rho_k - \sigma}{2}\|z^{k+1} - z^k\|^2 \\ &\geq \varphi_k(z^{k+1}) - \frac{1}{2\sigma}\|e_k\|^2 + \frac{2\rho_k - \beta_k L - \sigma}{2}\|z^{k+1} - z^k\|^2 \\ &= \varphi_k(z^{k+1}) - \frac{1}{2\sigma}\|e_k\|^2 + \frac{\sigma}{2}\|z^{k+1} - z^k\|^2.\end{aligned}$$

Then the conclusion follows immediately from (20). □

Similar to Theorem 1 and Proposition 5, by Lemma 2, the following convergence results of iPL-DCA can be derived easily. The proofs are purely technical and thus omitted.

Theorem 2 *Suppose f_0 is bounded below and the sequences $\{z^k\}$ and $\{\beta_k\}$ generated by iPL-DCA are bounded, functions g_0, h_1, h_0 are locally Lipschitz on set Σ. Then every accumulation point of $\{z^k\}$ is an KKT point for problem (DC).*

Proposition 6 *Suppose the sequence $\{z^k\}$ generated by iPL-DCA is bounded, functions g_0, h_1, h_0 are Lipschitz around at any accumulation point of $\{z^k\}$, and ENNAMCQ holds at any accumulation points of the sequence $\{z^k\}$. Then the sequence $\{\beta_k\}$ is bounded.*

Remark 1 In fact, if g_0 is further assumed to be differentiable and ∇g_0 is Lipschitz continuous, we can also linearize g_0 in iPL-DCA. The proof of convergence is similar.

4 DC algorithms for solving DCBP

In this section we will show how to solve problem (DCBP) numerically.

It is obvious that problem $(\mathrm{VP})_\epsilon$ is problem (DC) with

$$z := (x, y),\ f_0(x, y) := F_1(x, y) - F_2(x, y),\ f_1(x, y) := f(x, y) - v(x) - \epsilon,\ \Sigma = C.$$

According to Proposition 3, since $F_1(x, y), F_2(x, y), f(x, y)$ are Lipschitz continuous near every point on an open convex set containing C and hence Lipschitz continuous near every point on C. However our problem $(\mathrm{VP})_\epsilon$ involves the value function which is an extended-value function $v(x) : X \to [-\infty, \infty]$ defined by

$$v(x) := \inf_{y \in Y} \{f(x, y)\ s.t.\ g(x, y) \leq 0\},$$

with the convention of $v(x) = +\infty$ if the feasible region $\mathcal{F}(x)$ is empty. To apply the proposed DC algorithms, we need to answer the following questions.

(a) Is the value function convex and locally Lipschitz on the convex set X and how to obtain one element from $\partial V(x^k)$ in terms of problem data?
(b) Will the constraint qualification ENNAMCQ hold at any accumulation point of the iteration sequence?

We now give answers to these questions in the next two subsections.

4.1 Lipschitz continuity and the subdifferential of the value function

Thanks to the full convex structure of the lower level problem in (DCBP), the value function turns out to be convex and Lipschitz continuous under our problem setting as shown below.

Lemma 3 *The value function $v(x) : X \to \mathbb{R}$ is convex and Lipschitz continuous around any point in set X. Given $\bar{x} \in X$ and $\bar{y} \in S(\bar{x})$, we have*

$$\partial v(\bar{x}) = \{\xi \in \mathbb{R}^n : (\xi, 0) \in \partial\phi(\bar{x}, \bar{y})\},$$

where $\phi(x, y) := f(x, y) + \delta_D(x, y)$, $D := \{(x, y) \in \mathcal{O} \times Y \mid g(x, y) \le 0\}$, with $\mathcal{O}$ being the open set defined in the introduction.

Proof First we extend the definition of the value function from any element $x \in X$ to the whole space $\mathbb{R}^n$ as follows:

$$v(x) := \inf_{y \in \mathbb{R}^m} \phi(x, y), \qquad \forall x \in \mathbb{R}^n.$$

It follows that $v(x) = +\infty$ for $x \notin \mathcal{O}$. In our problem setting, f is fully convex on an open convex set containing the convex set C and hence we can assume without loss of generality that f is fully convex on the convex set D. Therefore the extended-valued function $\phi(x, y) : \mathbb{R}^n \times \mathbb{R}^m \to [-\infty, \infty]$ is convex. The convexity of the value function $v(x) = \inf_{y \in \mathbb{R}^m} \phi(x, y)$ then follows from [35, Theorem 1]. Hence the value function restricted on set X is convex. Next, according to [35, Theorem 24], we have the equation (3). By assumption stated in the introduction of the paper, the feasible region of the lower level program $\mathcal{F}(x) := \{y \in Y : g(x, y) \le 0\} \neq \emptyset$ and $v(x) \neq -\infty$ for all x in the open set $\mathcal{O}$. Hence $v(x) : \mathbb{R}^n \to [-\infty, \infty]$ is proper convex. Since $\mathrm{dom} v := \{x : v(x) < +\infty\} = \{x : \mathcal{F}(x) \neq \emptyset\} \supseteq \mathcal{O} \supseteq X$, we have $X \subseteq \mathrm{int}(\mathrm{dom} v)$. The result on Lipschitz continuity of the value function follows from Proposition 3. □

By using some sensitivity analysis techniques, an element of the subdifferential of the value function $v(x)$ can be expressed in terms of Lagrangian multipliers. In particular, given $\bar{y} \in S(\bar{x})$, we denote the set of KKT multipliers of the lower-level problem $(P_{\bar{x}})$ by

$$KT(\bar{x}, \bar{y}) := \left\{ \gamma \in \mathbb{R}^l_+ \middle| 0 \in \partial_y f(\bar{x}, \bar{y}) + \sum_{i=1}^{l} \gamma_i \partial_y g_i(\bar{x}, \bar{y}) + \mathcal{N}_Y(\bar{y}), \quad \sum_{i=1}^{l} \gamma_i g_i(\bar{x}, \bar{y}) = 0 \right\}.$$

Theorem 3 *Let $\bar{x} \in X$ and $\bar{y} \in S(\bar{x})$. Then*

$$\partial v(\bar{x}) \supseteq \left\{ \xi \middle| (\xi, 0) \in \partial f(\bar{x}, \bar{y}) + \sum_{i=1}^{l} \gamma_i \partial g_i(\bar{x}, \bar{y}) + \{0\} \times \mathcal{N}_Y(\bar{y}), \gamma \ge 0, \sum_{i=1}^{l} \gamma_i g_i(\bar{x}, \bar{y}) = 0 \right\}, \tag{22}$$

and the equality holds in (22) provided that

$$\mathcal{N}_E(\bar{x},\bar{y}) = \Big\{ \sum_{i=1}^{l} \gamma_i \partial g_i(\bar{x},\bar{y}) + \{0\} \times \mathcal{N}_Y(\bar{y}) \mid \gamma \geq 0, \sum_{i=1}^{l} \gamma_i g_i(\bar{x},\bar{y}) = 0 \Big\}, \quad (23)$$

where $E := \{(x,y) \in \mathbb{R}^n \times Y : g(x,y) \leq 0\}$.

Moreover if the partial derivative formula holds

$$\partial f(\bar{x},\bar{y}) = \partial_x f(\bar{x},\bar{y}) \times \partial_y f(\bar{x},\bar{y}), \;\; \partial g_i(\bar{x},\bar{y}) = \partial_x g_i(\bar{x},\bar{y}) \times \partial_y g_i(\bar{x},\bar{y}) \quad (24)$$

then

$$\bigcup_{\gamma \in KT(\bar{x},\bar{y})} \left(\partial_x f(\bar{x},\bar{y}) + \sum_{i=1}^{l} \gamma_i \partial_x g_i(\bar{x},\bar{y}) \right) \subseteq \partial v(\bar{x}), \quad (25)$$

and the equality in (25) holds provided that (23) holds.

Proof Let $\phi_E(x,y) := f(x,y) + \delta_E(x,y) = f(x,y) + \delta_Y(y) + \sum_{i=1}^{l} \delta_{C_i}(x,y)$ with $C_i := \{(x,y) | g_i(x,y) \leq 0\}$. Then by the sum rule (see Proposition 1) and the fact that $\mathcal{N}_E = \partial \delta_E$, we have

$$\partial f(\bar{x},\bar{y}) + \{0\} \times \mathcal{N}_Y(\bar{y}) + \sum_{i=1}^{l} \mathcal{N}_{C_i}(\bar{x},\bar{y}) \subseteq \partial \phi_E(\bar{x},\bar{y}). \quad (26)$$

When $g_i(\bar{x},\bar{y}) < 0$, we have $(\bar{x},\bar{y}) \in \text{int}C_i$ and hence $\gamma_i = 0 \in \mathcal{N}_{C_i}(\bar{x},\bar{y})$. Otherwise if $g_i(\bar{x},\bar{y}) = 0$, by definition of subdifferential and the normal cone we can show that for any $\gamma_i \geq 0$, $\gamma_i \partial g_i(\bar{x},\bar{y}) \subseteq \mathcal{N}_{C_i}(\bar{x},\bar{y})$. Hence together with (26), we have

$$\Big\{ \partial f(\bar{x},\bar{y}) + \sum_{i=1}^{l} \gamma_i \partial g_i(\bar{x},\bar{y}) + \{0\} \times \mathcal{N}_Y(\bar{y}) \mid \gamma \in \mathbb{R}^l, \gamma \geq 0, \sum_{i=1}^{l} \gamma_i g_i(\bar{x},\bar{y}) = 0 \Big\}$$
$$\subseteq \partial \phi_E(\bar{x},\bar{y}).$$

Since $\partial \phi_E(\bar{x},\bar{y}) = \partial \phi(\bar{x},\bar{y})$ where $\phi(x,y) = f(x,y) + \delta_D(x,y)$, it follows from Lemma 3 that

$$\begin{aligned} \partial v(\bar{x}) &= \{\xi \mid (\xi, 0) \in \partial \phi(\bar{x},\bar{y})\} = \{\xi \mid (\xi,0) \in \partial \phi_E(\bar{x},\bar{y})\} \\ &\supseteq \Bigg\{ \xi \mid (\xi,0) \in \partial f(\bar{x},\bar{y}) + \sum_{i=1}^{l} \gamma_i \partial g_i(\bar{x},\bar{y}) + \{0\} \times \mathcal{N}_Y(\bar{y}), \\ &\qquad\qquad \gamma \in \mathbb{R}^l, \gamma \geq 0, \sum_{i=1}^{l} \gamma_i g_i(\bar{x},\bar{y}) = 0 \Bigg\}. \end{aligned}$$

Hence (22) holds. Since f is Lipschitz continuous at $(\bar{x},\bar{y})$, by the sum rule in Proposition 1, we have $\partial \phi_E(\bar{x},\bar{y}) = \partial f(\bar{x},\bar{y}) + \mathcal{N}_E(\bar{x},\bar{y})$. Hence if (23) holds, then

$$\begin{aligned}\partial v(\bar{x}) &= \{\xi \mid (\xi,0) \in \partial\phi(\bar{x},\bar{y})\} = \{\xi \mid (\xi,0) \in \partial\phi_E(\bar{x},\bar{y})\}\\ &= \{\xi \mid (\xi,0) \in \partial f(\bar{x},\bar{y}) + \mathcal{N}_E(\bar{x},\bar{y})\}\\ &= \left\{\xi \mid (\xi,0) \in \partial f(\bar{x},\bar{y}) + \sum_{i=1}^{l}\gamma_i\partial g_i(\bar{x},\bar{y}) + \{0\}\times\mathcal{N}_Y(\bar{y}), \gamma\geq 0, \sum_{i=1}^{l}\gamma_i g_i(\bar{x},\bar{y}) = 0\right\}.\end{aligned}$$

This shows that the equality holds in (22).

Now suppose that (24) holds. Then for any $\gamma \in \mathbb{R}^l, \gamma \geq 0, \sum_{i=1}^{l}\gamma_i g_i(\bar{x},\bar{y}) = 0$, by the sum rule in Proposition 1 we have

$$\begin{aligned}&\partial\phi_E(\bar{x},\bar{y})\\ &\supseteq \partial f(\bar{x},\bar{y}) + \sum_{i=1}^{l}\gamma_i\partial g_i(\bar{x},\bar{y}) + \{0\}\times\mathcal{N}_Y(\bar{y})\\ &= \left\{\partial_x f(\bar{x},\bar{y}) + \sum_{i=1}^{l}\gamma_i\partial_x g_i(\bar{x},\bar{y})\right\} \times \left\{\partial_y f(\bar{x},\bar{y}) + \sum_{i=1}^{l}\gamma_i\partial_y g_i(\bar{x},\bar{y}) + \mathcal{N}_Y(\bar{y})\right\}.\end{aligned}$$

Combining with (22), we obtain (25). Similarly when (23) holds, the equality holds in (25). □

By the description in (25), $\partial v(\bar{x})$ can be calculated as long as (23) is satisfied. We claim that (23) is a mild condition. In fact, by convexity, there always holds the inclusion

$$\mathcal{N}_E(\bar{x},\bar{y}) \supseteq \left\{\sum_{i=1}^{l}\gamma_i\partial g_i(\bar{x},\bar{y}) + \{0\}\times\mathcal{N}_Y(\bar{y}) : \gamma\geq 0, \sum_{i=1}^{l}\gamma_i g_i(\bar{x},\bar{y}) = 0\right\}.$$

By virtue of [16, Theorem 4.1], the reverse inclusion also follows under standard assumptions. Some suffcient conditions for (23) are thus summarized in the following proposition.

Proposition 7 *Equation (23) holds provided that the set-valued map*

$$\Psi(\alpha) := \{(x,y) \in \mathbb{R}^n \times Y : g(x,y) + \alpha \leq 0\}$$

is calm at $(0,\bar{x},\bar{y})$, i.e., there exist $\kappa,\delta > 0$ such that

$$\mathrm{dist}_E(x,y) \leq \kappa\|\max\{g(x,y),0\}\| \qquad \forall (x,y) \in B_\delta(\bar{x},\bar{y}) \cap E;$$

in particular if one of the following conditions:

(a) The linear constraint qualification holds: $g(x,y)$ is an affine mapping of (x,y) and Y is convex polyhedral.
(b) The Slater condition holds: there exists a point $(x_0,y_0) \in \mathbb{R}^n \times Y$ such that $g(x_0,y_0) < 0$

holds.

Proof By virtue of [16, Theorem 4.1], the reverse inclusion

$$\mathcal{N}_E(\bar{x},\bar{y}) \subseteq \Big\{ \sum_{i=1}^{l} \gamma_i \partial g_i(\bar{x},\bar{y}) + \{0\} \times \mathcal{N}_Y(\bar{y}) : \gamma \geq 0, \sum_{i=1}^{l} \gamma_i g_i(\bar{x},\bar{y}) = 0 \Big\}$$

holds provided that the system $y \in Y, g(x,y) \leq 0$ is calm at $(0,\bar{x},\bar{y})$. It is well-known that (a) or (b) is a sufficient condition for calmness. □

4.2 Motivations for studying the approximate bilevel program

There are three motivations to consider the approximate program $(\mathrm{VP})_\epsilon$. First, as shown in [22], the solutions of $(\mathrm{VP})_\epsilon$ approximate a true solution of the original bilevel program (DCBP) as ϵ approaches zero. Second, the proximity from a local minimizer of $(\mathrm{VP})_\epsilon$ to the solution set of (DCBP) can be controlled by adjusting the value of ϵ. The third motivation is that the approximate program $(\mathrm{VP})_\epsilon$ when $\epsilon > 0$ would satisfy the required constraint qualification automatically. We present the second motivation in the following proposition.

Proposition 8 *Suppose $\mathcal{S}^*$, the solution set of problem* (VP)*, is compact and the value function $v(x)$ is continuous. Then for any $\delta > 0$, there exists $\bar{\epsilon} > 0$ such that for any $\epsilon \in (0,\bar{\epsilon}]$, there exists (x_ϵ, y_ϵ) which is a local minimum of ϵ-approximation problem* $(\mathrm{VP})_\epsilon$ *with* $\mathrm{dist}((x_\epsilon,y_\epsilon),\mathcal{S}^*) < \delta$.

Proof To the contrary, assume that there exist $\delta > 0$ and sequence $\{\epsilon^k\}$ with $\epsilon^k \downarrow 0$ as $k \to \infty$ such that there does not exist (x,y) being a local minimum of ϵ^k-approximation problem $(\mathrm{VP})_{\epsilon_k}$ satisfying $\mathrm{dist}((x^k,y^k),\mathcal{S}^*) < \delta$ for all k. Consider a point $(\hat{x}^k,\hat{y}^k)$, which is a global minimum to the following problem

$$\begin{aligned} \min_{(x,y)\in C} \quad & F(x,y) \\ s.t. \quad & f(x,y) - v(x) \leq \epsilon, \\ & \mathrm{dist}((x,y),\mathcal{S}^*) \leq \delta. \end{aligned}$$

Then by assumption, it holds that $\mathrm{dist}((\hat{x}^k,\hat{y}^k),\Sigma^*) = \delta$ and $F(\hat{x}^k,\hat{y}^k) \leq F^*$, where F^* is the optimal value of problem (VP). As $\mathcal{S}^*$ is compact, sequence $\{(\hat{x}^k,\hat{y}^k)\}$ is bounded and we can assume without lost of generality that $(\hat{x}^k,\hat{y}^k) \to (\bar{x},\bar{y})$ as $k \to \infty$. Since the value function v is continuous, by taking $k \to \infty$ in $(\hat{x}^k,\hat{y}^k) \in C$ and $f(\hat{x}^k,\hat{y}^k) - v(\hat{x}^k) \leq \epsilon^k$, we obtain the feasibility of the limit point $(\bar{x},\bar{y})$ for problem (VP). Next, by taking $k \to \infty$ in $\mathrm{dist}((\hat{x}^k,\hat{y}^k),\mathcal{S}^*) = \delta$ and $F(\hat{x}^k,\hat{y}^k) \leq F^*$, we obtain that $\mathrm{dist}((\bar{x},\bar{y}),\mathcal{S}^*) = \delta$ and $F(\bar{x},\bar{y}) \leq F^*$, a contradiction. □

Before we clarify the third motivation, we define the concepts of NNAMCQ for problem (VP) and ENNAMCQ for problem $(\mathrm{VP})_\epsilon$ where $\epsilon \geq 0$.

Definition 3 Let $(\bar{x}, \bar{y})$ be a feasible solution to problem (VP). We say that NNAMCQ holds at $(\bar{x}, \bar{y})$ for problem (VP) if

$$0 \notin \partial f(\bar{x}, \bar{y}) - \partial v(\bar{x}) \times \{0\} + \mathcal{N}_C(\bar{x}, \bar{y}). \tag{27}$$

Let $(\bar{x}, \bar{y}) \in C$, we say that ENNAMCQ holds at $(\bar{x}, \bar{y})$ for problem $(VP)_\epsilon$ if either $f(\bar{x}, \bar{y}) - v(\bar{x}) < \epsilon$ or $f(\bar{x}, \bar{y}) - v(\bar{x}) \geq \epsilon$ but (27) holds.

Proposition 9 *Let $(\bar{x}, \bar{y})$ be a feasible solution to problem (VP). Suppose that $v(x)$ is Lipschitz continuous around $\bar{x}$. Then NNAMCQ never holds at $(\bar{x}, \bar{y})$.*

Proof By definition of the value function, we can never have $f(\bar{x}, \bar{y}) - v(\bar{x}) < 0$ and we always have $f(\bar{x}, \bar{y}) - v(\bar{x}) = 0$. But since $(\bar{x}, \bar{y})$ is a feasible solution to problem (VP), it is easy to see that $(\bar{x}, \bar{y})$ must be a solution to the following problem

$$\min_{(x,y)\in C} \{f(x, y) - v(x)\}.$$

But by the optimality condition we must have

$$0 \in \partial f(\bar{x}, \bar{y}) - \partial v(\bar{x}) \times \{0\} + \mathcal{N}_C(\bar{x}, \bar{y}).$$

This means that (27) would never hold. □

Although the NNAMCQ never hold for (VP), thanks to the model structures, it holds automatically for $(\mathrm{VP})_\epsilon$ if $\epsilon > 0$. Hence, according to the DC theories established in the preceding section, powerful DCA can be employed to solve $(\mathrm{VP})_\epsilon$.

Proposition 10 *For any $(\bar{x}, \bar{y}) \in C$, problem $(\mathrm{VP})_\epsilon$ with $\epsilon > 0$ satisfies ENNAMCQ at $(\bar{x}, \bar{y})$.*

Proof If $f(x, y) - v(x) < \epsilon$ holds, then by definition, ENNAMCQ holds at $(\bar{x}, \bar{y})$. Now suppose that $f(\bar{x}, \bar{y}) - v(\bar{x}) \geq \epsilon$ and ENNAMCQ does not hold, i.e.,

$$0 \in \partial f(\bar{x}, \bar{y}) - \partial v(\bar{x}) \times \{0\} + \mathcal{N}_C(\bar{x}, \bar{y}).$$

It follows from (3) that

$$0 \in \begin{bmatrix} \partial_x f(\bar{x}, \bar{y}) - \partial v(\bar{x}) \\ \partial_y f(\bar{x}, \bar{y}) \end{bmatrix} + \mathcal{N}_C(\bar{x}, \bar{y}). \tag{28}$$

By (3) we have

$$\mathcal{N}_C(\bar{x}, \bar{y}) = \partial \delta_C(\bar{x}, \bar{y}) \subseteq \partial_x \delta_C(\bar{x}, \bar{y}) \times \partial_y \delta_C(\bar{x}, \bar{y}) \subseteq \mathbb{R}^n \times \mathcal{N}_{C(\bar{x})}(\bar{y}),$$

where $C(\bar{x}) := \{y \in Y \mid g_i(\bar{x}, y) \leq 0, i = 1, \ldots, l\}$. Thus, it follows from (28) that

$$0 \in \partial_y f(\bar{x}, \bar{y}) + \mathcal{N}_{C(\bar{x})}(\bar{y}),$$

which further implies that $\bar{y} \in \mathcal{S}(\bar{x})$. However, an obvious contradiction to the assumption that $f(\bar{x}, \bar{y}) - v(\bar{x}) \geq \epsilon$ occurs and thus the desired conclusion follows immediately. □

By using the definition of KKT points for (DC) in Definition 1, under the assumption that the value function is locally Lipschitz continiuous, we define KKT points for problem $(\mathrm{VP})_\epsilon$.

Definition 4 We say a point $(\bar{x},\bar{y})$ is a KKT point of problem $(\mathrm{VP})_\epsilon$ with $\epsilon \geq 0$ if there exists $\lambda \geq 0$ such that

$$\begin{cases} 0 \in \partial F_1(\bar{x},\bar{y}) - \partial F_2(\bar{x},\bar{y}) + \lambda \partial f(\bar{x},\bar{y}) - \lambda \partial v(\bar{x}) \times \{0\} + \mathcal{N}_C(\bar{x},\bar{y}), \\ f(\bar{x},\bar{y}) - v(\bar{x}) - \epsilon \leq 0, \quad \lambda\left(f(\bar{x},\bar{y}) - v(\bar{x}) - \epsilon\right) = 0. \end{cases}$$

By virtue of Proposition 10 and Theorem 1, we have the following necessary optimality condition. Since the issue of constraint qualifications for problem (VP) is complicated and it is not the main concern in this paper, we refer the reader to discussions on this topic in [2, 43].

Theorem 4 *Let $(\bar{x},\bar{y})$ be a local optimal solution to problem $(\mathrm{VP})_\epsilon$ with $\epsilon \geq 0$. Suppose either $\epsilon > 0$ or $\epsilon = 0$ and a constraint qualification holds. Then $(\bar{x},\bar{y})$ is an KKT point of problem $(\mathrm{VP})_\epsilon$.*

4.3 Inexact proximal DCA for solving $(\mathrm{VP})_\epsilon$

In this subsection we implement the proposed DC algorithms in Section 3 to solve $(\mathrm{VP})_\epsilon$. To proceed, let us describe iP-DCA to solve $(\mathrm{VP})_\epsilon$. Given a current iteration (x^k, y^k) for each $k = 0, 1, \ldots$, solving the lower level problem parameterized by x^k

$$\min_{y\in Y} f(x^k, y), \;\; s.t.\; g(x^k, y) \leq 0$$

leads to a solution $\tilde{y}^k \in S(x^k)$ and a corresponding KKT multiplier $\gamma^k \in KT(x^k, \tilde{y}^k)$. Select

$$\xi_0^k \in \partial F_2(x^k, y^k), \quad \xi_1^k \in \partial_x f(x^k, \tilde{y}^k) + \sum_{i=1}^{l} \gamma_i^k \partial_x g_i(x^k, \tilde{y}^k). \tag{29}$$

Note that according to (25) in Theorem 3, if the partial derivative formula holds, then $\xi_1^k \in \partial_x f(x^k, \tilde{y}^k) + \sum_{i=1}^{l} \gamma_i^k \partial_x g_i(x^k, \tilde{y}^k)$ is an element of the subdifferential $\partial v(x^k)$. Compute (x^{k+1}, y^{k+1}) as an approximate minimizer of the strongly convex subproblem for $(\mathrm{VP})_\epsilon$ given by

$$\begin{aligned} \min_{(x,y)\in C} \quad & F_1(x,y) - \langle \xi_0^k, (x,y)\rangle + \frac{\rho}{2}\|(x,y) - (x^k, y^k)\|^2 \\ & + \beta_k \max\{f(x,y) - f(x^k, \tilde{y}^k) - \langle \xi_1^k, x - x^k\rangle - \epsilon, 0\}, \end{aligned} \tag{30}$$

satisfying one of the two inexact criteria. Under the assumption that $KT(x^k, y) \neq \emptyset$ for all $y \in S(x^k)$, the description of iP-DCA on $(\mathrm{VP})_\epsilon$ with $\epsilon \geq 0$ now follows:

Algorithm 3 iP-DCA for solving $(\mathrm{VP})_\epsilon$

1: Take an initial point $(x^0, y^0) \in X \times Y$; $\delta_\beta > 0$; an initial penalty parameter $\beta_0 > 0$, $tol > 0$.
2: **for** $k = 0, 1, \dots$ **do**
   1. Solve the lower level problem P_{x^k} defined in (4.3) and obtain $\tilde{y}^k \in S(x^k)$ and $\gamma^k \in KT(x^k, \tilde{y}^k)$.
   2. Compute ξ_i^k, $i = 0, 1$ according to (29).
   2. Obtain an inexact solution (x^{k+1}, y^{k+1}) of (30).
   3. Stopping test. Compute $t^{k+1} = \max\{f(x^{k+1}, y^{k+1}) - f(x^k, \tilde{y}^k) - \langle \xi_1^k, x^{k+1} - x^k \rangle - \epsilon, 0\}$. Stop if $\max\{\|(x^{k+1}, y^{k+1}) - (x^k, y^k)\|, t^{k+1}\} < tol$.
   4. Penalty parameter update. Set
   $$\beta_{k+1} = \begin{cases} \beta_k + \delta_\beta, & \text{if } \max\{\beta_k, 1/t^{k+1}\} < \|(x^{k+1}, y^{k+1}) - (x^k, y^k)\|^{-1}, \\ \beta_k, & \text{otherwise.} \end{cases}$$
   5. Set $k := k + 1$.
3: **end for**

Thanks to Proposition 10, when $\epsilon > 0$, problem $(\mathrm{VP})_\epsilon$ satisfies ENNAMCQ automatically. Moreover, since the partial subgradient formula (24) holds, according to (25) in Theorem 3, the selection criteria in (29) implies that $\xi_1^k \in \partial v(x^k)$. Hence the convergence of iP-DCA for solving $(\mathrm{VP})_\epsilon$ (Algorithm 3) follows from Theorem 1 and Proposition 5.

Theorem 5 *Assume that F is bounded below on C. Let $\{(x^k, y^k)\}$ be an iteration sequence generated by Algorithm 3. Moreover assume that the partial subgradient formula (2) holds at every iteration point (x^k, y^k) and $KT(x^k, y) \neq \emptyset$ for all $y \in S(x^k)$. Suppose that either $\epsilon > 0$ or $\epsilon = 0$ and the penalty sequence $\{\beta_k\}$ is bounded. Than any accumulation point of $\{(x^k, y^k)\}$ is an KKT point of problem* $(\mathrm{VP})_\epsilon$.

We now assume that the lower level objective f is differentiable and ∇f is Lipschitz continuous with modulus L_f on set C. Given a current iterate (x^k, y^k), the next iterate (x^{k+1}, y^{k+1}) can be returned as an approximate minimizer of subproblem (30) with linearized f given by

$$\begin{aligned} \min_{(x,y) \in C} \quad & F_1(x, y) - \langle \xi_0^k, (x, y) \rangle + \frac{\rho_k}{2} \|(x, y) - (x^k, y^k)\|^2 \\ & + \beta_k \max\{f(x^k, y^k) + \langle \nabla f(x^k, y^k), (x, y) - (x^k, y^k) \rangle - f(x^k, \tilde{y}^k) - \langle \xi_1^k, x - x^k \rangle - \epsilon, 0\}. \end{aligned} \tag{31}$$

Under the assumption that $KT(x^k, y) \neq \emptyset$ for all $y \in S(x^k)$, the iteration scheme of iPL-DCA on $(\mathrm{VP})_\epsilon$ with $\epsilon \geq 0$ thus reads as:

Algorithm 4 iPL-DCA for solving $(\mathrm{VP})_\epsilon$

1: Take an initial point $(x^0, y^0) \in X \times Y$; $\delta_\beta, \sigma > 0$; $\beta_0 > 0$; $\rho_0 = \frac{1}{2}\beta_0 L_f + \sigma$; $tol > 0$.
2: **for** $k = 0, 1, \ldots$ **do**
   1. Solve the lower level problem P_{x^k} defined in (4.3) and obtain $\tilde{y}^k \in S(x^k)$ and $\gamma^k \in KT(x^k, \tilde{y}^k)$.
   2. Compute ξ_i^k, $i = 0, 1$ according to (29).
   2. Obtain an inexact solution (x^{k+1}, y^{k+1}) of (31).
   3. Stopping test. Compute

$$t^{k+1} = \max\{f(x^k, y^k) + \langle \nabla f(x^k, y^k), (x^{k+1}, y^{k+1}) - (x^k, y^k)\rangle - f(x^k, \tilde{y}^k) - \langle \xi_1^k, x^{k+1} - x^k\rangle - \epsilon, 0\}.$$

   Stop if $\max\{\|(x^{k+1}, y^{k+1}) - (x^k, y^k)\|, t^{k+1}\} < tol$.
   4. Penalty parameter update. Set

$$\beta_{k+1} = \begin{cases} \beta_k + \delta_\beta, & \text{if } \max\{\beta_k, 1/t^{k+1}\} < \|(x^{k+1}, y^{k+1}) - (x^k, y^k)\|^{-1}, \\ \beta_k, & \text{otherwies.} \end{cases}$$

$$\rho_{k+1} = \frac{1}{2}\beta_{k+1} L_f + \sigma.$$

   5. Set $k := k + 1$.
3: **end for**

The convergence of iPL-DCA follows from Theorem 2 and Proposition 6 directly.

Theorem 6 *Assume that F is bounded below on C, f is L_f smooth on C. Let the sequence $\{(x^k, y^k)\}$ be generated by Algorithm 4. Moreover assume that the partial subgradient formula (2) holds at every iteration point (x^k, y^k) and $KT(x^k, y) \neq \emptyset$ for all $y \in S(x^k)$. Suppose that either $\epsilon > 0$ or $\epsilon = 0$ and the penalty parameter sequence $\{\beta_k\}$ is bounded. Then any accumulation point of $\{(x^k, y^k)\}$ is an KKT point of problem* $(\mathrm{VP})_\epsilon$.

5 Numerical experiments on SV bilevel model selection

In this section, we will conduct numerical experiments on the bilevel model for SV classification which is equivalent to (1):

$$\begin{aligned} \min \quad & \Theta(\mathbf{w}^1, \ldots, \mathbf{w}^T, \mathbf{c}) := \frac{1}{T}\sum_{t=1}^{T} \frac{1}{|\Omega_{val}^t|} \sum_{j \in \Omega_{val}^t} \max(1 - b_j(\mathbf{a}_j^T \mathbf{w}^t - c_t), 0) \\ s.t. \quad & \frac{1}{\lambda_{ub}} \leq \mu \leq \frac{1}{\lambda_{lb}}, \quad \bar{\mathbf{w}}_{lb} \leq \bar{\mathbf{w}} \leq \bar{\mathbf{w}}_{ub}, \\ & (\mathbf{w}^1, \ldots, \mathbf{w}^T, \mathbf{c}) \in \operatorname*{argmin}_{-\bar{\mathbf{w}} \leq \mathbf{w}^t \leq \bar{\mathbf{w}}, c_t \in \mathbb{R}} \sum_{t=1}^{T} \left(\frac{\|\mathbf{w}^t\|^2}{2\mu} + \sum_{j \in \Omega_{trn}^t} \max(1 - b_j(\mathbf{a}_j^T \mathbf{w}^t - c_t), 0) \right). \end{aligned} \tag{32}$$

Here we use the hinge loss as the cross validation error. One can find other functions that can be used for cross validation error in [5,18,19]. Let $x :=$

$(\mu, \bar{\mathbf{w}})$, $y := (\mathbf{w}^1, \ldots, \mathbf{w}^T, \mathbf{c})$, $X = [\frac{1}{\lambda_{ub}}, \frac{1}{\lambda_{lb}}] \times [\bar{\mathbf{w}}_{lb}, \bar{\mathbf{w}}_{ub}]$, $Y = \mathbb{R}^{(n+1)T}$,

$$f(x,y) = \sum_{t=1}^{T} \left(\frac{\|\mathbf{w}^t\|^2}{2\mu} + \sum_{j \in \Omega_{trn}^t} \max(1 - b_j(\mathbf{a}_j^T \mathbf{w}^t - c_t), 0) \right),$$

and

$$g(x,y) = \begin{pmatrix} g_1(x,y) \\ \vdots \\ g_T(x,y) \end{pmatrix} \quad \text{and} \quad g_t(x,y) = \begin{pmatrix} -\bar{\mathbf{w}} - \mathbf{w}^t \\ \mathbf{w}^t - \bar{\mathbf{w}} \end{pmatrix}, t = 1, \ldots, T.$$

Obviously F, f and g are all convex functions defined an open set containing $X \times Y$. Both F and f are bounded below on $X \times Y$. When $\bar{\mathbf{w}}_{ub} \geq \bar{\mathbf{w}}_{lb} > 0$, $\mathcal{F}(x) \neq \emptyset$ for an open set containing X. And since $b_j \in \{-1, 1\}$, $f(x,y)$ is coercive and continuous with respect to lower-level variable y for any given x in an open set containing X, thus $S(x) \neq \emptyset$ for all x in an open set containing X. The function g is smooth and f is a sum of a smooth function and a function which is independent of variable x. Hence by Proposition 2, the partial differential formula (2) holds at each point (x,y). Since the lower level constraints are affine, KKT conditions holds at any $y \in S(x)$ for any $x \in X$. Therefore, all conditions required by the convergence results of iP-DCA in Theorem 5 are satisfied.

We will compare our proposed algorithms with the MPEC approach considered in [19], in which the authors reformulate such bilevel model into MPEC by replacing the lower level problem with its KKT optimality condition and then apply nonlinear program solver to solve the obtained MPEC. In numerical experiments, we will follow the suggestions given in [19] to replace the complementarity constraints with the relaxed complementarity constraints. As claimed by [19], such approach can facilitate an early termination of cross-validation and ease the difficulty of dealing the complementarity constraints for nonlinear program solver.

5.1 Numerical tests

All the numerical experiments are implemented on a laptop with Intel(R) Core(TM) i7-9750H CPU@ 2.60GHz and 32.00 GB memory. All the codes are implemented on MATLAB 2019b. The subproblems in iP-DCA are all convex optimization problems and we apply the Matlab software package SDPT3 [41,42] with default setting to solve them. MPEC problem is solved by using **fmincon** in Matlab optimization toolbox with setting $'\mathbf{Algorithm}'$ being $'\mathbf{interior-point}'$, $'\mathbf{MaxIterations}'$ being 200, $'\mathbf{MaxFunctionEvaluations}'$ being 10^6 and $'\mathbf{OptimalityTolerance}'$ being 10^{-6}. As **fmincon** needs extremely long time to solve large dimension MPEC problems, we first use small size datasets to conduct the numerical comparison between iP-DCA and

MPEC approach. We test here three real datasets "australian_scale", "breast-cancer_scale" and "diabetes_scale" downloaded from the SVMLib repository [6][1]. Each dataset is randomly split into a training set Ω with $|\Omega| = \ell_{train}$ data pairs, which is used in the cross-validation bilevel model and a hold-out test set $\mathcal{N}$ with $|\mathcal{N}| = \ell_{test}$ data pairs. We give the descriptions of datasets in Table 1. For each dataset, we use a three-fold cross-validation in SV classification bilevel model (32), i.e. $T = 3$, and that each training fold consists of two-thirds of the total training data and validation fold consists of one-third of the total training data. We repeat the experiments 20 times for each dataset. The values of parameters in SV classification bilevel model (32) are set as: $\lambda_{lb} = 10^{-4}$, $\lambda_{ub} = 10^{4}$, $\bar{\mathbf{w}}_{lb} = 10^{-6}$ and $\bar{\mathbf{w}}_{ub} = 1.5$. For our approach, we test three different values of relaxation parameter $\epsilon \in \{0, 10^{-2}, 10^{-4}\}$ in $(\text{VP})_\epsilon$. And the value of relaxation parameter of the relaxed complementarity constraints in MPEC is set to be 10^{-6}. The initial points for both iP-DCA and MPEC approach are chosen as $\lambda = 1$, $\bar{\mathbf{w}} = 0.1\mathbf{e}$, where $\mathbf{e}$ denotes the vector whose elements are all equal to 1, and the values of other variables are all equal to 0. These settings are used for all experiments. Parameters in iP-DCA are set as $\beta_0 = 1$, $\rho = 10^{-2}$ and $\delta_\beta = 5$. And we terminate iP-DCA when $t^{k+1} < 10^{-4}$ and $\|(x^{k+1}, y^{k+1}) - (x^k, y^k)\|/(1 + \|(x^k, y^k)\|) < tol$.

For each experiment, after we obtain the hyperparameters $\hat{\mu}$ and $\hat{\bar{\mathbf{w}}}$ from implementing our proposed algorithm for the bilevel SV classification model and MPEC approach, we calculate their corresponding cross-validation error (CV error) and test error for comparing the performances of these two methods. For calculating the CV error, we put $\hat{\mu}$ and $\hat{\bar{\mathbf{w}}}$ back to the lower level problem in (32) to get the corresponding lower level solution $(\hat{\mathbf{w}}^1, \ldots, \hat{\mathbf{w}}^T, \hat{\mathbf{c}})$ and calculate the corresponding cross-validation error $\Theta(\hat{\mathbf{w}}^1, \ldots, \hat{\mathbf{w}}^T, \hat{\mathbf{c}})$. Next, as in [19] we implement a post-processing procedure to calculate the generalization error on the hold-out data for each instance. In particular as suggested in [19], since only two thirds of the data in Ω was used in each fold while in testing we use all the training data from Ω, we should solve the following support vector classification problem with $\frac{3}{2}\hat{\lambda} = \frac{3}{2\hat{\mu}}$ and $\hat{\bar{\mathbf{w}}}$ as hyperparameter

$$\min_{\substack{-\hat{\bar{\mathbf{w}}} \leq \mathbf{w} \leq \hat{\bar{\mathbf{w}}} \\ c \in \mathbb{R}}} \left\{ \frac{3}{4\hat{\mu}} \|\mathbf{w}\|^2 + \sum_{j \in \Omega} \max(1 - b_j(\mathbf{a}_j^T \mathbf{w} - c), 0) \right\}$$

to obtain the final classifier $(\hat{\mathbf{w}}, \hat{c})$. Then the test (hold-out) error rate is calculated as:

$$\text{Test error} = \frac{1}{\ell_{test}} \sum_{i \in \mathcal{N}} \frac{1}{2} |\text{sign}(\mathbf{a}_i^T \hat{\mathbf{w}} - \hat{c}) - b_i|,$$

where $\text{sign}(x)$ denote the sign function. Note that for each $(\mathbf{a}_i, b_i)$ in the test set $\mathcal{N}$, $|\text{sign}(\mathbf{a}_i^T \hat{\mathbf{w}} - \hat{c}) - b_i|$ is either equal to zero or 2 and hence the test error is the average misclassification by the final classifier. The achieved numerical results averaged over 20 repetitions for each dataset are reported in Table 2.

[1] http://www.csie.ntu.edu.tw/ cjlin/libsvmtools/datasets/.

Observe from Table 2 that our proposed iP-DCA always achieves a smaller cross-validation error, which is exactly the value of upper level objective of bilevel problem (32). Furthermore, the time taken by our proposed iP-DCA is much shorter than MPEC approach. And the test errors of our proposed iP-DCA and MPEC approach are similar and iP-DCA achieves a smaller test error than MPEC approach on dataset “diabetes_scale”. And it is observed that different values of ϵ and tol do not influence the cross-validation error obtained by iP-DCA much. iP-DCA can obtain a relatively good solution without requiring a small tol.

Next, we are going to test our proposed iP-DCA on two large scale datasets “mushrooms” and “phishing” downloaded from the SVMLib repository. The descriptions of datasets are given in Table 1. We set $tol = 10^{-2}$ for these tests. The numerical results averaged over 20 repetitions for each dataest are reported in Table 3. It can be observed from Table 3 that different values of ϵ and tol do not influence the cross-validation error obtained by iP-DCA much but the case $\epsilon = 0$ takes more time to achieve desired tolerance. And iP-DCA can obtain a satisfactory solution within an acceptable time on large scale problems.

Table 1 Description of datasets used

Dataset	ℓ_{train}	ℓ_{test}	n	T
australian_scale	345	345	14	3
breast-cancer_scale	339	344	10	3
diabetes_scale	384	384	8	3
mushrooms	4062	4062	112	3
phishing	5526	5529	68	3

6 Concluding remarks

Motivated by hyperparameter selection problems, in this paper we develop two DCA type algorithms for solving the DC bilevel program. Our numerical experiments on the bilevel model of SV classification problems show that our approach is promising. Due to the space limit, we are not able to present more studies for more complicated models in hyperparameter selection problems. We hope to study these problems in our future work.

References

1. Allende, G., Still, G.: Solving bilevel programs with the KKT-approach. Mathematical Programming. **138**, 309-332 (2013)
2. Bai, K., Ye, J.J.: Directional necessary optimality conditions for bilevel programs. arXiv:2004.01783v2. (2020)

Table 2 Numerical results comparing iP-DCA and MPEC approach

Dataset	Method	CV error	Test error	Time(sec)
australian_scale	iP-DCA($\epsilon = 0$, $tol = 10^{-2}$)	0.28 ± 0.03	0.15 ± 0.01	73.7 ± 106.6
	iP-DCA($\epsilon = 0$, $tol = 10^{-3}$)	0.28 ± 0.03	0.15 ± 0.01	81.2 ± 110.8
	iP-DCA($\epsilon = 10^{-2}$, $tol = 10^{-2}$)	0.28 ± 0.03	0.15 ± 0.01	10.7 ± 6.3
	iP-DCA($\epsilon = 10^{-2}$, $tol = 10^{-3}$)	0.28 ± 0.03	0.15 ± 0.01	128.7 ± 74.4
	iP-DCA($\epsilon = 10^{-4}$, $tol = 10^{-2}$)	0.28 ± 0.03	0.15 ± 0.01	74.2 ± 123.8
	iP-DCA($\epsilon = 10^{-4}$, $tol = 10^{-3}$)	0.28 ± 0.03	0.15 ± 0.01	109.0 ± 141.0
	MPEC approach	0.29 ± 0.04	0.15 ± 0.01	491.2 ± 245.1
breast-cancer_scale	iP-DCA($\epsilon = 0$, $tol = 10^{-2}$)	0.06 ± 0.01	0.04 ± 0.00	53.1 ± 67.2
	iP-DCA($\epsilon = 0$, $tol = 10^{-3}$)	0.06 ± 0.01	0.04 ± 0.00	78.3 ± 73.9
	iP-DCA($\epsilon = 10^{-2}$, $tol = 10^{-2}$)	0.06 ± 0.01	0.04 ± 0.00	15.5 ± 2.1
	iP-DCA($\epsilon = 10^{-2}$, $tol = 10^{-3}$)	0.06 ± 0.01	0.04 ± 0.00	108.9 ± 40.4
	iP-DCA($\epsilon = 10^{-4}$, $tol = 10^{-2}$)	0.06 ± 0.01	0.04 ± 0.01	24.6 ± 17.5
	iP-DCA($\epsilon = 10^{-4}$, $tol = 10^{-3}$)	0.06 ± 0.01	0.04 ± 0.01	86.8 ± 59.3
	MPEC approach	0.08 ± 0.01	0.04 ± 0.01	294.5 ± 98.2
diabetes_scale	iP-DCA($\epsilon = 0$, $tol = 10^{-2}$)	0.56 ± 0.03	0.24 ± 0.02	12.0 ± 13.6
	iP-DCA($\epsilon = 0$, $tol = 10^{-3}$)	0.56 ± 0.03	0.24 ± 0.02	25.9 ± 33.2
	iP-DCA($\epsilon = 10^{-2}$, $tol = 10^{-2}$)	0.57 ± 0.03	0.24 ± 0.02	3.1 ± 0.6
	iP-DCA($\epsilon = 10^{-2}$, $tol = 10^{-3}$)	0.56 ± 0.03	0.24 ± 0.02	62.1 ± 31.7
	iP-DCA($\epsilon = 10^{-4}$, $tol = 10^{-2}$)	0.56 ± 0.03	0.24 ± 0.02	12.7 ± 19.7
	iP-DCA($\epsilon = 10^{-4}$, $tol = 10^{-3}$)	0.56 ± 0.03	0.24 ± 0.02	39.2 ± 45.7
	MPEC approach	0.59 ± 0.03	0.25 ± 0.02	346.7 ± 216.9

Table 3 Numerical results of iP-DCA on datasets "mushrooms" and "phishing" with $tol = 10^{-2}$

Dataset	Method	CV error	Test error	Time(sec)
mushrooms	iP-DCA($\epsilon = 0$)	6.36e-04 ± 5.94e-04	0 ± 0	334.3 ± 346.1
	iP-DCA($\epsilon = 10^{-2}$)	1.53e-03 ± 3.85e-03	3.57e-04 ± 1.34e-03	109.3 ± 35.2
	iP-DCA($\epsilon = 10^{-4}$)	6.38e-04 ± 6.08e-04	0 ± 0	162.9 ± 27.4
phishing	iP-DCA($\epsilon = 0$)	0.29 ± 0.00	0.09 ± 0.00	357.9 ± 95.2
	iP-DCA($\epsilon = 10^{-2}$)	0.29 ± 0.00	0.09 ± 0.00	222.1 ± 18.9
	iP-DCA($\epsilon = 10^{-4}$)	0.29 ± 0.00	0.09 ± 0.00	215.4 ± 46.5

3. Bard, J.: Practical Bilevel Optimization: Algorithms and Applications. Kluwer Academic Publishers, Dordrecht (1998)
4. Ben-Tal, A., Blair, C.: *Computational difficulties of bilevel linear programming.* Operations Research. **38**, 556-560 (1990)
5. Bennett, K.P., Hu, J., Ji., X., Kunapuli, G., Pang, J.-S.: Model selection via bilevel optimization, in The 2006 IEEE International Joint Conference on Neural Network Proceedings, 1922-1929 (2006)
6. Chang, C-C., Lin, C-J.: LIBSVM : a library for support vector machines. ACM Transactions on Intelligent Systems and Technology. **2**(3), 1-27 (2011)
7. Clarke, F.H.: Optimization and Nonsmooth Analysis. Society for Industrial and Applied Mathematics, Philadelphia (1990)
8. Clarke, F.H., Ledyaev, Y.S., Stern, R.J., Wolenski, P.R.: Nonsmooth Analysis and Control Theorey. Springer Science & Business Media, New York (1998)
9. Colson, B., Marcotte, P., Savard, G.: An overview of bilevel optimization. Annals of Operations Research. **153**(1-2), 235-256 (2007)
10. Dempe, S.: Foundations of Bilevel Programming. Kluwer Academic Publishers, Dordrecht (2002)
11. Dempe, S., Dutta, J.: Is bilevel programming a special case of mathematical programming with equilibrium constraints?. Mathematical Programming. **131**, 37-48 (2012)
12. Dempe, S., Zemkoho, A.: Bilevel optimization: advances and next challenges. Springer Optimization and its Applications, vol. 161 (2020)
13. Dempe, S., Kalashnikov, V., Pérez-Valdés, G., Kalashnykova, N.: Bilevel Programming Problems. Energy Systems, Springer Science & Business Media, Berlin (2015)
14. Franceschi, L., Frasconi, P., Salzo, S., Grazzi, R., Pontil, M.: Bilevel programming for hyperparameter optimization and meta-learning. In: International Conference on Machine Learning. **80**, 1568-1577 (2018)

15. Jourani, A.: Constraint qualifications and Lagrange multipliers in nondifferentiable programming problems. Journal of Optimization Theory and Applications. **81**, 533-548 (1994)
16. Henrion, R., Jourani, A., Outrata, J.V.: On the calmness of a class of multifunctions. SIAM Journal on Optimization. **13**, 603–618 (2002)
17. Horst, R., Thoai, N.V.: DC programming: overview. Journal of Optimization Theory and Applications. **103**(1), 1–43 (1999)
18. Kunapuli, G.: A bilevel optimization approach to machine learning. Ph.D Thesis. (2008)
19. Kunapuli, G., Bennett, K.P., Hu, J., Pang, J-S.: Classification model selection via bilevel programming. Optimization Methods and Software. **23**, 475-489 (2008)
20. Kunapuli, G., Bennett, K.P., Hu, J., Pang, J-S.: Bilevel model selection for support vector machines. In: CRM proceedings and lecture notes. **45**, 129-158 (2008)
21. Lampariello, L., Sagratella, S.: Numerically tractable optimistic bilevel problems. Computational Optimization and Applications. **76**, 277-303 (2020)
22. Lin, G., Xu, M., Ye, J.J.: On solving simple bilevel programs with a nonconvex lower level program. Mathematical Programming. **144**, 277-305 (2014)
23. Liu, R., Mu, P., Yuan, X., Zeng, S., Zhang, J.: A generic first-order algorithmic framework for bi-level programming beyond lower-level singleton. In: International Conference on Machine Learning. **119**, 6305-6315 (2020)
24. Liu, R., Mu, P., Yuan, X., Zeng, S., Zhang, J.: A generic descent aggregation framework for gradient-based bi-level optimization. arXiv:2102.07976. (2021)
25. Luo, Z-Q., Pang, J-S., Ralph, D.: Mathematical Programs with Equilibrium Constraints. Cambridge University Press, Cambridge (1996)
26. Mirrlees, J.A.: The theory of moral hazard and unobservable behaviour: Part I. The Review of Economic Studies. **66**, 3-21 (1999)
27. Moore, G.: Bilevel programming algorithms for machine learning model selection. Ph.D Thesis. (2010)
28. Moore, G., Bergeron, C., Bennett, K.P.: Model selection for primal SVM. Machine Learning. **85**, 175-208 (2011)
29. Nie, J., Wang, L., Ye, J.J.: Bilevel polynomial programs and semidefinite relaxation methods. SIAM Journal on Optimization. **27**, 1728-1757 (2017)
30. Nie, J., Wang, L., Ye, J.J., Zhong, S.: A Lagrange Multiplier Expression Method for Bilevel Polynomial Optimization, SIAM Journal on Optimization. revised.
31. Okuno, T., Kawana, A.: Bilevel optimization of regularization hyperparameters in machine learning. In: Bilevel Optimization: Advances and Next Challenges, Ch. 6. Springer Optimization and its Applications, vol. 161 (2020).
32. Outrata, J. V.: On the numerical solution of a class of Stackelberg problems, ZOR - Methods and Models of Operations Research. **34**, 255–277 (1990)
33. Outrata, J., Kocvara, M., Zowe, J.: Nonsmooth Approach to Optimization Problems with Equilibrium Constraints: Theory, Applications and Numerical Results. Kluwer Academic Publishers, Boston (1998)
34. Rockafellar, R.T.: Convex Anlysis. Princeton University Press, Princeton (1970)
35. Rockafellar, R.T.: Conjugate duality and optimization. CBMS-NSF Regional Conference Series in Applied Mathematics. **16**, 1-74 (1974)
36. Shimizu, K., Ishizuka, Y., Bard, J.: Nondifferentiable and Two-level Mathematical Programming. Kluwer Academic Publishers, Dordrecht (1997)
37. Stackelberg, H.: Market Structure and Equilibrium. Springer Science & Business Media, Berlin (2010)
38. Stephen, B., Vandenberghe, L.: Convex Optimization. Cambridge University Press, Cambridge (2004)
39. Thi, H.A.L., Dinh, D.T.: Advanced Computational Methods for Knowledge Engineering, DC programming and DCA for general DC programs. pp. 15-35. Springer, Cham, Switzerland (2014)
40. Thi, H.A.L., Dinh, T.P.: DC programming and DCA: thirty years of developments. Math. Program. **169**, 5-68 (2018)
41. Toh, K.C., Todd, M.J., Tutuncu, R.H.: SDPT3 — a Matlab software package for semidefinite programming, Optimization Methods and Software. **11**, 545–581 (1999)
42. Tutuncu, R.H., Toh, K.C., Todd, M.J.: Solving semidefinite-quadratic-linear programs using SDPT3. Mathematical Programming, Series B. **95**, 189–217 (2003)

43. Ye, J.J.: Constraint qualifications and optimality conditions in bilevel optimization. In: Bilevel Optimization: Advances and Next Challenges, Ch. 8. Springer Optimization and its Applications, vol. 161 (2020).
44. Ye J.J., Zhu, D. L.: Optimality conditions for bilevel programming problems. Optimization. **33**, 9–27 (1995)