Computational Optimization and Applications, 15, 145–166 (2000)

A Bundle Algorithm Applied to Bilevel Programming Problems with Non-Unique Lower Level Solutions

S. DEMPE dempe@math.tu-freiberg.de
Freiberg University of Mining and Technology, Freiberg, Germany

Received August 6, 1997; Accepted October 8, 1998

Abstract. In the paper, the question is investigated if a bundle algorithm can be used to compute approximate solutions for bilevel programming problems where the lower level optimal solution is in general not uniquely determined. To give a positive answer to this question, an appropriate regularization approach is used in the lower level. In the general case, the resulting algorithm computes an approximate solution. If the problem proves to have strongly stable lower level solutions for all parameter values in a certain neighborhood of the stationary solutions of the bilevel problem, convergence to stationary solutions can be shown.

Keywords: bilevel programming, parametric optimization, bundle algorithm, nondifferentiable optimization

1. Introduction

Bilevel programming problems are hierarchical optimization problems where a part of the unknowns is restricted to be an optimal solution of a second problem parametrized in the remaining variables. Formally, bilevel programming problems can be stated as follows:

$$\begin{cases} F(x, y) \to \text{“}\min_{y}\text{”} \\ y \in Y, \quad x \in \Psi(y), \end{cases} \tag{1}$$

where $\Psi(\cdot)$ is equal to the set of optimal solutions of a second problem

$$\begin{cases} f(x, y) \to \min_{x} \\ g(x, y) \leq 0. \end{cases} \tag{2}$$

Here, all functions used to formulate these problems are assumed to be sufficiently smooth: $F,\ f,\ g_i \in C^2(\mathbb{R}^n \times \mathbb{R}^m, \mathbb{R}),\ i = 1, \ldots, p$, and $Y \subseteq \mathbb{R}^m$ is a fixed closed set, $g(x, y) = (g_1(x, y), \ldots, g_p(x, y))^T$. Bilevel problems have a large number of potential applications (see e.g. [22] and the references therein).

In the formula (1) we have used quotation marks to express the uncertainty of minimization with respect to y only in the case when the lower level optimal solution $x \in \Psi(y)$ is not uniquely determined. Then, the value of $F(x, y)$ cannot be predicted in general without knowledge of the response $x \in \Psi(y)$ of the lower level problem.

In most of the references investigating this setting, one of two ways out of this situation are used: An optimistic approach [6, 11] and a pessimistic one [19, 21]. Both approaches lead to the necessity of minimizing a nonsmooth (in general even discontinuous) auxiliary function:

$$\phi_o(y) := \min_x\{ f(x, y) : x \in \Psi(y)\}$$

in the optimistic and

$$\phi_p(y) := \max_x\{ f(x, y) : x \in \Psi(y)\}$$

in the pessimistic cases. Moreover, it is difficult to decide if one of them is really adequate to the considered situation.

In this paper neither of these approaches is discussed. Instead, a regularization approach is used to circumvent the difficulties of non-uniquely determined lower level solutions: Replace the problem (2) by

$$\begin{cases} f(x, y) + \alpha\|x\|^2 \to \min\limits_x \\ g(x, y) \leq 0 \end{cases} \tag{3}$$

for $\alpha > 0$. Let $\Psi_\alpha(y)$ denote the set of optimal solutions for this problem. Some relations between the regularized and the original bilevel problems have been investigated in [20]. Another regularization has been used in [6]. We will show that, for fixed $\alpha > 0$, and under presumably not too restrictive assumptions, the optimal solution of problem (3) is uniquely determined by a locally Lipschitz continuous function $x_\alpha(y)$ with respect to α, y. Hence, the regularized problem

$$\min_y\{G_\alpha(y) := F(x_\alpha(y), y) : y \in Y\} \tag{4}$$

is a Lipschitz optimization problem for $\alpha > 0$ and can be solved by means of nondifferentiable minimization techniques as e.g. the bundle-trust region algorithm [16, 29]. This approach has been successfully applied to problem (1) in the case when the lower level problem's optimal solution is strongly stable [17] for all y and the linear independence constraint qualification is satisfied there [7, 24, 25]. Here we will apply this method to the case when neither assumption is valid. This will lead to a modified bundle algorithm which, in general, is only an approximation algorithm to the bilevel problem. If the lower level solutions are strongly stable for all parameter values in a small neighborhood of the stationary solutions, this algorithm is also exact.

2. Properties of the lower level optimal solution

To avoid the difficulties which arise when disconnected sets $\Psi_\alpha(y)$ of global optimal solutions arise in (3) we will assume that the lower level problem is a convex one satisfying some compactness assumptions:

(C) All functions $f(\cdot, y)$, $g_i(\cdot, y)$, $i = 1, \ldots, p$, are convex on $\mathbb{R}^n$ and the set $\{x : g(x, y) \leq 0\}$ is not empty and compact for each fixed $y \in Y$.

We will also use two constraint qualifications:

(MF) For each $y \in Y$ there exists $\bar{x}$ satisfying $g(\bar{x}, y) < 0$.

Let $L(x, y, \lambda) = f(x, y) + \sum_{i=1}^{p} \lambda_i g_i(x, y)$ denote the Lagrangean function of problem (2). If (C) and (MF) are satisfied it is well-known that a point x satisfying $g(x, y) \leq 0$ is an optimal solution of problem (2) if and only if the set of Lagrange multiplier vectors

$$\Lambda(x, y) = \{\lambda \geq 0 : \nabla_x L(x, y, \lambda) = 0,\ \lambda^T g(x, y) = 0\}$$

is not empty. In this case $\Lambda(x, y)$ is a bounded polyhedron [9].

(CR) For each $\alpha \geq 0$ and for each (x^0, y^0) satisfying $x^0 \in \Psi_\alpha(y^0)$ there exists an open neighborhood V of (x^0, y^0) such that: for all $I \subseteq \{j : g_j(x^0, y^0) = 0\}$ the family of gradients $\{\nabla_x g_i(x, y) : i \in I\}$ has constant rank on V.

This constant rank constraint qualification has been used e.g. in [12, 13].

Due to strong convexity of the objective function in the regularized lower level problem (3), the optimal solution of this problem is uniquely determined for positive regularization parameter $\alpha > 0$. This property is essentially used in the second assertion of Theorem 2.1 below. With respect to relations between the sets $\Psi_\alpha(y)$ and $\Psi(y)$ we have by straightforward application of the results in [1, Theorems 3.1.1, 3.1.5, and 4.2.2]

Theorem 2.1. *Consider the parametric problems* (2) *and* (3) *and let the assumptions* (*C*) *and* (*MF*) *be satisfied. Then*

1. *For each sequences* $\{y^k\}_{k=1}^{\infty} \subseteq Y$, $\{\alpha^k\}_{k=1}^{\infty} \subseteq \mathbb{R}_+$ *converging to* $\bar{y}$, $\bar{\alpha}$, *resp., and for each sequence* $\{x^k\}_{k=1}^{\infty}$ *satisfying* $x^k \in \Psi_{\alpha^k}(y^k)$ $\forall k$ *the sequence* $\{x^k\}_{k=1}^{\infty}$ *has accumulation points* $\bar{x}$ *and all these points satisfy* $\bar{x} \in \Psi_{\bar{\alpha}}(\bar{y})$.
2. *For* $\bar{\alpha} = 0$ *we have*

$$\lim_{\substack{y^k \to \bar{y} \\ \alpha^k \searrow 0}} x_{\alpha^k}(y^k) = x(\bar{y})$$

provided that $\Psi(\bar{y}) = \{x(\bar{y})\}$.

It should be noted that in general we do not have

$$\lim_{\substack{y^k \to \bar{y} \\ \alpha^k \searrow 0}} x_{\alpha^k}(y^k) \in \operatorname*{Argmin}_x \{\|x\| : x \in \Psi(\bar{y})\}$$

without the assumption that $\Psi(\bar{y})$ reduces to a singleton, even if this limit exists. This can be seen in

Example 2.2. Consider the problem

$$\min_x\{xy : x \in [-1, 1]\}.$$

Then,

$$\Psi(y) = \begin{cases} [-1, 1], & \text{if } y = 0, \\ \{1\}, & \text{if } y < 0, \\ \{-1\}, & \text{if } y > 0, \end{cases}$$

and

$$\Psi_\alpha(y) = \begin{cases} \{-y/(2\alpha)\} & \text{if } -y/(2\alpha) \in [-1, 1], \\ \{1\}, & \text{if } -y/(2\alpha) \geq 1, \\ \{-1\}, & \text{if } -y/(2\alpha) \leq -1. \end{cases}$$

for $y \in [-2\alpha, 2\alpha]$. Hence, depending on the limit of the sequence $-y^k/\alpha^k$, the sequence $x_{\alpha^k}(y^k)$ can have any limit point in $[-1, 1]$ for $y^k \to 0,\ \alpha^k \to 0$. □

This drawback results from the inherent difficulties in the bilevel problem with nonunique lower level solutions. When solving this problem we have either to minimize a discontinuous, only implicitly determined function (which is also not lower semicontinuous in the case of the pessimistic approach) or we have to minimize a certain continuous relaxation of it. Such a relaxation can be obtained by the way of regularizing the lower level problem, but then we loose the knowledge of some properties of the limit point $x_0(y^0)$ of the sequence of lower level solutions $\{x_{\alpha^k}(y^k)\}_{k\in N}$ computed in any iteration algorithm. We only know that the limit point $(x_0(y^0), y^0)$ of the sequence $\{(x_{\alpha^k}(y^k), y^k)\}_{k\in N}$ is feasible for the bilevel problem, i.e. $x_0(y^0) \in \Psi(y^0)$. Hopefully, the point $(x_0(y^0), y^0)$ has a good function value for the bilevel programming problem. It is even possible that this function value is better than that obtained when minimizing the functions arising in either the optimistic or the pessimistic approach. For theoretical investigations of how to get a solution $(x_0(y^0), y^0)$ such that $x_0(y^0)$ is the least norm element in the solution set of the lower level problem, the reader is referred to [20].

For a fixed value $\alpha > 0$, the optimal solution of problem (3) is uniquely determined and the function $x_\alpha(\cdot)$ of optimal solutions for these problems is continuous. In order to apply the bundle algorithm to problem (4) we need even more: The objective function of this problem should be locally Lipschitz continuous. To meet this property we have to add an additional assumption as (CR) [5]. Then, the next theorem shows that the local behavior of the thus well-defined function $x.(\cdot)$ is even better: the function $x.(\cdot)$ is locally a PC^1-function, i.e. a continuous selection of finitely many continuously differentiable functions:

Definition 2.3. A function $h : \mathbb{R}^q \to \mathbb{R}^r$ is called a PC^1-function locally at $w^0 \in \mathbb{R}^q$ if there exist an open neighborhood W of w^0 and finitely many continuously differentiable functions $h_i : W \to \mathbb{R}^r,\ i = 1, \ldots, t$, such that

1. h is continuous on W and
2. $h(w) \in \{h_1(w), \ldots, h_t(w)\}\ \forall w \in W$.

Theorem 2.4. *Let $x_\alpha(y)$ be an optimal solution of the problem* (3) *for $y \in Y,\ \alpha \geq 0$. Take $y^0 \in Y,\ \alpha^0 > 0$ and assume that* (*C*), (*MF*), *and* (*CR*) *are satisfied. Then, the function $x.(\cdot)$ is a PC^1-function locally at (α^0, y^0).*

This theorem is a direct consequence of [26] since the strong sufficient optimality condition of second order is implied by (C) and strong convexity of the Euclidean norm. The proof of this theorem in [26] has shown that in place of the selection functions used in the representation of $x.(\cdot)$ as a PC^1-function we can use functions $x_\alpha^I(y)$ describing local optimal solutions of the modified problems

$$\min_x \{ f(x, y) + \alpha \|x\|^2 : g_i(x, y) = 0,\ i \in I\}, \tag{5}$$

where I is a subset of $\{1, \ldots, p\}$ satisfying the following two conditions:

1. $\{j : \lambda_j > 0\} \subseteq I \subseteq \{j : g_j(x_{\alpha^0}(y^0), y^0) = 0\}$ for some Lagrange multiplier vector $\lambda \in \Lambda_{\alpha^0}(x_{\alpha^0}(y^0), y^0)$ and
2. the gradients $\nabla_x g_i(x_{\alpha^0}(y^0), y^0),\ i \in I$, are linearly independent.

Here, $\Lambda_\alpha(x, y) = \{\lambda \geq 0 : \nabla_x L(x, y, \lambda) + 2\alpha x = 0,\ \lambda^T g(x, y) = 0\}$ is the set of Lagrange multipliers for problem (3). Let $\mathcal{I} = \mathcal{I}(x_{\alpha^0}(y^0), y^0)$ denote the family of all sets having these two properties. Under (C) and (MF), this family is not empty. To see this, remember that, for each vertex $\lambda \in \Lambda_{\alpha^0}(x_{\alpha^0}(y^0), y^0)$, the set $I = \{j : \lambda_j > 0\} \in \mathcal{I}$.

PC^1-functions are locally Lipschitz continuous [10] Directional differentiability of the function $x.(\cdot)$ is guaranteed even without (CR) [3, 30]. These results can be applied for solving the bilevel programming problem (1) by means of a descent algorithm [6]. For the use of other algorithms of nondifferentiable optimization as e.g. the bundle algorithm as desired in this paper we need a tool for computing at least one generalized gradient of the locally Lipschitz continuous function $G_\alpha(\cdot) := F(x_\alpha(\cdot), \cdot)$ at each iteration point y. For this, the following theorem is very helpful which gives the principal possibility for the computation of the subdifferential of the function $G_\alpha(\cdot)$:

Theorem 2.5 [18, 27]**.** *Let $h : \mathbb{R}^q \to \mathbb{R}^r$ be a PC^1-function locally at w^0. Then, its generalized Jacobian in the sense of Clarke* [2] *is equal to*

$$\partial h(w^0) = \operatorname{conv}\{\nabla h_i(w^0) : w^0 \in \operatorname{cl}(\operatorname{int} T_i)\},$$

where

$$T_i = \{w : h(w) = h_i(w)\}, \quad i = 1, \ldots, t.$$

3. The generalized Jacobian of the lower level optimal solution

For our later investigations we need the generalized Jacobian of the function $x.(\cdot)$ with respect to y only. Due to Theorem 2.5, two questions are to be answered for its description: On the first hand a method for the computation of the Jacobian for each selection function is needed. On the second hand, a rule for verification if $y^0 \in \mathrm{cl}\,(\mathrm{int}\, T_i)$ is searched for.

The selection functions $x^I_\alpha(\cdot)$ are functions describing local optimal solutions of the modified lower level problems (5) for fixed $\alpha > 0$. By assumption (C) the objective function of this problem has a positive definite Hessian matrix with respect to x for each $y \in Y, \alpha > 0$ and the linear independence constraint qualification is satisfied for this equality-constrained problem by the second property of the set $I \in \mathcal{I}$. By [8] this implies that the Jacobian of the function $x^I_\alpha(\cdot)$ can be computed by solving the following system of equations:

$$M^0 \begin{pmatrix} \nabla x^I_\alpha(y^0) \\ \mu \end{pmatrix} = N^0, \tag{6}$$

where

$$M^0 = \begin{pmatrix} \nabla^2_{xx} L_I\big(x^I_\alpha(y^0), y^0, \lambda\big) + 2\alpha E & \nabla^T_x g_I\big(x^I_\alpha(y^0), y^0\big) \\ \nabla_x g_I\big(x^I_\alpha(y^0), y^0\big) & 0 \end{pmatrix},$$

$$N^0 = \begin{pmatrix} -\nabla^2_{yx} L_I\big(x^I_\alpha(y^0), y^0, \lambda\big) \\ -\nabla_y g_I\big(x^I_\alpha(y^0), y^0\big) \end{pmatrix}$$

for the uniquely determined Lagrange multiplier vector λ of problem (5) at the point $(x^I_\alpha(y^0), y^0)$, where E denotes the unit matrix of order n and $L_I(x, y, \lambda) = f(x, y) + \sum_{i \in I} \lambda_i g_i(x, y)$.

Remark 3.1. Formula (6) for the computation of the Jacobian for the selection function remains valid for $\alpha = 0$ if the strong sufficient optimality condition of second order is satisfied for the original problem:
(SSOC) [17] For each $\lambda \in \Lambda(x_0(y^0), y^0)$, for each direction $d \neq 0$ satisfying

$$\nabla_x g_i(x_0(y^0), y^0) d = 0 \quad \text{for all } i \text{ with } \lambda_i > 0$$

we have

$$d^T \nabla^2_{xx} L(x_0(y^0), y^0, \lambda) d > 0.$$

To answer the second posed question is not possible without an additional assumption:
(NE) For each vertex $\lambda^0 \in \Lambda_\alpha(x_\alpha(y^0), y^0)$, the matrix

$$\begin{pmatrix} \nabla^2_{xx} L(x_\alpha(y^0), y^0, \lambda^0) + 2\alpha E & \nabla^T_x g_{J^0}(x_\alpha(y^0), y^0) & \nabla^2_{yx} L(x_\alpha(y^0), y^0, \lambda^0) \\ \nabla_x g_{I^0}(x_\alpha(y^0), y^0) & 0 & \nabla_y g_{I^0}(x_\alpha(y^0), y^0) \end{pmatrix}$$

has full row rank $n + |I(x_\alpha(y^0), y^0)|$, where the abbreviations $J^0 = \{j : \lambda_j^0 > 0\}$ and $I^0 = I(x_\alpha(y^0), y^0) = \{j : g_j(x_\alpha(y^0), y^0) = 0\}$ have been used.

Remark 3.2. This assumption is trivially satisfied for $\alpha > 0$ e.g. if the Linear Independence Constraint Qualification (LICQ) together with the Strict Complementarity Slackness Condition (SCS) hold since then $J^0 = I^0$ and the left part of the above matrix is regular [17]. Condition (NE) simply means that the decrease of the rank of this matrix for violated (LICQ) and/or (SCS) is compensated by the increase due to addition of the right part of this matrix. This assumption is not assumed to be satisfied for all Lagrange multipliers but only for the vertices of the set $\lambda^0 \in \Lambda_\alpha(x_\alpha(y^0), y^0)$.

Theorem 3.3. *Consider problem* (3) *with* $\alpha > 0$ *at a point* $y^0 \in Y$. *Let* $x_\alpha(y^0)$ *be the optimal solution and assume* (*C*), (*MF*), (*CR*), *and* (*NE*). *Then*

$$\partial_y x_\alpha(y^0) = \operatorname{conv}\left\{\nabla_y x_\alpha^I(y^0) : I \in \mathcal{I}\right\}.$$

The proof of this theorem uses the directional derivative of the function $x_\alpha(\cdot)$

$$x'_\alpha(y^0; r) = \lim_{t \searrow 0} t^{-1}[x_\alpha(y^0 + tr) - x_\alpha(y^0)]$$

with respect to y only. This directional derivative is equal to the unique solution of the following strictly convex quadratic optimization problem

$$\begin{aligned} 0.5d^T\left(\nabla^2_{xx} L(x_\alpha(y^0), y^0, \lambda^0) + 2\alpha E\right)d + r^T \nabla^2_{yx} L(x_\alpha(y^0), y^0, \lambda^0)d \to \min_d \\ \nabla_x g_i(x_\alpha(y^0), y^0)d + \nabla_y g_i(x_\alpha(y^0), y^0)r = 0, \quad i \in J^0 \\ \nabla_x g_i(x_\alpha(y^0), y^0)d + \nabla_y g_i(x_\alpha(y^0), y^0)r \le 0, \quad i \in I^0 \setminus J^0 \end{aligned} \tag{7}$$

for each $\lambda^0 \in \Lambda_\alpha(x_\alpha(y^0), y^0)$ solving the linear problem

$$\max_\lambda \{\nabla_y L(x_\alpha(y^0), y^0, \lambda)r : \lambda \in \Lambda_\alpha(x_\alpha(y^0), y^0)\} \tag{8}$$

[26]. Moreover, $\lambda^0 \in \Lambda_\alpha(x_\alpha(y^0), y^0)$ solves this linear problem for a fixed direction r if and only if there exists a solution d of the following system

$$\begin{aligned} &\nabla_x g_i(x_\alpha(y^0), y^0)d + \nabla_y g_i(x_\alpha(y^0), y^0)r = 0, \quad i \in J^0 \\ &\nabla_x g_i(x_\alpha(y^0), y^0)d + \nabla_y g_i(x_\alpha(y^0), y^0)r \le 0, \quad i \in I^0 \setminus J^0, \end{aligned}$$

i.e. if the quadratic problem has a feasible solution for the fixed values of r and λ^0 [3].

Proof of Theorem 3.3.: Take an arbitrary set $I \in \mathcal{I}$ and let without loss of generality

$$J^0 = \{1, \ldots, s\}, \quad I = \{1, \ldots, u\}, \quad I^0 = \{1, \ldots, v\}$$

for some $s \leq u \leq v$, where J^0 corresponds to the vertex $\lambda^0 \in \Lambda_\alpha(x_\alpha(y^0), y^0)$ chosen according to the first condition in the definition of $\mathcal{I}$. Denote the ith unit vector of $\mathbb{R}^n$ by e_i. Then, by (NE) the matrix M^1 defined by

$$\begin{pmatrix} \nabla^2_{xx} L(x_\alpha(y^0), y^0, \lambda^0) + 2\alpha E & \nabla_x^T g_{I^0}(x_\alpha(y^0), y^0) & \nabla^2_{yx} L(x_\alpha(y^0), y^0, \lambda^0) \\ \nabla_x g_{I^0}(x_\alpha(y^0), y^0) & 0 & \nabla_y g_{I^0}(x_\alpha(y^0), y^0) \\ 0 & e_{s+1}^T & 0 \\ \vdots & \vdots & \vdots \\ 0 & e_v^T & 0 \end{pmatrix}$$

has full row rank: if the $v - s$ columns $(\nabla_x^T g_i(x_\alpha(y^0), y^0), e_i)$ with $i \in I^0 \setminus J^0$ together with the last $v - s$ rows are deleted, a matrix arises which by (NE) has full row rank. The deleted rows and columns contain a unit matrix of order $v - s$ in the last $v - s$ rows.

Hence, the system of linear equations

$$M^1(d, \nu, r)^T = a$$

has a solution for each right-hand-side vector a. Let $\varepsilon > 0$ be arbitrary and take $a_i = 0,\ i = 1, \ldots, n+u, a_i = -\varepsilon,\ i = n+u+1, \ldots, n+v, a_i = \varepsilon,\ i = n+v+1, \ldots, n+v+u-s$, and $a_i = 0,\ i = n+v+u-s+1, \ldots, n+2v-s$ and let (d^0, ν^0, r^0) be a solution of the resulting system. Then, (d^0, ν^0) solves the Karush-Kuhn-Tucker conditions of the quadratic problem (7) for the fixed values of $\lambda^0,\ r^0$. Since the problem (7) is strictly convex in d for each fixed r^0, λ^0, it has a unique optimal solution whenever it has one and the Karush-Kuhn-Tucker conditions are necessary and sufficient for optimality. Hence, $d^0 = x'_\alpha(y^0; r^0)$. Moreover, using Taylor's expansion up to the first order for the function $t \mapsto g_i(x_\alpha(y^0 + tr^0), y^0 + tr^0)$ at $t = 0$ with respect to $t \geq 0$ we get

$$g_i(x_\alpha(y^0 + tr^0), y^0 + tr^0) = g_i(x_\alpha(y^0), y^0) - t\varepsilon + o(t) < 0$$

for $i \in I^0 \setminus I = \{u + 1, \ldots, v\}$ and sufficiently small $t > 0$, where $\lim_{t \searrow 0} o(t)/t = 0$. Hence, the constraints $g_i(x_\alpha(y), y)$ are inactive for $i \in I^0 \setminus I, y = y^0 + tr^0$, and sufficiently small $t > 0$. Thus, they can be dropped.

By the second condition for $\mathcal{I}$ the linear independence constraint qualification is satisfied for the problem (5) which together with (C) and [8] implies that this problem has a continuously differentiable (local) optimal solution $x^I_\alpha(\cdot)$ with $x^I_\alpha(y^0) = x_\alpha(y^0)$ in a certain neighborhood of $(x_\alpha(y^0), y^0)$. The corresponding Langrange multiplier vector $\lambda^I_\alpha(\cdot)$ is also continuously differentiable at y^0. By exploiting the system of equations resulting from $M^1(d, \nu, r) = a$ after deleting all equations related to the index set $I^0 \setminus I$ (i.e. the lines $n + u + 1, \ldots, n + v$ and $n + v + u - s + 1, \ldots, n + 2v - s$) it is easy to see, that

$$d^0 = x'_\alpha(y^0; r^0) = \nabla_y x^I_\alpha(y^0) r^0, \qquad \nu^0 = \nabla_y \lambda^I_\alpha(y^0) r^0.$$

By Taylor's expansion up to the first order for the function $t \mapsto \lambda^I_{\alpha,i}(y^0 + tr^0)$ at $t = 0$

we get

$$\lambda^I_{\alpha,i}(y^0 + tr^0) = \lambda^I_{\alpha,i}(y^0) + t\varepsilon + o(t) > 0$$

for $i \in I \setminus J^0 = \{n + v + 1, \ldots, n + v + u - s\}$ and sufficiently small $t > 0$. Using that the Karush-Kuhn-Tucker conditions are necessary and sufficient optimality conditions for problem (3) under our assumptions, this implies that $x_\alpha(y^0 + tr^0) = x^I_\alpha(y^0 + tr^0)$ for sufficiently small $t > 0$ and

$$\{j : g_j(x_\alpha(y^0 + tr^0), y^0 + tr^0) = 0\} = I.$$

Thus, $\nabla x_\alpha(y^0+tr^0) = \nabla x^I_\alpha(y^0+tr^0)$ for sufficiently small $t > 0$. Hence, by Rademacher's Theorem [2], $\nabla_y x^I_\alpha(y^0) \in \partial_y x_\alpha(y^0)$.

By Theorem 2.4,

$$x_\alpha(y) \in \left\{x^I_\alpha(y) : I \in \mathcal{I}\right\}$$

for all y in a sufficiently small open neighborhood of y^0. Hence, by Theorem 2.5

$$\partial_y x_\alpha(y^0) \subseteq \operatorname{conv}\left\{\nabla_y x^I_\alpha(y^0) : I \in \mathcal{I}\right\}$$

and we are done. □

Remark 3.4. If the computation of only one generalized Jacobian of the function $x_\alpha(y)$ at $y = y^0$ is needed, assumption (NE) can be weakened [4]. If the assumption (SSOC) is satisfied at y^0, then this theorem remains true also for $\alpha = 0$.

Assumption (NE) is needed for proving the computability of generalized gradients of the function $x_\alpha(\cdot)$ not for the existence of generalized gradients. In [23] it is shown that for each set $I \in \mathcal{I}$ a generalized gradient of the function $x_\alpha(\cdot)$ can be computed provided that the problem of computing the projection on a parameter independent set is considered and (LICQ) is satisfied. For general parametric optimization problems satisfying (SSOC) and (LICQ) this is not true. Hence, we have to add a certain condition. In [7] the strong complementarity assumption with respect to (7) has been added (it is possible to replace (LICQ) by (MFCQ) [4]). In [24] the difficulties of the computation of generalized gradients of the $x_\alpha(\cdot)$ are overcome by using the assumptions of [23], and in [25] the condition (LICQ) is used together with a condition which is dual to (NE).

The following example shows that Theorem 3.3 is in general not valid if assumption (NE) is dropped. This example is concerned with problem (2) (i.e. with $\alpha = 0$) but examples showing the same effect for problem (3) can easily be found.

Example 3.5. Consider the simple problem

$$\min_x\{(x_1 - y)^2 + (x_2 - 1)^2 : x_1 + x_2 \le 1,\ -x_1 + x_2 \le 1\}$$

at the point $y^0 = 0$. Then, $x(y^0) = (0, 1)^T$ is the unique optimal solution and the assumptions (C), (MF), (SSOC), and (CR) are satisfied there. (NE) is not satisfied. It is easy to see that

$$x(y) \in \{(y/2, 1 - y/2)^T, (y/2, 1 + y/2)^T\}$$

for y near y^0, and $\Lambda(x(y^0), y^0) = \{(0, 0)^T\}$. The sets $I \in \{\emptyset, \{1\}, \{2\}, \{1, 2\}\}$ are to be considered in order to compute the generalized Jacobian of the function $x(\cdot)$.

For $I = \{1\}$ we compute $\nabla x^I(y^0) = (0.5, -0.5)^T$. For $I = \{2\}$ we get $\nabla x^I(0) = (0.5, 0.5)^T$. Hence, $\nabla x^I(0) \in \partial_y x(0)$ for both sets $I = \{1\}$ and $I = \{2\}$. But, for $I = \{1, 2\}$ we compute $x^I(0) = (0, 0)^T \notin \partial_y x(0)$. Last but not least we see that $\nabla x^I(0) = (1, 0)^T \notin \partial_y x(0)$ for $I = \emptyset$. □

Remark 3.6. If the assumptions (C), (MF), and (CR) are satisfied and $\alpha > 0$ is fixed, the function $x_\alpha(\cdot)$ is weakly semismooth in the sense that $x'_\alpha(y^0; r) = \lim_{t \searrow 0} w(t)r$, where $w(t) \in \partial_y x_\alpha(y^0 + tr)$ for $t > 0$ [4]. This property remains true for $\alpha = 0$ if the assumption (SSOC) is also satisfied.

4. A bundle algorithm

Now, we are prepared to describe the application of the bundle idea to the bilevel problem (4). The idea is to combine iterations of a bundle algorithm with decreasing values for α. The simplest way to do that is to use the following

Prototype bundle algorithm:
Step 1: Select $\varepsilon^0 > 0$, $\alpha^0 > 0$ and a starting point z^0. Set $s := 0$.
Step 2: Starting with z^s, compute by use of a bundle algorithm an ε^s-optimal solution z^{s+1} of the problem (4) for that fixed value of $\alpha = \alpha^s$.
Step 3: Set $\varepsilon^{s+1} := \varepsilon^s/2$, $\alpha^{s+1} := \alpha^s/2$, $s := s+1$ and repeat Step 2 until some termination criterion is satisfied.

At least if the value of α^s is not too small, Step 2 can be done in a finite number of iterations. This is also guaranteed for each iteration if the sequence $\{z^s\}_{s \in N}$ converges to some point $\bar{z}$ where the unique lower level optimal solution $x_0(\bar{z})$ is strongly stable. In this case, by upper semicontinuity of the ε-subdifferential of a locally Lipschitzian function, we easily derive that the limit point $(x_0(\bar{z}), \bar{z})$ of the iteration process is a stationary solution of the bilevel problem.

In what follows we will do two things. First a certain diagonalization process is initiated which means that we will not compute an ε^s-optimal solution of the regularized problem (4) before updating α^s but decrease α^s whenever a (short) serious step has been made. Hopefully this results in a smaller number of the (expensive) evaluations of the objective function value $G_{\alpha^s}(z^s)$. And, additionally, we intend to avoid difficulties in the calculation process which could arise if the value of α^s is decreased too rapidly at points which are far from the region where the problem behaves sufficiently well to guarantee that all calculations in the

algorithm can really be made, i.e. at points, where the Lipschitz constant of the function $x_\cdot(\cdot)$ is too large. Second, a special (but possible one within the settings of the bundle algorithm in [16]) rule for updating the trust region radius is used in order to catch the sequence of trial and iteration points generated by the algorithm in some neighborhood of the stationary points of the bilevel problem.

The above prototype algorithm consists of a sequence of applications of a bundle algorithm to different problems, with other words, the bundle algorithm is restarted infinitely often. When using the algorithm below, restarts will also be used if the result of one application of the modified algorithm is not satisfactory, yet (i.e. if for instance the value of α^s is too large). In the worst theoretical case, this can result in an infinite sequence of restarts of the algorithm.

We do not intend to describe the bundle algorithm in details, this has been done e.g. in [16, 28, 29] but rather to give only the necessary material for understanding our modification. For the sake of simplicity the algorithm is given for $Y = \mathbb{R}^m$. The case when Y is a polyhedron can be treated in an analogous way [29]. For ideas of how to include nonlinear constraints, see [14].

Let $\alpha > 0$ be fixed for the moment. Subsequently we intend to decrease the value of α in order to get step by step a better approximation of the original problem. Set

$$G_\alpha(y) = F(x_\alpha(y), y)$$

and let $v(y)$ denote an arbitrary generalized gradient of this function:

$$v(y) \in \{\nabla_x F(x_\alpha(y), y)d + \nabla_y F(x_\alpha(y), y) : d \in \partial_y x_\alpha(y)\}$$

(cf. the chain rule for the subdifferential [2]).

In bundle algorithms, a model of the function to be minimized is used to compute a direction of descent. Let two sequences of points, the trial points $\{y^k\}_{k=1}^s$, and the iteration points $\{z^k\}_{k=1}^s$ have already been computed. Then, for minimizing the nonconvex function $G_\alpha(y)$, this model has the form

$$\max_{k\in J_s}\{v(y^k)^T d - \beta_{k,s}\} + G_\alpha(z^s) + u^s d^T d/2, \tag{9}$$

where $J_s \subseteq \{1, \ldots, s\}$, $u^s > 0$ is a weight factor and

$$\beta_{k,s} = |G_\alpha(z^s) - v(y^k)^T(z^s - y^k) - G_\alpha(y^k)|.$$

In what follows we will adopt Algorithm 2.1 in [16]. It consists of a sequence of iterations, where directions of descent, step-sizes as well as new trial points are computed, and parameters are updated. We will not give the algorithm in details but only outline it in short. The interested reader is referred to the original paper [16]. Denote by $\varepsilon > 0$ a final optimality tolerance, by $\varrho^s > 0$ a trust region radius, and by m_L a positive line search parameter. Let $u^s \geq u_{\min} > 0$ be some weight and $\alpha^s > 0$ be some regularization coefficients. Let d^s be

an optimal solution of minimizing the function (9). Then, as in [16] the value

$$\varepsilon^s := \max\left\{ m_p\left(\| u^s d^s \| + \sum_{j\in J_s} \eta_j^s \beta(j,s)\right), \max_{\substack{j\in J_s \\ \eta_j^s\neq 0}} \left\{ \| y^j - z^j \| + \sum_{k=j}^{s-1} \| z^{k+1} - z^k \| \right\} \right\},$$

where η_j^s are the Lagrange multipliers of the minimization problem (9) and $J_s \supset \{j : \eta_j^s \neq 0\}$ contains the present index set of used iteration points, is used for measuring the quality of the present approximation z^s of a solution for the problem (4). Hence, during the procedure, ε^s should tend to zero. For practical reasons yet, the iteration process is terminated if $\varepsilon^s \leq \varepsilon$ for some positive ε. Let $\kappa \geq 1$ be some positive constant used for bounding the change of the trust region radius and δ^1 be some initial stationarity target. Let $\omega \in (0,1)$.

In the following algorithm all not explicitly described changes of the parameter values are as in the original Algorithm 2.1 in [16]. Summing up we have made only two small changes in the original algorithm: First, we update the value of α every time a (short) serious step has been made, and second we bound the trust region radius using one of the possible rules in the Algorithm 2.1 in [16]. The principal steps of the algorithm are as follows:

Bundle algorithm:

Step 1: Compute, if necessary by increasing the value of u^s, an optimal solution d^s of minimizing the model function:

$$\max_{k\in J_s}\{v(y^k)^T d - \beta_{k,s}\} + u^s d^T d/2 \tag{10}$$

together with some corresponding Lagrange multipliers $\eta_j^s,\ j \in J_s$, such that $\| d^s \| \leq \varrho^s$. Test the stopping criterion $\varepsilon^s \leq \varepsilon$ and diminish ε^s as well as J_s if necessary. If $\varepsilon^s \leq \delta^s$ then $\delta^s := \omega\delta^s$.

Step 2: Test, if a serious step can be made: If

$$G_{\alpha^s}(z^s + d^s) \leq G_{\alpha^s}(z^s) + m_L \max_{k\in J_s}\{v(y^k)^T d^s - \beta_{k,s}\}$$

then set $z^{s+1} = z^s + d^s$, $t_s = 1$. Else make a line search (by Algorithm 2.2 in [16]) to find either some $t_s > 0$ such that a short serious step is possible which leads to the new solution $z^{s+1} = z^s + t_s d^s$ satisfying

$$G_{\alpha^s}(z^s + t_s d^s) \leq G_{\alpha^s}(z^s) + m_L t_s \max_{k\in J_s}\{v(y^k)^T d^s - \beta_{k,s}\}$$

or make a null step which means that a new trial point $y^{s+1} = z^s + td^s$ is computed for some $t > 0$ and $t_s = 0$. Now, if $t_s > 0$, compute a new regularization coefficient $\alpha^{s+1} \in (0, \alpha^s]$ such that

$$G_{\alpha^{s+1}}(z^{s+1}) - G_{\alpha^s}(z^s) \leq 0.5(G_{\alpha^s}(z^{s+1}) - G_{\alpha^s}(z^s))$$

else $\alpha^{s+1} = \alpha^s$. Set $y^{s+1} = z^{s+1}$ if $t_s > 0$.

Step 3: Select a new trust region radius $\varrho^s \leq \kappa \cdot \delta^s$ maintaining the rules in [16], update all other parameters and u^s of the algorithm as well as the set J_s, compute a new generalized gradient of the function $G_{\alpha^{s+1}}(y)$ at the point $y = y^{s+1}$, set $\delta^{s+1} := \delta^s$, $s := s + 1$ and repeat Step 1. □

Some words with respect to the updating of the regularization parameter seem to be necessary. Our principal aim is to minimize α^s to get a closer approximation of the original problem. A decrease in its value is possible if we have a positive step length in Step 2. Then,

$$G_{\alpha^s}(z^s + t_s d^s) \leq G_{\alpha^s}(z^s) + m_L t_s \max_{k \in J_s}\{v(y^k)^T d^s - \beta_{k,s}\}.$$

Hence, since $G.(\cdot)$ is locally Lipschitz continuous there is some interval $[\alpha^0, \alpha^s] \subseteq [0, \alpha^s]$ with $\alpha^0 < \alpha^s$ such that the inequality

$$G_\alpha(z^s + t_s d^s) \leq G_{\alpha^s}(z^s) + m_L t_s \max_{k \in J_s}\{v(y^k)^T d^s - \beta_{k,s}\}/2 < G_{\alpha^s}(z^s)$$

is satisfied for each $\alpha \in [\alpha^0, \alpha^s]$. To find α^{s+1}, in the worst case, we can use dichotomy. It is also possible to let α^s be fixed during a (small) number of iterations if the algorithm seems to run into difficulties due to a too large Lipschitz constant of the objective function.

For the theoretical investigations take the optimality tolerance $\varepsilon = 0$. Then, in the above bundle algorithm, a sequence $\{(\alpha^s, x_{\alpha^s}(z^s))\}_{s \in N}$ is computed (which in general is infinite, but can also be finite due to finite termination or due to numerical reasons). If the sequence $\{\alpha^s\}_{s \in N}$ converges to zero, the problem is successfully solved as will be seen below. In the other case, the sequence $\{\alpha^s\}_{s \in N}$ will converge to some $\bar{\alpha}^m > 0$ and the sequence $\{z^s\}_{s \in N}$ converges to $\bar{z}^m$ with $0 \in \partial G_{\bar{\alpha}^m}(\bar{y}^m)$. Then (or if the last regularization parameter α^s is too large in practice) we use a restart of the bundle algorithm to produce a new sequence of points which reflects the original problem in a better way.

When the bundle algorithm is restarted we can use the limit point $\bar{y}^m$ (or the last point z^s of the computed sequence if finite termination occurs) together with a new smaller $\alpha^1 \in (0, \bar{\alpha}^m)$ (resp. $\alpha^1 \in (0, \alpha^s)$) as starting point.

The above bundle algorithm uses a modification in the control of the trust region radius by demanding that $\varrho^s \leq \kappa \cdot \delta^s$ with $\kappa \geq 1$ which is not used in [16]. This rule is compatible with all rules in [16] and applies only to parts (ii) to (iv) of Step 10 of the Algorithm 2.1 in [16] and is larger than the lower bounds for the update of ϱ used there. It implies that this radius will be minimized during the bundle algorithm (since convergence of δ^s to zero is shown in [16] provided that the sequence $\{z^s\}_{s=1}^{\infty}$ remains bounded). This rule has been added to guarantee that all iteration and trial points generated during the computations remain in some set of interest.

No modification of the update of other parameters of the algorithm is made.

5. Convergence of the modified bundle algorithm

Throughout this section let the assumptions (C), (MF), (CR), and (NE) be satisfied for the lower level problem and let $Y = \mathbb{R}^m$. Also assume that the sequence $\{z^s\}_{s=1}^{\infty}$ computed

by the bundle algorithm remains bounded. If the tolerance $\varepsilon = 0$ is chosen, then finite termination means that

$$0 \in \{\nabla_x F(x_{\alpha^s}(z^s), z^s)d + \nabla_y F(x_{\alpha^s}(z^s), z^s) : d \in \partial_y x_{\alpha^s}(z^s)\}$$

[16]. In this case, a restart of the algorithm is done with a new, smaller value of $\alpha < \alpha^s$. Hence, let the bundle algorithm compute an infinite sequence. Let v^s denote

$$v^s = \max_{k \in J_s}\{v(y^k)^T d^s - \beta_{k,s}\}.$$

Note that due to $\beta_{k,s} \geq 0$ and $u^s > 0$, $v^s \leq v^s + u^s(d^s)^T d^s/2 \leq -\beta_{k,s} \leq 0$ where the second inequality follows since $d = 0$ is feasible for the problem of minimizing the function (10). Then, as in [15], we get

Lemma 5.1. *If the function $F(x, y)$ is bounded from below on the set $\{(x, y) : y \in \mathbb{R}^m, g(x, y) \leq 0\}$, then $\sum_{s=1}^{\infty} t_s \mid v^s \mid < \infty$.*

Proof: First note that $t_s > 0$ iff a (short) serious step is made. Then, setting $G_\alpha(z) = F(x_\alpha(z), z)$,

$$\begin{aligned} -\infty &< \sum_{i=1}^{\infty}(G_{\alpha^{s+1}}(z^{s+1}) - G_{\alpha^s}(z^s)) \\ &\leq 0.5\sum_{i=1}^{\infty}(G_{\alpha^s}(z^{s+1}) - G_{\alpha^s}(z^s)) \leq 0.5m_L \sum_{i=1}^{\infty} t_s v^s < 0. \end{aligned}$$ □

Recall that the function $G.(\cdot)$ is locally Lipschitz continuous on $[\alpha^0, \infty) \times \mathbb{R}^m$ for each fixed $\alpha^0 > 0$. Consequently, the update of α in Step 2 of the bundle algorithm is possible.

Theorem 5.2. *Let $\{(x_{\alpha^s}(z^s), z^s, \alpha^s)\}_{s=1}^{\infty}$ be the sequence computed by the modified bundle algorithm with $\varepsilon = 0$. Let the sequence $\{\alpha^s\}_{s=1}^{\infty}$ be bounded from below by some $\alpha^0 > 0$. Then, every accumulation point $(x_{\bar\alpha}(\bar y), \bar y, \bar\alpha)$ of this sequence satisfies*

$$0 \in \{\nabla_x F(x_{\bar\alpha}(\bar y), \bar y)d + \nabla_y F(x_{\bar\alpha}(\bar y), \bar y) : d \in \partial_y x_{\bar\alpha}(\bar y)\}.$$

Proof: With respect to the Algorithm 2.1 in [16] we have changed only the evaluation of the objective function values in Steps 7 and 8 while maintaining the condition

$$G_{\alpha^{s+1}}(z^{s+1}) \leq G_{\alpha^s}(z^s) + 0.5m_L t_s \max_{k \in J_s}\{v(y^k)^T d^s - \beta_{k,s}\}$$

and the update of ϱ^s in Step 10 by introducing an additional upper bound, which can be satisfied in any of the cases in Algorithm 2.1 in [16].

First we have to show an analogous result to Lemma 3.1 in [16]: Let $\bar{y}$ be some accumulation point of the sequence $\{z^s\}_{s\in N}$ computed by the algorithm. Similarly to formula (2.12) in [16] we obtain that the optimal solution d^s of the minimization of the function (10) satisfies

$$d^s \in \partial\left(G_{\alpha^s}\left(z^s;\ \max_{\substack{j\in J_s\\ \eta_j^s\neq 0}}\left\{\| y^j - z^j \| + \sum_{k=j}^{s-1} \| z^{k+1} - z^k \|\right\}\right)\right),$$

where $\partial G_\alpha(z;\varepsilon)$ denotes the Goldstein ε-subdifferential

$$\partial G_\alpha(z;\varepsilon) = \operatorname{conv}\{\partial G_\alpha(y) : \|z-y\| \le \varepsilon\}$$

which is upper semicontinuous and locally bounded [14]. Thus, if $(\bar{\alpha}, \bar{y}, 0, 0)$ is an accumulation point of the sequence

$$\left\{\left(\alpha^s, z^s, d^s, \max_{\substack{j\in J_s\\ \eta_j^s\neq 0}}\left\{\| y^j - z^j \| + \sum_{k=j}^{s-1} \| z^{k+1} - z^k \|\right\}\right)\right\}_{s=1}^{\infty},$$

then $0 \in \partial G_{\bar{\alpha}}(\bar{y})$.

Second, since the value of α^s is changed only if a (short) serious step is made and since our modification coincides with the original algorithm in treating null steps, after a finite number of null steps a (short) serious step is made (cf. Lemmas 3.5 and 3.6 in [16]). Note, that our changed rule for updating the value of ϱ^s leads only to bounding the increase of this value but never the decrease. Hence, if the decrease of ϱ^s is need to be bounded this remains true also in the modified algorithm.

Since in the proof of Lemma 3.7 in [16] the function value $G_{\alpha^s}(z^s)$ is only used to show that $t_s|v^s|$ converges to zero, (which results also from Theorem 5.1) this lemma remains valid. The same is true for Lemma 3.8 of [16] which uses only the updating rules in the algorithm.

Hence, also Theorem 3.9 of [16] remains valid which gives the desired result. □

If the sequence $\{\alpha^s\}_{s=1}^{\infty}$ converges to zero then the strong sufficient optimality condition of second order at the accumulation points of the sequence $\{(x_{\alpha^s}(z^s), z^s)\}_{s=1}^{\infty}$ is necessary in order to guarantee that the bundle algorithm is able to compute all the data necessary. If this optimality condition is satisfied then the above theorem shows that the accumulation points are stationary. Note that the sequence $\{z^s\}_{s=1}^{\infty}$ converges due to the modified rule for the control of the trust region parameter provided that it remains bounded. Hence, we get the

Corollary 5.3. *Let the sequence $\{(x_{\alpha^s}(z^s), z^s, \alpha^s)\}_{s=1}^{\infty}$ be computed by the bundle algorithm, where $\{\alpha^s\}_{s=1}^{\infty}$ converges to zero. Let $\bar{y}$ be the limit point of the sequence $\{z^s\}_{s=1}^{\infty}$ and*

let $x_0(\bar{y}) \in \Psi(\bar{y})$. *If the additional assumption* (*SSOC*) *is satisfied at* $(x_0(\bar{y}), \bar{y})$ *then*

$$0 \in \{\nabla_x F(x_0(\bar{y}), \bar{y})d + \nabla_y F(x_0(\bar{y}), \bar{y}) : d \in \partial_y x_0(\bar{y})\}.$$

In this corollary, the restrictive assumption (SSOC) has been used. Together with the blanket assumptions (MF) and (CR) this assumption guarantees that the solution function $x(\cdot)$ of the original lower level problem (2) is locally Lipschitz continuous at $\bar{y}$. This implies also that α^s can tend to zero since we could minimize the function $G_0(\cdot)$ itself by use of the bundle algorithm. In the other case, if the function $G_0(\cdot)$ is not locally Lipschitz continuous, numerical difficulties make the decrease of α^s to zero impossible. In the following we intend to investigate the possibilities to get $\alpha^s \to 0$ during the bundle algorithm.

To reach this aim we need one more assumption. Recall that $x_0(y) \in \Psi(y)$ is an arbitrary optimal solution of the problem (2). Consider the set

$$\hat{Y} = \{y \in \mathbb{R}^m : \text{(SSOC) is satisfied at } (x_0(y), y)\}.$$

In general, the set $\hat{Y}$ is neither open nor closed nor connected, but $x_0(\cdot)$ is locally Lipschitz continuous on $\hat{Y}$. The following considerations are only useful if the set $\hat{Y}$ has a suitable structure which is the main assumption in what follows. Let $\theta > 0$ be a (small) constant and consider a set D such that $D + 2\theta\mathcal{B} := \{y : \inf_{w\in D} \|y - w\| < 2\theta\} \subseteq \hat{Y}$. Here, $\mathcal{B}$ denotes the unit ball in $\mathbb{R}^m$. Subsequently, we assume that $D \neq \emptyset$ exists. This is a weaker assumption than supposing (SSOC) throughout $\mathbb{R}^m$.

Note, that if the assumption (SSOC) is not satisfied at the limit point $(x_0(y^0), y^0)$ of the iteration sequence, then the function $x_0(\cdot)$ is generally not locally Lipschitz continuous at the point y^0. This in turn means that we intend to minimize a non-Lipschitzian function by a bundle algorithm. This probably will cause difficulties in the procedure which will result in a termination without computing a solution. Then, the best solution found so far will be an (presumably good) approximation of an optimal solution.

Lemma 5.4. *Let* $B \subseteq \mathbb{R}^m$ *be a bounded set and* $0 < \hat{\alpha} < \tilde{\alpha}$ *positive numbers. Let* $(z^s, \alpha^s) \in ((D + \theta\mathcal{B}) \cap B) \times [0, \tilde{\alpha}] \cup B \times [\hat{\alpha}, A)$, *where* A *is a sufficiently large constant. Assume that* $\|d^s\| \leq \theta/11$ *whenever* $\alpha^s \leq \tilde{\alpha}$. *Let* $u^s \geq u_{\min}$ *for all* s. *Then there exists* C (*depending only on* $B, \hat{\alpha}, \tilde{\alpha}$) *such that*

1. $\|d^s\| \leq C/\sqrt{u^s}$,
2. $\|v(z^s + td^s)\| \leq C$,
3. $|G_{\alpha^s}(z^s + td^s) - G_{\alpha^s}(z^s)| \leq Ct\|d^s\|$,
4. $|G_{\alpha^s}(z^s) - G_{\alpha^s}(z^s + td^s) - t\langle v(z^s + td^s), d^s\rangle| \leq Ct\|d^s\|$

for $0 \leq t \leq 10$.

Note, that in the proofs of Lemmas 3.5–3.8 in [16] convergence of $G_{\alpha^s}(z^s) - G_{\alpha^s}(z^s + td^s) - t\langle v(z^s + td^s), d^s\rangle$ to zero is used to investigate the null steps and to give the contradiction in Lemma 3.7(ii). Also, boundedness of $\|v(z^s + td^s)\|$ is needed only for unchanged α^s. The bound for the norm on d is taken such that, for small α, the new point $z^s + td^s$ belongs to $D + \theta\mathcal{B}$.

Proof of Lemma 5.4.: The first assertion is proved as the first one of Lemma 3.3 in [15]. The second one follows since $(z^{s+1}, \alpha^s) \in ((D+\theta\mathcal{B})\cap B+10/11\theta\mathcal{B})\times[0,\tilde{\alpha}]\cup B\times[\hat{\alpha}, A)$ and the function $G.(\cdot)$ is locally Lipschitz continuous there. C is the common Lipschitz constant of G on the set $((D+\theta\mathcal{B})\cap B+10/11\theta\mathcal{B})\times[0,\tilde{\alpha}]\cup B\times[\hat{\alpha}, A)$. The third assertion is due to Lipschitz continuity. For proving the last one first use the triangle inequality followed by application of the second assertion of this lemma. □

Note that the results in [16] prove that $\{\delta^s\}_{s=1}^{\infty}$ converges to zero. By the rules of the modified bundle algorithm this implies that the sequence $\{z^s\}_{s=1}^{\infty}$ is a bounded Cauchy sequence and, hence, itself convergent.

As already mentioned above, the control of the regularization parameter α^s is not clearly given. Due to the unknown behavior (besides Lipschitz continuity) of the function $G.(y)$ for fixed y it seems to be impossible to give a valid rule. It is especially possible that $G_\alpha(z^{s+1}) > G_{\alpha^s}(z^s)$ for $0 < \alpha \ll \alpha^s$, where $z^s,\ z^{s+1}$ have been computed as given in the bundle algorithm. This means that a sequence $\{\alpha^s\}_{s=1}^{\infty}$ computed by the rules of the bundle algorithm can converge to a positive limit $\hat{\alpha}$. In that case, a restart of the bundle algorithm is done with a smaller value for $0 < \alpha^1 < \hat{\alpha}$ (e.g. $\alpha^1 \le \max\{0, \hat{\alpha} - \xi\}$ for some constant $\xi > 0$). This results in a (finite or infinite) sequence of calls of the above bundle algorithm.

Clearly, in practice, the bundle algorithm itself uses only a finite number of iterates to compute an approximate solution and only a finite number of restarts can be used. For theoretical investigations let the bundle algorithm compute after each restart an infinite sequence of iterates or stop due to satisfaction of a stationarity condition after a finite number of iterations.

In the following theorem the resulting algorithm is investigated and we consider two cases: First the case when the bundle algorithm is restarted finitely many times. Then, the overall algorithm can be considered as being equal to one run of the bundle algorithm (namely the last one). Second the case when the bundle algorithm is restarted infinitely often. Again, for the investigation of the convergence of the algorithm, we use $\varepsilon = 0$ in the stopping criterion. Then, due to the assumed boundedness of the sequence $\{z^s\}_{s=1}^{\infty}$ and convergence of $\{\varrho^s\}_{s=1}^{\infty}$ to zero, the sequence $\{z^s\}_{s=1}^{\infty}$ computed during the m-th run of the above bundle algorithm converges to some $\bar{y}^m$. Hence, also $\{(x_{\alpha^s}(z^s), z^s, \alpha^s)\}_{s=1}^{\infty}$ converges to $(x_{\tilde{\alpha}}(\bar{y}^m), \bar{y}^m, \bar{\alpha}^m)$. The sequence of all iteration points computed during the infinite number of applications of the modified bundle algorithm is then being considered in the second case. One accumulation point of this sequence is also an accumulation point of the sequence $\{(x_{\tilde{\alpha}}(\bar{y}^m), \bar{y}^m, \bar{\alpha}^m)\}_{m\in N}$.

Now we are able to state the main convergence theorem for the modified bundle algorithm:

Theorem 5.5. *Consider the regularized bilevel problem* (4) *and let the assumptions* (*C*), (*MF*), (*CR*) *and* (*NE*) *be satisfied for all* $y \in \mathbb{R}^m, \alpha > 0$. *Let there exist* $\alpha^* > 0$, $\hat{\varepsilon} > 0$, *and* $\bar{F}$ *such that for all* $0 < \alpha < \alpha^*$ *all stationary points* $\tilde{y}_\alpha$ *of the functions* $F(x_\alpha(y), y)$ *have* $F(x_\alpha(\tilde{y}_\alpha), \tilde{y}_\alpha) \le \bar{F}$ *and*

$$\{y : F(x_\alpha(y), y) \le \bar{F} + \hat{\varepsilon}\} \subseteq D \quad \forall\, 0 < \alpha < \alpha^*,$$

where the set D is defined as above stated and bounded. Then we have the following:

1. *if only a finite number of restarts is needed to get convergence of* $\{\alpha^s\}_{s=1}^{\infty}$ *to zero, then the limit point* $(x_0(\bar{y}), \bar{y})$ *of the sequence computed by the algorithm satisfies*

$$0 \in \{\nabla_x F(x_0(\bar{y}), \bar{y})d + \nabla_y F(x_0(\bar{y}), \bar{y}) : d \in \partial_y x_0(\bar{y})\}.$$

2. *if an infinite number of restarts is necessary then there exist accumulation points* $(x_{\bar{\alpha}}(\bar{y}), \bar{y}, \bar{\alpha})$ *of the sequence computed by the algorithm with* $\bar{\alpha} = 0$ *satisfying*

$$0 \in \{\nabla_x F(x_{\bar{\alpha}}(\bar{y}), \bar{y})d + \nabla_y F(x_{\bar{\alpha}}(\bar{y}), \bar{y}) : d \in \partial_y x_{\bar{\alpha}}(\bar{y})\}.$$

Moreover, the sequence computed by the algorithm has at least one accumulation point $(x_{\bar{\alpha}}(\bar{y}), \bar{y}, \bar{\alpha})$ *with* $\bar{\alpha} = 0$.

Proof: By use of certain restarts and after dropping the first iterations if necessary we can assume that $\alpha^1 \leq \alpha^*$. By use of slow changes of α^s (or even by use of constant values of α^s in the first iterations) we can assume that $\alpha^s \leq \alpha^*$ and $F(x_{\alpha^s}(z^s), z^s) \leq \bar{F} + \hat{\varepsilon}$ for some iteration index s, since the sequence $\{z^s\}_{s=1}^{\infty}$ converges to some stationary point [16] and all stationary points y_{α^s} satisfy $F(x_{\alpha^s}(y_{\alpha^s}), y_{\alpha^s}) \leq \bar{F}$. But then, by the rules of the algorithm, both inequalities are satisfied in all later iterations, too. Hence we can assume without loss of generality that $z^s \in D$ for all s.

First assume that we are not forced to use restarts to guarantee that $\lim_{s\to\infty} \alpha^s = 0$.Then, this implies that the sequence $\{z^s\}_{s=1}^{\infty}$ converges to some $\bar{y} \in \text{cl } D$. Indeed, by the rules of the modified bundle algorithm, $\| z^{s+1} - z^s \| \leq c \| d^s \| \leq c \cdot \kappa \cdot \delta^s$ and δ^s converges to zero by the proof of Theorem 3.9 in [16]. Since (SSOC) is satisfied at $\bar{y} \in \text{cl } D$ by the assumptions of the theorem, the Theorem 3.9 in [16] can be used to show that

$$0 \in \{\nabla_x F(x_0(\bar{y}), \bar{y})d + \nabla_y F(x_0(\bar{y}), \bar{y}) : d \in \partial_y x_0(\bar{y})\}.$$

If a finite number $\bar{m}$ of restarts is used then the assertion of the theorem follows since convergence of $\{\alpha^s\}_{s=1}^{\infty}$ to zero means that only a finite number of iteration points of the first $\bar{m} - 1$ applications of the bundle algorithm can join each convergent subsequence of the iteration points $\{(x_{\alpha^s}(z^s), z^s, \alpha^s)\}_{s=1}^{\infty}$ with $\lim_{s\to\infty} \alpha^s = 0$. Hence, we can drop the first $\bar{m} - 1$ applications of the bundle algorithm and consider the sequence $\{(x_{\alpha^s}(z^s), z^s, \alpha^s)\}_{s=1}^{\infty}$ as being computed in the last one only. Then, the assertion follows as above.

Now, if an infinite number of restarts is necessary we can select the sequence $\{(x_{\bar{\alpha}^m}(\bar{y}^m), \bar{y}^m)\}_{m=1}^{\infty}$ of limit points in the applications of the bundle algorithm (i.e. the mth application produces a sequence $\{(x_{\alpha^s}(z^s), z^s, \alpha^s)\}_{s=1}^{\infty}$ converging to $(x_{\bar{\alpha}^m}(\bar{y}^m), \bar{y}^m, \bar{\alpha}^m)$, $m = 1, 2, \ldots$). Then, by [16]

$$0 \in \{\nabla_x F(x_{\bar{\alpha}^m}(\bar{y}^m), \bar{y}^m)d + \nabla_y F(x_{\bar{\alpha}^m}(\bar{y}^m), \bar{y}^m) : d \in \partial_y x_{\bar{\alpha}^m}(\bar{y}^m)\}.$$

The assertion now follows by upper semicontinuity of the point-to-set mapping $\partial_y x.(\cdot)$ and convergence of a subsequence of $\{(x_{\bar{\alpha}^m}(\bar{y}^m), \bar{y}^m, \bar{\alpha}^m)\}_{m=1}^{\infty}$.

The proof is completed after showing that the algorithm can indeed compute some sequence $\{(x_{\alpha^s}(z^s), z^s, \alpha^s)\}_{s=1}^{\infty}$ such that $\lim_{s\to\infty} \alpha^s = 0$ using a certain number of restarts.

This means that first, during one application of the bundle algorithm we are able to compute all information necessary for the bundle algorithm to proceed and, second that, if a restart is made, the value of α can be sufficiently decreased to guarantee convergence of the sequence $\{\alpha^s\}_{s=1}^{\infty}$ to zero. The first assertion follows since all iterates remain in the set $\{y : F(x_\alpha(y), y) \leq \bar{F} + \hat{\varepsilon}\} \subseteq \hat{Y}$ where the function $x.(\cdot)$ is locally Lipschitz continuous. The second one follows from local Lipschitz continuity of $G.(\bar{y})$ since $G_{\bar{\alpha}^m}(\bar{y}^m) \leq \bar{F}$ and a new value α^1 is searched for satisfying

$$G_{\alpha^1}(\bar{y}^m) \leq \bar{F} + \hat{\varepsilon}.$$

Hence, if K denotes a Lipschitz constant of $G.(\cdot)$ with respect to variations of α on the compact set $[0, \alpha^*] \times \text{cl } D$, then

$$\alpha^1 \leq \max\left\{0, \bar{\alpha}^m - \frac{\hat{\varepsilon}}{2K}\right\}$$

can be used. □

Since the value of $\bar{F}$ is not known a priori in general, the application of the above ideas needs some appropriate control of the changes of α^s. Here, some alternate strategies should be used: first take some iterations in which the values of α^s are changed only very slowly in order to come near to a stationary point for some positive α; then use a more rapid change of α^s in order to reach $\alpha^s \leq \alpha^*$; and so on. We are not able to give more concrete rules about this control since this is only possible after a larger number of numerical experiments which we have not done yet. Also, the numerical behavior of the algorithm in the case when the assumptions of the last theorem are not satisfied is to be investigated in the future.

Example 5.6. The main ideas for the control of α should be illustrated by means of the following very simple example: Let

$$\Psi(y) = \operatorname*{Argmin}_x \{xy : -y \leq x \leq y + 1\}$$

and thus

$$\Psi_\alpha(y) = \operatorname*{Argmin}_x \{xy + \alpha x^2 : -y \leq x \leq y + 1\}, \quad \alpha > 0.$$

Consider the bilevel problem

$$\min\{(x - y + d)^2 : x \in \Psi(y),\ -0.5 \leq y \leq 10\}, \quad d > 0 \text{ fixed.}$$

Then,

$$\Psi(y) = \begin{cases} [0, 1], & y = 0, \\ \{y + 1\}, & -0.5 \leq y < 0, \\ \{-y\}, & 0 < y \leq 10, \end{cases}$$

and, for $y \in [-0.5, 0]$,

$$\Psi_\alpha(y) = \begin{cases} \{-y\}, & \alpha \geq 0.5, \\ \{-\frac{y}{2\alpha}\}, & -\frac{y}{2y+2} \leq \alpha \leq 0.5, \\ \{y+1\}, & 0 < \alpha \leq -\frac{y}{2y+2}, \end{cases}$$

and for $y \in [0, 10]$,

$$\Psi_\alpha(y) = \begin{cases} \{-y\}, & \alpha \leq 0.5, \\ \{-\frac{y}{2\alpha}\}, & \alpha \geq 0.5. \end{cases}$$

If we start with $0 < \alpha < 0.5$ sufficiently large (e.g. $\alpha \geq -\frac{y}{2y+2}$) and $y < 0$, then the optimal solution is $y = 0.5d$ which, occasionally, coincides with the optimal solution of the original problem. The idea to solve the problem is now to use a relatively large α to reach a neighborhood of an optimal solution (which lies outside the region where the algorithm has a difficult behavior due to an increasing Lipschitz constant of the solution function of the lower level problem, if the assumptions of Theorem 5.5 are satisfied). This is a Lipschitz optimization problem. Here we need to compute a point $(x_\alpha(y), y)$ with $y > 0$ for large (fixed) α, say the algorithm stops with $(x_\alpha(y), y) = (-0.25d, 0.25d)$ due to a "non-sufficient" decrease in the next step. Then, after reaching this neighborhood, convergence to an optimal solution is started with decreasing α and the algorithm computes a sequence $\{(x_{\alpha^s}(y^s), y^s, \alpha^s)\}$ converging to $(-0.5d, 0.5d, 0)$. □

6. Conclusions

In the paper we have investigated the question if it is possible to use a standard nondifferentiable optimization algorithm for computing approximate solutions for bilevel programming problems with generally non-unique lower level solutions. We have shown that this is indeed possible if the lower level problem is appropriately regularized. By use of an additional assumption which means that at all points of interest in the bilevel problem the lower level solution is also strongly stable [17], convergence of the algorithm to a stationary solution of the bilevel problem has been shown. It is one task for future investigations to make numerical experiments with this algorithm. The results of the experiments in [24, 25] for problems satisfying stronger assumptions (all lower level solutions are strongly stable and the linear independence constraint qualification is satisfied) are very encouraging for this aim.

Acknowledgment

I would like to thank the two anonymous referees and the Editor-in-Chief, Prof. W.W. Hager, for their valuable comments and suggestions.

References

1. B. Bank, J. Guddat, D. Klatte, B. Kummer, and K. Tammer, Non-Linear Parametric Optimization, Akademie-Verlag: Berlin, 1982.
2. F.H. Clarke, Optimization and Nonsmooth Analysis, J. Wiley & Sons: New York, 1983.
3. S. Dempe, "Directional differentiability of optimal solutions under Slater's condition," Mathematical Programming, vol. 59, pp. 49–69, 1993.
4. S. Dempe, " On generalized differentiability of optimal solutions and its application to an algorithm for solving bilevel optimization problems," in Recent Advances in Nonsmooth Optimization, D.-Z. Du, L. Qi, and R.S. Womersley (Eds.), World Scientific Publishers: Singapore et al., 1995, pp. 36–56.
5. S. Dempe, "An implicit function approach to bilevel programming problems," In Multilevel Optimization: Algorithms and Applications, A. Migdalas, P.M. Pardalos, and P. Värbrand (Eds.), Kluwer Academic Publishers, Dordrecht, 1998. pp. 273–294.
6. S. Dempe and H. Schmidt, "On an algorithm solving two-level programming problems with nonunique lower level solutions," Computational Optimization and Applications, vol. 6, pp. 227–249, 1996.
7. J.F. Falk and J. Liu, "On bilevel programming, Part i: general nonlinear cases," Mathematical Programming, vol. 70, pp. 47–72, 1995.
8. A.V. Fiacco and G.P. McCormic, Nonlinear Programming: Sequential Unconstrained Minimization Techniques, J. Wiley & Sons: New York, 1968.
9. J. Gauvin, "A necessary and sufficient regularity condition to have bounded multipliers in nonconvex programming," Mathematical Programming, vol. 12, pp. 136–139, 1977.
10. W.W. Hager, "Lipschitz continuity for constrained processes," SIAM Journal on Control and Optimization, vol. 17, pp. 321–328, 1979.
11. P.T. Harker and J.-S. Pang, "Existence of optimal solutions to mathematical programs with equilibrium constraints," Operations Research Letters, vol. 7, pp. 61–64, 1988.
12. R. Janin, "Sensitivity for nonconvex optimization problems," in Convex Analysis and its Applications, A. Auslender (Ed.), Springer Verlag: Berlin, 1977. Lecture Notes in Economics and Mathematical Systems, vol. 144, pp. 115–119.
13. R. Janin, "Directional derivative of the marginal function in nonlinear programming," Mathematical Programming Study, vol. 21, pp. 110–126, 1984.
14. K.C. Kiwiel, Methods of Descent for Nondifferentiable Optimization, Springer-Verlag: Berlin, 1985.
15. K.C. Kiwiel, "Proximity control in bundle methods for convex nondifferentiable minimization," Mathematical Programming, vol. 46, pp. 105–122, 1990.
16. K.C. Kiwiel, "Restricted step and Levenberg-Marquardt techniques in proximal bundle methods for nonconvex nondifferentiable optimization," SIAM Journal on Optimization, vol. 6, pp. 227–249, 1996.
17. M. Kojima, "Strongly stable stationary solutions in nonlinear programs," in Analysis and Computation of Fixed Points, S.M. Robinson (Ed.), Academic Press: New York, 1980. pp. 93–138.
18. B. Kummer, "Newton's method for non-differentiable functions," in Advances in Mathematical Optimization, Akademie-Verlag: Berlin, 1988. Mathematical Research, vol. 45.
19. P. Loridan and J. Morgan, "Approximate solutions for two-level optimization problems," in Trends in Mathematical Optimization, K. Hoffman, J. Hiriart-Urruty, C. Lemarechal and J. Zowe (Eds.), Birkhäuser Verlag: Basel, 1988. International Series of Numerical Mathematics, vol. 84, pp. 181–196.
20. P. Loridan and J. Morgan, "Least-norm regularization for weak two-level optimization problems," in Optimization, Optimal Control and Partial Differential Equations, Birkhäuser Verlag, Basel, 1992. International Series of Numerical Mathematics, vol. 107, pp. 307–318.
21. R. Lucchetti, F. Mignanego, and G. Pieri, "Existence theorem of equilibrium points in Stackelberg games with constraints," Optimization, vol. 18, pp. 857–866, 1987.
22. Z.-Q. Luo, J.-S. Pang, and D. Ralph, Mathematical Programs with Equilibrium Constraints, Cambridge University Press, Cambridge, 1996.
23. K. Malanowski, " Differentiability with respect to parameters of solutions to convex programming problems," Mathematical Programming, vol. 33, pp. 352–361, 1985.
24. J. Outrata, "On the numerical solution of a class of Stackelberg problems," ZOR-Mathematical Methods of Operations Research, vol. 34, pp. 255–277, 1990.

25. J. Outrata and J. Zowe, “A numerical approach to optimization problems with variational inequality constraints,” Mathematical Programming, vol. 68, pp. 105–130, 1995.
26. D. Ralph and S. Dempe, “Directional derivatives of the solution of a parametric nonlinear program,” Mathematical Programming, vol. 70, pp. 159–172, 1995.
27. S. Scholtes, “Introduction to piecewise differentiable equations,” Technical Report, Universität Karlsruhe, Institut für Statistik und Mathematische Wirtschaftstheorie, No. 53, 1994.
28. H. Schramm, *Eine Kombination von bundle- und trust-region-Verfahren zur Lösung nichtdifferenzierbarer Optimierungsprobleme*, Bayreuther Mathematische Schriften, No. 30, 1989.
29. H. Schramm and J. Zowe, “A version of the bundle idea for minimizing a nonsmooth function: conceptual idea, convergence analysis, numerical results,” SIAM Journal on Optimization, vol. 2, pp. 121–152, 1992.
30. A. Shapiro, “Sensitivity analysis of nonlinear programs and differentiability properties of metric projections,” SIAM Journal on Control and Optimization, vol. 26, pp. 628–645, 1988.