A smoothing augmented Lagrangian method for solving simple bilevel programs

Mengwei Xu[*] and Jane J. Ye[†]

Dedicated to Masao Fukushima in honor of his 65th birthday

Abstract. In this paper, we design a numerical algorithm for solving a simple bilevel program where the lower level program is a nonconvex minimization problem with a convex set constraint. We propose to solve a combined problem where the first order condition and the value function are both present in the constraints. Since the value function is in general nonsmooth, the combined problem is in general a nonsmooth and nonconvex optimization problem. We propose a smoothing augmented Lagrangian method for solving a general class of nonsmooth and nonconvex constrained optimization problems. We show that, if the sequence of penalty parameters is bounded, then any accumulation point is a Karush-Kuch-Tucker (KKT) point of the nonsmooth optimization problem. The smoothing augmented Lagrangian method is used to solve the combined problem. Numerical experiments show that the algorithm is efficient for solving the simple bilevel program.

Key Words. Bilevel program, value function, smoothing method, augmented Lagrangian method, partial calmness, principal-agent problem.

2010 Mathematics Subject Classification. 65K10, 90C26.

[*] School of Mathematical Sciences, Dalian University of Technology, Dalian 116024, China. E-mail: xumengw@hotmail.com.

[†] Corresponding Author. Department of Mathematics and Statistics, University of Victoria, Victoria, B.C., Canada V8W 3R4. E-mail: janeye@uvic.ca. The research of this author was partially supported by NSERC.

1 Introduction.

In this paper, we propose a numerical algorithm for solving the following simple bilevel program:

$$\begin{array}{lll} \text{(SBP)} & \min & F(x,y) \\ & \text{s.t.} & g_i(x,y) \leq 0,\ i=m+1,\cdots,l, \\ & & x \in X, \quad y \in S(x), \end{array}$$

where $S(x)$ denotes the set of solutions of the lower level program

$$(\mathrm{P}_x) \qquad \min_{y \in Y} \ f(x,y),$$

where X is a closed and convex subset of $\mathbb{R}^n$, Y is a closed, convex and compact subset of $\mathbb{R}^m$, $f, F, g_i : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ are continuously differentiable functions, $i = m+1, \cdots, l$, and f is twice continuously differentiable in variable y. The simple bilevel program is a special case of a general bilevel program where the constraint set Y may depend on x. The reader is referred to [2, 10, 11, 27, 28] for applications and recent developments of general bilevel programs. (SBP) has many applications including a very important model in economics called the moral hazard model of the principal-agent problem [20].

The classical Karush-Kuhn-Tucker (KKT) approach (also called the first order approach) to solve bilevel program is to replace the solution set $S(x)$ by the set of KKT points of the lower level problem and consider the following single level optimization problem:

$$\begin{array}{lll} \text{(SP)} & \min & F(x,y) \\ & \text{s.t.} & 0 \in \nabla_y f(x,y) + \mathcal{N}_Y(y), \\ & & g_i(x,y) \leq 0,\ i=m+1,\cdots,l, \\ & & (x,y) \in X \times Y, \end{array}$$

where $\mathcal{N}_Y(y) := \{\xi : \langle \xi, y' - y \rangle \leq 0, \quad \forall y' \in Y\}$ denotes the normal cone of Y at y in the sense of convex analysis and $\nabla_y f$ denotes the gradient of f with respect to variable y. When the lower level constraint set Y has some structures, e.g. described by some equality and/or inequality constraints, problem (SP) is reduced to the so-called mathematical program with equilibrium constraints (MPEC) (see e.g. [19, 22, 23, 29]). If for all $x \in X$, $y \in S(x)$ lies in the interior of the set Y, then $\mathcal{N}_Y(y) = \{0\}^m$ and (SP) is a nonlinear program. For the case where the lower level objective function f is convex in variable y, (SBP) and its first order reformulation (SP) are equivalent. In the case where f is not convex in variable y, it is tempting to believe that a locally optimal solution of

(SBP) must be a KKT point of the problem (SP). This turns out to be wrong as it is shown by a counter example of Mirrlees [20]. Hence using the first order approach to solve (SBP) is not valid in the sense that the true optimal solution may be missed.

In recent years, most of the numerical algorithms for bilevel programs assume that the lower level program is convex with few exceptions [17, 21]. Since the first order approach is not valid for (SBP) in general, it remains a very difficult problem to solve theoretically and numerically. In this paper we will try to attack this difficult problem. In particular we do not assume that f is convex in variable y.

It is obvious that (SBP) can be reformulated as the following single level optimization problem involving the value function:

$$\begin{aligned} \text{(VP)} \qquad \min \quad & F(x,y) \\ \text{s.t.} \quad & f(x,y) - V(x) \leq 0, \qquad (1.1) \\ & g_i(x,y) \leq 0, \ i = m+1, \cdots, l, \\ & (x,y) \in X \times Y, \end{aligned}$$

where $V(x) := \inf_{y \in Y} f(x,y)$ is the value function of the lower level problem. This approach was first proposed by Outrata [22] for a numerical purpose and used to derive necessary optimality conditions by Ye and Zhu [30, 31]. Under the given assumptions, the value function is Lipschitz continuous and hence the Fritz John type necessary optimality condition of Clarke [6, Theorem 6.1.1] holds. However since the nonsmooth Mangasarian Fromovitz constraint qualification (MFCQ) for problem (VP) will never be satisfied; see [30, Proposition 3.2], the nonsmooth KKT condition may not hold at a local optimal solution. For the nonsmooth KKT condition to hold at an optimal solution, Ye and Zhu [30, 31] introduced the partial calmness condition under which the difficult constraint (1.1) is moved to the objective function. Based on the value function approach, recently [17] proposed a numerical algorithm to solve the problem (VP) when the problem (SBP) is partially calm and to solve an approximate bilevel problem $(\text{VP})_\varepsilon$ where the constraint (1.1) is replaced by $f(x,y) - V(x) \leq \varepsilon$ for small $\varepsilon > 0$ otherwise.

The partial calmness condition, however, is rather strong and hence a local optimal solution of a bilevel program may not satisfy the KKT condition of (VP). Recently Ye and Zhu [32] proposed to combine the first order and the value function approaches. For the problem (SBP), it amounts to consider the following combined program:

$$\begin{aligned} \text{(CP)} \qquad \min \quad & F(x,y) \\ \text{s.t.} \quad & f(x,y) - V(x) \leq 0, \qquad (1.2) \\ & 0 \in \nabla_y f(x,y) + \mathcal{N}_Y(y), \qquad (1.3) \end{aligned}$$

$$\begin{aligned} & g_i(x,y) \le 0, \ i = m+1, \cdots, l, \\ & (x,y) \in X \times Y. \end{aligned}$$

The advantages of solving (CP) instead of the problem (SP) or (VP) are twofold. On one hand since the value function constraint (1.2) is present, a locally optimal solution of (SBP) is guaranteed to be a KKT point of the problem (CP) (it may not be a KKT point of the problem (SP)). On the other hand since an extra constraint – the first order condition (1.3) is present, the necessary optimality condition for problem (CP) is much more likely to hold than the one for problem (VP) since there is more flexibility in choosing a multiplier.

In this paper, we propose an algorithm to solve the problem (CP). To concentrate on the main idea, we propose to solve (CP) under the following assumption.

Assumption 1.1 *For every $x \in X$, y is an interior point of set Y if $y \in S(x)$.*

Under Assumption 1.1, $\mathcal{N}_Y(y) = \{0\}^m$ and the constraint $0 \in \nabla_y f(x,y) + \mathcal{N}_Y(y)$ reduces to $\nabla_y f(x,y) = 0$. For some applications such as the principal-agent problem [20], it is a common practice to assume that the solution of the lower level problem lies in the interior of set Y which is a bounded interval since the solution of the lower level problem usually can be estimated to lie in the interior of certain bounded interval.

It is well-known that the value function is in general a nonsmooth function even though the function f is continuously differentiable. Danskin's theorem ([7, Page 99] or [9]) guaranteed that it is a Lipschitz continuous function with a computable Clarke generalized gradient. Hence (CP) is a nonsmooth optimization problem with all functions continuously differentiable except the value function $V(x)$. Lin, Xu and Ye [17] proposed the following integral entropy function to approximate the value function:

$$\gamma_\rho(x) \ := \ -\rho^{-1} \ln \left(\int_Y \exp[-\rho f(x,y)] dy \right). \tag{1.4}$$

It was shown [17] that the integral entropy function is a smoothing function for the value function in the sense that $\gamma_\rho(z) \to V(x)$ as $z \to x$ and $\rho \to +\infty$ and it satisfies the gradient consistent property. A smoothing projected gradient algorithm is then proposed to solve (VP) if the problem (VP) is partially calm and to solve $(\text{VP})_\varepsilon$ otherwise. In this paper we use the integral entropy function to approximate the value function and propose to solve (CP) under the partial calmness condition. Although the partial calmness condition is a very strong condition for (VP), it is likely to hold for (CP); see [32] and hence we are able to solve a large class of the original bilevel problem (SBP) instead of solving an approximate problem.

Problem (CP) is a nonsmooth and nonconvex constrained optimization problem. Recently smoothing methods for solving nonsmooth and nonconvex unconstrained optimization problems have been proposed [4, 34]. [17] combined the smoothing technique with the gradient projected method to solve a class of nonsmooth and nonconvex constrained optimization problems and used it to solve (VP). Smoothing technique has the advantage over other algorithms such as the sampling gradient algorithm [8] for solving nonsmooth and nonconvex problems in that one does not need to evaluate the function value or its gradient. Such an algorithm turns out to be useful for solving bilevel programs since one does not need to solve the lower level problem at every iteration. Solving problem (CP) means that there are a lot more constraints than problem (VP) due to the first order condition $\nabla_y f(x,y) = 0$. To handle more constraints we use the augmented Lagrangian method. The augmented Lagrangian method is also known as the method of multipliers. It has been studied from various angles in [14, 24, 25, 26] and is the basis for some successful softwares such as ALGENCAN [1] and LANCELOT [16]. One of the main contributions of this paper is the designing of a smoothing augmented Lagrangian method for solving a general *nonsmooth and nonconvex* constrained optimization problem.

The rest of the paper is organized as follows. In Section 2, we propose a new algorithm for solving a class of nonsmooth and nonconvex optimization problems where only one constraint is nonsmooth by combining the smoothing technique and the classical augmented Lagrangian algorithm and establish convergence for the algorithm. In Section 3, we use the entropy integral function as a smoothing function of the value function and apply the new algorithm to the problem (CP). We report our numerical experiments for some bilevel programs and a moral hazard problem in Section 4.

We adopt the following standard notation in this paper. For any two vectors a and b in $\mathbb{R}^n$, we denote by $a^T b$ their inner product. Given a function $G : \mathbb{R}^n \to \mathbb{R}^m$, we denote its Jacobian by $\nabla G(z) \in \mathbb{R}^{m\times n}$ and, if $m = 1$, the gradient $\nabla G(z) \in \mathbb{R}^n$ is considered as a column vector. For a set $\Omega \subseteq \mathbb{R}^n$, we denote by int$\Omega$, co$\Omega$, and dist$(x, \Omega)$ the interior, the convex hull, and the distance from x to Ω respectively. For a matrix $A \in \mathbb{R}^{n\times m}$, A^T denotes its transpose. In addition, we let $\mathbf{N}$ be the set of nonnegative integers and exp$[z]$ be the exponential function. For any sets $\Omega_i \subseteq \mathbb{R}^n, i = 1, \cdots, p$, we denote their direct product by

$$\bigotimes_{i=1}^{p} \Omega_i := \{(\omega_1, \cdots, \omega_p) : \omega_i \in \Omega_i, i = 1, \cdots, p\}.$$

2 Smoothing augmented Lagrangian algorithm for nonsmooth nonconvex programs

As discussed in the introduction, our aim is to solve the problem (CP) which is a nonsmooth constrained optimization problem where all functions except one is smooth. In this section, we propose an algorithm which combines the smoothing technique with the classical augmented Lagrangian algorithm to solve the following nonsmooth constrained optimization problem:

$$
\begin{array}{lll}
(\mathrm{P}) & \min & G(x) \\
 & \text{s.t.} & g_0(x) \leq 0, \\
 & & g_i(x) \leq 0,\ i = 1, \cdots, p, \\
 & & g_i(x) = 0,\ i = p+1, \cdots, q, \\
 & & x \in \Omega,
\end{array}
$$

where $\Omega \subseteq \mathbb{R}^n$ is a nonempty closed convex set, $G, g_i : \mathbb{R}^n \to \mathbb{R}$ are continuously differentiable, $i = 1, \cdots, q$, and $g_0 : \mathbb{R}^n \to \mathbb{R}$ is locally Lipschitzian but not necessarily differentiable.

We denote by $\partial g_0(\bar{x})$ the Clarke generalized gradient of g_0 at $\bar{x}$ and recall [6, Theorem 2.5.1] that $\partial g_0(\bar{x}) = \mathrm{co}\left\{\lim \nabla g_0(x_i) : x_i \to \bar{x}, g_0 \text{ is differentiable at } x_i\right\}$.

Definition 2.1 (KKT point) *We call a feasible point $\bar{x}$ of problem* (P) *a KKT point if there exists scalars $\mu_0, \mu_1, \ldots, \mu_q$ such that*

$$
\begin{aligned}
& 0 \in \nabla G(\bar{x}) + \mu_0 \partial g_0(\bar{x}) + \sum_{i=1}^{q} \mu_i \nabla g_i(\bar{x}) + \mathcal{N}_\Omega(\bar{x}), \\
& \mu_i \geq 0,\ \mu_i g_i(\bar{x}) = 0,\ i = 0, \cdots, p.
\end{aligned}
$$

We now give a condition to verify that a feasible point is a KKT point which will be used later to verify that an accumulation point is a KKT point.

Proposition 2.1 *Let $\bar{x}$ be a feasible point of problem* (P) *and v be an element of $\partial g_0(\bar{x})$. Suppose that*

$$
\nabla G(\bar{x})^T d \geq 0 \tag{2.1}
$$

for all d in the linearization cone of the feasible region

$$
\begin{aligned}
\mathcal{L}(\bar{x}) :\ = \ & \{d \in \mathcal{T}_\Omega(\bar{x}) : \nabla g_i(\bar{x})^T d = 0,\ i = p+1, \cdots, q, \\
& v^T d \leq 0,\ \text{if } g_0(\bar{x}) = 0, \nabla g_i(\bar{x})^T d \leq 0,\ i \in I(\bar{x})\},
\end{aligned}
$$

where $I(\bar{x}) := \{i = 1, \cdots, p : g_i(\bar{x}) = 0\}$. Then $\bar{x}$ is a KKT *point of* (P)*.*

Proof. By (2.1), $d = 0$ is an optimal solution to the following linearized problem:

$$\begin{array}{ll} \min\limits_{d} & \Phi(d) := \nabla G(\bar{x})^T d \\ \text{s.t.} & \nabla g_i(\bar{x})^T d = 0, \ i = p+1, \cdots, q, \\ & v^T d \leq 0, \ \text{if } g_0(\bar{x}) = 0, \\ & \nabla g_i(\bar{x})^T d \leq 0, \ i \in I(\bar{x}), \\ & d \in \mathcal{T}_\Omega(\bar{x}). \end{array}$$

Since the objective function and the constraint functions are all linear in variable d, the KKT condition holds at the optimal solution. Hence there exist multipliers μ_i, $i = 0, \cdots, q$ such that

$$\begin{aligned} & 0 \in \nabla G(\bar{x}) + \mu_0 v + \sum_{i=1}^{q} \mu_i \nabla g_i(\bar{x}) + \mathcal{N}_\Omega(\bar{x}), \qquad (2.2) \\ & \mu_i \geq 0, \ i \in I(\bar{x}), \ \mu_i = 0, \ i = 1, \cdots, p, \ i \notin I(\bar{x}), \\ & \mu_0 \geq 0, \ \text{if } g_0(\bar{x}) = 0, \ \mu_0 = 0, \ \text{if } g_0(\bar{x}) < 0. \end{aligned}$$

Since $v \in \partial g_0(\bar{x})$, $\bar{x}$ is a KKT point of (P). ■

Following from the Fritz John type necessary optimality condition [6, Theorem 6.1.1], we define the following constraint qualification, which is weaker than the nonsmooth MFCQ, but equivalent to the nonsmooth MFCQ if the interior of the tangent cone $\mathcal{T}_\Omega(\bar{x})$ is not empty [15].

Definition 2.2 (NNAMCQ) *We say that the no nonzero abnormal multiplier constraint qualification* (NNAMCQ) *holds at a feasible point $\bar{x}$ of problem* (P) *if there is no scalars $\mu_0, \mu_1, \ldots, \mu_q$ not all zero such that*

$$\begin{aligned} & 0 \in \mu_0 \partial g_0(\bar{x}) + \sum_{i=1}^{q} \mu_i \nabla g_i(\bar{x}) + \mathcal{N}_\Omega(\bar{x}), \\ & \mu_i \geq 0, \ \mu_i g_i(\bar{x}) = 0, \ i = 0, \cdots, p. \end{aligned}$$

In order to accommodate infeasible accumulation points in the numerical algorithm, we now extend the NNAMCQ to infeasible points.

Definition 2.3 (ENNAMCQ) *We say that the extended no nonzero abnormal multiplier constraint qualification* (ENNAMCQ) *holds at $\bar{x} \in \Omega$ if there is no scalars $\mu_0, \mu_1, \ldots, \mu_q$ not all zero such that*

$$\begin{aligned} & 0 \in \mu_0 \partial g_0(\bar{x}) + \sum_{i=1}^{q} \mu_i \nabla g_i(\bar{x}) + \mathcal{N}_\Omega(\bar{x}), \\ & \mu_i \geq 0, \ \mu_i (g_i(\bar{x}) - \max\{0, g_i(\bar{x})\}) = 0, \ i = 0, \cdots, p. \end{aligned}$$

Definition 2.4 *Assume that, for a given $\rho > 0$, $g_\rho : \mathbb{R}^n \to \mathbb{R}$ is a continuously differentiable function. We say that $\{g_\rho : \rho > 0\}$ is a family of smoothing functions of g_0 if $\lim\limits_{z\to x,\ \rho\uparrow\infty} g_\rho(z) = g_0(x)$ for any fixed $x \in \mathbb{R}^n$.*

Definition 2.5 [5] *We say that a family of smoothing functions $\{g_\rho : \rho > 0\}$ satisfies the gradient consistent property if for any $x \in \mathbb{R}^n$, $\limsup\limits_{z\to x,\, \rho\uparrow\infty} \nabla g_\rho(z)$ is nonempty and $\limsup\limits_{z\to x,\, \rho\uparrow\infty} \nabla g_\rho(z) \subseteq \partial g_0(x)$, where*

$$\limsup_{z\to x,\, \rho\uparrow\infty} \nabla g_\rho(z) := \left\{ \lim_{k\to\infty} \nabla g_{\rho_k}(z_k) : z_k \to x, \rho_k \uparrow \infty \right\}$$

is the set of all limit points.

We approximate the nonsmooth function $g_0(x)$ by its smoothing function $g_\rho(x)$, define the augmented Lagrangian function as

$$\begin{aligned} G_\rho^{\lambda,c}(x) \;\; := \;\; & G(x) + \frac{1}{2c}\sum_{i=1}^{p}\left(\max\{0, \lambda_i + cg_i(x)\}^2 - \lambda_i^2\right) \\ & + \sum_{i=p+1}^{q}\left(\lambda_i g_i(x) + \frac{c}{2}(g_i(x))^2\right) + \frac{1}{2c}\left(\max\{0, \lambda_0 + cg_\rho(x)\}^2 - \lambda_0^2\right) \end{aligned}$$

and consider the following unconstrained optimization problem for $\rho > 0, c > 0, \lambda \in R^{q+1}$:

$$(\mathrm{P}_\rho^{\lambda,c}) \qquad \min_{\mathrm{x}\in\Omega} \quad G_\rho^{\lambda,c}(x).$$

Since $(\mathrm{P}_\rho^{\lambda,c})$ is a smooth optimization problem with a convex constraint set for any fixed $\rho > 0, c > 0, \lambda \in R^{q+1}$, we suggest a projected gradient algorithm for problem $(\mathrm{P}_\rho^{\lambda,c})$. We then update the iteration by increasing the smoothing parameter ρ and the penalty parameter c and update the multiplier λ. We will show that any convergent subsequence of iteration points generated by the algorithm converges to a KKT point of problem (P) when ρ goes to infinity and the penalty parameter c is bounded. We will also show that, under the ENNAMCQ, the penalty parameter must be bounded.

We propose the following smoothing augmented Lagrangian algorithm. In the algorithm, we denote by P_Ω the projection operator onto Ω, that is,

$$P_\Omega[x] := \arg\min\{\|z - x\| : z \in \Omega\}$$

and the residual function to measure infeasibility and complementarity:

$$\begin{aligned} & \sigma_\rho^\lambda(x) := \\ & \max\left\{|g_j(x)|, j = p+1, \cdots, q,\ |\min\{\lambda_i, -g_i(x)\}|,\ i = 1, \cdots, p, |\min\{\lambda_0, -g_\rho(x)\}|\right\}. \end{aligned}$$

Algorithm 2.1 *Let $\{\beta, \gamma, \sigma_1, \sigma_2\}$ be constants in $(0,1)$ with $\sigma_1 \leq \sigma_2$, $\epsilon \geq 0, \epsilon_1 \geq 0$ be very small constants, $\{\sigma, \sigma', \hat{\eta}\}$ be constants in $(1, \infty)$. Choose an initial point $x^0 \in \Omega$, an initial smoothing parameter $\rho_0 > 0$, an initial penalty parameter $c_0 > 0$, an initial multiplier $\bar{\lambda}^0 \in \bigotimes_{i=0}^{p}[0, \lambda_{max}] \times \bigotimes_{i=p+1}^{q}[\lambda_{min}, \lambda_{max}]$, where $\lambda_{min} < 0$ and $\lambda_{max} > 0$ are given constants and set $k := 0, s := 0$.*

1. *Let $z_0^k := x^k$ and $z_{s+1}^k := P_\Omega[z_s^k - \alpha_s \nabla G_{\rho_k}^{\bar{\lambda}^k, c_k}(z_s^k)]$, where $\alpha_s := \beta^{l_s}$, $l_s \in \{0, 1, 2 \cdots\}$ is the smallest number satisfying*

$$G_{\rho_k}^{\bar{\lambda}^k, c_k}(z_{s+1}^k) - G_{\rho_k}^{\bar{\lambda}^k, c_k}(z_s^k) \leq \sigma_1 \nabla G_{\rho_k}^{\bar{\lambda}^k, c_k}(z_s^k)^T \left(z_{s+1}^k - z_s^k\right) \tag{2.3}$$

and $\beta^{l_s} \geq \gamma$, or $\bar{\alpha}_s := \beta^{l_s - 1}$ such that $\bar{z}_{s+1}^k := P_\Omega[z_s^k - \bar{\alpha}_s \nabla G_{\rho_k}^{\bar{\lambda}^k, c_k}(z_s^k)]$ satisfies

$$G_{\rho_k}^{\bar{\lambda}^k, c_k}(\bar{z}_{s+1}^k) - G_{\rho_k}^{\bar{\lambda}^k, c_k}(z_s^k) > \sigma_2 \nabla G_{\rho_k}^{\bar{\lambda}^k, c_k}(z_s^k)^T \left(\bar{z}_{s+1}^k - z_s^k\right). \tag{2.4}$$

If

$$\frac{\|z_{s+1}^k - z_s^k\|}{\alpha_s} < \hat{\eta} \rho_k^{-1}, \tag{2.5}$$

set $x^{k+1} := z_{s+1}^k$, $\rho_{k+1} := \sigma \rho_k$, $s := 0$, go to Step 2. Otherwise, set $s = s + 1$, and go to Step 1.

2. *Set*

$$\lambda_0^{k+1} = \max\{0, \bar{\lambda}_0^k + c_k g_{\rho_k}(x^{k+1})\}; \tag{2.6}$$
$$\lambda_i^{k+1} = \max\{0, \bar{\lambda}_i^k + c_k g_i(x^{k+1})\},\ i = 1, \cdots, p; \tag{2.7}$$
$$\lambda_i^{k+1} = \bar{\lambda}_i^k + c_k g_i(x^{k+1}),\ i = p + 1, \cdots, q. \tag{2.8}$$

Take $\bar{\lambda}^{k+1}$ as the Euclidean projection of λ^{k+1} onto $\bigotimes_{i=0}^{p}[0, \lambda_{max}] \times \bigotimes_{i=p+1}^{q}[\lambda_{min}, \lambda_{max}]$, and go to Step 3.

3. *If*

$$\sigma_{\rho_k}^{\lambda^{k+1}}(x^{k+1}) \leq \epsilon, \tag{2.9}$$

go to Step 4. Else if $k = 0$ or

$$\sigma_{\rho_k}^{\lambda^{k+1}}(x^{k+1}) \leq \gamma \sigma_{\rho_{k-1}}^{\lambda^k}(x^k), \tag{2.10}$$

set $k = k + 1$ and go to Step 1. Otherwise, set $c_{k+1} := \sigma' c_k$, $k = k + 1$ and go to Step 1.

4. If

$$\|P_\Omega[x^{k+1} - \nabla G_{\rho_k}^{\bar{\lambda}^k, c_k}(x^{k+1})] - x^{k+1}\| = 0, \tag{2.11}$$

or $\|x^{k+1} - x^k\| \leq \epsilon_1$*, terminate. Otherwise, set* $k = k + 1$*, and go to Step 1.*

It is easy to see that Step 1 of the algorithm is the classical projected gradient algorithm with Armijo line search as in [3] when $\gamma_2 = \beta$. Condition (2.3) on line search step size forces a sufficient decrease of the function value while condition (2.4) guarantees that the line search step size is not too small. In practice, only a small number of iterations are required to compute the Armijo step size. Since Dunn [12] has shown that if z_{s+1}^k is not stationary then (2.4) fails for all $\bar{\alpha}_s$ sufficiently small, it follows that a line search step α_s can be always found provided that $\sigma_1 \leq \sigma_2$. From the updating rule of c_k and λ^k, we know that the boundedness of $\{c_k\}$ implies the boundedness of $\{\lambda^k\}$.

Suppose that Algorithm 2.1 does not terminate within a finite number of iterations. The next theorem shows the global convergence of Algorithm 2.1. We first make the following standing assumption. Under the assumptions of this section the condition holds automatically if the set Ω is compact.

Assumption 2.1 *For any fixed* $\rho > 0$*,* $c > 0$ *and* λ*,* $G_\rho^{\lambda,c}(\cdot)$ *is bounded below and* $\nabla G_\rho^{\lambda,c}(\cdot)$ *is uniformly continuous on the level set* $\{x \in \Omega : G_\rho^{\lambda,c}(x) \leq \Gamma\}$ *for any* $\rho > 0, \lambda \in R^{q+1}, c > 0, \Gamma > 0$.

The following lemmas are well-known.

Lemma 2.1 [3, Lemma 2.1]

(a) *For any* $x \in \mathbb{R}^n$ *and* $z \in \Omega$*, we have* $(P_\Omega[x] - x)^T(z - P_\Omega[x]) \geq 0$.

(b) $P_\Omega[x]$ *is a monotone operator, that is,* $(P_\Omega[y] - P_\Omega[x])^T(y - x) \geq 0$ *for* $x, y \in \mathbb{R}^n$*. If* $P_\Omega[y] \neq P_\Omega[x]$*, then strict inequality holds.*

Lemma 2.2 [3, Lemma 2.2] or [13] *For any* $x \in \mathbb{R}^n$ *and* $d \in \mathbb{R}^n$*, the function* ψ *defined by*

$$\psi(\alpha) := \frac{\|P_\Omega(x + \alpha d) - x\|}{\alpha}, \quad \alpha > 0$$

is nonincreasing.

Note that by setting $x := z_s^k - \alpha_s \nabla G_{\rho_k}^{\bar{\lambda}^k, c_k}(z_s^k)$ and $z := z_s^k$ in Lemma 2.1 (a), the following inequality can be obtained immediately:

$$\nabla G_{\rho_k}^{\bar{\lambda}^k, c_k}(z_s^k)^T \left(z_{s+1}^k - z_s^k\right) \leq -\frac{\|z_{s+1}^k - z_s^k\|^2}{\alpha_s}. \tag{2.12}$$

Set $y = z_s^k - \alpha_s \nabla G_{\rho_k}^{\bar{\lambda}^k, c_k}(z_s^k)$ and $x = z_s^k - \bar{\alpha}_s \nabla G_{\rho_k}^{\bar{\lambda}^k, c_k}(z_s^k)$ in Lemma 2.1 (b), since $\bar{\alpha}_s - \alpha_s > 0$, we have

$$\nabla G_{\rho_k}^{\bar{\lambda}^k, c_k}(z_s^k)^T (z_{s+1}^k - \bar{z}_{s+1}^k) \geq 0. \tag{2.13}$$

Combining the proofs of [3, Theorem 2.3] and [34, Lemma 2.3], the following results can be shown. We show it for completeness.

Lemma 2.3 *Under Assumption 2.1, if Algorithm 2.1 does not terminate at Step 4, we have for each k,*

$$\lim_{s \to \infty} \frac{\|z_{s+1}^k - z_s^k\|}{\alpha_s} = 0, \tag{2.14}$$

and hence $\lim_{k \to \infty} \rho_k = +\infty$.

Proof. We assume for a contradiction that for any given k and any $\varepsilon > 0$, there is an subsequence $K \subseteq \mathbf{N}$ such that for $s \in K$,

$$\frac{\|z_{s+1}^k - z_s^k\|}{\alpha_s} \geq \varepsilon.$$

Without loss of generality, in the following proof let $K = \mathbf{N}$. Since $\{G_{\rho_k}^{\bar{\lambda}^k, c_k}(z_s^k)\}_{s \in K}$ converges, from condition (2.3) and (2.12), we have as $s \to \infty$

$$0 \leftarrow \nabla G_{\rho_k}^{\bar{\lambda}^k, c_k}(z_s^k)^T \left(z_s^k - z_{s+1}^k\right) \geq \varepsilon \max\{\varepsilon \alpha_s, \|z_{s+1}^k - z_s^k\|\}.$$

Hence,

$$\lim_{s \to \infty} \alpha_s = 0, \quad \text{and} \quad \lim_{s \to \infty} \|z_{s+1}^k - z_s^k\| = 0.$$

This also implies that $\alpha_s = \beta^{l_s} < \gamma$.

Lemma 2.2 implies that

$$\frac{\|z_{s+1}^k - z_s^k\|^2}{\alpha_s} \geq \alpha_s \left(\frac{\|z_{s+1}^k - z_s^k\|}{\alpha_s}\right) \left(\frac{\|\bar{z}_{s+1}^k - z_s^k\|}{\bar{\alpha}_s}\right) \geq \varepsilon\beta \|\bar{z}_{s+1}^k - z_s^k\|. \tag{2.15}$$

Condition (2.13) together with (2.12) and (2.15) imply that for $s \in K$

$$\nabla G_{\rho_k}^{\bar{\lambda}^k, c_k}(z_s^k)^T \left(z_s^k - \bar{z}_{s+1}^k\right) \geq \nabla G_{\rho_k}^{\bar{\lambda}^k, c_k}(z_s^k)^T \left(z_s^k - z_{s+1}^k\right) \geq \varepsilon\beta \|\bar{z}_{s+1}^k - z_s^k\|. \tag{2.16}$$

From $\nabla G_{\rho_k}^{\bar{\lambda}^k, c_k}(z_s^k)^T \left(z_s^k - z_{s+1}^k\right) \to 0$, we have $\|\bar{z}_{s+1}^k - z_s^k\| \to 0$.

Since $\{G_{\rho_k}^{\bar{\lambda}^k, c_k}(z_s^k)\}_{s \in K}$ is monotonously nonincreasing (since for any k, $\{z_s^k\}$ is a iteration sequence of the projected gradient method), there must exists $\Gamma > 0$ such that for $\theta \in [0, 1]$,

$$G_{\rho_k}^{\bar{\lambda}^k, c_k}(z_{s+1}^k) \leq \Gamma \quad \text{and} \quad G_{\rho_k}^{\bar{\lambda}^k, c_k}(\theta \bar{z}_{s+1}^k + (1 - \theta) z_s^k) \leq \Gamma.$$

From the Mean Value Theorem and the uniform continuity of $\nabla G_{\rho_k}^{\bar{\lambda}^k, c_k}(\cdot)$ on the level set, we have for each fixed k,

$$\begin{aligned}
&\left| G_{\rho_k}^{\bar{\lambda}^k, c_k}(\bar{z}_{s+1}^k) - G_{\rho_k}^{\bar{\lambda}^k, c_k}(z_s^k) - \nabla G_{\rho_k}^{\bar{\lambda}^k, c_k}(z_s^k)^T \left(\bar{z}_{s+1}^k - z_s^k \right) \right| \\
&= \left| \left(\nabla G_{\rho_k}^{\bar{\lambda}^k, c_k}(\theta \bar{z}_{s+1}^k + (1-\theta) z_s^k) - \nabla G_{\rho_k}^{\bar{\lambda}^k, c_k}(z_s^k) \right)^T \left(\bar{z}_{s+1}^k - z_s^k \right) \right| \\
&= o(\|\theta \bar{z}_{s+1}^k + (1-\theta) z_s^k - z_s^k\|) = o(\|\bar{z}_{s+1}^k - z_s^k\|),
\end{aligned}$$

for some $\theta \in [0, 1]$.

This equation guarantees that

$$\left| \frac{G_{\rho_k}^{\bar{\lambda}^k, c_k}(\bar{z}_{s+1}^k) - G_{\rho_k}^{\bar{\lambda}^k, c_k}(z_s^k)}{\nabla G_{\rho_k}^{\bar{\lambda}^k, c_k}(z_s^k)^T \left(\bar{z}_{s+1}^k - z_s^k \right)} - 1 \right| \to 0.$$

But this is impossible since from (2.12) and (2.16),

$$\nabla G_{\rho_k}^{\bar{\lambda}^k, c_k}(z_s^k)^T \left(\bar{z}_{s+1}^k - z_s^k \right) < 0$$

and hence

$$\frac{G_{\rho_k}^{\bar{\lambda}^k, c_k}(\bar{z}_{s+1}^k) - G_{\rho_k}^{\bar{\lambda}^k, c_k}(z_s^k)}{\nabla G_{\rho_k}^{\bar{\lambda}^k, c_k}(z_s^k)^T \left(\bar{z}_{s+1}^k - z_s^k \right)} < \sigma_2 < 1,$$

by (2.4). The contradiction proves that (2.14) holds.

Condition (2.14) implies that for any x^k, we can find some s such that condition (2.5) holds. Then $\lim_{k \to \infty} \rho_k = +\infty$ by the updating rule in Algorithm 2.1. ■

Theorem 2.1 *Let Assumption 2.1 hold and suppose that the Algorithm 2.1 does not terminate within finite iterations. Let x^* be an accumulation point of the sequence $\{x^k\}$ generated by Algorithm 2.1. If $\{c_k\}$ is bounded, then x^* is a* KKT *point of problem* (P).

Proof. Without loss of generality, assume that $\lim_{k \to \infty} x^k = x^*$. From the definition of the algorithm, the boundedness of $\{c_k\}$ is equivalent to saying that $\lim_{k \to \infty} \sigma_{\rho_k}^{\lambda^{k+1}}(x^{k+1}) = 0$. Thus it follows by the continuity of g_i and the fact that $\{g_\rho : \rho > 0\}$ is a family of smoothing functions of g_0 that $g_i(x^*) = 0,\ i = p+1, \cdots, q,\ g_i(x^*) \leq 0,\ i = 0, \cdots, p$. Let

$$\begin{aligned}
\mu_{\rho,0}^{\lambda,c}(x) &:= \max\{0, \lambda_0 + c g_\rho(x)\}, \\
\mu_{\rho,i}^{\lambda,c}(x) &:= \max\{0, \lambda_i + c g_i(x)\},\ i = 1, \cdots, p, \\
\mu_{\rho,i}^{\lambda,c}(x) &:= \lambda_i + c g_i(x),\ i = p+1, \cdots, q.
\end{aligned}$$

(i) Consider the case when there is a sequence $K_0 \subseteq \mathbf{N}$ such that (2.11) holds for all $k \in K_0$. From the definition of the projection and the normal cone, it is easy to see that, for each $k \in K_0$, x^{k+1} is a stationary point of $\min_{x\in\Omega} G_{\rho_k}^{\bar{\lambda}^k,c_k}(x)$, that is,

$$\begin{aligned} 0 &\in \nabla G_{\rho_k}^{\bar{\lambda}^k,c_k}(x^{k+1}) + \mathcal{N}_\Omega(x^{k+1}) \qquad (2.17)\\ &= \nabla G(x^{k+1}) + \sum_{i=1}^{q} \mu_{\rho_k,i}^{\bar{\lambda}^k,c_k}(x^{k+1})\nabla g_i(x^{k+1}) + \mu_{\rho_k,0}^{\bar{\lambda}^k,c_k}(x^{k+1})\nabla g_{\rho_k}(x^{k+1}) + \mathcal{N}_\Omega(x^{k+1}). \end{aligned}$$

From the definition of $\mu_\rho^{\lambda,c}(\cdot)$ and the updating rule (2.6) $-$ (2.8), we have $\mu_{\rho_k,i}^{\bar{\lambda}^k,c_k}(x^{k+1}) = \lambda_i^{k+1},\ i = 0,\cdots,q$. By the gradient consistent property of g_ρ, there exists a subsequence $\hat{K}_0 \subseteq K_0$ such that

$$\lim_{k\to\infty,\, k\in\hat{K}_0} \nabla g_{\rho_k}(x^{k+1}) \in \partial g_0(x^*).$$

Note that, by the boundedness of $\{\lambda^k\}$, there is a subsequence $\bar{K}_0 \subseteq \hat{K}_0$ such that $\{\lambda^k\}$ is convergent. Let $\lambda^* := \lim_{k\to\infty,\, k\in\bar{K}_0} \lambda^k$.

By letting $k \to \infty$ with $k \in \bar{K}_0$ in (2.17),

$$0 \in \nabla G(x^*) + \sum_{i=1}^{q} \lambda_i^* \nabla g_i(x^*) + \lambda_0^* \partial g_0(x^*) + \mathcal{N}_\Omega(x^*). \qquad (2.18)$$

It follows from (2.6) and (2.7) that $\lambda_i^* \geq 0, i = 0,\cdots,p$. We now show that the complementary slackness condition holds. Suppose that $g_i(x^*) < 0$ for certain $i \in \{1,\ldots,p\}$. Then by the continuity of $g_i(x)$ we have $g_i(x^{k+1}) < 0$. Thus $\lambda_i^* = \lim_{k\to\infty,\, k\in\bar{K}_0} \lambda_i^{k+1} = 0$ since $\lim_{k\to\infty} \sigma_{\rho_k}^{\lambda^{k+1}}(x^{k+1}) = 0$. Similarly by the fact that $\{g_\rho : \rho > 0\}$ is a family of smoothing functions of g_0, we conclude that $g_0(x^*) < 0$ implies $g_{\rho_k}(x^{k+1}) < 0$. Consequently we have $\lambda_0^{k+1} \to 0$ which implies that $\lambda_i^* = 0$ if $g_i(x^*) < 0$, for $i \in \{0,\cdots,p\}$. Therefore x^* is a KKT point of (P).

(ii) Consider the case when there is a sequence $K_1 \subseteq \mathbf{N}$ such that (2.11) fails for all $k \in K_1$. Then similar to [17, Lemma 3.3], one can show that for $k \in K_1$, there exists s_k such that $x^{k+1} = z_{s_k+1}^k$ and for each $x \in \Omega$,

$$\limsup_{k\to\infty} \nabla G_{\rho_k}^{\bar{\lambda}^k,c_k}(z_{s_k}^k)^T(z_{s_k}^k - x) \leq 0. \qquad (2.19)$$

We have from condition (2.5) that $\lim_{k\to\infty,\, k\in K_1} z_{s_k}^k = x^*$. By the gradient consistent property of g_ρ, there exists a subsequence $\hat{K}_1 \subseteq K_1$ such that

$$\lim_{k\to\infty,\, k\in\hat{K}_1} \nabla g_{\rho_k}(z_{s_k}^k) \in \partial g_0(x^*).$$

Note that, by the continuity of $g_i, i = 1, \cdots, q,\ g_{\rho_k}$ and condition (2.5), there exists $a_0, a_1 > 0$ such that

$$|g_i(z^k_{s_k}) - g_i(x^{k+1})| \leq a_0 |z^k_{s_k} - x^{k+1}| < \frac{a_1}{\rho_k},\ i = 1, \cdots, q,\ \rho_k. \tag{2.20}$$

Thus, by the definition of $\mu^{\lambda,c}_{\rho}(\cdot)$ and the definition of λ^{k+1} in (2.6)-(2.8) we have

$$\begin{aligned} \mu^{\bar{\lambda}^k, c_k}_{\rho_k, 0}(z^k_{s_k}) &\leq \max\{0, \bar{\lambda}^k_0 + c_k g_{\rho_k}(x^{k+1}) + \frac{a_1 c_k}{\rho_k}\} \leq \lambda^{k+1}_0 + \frac{a_1 c_k}{\rho_k}, & (2.21)\\ \mu^{\bar{\lambda}^k, c_k}_{\rho_k, i}(z^k_{s_k}) &\leq \max\{0, \bar{\lambda}^k_i + c_k g_i(x^{k+1}) + \frac{a_1 c_k}{\rho_k}\} \leq \lambda^{k+1}_i + \frac{a_1 c_k}{\rho_k},\ i = 1, \cdots, p, & (2.22)\\ \left|\mu^{\bar{\lambda}^k, c_k}_{\rho_k, i}(z^k_{s_k})\right| &\leq \left|\bar{\lambda}^k_i + c_k g_i(x^{k+1}) \pm \frac{a_1 c_k}{\rho_k}\right| \leq \left|\lambda^{k+1}_i \pm \frac{a_1 c_k}{\rho_k}\right|,\ i = p+1, \cdots, q. \end{aligned}$$

Since $\{c_k\}$ is bound while $\{\rho_k\}$ is unbounded by virtue of Lemma 2.3, $\{\mu^{\bar{\lambda}^k, c_k}_{\rho_k}(z^k_{s_k})\}$ is bounded. Hence, there is a subsequence $\bar{K}_1 \subseteq \hat{K}_1$ such that $\{\mu^{\bar{\lambda}^k, c_k}_{\rho_k}(z^k_{s_k})\}_{k \in \bar{K}_1}$ is convergent. Let $\bar{\mu} := \lim_{k \to \infty,\, k \in \bar{K}_1} \mu^{\bar{\lambda}^k, c_k}_{\rho_k}(z^k_{s_k})$. It follows from the definition of $\mu^{\lambda,c}_{\rho}(\cdot)$ that $\bar{\mu}_i \geq 0,\ i = 0, \cdots, p$.

On the other hand, let

$$V_k \ := \ \nabla G^{\bar{\lambda}^k, c_k}_{\rho_k}(z^k_{s_k}) = \nabla G(z^k_{s_k}) + \sum_{i=1}^{q} \mu^{\bar{\lambda}^k, c_k}_{\rho_k, i}(z^k_{s_k}) \nabla g_i(z^k_{s_k}) + \mu^{\bar{\lambda}^k, c_k}_{\rho_k, 0}(z^k_{s_k}) \nabla g_{\rho_k}(z^k_{s_k})$$

Then

$$V \ := \ \lim_{k \to \infty,\, k \in \bar{K}_1} V_k \in \nabla G(x^*) + \sum_{i=1}^{q} \bar{\mu}_i \nabla g_i(x^*) + \bar{\mu}_0 \partial g_0(x^*).$$

It follows from (2.19) that

$$V^T(x^* - x) \leq 0, \quad \forall x \in \Omega.$$

This means $-V \in \mathcal{N}_\Omega(x^*)$ and hence (2.18) holds with $\lambda^* = \bar{\mu}$. Now suppose that $g_0(x^*) < 0$. Then similarly as in the proof of part (i), we can show that $\lambda^{k+1}_0 \to 0$ which implies that $\mu^{\bar{\lambda}^k, c_k}_{\rho_k, 0}(z^k_{s_k}) \to 0$ and hence $\bar{\mu}_0 = 0$. Similarly we can show that $\bar{\mu}_i = 0$ if $g_i(x^*) < 0$, $i = 1, \cdots, p$. As a result, we always have $\bar{\mu}_i g_i(x^*) = 0,\ i = 0, \cdots, p$. Therefore x^* is a KKT point of (P). This completes the proof. ■

From the proof of part (i) and part (ii) in Theorem 2.1, we show that the stopping criteria in Step 4 of Algorithm 2.1 are reasonable since they lead to the KKT condition at any accumulation point.

The next theorem gives a sufficient condition for the boundedness of $\{c_k\}$.

Theorem 2.2 *Let Assumption 2.1 hold and suppose that the Algorithm 2.1 does not terminate within finite iterations. Let $\{x^k\}$ be a sequence generated by Algorithm 2.1. Suppose that $\lim_{k \to \infty} x^k = x^*$ and the* ENNAMCQ *holds at x^* for* (P). *Then $\{c_k\}$ is bounded.*

Proof. Assume for a contradiction that the conclusion is not true. Thus there exists a set $K \subseteq \mathbf{N}$ such that condition (2.9) fails for every $k \in K$ sufficiently large. Then there exist an index $i_1 \in \{0, \cdots, q\}$, such that when k is sufficient large one of the following holds:

$$\begin{aligned}
&(a) \text{ When } i_1 = 0, \ g_{\rho_k}(x^{k+1}) > \epsilon. \\
&(b) \text{ When } i_1 \in \{1, \cdots, p\}, \ g_{i_1}(x^{k+1}) > \epsilon. \\
&(c) \text{ When } i_1 \in \{p+1, \cdots, q\} \ |g_{i_1}(x^{k+1})| > \epsilon.
\end{aligned}$$

(i) First consider the case when there is a sequence $K_0 \subseteq K$ such that (2.11) holds for all $k \in K_0$. Similarly to the part (i) of Theorem 2.1, we know that condition (2.17) holds for every $k \in K_0$ with $\mu_{\rho_k,i}^{\bar{\lambda}^k,c_k}(x^{k+1}) = \lambda_i^{k+1}, \ i = 0, \cdots, q$.

By the gradient consistent property of g_ρ, there exists a subsequence $\hat{K_0} \subseteq K_0$ such that

$$v_0 = \lim_{k\to\infty,\, k\in\hat{K_0}} \nabla g_{\rho_k}(x^{k+1}) \in \partial g_0(x^*).$$

From the definition of $\mu_\rho^{\lambda,c}(\cdot)$, under any case of (a)-(c), we have $|\mu_{\rho_k,i_1}^{\bar{\lambda}^k,c_k}(x^{k+1})| \to \infty$. Thus $\|\mu_{\rho_k}^{\bar{\lambda}^k,c_k}(x^{k+1})\| \to \infty$. There exists a subsequence $\bar{K_0} \subseteq \hat{K_0}$ and $\mu_i \in \mathbb{R}, i = 0, \cdots, q$ not all equal to zero such that

$$\lim_{k\to\infty, k\in\bar{K_0}} \frac{\mu_{\rho_k,i}^{\bar{\lambda}^k,c_k}(x^{k+1})}{\|\mu_{\rho_k}^{\bar{\lambda}^k,c_k}(x^{k+1})\|} = \mu_i, \ i = 0, \cdots, q.$$

It follows from the definition of $\mu_\rho^{\lambda,c}(\cdot)$ that $\mu_i \geq 0, \ i = 0, \cdots, p$.

Dividing by $\|\mu_{\rho_k}^{\bar{\lambda}^k,c_k}(x^{k+1})\|$ in both sides of (2.17) and letting $k \to \infty$ in $\bar{K_0}$, we have

$$0 \in \sum_{i=1}^{q} \mu_i \nabla g_i(x^*) + \mu_0 v_0 + \mathcal{N}_\Omega(x^*) \subseteq \sum_{i=1}^{q} \mu_i \nabla g_i(x^*) + \mu_0 \partial g(x^*) + \mathcal{N}_\Omega(x^*). \quad (2.23)$$

Suppose that $g_i(x^*) < 0$ for certain $i \in \{1, \ldots, p\}$. Then by the continuity of $g_i(x)$ we have $g_i(x^{k+1}) < 0$. Thus $\mu_i = \lim_{k\to\infty,\, k\in\bar{K_0}} \mu_{\rho_k,i}^{\bar{\lambda}^k,c_k}(x^{k+1}) = 0$ by the definitions $\mu_\rho^{\lambda,c}(\cdot)$ and the unboundedness of $\{c_k\}$. Similarly by the fact that $\{g_\rho : \rho > 0\}$ is a family of smoothing functions of g_0 we conclude that $g_0(x^*) < 0$ implies $g_{\rho_k}(x^{k+1}) < 0$. Consequently we have $\mu_{\rho_k,0}^{\bar{\lambda}^k,c_k}(x^{k+1}) \to 0$ which implies that $\mu_i = 0$ if $g_i(x^*) < 0$, for $i \in \{0, \cdots, p\}$. Therefore (2.23) contradicts the ENNAMCQ assumption.

(ii) Now we consider the case where condition (2.11) fails for every $k \in K_1$ sufficiently large. Then (2.19) holds for $k \in K_1$. By the gradient consistent property of g_ρ, there exists a subsequence $\hat{K_1} \subseteq K_1$ such that

$$v := \lim_{k\to\infty, k\in\hat{K_1}} \nabla g_{\rho_k}(z_{s_k}^k) \in \partial g_0(x^*).$$

On the other hand, by the continuity of $g_i, i = 1, \cdots, q,\ g_{\rho_k}$ and condition (2.5), when k is sufficiently large, under any case (a)-(c), we have $g_{i_1}(z^k_{s_k}) > \frac{\epsilon}{2}$ for $i_1 \in \{1, \cdots, p,\ \rho_k\}$, $|g_{i_1}(z^k_{s_k})| > \frac{\epsilon}{2}$ for $i_1 \in \{p+1, \cdots, q\}$. Thus $\mu^{\lambda^k, c_k}_{\rho_k, i_1}(z^k_{s_k})$ is unbounded. Therefore, $\|\mu^{\bar{\lambda}^k, c_k}_{\rho_k}(z^k_{s_k})\| \to +\infty$ as $\hat{K}_1 \ni k \to \infty$. There exists a subsequence $\bar{K}_1 \subseteq \hat{K}_1$ and $\mu_i \in \mathbb{R}, i = 0, \cdots, q$ not all equal to zero such that

$$\lim_{k \to \infty, k \in \bar{K}_1} \frac{\mu^{\bar{\lambda}^k, c_k}_{\rho_k, i}(z^k_{s_k})}{\|\mu^{\bar{\lambda}^k, c_k}_{\rho_k}(z^k_{s_k})\|} = \mu_i,\ i = 0, \cdots, q$$

and $\mu_i \geq 0,\ i = 0, \cdots, p$ from the definition of $\mu^{\lambda, c}_{\rho}(\cdot)$.

Note that for any $x \in \Omega$ and $k \in \bar{K}_1$,

$$\frac{\nabla G^{\bar{\lambda}^k, c_k}_{\rho_k}(z^k_{s_k})}{\|\mu^{\bar{\lambda}_k, c_k}_{\rho_k}(z^k_{s_k})\|} = \frac{\nabla G(z^k_{s_k})}{\|\mu^{\bar{\lambda}^k, c_k}_{\rho_k}(z^k_{s_k})\|} + \sum_{i=1}^{q} \frac{\mu^{\bar{\lambda}^k, c_k}_{\rho_k, i}(z^k_{s_k})}{\|\mu^{\bar{\lambda}^k, c_k}_{\rho_k}(z^k_{s_k})\|} \nabla g_i(z^k_{s_k}) + \frac{\mu^{\bar{\lambda}^k, c_k}_{\rho_k, 0}(z^k_{s_k})}{\|\mu^{\bar{\lambda}^k c_k}_{\rho_k}(z^k_{s_k})\|} \nabla g_{\rho_k}(z^k_{s_k}),$$

dividing by $\|\mu^{\bar{\lambda}_k, c_k}_{\rho_k}(z^k_{s_k})\|$ in both sides of (2.19) and taking a limit within $\bar{K}_1$, we have

$$\left(\sum_{i=1}^{q} \mu_i \nabla g_i(x^*) + \mu_0 v \right)^T (x^* - x) \leq 0,$$

which means

$$0 = \sum_{i=1}^{q} \mu_i \nabla g_i(x^*) + \mu_0 v + \mathcal{N}_\Omega(x^*) \subseteq \sum_{i=1}^{q} \mu_i \nabla g_i(x^*) + \mu_0 \partial g(x^*) + \mathcal{N}_\Omega(x^*).$$

It remains to show that the complementary slackness condition holds. Suppose that $g_i(x^*) < 0$ for $i \in \{0, \cdots, p\}$. Then we have $g_{\rho_k}(z^k_{s_k}) < 0$ and $g_i(z^k_{s_k}) < 0$ for sufficiently large k, thus $\mu^{\bar{\lambda}^k, c_k}_{\rho_k, i}(z^k_{s_k}) \to 0$. Thus if $g_i(x^*) < 0$ for $i = 0, \cdots, p$, $\mu_i = 0$. This contradicts the ENNAMCQ assumption. From the above discussion, we know that $\{c_k\}$ is bounded.

■

The next corollary follows immediately from Theorems 2.1 and 2.2.

Corollary 2.1 *Let Assumption 2.1 hold and suppose that the Algorithm 2.1 does not terminate within finite iterations. Let x^* be a limiting point of the sequence $\{x^k\}$ generated by Algorithm 2.1. If the* ENNAMCQ *holds at x^*, then x^* is a* KKT *point of problem* (P)*.*

To derive the convergence result for any accumulation point, one needs to assume the ENNAMCQ holds for every infeasible point $x \in \Omega$ as shown in the following theorem.

Theorem 2.3 *Let Assumption 2.1 hold and suppose that the Algorithm 2.1 does not terminate within finite iterations. Let $\{x^k\}$ be a sequence generated by Algorithm 2.1. Assume that the* ENNAMCQ *holds for* (P) *at any infeasible point $x \in \Omega$. If $\{x^k\}$ is bounded, then $\{c_k\}$ is bounded and hence any accumulation point of $\{x^k\}$ is a* KKT *point of problem* (P)*.*

Proof. Suppose to the contrary that the sequence $\{c_k\}$ is unbounded. Let x^* be an accumulation point of $\{x^k\}$. Then there must exist an index $i_0 = 0, \cdots, p$ such that $g_{i_0}(x^*) \geq \epsilon$ or $i_0 = p+1, \cdots, q$ $|g_{i_0}(x^*)| \geq \epsilon$. Hence, by the assumption, the ENNAMCQ holds at x^* for (P). On the other hand, similarly as in the proof of Theorem 2.2, we can show that the ENNAMCQ fails at x^*. As a contradiction, we have shown that $\{c_k\}$ is bounded.

The second assertion follows from the boundedness of $\{c_k\}$ and Theorem 2.1 immediately. ■

3 Smoothing augmented Lagrangian algorithm for simple bilevel programs

In this section, we apply Algorithm 2.1 to the problem (CP). Since Y is assumed to be compact, Assumption 2.1 is satisfied automatically. Under Assumption 1.1, problem (CP) takes the form:

$$\begin{array}{lll} \text{(CP)} & \min & F(x,y) \\ & \text{s.t.} & f(x,y) - V(x) \leq 0, \\ & & \nabla_y f(x,y) = 0, \\ & & g_i(x,y) \leq 0, \ i = m+1, \cdots, l, \\ & & (x,y) \in X \times Y. \end{array}$$

Recently Lin, Xu and Ye [17] have shown that the integral entropy function (1.4) is a smoothing function of the value function and the gradient consistent property holds. Hence the new algorithm introduced in Section 2 is applicable to problem (CP). For given $\rho > 0, c > 0$ and $\lambda \in \mathbb{R}_+^{l+1-m} \times \mathbb{R}^m$, define the augmented Lagrange function:

$$\begin{aligned} G_\rho^{\lambda,c}(x,y) := {} & F(x,y) + \frac{1}{2c}\left(\max\{0, \lambda_0 + c(f(x,y) - \gamma_\rho(x))\}^2 - \lambda_0^2\right) \\ & + \sum_{j=1}^m \left(\lambda_j \nabla_{y_j} f(x,y) + \frac{c}{2}(\nabla_{y_j} f(x,y))^2\right) + \frac{1}{2c}\sum_{i=m+1}^l \left(\max\{0, \lambda_i + c g_i(x,y)\}^2 - \lambda_i^2\right) \end{aligned} \tag{3.1}$$

and the residual function:

$$\begin{aligned} \sigma_\rho^\lambda(x,y) := \max \ \{ \ & |\min\{\lambda_0, \gamma_\rho(x) - f(x,y)\}|, \ \ |\nabla_{y_j} f(x,y)|, j = 1, \cdots, m, \\ & |\min\{\lambda_i, -g_i(x,y)\}|, \ i = m+1, \cdots, l \ \ \}. \end{aligned}$$

The algorithm can be stated as follows:

Algorithm 3.1 *Let $\{\beta, \gamma, \sigma_1, \sigma_2\}$ be constants in $(0,1)$ with $\sigma_1 \le \sigma_2$, $\epsilon \ge 0$, $\epsilon_1 \ge 0$ be very small constants, $\{\sigma, \sigma', \hat{\eta}\}$ be constants in $(1, \infty)$. Choose an initial point $(x^0, y^0) \in X \times Y$, an initial smoothing parameter $\rho_0 > 0$, an initial penalty parameter $c_0 > 0$, an initial multiplier $\bar{\lambda}^0 \in [0, \lambda_{max}] \times \bigotimes_{j=1}^m [\lambda_{min}, \lambda_{max}] \times \bigotimes_{i=m+1}^l [0, \lambda_{max}]$, where $\lambda_{min} < 0$ and $\lambda_{max} > 0$ are given constants and set $k := 0,\ s := 0$.*

1. *Let $z_0^k := (x^k, y^k)$ and $z_{s+1}^k := P_{X \times Y}[z_s^k - \alpha_s \nabla G_{\rho_k}^{\bar{\lambda}^k, c_k}(z_s^k)]$, where $\alpha_s := \beta^{l_s}$, $l_s \in \{0, 1, 2 \cdots\}$ is the smallest number satisfying*

$$G_{\rho_k}^{\bar{\lambda}^k, c_k}(z_{s+1}^k) - G_{\rho_k}^{\bar{\lambda}^k, c_k}(z_s^k) \le \sigma_1 \nabla G_{\rho_k}^{\bar{\lambda}^k, c_k}(z_s^k)^T \left(z_{s+1}^k - z_s^k\right) \tag{3.2}$$

and $\beta^{l_s} \ge \gamma$, or $\bar{\alpha}_s := \beta^{l_s - 1}$ such that $\bar{z}_{s+1}^k := P_{X \times Y}[z_s^k - \bar{\alpha}_s \nabla G_{\rho_k}^{\bar{\lambda}^k, c_k}(z_s^k)]$ satisfies

$$G_{\rho_k}^{\bar{\lambda}^k, c_k}(\bar{z}_{s+1}^k) - G_{\rho_k}^{\bar{\lambda}^k, c_k}(z_s^k) > \sigma_2 \nabla G_{\rho_k}^{\bar{\lambda}^k, c_k}(z_s^k)^T \left(\bar{z}_{s+1}^k - z_s^k\right). \tag{3.3}$$

If

$$\frac{\|z_{s+1}^k - z_s^k\|}{\alpha_s} < \hat{\eta} \rho_k^{-1}, \tag{3.4}$$

set $(x^{k+1}, y^{k+1}) := z_{s+1}^k$, $\rho_{k+1} := \sigma \rho_k$, $s := 0$, and go to Step 2. Otherwise, set $s = s + 1$ and go to Step 1.

2. *Set*

$$\lambda_0^{k+1} = \max\{0, \bar{\lambda}_0^k + c_k(f(x^{k+1}, y^{k+1}) - \gamma_{\rho_k}(x^{k+1})\}, \tag{3.5}$$

$$\lambda_j^{k+1} = \bar{\lambda}_j^k + c_k \nabla_{y_j} f(x^{k+1}, y^{k+1}),\ j = 1, \cdots, m, \tag{3.6}$$

$$\lambda_i^{k+1} = \max\{0, \bar{\lambda}_i^k + c_k g_i(x^{k+1}, y^{k+1})\},\ i = m+1, \cdots, l. \tag{3.7}$$

Take $\bar{\lambda}^{k+1}$ as the Euclidean projection of λ^{k+1} onto $[0, \lambda_{max}] \times \bigotimes_{j=1}^m [\lambda_{min}, \lambda_{max}] \times \bigotimes_{i=m+1}^l [0, \lambda_{max}]$, and go to Step 3.

3. *If*

$$\sigma_{\rho_k}^{\lambda^{k+1}}(x^{k+1}, y^{k+1}) \le \epsilon, \tag{3.8}$$

go to Step 4. Else if $k = 0$ or

$$\sigma_{\rho_k}^{\lambda^{k+1}}(x^{k+1}, y^{k+1}) \le \gamma \sigma_{\rho_{k-1}}^{\lambda^k}(x^k, y^k), \tag{3.9}$$

set $k = k + 1$ and go to Step 1. Otherwise, set $c_{k+1} := \sigma' c_k$, $k = k + 1$ and go to Step 1.

4. If

$$\|P_{X\times Y}[(x^{k+1},y^{k+1})-\nabla G_{\rho_k}^{\bar{\lambda}^k,c_k}(x^{k+1},y^{k+1})]-(x^{k+1},y^{k+1})\|=0, \tag{3.10}$$

or $\|(x^{k+1},y^{k+1})-(x^k,y^k)\|\le\epsilon_1$, *terminate. Otherwise, set* $k=k+1$, *and go to Step 1.*

From Theorem 2.1, we have the following convergence result immediately.

Theorem 3.1 *Suppose that the Algorithm 3.1 does not terminate within finite iterations. If* (x^*,y^*) *is an accumulation point of* $\{(x^k,y^k)\}$ *which is the sequence generated by Algorithm 3.1 and the sequence* $\{c_k\}$ *is bounded, then* (x^*,y^*) *is a* KKT *point of problem* (CP).

The convergence results shown in Theorems 2.2 and 2.3 require the ENNAMCQ holds at the limiting point. Unfortunately, the combined program (CP) will never satisfy the ENNAMCQ [17] unless the lower level problem is replaced by an approximate problem. However, the problem (CP) is very likely to satisfy the partial calmness or the weakly calmness condition (see [32]). Hence the sequence $\{c_k\}$ is likely to be bounded. Therefore any accumulation point of the iteration sequence is likely to be a KKT point by virtue of Theorem 3.1.

In Proposition 2.1, we gave a condition to verify that a feasible point is a KKT point. We now specialize it to give a condition using which we can verify that an accumulation point is a KKT point. The following result follows from Proposition 2.1.

Proposition 3.1 *Let* $\{(x^k,y^k)\}$ *be a sequence generated by Algorithm 3.1. Suppose that there is a subsequence* K *such that* $\lim_{k\to\infty,k\in K}(x^k,y^k)=(x^*,y^*)$ *and*

$$v:=\lim_{k\to\infty,k\in K}\nabla\gamma_{\rho_k}(x^{k+1}).$$

If $\lim_{k\to\infty,k\in K}\sigma_{\rho_k}^{\lambda^{k+1}}(x^{k+1},y^{k+1})=0$ *and*

$$\nabla F(x^*,y^*)^Td\ge 0$$

for all d *in the linearization cone*

$$\begin{aligned}\mathcal{L}(x^*,y^*):\ &=\ \{d\in\mathcal{T}_X(x^*)\times\mathbb{R}^m:(\nabla f(x^*,y^*)-(v,0))^Td\le 0,\\ &\qquad \nabla(\nabla_{y_j})f(x^*,y^*)^Td=0,\ j=1,\dots,m,\ \nabla g_i(x^*,y^*)^Td\le 0,\ i\in I(x^*,y^*)\}\end{aligned}$$

where $I(x^*,y^*)=\{i=m+1,\cdots,l:g_i(x^*,y^*)=0\}$, *then* (x^*,y^*) *is a* KKT *point of* (CP).

Proof. $\lim\limits_{k\to\infty, k\in K} \sigma_{\rho_k}^{\lambda^{k+1}}(x^{k+1}, y^{k+1}) = 0$ implies that (x^*, y^*) is a feasible point of problem (CP) and hence $y^* \in S(x^*)$. From Assumption 1.1, y^* is an interior point of Y. Thus, $\mathcal{N}_Y(y^*) = \{0\}^m$. Since by the gradient consistent property of $\{\gamma_\rho : \rho > 0\}$, $v \in \partial V(x^*)$. Followed from Proposition 2.1, there exist multipliers μ_i, $i = 0, \cdots, l$ such that

$$\begin{aligned}
&0 \in \nabla F(x^*, y^*) + \sum_{j=1}^{m} \mu_j \nabla(\nabla_{y_j} f(x^*, y^*)) + \sum_{i=m+1}^{l} \mu_i \nabla g_i(x^*, y^*) \\
&\quad + \mu_0(\nabla f(x^*, y^*) - \partial V(x^*) \times \{0\}^m) + \mathcal{N}_X(x^*) \times \{0\}^m. \\
&\mu_0 \geq 0,\ \ \mu_i \geq 0,\ \ i \in I(x^*, y^*), \\
&\mu_i = 0,\ \ i = m+1, \cdots, l, i \notin I(x^*, y^*),
\end{aligned}$$

Therefore (x^*, y^*) is a KKT point of problem (CP). ■

4 Numerical examples

We first test Algorithm 3.1 on the following two simple bilevel programs.

Example 4.1 (Mirrlees' problem) [20] Consider Mirrlees' problem

$$\begin{aligned}
\min \quad & F(x, y) := (x-2)^2 + (y-1)^2 \\
\text{s.t.} \quad & y \in S(x),
\end{aligned}$$

where $S(x)$ is the solution set of the lower level program

$$\begin{aligned}
\min \quad & f(x, y) := -x \exp[-(y+1)^2] - \exp[-(y-1)^2] \\
\text{s.t.} \quad & y \in [-2, 2].
\end{aligned}$$

Table 1: Mirrlees' problem

ρ_0	(x^{k+1}, y^{k+1})	ρ_0	(x^{k+1}, y^{k+1})
10	(1, 0.95735)	100	(1, 0.95769)
30	(1, 0.95753)	150	(1, 0.95751)
50	(1, 0.95772)	180	(1, 0.95741)
70	(1, 0.95754)	200	(1, 0.95754)

In our test, we chose the initial point $(x^0, y^0) = (0.5, 0.5)$ and the parameters $\beta = 0.7,\ \gamma = 0.5,\ \sigma_1 = \sigma_2 = 10^{-6},\ \ c_0 = 100,\ \bar{\lambda}^0 = (100, 100),\ \hat{\eta} = 5 * 10^5,\ \sigma = \sigma' = 10$. To illustrate the influence of the initial smoothing parameter ρ_0, we select different ρ_0 in the test (see Table 1).

To show that our algorithm leading to a KKT point of the problem, we verify the case when $\rho_0 = 30$. Since the stopping criteria $\sigma_{\rho_k}^{\lambda^{k+1}}(x^{k+1}, y^{k+1}) \leq 5 * 10^{-5}$ and $\|(x^{k+1}, y^{k+1}) - (x^k, y^k)\| \leq 10^{-12}$ hold, we terminate at the 5th iteration. We obtain an accumulation point $(x^*, y^*) \approx (1, 0.95753)$.

Since

$$\begin{aligned}
&\nabla f(x^*, y^*) - (\lim_{k\to\infty} \nabla \gamma_{\rho_k}(x^{k+1}), 0) \approx (0.000107, 0.00004439),\\
&\nabla(\nabla_y f)(x^*, y^*) \approx (0.084831, 1.70041),
\end{aligned}$$

the linearization cone

$$\mathcal{L}(x^*, y^*) \quad \approx \quad \{d \in \mathbb{R}^2 : d = \alpha(-9.5417, 0.47602), \alpha \in \mathbb{R}_+\}.$$

It follows that for any $d \in \mathcal{L}(x^*, y^*)$,

$$\nabla F(x^*, y^*)^T d \geq 0.$$

Indeed if to the contrary that there exists $\hat{d} \in \mathcal{L}(x^*, y^*)$ such that

$$\nabla F(x^*, y^*)^T \hat{d} < 0.$$

Then since $\nabla F(x^*, y^*) \approx (-2, -0.0849396)$ and there exists $\hat{\alpha} > 0$ such that $\hat{d} \approx \hat{\alpha}(-9.5417, 0.47602)$, we have

$$(-2, -0.0849396) \cdot \hat{\alpha}(-9.5417, 0.47602) < 0$$

which is a contradiction since

$$(-2, -0.0849396) \cdot (-9.5417, 0.47602) = 19.043 > 0.$$

Therefore, by Proposition 3.1, (x^*, y^*) is a KKT point of problem (CP) for Mirrlees' problem. In fact it is the unique global minimum [20]. In [17], it was shown that the smoothing projected gradient algorithm fails for the Mirrlees problem but succeeds to find ε solutions. Hence the approach taken in this paper is better than the one in [17] in that the solution for the original problem is found.

Example 4.2 [21, Example 5.2]

$$\begin{aligned}
\min \quad & F(x, y) := x^2 - y\\
\text{s.t.} \quad & y \in \underset{y\in Y}{\operatorname{argmin}}\{f(x, y) := ((y - 1 - 0.1x)^2 - 0.5 - 0.5x)^2\},\\
& x \in X := [0, 1],\ y \in Y := [0, 3].
\end{aligned}$$

Mitsos et al. [21] found an approximate optimal solution for the problem to be $(\bar{x}, \bar{y}) = (0.2106, 1.799)$.

In our numerical experiment, we chose the initial point $(x^0, y^0) = (0, 0)$, the parameters $\beta = 0.9,\ \gamma = 0.5,\ \sigma_1 = \sigma_2 = 10^{-6},\ \rho_0 = 100,\ c_0 = 100,\ \bar{\lambda}^0 = (100, 100),\ \hat{\eta} = 2 * 10^5,\ \sigma = \sigma' = 10$. To illustrate the influence of the initial smoothing parameter ρ_0, we select different ρ_0 in the test (see Table 2).

Table 2: Example 2

ρ_0	(x^{k+1}, y^{k+1})	ρ_0	(x^{k+1}, y^{k+1})
30	(0.2054,1.7968)	75	(0.2047,1.7966)
50	(0.2054,1.7969)	100	(0.2027,1.7957)

To show that our algorithm leading to a KKT point of the problem, we verify the case when $\rho_0 = 100$. Since the stopping criteria $\sigma_{\rho_k}^{\lambda^{k+1}}(x^{k+1}, y^{k+1}) \leq 10^{-7}$ and $\|(x^{k+1}, y^{k+1}) - (x^k, y^k)\| \leq 10^{-12}$ hold, we terminate at the 8th iteration.

We obtain an accumulation point $(x^*, y^*) \approx (0.2027, 1.7957)$. Since

$$\begin{aligned} &\nabla f(x^*, y^*) - (\lim_{k \to \infty} \nabla \gamma_{\rho_k}(x^{k+1}), 0) \approx (-0.00017, 0), \\ &\nabla(\nabla_y f)(x^*, y^*) \approx (-2.032062, 4.81098), \end{aligned}$$

the linearization cone

$$\mathcal{L}(x^*, y^*) \quad \approx \quad \{d \in \mathbb{R}^2 : d = \alpha(5.753397, 2.43012), \alpha \in \mathbb{R}_+\}.$$

Since $\nabla F(x^*, y^*) \approx (3.5915, -1)$ and

$$(3.5915, -1) \cdot \alpha(5.753397, 2.43012) = 18.23\alpha > 0, \quad \forall \alpha \in \mathbb{R}_+.$$

It follows that for any $d \in \mathcal{L}(x^*, y^*)$,

$$\nabla F(x^*, y^*)^T d \geq 0.$$

Therefore, by Proposition 3.1, (x^*, y^*) is a KKT point of the problem (CP). Comparing our result with the result given by Mitsos et al. [21], our solution gives a lower objective function value.

The moral-hazard problem in Economics arises when a principal (leader) hires an agent (follower) to perform certain task in situations in which the principal can neither observe nor verify the agent's action (see e.g. [20]). The principal offers a contract to the agent. The agent will only accept the offer if it gives him a payoff that is not smaller than the minimal acceptable payoff. The agent takes an action to maximize his payoff that affects the principal's payoff as well. The principal now chooses a contract which is acceptable to the agent so as to maximize his payoff. We consider a moral-hazard model in which the agent chooses an action y from a interval $Y = [\underline{y}, \overline{y}]$. The outcome can be one of the N given alternatives $o_1, o_2, \cdots, o_N$. However, the probability for the agent to generate the outcome o_i is $p_i(y)$ when he takes action $y \in Y$. Let $x = (x_1. \cdots, x_N) \in \mathbb{R}^N$ denote a contract, where x_i is the money paid to the agent if the outcome o_i occurs.

The agent's and the principal's utilities are denoted by $v(x_i)$ and $u(o_i - x_i)$. $c(y)$ is the cost function of the agent's action y. The minimal acceptable payoff of the agent is X^*. The expected payoffs to the principal and agent with a contract x when the agent takes action y are as follows:

$$U_p(x, y) = \sum_{i=1}^{N} p_i(y) u(o_i - x_i),$$
$$U_a(x, y) = \sum_{i=1}^{N} p_i(y) v(x_i) - c(y).$$

Therefore, a mathematical program for finding an optimal deterministic contract (x^*, y^*) is

$$\begin{array}{lll} \text{(PA)} \qquad \max & U_p(x, y) \\ \text{s.t.} & y \in \arg\max_{y' \in Y} U_a(x, y'), \\ & U_a(x, y) \geq X^*. \end{array}$$

The principal-agent problem (PA) is a special case of the bilevel program (SBP). We now use Algorithm 3.1 to solve the following principal-agent problem.

Example 4.3 [18, Example 3.1] Consider a principal-agent problem with two possible outcomes $o_1 = 150, o_2 = 300$, $Y = [0.1, 3]$, $p_1(y) = 0.5^y, p_2(y) = 1 - 0.5^y$, $X^* = 5$, $u(o - x) = o - x$, $v(x) = \sqrt{x}$ and $c(y) = y$. Then the problem can be written as the following simple bilevel program:

$$\begin{array}{ll} \min & F(x, y) := -U_p(x, y) = 0.5^y(x_1 - 150) + (1 - 0.5^y)(x_2 - 300) \\ \text{s.t.} & g_1(x, y) := -U_a(x, y) + X^* = 5 - 0.5^y\sqrt{x_1} - (1 - 0.5^y)\sqrt{x_2} + y \leq 0, \\ & y \in \operatorname*{argmin}_{y \in Y}\{f(x, y) := -U_a(x, y) = -0.5^y\sqrt{x_1} - (1 - 0.5^y)\sqrt{x_2}\}, \\ & y \in Y. \end{array}$$

In [18], the author used the first order approach to obtain $\bar{x} = (3.0416, 75.9576)$, $\bar{y} = 2.2727$.

In our numerical experiment, we chose the initial point $x^0 = (5, 75)$, $y^0 = 2$ and the parameters $\beta = 0.9$, $\gamma = 0.5$, $\sigma_1 = \sigma_2 = 10^{-6}$, $c_0 = 100$, $\bar{\lambda}^0 = (100, 100, 100)$, $\hat{\eta} = 5 * 10^5$, $\sigma = \sigma' = 10$. To illustrate the influence of the initial smoothing parameter ρ_0, we select different ρ_0 in the test (see Table 3).

To show that our algorithm leading to a KKT point of the problem, we verify the case when $\rho_0 = 500$. Since the stopping criteria $\sigma_{\rho_k}^{\lambda^{k+1}}(x^{k+1}, y^{k+1}) \leq 5{*}10^{-6}$ and $\|(x^{k+1}, y^{k+1}) - (x^k, y^k)\| \leq 10^{-12}$ hold, we terminate at the 8th iteration. We obtain an accumulation point $(x^*, y^*) \approx (3.04536, 75.9527, 2.2724)$.

Table 3: Example 3

ρ_0	(x^{k+1}, y^{k+1})	ρ_0	(x^{k+1}, y^{k+1})
30	(3.04537,75.9528,2.2724)	100	(3.04536,75.9528,2.2724)
50	(3.04536,75.9527,2.2724)	300	(3.04536,75.9527,2.2724)
80	(3.04536,75.9528,2.2724)	500	(3.04556,75.9527,2.2724)

Since $g_1(x^*, y^*) \approx -1.36 * 10^{-6}$ (which can be consider as 0) and

$$\begin{aligned}
&\nabla f(x^*, y^*) - (\lim_{k\to\infty} \nabla\gamma_{\rho_k}(x^{k+1}), 0) \approx 1.0 * 10^{-5}(-0.5707, -0.0705, -0.4503),\\
&\nabla(\nabla_y f)(x^*, y^*) \approx (0.0411074, -0.0082313, 0.6932),\\
&\nabla g_1(x^*, y^*) \approx (-0.05930548, -0.04549653, -0.0000045),
\end{aligned}$$

the linearization cone

$$\begin{aligned}
\mathcal{L}(x^*, y^*) \approx \{&d \in \mathbb{R}^3 : \alpha_1, \alpha_2 \in \mathbb{R}_+,\\
&d = \alpha_1(2.246732, -2.928636, -0.16802) + \alpha_2(-0.374549, 2.6862, 0.0541118)\}.
\end{aligned}$$

Since $\nabla F(x^*, y^*) \approx (0.207, 0.7930, -11.0607)$ and

$$\begin{aligned}
&(0.207, 0.7930, -11.0607) \cdot \alpha_1(2.246732, -2.928636, -0.16802)\\
&+(0.207, 0.7930, -11.0607) \cdot \alpha_2(-0.374549, 2.6862, 0.0541118)\\
&= 0.0011\alpha_1 + 1.45\alpha_2 \geq 0, \qquad\qquad \forall \alpha_1, \alpha_2 \in \mathbb{R}_+.
\end{aligned}$$

It follows that for any $d \in \mathcal{L}(x^*, y^*)$,

$$\nabla F(x^*, y^*)^T d \geq 0.$$

Therefore, by Proposition 3.1, (x^*, y^*) is a KKT point for problem (CP).

The numerical examples show that taking different initial parameter ρ_0 leads to similar results. According to our experience, if a very large initial parameter ρ_0 is taken, one should select a smaller σ to let ρ_k goes to infinity slowly. We suggest to choose a smaller ρ_0 since otherwise the convergence rate may be slower.

References

[1] ALGENCAN. http://www.ime.usp.br/ ∼egbirgin/tango/.

[2] J.F. Bard, Practical Bilevel Optimization: Algorithms and Applications, Kluwer Academic Publications, Dordrecht, Netherlands, 1998.

[3] P.H. Calamai and J.J. Moré, *Projected gradient method for linearly constrained problems*, Math. Program., **39**(1987), 93–116.

[4] X. Chen, *Smoothing methods for nonsmooth, nonconvex optimization*, Math. Program., **134**(2012), 71–99.

[5] X. Chen, R.S. Womersley and J.J. Ye, *Minimizing the condition number of a gram matrix*, SIAM J. Optim., **21**(2011), 127–148.

[6] F.H. Clarke, Optimization and Nonsmooth Analysis, Wiley-Interscience, New York, 1983.

[7] F.H. Clarke, Yu.S. Ledyaev, R.J. Stern and P.R. Wolenski, Nonsmooth Analysis and Control Theory, Springer, New York, 1998.

[8] J.V. Burke, A.S. Lewis and M.L. Overton, *A robust gradient sampling algorithm for nonsmooth, nonconvex optimization*, SIAM J. Optim., **15**(2005), 751–779.

[9] J.M. Danskin, The Theory of Max-Min and its Applications to Weapons Allocation Problems, Springer, New York, 1967.

[10] S. Dempe, Foundations of Bilevel Programming, Kluwer Academic Publishers, 2002.

[11] S. Dempe, *Annotated bibliography on bilevel programming and mathematical programs with equilibrium constraints*, Optim., **52**(2003), 333–359.

[12] J.C. Dunn, *Global and asymptotic convergence rate estimates for a class of projected gradient processes*, SIAM J. Contr. Optim., **19**(1981), 368–400.

[13] E.M. Gafni and D.P. Bertsekas, *Two-metric projection methods for constrained optimization*, SIAM J. Contr. and Optim., **22**(1984), 936–964.

[14] M.R. Hestenes, *Multiplier and gradient methods*, J. Optim. Theo. and Appl., **4**(1969), 303–320.

[15] A. Jourani, *Constraint qualifications and Lagrange multipliers in nondifferentiable programming problems*, J. Optim. Theo. and Appl., **81**(1994), 533–548.

[16] LANCELOT. http://www.cse.scitech.ac.uk/nag/lancelot/lancelot.shtml.

[17] G.H. Lin, M. Xu and J.J. Ye, *On solving simple bilevel programs with a nonconvex lower level program*, Math. Program., series A, DOI 10.1007/s10107-013-0633–4.

[18] B. Liu, *The mathematics of Principal-Agent Problems*, MSc. Thesis, University of Victoria, 2008. 10.1007/s10107–013–0633–4.

[19] Z.-Q. Luo, J.-S. Pang and D. Ralph, Mathematical Programs with Equilibrium Constraints. Cambridge University Press, Cambridge, (1996).

[20] J. Mirrlees, *The theory of moral hazard and unobservable behaviour: Part I*, Rev. Econ. Stud., **66**(1999), 3–22.

[21] A. Mitsos, P. Lemonidis and P. Barton, *Global solution of bilevel programs with a nonconvex inner program*, J. Global Optim., **42**(2008), 475–513.

[22] J.V. Outrata, *On the numerical solution of a class of Stackelberg problems*, Z. Oper. Res., **34**(1990), 255–277.

[23] J.-S. Pang and M. Fukushima, *Complementarity constraint qualifications and simplified B-stationary conditions for mathematical programs with equilibrium constraints.* Comput. Optim. Appl., **13** (1999), 111–136.

[24] M.J.D. Powell, *A method for nonlinear constraints in minimization problems*, in Optimization, R. Fletcher, ed., Academic Press, New York, (1969), pp. 283–298.

[25] R.T. Rockafellar, *A dual approach for solving nonlinear programming problems by unconstrained optimization*, Math. Program., **5**(1973), pp. 354–373.

[26] R.T. Rockafellar, *Augmented Lagrange multiplier functions and duality in nonconvex programming*, SIAM J. Contr. Optim., **12**(1974), pp. 268–285.

[27] K. Shimizu, Y. Ishizuka and J.F. Bard, Nondifferentiable and Two-Level Mathematical Programming, Kluwer Academic Publishers, Boston, 1997.

[28] L.N. Vicente and P.H. Calamai, *Bilevel and multilevel programming: A bibliography review*. J. Global Optim., **5**(1994), 291–306.

[29] J.J. Ye, *Necessary and Sufficient conditions for Mathematical Programs with Equilibrium Constraints* J. Math. Anal. Appl., **307** (2005), 350–369.

[30] J.J. Ye and D.L. Zhu, *Optimality conditions for bilevel programming problems*, Optim., **33**(1995), 9–27.

[31] J.J. Ye and D.L. Zhu, *A note on optimality conditions for bilevel programming problems*, Optim., **39**(1997), 361–366.

[32] J.J. Ye and D.L. Zhu, *New necessary optimality conditions for bilevel programs by combining MPEC and the value function approach*, SIAM J. Optim., **20**(2010), 1885–1905.

[33] E.H. Zarantonello, Contributions to Nonlinear Functional Analysis, Academic Press, New York, 1971.

[34] C. Zhang and X. Chen, *Smoothing projected gradient method and its application to stochastic linear complementarity problems*, SIAM J. Optim., **20**(2009), 627–649.