On solving simple bilevel programs with a nonconvex lower level program*

Gui-Hua Lin†, Mengwei Xu‡ and Jane J. Ye§

December 2011, Revised September 2012

Abstract. In this paper, we consider a simple bilevel program where the lower level program is a nonconvex minimization problem with a convex set constraint and the upper level program has a convex set constraint. By using the value function of the lower level program, we reformulate the bilevel program as a single level optimization problem with a nonsmooth inequality constraint and a convex set constraint. To deal with such a nonsmooth and nonconvex optimization problem, we design a smoothing projected gradient algorithm for a general optimization problem with a nonsmooth inequality constraint and a convex set constraint. We show that, if the sequence of penalty parameters is bounded then any accumulation point is a stationary point of the nonsmooth optimization problem and, if the generalized sequence is convergent and the extended Mangasarian-Fromovitz constraint qualification holds at the limit then the limit point is a stationary point of the nonsmooth optimization problem. We apply the smoothing projected gradient algorithm to the bilevel program if a calmness condition holds and to an approximate bilevel program otherwise. Preliminary numerical experiments show that the algorithm is efficient for solving the simple bilevel program.

Key Words. Bilevel program, value function, partial calmness, smoothing function, gradient consistent property, integral entropy function, smoothing projected gradient algorithm.

2010 Mathematics Subject Classification. 65K10, 90C26.

*The first and second authors' work was supported in part by NSFC Grant #11071028. The third author's work was supported in part by NSERC.

†School of Mathematical Sciences, Dalian University of Technology, Dalian 116024, China. Current address: School of Management, Shanghai University, Shanghai 200444, China. E-mail: guihualin@shu.edu.cn.

‡School of Mathematical Sciences, Dalian University of Technology, Dalian 116024, China. Current address: Department of Mathematics and Statistics, University of Victoria, Victoria, B.C., Canada V8W 3R4. 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.

1 Introduction.

Consider the simple bilevel program

$$\text{(SBP)} \qquad \min_{x\in X, y\in S(x)} \; F(x,y),$$

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

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

X and Y are closed convex subsets of $\mathbb{R}^n$ and $\mathbb{R}^m$ respectively, and $F, f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ are continuously differentiable functions. To concentrate on main ideas, we omit possible constraints on the upper level variable since the analysis can be carried over to the case where there are such constraints without much difficulty. The simple bilevel program is a special case of a general bievel program where the constraint set Y may depend on x. The reader is referred to [1, 8, 9, 20, 22] for applications and recent developments of general bilevel program.

Let x and y denote the decision variables of the leader and the follower respectively. Problem (SBP) represents the so-called *optimistic approach* to the leader and follower's game in which the follower is assumed to be co-operative and is willing to use any optimal solution from $S(x)$. Another approach called *pessimistic approach* is to assume that the follower may not be co-operative and hence the leader will have to prepare for the worst and try to solve the following pessimistic bilevel program:

$$\min_{x\in X} \max_{y\in S(x)} F(x,y).$$

Although a simple bilevel program is simpler than the general bilevel program in that the constraint region of the lower level problem is independent of the upper level decision variable x, it has many applications including a very important model in economics called the moral hazard model of the principal-agent problem [15]. The moral hazard model studies the relationship between a principal (leader) and an agent (follower) in situations in which the principal can only observe the outcome of the agent's action but not the action itself. In this situation, it is a challenge for the principal to design an optimal incentive scheme as a function of the outcome of the agent's action.

In the case where the lower level program is a convex program in variable y, the general practice to solve a bilevel program is to replace the lower level program by its Karush-Kuhn-Tucker (KKT) condition and solve a mathematical program with equilibrium constraints (MPEC). Although the globally optimal solutions for the original bilevel program and its KKT reformulation coincide, the *locally optimal solutions* for the original

bilevel program and its KKT reformulation may not be the same in the case where the lower level program has multiple multipliers (see [10]). Hence, it is not guaranteed that the solutions by solving the KKT reformulation solves the original bilevel program.

For the simple bilevel program, the so-called first order approach replaces the solution set $S(x)$ of the lower level program by the set of stationary points of the lower level program. For the case where $f(x, y)$ is convex in y, (SBP) and its first order reformulation are equivalent in terms of both globally and locally optimal solutions. In the nonconvex case, it is tempting to believe a locally optimal solution of the original bilevel program must be a stationary point of its first order reformulation. However, Mirrlees [15] gave a very convincing example (see Example 4.1 below) to show that this belief is wrong. Since the first order approach may not be valid for (SBP) in general, (SBP) remains a very difficult problem to solve theoretically and numerically. In recent years, many numerical algorithms have been suggested for bilevel programs. However, most of the works assume that the lower level program is convex with few exceptions [16, 18]. In this paper, we will try to attack this difficult problem and, in particular, we do not assume that the lower level program is convex.

Taking the value function approach, we define *the value function* of the lower level program as

$$V(x) := \inf_{y \in Y} \ f(x, y)$$

and reformulate (SBP) as the following single level optimization problem:

$$\begin{array}{lll} \text{(VP)} \qquad \min & F(x, y) & \\ \text{s.t.} & f(x, y) - V(x) \leq 0, & (1.1) \\ & (x, y) \in X \times Y. & \end{array}$$

This reformulation was first proposed by Outrata [18] for a numerical purpose and subsequently used by Ye and Zhu [24] for the purpose of obtaining necessary optimality conditions. One may think that reformulating the bilevel program (SBP) as an equivalent single level program (VP) would solve the problem. This is not true since there are two issues to be resolved. First, is a local solution of (VP) a stationary point of (VP)? Second, is there an iterative algorithm that generates a sequence converging to a stationary point of (VP)? Problem (VP) is a nonsmooth problem since the value function $V(x)$ is generally nonsmooth even when the function $f(x, y)$ is smooth. If Y is compact, by the Danskin's theorem (see Proposition 2.1 below), the value function is Lipschitz continuous and its Clarke generalized gradients may be computed. To answer the first question, in general one needs to have some constraint qualification or calmness condition. Since the constraint (1.1) is actually an equality constraint and hence the nonsmooth Mangasarian

Fromovitz constraint qualification (MFCQ) for the single level problem (VP) will never be satisfied; see [24, Proposition 3.2]. Nevertheless, using the value function formulation, Ye and Zhu [24, 25] introduced the partial calmness condition, under which a necessary optimality condition for the general bilevel program was developed. For (SBP), the partial calmness condition reduces to the calmness condition [4] that is a sufficient condition under which a local solution of (VP) is a stationary point. To address the second issue, we propose to approximate the value function by a smooth function and design a smoothing projected gradient algorithm to solve the problem. We show that any accumulation point of the sequence generated by the algorithm is a stationary point of problem (VP) provided that the sequence of the penalization parameters is bounded. Under the calmness condition, it is known that there exists a constant $\lambda > 0$ such that any locally optimal solution of (VP) is also a locally optimal solution of the exact penalty problem

$$\min_{(x,y)\in X\times Y} \; F(x,y) + \lambda(f(x,y) - V(x)).$$

Due to the exactness of the penalization, the sequence of penalization parameters generated from our smoothing projected gradient algorithm is likely to be bounded and hence the algorithm would converge to a stationary point of (VP). Note that the calmness condition for (VP) is a very strong condition so that it does not hold for many bilevel programs. In [26], a new first order necessary optimality condition was derived by a combination of the first order condition and the value function. The resulting necessary optimality condition is much more likely to hold since it contains the ones derived by using the first order condition or the value function approach as special cases.

If the calmness condition does not hold, an optimal solution of (SBP) (or equivalently an optimal solution of (VP)) is not guaranteed to be a stationary point of the problem (VP). In this case, we consider the following approximate bilevel program, where the solution set for the lower level program is replaced by the set of ε-solutions for a given $\varepsilon > 0$:

$$\begin{array}{lll} (\mathrm{VP})_\varepsilon & \min & F(x,y) \\ & \text{s.t.} & f(x,y) - V(x) - \varepsilon \le 0, \\ & & (x,y) \in X \times Y. \end{array}$$

There are three incentives to consider the above approximate bilevel program. First, in practice, it is usually too much to ask for exact optimal solutions. The follower may be satisfied with an almost optimal solution. Second, as we will show in Theorem 4.1, the solutions of $(\mathrm{VP})_\varepsilon$ approximate a solution of the original bilevel program (VP) as ε approaches zero. Third, although the nonsmooth MFCQ does not hold for (VP), it may

hold for $(\mathrm{VP})_\varepsilon$ if $\varepsilon > 0$ and hence $(\mathrm{VP})_\varepsilon$ is much easier to solve than (VP). In particular, $(\mathrm{VP})_\varepsilon$ is calm under the nonsmooth MFCQ and, consequently, the smoothing projected gradient algorithm would converge. Here, we would like to point out that the strategy of studying the approximate bilevel program has been used to study the existence and stability of bilevel programs (see [14]).

One of the main contributions of this paper is the designing of a smoothing projected gradient algorithm for solving a general *nonsmooth and nonconvex* constrained optimization problem. Our smoothing projected gradient algorithm has the advantage over other algorithms such as the sampling gradient algorithm [6] for solving nonsmooth and nonconvex problems in that we do not need to evaluate the constraint 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 program at each iteration.

The rest of the paper is organized as follows. In Section 2, we present basic definitions as well as some preliminaries which will be used in this paper. In Section 3, we propose a smoothing projected gradient algorithm for a nonsmooth and nonconvex constrained optimization problem and establish convergence for the algorithm. Section 4 is mainly devoted to the study of approximate bilevel programming problems and sufficient conditions for calmness. In Section 5, we propose to use the entropy integral function as a smoothing function of the value function and show that the entropy integral function satisfies the gradient consistent property, which is required for the convergence of the algorithm presented in Section 3. We also report our numerical experiments for two simple examples. The final section contains some concluding remarks.

We adopt the following standard notation in this paper. For any two vectors a and b in 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 $\mathrm{int}\Omega$, $\mathrm{co}\Omega$, and $\mathrm{dist}(x, \Omega)$ 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.

2 Preliminaries

In this section, we present some background materials which will be used later on. Detailed discussions on these subjects can be found in [4, 5, 17, 19, 21].

For a convex set $C \subseteq \mathbb{R}^m$ and a point $z \in C$, the normal cone of C at z is given by

$$N_C(z) := \{\zeta \in \mathbb{R}^m : \zeta^T(z' - z) \le 0,\ \forall z' \in C\}$$

and the tangent cone of C at z is given by

$$T_C(z) := \{d \in \mathbb{R}^m : (z^\nu - z)/\tau_\nu \to d \text{ for some } z^\nu \in C,\ z^\nu \to z,\ \tau_\nu \searrow 0\},$$

respectively. Let $\varphi : \mathbb{R}^n \to \mathbb{R}$ be Lipschitz continuous near $\bar{x}$. The Clarke generalized directional derivative of φ at $\bar{x}$ in direction d is defined by

$$\varphi^\circ(\bar{x}; d) := \limsup_{x \to \bar{x},\ t \searrow 0} \frac{\varphi(x + td) - \varphi(x)}{t}.$$

The Clarke generalized gradient of φ at $\bar{x}$ is a convex and compact subset of $\mathbb{R}^n$ defined by

$$\partial\varphi(\bar{x}) := \{\xi \in \mathbb{R}^n : \xi^T d \le \varphi^\circ(\bar{x}; d),\ \ \forall d \in \mathbb{R}^n\}.$$

Note that, when φ is convex, the Clarke generalized gradient coincides with the subdifferential in the sense of convex analysis, i.e.,

$$\partial\varphi(\bar{x}) = \{\xi \in \mathbb{R}^n : \xi^T(x - \bar{x}) \le \varphi(x) - \varphi(\bar{x}),\ \ \forall x \in \mathbb{R}^n\}$$

and, when φ is continuously differentiable at $\bar{x}$, we have $\partial\varphi(\bar{x}) = \{\nabla\varphi(\bar{x})\}$.

Proposition 2.1 (Danskin's Theorem) ([5, Page 99] or [7]) *Let $Y \subseteq R^m$ be a compact set and $f(x, y)$ be a function defined on $\mathbb{R}^n \times \mathbb{R}^m$ that is continuously differentiable at $\bar{x}$. Then the value function*

$$V(x) := \min\{f(x, y) : y \in Y\}$$

is Lipschitz continuous near $\bar{x}$ and its Clarke generalized gradient at $\bar{x}$ is

$$\partial V(\bar{x}) = co\{\nabla_x f(\bar{x}, y) : y \in S(\bar{x})\}, \tag{2.1}$$

where $S(\bar{x})$ is the set of all minimizers of $f(\bar{x}, y)$ over $y \in Y$.

Consider the constrained optimization problem

$$\begin{array}{lll} \text{(P)} & \min & G(x) \\ & \text{s.t.} & g(x) \le 0, \\ & & x \in \Omega, \end{array}$$

where $\Omega \subseteq \mathbb{R}^n$ is a nonempty closed and convex set, $G : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable, and $g : \mathbb{R}^n \to \mathbb{R}$ is locally Lipschitzian but not necessarily differentiable.

Definition 2.1 (Nonsmooth MFCQ) *Let $\bar{x}$ be a feasible point of problem* (P)*. We say that the nonsmooth* MFCQ *holds at $\bar{x}$ if either $g(\bar{x}) < 0$ or $g(\bar{x}) = 0$ but there exists a direction $d \in \text{int } T_\Omega(\bar{x})$ such that*

$$v^T d < 0, \quad \forall v \in \partial g(\bar{x}).$$

Following from the Fritz John type necessary optimality condition [4, Theorem 6.1.1], we define the following constraint qualification, which is weaker than the nonsmooth MFCQ but equivalent to the nonsmooth MFCQ if int $T_\Omega(\bar{x}) \neq \emptyset$ [13, 23].

Definition 2.2 (NNAMCQ) *Let $\bar{x}$ be a feasible point of problem* (P)*. We say that the no nonzero abonormal multiplier constraint qualification* (NNAMCQ) *holds at $\bar{x}$ if either $g(\bar{x}) < 0$ or $g(\bar{x}) = 0$ but*

$$0 \notin \partial g(\bar{x}) + N_\Omega(\bar{x}). \tag{2.2}$$

Note that the above condition is equivalent to saying that there is no $\mu > 0$ such that

$$\begin{aligned} &0 \in \mu \partial g(\bar{x}) + N_\Omega(\bar{x}), \\ &\mu g(\bar{x}) = 0. \end{aligned}$$

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

Definition 2.3 (ENNAMCQ) *Let $\bar{x} \in \Omega$. We say that the extended no nonzero abnormal multiplier constraint qualification* (ENNAMCQ) *holds at $\bar{x}$ for problem* (P) *if either $g(\bar{x}) < 0$ or $g(\bar{x}) \geq 0$ but*

$$0 \notin \partial g(\bar{x}) + N_\Omega(\bar{x}).$$

The following is equivalent to the calmness given in [4].

Definition 2.4 (Calmness) *Let $\bar{x}$ be a locally optimal solution of problem* (P)*. We say that* (P) *is calm at $\bar{x}$ if $\bar{x}$ is also a locally optimal solution of the exact penalty problem*

$$(\mathrm{P}_\lambda) \qquad \begin{aligned} \min \quad & G(x) + \lambda \max\{g(x), 0\} \\ \text{s.t.} \quad & x \in \Omega \end{aligned}$$

for some $\lambda > 0$.

Definition 2.5 (Stationary point) *We call a feasible point $\bar{x}$ a stationary point of problem* (P) *if there exists $\mu \geq 0$ such that*

$$\begin{aligned} &0 \in \nabla G(\bar{x}) + \mu \partial g(\bar{x}) + N_\Omega(\bar{x}), \\ &\mu g(\bar{x}) = 0. \end{aligned}$$

It is not difficult to see from the above definitions that a feasible point $\bar{x}$ is a stationary point of (P) if and only if there is some $\mu \geq 0$ such that $\mu g(\bar{x}) = 0$ and

$$\|P_\Omega[\bar{x} - \nabla G(\bar{x}) - \mu\xi] - \bar{x}\| = 0 \quad \text{for some } \xi \in \partial g(\bar{x}),$$

where P_Ω denotes the projection operator onto Ω, that is,

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

The following property is well known.

Lemma 2.1 [27] *For any $x \in \mathbb{R}^n$ and $z \in \Omega$, we have $(P_\Omega[x] - x)^T(z - P_\Omega[x]) \geq 0$.*

We now review some results from measure theory and integration [21].

Definition 2.6 (Exterior measure) *If $E \subseteq \mathbb{R}^n$, the exterior measure of E is*

$$m_*(E) := \inf \sum_{j=1}^{\infty} |Q_j|,$$

where $|Q|$ denotes the volume of a closed cube Q and the infimum is taken over all countable closed cubes $\{Q_j\}_{j=1}^{\infty}$ such that $\cup_{j=1}^{\infty} Q_j \supseteq E$.

Definition 2.7 (Lebesgue measurability) *A set $E \subseteq \mathbb{R}^n$ is Lebesgue measurable if, for any $\epsilon > 0$, there exists an open set O with $E \subseteq O$ and*

$$m_*(O - E) \leq \epsilon.$$

For a measurable set E, $m^*(E)$ is called the Lebesgue measure of E.

Proposition 2.2 [21, Property 1.3.4] *All closed sets are Lebesgue measurable.*

Lemma 2.2 (Leibniz's rule) *Let $f : X \times Y \to \mathbb{R}$ be a function such that both f and $\nabla_x f$ are continuous and Y be a compact set. Then, for any $x \in X$,*

$$\nabla_x \int_Y f(x, y)dy = \int_Y \nabla_x f(x, y)dy.$$

3 Smoothing projected gradient algorithm for (P)

In this section, we propose a smoothing projected gradient algorithm, which combines a smoothing technique with a classical projected gradient algorithm to solve the constrained optimization problem (P) given in Section 2. Our algorithm can be regarded as a generalization of the one proposed in [28] for unconstrained nonsmooth optimization problems. We suppose that the function g in (P) is eventually not differentiable at some points. Our method can be easily extended to the case where the objective function is locally Lipschitz and the case where there are more than one nonsmooth constraint.

Definition 3.1 *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 if $\lim_{z \to x,\ \rho \uparrow \infty} g_\rho(z) = g(x)$ for any fixed $x \in \mathbb{R}^n$.*

Definition 3.2 [3] *We say that a family of smoothing functions $\{g_\rho : \rho > 0\}$ satisfies the gradient consistent property if $\limsup_{z \to x,\, \rho \uparrow \infty} \nabla g_\rho(z)$ is nonempty and $\limsup_{z \to x,\, \rho \uparrow \infty} \nabla g_\rho(z) \subseteq \partial g(x)$ for any $x \in \mathbb{R}^n$, where $\limsup_{z \to x,\, \rho \uparrow \infty} \nabla g_\rho(z)$ denotes the set of all limiting points*

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

Note that our definition of smoothing functions in Definition 3.1 is different from the one originally defined in [28] in that we do not assume that the set $\limsup_{z \to x,\, \rho \uparrow \infty} \nabla g_\rho(z)$ is bounded for any given $x \in \mathbb{R}^n$. Nevertheless, since the Clarke generalized gradient of a locally Lipschitz function is nonempty and compact, it is easy to see that, a family of smooth functions $\{g_\rho : \rho > 0\}$ satisfies the gradient consistent property in our sense if and only if it satisfies the gradient consistent property in the sense of [28].

In what follows, we approximate the function $\max\{x, 0\}$ by $\frac{1}{2}(\sqrt{x^2 + \rho^{-1}} + x)$ and the nonsmooth function $g(x)$ by its family of smoothing function $\{g_\rho(x) : \rho > 0\}$ which satisfies the gradient consistent property and get the following approximation problem of (P_λ):

$$(\mathrm{P}_\lambda^\rho) \qquad \begin{array}{ll} \min & G_\rho^\lambda(x) := G(x) + \dfrac{\lambda}{2}\left(\sqrt{g_\rho^2(x) + \rho^{-1}} + g_\rho(x)\right) \\ \text{s.t.} & x \in \Omega. \end{array}$$

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

Algorithm 3.1 *1. Let $\{\beta, \gamma, \sigma_1, \sigma_2\}$ be constants in $(0, 1)$ with $\sigma_1 \le \sigma_2$, $\{\hat{\eta}, \rho_0, \lambda_0\}$ be positive constants, and $\{\sigma, \sigma'\}$ be constants in $(1, \infty)$. Choose an initial point $x^0 \in \Omega$ and set $k := 0$.*

2. Compute the stepsize β^{l_k}, where $l_k \in \{0, 1, 2 \cdots\}$ is the smallest number satisfying

$$\begin{aligned} & G_{\rho_k}^{\lambda_k}(P_\Omega[x^k - \beta^{l_k} \nabla G_{\rho_k}^{\lambda_k}(x^k)]) - G_{\rho_k}^{\lambda_k}(x^k) \\ & \le \sigma_1 \nabla G_{\rho_k}^{\lambda_k}(x^k)^T \left(P_\Omega[x^k - \beta^{l_k} \nabla G_{\rho_k}^{\lambda_k}(x^k)] - x^k\right) \end{aligned} \tag{3.1}$$

and $\beta^{l_k} \geq \gamma$, *or*

$$\begin{aligned} & G_{\rho_k}^{\lambda_k}(P_\Omega[x^k - \beta^{l_k-1}\nabla G_{\rho_k}^{\lambda_k}(x^k)]) - G_{\rho_k}^{\lambda_k}(x^k) \\ > & \sigma_2 \nabla G_{\rho_k}^{\lambda_k}(x^k)^T \left(P_\Omega[x^k - \beta^{l_k-1}\nabla G_{\rho_k}^{\lambda_k}(x^k)] - x^k\right). \end{aligned} \tag{3.2}$$

Go to Step 3.

3. *If*

$$\frac{\|P_\Omega[x^k - \beta^{l_k}\nabla G_{\rho_k}^{\lambda_k}(x^k)] - x^k\|}{\beta^{l_k}} < \hat{\eta}\rho_k^{-1}, \tag{3.3}$$

set $x^{k+1} := P_\Omega[x^k - \beta^{l_k}\nabla G_{\rho_k}^{\lambda_k}(x^k)]$ *and go to Step 4. Otherwise, set* $x^{k+1} := P_\Omega[x^k - \beta^{l_k}\nabla G_{\rho_k}^{\lambda_k}(x^k)]$, $k := k+1$, *and go to Step 2.*

4. *If*

$$g_{\rho_k}(x^{k+1}) \leq 0 \tag{3.4}$$

and

$$\|P_\Omega[x^{k+1} - \nabla G_{\rho_k}^{\lambda_k}(x^{k+1})] - x^{k+1}\| = 0, \tag{3.5}$$

go to Step 6. Else if (3.4) *holds while* (3.5) *fails, go to Step 5. Otherwise, if* (3.4) *fails, set* $\lambda_{k+1} := \sigma'\lambda_k$ *and go to Step 5.*

5. *Set* $\rho_{k+1} := \sigma\rho_k$, $k := k+1$, *and go to Step 2.*

6. *If a stopping criterion leading to the stationary condition for* (P) *holds at* x^{k+1}, *terminate. Otherwise, go to Step 5.*

We make some remarks on Algorithm 3.1. First of all, it is easy to see that Step 2 of the algorithm is the Armijo line search. In practice, only a small number of iterations are required to compute the Armijo stepsize. Note, in particular, that the Armijo procedure (3.1) – (3.2) satisfies the conditions in [28] with $\gamma_2 = \beta$. The search for a stepsize is a finite process under the continuous differentiability of G_ρ^λ, which can be seen from [28].

Moreover, for a given tolerance $\epsilon > 0$, we suggest the condition

$$|G_{\rho_k}^{\lambda_k}(x^{k+1}) - G(x^{k+1})| \leq \epsilon \tag{3.6}$$

as a stopping criterion of the above algorithm. To justify the stopping criterion (3.6), we assume without loss of generality that $x^k \to x^*$ as $k \to \infty$ and denote

$$\mu_\rho^\lambda(x) := \frac{\lambda}{2}\left(\frac{g_\rho(x)}{\sqrt{g_\rho^2(x) + \rho^{-1}}} + 1\right). \tag{3.7}$$

Then $\nabla G_{\rho_k}^{\lambda_k}(x^{k+1})$ is equal to $\nabla G(x^{k+1}) + \mu_{\rho_k}^{\lambda_k}(x^{k+1})\nabla g_{\rho_k}(x^{k+1})$. Consider the stopping criterion (3.6). If

$$G_{\rho_k}^{\lambda_k}(x^{k+1}) - G(x^{k+1}) = \frac{\lambda_k}{2}\Big(\sqrt{g_{\rho_k}^2(x^{k+1}) + \rho_k^{-1}} + g_{\rho_k}(x^{k+1})\Big) \to 0 \ \text{ as } \ k \to \infty,$$

it follows that

$$\begin{aligned} \mu_{\rho_k}^{\lambda_k}(x^{k+1})g_{\rho_k}(x^{k+1}) &= \frac{\lambda_k}{2}\Big(\sqrt{g_{\rho_k}^2(x^{k+1}) + \rho_k^{-1}} + g_{\rho_k}(x^{k+1})\Big)\frac{g_{\rho_k}(x^{k+1})}{\sqrt{g_{\rho_k}^2(x^{k+1}) + \rho_k^{-1}}} \\ &\to 0 \quad \text{as} \quad k \to \infty. \end{aligned} \tag{3.8}$$

Therefore, letting μ^* be an accumulation point of $\{\mu_{\rho_k}^{\lambda_k}(x^{k+1})\}$, we have from (3.4) and (3.7) – (3.8) that

$$\mu^* \geq 0, \quad g(x^*) \leq 0, \quad \mu^* g(x^*) = 0. \tag{3.9}$$

Since any limit of $\{\nabla g_{\rho_k}(x^{k+1})\}$ must be an element of $\partial g(x^*)$ by Definition 3.2, we have from (3.5) and the definition of $\nabla G_{\rho_k}^{\lambda_k}(x^{k+1})$ that any limit d^* of $\{\nabla G_{\rho_k}^{\lambda_k}(x^{k+1})\}$ must satisfy

$$d^* \in \nabla G(x^*) + \mu^* \partial g(x^*), \quad \|P_\Omega[x^* - d^*] - x^*\| = 0,$$

which means

$$0 \in \nabla G(x^*) + \mu^* \partial g(x^*) + N_\Omega(x^*).$$

This, together with (3.9), indicates that x^* is a stationary point of (P). Therefore, (3.6) is a reasonable stopping criterion.

In addition, if Algorithm 3.1 does not terminate at Step 6, the assumption given below guarantees that $\rho_k \to +\infty$ as $k \to \infty$, which is shown in the next lemma.

Assumption 3.1 *For any $\rho > 0$ and $\lambda > 0$, $G_\rho^\lambda(\cdot)$ is bounded below and $\nabla G_\rho^\lambda(\cdot)$ is uniformly continuous on the nonempty closed convex set Ω, that is, for any $\epsilon > 0$, there exists $\delta > 0$ such that*

$$x \in \Omega, \ y \in \Omega, \ \|x - y\| < \delta \quad \Longrightarrow \quad \|\nabla G_\rho^\lambda(x) - \nabla G_\rho^\lambda(y)\| < \epsilon.$$

Lemma 3.1 *Under Assumption 3.1, if Algorithm 3.1 does not terminate at Step 6, we have* $\lim_{k\to\infty} \rho_k = +\infty$.

Proof. Note that, for any $\rho > 0$ and $\lambda > 0$, G_ρ^λ is continuously differentiable and the Armijo procedure (3.1) – (3.2) satisfies conditions (2.1) – (2.3) in [2] with $\gamma_2 = \beta$. Then, following the proof of [2, Theorem 2.3], we have

$$\lim_{k\to\infty} \frac{\|P_\Omega[x^k - \beta^{l_k}\nabla G_{\rho_k}^{\lambda_k}(x^k)] - x^k\|}{\beta^{l_k}} = 0,$$

which means that, for any $\rho_k > 0$, we can find some x^k such that condition (3.3) holds. Then $\lim_{k\to\infty} \rho_k = +\infty$ by Algorithm 3.1. ■

We now introduce an inequality which was proposed by Dunn [11] in his analysis for projected gradient methods.

Lemma 3.2 *Suppose that $\{x^k\}$ is a sequence generated by Algorithm 3.1. Then, for each k, we have*

$$\nabla G_{\rho_{k-1}}^{\lambda_{k-1}}(x^{k-1})^T(x^k - x^{k-1}) \le -\frac{\|x^k - x^{k-1}\|^2}{\beta^{l_{k-1}}}. \tag{3.10}$$

Proof. By setting $x := x^{k-1} - \beta^{l_{k-1}}\nabla G_{\rho_{k-1}}^{\lambda_{k-1}}(x^{k-1})$ and $z := x^{k-1}$ in Lemma 2.1, the following inequality can be obtained immediately:

$$\begin{aligned}&\nabla G_{\rho_{k-1}}^{\lambda_{k-1}}(x^{k-1})^T\left(P_\Omega[x^{k-1} - \beta^{l_{k-1}}\nabla G_{\rho_{k-1}}^{\lambda_{k-1}}(x^{k-1})] - x^{k-1}\right)\\ &\le -\frac{\|P_\Omega[x^{k-1} - \beta^{l_{k-1}}\nabla G_{\rho_{k-1}}^{\lambda_{k-1}}(x^{k-1})] - x^{k-1}\|^2}{\beta^{l_{k-1}}}.\end{aligned}$$

This implies the required inequality since $x^k = P_\Omega[x^{k-1} - \beta^{l_{k-1}}\nabla G_{\rho_{k-1}}^{\lambda_{k-1}}(x^{k-1})]$. ■

Lemma 3.3 *Suppose that Algorithm 3.1 does not terminate at Step 6 and $\{x^k\}$ is a sequence generated by the algorithm such that condition* (3.5) *fails for each k. Then, under Assumption 3.1, for each $x \in \Omega$, we have*

$$\begin{aligned}&\nabla G_{\rho_{k-1}}^{\lambda_{k-1}}(x^{k-1})^T(x^{k-1} - x)\\ &\quad\le \nabla G_{\rho_{k-1}}^{\lambda_{k-1}}(x^{k-1})^T(x^{k-1} - x^k) + \frac{1}{\beta^{l_{k-1}}}\|x^k - x^{k-1}\|\|x^{k-1} - x\|\end{aligned} \tag{3.11}$$

and

$$\limsup_{k\to\infty} \nabla G_{\rho_{k-1}}^{\lambda_{k-1}}(x^{k-1})^T(x^{k-1} - x) \le 0. \tag{3.12}$$

Proof. Note that condition (3.3) implies

$$\lim_{k\to\infty} \frac{\|x^k - x^{k-1}\|}{\beta^{l_{k-1}}} \le \lim_{k\to\infty} \hat{\eta}\frac{1}{\rho_{k-1}} = 0, \tag{3.13}$$

while condition (3.10) together with (3.1) implies

$$\begin{aligned} G_{\rho_{k-1}}^{\lambda_{k-1}}(x^k) - G_{\rho_{k-1}}^{\lambda_{k-1}}(x^{k-1}) &\leq \sigma_1 \nabla G_{\rho_{k-1}}^{\lambda_{k-1}}(x^{k-1})^T (x^k - x^{k-1}) \\ &\leq -\sigma_1 \frac{\|x^k - x^{k-1}\|^2}{\beta^{l_{k-1}}} \end{aligned} \tag{3.14}$$

for each k. Since $G_{\rho_{k-1}}^{\lambda_{k-1}}$ is smooth, we have $\lim_{k\to\infty} G_{\rho_{k-1}}^{\lambda_{k-1}}(x^k) - G_{\rho_{k-1}}^{\lambda_{k-1}}(x^{k-1}) = 0$. This together with (3.14) yields

$$\lim_{k\to\infty} \nabla G_{\rho_{k-1}}^{\lambda_{k-1}}(x^{k-1})^T (x^k - x^{k-1}) = 0. \tag{3.15}$$

For any $z \in \Omega$, by setting $x := x^{k-1} - \beta^{l_{k-1}} \nabla G_{\rho_{k-1}}^{\lambda_{k-1}}(x^{k-1})$ in Lemma 2.1, we have that, for each k,

$$\begin{aligned} \beta^{l_{k-1}} \nabla G_{\rho_{k-1}}^{\lambda_{k-1}}(x^{k-1})^T (x^k - z) &\leq (x^k - x^{k-1})^T (z - x^k) \\ &\leq (x^k - x^{k-1})^T (z - x^{k-1}) \\ &\leq \|x^k - x^{k-1}\| \|x^{k-1} - z\|. \end{aligned}$$

Thus, we obtain (3.11) by setting $z := x \in \Omega$ in the above inequality. Furthermore, we have (3.12) from (3.13) and (3.15). ■

Suppose that Algorithm 3.1 does not terminate within finite iterations. The next theorem shows the global convergence of Algorithm 3.1.

Theorem 3.1 *Let Assumption 3.1 hold and x^* be an accumulation point of the sequence $\{x^k\}$ generated by Algorithm 3.1. If $\{\lambda_k\}$ is bounded, then x^* is a stationary point of* (P).

Proof. Since $\{\lambda_k\}$ is bounded, there exist $\bar{k}$ and $\hat{\lambda}$ such that $\lambda_k = \hat{\lambda}$ and condition (3.4) hold for all $k \geq \bar{k}$. Let $\mu_\rho^\lambda(x)$ be defined as (3.7). We consider the following two cases.

(i) Consider the case where there is a sequence $K_0 \subseteq \mathbf{N}$ such that both (3.4) and (3.5) hold for all $k \in K_0$. It is easy to see that, for each $k \in K_0$, by the discussions in Section 2, x^k is a stationary point of $\min_{x\in\Omega} G_{\rho_{k-1}}^{\lambda_{k-1}}(x)$, that is,

$$0 \in \nabla G_{\rho_{k-1}}^{\lambda_{k-1}}(x^k) + N_\Omega(x^k) = \nabla G(x^k) + \mu_{\rho_{k-1}}^{\lambda_{k-1}}(x^k) \nabla g_{\rho_{k-1}}(x^k) + N_\Omega(x^k). \tag{3.16}$$

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

$$\lim_{k\to\infty, k\in\hat{K_0}} \nabla g_{\rho_{k-1}}(x^k) \in \partial g(x^*).$$

Note that, by (3.7), $\{\mu_{\rho_{k-1}}^{\lambda_{k-1}}(x^k)\}$ is bounded. Hence, there is a subsequence $\bar{K_0} \subseteq \hat{K_0}$ such that $\{\mu_{\rho_{k-1}}^{\lambda_{k-1}}(x^k)\}_{k\in\bar{K_0}}$ is convergent. Let $\bar{\mu} := \lim_{k\to\infty, k\in\bar{K_0}} \mu_{\rho_{k-1}}^{\lambda_{k-1}}(x^k)$. It follows from (3.7) that $\bar{\mu} \geq 0$ and, by letting $k \to \infty$ with $k \in \bar{K_0}$ in (3.16),

$$0 \in \nabla G(x^*) + \bar{\mu} \partial g(x^*) + N_\Omega(x^*). \tag{3.17}$$

On the other hand, note that $g(x^*) = \lim_{k\to\infty} g_{\rho_{k-1}}(x^k) \le 0$ by (3.4). Therefore, if $g(x^*) < 0$, there holds $\bar{\mu} = 0$ from (3.7) and Lemma 3.1. As a result, we always have $\bar{\mu} g(x^*) = 0$. From the above discussion, we know that x^* is a stationary point of (P).

(ii) Consider the case where there is a sequence $K_1 \subseteq \mathbf{N}$ such that (3.4) holds while (3.5) fails for all $k \in K_1$. We have from condition (3.3) and Lemma 3.1 that $\lim_{k\to\infty,\, k\in K_1} x^{k-1} = x^*$. By the gradient consistent property of g_ρ, there exists a subsequence $\hat{K}_1 \subseteq K_1$ such that

$$\lim_{k\to\infty,\, k\in \hat{K}_1} \nabla g_{\rho_{k-1}}(x^{k-1}) \in \partial g(x^*).$$

Note that, by (3.7), $\{\mu^{\lambda_{k-1}}_{\rho_{k-1}}(x^{k-1})\}$ is bounded. Hence, there is a subsequence $\bar{K}_1 \subseteq \hat{K}_1$ such that $\{\mu^{\lambda_{k-1}}_{\rho_{k-1}}(x^{k-1})\}_{k\in\bar{K}_1}$ is convergent. Let $\bar{\mu} := \lim_{k\to\infty,\, k\in\bar{K}_1} \mu^{\lambda_{k-1}}_{\rho_{k-1}}(x^{k-1})$. It follows from (3.7) that $\bar{\mu} \ge 0$. Note also that $g(x^*) = \lim_{k\to\infty} g_{\rho_{k-1}}(x^k) \le 0$ by (3.4). Therefore, if $g(x^*) < 0$, we have $g_{\rho_{k-1}}(x^{k-1}) < 0$ by Definition 3.1. Hence, there holds $\bar{\mu} = 0$ from (3.7) and Lemma 3.1. As a result, we always have $\bar{\mu} g(x^*) = 0$. On the other hand, let

$$\begin{aligned} V_{k-1} &:= \nabla G^{\lambda_{k-1}}_{\rho_{k-1}}(x^{k-1}) = \nabla G(x^{k-1}) + \mu^{\lambda_{k-1}}_{\rho_{k-1}}(x^{k-1}) \nabla g_{\rho_{k-1}}(x^{k-1}), \\ V &:= \lim_{k\to\infty,\, k\in\bar{K}_1} V_{k-1} \in \nabla G(x^*) + \bar{\mu}\partial g(x^*). \end{aligned}$$

It follows from Lemma 3.3 that

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

This means $-V \in N_\Omega(x^*)$ and hence (3.17) holds. From the above discussion, we know that x^* is a stationary point of (P). This completes the proof. ■

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

Theorem 3.2 *Let Assumption 3.1 hold and $\{x^k\}$ be a sequence generated by Algorithm 3.1. Suppose that $\lim_{k\to\infty} x^k = x^*$ and the* ENNAMCQ *holds at x^* for* (P)*, then $\{\lambda_k\}$ is bounded.*

Proof. Assume for a contradiction that the conclusion is not true. This means that there is a sequence $K_1 \subseteq \mathbf{N}$ such that condition (3.4) fails for all $k \in K_1$. Let $\mu^\lambda_\rho(x)$ be defined as (3.7).

First consider the case where there is a subsequence $K_2 \subseteq K_1$ such that condition (3.5) holds for every $k \in K_2$. Similarly to Part (i) of the proof of Theorem 3.1, we know that condition (3.16) holds for every $k \in K_2$ and, since $g_{\rho_{k-1}}(x^k) > 0$ for all $k \in K_2$,

$$\mu^{\lambda_{k-1}}_{\rho_{k-1}}(x^k) \to +\infty \ \text{ as } \ K_2 \ni k \to \infty. \tag{3.18}$$

By the gradient consistent property of g_ρ, there exists a subsequence $\hat{K}_2 \subseteq K_2$ such that

$$\lim_{k\to\infty, k\in\hat{K}_2} \nabla g_{\rho_{k-1}}(x^k) \in \partial g(x^*).$$

Dividing by $\mu_{\rho_{k-1}}^{\lambda_{k-1}}(x^k)$ in both sides of (3.16), we have

$$0 \in \frac{1}{\mu_{\rho_{k-1}}^{\lambda_{k-1}}(x^k)} \nabla G(x^k) + \nabla g_{\rho_{k-1}}(x^k) + N_\Omega(x^k). \tag{3.19}$$

Letting $k \to \infty$ with $k \in \hat{K}_2$ in (3.19), we have from (3.18) that

$$0 \in \partial g(x^*) + N_\Omega(x^*), \tag{3.20}$$

which contradicts the ENNAMCQ assumption.

Now we consider the case where condition (3.5) fails for every $k \in K_1$ sufficiently large. By the gradient consistent property of g_ρ, there exists a subsequence $\hat{K}_1 \subseteq K_1$ such that

$$v := \lim_{k\to\infty, k\in\hat{K}_1} \nabla g_{\rho_{k-1}}(x^{k-1}) \in \partial g(x^*).$$

On the other hand, we have from (3.3) that, for each k,

$$\|x^k - x^{k-1}\| \le \hat{\eta}\rho_{k-1}^{-1}.$$

Moreover, it follows from the gradient consistent property and the fact that the Clarke generalized gradient is nonempty and compact that the set $\limsup_{z\to x^*,\, \rho\uparrow\infty} \nabla g_\rho(z)$ is nonempty and bounded. Thus, from the mean-value theorem, there exist a constant $c > 0$ and a positive integer k_0 such that

$$|g_{\rho_{k-1}}(x^k) - g_{\rho_{k-1}}(x^{k-1})| \le c\rho_{k-1}^{-1}$$

holds for each $k \in \hat{K}_1$ with $k \ge k_0$. For each $k \in \hat{K}_1$ with $k \ge k_0$, since $g_{\rho_{k-1}}(x^k) > 0$, we have

$$g_{\rho_{k-1}}(x^{k-1}) \ge g_{\rho_{k-1}}(x^k) - c\rho_{k-1}^{-1} > -c\rho_{k-1}^{-1}$$

and hence

$$\frac{g_{\rho_{k-1}}(x^{k-1})}{\sqrt{g_{\rho_{k-1}}^2(x^{k-1}) + \rho_{k-1}^{-1}}} > \frac{-c\rho_{k-1}^{-1}}{\sqrt{g_{\rho_{k-1}}^2(x^{k-1}) + \rho_{k-1}^{-1}}} = \frac{-c}{\sqrt{\rho_{k-1}^2 g_{\rho_{k-1}}^2(x^{k-1}) + \rho_{k-1}}} \to 0$$

as $k \to \infty$. This implies

$$\mu_{\rho_{k-1}}^{\lambda_{k-1}}(x^{k-1}) = \frac{\lambda_{k-1}}{2}\left(\frac{g_{\rho_{k-1}}(x^{k-1})}{\sqrt{g_{\rho_{k-1}}^2(x^{k-1}) + \rho_{k-1}^{-1}}} + 1\right) \to +\infty$$

as $k \to \infty$. Then, for any $x \in \Omega$ and $k \in \hat{K}_1$, dividing by $\mu_{\rho_{k-1}}^{\lambda_{k-1}}(x^{k-1})$ in both sides of (3.11), we have

$$\begin{aligned} & \left(\frac{1}{\mu_{\rho_{k-1}}^{\lambda_{k-1}}(x^{k-1})} \nabla G(x^{k-1}) + \nabla g_{\rho_{k-1}}(x^{k-1}) \right)^T (x^{k-1} - x) \\ \leq & \frac{1}{\mu_{\rho_{k-1}}^{\lambda_{k-1}}(x^{k-1})} \nabla G_{\rho_{k-1}}^{\lambda_{k-1}}(x^{k-1})^T (x^{k-1} - x^k) + \frac{1}{\beta^{l_{k-1}} \mu_{\rho_{k-1}}^{\lambda_{k-1}}(x^{k-1})} \|x^k - x^{k-1}\| \|x^{k-1} - x\|. \end{aligned}$$

Taking a limit within $\hat{K}_1$, we have from (3.13) and (3.15) that, for any $x \in \Omega$,

$$\begin{aligned} v^T(x^* - x) & \leq \lim_{k \to \infty, k \in \hat{K}_1} \frac{1}{\mu_{\rho_{k-1}}^{\lambda_{k-1}}(x^{k-1})} \nabla G_{\rho_{k-1}}^{\lambda_{k-1}}(x^{k-1})^T (x^{k-1} - x^k) \\ & \quad + \lim_{k \to \infty, k \in \hat{K}_1} \frac{\|x^{k-1} - x\|}{\mu_{\rho_{k-1}}^{\lambda_{k-1}}(x^{k-1})} \frac{\|x^k - x^{k-1}\|}{\beta^{l_{k-1}}} \\ & = 0, \end{aligned}$$

which means

$$0 \in v + N_\Omega(x^*) \subseteq \partial g(x^*) + N_\Omega(x^*).$$

This contradicts the ENNAMCQ assumption.

From the above discussion, we know that $\{\lambda_k\}$ is bounded. ■

The next corollary follows immediately from Theorems 3.1 and 3.2.

Corollary 3.1 *Let Assumption 3.1 hold. Suppose that $\{x^k\}$ is a sequence generated by Algorithm 3.1 and $\lim_{k \to \infty} x^k = x^*$. If the* ENNAMCQ *holds at x^*, then x^* is a stationary point of* (P).

Notice that, in Theorem 3.2 and Corollary 3.1, x^* must be a limit point of the sequence generated by the algorithm. It is not enough to just assume that x^* is an accumulation point of the sequence generated by the algorithm. The reason is that the subsequence K_1 in the proof of Theorem 3.2 may not be included in any subsequence converging to the accumulation point and hence the contradiction to the NNAMCQ in the proof may not be true. 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 3.3 *Let Assumption 3.1 hold and $\{x^k\}$ be a sequence generated by Algorithm 3.1. Assume that the* ENNAMCQ *holds for* (P) *at any point x satisfying $g(x) \geq 0$. If $\{x^k\}$ is bounded, then $\{\lambda^k\}$ is bounded and hence any accumulation point of $\{x^k\}$ is a stationary point of* (P).

Proof. Suppose to the contrary that the sequence $\{\lambda^k\}$ is unbounded. Then there is a sequence $K \subseteq \mathbf{N}$ such that condition (3.4) fails for all $k \in K$. Let x^* be an accumulation point of $\{x^k\}_{k\in K}$. Then we must have $g(x^*) \geq 0$ and hence, by the assumption, the ENNAMCQ holds at x^* for (P). On the other hand, similarly as in the proof of Theorem 3.2, we can show that the ENNAMCQ fails at x^*. As a contradition, we have shown that $\{\lambda^k\}$ is bounded.

The second assertion follows from the boundedness of $\{\lambda^k\}$ and Theorem 3.1 immediately. ■

We now discuss the situations when the NNAMCQ does not hold but problem (P) is calm at a local optimal solution. Let x^* be a locally optimal solution of (P) and (P) be calm at x^*. Without loss of generality, we assume $g(x^*) = 0$. Since there exists $\lambda^* > 0$ sufficiently large such that x^* is also a local solution to the exact penalty problem (P_{λ^*}), we have

$$0 \in \nabla G(x^*) + \lambda^* \mu \partial g(x^*) + N_\Omega(x^*), \quad \mu \in [0, 1].$$

Fix $\lambda > 0$. For any $\rho > 0$, let x_ρ be a stationary point of problem $(\mathrm{P}^\rho_\lambda)$. If $x_\rho \to x^*$ as $\rho \to \infty$, we can derive

$$0 \in \nabla G(x^*) + \lambda \mu \partial g(x^*) + N_\Omega(x^*), \quad \mu \in [0, 1]$$

by the gradient consistent property of g_ρ. Hence, it is easy to see that, if (P) is calm at all locally optimal solutions, then the sequence of penalty parameters of any convergent sequence generated by the algorithm will be likely to be bounded.

4 Approximate bilevel programs

Consider the approximate bilevel program $(\mathrm{VP})_\varepsilon$ introduced in Section 1. We first investigate its limiting behavior.

Theorem 4.1 *Let F be continuous, both f and $\nabla_x f$ be continuously differentiable and X, Y be closed sets. For each $\varepsilon > 0$, suppose that $(x^\varepsilon_\delta, y^\varepsilon_\delta)$ is a δ-solution of problem $(\mathrm{VP})_\varepsilon$, i.e., for any feasible point (x, y) of $(\mathrm{VP})_\varepsilon$, $F(x, y)$ is not less than $F(x^\varepsilon_\delta, y^\varepsilon_\delta) - \delta$. Then any accumulation point of the net $\{(x^\varepsilon_\delta, y^\varepsilon_\delta)\}$ as ε and δ approach zero is an optimal solution of the bilevel program* (SBP).

Proof. Without loss of generality, suppose that $\lim\limits_{\varepsilon\downarrow 0, \delta\downarrow 0} (x^\varepsilon_\delta, y^\varepsilon_\delta) = (x^*, y^*)$. By the continuity of the functions f and V (see Proposition 2.1), it is easy to verify that (x^*, y^*) is a feasible

point of problem (SBP). Suppose that (x^*, y^*) is not an optimal solution of (SBP). Then there must exist a feasible point $(\bar{x}, \bar{y}) \neq (x^*, y^*)$ such that

$$F(\bar{x}, \bar{y}) < F(x^*, y^*). \tag{4.1}$$

Since $(x_\delta^\varepsilon, y_\delta^\varepsilon)$ is a δ-solution of $(\mathrm{VP})_\varepsilon$ and $(\bar{x}, \bar{y})$ is a feasible point of $(\mathrm{VP})_\varepsilon$, we have

$$F(\bar{x}, \bar{y}) \geq F(x_\delta^\varepsilon, y_\delta^\varepsilon) - \delta.$$

Letting ε and δ tend to zero, we have

$$F(\bar{x}, \bar{y}) \geq F(x^*, y^*).$$

This contradicts (4.1) and hence (x^*, y^*) is an optimal solution of the bilevel program (SBP). ■

For any $\varepsilon > 0$, the approximate bilevel program $(\mathrm{VP})_\varepsilon$ is relatively easy to solve since, unlike the original bilevel program, it is possible to satisfy the NNAMCQ. Indeed, if $(x^\varepsilon, y^\varepsilon)$ is a feasible point of $(\mathrm{VP})_\varepsilon$ with $f(x^\varepsilon, y^\varepsilon) - V(x^\varepsilon) = \varepsilon$, then y^ε is not a solution of the lower level program $(\mathrm{P}_{x^\varepsilon})$ and hence it is possible to satisfy condition (2.2).

Proposition 4.1 *Let $f(x, y)$ be continuously differentiable and X, Y be closed sets. Problem $(\mathrm{VP})_\varepsilon$ satisfies the* ENNAMCQ *at $(x^\varepsilon, y^\varepsilon)$ if one of the following conditions holds:*

(1) $f(x^\varepsilon, y^\varepsilon) - V(x^\varepsilon) < \varepsilon$.

(2) $f(x^\varepsilon, y^\varepsilon) - V(x^\varepsilon) \geq \varepsilon$, $(x^\varepsilon, y^\varepsilon)$ *is an interior point of* $X \times Y$, *and* $\nabla_y f(x^\varepsilon, y^\varepsilon) \neq 0$.

(3) $f(x^\varepsilon, y^\varepsilon) - V(x^\varepsilon) \geq \varepsilon$, $(x^\varepsilon, y^\varepsilon)$ *is an interior point of* $X \times Y$, *and* $\nabla_x f(x^\varepsilon, y^\varepsilon) \notin \partial V(x^\varepsilon)$.

Furthermore, if $(x^\varepsilon, y^\varepsilon)$ is a locally optimal solution of $(\mathrm{VP})_\varepsilon$, then $(\mathrm{VP})_\varepsilon$ is calm at $(x^\varepsilon, y^\varepsilon)$.

Proof. Since $(x^\varepsilon, y^\varepsilon)$ is an interior point of the feasible set $X \times Y$, we have $N_{X\times Y}(x^\varepsilon, y^\varepsilon) = \{(0, 0)\}$. Then condition (2.2) for $(\mathrm{VP})_\varepsilon$ reduces to either $\nabla_x f(x^\varepsilon, y^\varepsilon) \notin \partial V(x^\varepsilon)$ or $\nabla_y f(x^\varepsilon, y^\varepsilon) \neq 0$. Hence, the ENNAMCQ holds at $(x^\varepsilon, y^\varepsilon)$ by the assumptions. Furthermore, if $(x^\varepsilon, y^\varepsilon)$ is a locally optimal solution of $(\mathrm{VP})_\varepsilon$, then $(x^\varepsilon, y^\varepsilon)$ is feasible for $(\mathrm{VP})_\varepsilon$ and the NNAMCQ holds at it. Since it is well known that the NNAMCQ is a sufficient condition for calmness, problem $(\mathrm{VP})_\varepsilon$ is calm at $(x^\varepsilon, y^\varepsilon)$. ■

We next use some examples to illustrate the above result.

Example 4.1 (Mirrlees' problem) Consider

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

The first order optimality condition for the lower level program is

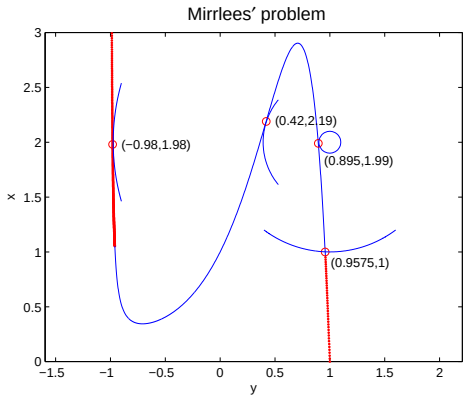
$$x(y+1)\exp[-(y+1)^2] + (y-1)\exp[-(y-1)^2] = 0.$$

Hence, the relation between x and any stationary point y of the lower level program is given by

$$(1+y)x = (1-y)\exp[4y], \tag{4.2}$$

which is a smooth and connected curve as shown in Figure 1. Since the objective of the lower level program is not convex in y, for each fixed x, not all corresponding $y's$ lying on the curve are globally optimal solutions of the lower level program. The true globally optimal solutions for the lower level program run as a disconnected curve with a jump at $\bar{x} = 1$ (see the darker curve in Figure 1), which represents the feasible region of the bilevel program.

Figure 1:



By the value function approach, Mirrlees' problem is equivalent to the single level optimization problem

$$\begin{aligned} \min \quad & F(x,y) \\ \text{s.t.} \quad & f(x,y) - V(x) \le 0. \end{aligned} \tag{4.3}$$

As shown by Mirrlees [15], at $\bar{x} = 1$, both $\bar{y}_1 \approx 0.9575$ and $\bar{y}_2 \approx -0.9575$ are optimal solutions of the lower level program $(\mathrm{P}_{\bar{x}})$. By Danskin's theorem, we have

$$\partial V(\bar{x}) = co\{\nabla_x f(\bar{x}, \bar{y}_1), \nabla_x f(\bar{x}, \bar{y}_2)\}.$$

As shown in [26], problem (4.3) is not calm at the solution $(\bar{x}, \bar{y}) \approx (1, 0.9575)$ and the optimal solution $(\bar{x}, \bar{y})$ is not a stationary point of problem (4.3). We now consider the approximate bilevel program

$$\begin{array}{ll} \min & F(x,y) \\ \text{s.t.} & f(x,y) - V(x) \le \varepsilon. \end{array} \tag{4.4}$$

Let ε be a positive number that is not equal to $f(1,0) - V(1)$. If $f(x^\varepsilon, y^\varepsilon) - V(x^\varepsilon) < \varepsilon$ then ENNAMCQ (equivalently NNAMCQ) holds at $(x^\varepsilon, y^\varepsilon)$. Otherwise suppose $f(x^\varepsilon, y^\varepsilon) - V(x^\varepsilon) \ge \varepsilon$. Then $y^\varepsilon \notin S(x^\varepsilon)$ and hence $(x^\varepsilon, y^\varepsilon)$ does not lie on the darker curve. If $(x^\varepsilon, y^\varepsilon)$ does not lie on the lighter curve as well, then $f_y(x^\varepsilon, y^\varepsilon) \neq 0$. If $(x^\varepsilon, y^\varepsilon)$ lies on the lighter curve and $x^\varepsilon \neq 1$, then $S(x^\varepsilon)$ is a singleton, say $\{y(x^\varepsilon)\}$, and hence by Danskin's theorem the value function is differentiable at x^ε with $V'(x^\varepsilon) = f'_x(x^\varepsilon, y(x^\varepsilon))$ and so $f'_x(x^\varepsilon, y^\varepsilon) \notin \partial V(x^\varepsilon)$. Note the choice of ε has ruled out the possibility that $x^\varepsilon = 1$ and $(x^\varepsilon, y^\varepsilon)$ lies on the lighter curve which means that $(x^\varepsilon, y^\varepsilon) = (0, 1)$. By Proposition 4.1, the ENNAMCQ holds at all points $(x^\varepsilon, y^\varepsilon)$. Furthermore, suppose that $(x^\varepsilon, y^\varepsilon)$ is a local solution of problem (4.4). Then, by Proposition 4.1, $(x^\varepsilon, y^\varepsilon)$ is a locally optimal solution of the exact penalty problem

$$\min_{x,y} \quad F(x,y) + \lambda \max\{f(x,y) - V(x) - \varepsilon, 0\}$$

for some $\lambda > 0$ sufficiently large.

5 Smoothing projected gradient algorithm for bilevel programs

In this section, we first present a smoothing approximation for the value function $V(x)$ and then apply the smoothing projected algorithm presented in Section 3 to the approximate bilevel program $(\mathrm{VP})_\varepsilon$ with $\varepsilon \ge 0$.

Throughout this section, we suppose that the set Y is a nonempty and compact set with $m^*(Y) \neq 0$. For given $\rho > 0$ and an integrable function $f(x,y)$, we define the integral entropy function as

$$\begin{aligned} \gamma_\rho(x) &:= -\rho^{-1} \ln \left(\int_Y \exp[-\rho f(x,y)] dy \right) \\ &\equiv V(x) - \rho^{-1} \ln \left(\int_Y \exp[-\rho (f(x,y) - V(x))] dy \right). \end{aligned}$$

As shown in the next theorem, the above function is a smoothing approximation of the value function of the lower level program.

Theorem 5.1 *Let $f(x,y)$ be continuous in (x,y) and continuously differentiable in x. The family of entropy integral functions $\{\gamma_\rho(x) : \rho > 0\}$ is a family of smoothing functions for the value function $V(x)$.*

Proof. The continuous differentiability of $\gamma_\rho(x)$ is obvious by its definition. From the proof of [12, Theorem 1], it is easy to get that, for any $\epsilon > 0$, there exist $l \in (\exp[-\epsilon], 1)$ and $\tilde{\rho} > 0$ such that, for any $\rho > \tilde{\rho}$ and $(x,y) \in X \times Y$, there holds

$$l\ \left(\rho^{-1} m^*(Y)\right)^{1/\rho} \max_{y\in Y} \exp[-f(x,y)] \le \left(\int_Y \exp[-\rho f(x,y)]dy\right)^{1/\rho} \le m^*(Y)^{1/\rho} \max_{y\in Y} \exp[-f(x,y)].$$

By the monotonicity of the logarithmic function, we have

$$V(x) - \rho^{-1} \ln m^*(Y) \le \gamma_\rho(x) \le V(x) - \rho^{-1} \ln(\rho^{-1} m^*(Y)) + \epsilon$$

for any x, where ρ is sufficiently large. From the Squeeze law, we have

$$\lim_{z\to x,\, \rho\to\infty} \gamma_\rho(z) = V(x)$$

for any x. This completes the proof. ■

To show the gradient consistent property of the family of entropy integral functions, we first derive some preliminary results. The next theorem gives an integral representation for the gradient of the integral entropy function.

Theorem 5.2 *Let $f(x,y)$ be a continuous function which is continuously differentiable in variable x. For fixed $\rho > 0$, $\gamma_\rho(x)$ is differentiable and*

$$\nabla\gamma_\rho(x) = \int_Y \mu_\rho(x,y) \nabla_x f(x,y) dy,$$

where

$$\mu_\rho(x,y) := \frac{\exp[-\rho f(x,y)]}{\int_Y \exp[-\rho f(x,z)]dz}.$$

Proof. By the definition of γ_ρ, we have

$$\nabla\gamma_\rho(x) = -\rho^{-1} \frac{\nabla_x \int_Y \exp[-\rho f(x,y)]dy}{\int_Y \exp[-\rho f(x,y)]dy}.$$

Form the continuous differentiability of f, we know that $\exp[-\rho f(x,y)]$ is continuously differentiable in x. Thus, from the Leibniz's rule for differentiating an integral, we have

$$\begin{aligned}\nabla_x \int_Y \exp[-\rho f(x,y)]dy &= \int_Y \nabla_x \exp[-\rho f(x,y)]dy \\ &= \int_Y -\rho \exp[-\rho f(x,y)]\nabla_x f(x,y)dy.\end{aligned}$$

We obtain the conclusion from the above two equations immediately. ■

Note that $\mu_\rho(x,y)$ is positive-valued and $\int_{y\in Y} \mu_\rho(x,y)dy = 1$. Thus, for each x, $\mu_\rho(x,y)$ and $\nabla\gamma_\rho(x)$ can be regarded as the probability density function and the expected value of $\nabla_x f(x,y)$ over Y, respectively. The next theorem gives the expression for the limits $\lim_{\rho\to\infty} \mu_\rho(x,y)$.

Theorem 5.3 *Assume that f is a continuously differentiable function and Y is compact. For any $x \in X$, the solution set $S(x)$ of $\min_{y\in Y} f(x,y)$ is Lebesgue measurable. Furthermore, we have*

$$\lim_{\rho\to\infty} \mu_\rho(x,y) = \begin{cases} m^*(S(x))^{-1}, & y \in S(x), \\ 0, & y \in Y \setminus S(x). \end{cases}$$

Here, $m^*(S(x))^{-1} := +\infty$ if $m^*(S(x)) = 0$.

Proof. Since $S(x)$ is nonempty and closed, it is Lebesgue measurable by Proposition 2.2. From the definition of $V(x)$, we have

$$\exp[-\rho(f(x,y) - V(x))] = 1 \tag{5.1}$$

for any $y \in S(x)$ and $f(x,y) > V(x)$ for any $y \in Y \setminus S(x)$. Hence, $\exp[-\rho(f(x,y)-V(x))]$ is never greater than 1 and approaches 0 as ρ tends to infinity for any $y \in Y \setminus S(x)$. This, together with the Lebesgue dominated convergence theorem, implies

$$\begin{aligned}&\lim_{\rho\to\infty} \int_{Y\setminus S(x)} \exp[-\rho(f(x,z) - V(x))]\, dz \\ &\quad = \int_{Y\setminus S(x)} \lim_{\rho\to\infty} \exp[-\rho(f(x,z) - V(x))]\, dz \\ &\quad = 0.\end{aligned} \tag{5.2}$$

From the definition of $\mu_\rho(x,y)$, we get

$$\begin{aligned}\mu_\rho(x,y) &= \frac{\exp[-\rho(f(x,y) - V(x))]}{\int_Y \exp[-\rho(f(x,z) - V(x))]dz} \\ &= \frac{\exp[-\rho(f(x,y) - V(x))]}{\int_{S(x)} \exp[-\rho(f(x,z) - V(x))]\, dz + \int_{Y\setminus S(x)} \exp[-\rho(f(x,z) - V(x))]\, dz}.\end{aligned} \tag{5.3}$$

(i) If $m^*(S(x)) \neq 0$, it follows from (5.1)–(5.3) that

$$\lim_{\rho\to\infty} \mu_\rho(x,y) = \begin{cases} m^*(S(x))^{-1}, & y \in S(x), \\ 0, & y \in Y \setminus S(x). \end{cases}$$

(ii) If $m^*(S(x)) = 0$, from the above proof process, we can get $\mu_\rho(x,y) \to \infty$ for any $y \in S(x)$. When $y \in Y \setminus S(x)$, let

$$\begin{aligned} Y_1 &:= \{z \in Y : f(x,z) > f(x,y)\}, \\ Y_2 &:= \{z \in Y : f(x,z) = f(x,y)\}, \\ Y_3 &:= \{z \in Y : f(x,z) < f(x,y)\}. \end{aligned}$$

It is easy to get that $S(x) \subseteq Y_3$. By the continuity of $f(x,\cdot)$, we have $m^*(Y_3) \neq 0$ and

$$\mu_\rho(x,y) = \frac{1}{\int_{Y_1+Y_2+Y_3} \exp[-\rho(f(x,z) - f(x,y))]dz}.$$

Therefore, we have

$$\begin{aligned} \lim_{\rho\to\infty} \int_{Y_1} \exp[-\rho(f(x,z) - f(x,y))]dz &= 0, \\ \lim_{\rho\to\infty} \int_{Y_2} \exp[-\rho(f(x,z) - f(x,y))]dz &= m^*(Y_2), \\ \lim_{\rho\to\infty} \int_{Y_3} \exp[-\rho(f(x,z) - f(x,y))]dz &= \infty. \end{aligned}$$

It follows that $\lim_{\rho\to\infty} \mu_\rho(x,y) = 0$ when $y \in Y \setminus S(x)$. This completes the proof. ■

It follows from Danskin's theorem and the continuity of $\nabla_x f(x,y)$ that $\partial V(x)$ is a bounded set for any x. The following theorem shows that the distance between $\nabla\gamma_\rho(z)$ and $\partial V(x)$ approaches 0 when $\rho \to \infty$ and $z \to x$.

Theorem 5.4 *Assume that f is a continuously differentiable function, X and Y are compact sets. For any $x \in X$, we have*

$$\lim_{\rho\to\infty,\, z\to x} \mathrm{dist}(\nabla\gamma_\rho(z), \partial V(x)) = 0.$$

Proof. Since both X and Y are compact and $\nabla_x f(x,y)$ is continuous on $X \times Y$, $\nabla_x f(x,y)$ is uniformly continuous on $X \times Y$. Thus, for any $\epsilon > 0$, there exists $\delta > 0$ such that, for any (z_1, y_1) and (z_2, y_2) satisfying $\|(z_1,y_1) - (z_2,y_2)\| \leq 3\delta$,

$$\|\nabla_x f(z_1,y_1) - \nabla_x f(z_2,y_2)\| \leq \epsilon. \tag{5.4}$$

Due to the fact that $S(x)$ is compact and $\bigcup_{y\in S(x)}(B(y,\delta)\cap Y)\supseteq S(x)$, by letting $\hat{B}(y,\delta)=B(y,\delta)\cap Y$, we get from the Heine-Borel covering theorem that there exist $N>0$ and $y_i\in S(x)$ such that

$$\bigcup_{i=1}^{N}\hat{B}(y_i,\delta)\supseteq S(x).$$

Let $\Omega_1:=\hat{B}(y_1,\delta)$, $\Omega_i:=\hat{B}(y_i,\delta)\setminus\hat{B}(y_i,\delta)\cap(\cup_{j=1}^{i-1}\hat{B}(y_j,\delta))$ for $i=2,\ldots,N$ and $\Omega_{N+1}:=Y\setminus\cup_{i=1}^{N}\Omega_i$. It is obvious that $\Omega_1\cap\cdots\cap\Omega_{N+1}=\emptyset$ and $\cup_{i=1}^{N+1}\Omega_i=Y$.

For any $z\in B(x,\delta)$, let $\lambda_i^z:=\int_{\Omega_i}\mu_\rho(z,y)\,dy$ for $1\le i\le N-1$ and $\lambda_N^z:=\int_{Y\setminus\cup_{i=1}^{N-1}\Omega_i}\mu_\rho(z,y)\,dy$. It follows that $\lambda_i^z\ge 0$ for $1\le i\le N$ and $\sum_{i=1}^{N}\lambda_i^z=1$. Since $f(z,y)$ is continuously differentiable on compact set $X\times\Omega_{N+1}$,

$$\sup_{(z,y)\in X\times\Omega_{N+1}}|f(z,y)-V(z)|$$

is bounded and hence, by Theorem 5.3,

$$\lim_{\rho\to\infty}\sup_{(z,y)\in X\times\Omega_{N+1}}\mu_\rho(z,y)=0.$$

It is easy to see that $\mu_\rho(z,y)$ is uniformly convergent to 0 on $X\times\Omega_{N+1}$. Thus, there exists $\rho_0>0$ such that, for any $(z,y)\in X\times\Omega_{N+1}$ and $\rho>\rho_0$,

$$\|\mu_\rho(z,y)(\nabla_x f(z,y)-\nabla_x f(x,y_N))\|\le\epsilon m^*(Y)^{-1}. \tag{5.5}$$

Therefore, it follows from (5.4) and (5.5) that, for $\rho>\rho_0$ and $z\in B(x,\delta)$,

$$\begin{aligned}
&\Big\|\nabla\gamma_\rho(z)-\sum_{i=1}^{N}\lambda_i^z\nabla_x f(x,y_i)\Big\|\\
=\;&\Big\|\int_Y\mu_\rho(z,y)\nabla_x f(z,y)dy-\sum_{i=1}^{N}\int_{\Omega_i}\mu_\rho(z,y)\nabla_x f(x,y_i)\,dy\\
&-\int_{\Omega_{N+1}}\mu_\rho(z,y)\nabla_x f(x,y_N)\,dy\Big\|\\
\le\;&\sum_{i=1}^{N}\int_{\Omega_i}\|\mu_\rho(z,y)(\nabla_x f(z,y)-\nabla_x f(x,y_i))\|\,dy\\
&+\Big\|\int_{\Omega_{N+1}}\mu_\rho(z,y)(\nabla_x f(z,y)-\nabla_x f(x,y_N))\,dy\Big\|\\
\le\;&N\epsilon\int_Y\|\mu_\rho(z,y)\|\,dy+\int_{\Omega_{N+1}}\epsilon m^*(Y)^{-1}\,dy\\
\le\;&(N+1)\epsilon,
\end{aligned}$$

from which and (2.1) we have

$$\mathrm{dist}(\nabla\gamma_\rho(z), \partial V(x)) \leq (N+1)\epsilon, \qquad \forall \rho > \rho_0,\ \forall z \in B(x,\delta).$$

Since $\epsilon > 0$ is arbitrary, the conclusion follows from the above inequality. ■

The next result reveals the fact that the family of entropy integral functions possesses the gradient consistent property.

Theorem 5.5 *Assume that f is a continuously differentiable function, X and Y are compact sets. Then the family of entropy integral functions satisfies the gradient consistent property. That is, for any $x^* \in X$, we have*

$$\emptyset \neq \limsup_{\rho\to\infty, z\to x^*} \nabla\gamma_\rho(z) \subseteq \partial V(x^*).$$

Proof. By Theorem 5.4, for any $\epsilon > 0$, there exist $\rho_0 > 0$ and $\delta > 0$ such that

$$\mathrm{dist}(\nabla\gamma_\rho(z), \partial V(x^*)) < \epsilon, \qquad \forall \rho > \rho_0,\ \forall z \in B(x^*,\delta).$$

It follows that $\nabla\gamma_\rho(z) \in \partial V(x^*) + \epsilon B(0,1)$ for $\rho > \rho_0$ and $z \in B(x^*,\delta)$. By the integral representation in Theorem 5.2 and the Lebesgue dominated convergence theorem, we have $\lim_{\rho\to\infty} \nabla\gamma_\rho(x^*)$ exists. The compactness of $\partial V(x^*)$ yields

$$\limsup_{\rho\to\infty, z\to x^*} \nabla\gamma_\rho(z) \subseteq \partial V(x^*).$$

This completes the proof. ■

Now we apply the smoothing projected gradient algorithm presented in Section 3 to solve $(\mathrm{VP})_\varepsilon$. To this end, for given $\rho > 0$ and $\lambda > 0$, let

$$G_\rho^\lambda(x,y) := F(x,y) + \frac{\lambda}{2}\left(\sqrt{(f(x,y) - \gamma_\rho(x) - \varepsilon)^2 + \rho^{-1}} + (f(x,y) - \gamma_\rho(x) - \varepsilon)\right). \quad (5.6)$$

The algorithm can be stated as follows:

Algorithm 5.1 *1. Let $\{\beta, \gamma, \sigma_1, \sigma_2\}$ be constants in $(0,1)$ with $\sigma_1 \leq \sigma_2$, $\{\hat{\eta}, \rho_0, \lambda_0\}$ be positive constants, and $\{\sigma, \sigma'\}$ be constants in $(1,\infty)$. Choose an initial point $(x^0, y^0) \in X \times Y$ and set $k := 0$.*

2. Compute the stepsize β^{l_k}, where $l_k \in \{0, 1, 2\cdots\}$ is the smallest number satisfying

$$\begin{aligned} &G_{\rho_k}^{\lambda_k}(P_{X\times Y}[(x^k,y^k) - \beta^{l_k}\nabla G_{\rho_k}^{\lambda_k}(x^k,y^k)]) - G_{\rho_k}^{\lambda_k}(x^k,y^k) \qquad (5.7)\\ &\leq \sigma_1 \nabla G_{\rho_k}^{\lambda_k}(x^k,y^k)^T \left(P_{X\times Y}[(x^k,y^k) - \beta^{l_k}\nabla G_{\rho_k}^{\lambda_k}(x^k,y^k)] - (x^k,y^k)\right) \end{aligned}$$

and $\beta^{l_k} \geq \gamma$, *or*

$$\begin{aligned} & G_{\rho_k}^{\lambda_k}(P_{X\times Y}[(x^k, y^k) - \beta^{l-1}\nabla G_{\rho_k}^{\lambda_k}(x^k, y^k)]) - G_{\rho_k}^{\lambda_k}(x^k, y^k) \\ > \sigma_2 \nabla G_{\rho_k}^{\lambda_k}(x^k, y^k)^T & \left(P_{X\times Y}[(x^k, y^k) - \beta^{l-1}\nabla G_{\rho_k}^{\lambda_k}(x^k, y^k)] - (x^k, y^k)\right). \end{aligned} \tag{5.8}$$

Go to Step 3.

3. *If*

$$\frac{\|P_{X\times Y}[(x^k, y^k) - \beta^{l_k}\nabla G_{\rho_k}^{\lambda_k}(x^k, y^k)] - (x^k, y^k)\|}{\beta^{l_k}} < \hat{\eta}\rho_k^{-1}, \tag{5.9}$$

set $(x^{k+1}, y^{k+1}) := P_{X\times Y}[(x^k, y^k) - \beta^{l_k}\nabla G_{\rho_k}^{\lambda_k}(x^k, y^k)]$ *and go to Step 4. Otherwise, set* $(x^{k+1}, y^{k+1}) := P_{X\times Y}[(x^k, y^k) - \beta^{l_k}\nabla G_{\rho_k}^{\lambda_k}(x^k, y^k)]$, $k := k + 1$, *and go to Step 2.*

4. *If*

$$f(x^{k+1}, y^{k+1}) - \gamma_{\rho_k}(x^{k+1}) - \varepsilon \leq 0 \tag{5.10}$$

and

$$\|P_{X\times Y}[(x^{k+1}, y^{k+1}) - \nabla G_{\rho_k}^{\lambda_k}(x^{k+1}, y^{k+1})] - (x^{k+1}, y^{k+1})\| = 0, \tag{5.11}$$

go to Step 6. Else if (5.10) *holds while* (5.11) *fails, go to Step 5. Otherwise, if* (5.10) *fails, set* $\lambda_{k+1} := \sigma'\lambda_k$ *and go to Step 5.*

5. *Set* $\rho_{k+1} := \sigma\rho_k$, $k = k + 1$, *and go to Step 2.*

6. *If a stopping criterion is satisfied, terminate. Otherwise, go to Step 5.*

A stopping criterion for Algorithm 5.1 can be taken as

$$\left|\frac{\lambda_k}{2}\left(\sqrt{(f(x^{k+1}, y^{k+1}) - \gamma_{\rho_k}(x^{k+1}) - \varepsilon)^2 + \rho_k^{-1}} + (f(x^{k+1}, y^{k+1}) - \gamma_{\rho_k}(x^{k+1}) - \varepsilon)\right)\right| < \epsilon,$$

where $\epsilon > 0$ is a given tolerance. Moreover, note that Assumption 3.1 must hold by the compactness of $X \times Y$.

Suppose that Algorithm 5.1 does not terminate within finite iterations. Then, from Theorem 3.1, Corollary 3.1 and Theorem 3.3, we have the following convergence results immediately.

Theorem 5.6 *Assume that* F *and* f *are continuously differentiable functions,* X *and* Y *are compact and convex sets. Let* $\{(x^k, y^k)\}$ *be a sequence generated by Algorithm 5.1.*

(1) *If* $(x^\varepsilon, y^\varepsilon)$ *is an accumulation point of* $\{(x^k, y^k)\}$ *and the sequence* $\{\lambda_k\}$ *is bounded, then* $(x^\varepsilon, y^\varepsilon)$ *is a stationary point of* $(\mathrm{VP})_\varepsilon$.

(2) *If* $\lim_{k\to\infty}(x^k, y^k) = (x^\varepsilon, y^\varepsilon)$ *and the* ENNAMCQ *holds at* $(x^\varepsilon, y^\varepsilon)$, *then* $(x^\varepsilon, y^\varepsilon)$ *is a stationary point of* $(\mathrm{VP})_\varepsilon$.

(3) *If the* ENNAMCQ *holds for* $(\mathrm{VP})_\varepsilon$ *at any point* $(x, y) \in X \times Y$ *satisfying* $f(x, y) - V(x) - \varepsilon \geq 0$, *then any accumulation point of* $\{(x^k, y^k)\}$ *is a stationary point of* $(\mathrm{VP})_\varepsilon$.

We have tested Algorithm 5.1 on the following two examples.

Example 5.1 Consider the Mirrlees' problem. Note that the solution of Mirrlees' problem does not change if we add the constraint $x, y \in [-2, 2]$ into the problem. Hence, $(\bar{x}, \bar{y}) = (1, 0.9575)$ is the optimal solution to the bilevel programming program

$$\begin{array}{ll} \min & (x-2)^2 + (y-1)^2 \\ \text{s.t.} & x \in [-2, 2],\ y \in S(x), \end{array}$$

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

$$\begin{array}{ll} \min & -x\exp[-(y+1)^2] - \exp[-(y-1)^2] \\ \text{s.t.} & y \in [-2, 2]. \end{array}$$

In our test, we chose the initial point $(x^0, y^0) = (-0.8, -0.8)$ and the parameters $\beta = 0.9,\ \gamma = 0.5,\ \sigma_1 = 0.9,\ \sigma_2 = 0.95,\ \rho_0 = 10,\ \lambda_0 = 10,\ \hat{\eta} = 200,\ \sigma = \sigma' = 10$.

- We first considered the case where $\varepsilon = 0$. The numerical results show that, after finite iterations, the iteration point (x^k, y^k) does not change and equals to $(0.99756, 0.95788)$. Actually, at this point, condition (5.10) can not be satisfied and so the sequence $\{\lambda_k\}$ is unbounded. Hence, Theorem 5.6 can not be used to guarantee the convergence of the algorithm.

- We next considered the case where $\varepsilon > 0$. The results are reported in Table 1, in which $d(x^\varepsilon, y^\varepsilon)$ means the distance between $(x^\varepsilon, y^\varepsilon)$ and the optimal point $(1, 0.9575)$ defined by

$$d(x^\varepsilon, y^\varepsilon) := |x^\varepsilon - 1| + |y^\varepsilon - 0.9575|\,.$$

Table 1: Mirrlees' problem

ε	$(x^\varepsilon, y^\varepsilon)$	$d(x^\varepsilon, y^\varepsilon)$
10^{-2}	(1.00818,0.95647)	9.21e-003
10^{-3}	(1.00055,0.95972)	2.78e-003
10^{-4}	(0.99757,0.95779)	2.71e-003
10^{-5}	(0.99756,0.95787)	2.80e-003

For this example, we observe that the smoothing projected gradient algorithm fails when $\varepsilon = 0$ but succeeds in finding the ε solutions when $\varepsilon > 0$. The numerical results are consistent with the fact that the calmness condition fails for (4.3) (see [26]) but the ENNAMCQ holds for $(\mathrm{VP})_\varepsilon$ at any point $(x, y) \in X \times Y$ satisfying $f(x, y) - V(x) - \varepsilon \geq 0$ (see Section 4).

Example 5.2 [16] Consider

$$\begin{array}{ll} \min & F(x, y) := x + y \\ \text{s.t.} & x \in [-1, 1], \\ & y \in S(x) := \underset{y \in [-1,1]}{\operatorname{argmin}} f(x, y) := \frac{xy^2}{2} - \frac{y^3}{3}. \end{array}$$

The value function of the lower level program can be easily formulated as

$$V(x) = \begin{cases} 0 & \text{if } x \in [\frac{2}{3}, 1], \\ \frac{x}{2} - \frac{1}{3} & \text{if } x \in [-1, \frac{2}{3}), \end{cases}$$

and the solution set is

$$S(x) = \begin{cases} \{0\} & \text{if } x \in (\frac{2}{3}, 1], \\ \{0, 1\} & \text{if } x = \frac{2}{3}, \\ \{1\} & \text{if } x \in [-1, \frac{2}{3}). \end{cases}$$

It is easy to see that the unique optimal solution of the bilevel program is $(\bar{x}, \bar{y}) = (-1, 1)$. In addition, setting $\lambda = 3$, we can verify that $(\bar{x}, \bar{y})$ is a local minimizer of the following problem:

$$\begin{array}{ll} \min & F(x, y) + \lambda(f(x, y) - V(x)) \\ \text{s.t.} & x \in [-1, 1],\ y \in [-1, 1]. \end{array}$$

This means that the original bilevel program is calm at $(\bar{x}, \bar{y})$.

Thus, in our test for this example, we set $\varepsilon = 0$, the initial point $(x^0, y^0) = (-0.7, 0.7)$ and the parameters $\beta = 0.9,\ \gamma = 0.5,\ \sigma_1 = 0.9,\ \sigma_2 = 0.95,\ \sigma = 10,\ \sigma' = 10,\ \hat{\eta} = 3 * 10^8,\ \rho_0 = 100,\ \lambda_0 = 200$ and the given tolerance $\epsilon = 2.50\mathrm{e} - 005$. Fortunately, the algorithm terminated at $(x^k, y^k) = (-1, 1)$ within finite iterations.

6 Conclusions

We have presented an implementable algorithm for constrained optimization problems with a convex set and a nonsmooth constraint. The key idea of the algorithm is to use a smoothing approximation function. We have applied the algorithm to solve the simple bilevel program and its approximate problems. Our algorithm has advantage over other nonsmooth algorithms such as gradient sampling algorithms in that there is no need to solve the lower level program at each iteration. Theoretical and numerical results show that the algorithm may perform well.

Acknowledgements. The authors are grateful to the two anonymous referees for their helpful comments and suggestions. Current address of the first author: School of Management, Shanghai University, Shanghai 200444, China.

References

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

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

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

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

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

[6] 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.

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

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

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

[10] S. Dempe and J. Dutta, *Is bilevel programming a special case of a mathematical program with complementarity constraints?* Math. Program., **131**(2012), 37-48.

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

[12] S.C. Fang and S.Y. Wu, *Solving min-max problems and linear semi-infinite programs*, Comput. Math. Appl., **32**(1996), 87-93.

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

[14] M.B. Lignola and J. Morgan, *Stability of regularized bilevel programming problems*, J. Optim. Theo. Appl., **93**(1997), 575-596.

[15] J. Mirrlees, *The theory of moral hazard and unobservable behaviour: Part I*, Review of Economic Studies, **66**(1999), 3-22.

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

[17] B.S. Mordukhovich, *Variational Analysis and Generalized Differentiation, Vol.1: Basic Theory, Vol.2: Applications*, Springer, 2006.

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

[19] R.T. Rockafellar and R.J.-B. Wets, *Variational Analysis*, Springer, Berlin, 1998.

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

[21] E.M. Stein and R. Shakarchi, *Real Analysis: Measure Theory, Integration, and Hilbert Spaces*, Springer, 2005.

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

[23] J.J. Ye, *Multiplier rules under mixed assumptions of differentiability and Lipschitz continuity*, SIAM J. Control Optim., **39**(2001), 1441-1460.

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

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

[26] 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.

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

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