Journal of Convex Analysis
Vol. 14, No. 2, March 2007

AN EXPLICIT DESCENT METHOD FOR BILEVEL CONVEX OPTIMIZATION*

Mikhail Solodov†

September 12, 2005

ABSTRACT

We consider the problem of minimizing a smooth convex function over the set of constrained minimizers of another smooth convex function. We show that this problem can be solved by a simple and explicit gradient descent type method. Standard constrained optimization is a particular case in this framework, corresponding to taking the lower level function as a penalty of the feasible set. We note that in the case of standard constrained optimization, the method does not require solving any penalization (or other optimization) subproblems, not even approximately, and does not perform projections (although explicit projections onto simple sets can be incorporated).

Key words. convex minimization, bilevel optimization, penalty methods, descent methods.
AMS subject classifications. 90C30, 65K05.

* The author is supported in part by CNPq Grants 300734/95-6(RN) and 471780/2003-0, by PRONEX-Optimization and by FAPERJ.
† Instituto de Matemática Pura e Aplicada, Estrada Dona Castorina 110, Jardim Botânico, Rio de Janeiro, RJ 22460-320, Brazil.
Email: solodov@impa.br

1 Introduction

We consider the *bilevel* problem

$$\begin{array}{ll} \text{minimize} & f_1(x) \\ \text{subject to} & x \in S_2 = \arg\min\{f_2(x) \mid x \in D\}, \end{array} \tag{1.1}$$

where $f_1 : \Re^n \to \Re$ and $f_2 : \Re^n \to \Re$ are smooth convex functions and D is a closed convex subset of $\Re^n$.

The above is a special case of the *mathematical program with generalized equation (or equilibrium) constraint* [11, 8], which is

$$\begin{array}{ll} \text{minimize} & f_1(x) \\ \text{subject to} & x \in \{x \in \Re^n \mid 0 \in F(x) + Q(x)\}, \end{array}$$

where $F : \Re^n \to \Re^n$ and Q is a set-valued mapping from $\Re^n$ to the subsets of $\Re^n$. The bilevel problem (1.1) is obtained by setting $F(x) = f_2'(x)$ and $Q(x) = N_D(x)$, the normal cone of the set D at the point $x \in \Re^n$. In the formulation of the problem considered here, there is only one (decision) variable $x \in \Re^n$, and we are interested in identifying specific solutions of the generalized equation $0 \in F(x)+Q(x)$ (equivalently, of the lower level minimization problem in (1.1)), see [8].

Note that, as a special case, (1.1) contains the standard constrained optimization problem

$$\begin{array}{ll} \text{minimize} & f_1(x) \\ \text{subject to} & x \in \{x \in D \mid Ax = a,\, g(x) \leq 0\}, \end{array} \tag{1.2}$$

where $g : \Re^n \to \Re^m$ is a smooth convex function, A is an $l \times n$ matrix and $a \in \Re^l$. Indeed, (1.2) is obtained from (1.1) by taking $f_2(x) = p(x)$, where p penalizes functional constraints, e.g.,

$$f_2(x) = p(x) = \|Ax - a\|^2 + \|\max\{0,\, g(x)\}\|^2, \tag{1.3}$$

where the maximum is taken coordinate-wise.

In this paper, we show that the bilevel problem (1.1) can be solved by a very simple gradient projection method (where projection is onto the set D), iteratively applied to the parametrized family of functions

$$\psi_\sigma(x) = \sigma f_1(x) + f_2(x), \quad \sigma > 0, \tag{1.4}$$

where σ varies along the iterations. Specifically, if $x^k \in D$ is the current iterate and σ_k is the current parameter, it is enough to make *just one* projected gradient step for ψ_{σ_k} from the point x^k, after which the parameter σ_k can be immediately updated. For convergence of the resulting algorithm to the solution set of (1.1), it should hold that

$$\lim_{k\to\infty} \sigma_k = 0, \quad \sum_{k=0}^{\infty} \sigma_k = +\infty. \tag{1.5}$$

In some ways, our proposal is related to [6], where a proximal point method for (nonsmooth) two-level problem has been considered, and (1.5) is referred to as *slow control*. However, as any proximal method, the method of [6] is *implicit*: it requires solving nontrivial subproblems of minimizing regularized functions ψ_{σ_k} at every iteration, even if approximately. Computationally, proximal point iterations are impractical, unless accompanied by numerically realistic approximation rules (e.g., such as [16, 17]) and specific implementable schemes for satisfying those rules (e.g., such as [18, 14]). By contrast, the method proposed in this paper is completely explicit. Furthermore, it has a very low cost per iteration, especially when projection onto D can be computed in closed form. We emphasize that the latter property can always be achieved in the important case of standard optimization (1.2), by choosing appropriately the constraints to be handled directly via the set D and constraints to be penalized by (1.3). In fact, although D may be of arbitrary structure in our convergence analysis, we prefer to (implicitly) assume that all "hard" constraints in (1.2) are represented by $Ax = a$ and $g(x) \leq 0$, while D is a "simple" set, in the sense that projection onto D can be computed explicitly (e.g., D is defined by bound constraints, such as the nonnegative orthant $\Re^n_+$; D is a ball; or perhaps, the whole space $\Re^n$).

It should be noted that gradient methods are certainly not competitive for many classes of problems. Nevertheless, it is now realized (see, e.g., [1]) that in extremely large-scale problems gradient-based methods are sometimes the only computationally viable choice, as more advanced methods (e.g., those requiring solving linear systems of equations, optimization subproblems, etc.) are simply not applicable at all [1]. The method to be presented can be considered as another step in the direction of revival of simple gradient methods, motivated by extremely large problems. In this respect, we stress once again that the method does not require solving any general optimization subproblems or systems of equations. Furthermore, at least in the case of standard optimization (1.2), it does not perform any projections which are not explicit.

The special case of the optimization problem (1.2) deserves some further comments. We believe that, in this case, our method is closely related to the one in [13], if all the constraints are being penalized. We note, however, that the possibility of treating simple constraints (such as bounds) directly rather than through penalization is a well recognized necessity for efficient computational methods. This feature gives an advantage to our proposal as compared to [13]. It is also interesting to comment on the relation between our method (and that of [13]) and the classical [9, 12] penalty approximation scheme. The penalty scheme consists of solving a sequence of subproblems

$$\begin{array}{ll} \text{minimize} & \psi_\sigma(x) \\ \text{subject to} & x \in D, \end{array} \tag{1.6}$$

where ψ_σ is given by (1.4) with f_2 being the penalty term p, such as (1.3) (in the literature, it is more common to minimize $f_1(x)+\sigma^{-1}p(x)$, but the resulting subproblem is clearly equivalent to (1.6)). As is well-known, under mild assumptions optimal paths

of solutions $x(\sigma)$ of penalized problems (1.6) tend to the solution set of (1.2) as $\sigma \to 0$. We note that the requirement that penalty parameters should tend to zero is, in general, indispensable. Even if nonsmooth penalty is used, to guarantee exactness of the penalty function (i.e., to guarantee that a solution of (1.6), for some fixed $\sigma > 0$, is a solution of the original problem (1.2)), some regularity assumptions on constraints are needed (e.g., see [5, Section 14.4]). No assumptions of this type are made in this paper. The fundamental issue is approximating $x(\sigma_k)$ for some sequence of parameters $\sigma_k \to 0$. It is clear that approximating $x(\sigma_k)$ with precision is computationally expensive and is one of the limitations of basic forms of penalty methods in general. Another drawback is ill-conditioning of subproblems (1.6) for small values of σ. To deal with the first issue, it is attractive to trace the optimal path in a loose (and computationally cheap) manner, while still safeguarding convergence. In a sense, this is what our method does: instead of solving subproblems (1.6) to some prescribed accuracy, it makes just one steepest descent step for ψ_{σ_k} from the current iterate x^k, and immediately updates the parameter. We emphasize that this results in meaningful progress (and ultimately produces iterates converging to solutions of the problem) for arbitrary points x^k, and not just for points close to the optimal path, i.e., points close to $x(\sigma_k)$. We therefore obtain a very simple and computationally cheap algorithm for tracing optimal paths of penalty schemes. Of course, it still does not solve the problem of ill-conditioning of (1.6) in the limit. However, ill-conditioning is more of a problem for sophisticated (Newton-type) methods, where systems of equations need to be solved. For our method, we show that the stepsize stays uniformly bounded away from zero in the limit, which means that the cost of iterations (which is here the cost of the linesearch procedure) does not increase. Thus, in the context of the given method, ill-conditioning does not seem to make iterations increasingly more difficult, although it may affect the speed of convergence in the limit. In any case, the presented proposal can be interesting as a cheap global scheme for approaching the solution set, while more sophisticated techniques can come into play locally, if needed.

For the standard optimization setting (1.2), this paper is also somewhat related to [7], where interior penalty schemes are coupled with continuous-time steepest descent to produce a family of paths converging to solution set. However, concrete numerical schemes in [7] arise from *implicit* discretization and, thus, result in implicit proximal-point iterations, just as in [6]. Nevertheless, it was conjectured in [7] that "an economic algorithm performing a single iteration of some descent method for each value of σ_k could be enough to generate a sequence of iterates converging to a solution of the problem". This is what the presented method does, although we use exterior rather than interior penalties. Finally, we note that the idea of making some descent step for a penalty function and then changing the penalty parameter is certainly not new in itself. For example, globalizations of the sequential quadratic programming [4, 2, 5] and sequential quadratically constrained quadratic programming [10, 15] methods do precisely that. However, in those methods descent directions are computed by solving

optimization subproblems. Also, some regularity of constraints (sometimes assumed in the form of boundedness of the multiplier estimates) is needed for showing convergence.

Our notation is quite standard. By $\langle x, y\rangle$ we denote the inner product of x and y, and by $\|\cdot\|$ the associated norm, where the space is always clear from the context. For a differentiable function ϕ, its gradient is denoted by ϕ'. If D is a closed convex set, $P_D(x)$ stands for the orthogonal projection of the point x onto D, and $\mathrm{dist}(x, D) = \|x - P_D(x)\|$ is the distance from x to D.

We conclude this section with stating some well-known facts, to be used in the sequel.

Theorem 1.1 *Let $D \neq \emptyset$ be a closed convex set.*

(i) It holds that $y = P_D(x)$ if, and only if, $\langle x - y, z - y\rangle \leq 0$ for all $z \in D$.

(ii) A point $\bar{x}$ is a minimizer for a convex function ϕ on the set D if, and only if, $\bar{x} = P_D(\bar{x} - \alpha\phi'(\bar{x}))$, where $\alpha > 0$.

Lemma 1.1 *If ϕ is a differentiable function whose derivatives are Lipschitz-continuous (with modulus $L > 0$) on the set Ω, then*

$$|\phi(y) - \phi(x) - \langle \phi'(x), y - x\rangle| \leq \frac{L}{2}\|y - x\|^2 \quad \forall\, x, y \in \Omega.$$

Lemma 1.2 *If $\{a_k\}$ and $\{b_k\}$ are two sequences of nonnegative real numbers satisfying*

$$a_{k+1} \leq a_k + b_k \;\forall\, k, \quad \sum_{k=0}^{\infty} b_k < +\infty,$$

then the sequence $\{a_k\}$ converges.

2 The algorithm

As already outlined above, the algorithm is very simple. Each iteration is just a step of one of the standard variants of the gradient projection method, e.g., [2, Section 2.3], applied to the function ψ_{σ_k} at the point x^k, with D being the set onto which the iterates are projected.

Algorithm 2.1 *Choose parameters $\bar{\alpha} > 0$, $\theta \in (0, 1)$ and $\eta \in (0, 1)$. Choose starting values $x^0 \in D$ and $\sigma_0 > 0$; set $k := 0$.*
Given x^k, compute $x^{k+1} = z^k(\alpha_k)$, where

$$z^k(\alpha) = P_D(x^k - \alpha\psi'_{\sigma_k}(x^k)), \tag{2.1}$$

and $\alpha_k = \eta^{m_k}\bar{\alpha}$, with m_k being the smallest nonnegative integer m satisfying

$$\psi_{\sigma_k}(z^k(\eta^m\bar{\alpha})) \leq \psi_{\sigma_k}(x^k) + \theta\langle\psi'_{\sigma_k}(x^k), z^k(\eta^m\bar{\alpha}) - x^k\rangle. \tag{2.2}$$

Choose $0 < \sigma_{k+1} \leq \sigma_k$; set $k := k + 1$ and repeat.

We note that there is certain freedom in updating or not the parameter σ_k after every iteration. While our goal is to show that we can update it after a single descent step, note that in principle, we are not obliged to do so ($\sigma_{k+1} = \sigma_k$ is allowed). For convergence, it would be required that σ_k does not go to zero too fast, in the sense of condition (1.5) stated above.

Consider, for the moment, the case of standard optimization (1.2). Condition (1.5) would certainly have been undesirable if imposed on the classical penalty scheme (1.6). Indeed, since σ should be changing relatively slowly, a lot of optimization subproblems (1.6) would need to be solved, making the penalty method even less attractive. In the setting of Algorithm 2.1, however, this does not seem to be such a drawback, since every iteration is extremely cheap. It is only natural that in order to be able to trace the optimal path with such a relaxed precision and simple tools, we should not be jumping too far from the target $x(\sigma_k)$ on the path to the next target $x(\sigma_{k+1})$ as we move along. On the other hand, if σ_k is kept constant over a few iterations, this opens space for a more rapid change in the parameter for the next iteration, while still guaranteeing the second condition in (1.5). This is also intuitively reasonable: if we get closer to the optimal path then the target can be moved further.

Another observation is that it is formally possible that at some iteration k, it may happen that $x^k = z^k(\alpha) = P_D(x^k - \alpha\psi'_{\sigma_k}(x^k))$ for $\alpha = \bar{\alpha}$. In this case, by Theorem 1.1(ii), x^k is a minimizer of ψ_{σ_k} on D (and in fact, $x^k = z^k(\alpha)$ for all $\alpha > 0$). In the setting of penalty schemes, this means that the current iterate is exactly on the optimal path: $x^k = x(\sigma_k)$. Naturally, in this case there is nothing more that needs to be done. We just decrease the parameter and proceed. Note that there is no need to specify this case in Algorithm 2.1 explicitly. Indeed, when x^k is a minimizer of ψ_{σ_k} on D, the stepsize condition (2.2) is trivially satisfied for $m = 0$ (as an equality), so that we declare $x^{k+1} = z^k(\bar{\alpha}) = x^k$ and proceed. This case, however, is very unlikely to occur, since for no iteration k the function ψ_{σ_k} is being minimized on D with any prescribed precision.

3 Convergence Analysis

In our convergence analysis, we assume that the objective function f_1 is bounded below on the set D, i.e.,

$$-\infty < \bar{f}_1 = \inf\{f_1(x) \mid x \in D\}.$$

Since we also assume that the problem is solvable, the function f_2 is automatically bounded below on D, and we define

$$-\infty < \bar{f}_2 = \min\{f_2(x) \mid x \in D\}.$$

Observe that Algorithm 2.1 would generate the same iterates when "applied" to the function

$$\psi_\sigma(x) = \sigma(f_1(x) - \bar{f}_1) + (f_2(x) - \bar{f}_2), \tag{3.1}$$

as when applied to the function $\psi_\sigma(x)$ given by (1.4) (as stated originally). This is because the two functions have the same gradient and the same difference for function values at any two points (hence, the relations in (2.1) and (2.2) do not change). From now on, we consider that the method is "applied" to function $\psi_\sigma(x)$ defined by (3.1) (even though the function from (1.3) is used in reality, of course). This is convenient for the subsequent analysis and should not lead to any confusion.

The proof of the fact that the stepsize procedure is well-defined is essentially standard. We include it here for completeness, and because some intermediate relations will be needed later on.

Proposition 3.1 *Suppose that D is a closed convex set and that f_1 and f_2 are differentiable functions with locally Lipschitz-continuous derivatives around x^k (with modulus $L_k > 0$).*

Then the stepsize procedure of Algorithm 2.1 terminates with some finite integer m_k such that

$$\alpha_k = \eta^{m_k}\bar{\alpha} \geq \min\left\{\bar{\alpha}\,;\, \frac{2(1-\theta)}{(1+\sigma_k)L_k}\right\} > 0.$$

In particular, Algorithm 2.1 is well-defined.

Proof. By Theorem 1.1(i), since $x^k \in D$, for any $\alpha > 0$ it holds that

$$\langle x^k - \alpha\psi'_{\sigma_k}(x^k) - z^k(\alpha), x^k - z^k(\alpha)\rangle \leq 0,$$

implying that

$$\|z^k(\alpha) - x^k\|^2 \leq \alpha\langle\psi'_{\sigma_k}(x^k), x^k - z^k(\alpha)\rangle. \tag{3.2}$$

By the hypothesis, ψ'_{σ_k} is locally Lipschitz-continuous with modulus $(1+\sigma_k)L_k$. We shall assume, for simplicity, that $\bar{\alpha}$ is small enough, so that $z^k(\bar{\alpha})$ belongs to the relevant neighbourhood Ω_k of x^k. Then $z^k(\alpha) \in \Omega_k$ for all $\alpha \leq \bar{\alpha}$. Using Lemma 1.1, we obtain that for all $\alpha \leq \bar{\alpha}$, it holds that

$$\begin{aligned}\psi_{\sigma_k}(z^k(\alpha)) &\leq \psi_{\sigma_k}(x^k) + \langle\psi'_{\sigma_k}(x^k), z^k(\alpha) - x^k\rangle + \frac{(1+\sigma_k)L_k}{2}\|z^k(\alpha) - x^k\|^2 \\ &\leq \psi_{\sigma_k}(x^k) + (1 - L_k(1+\sigma_k)\alpha/2)\langle\psi'_{\sigma_k}(x^k), z^k(\alpha) - x^k\rangle,\end{aligned}$$

where (3.2) was used for the second inequality. The relation above shows that condition (2.2) is guaranteed to be satisfied whenever $1 - L_k(1+\sigma_k)\alpha/2 \leq \theta$. Taking into account also the fact that $\alpha_k \leq \bar{\alpha}$ by construction, we obtain the assertion. ∎

We proceed to prove convergence of the algorithm.

Theorem 3.1 *Let f_1 and f_2 be convex differentiable functions, whose derivatives are Lipschitz-continuous on bounded sets. Suppose that f_1 is bounded below on the closed convex set D, and that the solution set S_1 of problem* (1.1) *is nonempty and bounded.*

Then for any sequence $\{x^k\}$ generated by Algorithm 2.1 satisfying condition (1.5)*, it holds that* $\mathrm{dist}(x^k, S_1) \to 0$ *as* $k \to \infty$.

Proof. By (2.2), it holds that

$$\begin{aligned}\theta\langle\psi'_{\sigma_k}(x^k), x^k - x^{k+1}\rangle &\leq \psi_{\sigma_k}(x^k) - \psi_{\sigma_k}(x^{k+1}) \\ &= \sigma_k(f_1(x^k) - \bar{f}_1) - \sigma_k(f_1(x^{k+1}) - \bar{f}_1) \\ &\quad +(f_2(x^k) - \bar{f}_2) - (f_2(x^{k+1}) - \bar{f}_2).\end{aligned}$$

Summing up the latter inequalities for $k = 0, \ldots, \bar{k}$, we obtain that

$$\begin{aligned}\theta\sum_{k=0}^{\bar{k}}\langle\psi'_{\sigma_k}(x^k), x^k - x^{k+1}\rangle &\leq \sigma_0(f_1(x^0) - \bar{f}_1) + \sum_{k=0}^{\bar{k}-1}(\sigma_{k+1} - \sigma_k)(f_1(x^{k+1}) - \bar{f}_1) \\ &\quad -\sigma_{\bar{k}}(f_1(x^{\bar{k}+1}) - \bar{f}_1) + (f_2(x^0) - \bar{f}_2) - (f_2(x^{\bar{k}+1}) - \bar{f}_2) \\ &\leq \sigma_0(f_1(x^0) - \bar{f}_1) + (f_2(x^0) - \bar{f}_2),\end{aligned}$$

where we have used the facts that, for all k, $f_1(x^k) \geq \bar{f}_1$ and $f_2(x^k) \geq \bar{f}_2$ (because $x^k \in D$), and $0 < \sigma_{k+1} \leq \sigma_k$.

Letting $\bar{k} \to \infty$, we conclude that

$$\sum_{k=0}^{\infty}\langle\psi'_{\sigma_k}(x^k), x^k - x^{k+1}\rangle \leq \theta^{-1}(\sigma_0(f_1(x^0) - \bar{f}_1) + (f_2(x^0) - \bar{f}_2)) < +\infty. \tag{3.3}$$

In particular,

$$\langle\psi'_{\sigma_k}(x^k), x^k - x^{k+1}\rangle \to 0 \text{ as } k \to \infty. \tag{3.4}$$

We next prove that if $\{x^k\}$ is bounded, then all of its accumulation points belong to the set S_2, the set of solutions of the lower-level problem in (1.1) (boundedness itself will be established later, considering certain cases separately).

Taking instead of L_k a uniform Lipschitz constant $L > 0$ (valid for the bounded set containing $\{x^k\}$) in Proposition 3.1, and recalling also that $\sigma_k \leq \sigma_0$, we conclude that

$$\alpha_k \geq \min\left\{\bar{\alpha}\,; \frac{2(1-\theta)}{(1+\sigma_0)L}\right\} = \beta > 0 \quad \forall\, k. \tag{3.5}$$

Using (3.4) and (3.2), we obtain that $(x^k - x^{k+1}) \to 0$, i.e.,

$$x^k - P_D(x^k - \alpha_k(\sigma_k f'_1(x^k) + f'_2(x^k))) \to 0 \text{ as } k \to \infty. \tag{3.6}$$

Let $\hat{x}$ be any accumulation point of $\{x^k\}$. Taking into account the continuity of the projection operator and the facts that $\sigma_k \to 0$ and $0 < \beta \leq \alpha_k \leq \bar{\alpha}$ for all k, and extracting appropriate subsequences (if necessary), we conclude from (3.6) that

$$\hat{x} = P_D(\hat{x} - \hat{\alpha} f_2'(\hat{x})), \quad \hat{\alpha} > 0.$$

By Theorem 1.1(ii), this means that $\hat{x}$ is a minimizer of f_2 on the set D, i.e., $\hat{x} \in S_2$. We have established therefore that whenever $\{x^k\}$ is bounded, all its accumulation points belong to the set S_2.

Take any $\bar{x} \in S_1 \neq \emptyset$. By the convexity of ψ_{σ_k}, we obtain that

$$\begin{aligned} \langle \psi'_{\sigma_k}(x^k), \bar{x} - x^k \rangle &\leq \psi_{\sigma_k}(\bar{x}) - \psi_{\sigma_k}(x^k) \\ &= \sigma_k(f_1(\bar{x}) - \bar{f}_1) + (f_2(\bar{x}) - \bar{f}_2) \\ &\quad -\sigma_k(f_1(x^k) - \bar{f}_1) - (f_2(x^k) - \bar{f}_2) \\ &\leq \sigma_k(f_1(\bar{x}) - f_1(x^k)), \end{aligned} \tag{3.7}$$

where we used the facts that $f_2(\bar{x}) \leq f_2(x^k)$, since $\bar{x} \in S_1 \subset S_2$ and $x^k \in D$.

We further have that

$$\begin{aligned} \|x^{k+1} - \bar{x}\|^2 &= \|x^k - \bar{x}\|^2 + 2\langle x^{k+1} - x^k, x^k - \bar{x} \rangle + \|x^{k+1} - x^k\|^2 \\ &= \|x^k - \bar{x}\|^2 - \|x^{k+1} - x^k\|^2 + 2\langle x^{k+1} - x^k, x^{k+1} - \bar{x} \rangle. \end{aligned} \tag{3.8}$$

Note that

$$\begin{aligned} \langle x^{k+1} - x^k, x^{k+1} - \bar{x} \rangle &= \langle x^{k+1} - x^k + \alpha_k \psi'_{\sigma_k}(x^k), x^{k+1} - \bar{x} \rangle - \alpha_k \langle \psi'_{\sigma_k}(x^k), x^{k+1} - \bar{x} \rangle \\ &\leq -\alpha_k \langle \psi'_{\sigma_k}(x^k), x^{k+1} - \bar{x} \rangle \\ &= \alpha_k \langle \psi'_{\sigma_k}(x^k), x^k - x^{k+1} \rangle + \alpha_k \langle \psi'_{\sigma_k}(x^k), \bar{x} - x^k \rangle \\ &\leq \bar{\alpha} \langle \psi'_{\sigma_k}(x^k), x^k - x^{k+1} \rangle + \alpha_k \sigma_k (f_1(\bar{x}) - f_1(x^k)), \end{aligned}$$

where the first inequality follows from Theorem 1.1(i), and the last is by (3.7) and the facts that $\langle \psi'_{\sigma_k}(x^k), x^k - x^{k+1} \rangle \geq 0$ (by (3.2)) and $\alpha_k \leq \bar{\alpha}$.

Combining the last relation with (3.8), we obtain

$$\|x^{k+1} - \bar{x}\|^2 \leq \|x^k - \bar{x}\|^2 + 2\bar{\alpha} \langle \psi'_{\sigma_k}(x^k), x^k - x^{k+1} \rangle + 2\alpha_k \sigma_k (f_1(\bar{x}) - f_1(x^k)). \tag{3.9}$$

We next consider separately the following two possible cases:

Case 1. There exists k_0 such that $f_1(\bar{x}) \leq f_1(x^k)$ for all $k \geq k_0$.

Case 2. For each k, there exists $k_1 \geq k$ such that $f_1(\bar{x}) > f_1(x^{k_1})$.

Case 1. For $k \geq k_0$, we obtain from (3.9) that

$$\|x^{k+1} - \bar{x}\|^2 \leq \|x^k - \bar{x}\|^2 + 2\bar{\alpha}\langle \psi'_{\sigma_k}(x^k), x^k - x^{k+1}\rangle.$$

Recalling (3.3) and using Lemma 1.2, we conclude that $\{\|x^k - \bar{x}\|^2\}$ converges. Hence, $\{x^k\}$ is bounded.

We next show that $\liminf_{k\to\infty} f_1(x^k) = f_1(\bar{x})$. Assume the contrary, i.e., that there exists $\varepsilon > 0$ such that $f_1(\bar{x}) \leq f_1(x^k) - \varepsilon$ for all $k \geq k_2$. Recalling (3.5) (which holds by boundedness of $\{x^k\}$, already established), we then obtain from (3.9) that for $k > k_2$, it holds that

$$\begin{aligned} \|x^{k+1} - \bar{x}\|^2 &\leq \|x^k - \bar{x}\|^2 + 2\bar{\alpha}\langle \psi'_{\sigma_k}(x^k), x^k - x^{k+1}\rangle - 2\beta\varepsilon\sigma_k \\ &\leq \|x^{k_2} - \bar{x}\|^2 + 2\bar{\alpha}\sum_{i=k_2-1}^{k} \langle \psi'_{\sigma_i}(x^i), x^i - x^{i+1}\rangle - 2\beta\varepsilon\sum_{i=k_2-1}^{k} \sigma_i. \end{aligned}$$

Passing onto the limit when $k \to \infty$ in the latter relation, we obtain

$$2\beta\varepsilon\sum_{i=k_2-1}^{\infty} \sigma_i \leq \|x^{k_2} - \bar{x}\|^2 + 2\bar{\alpha}\sum_{i=k_2-1}^{\infty} \langle \psi'_{\sigma_i}(x^i), x^i - x^{i+1}\rangle,$$

which is a contradiction, due to (3.3) and (1.5).

Hence, $\liminf_{k\to\infty} f_1(x^k) = f_1(\bar{x})$. Since $\{x^k\}$ is bounded, it must have an accumulation point $\hat{x}$ such that $f_1(\hat{x}) = f_1(\bar{x})$. Taking into account that $\hat{x} \in S_2$ (as already established above), this means that $\hat{x} \in S_1$. Now choosing $\bar{x} = \hat{x}$ in the preceding analysis, we obtain that $\{\|x^k - \hat{x}\|\}$ converges. Since it has a subsequence converging to zero, this means that the whole sequence $\{\|x^k - \hat{x}\|\}$ converges to zero, i.e., $\{x^k\} \to \hat{x} \in S_1$.

Case 2. For each k, define

$$i_k = \max\{i \leq k \mid f_1(\bar{x}) > f_1(x^i)\}.$$

In the case under consideration, it holds that $i_k \to \infty$ when $k \to \infty$.

We first show that $\{x^{i_k}\}$ is bounded. Observe that

$$\begin{aligned} S_1 &= \{x \in S_2 \mid f_1(x) \leq f_1(\bar{x})\} \\ &= \{x \in D \mid \max\{f_2(x) - \bar{f}_2, f_1(x) - f_1(\bar{x})\} \leq 0\}. \end{aligned}$$

By assumption, the set S_1 is nonempty and bounded. Therefore, the convex function

$$\phi : D \to \Re, \quad \phi(x) = \max\{f_2(x) - \bar{f}_2, f_1(x) - f_1(\bar{x})\}$$

has a particular level set $\{x \in D \mid \phi(x) \leq 0\}$ which is nonempty and bounded. It follows that all level sets of ϕ are bounded (e.g., [3, Proposition 2.3.1]), i.e., $L(c) = \{x \in D \mid \phi(x) \leq c\}$ is bounded for any $c \in \Re$.

Since $f_1(x) - \bar{f}_1 \geq 0$ for all $x \in D$ and $\sigma_{k+1} \leq \sigma_k$, it holds that $\psi_{\sigma_{k+1}}(x) \leq \psi_{\sigma_k}(x)$ for all $x \in D$. Hence,

$$0 \leq \psi_{\sigma_{k+1}}(x^{k+1}) \leq \psi_{\sigma_k}(x^{k+1}) \leq \psi_{\sigma_k}(x^k),$$

where the third inequality follows from (2.2). The above relations show that $\{\psi_{\sigma_k}(x^k)\}$ is nonincreasing and bounded below. Hence, it converges. It then easily follows that $\{f_2(x^k) - \bar{f}_2\}$ is bounded (because both terms in $\psi_{\sigma_k}(x^k) = \sigma_k(f_1(x^k) - \bar{f}_1) + (f_2(x^k) - \bar{f}_2)$ are nonnegative).

Fix any $c \geq 0$ such that $f_2(x^k) - \bar{f}_2 \leq c$ for all k. Since $f_1(x^{i_k}) - f_1(\bar{x}) < 0 \leq c$, we have that $x^{i_k} \in L(c)$, which is a bounded set. This shows that $\{x^{i_k}\}$ is bounded.

By the definition of i_k, it holds that

$$f_1(\bar{x}) \leq f_1(x^i), \quad i = i_k + 1, \ldots, k \quad (\text{if } k > i_k).$$

Hence, from (3.9), we have that

$$\|x^{i+1} - \bar{x}\|^2 \leq \|x^i - \bar{x}\|^2 + 2\bar{\alpha}\langle \psi'_{\sigma_i}(x^i), x^i - x^{i+1}\rangle, \quad i = i_k + 1, \ldots, k.$$

Therefore, for any k, it holds that

$$\begin{aligned} \|x^k - \bar{x}\|^2 &\leq \|x^{i_k} - \bar{x}\|^2 + 2\bar{\alpha} \sum_{i=i_k+1}^{k-1} \langle \psi'_{\sigma_i}(x^i), x^i - x^{i+1}\rangle \\ &\leq \|x^{i_k} - \bar{x}\|^2 + 2\bar{\alpha} \sum_{i=i_k+1}^{\infty} \langle \psi'_{\sigma_i}(x^i), x^i - x^{i+1}\rangle. \end{aligned} \tag{3.10}$$

Recalling that $i_k \to \infty$, by (3.3) we have that

$$\sum_{i=i_k+1}^{\infty} \langle \psi'_{\sigma_i}(x^i), x^i - x^{i+1}\rangle \to 0 \text{ as } k \to \infty. \tag{3.11}$$

Taking also into account boundedness of $\{x^{i_k}\}$, (3.10) implies that the whole sequence $\{x^k\}$ is bounded.

Since all accumulation points of $\{x^k\}$ belong to S_2 (as established above), and for any accumulation point $\hat{x}$ of $\{x^{i_k}\}$ we have that $f_1(\bar{x}) \geq f_1(\hat{x})$, it must be the case that all accumulation points of $\{x^{i_k}\}$ are solutions of the problem. In particular,

$$\operatorname{dist}(x^{i_k}, S_1) \to 0 \text{ as } k \to \infty. \tag{3.12}$$

For each k, define $\bar{x}^k = P_{S_1}(x^{i_k})$. Using (3.10) with $\bar{x} = \bar{x}^k$ gives

$$\begin{aligned} \operatorname{dist}(x^k, S_1)^2 &\leq \|x^k - \bar{x}^k\|^2 \\ &\leq \operatorname{dist}(x^{i_k}, S_1)^2 + 2\bar{\alpha} \sum_{i=i_k+1}^{\infty} \langle \psi'_{\sigma_i}(x^i), x^i - x^{i+1}\rangle. \end{aligned}$$

Passing onto the limit in the latter relation as $k \to \infty$, and using (3.11) and (3.12), we obtain that $\operatorname{dist}(x^k, S_1) \to 0$. ∎

References

[1] A. Ben-Tal, T. Margalit, and A. Nemirovski. The ordered subsets mirror descent optimization method with applications to tomography. *SIAM Journal on Optimization*, 12:79–108, 2001.

[2] D.P. Bertsekas. *Nonlinear programming*. Athena Scientific, Belmont, Massachusetts, 1995.

[3] D.P. Bertsekas. *Convex Analysis and Optimization*. Athena Scientific, Belmont, Massachusetts, 2003.

[4] B.T. Boggs and J.W. Tolle. Sequential quadratic programming. *Acta Numerica*, 4:1–51, 1996.

[5] J.F. Bonnans, J.Ch. Gilbert, C. Lemaréchal, and C. Sagastizábal. *Numerical Optimization: Theoretical and Practical Aspects*. Springer–Verlag, Berlin, Germany, 2003.

[6] A. Cabot. Proximal point algorithm controlled by a slowly vanishing term: Applications to hierarchial minimization. *SIAM Journal on Optimization*, 15:555–572, 2005.

[7] R. Cominetti and M. Courdurier. Coupling general penalty schemes for convex programming with the steepst descent and the proximal point algorithm. *SIAM Journal on Optimization*, 13:745–765, 2002.

[8] M. Kočvara and J.V. Outrata. Optimization problems with equilibrium constraints and their numerical solution. *Mathematical Programming*, 101:119–149, 2004.

[9] A.V. Fiacco and G.P. McCormick. *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*. John Wiley & Sons, New York, 1968.

[10] M. Fukushima, Z.-Q. Luo, and P. Tseng. A sequential quadratically constrained quadratic programming method for differentiable convex minimization. *SIAM Journal on Optimization*, 13:1098–1119, 2003.

[11] Z.-Q. Luo, J.-S. Pang, and D. Ralph. *Mathematical programs with equilibrium constraints*. Cambridge University Press, 1996.

[12] B. T. Polyak. *Introduction to Optimization*. Optimization Software, Inc., Publications Division, New York, 1987.

[13] A.V. Rodionov and A.A. Tret'yakov. A method for solving the convex programming problem. *Voprosy Kibernet.*, (136):111–117, 1988. In Russian.

[14] C.A. Sagastizábal and M.V. Solodov. On the relation between bundle methods for maximal monotone inclusions and hybrid proximal point algorithms. In D. Butnariu, Y. Censor, and S. Reich, editors, *Inherently Parallel Algorithms in Feasibility and Optimization and their Applications*, volume 8 of *Studies in Computational Mathematics*, pages 441–455. Elsevier Science B.V., 2001.

[15] M.V. Solodov. On the sequential quadratically constrained quadratic programming methods. *Mathematics of Operations Research*, 29:64–79, 2004.

[16] M.V. Solodov and B.F. Svaiter. A hybrid projection–proximal point algorithm. *Journal of Convex Analysis*, 6:59–70, 1999.

[17] M.V. Solodov and B.F. Svaiter. Error bounds for proximal point subproblems and associated inexact proximal point algorithms. *Mathematical Programming*, 88:371–389, 2000.

[18] M.V. Solodov and B.F. Svaiter. A truly globally convergent Newton-type method for the monotone nonlinear complementarity problem. *SIAM Journal on Optimization*, 10:605–625, 2000.