Optimization and Engineering, 6, 257–271, 2005

Inexact Proximal Point Algorithms and Descent Methods in Optimization

CARLOS HUMES JR.*
PAULO J. S. SILVA[†,‡]
Department of Computer Science, University of São Paulo
email: chumes@usp.br
email: rsilva@ime.usp.br

Received November 21, 2000; Revised August 4, 2004

Abstract. Proximal point methods have been used by the optimization community to analyze different algorithms like multiplier methods for constrained optimization, and bundle methods for nonsmooth problems. This paper aims to be an introduction to the theory of proximal algorithms borrowing ideas from descent methods for unconstrained optimization. This new viewpoint allows us to present a simple and natural convergence proof. We also improve slightly the results from Solodov and Svaiter (1999).

Keywords: proximal methods, convex programming, monotone operators

1. Introduction

Proximal point methods are tools for the solution of a wide class of problems whose mathematical representations may range from Nonlinear Programming models to the computation of fixed points, or zeroes, for some classes of operators.

These methods were originally introduced by Martinet in the context of fixed points and variational inequalities (Martinet, 1970, 1972). Afterwards, Rockafellar presented the proximal algorithm to the optimization community and uncovered their tight relation to multiplier methods for constrained optimization (Rockafellar, 1976a, 1976b). Henceforth proximal methods have continually appeared in the nonlinear programming literature, with a perceptible increase in the last decade. Many authors have shown that proximal methods can be used as a framework to better understand many algorithms like splitting methods used to exploit parallelism (Eckstein and Bertsekas, 1992), bundle methods for nonsmooth optimization (Auslender, 1987), and multiplier methods. Just in that last area, a recent survey lists dozens of references (Iusem, 1999). In spite of the increase in the number of papers and computational reports on proximal point methods, there are few didactical references on the subject and usually they correspond to reasonably advanced texts like (Bertsekas, 1996).

*Supported by CNPq and PRONEX Optimization.
†Supported by FAPESP grant 96/09939-0, CNPq and PRONEX Optimization.
‡Corresponding author.

The objective of this paper is twofold. First, we show how to interpret the proximal point algorithm as a descent method whenever it is applied to optimization problems. This goal is supported by three reasons: the conceptual simplicity of descent methods, their familiarity to optimization practitioners, and the easy way this point of view can handle questions like realistic stopping rules for the proximal subproblems. This approach is used to present a very simple convergence proof for Solodov-Svaiter's hybrid projection-proximal algorithm (Solodov and Svaiter, 1999) and to show the new result that no projection is needed in the optimization case. The report of the last result is our second objective.

In order to present both a common language and a motivation for the study of this area, we start our exposition by an introductory section on proximal methods. Readers familiar with such methods may skip this section. Next, we present the descent view for proximal methods and we show how it simplifies the analysis of the proximal algorithm for optimization problems. The nonsmooth optimization case is analyzed, as it is the natural setting for multiplier methods. Finally, we study zeroes of operators along with the projection step required for convergence in this case.

2. The proximal point algorithm

Let us consider the unconstrained optimization problem

$$\begin{aligned} &\text{minimize} \quad f(x) \\ &\quad \text{s.t.} \quad x \in \mathbb{R}^n, \end{aligned} \tag{1}$$

where $f : \mathbb{R}^n \mapsto \mathbb{R}$ is convex and twice continuously differentiable. A differentiable function is convex if and only if

$$\forall x, y \in \mathbb{R}^n, \quad f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle.$$

Then a necessary and sufficient condition for x to be a minimizer of f is

$$\nabla f(x) = 0.$$

Solving the above system of nonlinear equations by Newton's method is probably the best known algorithm for unconstrained optimization problems. Its iteration is given by

$$h^k \stackrel{\text{def}}{=} -\nabla^2 f(x^k)^{-1} \nabla f(x^k); \quad x^{k+1} \stackrel{\text{def}}{=} x^k + h^k.$$

If the initial point is close to a minimizer of f with a positive-definite (and hence nonsingular) Hessian, then Newton's method converges quadratically to such minimizer. However, the convergence is not global and the method may face numerical instability if the Hessian is nearly singular at the solution.

To overcome such difficulties, many strategies have been proposed. Among others, Dennis and Schnabel recommended in their classical book (Dennis and Schanbel, 1996) the use of a modified Newton's method based on a slightly different system of linear equations:

$$h^k \stackrel{\text{def}}{=} -(H^k)^{-1}\nabla f(x^k); \quad H^k \stackrel{\text{def}}{=} \nabla^2 f(x^k) + \alpha_k I,$$

where α_k is large enough to make H^k safely positive definite. This change assures both that the step is taken towards a good descent direction[1] and that the system is well conditioned.

The modified Newton's method can be interpreted in two equivalent viewpoints. The first one is that the next iterate is restricted to a circular trust region around the current point. The other perspective is to consider that the modified algorithm computes an iteration of the Newton's method for the regularized problem

$$\begin{aligned} \text{minimize} \quad & f(x) + \frac{\alpha_k}{2}\|x - x^k\|^2 \\ \text{s.t} \quad & x \in I\!R^n. \end{aligned} \tag{2}$$

This may be verified by direct calculations. Observe that the regularized problem has nicer properties than (1), like uniqueness of solutions and better conditioned Hessians.

Proximal point methods follow this regularization idea to the extreme: instead of taking a single Newton step, the classical proximal point algorithm[2] asks for a full solution of (2). That is, it defines its iterative step by

$$x^{k+1} \stackrel{\text{def}}{=} \operatorname*{argmin}_{x \in I\!R^n} \left\{ f(x) + \frac{\alpha_k}{2}\|x - x^k\|^2 \right\}.$$

Note that, if we apply the proximal algorithm as above, the computation of the full proximal step can be computationally expensive. Therefore, it is natural to study means to tolerate errors in the proximal steps, leading to inexact proximal methods.

Following this line, the modified Newton's method proposed by Dennis and Schnabel can be understood as an inexact proximal method where the required precision should be achieved after a single Newton step. In this spirit, Solodov and Svaiter (1999) showed that it is possible to couple Newton's method with a proximal algorithm to obtain a *globally* convergent algorithm for nonlinear equations.[3]

Proximal methods have a wider applicability in Nonlinear Programming than outlined above. For instance, the classical method of multipliers (Bertsekas, 1995, 1996) is equivalent to the proximal algorithm applied to the dual of a given constrained optimization problem (Rockafellar, 1976b).

At first, this approach may seem odd and worthless. The dual objective is very difficult to compute: it is the minimum value of the Lagrangian function with a fixed multiplier. Fortunately, it turns out that a proximal step is equivalent to the unconstrained minimization of a modified Lagrangian function, called the augmented Lagrangian. This fact is remarkable since the proximal steps happen to be much easier to solve than the original constrained problem. They can be carried out by standard unconstrained solvers. The convergence of multiplier methods can then be readily derived from the convergence properties of proximal

methods. We leave the proof of this assertion, and the detailed formulas needed to implement the method of multipliers, to the Appendix.

Finally, we recall that the dual problem is always convex. But its objective function is not necessarily differentiable, or even finite, everywhere. This is the reason why we must deal with such questions in Section 4.

We can now proceed to study the classical proximal point algorithm.

3. The descent property

Let us consider the unconstrained optimization problem

$$\begin{aligned} &\text{minimize} \quad f(x) \\ &\quad \text{s.t.} \quad x \in \mathbb{R}^n, \end{aligned} \tag{P1}$$

where $f : \mathbb{R}^n \mapsto \mathbb{R}$ is convex and continuously differentiable.

Many algorithms used to solve (P1) generate a minimizing sequence, i.e. a sequence $\{x^k\}$ such that $\{f(x^k)\}$ is decreasing. Convergence towards an optimal solution is then proved by showing that the objective function is decreasing fast enough.[4]

Suppose that we want to relate the decrease of f on a given sequence, $\{x^k\}$, and the norm of the gradients of f on this same sequence. This may be done using the fundamental inequality

$$f(x^k) \geq f(x^{k+1}) + \langle g^{k+1}, x^k - x^{k+1} \rangle, \tag{3}$$

where g^{k+1} denotes $\nabla f(x^{k+1})$. This is equivalent to

$$f(x^k) \geq f(x^{k+1}) \ + \|g^{k+1}\| \|x^k - x^{k+1}\| \cos\theta, \tag{4}$$

where θ is the angle between g^{k+1} and $x^k - x^{k+1}$, see Figure 1.

Then x^{k+1} is surely "better" than the previous point whenever θ is acute. If we consider the triangle $[x^k, x^{k+1}, x^{k+1} + g^{k+1}]$, shown in Figure 1, it is easy to see that θ will be acute if the side opposite to it is not the largest one.

One way to ensure such a property is to focus on sequences that verify the following *acceptance criterion*:

$$\|g^{k+1} + (x^{k+1} - x^k)\| \leq \sigma \max\{\|g^{k+1}\|, \|x^{k+1} - x^k\|\}, \tag{AC1}$$

where σ is a positive number smaller than 1. Below we will show that the proximal point method generates a sequence that satisfies (AC1).

We start our analysis by studying some simple properties of the triangle $[x^k, x^{k+1}, x^{k+1} + g^{k+1}]$.

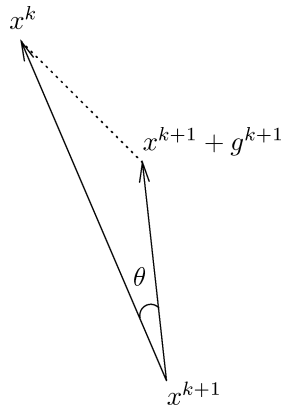

Figure 1. The θ angle.

Lemma 1. *If* $(AC1)$ *holds true, then*

$$0 \leq \sin\theta \leq \sigma \ \ (\textit{and thus} \ \cos\theta \geq \sqrt{1-\sigma^2}).$$

Proof: Let M denote the length of the largest side of the triangle ($M \stackrel{\text{def}}{=} \max\{\|g^{k+1}\|, \|x^{k+1} - x^k\|\}$) and let α be its opposite angle. Then, from the sine rule:

$$\frac{M}{\sin\alpha} = \frac{\|g^{k+1} + x^{k+1} - x^k\|}{\sin\theta}.$$

Therefore,

$$\frac{\sin\theta}{\sin\alpha} = \frac{\|g^{k+1} + x^{k+1} - x^k\|}{M} \leq \sigma,$$

and the result follows. □

Note that (AC1) not only implies that θ is acute; it actually asserts that θ is bounded away from $\frac{\pi}{2}$.

The acceptance criterion also implies that the ratio between $\|g^{k+1}\|$ and $\|x^{k+1} - x^k\|$ is well behaved.

Lemma 2. *Let* $M \stackrel{\text{def}}{=} \max\{\|g^{k+1}\|, \|x^{k+1} - x^k\|\}$ *and* m *be the minimum of these two numbers. Then, if* $(AC1)$ *is satisfied,*

$$(1-\sigma)M \leq m \leq M.$$

Proof: Clearly,

$$M - m \leq \|x^k - x^{k+1} - g^{k+1}\| \leq \sigma M \Rightarrow$$
$$(1-\sigma)M \leq m \leq M.$$

□

We have now all the ingredients to derive a very useful bound on the decrease of f on any sequence satisfying (AC1):

$$\begin{aligned} f(x^k) &\geq f(x^{k+1}) + \|g^{k+1}\|\|x^k - x^{k+1}\|\cos\theta \\ &= f(x^{k+1}) + mM\cos\theta \\ &\geq f(x^{k+1}) + mM\sqrt{1-\sigma^2} \\ &\geq f(x^{k+1}) + (1-\sigma)\sqrt{1-\sigma^2}M^2 \\ &\geq f(x^{k+1}) + (1-\sigma)\sqrt{1-\sigma^2}\|g^{k+1}\|^2. \end{aligned} \tag{5}$$

This inequality will be the key to prove the convergence of the proximal algorithm:

Algorithm 1. *Inexact Proximal Point Algorithm for Differentiable Functions*

Let $f : \mathbb{R}^n \mapsto (-\infty, \infty]$ be a continuously differentiable convex function. Let σ be a number in $[0, 1)$.

1. *Initialization*: *Let x^1 be any point in $\mathbb{R}^n$*;
2. *Iteration*: *Find x^{k+1} such that, for $g^{k+1} \stackrel{\text{def}}{=} \nabla f(x^{k+1})$ and $e^k \stackrel{\text{def}}{=} g^{k+1} + x^{k+1} - x^k$,*

$$\|e^k\| \leq \sigma \max\{\|g^{k+1}\|, \|x^{k+1} - x^k\|\}.$$

It is important to observe that the acceptance condition presented in the iterative step is exactly (AC1). Moreover, if we approximately minimize the regularized function

$$f(x) + \frac{1}{2}\|x - x^k\|^2$$

with enough accuracy, we can find a point that satisfies this condition.

Applying (5) recursively to a sequence computed by the inexact proximal algorithm it follows that

$$f(x^1) \geq f(x^k) + (1-\sigma)\sqrt{1-\sigma^2}\sum_{j=2}^{k}\|g^j\|^2. \tag{6}$$

Therefore, the sequence $\left\{f(x^k)\right\}$ is nonincreasing, being either unbounded below or convergent to some real value. In the later case it is clear by (6) that $\nabla f(x^k) = g^k \to 0$. We can now state:

Theorem 1. *Assume that f is convex, continuously differentiable, and bounded below. Let $\{x^k\}$ be any sequence generated by the inexact proximal point algorithm. Then all of its cluster points are solutions to* $(P1)$.

Proof: The discussion following the inequality (5) showed that $\nabla f(x^k) \to 0$. Hence, if $\bar{x}$ is an accumulation point of $\{x^k\}$, the continuity of ∇f implies that $\nabla f(\bar{x}) = 0$. It follows that $\bar{x}$ is a minimizer of f. □

4. The nonsmooth case

In this section, we will further explore the convexity of f. This allows us to drop the differentiability assumptions from Section 3. We will also introduce a parameter to control the level of regularization.

Let us consider the unconstrained optimization problem:

$$\begin{aligned} &\text{minimize} \quad f(x) \\ &\quad \text{s.t} \quad x \in I\!R^n, \end{aligned} \tag{P2}$$

where $f : I\!R^n \mapsto (-\infty, +\infty]$ is a lower semi-continuous convex function but not necessarily differentiable, or even finite, everywhere. Moreover we assume that f is *proper*, that is it must be finite at least at one point.

What can we use to replace the gradient of f? The natural choice, found in the convex analysis literature, is the subdifferential operator, ∂f. We say that a vector g is a subgradient of f at x, $g \in \partial f(x)$, if

$$\forall y \in I\!R^n, \quad f(y) \geq f(x) + \langle g, y - x \rangle.$$

This inequality says that a subgradient defines an affine lower bound for f whose value at x is exactly $f(x)$. If f is differentiable at x then this affine bound is unique and defined by the gradient. However, it may not be unique, as for $|x|$ at the origin, or it may even not exist at some points. In any case, there is an one-to-one relation between lower affine bounds and subgradients. For more details see Hiriart-Urruty and Lemarechal (2002). Therefore, if $g^{k+1} \in \partial f(x^{k+1})$, the inequality (3) still holds. Following the sequence of inequalities presented in the last section, we may derive again the relation given by (5):

$$f(x^k) \geq f(x^{k+1}) + (1 - \sigma)\sqrt{1 - \sigma^2}\|g^{k+1}\|^2.$$

Once more, using this inequality to motivate an algorithm, we devise an iterative process where a better point x^{k+1} is computed from x^k by minimizing approximately:

$$f(x) + \frac{1}{2}\|x - x^k\|^2.$$

Note that the Euclidean norm above regularizes the original objective function, guaranteeing strong convexity and, therefore, improving the convergence properties of standard

minimization algorithms. Nevertheless, it forces the next iterate to remain close to the actual point and we may want to reduce this effect. To achieve this we introduce a regularization parameter, α_k, and we minimize approximately

$$f(x) + \frac{\alpha_k}{2}\|x - x^k\|^2.$$

A small value for α_k gives more liberty to explore regions far from x^k, while a large α_k ensures that x^{k+1} remains close to x^k.

Now, it is natural to introduce the following proximal method:

Algorithm 2. *Inexact Proximal Point Algorithm*

Let $f : \mathbb{R}^n \mapsto (-\infty, \infty]$ be a proper convex function. Choose $\sigma \in [0, 1)$ and $\bar{\alpha} > 0$.

1. *Initialization*: *Let x^1 be any point in $\mathbb{R}^n$*;
2. *Iteration*: *Let $0 < \alpha_k \leq \bar{\alpha}$. Find x^{k+1} such that, for a $g^{k+1} \in \partial f(x^{k+1})$ and $e^k \stackrel{\text{def}}{=} g^{k+1} + \alpha_k(x^{k+1} - x^k)$*

$$\|e^k\| \leq \sigma \max\{\|g^{k+1}\|, \alpha_k\|x^{k+1} - x^k\|\} \tag{AC2}$$

holds.

Once again, a point that satisfies (AC2) may be computed if we approximately minimize

$$f(x) + \frac{\alpha_k}{2}\|x - x^k\|^2.$$

Theorem 2. *Assume that f is convex, lower semi-continuous, proper, and bounded from below. Let $\{x^k\}$ be any sequence generated by the inexact proximal point algorithm. Then all of its cluster points are solutions to $(P2)$.*

Proof: Using a reasoning analog to the last section, we may define

$$M \stackrel{\text{def}}{=} \max\{\|g^{k+1}\|,\ \alpha^k\|x^{k+1} - x^k\|\},\ \ m \stackrel{\text{def}}{=} \min\{\|g^{k+1}\|,\ \alpha^k\|x^{k+1} - x^k\|\}.$$

Then we may follow the steps that took us to (5) and, using the acceptance criterion (AC2), we can conclude that a small variation of this inequality holds:

$$f(x^k) \geq f(x^{k+1}) + \frac{1}{\alpha_k}\sqrt{1 - \sigma^2}(1 - \sigma)\|g^{k+1}\|^2.$$

Finally, as $\{\alpha^k\}$ is bounded above, it follows that, if $\{f(x^k)\}$ is bounded below, then $g^k \to 0$. As the graph of the subdifferential of a lower semi-continuous convex function is outer semi-continuous (Hiriart-Urruty and Lemarechal, 2002, Theorem 6.2.4), it follows that any accumulation point of $\{x^k\}$ must be a minimizer of f (its subdifferential contains the origin). □

5. Maximal monotone operators

Proximal point methods are important in the more general framework of finding zeroes of maximal monotone operators.[5] In this context, it is still possible to use the geometric ideas presented above to prove convergence of a slightly modified proximal method. Let us introduce the definition of a monotone operator.

Definition 1. An operator $T : \mathbb{R}^n \mapsto 2^{\mathbb{R}^n}$ is called *monotone* if for all $x, y \in \mathbb{R}^n$, and all $g \in T(x), \zeta \in T(y)$

$$\langle x - y, g - \zeta \rangle \geq 0.$$

Moreover, a monotone operator T will be called *maximal monotone* if it cannot be extended preserving monotonicity (*i.e.*, whenever T' is a monotone operator such that $T'(x) \supseteq T(x),\ \forall x$, then $T' \equiv T$).

The problem we want to solve is to find a zero of a maximal monotone operator, T, i.e., find $x \in \mathbb{R}^n$ such that

$$0 \in T(x).$$

The *exact* version of the proximal point method computes a sequence using the following recursion:

$$g^{k+1} \in T(x^{k+1}),$$
$$g^{k+1} + \alpha_k(x^{k+1} - x^k) = 0.$$

In Rockafellar (1976a), the author introduced error bounds to relax the equation above. In Solodov and Svaiter (1999), using an extra projection after each proximal step, less stringent error bounds were presented. We will prove the convergence of Solodov-Svaiter's method. Formally the algorithm is given below:

Algorithm 3. *Inexact Hybrid Projection-Proximal Point Algorithm*

Let $T : \mathbb{R}^n \mapsto 2^{\mathbb{R}^n}$ be a maximal monotone operator. Choose $\sigma \in [0, 1)$ and $\bar{\alpha} > 0$.

1. *Initialization*: *Let x^1 be any point in $\mathbb{R}^n$;*
2. *Iteration*: *Let $0 < \alpha_k \leq \bar{\alpha}$. Find $\tilde{x}^k$ such that, for a $\tilde{g}^k \in T(\tilde{x}^k)$ and $e^k \stackrel{\text{def}}{=} \tilde{g}^k + \alpha_k(\tilde{x}^k - x^k)$,*

$$\|e^k\| \leq \sigma \max\{\|\tilde{g}^k\|, \alpha_k\|\tilde{x}^k - x^k\|\}. \tag{AC3}$$

We call $\tilde{x}^k$ the partial iterate.

Finally, let x^{k+1} be the projection of x^k onto the hyperplane that crosses $\tilde{x}^k$ and that has $\tilde{g}^k$ as normal.

This is graphically explained by Figure 2.

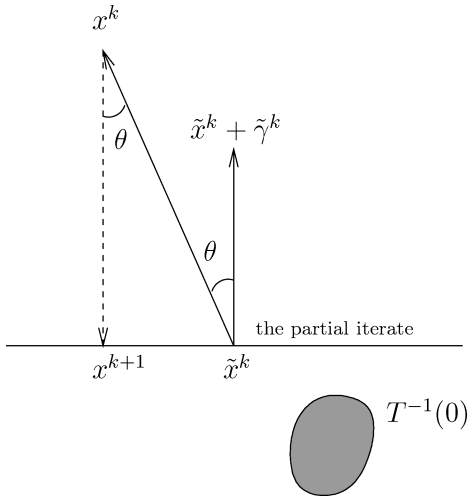

Figure 2. The projection after the inexact proximal step.

Note that the acceptance criterion (AC3) is analogous to (AC2). On the other hand, we do not have the projection in the optimization case. The need for the projection resides in the absence of an explicit merit function: there is no function being minimized. Then we try to ensure that the distance to the solution set decreases at each iteration. The projection guarantees that the new iterate, x^{k+1}, is closer to any zero of T than x^k. This is a direct consequence of the monotonicity of T, since for any $x^* \in T^{-1}(0)$,

$$\langle x^* - \tilde{x}^k, \tilde{g} - 0^k \rangle = \langle x^* - \tilde{x}^k, \tilde{g}^k \rangle \leq 0,$$

and, as we already know, the angle θ between $x^k - \tilde{x}^k$ and $\tilde{g}^k$ is acute due to (AC3). This asserts that

$$\langle x^k - \tilde{x}^k, \tilde{g}^k \rangle > 0.$$

We conclude that the projection hyperplane strictly separates x^k from the set of zeroes. Following these geometrical ideas, it should be easy to the reader to verify that

$$\begin{aligned} x^{k+1} &= x^k - \frac{\langle \tilde{g}^k, x^k - \tilde{x}^k \rangle}{\|\tilde{g}^k\|^2} \tilde{g}^k, \\ \|x^{k+1} - x^*\|^2 &\leq \|x^k - x^*\|^2 - \|x^{k+1} - x^k\|^2. \end{aligned} \tag{7}$$

We can state the main result of this section:

Theorem 3. *Let $T : \mathbb{R}^n \mapsto 2^{\mathbb{R}^n}$ be a maximal monotone operator. Assume that T has at least one zero. Let $\{x^k\}$ be any sequence generated by the hybrid projection-proximal point algorithm. Then $\{x^k\}$ converges to a zero of T.*

Proof: Let x^* be a zero of T. From (7) we may conclude that $\|x^k - x^*\|$ is a decreasing sequence bounded below by 0 and that

$$\|x^{k+1} - x^k\| \to 0.$$

Since θ is bounded away from $\frac{\pi}{2}$ (from Lemma 1), it follows that

$$\|x^k - \tilde{x}^k\| \to 0, \tag{8}$$

And, using the boundedness of $\{\alpha_k\}$ and Lemma 2, we conclude that

$$\begin{aligned} \alpha_k(x^k - \tilde{x}^k) &\to 0, \\ \tilde{g}^k &\to 0. \end{aligned} \tag{9}$$

Moreover, $\{x^k\}$ is bounded since it is contained in the ball with center x^* and radius $\|x^1 - x^*\|$. Let $\bar{x}$ be any accumulation point $\{x^k\}$. From (8) we see that the partial iterates, $\{\tilde{x}^k\}$, have the same subsequence converging to $\bar{x}$. Hence, (9) and the outer semicontinuity of T imply that $0 \in T(\bar{x})$. That is, any accumulation point of $\{x^k\}$ is a zero of T.

Finally, let us show that $\{x^k\}$ has only one accumulation point, and therefore it is convergent. Given any accumulation point $\bar{x}$, we can see from (7) that $\|x^k - \bar{x}\|$ is decreasing. Hence, it must go to zero and then $x^k \to \bar{x}$. □

6. Final comments

This article is based on Solodov and Svaiter's results for computing zeroes of maximal monotone operators (Solodov and Svaiter, 1999). It has emerged from our efforts to understand the convergence proof therein and to present simpler proofs. The descent approach is presented in a more general framework in Humes and Silva (2000). Another reference where a similar reasoning is used to analyze proximal methods is Birge et al. (1998). In this interesting paper, the authors use a slightly different acceptance criterion and do not consider the case of maximal monotone operators. The result without projection cannot be extended to general monotone operators as shown in Solodov and Svaiter (1999).

Appendix

Consider the following Nonlinear Programming problem:

$$\begin{aligned} \min \quad & f(x) \\ \text{s.t.} \quad & g(x) \leq 0 \\ & x \in I\!R^n, \end{aligned} \tag{10}$$

where $f : I\!R^n \to I\!R$ and $g : I\!R^n \to I\!R^m$ are given functions.

We define the Lagrangian function associated to (10) by

$$L(x,\lambda) \stackrel{\text{def}}{=} \begin{cases} f(x) + \sum_{i=1}^{m} \lambda_i g_i(x), & \text{if } \lambda \geq 0 \\ -\infty, & \text{otherwise.} \end{cases}$$

Moreover, the respective Lagrangian dual problem consists on the maximization of

$$F(\lambda) \stackrel{\text{def}}{=} \inf_{x \in I\!R^n} \{L(x,\lambda)\}. \tag{11}$$

This optimization problem, re-written as a minimization problem, is known to be convex and carries an intimate relation to (10). In particular, under reasonable convexity assumptions, its solutions are the Lagrange multipliers for (10). Many algorithms for constrained optimization work with the dual problem as a way to solve the original, primal, problem (Bertsekas, 1995, Chapter 4–6).

One of such algorithms, is the method of multipliers, which starts with an initial multiplier guess, $\lambda^0 \geq 0$, and tries to to improve it solving the following a sequence of unconstrained problems:

$$x^{k+1} \in \operatorname*{argmin}_{x \in I\!R^n} \left\{ f(x) + \frac{1}{2} \sum_{i=1}^{m} \max\left\{0, \lambda_i^k + g_i(x)\right\}^2 - \left(\lambda_i^k\right)^2 \right\};$$
$$\lambda_i^{k+1} \stackrel{\text{def}}{=} \max\left\{0, \lambda_i^k + g_i(x^{k+1})\right\} \; i = 1, \ldots, m.$$

Theorem 4. *Assume that f and each component of g are differentiable convex functions. Let $\lambda \in I\!R^m$, and define*

$$\bar{x} \in \operatorname*{argmin}_{x \in I\!R^n} \left\{ f(x) + \frac{1}{2} \sum_{i=1}^{m} \max\left\{0, \lambda_i + g_i(x)\right\}^2 - \lambda_i^2 \right\};$$
$$\bar{\lambda}_i \stackrel{\text{def}}{=} \max\left\{0, \lambda_i + g_i(\bar{x})\right\}, \; i = 1, \ldots, m.$$

Then,

$$\bar{\lambda} = \operatorname*{argmin}_{\gamma \in I\!R^m} \left\{ -F(\gamma) + \frac{1}{2}\|\gamma - \lambda\|^2 \right\}.$$

Therefore, the multiplier sequence computed by the method of multipliers is actually the same sequence that would be computed by the proximal point algorithm used to solve the dual problem and starting from λ^0.

Proof: We use the proof from Iusem (1995).

Let $\tilde{\lambda}$ be defined by

$$\tilde{\lambda} \stackrel{\text{def}}{=} \operatorname*{argmin}_{\gamma \in \mathbb{R}^m} \left\{ -F(\gamma) + \frac{1}{2}\|\gamma - \lambda\|^2 \right\}.$$

Note that $\tilde{\lambda}$ is uniquely determined, since the function being minimized is strictly convex. Using subgradient notation, this condition can be rewritten as $\lambda - \tilde{\lambda} \in \partial(-F)(\tilde{\lambda})$, or equivalently,

$$-F(\gamma) \geq -F(\tilde{\lambda}) + \langle \lambda - \tilde{\lambda}, \gamma - \tilde{\lambda} \rangle, \quad \forall \gamma \geq 0. \tag{12}$$

Since this last inequality fully characterizes $\tilde{\lambda}$, all we need to show is that it holds with $\bar{\lambda}$ replacing $\tilde{\lambda}$.

First, the definitions of $\bar{x}$, $\bar{\lambda}$, and simple calculus rules imply that

$$0 = \nabla f(\bar{x}) + \sum_{i=1}^{m} \bar{\lambda}_i \nabla g_i(\bar{x}).$$

Therefore, the convexity assumptions ensure that $\bar{x}$ minimizes $L(\cdot, \bar{\lambda})$, and hence

$$F(\bar{\lambda}) = f(\bar{x}) + \sum_{i=1}^{m} \bar{\lambda}_i g_i(\bar{x}). \tag{13}$$

This is the first term in the right hand side of (12) with $\bar{\lambda}$ replacing $\tilde{\lambda}$.

Let us turn on attention to $\langle \lambda - \bar{\lambda}, \gamma - \bar{\lambda} \rangle$. For each $i = 1, \ldots, m$,

$$\begin{aligned} \bar{\lambda}_i &= \max\{0, \lambda_i + g_i(\bar{x})\} \Rightarrow \\ \bar{\lambda}_i - \lambda_i &= \max\{-\lambda_i, g_i(\bar{x})\} \geq g_i(\bar{x}). \end{aligned} \tag{14}$$

It follows from the last equation that if $\bar{\lambda}_i > 0$,

$$\max\{-\lambda_i, g_i(\bar{x})\} = \bar{\lambda}_i - \lambda_i > -\lambda_i.$$

And then, $\max\{-\lambda_i, g_i(\bar{x})\} = g_i(\bar{x})$. Hence, as $\bar{\lambda}_i \geq 0$

$$(\lambda_i - \bar{\lambda}_i)(-\bar{\lambda}_i) = (\bar{\lambda}_i - \lambda_i)\bar{\lambda}_i = \max\{-\lambda_i, g_i(\bar{x})\}\bar{\lambda}_i = g_i(\bar{x})\bar{\lambda}_i.$$

Moreover, using the inequality in (14), it follows that for any $\gamma_i \geq 0$,

$$(\lambda_i - \bar{\lambda}_i)(\gamma_i - \bar{\lambda}_i) \leq g_i(\bar{x})\bar{\lambda}_i - g_i(\bar{x})\gamma_i.$$

Finally, if follows from the definition of F and (13) that, for $\gamma \geq 0$:

$$\begin{aligned} -F(\bar{\lambda}) + \langle \lambda - \bar{\lambda}, \gamma - \bar{\lambda} \rangle &= -f(\bar{x}) - \langle \bar{\lambda}, g_i(\bar{x}) \rangle + \langle \lambda - \bar{\lambda}, \gamma - \bar{\lambda} \rangle \\ &\leq -f(\bar{x}) - \langle \bar{\lambda} - \bar{\lambda} + \gamma, g_i(\bar{x}) \rangle \\ &= -f(\bar{x}) - \langle \gamma, g_i(\bar{x}) \rangle \\ &\leq -F(\gamma), \end{aligned}$$

which is exactly (12) for $\bar{\lambda} = \tilde{\lambda}$. □

Acknowledgments

We would like to thank the anonymous referees for suggestions that greatly improved the text. We would also like to thank the student Thiago A. André for carefully reading the manuscript.

Notes

1. Actually it can be used to ensure that the computed direction is a descent direction even in the nonconvex case.
2. In this paper we deal only with the classical version of the proximal algorithm, where the regularization term is quadratic. Algorithms based on different regularizations, like Bregman distances or φ-divergences, are the subject of more advanced texts.
3. This work is based on the hybrid-projection proximal algorithm that is presented in Section 5.
4. Here, *fast enough* is not clearly defined on purpose. There are several approaches that try to give a precise meaning to this property. For example, see the basic algorithm models in Polak (1997).
5. Note that the subdifferential map of a closed convex function is maximal monotone (Rockafellar, 1970). However, there are maximal monotone operators that are not a subdifferential map of a function (Rockafellar and Wets, 1998). Actually, the problem of finding a zero of a maximal monotone operator generalizes different problems like convex optimization, saddle point problems and variational inequalities. A good introduction on maximal monotone operators is Rockafellar and Wets (1998) [Chapter 12].

References

A. Auslender, "Numerical methods for nondifferentiable convex optimization," *Mathematical Programming Studies* vol. 30, pp. 102–126, 1987.

D. Bertsekas, *Nonlinear Programming*, Athena Scientific, 1995.

D. Bertsekas, *Constrained Optimization and Lagrange Multipliers*, Athena Scientific, 1996.

D. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*, Prentice Hall, 1989.

J. R. Birge, L. Qi, and Z. Wei, "A general approach to convergence properties of some methods for nonsmooth convex optimization," *Applied Mathematics and Optimization* vol. 38, pp. 141–158, 1998.

J. E. Dennis Jr. and R. B. Schnabel, "Numerical methods for unconstrained optimization and nonlinear equations," SIAM, 1996.

J. Eckstein and D. Bertsekas, "On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators," *Mathematical Programming* vol. 55, pp. 293–318, 1992.

C. Humes and P. J. S. Silva, "Strict convex regularizations, proximal point and augmented Lagrangians," *RAIRO Operations Research* vol. 34, no. 3, pp. 283–303, 2000.

C. Humes, P. J. S. Silva, and B. F. Svaiter, "An inexact projection-proximal augmented Lagrangian algorithm,"

in *Proceedings of the 36th Annual Allerton Conference on Communication, Control, and Computing*, 1998, pp. 450–459.

J.-B. Hiriart-Urruty and C. Lemarechal, *Fundamental of Convex Analysis*, Springer-Verlag, 2002.

A. N. Iusem, *Proximal Point Methods in Optimization*. Instituto de Matemática Pura e Aplicada - CNPq, 1995.

A. N. Iusem, B, ''Augmented Lagrangian methods and proximal point methods for convex optimization," *Investigación Operativa* vol. 8, pp. 11–49, 1999.

B. Martinet, "Regularisation d'inéquations variationelles par approximations successives," *Revue Française Informatique Recherche Opérationnelle* vol. 4, pp. 154–158, 1970.

B. Martinet, "Détermination approchée d'un point fixe d'une application pseudo-contractante," *Comptes Rendus Hebdomadaires des Séances de l'Académie des Sciences. Séries A et B* vol. 274, pp. A163–A165, 1972.

L. Polak, *Optimization: Algorithms and Consistent Approximations*, Springer Verlag, 1997.

R. T. Rockafellar, *Convex Analysis*, Princeton University Press, 1970.

R. T. Rockafellar, "'Monotone operators and the proximal point algorithm," *SIAM Journal on Control and Optimization* vol. 14, pp. 887–898, 1976a.

R.T. Rockafellar, "Augmented Lagrangians and applications of the proximal point algorithm in convex programming," *Mathematics of Operations Research* vol. 1, pp. 97–116, 1976b.

R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Berlin: Springer-Verlag, 1998.

M. Solodov and B. F. Svaiter, "A Hybrid projection-proximal point algorithm," *Journal of Convex Analysis* vol. 6, pp. 59–70, 1999.

M. Solodov and B. F. Svaiter, "A globally convergent inexact Newton's method for systems of monotone equations," in M. Fukushima and L. Qi, eds., *Reformulation - Nonsmooth, Piecewise Smooth, Semismooth and Smoothing Methods*, *Applied Optimization*, Kluwer Academic Publishers, vol. 22, pp. 355–369, 1999.