REVUE FRANÇAISE D'AUTOMATIQUE, D'INFORMATIQUE ET DE RECHERCHE OPÉRATIONNELLE. RECHERCHE OPÉRATIONNELLE

A. N. IUSEM
B. F. SVAITER

A proximal regularization of the steepest descent method

Revue française d'automatique, d'informatique et de recherche opérationnelle. Recherche opérationnelle, tome 29, n° 2 (1995), p. 123-130.

<http://www.numdam.org/item?id=RO_1995__29_2_123_0>

Recherche opérationnelle/Operations Research
(vol. 29, n° 2, 1995, p. 123 à 130)

A PROXIMAL REGULARIZATION OF THE STEEPEST DESCENT METHOD (*)

by A. N. IUSEM ([1]) (**) and B. F. SVAITER ([1])

Communicated by Jean-Pierre CROUZEIX

Abstract. – *We introduce a quadratic regularization term (in the spirit of the proximal point method) in the line searches of the steepest descent method, obtaining thus better convergence results. While the convergence analysis of the steepest descent method requires bounded level sets of the minimand to get a bounded sequence, and establishes, even for convex objectives, only optimality of the cluster points, our approach guarantees convergence of the whole sequence to a minimizer when the objective function is pseudo-convex, whether its level sets are bounded or not.*

Keywords: Convex programming, Steepest descent method, Proximal point method.

Résumé. – *On introduit un terme de régularisation quadratique (dans l'esprit de la méthode du point proximal) dans les minimisations unidimensionnelles de la méthode du gradient, et on obtient ainsi des résultats de convergence plus forts. Tandis que l'analyse de la convergence de la méthode du gradient demande des ensembles de niveau bornés, et démontre, même pour des fonctions convexes, tout seulement l'optimalité des points d'accumulation, notre régularisation permet de démontrer la convergence de la suite toute entière à un minimisateur quand la fonction objectif est pseudo-convexe, même dans le cas où les ensembles de niveau ne sont pas bornés.*

1. INTRODUCTION

The steepest descent method (also called Cauchy's method, or gradient method) is one of the oldest and simplest algorithms for minimizing a real function defined on $\mathbf{R}^n$. It is also the departure point for many other more sophisticated optimization procedures. Despite its simplicity and notoriety (practically no optimization book fails to discuss it), its convergence theory is not fully satisfactory from a theoretical point of view (from a practical point of view the situation is even worse, but here we are not concerned with this issue). More precisely, standard convergence results (e.g. [1]) demand

(*) Received September 1993.
(**) Research of this author was partially supported by CNP_q grant n° 301280/86.
([1]) Instituto de Matemática Pure e Aplicada, Estrada Dona Castorina, 110, 22.460, Rio de Janeiro, RJ, Brazil.

Recherche opérationnelle/Operations Research, 0399-0559/95/02/$ 4.00

that the initial point belong to a bounded level set of the objective function f (and henceforth that f have at least one bounded level set) and fail to prove convergence of the sequence generated by the method to a stationary point of f, establishing only that all its cluster points are stationary. Even when f is convex (in which case the stationary points are the minimizers of f) the level set boundedness assumption is required, and the result is just what has been called *weak convergence* to the set of minimizers of f (a sequence $\{x^k\}$ is said to be weakly convergent to a set S if $\{x^k\}$ is bounded, $x^{k+1} - x^k$ converges to zero and every cluster point of $\{x^k\}$ belongs to S).

It is true that from a computational point of view weak convergence is almost undistinguishable from full convergence, but failure to prove full convergence is theoretically insatisfactory. On the other hand, the condition of bounded level sets is quite restrictive both theoretically and practically.

In fact, the hypothesis of bounded level sets is indeed essential not only for convergence but also for well-definedness of the steepest descent method: we give next an example of a convex and differentiable function with nonempty set of minimizers and unbounded level sets for which the steepest descent method fails in the first iteration.

It is well known that if C is a closed convex subset of $\mathbf{R}^n$, $P : \mathbf{R}^n \to C$ is the orthogonal projection onto C and $g(x) = \| P(x) - x \|^2$, then g is convex and differentiable, C is the set of minimizers of g and $\nabla g(x) = 2(x - P(x))$. Consider C_1, $C_2 \subset \mathbf{R}^2$ defined as $C_1 = \{(x_1, x_2) : x_1 \geq 0, x_2 \geq 1/x_1\}$, $C_2 = \{(x_1, x_2) : x_2 \leq x_1 - 2\}$, let P_i be the orthogonal projection onto C_i $(i = 1, 2)$ and take $f(x) = \| P_1(x) - x \|^2 + \| P_2(x) - x \|^2$. It follows that f is convex and differentiable, the set of minimizers of f is $C_1 \cap C_2 = \{(x_1, x_2) : x_1 \geq 1 + \sqrt{2}, 1/x_1 \leq x_2 \leq x_1 - 2\}$ where f vanishes and $\nabla f(x) = 4x - 2P_1(x) - 2P_2(x)$. It is easy to check that $P_1(0, 0) = (1, 1)$, $P_2(0, 0) = (1, -1)$ so that $\nabla f(0, 0) = (-4, 0)$. If we start the steepest descent method at $x^0 = (0, 0)$, we are required to minimize f on the halfline $L = \{(t, 0) : t \geq 0\}$, but the minimum is not achieved, because $L \cap C_1 \cap C_2 = \varnothing$ and the distance from L to $C_1 \cap C_2$ is 0 so that $f(x) > 0$ for all $x \in L$ and $\inf_{x \in L} f(x) = 0$.

In this paper we show that addition of a quadratic regularization term (in the spirit of the proximal point method) to the linear searches of the steepest descent method removes such obstacles and, for the case of a pseudo-convex objective f, it is possible both to prove full convergence of the sequence to a minimizer of f and to eliminate the assumption of boundedness of the level sets.

We start by recalling the proximal point method for minimizing $f : \mathbf{R}^n \to \mathbf{R}$. Starting from an arbitrary $x^0 \in \mathbf{R}^n$, a sequence $\{x^k\} \subset \mathbf{R}^n$ is defined by

$$x^{k+1} = \arg\min_{x \in \mathbf{R}^n} \{ f(x) + \lambda_k \| x^k - x \|^2 \} \tag{1}$$

where $\{\lambda_k\}$ is a bounded sequence of positive real numbers called regularization parameters. The origin of this algorithm can be traced back to [5], and its applications in the area of convex optimization were fully developed in [7], [8], where it is proved that under reasonable assumptions of f the sequence $\{x^k\}$ defined by (1) converges to a minimizer of f. See [4] for a recent survey on the proximal point method and its extensions. We remak that this method is in general a theoretical device rather than an implementable procedure, because the minimization subproblems in (1) are in principle as hard as the basic problem $\min f(x)$, and some numerical algorithm is required to solve them in order to get each iterate x^k.

In this paper we do not analyze the proximal point method as given by (1). Rather we limit ourselves to add a regularization term of the form $\lambda_k \| x - x^k \|^2$ to the line searches of the steepest descent method.

Our proof is based upon the notion of quasi-Frejér convergence, introduced for the first time [2], which will be discussed in the next section.

2. THE PROXIMAL REGULARIZATION OF THE STEEPEST DESCENT METHOD

Given a continuously differentiable function $f : \mathbf{R}^n \to \mathbf{R}$, the steepest descent method for the problem

$$\min_{x \in \mathbf{R}^n} f(x) \tag{2}$$

is given by

Initialization:

$$x^0 \in \mathbf{R}^n$$

Iterative step:

$$\alpha_k = \arg\min_{\alpha \geq 0} f(x^k - \alpha \nabla f(x^k)) \tag{3}$$

$$x^{k+1} = x^k - \alpha_k \nabla f(x^k) \tag{4}$$

i.e. α_k minimizes the restriction of $f(x)$ to the halfline starting at x^k and passing through $x^k - \nabla f(x^k)$.

Our method consist of using instead the restriction of $f(x)+\lambda_k \| x-x^k \|^2$ to this halfline. If $x = x^k - \alpha \nabla f(x^k)$ then $f(x)+\lambda_k \| x-x^k \|^2 = f(x^k - \alpha \nabla f(x^k)) + \alpha^2 \lambda_k \| \nabla f(x^k) \|^2$. Therefore the proximal regularization of the steepest descent method replaces (3), (4) by

$$\alpha_k = \arg\min_{\alpha \geq 0} \{ f(x^k - \alpha \nabla f(x^k)) + \alpha^2 \lambda_k \| \nabla f(x^k) \|^2 \} \tag{5}$$

$$x^{k+1} = x^k - \alpha_k \nabla f(x^k) \tag{6}$$

where $\| \cdot \|$ is the Euclidean norm and $\{\lambda_k\} \subset \mathbf{R}$ satisfy

$$\hat{\lambda} \leq \lambda_k \leq \tilde{\lambda} \tag{7}$$

for some $\hat{\lambda}$, $\tilde{\lambda}$ such that $0 < \hat{\lambda} \leq \tilde{\lambda}$.

We proceed to the convergence analysis. From now on $\{x^k\}$ refers to the sequence generated by (5)-(6). We assume that problem (2) has solutions. Let $f^* = \min_{x \in \mathbf{R}^n} f(x)$.

PROPOSITION 1: *The sequence $\{x^k\}$ is well defined and, for all k,*

$$\nabla f(x^{k+1})^t \nabla f(x^k) = 2 \alpha_k \lambda_k \| \nabla f(x^k) \|^2 \tag{8}$$

Proof: Let $\varphi_k(\alpha)$ be the minimand of (5), *i.e.* $\varphi_k(\alpha) = f(x^k - \alpha \nabla f(x^k)) + \alpha^2 \lambda_k \| \nabla f(x^k) \|^2$. We claim that problem (5) always has solutions. If $\nabla f(x^k) = 0$ then φ_k is constant and from (6) $x^{k+1} = x^k$ for any choice of α_k. Otherwise $\varphi_k(\alpha) \geq f^* + \alpha^2 \hat{\lambda} \| \nabla f(x^k) \|^2$, implying $\lim_{\alpha \to \infty} \varphi_k(\alpha) = \infty$, so that the minimization in (5) reduces to a bounded interval and existence of α_k follows from continuity of φ_k.

Observe that $\varphi_k'(\alpha) = -\nabla f(x^k - \alpha \nabla f(x^k))^t \nabla f(x^k) + 2 \alpha \lambda_k \| \nabla f(x^k) \|^2$, so that $\varphi_k'(0) = -\| \nabla f(x^k) \|^2 \leq 0$. It follows that either $\nabla f(x^k) = 0$, in which case (8) holds trivially, or $\alpha_k > 0$, in which case, using (6),

$$0 = \varphi_k'(\alpha_k) = -\nabla f(x^{k+1})^t \nabla f(x^k) + 2 \alpha_k \lambda_k \| \nabla f(x^k) \|^2 \tag{9}$$

and (8) follows from (9).

PROPOSITION 2:

(i) $f^* \leq f(x^{k+1}) + \lambda_k \| x^{k+1} - x^k \|^2 \leq f(x^k)$,

(ii) $\{f(x^k)\}$ *is decreasing and convergent,*

(iii) $\sum_{k=0}^{\infty} \| x^{k+1} - x^k \|^2 < \infty$,

(iv) $\lim_{k\to\infty} (x^{k+1} - x^k) = 0.$

Proof:

(i) With φ_k as defined in the proof of Proposition 1, use (5) and (6) to get

$$f(x^k) = \varphi_k(0) \geq \varphi_k(\alpha_k) = f(x^k - \alpha_k \nabla f(x^k)) + \alpha_k^2 \lambda_k \| \nabla f(x^k) \|^2$$
$$= f(x^{k+1}) + \lambda_k \| x^{k+1} - x^k \|^2$$

which proves the rightmost inequality. The leftmost one is trivial.

(ii) Follows from (i) using $\lambda_k > 0$.

(iii) From (i) and (7), $\hat{\lambda} \| x^{l+1} - x^l \|^2 \leq \lambda_l \| x^{l+1} - x^l \|^2 \leq f(x^l) - f(x^{l+1})$. Summing on l, $\hat{\lambda} \sum_{l=0}^{k} \| x^{l+1} - x^l \|^2 \leq f(x^0) - f(x^{k+1}) \leq f(x^0) - f^*$ so that $\sum_{k=0}^{\infty} \| x^{k+1} - x^k \|^2 \leq \frac{1}{\hat{\lambda}} (f(x^0) - f^*) < \infty$.

(iv) Follows from (iii).

PROPOSITION 3: *If $\bar{x}$ is a cluster point of $\{x^k\}$ then $\nabla f(\bar{x}) = 0$.*

Proof: Let $\{x^{l_k}\}$ be a subsequence of $\{x^k\}$ such that $\lim_{k\to\infty} x^{l_k} = \bar{x}$. By Proposition 1 and (6)

$$\nabla f(x^{l_k+1})^t \nabla f(x^{l_k}) = 2\alpha_{l_k} \lambda_{l_k} \| \nabla f(x^{l_k}) \|^2$$
$$= 2\lambda_{l_k} \| x^{l_k+1} - x^{l_k} \| \, \| \nabla f(x^{l_k}) \| \qquad (10)$$

By Proposition 2 (iv), (7) and continuous differentiability of f, the left hand side of (10) converges to $\| \nabla f(\bar{x}) \|^2$ as k goes to ∞, and the right hand side of (10) converges to 0, *i.e.* $\nabla f(\bar{x}) = 0$.

We remark that up to now we have not established existence of cluster points of $\{x^k\}$, which will be proved below with a convexity hypothesis on f, but we chose to prove stationarity of the limit points before existence in order to complete the set of statements that hold without demanding convexity. Of course, if we assume that x^0 belongs to a bounded level set f, then by Proposition 2 (ii) such set contains the whole sequence, which is therefore bounded, in which case Proposition 3 implies the standard result of weak convergence to the set of stationary points, *i.e.* even without convexity our algorithm fares as well as the usual steepest descent method. But our main goal is to get rid of the boundedness assumption and nevertheless prove full convergence, which we do next for the case of pseudo-convex f.

We present now the notion of quasi-Fejér convergence, introduced in [2].

DEFINITION 1: A sequence $\{y^k\} \subset \mathbf{R}^n$ is quasi-Fejér convergent to a set $U \subset \mathbf{R}^n$ if for every $u \in U$ there exists a sequence $\{\varepsilon_k\} \subset \mathbf{R}$ such that $\varepsilon_k \geq 0$, $\sum_{k=0}^{\infty} \varepsilon_k < \infty$ and $\| y^{k+1} - u \|^2 \leq \| y^k - u \|^2 + \varepsilon_k$.

The following result was proved in [3, Theorem 4.1] for rather general distances. We give here a proof for the Euclidean distance in order to make this paper self-contained.

PROPOSITION 4: *If $\{y^k\}$ is quasi-Fejér convergent to a nonempty set U then $\{y^k\}$ is bounded. If furthermore a cluster point y of $\{y^k\}$ belongs to U then $y = \lim_{k \to \infty} y^k$.*

Proof: Take $u \in U$. Apply iteratively Definition 1 and get $\| y^k - u \|^2 \leq \| y^0 - u \|^2 + \sum_{l=0}^{k-1} \varepsilon_l \leq \| y^0 - u \|^2 + \beta$, where $\beta = \sum_{k=0}^{\infty} \varepsilon_k < \infty$. It follows that $\{y^k\}$ is bounded. Let now $y \in U$ be a cluster point of $\{y^k\}$ and take any $\delta > 0$. Let $\{y^{l_k}\}$ be a subsequence of $\{y^k\}$ such that $\lim_{k \to \infty} y^{l_k} = y$. Since $y \in U$ there exists $\{\varepsilon_k\}$ satisfying the properties of Definition 1. Take k_0 such that $\sum_{k=k_0}^{\infty} \varepsilon_k \leq \frac{\delta}{2}$ and $\bar{k}$ such that $l_{\bar{k}} \geq k_0$ and $\| y^{l_{\bar{k}}} - y \|^2 \leq \frac{\delta}{2}$. Then, for any $k \geq l_{\bar{k}}$:

$$\| y^k - y \|^2 \leq \| y^{l_{\bar{k}}} - y \|^2 + \sum_{i=l_{\bar{k}}}^{k-1} \varepsilon_i \leq \frac{\delta}{2} + \sum_{i=k_0}^{\infty} \varepsilon_i \leq \frac{\delta}{2} + \frac{\delta}{2} = \delta$$

Since δ is arbitrary, it follows that $y = \lim_{k \to \infty} y^k$. □

We present next our quasi-Fejér convergence result. We remind that f is pseudo-convex if and only if it is differentiable and $(y - x)^t \nabla f(x) \geq 0$ implies $f(y) \geq f(x)$. It is immediate that differentiable convex functions are pseudo-convex.

PROPOSITION 5: *If f is pseudo-convex then the sequence $\{x^k\}$ is quasi-Fejér convergent to the set of minimizers of f.*

Proof: Let z be a minimizer of f. Then

$$\| z - x^{k+1} \|^2 - \| z - x^k \|^2 - \| x^k - x^{k+1} \|^2$$
$$= -2 (z - x^k)^t (x^{k+1} - x^k)$$
$$= 2 \alpha_k (z - x^k)^t \nabla f (x^k) \qquad (11)$$

using (6) in the second equality. We claim that $(z - x^k)^t \nabla f (x^k) \leq 0$. Otherwise, $(z - x^k)^t \nabla f (x^k) > 0$ and we get $f(z) \geq f(x^k)$ by pseudo-convexity of f. Then x^k is also a minimizer of f, and therefore $\nabla f(x^k) = 0$, in contradiction with $(z - x^k)^t \nabla f (x^k) > 0$. The claim is established. Then, we get from (11) $\| z - x^{k+1} \|^2 \leq \| z - x^k \|^2 + \| x^k - x^{k+1} \|^2$ and the result follows from Proposition 2 (iii) and Definition 1, with $\varepsilon_k = \| x^{k+1} - x^k \|^2$. □

Finally we give our main result.

THEOREM 1: *If f is pseudo-convex and attains its minimum on $\mathbf{R}^n$ then the sequence $\{x^k\}$ generated by (5), (6) converges to a minimizer of f.*

Proof: By Propositions 4 and 5 the sequence is bounded, so it has cluster points. By Proposition 3 any such cluster point is a stationary point of f, and therefore, by pseudo-convexity of f a minimizer of f. The result follows from the second statement of Proposition 4. □

3. FINAL REMARKS

The computational performance of the steepest descent method has been up to now quite disappointing. Not only it has a best a linear convergence rate (e.g. [6], though this happens with all first order methods) but the fact the gradients at consecutive iterates are mutually orthogonal leads to poor performance ("hemstitching" phenomena, *see* [1]), unless f has almost spherical level sets.

We do not make any claim in the sense that our regularization is not subject to the same limitations (though in our scheme consecutive gradients form an acute angle, which in principle looks better, and the freedom in the choice of the regularization parameters λ_k, which control the size of such angle, could have interesting computational effects); we just present a very simple variant of the method for which much neater convergence results can be obtained, eliminating at same time the rather annoying hypothesis of bounded level sets.

Nevertheless, we remark that the addition of the quadratic term to the line searches represents a negligeable additional computational burden, and in compensation, besides the better theoretical results presented above, the minimand φ_k of the line search will be in general better behaved than the restriction of f to the halfline (e.g. if f is convex then φ_k is strictly convex) so that numerical procedures used to perform the line searches can be expected to be more efficient for our proposal, whether they are nonderivative ones (e.g. Fibonacci search) or derivative ones (e.g. Newton's method).

REFERENCES

1. M. AVRIEL, *Nonlinear Programming, Analysis and Methods,* Prentice Hall, New Jersey, 1976.
2. Yu. M ERMOL'EV, On the method of generalized stochastic gradients and quasi-Fejér sequences, *Cybernetics,* 1969, *5*, p. 208-220.
3. A. N. IUSEM, B. F. SVAITER and M. TEBOULLE, Entropy-like proximal methods in convex programming (to be published in Mathematics of Operations Research).
4. B. LEMAIRE, The proximal algorithm, in *International Series of Numerical Mathematics,* 1989, *87*, (J. P. Penot, ed), *Birkhauser,* Basel, p. 73-87.
5. J.-J. MOREAU, Proximité et dualité dans un espace Hibertien, *Bull. Soc. Math. France,* 1965, *93*, p. 273-299.
6. B. POLYAK, *Introduction to Optimization,* Optimization Software, New York, 1987.
7. R. T. ROCKAFELLAR, Augmented Lagragians and applications of the proximal point algorithm in convex programming, *Mathematics of Operations Research,* 1976, *1*, p. 97-116.
8. R. T. ROCKAFELLAR, Monotone operators and the proximal point algorithm, *SIAM Journal on Control and Optimization,* 1976, *14*, p. 877-898.