GENERALIZED PROXIMAL POINT ALGORITHMS AND BUNDLE IMPLEMENTATIONS

STÉPHANE CHRÉTIEN* AND ALFRED O. HERO*

Abstract. In this paper, we present a study of the *proximal point* algorithm using very general regularizations for minimizing possibly nondifferentiable and nonconvex locally Lipschitz functions. We deduce from the *proximal point* scheme a simple and implementable bundle method. The originality of our bundle method is that the bundle information incorporates the subgradients of both the objective and the regularization function. The resulting method opens up a broad class of regularizations which are not restricted to quadratic or even differentiable functions.

Key words. mathematical programming, proximal point, bundle methods, nonsmooth regularization

AMS subject classifications. 26B05, 90C26, 90C30, 90C90, 65U05

1. Introduction.

In this paper, we address the problem of minimizing a locally Lipschitz possibly nondifferentiable and nonconvex function $f(x)$ on $\mathbb{R}^n$, i.e.

$$\min_{x\in\mathbb{R}^n} f(x). \tag{1.1}$$

One of the most widely studied methods for solving nondifferentiable problems is the bundle method first proposed by Lemarechal [13] and Wolfe [29] for convex minimization and further developed by Mifflin [17, 18] and Kiwiel [9, 10, 12] for the nonconvex case; see also [2], [15], [24] and the references therein. The bundle method can be interpreted as a *cutting plane* algorithm stabilized by a quadratic penalty or regularization. In its simplest form, for f convex, the bundle method generates a sequence of iterates, starting from x^1 and defined by

$$x^{k+1} = \operatorname{argmin}_{x\in\mathbb{R}^n}\{\hat{f}(x) + \frac{\lambda}{2}|x^k - x|^2\}. \tag{1.2}$$

where

$$\hat{f}(x) = \max_{j\in J^k}\{f(y_j) + \langle g(y_j), x - y_j\rangle\} \tag{1.3}$$

is the *cutting plane* model, y_j, $j \in J^k$ are some points in a neighborhood of the current iterate x^k and $g(y)$ is a subgradient of f at the point y. In the case where f is nonconvex, the following polyhedral approximation is usually chosen, as in [9, 10, 12],

$$\hat{f}(x) = f(x^k) + \max_{j\in J^k}\{-\alpha_j^k + \langle g^j, x - x^k\rangle, \quad j \in J^k\}, \tag{1.4}$$

where

$$\alpha_j^k = \alpha(x^k, y^j)$$
$$\alpha(x, y) = |f(x) - f(y) - \langle g(y), x - y\rangle|.$$

*Dpt. of Electrical Engineering and Computer Science, The University of Michigan, 1301 Beal Avenue, Ann Arbor, Michigan 48109-2122, USA. (chretien@eecs.umich.edu and hero@eecs.umich.edu). This work has been partially supported by AFOSR F49620-96-0028.

One fruitful interpretation of the bundle method is to consider iteration (1.2) as an implementable approximation of the well known *proximal point* algorithm using a *cutting plane* model $\hat{f}$ of the objective function. In the original form [16, 23], the *proximal point* algorithm is defined by the reccurence

$$x^{k+1} = \operatorname{argmin}_{y \in \mathbb{R}^n} \{f(y) + \frac{\lambda_k}{2}|x-y|^2\}. \quad (1.5)$$

where $(\lambda_k)_{k \in \mathbb{N}}$ is a sequence of positive relaxation parameters. The proximal point algorithm and the bundle method share the same property of solving a sequence of minimization subproblems incorporating a quadratic penalty, also denoted Moreau-Yosida regularization [14].

In this paper, we address the study of the proximal point algorithm and its bundle implementations using very general nonquadratic penalty functions. In particular, we establish convergence for a class of locally Lipschitz regularizations without any convexity nor differentiability assumptions. The utility of such nonquadratic regularization is motivated by the following examples.

EXAMPLE 1 (EM-type algorithms for maximum likelihood estimation). We show here that the case of Kullback regularization results in a proximal point method which is equivalent to the well known Expectation Maximization (EM) algorithm for maximum likelihood estimation [5]. Consider as in [5] the sample spaces Ω_1 and Ω_2 and the random variables V_1 and V_2 with respective probability densities $p_1(v_1; x)$ and $p_2(v_2; x)$, both indexed by an unknown parameter $x \in \mathbb{R}^n$ to be estimated. Assume that V_2 is obtained from V_1 through a many-to-one mapping $V_1 \to V_2 = h(V_1)$ and define $p(v_1 \mid v_2; x) = \frac{p_1(v_1;x)}{p_2(v_2;x)}$ the density of V_1 conditioned on $V_2 = v_2$. Then, the Kullback information measure between $p(v_1 \mid v_2; x)$ and $p(v_1 \mid v_2; y)$ for two parameter values x and y takes the following form

$$I(x, y; v_2) = \int_{\{h^{-1}(v_2)\}} \log\left(\frac{p(v_1 \mid v_2; y)}{p(v_1 \mid v_2; x)}\right) \; p(v_1 \mid v_2; y) dv_1 \quad (1.6)$$

Now, consider the following proximal point algorithm for maximizing the function $f(x) = \log p_2(v_2; x)$.

$$x^{k+1} = \operatorname{argmin}_{y \in \mathbb{R}^n} \{-\log p_2(v_2; x) + \lambda_k I(x^k, y; v_2)\}. \quad (1.7)$$

When x is an unknown parameter governing a random observation V_2, the maximum of $f(x)$ is the well known maximum likelihood estimate of x. Now, taking $\lambda_k = 1$ in (1.7), after some simplifications we obtain

$$x^{k+1} = \operatorname{argmin}_{y \in \mathbb{R}^n} \{-\int_{\{h^{-1}(v_2)\}} \log(p_1(v_1 \mid v_2; y)) \; p_1(v_1 \mid v_2; x^k) dv_1\}, \quad (1.8)$$

which is exactly the EM algorithm introduced in [5], where V_2 is the incomplete data and V_1 is the complete data. A special case of (1.7) is the case of Laplacian data,

$$p_2(v_2; x) = \frac{x}{2} \exp(-x|v_2|), \;\; x \geq 0. \quad (1.9)$$

When the complete data V_1 is chosen as Laplacian, it is easy to show that the Kullback regularization given by (1.6) is nonsmooth and nonconvex.

EXAMPLE 2 (Methods of multipliers). In [22] Rockafellar shows that the proximal point approach can be applied to the dual of a constrained optimization problem to yield interesting classes of multiplier methods. Subsequent studies [26, 27, 7] have demonstrated the benefit of using nonquadratic regularization functions. Among the possible choices for regularization functions is the $\phi-$divergence

$$d_\phi(x, x') = \sum_{j=1}^{n} x_j \phi(\frac{x'_j}{x_j}). \tag{1.10}$$

In particular, consider the convex program

$$\min_{x \in \mathbb{R}^n} f(x) \qquad \text{subject to} \qquad g_i(x) \leq 0, \quad i \in \{1, \ldots, m\} \tag{1.11}$$

where f and $g_1, \ldots, g_m$ are convex functions. The proximal point algorithm applied to the dual takes the form (see [7, section 6])

$$\begin{aligned} &p^0 \in \mathbb{R}^m_+ \\ &p^{k+1} = \operatorname{argmax}_{p \geq 0} \{c(p) - \lambda_k d_\phi(p^k, p)\} \end{aligned}$$

with $c(p)$ being the dual functional defined by $\inf_{x \in \mathbb{R}^n} L(x, p)$ where $L(x, p)$ is the Lagrangian

$$L(x, p) = \begin{cases} f(x) + \sum_{i=1}^m p_i g_i(x) & \text{if } p_i \geq 0, \ \forall i \in \{1, \ldots, m\} \\ -\infty & \text{otherwise} \end{cases}$$

Thus, one obtains convergence of $(x^k)_{k \in \mathbb{N}}$ and $(p^k)_{k \in \mathbb{N}}$ to the solution of problem (1.11) (for instance, see [7]). Using different choices for the function ϕ, some well known multiplier methods can be recovered and many original ones can be generated.

Further examples of nonsmooth and nonconvex regularizations were also recently studied in the context of inverse problems in [19] and [20].

The outline of the paper is the following. In section 2 the generalized proximal point algorithm is introduced for a wide class of possible regularizations. Global convergence of the method is established in section 3. In particular an analysis of nondifferentiable nonconvex regularization is provided which seems to have no precedent in the literature. We then demonstrate local convergence when f is strictly convex in an open neighborhood of an accumulation point of the minimizing sequence. In section 4 we turn to implementable bundle methods for approximation of the generalized proximal point iterations by incorporating the successive subgradients of the regularization function into the bundle information.

We recall the following notations and definitions [4]. The inner product on $\mathbb{R}^n$ is denoted by $\langle ., . \rangle$ and the associated norm is $|.|$. The convex hull of a set S is denoted $\text{conv}(S)$. The function f: $\mathbb{R}^n \rightarrow R$ is locally lipschitzian if for any bounded set $B \subset \mathbb{R}^n$ there exists a constant $L < \infty$ such that $|f(x) - f(y)| \leq L|x - y| \ \forall x, y \in B$. The effective domain of f is $C_f = \{x \ \mid \ f(x) < \infty\}$. The generalized derivative of f at x in the direction of d is

$$f'(x, d) = \limsup_{\substack{h \to 0 \\ t \downarrow 0}} \frac{f(x + h + td) - f(x)}{t} \tag{1.12}$$

The subdifferential of f at x is the set

$$\delta f(x) = \text{conv}\{\lim_{x^i \to x} \nabla f(x^i) \ \mid \ \nabla f(x^i) \text{ exists}\}, \tag{1.13}$$

Where ∇f denotes the gradient of f. In addition, we have the property that the subdifferential is a closed and convex set. An equivalent definition of $\delta f(x)$ which will be useful is

$$\delta f(x) = \{g \in \mathbb{R}^n \mid \langle g, d\rangle \leq f'(x,d) \ \forall d \in \mathbb{R}^n\}. \tag{1.14}$$

The multivalued function $x \mapsto \delta f(x)$ is upper semicontinuous and locally bounded. Notice that if a point x is a local minimum of the function f we have $f'(x,d) \geq 0$ for all d. In this case, following the second definition (1.14) of the subdifferential, $0 \in \delta f(x)$. Moreover, we say that x is a stationary point for f if $0 \in \delta f(x)$. For any $x \in \mathbb{R}^n$ and any $\epsilon \geq 0$ the Goldstein $a-$subdifferential [9] of f at x is the set

$$\delta_a f(x) = \mathrm{conv}\{\delta f(y) \mid |y-x| \leq a\}. \tag{1.15}$$

The multivalued function $(x,a) \mapsto \delta_a f(x)$ is locally bounded and upper semicontinuous, i.e., $x^k \to x$, $a^k \to a$, $p^k \in \delta_{a^k} f(x^k)$ and $p^k \to p$ imply $p \in \delta_a f(x)$. Finally, for any function $p(x,x')$ satisfying the local Lipschitz property in x', we will use the notation $\delta p(x,x')$ for the subdifferential in the second variable at x'.

2. Generalized Proximal Point Algorithms. In what follows, the regularization function is denoted $\Psi(.,.)$. We first state some assumptions on the objective function f and the regularization function Ψ.

Assumptions 1 (Objective function). *(i)* f is inf-compact, i.e. the $\alpha-$level sets $\mathcal{L}_f(\alpha) = \{x \in \mathbb{R}^n \mid f(x) \leq \alpha\}$ are bounded for any $\alpha \in R$
(ii) f *is locally lipschitz over* $\mathbb{R}^n$.

Notice that Assumptions 1 (i) and (ii) together imply that f is bounded from below.

Assumptions 2 (Regularization function). *(i)* (Positivity) $\Psi(x,y) \geq 0$ for all x, $y \in \mathbb{R}^n$.
(ii) (Identifiability) $\Psi(x,y) = 0 \Leftrightarrow x = y$.
(iii) $\Psi(x,y)$ is locally lipschitz over $\mathbb{R}^n \times \mathbb{R}^n$.
(iv) The effective domain $C_{\Phi(x,.)} = \mathbb{R}^n$ for any $x \in \mathbb{R}^n$.

Define the *generalized proximal point* algorithm

Definition 2.1. *Assume that f and $\Psi(.,.)$ satisfy assumptions 1 and 2 respectively. Then, the following recursion starting at x^1*

$$x^{k+1} = \mathrm{argmin}_{x\in\mathbb{R}^n}\{f(x) + \lambda_k \Psi(x^k, x)\}, \tag{2.1}$$

is well defined and will be called the generalized proximal point *algorithm.*

For the sake of notational convenience, in the remainder of this paper, the regularized function with relaxation parameter t will be denoted $p_t(x,y) = f(y) + t\Psi(x,y)$ and the *generalized proximal* operator will be defined by $P_t(x) = \mathrm{argmin}_{y\in\mathbb{R}^n} p_t(x,y)$.

3. Convergence. In this section, we give an asymptotic analysis of the *generalized proximal point* algorithm which does not require differentiability nor convexity assumptions. We first investigate theoretical difficulties concerning optimality conditions for nonsmooth regularization. Indeed the problem which one encounters in nonsmooth situations is that a stationary point of the regularized objective function is not in general a stationary point of the unregularized objective function itself. Thus, it is important to establish sufficient conditions which guarantee that stationary points of the objective function coincide with those of the regularized problem. For that purpose, we introduce the condition of *subdifferential domination*. Under this condition, we show that the fixed points of the generalized proximal point mapping are locally optimal. Then, we show that the accumulation points of the sequence defined by (2.1) are also locally optimal. Under a strict local convexity assumption local convergence is established.

3.1. Local analysis and optimality conditions.

3.1.1. Nondifferentiable regularization. Consider the regularized function where, with no loss of generality, λ is set to 1

$$p(x,y) = f(y) + \Psi(x,y). \tag{3.1}$$

Using the calculus of subdifferentials, we have

$$\delta p(x,y) \subset \delta f(y) + \delta\Psi(x,y), \tag{3.2}$$

which is the sum of two (closed convex) sets. One easily deduces from Assumptions 2 that $0 \in \delta\Psi(x,x)$ for any $x \in \mathbb{R}^n$. Now, consider the following situation, where x^* is a stationary point of $p(x^*,.)$, i.e.

$$\begin{cases} 0 \in \delta p(x^*,x^*) \subset \delta f(x^*) + \delta\Psi(x^*,x^*) \\ 0 \in \delta\Psi(x^*,x^*). \end{cases} \tag{3.3}$$

The question is to know when such a point x^* is also a stationary point of f.

Note that if $\Psi(x^*,y)$ is smooth in the second variable, e.g. $\Psi(x^*,y) = |x^* - y|^2$, then $\delta\Psi(x^*,x^*)$ reduces to $\{0\}$, and we deduce from (3.3) that $0 \in \delta f(x^*)$ which proves that x^* is at least a stationary point of the objective function. For $\Psi(x^*,y)$ nonsmooth in the second variable, x^* is no longer guaranteed to be a stationary point of f. To illustrate, consider the following monodimensional example,

$$f(y) = \begin{cases} x^* - y, \text{ if } y \le x^* \\ \frac{1}{2}(x^* - y), \text{ if } y > x^* \end{cases}$$

$$\Psi(x^*,y) = |x^* - y|.$$

We then have $\delta f(x^*) = [-1, -\frac{1}{2}]$ and $\delta\Psi(x^*,x^*) = [-1,1]$. On the other hand, $\delta p(x^*,x^*) = [-2, \frac{1}{2}]$. Thus, in this case, we have $0 \in \delta p(x^*,x^*)$ whereas x^* is not a stationary point of f (indeed, f has no stationary point).

The reason that condition (3.3) fails in this example is that the subdifferential of the regularization Ψ is "bigger" than the subdifferential of the objective f. To overcome this difficulty, we will impose that the objective have "greater" variations than the regularization. The following definition makes precise the idea of "greater" variations.

Definition 3.1. *Let f_1 and f_2 be locally Lipschitzian functions. The function f_1 subdifferentially dominates f_2 at the point x if*

$$|f_1'(x,d)| > |f_2'(x,d)| \quad \forall d \in \mathbb{R}^n. \tag{3.4}$$

An important consequence is that if f_1 subdifferentially dominates f_2 at x, then f_1 subdifferentially dominates tf_2 for any t satisfying $0 \le t \le 1$ at x. The following lemma establishes that the conditions given by (3.3) specify a stationary point of the objective function under the subdifferential domination hypothesis.

LEMMA 3.2. *Let $p(x,y) = f(y) + \Psi(x,y)$ and assume the existence of a point x^* satisfying the following stationarity conditions*

$$\begin{cases} 0 \in \delta f(x^*) + \delta\Psi(x^*, x^*) \\ 0 \in \delta\Psi(x^*, x^*). \end{cases} \tag{3.5}$$

If f subdifferentially dominates $\Psi(x^*, .)$ at x^*, then x^* is a stationary point of f.

Proof. For simplicity denote $\Psi(x^*, y)$ by $\Psi_{x^*}(y)$. Since $0 \in \delta\Psi_{x^*}(x^*)$, we have $\Psi'_{x^*}(x^*, d) \geq 0$ for all $d \in \mathbb{R}^n$. Due to the first equation in (3.5), we deduce the existence of two vectors g_1 and g_2 such that

$$g_1 + g_2 = 0,$$
$$g_1 \in \delta f(x^*),$$
$$g_2 \in \delta\Psi_{x^*}(x^*),$$

and thus

$$\langle g_1, d\rangle + \langle g_2, d\rangle = 0,$$

for all $d \in \mathbb{R}^n$. Since, by definition, $f'(x,d) \geq \langle g_1, d\rangle$ and $\Psi_{x^*}(x^*, d) \geq \langle g_2, d\rangle$, we have $f'(x^*, d) + \Psi'_{x^*}(x^*, d) \geq 0$. Then, we obtain

$$f'(x^*, d) \geq -\Psi'_{x^*}(x^*, d). \tag{3.6}$$

We now proceed by contradiction. Suppose that $f'(x^*, d) < 0$ for some $d \in \mathbb{R}^n$. Equation (3.6) implies in this particular case

$$|f'(x^*, d)| \leq |\Psi'_{x^*}(x^*, d)|. \tag{3.7}$$

Since this last equation contradicts the domination assumption, we deduce that $f'(x^*, d) \geq 0$ for all $d \in \mathbb{R}^n$. Using the (second) definition (1.14) of the subdifferential, we conclude that $0 \in \delta f(x^*)$. □

3.1.2. Fixed points of the generalized proximal mapping. Using lemma 3.2 we can study the optimality of the fixed points of the generalized proximal point mapping P_t with relaxation parameter $t \geq 0$. The following lemma characterizes the fixed points of P_t as solutions to the minimization problem (1.1).

LEMMA 3.3. *Assume that $f(x)$ subdifferentially dominates $t\Psi(x,x)$ at each fixed point x^* of the* generalized proximal *mapping P_t with relaxation parameter $t \geq 0$. Then any fixed point x^* of P_t is a stationary point of $f(x)$.*

Proof. The case $t = 0$ is trivial. Thus fix $t > 0$. Let x^* be a fixed point of the proximal mapping P_t, i.e. x^* is a solution of $\text{argmin}_{x\in\mathbb{R}^n} p_t(x^k, x)$ with $p_t(x,y) = f(y) + t\Psi(x,y)$. Then, the optimality condition gives

$$0 \in \delta f(x^*) + t\delta\Psi(x^*, x^*), \tag{3.8}$$

Recalling that $0 \in \delta\Phi(x^*, x^*)$, we can derive from lemma 3.2 that $0 \in \delta f(x^*)$, and the proof is completed. □

With this local result in hand, the next step in the convergence analysis is to establish that the accumulation points of the generalized proximal point algorithm are fixed points of the general proximal mapping.

3.2. Convergence analysis. In this section, we consider global and local convergence. A Lyapunov method is the guideline of the proof where the Lyapunov function is simply the objective $f(x)$.

3.2.1. Main results. We start with the following result.

LEMMA 3.4. *For any iteration $k > 1$, we have*

$$f(x^{k+1}) - f(x^k) \leq -\lambda_k \Psi(x^k, x^{k+1}) \leq 0. \tag{3.9}$$

Proof. Iteration (2.1) implies that $f(x^{k+1}) + \lambda_k \Psi(x^k, x^{k+1}) \leq f(x^k) + \lambda_k \Psi(x^k, x^k)$. Recall that $\Psi(x^k, x^k) = 0$ due to identifiability assumption 2 (ii) and $\Psi(x^k, x^{k+1}) \geq 0$ by positivity assumption 2 (i). Thus (3.9) follows. □

We next deduce the following important property which is sometimes referred to as "asymptotic regularity" [1].

LEMMA 3.5. *Assume that there exists a real number λ such that $\lambda_k \geq \lambda > 0$. Then, the sequence of iterates $(x^k)_{k\in\mathbb{N}}$ satisfies $\lim_{k\to\infty} |x^k - x^{k+1}| = 0$.*

Proof. By lemma 3.4, $(f(x^k))_{k\in\mathbb{N}}$ is decreasing and thus the fact that f is bounded from below implies that $(f(x^k))_{k\in\mathbb{N}}$ converges. Thus, the left hand side of (3.9) tends towards zero and since $\lambda_k \geq \lambda > 0$, for all k, we have $\lim_{k\to\infty} \Psi(x^k, x^{k+1}) = 0$.

We now prove that $\lim_{k\to\infty} |x^k - x^{k+1}| = 0$ by contradiction. Assume that there exits a subsequence $(x^{\sigma(k)})_{k\in\mathbb{N}}$ such that $|x^{\sigma(k)} - x^{\sigma(k)+1}| \geq 3\epsilon$ for some $\epsilon > 0$ and for all large k. The fact that $(f(x)^{\sigma(k)})_{k\in\mathbb{N}} \leq f(x^1)$ implies that $(x^{\sigma(k)})_{k\in\mathbb{N}}$ is bounded due to the inf-compactness assumption 1 (ii). Since $(x^{\sigma(k)})_{k\in\mathbb{N}}$ is bounded, one can extract a convergent subsequence, and thus, we may assume without any loss of generality that $(x^{\sigma(k)})_{k\in\mathbb{N}}$ is convergent with limit x'. Using the triangle inequality, we have $|x^{\sigma(k)} - x'| + |x' - x^{\sigma(k)+1}| \geq 3\epsilon$. Since $(x^{\sigma(k)})_{k\in\mathbb{N}}$ converges to x', there exists a integer K such that $k \geq K$ implies $|x^{\sigma(k)} - x'| \leq \epsilon$. Thus for $k \geq K$ we have $|x' - x^{\sigma(k)+1}| \geq 2\epsilon$. Now extract from $(x^{\sigma(k)+1})_{k\geq K}$ a convergent subsequence $(x^{\sigma(\gamma(k))+1})_{k\geq K}$ with limit x''. Then, using the same arguments as above, we obtain $|x' - x''| \geq \epsilon$. Finally, recall that $\lim_{k\to\infty} \Psi(x^k, x^{k+1}) = 0$. We thus have $\lim_{k\to\infty} \Psi(x^{\sigma(\gamma(k))}, x^{\sigma(\gamma(k))+1}) = 0$, and, due to the fact that the sequences are bounded and $\Psi(.,.)$ is locally Lipschitz (and therefore continuous in both variables), we have $\Psi(x', x'') = 0$. Thus assumption 2 (ii) implies that $|x' - x''| = 0$ and we obtain a contradiction. Hence, $\lim_{k\to\infty} |x^k - x^{k+1}| = 0$ as claimed. □

We are now ready to establish our global convergence theorem.

THEOREM 3.6. *Assume that the sequence $(\lambda_k)_{k\in\mathbb{N}}$ is bounded and satisfies $\lambda^k \geq \lambda > 0$, $\forall k \in \mathbb{N}$ for a given real number λ. Define $\lambda^+ = \limsup_{k\to\infty} \lambda_k$. If f subdifferentially dominates $\lambda^+\Psi(x^*, .)$ at any fixed point x^* of the operator P_{λ^+}, then every accumulation point of the sequence $(x^k)_{k\in\mathbb{N}}$ is a stationary point of $f(x)$.*

Proof. Take a convergent subsequence $(x^{\sigma(k)})_{k\in\mathbb{N}}$ of $(x^k)_{k\in\mathbb{N}}$ with limit point x^*. Lemma 3.5 implies that $(x^{\sigma(k)+1})_{k\in\mathbb{N}}$ also converges to x^*. In accordance with lemma 3.3, we need to prove the existence of a real $\lambda^+ \geq t \geq 0$, such that $p_t(x^*, x^*) \leq p_t(x^*, x)$ for all $x \in \mathbb{R}^n$. In the following, we prove that this result holds with $t = \lambda^+$. Recall that $p_{\lambda_k}(x^{\sigma(k)}, x^{\sigma(k)+1}) \geq p_{\lambda_k}(x^{\sigma(k)}, x)$ for all $x \in \mathbb{R}^n$. Therefore, for all x

$$\begin{aligned} p_{\lambda^+}(x^{\sigma(k)}, x^{\sigma(k)+1}) &+ (\lambda_k - \lambda^+)\Psi(x^{\sigma(k)}, x^{\sigma(k)+1}) \\ &\leq p_{\lambda^+}(x^{\sigma(k)}, x) + (\lambda_k - \lambda^+)\Psi(x^{\sigma(k)}, x). \end{aligned} \tag{3.10}$$

Since $\lim_{k\to\infty} \Psi(x^{\sigma(k)}, x^{\sigma(k)+1}) = 0$, for any $\epsilon > 0$, there exits an integer K_1 such that $\Psi(x^{\sigma(k)}, x^{\sigma(k)+1}) \leq \epsilon$ for all $k \geq K_1$. On the other hand, p_{λ^+} is continuous in both variables, due to the locally Lipschitz property. Fix $x \in \mathbb{R}^n$. By continuity in the first and second arguments of $p_{\lambda^+}(.,.)$, respectively, we have for any $\epsilon > 0$ there exists $K_2 \in \mathbb{N}$ such that for all $k \geq K_2$

$$p_{\lambda^+}(x^*, x) \geq p_{\lambda^+}(x^{\sigma(k)}, x) - \epsilon, \tag{3.11}$$

and

$$p_{\lambda^+}(x^*, x^*) \leq p_{\lambda^+}(x^{\sigma(k)}, x^{\sigma(k)+1}) + 2\epsilon. \tag{3.12}$$

Combining equations (3.11) and (3.12) with (3.2.1), we obtain

$$p_{\lambda^+}(x^*, x^*) \leq p_{\lambda^+}(x^*, x) - (\lambda^+ - \lambda_k)(\Psi(x^{\sigma(k)}, x) - \Psi(x^{\sigma(k)}, x^{\sigma(k)+1})) + 3\epsilon. \tag{3.13}$$

Now, since $\lambda^+ = \limsup_{k\to\infty} \lambda_k$, there exists an integer K_3 such that $\lambda^+ - \lambda_k \geq -\epsilon$ for all $k \geq K_3$. Then for all $k \geq \max\{K_1, K_2, K_3\}$, we easily obtain

$$p_{\lambda^+}(x^*, x^*) \leq p_{\lambda^+}(x^*, x) + \epsilon\Psi(x^{\sigma(k)}, x) - \epsilon^2 + 3\epsilon.$$

Since $(x^{\sigma(k)})_{k\in\mathbb{N}}$ is bounded and due to Assumption 2(iv) and Assumption 2(iii), there exists an upper bound C such that $\Psi(x^{\sigma(k)}, x) \leq C$ for all $k \geq \max\{K_1, K_2, K_3\}$. Thus, we have

$$p_{\lambda^+}(x^*, x^*) \leq p_{\lambda^+}(x^*, x) + \epsilon C - \epsilon^2 + 3\epsilon.$$

Since no assumption was made on x, this holds for any $x \in \mathbb{R}^n$. Thus, letting ϵ tend to zero, we see that x^* is a fixed point of $\mathrm{argmin}_{x\in\mathbb{R}^n} p_{\lambda^+}(x^k, x)$. Furthermore, recall that $f(x)$ subdifferentially dominates $\lambda^+\Psi(x, x)$ at the point x^*. Therefore, Lemma 3.3 implies that x^* is a stationary point of $f(x)$. ◻

3.2.2. Convergence to local optima under additional convexity assumptions. Let S^* be the set of accumulation points of the sequence $(x^k)_{k\in\mathbb{N}}$. We first establish the following lemma

LEMMA 3.7. *The set x^* of accumulation points of the sequence $(x^k)_{k\in N}$ is compact and connected.*

Proof. This result follows directly from the fact that $\lim_{k\to\infty} |x^{k+1} - x^k| = 0$ and from [21, Theorem 28.1]. ◻

COROLLARY 3.8. *Suppose, in addition to the assumptions of Theorem 3.6, that $f(x)$ is strictly convex in an open neighborhood $\mathcal{N}$ of an accumulation point x^* of the sequence $(x^k)_{k\in\mathbb{N}}$. Then the sequence $(x^k)_{k\in\mathbb{N}}$ converges to a local minimizer of $f(x)$.*

Proof. We obtained in Theorem 3.6 that every accumulation point of $(x^k)_{k\in\mathbb{N}}$ is a stationary point of $f(x)$. Since $f(x)$ is strictly convex over $\mathcal{N}$, the set of stationary points of $f(x)$ belonging to $\mathcal{N}$ reduces to singleton. Thus x^* is the unique stationary point of $f(x)$, and *a fortiori*, the unique accumulation point of $(x^k)_{k\in\mathbb{N}}$ belonging to $\mathcal{N}$. To complete the proof, it remains to show that there is no accumulation point in the exterior of $\mathcal{N}$. For that purpose, consider an open ball $\mathcal{B}$ of center x^* and radius ϵ included in $\mathcal{N}$. Hence, x^* is the unique accumulation point in $\mathcal{B}$. Moreover, any accumulation point x', lying in the exterior of $\mathcal{N}$ must satisfy $|x^* - x'| \geq \epsilon$, and we obtain a contradiction with the fact that S^* is connected. Thus every accumulation point lies in $\mathcal{N}$, from which we conclude that x^* is the only accumulation point of $(x^k)_{k\in\mathbb{N}}$ or, in other words, that $(x^k)_{k\in\mathbb{N}}$ converges towards x^*. Finally, notice that the strict convexity of $f(x)$ over $\mathcal{N}$ implies that x^* is a local minimizer and the proof is completed. ◻

4. Bundle implementations. The study of the generalized proximal point algorithm gives an elegant framework for the exploration of a large class of regularizations. In this section, we introduce a bundle implementation in order to make this approach more tractable in applications. The bundle approximation we propose is based on the work of Kiwiel [9].

4.1. Background. The bundle method has been widely recognized as a very efficient technique for minimization of nondifferentiable functions. Bundle methods can be interpreted as stepwise approximation of the proximal point algorithm [11] and they can be used to obtain tractable implementations of proximal point methods with variable metric [3]. However, bundle methods for general nonquadratic regularizations considered in this a paper have not been considered. This may be because, for computational tractability, bundle methods are essentially sequential quadratic programs, which would seem to exclude the possibility of nonquadratic regularization.

The main idea behind application of bundle methods to nonquadratic proximal point algorithms is the following. Consider the following polyhedral approximation of the regularized proximal function $p_{\lambda_k}(x^k, x)$ at iteration k,

$$\hat{p}_{\lambda_k}(x^k, x) = p_{\lambda_k}(x^k, x^k) + \max_{j \in J^k}\{-\alpha_j^k + \langle g^j, x - x^k\rangle\}, \tag{4.1}$$

with

$$\alpha_j^k = \alpha(x^k, y^j)$$

where

$$\alpha(x, y) = |p_{\lambda_k}(x, x) - p_{\lambda_k}(x, y) - \langle g, x - y\rangle|$$

and $g \in \delta p_{\lambda_k}(x, y)$. This model is simply the penalized polyhedral approximation given in (1.4), adapted to the current regularized objective $p_{\lambda_k}(x^k, x)$. Now, an approximation of the generalized proximal point iteration (2.1) is obtained by computing

$$x^{k+1} = \mathrm{argmin}_{x\in\mathbb{R}^n}\ \{\hat{p}_{\lambda_k}(x^k, x) + |x^k - x|^2\}, \tag{4.2}$$

where the quadratic term is the usual stabilization of the polyhedral approximation as in (1.2).

Below, a detailed algorithm is given which has similar algorithmic structure as the one proposed in [9]. Lemma 4.2 and Lemma 4.3 are original and apply to more refined algorithms as in [12] which uses a trust region approach. Before stating the detailed algorithm, we give a high level explanation of the method. As in any bundle algorithm, the actual minimization at iteration k is divided into two steps. First an attempt is made at finding a direction of descent d^k, using the following reformulation of (4.2)

$$d^k = \mathrm{argmin}_{d\in\mathbb{R}^n}\ \hat{p}_{\lambda_k}(x^k, x^k + d) + |d|^2, \tag{4.3}$$

and an approximate line search is performed along the direction d^k to produce a point y^k about the current iterate x^k. If a sufficient decrease of $p_{\lambda_k}(x^k, x)$ is achieved at y^k, then a serious step is made, i.e. $x^{k+1} = y^k$. Otherwise, a null step occurs, i.e. $x^{k+1} = x^k$ and the local subgradient $g^{k+1} \in \delta p_{\lambda_k}(x^k, y^k)$ is used to enrich the polyhedral model (4.1) so that a better direction of descent can be computed. The decrease is evaluated in terms of the regularization function as for the general proximal point described in the former sections. Hence, we only require that

$$p_{\lambda_k}(x^k, y^k) \leq p_{\lambda_k}(x^k, x^k), \tag{4.4}$$

in order for a serious step to be achieved, implying the following decrease in the objective function

$$f(y^k) - f(x^k) \leq -\lambda_k \Psi(x^k, y^k) \leq 0, \tag{4.5}$$

using (3.9). To ensure convergence, two conditions must be enforced. Firstly, one must enforce that if a null step occurs, the subgradient g^{k+1} will provide additional information about the subdifferential of $p_{\lambda_k}(x^k, x)$ at x^k. For this purpose, it is sufficient that $\langle g^{k+1}, d^k \rangle > 0$. Depending on the line search, this can be guaranteed by the Bihain semidifferentiability criterion defined below. Secondly, the approximation must be refined about accumulation points by storing only the local subgradients. For this purpose, a deletion rule is used, which deletes past subgradients in accordance to a locality measure a^k.

We now define some additional assumptions needed in the sequel.

Assumptions 3 (Additional assumptions). (i) (Bihain's upper semidifferentiability) *For any* $x, y, d \in \mathbb{R}^n$,

$$\limsup_{t \to 0} \{ \frac{p_\lambda(y, x+td) - p_\lambda(y, x)}{t} - \langle g(x+td), d \rangle \\ \mid \quad g(x+td) \in \delta p_\lambda(y, x+td) \ \} \leq 0.$$

(ii) (Quazisymmetry) *There exists a function* $q(.,.)$, $\mathbb{R}_+^2 \mapsto \mathbb{R}_+$ *satisfying*
a. $q(.,b)$ *is increasing for all* $b \in \mathbb{R}_+$ *and* $q(a,.)$ *is increasing for all* $a \in \mathbb{R}_+$,
b. $q(.,.)$ *is continuous at the point* $(0,0)$,
c. $\lim_{(a,b)\to(0,0)} q(a,b) = 0$,
and such that if $g \in \delta_a \Psi(x, y)$, $a \geq 0$ *then* $g \in -\delta_{q(a,|x-y|)} \Psi(y, x)$.

These assumptions are not restrictive since (i) is natural in the nonconvex nondifferentiable context [12] and (ii) is satisfied with $q(a, |x-y|) = a$ for the Moreau-Yosida regularization $\Psi(x, y) = |x-y|^2$. This assumption of quazisymmetry permits utilization of nonsymmetric regularization functions. However, notice that quazisymmetry implies that the subgradient $\delta\Psi(x, x)$ be symmetric for all $x \in \mathbb{R}^n$.

4.2. The regularized bundle method. We now describe our regularized bundle method.

Algorithm 1 (Regularized Bundle Method).

Step 0 *(Initialization) Select the starting point* x^1 *and a final accuracy tolerance* $\epsilon_s > 0$. *Choose the line search parameters* $m_R \in (0,1)$, $m_\alpha \in (0, m_R)$, $\kappa \in (0,1)$ *and a positive reset tolerance* m_a. *Set the threshold stepsize* $t^1 = 1$ *and the reset indicator* $r_a^1 = 1$. *Set* $J^1 = \{1\}$, $y^1 = x^1$, $g^1 \in \delta p(x^1, x^1)$, $G^1 = \{g^1\}$ *and* $s_1^1 = 0$. *Set the counters* $k = 1$, $l = 0$ *and* $k(0) = 1$.

Step 1 *(Direction finding) Find the solution* $u^k = (u_j^k)_{j \in J^j}$ *to the following quadratic subproblem*

$$\begin{aligned} &\min_u \{ \tfrac{1}{2} \Big| \textstyle\sum_{j \in J^k} u_j g^j \Big|^2 + \sum_{j \in J^k} u_j \alpha_j^k \} \\ &\textit{subject to } u_j \geq 0, \quad \textstyle\sum_{j \in J^k} u_j = 1. \end{aligned} \tag{4.6}$$

Set

$$
\begin{aligned}
d^k &= -\sum_{j\in J^k} u_j^k g^j \\
\alpha^k &= \sum_{j\in J^k} u_j^k \alpha_j^k \\
v^k &= -\left(|d^k|^2 + \alpha^k\right)
\end{aligned}
\tag{4.7}
$$

If $r_a^k = 1$*, set* $a^k = \max\{s_j^k \mid j \in J^k\}$.

Step 2 *(Stopping criterion) If* $\max\{|d^k|, m_a a^k\} \le \epsilon_s$*, terminate. Otherwise go to Step 3.*

Step 3 *(Resetting test) If* $|d^k| > m_a a^k$ *then go to Step 5. Otherwise go to Step 4.*

Step 4 *(Resetting) (i) If* $r_a^k = 0$ *then set* $r_a^k = 1$ *and go to step 1.*

(ii) If $J^k \neq \{k(l)\}$ *then delete from* J^k *one among the elements* j *with the largest* $s_j^k > 0$ *and go to Step 1.*

Step 5 *(Line search) Compute two stepsizes* t_L^k *and* t_R^k*,* $t_L^k \le t_R^k$ *using procedure 1 below and such that the two corresponding points* $x^{k+1} = x^k + t_L^k d^k$ *and* $y^{k+1} = x^k + t_R^k d^k$ *satisfy either*

(serious step)

$$
(f(x^k) + \lambda_k \Psi(x^k, x^{k+1}) =) \quad p_{\lambda_k}(x^k, x^{k+1}) < p_{\lambda_k}(x^k, x^k) \quad (= f(x^k)),
\tag{4.8}
$$

and $t_L = t_R > t^k$ *or* $\alpha(x^k, x^{k+1}) > m_\alpha |v^k|$*, or*

(null step)

$$
t_R^k \le t^k \tag{4.9}
$$

$$
-\alpha(x^{k+1}, yk+1) + \langle g^{k+1}, d^k\rangle \ge m_R v^k, \quad g^{k+1} \in \delta p_{\lambda_k}(x^k, y^{k+1}) \tag{4.10}
$$

and $t_L^k = 0$*. Set*

$$
g^{k+1} \in \delta p_{\lambda^k}(x^k, y^{k+1}). \tag{4.11}
$$

Step 6 *(Threshold stepsize updating) If* $t_L^k = 0$*, set* $t^{k+1} = \kappa t^k$*. Otherwise, set* $t^{k+1} = 1$*,* $k(l+1) = k+1$ *and increase* l *by* 1*.*

Step 7 *(Updating) Set*

$$
s_j^{k+1} = s_j^k + |x^{k+1} - x^k|, \tag{4.12}
$$

$$
s_{k+1}^{k+1} = |x^{k+1} - y^{k+1}| \tag{4.13}
$$

$J^{k+1} = J^k \cup \{k+1\}$*, calculate* $a^{k+1} = \max\{a^k + |x^{k+1} - x^k|, s_{k+1}^{k+1}\}$ *and set* $r_a^k = 0$*.*

Step 8 *Increase* k *by* 1 *and go to Step 1.*

Notice that in the algorithm, the direction search (4.2) is replaced by its dual formulation (see [9] for details). We now describe the line search procedure.

PROCEDURE 1 (Line search).

(i) Set $t_L = 0$ *and* $t = t_U = 1$*.*

(ii) If $p_{\lambda_k}(x^k, x^k + td^k) \le p_{\lambda_k}(x^k, x^k)$ *set* $t_L = t$*. Otherwise set* $t_U = t$*.*

(iii) If $p_{\lambda_k}(x^k, x^k+td^k) \leq p_{\lambda_k}(x^k, x^k)$ and either $t_L \geq t^k$ or $\alpha(x^k, x^{k+1} > m_\alpha |v^k|$, set $t_L^k = t_R^k = t$ and return.

(iv) If $t < t^k$ and

$$-\alpha(x^k, x^k+td^k) + \langle g, d^k \rangle \geq m_R v^k, \quad g \in \delta p_{\lambda_k}(x^k, x^k+td^k) \tag{4.14}$$

set $t_R^k = t$, $t_L^k = 0$ and return.

(v) Choose $t = \frac{t_L+t_U}{2}$ and go to (ii).

At this point, we would like to make the following comment. The algorithm presented here requires computation of subgradients in $\delta p_{\lambda_k}(x^k, y)$ at certain points y involved in the line search. However, all the results in the sequel clearly hold in the case where the subgradients are taken in $\delta f(y) + \lambda_k \delta\Psi(x^k, y)$. We now establish convergence results.

4.3. Convergence. We assume that the functions f and Ψ satisfy assumptions 1, 2 and 3. Convergence of the line search procedure is a standard result when assumption 3 (i) is satisfied (see [12]). The convergence theorem below relies on the following lemmas

LEMMA 4.1. *All the sequences $(x^k)_{k\in\mathbb{N}}$, $(y^k)_{k\in\mathbb{N}}$, $(g^k)_{k\in\mathbb{N}}$ and $(d^k)_{k\in\mathbb{N}}$ defined in the regularized bundle algorithm are bounded.*

Proof. Due to assumption 1 (i) of inf-compactness, and the fact that algorithm 1 is a descent method, the sequence $(x^k)_{k\in\mathbb{N}}$ is bounded. The remainder of the proof can be easily adapted from [9, Lemma 3.3]. ◻

LEMMA 4.2. *At any iteration k, define $\lambda_k^+ = \sup_{j\in J^k} \lambda_j$. Fix a point x^* in the level set $\{x \mid f(x) \leq f(x^1)\}$ and define $b^k = |x^k - x^*|$. Then, under Assumptions 3, for any $k \in \mathbb{N}$, the direction d^k defined in algorithm 1 satisfies*

$$d^k \in -\Big(\delta_{a^k} f(x^k) - \lambda_k^+ \delta_{q(a^k+b^k, a^k+b^k)+a^k} \Psi(x^*, x^k)\Big), \tag{4.15}$$

where a^k is the locality measure defined in steps 1 and 7 of the regularized bundle method.

Proof. Using (4.11), we have for any $j \in J^k$,

$$g^j \in \delta f(y^j) + \lambda_j \delta\Psi(x^j, y^j). \tag{4.16}$$

Therefore, using the fact that $|x^k - y^j| \leq a^k$ and the definition of the Goldstein $a-$subdifferential

$$g^j \in \delta_{a^k} f(x^k) + \lambda_j \delta_{a^k} \Psi(x^j, x^k), \tag{4.17}$$

and

$$g^j \in \delta_{a^k} f(x^k) + \lambda_j \delta_{a^k+b^k} \Psi(x^j, x^*). \tag{4.18}$$

Now, by assumption 3 (ii) of quazisymmetry,

$$g^j \in \delta_{a^k} f(x^k) - \lambda_j \delta_{q(a^k+b^k, |x^*-x^j|)} \Psi(x^*, x^j), \tag{4.19}$$

Thus, using the fact that $|x^k - x^j| \leq a^k$, $|x^* - x^j| \leq |x^* - x^k| + a^k$, and that $q(a, .)$ is increasing for all fixed $a \in \mathbb{R}_+$ as required by Assumptions 3(ii) a, we obtain

$$g^j \in \delta_{a^k} f(x^k) - \lambda_j \delta_{q(a^k+b^k, |x^*-x^k|+a^k)+a^k} \Psi(x^*, x^k), \tag{4.20}$$

and, recalling that $b^k = |x^* - x^k|$,

$$g^j \in \delta_{a^k} f(x^k) - \lambda_j \delta_{q(a^k+b^k, a^k+b^k)+a^k} \Psi(x^*, x^k), \tag{4.21}$$

Since $\lambda_k^+ \geq \lambda_j$, we have

$$-\lambda_j \delta_{q(a^k+b^k, a^k+b^k)+a^k} \Psi(x^*, x^k) \subset -\lambda_k^+ \delta_{q(a^k+b^k, a^k+b^k)+a^k} \Psi(x^*, x^k) \tag{4.22}$$

which, combined with (4.21), gives

$$g^j \in \delta_{a^k} f(x^k) - \lambda_k^+ \delta_{q(a^k+b^k, a^k+b^k)+a^k} \Psi(x^*, x^k), \tag{4.23}$$

for any $j \in J^k$. Now, due to equation (4.7), $-d^k$ is a convex combination of the subgradients g^j, $j \in J^k$. Recalling that $a-$Goldstein subdifferential are closed and convex sets, we have

$$d^k \in -\Big(\delta_{a^k} f(x^k) - \lambda_k^+ \delta_{q(a^k+b^k, a^k+b^k)+a^k} \Psi(x^*, x^k)\Big) \tag{4.24}$$

as desired. □

Lemma 4.3. *Suppose that there exist sequences $(x^k)_{k\in\mathbb{N}}$, $(d^k)_{k\in\mathbb{N}}$ and $(a^k)_{k\in\mathbb{N}}$ and a point $x^* \in \mathbb{R}^n$ such that $\lim_{k\to\infty} x^k = x^*$, $\lim_{k\to\infty} d^k = 0$, $\lim_{k\to\infty} a^k = 0$. Define $\lambda^+ = \limsup_{k\to\infty} \lambda_k^+$, where $\lambda_k^+ = \max_{j\in J^k} \lambda_j$. Assume in addition that $\lambda^+ < \infty$ and f subdifferentially dominates $\lambda^+\Psi(x^*, .)$ at any point x^* where $0 \in \delta f(x^*) + \delta\Psi(x^*, x^*)$ (e.g. x^* is a stationary point of $p_{\lambda^+}(x^*, y)$). Then x^* is a stationary point of $f(x)$.*

Proof. Lemma 4.2 above implies the existence of two sequences $(d_1^k)_{k\in\mathbb{N}}$ and $(d_2^k)_{k\in\mathbb{N}}$ such that $d^k = d_1^k + d_2^k$ and

$$-d_1^k \in \delta_{a^k} f(x^k),$$
$$-d_2^k \in -\lambda_k^+ \delta_{q(a^k+b^k, a^k+b^k)+a^k} \Psi(x^*, x^k).$$

Due to lemma 4.1, $(x^k)_{k\in\mathbb{N}}$ is bounded and thus the local boundedness of the subgradients implies that the sequences $(d_1^k)_{k\in\mathbb{N}}$ and $(d_2^k)_{k\in\mathbb{N}}$ are bounded. Thus, we can extract convergent subsequences $(d_1^{\sigma(k)})_{k\in\mathbb{N}}$, and $(d_2^{\sigma(\gamma(k))})_{k\in\mathbb{N}}$ with respective limits d_1^* and d_2^* such that

$$d_1^* + d_2^* = 0. \tag{4.25}$$

On the other hand, using [12, lemma 3.1], we obtain

$$-d_1^* \in \delta f(x^*). \tag{4.26}$$

Furthermore, since $\lim_{k\to\infty} x^k = x^*$, we have $\lim_{k\to\infty} b^k = 0$ and thus $\lim_{k\to\infty} q(a^k + b^k, b^k + a^k) + a^k = 0$, due to Assumption 3(ii) c. Using [12, lemma 3.1] and the fact that $\lambda^+ = \limsup_{k\to\infty} \lambda_k^+$ one easily obtains

$$-d_2^* \in -\lambda^+ \delta\Psi(x^*, x^*). \tag{4.27}$$

Now, due to quazisymmetry Assumption 3 (ii) in the case $a = 0$ and $x = y = x^*$, we have

$$-d_2^* \in \lambda^+ \delta\Psi(x^*, x^*). \tag{4.28}$$

Finally, (4.25), (4.26) and (4.28) together give

$$0 \in \delta f(x^*) + \lambda^+ \delta\Psi(x^*, x^*). \tag{4.29}$$

The proof is then easily completed using lemma 3.2. □

The following lemma due to Kiwiel [8] will be used in the proof of Theorem 4.5. In particular, given $\epsilon > 0$, this lemma implies existence of a finite number of successive null steps without reset such that $|d^k| < \epsilon$.

LEMMA 4.4. [8, 9] *Let* $w^k = \frac{1}{2}|d^k|^2 + \alpha^k$ *and assume* $t_L^k = 0$ *and* $r_a^k = 0$. *Then, there exists a constant* C *independent of* k *such that*

$$0 \le w^{k+1} \le w^k - (1 - m_R)^2 (w^k)^2 / 8C^2. \tag{4.30}$$

We are now in position to establish convergence of the regularized bundle method in the following theorem.

THEOREM 4.5. *Assume that* f *subdifferentially dominates* $\lambda^+\Psi(x^*, .)$ *at any point* x^* *where* $0 \in \delta f(x^*) + \delta\Psi(x^*, x^*)$ *(e.g.* x^* *is a stationary point of* $p_{\lambda^+}(x^*, y)$*), with* λ^+ *defined as in lemma 4.3 and* $\lambda^+ < \infty$. *In addition, assume that* $\lambda_k \ge \lambda > 0$ *for a given* λ *and for any* $k \in \mathbb{N}$. *Then, any accumulation point of the sequence* $(x^k)_{k\in\mathbb{N}}$ *is a stationary point of* $f(x)$.

Proof. Let x^* be an accumulation point of $(x^k)_{k\in\mathbb{N}}$, and let $(x^{\sigma(k)})_{k\in\mathbb{N}}$ be a subsequence of $(x^k)_{k\in\mathbb{N}}$ converging to x^*.

In the case when a finite number of serious steps occurs, the proof of [9, lemma 3.4] can easily be adapted to our method, which establishes that $\lim_{k\to\infty} a^k = 0$ and $\lim_{k\to\infty} d^k = 0$. Then, using lemma 4.3, we easily conclude that x^* is a stationary point of f.

Now, consider the case when an infinite number of serious steps occurs. Although parts of the proof are similar to Lemma 3.5 in [9], all the details are provided for the sake of completeness. Notice that we may reduce the study to the sequence $(x^{k(\sigma(l))})_{l\in\mathbb{N}}$, consisting of serious steps only. Let K_l denote the set of indices between two successive serious steps defined by $K_l = \{k(\sigma(l)), \ldots, k(\sigma(l)+1) - 1\}$.

First, suppose that $\liminf_{l\to\infty} \min_{K_l}\{\max\{|d^k|, a^k\}\} = 0$. Thus, there exists a subsequence of indices $k(\gamma(l))$ such that $\lim_{l\to\infty} |d^{k(\gamma(l))}| = 0$ and $\lim_{l\to\infty} a^{k(\gamma(l))} = 0$. Now consider the sequence $(x^{k(\gamma(l))})_{l\in\mathbb{N}}$. Since $k(\gamma(l)) \in K_l$ and K_l contains the indices preceding the next serious step, we have $x^{k(\gamma(l))} = x^{k(\sigma(l))}$, and therefore $(x^{k(\gamma(l))})_{l\in\mathbb{N}}$ converges to x^*. Hence, there exists subsequences $(x^{k(\gamma(l))})_{l\in\mathbb{N}}$, $(a^{k(\gamma(l))})_{l\in\mathbb{N}}$ and $d^{k(\gamma(l))})_{l\in\mathbb{N}}$ satisfying the conditions of lemma 4.3, from which we deduce that x^* is a stationary point of f.

Next, suppose that $\max\{|d^k|, a^k\} \ge \epsilon > 0$ for all large k. In what follows, we show that, in this case, we obtain a contradiction. Consider the subsequence $(x^{k(l)})_{l\in\mathbb{N}}$ consisting of every serious steps. Then, equation (3.9) gives

$$f(x^{k(l+1)}) - f(x^{k(l+1)-1}) \le -\lambda_k \Psi(x^{k(l+1)-1}, x^{k(l+1)}) \le 0. \tag{4.31}$$

Now since only null steps occur between iterations $k(l)$ and $k(l+1)$, we have $x^{k(l)} = x^{k(l+1)-1}$, and

$$f(x^{k(l+1)}) - f(x^{k(l)}) \le -\lambda_k \Psi(x^{k(l)}, x^{k(l+1)}) \le 0. \tag{4.32}$$

Subsequently, $(f(x^{k(l)}))_{l\in\mathbb{N}}$ is decreasing and thus convergent due to assumption 1 (ii). Hence, we obtain from (4.32) and the assumption $\lambda_k \ge \lambda > 0$, that

$$\lim_{l\to\infty} \Psi(x^{k(l)}, x^{k(l+1)}) = 0. \tag{4.33}$$

Now, using the same method of proof as in Lemma 3.5, we obtain that $\lim_{l\to\infty} |x^{k(l)} - x^{k(l+1)}| = 0$. As a result, we deduce that $\lim_{l\to\infty} t_L^{k(l)} d^{k(l)} = 0$.

Recall that, at any serious step, we have $|d^{k(l)}| > m_a a^{k(l)}$, and thus $|d^{k(l)}| \geq \epsilon$ and $|v^{k(l)}| \geq \epsilon$. The fact that $\lim_{l\to\infty} t_L^{k(l)} d^{k(l)} = 0$, as proved above, implies that $\lim_{l\to\infty} t_L^{k(l)} = 0$. Therefore, $\alpha(x^{k(l-1)}, x^{k(l)})$ tends to zero, and thus, due to the rules of step (iii) in the line search, $t_L^{k(l)} \geq t^{k(l)}$ for l large enough. It follows that $\lim_{l\to\infty} t^{k(l)} = 0$. The remainder of the proof is then similar to the one of Lemma 3.5 in [9]. Indeed, since $\lim_{l\to\infty} t^{k(l)} = 0$ and $t^k = \kappa^{k(l+1)-1-k(l)}$, we have $\lim_{l\to\infty} |k(l+1) - k(l)| = \infty$. In other words, the interval between two successive serious steps tends to infinity. On the other hand, Lemma 9 implies that only a finite number of null steps is sufficient in our case for having $|d^k| < m_a a^k$ and thus obtaining a reset. We conclude that the number of resets occurring between two successive serious steps grows to infinity. Lemma 3.4 (ii) in [9] thus implies that a^k tends towards zero. Since during a reset $|d^k| \leq m_a a^k$, there exits a subsequence $(|d^{\gamma(k)}|)_{k\in\mathbb{N}}$, taken at reset steps, converging towards zero. Hence both subsequences $(|d^{\gamma(k)}|)_{k\in\mathbb{N}}$ and $(a^{\gamma(k)})_{k\in\mathbb{N}}$ tend towards zero which leads to a contradiction. Hence, $\liminf_{l\to\infty} \min_{K_l}\{\max\{|d^k|, a^k\}\} = 0$, and, as already proved above, x^* is a stationary point of f. ☐

5. Numerical experiments. In this section, we illustrate the bundle implementation of Section 4 for a particular nonsmooth optimization problem arising in adaptive digital signal processing for channel equalization and which may be formulated as follows

$$\min_{x\in\mathbb{R}^n} f(x), \qquad f(x) = \sum_{i=1}^{n} |1 - |a_i^T x|^p|^q. \tag{5.1}$$

for a given set of $n-$dimensional vectors $(a_i)_{1\leq i\leq n}$. The usual algorithm for implementing (5.1) is the well known constant modulus algorithm (CMA) [6, 28, 25]. The objective function has the shape represented in Figure 6.1, for the case $n = 1$, $p = 2$, $q = 1$.

Numerical experiments were performed in order to compare the computational properties of the three following regularization functions.

$$\Psi_1(x, y) = \frac{1}{2}|x - y|^2 \text{ (Moreau-Yosida)},$$
$$\Psi_2(x, y) = \sum_{i=1}^{n} |x_i - y_i| \ (l_1 \text{ norm})$$
$$\Psi_3(x, y) = \sum_{i=1}^{n} \log(1 + |x_i - y_i|),$$

These three functions offer smooth convex, nonsmooth convex and nonsmooth nonconvex alternatives; see Figures 6.2, 6.3 and 6.4.

We will compare the three regularizations on the basis of the number of iterations required, the number of function/subgradient calls and the computed optimum value of the objective function. The algorithms were run over a range of the relaxation parameter $\lambda_k = \lambda$ which we kept constant along iterations. The parameters of the regularized bundle method are set as in Table 6.1. Our results are reported in Table 6.2. The symbols it, f^* and nb denote, respectively, the number of iterations necessary to achieve an accuracy smaller that ϵ_s, the final objective value and the total number of function/subgradient calls.

The table clearly demonstrates that the use of nonsmooth regularization functions is competitive with the standard Moreau-Yosida regularization. Moreover, the best candidate often appears to be one of the nonsmooth regularizations Ψ_2 and Ψ_3 with regard to the number of function/subgradient calls and the number of iterations required for the algorithm to terminate.

Algorithmic refinements of the basic regularized bundle method proposed in this paper will be investigated in future works. Notice that the theoretical results demonstrated in section 4 may apply to more sophisticated approaches recently developed for the standard Moreau-Yosida regularization, including trust region techniques [12] and second order methods [14] in the case of convex objective and regularization functions.

6. Conclusion. This paper developed a convergence analysis for the proximal point algorithm with general regularization functions. Strong convergence results were obtained without any assumption on differentiability nor convexity of the regularization. A sufficient condition, denoted "subdifferential domination", on the regularization was defined in order for the method to converge toward a stationary point. Similar convergence results were given for regularized bundle implementation of the proximal method. The resulting bundle algorithm enjoys additional degrees of freedom when compared to the standard procedures, since many types of regularization may be easily implemented in the procedure. Numerical experiments demonstrate that nonconvex and even nonsmooth regularizations are competitive with the standard Moreau-Yosida.

REFERENCES

[1] H. H. Bauschke and J. M. Borwein. On projection algorithms for solving convex feasibility problems. *SIAM Review*, 38(3):367–426, 1996.

[2] A. Bihain. Optimization of upper semidifferentiable functions. *Journal of Optimization Theory and Applications*, 44:545–568, 1984.

[3] J. Bonnans, J. Gilbert, C. Lemarechal, and C. Sagatizabal. A familly of variable metric proximal methods. *Mathematical Programming*, 68:15–48, 1995.

[4] F. H. Clarke. *Optimization and nonsmooth analysis*. Wiley-Interscience, 1983.

[5] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood for incomplete data via the em algorithm (with discussion). *Journal of the Royal Statistical Society*, 39:1–38, 1977.

[6] M. Godard. Self-recovering equalization and carrier tracking in two-dimensional data communication systems. *IEEE Trans. Commun.*, 28:1867–1875, 1980.

[7] A. N. Iusem, B. Svaiter, and M. Teboulle. Entropy-like proximal methods in convex programming. *Mathematics of Operations Research*, 19:790–814, 1994.

[8] K. C. Kiwiel. An aggregate subgradient method for nonsmooth convex minimization. *Mathematical Programming*, 27:320–341, 1983.

[9] K. C. Kiwiel. A linearization algorithm for nonsmooth minimization. *Mathematics of Operations Research*, 10(2):185–194, 1985.

[10] K. C. Kiwiel. A method of linearization for linearly constrained nonconvex nonsmooth minimization. *Mathematical Programming*, 34:175–187, 1986.

[11] K. C. Kiwiel. Proximity control in bundle methods for convex nondifferentiable minimization. *Mathematical Programming*, 46:105–122, 1990.

[12] K. C. Kiwiel. Restricted step and Levenberg-Marquard techniques in proximal bundle methods for nonconvex nondifferentiable optimization. *SIAM Journal on Optimization*, 6(1):227–249, 1996.

[13] C. Lemarechal. An extention of Davidon methods to nondifferentiable problems. *Mathematical Programming Study*, 3:145–173, 1975. M. Balinski and P. Wolfe, Eds.

[14] C. Lemarechal and C. Sagastizabal. Practical aspects of the moreau yosida regularization: Theoretical prelimiaries. *SIAM Journal on Optimization*, 7(2):867–895, 1997.

[15] C. Lemarechal, J. J. Strodiot, and A. Bihain. On a bundle algorithm for nonsmooth optimization. *Nonlinear Programming*, 4:245–281, 1981. O. L. Mangasarian, R. R. Meyer and S. M. Robinson, Eds.

[16] B. Martinet. Régularisation d'inéquation variationnelles par approximations successives. *Revue Francaise d'Informatique et de Recherche Operationnelle*, 3:154–179, 1970.

[17] R. Mifflin. An algorithm for constrained optimization with semismooth functions. *Mathematics of Operations Research*, 2:191–207, 1977.

[18] R. Mifflin. A modification and extension of Lemarechal's algorithm for nonsmooth minimization. nondifferential and variational techniques in optimization. *Mathematical Programming Study*, 17:77–90, 1982. D.C. Sorensen and R.-J.B. Wets, Eds.

[19] M. Nikolova. Estimées localement fortement homogènes. *Comptes Rendus de l'Académie des Sciences, Paris, Série I*, 325:665–670, 1997.

[20] M. Nikolova. Local strong homogeneity using a regularized estimator. *Internal Report, UFR Mathématiques et Informatiques, Université René Descartes, Paris 5*, 1997.

[21] A. M. Ostrowski. *Solution of equations and systems of equations*. Academic, New York, 1966.

[22] R. T. Rockafellar. Augmented lagragians and application of the proximal point algorithm in convex programming. *Mathematics of Operations Research*, 1:96–116, 1976.

[23] R. T. Rockafellar. Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization*, 14:877–898, 1976.

[24] H. Schramm and J. Zowe. A version of the bundle idea for minimizing a nonsmooth function: conceptual idea, convergence analysis, numerical results. *SIAM Journal on Optimization*, 2(1):121–152, 1992.

[25] O. Tanrikulu, A. G. Constantinides, and J. A. Chambers. New normalized constant modulus algorithms with relaxation. *IEEE Sig. Proc. Letters*, 4(9):256–258, 1997.

[26] M. Teboulle. Entropic proximal mappings with application to nonlinear programming. *Mathematics of Operations Research*, 17:670–690, 1992.

[27] M. Teboulle. Convergence of proximal-like algorithms. *SIAM Journal on Optimization*, 7:1069–1083, 1997.

[28] J. R. Treichler and M. G. Larimore. New processing techniques based on the constant modulus adaptive algorithm. *IEEE Trans. ASSP*, 33:420–431, 1985.

[29] P. Wolfe. A method of conjugate subgradients for minimizing nondifferentiable funtions. *Mathematical Programming Study*, 3:145–173, 1975. M. Balinski and P. Wolfe, Eds.

TABLE 6.1
Parameters of the regularized bundle algorithm

n	2
p	2
q	1
ϵ_s	10^{-5}
m_R	.3
m_α	.15
κ	.8

TABLE 6.2
Numerical results for the three regularizations

	Ψ_1			Ψ_2			Ψ_3		
λ	it	f^*	nb	it	f^*	nb	it	f^*	nb
.05	155	0.0001	278	183	0.0000	353	203	0.0000	347
.15	198	0.0000	314	96	0.0002	180	90	0.0000	164
.20	220	0.0005	360	73	0.0000	137	121	0.0000	227
.25	182	0.0002	336	102	0.0000	194	145	0.0000	280
.30	138	0.0002	257	90	0.0000	172	79	0.0001	149
.35	79	0.0003	141	124	0.0000	245	88	0.0001	166
.40	136	0.0002	208	136	0.0000	267	118	0.0000	229
.45	148	0.0014	287	112	0.0000	219	78	0.0000	148
.50	74	0.0004	139	129	0.0000	254	147	0.0000	288
.55	83	0.0000	154	115	0.0000	222	84	0.0001	161
.60	99	0.0008	186	65	0.0000	127	75	0.0000	147
.65	123	0.0003	267	96	0.0000	186	96	0.0000	186
.70	165	0.0005	269	84	0.0000	163	66	0.0000	127
.75	106	0.0001	203	102	0.0000	199	58	0.0000	108
.80	82	0.0000	158	74	0.0000	145	101	0.0000	210
.85	95	0.0003	179	66	0.0001	129	67	0.0001	129

FIG. 6.1. *The objective f and its derivative f'.*

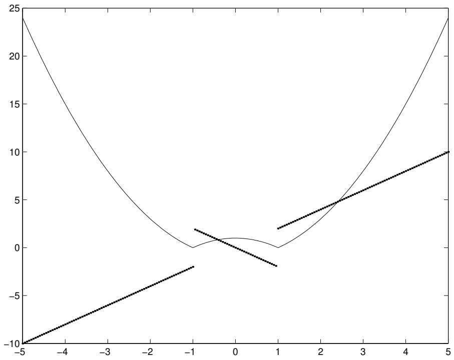

FIG. 6.2. *Moreau-Yosida regularization* Ψ_1

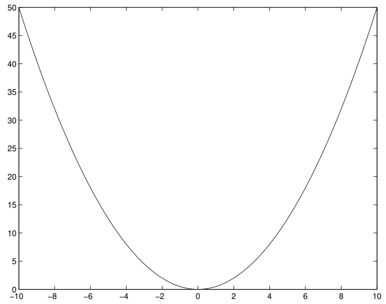

FIG. 6.3. l_1 *norm regularization* Ψ_2

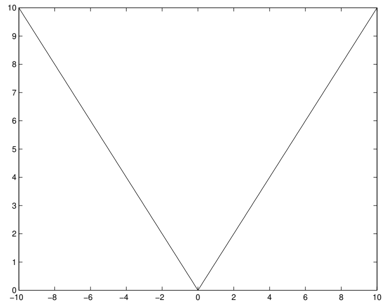

FIG. 6.4. $log(1 + |.|)$ *regularization* Ψ_3