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 simple and implementable bundle methods for the convex and nonconvex cases. 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, convex 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 optimization problems is the bundle method first proposed by Lemarechal [14] and Wolfe [31] for convex minimization and further developed by Mifflin [19, 20] and Kiwiel [10, 11, 13] for the nonconvex case; see also [2], [17], [26] 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

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

is a piecewise linear approximation called 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 [10, 11, 13],

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

where

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

*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 [18, 25], the *proximal point* algorithm is defined by the reccurence

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

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 [16].

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 a generalization of the well known Expectation Maximization (EM) algorithm for maximum likelihood estimation [5]. Consider as in [5] the sample spaces Ω_1 and Ω_2 on which one defined 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

$$I(x,y \mid v_2) = \int_{\{h^{-1}(v_2)\}} \log\left(\frac{p(v_1 \mid v_2;y)}{p(v_1 \mid v_2;x)}\right)\ p(v_1 \mid v_2;y)dv_1 \tag{1.4}$$

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

$$x^{k+1} = \mathrm{argmin}_{y\in\mathbb{R}^n}\{-\log p_2(v_2;y) + \lambda_k I(x^k,y \mid v_2)\}. \tag{1.5}$$

Using the facts that

$$p(v_1 \mid v_2;y) = \frac{p(v_1;y)}{p(v_2;y)} \tag{1.6}$$

and

$$\int p(v_1 \mid v_2;y)dv_1 = 1 \tag{1.7}$$

it can be shown that (1.5) takes the equivalent form

$$x^{k+1} = \mathrm{argmax}_{y\in\mathbb{R}^n}\Big\{-(1-\lambda_k)\log p_2(v_2;y) \tag{1.8}$$

$$1cm + \lambda_k \mathsf{E}[\log p_1(v_1;y) \mid V_2 = v_2; x^k]\Big\} \tag{1.9}$$

where $\mathsf{E}[\log p_1(v_1; y) \mid V_2 = v_2; x^k] = \int \log p_1(v_1; y) p(v_1 \mid v_2; x^k) dv_1$ denotes conditional expectation. When $\lambda_k = 1$, recursion (1.8) is identical to the EM algorithm introduced in [5] where V_2 is the incomplete data and V_1 is the complete data. Furthermore, as implied by Lemma 4.1 below the recursion (1.8) monotonically increases the log-likelihood $\log p_2(v_2; y)$ as does the standard EM algorithm of [5]. A special case of (1.5) is the case of Laplacian data,

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

When the complete data V_1 is also chosen as Laplacian, it is easy to show that the Kullback regularization given by (1.4) is nonsmooth and nonconvex.

Example 2 (Methods of multipliers). In [24] 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 [28, 29, 8] have demonstrated the benefit of using nonquadratic regularization functions. Among the possible choices for regularization functions proposed in [28] is the $\phi-$divergence

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

where, in [28], ϕ was assumed strictly convex. In particular, consider the convex program

$$\min_{x \in \mathbb{R}^n} f(x) \qquad \text{subject to} \qquad g_i(x) \leq 0, \quad i \in \{1, \ldots, m\} \tag{1.10}$$

where f and $g_1, \ldots, g_m$ are convex functions. The proximal point algorithm applied to the dual takes the form (see [8, section 6])

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

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 (1.10) (for instance, see [8]). Using different choices for the function ϕ, some well known multiplier methods can be recovered. Our generalization of ϕ to nonconvex functions opens up many new possibilities.

Further examples of nonsmooth and nonconvex regularizations have also recently been studied in the context of inverse problems in [21] and [22].

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. In Section 3, the fixed points of the method are studied. In particular an analysis of nondifferentiable nonconvex regularization is provided which seems to have no precedent in the literature. In Section 4, global convergence of the method is established. We then demonstrate local convergence when f is strictly convex in an open neighborhood of an accumulation point of the minimizing sequence. In sections 5 and 6, we turn to implementable bundle methods for approximation of the generalized proximal point iterations. The case of convex function is discussed first in Section 5 for the sake of clarity in the exposition. Then, algorithmic refinements, including a linesearch, are introduced in section 6 in order to accomodate the nonconvex case.

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 conv(S). The function f: $\mathbb{R}^n \to R$ is locally Lipschitz 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

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

The subdifferential of f at x is the set

$$\partial f(x) = \mathrm{conv}\{\lim_{x^i \to x} \nabla f(x^i) \mid \nabla f(x^i) \text{ exists}\},$$

Where ∇f denotes the gradient of f. The subdifferential has the property that it is a closed and convex set. An equivalent definition of $\partial f(x)$ which will be useful is

$$\partial f(x) = \{g \in \mathbb{R}^n \mid \langle g, d \rangle \leq f'(x, d) \ \forall d \in \mathbb{R}^n\}. \tag{1.11}$$

The multivalued function $x \mapsto \partial f(x)$ 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 (1.11) of the subdifferential, $0 \in \partial f(x)$. More generally, we say that x is a stationary point for f if $0 \in \partial f(x)$. For convex functions $f(x)$ the subdifferential is equivalently defined by

$$\partial f(x) = \{g \in \mathbb{R}^n \mid \langle g, y - x \rangle \leq f(y) - f(x) \ \forall y \in \mathbb{R}^n\}. \tag{1.12}$$

An extention of the subdifferential, called the ϵ-subdifferential, can be defined for convex function as follows

$$\partial_\epsilon f(x) = \{g \in \mathbb{R}^n \mid f(y) \geq f(x) + \langle g, y - x \rangle - \epsilon \ \forall y \in \mathbb{R}^n\}. \tag{1.13}$$

In the nonconvex case, we will need to introduce another type of approximate subdifferential called the Goldstein ϵ-subdifferential. For any $x \in \mathbb{R}^n$ and any $\epsilon \geq 0$ the Goldstein ϵ-subdifferential [10] of f at x is the set

$$\partial^\epsilon f(x) = \mathrm{conv}\{\partial f(y) \mid |y - x| \leq \epsilon\}.$$

The multivalued function $(x, \epsilon) \mapsto \partial^\epsilon f(x)$ is locally bounded and upper semicontinuous, i.e., $x^k \to x$, $\epsilon^k \to \epsilon$, $p^k \in \partial_{\epsilon^k} f(x^k)$ and $p^k \to p$ imply $p \in \partial_\epsilon f(x)$. We will also need the notion of *weak upper semismoothness* introduced by Mifflin [19]. A function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be weakly upper semismooth at $x \in \mathbb{R}^n$ if

a. f is Lipschitz on a ball about x and

b. for each $d \in \mathbb{R}^n$ and for any sequences $\{t^k\} \subset \mathbb{R}_+$ and $\{g^k\} \subset \mathbb{R}^n$ such that $\{t^k\} \downarrow 0$ and $g^k \in \partial f(x + t^k d)$ it follows that

$$\liminf_{k \to \infty} \langle g^k, d \rangle \geq \limsup_{t \downarrow 0} \frac{f(x + td) - f(x)}{t}.$$

Finally, for any function $F(x, x')$ satisfying the local Lipschitz property in x', we will use the notation $\partial F(x, x')$ 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 $\alpha \in \mathbb{R}$
(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 $x \in \mathbb{R}^n$. A usefull property of the regularizing function Ψ is now given

LEMMA 2.1. *Assume that $\Psi(x, y)$ satisfies Assumptions 2. Then,*

$$\partial\Psi(x, x) = \{0\},$$

for all x in $\mathbb{R}^n$.

Proof. Fix x in $\mathbb{R}^n$. Due to Assumptions 2, $\Psi(x, y) > \Psi(x, x)$ for all y in $\mathbb{R}^n$. Therefore, the first order optimality condition gives

$$0 \in \partial\Psi(x, x),$$

as desired. □

We now define the *generalized proximal point* algorithm

DEFINITION 2.2. *Assume that f and $\Psi(.,.)$ satisfy assumptions 1 and 2 respectively. Then, the* generalized proximal point *algorithm is defined by the following recursion starting at x^1*

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

The following result proves that recursion (2.1) definition is well defined.

PROPOSITION 2.3. *The set of minimizers of (2.1) is nonempty for any x^k in $\mathbb{R}^n$.*

Proof. This result is a straightforward consequence of Assumptions 1 and 2. Indeed, the function $f(x)+\lambda_k\Psi(x', x)$ inherits the inf-compactness and local Liptschitz properties of f and Ψ and thus possesses at least one bounded minimizer. □

Notice that, due to nonconvexity of the functions involved, the minimum in (2.1) may not be unique. In such cases, x^{k+1} in (2.1) can be arbitrarily chosen among the set of minimizers of (2.1).

For the sake of notational convenience, in the remainder of this paper, the regularized objective function with relaxation parameter t will be denoted

$$F_t(x, y) = f(y) + t\Psi(x, y) \tag{2.2}$$

and the *generalized proximal* operator will be defined by

$$P_t(x) = \mathrm{argmin}_{y \in \mathbb{R}^n} F_t(x, y).$$

3. Optimality for nonsmooth regularization. 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.

3.1. Fixed points and nonsmooth regularization. Consider the regularized function (2.2) where, with no loss of generality, λ is set to 1, i.e.

$$F(x,y) \triangleq F_1(x,y) = f(y) + \Psi(x,y),$$

and the associated proximal operator

$$P(x) = \mathrm{argmin}_{y\in\mathbb{R}^n} F(x,y).$$

The success of the proximal point approach relies on the condition that fixed points of P also be stationary points of f. In the case where Ψ is quadratic, say $\Psi(x,y) = \frac{1}{2}|x-y|^2$, this property is well known and straightforward. In the case of the general possibly nonconvex regularization considered in this paper, we only have

$$x^* = \mathrm{argmin}_{y\in\mathbb{R}^n} F(x^*,y),$$

which is equivalent to

$$0 \in \partial F(x^*,x^*). \tag{3.1}$$

Using the calculus of subdifferentials, we have

$$\partial F(x,y) \subset \partial f(y) + \partial\Psi(x,y),$$

and thus, (3.1) implies

$$0 \in \partial F(x^*,x^*) \subset \partial f(x^*) + \partial\Psi(x^*,x^*) \tag{3.2}$$

Hence, consistence of the generalized proximal point approach reduces to the question of knowing in which circumstances a point x^* satisfying (3.2) is a stationary point of f.

In the case where $\Psi(x^*,y)$ is smooth in the second variable, e.g. $\Psi(x^*,y) = |x^*-y|^2$, then $\partial\Psi(x^*,x^*)$ reduces to $\{0\}$, and we deduce from (3.2) that $0\in\partial f(x^*)$ which proves that x^* is 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 one dimensional 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 $\partial f(x^*) = [-1,-\frac{1}{2}]$ and $\partial\Psi(x^*,x^*) = [-1,1]$. On the other hand, $\partial F(x^*,x^*) = [-2,\frac{1}{2}]$. Thus, in this case, we have $0\in\partial F(x^*,x^*)$ whereas x^* is not a stationary point of f (indeed, f has no stationary point).

3.2. Subdifferential domination. The reason that condition (3.2) fails in the latter 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" variation than the regularization. The following definition makes precise the idea of "greater" variation.

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*

$$|f_1'(x,d)| > |f_2'(x,d)| \quad \forall d \in \mathbb{R}^n. \tag{3.3}$$

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 $0 \leq t \leq 1$ at x. The following lemma establishes that the conditions given by (3.2) specify a stationary point of the objective function under the subdifferential domination hypothesis.

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

$$0 \in \partial f(x^*) + \partial \Psi(x^*, x^*), \tag{3.4}$$

with Ψ satisfying Assumptions 2. 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 $0 \in \partial \Psi_{x^*}(x^*)$, due to Lemma 2.1, we have $\Psi_{x^*}'(x^*,d) \geq 0$ for all $d \in \mathbb{R}^n$. Due to the first equation in (3.4), we deduce the existence of two vectors g_1 and g_2 such that

$$\begin{aligned}
&g_1 + g_2 = 0,\\
&g_1 \in \partial f(x^*),\\
&g_2 \in \partial \Psi_{x^*}(x^*),
\end{aligned}$$

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 $f'(x^*,d) + \Psi_{x^*}'(x^*,d) \geq 0$. Thus, we obtain

$$f'(x^*,d) \geq -\Psi_{x^*}'(x^*,d). \tag{3.5}$$

We now proceed by contradiction. Suppose that $f'(x^*,d) < 0$ for some $d \in \mathbb{R}^n$. In this case, as $\Psi_{x^*}'(x^*,d) \geq 0$, equation (3.5) implies

$$|f'(x^*,d)| \leq |\Psi_{x^*}'(x^*,d)|.$$

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 (1.11) of the subdifferential, we conclude that $0 \in \partial f(x^*)$. ☐

With this result in hand, we are now ready to discuss the case of the generalized proximal point operator. Using Lemma 3.2 we first deduce the optimality of the fixed points of the generalized proximal point mapping P_t with relaxation parameter $t \geq 0$ in the following straightforward lemma.

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 $f(x)$.*

A last question of computational importance remains to be discussed. In real life situations, one may not know whether a given regularization is subdifferentially dominated by the function to be minimized. This problem is easily overcome by forcing the relaxation parameter t towards zero in the generalized proximal point operator P_t. Indeed, the definition of subdifferential domination and Lemma 3.3 prescribe that

$$|f'(x^*,d)| > t|\Psi'_{x^*}(x^*,d)|, \quad \forall d \in \mathbb{R}^n,$$

where, as above, x^* is a fixed point of P_t and $\Psi'_{x^*}(x^*,d)$ is the directional derivative of $\Psi(x^*,.)$ in the direction of d at x^*. Therefore, one easily checks that, given f and Ψ, it is sufficient to take a small enough relaxation parameter t to guarantee (3.2). As a consequence, we may conclude that a safe strategy, when performing the generalized proximal point algorithm with a nonsmooth regularization, is to take a sequence of relaxation parameters $t = \lambda_k$ indexed by iteration k converging towards a sufficiently small value[1].

4. Convergence analysis.

In this section, we give an asymptotic analysis of the *generalized proximal point* algorithm which does not require differentiability nor convexity assumptions. A Lyapunov method is the guideline of the proof where the Lyapunov function is simply the objective $f(x)$. We show that the accumulation points of the sequence defined by (2.1) are locally optimal. Under a strict local convexity assumption convergence is established.

4.1. Main results.

We start with the following monotonicity result.

LEMMA 4.1. *Let f and Ψ satisfy Assumptions 1 and 2. For any iteration $k > 1$, the sequence $(x^k)_{k\in\mathbb{N}}$ satisfies*

$$f(x^{k+1}) - f(x^k) \leq -\lambda_k \Psi(x^k, x^{k+1}) \leq 0. \tag{4.1}$$

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 (4.1) follows. □

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

LEMMA 4.2. *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 4.1, $(f(x^k))_{k\in\mathbb{N}}$ 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 (4.1) 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

[1] We will require nevertheless that $(\lambda_k)_{k\in\mathbb{N}}$ does not vanish to push through our convergence analysis

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 4.3. *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 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 4.2 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 $F_t(x^*, x^*) \leq F_t(x^*, x)$ for all $x \in \mathbb{R}^n$. In the following, we prove that this result holds with $t = \lambda^+$. By definition of $(x^k)_{k\in\mathbb{N}}$, $F_{\lambda_k}(x^{\sigma(k)}, x^{\sigma(k)+1}) \leq F_{\lambda_k}(x^{\sigma(k)}, x)$ for all $x \in \mathbb{R}^n$. Therefore, for all x

$$\begin{aligned} &F_{\lambda^+}(x^{\sigma(k)}, x^{\sigma(k)+1}) + (\lambda_k - \lambda^+)\Psi(x^{\sigma(k)}, x^{\sigma(k)+1}) \\ &\qquad \leq F_{\lambda^+}(x^{\sigma(k)}, x) + (\lambda_k - \lambda^+)\Psi(x^{\sigma(k)}, x). \end{aligned} \tag{4.2}$$

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, F_{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 $F_{\lambda^+}(.,.)$, respectively, we have for any $\epsilon > 0$ there exists $K_2 \in \mathbb{N}$ such that for all $k \geq K_2$

$$F_{\lambda^+}(x^*, x) \geq F_{\lambda^+}(x^{\sigma(k)}, x) - \epsilon, \tag{4.3}$$

and

$$F_{\lambda^+}(x^*, x^*) \leq F_{\lambda^+}(x^{\sigma(k)}, x^{\sigma(k)+1}) + 2\epsilon. \tag{4.4}$$

Combining equations (4.3) and (4.4) with (4.1), we obtain

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

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

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

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

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

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 $\mathrm{argmin}_{x\in\mathbb{R}^n} F_{\lambda^+}(x^k, x)$. 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 $f(x)$. ☐

4.2. Convergence to local minima 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 4.4. *Let f and Ψ satisfy Assumptions 1 and 2. Then, for a given starting point x^1, the set x^* of accumulation points of the sequence $(x^k)_{k\in\mathbb{N}}$ 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 [23, Theorem 28.1]. ☐

COROLLARY 4.5. *Suppose, in addition to the assumptions of Theorem 4.3, that $f(x)$ is strictly convex in an open neighborhood $\mathcal{N}$ of an accumulation point x^* of $(x^k)_{k\in\mathbb{N}}$. Then the sequence $(x^k)_{k\in\mathbb{N}}$ converges to a local minimizer of $f(x)$.*

Proof. We obtained in Theorem 4.3 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 in $\mathbb{N}$ 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 $\mathcal{N}$. Then, 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. ☐

5. Bundle implementations: the convex case. The study of the generalized proximal point algorithm gives an elegant framework for the exploration of a large class of regularizations. In the two next sections, we use the bundle framework introduced by Lemaréchal [7] to make this approach tractable in applications. The present section introduces the main ideas governing incorporation of nonquadratic regularization functions into the bundle mechanism in the case where the functions are convex. More precisely we will require that $f(x)$ be convex and $\Psi(x',x)$ be convex with respect to x for every x' in $\mathbb{R}^n$. The case where the functions f and Ψ are only required to be locally Lipschitz will be discussed in the next section.

5.1. Background. Bundle methods have been widely recognized as a very efficient technique for minimization of nondifferentiable functions. They can be interpreted as implementations of stepwise approximations of the proximal point algorithm [16, 12]. However, to our knowledge, bundle implementations of general nonquadratic regularizations have not been considered in the litterature. This may be due to be fact that bundle methods are essentially sequential quadratic programs, which seems to exclude the possibility of more general regularization functions.

The main idea behind application of bundle methods to nonquadratic proximal algorithms is the following. The generalized proximal iteration is approximated using subgradient information with respect to f and Ψ about the current iterate x^k. Let $\{y^j\}_{J^k}$ be a set of points in a neighborhood of x^k, indexed by $j\in J^k$. For any point $y\in\mathbb{R}^n$, let $g(y)$ (resp. $h_k(y)$) denote an arbitrary subgradient in $\partial f(y)$ (resp. $\partial\Psi(x^k,y)$). Hence, any point y^j, $j\in J^k$ allows to define approximate models

$$\hat{f}_j(x)=f(y^j)+\langle g(y^j),x-y^j\rangle$$

and

$$\hat{\Psi}_j(x^k, x) = \Psi(x^k, y^j) + \langle h_k(y^j), x - y^j \rangle$$

of f and Ψ respectively. To stabilize the model $\hat{\Psi}_j(x^k, x)$ a quadratic term $\frac{1}{2}|x - x^k|^2$ is added and the following approximation is obtained for the function $F_{\lambda_k}(x^k, x)$ in the proximal point algorithm

$$\begin{aligned}\hat{F}_{\lambda_k,j}(x^k, x) =& f(y^j) + \langle g(y^j), x - y^j \rangle + \lambda_k \big(\Psi(x^k, y^j) + \langle h_k(y^j), x - y^j \rangle \\ &+ \frac{1}{2}|x - x^k|^2\big).\end{aligned}$$

Note that convex functions are bounded from below by their local first order approximations and that the quadratic stabilization term is independent from y^j. Hence, the best approximation of $F_{\lambda_k}(x^k, x)$ can be obtained by gathering the information at each y^j in the neighborhood of x^k via the following max-function,

$$\begin{aligned}\hat{F}_{\lambda_k}(x^k, x) =& \max_{j \in J^k}\{f(y^j) + \langle g(y^j), x - y^j \rangle + \lambda_k \big(\Psi(x^k, y^j) + \langle h_k(y^j), x - y^j \rangle \\ &+ \frac{1}{2}|x - x^k|^2\},\end{aligned}$$

Equivalently, $\hat{F}_{\lambda_k}(x^k, x)$ can be written

$$\hat{F}_{\lambda_k}(x^k, x) = \max_{j \in J^k}\{f(x^k) - \alpha_j^k + \langle s^k(y^j), x - x^k \rangle\} + \frac{1}{2}\lambda_k |x - x^k|^2,$$

where α_j^k is the accumulated linearization error due to the fact that the subgradients are computed at the points y^j instead of x^k, i.e.

$$\alpha_j^k = f(x^k) - f(y^j) - \langle g(y^j), x^k - y^j \rangle - \Psi(x^k, y^j) - \langle h_k(y^j), x^k - y^j \rangle, \quad (5.1)$$

and $s(y^j)$ is the accumulated subgradient

$$s^k(y^j) = g(y^j) + \lambda_k h_k(y^j).$$

With this model in hand, we are now ready to introduce the bundle mechanism. First, an approximate proximal step is taken which defines a candidate for the next iterate,

$$y^{k+1} = \mathrm{argmin}_{x \in \mathbb{R}^n} \hat{F}_{\lambda_k}(x^k, x), \quad (5.2)$$

or, equivalently,

$$y^{k+1} = \mathrm{argmin}_{x \in \mathbb{R}^n} \max_{j \in J^k}\{f(x^k) - \alpha_j^k + \langle s(y^j), x - x^k \rangle\} + \frac{\lambda_k}{2}|x - x^k|^2\}.$$

The bundle approach is a strategy in order to decide wether the accumulated subgradient information at every point of the set $\{y^j\}$, $j \in J^k$, is sufficiently accurate so that a reliable proximal step can be achieved, i.e. $x^{k+1} = y^{k+1}$. It is well known that iteration (5.2) may even not provide a descent step with regard to f if the subgradient information is innaccurate Hence, a reasonable selection of candidates $\{y^{k+1}\}$ for a proximal step should be based on a test of descent in the objective function f. In order to implement such a test, a parameter δ_k is computed, representing the expected decrease given the model $\hat{F}_{\lambda_k}$,

$$\delta_k = f(x^k) - \hat{F}_{\lambda_k}(y^{k+1}, x^k),$$

or equivalently

$$\delta_k = f(x^k) - \max_{j \in J^k}\{f(x^k) - \alpha_k^j + \langle s^k(y^j), y^{k+1} - x^k\rangle\} - \frac{\lambda_k}{2}|y^{k+1} - x^k|^2.$$

The decrease obtained at y^{k+1} is then compared to a fraction $m \in (0,1)$ of the expected decrease δ_k, following the rule

- if $f(x^k) - f(y^{k+1}) \geq m\delta_k$, then a descent step is taken, i.e. $x^{k+1} = y^{k+1}$,
- otherwise $x^{k+1} = x^k$.

In either case, the subgradient information at y^{k+1} is collected and is incorporated to the polyhedral approximation of f and Ψ at iteration $k+1$. In this manner the accuracy of the approximation $\hat{F}(x^k, x)$ to $f(x) + \lambda_k \Psi(x^k, x)$ improves at each iteration.

5.2. A generalized proximal bundle method. In this section, we present the details of our bundle implementation.

ALGORITHM 1 (Generalized Proximal Bundle Method for Convex Functions).

Step 0 (Initialization) *Choose a final accuracy parameter* $\delta_s > 0$ *and a parameter* $m \in (0,1)$. *Choose the starting point* $x^1 \in \mathbb{R}^n$. *Set* $k = 1$, $y^1 = x^1$, $J^1 = \{1\}$, $g^1 = g(y^1)$, $h^1 = h_1(y^1)$, $s^1 = g^1 + h_1^1$ *and* $\alpha_1^1 = 0$.

Step 1 (Proximal step) *Compute*

$$y^{k+1} = \mathrm{argmin}_{x \in \mathbb{R}^n} \max_{j \in J^k} \{f(x^k) - \alpha_k^j + \langle s^j, x - x^k\rangle\} + \frac{\lambda_k}{2}|x - x^k|^2\}. \qquad (5.3)$$

Step 2 (Descent test) *Set*

$$\delta_k = f(x^k) - \max_{j \in J^k}\{f(x^k) - \alpha_k^j + \langle s^j, y^{k+1} - x^k\rangle\} - \frac{\lambda_k}{2}|y^{k+1} - x^k|^2.$$

- *if* $f(x^k) - f(y^{k+1}) \geq m\delta_k$, *then set* $x^{k+1} = y^{k+1}$ (**descent step**),
- *otherwise set* $x^{k+1} = x^k$ (**null step**).

Step 3 (Stopping criterion) *If* $\delta_k < \delta_s$, *then stop.*

Step 4 (Variable updating) *Set* $J^{k+1} = J^k \cup \{k+1\}$, *choose* λ_{k+1}, *set* $g^{k+1} = g(y^{k+1})$, $h^{k+1} = h_{k+1}(y^{k+1})$, $s^{k+1} = g^{k+1} + \lambda_{k+1}h^{k+1}$, *set* $\alpha_{k+1}^j = \alpha_k^j$, *for* $j \in \{1, \dots, k\}$, $\alpha_{k+1}^{k+1} = f(x^{k+1}) - \langle s^{k+1}, x^{k+1} - y^{k+1}\rangle$. *Finally, increase k by 1 and go to Step 1.*

We now establish some preliminary results concerning this algorithm. The following lemmas are standard in the analysis of bundle methods. In particular, the following lemma expresses proximal step (5.3) in the form of a dual quadratic program:

LEMMA 5.1. [3] *Consider the constrained minimization problem*

$$\bar{u} = \mathrm{argmin}_{u \in \mathbb{R}^k}\Big\{\frac{1}{2}|\sum_{j \in J^k} u_j s^j|^2 + \lambda_k \sum_{j \in J^k} u_j \alpha_k^j\Big\} \qquad (5.4)$$

subject to (5.5)

$$u \in \Delta_k = \{z \in (0,1)^k \mid \sum_{j \in J^k} z_j = 1\}.$$

Then,

$$\text{(i)} \qquad y^{k+1} = x^k - \frac{1}{\lambda_k} \sum_{j \in J^k} \bar{u}_j s^j, \tag{5.6}$$

where y^{k+1} solves the proximal step (5.3) in Step 1 of Algorithm 1, and the polyhedral component $\Phi_k(x) = \max_{j \in J^k}\{f(x^k) - \alpha_k^j + \langle s^j, x - y^j\rangle\}$ of $\hat{F}_{\lambda_k}(x^k, x)$ satisfies

$$\text{(ii)} \qquad \Phi_k(y^{k+1}) = f(x^k) - \frac{1}{\lambda_k}\Big| \sum_{j \in J^k} u_j s^j \Big|^2 - \sum_{j \in J^k} \bar{u}_j \alpha_k^j,$$

.

In the sequel, we adopt the notation

$$p^k = \sum_{j \in J^k} \bar{u}_j s^j. \tag{5.7}$$

We must also introduce another dual variable which plays an important role in the bundle method. For this purpose, notice that the quadratic program given by (5.4) is equivalent to

$$\bar{u} = \text{argmin}_{u \in \mathbb{R}^k} \frac{1}{2}\Big| \sum_{j \in J^k} u_j s^j \Big|^2 \tag{5.8}$$

$$\text{subject to} \tag{5.9}$$

$$u \in \Delta_k$$

$$\sum_{j \in J_k} u_j \alpha_j \leq \epsilon_k, \tag{5.10}$$

where λ_k in (5.4) is identified with the Lagrange multiplier associated with the contraint (5.10). Using lemma 5.1 (ii), and the definition of δ_k in step 2 of Algorithm 1

$$f(x^k) - \Phi(y^{k+1}) - \frac{\lambda_k}{2}|y^{k+1} - x^k|^2 = \frac{1}{2\lambda_k}\Big| \sum_{j \in J^k} \bar{u}_j s^j \Big|^2 + \sum_{j \in J^k} \bar{u} \alpha_k^j. \tag{5.11}$$

Recalling that $\bar{u}$ is solution to the quadratic program (5.8), we have

$$\epsilon_k = \sum_{j \in J^k} \bar{u}_j \alpha_k^j. \tag{5.12}$$

Therefore (5.11) is equivalent to (see also [3])

$$\delta_k = \epsilon_k + \frac{1}{2\lambda_k}|p^k|^2. \tag{5.13}$$

With this result in hand, we obtain the following lemma.

LEMMA 5.2. *Let p^k be the direction defined by (5.7). Then,*

$$p^k \in \partial_{\epsilon_k}\big(f(x^k) + \lambda_k \Psi(x^k, x^k)\big).$$

Proof. The polyhedral function $\Phi_k(x)$ may be written

$$\Phi_k(x) = \max_{j\in J^k}\{f(y^j) + \langle g^j, x - y^j\rangle + \lambda_k\big(\Psi(x^k, y^j) + \langle h^j, x - y^j\rangle\big)\}.$$

Then, due to convexity, we have

$$f(x) \geq f(y^j) + \langle g^j, x - y^j\rangle$$

and

$$\Psi(x^k, x) \geq \Psi(x^k, x^k) + \langle h^j, x - y^j\rangle.$$

Therefore,

$$f(x) + \lambda_k\Psi(x^k, x) \geq \Phi_k(x). \tag{5.14}$$

On the other hand, optimality in the proximal step (5.3) (Step 1 of Algorithm 1), gives

$$0 \in \partial\Phi_k(y^{k+1}) + \lambda_k(y^{k+1} - x^k).$$

Noticing that $\lambda_k(y^{k+1} - x^k) = p^k$, we thus obtain that

$$p^k = \partial\Phi_k(y^{k+1}).$$

Hence, using the subgradient inequality, we obtain

$$\Phi_k(x) \geq \Phi_k(y^{k+1}) + \langle p^k, x - y^{k+1}\rangle. \tag{5.15}$$

Recalling that, due to Lemma 5.1 (ii)

$$\Phi_k(y^{k+1}) = f(x^k) - \frac{1}{\lambda_k}|p^k|^2 - \sum_{j\in J_k}\bar{u}_j\alpha_k^j,$$

and combining this result with (5.15), (5.14) becomes

$$f(x) + \lambda_k\Psi(x^k, x) \geq f(x^k) - \frac{1}{\lambda_k}|p^k|^2 - \sum_{j\in J^k}\bar{u}_j\alpha_k^j + \langle p^k, x - y^{k+1}\rangle.$$

Since $-\frac{1}{\lambda_k}|p^k|^2 = \langle p^k, y^{k+1} - x^k\rangle$, we obtain

$$f(x) + \lambda_k\Psi(x^k, x) \geq f(x^k) + \langle p^k, x - x^k\rangle - \sum_{j\in J^k}\bar{u}_j\alpha_k^j.$$

Furthermore, recalling that

$$\epsilon_k = \sum_{j\in J^k}\bar{u}_j\alpha_k^j,$$

and that $\Psi(x^k, x^k) = 0$, we easily obtain

$$f(x) + \lambda_k\Psi(x^k, x) \geq f(x^k) + \lambda_k\Psi(x^k, x^k) + \langle p^k, x - x^k\rangle - \epsilon_k,$$

which is equivalent to

$$p^k \in \partial_{\epsilon_k}\big(f(x^k) + \lambda_k\Psi(x^k, x^k)\big).$$

□

This last result shows in particular the well known fact that bundle methods may be interpreted as $\epsilon-$subgradient methods (see [7]). The main feature of bundle methods is therefore the control of the parameter ϵ_k via (5.13) and the descent step/null step strategy using the expected decrease δ_k. We now discuss convergence of this method to a minimizer of f.

5.3. Convergence. For the convex case, the proof of convergence of the generalized proximal bundle method is similar to the proof of convergence of the standard form of bundle methods. In the sequel, K denotes the set of indices k where a descent step is taken. We start with the following lemma.

LEMMA 5.3. *Consider the following convergent sequences,* $x^k \to \bar{x}$, $y^k \to \bar{y}$, $\epsilon_k \to \bar{\epsilon}$, $\lambda^k \to \bar{\lambda}$ *and* $p^k \to \bar{d}$. *Assume that*

$$p^k \in \partial_{\epsilon_k}\big(f(y^k) + \lambda_k \Psi(x^k, y^k)\big)$$

for all $k \in \mathbb{N}$. *Then,* $\bar{d} \in \partial_{\bar{\epsilon}}\big(f(\bar{y}) + \bar{\lambda}\Psi(\bar{x}, \bar{y})\big)$.

Proof. Fix y in $\mathbb{R}^n$. The fact that $p^k \in \partial_{\epsilon_k}\big(f(y^k) + \lambda_k \Psi(x^k, y^k)\big)$ implies that for fixed y

$$f(y) + \lambda_k \Psi(x^k, y) \geq f(y^k) + \lambda_k \Psi(x^k, y^k) + \langle p^k, y - y^k \rangle - \epsilon_k.$$

Therefore,

$$\begin{aligned} f(y) + \lambda_k \Psi(x^k, y) \geq & f(y^k) + \lambda_k \Psi(x^k, \bar{y}) + \lambda_k \big(\Psi(x^k, y^k) - \Psi(x^k, \bar{y})\big) \\ & + \langle p^k, y - y^k \rangle - \epsilon_k. \end{aligned} \tag{5.16}$$

Now, since Ψ^k is assumed locally Lipschitz over $\mathbb{R}^n \times \mathbb{R}^n$, and since $(x^k)_{k\in\mathbb{N}}$ and $(y^k)_{k\in\mathbb{N}}$ are bounded due to convergence, there exists a constant C such that

$$|\Psi(x^k, y^k) - \Psi(x^k, \bar{y})| \leq C\sqrt{|x^k - x^k|^2 + |y^k - \bar{y}|^2} = C|y^k - \bar{y}|.$$

Therefore, as y^k converges to $\bar{y}$

$$\lim_{k\to\infty} \Psi(x^k, y^k) - \Psi(x^k, \bar{y}) = 0.$$

Using a similar argument one easily deduces that $\Psi(x^k, y)$ converges to $\Psi(\bar{x}, y)$ and that $\Psi(x^k, \bar{y})$ converges to $\Psi(\bar{x}, \bar{y})$. Furthermore, Assumption 1 (ii) implies (lower semi-) continuity of f and thus, passing to the limit in (5.16) gives

$$f(y) + \bar{\lambda}\Psi(\bar{x}, y) \geq f(\bar{y}) + \bar{\lambda}\Psi(\bar{x}, \bar{y}) + \langle \bar{d}, y - \bar{y} \rangle - \bar{\epsilon},$$

which proves the desired result. ∎

We will also need a more technical result the proof of which can be found in [3].

LEMMA 5.4. *Let f and Ψ satisfy assumptions 1 and 2. Let K be the set of indices where a descent step is taken. Let* $f_* = \lim_{k\to\infty, k\in K} f(x^k)$. *Then*

$$\sum_{k\in K} \delta_k \leq \frac{f(x^1) - f_*}{m}.$$

The following convergence analysis is divided into two parts. In the first part, we consider the case where an infinite number of descent steps are taken. Then, we will turn to the finite case.

LEMMA 5.5. *Assume that f and Ψ satisfy Assumptions 1 and 2. Assume that f subdifferentially dominates $\lambda^+\Psi$ at every point x^* such that* $0 \in \partial f(x^*) + \lambda^+ \partial\Psi(x^*, x^*)$ *for some real number* $\lambda^+ > 0$. *Assume in addition that* $\limsup_{k\to\infty} \lambda_k \leq \lambda^+$ *and that* $\liminf_{k\to\infty} \lambda_k \geq \lambda^- > 0$ *for some real number λ^- and for all k in $\mathbb{N}$. Finally, assume* *that an infinite number of descent steps is taken. Then, every accumulation point of $(x^k)_{k\in\mathbb{N}}$ is a minimizer of f.*

Proof. Using (5.13), we obtain

$$\frac{1}{2\lambda_k}|p^k|^2 = \delta_k - \epsilon_k \geq \delta_k,$$

for all $k \in K$. Hence, we have

$$\sum_{k\in K} |p^k|^2 \leq 2\lambda_k \sum_{k\in K} \delta_k. \tag{5.17}$$

Recall that Assumption 1 implies that f is bounded from below. Thus, Lemma 5.4 implies that $\lim_{k\to\infty, k\in K} \delta_k = 0$, and therefore, (5.13) implies that $\lim_{k\to\infty, k\in K} \epsilon_k = 0$. The assumptions on the sequence $(\lambda_k)_{k\in\mathbb{N}}$ imply boundedness of λ_k and thus, equation (5.17) gives

$$\sum_{k\in K} |p^k|^2 < +\infty,$$

which proves that $\lim_{k\to\infty, k\in K} |p^k| = 0$. On the other hand, inf-compactness of f and the fact that $(f(x^k))_{k\in K}$ is strictly decreasing imply that the sequence $(x^k)_{k\in K}$ is bounded. Now, take a convergent subsequence $(x^{\sigma(k)})_{k\in K}$ and let $\bar{x}$ be its limit. Since $(\lambda_k)_{k\in K}$ is bounded, a convergent subsequence can be taken, say $(\lambda_{\sigma(\gamma(k))})_{k\in K}$, tending towards $\bar{\lambda}$. Then, Lemma 5.2 and Lemma 5.3 together imply

$$0 \in \partial\big(f(\bar{x}) + \bar{\lambda}\Psi(\bar{x}, \bar{x})\big), \text{ and } 0 \in \partial f(\bar{x}) + \bar{\lambda}\partial\Psi(\bar{x}, \bar{x}).$$

Finally, since every accumulation point $\bar{\lambda}$ satisfies $\bar{\lambda} \leq \lambda^+$, subdifferential domination and lemma 3.2 imply that $0 \in \partial f(\bar{x})$. Finally, convexity implies that $\bar{x}$ is a minimizer of f is convex, which finishes the proof. ☐

We now discuss the case of a finite number of descent steps.

LEMMA 5.6. *Let the functions f and Ψ satisfy the assumptions in Lemma 5.5 Assume that only a finite number of descent steps is taken and let k_0 denote the index of the last descent step. Then, x^{k_0} is a minimizer of f.*

Proof. The proof is easily adapted from [15]. The definition of δ_k gives

$$f(x^k) - \delta^k = \max_{j\in J^k}\{f(y^j) + \lambda_k\Psi(x^k, y^j) + \langle s^j, y^{k+1} - y^j\rangle\} + \frac{\lambda_k}{2}|y^{k+1} - x^k|^2.$$

Take $k \geq k_0$. Since $x^k = x^{k_0}$ for all $k \geq k_0$, we obtain

$$f(y^j) + \lambda_k\Psi(x^{k_0}, y^j) + \langle s^j, y^{k+1} - y^j\rangle + \frac{\lambda_k}{2}|y^{k+1} - x^{k_0}|^2 \leq f(x^{k_0}) - \delta^k, \tag{5.18}$$

for all j in J^k. On the other hand, nondescent steps imply

$$f(x^k) - m\delta^j \leq f(y^j),$$

for $k \geq j \geq k_0$. Thus, as $f(x^k) = f(x^{k_0})$ for $k \geq k_0$, (5.18) gives

$$\lambda_k\Psi(x^{k_0}, y^j) + \langle s^j, y^{k+1} - y^j\rangle + \frac{\lambda_k}{2}|y^{k+1} - x^{k_0}|^2 \leq m\delta^j - \delta^k,$$

for $k \geq j \geq k_0$, which implies, due to positivity of both Ψ and the quadratic term, that

$$\langle s^j, y^{k+1} - y^j \rangle \leq m\delta^j - \delta^k, \tag{5.19}$$

for $k \geq j \geq k_0$. The case $j = k_0$ gives $y^j = y^{k_0} = x^{k_0}$, the last equality coming from the fact that k_0 is the index of the last descent step. Using this fact along with (5.18)

$$\langle s^{k_0}, y^{k+1} - x^{k_0} \rangle + \frac{\lambda_k}{2}|y^{k+1} - x^{k_0}|^2 \leq -\delta_k \leq 0.$$

The left term of this last equation is a quadratic form in the variable y^{k+1}, which cannot take negative values except on a bounded neighborhood of x^{k_0}. Hence, the sequence $(y^k)_{k\in\mathbb{N}}$ is bounded. As a consequence, s^j is also bounded, due to the local boundedness property of the subgradients of $f(.)$ and $\Psi(x^{k_0}, .)$ and the boundedness of λ_k. On the other hand, due to refinement of the polyhedral approximation, and since $\lambda_k = \lambda_{k_0}$ and $x^k = x^{k_0}$ for $k \geq k_0$,

$$\hat{F}_{\lambda_k}(x^k, x) \leq \hat{F}_{\lambda_{k+1}}(x^{k+1}, x),$$

for all x in $\mathbb{R}^n$, which proves that $(\delta_k)_{k\geq k_0}$ is decreasing and thus, has a limit. We now prove that this limit is zero. Indeed, take a convergent subsequence $(y^{\sigma(k)})_{k\geq k_0}$. Hence, for any j satisfying $\sigma(j) \geq k_0$, we have, using (5.19),

$$\langle s^{\sigma(j)}, y^{\sigma(j+1)} - y^{\sigma(j)} \rangle \leq m\delta^{\sigma(j)} - \delta^{\sigma(j+1)-1}.$$

Since $\lim_{j\to\infty} |y^{\sigma(j+1)} - y^{\sigma(j)}| = 0$ and due to boundedness of s^j, we have

$$\lim_{j\to\infty} m\delta^{\sigma(j)} - \delta^{\sigma(j+1)-1} \geq 0. \tag{5.20}$$

Recall that $(\delta^k)_{k\in\mathbb{N}}$ is convergent. Let $\bar{\delta}$ be its limit. By continuity, (5.20) gives

$$(m-1)\bar{\delta} \geq 0.$$

Since $0 < m < 1$, and $\bar{\delta} \geq 0$ by positivity of δ^k for all $k \in \mathbb{N}$, we have $(m-1)\bar{\delta} = 0$ which implies $\bar{\delta} = 0$, as desired. Now recall that

$$\delta^k = \epsilon_k + \frac{1}{2\lambda_k}|p^k|^2,$$

to conclude, using lower and upper boundedness of λ_k, that $\lim_{k\to\infty} \epsilon_k = 0$ and $\lim_{k\to\infty} |p^k| = 0$. As in lemma 5.5, we therefore obtain that

$$0 \in \partial\big(f(x^{k_0}) + \bar{\lambda}\Psi(x^{k_0}, x^{k_0})\big)$$

for every accumulation point $\bar{\lambda}$ of $(\lambda_k)_{k\in\mathbb{N}}$. Since every accumulation point $\bar{\lambda}$ satisfies $\bar{\lambda} \leq \lambda^+$, subdifferential domination gives

$$0 \in \partial f(x^{k_0}),$$

which proves optimality of x^{k_0} since f is convex. □

Combining these two lemmas yields the following convergence theorem for our bundle implementation of the generalized proximal point algorithm for convex optimization.

THEOREM 5.7. *Let the functions f and Ψ satisfy the assumptions in Lemma 5.5 Then, every accumulation point of $(x^k)_{k\in\mathbb{N}}$ defined by Algorithm 1 is a minimizer of f.*

6. Bundle implementations: the nonconvex case.

6.1. Preliminary comments. In this section, a bundle approach for implementing generalized proximal steps is developed for the case of nonconvex functions. The main ideas remain the same as in the convex situation. Nevertheless, several modifications need to be introduced in order to overcome the difficulties associated with nonconvexity. The algorithmic structure used in the sequel is similar to the one proposed by Kiwiel [10] and therefore inherits useful convergence properties. This allows us to concentrate on the particular problems induced by the use of our generalized regularization. We first introduce the main characteristics of the method.

One important property satisfied in the convex case by the polyhedral approximations to f and Ψ is that they are lower approximations, i.e. they lie below the original functions f and Ψ. Therefore α_k^j in (5.1) is always positive. This property no longer holds in the nonconvex case. Nevertheless, a similar property can be obtained when α_k^j is defined by

$$\alpha_k^j = |f(x^k) - f(y^j) - \langle g^j, x^k - y^j\rangle| + \lambda_k|\Psi(x^k, y^j) + \langle h^j, x^k - y^j\rangle|$$

with $g^j \in \partial f(y^j)$ and $h^j \in \partial\Psi(x^k, y^j)$ and recalling that $\Psi(x^k, x^k) = 0$. The approximate model is therefore written in the same manner as in the convex case, i.e.

$$\hat{F}_{\lambda_k}(x^k, x) = \max_{j\in J^k}\{f(x) - \alpha_j^k + \langle s^j, x - x^k\rangle\} + \frac{1}{2}\lambda_k|x - x^k|^2,$$

using the new definition of α_k^j and where as before s^j is defined by $s^j \in g^j + \lambda_k h^j$ and $\{y^j\}_{j\in J^k}$ is a collection of points in a neighborhood of x^k.

With this lower approximation in hand, an approximate proximal step is taken, yielding a precandidate for x^{k+1}

$$z^{k+1} = \mathrm{argmin}_{z\in\mathbb{R}^n}\max_{j\in J^k}\{f(x^k) - \alpha_k^j + \langle s^j, x - x^k\rangle\} + \frac{\lambda_k}{2}|x - x^k|^2.$$

Let d^k denote the direction given by $d^k = z^{k+1} - x^k$ or, in other words,

$$d^k = \frac{1}{\lambda_k}\sum_{j\in J^k}\bar{u}_j s^j,$$

and let p^k denote the convex combination of collected subgradients

$$p^k = \sum_{j\in J^k}\bar{u}_j s^j.$$

In the following implementation, z^{k+1} does not directly provide a candidate for a descent step. Rather, a linesearch is performed along the direction d^k. The goal of this linesearch is twofold. Firstly, the linesearch provides a improvement of the result obtained using only first order approximation of f and Ψ, and therefore refines the implementation given in section 5 for the convex case. Secondly, if the linesearch only leads to a null step, the linesearch provides a systematic refinement of the local subgradient information. Indeed, as discussed in [7] and[19], under some assumptions of weak upper semi-smothness, a new candidate is obtained which satisfies

$$\langle s^{k+1}, d^k\rangle > 0$$

when a null step is taken.

The question of optimality is solved in the following manner. The main difference between the convex case and the nonconvex case is the manner in which an approximate subdifferential is generated. In the convex case, the same threshold δ_k was used to control both descent tests and the size of the ϵ-subdifferential (recall relation (5.12) and Lemma 5.2). In the nonconvex case, the use of the Goldstein ϵ-subdifferential, based on the distance from past subgradients to the current iterate, leads to a different strategy. A locality measure a^k is chosen for bounding from above the "size" of the Golstein ϵ-subdifferentials. In Lemma 6.1 below, we show that p^k belongs to the sum of Goldstein ϵ-subdifferentials at x^k, whose size is controled by a^k. For this purpose, we introduce a property called *quazi-symmetry*. *Quazi-symmetry* is then added to the assumptions on Ψ. Along the iterations, past subgradients are deleted from the bundle information when the distance from the corresponding neighbor y^j to the current iterate becomes excessive. More precisely, deletion occurs when

$$|p^k| > m_a a^k,$$

where m_a is a fixed parameter supplied to the algorithm. In other words, $|p^k|$ should not vanish faster than a fraction of the locaty measure. Convergence of p^k towards zero is obtained using the same type of proof as in [10] which implies convergence of a^k to zero.

6.2. A generalized proximal bundle method for nonconvex functions. In the sequel, at iteration k we will use the notation $s(y)$ for any element of $\partial f(y) + \lambda_k \partial \Psi(x^k, y)$ and

$$\alpha(x,y) = |f(x) - f(y) - \lambda_k \Psi(x,y) - \langle s(y), x - y\rangle|$$

ALGORITHM 2.

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

Step 1 (Proximal step) *Solve the following quadratic subproblem*

$$\begin{aligned} &\bar{u} = \mathrm{argmin}_{u \in \mathbb{R}^{\mathrm{card}(J^k)}} \{\tfrac{1}{2}\Big|\textstyle\sum_{j\in J^k} u_j s^j\Big|^2 + \sum_{j \in J^k} u_j \alpha_j^k\} \\ &\textit{subject to} \\ &u_j \geq 0, \quad \textstyle\sum_{j \in J^k} u_j = 1. \end{aligned} \tag{6.1}$$

Set

$$\begin{aligned} p^k &= -\sum_{j \in J^k} \bar{u}_j s^j \\ \alpha^k &= \sum_{j\in J^k} \bar{u}_j \alpha_k^j \\ v^k &= |p^k|^2 + \alpha^k \end{aligned} \tag{6.2}$$

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

Step 2 (Stopping criterion) *If* $\max\{|d^k|, m_a a^k\} \leq \epsilon_s$, *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 (Linesearch) *Compute two stepsizes* t_L^k *and* t_R^k, $t_L^k \leq t_R^k$ *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*

- $f(x^k) - f(x^{k+1}) - \lambda_k \Psi(x^k, x^{k+1}) > m_L t|v^k|$ *and*

$$\begin{cases} t_L = t_R > t^k \text{ or} \\ \alpha(x^k, x^{k+1}) > m_\alpha |v^k|, \end{cases}$$

(descent step)

or

- $-\alpha(x^{k+1}, yk+1) + \langle s^{k+1}, d^k \rangle \geq m_R v^k$, $t_R^k \leq t^k$ *and* $t_L^k = 0$, **(null step)**
with

$$s^{k+1} = g(y^{k+1}) + \lambda_k h_k(y^{k+1}). \tag{6.3}$$

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) *Update* λ_k. *Set*

$$n_j^{k+1} = n_j^k + |x^{k+1} - x^k|, \tag{6.4}$$

$$n_{k+1}^{k+1} = |x^{k+1} - y^{k+1}| \tag{6.5}$$

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

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

We now describe the line search procedure.

PROCEDURE 1 (Linesearch).

(i) Set $t_L = 0$, $t = t_U = 1$ *and* $m = (m_R - m_\alpha + m_L)$.

(ii) If $f(x^k) - f(x^k + td^k) - \lambda_k \Psi(x^k, x^t + td^k) > mt|v^k|$ *set* $t_L = t$. *Otherwise set* $t_U = t$.

(iii) If $f(x^k) - f(x^k + td^k) - \lambda_k \Psi(x^k, x^k + td^k) > m_L t|v^k|$ *and either* $t_L \geq t^k$ *or* $\alpha(x^k, x^k + td^k > m_\alpha |v^k|$, *set* $t_L^k = t_R^k = t$, $s^{k+1} = s(x^k + t_L d^k)$ *and return.*

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

$$-\alpha(x^k, x^k + td^k) + \langle s(x^k + td^k), d^k \rangle \geq m_R v^k,$$

set $t_R^k = t$, $t_L^k = 0$, $s^{k+1} = s(x^k + t_R d^k)$ *and return.*

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

6.3. Intermediate results. The convergence analysis for Algorithm 2 will require the additional assumptions below.

ASSUMPTIONS 3. (i) *The functions* $f(y)$ *and* $\Psi(x, y)$ *are semi-smooth in the variable* y *for any* x *in* $\mathbb{R}^n$.

(ii) *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 $g \in \partial^a \Psi(x,y)$, $a \geq 0$ *then* $g \in -\partial^{q(a,|x-y|)}\Psi(y,x)$. We assume that the functions f and Ψ satisfy assumptions 1, 2 and 3. We now prove two essential lemmas. The first one establishes that p^k belongs to certain Goldstein ϵ-subdifferentials at x^k. The second lemma proves an important continuity result at stationary points.

LEMMA 6.1. *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* p^k *defined in Step 1 of Algorithm 2 satisfies*

$$p^k \in \partial^{a^k} f(x^k) - \lambda_k^+ \partial^{q(a^k+b^k,a^k+b^k)+a^k}\Psi(x^*, x^k),$$

where a^k *is the locality measure defined in steps 1 and 6.*

Proof. Using (6.3), we have for any $j \in J^k$,

$$s^j \in \partial f(y^j) + \lambda_j \partial \Psi(x^j, y^j).$$

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

$$s^j \in \partial^{a^k} f(x^k) + \lambda_j \partial^{a^k}\Psi(x^j, x^k),$$

and

$$s^j \in \partial^{a^k} f(x^k) + \lambda_j \partial^{a^k+b^k}\Psi(x^j, x^*).$$

Now, by assumption 3 (ii) of quazisymmetry,

$$s^j \in \partial^{a^k} f(x^k) - \lambda_j \partial^{q(a^k+b^k,|x^*-x^j|)}\Psi(x^*, x^j),$$

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

$$s^j \in \partial^{a^k} f(x^k) - \lambda_j \partial^{q(a^k+b^k,|x^*-x^k|+a^k)+a^k}\Psi(x^*, x^k),$$

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

$$s^j \in \partial^{a^k} f(x^k) - \lambda_j \partial^{q(a^k+b^k,a^k+b^k)+a^k}\Psi(x^*, x^k), \tag{6.6}$$

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

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

which, when combined with (6.6), gives

$$s^j \in \partial^{a^k} f(x^k) - \lambda_k^+ \partial^{q(a^k+b^k,a^k+b^k)+a^k}\Psi(x^*, x^k),$$

for any $j \in J^k$. Now, due to (6.2), p^k is a convex combination of the subgradients s^j, $j \in J^k$. Recalling that $\epsilon-$Goldstein subdifferentials are closed and convex sets, we have

$$p^k \in \partial^{a^k} f(x^k) - \lambda_k^+ \partial^{q(a^k+b^k,a^k+b^k)+a^k}\Psi(x^*, x^k)$$

as desired. □

The following lemma will be needed in the sequel.

LEMMA 6.2. *All the sequences* $(x^k)_{k\in\mathbb{N}}$, $(y^k)_{k\in\mathbb{N}}$, $(s^k)_{k\in\mathbb{N}}$ *and* $(p^k)_{k\in\mathbb{N}}$ *defined in the regularized bundle algorithm are bounded.*

Proof. Due to assumption 1 (i) of inf-compactness, and the fact that algorithm 2 is a descent method, the sequence $(x^k)_{k\in\mathbb{N}}$ is bounded. The remainder of the proof can be easily adapted from [10, Lemma 3.3]. □

LEMMA 6.3. *Suppose that there exist sequences* $(x^k)_{k\in\mathbb{N}}$, $(p^k)_{k\in\mathbb{N}}$ *and* $(a^k)_{k\in\mathbb{N}}$ *and a point* $\bar{x}\in\mathbb{R}^n$ *such that* $\lim_{k\to\infty} x^k=\bar{x}$, $\lim_{k\to\infty} p^k=0$, $\lim_{k\to\infty} a^k=0$. *Define* $\lambda_k^+=\max_{j\in J^k}\lambda_j$ *for all* k *in* $\mathbb{N}$. *Assume that* f *subdifferentially dominates* $\lambda^+\Psi$ *at every point* x^* *such that* $0\in\partial f(x^*)+\lambda^+\partial\Psi(x^*,x^*)$ *for some real number* $\lambda^+>0$. *Assume in addition that* $\limsup_{k\to\infty}\lambda_k^+\leq\lambda^+$ *and that* $\liminf_{k\to\infty}\lambda_k\geq\lambda^->0$ *for some real number* λ^- *and for every* $k\in\mathbb{N}$. *Then* $\bar{x}$ *is a stationary point of* $f(x)$.

Proof. Lemma 6.1 above implies the existence of two sequences $(p_1^k)_{k\in\mathbb{N}}$ and $(p_2^k)_{k\in\mathbb{N}}$ such that $p^k=p_1^k+p_2^k$ and

$$p_1^k\in\partial^{a^k}f(x^k),$$
$$p_2^k\in-\lambda_k^+\partial^{q(a^k+b^k,a^k+b^k)+a^k}\Psi(x^*,x^k).$$

Due to lemma 6.2, $(x^k)_{k\in\mathbb{N}}$ is bounded and thus the local boundedness of the subgradients implies that the sequences $(p_1^k)_{k\in\mathbb{N}}$ and $(p_2^k)_{k\in\mathbb{N}}$ are bounded. Thus, we can extract convergent subsequences $(p_1^{\sigma(k)})_{k\in\mathbb{N}}$, and $(p_2^{\sigma(\gamma(k))})_{k\in\mathbb{N}}$ with respective limits p_1^* and p_2^* such that

$$p_1^*+p_2^*=0. \tag{6.7}$$

On the other hand, using [13, lemma 3.1], we obtain

$$p_1^*\in\partial f(\bar{x}). \tag{6.8}$$

Furthermore, since $\lim_{k\to\infty}x^k=\bar{x}$, 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 [13, lemma 3.1] and the fact that $\lambda^+=\limsup_{k\to\infty}\lambda_k^+$ one easily obtains

$$p_2^*\in-\lambda^+\partial\Psi(\bar{x},\bar{x}).$$

Now, due to quazisymmetry Assumption 3 (ii) in the case $a=0$ and $x=y=\bar{x}$, we have

$$p_2^*\in\lambda^+\partial\Psi(\bar{x},\bar{x}). \tag{6.9}$$

Finally, (6.7), (6.8) and (6.9) together give

$$0\in\partial f(\bar{x})+\lambda^+\partial\Psi(\bar{x},\bar{x}). \tag{6.10}$$

The proof is then easily completed using Lemma 3.2. □

6.4. Convergence. Convergence of Algorithm 2 is established using the properties of Algorithm 2.1 in [10], to which it is structurally very close, and the preliminary properties established in the previous section. In order to adapt the study of [10] to our generalized proximal bundle method, we make the following important observation.

The behavior of the method principally relies on the output of the linesearch (Procedure 1). In particular, in the case of descent steps, the linesearch procedure gives

$$f(x^k) - f(y^{k+1}) - \lambda_k \geq m_L t_L v^k.$$

Due to positivity of Ψ, we obtain

$$f(x^k) - f(y^{k+1}) \geq m_L t_L v^k. \tag{6.11}$$

This last equation corresponds exactly to the output given by Algorithm 2.1 of [10] in the descent case. On the other hand, the case of a null step gives

$$f(x^k) - f(y^{k+1}) - \lambda_k \Psi(x^k, y^{k+1}) \leq m_R t_R v^k,$$

which is equivalent to

$$F_{\lambda_k}(x^k, x^k) - F_{\lambda_k}(x^k, y^{k+1}) \leq m_R t_R v^k.$$

This corresponds exactly to the output of the linesearch procedure in Algorithm 2.1 of [10] when $f(.)$ is replaced by $F_{\lambda_k}(x^k, .)$. Using these strong similarities between the two methods, we now establish convergence of our procedure. The first step is study the linesearch procedure. In particular, we will need the following simple lemma.

LEMMA 6.4. *Let f_1 and f_2 be two weakly upper semismooth functions. Then $f_1 + f_2$ is also weakly upper semismooth.*

Proof. The proof is an immediate consequence of the definition. Indeed, take $p^k \in \partial\big(f_1 + f_2\big)(x + t^k d)$. Then, we have $p^k = p_1^k + p_2^k$ with

$$p_1^k \in \partial f_1(x + t^k d),$$

and

$$p_2^k \in \partial f_2(x + t^k d).$$

Since f_1 and f_2 are weakly upper semismooth, we have

$$\liminf_{k\to\infty} \langle p_1^k, d\rangle \geq \limsup_{t\downarrow 0} \frac{f_1(x + td) - f_1(x)}{t}$$

and

$$\liminf_{k\to\infty} \langle p_2^k, d\rangle \geq \limsup_{t\downarrow 0} \frac{f_2(x + td) - f_2(x)}{t}.$$

Therefore, we easily obtain

$$\liminf_{k\to\infty} \langle p_1^k + p_2^k, d\rangle \geq \limsup_{t\downarrow 0} \frac{f_1(x + td) + f_2(x + td) - f_1(x) - f_2(x)}{t}.$$

Recalling that the Lipschitz property is preserved by summation, this last equation proves the desired result. ◻

LEMMA 6.5. *The linesearch (Procedure 1) terminates in a finite number of steps.*

Proof. Due to Lemma 6.4 and Assumption 3(i), the function $f(.) + \lambda_k \Psi(x^k, .)$ is weakly upper semismooth. Therefore, the proof of Theorem 4.1 (a) in [19] can be easily adapted to our linesearch procedure. ◻

The following lemma due to Kiwiel [9] will be used in the proof of Theorem 6.9. 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 6.6. [9, 10] *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*

$$0 \leq w^{k+1} \leq w^k - (1 - m_R)^2 (w^k)^2 / 8C^2. \tag{6.12}$$

We are now in a position to establish convergence of the generalized proximal bundle method. We first study the case of a finite number of descent steps.

LEMMA 6.7. *Assume that* f *and* Ψ *satisfy Assumptions 1, 2 and 3. Assume that* f *subdifferentially dominates* $\lambda^+\Psi$ *at every point* x^* *such that* $0 \in \partial f(x^*) + \lambda^+ \partial\Psi(x^*, x^*)$ *for some real number* $\lambda^+ > 0$. *Assume in addition that* $\limsup_{k\to\infty} \lambda_k \leq \lambda^+$ *and that* $\liminf_{k\to\infty} \lambda_k \geq \lambda^- > 0$ *for some real number* λ^- *and for every* $k \in \mathbb{N}$. *Assume in addition that only a finite number of descent step is taken and let* k_0 *denote the index of the last descent step. Then* x^{k_0} *is a stationary point of* f.

Proof. Using the same type of proof as in Lemma 3.4 of [10], we obtain that

$$\lim_{k\to\infty} p^k = 0 \text{ and } \lim_{k\to\infty} a^k = 0.$$

Therefore, Lemma 6.3 implies the desired result. □

The case of an infinite number of descent steps follows.

LEMMA 6.8. *Assume that* f *and* Ψ *satisfy the assumptions in lemma 6.7. Assume that a infinite number of descent steps is taken. Then, every accumulation point of* $(x^k)_{k\in\mathbb{N}}$ *is a minimizer of* f.

Proof. Since the linesearch procedure 1 gives equation (6.11) in the case of a descent step, the sequence $(f(x^k))_{k\in\mathbb{N}}$ is nondecreasing. Furthermore, one can easily check that the proof of lemma 3.5 in [10] can also be adapted to our method. Therefore, we obtain

$$\lim_{k\to\infty} p^k = 0 \text{ and } \lim_{k\to\infty} a^k = 0.$$

Now, take any convergent subsequence from $(x^k)_{k\in\mathbb{N}}$ and let $\bar{x}$ denote its limit. Then, Lemma 6.3 implies the desired result. □

Combining these two last lemmas, we obtain our convergence theorem for the generalized proximal bundle method 2.

THEOREM 6.9. *Let* f *and* Ψ *satisfy the assumptions in lemma 6.7 Then, every accumulation point of* $(x^k)_{k\in\mathbb{N}}$ *defined by Algorithm 2 is a minimizer of* f.

7. Numerical experiments. In this section, we illustrate the bundle implementation of Section 6 for a particular nonsmooth optimization problem arising in adaptive digital signal processing for channel equalization. This problem is formulated as follows

$$\min_{x\in\mathbb{R}^n} f(x), \qquad \text{where } f(x) = \sum_{i=1}^{n} |1 - |a_i^T x|^p|^q. \tag{7.1}$$

for a given set of $n-$dimensional vectors $(a_i)_{1\leq i\leq n}$. For odd p or q, the function is nondifferentiable. A popular algorithm for implementing (7.1) is the well known constant modulus algorithm (CMA) [6, 30, 27]. The objective function has the shape represented in Figure 8.1, 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 8.2, 8.3 and 8.4.

We will compare the three regularizations on the basis of the number of iterations required, the number of function/subgradient calls and the computed minimum value of the objective function. The algorithms were run over a range of the relaxation parameters $\lambda_k = \lambda$ which we kept constant along iterations. The parameters of the regularized bundle method are set as in Table 8.1. Our results are reported in Table 8.2. The symbols it, f^* and nb denote, respectively, the minimum number of iterations N necessary to achieve $f(x^N) \leq \epsilon_s$, 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 work. Notice that the theoretical results demonstrated in Section 6 may apply to more sophisticated approaches recently developed for the standard Moreau-Yosida regularization, including trust region techniques [13] and second order methods [16] in the case of convex objective and regularization functions.

8. 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 implementations of the proximal method. The resulting bundle algorithms 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 regularization.

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. F. Bonnans, J.-Ch. Gilbert, C. Lemaréchal, and C. Sagastizabal. *Optimization numérique. Aspects théoriques et pratiques*, volume 27. Springer Verlag, 1997. Series : Mathématiques et Applications.

[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] J. B. Hirriart Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms I-II.* Springer Verlag, 1993. Grundlehren der mathematischen Wissenschaften 306.

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

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

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

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

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

[13] 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.

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

[15] C. Lemarechal and C. Sagastizabal. A class of variable metric bundle methods. *Research Report 2128*, 1993.

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

[17] 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.

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

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

[20] 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.

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

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

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

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

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

[26] 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.

[27] 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.

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

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

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

[31] 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 8.1
Parameters of the regularized bundle algorithm

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

TABLE 8.2
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. 8.1. *The objective f and its derivative f'.*

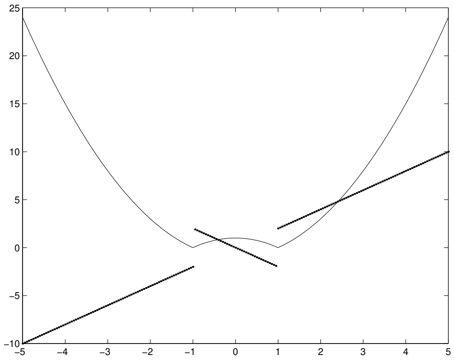

FIG. 8.2. *Moreau-Yosida regularization* Ψ_1

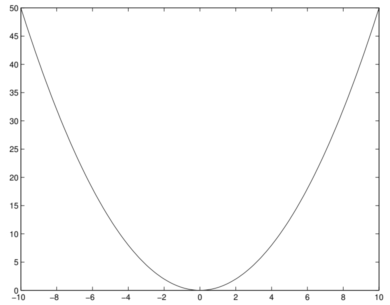

FIG. 8.3. l_1 *norm regularization* Ψ_2

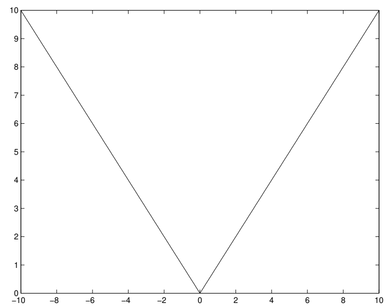

FIG. 8.4. $log(1+|.|)$ *regularization* Ψ_3