GENERALIZED PROXIMAL POINT ALGORITHMS *

STÉPHANE CHRÉTIEN† AND ALFRED O. HERO III‡

Abstract. In this paper, we address the problem of minimizing a continuous function f over $\mathbb{R}^n$. More precisely, we study the convergence properties of a general proximal point scheme of the form

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

where Ψ is a distance-like bivariate function. The analysis is performed without convexity nor differentiability assumptions. Applications of this work are to extensions of EM-type algorithms in statistical estimation. We also extend the study to the problem of minimizing f over the non-negative orthant via entropic-like proximal methods. In particular, convergence of entropic-type methods is obtained with greatly reduced regularity assumptions in comparison to previous published works.

Keywords: mathematical programming, proximal point algorithm, non-smooth regularization, EM algorithm, maximum likelihood estimation, entropic methods

AMS: 26B05, 90C26, 90C30, 90C90, 65U05

1. Introduction. We address the following minimization problem

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

where f is a continuous function. The goal of this paper is to establish the convergence of a general proximal point scheme of the form

$$x^{k+1} \in \mathrm{argmin}_{y\in\mathbb{R}^n} f(y) + \lambda_k \Psi(x^k, y), \tag{1.2}$$

where Ψ is a distance-like bivariate function (i.e Ψ is continuous, positive, $\Psi(x, y) = 0$ iff $x = y$), to a solution of (1.1). At the same time, we provide a general framework for the study of some algorithms coming from statistical estimation, like the EM algorithm, and nonlinear programming, like "entropic"-like methods. The analysis in this paper shows that neither convexity nor differentiability of f and Ψ are natural assumptions for convergence of recursion (1.2) for a general class of entropic penalties Ψ.

The original proximal point method, defined by the recursion

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

was first studied by Martinet [22] and Rockafellar [30], in the case where f is convex. When applied to the dual of a constrained maximization problem, the proximal point algorithm is equivalent to an augmented Lagrangian procedure [29]. Each iteration of the proximal point method is equivalent to a steepest descent algorithm applied to the convex differentiable Moreau-Yosida regularization of f defined by

$$F(x) = \min_{y\in\mathbb{R}^n} f(y) + \frac{1}{2}\|x - y\|^2.$$

*This work has been partially supported by AFOSR F49620-96-0028.

†Service Mathématique de la Gestion, Université Libre de Bruxelles, Bd du Triomphe, 1050 Bruxelles, Belgium. The main part of this work was done during a postdoctoral stay at the EECS Dpt., The University of Michigan, Ann Arbor, USA. (schretie@smg.ulb.ac.be)

‡Dpt. of Electrical Engineering and Computer Science, The University of Michigan, 1301 Beal Avenue, Ann Arbor, Michigan 48109-2122, USA. (hero@eecs.umich.edu).

Passing to practical implementation, most bundle methods [12] for non-smooth optimization can be viewed as approximate proximal point iterations, using the polyhedral model at iteration k

$$\hat{f}_k(y) = \max_{j \in J_{x^k}} f(y_j) + \langle g_j, y - y_j \rangle$$

instead of the objective f and where $\{y_j\}_{j \in J_x}$ is a collection of neighbor points about x^k and g_j is a subgradient of f at y_j for all $j \in J_{x^k}$. Moreover Moreau-Yosida regularization and proximal bundle methods have been much studied recently, due to the connections with second order analysis of nonsmooth functions; we refer to [16], [19], [2], [17],[20], [21].

Convergence of generalized proximal point iterations, for which Ψ is not quadratic, have recently received much attention. Generalized bundle methods using non-quadratic regularization functions have also been studied, the main contribution in this direction being the work of Frangioni [11]; see also [15]. For generalized versions of the "pure" proximal point algorithm, the regularization function Ψ is usually of the form

$$\Psi(x, y) = \sum_{i=1}^{n} x_i \phi(\frac{y_i}{x_i}), \tag{1.3}$$

where ϕ satisfies some properties insuring that Ψ is a convex differentiable distance-like function. An important instance is the case

$$\phi(t) = t \log(t) - t + 1$$

for which the corresponding Ψ is the well known Kullback-Leibler entropy function from statistics. For this reason, most generalized proximal point methods in the literature are called "entropic" proximal point methods. The main feature of entropic methods is that the domain of $\Psi(x, \cdot)$ is the positive orthant $\mathbb{R}^n_{++}$. Hence, entropic procedures automatically handle positivity constraints in the dual formulation of inequality constrained minimization problems.

The main motivation of this paper is the study of a class of algorithms for statistical estimation with application to computerized tomography and digital communication problems. The authors have recently introduced in [5] a class of methods for statistical estimation which can be recast into the generalized proximal point framework (1.2). As an important instance, it was shown in [5, 4] that the well known EM algorithm of Dempster *et al.* [7] extensively used for maximum likelihood estimation [23, 9] is a generalized proximal point procedure of the form (1.2), where $\Psi(x, y)$ is a Kullback-Leibler divergence between two density functions $k(u; x)$ and $k(u; y)$ parameterized by deterministic vectors x and y respectively, i.e.

$$\Psi(x, y) = \int_D k(u; x) \log \frac{k(u; y)}{k(u; x)} du, \tag{1.4}$$

where D is the "complete data space" in the EM literature [1]. The EM iteration, although non-quadratic, enjoys *closed form* expressions in many important applications, e.g. computerized tomography [34, 23], or mixture densities estimation [23], [3] and

[1] More precisely, the quantity Ψ is well defined if the measures $k(u; x)du$ and $k(u; y)du$ are absolutely continuous with respect to each other. In most applications in statistics, this assumption is satisfied.

[28]. Other extensions of the EM approach like the SAGE algorithm of Fessler and Hero [10] or the AECM of Meng and van Dyck [24], [25] can be viewed as coordinate-wise generalized proximal point procedures and their convergence analysis is considerably simplified in the proximal framework (see [3] in the context of mixture models). Here, we develop a convergence analysis for regularization functions Ψ which generalizes previous results to the case where

- the function Ψ *cannot*, in general, be expressed in the form of an entropic regularization given by (1.3),
- the function Ψ is not assumed to be convex and/or differentiable.

Conversely, we show in section 5.2 that our results provide at the same time an alternative approach to convergence analysis of entropic proximal methods.

The paper is organized as follows. In section 2, the generalized proximal point algorithm is presented. Convergence of the method is established in Section 3 under the assumption that either the relaxation sequence $\{\lambda_k\}$ converges to zero, or the regularization function $\Psi(x, .)$ is differentiable at x for any $x \in \mathbb{R}^n$. Approximate iterations are also considered and the approximate method is shown to converge under unrestrictive assumptions. In Section 4, we discuss the case where Ψ is nonsmooth and we describe a sufficient condition, called *subdifferential domination* for convergence of the generalized proximal iterations towards a stationary point of f. In Section 5, applications of our generalized proximal point framework are discussed. First, we present an extension of the EM algorithm for maximum likelihood estimation [7]. Next, we show how our general approach may contribute to simplifying the analysis of entropic proximal-type methods. In particular, convergence of entropic methods is established without assuming convexity or differentiability. Finally, we return to EM type algorithms and focus on the case of inverse problems with Poisson distribuded data.

We recall the following notations and definitions [6]. 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: $\mathbb{R}^n \mapsto \mathbb{R}$ 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 $\mathrm{dom} f = \{x \mid f(x) < \infty\}$. 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

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

where ∇f denotes the gradient of f. The subdifferential has the property that it is a closed and convex set. The multivalued function $x \mapsto \partial f(x)$ is upper semicontinuous and locally bounded.

2. Generalized Proximal Point Algorithms. In the remainder of this paper[2], we will make the following assumptions.

ASSUMPTIONS 1. (i) f is continuous and inf-compact.
(ii) $\Psi(x, y)$ is continuous on $\mathbb{R}^{2n}$.
(iii) $\lim_{\|x-y\| \to \infty} \Psi(x, y) = +\infty$.

[2]except in Section 5.2 and 5.3

(iv) (Positivity) $\Psi(x,y) \geq 0$ for all x, $y \in \mathbb{R}^n$.
(v) (Identifiability) $\Psi(x,y) = 0 \Leftrightarrow x = y$.
(vi) $(\lambda_k)_{k\in\mathbb{N}}$ is a nonnegative sequence of real numbers.

We now define the *generalized proximal point* algorithm

DEFINITION 2.1. *Assume that f and $\Psi(\cdot,\cdot)$ satisfy Assumptions 1 respectively. Then, the* generalized proximal point *algorithm is defined by the following recurrence starting at x^1,*

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

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

$$F_\lambda(x,y) = f(y) + \lambda\Psi(x,y), \tag{2.2}$$

and the *generalized proximal* operator will be defined by

$$P_\lambda \in \operatorname{argmin}_{y\in\mathbb{R}^n} F_\lambda(x,y). \tag{2.3}$$

REMARK 1. *First notice that minimizers in the generalized proximal iteration always exist due to the fact that $f(\cdot) + \Psi(x^k,\cdot)$ is inf-compact. However, due to non-convexity, multiple minimizers may achieve the minimum value in (2.1). In such case, the next iterate is chosen arbitrarily among these minimizers.*

REMARK 2. *Inf-compactness of f seems to be a strong assumption for the analysis of the generalized proximal point method. Indeed, one might prefer the weaker assumption of functions with unbounded level sets. The following simple one dimensional counterexample shows however that even the standard quadratic proximal point method may not attain the infimum of f when inf-compactness is dropped: define f by*

$$f(x) = \begin{cases} \exp(-x) - 1, & \text{if } x \geq 0 \\ x \text{ otherwise.} \end{cases}$$

and consider the proximal procedure given by: $x^1 = 1$ and

$$x^{k+1} = \operatorname{argmin}_{y\in\mathbb{R}} f(y) + \frac{1}{2}(y - x^k)^2. \tag{2.4}$$

Now, one easily checks that for all $a \geq 1$ and all $y \leq a$, $\Big(f(y) + \frac{1}{2}(y-a)^2\Big)' < 0$, where $'$ denotes the derivative w.r.t. y, which excludes that the proximal point algorithm (2.4) produce negative iterates whereas any minimizing sequence is negative and tends to $-\infty$.

3. Convergence analysis. In this section, we give an asymptotic analysis of the *generalized proximal point* algorithm. Although differentiability of the regularization function is not assumed in the preliminary lemmas, our main convergence theorem (Theorem 3.5) requires this assumption in the case where the chosen relaxation sequence $(\lambda_k)_{k\in\mathbb{N}}$ does not converge to zero. The case where Ψ is nonsmooth and $(\lambda_k)_{k\in\mathbb{N}}$ is bounded away from zero deserves a special treatment and is studied in Section 4.3 under an additional local Lipschitz assumption. We show that the accumulation points of the sequence defined by (2.1) are global minimizers of f in the case where $(\lambda_k)_{k\in\mathbb{N}}$ converges to zero. Finally, under an additional strict local convexity assumption, convergence is established.

3.1. Main results. We start with the following monotonicity result.

LEMMA 3.1. *Let f and Ψ satisfy Assumptions 1. For any iteration $k > 1$, the sequence $(x^k)_{k\in\mathbb{N}}$ satisfies*

$$f(x^{k+1}) - f(x^k) \le -\lambda_k \Psi(x^k, x^{k+1}) \le 0. \tag{3.1}$$

Proof. Iteration (2.1) implies that

$$f(x^{k+1}) + \lambda_k \Psi(x^k, x^{k+1}) \le f(x^k) + \lambda_k \Psi(x^k, x^k).$$

Recall that $\Psi(x^k, x^k) = 0$ due to identifiability assumption 1 (v) and $\Psi(x^k, x^{k+1}) \ge 0$ by positivity assumption 1 (iv). Thus (3.1) follows. ☐

LEMMA 3.2. *The sequence $(x^k)_{k\in\mathbb{N}}$ is bounded.*

Proof. Due to Lemma 3.1, $(x^k)_{k\in\mathbb{N}} \subset \{x \in \mathcal{R}^n \mid f(x) \le f(x^1)\}$. Then, inf-compactness of f implies the desired result. ☐

The following Lemma gives an asymptotic relation between Ψ and the euclidian norm.

LEMMA 3.3. *Let $(y^k)_{k\in\mathbb{N}}$ be any bounded sequence in $\mathbb{R}^n$ and let $(z^k)_{k\in\mathbb{N}}$ be a sequence in $\mathbb{R}^n$ satisfying $\lim_{k\to\infty} \Psi(y^k, z^k) = 0$. Then,*

$$\lim_{k\to\infty} \|y^k - z^k\| = 0.$$

Proof. First, one easily obtains that $(z^k)_{k\in\mathbb{N}}$ is bounded (use a contradiction argument and Assumption 1 (iii)). Assume that there exits a subsequence $(y^{\sigma(k)})_{k\in\mathbb{N}}$ such that $\|y^{\sigma(k)} - z^{\sigma(k)}\| \ge 3\epsilon$ for some $\epsilon > 0$ and for all large k. Since $(y^{\sigma(k)})_{k\in\mathbb{N}}$ is bounded, one can extract a convergent subsequence, and thus, we may assume without any loss of generality that $(y^{\sigma(k)})_{k\in\mathbb{N}}$ is convergent with limit y^*. Using the triangle inequality, we have $\|y^{\sigma(k)} - y^*\| + \|y^* - z^{\sigma(k)}\| \ge 3\epsilon$. Since $(y^{\sigma(k)})_{k\in\mathbb{N}}$ converges to y^*, there exists a integer K such that $k \ge K$ implies $\|y^{\sigma(k)} - y^*\| \le \epsilon$. Thus for $k \ge K$ we have $\|y^* - z^{\sigma(k)}\| \ge 2\epsilon$. Now recall that $(z^k)_{k\in\mathbb{N}}$ is bounded and extract a convergent subsequence $(z^{\sigma(\gamma(k))})_{k\ge K}$ with limit denoted by z^*. Then, using the same arguments as above, we obtain $\|y^* - z^*\| \ge \epsilon$. Finally, recall that $\lim_{k\to\infty} \Psi(y^k, z^k) = 0$. We thus have $\lim_{k\to\infty} \Psi(y^{\sigma(\gamma(k))}, z^{\sigma(\gamma(k))}) = 0$, and, due to the fact that the sequences are bounded and $\Psi(\cdot,\cdot)$ is continuous in both variables, we have $\Psi(y^*, z^*) = 0$. Thus assumption 1 (v) implies that $\|y^* - z^*\| = 0$ and we obtain a contradiction. Hence, $\lim_{k\to\infty} \|y^k - z^k\| = 0$ as claimed. ☐

The following lemma is standard.

LEMMA 3.4. *We have*

$$\lim_{k\to\infty} \lambda_k \Psi(x^{\sigma(k)}, x^{\sigma(k)+1}) = 0.$$

Proof. Lemma 3.1 implies that the sequence $(f(x^k))_{k\in\mathbb{N}}$ is monotone nonincreasing. Since f is inf-compact and continuous over $\mathbb{R}^n$, $\inf_{x\in\mathbb{R}^n} f(x) > -\infty$. Thus, $(f(x^k))_{k\in\mathbb{N}}$ is convergent and $\lim_{k\to\infty} f(x^k) - f(x^{k+1}) = 0$. Applying Lemma 3.1 again, we obtain the desired result. ☐

We are now ready to establish our global convergence results.

THEOREM 3.5. *Assume that the sequence $(\lambda_k)_{k\in\mathbb{N}}$ is convergent with limit λ^*. Let x^* be any accumulation point of $(x^k)_{k\in\mathbb{N}}$. Then, if $\lambda^* > 0$, f is locally Lipschitz* *and $\Psi(x,.)$ is differentiable at x for all x, x^* is a stationary point of $f(x)$. Moreover, without assuming local Lipschitzianity of f neither differentiability of Ψ, if $(\lambda_k)_{k\in\mathbb{N}}$ converges to zero, x^* is a global minimizer of f.*

Proof. *First part.* We first consider the case where $\lambda^*>0$. Thus, without loss of generality, we may reduce the analysis to the case where $\lambda_k\geq\lambda>0$ for a certain λ. Take a convergent subsequence $(x^{\sigma(k)})_{k\in\mathbb{N}}$ of $(x^k)_{k\in\mathbb{N}}$ with limit point x^*. Since $\lambda_k>\lambda>0$ for all k, Lemma 3.4 implies that $\lim_{k\to\infty}\Psi(x^{\sigma(k)},x^{\sigma(k)+1})=0$. Using, Lemma 3.3 we obtain that $(x^{\sigma(k)+1})_{k\in\mathbb{N}}$ also converges to x^*. 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 $x\in\mathbb{R}^n$,

$$\begin{aligned}F_{\lambda^*}(x^{\sigma(k)},x^{\sigma(k)+1})&+(\lambda_k-\lambda^*)\Psi(x^{\sigma(k)},x^{\sigma(k)+1})\\&\leq F_{\lambda^*}(x^{\sigma(k)},x)+(\lambda_k-\lambda^*)\Psi(x^{\sigma(k)},x).\end{aligned}\tag{3.2}$$

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 Assumptions 1 (i) and (ii). Fix $x\in\mathbb{R}^n$. By continuity in the first and second arguments of $F_{\lambda^*}(\cdot,\cdot)$, for any $\epsilon>0$ there exists $K_2\in\mathbb{N}$ such that for all $k\geq K_2$

$$F_{\lambda^*}(x^*,x)\geq F_{\lambda^*}(x^{\sigma(k)},x)-\epsilon,\tag{3.3}$$

and

$$F_{\lambda^*}(x^*,x^*)\leq F_{\lambda^*}(x^{\sigma(k)},x^{\sigma(k)+1})+2\epsilon.\tag{3.4}$$

Combining equations (3.3) and (3.4) with (3.1), we obtain

$$\begin{aligned}F_{\lambda^*}(x^*,x^*)\leq\quad &F_{\lambda^*}(x^*,x)+(\lambda_k-\lambda^*)\Psi(x^{\sigma(k)},x)\\&-(\lambda_k-\lambda^*)\Psi(x^{\sigma(k)},x^{\sigma(k)+1}))+3\epsilon.\end{aligned}\tag{3.5}$$

Now, since $\lambda^*=\lim_{k\to\infty}\lambda_k$, there exists an integer K_3 such that $\lambda_k-\lambda^*\leq\epsilon$ 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 Ψ is continuous and $(x^{\sigma(k)})_{k\in\mathbb{N}}$ is bounded, there exists a real constant C such that $\Psi(x^{\sigma(k)},x)\leq C$, for all $n\in\mathbb{N}$. Thus, for all $k\geq\max\{K_1,K_2,K_3\}$

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

Since no assumption was made on x, this holds for any $x\in\mathbb{R}^n$ and, letting ϵ tend to zero, we see that $x^*\in\mathrm{argmin}_{x\in\mathbb{R}^n}F_{\lambda^*}(x^*,x)$. Thus, recalling that in this part of the proof f is assumed locally Lipschitz (and thus, admits a subdifferential in the sense of Clarke) and $\Psi(x^*,.)$ is assumed differentiable at x^*,

$$0\in\partial f(x^*)+\lambda_*\nabla\Psi(x^*,x^*).$$

Since $\Psi(x^*,.)$ is minimum at x^*, $\nabla\Psi(x^*,x^*)=0$ and we thus obtain that x^* is a stationary point of $f(x)$.

Second part. We now prove that if $(\lambda_k)_{k\in\mathbb{N}}$ vanishes, x^* is a global minimizer of f without any differentiability assumption on Ψ. Let $(x^{\sigma(k)})_{k\in\mathbb{N}}$ be a convergent subsequence of $(x^k)_{k\in\mathbb{N}}$ with limit denoted x^*. By continuity of f, for any $\epsilon > 0$, there exist $K \in \mathbb{N}$ such that for all $k \geq K$,

$$f(x^*) \leq f(x^{\sigma(k)}) + \epsilon.$$

On the other hand, the proximal iteration implies that

$$f(x^{\sigma(k)}) + \lambda_{\sigma(k)-1}\Psi(x^{\sigma(k)-1}, x^{\sigma(k)}) \leq f(x) + \lambda_{\sigma(k)-1}\Psi(x^{\sigma(k)-1}, x),$$

for all $x \in \mathbb{R}^n$. Fix $x \in \mathbb{R}^n$. Thus, for all $k \geq K$,

$$f(x^*) \leq f(x) - \lambda_{\sigma(k)-1}\Psi(x^{\sigma(k)-1}, x^{\sigma(k)}) + \lambda_{\sigma(k)-1}\Psi(x^{\sigma(k)-1}, x) + \epsilon.$$

Since Ψ is a non-negative function and $(\lambda_k)_{k\in\mathbb{N}}$ is a non-negative sequence, we obtain

$$f(x^*) \leq f(x) + \lambda_{\sigma(k)-1}\Psi(x^{\sigma(k)-1}, x) + \epsilon.$$

Recall that $(x^k)_{k\in\mathbb{N}}$ is bounded due to Lemma 3.2. Thus, since Ψ is continuous, there exists a constant C such that $\Psi(x^{\sigma(k)-1}, x) \leq C$ for all k. Therefore, for k greater than K,

$$f(x^*) \leq f(x) + \lambda_{\sigma(k)-1}C + \epsilon.$$

Passing to the limit, and recalling that $(\lambda_k)_{k\in\mathbb{N}}$ tends to zero, we obtain that

$$f(x^*) \leq f(x).$$

Since no assumption was made on x, this latter equation holds for any $x \in \mathbb{R}^n$, which concludes the proof. □

3.2. Convergence to local minima under a local strict convexity assumption. Let S^* be the set of accumulation points of the sequence $(x^k)_{k\in\mathbb{N}}$. We first give the following lemma

LEMMA 3.6. *Let f and Ψ satisfy Assumptions 1. Then, for a given starting point x^1, the set S^* 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 [26, Theorem 28.1]. □

COROLLARY 3.7. *Suppose, in addition to the assumptions of Theorem 3.5, 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, for any relaxation sequence, the sequence $(x^k)_{k\in\mathbb{N}}$ converges to a local minimizer of $f(x)$.*

Proof. We obtained in Theorem 3.5 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 $\mathcal{N}$ 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. □

3.3. Approximate iterations. In this section, approximate generalized proximal methods are studied. Convergence of $\epsilon-$approximate iterations is established under quite unrestrictive assumptions. The $\epsilon-$approximate method is defined as follows.

DEFINITION 3.8. *Let $\{\epsilon^k\}_{k\in\mathbb{N}}$ be any positive sequence converging to zero. Then, given an initial point x^1, the approximate generalized proximal point algorithm is defined by the following recurrence*

$$\mathrm{val}^{k+1} = \min_{y\in\mathbb{R}^n}\{f(y)+\lambda_k\Psi(x^k,y)\}$$

and

$$\textit{choose } x^{k+1} \textit{ such that } \{f(x^{k+1})+\lambda_k\Psi(x^k,x^{k+1})\} \leq val^{k+1}+\epsilon^k. \tag{3.6}$$

We have the following convergence result

THEOREM 3.9. *Assume that the sequence $(\lambda_k)_{k\in\mathbb{N}}$ is convergent with limit λ^*. Let x^* be any accumulation point of the sequence $(x^k)_{k\in\mathbb{N}}$ generated by the approximate generalized proximal point algorithm (3.6). Then, if $\lambda^*>0$ and $\Psi(x,.)$ is differentiable at x for all x, x^* is a stationary point of $f(x)$. Moreover, if $(\lambda_k)_{k\in\mathbb{N}})$ converges to zero, x^* is a global minimizer of f.*

Proof. The result can be obtained with a slight modification in the proof of Theorem 3.5. First, assume that $\lambda^*>0$. Let K_4 be an integer such that $\epsilon_k<\epsilon$ for all $k\geq K_4$. We get from the approximate iteration that

$$F_{\lambda_k}(x^{\sigma(k)},x^{\sigma(k)+1}) \leq F_{\lambda_k}(x^{\sigma(k)},x)+\epsilon.$$

The rest of the proof is similar to the *First part* in the proof of Theorem 3.5. Now, the *Second part* of the proof of Theorem 3.5 can be easily adapted in the same manner in order to obtain that x^* is a global minimizer when $(\lambda_k)_{k\in\mathbb{N}}$ vanishes at infinity □

4. Optimality for nonsmooth regularization. In this section, both f and Ψ are assumed locally Lipschitz. Our goal is to investigate theoretical difficulties concerning first order optimality conditions for nonsmooth regularization when the relaxation parameter is bounded away from zero. 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 stationary points of f.

4.1. 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

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

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

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

which is equivalent to the stationarity condition

$$0 \in \partial F(x^*,x^*). \tag{4.1}$$

Using the calculus of subdifferentials

$$0 \in \partial F(x^*,x^*) \subset \partial f(x^*) + \partial\Psi(x^*,x^*) \tag{4.2}$$

Thus, to prove that the proximal sequence is optimal, it will be necessary to show that (4.2) implies $0 \in \partial f(x)$.

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 (4.2) 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).

4.2. Subdifferential domination. The reason that condition (4.2) 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 4.1. *Let the functions f_1 and f_2 be locally Lipschitz. The function f_1 subdifferentially dominates f_2 at the point x if the associated directional derivatives satisfy*

$$|f_1'(x,d)| > |f_2'(x,d)| \quad \forall d \in \mathbb{R}^n. \tag{4.3}$$

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 $0 \le t \le 1$ at x. The following lemma establishes that the conditions given by (4.2) specify a stationary point of the objective function under the subdifferential domination hypothesis.

Lemma 4.2. *Let $F(x) = f_1(x) + f_2(x)$ where f_1 and f_2 are two locally Lipschitz functions and assume the existence of a point x^* satisfying the following stationarity condition*

$$0 \in \partial f_1(x^*) + \partial f_2(x^*), \tag{4.4}$$

If f_1 subdifferentially dominates f_2 at x^*, and $0 \in \partial f_2(x^*)$, then x^* is a stationary point of f_1.

Proof. Since $0 \in \partial f_2(x^*)$ we have $f_1'(x^*, d) \geq 0$ for all $d \in \mathbb{R}^n$. Due to equation (4.4), we deduce the existence of two vectors g_1 and g_2 such that

$$\begin{aligned} & g_1 + g_2 = 0, \\ & g_1 \in \partial f_1(x^*), \\ & g_2 \in \partial f_2(x^*), \end{aligned}$$

and thus

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

for all $d \in \mathbb{R}^n$. Since, by definition of subgradients, $f_1'(x, d) \geq \langle g_1, d\rangle$ and $f_2(x^*, d) \geq \langle g_2, d\rangle$, we have $f_1'(x^*, d) + f_2'(x^*, d) \geq 0$. Thus, we obtain

$$f_1'(x^*, d) \geq -f_2'(x^*, d) \tag{4.5}$$

for all d in $\mathbb{R}^n$. We now proceed by contradiction. Suppose that $f_1'(x^*, d) < 0$ for some $d \in \mathbb{R}^n \setminus \{0\}$. For this d, since $f_2'(x^*, d) \geq 0$, equation (4.5) shows that

$$|f_1'(x^*, d)| \leq |f_2'(x^*, d)|.$$

Since this last equation contradicts the domination assumption, we deduce that $f_1'(x^*, d) \geq 0$ for all $d \in \mathbb{R}^n$. □

REMARK 3. *As noticed by one referee, this result is a consequence of Theorem V.3.3.1 in [12] in the convex case. Indeed, for f_1 and f_2 convex, the subdifferential domination condition implies $\partial f_2 \subset \partial f_1$.*

4.3. Application to generalized proximal iterations. Using Lemma 4.2 we immediately deduce the optimality of the fixed points of the generalized proximal point mapping P_λ with relaxation parameter $\lambda \geq 0$.

LEMMA 4.3. *Let x^* be a fixed point of P_λ, $\lambda > 0$. Assume that $f(x)$ subdifferentially dominates $\lambda\Psi(x^*, \cdot)$ at x^*. Then x^* is a stationary point of $f(x)$.*

Therefore, using the concept of subdifferential domination, we obtain the following straightforward generalizations of Theorem 3.5, Corollary 3.7 and Theorem 3.9 of the previous section.

THEOREM 4.4. *(i) Assume that the sequence $(\lambda_k)_{k\in\mathbb{N}}$ is convergent with limit λ^*. Let x^* be any accumulation point of $(x^k)_{k\in\mathbb{N}}$ generated by the generalized proximal point algorithm (2.1). Then, if f subdifferentially dominates $\lambda^*\Psi(x^*, \cdot)$ at x^*, then x^* is a stationary point of $f(x)$. Moreover, if $(\lambda_k)_{k\in\mathbb{N}}$ converges to zero, x^* is a global minimizer of f.*

(ii) Suppose, in addition to the assumptions of Theorem 3.5, 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 $f(x)$. Moreover, if $(\lambda_k)_{k\in\mathbb{N}}$ converges to zero, x^* is a global minimizer of f.

(iii) Assume that the sequence $(\lambda_k)_{k\in\mathbb{N}}$ is convergent with limit λ^*. Let x^* be any accumulation point of the sequence $(x^k)_{k\in\mathbb{N}}$ generated by the approximate generalized proximal point algorithm (3.6). Then, if f subdifferentially dominates $\lambda^*\Psi(x^*, \cdot)$ at x^*, then x^* is a stationary point of $f(x)$. Moreover, if $(\lambda_k)_{k\in\mathbb{N}}$ converges to zero, x^* is a global minimizer of f.

5. Applications. In this section, we present the main application which motivated this research, i.e. EM-type algorithms for statistical estimation and their extensions. Next, we show how the framework developed in this paper adapts to the study of entropic-type algorithms. In particular, a new convergence result is established using less restictive assumptions on ϕ than in the existing literature.

5.1. EM-type algorithms for statistical estimation. Consider the following standard problem in statistical estimation theory. Let $(\Omega, \mathcal{F}, P)$ be a probability space. Let $U(\omega) \in \mathbb{R}^m$ be a vector of random observed data defined on this probability space. Assume that the distribution of $U(\omega)$ is governed by a certain parameter $x \in \mathbb{R}^n$. The problem of interest is to estimate in a certain sense the parameter x. Estimation of x can be performed using a Maximum Likelihood strategy (ML). Given a sample y of the random observation data, the ML approach constists of maximizing the logarithm of the density function of $U(\omega)$, taken at the given sample u, as a function of the parameter x, i.e. find a solution to

$$\hat{x}_{ML} = \mathrm{argmax}_{x \in \mathbb{R}^n} \log f_U(u; x), \tag{5.1}$$

where $f_U(u;x)$ is the density of $U(\omega)$ with parameter x. In other words, x is chosen to be the parameter of the density of $U(\omega)$ with renders observation of the sample u the most likely.

The EM algorithm is one of the most popular methods for computing maximum likelihood estimates in the statistical literature. The EM was introduced in its current form by Dempster *et al.* in [7] and has been extensively used in many areas of statistics and signal processing; see [23]. It can be briefly described in three steps:

- a "more informative" data $V(\omega)$ is introduced, with density $f_V(v;x)$ also parameterized by x, and such that there exists a possibly non-injective map h with $U(\omega) = h(V(\omega))$ for any ω. For instance, many problems in statistical estimation are concerned with missing data, i.e. data which are not observed. In such a case, $V(\omega)$ is the random vector consisting in observed and unobserved data ($V(\omega)$ is called "complete data" in [7]) and $U(\omega)$ contains only the data which are actually observed.
- the maximization problem (5.1) is replaced by a minimum mean square estimator of $\log f_V(v;x)$ given by

$$\mathsf{E}\Big[\log f_V(v;x) \mid u; \tilde{x}\Big] \tag{5.2}$$

  where $\tilde{x}$ is an estimate of $\hat{x}_{ML}$. This estimator of the likelihood can be written in integral form

$$\int_{h^{-1}(u)} \log f_V(v;x) \frac{f_V(v;\tilde{x})}{f_U(u;\tilde{x})} d\mu(v),$$

  where μ is an appropriate measure on the range of V.
- now, the EM algorithm is obtained by maximizing the estimate of the likelihood (5.2) with $\tilde{x} = x^k$. It is thus defined by the following recurrence

$$x^{k+1} = \mathrm{argmax}_{x \in \mathbb{R}^n} \mathsf{E}\Big[\log f_V(v;x) \mid u; x^k\Big].$$

The most interesting feature of the EM methodology is the simplicity of its implementation. Closed forms can be obtained in many interesting applications such as computerized tomography or mixture densities estimation. The EM procedure also enjoys the property of fast convergence in the first iterations often observed in practice. In [5, Proposition 1], we proposed a proximal interpretation of the EM. More precisely, we showed that the EM is a generalized proximal point algorithm with unit relaxation parameter ($\lambda_k = 1\ \forall k$) and with Ψ given by

$$\Psi(x,y) = \mathsf{E}\Big[\log\frac{k(v|u;x)}{k(v|u;y)}|u;x\Big], \tag{5.3}$$

where $k(v \mid u; x)$ is the conditional density given by

$$k(v \mid u;x) = \frac{f_V(v;x)}{f_U(u;x)},$$

and which is well defined if the family of measures $\Big(f_V(v;x)\mu(dv)\Big)_{x\in\mathbb{R}^n}$ is absolutely continuous. In particular the family of regularization functions Ψ defined by (5.3) is well known as the Kullback Liebler divergence (also "entropy", or "information measure").

The results of Sections 3 and 4, when applied to the EM methodology, improve in particular on the work of Wu [35], where no relaxation of the EM is proposed and where differentiability is required for some proofs). Relaxation of the EM algorithm was first proposed by Iusem [14], but only for the specific case of Poisson distributed data. However, the important case of Poisson data is not completely covered by the theory presented up to this point since, in this case, the likelihood function and the Kullback-distance regularization function are defined on the nonnegative orthant only. The Poisson case is studied in Section 5.3 after a detour on entropic-type methods for nonlinear programming in Section 5.2. A class of EM-type methods is also proposed in [5] which extends the EM methodology to generalized proximal methods with variable relaxation parameters λ_k. We showed in [5] that superlinear convergence can be obtained when using relaxation sequences $(\lambda_k)_{k\in\mathbb{N}}$ converging to zero, in the case where f is twice continuously differentiable and elliptic. Hence, the stable properties of EM in the first iterations can be preserved while superlinear acceleration can be insured by appropriate relaxation.

5.2. Entropic methods. Entropic proximal methods have recently received much attention; for instance, see [8], [13], [27], [31], [33], [1], [18], and [32]. In this section, we propose a new analysis of entropic-type methods in the spirit of Section 3. In particular, the convergence of entropic-type methods is established under very unrestrictive assumptions on the divergence function. Moreover, neither convexity nor differentiability are required.

5.2.1. Entropic methods: Description. Consider the nonlinear programming problem

$$(P) \qquad \max_{u\in\mathbb{R}^m} g_0(u) \mid g_i(u) \le 0,\ i = 1,\dots,n.$$

Let $\mathbb{R}^n_+$ (resp. $\mathbb{R}^n_{++}$ denote the nonnegative (resp. positive) orthant. Problem (P) can be handled via Lagrangian duality which leads to the following unconstrained problem

$$(D)\ \min_{x\in\mathbb{R}^n_+} f(x), \tag{5.4}$$

where

$$f(x) = \max_{u \in \mathbb{R}^m} g_0(u) - x^{\mathsf{T}} g(u).$$

is the dual function and where x^{T} is the transpose of x and $g(u) = (g_1(u), \dots, g_n(u))^{\mathsf{T}}$.

The entropic proximal method for solving the dual problem (5.4) is given by

$$x^{k+1} = \mathrm{argmin}_{x \in \mathbb{R}^n} f(x) + \lambda_k \frac{1}{2} d_\phi(x, x^k). \tag{5.5}$$

where d_ϕ is called a ϕ-divergence and is of the form

$$d_\phi(x, y) = \sum_{i=1}^{n} x_i \phi(\frac{y_i}{x_i}), \tag{5.6}$$

where ϕ has domain the positive real half line. Thus, entropic proximal methods enjoy the stable property of automatically handling the positivity constraint coming from the Lagrangian duality framework. The associated approximate scheme is given by

$$x^{k+1} = \mathrm{argmax}_{x \in \mathbb{R}^n} f(x) + \lambda_k \frac{1}{2} d_\phi(x, x^k) + \epsilon_k, \tag{5.7}$$

for a sequence ϵ_k given *a priori* and converging to zero. For this approximate scheme, it was proved in [32] that $\lim_{k\to\infty} f(x^k) = f^* = \inf\{f(x) : x \in \mathbb{R}^n_+\}$ under the following assumptions:

ASSUMPTIONS 2. *(i) ϕ is twice differentiable on* int*(*dom *ϕ)=$(0, +\infty)$.*
(ii) ϕ is strictly convex on its domain.
(iii) $\lim_{t\to 0^+} \phi'(t) = -\infty$.
(iv) $\phi(1) = \phi'(1) = 0$ and $\phi''(1) > 0$.
(v) either $\phi'(t) \leq \phi''(1) \log t \;\; \forall t > 0$ or $\phi''(1)(1 - \frac{1}{t}) \leq \phi'(t) \;\; \forall t > 0$.
and the following conditions on the parameters λ_k and ϵ_k:

$$\sum_{k=1}^{\infty} \lambda_k^{-1} = \infty \text{ and } \lim_{k\to\infty} \lambda_k \epsilon_k (\sum_{i=1}^{k} \lambda_i^{-1}) = 0. \tag{5.8}$$

Moreover, the sequence $(x^k)_{k\in\mathbb{N}}$ converges to an optimal solution of (D) if

$$\sum_{i=1}^{\infty} \frac{\epsilon_k}{\lambda_k} < \infty. \tag{5.9}$$

5.2.2. Main results. In this section, we study the asymptotic behavior of entropic methods using an approach similar to the one developed in Section 3. In order to recast entropic-type proximal methods into our framework, we first list a new set of assumptions

ASSUMPTIONS 3. (i) f is continuous on $\mathbb{R}^n$ and inf-compact.
(ii) Ψ takes the form

$$\Psi(x, y) = \sum_{j=1}^{m} a_j^{\mathsf{T}} x \phi\Big(\frac{a_j^{\mathsf{T}} y}{a_j^{\mathsf{T}} x}\Big).$$

where $a_j \in \mathbb{R}^n$, $j = 1, \dots, m$, $m \geq n$
(iii) dom ϕ=$(0, +\infty)$.

(iv) Ψ is continuous on $(0,+\infty)$
(v) ϕ is decreasing on $(0,1)$ and increasing on $(0,+\infty)$ and $\phi(1)=0$

Let A denote the matrix whose lines are the $a_j^\top$, $j=1,\dots,m$. In the standard theory of entropic proximal point methods, A is taken as the identity matrix. However, it can also be useful to handle constraints which are linear transformations of the non-negative orthant.

We now propose in passing the following non-degeneracy result which is similar to Lemma 3.3 and does not seem to have been previously noticed in the literature. However, its proof requires stronger assumptions than necessary for the convergence Theorem 5.2 below and is relegated to the appendix (Section 7).

LEMMA 5.1. *Assume that $\Psi = d_\phi$ as given by (5.6) and satisfies Assumptions 2 (i)-(iv). Then, for any bounded sequence $(x^k)_{k\in\mathbb{N}} \subset \mathbb{R}^n_{++}$ and for any sequence $(y^k)_{k\in\mathbb{N}} \subset \mathbb{R}^n_{++}$, $\lim_{k\to\infty}\Psi(x^k,y^k)=0$ implies $\lim_{k\to\infty}\|x^k-y^k\|=0$.*

We now establish our main result on generalized proximal methods with Assumptions 3. The following theorem proves a convergence result towards a global minimizer over the non-negative orthant with greatly reduced regularity conditions on ϕ when compared to Assumptions 2(i)-(iv). A new rule for updating the sequence of relaxation parameters is introduced. The proof is adapted from the *Second part* of the proof of Theorem 3.5.

THEOREM 5.2. *Let f and Ψ satisfy Assumptions 3. Let $(x^k)_{k\in\mathbb{N}}$ be a sequence generated by the approximate proximal method (3.6), with starting point in $\mathbb{R}^n_{++}$ and with a sequence of errors $(\epsilon_k)_{k\in\mathbb{N}}$ converging to zero. Assume that the relaxation sequence $(\lambda_k)_{k\in\mathbb{N}}$ converges to zero and that*

$$\lambda_k \le \delta_k \max_{1\le i\le n} \frac{1}{a_j^\top x^k \phi\left(\frac{\alpha_k}{a_j^\top x^k}\right)}, \tag{5.10}$$

where $(\delta_k)_{k\in\mathbb{N}}$ is an arbitrary positive sequence converging to zero and where $(\alpha_k)_{k\in\mathbb{N}}$ is an arbitrary positive sequence tending to infinity. Then, any accumulation point of the sequence $(x^k)_{k\in\mathbb{N}}$ is a global minimizer of f over $A^{-1}(\mathbb{R}^m_+)$, the inverse image through A of the nonnegative orthant.

Proof. Clearly, by Assumption 3 (iii), the sequence $(Ax^{\sigma(k)})_{k\in\mathbb{N}}$ is positive. Notice that Lemma 3.2 applies to the present context and thus, $(x^k)_{k\in\mathbb{N}}$ is bounded. Let $(x^{\sigma(k)})_{k\in\mathbb{N}}$ be a a convergent subsequence of $(x^k)_{k\in\mathbb{N}}$ with limit point x^*. By continuity of f, for any $\epsilon>0$, there exists $K\in\mathbb{N}$ such that for all $k\ge K$,

$$f(x^*) \le f(x^{\sigma(k)}) + \epsilon.$$

Fix x such that Ax belongs to $\mathbb{R}^m_{++}$. Thus, as in the *Second part* of the proof of Theorem 3.5, for k sufficiently large,

$$f(x^*) \le f(x) + \lambda_{\sigma(k)-1}\Psi(x^{\sigma(k)-1},x) + \epsilon. \tag{5.11}$$

We now study the behavior of $\lambda_{\sigma(k)-1}\Psi(x^{\sigma(k)-1},x)$ when k tends to infinity. More precisely, we have to prove that for all $j\in\{1,m\}$, $\lambda_{\sigma(k)-1}a_j^\top x^{\sigma(k)-1}\phi(a_j^\top x/a_j^\top x^{\sigma(k)-1})$ tends to zero. Fix $j\in\{1,m\}$. We distinguish two cases

(i) First, assume that $(a_j^\top x^{\sigma(k)-1})_{k\in\mathbb{N}}$ is bounded away from zero, i.e. there exists a positive constant γ_1 such that $a_j^\top x/a_j^\top x^{\sigma(k)-1} \le \gamma_1$. Thus, there exists a positive real b such that $a_j^\top x/a_j^\top x^{\sigma(k)-1} \le b$ for all k. Moreover, since $(x^k)_{k\in\mathbb{N}}$ is bounded, there exist a positive constant γ_2 such that $\gamma_2 \leq a_j^\top x / a_j^\top x^{\sigma(k)-1}$. Thus, since ϕ is continuous on $(0, +\infty)$, and since

$$0 < \gamma_2 \leq a_j^\top x / a_j^\top x^{\sigma(k)-1} \leq \gamma_1, \tag{5.12}$$

there exists a constant C such that

$$\phi\Big(\frac{a_j^\top x}{a_j^\top x^{\sigma(k)-1}}\Big) \leq C.$$

Now, since $a_j^\top x^{\sigma(k)-1}$ is positive and bounded and since $(\lambda_k)_{k\in\mathbb{N}}$ converges to zero, we get that $\lambda_{\sigma(k)-1} a_j^\top x^{\sigma(k)-1} \phi\Big(\frac{a_j^\top x}{a_j^\top x^{\sigma(k)-1}}\Big)$ tends to zero.

(ii) Next, assume that there exists a subsequence $(a_j^\top x^{\sigma(\gamma(k))-1})_{k\in\mathbb{N}}$ tending to zero. Thus, for k sufficiently large,

$$\frac{a_j^\top x}{a_j^\top x^{\sigma(\gamma(k))-1}} \geq 1.$$

On the other hand, since $(\alpha^{\sigma(\gamma(k))-1})_{k\in\mathbb{N}}$ is increasing, for k sufficiently large,

$$\frac{\alpha^{\sigma(\gamma(k))-1}}{a_j^\top x^{\sigma(\gamma(k))-1}} \geq \frac{a_j^\top x}{a_j^\top x^{\sigma(\gamma(k))-1}}.$$

We conclude that for k great enough, since $\phi(t)$ is increasing for $t \in [1, +\infty)$,

$$a_j^\top x^{\sigma(\gamma(k))-1} \phi\Big(\frac{\alpha^{\sigma(\gamma(k))-1}}{a_j^\top x^{\sigma(\gamma(k))-1}}\Big) \geq a_j^\top x^{\sigma(\gamma(k))-1} \phi\Big(\frac{a_j^\top x}{a_j^\top x^{\sigma(\gamma(k))-1}}\Big).$$

Now, due to the choice of $\lambda_{\sigma(\gamma(k))-1}$ prescribed by equation (5.10), we obtain that

$$\lambda_{\sigma(\gamma(k))-1} a_j^\top x^{\sigma(\gamma(k))-1} \phi\Big(\frac{a_j^\top x}{a_j^\top x^{\sigma(\gamma(k))-1}}\Big) \leq \delta_{\sigma(\gamma(k))-1}.$$

and thus, since $(\delta_k)_{k\in\mathbb{N}}$ converges to zero, we obtain that

$$\lim_{k\to\infty} \lambda_{\sigma(\gamma(k))-1} a_j^\top x^{\sigma(\gamma(k))-1} \phi\Big(\frac{a_j^\top x}{a_j^\top x^{\sigma(\gamma(k))-1}}\Big) = 0.$$

Combining cases (i) and (ii), we obtain that $\lambda_{\sigma(k)-1}\Psi(x^{\sigma(k)-1}, x)$ tends to zero as k tends to infinity. Therefore, for k great enough,

$$\lambda_{\sigma(k)-1}\Psi(x^{\sigma(k)-1}, x) \leq \epsilon.$$

Combining this result with (5.11), we obtain for k great enough

$$f(x^*) \leq f(x) + 2\epsilon.$$

Passing to the limit, we conclude that

$$f(x^*) \leq f(x). \tag{5.13}$$

Since no assumption was made on x, this latter equation holds for any $x \in A^{-1}(\mathbb{R}^m_{++})$. Since f is continuous, (5.13) also holds for all $x \in A^{-1}(\mathbb{R}^m_+)$ which concludes the proof. $\square$

REMARK 4. *The latter convergence result contained in Theorem 5.2 requires neither convexity nor differentiability of f and Ψ, although in pratice f is convex due to Lagrangian duality and Ψ will often be chosen with "better" properties. Theorem 5.2 seems however to indicate that neither convexity nor differentiability are "natural" assumptions in the convergence properties.*

Let us now comment on the rule for updating the relaxation parameter λ_k given by (5.10). First, rule (5.10) does not depend on the approximation tolerance parameter ϵ_k as thus differ from rules (5.8) and (5.9) of [32]. Moreover, the speed of convergence of the sequence $(\lambda_k)_{k\in\mathbb{N}}$ towards zero is never constrained to be slow, but can be imposed to be fast when $a_j^\top x^k$ tends to zero for some j and $\phi(\alpha_i^k/a_j^\top x^k)$ "explodes". Although intuitively favorable for the convergence speed of the sequence of iterates $(x^k)_{k\in\mathbb{N}}$, the consequences of such a phenomenon have to be investigated in practical implementations of the method.

5.3. Return to EM-type algorithms: Computerized tomography with Poisson data. We now return to EM-type algorithms and focus on the special case where the data U and V are Poisson distributed. We are interested in solving maximum likelihood problems for computerized tomography drawn from the mathematical model given by Vardi *et al.* [34]. The function to be maximized is the log-likelihood function given by

$$f(x) = \sum_{j=1}^m \sum_{i=1}^n -A_{j,i}x_i + \sum_{j=1}^m u_i \log(\sum_{i=1}^n A_{j,i}x_i),$$

Where this time A is a matrix of transition probabilities, which implies that all $A_{j,i}$ are non-negative. Let $a_j^\top$, $j = 1, \dots, m$ denote the columns of A. The conditional expectation of the complete data likelihood (5.2) is well known (see [34] for instance) to be given by

$$\mathsf{E}\Big[\log f_V(v;y) \mid u;x\Big] = -\sum_{j=1}^m a_j^\top y + \sum_{j=1}^m \sum_{i=1}^n \frac{u_j x_i A_{j,i}}{a_j^\top x} \log(A_{j,i}y_i) \tag{5.14}$$

Applying the approach of [5] (briefly described in Section 5.1), one easily checks that the regularization distance-like function Ψ is given by

$$\Psi(x,y) = f(y) - \mathsf{E}\Big[\log f_V(v;y) \mid u;x\Big] - \Big(f(x) - \mathsf{E}\Big[\log f_V(v;x) \mid u;x\Big]\Big), \tag{5.15}$$

which gives

$$\Psi(x,y) = \sum_{j=1}^m \frac{1}{a_j^\top x}\Big\{\sum_{i=1}^n u_j A_{j,i}x_i \log\frac{x_i}{y_i} + u_j a_j^\top x \log\Big(\frac{a_j^\top y}{a_j^\top x}\Big)\Big\}. \tag{5.16}$$

The expression of Ψ implicitely implies that all x_i be non-negative, with the convention $0\log 0 = 0$. One can easily check that $\Psi \geq 0$: using convexity of $-\log$,

$$\log\big(\frac{a_j^\top x}{a_j^\top y}\big) = -\log\Big(\sum_{i=1}^n \big(\frac{A_{j,i}x_i}{\sum_{i=1}^n A_{j,i}x_i}\frac{y_i}{x_i}\big)\Big) \geq -\sum_{i=1}^n \frac{A_{j,i}x_i}{\sum_{i=1}^n A_{j,i}x_i}\log\frac{y_i}{x_i},$$

and the desired result follows by multiplying both terms by $\sum_{i=1}^n A_{j,i}x_i = a_j^\top x$. Now, following the analysis of Section 5.2, it remains to find a condition on the sequence of relaxation parameters $(\lambda_k)_{k\in\mathbb{N}}$ such that

$$\lim_{k\to\infty} \lambda_k \Psi(x^k, x) = 0,$$

for any positive x and any bounded sequence $(x^k)_{k\in\mathbb{R}_+^n}$. This is easily obtained by taking

$$\lambda_k = \min_{1\le j\le m} a_j^\top x^k \delta_k,$$

where $(\delta_k)_{k\in\mathbb{N}}$ is a positive sequence converging to zero.

REMARK 5. *Alternative proofs of convergence for the standard EM algorithm ($\lambda_k = 1$) for the computerized tomography problem can be found in [34, Appendix], and in [9] using a different relaxed formulation. However, our approach offers a systematic way of handling more general non-convex likelihood functions in a proximal methodology using the regularization Ψ defined by (5.16) while insuring global convergence towards global minimizers, whereas the analysis in [34] and [9] requires convexity of the objective function.*

6. Conclusion. In this paper, generalized proximal point methods were studied and their convergence was obtained without convexity nor differentiability assumptions. Applications to maximum likelihood estimation in statistics and nonlinear programming with non-negativity constraints were presented. The results of this paper weaken the standard assumptions usually required for both applications.

7. Appendix. Proof of Lemma 5.1. Fix i in $\{1,n\}$. First, assume that for K great enough, $(x_i^k)_{k\ge K}$ is bounded away from zero. Then, $\lim_{k\to\infty}\Psi(x^k,y^k) = 0$ implies that

$$\lim_{k\to\infty} \phi\Big(\frac{y_i^k}{x_i^k}\Big) = 0.$$

Since ϕ is continuous and, due to Assumptions 2 (iv), ϕ is decreasing in $(0,1)$ and increasing in $[1,+\infty)$, we obtain that

$$\lim_{k\to\infty} \frac{y_i^k}{x_i^k} = 1,$$

which implies that

$$\lim_{k\to\infty} \frac{y_i^k - x_i^k}{x_i^k} = 0$$

and since x_i^k is bounded, $\lim_{k\to\infty} y_i^k - x_i^k = 0$.

Next, consider the case where there exists a subsequence $(x_i^{\sigma(k)})_{k\in\mathbb{N}}$ which tends towards zero. For contradiction, assume the existence of a subsequence $(y_i^{\sigma(\gamma(k))})_{k\in\mathbb{N}}$ which remains bounded away from zero for k large enough, i.e. there exists $K\in\mathbb{N}$ and $a>0$ such that $y_i^{\sigma(\gamma(k))}\ge a$ for all $k\ge K$. Thus, for k large enough,

$$\frac{y_i^{\sigma(\gamma(k))}}{x_i^{\sigma(\gamma(k))}} \ge \frac{a}{x_i^{\sigma(\gamma(k))}} > 1,$$

and, due to the fact that ϕ is increasing on $(1,+\infty)$, we get

$$x_i^{\sigma(\gamma(k))}\phi\Big(\frac{y_i^{\sigma(\gamma(k))}}{x_i^{\sigma(\gamma(k))}}\Big) \geq x_i^{\sigma(\gamma(k))}\phi\Big(\frac{a}{x_i^{\sigma(\gamma(k))}}\Big). \tag{7.1}$$

On the other hand, since ϕ is twice continuously differentiable $\phi'(0)=0$ and $\phi''(0)>0$, there exist a real $b>1$ such that $\phi'(b)>0$. Since ϕ is convex, we have for all $t>0$,

$$\phi(t) \geq \phi(b)+(t-b)\phi'(b).$$

Take $t=a/x_i^k$ in this last expression and, combining with (7.1), obtain

$$x_i^{\sigma(\gamma(k))}\phi\Big(\frac{y_i^{\sigma(\gamma(k))}}{x_i^{\sigma(\gamma(k))}}\Big) \geq x_i^{\sigma(\gamma(k))}\Big(\phi(b)+\Big(\frac{a}{x_i^{\sigma(\gamma(k))}}-b\Big)\phi'(b)\Big).$$

Passing to the limit, we get

$$\lim_{k\to\infty} x_i^{\sigma(\gamma(k))}\phi\Big(\frac{y_i^{\sigma(\gamma(k))}}{x_i^{\sigma(\gamma(k))}}\Big) \geq a\phi'(b).$$

Since $\phi'(b)>0$, we thus obtain a contradiction with the fact that $\lim_{k\to\infty}\Psi(x^k,y^k)=0$ which ends the proof. $\square$

Acknowledgement. The authors are grateful to two referees for their valuable comments and remarks which contributed greatly to improve the results and presentation of this paper.

REFERENCES

[1] Auslender, Alfred and Haddou, Mounir. An interior-proximal method for convex linearly constrained problems and its extension to variational inequalities. *Math. Program. 71A, No.1, 77-100 (1995). [ISSN 0025-5610].*

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

[3] G. Celeux, S. Chrétien, F. Forbes, and A. Mkhadri. A component-wise EM algorithm for mixtures. *Research Report 3746, INRIA*, 1999.

[4] S. Chrétien and A. O. Hero. Acceleration of the EM algorithm via proximal point iterations. In *IEEE International Symposium on Information Theory, MIT Boston*, 1998.

[5] S Chrétien and A. O. Hero. Kullback proximal algorithms for maximum likelihood estimation. *Submitted to IEEE Trans. Information Theory. Also Research Report RR-3756, INRIA*, 1998.

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

[7] 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.

[8] Eckstein, J. Nonlinear proximal point algorithms using Bregman functions, with applications to convex programming. *Math. Oper. Res. 18, No.1, 202-226 (1993).*

[9] P.P.B. Eggermont. Multiplicative iterative algorithms for convex programming. *Linear Algebra Appl.*, 130:25–42, 1990.

[10] J. A. Fessler and A. O. Hero. Space-alternating generalized expectation-maximisation algorithm. *IEEE Trans. Signal Processing*, 42:2664–2677, 1994.

[11] A. Frangioni. Generalized bundle methods. *Technical Report TR-98-04, Dipartimento di Informatica, Università di Pisa*, 1998.

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

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

[14] Alfredo N. Iusem. Convergence analysis for a multiplicatively relaxed EM algorithm. *Math. Methods Appl. Sci.*, 14(8):573–593, 1991.

[15] Kim, Sehun and Chang, Kun-Nyeong and Lee, Jun-Yeon. A descent method with linear programming subproblems for nondifferentiable convex optimization. *Math. Program. 71A, No.1, 17-28. [ISSN 0025-5610]*, 1995.

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

[17] 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.

[18] Kiwiel, K. C. Proximal minimization methods with generalized Bregman functions. *SIAM J. Control Optimization 35, No.4, 1142-1168 (1997). [ISSN 0363-0129]*.

[19] C. Lemarechal and C. Sagastizàbal. A class of variable metric bundle methods. *Research Report INRIA 2128*, 1993.

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

[21] C. Lemarechal and C. Sagastizàbal. Variable metric bundle methods: form conceptual to implementable forms. *Math. Programming*, 76B(3):393–410, 1997.

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

[23] G.J. McLachlan and T. Krishnan. *The EM algorithm and extensions.* Wiley Interscience, 1997. Wiley series in probability and statistics.

[24] T. A. Meng, X.-L. and D. A. van Dyk. The EM algorithm - an old folk song sung to a fast new tune (with discussion). *J. Roy. Stat. Soc. Ser. B*, 59:511–567, 1997.

[25] T. A. Meng, X.-L. and D. A. van Dyk. Fast EM-type implementations for mixed effects models. *J. Roy. Stat. Soc. Ser. B*, 60:559–578, 1998.

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

[27] Polyak, R. Modified barrier functions (theory and methods). *Math. Program., Ser. A 54, No.2, 177-222 (1992).*

[28] R. A. Redner and H. F. Walker. Mixture densities, maximum lilelihood and the EM algorithm. *SIAM Review*, 26:195–239, 1984.

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

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

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

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

[33] Tseng, Paul and Bertsekas,Dimitri P. On the convergence of the exponential multiplier method for convex programming. *Math. Program. 60A, No.1, 1-19 (1993).*

[34] Y. Vardi, L. A. Shepp, and L. Kaufman. A statistical model for positron emission tomography. *J. Am. Stat. Assoc.*, 80:8–37, 1985.

[35] C. F. Wu. On the convergence of the EM algorithm. *Ann. Statis.*, 11(1):95–103, 1983.