Augmented Lagrangian Methods and Proximal Point Methods for Convex Optimization

A. N. Iusem

Instituto de Matemática Pura e Aplicada (IMPA)
Estrada Dona Castorina 110, Rio de Janeiro, RJ
CEP 22460-320, Brazil
iusp@impa.br

Abstract

We present a review of the classical proximal point method for finding zeroes of maximal monotone operators, and its application to augmented Lagrangian methods, including a rather complete convergence analysis. Next we discuss the generalized proximal point methods, either with Bregman distances or ϕ-divergences, which in turn give raise to a family of generalized augmented Lagrangians, as smooth in the primal variables as the data functions are. We give a sketch of the convergence analysis for the case of the proximal point method with Bregman distances for variational inequality problems. The difficulty with these generalized augmented Lagrangians lies in establishing optimality of the cluster points of the primal sequence, which is rather immediate in the classical case. In connection with this issue we present two results. First we prove optimality of such cluster points under a strict complementarity assumption (basically that no tight constraint is redundant at any solution). In the absence of this assumption, we establish an ergodic convergence result, namely optimality of the cluster points of a sequence of weighted averages of the primal sequence given by the method, improving over weaker ergodic results previously known. Finally we discuss similar ergodic results for the augmented Lagrangian method with ϕ-divergences and give the explicit formulae of generalized augmented Lagrangian methods for different choices of the Bregman distances and the ϕ-divergences.

1 Introduction

The proximal point method can be seen as a procedure to solve convex optimization problems, or more generally monotone variational inequality problems. It replaces the original problem by a sequence of more regular subproblems. When applied to the dual of a convex optimization problem, it becomes equivalent to the augmented Lagrangian method. One feature of the proximal point method is that the subproblems

are structurally similar to the original problem, e.g. the subproblems are constrained when the original problem is constrained. Recently, the proximal point method has been generalized by changing the regularization term so that it plays also a penalization role, making the subproblems unconstrained. The regularization/penalization terms of these new methods are based on generalized distances (e.g. Bregman distances or ϕ-divergences) which replace the Euclidean distance and force the solution of the subproblems to remain in the interior of the feasible set. These generalized proximal point methods, when applied to the dual of a convex optimization problem, give raise in turn to generalized augmented Lagrangian methods. The most significant difference between these and the classical augmented Lagrangian method is that the objective of the subproblems becomes as smooth as the data function, while in the classical method the objective of the subproblems is differentiable but never twice differentiable. On the other hand, the issue of convergence of the sequence of minimizers of the subproblems is much more complex for the new methods than for the classical one. This work reviews both the classical proximal point method and the classical augmented Lagrangian method, including a fairly complete convergence analysis for both. Then it introduces the generalized proximal point and augmented Lagrangian methods, sketching their convergence analysis, and presenting new convergence results for the primal sequence generated for the proximal point method with Bregman distances, which are stronger than those previously known. We also present a version of the method which is applied not to the dual of the original convex optimization problem, but to the saddle point problem of the Lagrangian, seen as a variational inequality problem. This variant has better convergence properties than the generalized augmented Lagrangian methods originated in the application of the generalized proximal point method to the dual of the original problem. For instance, the sequence of primal iterates is authomatically bounded.

The work is organized as follows: in Section 2 we review the classical augmented Lagrangian method, tracing back its origin to "tatônnement" methods for partial equilibrium problems. In Section 3 we present the classical proximal point method for finding zeroes of point-to-set maximal monotone operators and prove its convergence. In Section 4 we establish the relation between the classical proximal point method applied to the dual of a convex optimization problem and the classical augmented Lagrangian method, and establish the convergence properties of the latter based on the results in Section 3 for the former. Sections 5, 6 and 7 introduce the concepts of Bregman distance, ϕ-divergence and variational inequality problem respectively. Section 8 presents the proximal point method with Bregman distances, which, applied either to the dual of the original problem or to the saddle point problem for its Lagrangian, genarates two generalized augmented Lagrangian methods. The connection between these methods is established in Section 9, and the convergence analysis for the dual sequences they generate is presented in Section 10. The generalized proximal point and augmented Lagrangian method with ϕ-divergences are discussed in Section 11, where their convergence properties are commented upon but not proved. Section 12 contains some new results, related to the sequence of primal iterates generated by the two generalized augmented Lagrangian methods with Bregman distances: first we introduce a strict complementarity assumption, which basically says that no constraint,

which is tight at a solution of the problem, is redundant, and under this assumption we prove optimality of the cluster points of the primal sequence (if any) for the first method, and convergence of the full primal-dual sequence to an optimal primal-dual pair for the second one. In the absence of this strict complementarity assumption, we present some ergodic convergence results for the primal sequence generated by both methods. These results refer to the cluster points of a sequence of weighted averages of the iterates of such primal sequences, which are proved to be primal optimal solutions. Differently from previous results of the same nature, we make no additional assumptions on the behavior of the sequences. Section 13 contains the explicit iterative formulae of the generalized augmented Lagrangian methods for several choices of the Bregman distance and the ϕ-divergence.

2 Classical Augmented Lagrangians

Possibly, the motivation behind augmented Lagrangian methods can be traced back to the first equilibrium models proposed by Léon Walras in the 1890's, which in turn were among the first attempts to introduce mathematical analysis in economics. Dressed in modern garments, Walras' approach could be formulated in the following way. Consumers (aggregated into just one for the sake of simplicity) have to acquire a quantity x_j of the j-th good ($1 \leq j \leq n$), so as to maximize their utility $u(x)$, subject to the constraint that their disposable income is r. In other words, given the price π_j of the j-th good, they must solve the problem

$$\max u(x) \tag{1}$$

$$\text{s.t.} \qquad \pi^t x = r, \tag{2}$$

$$x \geq 0. \tag{3}$$

On the other hand, producers (also aggregated into one) face costs $c(y)$ in order to produce quantities y_j of each good ($1 \leq j \leq n$), and would like to maximize their returns, i.e., given prices π_j, their problem is

$$\max \pi^t y - c(y) \tag{4}$$

$$\text{s.t.} \qquad y \geq 0.$$

The system is in equilibrium when x, y and π are such that both optimization problems are simultaneously solved, subject to the market clearing equation, which in this case takes the particularly simple form

$$x = y. \tag{5}$$

Walras assumed that in real life, consumers solve their problem (1)–(3) given prices π, and so they determine the quantities $x = y$, in view of (5). Now given y,

producers fix certain new prices π' (possibly increasing π_j if y_j is big and decreasing it otherwise), so as to improve the value of their objective in (4). Next consumers modify their consumption basket, given π', solving (1)–(3) with the new prices, and opt to consume x' instead of x. Producers then modify π' to π'', and the process continues until (hopefully) an equilibrium is attained. Walras called this process "tatônnement". To decide whether this process indeed works in real life or not is beyond the author's humble economical knowledge; rather we are interested in this kind of process as a way to solve the equilibrium problem, and furthermore optimization problems. It is clear that, assuming continuous differentiability of u and c, the first order optimality conditions for the problems above, plus the market clearing equation, produce the following system in x, y, π and ω:

$$-\nabla u(x) + \omega\pi \geq 0, \tag{6}$$

$$x \geq 0, \tag{7}$$

$$\pi^t x = r, \tag{8}$$

$$x^t[-\nabla u(x) + \omega\pi] = 0, \tag{9}$$

$$\nabla c(y) - \pi \geq 0, \tag{10}$$

$$y \geq 0, \tag{11}$$

$$y^t[\nabla c(y) - \pi] = 0, \tag{12}$$

$$x = y, \tag{13}$$

where ω is the multiplier for the budget constraint (2).

The connection with optimization comes from a well known theorem by Samuelson ([45]), which ensures that the problem above can be cast as an optimization problem in the following way: let us call $x(\pi)$ the solution x of problem (1)–(3) as a function of π, and let $\pi(x)$ be the inverse function. Define $U : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ as

$$U(x, y) = -c(y) + \int_{x_0}^{x} \pi(\xi) d\xi.$$

and consider the optimization problem

$$\max U(x, y) \tag{14}$$

$$\text{s.t.} \qquad x \geq 0, \qquad y \geq 0, \qquad x = y. \tag{15}$$

Samuelson's theorem states that the equilibrium problem (6)–(13) and the optimization problem (14)–(15) are equivalent, in the sense that the first order conditions of (14)–(15) are equivalent to (6)–(13). As a consequence, both problems are indeed equivalent if u is concave and c is convex. For this very simple case, Samuelson's result is a matter of almost trivial verification.

We mention, parenthetically, that there is some redundancy above due to the too simple form of the market clearing equation: it's more realistic to assume that producers have different options in order to produce the final goods x, in which case y belongs to $\mathbb{R}^m$ $(m > n)$, giving raise to an input-output matrix $B \in \mathbb{R}^{n\times m}$. In such a case, $By = x$ substitutes for $x = y$ in (5), (13) and (15) and everything is slightly less trivial. In fact, Samuelson's result covers equilibrium models considerably more involved than this one.

Whether the problem is formulated as an equilibrium one or an optimization one, the idea of alternatingly adjusting quantities (solving an optimization problem given prices) and then prices (so that the "dual" objective increases; see (19) below), proved to be quite efficient to solve this type of equilibrium problems, and "tatônnement" methods are still being proposed, analyzed and used to solve real-life equilibrium problems (see e.g. [27]).

Here we will study a particular form of "tatônnement", namely augmented Lagrangian methods for convex optimization. Before introducing this family of methods we introduce some notation and basic convex optimization results.

From now on $\mathbb{R}^p_+ = \{z \in \mathbb{R}^p : z_\ell \geq 0 \ (1 \leq \ell \leq p)\}$, $\mathbb{R}^p_{++} = \{z \in \mathbb{R}^p : z_\ell > 0 \ (1 \leq \ell \leq p)\}$, $\langle \cdot, \cdot \rangle$ and $\|\cdot\|$ denote the Euclidean inner product and norm respectively. For $C \subset \mathbb{R}^p$, $\mathrm{int}(C)$ denotes the interior of C, e^ℓ is the ℓ-th vector in the canonical basis of $\mathbb{R}^p$ and ∇ and ∂ indicate the gradient and subdifferential of a convex function.

More precisely, for a convex $\varphi : \mathbb{R}^p \to \mathbb{R} \cup \{\infty\}$, the *subdifferential of* φ *at* z is defined as $\partial\varphi(z) = \{\xi \in \mathbb{R}^p : \langle \xi, z' - z\rangle \leq \varphi(z') - \varphi(z) \text{ for all } z' \in \mathbb{R}^p\}$. The elements of $\partial\varphi(z)$ are said to be *subgradients* of φ at z.

We deal with the convex optimization problem (P) defined as

$$\min f_0(x)$$

$$\text{s.t.} \quad f_i(x) \leq 0 \qquad (1 \leq i \leq m), \tag{16}$$

where $f_i : \mathbb{R}^n \to \mathbb{R}$ is convex and finite valued for $0 \leq i \leq m$. Note that the effective domain of the f_i's $(0 \leq i \leq m)$ is $\mathbb{R}^n$ and so they are continuous on $\mathbb{R}^n$ (e.g. [19], Vol. I, p. 175). The Lagrangian associated with (P) is $\mathcal{L} : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ defined as

$$\mathcal{L}(x, y) = f_0(x) + \sum_{i=1}^{m} y_i f_i(x), \tag{17}$$

and the dual problem associated with (P) is the convex optimization problem (D) given by

$$\min -\psi(y)$$

$$\text{s.t.} \quad y \geq 0, \tag{18}$$

with $\psi : \mathbb{R}^m \to \mathbb{R} \cup \{-\infty\}$ defined as

$$\psi(y) = \inf_{x \in \mathbb{R}^n} \mathcal{L}(x, y). \tag{19}$$

The set of solutions of (P) and(D) will be denoted as S_P^* and S_D^* respectively, and $S_P^* \times S_D^* \subset \mathbb{R}^n \times \mathbb{R}^m$ will be called S^*. We assume that some basic constraint qualification condition holds (e.g. the condition in [19], Vol. I, p. 307), so that S^* consists of the set of pairs (x, y) which satisfy the Karush-Kuhn-Tucker conditions for (P) (see [19], Vol. I, p. 305), namely

$$0 \in \partial f_0(x) + \sum_{i=1}^{m} y_i \partial f_i(x), \tag{20}$$

$$y \geq 0, \tag{21}$$

$$y_i f_i(x) = 0 \qquad (1 \leq i \leq m), \tag{22}$$

$$f_i(x) \leq 0 \qquad (1 \leq i \leq m). \tag{23}$$

It is easy to check that S^* coincides with the set of saddle points of $\mathcal{L}$, i.e. (x^*, y^*) belongs to S^* if and only if

$$\mathcal{L}(x^*, y) \leq \mathcal{L}(x^*, y^*) \leq \mathcal{L}(x, y^*),$$

for all $x \in \mathbb{R}^n$ and all $y \in \mathbb{R}^m_+$.

We comment that we have included the case of nonsmooth f_i's for the sake of wider generality, though this will cause some technical problems. Readers unfamiliar with nonsmooth convex analysis should replace ∂f_i by ∇f_i in all cases, and disregard specific nonsmooth discussions. For instance, in the smooth case no constraint qualification like the one mentioned above is required.

Following the "tatônnement" spirit, it seems reasonable to try a Lagrangian method consisting of, given a dual vector $y^k \geq 0$, determining x^{k+1} as a minimizer of $\mathcal{L}(\cdot, y^k)$, and then using somehow x^{k+1} in order to compute some $y^{k+1} \geq 0$ such that $\psi(y^{k+1}) > \psi(y^k)$. The problem lies in the nonnegativity constraints in (18). In other words, the Lagrangian, in order to be convex in x, must be defined as $-\infty$ when y is not nonnegative, i.e. it is not smooth in y at the boundary of $\mathbb{R}^m_+$. As a consequence,

a minimizer x^{k+1} of $\mathcal{L}(\cdot, y^k)$ does not provide in an easy way an increase direction for ψ. The solution consists of augmenting the Lagrangian, defining it over the whole $\mathbb{R}^m$, while keeping it convex in x. This augmentation, as we will see, will easily provide an increase direction for ψ, and even more, an appropriate closed formula for updating the sequence $\{y^k\}$. The price to be paid will be the loss of second differentiability of the augmented Lagrangian in x, even when the problem data (i.e. the functions f_i's $(0 \leq i \leq m)$) are as smooth as desired. Also, it is totally harmless and indeed convenient, to add a positive parameter multiplying the summation in the definition of $\mathcal{L}$.

We define then the augmented Lagrangian $\overline{\mathcal{L}} : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}_{++} \to \mathbb{R}$ as

$$\overline{\mathcal{L}}(x, y, \rho) = f_0(x) + \rho \sum_{i=1}^{m} \left[\left(\max\{0, y_i + (2\rho)^{-1} f_i(x)\} \right)^2 - y_i^2 \right]. \tag{24}$$

Now we can formally introduce the *augmented Lagrangian method* (AL from now on) for problems (P) and (D).

AL generates a sequence $\{(x^k, y^k)\} \subset \mathbb{R}^n \times \mathbb{R}^m$, starting from any $y^0 \in \mathbb{R}^m$, through the following iterative formulae:

$$x^{k+1} \in \operatorname{argmin}_{x \in \mathbb{R}^n} \overline{\mathcal{L}}(x, y^k, \lambda_k), \tag{25}$$

$$y_i^{k+1} = \max\left\{0, y_i^k + (2\lambda_k)^{-1} f_i(x^{k+1})\right\}, \tag{26}$$

where $\{\lambda_k\} \subset [\underline{\lambda}, \overline{\lambda}]$ for some $\overline{\lambda} \geq \underline{\lambda} > 0$.

Of course, one expects the sequence $\{(x^k, y^k)\}$ generated by (25)–(26) to converge to a point $(x^*, y^*) \in S^*$. There are some caveats, however. In the first place, the augmented Lagrangian $\overline{\mathcal{L}}$ is convex in x, but in principle $\overline{\mathcal{L}}(\cdot, y^k, \lambda^k)$ may fail to attain its minimum, in which case x^{k+1} is not defined. We remark that this situation may happen even if problems (P) and (D) have solutions. For instance, consider the problem of minimizing a constant function of one variable over the the halfline $\{x \in \mathbb{R} : e^x \leq 1\}$. Clearly any nonpositive real is a primal solution, and it is immediate that any pair $(x, 0)$ with $x \leq 0$ satisfies (20)-(23), so that 0 is a dual solution, but taking $y = \rho = 1$ and the constant value of the function also equal to 1, we get $\overline{\mathcal{L}}(x, y, \rho) = 1 + \max\{0, 1 + e^x - 1\}^2 - 1 = e^{2x}$, which obviously does not attains its minimum. Even if $\overline{\mathcal{L}}(\cdot, y^k, \lambda_k)$ has minimizers, they might be multiple, because, $\overline{\mathcal{L}}$ though convex in x, is not strictly convex. Thus, there is no way in principle to ensure that the sequence $\{x^k\}$ will be bounded. To give a trivial example, if none of the f_i's $(0 \leq i \leq m)$ depends upon x_n, we may choose the last component of x^{k+1} arbitrarily (e.g. $x_n^{k+1} = k$), making $\{x^k\}$ unbounded. As a consequence, all convergence results on the sequence $\{x^k\}$ will have to be stated under the assumption that such a sequence exists and is bounded (later on we will see a variant of the method which automatically ensures existence and boundedness of $\{x^k\}$). On the other hand, we will establish that the whole sequence $\{y^k\}$ converges to a point in S_D^* under the sole

condition of existence of solutions of (P)–(D). Of course, existence and boundedness of $\{x^k\}$ can be ensured by imposing additional conditions on the problem data, like coerciveness of f_0, for instance (meaning that its level sets are bounded) or of any of the constraint functions f_i $(1 \leq i \leq m)$, in which case the feasible set for problem (P) is necessarily bounded.

In order to provide a rationale for the method, we advance the facts, to be proved later on, that y^{k+1} has been chosen so that it is nonnegative (this is immediate) and that x^{k+1} minimizes $\mathcal{L}(\cdot, y^{k+1})$ (this is not so immediate). As a consequence, we have $0 \in \partial f_0(x^k) + \sum_{i=1}^m y_i^k \partial f_i(x^k)$ and $y^k \geq 0$ for all k. We may say that (20) and (21) are satisfied for all k. On the other hand, as we will see, (22) and (23) will hold only at cluster points of $\{(x^k, y^k)\}$. In this sense this sequence is dual feasible and satisfies the Lagrangian condition at all iterates, but it is primal infeasible and fails to satisfy complementarity along the iterates.

Augmented Lagrangian method fos solving equality constrained nonlinear optimization problems (nonconvex in general), were introduced in [18] and [39]. The first method of this kind for inequality constrained problems appeared in [8], and was furtherly developed in, e.g., [2] and [29]. Our presentation follows the formulation in [41] and [42]. A deep analysis of AL and its convergence properties can be found in [3]. Here we will follow an alternative procedure, consistent of reducing AL to a particular case of the proximal point method, which we describe next. This approach leads to a considerably simpler analysis and also gives slightly more robust results.

3 The Classical Proximal Point Method

The classical *proximal point method* (PP from now on) can be seen as an algorithm for finding a zero of a maximal monotone operator $T : \mathbb{R}^p \to \mathcal{P}(\mathbb{R}^p)$.

We recall that T is *monotone* if $\langle v - v', z - z' \rangle \geq 0$ for all $v \in T(z)$, all $v' \in T(z')$, and all z, $z' \in \mathbb{R}^p$. T is *maximal monotone* if it is monotone, and furthermore for all monotone operator T' such that $T(z) \subset T'(z)$ for all $z \in \mathbb{R}^p$, it holds that $T = T'$. A zero of T is a point $z \in \mathbb{R}^p$ such that $0 \in T(z)$.

Starting from any $z^0 \in \mathbb{R}^p$, PP generates a sequence $\{z^k\} \subset \mathbb{R}^p$ through the iterative step

$$0 \in \widetilde{T}_k(z^{k+1}), \tag{27}$$

with

$$\widetilde{T}_k(z) = T(z) + \lambda_k(z - z^k), \tag{28}$$

where $\{\lambda_k\}$ is a bounded sequence of positive real numbers. When $T = \partial\varphi$ for some convex $\varphi : \mathbb{R}^m \to \mathbb{R} \cup \{\infty\}$, the problem of finding a zero of T is equivalent to $\min \varphi(y)$

and the iterative formulae (27), (28) become

$$y^{k+1} = \operatorname{argmin}\left\{\varphi(y) + (\lambda_k/2)\left\|y - y^k\right\|^2\right\}.$$

It is not easy to determine the exact origin of PP. Possibly, it started with works by members of Thikonov's school dealing with regularization methods for solving ill-posed problems, and one of the earliest relevant references is [30]. It received its current name in the 60's, through the works of Moreau, Yoshida and Martinet, among others (see [36], [33], [34]) and attained the form of (27)–(28) in Rockafellar's works in the 70's ([43], [44]). A more recent survey on the proximal point method can be found in [31].

We will give next a convergence proof for PP, different to some extent from the proof in [43], which uses the concept of firm nonexpansiveness. Here we follow an approach which can also be applied to the generalized PP to be discussed in forthcoming sections.

We need first a classical result, due to Minty. We recall that given $Q : \mathbb{R}^p \to \mathcal{P}(\mathbb{R}^p)$, the inverse operator $Q^{-1} : \mathbb{R}^p \to \mathcal{P}(\mathbb{R}^p)$ is always defined through the relation $w \in Q^{-1}(v)$ if and only if $v \in Q(w)$.

Theorem 1. *If $Q : \mathbb{R}^p \to \mathcal{P}(\mathbb{R}^p)$ is maximal monotone then $I + Q$ is onto (i.e. for all $v \in \mathbb{R}^p$ there exists $w \in \mathbb{R}^p$ such that $v \in Q(w)$) and $(I + Q)^{-1}$ is point-to-point.*

Proof. See [35]. □

Minty's theorem is needed in the following convergence theorem for PP.

Theorem 2. *If the maximal monotone operator T has zeroes, then the sequence $\{z^k\}$ generated by PP (i.e. by (27)–(28)) converges to a zero of T.*

Proof. First we must establish that the sequence $\{z^k\}$ is well defined. Assuming inductively that z^k is well defined, we observe that (27)–(28) are equivalent to saying that z^{k+1} is such that $z^k \in (I + \lambda_k^{-1}T)(z^{k+1})$. Such a z^{k+1} exists by Minty's Theorem applied to the maximal monotone operator $\lambda_k^{-1}T$, and is unique because $(I + \lambda_k^{-1}T)^{-1}$ is point-to-point.

Next, we take any zero z^* of T, which exists by assumption, and claim that the following relation holds for all $k \geq 0$:

$$0 \leq 2\langle z^* - z^{k+1}, z^{k+1} - z^k\rangle = \left\|z^* - z^k\right\|^2 - \left\|z^* - z^{k+1}\right\|^2 - \left\|z^{k+1} - z^k\right\|^2. \quad (29)$$

The equality in (29) is simple algebra and the inequality follows from the facts that (27)–(28) can be written as $z^k - z^{k+1} \in \lambda_k^{-1} T(z^{k+1})$ and that $0 \in \lambda_k^{-1} T(z^*)$. Therefore we have $\langle z^* - z^{k+1}, z^{k+1} - z^k \rangle = \langle z^* - z^{k+1}, 0 - (z^k - z^{k+1}) \rangle \geq 0$, using monotonicity of $\lambda_k^{-1} T$ in the rightmost equality.

By (29), $\left\| z^* - z^{k+1} \right\| \leq \left\| z^* - z^k \right\|$, so that $\{\left\| z^* - z^k \right\|\}$ is nonincreasing, and henceforth convergent, since it is nonnegative. As a consequence $\left\| z^* - z^k \right\| \leq \left\| z^* - z^0 \right\|$ for all k, and thus $\{z^k\}$ is bounded. (29) also gives

$$0 \leq \left\| z^{k+1} - z^k \right\|^2 \leq \left\| z^* - z^k \right\|^2 - \left\| z^* - z^{k+1} \right\|^2 . \tag{30}$$

Since the righmost expression in (30) converges to 0 as k goes to ∞ because $\{\left\| z^* - z^k \right\|\}$ is convergent, it follows that

$$\lim_{k \to \infty} \left\| z^{k+1} - z^k \right\| = 0. \tag{31}$$

We mention that maximality of T easily implies that the graph of T is closed, i.e. that if $(w^k, v^k) \in \mathbb{R}^p \times \mathbb{R}^p$ satisfies $v^k \in T(w^k)$ for all k, and $\lim_{k \to \infty} (w^k, v^k) = (\bar{w}, \bar{v}) \in \mathbb{R}^p \times \mathbb{R}^p$, then $\bar{v} \in T(\bar{w})$. This property is also called *upper semicontinuity* of T. Next we observe that (27)–(28) can also be rewritten as

$$\lambda_k (z^k - z^{k+1}) \in T(z^{k+1}). \tag{32}$$

Let $\bar{z}$ be any cluster point of $\{z^k\}$ (which exists because $\{z^k\}$ is bounded). Taking limits in (32) as k goes to ∞ along an appropriate subsequence and using (31), we get $0 \in T(\bar{z})$, because $\{\lambda_k\}$ is bounded, so that $\bar{z}$ is a zero of T. As a consequence $\{\left\| \bar{z} - z^k \right\|\}$ is nonincreasing. Since it has a subsequence which converges to 0, it follows that the whole sequence converges to 0, i.e. $\bar{z} = \lim_{k \to \infty} z^k$. □

4 The Connection between PP and AL

We will prove here that the sequences $\{y^k\}$ generated by AL applied to problem (P) and by PP applied to problem (D) are essentially the same. This result appeared for the first time in [42]. The convergence analysis of AL will then be an easy consequence of Theorems 2 and 3. In order to apply PP to problem (D) we define $\bar{\psi} : \mathbb{R}^m \to \mathbb{R} \cup \{\infty\}$ as $\bar{\psi}(y) = -\psi(y)$ if $y \geq 0$, $\bar{\psi}(y) = +\infty$ otherwise, and take $T : \mathbb{R}^m \to \mathcal{P}(\mathbb{R}^m)$ defined as $T(y) = \partial \bar{\psi}(y)$.

Theorem 3. *Let $\{y^k\}$ be the sequence generated by PP applied to problem (D) (i.e. by (27)–(28) with $T = \partial \bar{\psi}$) and $\{(\widehat{x}^k, \widehat{y}^k)\}$ the sequence generated by AL (i.e. by (25)–(26)). If $\widehat{y}^0 = y^0$ then $\widehat{y}^k = y^k$ for all $k \geq 0$.*

Proof. We proceed by induction. Assume that $y^k = \widehat{y}^k$. First we claim that $\widehat{x}^{k+1}$ minimizes $\mathcal{L}(\cdot, \widehat{y}^{k+1})$ with $\mathcal{L}$ as in (17). Observe that (26) guarantees that $\{\widehat{y}^k\} \subset \mathbb{R}^m_+$. Define $s_i^k(x) = \lambda_k \left[\left(\max\{0, \widehat{y}_i^k + (2\lambda_k)^{-1} f_i(x)\} \right)^2 - (y_i^k)^2 \right]$. Then we have

$$0 \in \partial_x \overline{\mathcal{L}}(\widehat{x}^{k+1}, \widehat{y}^k, \lambda_k) = \partial f_0(\widehat{x}^{k+1}) + \sum_{i=1}^m \partial s_i^k(\widehat{x}^{k+1}) =$$

$$\partial f_0(\widehat{x}^{k+1}) + \sum_{i=1}^m \max\left\{0, \widehat{y}_i^k + (2\lambda_k)^{-1} f_i(x^{k+1})\right\} \partial f_i(\widehat{x}^{k+1}) =$$

$$\partial f_0(\widehat{x}^{k+1}) + \sum_{i=1}^m \widehat{y}^{k+1} \partial f_i(\widehat{x}^{k+1}) = \partial_x \mathcal{L}(\widehat{x}^{k+1}, \widehat{y}^{k+1}), \tag{33}$$

using (25) in the inclusion, (24), linearity of the subdifferential (e.g. [19], Vol. I, p. 261) and definition of s_i^k in the first equality, the chain rule of subdifferential calculus (e.g. [19], Vol. I, p. 264) and some elementary calculus to differentiate s_i^k in the second equality, (26) in the third equality and (17) in the fourth one. Since $\mathcal{L}(\cdot, \widehat{y}^{k+1})$ is convex, because $\widehat{y}^{k+1} \geq 0$, we conclude from (33) that the claim is established.

Thus, by (17) and (19),

$$\psi(\widehat{y}^{k+1}) = \mathcal{L}(\widehat{x}^{k+1}, \widehat{y}^{k+1}) = f_0(\widehat{x}^{k+1}) + \sum_{i=1}^m \widehat{y}_i^{k+1} f_i(\widehat{x}^{k+1}). \tag{34}$$

By (19), (34) and (17), for all $y \in \mathbb{R}^m$,

$$\psi(\widehat{y}^{k+1}) - \psi(y) \geq \psi(\widehat{y}^{k+1}) - \mathcal{L}(\widehat{x}^{k+1}, y) = \sum_{i=1}^m (\widehat{y}_i^{k+1} - y_i) f_i(\widehat{x}^{k+1}). \tag{35}$$

Next observe that (26) implies

$$\widehat{y}_i^{k+1} - \widehat{y}_i^k = \max\left\{ -\widehat{y}_i^k, (2\lambda_k)^{-1} f_i(\widehat{x}^{k+1}) \right\} \geq (2\lambda_k)^{-1} f_i(\widehat{x}^{k+1}), \tag{36}$$

which in turn implies

$$\widehat{y}_i^{k+1}(\widehat{y}_i^{k+1} - \widehat{y}_i^k) = (2\lambda_k)^{-1} \widehat{y}_i^{k+1} f_i(\widehat{x}^{k+1}). \tag{37}$$

By (36), (37), for all $y \in \mathbb{R}^m_+$,

$$2\lambda_k(\widehat{y}_i^{k+1} - \widehat{y}_i^k)(\widehat{y}_i^{k+1} - y_i) = \widehat{y}_i^{k+1} f_i(\widehat{x}^{k+1}) - 2\lambda_k(\widehat{y}_i^{k+1} - \widehat{y}_i^k) y_i \leq$$

$$\widehat{y}_i^{k+1} f_i(\widehat{x}^{k+1}) - y_i f_i(\widehat{x}^{k+1}). \tag{38}$$

Combining (38) and (35)

$$\psi(\widehat{y}^{k+1}) - \psi(y) \geq 2\lambda_k \langle \widehat{y}^{k+1} - y, \widehat{y}^k - \widehat{y}_i^{k+1} \rangle, \tag{39}$$

for all $y \in \mathbb{R}^m_+$. In view of the definition of $\bar{\psi}$, (39) and the inductive assumption imply that

$$2\lambda_k (y^k - \widehat{y}^{k+1}) = 2\lambda_k (\widehat{y}^k - \widehat{y}^{k+1}) \in \partial \bar{\psi}(\widehat{y}^{k+1}). \tag{40}$$

On the other hand, (27) and (28) imply, in view of Minty's Theorem, that y^{k+1} is the only vector satisfying

$$2\lambda_k (y^k - y^{k+1}) \in \partial \bar{\psi}(y^{k+1}). \tag{41}$$

Thus, we get from (40) and (41) that $\widehat{y}^{k+1} = y^{k+1}$, completing the induction step. □

With the help of Theorem 3, we get the following convergence result for AL.

Theorem 4. *Assume that problems (P) and (D) have solutions and that x^{k+1} as defined by (25) exists for all k. If $\{(x^k, y^k)\}$ is the sequence generated by AL (i.e. by (25)–(26)), then*

i) The sequence $\{y^k\}$ converges, as k goes to ∞, to a solution y^* of problem (D).

ii) All cluster points of $\{x^k\}$ (if any) are solutions of problem (P).

Proof. (i) follows directly from Theorems 2 and 3. We proceed to prove (ii). Let $\bar{x}$ be a cluster point of $\{x^k\}$. It suffices to prove that $(y^*, \bar{x})$ satisfy (20)–(23). Consider the operator $G : \mathbb{R}^n \times \mathbb{R}^m \to \mathcal{P}(\mathbb{R}^n \times \mathbb{R}^m)$ defined as $G = (\partial_x \mathcal{L}, 0)$. It is immediate that G is maximal monotone. It follows from (33) that

$$(0, 0) \in G(x^k, y^k) \tag{42}$$

for all k. Since G is upper semicontinuous, taking limits in (42) as k goes to ∞ along an appropriate subsequence, we get that $(0, 0) \in G(\bar{x}, y^*)$. In particular $0 \in \partial_x \mathcal{L}(\bar{x}, y^*)$, i.e. (20) holds. We already know that (21) holds by Theorem 2. We claim that (22) and (23) follow from (26). Taking limits in (26) as k goes to ∞ along an appropriate subsequence, we get

$$y_i^* = \max\{0, y_i^* + \widehat{\lambda} f_i(\bar{x})\}, \tag{43}$$

where $\widehat{\lambda} \in [\underline{\lambda}, \overline{\lambda}]$ is some cluster point of $\{\lambda_k\}$. Since $\underline{\lambda} > 0$, it follows easily from (43) that $f_i(\bar{x}) \leq 0$, i.e. (23) holds. Also, if $y_i^* > 0$, we get from (43) that $y_i^* = y_i^* + \widehat{\lambda} f_i(\bar{x})$, implying that $f_i(\bar{x}) = 0$, so that (22) also holds. □

One motivation for considering variants of AL lies in the fact that $\overline{\mathcal{L}}$, as defined by (24) is not twice differentiable in x even when the data functions f_i $(0 \leq i \leq n)$ are suficiently smooth. Another motivation is the lack of regularity of the sequence $\{x^k\}$, which in principle can be divergent. We will overcome the first limitation by introducing two kinds of generalized distances, called Bregman distances and ϕ-divergences, which we will do in the next two sections. In order to eliminate the second limitation, we will look at problems (P) and (D) in the form of variational inequality problems, to be defined in Section 7. This approach has also been developed in [42] within the classical framework, i.e. without generalized distances, as discussed in Section 8.

5 Bregman Functions and Distances

Take $C \subset \mathbb{R}^p$ closed, convex and with nonempty interior, and consider $c : C \to \mathbb{R}$, differentiable in $\mathrm{int}(C)$. Define $D_c : C \times \mathrm{int}(C) \to \mathbb{R}$ as

$$D_c(z, z') = c(z) - c(z') - \langle \nabla c(z'), z - z' \rangle. \tag{44}$$

The function c is said to be a *Bregman function with zone* C, and D_c the *Bregman distance* associated with c, if the following conditions hold:

B1. c is continuously differentiable on $\mathrm{int}(C)$.

B2. c is strictly convex and continuous on C.

B3. For all $\gamma \in \mathbb{R}$ and all $z \in C$ the partial level sets $\Gamma(z, \gamma) = \{w \in \mathrm{int}(C) : D_c(z, w) \leq \gamma\}$ are bounded.

B4. If $\{z^k\} \subset \mathrm{int}(C)$ and $\lim_{k\to\infty} z^k = \widetilde{z}$ then $\lim_{k\to\infty} D_c(\widetilde{z}, z^k) = 0$.

B5. If $\{w^k\} \subset C$ and $\{z^k\} \subset \mathrm{int}(C)$ are sequences such that $\{w^k\}$ is bounded, $\lim_{k\to\infty} z^k = \widetilde{z}$ and $\lim_{k\to\infty} D_c(w^k, z^k) = 0$ then $\lim_{k\to\infty} w^k = \widetilde{z}$.

B6. If $\{z^k\} \subset C$ is such that $\lim_{k\to\infty} z^k = \widetilde{z}$ and $\widetilde{z}$ belongs to the boundary of C, then $\lim_{k\to\infty} D_c(w, z^k) = \infty$ for all $w \in \mathrm{int}(C)$.

It follows easily from (44) and B1–B2 that $D_c(z, z') \geq 0$ for all $z \in C$, $z' \in \mathrm{int}(C)$, and that $D_c(z, z') = 0$ if and only if $z = z'$. However, D_c in general is not symmetric and it does not satisfy the triangular inequality. Bregman functions were introduced in [4], only with conditions B1–B5 above. B6 was introduced in [20], where it is called *boundary coerciveness*. Conditions B4–B6 hold automatically, as a consequence of B1–B3, when $C = \mathbb{R}^p$. For our applications to augmented Lagrangian methods, we are interested only in the case in which C is the whole space or an orthant, so that we present next some relevant examples of Bregman functions for these cases. Examples of Bregman functions for other sets C, e.g. balls, boxes or polyhedra with nonempty interior, can be found in [9].

Example 1: $C = \mathbb{R}^p$, $c(z) = \|z\|^2$, in which case $D_c(z, z') = \|z - z'\|^2$. More generally $c(z) = z^t M z$ with $M \in \mathbb{R}^{p \times p}$ symmetric and positive definite, in which case $D_c(z, z') = (z - z')^t M (z - z')$.

Example 2: $C = \mathbb{R}^p_+$, $c(z) = \sum_{\ell=1}^p z_\ell \log z_\ell$, continuously extended to the boundary of $\mathbb{R}^p_+$ with the convention that $0 \log 0 = 0$. In this case

$$D_c(z, z') = \sum_{\ell=1}^p [z_\ell \log(z_\ell / z'_\ell) + z'_\ell - z_\ell],$$

which is called the Kullback-Leibler distance, widely used in statistics (see [32]).

Example 3: $C = \mathbb{R}^p_+$, $c(z) = \sum_{\ell=1}^p (z_\ell^\alpha - z_\ell^\beta)$ with $\alpha \geq 1$, $\beta \in (0, 1)$. For $\alpha = 2$, $\beta = 1/2$ we get

$$D_c(z, z') = \|z - z'\|^2 + (1/2) \sum_{\ell=1}^p \frac{\left(\sqrt{z_\ell} - \sqrt{z'_\ell}\right)^2}{\sqrt{z'_\ell}};$$

for $\alpha = 1$, $\beta = 1/2$ we have

$$D_c(z, z') = (1/2) \sum_{\ell=1}^p \frac{\left(\sqrt{z_\ell} - \sqrt{z'_\ell}\right)^2}{\sqrt{z'_\ell}}.$$

6 ϕ-divergences

In this section we discuss another class of "distances", which will be denoted as $d_\phi(\cdot, \cdot)$, defined on the positive orthant of $\mathbb{R}^p$. Take $\phi \colon \mathbb{R}_{++} \to \mathbb{R}$, convex and thrice continuously differentiable, satisfying

$$\phi(1) = \phi'(1) = 0, \quad \phi''(1) > 0, \quad \lim_{t \to 0^+} \phi'(t) = -\infty. \tag{45}$$

If ϕ satisfies (45) then $d_\phi \colon \mathbb{R}^p_{++} \times \mathbb{R}^p_{++} \to \mathbb{R}$, defined by

$$d_\phi(z, z') = \sum_{\ell=1}^p z'_\ell \phi\left(z_\ell / z'_\ell\right)$$

is said to be a *ϕ-divergence*. The next properties easily follow from (45) and the definition of ϕ-divergences.

Proposition 1.

i) $d_\phi(z, z') \geq 0$ *for all* z, $z' \in \mathbb{R}^p_{++}$,

ii) $d_\phi(z,z')=0$ *iff* $z=z'$,

iii) the level sets of $d_\phi(\cdot,z')$ *are bounded for all* $z'\in\mathbb{R}^p_{++}$,

iv) the level sets of $d_\phi(z,\cdot)$ *are bounded for all* $z\in\mathbb{R}^p_{++}$,

v) $d_\phi(z,z')$ *is jointly convex in* z, z', *and strictly convex in* z,

vi) $\lim_{k\to\infty} d_\phi(z,z^k)=0$ *iff* $\lim_{k\to\infty} z^k=z$.

Proof. Elementary. □

We present next some relevant examples of ϕ-divergences.

Example 4: $\phi_1(t)=t\log t-t+1$. Then

$$d_{\phi_1}(z,z')=\sum_{\ell=1}^{p}[z_\ell\log(z_\ell/z'_\ell)+z'_\ell-z_\ell],$$

i.e. d_{ϕ_1} is the Kullback-Leibler distance of Example 2 and can therefore be extended to $\mathbb{R}^p_+\times\mathbb{R}^p_{++}$. Up to additive linear terms in c and multiplicative constants in ϕ, the pair (ϕ_1,c_1) with $c_1(x)=\sum_{\ell=1}^p z_\ell\log z_\ell$ is the only pair (ϕ,c) such that $d_\phi=D_c$.

Example 5: $\phi_2(t)=t-\log\ t-1$. Then

$$d_{\phi_2}(z,z')=d_{\phi_1}(z,z').$$

Example 6: $\phi_3(t)=(\sqrt{t}-1)^2$. Then

$$d_{\phi_3}(z,z')=\sum_{\ell=1}^{p}\left(\sqrt{z_\ell}-\sqrt{z'_\ell}\right)^2.$$

ϕ-divergences were introduced in [46], and have been recently extended in [1] to other open polyhedra besides $\mathbb{R}^n_{++}$. We do not discuss this extension in the sequel, since it is not of interest for our application to augmented Lagrangians. For convergence of the proximal point method with ϕ-divergences, an additional condition on ϕ is required, namely

$$\phi'(t)\le\phi''(1)\log t, \tag{46}$$

for all $t\in\mathbb{R}_{++}$. The class of ϕ-divergences satisfying (46) is called Φ_4 in [23], where it was introduced. The ϕ-divergences of Examples 4–6 belong to Φ_4.

7 Variational Inequality Problems

Given a maximal monotone point-to-set operator $T : \mathbb{R}^p \to \mathcal{P}(\mathbb{R}^p)$ and a closed and convex set $C \subset \mathbb{R}^p$, the variational inequality problem VIP(T,C) consists of finding $z^* \in C$ such that there exists $u^* \in T(z^*)$ satisfying

$$\langle u^*, z - z^* \rangle \geq 0 \tag{47}$$

for all $z \in C$. The set of solutions of VIP(T,C), which is closed and convex (see e.g. [17]), will be denoted as $S(T, C)$. It can be easily verified that when $T = \partial\varphi$ for some convex $\varphi : \mathbb{R}^p \to \mathbb{R} \cup \{\infty\}$, VIP($T$,$C$) reduces to min$\varphi(z)$ subject to $z \in C$. When $C = \mathbb{R}^p$, VIP(T,C) reduces to the problem of finding a zero of T. When $C = \mathbb{R}^p_+$, VIP(T,C) becomes the *nonlinear complementarity problem*, consisting of finding $z \in \mathbb{R}^p_+$ such that there exists $u \in \mathbb{R}^p_+ \cap T(z)$ satisfying $u^t z = 0$.

If we define $I_C : \mathbb{R}^p \to \mathbb{R} \cup \{\infty\}$ as the *indicator function* of C, i.e. $I_C(z) = 0$ if $z \in C$, $I_C(z) = +\infty$ otherwise, and then $N_C : \mathbb{R}^p \to \mathcal{P}(\mathbb{R}^p)$ as the *normalizing operator* of C, i.e. $N_C = \partial I_C$, then VIP(T,C) becomes the problem of finding a zero of the maximal monotone operator $T + N_C$. Thus, PP can be used to solve VIP(T,C), by applying it to $T + N_C$. The inconvenience is that the subproblems consist now of finding a zero of $\widetilde{T}_k + N_C$ with $\widetilde{T}_k$ as in (28), i.e. of solving VIP($\widetilde{T}_k$,C). In other words, the constraint $z \in C$ remains in the subproblems, which are in principle as difficult (though in general better conditioned, thanks to the additional regularization term in (28)) as the original problem VIP(T,C). For the case in which C has nonempty interior, we will use in the following sections Bregman distances and ϕ-divergences to construct generalized proximal point methods, where the regularization term plays also a penalization role, forcing the generated sequence $\{z^k\}$ to remain in the interior of C, so that the subproblems are "authentically" unconstrained, i.e. that the constraint $z \in C$ becomes superfluous. These generalized PP's will give raise in turn to generalized augmented Lagrangian methods where the augmented Lagrangian is as smooth in x as the data functions f_i's.

We close this section by establishing the connection between problems (P)–(D) and VIP(T,C). If we take $s = n + m$, $C = \mathbb{R}^n \times \mathbb{R}^m_+$ and

$$T(x, y) = \big(\partial_x \mathcal{L}(x, y), -\partial_y \mathcal{L}(x, y)\big) = \left(\partial f_0(x) + \sum_{i=1}^{m} y_i \partial f_i(x), -f(x)\right), \tag{48}$$

with $f(x) = (f_1(x), \ldots, f_m(x))$, then it can be easily verified that T is maximal monotone and that $S(T, C) = S^* = S^*_P \times S^*_D$. The rightmost equality in (48) follows from [19], Vol. I, p. 261.

8 Generalized Proximal Point Methods with Bregman Functions

We introduce now the *generalized proximal point method with Bregman functions* (GPPB from now on) for solving VIP(T,C). Consider a Bregman function with zone

C and assume that $\mathrm{int}(C)$ is nonempty. Starting from $z^0 \in \mathrm{int}(C)$, GPPB generates a sequence $\{z^k\}$ through the iterative formula

$$0 \in T_k(z^{k+1}),$$

where $T_k : \mathbb{R}^p \to \mathcal{P}(\mathbb{R}^p)$ is defined as

$$T_k(z) = T(z) + \lambda_k[\nabla c(z) - \nabla c(z^k)],$$

with λ_k as in PP. In this case, there is no need to modify T by adding the normalizing operator N_C, because divergence of ∇c at the boundary of C guarantees that the whole sequence $\{z^k\}$ is contained in the interior of C. This is more evident if we consider the case of $T = \partial\varphi$ with a convex $\varphi : \mathbb{R}^m \to \mathbb{R}$. If the problem of interest is $\min \varphi(x)$ subject to $x \in E$, and h is a Bregman function with zone $E \subset \mathbb{R}^m$, then the iteration of GPPB becomes

$$y^{k+1} = \mathrm{argmin}_{y \in \mathbb{R}^m}\{\varphi(y) + \lambda_k D_h(y, y^k)\}, \tag{49}$$

while the subproblems of PP for the same problem, after adding the normalizing operator N_C to the operator $\partial\varphi$, become

$$y^{k+1} = \mathrm{argmin}_{y \in E}\{\varphi(y) + (\lambda_k/2)\left\|y - y^k\right\|^2\}. \tag{50}$$

Note that the subproblems given by (49) are unconstrained, while the subproblems given by (50) are subject to the constraints $y \in E$. In the case of GPPB, these constraints are taken care of by D_h which, besides its regularization role, as in PP, has also a penalization effect. This is one advantage of GPPB over PP. Another one will be discussed when we consider the generalized augmented Lagrangians related to GPPB.

GPPB can be traced back to [15] and [16], which considered methods related to GPPB with the Bregman function of Example 2 applied to linear programming. The next step was [14], which considered GPPB with the same Bregman function applied to general convex optimization problems. Relevant works on GPPB in its current formulation include [13], which considers GPPB for finding zeroes of monotone operators (or VIP(T,C) with solutions in the interior of C) and [10], [11], [20], [28], which study GPPB for the convex optimization problem under progressively weaker assumptions on the problem data or the Bregman function. GPPB for variational inequality problems has been analyzed in [5].

GPPB can be applied for solving problems (P) and (D) in two different ways. One is to apply it directly to problem (D), using (49) with $\varphi = -\psi$. We will prove that the sequence $\{y^k\}$ generated in this way coincides with the sequence $\{y^k\}$ generated by a generalized augmented Lagrangian method (to be called GALB), which has the same structure of AL, in the sense that x^{k+1} minimizes a generalized augmented Lagrangian $\widehat{\mathcal{L}}$ evaluated at y^k and then y^k is updated through a closed formula. An advantage of GALB over AL is that $\widehat{\mathcal{L}}$ is as many times differentiable as a function of x as the problem data and the Bregman function are, while the augmented Lagrangian $\overline{\mathcal{L}}$ of AL is not twice differentiable. Thus, fast second order methods, like

Newton's one, can be used to minimize $\widehat{\mathcal{L}}$. Each choice of the Bregman function h in (49) generates a different GALB method. For $h(y) = \sum_{i=1}^{m} y_i \log y_i$ we recover the exponential multipliers method (see [3]).

The second option is to apply GPPB to VIP(T,$\mathbb{R}^n \times \mathbb{R}^m_+$), with T as in (48). This version of GPPB turns out to be equivalent to a generalized doubly augmented Lagrangian method (to be called GDALB), with a Lagrangian $\widetilde{\mathcal{L}}$, equal to $\widehat{\mathcal{L}}$ plus an additional regularization term in x, which ensures existence and uniqueness of x^k for all k and boundedness of the sequence $\{x^k\}$.

We introduce next the methods resulting from applying GPPB to problems (P)–(D), according to both alternatives. In Section 10 we will discuss the convergence properties of GPPB.

Given a sequence $\{\lambda_k\} \subset [\underline{\lambda}, \overline{\lambda}]$ for some $\overline{\lambda} \geq \underline{\lambda} > 0$, the generalized proximal point method with Bregman distances (GPPB) applied to problem (D), with a Bregman function h whose zone is $\mathbb{R}^m_+$, generates a sequence $\{y^k\}$ given by

$$y^0 \in \mathbb{R}^m_{++}, \tag{51}$$

$$y^{k+1} = \operatorname{argmin}\{-\psi(y) + \lambda_k D_h(y, y^k)\}, \tag{52}$$

with ψ as in (20). We will assume that h is *separable*, i.e. $h(y) = \sum_{i=1}^{m} h_i(y_i)$ with $h_i : \mathbb{R}_+ \to \mathbb{R}$. By B1–B2, all the h_i's are convex and continuous in $\mathbb{R}_+$, continuously differentiable in $\mathbb{R}_{++}$, and condition B6 can be written in terms of the h_i's as

B6'. $\lim_{t \to 0^+} h_i'(t) = -\infty$ for $1 \leq i \leq m$.

Assuming that (D) has solutions, it has been proved in Theorem 4.1 of [20] that y^{k+1} as defined by (52) exists, is unique, belongs to $\mathbb{R}^m_{++}$ and is the only solution of the first order optimality condition for (52), i.e.

$$\lambda_k \left[\nabla h(y^k) - \nabla h(y^{k+1})\right] \in \partial(-\psi(y^{k+1})). \tag{53}$$

Now we look at GPPB applied to VIP(T,$\mathbb{R}^n \times \mathbb{R}^m_+$), with T as in (48). We take a sequence $\{\lambda_k\}$ as above, a Bregman function g with zone $\mathbb{R}^n$ and a separable Bregman function $h(y) = \sum_{i=1}^{m} h_i(y_i)$ with zone $\mathbb{R}^m_+$. Note that g needs to satisfy only conditions B1–B3. We define $c(z) = g(x) + h(y)$. It is immediate that c is a Bregman function with zone $\mathbb{R}^n \times \mathbb{R}^m_+$.

GPPB applied to this problem generates a sequence $\{z^k\} = \{(x^k, y^k)\}$ given by

$$x^0 \in \mathbb{R}^n, \;\; y^0 \in \mathbb{R}^m_{++}, \tag{54}$$

$$0 \in T(z^{k+1}) + \lambda_k \left[\nabla c(z^{k+1}) - \nabla c(z^k)\right]. \tag{55}$$

In view of (48) and the definitions of c and h, (55) can be rewritten as

$$\lambda_k \big[\nabla g(x^k) - \nabla g(x^{k+1})\big] \in \partial f_0(x^{k+1}) + \sum_{i=1}^{m} y_i^{k+1} \partial f_i(x^{k+1}), \tag{56}$$

$$f_i(x^{k+1}) = \lambda_k \big[h_i'(y_i^{k+1}) - h_i'(y_i^k)\big] \qquad (1 \le i \le m). \tag{57}$$

Next we introduce the generalized augmented Lagrangian method (GALB) for problems (P) and (D). We remind that, given a convex $\varphi : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$, the convex conjugate $\varphi^* : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ is defined as $\varphi^*(t) = \sup_{\tau \in \mathbb{R}} \{\tau t - \varphi(\tau)\}$. It is well known that, when they exist, the derivatives of φ and φ^* are mutual inverses (e.g. [19], V. II, p. 48). Take $\{\lambda_k\}$ as above and a separable Bregman function $h(y) = \sum_{i=1}^{m} h_i(y_i)$ with zone $\mathbb{R}^m_+$. The generalized augmented Lagrangian $\widehat{\mathcal{L}} : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}_{++} \to \mathbb{R}$ is defined as

$$\widehat{\mathcal{L}}(x, y, \rho) = f_0(x) + \rho \sum_{i=1}^{m} h_i^*\big(h_i'(y_i) + \rho^{-1} f_i(x)\big). \tag{58}$$

GALB generates a sequence $\{(x^k, y^k)\} \subset \mathbb{R}^n \times \mathbb{R}^m$ given by

$$x^0 \in \mathbb{R}^n, \ y^0 \in \mathbb{R}^m_{++}, \tag{59}$$

$$x^{k+1} \in \operatorname{argmin}_{x \in \mathbb{R}^n} \widehat{\mathcal{L}}(x, y^k, \lambda_k), \tag{60}$$

$$y_i^{k+1} = (h_i^*)'\big(h_i'(y_i^k) + \lambda_k^{-1} f_i(x^{k+1})\big) \quad (1 \le i \le m). \tag{61}$$

Finally, we present the generalized doubly augmented Lagrangian method (GDALB) for problems (P) and (D). We take $\{\lambda_k\}$ and h as in GALB, and additionally a Bregman function g with zone $\mathbb{R}^n$. The generalized doubly augmented Lagrangian $\widetilde{\mathcal{L}} : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}_{++} \times \mathbb{R}^n \to \mathbb{R}$ is defined as

$$\widetilde{\mathcal{L}}(x, y, \rho, w) = f_0(x) + \rho \sum_{i=1}^{m} h_i^*\big(h_i'(y_i) + \rho^{-1} f_i(x)\big) + \rho D_g(x, w), \tag{62}$$

GDALB generates a sequence $\{(x^k, y^k)\} \subset \mathbb{R}^n \times \mathbb{R}^m$ given by

$$x^0 \in \mathbb{R}^n, \ y^0 \in \mathbb{R}^m_{++}, \tag{63}$$

$$x^{k+1} \in \operatorname{argmin}_{x \in \mathbb{R}^n} \widetilde{\mathcal{L}}(x, y^k, \lambda_k, x^k), \tag{64}$$

$$y_i^{k+1} = (h_i^*)'\big(h_i'(y_i^k) + \lambda_k^{-1} f_i(x^{k+1})\big) \quad (1 \le i \le m). \tag{65}$$

As in the case of AL, minimizers in (60), may fail to exist, even when problems (P) and (D) have solutions. If we take a similar example, namely the one-dimensional problem min 1 s.t. $e^x \leq 1$, and consider GALB with $h(x) = x \log x$, we get $\widetilde{\mathcal{L}}(x, 1, 1) = 1 + e^x$ (see (102)), which has no minimizers, even though, as in the case of the former example, any pair $(x, 0)$ with $x \leq 0$ belongs to $S_P^* \times S_D^*$. Again, boundedness of the level sets of any of the f_i's $(0 \leq i \leq m)$, for instance, is enough to ensure existence and uniqueness of x^k. Also, even if $\{x^k\}$ is well defined, it might be unbounded (e.g., the same example as used for AL). GDALB, on the other hand, guarantees existence and uniqueness of x^k for all k, and boundedness of the sequence $\{x^k\}$, as we prove below.

GALB was introduced for the first time in [13] and further analyzed in [28]. GDALB in fact has also a version within the classical framework, which could be called DAL, introduced in [42]. It modifies AL through the addition of a quadratic regularization term in x to the augmented Lagrangian, i.e. by changing (25) to

$$x^{k+1} \in \operatorname{argmin}_{x \in \mathbb{R}^n} \overline{\mathcal{L}}(x, y^k, \lambda_k) + \lambda_k \left\| x - x^k \right\|^2,$$

with $\overline{\mathcal{L}}$ as in (24), while (26) remains unchanged. In this case, the minimand in the definition of x^{k+1} is strictly convex and coercive (due to the presence of the quadratic term), so that x^{k+1} exists and is unique. Also, it can be proved that the whole sequence $\{x^k\}$ converges to a point in S_P^*, if this set is nonempty. The proof, similar to the convergence proof for AL presented in Theorem 2, can be found in [42]. GDALB as presented here, i.e. with Bregman distances, was also introduced in [13]. In this reference however, the convergence result for GDALB, appearing in Theorem 8, is wrong, because the author assumes that the interior of the zone of h contains $\mathbb{R}^m_+$. In such a case (65) does not guarantee nonnegativity of y^{k+1}. In fact $h(y) = \|y\|^2$ satisfies the asumptions in [13], but for this h we obtain $y_i^{k+1} = y_i^k + \lambda_k^{-1} f_i(x^{k+1})$, which might be negative. The correct formula associated with this h would be precisely (26), but in (26) the "max" forces nonnegativity of $\{y^k\}$. A correct analysis, which avoids this pitfall, is presented in Theorem 8 below.

Next we establish convexity of $\widehat{\mathcal{L}}$, $\widetilde{\mathcal{L}}$ in x.

Proposition 2. *Both $\widehat{\mathcal{L}}$ and $\widetilde{\mathcal{L}}$ are convex functions of x.*

Proof. Observe that the image of $(h_i^*)'$ is the domain of h_i', equal to $\mathbb{R}_{++}$ by B6'. Thus, $(h_i^*)'$ is positive, i.e. h_i^* is increasing. Since f_i is convex and ρ is positive, $h_i'(y_i) + \rho^{-1} f_i(\cdot)$ is convex. Since h_i^* is convex and increasing, we get that $h_i^*\big(h_i'(y_i) + \rho^{-1} f_i(\cdot)\big)$ is convex (see, e.g., [19], Vol. I, p. 264). Since f_0 and $D_g(\cdot, w)$ are convex, the result follows in view of (58), (62). □

Examples of $\widehat{\mathcal{L}}$, $\widetilde{\mathcal{L}}$ for specific choices of g, h are presented in Section 13.

9 The Connection between GPPB, GALB and GDALB

We will prove that, starting from the same y^0, the sequences $\{y^k\}$ generated by (51), (53) and by GALB coincide. This result was proved for the first time in [13]. The same holds for the sequences $\{y^k\}$ generated by (56)–(57) and by GDALB. The proof of the following theorem follows the same line as the proof of Theorem 3.

Theorem 5. *Let $\{y^k\}$ be the sequence generated by (51) and (53), and $\{(\widehat{x}^k, \widehat{y}^k)\}$ the sequence generated by GALB (i.e. by (59)–(61)). Assume that $\widehat{x}^k$ as defined by (60) exists for all k. If $y^0 = \widehat{y}^0$ then $y^k = \widehat{y}^k$ for all $k \geq 0$.*

Proof. We proceed by induction. Assume that $y^k = \widehat{y}^k$. We will prove that $y^{k+1} = \widehat{y}^{k+1}$. First we claim that $\widehat{x}^{k+1}$ minimizes $\mathcal{L}(\cdot, \widehat{y}^{k+1})$ with $\mathcal{L}$ as in (17). Let $q_i^k(x) = h_i^*\big(h_i'(\widehat{y}_i^k) + \lambda_k^{-1} f_i(x)\big)$. By (60), (58),

$$0 \in \partial_x \widehat{\mathcal{L}}\big(\widehat{x}^{k+1}, \widehat{y}^k, \lambda_k\big) = \partial f_0(\widehat{x}^{k+1}) + \lambda_k \sum_{i=1}^m \partial q_i^k(\widehat{x}^{k+1}) =$$

$$\partial f_0(\widehat{x}^{k+1}) + \sum_{i=1}^m (h_i^*)'\big(h_i'(\widehat{y}_i^k) + \lambda_k^{-1} f_i(\widehat{x}^{k+1})\big)\partial f_i(\widehat{x}^{k+1}), \tag{66}$$

using linearity of the subdifferential and the chain rule of subdifferential calculus (see [19], Vol. I, p. 261 and p. 264). By (66), (61) and (17),

$$0 \in \partial f_0(\widehat{x}^{k+1}) + \sum_{i=1}^m \widehat{y}_i^{k+1} \partial f_i(\widehat{x}^{k+1}) = \partial_x \mathcal{L}(\widehat{x}^{k+1}, \widehat{y}^{k+1}). \tag{67}$$

The claim is established. Thus, by (17) and (20),

$$\psi(\widehat{y}^{k+1}) = \mathcal{L}(\widehat{x}^{k+1}, \widehat{y}^{k+1}) = f_0(\widehat{x}^{k+1}) + \sum_{i=1}^m \widehat{y}_i^{k+1} f_i(\widehat{x}^{k+1}). \tag{68}$$

By (20), (68) and (17), for all $y \in \mathbb{R}^m$,

$$\psi(\widehat{y}^{k+1}) - \psi(y) \geq \psi(\widehat{y}^{k+1}) - \mathcal{L}(\widehat{x}^{k+1}, y) = \sum_{i=1}^m (\widehat{y}_i^{k+1} - y_i) f_i(\widehat{x}^{k+1}). \tag{69}$$

Next observe that, in view of (61) and the relation between the derivatives of h_i and h_i^*,

$$[\nabla h(\widehat{y}^k) - \nabla h(\widehat{y}^{k+1})]_i = h_i'(\widehat{y}_i^k) - h_i'(\widehat{y}_i^{k+1}) = h_i'(\widehat{y}_i^k) - h_i'\big[(h_i^*)'\big(h_i'(\widehat{y}_i^k) + \lambda_k^{-1} f_i(\widehat{x}^{k+1})\big)\big] =$$

$$h_i'(\widehat{y}_i^k) - h_i'(\widehat{y}_i^k) - \lambda_k^{-1} f_i(\widehat{x}^{k+1}) = -\lambda_k^{-1} f_i(\widehat{x}^{k+1}). \tag{70}$$

By inductive assumption, (70) and (69), for all $y \in \mathbb{R}^m_+$,

$$\lambda_k \langle \nabla h(y^k) - \nabla h(\widehat{y}^{k+1}), y - \widehat{y}^{k+1} \rangle = \lambda_k \langle \nabla h(\widehat{y}^k) - \nabla h(\widehat{y}^{k+1}), y - \widehat{y}^{k+1} \rangle \leq \psi(\widehat{y}^{k+1}) - \psi(y). \tag{71}$$

It follows from (71) that $\lambda_k[\nabla h(y^k) - \nabla h(\widehat{y}^{k+1})] \in \partial(-\psi(\widehat{y}^{k+1}))$, i.e. that $\widehat{y}^{k+1}$ solves (53). Since (53) is equivalent to (52), which uniquely determines the next iterate y^{k+1} of the sequence $\{y^k\}$, we conclude that such next iterate is precisely $\widehat{y}^{k+1}$. Thus, we have established that $y^{k+1} = \widehat{y}^{k+1}$. □

Theorem 6. *tarting from the same $z^0 = (x^0, y^0)$, the sequence $\{z^k\}$ generated by (56)–(57) and the sequence $\{(x^k, y^k)\}$ generated by GDALB coincide.*

Proof. It suffices to show that (56)–(57) are equivalent to (64)–(65). Note that, in view of the relation between the derivatives of h_i, h_i^*, (57) is equivalent to

$$y_i^{k+1} = (h_i^*)'\big(h_i'(y_i^k) + \lambda_k^{-1} f_i(x^{k+1})\big), \tag{72}$$

which is precisely (65). In view of (72), (56) is equivalent to

$$0 \in \lambda_k \big[\nabla g(x^{k+1}) - \nabla g(x^k)\big] + \partial f_0(x^{k+1}) + \sum_{i=1}^m (h_i^*)'\big(h_i'(y_i^k) + \lambda_k^{-1} f_i(x^{k+1})\big) \partial f_i(x^{k+1}) =$$

$$\lambda_k \big[\nabla g(x^{k+1}) - \nabla g(x^k)\big] + \partial f_0(x^{k+1}) + \sum_{i=1}^m \partial q_i^k(x^{k+1}). \tag{73}$$

Observe that, by (44), $\nabla g(x^{k+1}) - \nabla g(x^k)$ is $\nabla D_g(\cdot, x^k)$ evaluated at x^{k+1}. Thus, (73) is equivalent to saying that $0 \in \partial_x \widetilde{\mathcal{L}}(x^{k+1}, y^k, \lambda_k, x^k)$ because of (62), i.e., in view of Proposition 2, that $x^{k+1} \in \operatorname{argmin} \widetilde{\mathcal{L}}(\cdot, y^k, \lambda_k, x^k)$, which is precisely (64). □

10 Convergence Analysis for the Dual Sequence $\{y^k\}$ of GALB and GDALB

In this section we establish the convergence properties of the sequence $\{y^k\}$ generated either by GALB or by GDALB. For $y \in \mathbb{R}^m_+$, let $J(y) = \big\{i \in \{1, \ldots, m\} : y_i = 0\big\}$, $I(y) = \{1, \ldots, m\} \setminus J(y)$.

Theorem 7. *If problem (D) has solutions, then*

i) the sequence $\{y^k\}$ generated by GALB (i.e. by (59)–(61)) converges to a point y^* belonging to S_D^* and $J(y^*) \subset J(\widehat{y})$ for all $\widehat{y} \in S_D^*$,

ii) if x^k as defined by (60) *exists for all k and $\{x^k\}$ has cluster points, then any pair $(\bar{x}, y^*)$, where $\bar{x}$ is a cluster point of $\{x^k\}$, satisfies conditions* (20)–(22).

Proof. i) Since, by Theorem 5, $\{y^k\}$ coincides with the sequence generated by the proximal point with Bregman distances applied to problem (D), the result follows from already established results for this method applied to convex optimization, e.g. Theorems 4.1 and 5.1 in [20].

ii) Define $G : \mathbb{R}^n \times \mathbb{R}^m \to \mathcal{P}(\mathbb{R}^n \times \mathbb{R}^m)$ as $G(x, y) = \big(\partial_x \mathcal{L}(x, y), 0\big)$. It is immediate that G is maximal monotone. By maximality, it is upper semicontinuous. Note that (67) can be rewritten as

$$(0, 0) \in G(\widehat{x}^{k+1}, \widehat{y}^{k+1}). \tag{74}$$

By upper semicontinuity of G, taking limits in (74) as k goes to ∞ along an appropriate subsequence, we get, in view of Theorem 5, that $(\bar{x}, y^*)$ satisfies (20). (21) holds because y^* belongs to S_D^* by (i). (22) certainly holds if $y_i^* = 0$. If $y_i^* > 0$, we get from (70)

$$f_i(x^{k+1}) = \lambda_k \big[h_i'(y_i^{k+1}) - h_i'(y_i^k)\big]. \tag{75}$$

Taking limits in (75) as k goes to ∞ along an appropriate subsequence, and remembering that $\{\lambda_k\}$ is bounded, the right hand side of (75) converges to 0, because $\lim_{k\to\infty} y_i^k = \lim_{k\to\infty} y_i^{k+1} = y_i^* > 0$, and h_i' is continuous in $\mathbb{R}_{++}$ by B1. Thus $f_i(\bar{x}) = 0$ and (22) holds. □

Convergence of the proximal method with Bregman distances for convex optimization has also been proved in [11], but in this reference a condition on the Bregman function c stronger than B6 is assumed, namely that the image of $\mathrm{int}(C)$ through ∇c is the whole space $\mathbb{R}^p$. Some relevant Bregman functions, like the one in Example 3 of Section 5, satisfy B6 but not this stronger condition.

Now we analyze the sequence $\{y^k\}$ generated by GDALB, which coincides with the sequence $\{y^k\}$ generated by (54)–(55), by virtue of Theorem 6. Convergence of the sequence $\{(x^k, y^k)\}$ generated by GPPB applied to VIP$(T, \mathbb{R}^n \times \mathbb{R}^m_+)$, with T as in (48), to a point in S^*, would follow directly from the results in [5], excepting for the following obstacle. Several technical assumptions on T are made in [5] to ensure convergence. All but one of them hold for T as in (48). The missing one is *paramonotonicity*, introduced in [9]. T is said to be paramonotone if it is monotone and additionally $\langle v - v', z - z'\rangle = 0$ with $v \in T(z)$, $v' \in T(z')$ implies $v' \in T(z)$, $v \in T(z')$. We show next that T as in (48) is never paramonotone (unless (P) is unconstrained): take x, i such that there exists $0 \neq \xi^i \in \partial f_i(x)$, take $\xi^0 \in \partial f_0(x)$ such that $\xi^0 + \xi^i \notin \partial f_0(x)$ (ξ^0 exists because $\partial f_0(x)$ is bounded, see [19], Vol. I, p. 283) and let $z = (x, e^i)$, $z' = (x, 0)$. Then $(\xi^0 + \xi^i, -f(x)) \in T(z)$, $(\xi^0, -f(x)) \in T(z')$ and $\langle(\xi^0 + \xi^i, -f(x)) - -(\xi^0, -f(x)), z - z'\rangle = \langle(\xi^i, 0) - -(0, e^i)\rangle = 0$, but $(\xi^0 + \xi^i, -f(x)) \notin T(z') = (\partial f_0(x), -f(x))$. Thus this application of GPPB requires a

new convergence proof. Our next result puts together several partial results from [5], which do not require paramonotonicity, and others, which are related to the specific form of T as given by (48). The proof of the following theorem follows the same line as the proof of Theorem 2, and can be seen also as a corrected version of Theorem 8 in [13], avoiding the mistake mentioned above.

Theorem 8. *Let $\{z^k\} = \{(x^k, y^k)\}$ be the sequence generated by (54)–(55) (or, equivalently, by (56)–(57)). If $S^* \neq \emptyset$ then the following results hold:*

i) $\{D_c(z^*, z^k)\}$ is nonincreasing (and convergent) for all $z^* \in S^*$.

ii) The sequence $\{z^k\}$ is bounded.

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

iv) All cluster points $\bar{z} = (\bar{x}, \bar{y})$ of $\{z^k\}$ satisfy conditions (20)–(22).

v) All cluster points of $\{y^k\}$ belong to S_D^*.

vi) If $\bar{y}$ is a cluster point of $\{y^k\}$ then $J(\bar{y}) \subset J(\widehat{y})$ for all $\widehat{y} \in S_D^*$.

Proof. Items (i)–(iii) follows from the results in [5], since their proofs in this reference do not use paramonotonicity. For the sake of completeness we include their proof, without too many details. The facts that z^{k+1} is uniquely determined by (55) and that $\{z^k\} \subset \text{int}(C)$ (i.e. that $\{y^k\} \subset \mathbb{R}^m_{++}$) have been proved in Theorem 1 of [5], since all its assumptions hold in our case. It follows easily from (44) that

$$0 \leq \langle \nabla c(z^k) - \nabla c(z^{k+1}), z^{k+1} - z^* \rangle = D_c(z^*, z^k) - D_c(z^*, z^{k+1}) - D_c(z^{k+1}, z^k), \tag{76}$$

where the inequality follows from the fact that $\lambda_k \left[\nabla c(z^k) - \nabla c(z^{k+1})\right] \in T(z^{k+1})$ by (55), and so $\lambda_k \langle \nabla c(z^k) - \nabla c(z^{k+1}), z^{k+1} - z^* \rangle \geq \lambda_k \langle u^*, z^{k+1} - z^* \rangle \geq 0$ by monotonicity of T, where $u^* \in T(z^*)$ is the point which satisfies (47). Since D_c is nonnegative, it follows from (76) that $D_c(z^*, z^{k+1}) \leq D_c(z^*, z^k)$, establishing (i). By (i), $D_c(z^*, z^k) \leq D_c(z^*, z^0)$ for all $k \geq 0$, and thus (ii) follows from B3. By (76)

$$0 \leq D_c(z^{k+1}, z^k) \leq D_c(z^*, z^k) - D_c(z^*, z^{k+1}). \tag{77}$$

By (i) the right hand side of (77) converges to 0 as k goes to ∞, so that

$$\lim_{k\to\infty} D_c(z^{k+1}, z^k) = 0, \tag{78}$$

and then (iii) follows easily from (78) and B5, using (ii).

By (ii), $\{z^k\}$ has cluster points. Let $\bar{z} = (\bar{x}, \bar{y})$ be any of them. Using G as defined in the proof of Theorem 7, we may rewrite (56) as

$$\left(\lambda_k \left[\nabla g(z^k) - \nabla g(z^{k+1})\right], 0\right) \in G(x^{k+1}, y^{k+1}). \tag{79}$$

Since $\{\lambda_k\} \subset [\underline{\lambda}, \overline{\lambda}]$, the limit of the right hand side of (79) as k goes to ∞ along an appropriate subsequence, is, in view of (iii), $\widehat{\lambda}\big(\big[\nabla g(\bar{x}) - \nabla g(\bar{x})\big], 0\big) = (0,0)$, where $\widehat{\lambda} > 0$ is some cluster point of $\{\lambda_k\}$, because ∇g is continuous at $\bar{x}$, since the zone of g is $\mathbb{R}^n$. It follows from (79) and upper semicontinuity of G that

$$0 \in \partial_x \mathcal{L}(\bar{x}, \bar{y}) = \partial f_0(\bar{x}) + \sum_{i=1}^{m} \bar{y}_i \partial f_i(\bar{x}),$$

i.e. $\bar{z}$ satisfies (20). (21) holds at $\bar{y}$ because $\{y^k\} \subset \mathbb{R}^m_{++}$.

We look now at (22), which certainly holds if $\bar{y}_i = 0$. If $\bar{y}_i > 0$, then h_i' is continuous at $\bar{y}_i$ and, taking limits as k goes to ∞ along an appropriate subsequence in (57), we get, using (iii) and the fact that $\{\lambda_k\}$ is bounded,

$$f_i(\bar{x}) = \widetilde{\lambda}\big[h_i'(\bar{y}_i) - h_i'(\bar{y}_i)\big] = 0, \tag{80}$$

where $\widetilde{\lambda}$ is some cluster point of $\{\lambda_k\}$. By (80), $\bar{z}$ satisfies (22) and therefore (iv) holds.

We proceed to prove (v). By (78), (76), (i) and (57)

$$0 = \lim_{k\to\infty} \langle \nabla c(z^k) - \nabla c(z^{k+1}), z^{k+1} - z^* \rangle =$$

$$\lim_{k\to\infty} \big\{ \langle \nabla g(x^k) - \nabla g(x^{k+1}), x^{k+1} - x^* \rangle + \lambda_k^{-1} \sum_{i=1}^{m} f_i(x^{k+1})(y_i^* - y_i^{k+1}) \big\}. \tag{81}$$

Considering an appropriate subsequence in (81),

$$0 = \langle \nabla g(\bar{x}) - \nabla g(\bar{x}), \bar{x} - x^* \rangle + \widetilde{\lambda}^{-1} \sum_{i=1}^{m} f_i(\bar{x})(y_i^* - \bar{y}_i) = \widetilde{\lambda}^{-1} \sum_{i=1}^{m} f_i(\bar{x})(y_i^* - \bar{y}_i),$$

where $\widetilde{\lambda}$ is a cluster point of $\{\lambda_k\}$. Thus,

$$0 = \sum_{i=1}^{m} f_i(\bar{x})(y_i^* - \bar{y}_i) = \mathcal{L}(\bar{x}, y^*) - \mathcal{L}(\bar{x}, \bar{y}) \geq \psi(y^*) - \mathcal{L}(\bar{x}, \bar{y}), \tag{82}$$

using (17) and (20). Since $(\bar{x}, \bar{y})$ satisfies (20) by (iv), $0 \in \partial_x \mathcal{L}(\bar{x}, \bar{y})$, i.e. $\bar{x} \in \operatorname{argmin}_{x\in\mathbb{R}^n} \mathcal{L}(x, \bar{y})$, implying, in view of (19), that

$$\psi(\bar{y}) = \mathcal{L}(\bar{x}, \bar{y}). \tag{83}$$

By (82) and (83), $\psi(y^*) \leq \psi(\bar{y})$. Since $\bar{y} \geq 0$ by (iv) and y^* maximizes ψ in $\mathbb{R}^m_+$ because $S^* = S_P^* \times S_D^*$, we get that $\bar{y}$ also maximizes ψ in $\mathbb{R}^m_+$, i.e., it belongs to S_D^*.

In order to establish (vi), we must prove that $\bar{y}_i > 0$ if $\widehat{y}_i > 0$ for some $\widehat{y} \in S_D^*$. Suppose, by contradiction, that for some i and some $\widehat{y} \in S_D^*$ it holds that $\widehat{y}_i > 0$ and $\bar{y}_i = 0$. Then, taking any $\widehat{x} \in S_P^*$, it holds that $\widehat{z} := (\widehat{x}, \widehat{y}) \in S^*$, and we get

$$h_i(\widehat{y}_i) - h_i(y_i^k) - h_i'(y_i^k)(\widehat{y}_i - y_i^k) = D_{h_i}(\widehat{y}_i, y_i^k) \leq \sum_{\ell=1}^{m} D_{h_\ell}(\widehat{y}_\ell, y_\ell^k) =$$

$$D_h(\widehat{y}, y^k) \leq D_h(\widehat{y}, y^k) + D_g(\widehat{x}, x^k) = D_c(\widehat{z}, z^k) \leq D_c(\widehat{z}, z^0), \tag{84}$$

using (i) in the rightmost inequality and nonnegativity of Bregman distances. By (84),

$$h_i'(y_i^k)(y_i^k - \widehat{y}_i) \leq D_c(\widehat{z}, z^0) + h_i(y_i^k) - h_i(\widehat{y}_i). \tag{85}$$

Taking limits in (85) as k goes to ∞ along an appropriate subsequence, the right hand side converges to the finite value $D_c(\widehat{z}, z^0) + h_i(\bar{y}_i) - h_i(\widehat{y}_i)$, while the right hand side diverges to $+\infty$, because $y_i^k - \widehat{y}_i$ converges, along the subsequence, to $\bar{y}_i - \widehat{y}_i = -\widehat{y}_i < 0$, while $h_i'(y_i^k)$ diverges, along the subsequence, to $\lim_{t \to 0^+} h_i'(t) = -\infty$, by B6'. This contradiction establishes the result. □

Corollary 1. *If $S^* \neq \emptyset$ then the sequence $\{(x^k, y^k)\}$ generated by GDALB (i.e. by (63)–(65)) is bounded, all its cluster points satisfy (20)–(22), all cluster points of $\{y^k\}$ belong to S_D^* and $J(\bar{y}) \subset J(\widehat{y})$ for all $\widehat{y} \in S_D^*$. In particular, x^k, as defined by (64) exists and is unique for all k.*

Proof. Follows from Theorems 6 and 8. We mention again that existence and uniqueness of x^k follow from existence and uniqueness of z^k, established in Theorem 1 of [5]. □

11 Proximal Point and Augmented Lagrangian Methods with ϕ-divergences

We consider the problem

$$\min \varphi(y) \tag{86}$$

$$\text{s.t.} \qquad y \geq 0, \tag{87}$$

with a convex $\varphi : \mathbb{R}^m \to \mathbb{R} \cup \{\infty\}$. Given a ϕ-divergence d_ϕ satisfying (46), the *generalized proximal point method with ϕ-divergences* (GPPϕ from now on) generates a sequence starting from an arbitrary $y^0 \in \mathbb{R}^m_{++}$ through the iteration formula

$$y^{k+1} = \operatorname{argmin}\{\varphi(y) + \lambda_k d_\phi(y, y^k)\}, \tag{88}$$

with $\{\lambda_k\} \subset [\underline{\lambda}, \overline{\lambda}]$ for some $\overline{\lambda} \geq \underline{\lambda} > 0$.

GPPϕ has been introduced in [23], and further analyzed in [25] and [24]. GPPϕ applied to general nonlinear complementarity problems, rather than convex optimization problems, is studied in [6]. It has been proved in [23] that when problem (86)–(87) has solutions, the sequence $\{y^k\}$ generated by GPPϕ (i.e. by (88)) converges to a solution of problem (86)–(87). We will not include here a proof of this result, which is considerably harder than the similar result for GPPB, e.g. Theorem 4.1 of [20], and also than Theorem 8.

We present next the *augmented Lagrangian method with ϕ-divergences* (GALϕ form now on) for solving problems (P). Define $\mathcal{L}^* : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}_{++}$ as

$$\mathcal{L}^*(x, y, \rho) = f_0(x) + \rho \sum_{i=1}^{m} y_i \phi^*\big(\rho^{-1} f_i(x)\big),$$

where ϕ^* is the convex conjugate of ϕ. Starting from $(x^0, y^0) \in \mathbb{R}^n \times \mathbb{R}^m_{++}$, GAL$\phi$ generates a sequence $\{(x^k, y^k)\}$ through the following formulae:

$$x^{k+1} = \operatorname{argmin}_{x \in \mathbb{R}^n} \mathcal{L}^*(x, y^k, \lambda_k), \tag{89}$$

$$y_i^{k+1} = y_i^k (\phi^*)'\big(\lambda_k^{-1} f_i(x^{k+1})\big). \tag{90}$$

It has been proved in [23] that the sequences $\{y^k\}$ generated by GPPϕ applied to problem (D) and by GALϕ applied to problem (D) coincide when both start from the same y^0. The proof is similar to those of Theorems 3 and 5 and will be omitted here. We mention that for ϕ as in Example 5 of Section 6 we recover Polyak's modified barrier method (see [37]). In view of the results above, the sequence $\{y^k\}$ generated by GALϕ converges to a point $y^* \in S_D^*$ whenever S^* is nonempty. It is also easy to prove that if the sequence $\{x^k\}$ generated by GALϕ has cluster points, $(\bar{x}, y^*)$ satisfies (20)–(22) for any cluster point $\bar{x}$ of $\{x^k\}$, as we have proved in Theorem 7(ii) for GALB. In the following section we discuss more accurate convergence results for the sequence $\{x^k\}$ of GALϕ.

12 Convergence Analysis for the Primal Sequence $\{x^k\}$ of GALB and GDALB

Of course, one expects to prove that the sequence $\{x^k\}$ generated by either GALB, GALϕ or GDALB converges to a point in S_P^*, or at least that one of its cluster points belongs to S_P^*. In the case of GALB or GALϕ, we must confront the already discussed fact that the sequence $\{x^k\}$ may very well be unbounded. The situation is different for GDALB, thanks to the additional regularization term $D_g(x, w)$ in (62). In fact, $\{z^k\}$ (and henceforth $\{x^k\}$) is uniquely determined by (56)–(57) (or (64)–(65)), and $\{x^k\}$ is bounded by Corollary 1. We mention that, as a consequence, (64) can be

written as $x^{k+1} = \text{argmin}_{x \in \mathbb{R}^n} \widetilde{\mathcal{L}}(x, y^k, \lambda_k, x^k)$. The additional regularization term in $\widetilde{\mathcal{L}}$ makes it strictly convex, and this is the main advantage of GDALB over GALB or GALϕ. From now on we restrict our discussion to GALB and GDALB.

Assuming then that $\{x^k\}$ is bounded, it follows from Theorem 7(ii) and Corollary 1 that in order to establish optimality of a cluster point $\bar{x}$ of $\{x^k\}$ it suffices to check that $\bar{x}$ is primal feasible, i.e. that it satisfies (23) for all i, which is true for i such that $\bar{y}_i > 0$, since $(\bar{x}, \bar{y})$ satisfies (22). The problem lies in the case of $\bar{y}_i = 0$, because, if we try to take limits in (57) or (75) for some i such that $\bar{y}_i = 0$, we get an undetermined right hand side, since $\lim_{t \to 0^+} h_i'(t) = -\infty$ by B6'. We mention that, if we knew that the whole sequence $\{x^k\}$ converges, say to x^*, then it follows easily that x^* satisfies (23), because, since $y^k \geq 0$, for each i such that $\bar{y}_i = 0$ there exists a subsequence $\{y^{\ell_k}\}$ of $\{y^k\}$ such that $y_i^{\ell_k + 1} \leq y_i^{\ell_k}$, and, since h_i' is nondecreasing by B2, the right hand side of (57) or (75) with ℓ_k instead of k is nonpositive, so that, taking limits along this subsequence, we get that $f_i(x^*) \leq 0$. The problem is that we do not know whether the whole sequence $\{x^k\}$ converges or not. In fact, the argument above allows us to establish that for each i there exists a cluster point of $\{x^k\}$ which satisfies the i-th constraint in (16), but not that there exists a cluster point that satisfies all of them.

We will deal with this obstacle in two ways. First we will see that under a strict complementarity assumption on problem (P), and an additional condition on $\{z^k\}$ which our sequences do satisfy, conditions (20)–(22) automatically imply (23), and all cluster points of $\{x^k\}$ (if any) belong to S_P^*. In the case of GDALB, we get as a consequence that the whole sequence $\{(x^k, y^k)\}$ converges to some $(x^*, y^*) \in S^*$. In the absence of this strict complementarity assumption, we will prove primal optimality of the cluster points of the averaged sequence $\{\bar{x}^k\}$ introduced in Section 2. The remaining results in this section are taken from [7], and are, to our knowledge, new.

Before introducing the new assumption we need a preliminary lemma on the solution set of a certain variational inequality problem and some further notation. Let V be the set of pairs $(x, y) \in \mathbb{R}^n \times \mathbb{R}^m$ which satisfy (20)–(22). For any $J \subset \{1, \ldots, m\}$, let $U_J = \{y \in \mathbb{R}^m_+ : J(y) \supset J\}$, $V_J = \{(x, y) \in V : J(y) = J\}$.

Lemma 1. *With the notation above, if T is as defined by* (48)*, then $V_J \subset S(T, \mathbb{R}^n \times U_J) \subset V$ for any $J \subset \{1, \ldots, m\}$.*

Proof. By (47) and (48), $S(T, \mathbb{R}^n \times U_J)$ is the set of pairs $(\bar{x}, \bar{y}) \in \mathbb{R}^n \times U_J$ such that

$$\langle \bar{\xi}, x - \bar{x} \rangle + \sum_{i=1}^{m} f_i(\bar{x})(\bar{y}_i - y_i) \geq 0 \tag{91}$$

for some $\bar{\xi} \in \partial f_0(\bar{x}) + \sum_{i=1}^m \bar{y}_i \partial f_i(\bar{x})$ and all $(x, y) \in \mathbb{R}^n \times U_J$. Let $I = \{1, \ldots, m\} \setminus J$. Take $(\bar{x}, \bar{y}) \in V_J$ (so that $\bar{y} \in U_J$). Since $(\bar{x}, \bar{y})$ belongs to $V_J \subset V$, it satisfies (20), so

that we can take $\bar{\xi} = 0$. Then, for all $(x, y) \in \mathbb{R}^n \times U_J$,

$$\langle \bar{\xi}, x - \bar{x} \rangle + \sum_{i=1}^{m} f_i(\bar{x})(\bar{y}_i - y_i) = \sum_{i=1}^{m} f_i(\bar{x})(\bar{y}_i - y_i) = -\sum_{i=1}^{m} f_i(\bar{x}) y_i = -\sum_{i \in I} f_i(\bar{x}) y_i, \tag{92}$$

using (22) and the fact that $y_i = 0$ for $i \in J$ because $y \in U_J$. For $i \in I$, $\bar{y}_i > 0$, because $(\bar{x}, \bar{y}) \in V_J$, and thus, by (22), $f_i(\bar{x}) = 0$. It follows that the rightmost expression in (92) vanishes, and therefore (91) holds. Thus, $(\bar{x}, \bar{y}) \in S(T, \mathbb{R}^n \times U_J)$ and we have proved that $V_J \subset S(T, \mathbb{R}^n \times U_J)$.

Take now $(\bar{x}, \bar{y}) \in S(T, \mathbb{R}^n \times U_J)$, so that (91) holds for some $\xi \in \partial f_0(\bar{x}) + \sum_{i=1}^{m} \bar{y}_i \partial f_i(\bar{x})$ and for all $(x, y) \in \mathbb{R}^n \times U_J$. Choosing $(x, y) = (\bar{x} \pm e^j, \bar{y}) \in \mathbb{R}^n \times U_J$ with $1 \leq j \leq n$, we get from (91) that $\bar{\xi} = 0$, and thus $(\bar{x}, \bar{y})$ satisfies (20). Condition (21) holds because $(\bar{x}, \bar{y}) \in S(T, \mathbb{R}^n \times U_J) \subset \mathbb{R}^n \times \mathbb{R}^m_+$. It remains to be checked that $(\bar{x}, \bar{y})$ satisfies (22), which is certainly the case if $\bar{y}_i = 0$. If $\bar{y}_i > 0$ then take $(x, y) = (\bar{x}, \bar{y} - (1/2)\bar{y}_i e^i) \in U_J$, and get from (91) $(1/2) f_i(\bar{x}) \bar{y}_i \geq 0$, implying that $f_i(\bar{x})_i \geq 0$. Take next $(x, y) = (\bar{x}, \bar{y} + 2\bar{y}_i e^i) \in U_J$, and get from (91) $-f_i(\bar{x})\bar{y}_i \geq 0$, implying that $f_i(\bar{x}) \leq 0$. It follows that $f_i(\bar{x}) = 0$ and therefore $(\bar{x}, \bar{y})$ satisfies (22). Since $(\bar{x}, \bar{y})$ satisfies (20)–(22), we get that $(\bar{x}, \bar{y}) \in V$ and so we have proved that $S(T, \mathbb{R}^n \times U_J) \subset V$ □

Usefulness of Lemma 1 lies in the fact that in general V is not convex, but $S(T, \mathbb{R}^n \times U_J)$ is. The lemma allows us to conclude that at least the convex hull of V_J is contained in V (in fact, it follows easily from the lemma that V_J is indeed convex). Next we introduce the strict complementarity assumption, called SCA.

Definition 1. *The strict complementarity assumption SCA holds for problems (P) and (D) if for all $(\bar{x}, \bar{y}) \in S^*$ and for all i $(1 \leq i \leq m)$ it holds that either $\bar{y}_i > 0$ or $f_i(\bar{x}) < 0$.*

Loosely speaking, SCA says that no constraint in (16) which is tight at some solution of (P) is redundant.

The following lemma, of some interest on its own, is closely related to Proposition 4 in [21], which deals with nonlinear complementarity problems.

Lemma 2. *If $(\bar{x}, \bar{y}) \in \mathbb{R}^n \times \mathbb{R}^m$ satisfies (20)–(22), $J(\bar{y}) \subset J(y^*)$ for some $y^* \in S_D^*$ and SCA holds, then $(\bar{x}, \bar{y})$ belongs to S^*.*

Proof. We use Lemma 1 with $J = J(\bar{y})$. Clearly $(\bar{x}, \bar{y}) \in V_J$ so that, by Lemma 1, $(\bar{x}, \bar{y}) \in S(T, \mathbb{R}^n \times U_J)$ with T as in (48). Note that $y^* \in U_J$ because $J \subset J(y^*)$. Take

$x^* \in S_P^*$. Thus, (x^*, y^*) belongs to S^*, and therefore it solves VIP(T,$\mathbb{R}^n \times \mathbb{R}^m_+$), so that, a fortiori, it solves VIP(T,$\mathbb{R}^n \times U_J$). Thus, both $(\bar{x}, \bar{y})$ and (x^*, y^*) belong to $S(T, \mathbb{R}^n \times U_J)$. For $t \in [0,1]$, define

$$x(t) = \bar{x} + t(x^* - \bar{x}),$$

$$y(t) = \bar{y} + t(y^* - \bar{y}). \tag{93}$$

By convexity of $S(T, \mathbb{R}^n \times U_J)$, $(x(t), y(t)) \in S(T, \mathbb{R}^n \times U_J)$. By Lemma 1, $(x(t), y(t))$ belongs to V, i.e. it satisfies (20)–(22) for all $t \in [0,1]$. We need to show that it also satisfies (23) for $t = 0$. For $i \in I(\bar{y})$, $\bar{y}_i > 0$, implying, in view of (93), $y(t)_i > 0$ for all $t \in [0,1)$, so that, by (22) with $(x, y) = (x(t), y(t))$, we get $f_i(x(t)) = 0$ for all $t \in [0,1)$ and all $i \in I(\bar{y})$. Since $J = J(\bar{y}) \subset J(y^*)$, for $i \in J$ we have $y(1)_i = y_i^* = 0$, so that, by SCA, $f_i(x(1)) = f_i(x^*) < 0$. Let $A = \{t \in [0,1) : f_i(x(t)) < 0 \text{ for all } i \in J\}$. By continuity of the f_i's, A is nonempty. Note that for $t \in A$ we have $f_i(x(t)) \leq 0$ for all i, i.e. (23) holds, and since $(x(t), y(t))$ satisfies (20)–(22) for all $t \in [0,1]$, we get that $(x(t), y(t)) \in S^*$ for all $t \in A$. Let $\bar{t} = \inf A$. By continuity of the f_i's again, $(x(\bar{t}), y(\bar{t})) \in S^*$. We claim that $\bar{t} = 0$. If $\bar{t} > 0$, then, by definition of A and $\bar{t}$, we have $f_i(x(\bar{t})) = 0$ for some $i \in J = J(\bar{y})$. Since $y(t)_i = 0$ for all $t \in [0,1]$ and all $i \in J$ by (93), SCA is violated at $(x(\bar{t}), y(\bar{t}))$. Thus the claim is established, and therefore $(\bar{x}, \bar{y}) = (x(0), y(0)) = (x(\bar{t}), y(\bar{t})) \in S^*$. □

Next we present the convergence result under SCA.

Theorem 9. *Assume that $S^* \neq \emptyset$, that x^k as defined by (60) exists for all k and that SCA holds. Then,*

i) All cluster points of the sequence $\{x^k\}$ generated by GALB (i.e. by (59)–(61)), if any, belong to S_P^*.

ii) The sequence $\{z^k\} = \{(x^k, y^k)\}$ generated by GDALB (i.e. by (63)–(65)) converges to some point $\bar{z} = (\bar{x}, \bar{y}) \in S^*$.

Proof. i) Let $\bar{x}$ be a cluster point of $\{x^k\}$. By Theorem 7, $\lim_{k\to\infty} y^k = y^* \in S_D^*$ and $(\bar{x}, y^*)$ satisfies (20)–(22). The result follows from Lemma 2 with $\bar{y} = y^*$.

ii) Let $\bar{z} = (\bar{x}, \bar{y})$ be a cluster point of $\{z^k\}$, which exists by Corollary 1. $(\bar{x}, \bar{y})$ satisfies (20)–(22) and $\bar{y} \in S_D^*$ by Corollary 1. Thus we may apply Lemma 2 and conclude that $\bar{z} \in S^*$. Let $\{z^{\ell_k}\}$ be a subsequence of $\{z^k\}$ such that $\lim_{k\to\infty} z^{\ell_k} = \bar{z}$. $\lim_{k\to\infty} D_c(\bar{z}, z^{\ell_k}) = 0$ by B4. Since $\bar{z} \in S^*$, $\{D_c(\bar{z}, z^k)\}$ is nonincreasing by Theorem 5(i). Thus, $\{D_c(\bar{z}, z^k)\}$ is a nonnegative and nonincreasing sequence with a subsequence which converges to 0. It follows that $\lim_{k\to\infty} D_c(\bar{z}, z^k) = 0$, and therefore $\lim_{k\to\infty} z^k = \bar{z} \in S^*$ by B5. □

SCA is a rather strong assumption. It implies, for instance, uniqueness of the dual solution in the differentiable case, as the following proposition shows.

Proposition 3. *If S^* is nonempty, the f_i's are continuously differentiable ($0 \leq i \leq m$) and SCA holds then S_D^* is a singleton.*

Proof. Looking at (22) for a fixed $x^* \in S_P^*$, it follows easily from SCA that for all $y \in S_D^*$ it holds that $J(y) = \{i : f_i(x^*) < 0\}$. Assume that y, y' belong to S_D^* with $y \neq y'$. Since S_D^* is the intersection of $\mathbb{R}_+^m$ with the affine manifold $L = \{y \in \mathbb{R}^m : \nabla f_0(\bar{x}) + \sum_{i=1}^m y_i \nabla f_i(\bar{x}) = 0,\ f_i(\bar{x}) y_i = 0\ (1 \leq i \leq m)\}$, the intersection of the line through y, y' with the relative boundary of S_D^* provides a point in S_D^* with more zero components than y, y'. □

Proposition 3 does not hold in the nondifferentiable case. For instance, for the two dimensional problem $\min \|x + e\|_2^2$ subject to $x_1 \geq 0$, $2x_2 \geq \|x - e\|_\infty - 1$, with $e^t = (1, 1)$, one gets, after some subdifferential calculus, that $S_p^* = \{(0, 0)\}$, and that S_D^* is the whole segment between $(2/3, 2)$ and e, contained in $\mathbb{R}_{++}^2$, so that SCA certainly holds. Nevertheless, Proposition 3 indicates that it is worthwhile to look for convergence results without SCA. In this case we will obtain only so called ergodic results, i.e. they refer to a sequence $\{\bar{x}^k\}$ of weighted averages of the x^k's, defined as

$$\bar{x}^k = \sum_{\ell=1}^{k} \mu_{k\ell} x^\ell, \tag{94}$$

with

$$\mu_{k\ell} = \frac{\lambda_{\ell-1}^{-1}}{\sum_{j=0}^{k-1} \lambda_j^{-1}}. \tag{95}$$

We will prove that for GDAL, the sequence $\{\bar{x}^k\}$ is bounded and all its cluster points belong to S_P^* if $S^* \neq \emptyset$. The same result is proved for the sequence $\{\bar{x}^k\}$ generated by GAL under the assumption that $\{x^k\}$ is bounded. A weaker ergodic result on the sequence $\{\bar{x}^k\}$ generated by GALB can be found in Lemma 8.10 of [28], where optimality of the cluster points of $\{\bar{x}^k\}$ is proved, but under the asumption that $(\bar{x}, y^*)$ satisfies (22), where $\bar{x}$ is a cluster point of $\{\bar{x}^k\}$ and $y^* = \lim_{k\to\infty} y^k$. We prove, using Lemma 1, that this condition indeed holds, and thus needs not be required as an assumption. We need first a preliminary lemma on weighted averaged sequences.

Lemma 3. *Take $\{v^k\} \subset \mathbb{R}^p$, $\theta_{k\ell} \in \mathbb{R}_{++}$ ($k \geq 1$, $1 \leq \ell \leq k$) such that $\sum_{\ell=1}^k \theta_{k\ell} = 1$ for all $k \geq 1$, $\lim_{k\to\infty} \theta_{k\ell} = 0$ for all $\ell \geq 1$, and define $\bar{v}^k = \sum_{\ell=1}^k \theta_{k\ell} v^\ell$. Then*

i) If $\{v^k\}$ is bounded then $\{\bar{v}^k\}$ is bounded.

ii) If $v^* = \lim_{k\to\infty} v^k$ then $v^* = \lim_{k\to\infty} \bar{v}^k$.

iii) If $\{v^k\}$ is bounded, H is the set of its cluster points and $\widehat{H}$ is its convex hull, then all cluster points of $\{\bar{v}^k\}$ belong to $\widehat{H}$.

Proof. (i) and (ii) are elementary. We proceed to prove (iii). For $v \in \mathbb{R}^p$, $W \subset \mathbb{R}^p$, let $d(v, W)$ be the Euclidean distance from v to W. If W is convex then $d(\cdot, W)$ is convex. Since $\{v^k\}$ is bounded, $\lim_{k\to\infty} d(v^k, H) = 0$. Since $H \subset \widehat{H}$, we have that $0 \leq d(v^k, \widehat{H}) \leq d(v^k, H)$ for all k. It follows that

$$\lim_{k\to\infty} d(v^k, \widehat{H}) = 0. \tag{96}$$

Then

$$0 \leq d(\bar{v}^k, \widehat{H}) = d\Big(\sum_{\ell=1}^{k} \theta_{k\ell} v^\ell, \widehat{H}\Big) \leq \sum_{\ell=1}^{k} \theta_{k\ell} d(v^k, \widehat{H}), \tag{97}$$

using convexity of $\widehat{H}$ and of $d(\cdot, \widehat{H})$. By (ii) and (96), $\lim_{k\to\infty} \sum_{\ell=1}^{k} \theta_{k\ell} d(v^\ell, \widehat{H}) = 0$, and then $\lim_{k\to\infty} d(\bar{v}^k, \widehat{H}) = 0$ by (97). The result follows. □

The next theorem presents ou ergodic convergence result.

Theorem 10. *Let $\{z^k\} = \{(x^k, y^k)\}$ be the sequence generated either by GALB or by GDALB, and $\{\bar{x}^k\}$ the sequence defined by (94)–(95). In the case of GALB, assume that x^k as defined by (60) exists for all k and that $\{x^k\}$ is bounded. Then all cluster points of $\{\bar{x}^k\}$ belong to S_P^*.*

Proof. We consider also the auxiliary averages $\bar{y}^k = \sum_{\ell=1}^{k} \mu_{k\ell} y^\ell$, with $\mu_{k\ell}$ as in (95), and define $\bar{z}^k = (\bar{x}^k, \bar{y}^k)$. Note that $\{z^k\}$ is bounded, by Corollary 1 in the case of GDALB, and by Theorem 7(i) and boundedness of $\{x^k\}$ in the case of GALB. Let H be the set of cluster points of $\{z^k\}$, $\widehat{H}$ its convex hull and $J = \cap_{y \in S_D^*} J(y)$. Take any $\widetilde{z} = (\widetilde{x}, \widetilde{y}) \in H$. It follows easily from Corollary 1 or Theorem 7(i) that

$$J(\widetilde{y}) = J. \tag{98}$$

Now we apply Lemma 1 with this choice of J. By Corollary 1 or Theorem 7(i), $\widetilde{z} \in V$, and then $\widetilde{z} \in V_J$ by (98). By Lemma 1, $\widetilde{z} \in S(T, \mathbb{R}^n \times U_J)$ with T as in (48), i.e. we have proved that $H \subset S(T, \mathbb{R}^n \times U_J)$. Since $S(T, \mathbb{R}^n \times U_J)$ is convex, we get that $\widehat{H} \subset S(T, \mathbb{R}^n \times U_J)$. By Lemma 1 again, $\widehat{H} \subset V$. By Lemma 3(i), $\{\bar{z}^k\}$ is bounded. Let $z^* = (x^*, y^*)$ be a cluster point of $\{\bar{z}^k\}$. Since, by (95), $\sum_{\ell=1}^{k} \mu_{k\ell} = 1$ for all k, and also $\lim_{k\to\infty} \mu_{k\ell} = 0$ for all ℓ, because $0 \leq \mu_{k\ell} \leq \overline{\lambda}/(k\lambda_\ell)$, we can apply Lemma 3(iii) and conclude that $z^* \in \widehat{H} \subset V$. Thus, (x^*, y^*) satisfies (20)–(22). In order to prove that x^* belongs to S_P^*, it suffices to show that (x^*, y^*) also satisfies (23), which we do next.

By (57) or (70), for all $\ell \geq 0$,

$$h_i'(y_i^{\ell+1}) - h_i'(y_i^\ell) = \lambda_\ell^{-1} f_i(x^{\ell+1}) \qquad (1 \leq i \leq m). \tag{99}$$

Summing (99) with ℓ between 0 and $k-1$, and dividing by $\sum_{j=0}^{k-1}\lambda_j^{-1}$, we get

$$\frac{h_i'(y_i^k)-h_i'(y_i^0)}{\sum_{j=0}^{k-1}\lambda_j^{-1}}=\sum_{\ell=1}^{k}\mu_{k\ell}f_i(x^\ell)\geq f_i\left(\sum_{\ell=1}^{k}\mu_{k\ell}x^\ell\right)=f_i(\bar{x}^k), \tag{100}$$

using convexity of f_i in the inequality and (94) in the last equality. Since $\lambda_j\leq\overline{\lambda}$ for all $j\geq 0$, we get from (100)

$$f_i(\bar{x}^k)\leq\frac{\left|h_i'(y_i^k)-h_i'(y_i^0)\right|}{\sum_{j=0}^{k-1}\lambda_j^{-1}}\leq\frac{\overline{\lambda}}{k}\left|h_i'(y_i^k)-h_i'(y_i^0)\right|. \tag{101}$$

Now we look at a subsequence $\{\bar{x}^{j_k}\}$ of $\{\bar{x}^k\}$ converging to x^*. Without loss of generality, i.e. refining the subsequence if necessary, we may assume that $\{y^{j_k}\}$ converges, say to $\bar{y}$, because $\{y^k\}$ is bounded. If $\bar{y}_i>0$, then $\lim_{k\to\infty}h_i'(y_i^{j_k})=h_i'(\bar{y}_i)$ and thus the right hand side of (101) converges to 0 along the subsequence because of the denominator. It follows that $f_i(x^*)\leq 0$. If $\bar{y}_i=0$, then $\lim_{k\to\infty}h_i'(y_i^{j_k})=\lim_{t\to 0+}h_i'(t)=-\infty$ by B6', so that the left hand side of (100) is negative along the subsequence for large enough k, and we conclude again that $f_i(x^*)\leq 0$. We have proved that $f_i(x^*)\leq 0$ for all i, i.e. that (x^*,y^*) satisfies (23). It follows that x^* belongs to S_P^*. □

We remark that, in view of (94), (95), $\bar{x}^{k+1}$ can be computed as

$$\bar{x}^{k+1}=\delta_k\bar{x}^k+(1-\delta_k)x^{k+1},$$

with

$$\delta_k=\left(1+\frac{1}{\sigma_k\lambda_k}\right)^{-1},$$

where $\{\sigma_k\}$ is defined through the recurrence

$$\sigma_1=\lambda_0^{-1},\qquad\sigma_{k+1}=\sigma_k+\lambda_k^{-1}.$$

Thus, in actual implementations the additional cost of computing $\bar{x}^k$ (in addition to x^k, y^k) is negligeable, and we can look at $\{\bar{x}^k\}$ as the primal sequence generated by the algorithm, with $\{x^k\}$ being just an auxilary sequence. Nevertheless, the issue of primal optimality of the cluster points of $\{x^k\}$ is an interesting mathematical question which remains as an open problem.

Finally, we comment on the ergodic convergence results available for the sequence $\{x^k\}$ generated by GALϕ. For the case of $\phi(t)=t\log t-t+1$, i.e. ϕ as in Example 4 of Section 6, an ergodic convergence result for the sequence $\{x^k\}$ generated by (89)–(90) was proved in [47]. This is a particular case of Theorem 10, since this ϕ-divergence is also a Bregman distance. For $\phi(t)=t-\log t-1$, i.e. ϕ as in Example 5 of Section 6, an ergodic convergence result with an averaged sequence $\{\widetilde{x}^k\}$ defined with a formula more involved that (94)–(95) was established in [26], and then optimality of the cluster points of $\{\bar{x}^k\}$ as defined in (94)–(95) was proved in [40]. This result

was later extended to a larger class of ϕ-divergences in [38]. More precisely, it is required that ϕ belong to Φ_4 and furthermore that $\log\left(\phi^*(t)\right)$ be convex. This is a rather restrictive assumption, which does not hold for the ϕ-divergences in Examples 5 and 6 of Section 6, for instance. The results in [26] and [40] have no intersection with Theorem 10, since they correspond to a ϕ-divergence which is not a Bregman distance. The same is true for the results in [38], excepting for the case of $\phi(t) = t\log t - t + 1$.

13 Examples of Generalized Augmented Lagrangians

We give next the explicit expressions of the iterative formulae of GALB (i.e. of (60), (61)) for some specific choices of the Bregman function h.

Example 7: $h(y) = \sum_{i=1}^m y_i \log y_i$, with the convention that $0\log 0 = 0$, i.e. as in Example 2 of Section 5. For this h the iterative formulae of GALB are

$$x^{k+1} = \operatorname{argmin}_{x\in\mathbb{R}^n}\left\{f_0(x) + \lambda_k \sum_{i=1}^m y_i^k \exp\left(f_i(x)/\lambda_k\right)\right\}, \tag{102}$$

$$y_i^{k+1} = y_i^k \exp\left(f_i(x^{k+1})/\lambda_k\right). \tag{103}$$

This is the exponential multipliers method (see [3]) and the results of Theorem 10 for this particular case have been established in [47].

Example 8: $h(y) = \sum_{i=1}^m \left(y_i - y_i^\beta\right)$ with $\beta \in (0,1)$, i.e. h as in Example 3 of Section 5 with $\alpha = 1$. The iterative formulae for GALB with this h are

$$x^{k+1} = \operatorname{argmin}_{x\in\mathbb{R}^n}\left\{f_0(x) + (1-\beta)\lambda_k \sum_{i=1}^m (y_i^k)^\beta \left[\frac{\beta\lambda_k}{\beta\lambda_k - (y_i^k)^{1-\beta} f_i(x)}\right]^{\beta(1-\beta)}\right\}, \tag{104}$$

$$y_i^{k+1} = y_i^k \left[\frac{\beta\lambda_k}{\beta\lambda_k - (y_i^k)^{1-\beta} f_i(x^{k+1})}\right]^{1/(1-\beta)}. \tag{105}$$

The choice $\beta = 1/2$ makes (104) and (105) specially simple. For the general Bregman function in Example 3 of Section 5, i.e. with $\alpha > 1$, there are no explicit formulae for h_i^*, and henceforth for $\widehat{\mathcal{L}}$, excepting for $\alpha = 3/2$, $\beta = 1/2$, which we ommit here for the sake of conciseness.

Example 9: $h(y) = -\sum_{i=1}^m \log y_i$, which is just the logarithmic barrier, widely used in interior point methods for convex programming. In this case $D_h(y, y') = \sum_{i=1}^m [(y_i/y_i') - \log(y_i/y_i') - 1]$, used in statistics under the name of Itakura-Saitu distance (see e.g. [12]). The iterative formulae of GALB for this h are

$$x^{k+1} = \operatorname{argmin}_{x\in\mathbb{R}^n}\left\{f_0(x) - \lambda_k \sum_{i=1}^m \log\left(1 - y_i^k f_i(x)/\lambda_k\right)\right\}, \tag{106}$$

$$y_i^{k+1} = y_i^k \left[\frac{\lambda_k}{\lambda_k - y_i^k f_i(x^{k+1})} \right]. \tag{107}$$

The sequence $\{(x^k, y^k)\}$ generated by (106)–(107) starting with $y^0 \in \mathbb{R}^m_{++}$ is well defined, but the convergence results for GPPB do not apply to this case because h is not exactly a Bregman function, since the h_i's (and not just the h_i''s) diverge at 0, so that B2 fails. Thus, for instance, $D_c(z^*, z^k)$ in (76) with $z^* = (x^*, y^*)$ is not defined if $y_i^* = 0$ for some i. Convergence results for this h remains as an open problem (some results on the proximal point method with this h, applied directly to (P), rather than to (D), appear in [22]).

Next we present three examples of the iterative formulae of GALϕ, i.e. of the methods given by (89)–(90), for specific choices of ϕ.

Example 10: We consider GALϕ with $\phi(t) = t \log t - t + 1$, i.e. with ϕ as in Example 4 of Section 6. This particular case of GALϕ coincides with GALB as in Example 8.

Example 11: We consider GALϕ with ϕ as is Example 5 of Section 6. The iterative formulae of GALϕ for this ϕ are

$$x^{k+1} = \operatorname{argmin}_{x \in \mathbb{R}^n} \left\{ f_0(x) - \lambda_k \sum_{i=1}^m y_i^k \log\left(1 - f_i(x)/\lambda_k\right) \right\}, \tag{108}$$

$$y_i^{k+1} = y_i^k \left[\frac{\lambda_k}{\lambda_k - f_i(x^{k+1})} \right]. \tag{109}$$

This is one of the modified barrier methods presented in [37]. The ergodic results in [40] apply to the sequence $\{x^k\}$ generated by (108)–(109).

Example 12: We consider GALϕ with ϕ as in Example 6 of Section 6. The iterative formulae of GALϕ for this ϕ are

$$x^{k+1} = \operatorname{argmin}_{x \in \mathbb{R}^n} \left\{ f_0(x) - \lambda_k \sum_{i=1}^m y_i^k \left[1 - f_i(x)/\lambda_k\right]^{-1} \right\}, \tag{110}$$

$$y_i^{k+1} = y_i^k \left[\frac{\lambda_k}{\lambda_k - f_i(x^{k+1})} \right]^2. \tag{111}$$

To our knowledge, no convergence results are available for the sequence $\{x^k\}$ defined by (110)–(111), since ϕ^* is not log-convex for this ϕ, and so the results of [38] do not apply to this case.

Regarding GDALB, for the same h's as in Examples 7–9, formulae (103), (105) and (107) remain unchanged, while in formulae (102), (104) and (106) the only change

is an additional term $\lambda_k D_g(x, x^k)$ in the minimands. Since only B1–B3 are required for g, any continuously differentiable, strictly convex and coercive function defined on the whole $\mathbb{R}^n$ will do the job, but the most sensible choice seems to be a quadratic, as in Example 1 of Section 5, e.g. $g(x) = \nu \|x\|^2$ with $\nu > 0$, in which case the additional term is just $\lambda_k \nu \left\|x - x^k\right\|^2$.

References

[1] Auslender, A.A., Haddou, M. An interior-proximal method for convex linearly constrained problems and its extension to variational inequalities. *Mathematical Programming* **71** (1995) 77-100.

[2] Bertsekas, D.P. On penalty and multiplier methods for constrained optimization problems. *SIAM Journal on Control and Optimization* **14** (1976) 216-235.

[3] Bertsekas, D.P. *Constrained Optimization and Lagrange Multipliers*. Academic Press, New York (1982).

[4] Bregman, L.M. The relaxation method of finding the common points of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics* **7**,3 (1967) 200-217.

[5] Burachik, R.S., Iusem, A.N. A generalized proximal point algorithm for the variational inequality problem in a Hilbert space. *SIAM Journal on Optimization* **8** (1998) 197-216.

[6] Burachik, R.S., Iusem, A.N. A generalized proximal point algorithm for the nonlinear complementarity problem (to be published in *RAIRO, Recherche Opérationelle*).

[7] Burachik, R.S., Iusem, A.N. Generalized proximal point and augmented Lagrangian methods for convex optimization (to be published).

[8] Buys, J.D. Dual algoritmhs for constrained optimization problems. PhD. Thesis. University of Leiden, Leiden, The Netherlands (1972).

[9] Censor, Y., Iusem, A.N., Zenios, S.A. An interior point method with Bregman functions for the variational inequality problem with paramonotone operators. *Mathematical Programming* **81** (1998) 373-400.

[10] Censor, Y., Zenios, S.A. The proximal minimization algorithm with D-functions. *Journal of Optimization Theory and Applications* **73** (1992) 451-464.

[11] Chen, G., Teboulle, M. Convergence analysis of a proximal-like optimization algorithm using Bregman functions. *SIAM Journal on Optimization* **3** (1993) 538-543.

[12] Csiszár, I. An axiomatic approach to inference for linear inverse problems. *Annals of Statistics* **19** (1991) 2032-2066.

[13] Eckstein, J. Nonlinear proximal point algorithms using Bregman functions, with applications to convex programming. *Mathematics of Operations Research* **18** (1993) 202-226.

[14] Eggermont, P.P.B. Multiplicative iterative algorithms for convex programming. *Linear Algebra and Its Applications,* **130** (1990) 25-42.

[15] Eriksson, J. An iterative primal-dual algorithm for linear programming. Technical Report 85-10, Department of Mathematics, Linköping University (1985).

[16] Erlander, S. Entropy in linear programs. *Mathematical Programming* **21** (1981) 137-151.

[17] Harker, P.T., Pang, J.S. Finite dimensional variational inequalities and nonlinear complementarity problems: a survey of theory, algorithms and applications. *Mathematical Programming* **48** (1990) 161-220.

[18] Hestenes, M.R. Multiplier and gradient methods. *Journal of Optimization Theory and Applications* **4** (1969) 303-320.

[19] Hiriart-Urruty, J.-B., Lemaréchal, C. *Convex Analysis and Minimization Algorithms*. Springer, Berlin (1993).

[20] Iusem, A.N. On some properties of generalized proximal point methods for quadratic and linear programming. *Journal of Optimization Theory and Applications* **85** (1995) 593-612.

[21] Iusem, A.N., Kallio, M. An interior point method for constrained saddle point problems (to be published).

[22] Iusem, A.N., Monteiro, R.D.C. On dual convergence of the generalized proximal point method with Bregman distances (to be published).

[23] Iusem, A.N., Svaiter, B.F., Teboulle, M. Entropy-like proximal methods in convex programming. *Mathematics of Operations Research* **19** (1994) 790-814.

[24] Iusem, A.N., Teboulle, M. On the convergence rate of entropic proximal optimization algorithms. *Computational and Applied Mathematics* **12** (1993) 153-168.

[25] Iusem, A.N., Teboulle, M. Convergence rate analysis of nonquadratic proximal and augmented Lagrangian methods for convex and linear programming. *Mathematics of Operations Research* **20** (1995) 657-677.

[26] Jensen, D.L., Polyak, R.A.The convergence of a modified barrier method for convex programming. *IBM Journal of Research and Development* **38** (1994) 307-321.

[27] Kallio, M., Salo, S. Tatonnement procedures for linearly constrained convex optimization. *Management Science* **40** (1994) 788-797.

[28] Kiwiel, K. Proximal minimization methods with generalized Bregman functions. *SIAM Journal on Control and Optimization* **35** (1997) 1142-1168.

[29] Kort, B.W., Bertsekas, D.P. Combined primal-dual and penalty methods for convex programming. *SIAM Journal on Control and Optimization* **14** (1976) 268-294.

[30] Krasnoselskii, M.A. Two observations about the method of succesive approximations. *Uspekhi Matematicheskikh Nauk* **10** (1955) 123-127.

[31] Lemaire, B. The proximal algorithm. In *International Series of Numerical Mathematics* (J.P. Penot, editor). Birkhauser, Basel, **87** (1989) 73-87.

[32] Liese, F., Vajda, I. *Convex Statistical Distances*. Teubner, Leipzig (1987).

[33] Martinet, B. Régularisation d'inéquations variationelles par approximations succesives. *Revue Française de Informatique et Recherche Opérationelle* **2** (1970) 154-159.

[34] Martinet, B. *Algorithmes pour la résolution de problèmes d'optimisation et minimax*. Thèse d'état, Université de Grenoble (1972).

[35] Minty, G. Monotone nonlinear operators in Hilbert space. *Duke Mathematical Journal* **29** (1978) 341-346.

[36] Moreau, J. Proximité et dualité dans un espace hilbertien. *Bulletin de la Societé Mathématique de France* **93** (1965) 273-299.

[37] Polyak, R.A. Modified barrier functions, theory and applications. *Mathematical Programming* **54** (1992) 177-222.

[38] Polyak, R.A., Teboulle, M. Nonlinear rescaling and proximal-like methods in convex optimization. *Mathematical Programming* **76** (1997) 265-284.

[39] Powell, M.J.D. A method for nonlinear constraints in minimization problems. In *Optimization* (R. Fletcher, editor). Academic Press, London (1969).

[40] Powell, M.J.D. Some convergence properties of the shifted log barrier method for linear programming. *SIAM Journal on Optimization* **5** (1995) 697-739.

[41] Rockafellar, R.T. The multiplier method of Hestenes and Powell applied to convex programming. *Journal of Optimization Theory and Applications* **12** (1973) 555-562.

[42] Rockafellar, R.T. A dual approach to solving nonlinear programming problems by unconstrained optimization. *Mathematical Programming* **5** (1973) 354-373.

[43] Rockafellar, R.T. Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization* **14** (1976) 877-898.

[44] Rockafellar, R.T. Augmented Lagrangians and applications of the proximal point algorithm in convex programming. *Mathematics of Operations Research* **1** (1976) 97-116.

[45] Samuelson, P. Spatial price equilibrium and linear programming. *The American Economic Review* **42** (1952) 283-303.

[46] Teboulle, M. Entropic proximal mappings with applications to nonlinear programming. *Mathematics of Operations Research* **17** (1992) 97-116.

[47] Tseng, P., Bertsekas, D. On the convergence of the exponential multiplier method for convex programming. *Mathematical Programming* **60** (1993) 1-19.