20º COLÓQUIO BRASILEIRO DE MATEMÁTICA

MÉTODOS DE PONTO PROXIMAL EM OTIMIZAÇÃO

ALFREDO IUSEM

IMPA 24 - 28 JULHO, 1995

impa

ALFREDO N. IUSEM (IMPA, RJ)

ISBN 85-244-0095-1

Conselho Nacional de Desenvolvimento Científico e Tecnológico

INSTITUTO DE MATEMÁTICA PURA E APLICADA

Estrada Dona Castorina, 110 - Jardim Botânico

22460-320 - Rio de Janeiro, RJ, Brasil

Proximal Point Methods in Optimization

Alfredo N. Iusem

May 1995

Table of Contents

1. The concept of regularization.

The idea of regularization arose in connection with ill-posed problems. Given a problem of the form

$$L(f) = 0 \tag{1.1}$$

where f is an element of a set X (usually a function space) and $L: X \to X$ is an operator (usually differential, or integro-differential), (1.1) is said to be ill-posed when it has no solutions, or has more than one solution, or has a unique solution, but this solution does not depend in a continuous way upon some parameters of the operator L. The idea is to replace L by a regularized operator $L + \lambda M$ (with $\lambda \in \mathbf{R}$, $M: X \to X$), where M is such that the problem

$$L(f) + \lambda M(f) = (L + \lambda M)(f) = 0 \tag{1.2}$$

is well-posed (i.e. it is not ill-posed) for any $\lambda > 0$. In such a case (1.2) has a unique solution f_λ, and one expects that as λ approaches 0, f_λ provides some sort of approximation of a solution of (1.1) (see [**33**]).

This concept applies to optimization problems if we take $X = \mathbf{R}^n$ and $L = \nabla f$ where f is a convex function ($f: \mathbf{R}^n \to \mathbf{R}$), in which case (1.1) becomes

$$\nabla f(x) = 0 \tag{1.3}$$

or equivalently

$$\min_{x \in \mathbf{R}^n} f(x). \tag{1.4}$$

Assume that f is bounded below and take $g: \mathbf{R}^n \to \mathbf{R}$ strictly convex and coercive (i.e. $\lim_{\|x\| \to \infty} g(x) = +\infty$). Problem (1.3) may have no solution or more than one solution but the regularized problem

$$\min_{x \in \mathbf{R}^n} f(x) + \lambda g(x) \tag{1.5}$$

has a unique solution for each $\lambda > 0$, because the minimand $f + \lambda g$ is coercive (using the fact that f is bounded below) which reduces the problem to a compact set, so guaranteeing existence of solutions, and also strictly convex, implying uniqueness of the solution. (1.5)

has a unique solution $x(\lambda)$ and under some reasonable hypotheses (including existence of solutions of (1.4)) it can be proved that $\lim_{\lambda \to 0+} x(\lambda)$ exists and solves (1.4). The problem in this regularization approach is that, although $f + \lambda g$ is strictly convex and coercive for any $\lambda > 0$ however small, for very small λ this function will be numerically almost as ill behaved as f, or, in other words, if the system $\nabla f(x) = 0$ is ill conditioned then the system $(\nabla f + \lambda \nabla g)(x) = 0$ will also be ill conditioned when λ approaches 0, despite the fact that it has a unique solution for all $\lambda > 0$.

2. The proximal point algorithm for optimization in $\mathbf{R}^n$.

In order to overcome the difficulty just mentioned, it would be desirable to develop a regularization approach which does not require the regularization parameter λ to aproach 0 (say that it works with a constant λ). The proximal point algorithm attains such goal. It generates a sequence $\{x^k\} \subset \mathbf{R}^n$ in the following way:

$$x^0 \in \mathbf{R}^n \tag{2.1}$$

$$x^{k+1} = \operatorname*{argmin}_{x \in \mathbf{R}^n} \{ f(x) + \lambda_k \left\| x - x^k \right\|^2 \} \tag{2.2}$$

where λ_k is a real number satisfying

$$0 < \lambda_k \leq \tilde{\lambda} \tag{2.3}$$

for some $\tilde{\lambda} > 0$ (which includes the case of λ_k constant), and $\|\cdot\|$ is the Euclidean norm.

We will show next that under some reasonable hypotheses the sequence generated by (2.1), (2.2) converges to a minimizer of f. One approach for this convergence proof goes through the concept of firm nonexpansiveness, and works in fact for a problem much more general than (1.3), as we will see in section 4. We follow here an approach based upon the weaker notion of Fejér convergence, which works also for the nonquadratic extensions of the algorithm to be discussed in sections 11 – 15, for which the nonexpansiveness property does not hold.

A sequence $\{y^k\}$ in $\mathbf{R}^n$ is said to be Fejér convergent to a set $U \subset \mathbf{R}^n$ with respect to the Euclidean distance if

$$\|y^{k+1} - u\| \leq \|y^k - u\| \text{ for all } k \geq 0, \quad \text{for all } u \in U. \tag{2.4}$$

We have the following result

PROPOSITION 2.1. *If $\{y^k\}$ is Fejér convergent to $U \neq \emptyset$ then $\{y^k\}$ is bounded. If a cluster point y of $\{y^k\}$ belongs to U then $y = \lim_{k\to\infty} y^k$.*

PROOF: (2.4) implies $\|y^k - u\| \leq \|y^0 - u\|$ for any $u \in U$ so that the sequence $\{y^k\}$ is contained in a ball of center u and radius $\|y^0 - u\|$, henceforth it is bounded. For the second statement, let $\{y^{j_k}\}$ be a subsequence of $\{y^k\}$ such that $\lim_{k\to\infty} y^{j_k} = y$. Since $y \in U$, by (2.4) the sequence $\{\|y^k - y\|\}$ is decreasing and nonnegative, and it has a subsequence (namely $\{\|y^{j_k} - y\|\}$) which converges to 0. Then the whole sequence converges to 0, i.e. $0 = \lim_{k\to\infty} \|y^k - y\|$ implying $y = \lim_{k\to\infty} y^k$. ∎

Now we can prove the convergence of the proximal point algorithm.

THEOREM 2.1. *Let $f: \mathbf{R}^n \to \mathbf{R}$ be convex and continuously differentiable. Assume the set U of minimizers of f on $\mathbf{R}^n$ is nonempty. Then the sequence $\{x^k\}$ generated by (2.1), (2.2) converges to a point $x^* \in U$.*

PROOF: We divide the proof in 4 steps. In Step 1 we prove that $\{x^k\}$ is well defined. In Step 2 we prove that $\{x^k\}$ is Fejér convergent to U. In Step 3 we establish the useful fact that $\lim_{k\to\infty} (x^{k+1} - x^k) = 0$, to be used in Step 4, where we prove that any cluster point of $\{x^k\}$ belongs to U. The results of Steps 2 and 4, together with Proposition 2.1, imply the statement of the theorem.

STEP 1. *The sequence $\{x^k\}$ is well defined.*

By induction. Let $f_k(x) = f(x) + \lambda_k \|x - x^k\|^2$. Since f attains its minimum, it is bounded below, so that $\lim_{\|x\|\to\infty} f_k(x) = \infty$. Since f_k is continuous and the minimization in (2.2) reduces to a compact set, f_k attains its minimum. Since f is convex and $\lambda_k \|x - x^k\|^2$

is strictly convex, f_k is strictly convex and so it has a unique minimizer, i.e. x^{k+1} is uniquely determined.

STEP 2. $\|x^{k+1} - \overline{x}\|^2 \leq \|x^k - \overline{x}\|^2 - \|x^{k+1} - x^k\|^2$ *for all* $k \geq 0$ *and all* $\overline{x} \in U$.

$$\begin{aligned}\|x^k - \overline{x}\|^2 &= \|x^k - x^{k+1} + x^{k+1} - \overline{x}\|^2 \\ &= \|x^k - x^{k+1}\|^2 + \|x^{k+1} - \overline{x}\|^2 + 2\langle x^k - x^{k+1}, x^{k+1} - \overline{x}\rangle. \end{aligned} \tag{2.5}$$

Since x^{k+1} solves (2.2) we have

$$0 = \nabla f_k(x^{k+1}) = \nabla f(x^{k+1}) + 2\lambda_k(x^{k+1} - x^k). \tag{2.6}$$

From (2.5), (2.6) and convexity of f

$$\begin{aligned}&\|x^k - \overline{x}\|^2 - \|x^{k+1} - x^k\|^2 - \|x^{k+1} - \overline{x}\|^2 = 2\langle x^k - x^{k+1}, x^{k+1} - \overline{x}\rangle \\ &= \frac{1}{\lambda_k}\langle \nabla f(x^{k+1}), x^{k+1} - \overline{x}\rangle \geq \frac{1}{\lambda_k}[f(x^{k+1}) - f(\overline{x})] \geq 0 \end{aligned} \tag{2.7}$$

using the fact that $\overline{x}$ is a minimizer of f. The result follows from (2.7).

STEP 3. $\lim_{k\to\infty}(x^{k+1} - x^k) = 0$.

From Step 2

$$0 \leq \|x^{k+1} - x^k\|^2 \leq \|x^k - \overline{x}\|^2 - \|x^{k+1} - \overline{x}\|^2. \tag{2.8}$$

Since $\{\|x^k - \overline{x}\|\}$ is decreasing and nonnegative, it is convergent; so the right hand side of (2.8) converges to 0. The result follows.

STEP 4. $\{x^k\}$ *has cluster points and all of them belong to* U.

Existence of cluster points follows from Step 2 and the first statement of Proposition (2.1). Let $\overline{x}$ be a cluster point of $\{x^k\}$ and $\{x^{j_k}\}$ a subsequence of $\{x^k\}$ such that $\lim_{k\to\infty} x^{j_k} = \overline{x}$. By (2.6)

$$\nabla f(x^{j_k+1}) = 2\lambda_{j_k}(x^{j_k} - x^{j_k+1}). \tag{2.9}$$

By Step 3, $\lim_{k\to\infty} x^{j_k+1} = \lim_{k\to\infty} x^{j_k} = \overline{x}$. Taking limits in (2.9) as $k \to \infty$, and using $\lambda_k \leq \tilde{\lambda}$ and continuous differentiability of f,we get $\nabla f(\overline{x}) = 0$. By convexity of f, $\overline{x} \in U$.

Steps 2 and 4 indicate that both statements of Proposition 2.1 hold and therefore there exists $x^* \in U$ such that $x^* = \lim_{k\to\infty} x^k$. ∎

3. Maximal monotone operators.

A maximal monotone operator is a generalization of a positive semidefinite linear transformation to the nonlinear case, which includes the gradient of a convex differentiable function.

$A \in \mathbf{R}^{n\times n}$ is positive semidefinite iff $0 \leq x^t Ax = \langle x, Ax\rangle$ for all x. Let now $T\colon \mathbf{R}^n \to \mathbf{R}^n$ (not necessarily linear). $0 \leq \langle x, T(x)\rangle$does not work. Since $0 \leq \langle x, Ax\rangle$ for all x iff $0 \leq \langle (x-y), A(x-y)\rangle = \langle x-y, Ax-Ay\rangle$ for all x, y, we define:

DEFINITION 3.1: $T\colon \mathbf{R}^n \to \mathbf{R}^n$ is monotone iff

$$0 \leq \langle x-y, T(x)-T(y)\rangle \quad \text{for all } x, y \in \mathbf{R}^n. \tag{3.1}$$

EXAMPLE 3.1: $T = \nabla f$ with f convex and differentiable.

We like to cover also the nondifferentiable case, in which we use subgradients.

DEFINITION 3.2: ξ is a subgradient of f at x iff

$$\langle \xi, y-x\rangle \leq f(y) - f(x) \quad \forall y \in \mathbf{R}^n. \tag{3.2}$$

Let $\partial f(x) = \{\xi : \xi \text{ is a subgradient of } f \text{ at } x\}$. The two following properties are well known.

1. If f is differentiable and convex then $\partial f(x) = \{\nabla f(x)\}$.
2. If f is convex then $\partial f(x) \neq \emptyset$ for all x in the relative interior of the effective domain of f.

∂f associates to each x not just a vector but a subset of $\mathbf{R}^n$, so we need to extend the notion of monotone operator to point-to-set operators. Let $T\colon \mathbf{R}^n \to \mathcal{P}(\mathbf{R}^n)$.

DEFINITION 3.3: T is monotone iff

$$0 \leq \langle x - y, u - v \rangle \quad \text{for all } x, y \in \mathbf{R}^n, \quad u \in T(x), \quad v \in T(y). \tag{3.3}$$

EXAMPLE 3.2: $T = \partial f$ with f convex. Take $\xi \in T(x)$, $\eta \in T(y)$. By Definition 3.2 $\langle \xi, y - x \rangle \leq f(y) - f(x)$ and $\langle -\eta, y - x \rangle \leq f(x) - f(y)$, implying $0 \leq \langle \xi - \eta, x - y \rangle$.

DEFINITION 3.4: T is maximal monotone iff

i) T is monotone.

ii) For all T' monotone such that $T(x) \subset T'(x)$ for all x, it holds that $T = T'$.

It can be verified that ∂f is maximal monotone for convex f. If we take $\overline{T}: \mathbf{R}^n \to \mathcal{P}(\mathbf{R}^n)$ with $\overline{T}(x) \subset \partial f(x)$ but $\overline{T}(x) \neq \partial f(x)$ for some x then $\overline{T}$ is monotone but not maximal.

Since we admit $T(x) = \emptyset$ for some x, point-to-set operators can always be inverted: given $T: \mathbf{R}^n \to \mathcal{P}(\mathbf{R}^n)$ we define $T^{-1}: \mathbf{R}^n \to \mathcal{P}(\mathbf{R}^n)$ as $y \in T^{-1}(x)$ iff $x \in T(y)$.

4. The proximal point algorithm for maximal monotone operators.

The problem of interest is finding a zero of a maximal monotone operator. If $T: \mathbf{R}^n \to \mathbf{R}^n$ then x is a zero of T iff $T(x) = 0$. If $T: \mathbf{R}^n \to \mathcal{P}(\mathbf{R}^n)$, x is a zero of T iff $0 \in T(x)$.

For f convex and $T = \partial f$, we have that x is a zero of T iff x is a minimizer of f, because for $0 \in T(x)$ we have that, for all y,

$$0 = \langle 0, y - x \rangle \leq f(y) - f(x)$$

and then $f(x) \leq f(y)$ for all $y \in \mathbf{R}^n$, so that the problem of finding the zeroes of maximal monotone operators generalizes the problem of minimizing convex functions.

According to (2.6) the proximal point iteration is characterized by

$$\nabla f(x^{k+1}) = 2\lambda_k (x^k - x^{k+1})$$

which leads to the natural extension

$$\lambda_k (x^k - x^{k+1}) \in T(x^{k+1})$$

which is equivalent to

$$x^k \in (I + \frac{1}{\lambda_k} T)(x^{k+1})$$

i.e.

$$x^{k+1} \in (I + \frac{1}{\lambda_k} T)^{-1}(x^k). \tag{4.1}$$

(4.1) is the iteration of the proximal point method for finding zeroes of maximal monotone operators.

Before presenting an important result on maximal monotone operators, we introduce the concept of firmly nonexpansive operator.

DEFINITION 4.1: $P: \mathbf{R}^n \to \mathbf{R}^n$ is firmly nonexpansive iff

$$\|P(x) - P(y)\|^2 \leq \|x - y\|^2 - \|(x - y) - (P(x) - P(y))\|^2 . \tag{4.2}$$

PROPOSITION 4.1. *If P is firmly nonexpansive then the sequence defined by $x^0 \in \mathbf{R}^n$, $x^{k+1} = P(x^k)$ is Fejér convergent to the set of fixed points of P.*

PROOF: Apply (4.2) with $y = \overline{x}$ such that $P(\overline{x}) = \overline{x}$, $x = x^k$, $P(x) = x^{k+1}$ and get

$$\left\|x^{k+1} - \overline{x}\right\|^2 \leq \left\|x^k - \overline{x}\right\|^2 - \left\|x^k - x^{k+1}\right\|^2 . \tag{4.3}$$

Note that (4.3) is just Step 2 of Theorem 2.1.

For $P : \mathbf{R}^n \to \mathcal{P}(\mathbf{R}^n)$, we say that P is onto if for all $y \in \mathbf{R}^n$ there exists $x \in \mathbf{R}^n$ such that $y \in P(x)$ and that P is one to one if for $x \neq y$ it holds that $P(x) \cap P(y) = \emptyset$. The proof of the next theorem can be found in [27].

THEOREM 4.1. *(Minty's Theorem). If $T: \mathbf{R}^n \to \mathcal{P}(\mathbf{R}^n)$ is monotone maximal and $\mu > 0$ then $I + \mu T$ is one to one and onto and $(I + \mu T)^{-1}$ is firmly nonexpansive.*

The first statement ensures that (4.1) is a true regularization, in the sense that x^{k+1} is uniquely determined; the second one implies convergence of (4.1). However we will continue with our approach, and so we will use only the fact that $I + \mu T$ is onto, which guarantees existence of $x^{k+1} \in (I + \mu T)^{-1}(x^k)$. We will use neither uniqueness of x^{k+1} nor firm nonexpasiveness of $(I + \mu T)^{-1}$. We will need also the following lemma, which can be rephrased as saying that maximal monotone operators are closed.

LEMMA 4.1. *If* $\lim_{k\to\infty} y^k = \overline{y}$, $\lim_{k\to\infty} z^k = \overline{z}$, *$T$ is maximal monotone and* $y^k \in T(z^k)$ *then* $\overline{y} \in T(\overline{z})$.

PROOF: Define T' as

$$T'(z) = \begin{cases} T(z) & \text{if } z \neq \overline{z} \\ T(\overline{z}) \cup \{\overline{y}\} & \text{if } z = \overline{z}. \end{cases}$$

We claim that T' is monotone. We need to check

$$\langle y - y', z - z' \rangle \geq 0 \quad \forall z, z', \quad \forall y \in T'(z), \quad \forall y' \in T'(z') \tag{4.4}$$

and by monotonicity of T it suffices to check (4.4) for $y' = \overline{y}$, $z' = \overline{z}$. By monotonicity of T:

$$\langle y - y^k, z - z^k \rangle \geq 0 \quad \forall z, \quad \forall y \in T(z). \tag{4.5}$$

Taking limits in (4.5):

$$\langle y - \overline{y}, z - \overline{z} \rangle \geq 0 \quad \forall z, \quad \forall y \in T(z). \tag{4.6}$$

So (4.4) holds and T' is monotone. Since $T(x) \subset T'(x)$ for all x and T is maximal we conclude that $T = T'$, in particular $T(\overline{z}) = T'(\overline{z}) = T(\overline{z}) \cup \{\overline{y}\}$, i.e. $\overline{y} \in T(\overline{z})$ and the lemma is proved. ∎

THEOREM 4.2. *If* $T: \mathbf{R}^n \to \mathcal{P}(\mathbf{R}^n)$ *is maximal monotone and there exists* $\overline{x}$ *such that* $0 \in T(\overline{x})$ *then the sequence* $\{x^k\}$ *defined by*

$$x^0 \in \mathbf{R}^n \tag{4.7}$$

$$x^{k+1} \in (I + \frac{1}{\lambda_k} T)^{-1}(x^k) \tag{4.8}$$

with $0 < \lambda_k < \tilde{\lambda}$ *converges to a vector* x^* *such that* $0 \in T(x^*)$.

PROOF: We repeat the steps of the proof of Theorem 2.1.

STEP 1. *The sequence* $\{x^k\}$ *is well defined.*

$\{x^k\}$ is well defined by Minty's Theorem.

STEP 2. $\left\|x^{k+1}-\overline{x}\right\|^2 \leq \left\|x^k-\overline{x}\right\|^2 - \left\|x^{k+1}-x^k\right\|^2$ *for all* k *and all* $\overline{x}$ *such that* $0 \in T(\overline{x})$.

As in Theorem 2.1

$$\left\|x^k-\overline{x}\right\|^2 - \left\|x^{k+1}-\overline{x}\right\|^2 - \left\|x^{k+1}-x^k\right\|^2 = \frac{2}{\lambda_k}\left\langle \lambda_k(x^k-x^{k+1})-0, x^{k+1}-\overline{x}\right\rangle \geq 0 \quad (4.9)$$

because $0 \in T(\overline{x})$, and (4.9) follows from monotonicity of T.

STEP 3. $\lim_{k\to\infty}(x^{k+1}-x^k)=0$.

Follows from Step 2 as in Theorem 2.1.

STEP 4. $\{x^k\}$ *has cluster points and all of them are zeroes of* T.

Existence of cluster points follows from Step 2 as in Theorem 2.1. Let $\hat{x}$ be a cluster point of $\{x^k\}$ and $\{x^{j_k}\}$ a subsequence of $\{x^k\}$ such that $\lim_{k\to\infty} x^{j_k} = \hat{x}$. From Step 3 $\lim_{k\to\infty} x^{j_k+1} = \hat{x}$ and from (4.8)

$$\lambda_{j_k}(x^{j_k}-x^{j_k+1}) \in T(x^{j_k+1}). \quad (4.10)$$

The left hand side of (4.10) converges to 0, because $\lambda_k < \tilde{\lambda}$, but we cannot take limits in the right hand side in a straightforward way. This is the point where maximality of T has to be used. We invoke Lemma 4.1 in (4.10) with $y^k = \lambda_{j_k}(x^{j_k}-x^{j_k+1})$, $\overline{y}=0$, $z^k = x^{j_k+1}$, $\overline{z}=\hat{x}$, and conclude that $0 \in T(\hat{x})$, so Step 4 is proved. As in Theorem 2.1, Steps 2 and 4 and the Fejér convergence theorem establish the result. ∎

We proceed now to extend Theorem 4.2 to a Hilbert space. Steps 1-3 of Theorem 4.2 hold with the same proof but it is not true any more that a bounded sequence has cluster points. We need to introduce the concept of weak convergence.

Let H be a Hilbert space.

DEFINITION 4.2:

i) A sequence $\{x^k\} \subset H$ is strongly convergent to $x \in H$ ($x^k \underset{k\to\infty}{\longrightarrow} x$) iff

$$0 = \lim_{k\to\infty} \left\|x^k-x\right\|^2 = \lim_{k\to\infty}\left\langle x^k-x, x^k-x\right\rangle.$$

ii) A sequence $\{x^k\} \subset H$ is weakly convergent to $x \in H$ ($x^k \underset{k\to\infty}{\overset{w}{\longrightarrow}} x$) iff

$$0 = \lim_{k\to\infty} \langle x^k - x, y\rangle \quad \text{for all } y \in H.$$

$x^k \underset{k\to\infty}{\longrightarrow} x$ implies trivially (via Cauchy-Schwartz) that $x^k \underset{k\to\infty}{\overset{w}{\longrightarrow}} x$ but the opposite implication does not hold: take $H = \ell_2 = \{x = \{x_n\}_{n\in\mathbf{N}} : \sum_{n=1}^{\infty} x_n^2 < \infty\}$ with $\langle x, y\rangle = \sum_{n=1}^{\infty} x_n y_n$. Take $\{e^k\} \subset H$ with $e_n^k = \delta_{kn}$ (Kronecker's delta). Then $e^k \underset{k\to\infty}{\overset{w}{\longrightarrow}} 0$ because $\lim_{k\to\infty} \langle e^k - 0, y\rangle = \lim_{k\to\infty} \langle e^k, y\rangle = \lim_{k\to\infty} y_k = 0$, using $\sum_{k=1}^{\infty} y_k^2 = 0$. Note that $\|e^k\| = 1$ for all k and $\langle e^k, e^j\rangle = 0$, $\|e^k - e^j\| = \sqrt{2}$ for all $j \neq k$. For weak convergence we have the following result (see, e.g. [31 pp.141-143,177]).

THEOREM 4.3. *(Bourbaki-Alaoglu). If $\{x^k\} \subset H$ is bounded then $\{x^k\}$ has a weakly convergent subsequence.*

We also need an extension of Lemma 4.1 to weak convergence.

LEMMA 4.2. *If $y^k \underset{k\to\infty}{\longrightarrow} \overline{y}$, $z^k \underset{k\to\infty}{\overset{w}{\longrightarrow}} \overline{z}$, T is maximal monotone and $y^k \in T(z^k)$ then $\overline{y} \in T(\overline{z})$.*

PROOF: As in the proof of Lemma 4.1 it suffices to prove:

$$\langle y - \overline{y}, z - \overline{z}\rangle \geq 0 \text{ for all } z, \text{ for all } y \in T(z) \tag{4.11}$$

and we know

$$\langle y - y^k, z - z^k\rangle \geq 0 \text{ for all } z \text{ and all } y \in T(z). \tag{4.12}$$

By Banach–Steinhaus' Theorem (see, e.g., [31]) the sequence $\{z^k\}$ is bounded. Let $u^k = y - y^k$, $v^k = z - z^k$, $u = y - \overline{y}$, $v = z - \overline{z}$, so that $\{v^k\}$ is bounded, $u^k \underset{k\to\infty}{\longrightarrow} u$, $v^k \underset{k\to\infty}{\overset{w}{\longrightarrow}} v$. We claim that $\lim_{k\to\infty} \langle u^k, v^k\rangle = \langle u, v\rangle$:

$$|\langle u^k, v^k\rangle - \langle u, v\rangle| = |\langle u^k - u, v^k\rangle + \langle u, v^k - v\rangle| \leq \|u^k - u\|\,\|v^k\| + |\langle u, v^k - v\rangle|. \tag{4.13}$$

The right hand side of (4.13) converges to 0 and the claim is established, so that the left hand side of (4.12) converges to the left hand side of (4.11) and then (4.11) follows from (4.12). The lemma is proved. ∎

With the help of Theorem 4.3 and Lemma 4.2 we extend Theorem 4.2 to:

THEOREM 4.4. *If $T: H \to \mathcal{P}(H)$ is maximal monotone and there exists $\overline{x} \in H$ such that $0 \in T(\overline{x})$ then the sequence*

$$x^0 \in H \tag{4.14}$$

$$x^{k+1} \in (I + \frac{1}{\lambda_k} T)^{-1}(x^k), \tag{4.15}$$

with $0 < \lambda_k < \tilde{\lambda}$, is weakly convergent to a point $x^* \in H$ such that $0 \in T(x^*)$.

PROOF: Steps 1, 2 and 3 of Theorem 4.2 hold without changes. For Step 4 we must restrict the result to weak cluster points (i.e. weak limits of subsequences) and prove that any weak cluster point is a 0 of T. Existence of weak cluster points follows from Theorem 4.3, and by (4.15)

$$\lambda_{j_k}(x^{j_k} - x^{j_k+1}) \in T(x^{j_k+1}) \tag{4.16}$$

where $\{x^{j_k}\}$ is a subsequence of $\{x^k\}$ such that $x^{j_k} \underset{k\to\infty}{\overset{w}{\longrightarrow}} \hat{x}$. By Step 3 we have $x^{j_k+1} \underset{k\to\infty}{\overset{w}{\longrightarrow}} \hat{x}$. In order to take limits in (4.16) we apply Lemma 4.2 to (4.16), noting that $\lambda_{j_k}(x^{j_k} - x^{j_k+1}) \underset{k\to\infty}{\longrightarrow} 0$ by Step 3, and we get that $0 \in T(\hat{x})$, completing the proof of Step 4.

Unfortunately, the Fejér convergence theorem does not hold with weak limits, so we must prove that there is only one cluster point in a direct way. Assume $\hat{x}$ and $\tilde{x}$ are two weak cluster points. By Step 4, $0 \in T(\tilde{x})$, $0 \in T(\hat{x})$ and therefore, by Step 2, $\{\|x^k - \tilde{x}\|\}$ and $\{\|x^k - \hat{x}\|\}$ converge, say to α and β. Then

$$\|x^k - \tilde{x}\|^2 = \|x^k - \hat{x}\|^2 + \|\hat{x} - \tilde{x}\|^2 + 2\langle x^k - \hat{x}, \hat{x} - \tilde{x}\rangle . \tag{4.17}$$

(4.17) implies

$$2\langle x^k - \hat{x}, \hat{x} - \tilde{x}\rangle \underset{k\to\infty}{\longrightarrow} \alpha^2 - \beta^2 - \|\tilde{x} - \hat{x}\|^2 . \tag{4.18}$$

Since $x^{\ell_k} \xrightarrow[k\to\infty]{w} \hat{x}$ for some subsequence $\{x^{\ell_k}\}$ we have $\lim_{k\to\infty} \langle x^{\ell_k} - \hat{x}, \tilde{x} - \hat{x}\rangle = 0$ and then, from (4.18)

$$\|\hat{x} - \tilde{x}\| = \alpha^2 - \beta^2. \tag{4.19}$$

In a similar fashion $\|x^k - \hat{x}\|^2 = \|x^k - \tilde{x}\|^2 + \|\tilde{x} - \hat{x}\|^2 + 2\langle x^k - \tilde{x}, \tilde{x} - \hat{x}\rangle$, so that $2\langle x^{\ell_k} - \tilde{x}, \tilde{x} - \hat{x}\rangle \xrightarrow[k\to\infty]{} \beta^2 - \alpha^2 - \|\tilde{x} - \hat{x}\|^2$, implying

$$\|\tilde{x} - \hat{x}\|^2 = \beta^2 - \alpha^2. \tag{4.20}$$

From (4.19), (4.20), $\|\tilde{x} - \hat{x}\|^2 = 0$, i.e. $\tilde{x} = \hat{x}$, so that the weak cluster point is unique. ∎

5. Convergence rate of the proximal point algorithm.

It is not difficult to establish convergence rate results (basically linearity of the convergence rate) but this easiness is somewhat tricky; the hypotheses which allow such a simple proof mean basically that the operator T is quite regular to begin with: for the case of $T = \nabla f$ with f convex, these hypotheses are almost equivalent to demanding that f have a unique minimizer x^* and that $\nabla^2 f(x^*)$ be positive definite, in which case it is not clear why one should regularize it. Later on we will present linear convergence results for some cases where such hypotheses do not hold, as linear programming, i.e. with f defined as

$$f(x) = \begin{cases} c^t x & \text{if } Ax = b, \quad x \geq 0 \\ +\infty & \text{otherwise} \end{cases} \tag{5.1}$$

without demanding uniqueness of the solution, but the proof is much harder. We start by recalling that a function g is Lipschitz continuous at x if there exists σ such that $\|g(x) - g(y)\| \leq \sigma \|x - y\|$ for y in some neighborhood of x. The natural extension to operators $T: H \to \mathcal{P}(H)$ is to say that given x and $u \in T(x)$ there exist σ such that $\|u - v\| \leq \sigma \|x - y\|$ for all y in a neighborhood of x and all $v \in T(y)$. In particular we have

DEFINITION 5.1: If T is a maximal monotone operator ($T: H \to \mathcal{P}(H)$), we say that T^{-1} is Lipschitz continuous at 0 with constant σ if there exists β such that $\|x - \overline{x}\| \leq \sigma \|y\|$ whenever $0 \in T(\overline{x})$, $y \in T(x)$ and $\|y\| \leq \beta$.

Note that Lipschitz continuity of T^{-1} at 0 implies that there exists at most one x such that $0 \in T(x)$, because if we have $0 \in T(\overline{x})$, $0 \in T(\tilde{x})$, we get $\|x - \overline{x}\| \leq \sigma \|y\|$ and we can take $x = \tilde{x}$, $y = 0$ which gives $\|\overline{x} - \tilde{x}\| = 0$, i.e. $\overline{x} = \tilde{x}$. If $T = \nabla f$ with f convex, Lipschitz continuity of T^{-1} at 0 implies uniqueness of the minimizer of f and $\|x - \overline{x}\| \leq \sigma \|\nabla f(x)\|$ where $\overline{x}$ is the minimizer and x is close to $\overline{x}$. If $\nabla^2 f(\overline{x})$ is positive definite and its smaller eigenvalue is $\eta > 0$, we have Lipschitz continuity with $\sigma = \frac{2}{\eta}$, because $\nabla f(x) = \nabla^2 f(\overline{x} + \theta(x - \overline{x}))(x - \overline{x})$ with $0 \leq \theta \leq 1$, so that $\|\nabla f(x)\| \geq \overline{\eta} \|x - \overline{x}\|$, where $\overline{\eta}$ is the smallest eigenvalue of $\nabla^2 f(\overline{x} + \theta(x - \overline{x}))$ and we have $\overline{\eta} \geq \frac{1}{2}\eta$ for x close enough to $\overline{x}$, so that $\|x - \overline{x}\| \leq \frac{2}{\eta} \|\nabla f(x)\|$. This shows that when T^{-1} is Lipschitz continuous at 0 it hardly needs regularization, but this hypothesis allows a very simple linear convergence proof.

THEOREM 5.1. *If T is maximal monotone, T^{-1} is Lipschitz continuous at x with constant σ and there exists $\overline{x}$ such that $0 \in T(\overline{x})$ then $\overline{x}$ is the only zero of T and the sequence defined by the proximal point method (i.e. (4.11), (4.12)) satisfies, for k large enough,*

$$\left\|x^{k+1} - \overline{x}\right\| \leq \frac{\sigma\lambda_k}{\sqrt{1 + (\sigma\lambda_k)^2}} \left\|x^k - \overline{x}\right\| \tag{5.2}$$

so $\{x^k\}$ converges to $\overline{x}$ superlinearly if $\lim_{k\to\infty} \lambda_k = 0$ and linearly otherwise, with an asymptotic error constant bounded by $\frac{\sigma\tilde{\lambda}}{\sqrt{1+\sigma\tilde{\lambda}}}$.

PROOF: Using Definition 5.1 with $x = x^{k+1}$, we get

$$\left\|x^{k+1} - \overline{x}\right\| \leq \sigma \|y\| \text{ for all } y \in T(x^{k+1}) \text{ such that } \|y\| \leq \beta. \tag{5.3}$$

By (4.12) $\lambda_k(x^k - x^{k+1}) \in T(x^{k+1})$. Since $\lambda_k < \tilde{\lambda}$ and $\lim_{k\to\infty} (x^{k+1} - x^k) = 0$ by Step 3 of Theorem 4.4, we have $\left\|\lambda_k(x^k - x^{k+1})\right\| \leq \beta$ for large enough k, so that we can take $y = \lambda_k(x^k - x^{k+1})$ in (5.3) obtaining:

$$\left\|x^{k+1} - \overline{x}\right\|^2 \leq \sigma^2\lambda_k^2 \left\|x^{k+1} - x^k\right\|^2 \leq \sigma^2\lambda_k^2(\left\|x^k - \overline{x}\right\|^2 - \left\|x^{k+1} - \overline{x}\right\|^2) \tag{5.4}$$

using Step 2 of Theorem 4.4 in the rightmost inequality of (5.4). (5.2) follows immediately from (5.4), implying superlinear convergence when $\lim_{k\to\infty} \lambda_k = 0$. For the last statement of the theorem, get from (5.4)

$$\left\|x^{k+1} - \overline{x}\right\|^2 \leq \sigma^2\tilde{\lambda}^2(\left\|x^k - \overline{x}\right\|^2 - \left\|x^{k+1} - \overline{x}\right\|^2) \tag{5.5}$$

which implies $\left\|x^{k+1} - \overline{x}\right\| \leq \frac{\sigma\tilde{\lambda}}{\sqrt{1+(\sigma\tilde{\lambda})^2}} \left\|x^k - \overline{x}\right\|$. ∎

Some results are available for special cases where T^{-1} is not Lipschitz continuous at 0. For instance, if f is polyhedral (i.e. its epigraph is a polyhedron, e.g. problem (5.1)) and bounded below on $\mathbf{R}^n$, it has been proved in [29] that convergence is indeed finite.

6. Approximate versions of the proximal point algorithm.

Let $P_k = (I + \frac{1}{\lambda_k}T)^{-1}$, so that the proximal point algorithm for finding the zeroes of the maximal monotone operator T can be written as

$$x^0 \in \mathbf{R}^n \tag{6.1}$$

$$x^{k+1} = P_k(x^k). \tag{6.2}$$

Note that the equality in (6.2) is legitimate by the one-to-one statement of Minty's Theorem. Two approximate versions of the algorithm can be obtained by requiring that, instead of (6.2), x^{k+1} is chosen so that it satisfies just

$$\left\|x^{k+1} - P_k(x^k)\right\| \leq \varepsilon_k \tag{6.3}$$

or

$$\left\|x^{k+1} - P_k(x^k)\right\| \leq \varepsilon_k \left\|x^{k+1} - x^k\right\|. \tag{6.4}$$

All previous convergence results (including the convergence rate) can be established for algorithms (6.3) and (6.4), when ε_k is such that $\sum_{k=0}^{\infty} \varepsilon_k < \infty$. The price to be paid is a considerable complication in some of the proofs (see [29]).

7. Augmented Lagrangians.

After establishing the convergence properties of the proximal point method, it is worthwhile to discuss its usefulness in optimization. The first observation is that the proximal point method should be seen more as a conceptual scheme than as an implementable algorithm. Though in general $f_k(x) = f(x) + \lambda_k \left\|x^k - x\right\|^2$ is easier to minimize than f, because of its more regular behavior, each iteration of the proximal point method requires minimization of a function on $\mathbf{R}^n$, for which some other numerical procedure must be used. The virtues of the proximal point method will make themselves evident only when the subproblems are substantially easier than the original problem.

Now, there are many situations in which it pays to replace one minimization problem by a sequence of minimization problems. One of them is the case of constrained original problem and unconstrained subproblems. Another one could be inequality constrained original problem and equality constrained subproblems. All the applications of the proximal point method and its extensions which we will consider in the remainder of the text will be of one of these two types. In this section we discuss the so called augmented Lagrangian method, which substitutes a sequence of unconstrained subproblems for a constrained original problem. We will show that this method is just a particular instance of the proximal point method, and our results of the previous sections will provide a convergence analysis for the augmented Lagrangian algorithm far easier than a direct study.

The problem under consideration is

$$\min f(x) \tag{7.1}$$

$$\text{s.t. } g_i(x) \leq 0 \qquad (1 \leq i \leq m) \tag{7.2}$$

with $f, g_i: \mathbf{R}^n \to \mathbf{R}$ convex and differentiable. The standard Lagrangian for this problem is $L: \mathbf{R}^n \times \mathbf{R}^m \to \mathbf{R}$ defined by

$$L(x,y) = \begin{cases} f(x) + \sum_{i=1}^{m} y_i g_i(x) & \text{if } y \geq 0 \\ +\infty & \text{otherwise.} \end{cases} \tag{7.3}$$

One approach to solve (7.1)-(7.2) is to generate two sequences, $\{x^k\} \subset \mathbf{R}^n$, $\{y^k\} \subset \mathbf{R}^m$ so that, given $y^k \geq 0$,

$$x^k = \operatorname*{argmin}_{x \in \mathbf{R}^n} L(x, y^k) \tag{7.4}$$

and then y^k is updated in some appropriate way. Computationally, the trouble lies in the discontinuity of $L(x, \cdot)$. This leads to the notion of an augmented Lagrangian with better differentiability properties. For $x \in \mathbf{R}^n$, let $x^+ \in \mathbf{R}^n$ be defined as $x_j^+ = \max\{x_j, 0\}$. Take $\alpha > 0$ and define the augmented Lagrangian $L_\alpha : \mathbf{R}^n \times \mathbf{R}^m \to \mathbf{R}$ as

$$L_\alpha(x, y) = f(x) + \frac{1}{4\alpha} \sum_{i=1}^{m} \{[(y_i + 2\alpha g_i(x))^+]^2 - y_i^2\}. \tag{7.5}$$

A simple computation shows that $L_\alpha(x, y)$ is differentiable if f and g_i are differentiable (however L_α is not twice differentiable, even when this holds for f and g_i). In fact

$$\nabla_x L_\alpha(x, y) = \nabla f(x) + \sum_{i=1}^{m} (y_i + 2\alpha g_i(x))^+ \nabla g_i(x). \tag{7.6}$$

The augmented Lagrangian method generates $\{x^k\} \subset \mathbf{R}^n$, $\{y^k\} \subset \mathbf{R}^m$ through:

$$y^0 \in \mathbf{R}^m_+ \tag{7.7}$$

and given $y^k \in \mathbf{R}^m$

$$x^k = \operatorname*{argmin}_{x \in \mathbf{R}^n} L_\alpha(x, y^k) \tag{7.8}$$

$$y_i^{k+1} = (y_i^k + 2\alpha g_i(x^k))^+. \tag{7.9}$$

Consider now the dual objective $\varphi : \mathbf{R}^m \to \mathbf{R}$ defined by

$$\varphi(y) = \min_{x \in \mathbf{R}^n} L(x, y) \tag{7.10}$$

with L as in (7.3). It is easy to see that φ is concave.

We prove next that the sequence $\{y^k\}$ generated by (7.7)-(7.9) is the same as the sequence generated by the proximal point method applied to $-\varphi$.

THEOREM 7.1. *Let* $\{\overline{y}^k\} \subset \mathbf{R}^m$ *be the sequence generated by the proximal point method for* $\min_{y \in \mathbf{R}^m} \{-\varphi(y)\}$, *with* $\lambda_k = \frac{1}{4\alpha}$, *and* $\{y^k\}$ *the sequence given by (7.7)-(7.9). If* $\overline{y}^0 = y^0$ *then* $\overline{y}^k = y^k$ *for all* k.

PROOF: By induction. True for $k = 0$. Assume $y^k = \overline{y}^k$. Since

$$\overline{y}^{k+1} = \operatorname{argmin}\{-\varphi(y) + \frac{1}{4\alpha} \left\| y - y^k \right\|^2\}$$

we have

$$\frac{1}{2\alpha}(\overline{y}^{k+1} - y^k) \in \partial\varphi(\overline{y}^{k+1}). \tag{7.11}$$

By definition of ∂f and φ, (7.11) is equivalent to

$$\frac{1}{2\alpha} \left\langle \overline{y}^{k+1} - y^k, \overline{y}^{k+1} - u \right\rangle \leq \varphi(\overline{y}^{k+1}) - \varphi(u) \text{ for all } u \geq 0 \tag{7.12}$$

i.e.

$$\varphi(u) \leq \varphi(\overline{y}^{k+1}) - \frac{1}{2\alpha} \left\langle \overline{y}^{k+1} - y^k, \overline{y}^{k+1} - u \right\rangle \text{ for all } u \geq 0. \tag{7.13}$$

Since (7.13) determines $\overline{y}^{k+1}$ uniquely, it suffices to check that (7.13) holds with y^{k+1} substituting for $\overline{y}^{k+1}$. By (7.6), (7.8), (7.9), (7.3)

$$\begin{aligned} 0 = \nabla_x L_\alpha(x^k, y^k) &= \nabla f(x^k) + \sum_{i=1}^m (y_i^k + 2\alpha g_i(x^k))^+ \nabla g_i(x^k) \\ &= \nabla f(x^k) + \sum_{i=1}^m y_i^{k+1} \nabla g_i(x^k) \in \partial_x L(x^k, y^{k+1}). \end{aligned} \tag{7.14}$$

(7.14) implies that x^k minimizes $L(\ \cdot\ , y^{k+1})$ and, by (7.10)

$$\varphi(y^{k+1}) = f(x^k) + \sum_{i=1}^m y_i^{k+1} g_i(x^k). \tag{7.15}$$

(7.15) gives an expression for the first term in the right hand side of (7.13) with y^{k+1} instead of $\overline{y}^{k+1}$. Next we evaluate the second term, i.e. $\frac{1}{2\alpha} \left\langle y^{k+1} - y^k, y^{k+1} - u \right\rangle$. By (7.9)

$$y_i^{k+1} - y_i^k = \max\{-y_i^k, 2\alpha g_i(x^k)\} \geq 2\alpha g_i(x^k) \tag{7.16}$$

which implies

$$y_i^{k+1}(y_i^{k+1} - y_i^k) = 2\alpha y_i^{k+1} g_i(x^k). \tag{7.17}$$

So, for any $u \geq 0$

$$\frac{1}{2\alpha}(y_i^{k+1} - y_i^k)(y_i^{k+1} - u_i) = y_i^{k+1} g_i(x^k) - \frac{1}{2\alpha}(y_i^{k+1} - y_i^k)u_i \leq y_i^{k+1} g_i(x^k) - u_i g_i(x^k). \tag{7.18}$$

From (7.15), (7.18)

$$\begin{aligned}
&\varphi(y^{k+1}) - \frac{1}{2\alpha}\left\langle y^{k+1} - y^k, y^{k+1} - u\right\rangle \\
&\geq f(x^k) + \sum_{i=1}^m y_i^{k+1} g_i(x^k) - \sum_{i=1}^m y_i^{k+1} g_i(x^k) + \sum_{i=1}^m u_i g_i(x^k) \\
&= L(x^k, u) \geq \min_{x \in \mathbf{R}^n} L(x, u) = \varphi(u).
\end{aligned} \tag{7.19}$$

(7.19) shows that (7.13) holds with y^{k+1} substituting for $\overline{y}^{k+1}$, so that $y^{k+1} = \overline{y}^{k+1}$ and the induction step is complete. ∎

Theorem 7.1, combined with Theorem 4.2, ensures convergence of the sequence $\{y^k\}$ generated by the augmented Lagrangian method to a maximizer y^* of the dual objective φ (whenever problem (7.1)-(7.2) has solutions), i.e., by standard duality results, to a vector y^* of optimal Karush-Kuhn-Tucker multipliers for problem (7.1)-(7.2). Convergence of the sequence $\{x^k\}$ cannot be immediately obtained from the proximal point theory (this requires some additional hypotheses, including a Slater condition, i.e. existence of $\overline{x}$ such that $g_i(\overline{x}) < 0$ for all i). However, it is immediate from the fact that x^k minimizes $L(\,\cdot\,, y^{k+1})$ that if the sequence $\{x^k\}$ converges to x^* then the pair (x^*, y^*) is a saddle point of L, and again standard convex duality results imply that x^* is a solution of (7.1)-(7.2).

8. Penalization.

Consider problem

$$\begin{aligned}
&\min\ f(x) \\
&\text{s.t. } x \in \overline{S}
\end{aligned} \tag{8.1}$$

with $\overline{S} \subset \mathbf{R}^n$ closed. One approach to deal with (8.1) is the introduction of a penalty function g, satisfying $g(x) = \infty$ if $x \notin S$. For $\lambda > 0$, the problem

$$\min_{x \in \mathbf{R}^n} \{f(x) + \lambda g(x)\} \tag{8.2}$$

will have its solutions in $\overline{S}$ (if it has solutions at all). If (8.1) has solutions and g is appropriately chosen, then (8.2) will have a unique solution $x(\lambda)$ and under some additional hypotheses it can be proved that $\lim_{\lambda \to 0} x(\lambda)$ exists and solves (8.1). As in the case of regularization, the trouble is that (8.2) becomes ill conditioned for very small λ (i.e., for points x close to the boundary $\partial\overline{S}$ of $\overline{S}$, $g(x)$ is very large and $\lambda g(x)$ becomes sort of indetermined when λ approaches 0). The idea here is to combine penalization with the proximal point approach so as to get convergence even when λ is far from 0. This is related somehow to the so called exact penalty functions, where g is chosen so that $x(\overline{\lambda})$ obtained from (8.2) solves (8.1) for some fixed $\overline{\lambda}$. The difference is that in the exact penalty function approach existence of $\overline{\lambda}$ is theoretically guaranteed but the value of $\overline{\lambda}$ is not known beforehand, so that it may be so small that the indetermination discussed above is still present. The proximal point approach on the other hand will work with an arbitrary λ (or a sequence λ_k bounded above). We will attain this goal through the introduction of a distance-like function $D: S \times S \to \mathbf{R}_+$ (where S is the interior of $\overline{S}$) such that $D(x, y) = 0$ iff $x = y$ and $D(x, y)$ approaches infinity as y approaches the boundary ∂S of S. We will consider two classes of such "distances": Bregman distances (which in fact are defined on $\overline{S} \times S$) and φ-divergences. In both cases the proximal point approach consists of generating a sequence $\{x^k\}$ with $x^0 \in S$ and

$$x^{k+1} = \operatorname*{argmin}_{x \in \mathbf{R}^n} \{f(x) + \lambda_k D(x, x^k)\}. \tag{8.3}$$

9. Bregman functions and distances.

Let S be an open and convex subset of $\mathbf{R}^n$ and $\overline{S}$ its closure. Consider a convex real function h defined on $\overline{S}$ and let $D_h: \overline{S} \times S \to \mathbf{R}$ be

$$D_h(x, y) = h(x) - h(y) - \nabla h(y)^t (x - y). \tag{9.1}$$

h is said to be a *Bregman function* (and D_h the *Bregman distance* induced by h) if the following conditions hold:

B1) h is continuously differentiable on S.

B2) h is strictly convex and continuous on $\overline{S}$.

B3) For all $\delta \in \mathbf{R}$ the partial level sets $\Gamma_1(y,\delta) = \{x \in \overline{S} : D_h(x,y) \leq \delta\}$, $\Gamma_2(x,\delta) = \{y \in S : D_h(x,y) \leq \delta\}$ are bounded for all $y \in S$, all $x \in \overline{S}$ respectively.

B4) If $\{y^k\} \subset S$ converges to y^* then $D_h(y^*, y^k)$ converges to 0.

B5) If $\{x^k\} \subset \overline{S}$ and $\{y^k\} \subset S$ are sequences such that $\{x^k\}$ is bounded, $\lim_{k\to\infty} y^k = y^*$ and $\lim_{k\to\infty} D_h(x^k, y^k) = 0$ then $\lim_{k\to\infty} x^k = y^*$.

S is called the *zone* of h. It is easy to check that $D_h(x,y) \geq 0$ for all $x \in \overline{S}$, $y \in S$ and $D_h(x,y) = 0$ if and only if $x = y$. We remark that B4 and B5 hold automatically when x^k, y^* are in S, as a consequence of B1, B2 and B3, and so they need to be checked only at points in the boundary ∂S of S. It has been proved in [13] that when $S = \mathbf{R}^n$ a sufficient condition for a convex and differentiable function h to be a Bregman function is $\lim_{\|x\|\to\infty} \frac{h(x)}{\|x\|} = \infty$.

Before presenting examples of Bregman functions, we introduce two subclasses to be used in the sequel.

A Bregman function h is said to be *boundary coercive* if:

B6) If $\{y^k\} \subset S$ is such that $\lim_{k\to\infty} y^k = y \in \partial S$, then $\lim_{k\to\infty} \nabla h(y^k)^t(x - y^k) = -\infty$ for all $x \in S$.

A Bregman function h is said to be *zone coercive* if:

B7) For every $y \in \mathbf{R}^n$ there exists $x \in S$ such that $\nabla h(x) = y$.

B6 will be a key concept in connection with proximal point methods for the following reason. It is clear from B1-B5 that if h is Bregman function with zone S and S' is an open subset of S then h is also a Bregman function with zone S', i.e. we cannot recover S from h. On the other hand, we want to use D_h for penalization purposes, in order to minimize functions on a closed convex set C. The information about the set C in the algorithms considered in the sequel will be encapsulated in D_h, so that C will have to

be recoverable from h. B6 fits this situation, because divergence of ∇h at ∂S makes S univocally determined by h. In all our algorithms we will take C equal to the closure $\overline{S}$ of the zone S of the Bregman function h.

B7 is required in the convergence analyses of the proximal point method for variational inequality problems discussed in section 11 and of the approximate version of the proximal point method discussed in section 15. It is equivalent to Rockafellar's concept of essential smoothness.

We will give now some examples of Bregman functions.

EXAMPLE 9.1: $S = \mathbf{R}^n$, $h(x) = x^t M x$, with $M \in \mathbf{R}^{n\times n}$ symmetric and positive definite. In this case $D_h(x,y) = (x-y)^t M(x-y) = \|x-y\|_M^2$.

EXAMPLE 9.2: $S = \mathbf{R}^n_{++}$, $h(x) = \sum_{j=1}^{n} x_j \log x_j$, extended with continuity to $\partial\mathbf{R}^n_+$ with the convention that $0\log 0 = 0$. In this case $D_h(x,y) = \sum_{j=1}^{n}(x_j \log \frac{x_j}{y_j} + y_j - x_j)$, which is the Kullback-Leibler divergence, widely used in statistics.

EXAMPLE 9.3: $S = \mathbf{R}^n_{++}$, $h(x) = \sum_{j=1}^{n}(x_j^\alpha - x_j^\beta)$ with $\alpha \geq 1$, $0 < \beta < 1$. For $\alpha = 2$, $\beta = \frac{1}{2}$ we get $D_h(x,y) = \|x-y\|^2 + \sum_{j=1}^{n} \frac{1}{2\sqrt{y_j}}(\sqrt{x_j} - \sqrt{y_j})^2$, and for $\alpha = 1$, $\beta = \frac{1}{2}$ we get

$$D_h(x,y) = \sum_{j=1}^{n} \frac{1}{2\sqrt{y_j}}(\sqrt{x_j} - \sqrt{y_j})^2.$$

The Bregman functions of Examples 9.1 and 9.2 are zone and boundary coercive, and the same holds for Example 9.3 for $\alpha > 1$. For $\alpha = 1$ this h is boundary but not zone coercive. h as in Example 9.1 but with $S = \mathbf{R}^n_{++}$ instead of $\mathbf{R}^n$ is neither zone nor boundary coercive.

The following elementary property of Bregman distances follows easily from (9.1) and conditions B1-B5.

PROPOSITION 9.1. *If h is a Bregman function with zone S then*

i) $D_h(x,y) - D_h(x,z) - D_h(z,y) = \langle \nabla h(y) - \nabla h(z), z - x\rangle$ *for all* $x \in \overline{S}; y, z \in S$,

ii) $\nabla_x D_h(x,y) = \nabla h(x) - \nabla h(y)$ *for all* $x, y \in S$,

iii) $D_h(\,\cdot\,, y)$ *is strictly convex for all* $y \in S$.

B6 and B7 are related. We conjecture that B7 implies B6. This is true at least in one significant case, as the following proposition shows. Assume that S is a *box*, i.e. $S = (a_1, b_1) \times \cdots \times (a_n, b_n)$, with $a_j \in \mathbf{R} \cup \{-\infty\}$, $b_j \in \mathbf{R} \cup \{+\infty\}$, $a_j < b_j$ $(1 \leq j \leq n)$, and that h is *separable*, i.e. $h(x) = \sum_{j=1}^n h_j(x_j)$, with $h_j : [a_j, b_j] \to \mathbf{R}$.

PROPOSITION 9.2. *If* S *is a box and* h *is separable then*

i) D_h *is zone coercive if and only if* $\lim_{t \to a_j} h_j(t) = -\infty$, $\lim_{t \to b_j} h_j(t) = +\infty$ *for all* j.

ii) D_h *is boundary coercive if and only if* $\lim_{t \to a_j} h_j(t) = -\infty$ *for all* j *such that* $a_j > -\infty$, *and* $\lim_{t \to b_j} h_j(t) = +\infty$ *for all* j *such that* $b_j < +\infty$.

PROOF:

i) It is immediate that h_j is strictly convex for all j, and also that ∇h is onto if and only if h_j' is onto for all j. Since h_j is strictly convex, h_j' is increasing and the result follows.

ii) Note that $\nabla h(y^k)^t(x - y^k) = \sum_{j=1}^n h_j'(y_j^k)(x_j - y_j^k)$. If h is boundary coercive and $a_i > -\infty$, take $x \in S$, $y_j^k = x_j$ if $j \neq i$, $y_i^k = a_i + 1/k$, so that $\lim_{k \to \infty} y^k = y$ with $y_j = x_j$ for $j \neq i$, $y_i = a_i$. Then $\nabla h(y^k)^t(x - y^k) = h_i'(a_i + 1/k)(x_i - a_i + 1/k)$ and $\lim_{k \to \infty} \nabla h(y^k)^t(x - y^k) = (x_i - a_i) \lim_{k \to \infty} h_i'(a_i + 1/k) = (x_i - a_i) \lim_{t \to a_i} h_i'(t)$. Since $x \in S$, we have $x_i > a_i$ and so $\lim_{t \to a_i} h_i'(t) = -\infty$. We can prove in a similar way that $\lim_{t \to b_i} h_i'(t) = +\infty$ when $b_i < +\infty$. The reverse implication is immediate. ∎

COROLLARY 9.1. *If* S *is a box and* h *is separable then zone coerciveness implies boundary coerciveness.*

10. The proximal point method with Bregman distances.

The problem of interest is:

$$\begin{aligned} &\min\ f(x) \\ &\text{s.t. } x \in \overline{S} \end{aligned} \tag{10.1}$$

with $S \subset \mathbf{R}^n$ open and convex, $\overline{S}$ the closure of S and f convex and continuous on $\overline{S}$. The proximal point method with Bregman distances is defined as:

$$x^0 \in S \tag{10.2}$$

$$x^{k+1} = \operatorname*{argmin}_{x \in \overline{S}} \{f(x) + \lambda_k D_h(x, x^k)\} \tag{10.3}$$

where h is a Bregman function with zone S and λ_k satisfies

$$0 < \lambda_k \leq \tilde{\lambda} \tag{10.4}$$

for some $\tilde{\lambda} > 0$.

We proceed to the convergence analysis. We assume from now on that (10.1) has solutions, so that f is bounded below on $\overline{S}$.

THEOREM 10.1. *If problem (10.1) has solutions and h is boundary coercive with respect to S then the sequence $\{x^k\}$ generated by (10.2),(10.3) converges to a solution x^* of problem (10.1).*

PROOF: We follow the same line as in Theorem 2.1, with D_h instead of $\|\cdot\|^2$.

STEP 1. *The sequence $\{x^k\}$ is well defined and contained in S.*

Let β be a lower bound for f on $\overline{S}$ and $f_k(x) = f(x) + \lambda_k D_h(x, x^k)$. Then $f_k(x) \geq \beta + \lambda_k D_h(x, x^k)$ and it follows from B3 that the level sets of f_k are bounded, so that the minimization in (10.3) reduces to a compact set and the minimum is attained. f_k is strictly convex by convexity of h and Proposition 9.1(iii), so that the minimum is unique and x^{k+1} is uniquely determined.

We prove next that $x^{k+1} \in S$. It is easy to check from (10.3) that x^{k+1} is the only $x \in \overline{S}$ such that

$$\lambda_k \nabla h(x^k) \in \partial(f + \lambda_k h)(x). \tag{10.5}$$

We will show that, under B6, $\partial(f + \lambda_k h)(x) = \emptyset$ for all $x \in \partial S$, which implies, in view of (10.5), that $x^{k+1} \in S$. Take $x \in \partial S$ and assume that there exists $\xi \in \partial(f + \lambda_k h)(x)$. Take $z \in S$ and define

$$y^\ell = (1 - \epsilon_\ell)x + \epsilon_\ell z \tag{10.6}$$

with $\lim_{\ell\to\infty} \varepsilon_\ell = 0$. Then $y^\ell \in S$, by convexity of S, and $\lim_{k\to\infty} y^\ell = x$. So

$$\begin{aligned}\varepsilon_\ell \xi^t(z-x) = \xi^t(y^\ell - x) &\le f(y^\ell) - f(x) + \lambda_k(h(y^\ell) - h(x)) \\ &\le f(y^\ell) - f(x) + \lambda_k \nabla h(y^\ell)^t(y^\ell - x) \\ &\le \varepsilon_\ell(f(z) - f(x)) + \lambda_k \frac{\varepsilon_\ell}{1-\varepsilon_\ell} \nabla h(y^\ell)^t(z - y^\ell) \end{aligned} \tag{10.7}$$

using (10.6) in the first equality, definition of $\partial(f + \lambda_k h)$ in the first inequality, B2 and convexity of h in the second inequality, (10.6) again and convexity of f in the third one. From (10.7)

$$\frac{1-\varepsilon_\ell}{\lambda_k}[f(x) - f(z) + \xi^t(z-x)] \le \nabla h(y^\ell)^t(z - y^\ell). \tag{10.8}$$

Since $\lim_{\ell\to\infty} y^\ell = x \in \partial S$, B6 implies that the right hand side of (10.8) goes to $-\infty$ as ℓ goes to ∞, while the left hand side has a finite limit. This contradiction implies that $\partial(f + \lambda_k h) = \emptyset$ for all $x \in \partial S$ and so $x^{k+1} \in S$.

STEP 2. *$D_h(\overline{x}, x^{k+1}) \le D_h(\overline{x}, x^k) - D_h(x^{k+1}, x^k)$ for all k and every solution $\overline{x}$ of (10.1).*

We use Proposition 9.1(i) with $x = \overline{x}, y = x^k, z = x^{k+1}$ and get

$$D_h(\overline{x}, x^k) - D_h(\overline{x}, x^{k+1}) - D_h(x^{k+1}, x^k) = \left\langle \nabla h(x^k) - \nabla h(x^{k+1}), x^{k+1} - \overline{x} \right\rangle. \tag{10.9}$$

From (10.3)

$$0 \in \partial[f + \lambda_k D_h(\cdot, x^k)](x^{k+1}). \tag{10.10}$$

From (10.10) and Proposition 9.1(ii)

$$\lambda_k[\nabla h(x^k) - \nabla h(x^{k+1})] \in \partial f(x^{k+1}). \tag{10.11}$$

Let $y^k = \lambda_k(\nabla h(x^k) - \nabla h(x^{k+1}))$. From (10.9) and the definition of subgradient

$$D_h(\overline{x}, x^k) - D_h(\overline{x}, x^{k+1}) - D_h(x^{k+1}, x^k) = \frac{1}{\lambda_k}\left\langle y^k, x^{k+1} - \overline{x} \right\rangle \ge \frac{1}{\lambda_k}(f(x^{k+1}) - f(\overline{x})) \tag{10.12}$$

and the result follows because $\overline{x}$ minimizes f on $\overline{S}$.

STEP 3. $\{x^k\}$ *is bounded, and* $\lim_{k\to\infty} x^{j_k} = \hat{x}$ *implies* $\lim_{k\to\infty} x^{j_k+1} = \hat{x}$.

From Step 2 $\{D_h(\overline{x}, x^k)\}$ is decreasing and nonnegative, hence convergent, and $D_h(x^{k+1}, x^k) \leq D_h(\overline{x}, x^k) - D_h(\overline{x}, x^{k+1})$ so that

$$\lim_{k\to\infty} D_h(x^{k+1}, x^k) = 0. \tag{10.13}$$

Since $\{D_h(\overline{x}, x^k)\}$ is decreasing, we have $D_h(\overline{x}, x^k) \leq D_h(\overline{x}, x^0)$ and therefore $\{x^k\}$ is bounded by B3. If $\lim_{k\to\infty} x^{j_k} = \hat{x}$ for a subsequence $\{x^{j_k}\}$ of $\{x^k\}$ then $\lim_{k\to\infty} x^{j_k+1} = \hat{x}$ by B5.

STEP 4. *The sequence* $\{x^k\}$ *has cluster points all of which are solutions of (10.1).*

Take a solution $\overline{x}$ of (10.1). Let $\hat{x}$ be a cluster point of $\{x^k\}$ and $\{x^{j_k}\}$ a subsequence of $\{x^k\}$ such that $\lim_{k\to\infty} x^{j_k} = \hat{x}$. Existence of $\hat{x}$ follows from Step 3 which also ensures that $\lim_{k\to\infty} x^{j_k+1} = \hat{x}$. From (10.12)

$$\begin{aligned} 0 \leq \frac{1}{\overline{\lambda}}(f(x^{j_k+1}) - f(\overline{x})) &\leq \frac{1}{\lambda_k}(f(x^{j_k+1}) - f(\overline{x})) \\ &\leq D_h(\overline{x}, x^{j_k}) - D_h(\overline{x}, x^{j_k+1}) - D_h(x^{j_k+1}, x^{j_k}) \underset{k\to\infty}{\longrightarrow} 0 \end{aligned} \tag{10.14}$$

using convergence of $\{D_h(\overline{x}, x^k)\}$ and B5. From (10.14), taking limits as k goes to ∞, $f(\hat{x}) = f(\overline{x})$. Since $\overline{S}$ is closed and $\{x^k\} \subset \overline{S}$, we have $\hat{x} \in \overline{S}$ and so $\hat{x}$ solves (10.1).

In order to complete the proof we need a Fejér convergence theorem for Bregman distances, which in fact holds, but we can proceed directly: let $\hat{x}$ be a cluster point of $\{x^k\}$ and take a subsequence $\{x^{j_k}\}$ of $\{x^k\}$ such that $\lim_{k\to\infty} x^{j_k} = \hat{x}$. Then by B4 $\lim_{k\to\infty} D_h(\hat{x}, x^{j_k}) = 0$. By Step 4 $\hat{x}$ solves (10.1) and so by Step 2 $\{D_h(\hat{x}, x^k)\}$ is a nonnegative and decreasing sequence with a subsequence converging to 0. It follows that the whole sequence converges to 0 and by B4 again we get $\lim_{k\to\infty} x^k = \hat{x}$. ∎

We remark that for this case we cannot use firm nonexpansiveness instead of Fejér convergence. If we define $P_k: S \to S$ as

$$P_k(x) = \operatorname*{argmin}_{y\in\mathbf{R}^n} \{f(y) + \lambda_k D_h(y, x)\}$$

it is not true that

$$D_h(P_k(x), P_k(y)) \leq D_h(x, y) - D_h(x - y, P_k(x) - P_k(y)) \text{ for all } x, y \in S$$

and the same happens with

$$D_h(P_k(x), P_k(y)) \leq D_h(x, y) - D_h(x - P_k(x), y - P_k(y)) \text{ for all } x, y \in S$$

which would be the natural extensions of firm nonexpansiveness to Bregman distances. In fact P_k is not even nonexpansive with respect to D_h, i.e. it is not true that $D_h(P_k(x), P_k(y)) \leq D_h(x, y)$ for all $x, y \in S$.

We present now an application of the proximal point method with Bregman distances, similar to the case discussed in section 7. We show that the proximal point method allows us to recover another augmented Lagrangian method, namely Bertsekas' exponential multipliers method. We consider again

$$\begin{aligned} &\min f(x) \\ &\text{s.t. } g_i(x) \leq 0 \qquad (1 \leq i \leq m) \end{aligned} \tag{10.15}$$

with $f, g_i : \mathbf{R}^n \to \mathbf{R}$ convex and differentiable. The augmented Lagrangian in this case is

$$L_\alpha(x, y) = f(x) + \frac{1}{\alpha} \sum_{i=1}^{m} y_i e^{\alpha g_i(x)} \tag{10.16}$$

with $\alpha > 0$. This L_a is as many times differentiable in x as the data f, g_i, which was not the case for L_α as given by (7.5). The exponential multipliers method generates a sequence $\{x^k\} \subset \mathbf{R}^n$, $\{y^k\} \subset \mathbf{R}^m$ through

$$y^0 \in \mathbf{R}^m_+ \tag{10.17}$$

and, given y^k,

$$x^k = \operatorname*{argmin}_{x \in \mathbf{R}^n} L_\alpha(x, y^k) \tag{10.18}$$

$$y_i^{k+1} = y_i^k \; e^{\alpha g_i(x^k)}. \tag{10.19}$$

As before, we consider the dual objective $\varphi(y) = \min_{x \in \mathbf{R}^n} \{f(x) + \sum_{i=1}^{m} y_i g_i(x)\} = \min_{x \in \mathbf{R}^n} L(x, y)$, defined for $y \geq 0$ and the sequence $\{y^k\}$ defined by (10.17)-(10.19) coincides with the sequence generated by the proximal point method with Bregman distances (with $\lambda_k = \frac{1}{\alpha}$, $h(x) = \sum_{j=1}^{n} x_j \log x_j$), applied to $\min\{-\varphi(y)\}$ s.t. $y \geq 0$.

THEOREM 10.2. *Let $\{g^k\}$ be the sequence generated by (10.17)-(10.19), and $\{\overline{y}^k\}$ be the sequence obtained through*

$$\overline{y}^{k+1} = \operatorname*{argmin}_{y \geq 0} \{-\varphi(y) + \frac{1}{\alpha} D_h(y, y^k)\} \tag{10.20}$$

with $h: \mathbf{R}^n_+ \to \mathbf{R}$ given by $h(x) = \sum_{j=1}^{n} x_j \log x_j$. If $y^0 = \overline{y}^0$ then $y^k = \overline{y}^k$ for all k.

PROOF: By induction. True for $k = 0$. Assume $y^k = \overline{y}^k$. For the h under consideration (10.11) becomes

$$u^k \in \partial(-\varphi)(\overline{y}^{k+1}) \tag{10.21}$$

with

$$u_i^k = \frac{1}{\alpha} \log\left(\frac{y_i^k}{\overline{y}_{i+1}^k}\right). \tag{10.22}$$

By (10.22) and definition of ∂f

$$\frac{1}{\alpha} \sum_{i=1}^{m} \log\left(\frac{y_i^k}{\overline{y}_i^{k+1}}\right) \leq \varphi(\overline{y}^{k+1}) - \varphi(u) \text{ for all } u \geq 0. \tag{10.23}$$

Since (10.23) determines $\overline{y}^{k+1}$ uniquely, it suffices to check that (10.23) holds with y^{k+1} substituting for $\overline{y}^{k+1}$, i.e.

$$\varphi(y^{k+1}) - \frac{1}{\alpha} \sum_{i=1}^{m} \log\left(\frac{y_i^k}{y_i^{k+1}}\right)\left(y_i^{k+1} - u_i\right) \geq \varphi(u) \text{ for all } u \geq 0. \tag{10.24}$$

From (10.16)

$$\nabla_x L_\alpha(x,y) = \nabla f(x) + \sum_{i=1}^{m} y_i e^{\alpha g_i(x)} \nabla g_i(x) \tag{10.25}$$

and from (10.18), (10.19)

$$\begin{aligned} 0 = \nabla_x L_\alpha(x^k, y^k) &= \nabla f(x^k) + \sum_{i=1}^{m} y_i^k e^{\alpha g_i(x^k)} \nabla g_i(x^k) \\ &= \nabla f(x^k) + \sum_{i=1}^{m} y_i^{k+1} \nabla g_i(x^k) = \nabla_x L(x^k, y^{k+1}) \end{aligned} \tag{10.26}$$

so that x^k minimizes $L(\cdot, y^{k+1})$, i.e.

$$\varphi(y^{k+1}) = f(x^k) + \sum_{i=1}^{m} y_i^{k+1} g_i(x^k). \tag{10.27}$$

Now we evaluate the second term in the left hand side of (10.24). From (10.19)

$$\frac{1}{\alpha} \log\left(\frac{y_i^k}{y_i^{k+1}}\right) = g_i(x^k). \tag{10.28}$$

Substituting (10.28) in (10.24):

$$\begin{aligned} &\varphi(y^{k+1}) - \frac{1}{\alpha} \sum_{i=1}^{m} \log\left(\frac{y_i^k}{y_i^{k+1}}\right)\left(y_i^{k_1} - u_i\right) = f(x^k) + \sum_{i=1}^{m} u_i g_i(x) \\ &= L(x^k, u) \geq \min_{x \in \mathbf{R}^n} L(x,u) = \varphi(u). \end{aligned} \tag{10.29}$$

So (10.24) holds and $y^{k+1} = \overline{y}^{k+1}$. ∎

As in section 7, our convergence theorem for the proximal point method guarantees that $\{y^k\}$ converges to a maximizer of φ on $\overline{S} = \mathbf{R}_+^n$, i.e. to an optimal vector of Karush-Kuhn-Tucker multipliers for problem (10.15). It also follows from (10.26) that all cluster points of $\{x^k\}$ (if any) are solutions of (10.15), via standard convex duality results, but in this case also convergence of $\{x^k\}$ does not follow directly from the proximal point convergence

theory, which deals only with $\{y^k\}$, and demands some additional hypotheses, like a Slater condition on problem (10.15). This approach provides a wider and easier convergence analysis for the exponential multiplier method.

We end this section with an interesting property of the proximal point method with Bregman functions applied to linear or quadratic programming, namely that the limit of the generated sequence is the solution of the problem which is closest, in the sense of D_h, to the initial iterate x^0. Consider problem P:

$$\min f(x) = \frac{1}{2}x^t Q x + c^t x + \gamma \tag{10.30}$$

$$\text{s.t. } Ax = b \tag{10.31}$$

$$x \geq 0 \tag{10.32}$$

with Q symmetric and positive semidefinite, $A \in \mathbf{R}^{m\times n}$, $b \in \mathbf{R}^m$, $c \in \mathbf{R}^n$ and $\gamma \in \mathbf{R}$. The proximal point method can be applied to problem P, defining $\overline{f}$ as:

$$\overline{f}(x) = \begin{cases} f(x) & \text{if } x \text{ satisfies (10.31)-(10.32)} \\ +\infty & \text{otherwise} \end{cases}$$

and using (2.2) or (10.3) with $\overline{f}$ substituting for f. The property mentioned above holds when we use the proximal point method with a boundary coercive Bregman function h with zone $S = \mathbf{R}^n_{++}$, which excludes the use of (2.2), corresponding to $h(x) = 1/2\,\|x\|^2$ whose zone is $\mathbf{R}^n$.

Using $\overline{f}$, subproblem (10.3) becomes

$$\min f(x) + \lambda_k D_h(x, x^k) \tag{10.33}$$

$$\text{s.t. } Ax = b \tag{10.34}$$

$$x \geq 0. \tag{10.35}$$

If h is boundary coercive with respect to $S = \mathbf{R}^n_{++}$ then $x^{k+1} > 0$ by Step 1 of Theorem 10.1, so that (10.35) is superfluous and $x^{\ell+1}$ solves

$$\min f(x) + \lambda_\ell D_h(x, x^\ell) \tag{10.36}$$

$$\text{s.t. } Ax = b \tag{10.37}$$

whose Karush-Kuhn-Tucker conditions are:

$$\lambda_\ell[\nabla h(x^{\ell+1}) - \nabla h(x^\ell)] + Qx^{\ell+1} + c + A^t u^\ell = 0 \tag{10.38}$$

for some $u^\ell \in \mathbf{R}^m$. Dividing (10.38) by λ_ℓ, summing from $\ell = 0$ to k and defining $v^k = \sum_{\ell=o}^k (\lambda_\ell)^{-1} x^{\ell+1}$, $\eta_k = \sum_{\ell=0}^k (\lambda_\ell)^{-1}$ and $w^k = \sum_{\ell=0}^k (\lambda_\ell)^{-1} u^\ell$ we get

$$\nabla h(x^{k+1}) - \nabla h(x^0) + Qv^k + \eta_k c + A^t w^k = 0 \tag{10.39}$$

which allows us to prove the following intermediate optimality property:

PROPOSITION 10.1. *If h is boundary coercive with respect to $\mathbf{R}^n_{++}$, then x^{k+1} is the solution of*

$$\min D_h(x, x^0) \tag{10.40}$$

$$\text{s.t. } Qx = Qx^{k+1} \tag{10.41}$$

$$c^t x = c^t x^{k+1} \tag{10.42}$$

$$Ax = b \tag{10.43}$$

$$x \geq 0. \tag{10.44}$$

PROOF: . Since $x^{k+1} > 0$ by Step 1 of Theorem 10.1, (10.44) is superfluous. x^{k+1} trivially satisfies (10.41)-(10.43), and the Karush-Kuhn-Tucker conditions of problem (10.40)-(10.43), sufficient by convexity of $D_h(\cdot, x^k)$ and linearity of (10.41)-(10.43), are precisely (10.39). ∎

Now we use a result due to Hoffman [17], which implies that if $\{p^k\} \subset \mathbf{R}^q$ converges to p, and the polyhedra $V_k = \{x \in \mathbf{R}^n : Hx \leq p^k\}$, $V = \{x \in \mathbf{R}^n : Hx \leq p\}$, with $H \in \mathbf{R}^{q \times n}$, are nonempty, then for all $\overline{x} \in V$ there exists $\overline{x}^k \in V_k$ such that $\overline{x} = \lim_{k \to \infty} \overline{x}^k$. Let V_k be the feasible set for (10.41)-(10.44), $x^* = \lim_{k \to \infty} x^k$, and $V = \{x \in \mathbf{R}^n : Qx = Qx^*, c^t x = c^t x^*, Ax = b, x \geq 0\}$. V is nonempty because $x^* \in V$. Take any $\overline{x}$ in V and let $\{\overline{x}^k\}$ be the sequence resulting from Hoffman's result. By Proposition 10.1

$$D_h(x^{k+1}, x^0) \leq D_h(\overline{x}^k, x^0). \tag{10.45}$$

Taking limits as $k \to \infty$ in (10.45) and using continuity of D_h in its first variable, we conclude that $D_h(x^*, x^0) \leq D_h(\overline{x}, x^0)$. So we have proved:

PROPOSITION 10.2. *x^* solves $\min D_h(x, x^0)$ s.t. $x \in V$.*

We claim now that V is precisely the set T of solutions of P. Clearly $V \subset T$: for any $x \in V$ we have $Qx = Qx^*$, $c^t x = c^t x^*$ implying $f(x) = f(x^*)$ and, since x^* solves P by Theorem 10.1, so does x, i.e. $x \in T$. We prove the remaining inclusion in the following proposition.

PROPOSITION 10.3. *If x is a solution of P then $x \in V$.*

PROOF: Since both x and x^* are solutions of P, f is constant in the segment between x and x^* (otherwise, by convexity of f, there is a point in the segment with strictly lower functional value). It follows that $0 = \langle \nabla f(x) - \nabla f(x^*), x - x^* \rangle = (x - x^*)^t Q(x - x^*)$, implying, by symmetry and positive semidefiniteness of Q,

$$Qx = Qx^*. \tag{10.46}$$

Since $f(x) = f(x^*)$, it follows from (10.46) that

$$c^t x = c^t x^*. \tag{10.47}$$

(10.46) and (10.47) imply that $x \in V$. ∎

We have shown that $T = V$, which, together with Proposition 10.2, establishes

THEOREM 10.3. *If h is boundary coercive with respect to $S = R^n_{++}$ and x^* is the limit of the sequence generated by the proximal point method with D_h applied to problem P then x^* is the solution of*

$$\min D_h(x, x^k)$$

$$s.t. \quad x \in U$$

where U is the solution set of problem P.

We remark that this result depends upon the fact that x^k is strictly positive for all k, so that (10.35) is superfluous for the subproblem. Strict positivity of x^k results from boundary coerciveness of h (through Step 1 of Theorem 10.1). This result does not hold for the standard proximal point method (i.e. with $h(x) = 1/2 \|x\|^2$), because in such a

case x^k can have zero components. Observe that for $Q = 0$ problem P reduces to a linear programming problem.

Extension of Theorem 10.3 to convex programming problems other than quadratic seems difficult, because in order to use Hoffman's result we need that the solution set of the problem be a polyhedron. Note that by Step 1 of Theorem 10.1 the proximal point method with boundary coercive Bregman functions with zone $S = R^n_{++}$ is an interior point method. Results similar to Theorem 10.3 have been established for other interior point methods for linear programming (see [1]).

11. The proximal method with Bregman functions for the variational inequality problem.

The natural extension of the problem $\min f(x)$ s.t. $x \in C$ (with $C \subset \mathbf{R}^n$ closed and convex) to monotone operators is the so called variational inequality problem, defined in the following way:

DEFINITION 11.1: Given $T : \mathbf{R}^n \to \mathcal{P}(\mathbf{R}^n)$ maximal monotone and $C \subset \mathbf{R}^n$ closed and convex, the problem VIP(T, C) consists of finding $z \in C$ such that there exists $u \in T(z)$ satisfying

$$\langle u, x - z \rangle \geq 0 \tag{11.1}$$

for all $x \in C$.

When $T = \partial f$ with f convex, we have $0 \leq \langle u, x - z \rangle \leq f(x) - f(z)$ for all $x \in C$ and therefore z minimizes f on C. Variational inequality problems arise also in other contexts; see [24].

In the same way as the standard proximal point method of section 4 can be used to find zeroes of maximal monotone operators, we can use the proximal point method with a boundary coercive Bregman with zone S to solve VIP$(T, \overline{S})$. The algorithm can be written as:

$$x^0 \in S \tag{11.2}$$

Find x^{k+1} such that

$$0 \in [\lambda_k \nabla_x D_h(\cdot, x^k) + T](x^{k+1}). \tag{11.3}$$

(11.3) is equivalent to

$$x^{k+1} \in [\nabla_x D_h(\cdot, x^k) + \frac{1}{\lambda_k} T]^{-1}(x^k) \tag{11.4}$$

and is also equivalent, by Proposition 9(ii), to

$$\lambda_k[\nabla h(x^k) - \nabla h(x^{k+1})] \in T(x^{k+1}). \tag{11.5}$$

Two additional condition must be imposed on T in order to establish convergence, namely paramonotonicity and pseudomonotonicity.

DEFINITION 11.2: $T : \mathbf{R}^n \to \mathcal{P}(\mathbf{R}^n)$ is paramonotone if it is maximal monotone and $\langle u - v, x - y\rangle = 0$ with $u \in T(x)$, $v \in T(y)$ implies $u \in T(y)$ and $v \in T(x)$.

It has been proved in [9, Lemma 1] that ∂f is paramonotone when f is convex.

DEFINITION 11.3: $T : \mathbf{R}^n \to \mathcal{P}(\mathbf{R}^n)$ is pseudomonotone if for any sequence $\{x^k\}$ converging to a point $\bar{x}$, for any $u^k \in T(x^k)$ such that $\limsup_{k\to\infty}\langle u^k, x^k - \bar{x}\rangle \geq 0$ and for any y there exists $\bar{u} \in T(\bar{x})$ such that $\liminf_{k\to\infty}\langle u^k . x^k - y\rangle \geq \langle \bar{u}, \bar{x} - y\rangle$.

It can be shown that when $T = \partial f$ for some convex f then T is pseudomonotone. The same holds true is T is point-to-point anc continuous.

We will sketch next the convergence analysis of algorithm (11.2)-(11.3). A full proof can be found in [6].

THEOREM 11.1. *Let $C \subset \mathbf{R}^n$ be a closed convex set with nonempty interior, T a paramonotone and pseudomonotone operator such that $T(x) \neq \emptyset$ for all $x \in C$ and h a boundary coercive Bregman function whose zone S is the interior of C. If VIP(T, C) has solutions and either*

a) h is zone coercive, or

b) $\sup\{\langle v, x - y\rangle : y \in \mathbf{R}^n, v \in T(y)\} < \infty$ for all $x \in C$, then the sequence $\{x^k\}$ generated by (11.2)-(11.3) converges to a solution of VIP(T, C).

PROOF: We follow the steps of Theorem 10.1. The first difficulty arises in Step 1. We cannot use anymore a compactness argument to ensure existence of x^{k+1}. In view of (11.2) it suffices to prove that $\lambda_k \nabla_x D_h(\cdot, x^k) + T$ is onto. For the standard proximal point method, i.e. for $h(x) = 1/2\,\|x\|^2$, this follows, as discussed in section 4, from Minty's

theorem. In this case the following result can be proved: if either of hypotheses (a) and (b) of the theorem holds, then there exists $y \in S$ such that 0 belongs to $\lambda_k \nabla_x D_h(y, x^k) + T(y)$ and therefore existence of x^{k+1} is guaranteed. Boundary coerciveness ensures that x^{k+1} belongs to the interior of C. Hypothesis (b) holds in several important cases, e.g. when $T = \partial f$ with f convex, or when T is strongly monotone (i.e. $\langle u - v, x - y \rangle \geq \beta \|x - y\|^2$ for some $\beta > 0$ and all x, y, $u \in T(x)$, $v \in T(y)$). Uniqueness of x^{k+1} follows from strict convexity of h and the fact that x^{k+1} belongs to S follows from boundary coerciveness of h, as in the proof of Theorem 10.1.

For Step 2, we take a solution $\overline{x}$ of VIP(T, C), use T instead of ∂f, and follow the proof of Theorem 10.1 up to (10.11), obtaining

$$D_h(\overline{x}, x^k) - D_h(\overline{x}, x^{k+1}) - D_h(x^{k+1}, x^k) = \langle u^k, x^{k+1} - \overline{x} \rangle \tag{11.6}$$

with $u^k \in T(x^{k+1})$. Now we take $v \in T(\overline{x})$ such that $\langle v, x^{k+1} - \overline{x} \rangle \geq 0$, which exists because $\overline{x}$ solves VIP(T, C), and by monotonicity of T we have

$$\langle u^k, x^{k+1} - \overline{x} \rangle \geq \langle v, x^{k+1} - \overline{x} \rangle \geq 0 \tag{11.7}$$

because $u^k \in T(x^{k+1})$. From (11.6) and (11.7):

$$D_h(\overline{x}, x^k) - D_h(\overline{x}, x^{k+1}) - D_h(x^{k+1}, x^k) \geq \langle u^k, x^{k+1} - \overline{x} \rangle \geq 0 \tag{11.8}$$

and then Step 2 holds. Step 3 is proved exactly as in Theorem 10.1, and it follows then that the leftmost expression of (11.8) converges to 0, i.e.

$$\lim_{k \to \infty} \langle u^k, x^{k+1} - \overline{x} \rangle = 0 \tag{11.9}$$

for all solution $\overline{x}$ of VIP(T, C), where $u^k \in T(x^{k+1})$.

Now we proceed to Step 4. Given a cluster point $\hat{x}$ of $\{x^k\}$ we take a subsequence$\{x^{j_k}\}$ of $\{x^k\}$ which converges to $\hat{x}$. It can be proved from (11.9) that, if $T$is pseudomonotone then there exists $\hat{u} \in T(\hat{x})$ such that

$$\langle \hat{u}, \hat{x} - \overline{x} \rangle = 0. \tag{11.10}$$

(if $\hat{x}$ does not belong to S, such an $\hat{u}$ may not be the limit of a convergent subsequence of $\{u^{j_k}\}$, which may even be unbounded).

Let $\overline{x}$ be a solution of VIP(T, C). Then there exists $\overline{u} \in T(\overline{x})$ such that

$$\langle \overline{u}, x - \overline{x} \rangle \geq 0 \tag{11.11}$$

for all $x \in \overline{S}$ and therefore

$$\langle \overline{u}, \hat{x} - \overline{x} \rangle \geq 0. \tag{11.12}$$

By (11.10)-(11.12) and monotonicity of T, since $\hat{u} \in T(\hat{x})$, we get

$$0 \leq \langle \overline{u}, \hat{x} - \overline{x} \rangle \leq \langle \hat{u}, \hat{x} - \overline{x} \rangle = 0 \tag{11.13}$$

which implies

$$0 = \langle \overline{u}, \hat{x} - \overline{x} \rangle = \langle \hat{u}, \hat{x} - \overline{x} \rangle \tag{11.14}$$

and therefore $0 = \langle \hat{u} - \overline{u}, \hat{x} - \overline{x} \rangle$. By paramonotonicity of T we conclude that $\overline{u} \in T(\hat{x})$. Then, using (11.11)-(11.14), for any $x \in \overline{S}$,

$$\langle \overline{u}, x - \hat{x} \rangle = \langle \overline{u}, x - \overline{x} \rangle + \langle \overline{u}, \overline{x} - \hat{x} \rangle = \langle \overline{u}, x - \overline{x} \rangle \geq 0. \tag{11.15}$$

Since $\overline{u} \in T(\hat{x})$, we have shown that $\hat{x}$ solves VIP(T, C). Since $\hat{x}$ is any cluster point of $\{x^k\}$, we conclude that all cluster points of $\{x^k\}$ are solutions of VIP(T, C). Using the same argument as in the end of Theorem 10.1, we conclude that there is only one cluster point, i.e. the whole sequence converges to a solution of VIP(T, C). ∎

12. φ-divergences.

In this section we discuss another class of "distances", which will be called $d_\varphi(\cdot, \cdot)$, defined on the positive orthant of $\mathbf{R}^n$ (i.e. corresponding to $S = \mathbf{R}^n_+$ of section 10). Take $\varphi: \mathbf{R}_{++} \to \mathbf{R}$, convex and thrice continuously differentiable, satisfying

$$\varphi(1) = \varphi'(1) = 0, \;\; \varphi''(1) > 0, \;\; \lim_{t \to 0} \varphi'(t) = -\infty. \tag{12.1}$$

DEFINITION 12.1: If φ satisfies (12.1) and the hypotheses above then $d_\varphi: \mathbf{R}^n_{++} \times \mathbf{R}^n_{++} \to \mathbf{R}$ defined by

$$d_\varphi(x,y) = \sum_{j=1}^{n} y_i \varphi\left(\frac{x_i}{y_j}\right) \tag{12.2}$$

is said to be a φ-divergence.

The next properties follow easily from Definition 12.1 and (12.1),(12.2).

PROPOSITION 12.1.

i) $d_\varphi(x,y) \geq 0$ *for all* $x, y \in \mathbf{R}^n_{++}$,

ii) $d_\varphi(x,y) = 0$ *iff* $x = y$,

iii) the level sets of $d_\varphi(\cdot, y)$ *are bounded for all* $y \in \mathbf{R}^n_{++}$,

iv) the level sets of $d_\varphi(x, \cdot)$ *are bounded for all* $x \in \mathbf{R}^n_{++}$,

v) $d_\varphi(x,y)$ *is jointly convex on* x, y, *and strictly convex in* x,

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

EXAMPLE 12.1: $\varphi_1(t) = t \log t - t + 1$. Then

$$d_{\varphi_1}(t) = \sum_{j=1}^{n} (x_j \log \frac{x_j}{y_j} + y_j - x_j) \tag{12.3}$$

i.e. d_{φ_1} is the Kullback-Leibler divergence and can therefore be extended to $\mathbf{R}^n_+ \times \mathbf{R}^n_{++}$. Up to additive linear terms in h and multiplicative constants in φ, h, the pair (φ_1, h_1) with $h_1(x) = \sum_{j=1}^{n} x_j \log x_j$ is the only pair (φ, h) such that $d_\varphi = D_h$.

EXAMPLE 12.2: $\varphi_2(t) = t - \log\ t - 1$. Then

$$d_{\varphi_2}(x,y) = d_{\varphi_1}(y,x). \tag{12.4}$$

EXAMPLE 12.3: $\varphi_3(t) = (\sqrt{t} - 1)^2$. Then

$$d_{\varphi_3}(x,y) = \sum_{j=1}^{n} (\sqrt{x_j} - \sqrt{y_j})^2. \tag{12.5}$$

φ-divergences have been recently extended in [2] to other open polyhedra besides $\mathbf{R}^n_{++}$. Let $E = \{x \in \mathbf{R}^n : Ax < b\}$, where b belongs to $\mathbf{R}^m$, $A \in \mathbf{R}^{m\times n}$ has full column rank and E has nonempty interior (which implies $m \geq n$). Given φ satisfying (12.1) we define $\Delta_\varphi : E \times E \to \mathbf{R}$ as:

$$\Delta_\varphi(x,y) = \sum_{i=1}^{m}(b_i - \langle a^i, y\rangle)\varphi\left[\frac{b_i - \langle a^i, x\rangle}{b_i - \langle a^i, y\rangle}\right] \tag{12.6}$$

where a^i $(1 \leq i \leq m)$ are the rows of A. We will not discuss this extension in the sequel, but most results on φ-divergences given in sections 13-15 can be extended to this situation.

13. The proximal point method with φ-divergences.

Now the problem of interest is

$$\begin{aligned} &\min\ f(x) \\ &\text{s.t.}\quad x \geq 0 \end{aligned} \tag{13.1}$$

with $f: \mathbf{R}^n \to \mathbf{R}$ convex, and the proximal point method with φ-divergences for problem (13.1) generates a sequence $\{x^k\} \subset \mathbf{R}^n$ given by

$$x^0 > 0 \tag{13.2}$$

$$x^{k+1} = \operatorname*{argmin}_{x\in\mathbf{R}^n}\{f(x) + \lambda_k d_\varphi(x, x^k)\} \tag{13.3}$$

with λ_k satisfying $0 < \lambda_k \leq \tilde{\lambda}$ for some $\tilde{\lambda} > 0$. The optimality conditions for (13.3) say that $u_k \in \partial f(x^{k+1})$ with $u^k_j = -\lambda_k\varphi'\left(\frac{x^{k+1}_j}{x^k_j}\right)$. If f is differentiable, x^{k+1} is the solution x of the system

$$\nabla f(x)_j + \lambda_k\varphi'\left(\frac{x_j}{x^k_j}\right) = 0. \tag{13.4}$$

(13.4) is a nonlinear system of n equations in the n unknowns $x_1, \ldots, x_n$.

As in the previous cases, when (13.1) has solutions, the minimization problem in (13.3) reduces to a compact subset ($\lim_{t\to\infty}\varphi(t) = \infty$ by (12.1)), guaranteeing existence of x^{k+1},

and uniqueness of x^{k+1} follows from convexity of f and Proposition 12.1(v), so that Step 1 in the convergence analysis of sections 2, 4 and 10 holds for this case. The problem is that the sequence $\{x^k\}$ generated by (13.2)-(13.3) is not Fejér convergent to the set of solutions of (13.1) with respect to d_φ, i.e., we may have a solution $\overline{x}$ of (13.1) such that $d_\varphi(\overline{x}, x^k) > d_\varphi(\overline{x}, x^{k+1})$. The convergence analysis of the proximal point method with φ-divergences is much harder and we will just give an outline of the convergence proof.

In the first place we will relax the notion of Fejér convergence to quasi-Fejér convergence.

DEFINITION 13.1: A sequence $\{y^k\} \subset \mathbf{R}^n_{++}$ is quasi-Fejér convergent to a set $U \subset \mathbf{R}^n_{++}$ with respect to a φ-divergence d_φ if for each $u \in U$ there exists a sequence $\{\varepsilon_k\} \subset \mathbf{R}_{++}$ such that $\sum_{k=0}^{\infty} \varepsilon_k < \infty$ and, for all $k \geq 0$

$$d_\varphi(u, y^{k+1}) \leq d_\varphi(u, y^k) + \varepsilon_k. \tag{13.5}$$

Proposition 2.1 holds with quasi-Fejér convergence substituing for Fejér convergence, as the next proposition shows.

PROPOSITION 13.1. *If $\{y^k\} \subset \mathbf{R}^n_{++}$ is quasi-Fejér convergent to $U \subset \mathbf{R}^n_{++}$ with respect to a φ-divergence d_φ then $\{y^k\}$ is bounded. If a cluster point $\overline{y}$ of $\{y^k\}$ belongs to U then $\overline{y} = \lim_{k\to\infty} y^k$.*

PROOF:

i) Take $u \in U$ and ε_k as in Definition 13.1. Let $\beta = \sum_{k=0}^{\infty} \varepsilon_k$. Then $d_\varphi(u, y^k) \leq d_\varphi(u, y^0) + \sum_{j=0}^{k-1} \varepsilon_k \leq d_\varphi(y, y^0) + \beta$ and the result follows from Proposition 12.1(iv).

ii) Let $\overline{y} \in U$ be a cluster point of $\{y^k\}$ and $\{y^{j_k}\}$ a subsequence of $\{y^k\}$ such that $\lim_{k\to\infty} y^{j_k} = \overline{y}$. Given any $\delta > 0$, take $\overline{k}$ such that $\sum_{k=\overline{k}}^{\infty} \varepsilon_k \leq \frac{\delta}{2}$ and $\hat{k}$ such that $j_{\hat{k}} > \overline{k}$ and $d_\varphi(\overline{y}, y^{j_{\hat{k}}}) \leq \frac{\delta}{2}$ ($\hat{k}$ exists by Proposition 12.1(vi)). Then for $k > j_{\hat{k}}$ we have

$$d_\varphi(\overline{y}, y^k) \leq d_\varphi(\overline{y}, y^{j_{\hat{k}}}) + \sum_{\ell=j_{\hat{k}}}^{k} \varepsilon_\ell \leq \frac{\delta}{2} + \sum_{\ell=\overline{k}}^{\infty} \varepsilon_\ell \leq \frac{\delta}{2} + \frac{\delta}{2} = \delta. \tag{13.6}$$

Since δ is arbitrary, (13.6) implies $\lim_{k\to\infty} d_\varphi(\overline{y}, y^k) = 0$ and then $\lim_{k\to\infty} y^k = \overline{y}$ by Proposition 12.1(vi). ∎

It turns out to be the case that the sequence $\{x^k\}$ generated by (13.2)-(13.3) is not even quasi-Fejér convergent to the set of solutions of (13.1) with respect to d_φ (in fact, if a solution $\overline{x}$ of (13.2) belongs to $\partial\mathbf{R}^n_+$ then $d_\varphi(\overline{x}, x^k)$ is not even defined). However, under some conditions on φ, $\{x^k\}$ is quasi-Fejér convergent to the set of solutions of (13.1) with respect to the Kullback-Leibler divergence, i.e. with respect to d_ψ with $\psi(t) = \varphi_1(t) = t\log t - t + 1$, which admits points in $\partial\mathbf{R}^n_+$ as a first argument and for which Proposition 13.1 holds with $U \subset \mathbf{R}^n_+$, $\overline{y} \in \mathbf{R}^n_+$. More precisely, consider the sequence $\{x^k\}$ generated by (13.2)-(13.3) with some d_φ, let U be be the set of solutions of (13.1), take $\overline{x} \in U$ and define δ_k as

$$\delta_k = d_\psi(\overline{x}, x^{k+1}) - d_\psi(\overline{x}, x^k). \tag{13.7}$$

PROPOSITION 13.3.

i) If $\varphi'(t) \le \varphi''(1)\log\ t$ and there exists $\hat{\lambda}$ such that $\lambda_k \ge \hat{\lambda} > 0$ for all k then $\sum_{k=0}^{\infty} \delta_k < \infty$ (and so $\{x^k\}$ is quasi-Fejér convergent to U).

ii) If $\frac{\varphi''(1)}{t} \le \varphi'(t) \le \varphi''(1)\log\ t$ then $\delta_k \le 0$ (and so $\{x^k\}$ is Fejér convergent, and a fortiori quasi-Fejér convergent, to U).

The proof of this proposition is rather involved and can be found in [**19**, Prop. 4.1] and [**21**, Prop. 3]. We remark that $\varphi_1, \varphi_2, \varphi_3$ of Examples 13.1, 13.2 and 13.3 satisfy the inequalities in the hypotheses of Proposition 13.2. In view of the first statement of Proposition 13.1 and of Proposition 13.2, the sequence $\{x^k\}$ is bounded when $\varphi'(t) \le \varphi''(1)\log\ t$ and $\lambda_k \ge \hat{\lambda} > 0$ or when $\frac{\varphi''(1)}{t} \le \varphi'(t) \le \varphi''(1)\log\ t$ and by Proposition 13.1, it suffices to prove that the cluster points of $\{x^k\}$ belong to U to ensure convergence of $\{x^k\}$ to a solution of (13.1). We mention that if the sequence $\{x^k\}$ converges to a point $\overline{x} > 0$ and f is differentiable then it is immediate that $\overline{x}$ solves (13.1), because from (13.4).

$$\nabla f(x^{k+1}) = -\lambda_k \varphi'\left(\frac{x_j^{k+1}}{x_j^k}\right)$$

and $\lim_{k\to\infty} x_j^{k+1} = \lim_{k\to\infty} x_j^k = \overline{x}_j > 0$ so that $\lim_{k\to\infty} \varphi'\left(\frac{x_j^{k+1}}{x_j^k}\right) = \varphi'(1) = 0$ and therefore $\nabla f(\overline{x})_j = 0$ (using $\lambda_k \leq \tilde{\lambda}$) for all j, i.e. $\overline{x}$ is the unrestricted minimizer of f, but in general $\overline{x}$ is just a cluster point which can have some zero components and f is not differentiable. Under the hipotheses of Proposition 13.1(i) and (ii) it can be proved that all cluster points of $\{x^k\}$ belong to U, but the proofs ([**19**, Prop. 4.3] and [**21**, Prop. 6]) are also rather involved. The final result is

THEOREM 13.1. *If either*

i) $\varphi'(t) \leq \varphi'(1)\log\ t$ *and* $\lambda_k \geq \hat{\lambda}$ *for some* $\hat{\lambda} > 0$, *or*

ii) $\varphi'(1) < \varphi'(t) \leq \varphi''(1)\log\ t$,

then the sequence $\{x^k\}$ *generated by (13.2)-(13.3) converges to a solution of (13.1).*

We remark that convergence of $\{x^k\}$ to a solution of (13.1) can be proved also when φ does not satisfy (i),(ii) of Theorem 13.1 but then some conditions must be imposed upon problem (13.1) (e.g. boundedness of the set U) which is undesirable, since the proximal point method is devised for ill-posed problems, and so it is important to establish convergence under minimal hypotheses on f (φ, on the other hand, can be chosen at will).

The proximal point method with φ-divergences can be used to solve variational inequality problems when the set C is equal to $\mathbf{R}^n_{++}$. It is easy to verify that in such a case VIP(T, C), with $T : \mathbf{R}^n \to \mathbf{R}^n$, reduces to finding $z \in \mathbf{R}^n$ such that

$$T(z) \geq 0 \tag{13.8}$$

$$z \geq 0 \tag{13.9}$$

$$\langle z, T(z)\rangle = 0 \tag{13.10}$$

which is called the nonlinear complementarity problem for T (NCP(T) from now on). The algorithm is:

$$x^0 > 0 \tag{13.11}$$

and given x^k, find $x^{k+1} > 0$ such that

$$0 \in [\lambda_k \nabla_x d_\varphi(\cdot, x^k) + T](x^{k+1}). \tag{13.12}$$

The convergence analysis combines the arguments in the proof of Theorem 13.1 (including quasi-Fejér convergence with respect to d_ψ) with those in the proof of Theorem 11.1. As in Theorem 11.1, something akin to zone coerciveness is needed for a general T. The equivalent property is $\lim_{t\to\infty}\varphi'(t) = +\infty$ (we remark that the fact that $\lim_{t\to 0}\varphi'(t) = -\infty$, which follows from (12.1), ensures that all φ-divergences satisfy a property similar to boundary coerciveness). The following result has been proved in [6]:

THEOREM 13.2. *If T is paramonotone and pseudomonotone, and either*

a) $\lim_{t\to\infty}\varphi'(t) = +\infty$, or

b) $\sup\{\langle v, x-y\rangle : y \in \mathbf{R}^n, v \in T(y)\} < \infty$ *for all* $x \geq 0$,

then the sequence generated by (13.11)-(13.12) converges to a solution of NCP(T).

Observe that hypothesis (a) of Theorem 13.2 holds for φ_1 and φ_3 of Examples 12.1 and 12.3, but not for φ_2 of Example 12.2.

Theorem 13.2 can be extended to VIP(T,C) for the case in which C is a polyhedron of the form $\{x \in \mathbf{R}^n : Ax \leq b\}$ where A is an $m \times n$ matrix of rank n, using the extensions of *varphi*-divergences to this class of polyhedra introduced in [2].

14. Convergence rate results for the proximal point method with Bregman distances or φ-divergences.

Convergence rate results similar to those of section 5 can be established for these extensions. Consider first the proximal point method with Bregman distances for the problem $\min f(x)$ s.t. $x \in \overline{S}$. It has been proved in [23] the if f is twice continuously differentiable at the limit x^* of $\{x^k\}$ and $\nabla^2 f(x^*)$ is positive definite then x^k converges linearly to x^* and superlinearly if $\lim_{k\to\infty}\lambda_k = 0$. Similar results hold for the proximal method with φ-divergences when $\varphi'(t) \leq \varphi''(1)\log\ t$ and $\lambda_k \geq \hat{\lambda} > 0$, in which case the convergence rate is linear ([23]) or when $\frac{\varphi''(1)}{t} \leq \varphi'(t) \leq \varphi''(1)\log\ t$ in which case we get a linear convergence rate, and a superlinear one when $\lim_{k\to\infty}\lambda_k = 0$.

In a similar way, we have linear or superlinear convergence rates for the proximal methods with Bregman functions or φ-divergences applied to variational inequality or nonlinear

complementarity problems respectively, under rather strong assumptions on the operator T. We must assume, as in section 5, that T^{-1} is Lipschitz continuous at 0 (which already implies existence of at most one solution) and that the limit x^* of the sequence $\{x^k\}$ belongs to S in the case of Bregman functions or to $\mathbf{R}^n_{++}$ in the case of φ-divergences, in which case it is easy to prove that $0 \in T(x^*)$, i.e. that x^* is a zero of T. In the case of Bregman functions, it is also required that $\nabla^2 h(x)$ be continuous, and positive definite for all $x \in S$ (this is a very mild assumption, satisfied by all interesting Bregman functions). The following two theorems have been proved in [6]. $\rho(M)$ denotes the spectral radius of a square matrix M.

THEOREM 14.1. *Assume that the limit x^* of the sequence $\{x^k\}$ generated by (11.2)-(11.3) belongs to S and that T^{-1} is Lipschitz continuous at 0. Then,*

i) if $\lim_{k\to\infty} \lambda_k = 0$ then X^k converges superlinearly to x^*,

ii) if $\lambda_k > \hat{\lambda}$ for all k and some $\hat{\lambda} > 0$ then x^k converges linearly to x^* and

$$\lim_{k\to\infty} \frac{\|x^{k+1} - x^*\|_H}{\|x^k - x^*\|_H} = \frac{\sigma\hat{\lambda}\rho(H)}{(1+\sigma^2\hat{\lambda}^2\rho(H)^2)^{1/2}}$$

where σ is the Lipschitz constant and $H = \nabla^2 h(x^*)^{-1}$.

THEOREM 14.2. *Assume that the limit x^* of the sequence $\{x^k\}$ generated by (13.11)-(13.12) is fully positive and that T^{-1} is Lipschitz continuous at 0. Then*

i) if $\lim_{k\to\infty} \lambda_k = 0$ then x^k converges superlinearly to x^*,

ii) if $\lambda_k > \hat{\lambda}$ for all k and some $\hat{\lambda} > 0$ then x^k converges linearly to x^* and

$$\lim_{k\to\infty} \frac{\|x^{k+1} - x^*\|_B}{\|x^k - x^*\|_B} = \left[1 + \left(\frac{\sigma\hat{\lambda}\varphi''(1)}{x_i^*}\right)^{-2}\right]^{-1/2}$$

where B is the diagonal matrix with diagonal entries $\varphi''(1)/x_j^*$ $(1 \le j \le n)$, $x_i^* = \min_j\{x_j^*\}$ and σ is the Lipschitz constant.

If we compare the asymptotic error constants of Theorems 14.1(ii) and 14.2(ii) with that of Theorem 5.1, we observe that they differ just by the presence of the spectral radiuses of H and B in the former. The reason is that when x^* belongs to S or $\mathbf{R}^n_{++}$, $D_h(x,y)$

and $d_\varphi(x,y)$ can be approximated, in a neighborhood of x^*, by $\|x-y\|_H^2$ and $\|x-y\|_B^2$, so that the sequences $\{x^k\}$ generated by (11.1)-(11.2) or (13.11)-(13.12) behave, for large enough k, as the sequence $\{x^k\}$ of the standard proximal point method (4.1) with $\|\cdot\|_H$, $\|\cdot\|_B$ respectively, instead of $\|\cdot\|$.

As discussed in section 5, the results above are not too interesting, because positive definiteness of $\nabla^2 f(x^*)$ is almost as much as strong convexity of f, and Lipschitz continuity of T^{-1} at 0 is almost as much as strong monotonicity of T, in which case the problems are quite well behaved to begin with. Also the interior point condition on the solution is quite restrictive. Such conditions, however, cannot be easily removed, because we have examples of strictly convex functions such that $\nabla^2 f(x^*)$ is not positive definite at the solution of problems 10.1 or 13.1 (e.g. $f(x) = \sum_{j=1}^{n} (x_j - 1)^4$ with $\overline{S} = \mathbf{R}^n_+$) for which convergence is sublinear for any φ-divergence and any Bregman distance. Nevertheless, there is an interesting case which does not satisfy such conditions and for which a linear (or superlinear) convergence rate can still be established: Linear Programming. Consider

$$f(x) = \begin{cases} c^t x & \text{if} \quad Ax = b, \ x \geq 0 \\ +\infty & \text{otherwise} \end{cases} \tag{14.1}$$

and assume that there exists $\overline{x} > 0$ such that $A\overline{x} = b$. In this case (13.2)-(13.3) becomes

$$x^0 > 0 \tag{14.2}$$

$$x^{k+1} = \operatorname*{argmin}_x \{c^t x + \lambda_k d_\varphi(x, x^k)\} \tag{14.3}$$

$$\text{s.t.} \quad Ax = b. \tag{14.4}$$

The convergence rate results above can be extended to this case, even when the linear programming problem has many solutions (in which case not even the slightly weaker condition of Lipschitz continuity of T^{-1} at 0 holds), though the finite convergence property of the standard proximal point method mentioned in section 5 cannot be attained. Namely, the sequence defined by (14.2)-(14.4) converges to a solution x^* of the linear program with a linear convergence rate under hypothesis (i) of Theorem 13.1 and superlinearly under

hypothesis (ii) of the same theorem when $\lim_{k\to\infty} \lambda_k = 0$. The proof is relatively simple if the linear programming problem has a unique solution and quite harder otherwise (a key element is the fact that the set of solutions of a linear program is a polyhedron, see [22]). Under a weak nondegeneracy assumption (existence of at least one nondegenerate primal-dual optimal pair (x^*, y^*)) it can be proved that any sequence $\{y^k\}$ of Lagrange multipliers for constraints (14.4) in the subproblem (14.3)-(14.4) converges to a dual solution of the linear program.

The situation is more complicated for the proximal point method for Bregman functions. In the first place, in order to eliminate the positivity constraints from the subproblems, as in (14.3)-(14.4), we must take a boundary coercive Bregman function with zone $S = \mathbf{R}^n_{++}$. It is then natural to consider a separable h, as defined in section 9, i.e. $h(x) = \sum_{i=1}^n h_j(x_j)$ with $h_j : \mathbf{R}_{++} \to \mathbf{R}$. It has been proved in [6] that the proximal method with a separable and boundary coercive Bregman function converges linearly when applied to a linear programming problem if $\lim_{t\to 0} th_j''(t) < \infty$ for all j, but for any h such that $\lim_{t\to 0} th_j''(t) = \infty$ for some j, there exist linear programming problems for which convergence is sublinear. $\lim_{t\to 0} th_j''(t) < \infty$ for all j holds for $h(x) = \sum_{j=1}^n x_j \log x_j$ (this case is already covered in the results for φ-divergences because $D_h = d_\varphi$ with $\varphi(t) = t\log t - t + 1$) and also for $h(x) = \sum_{j=1}^n x_j \log(P_j(x_j))$ where P_j is a polynomial with nonnegative coefficients, but almost all other examples of boundary coercive Bregman functions with zone $S = \mathbf{R}^n_{++}$ fail to satisfy this condition and therefore may generate sublinearly convergent sequences when applied to linear programming problems.

The proximal point method with φ-divergences for linear programming has been applied to multicommodity transportation problems with excellent computational results, using a highly parallelizable computer. In this application (developed in [34]) the linear program is of the form $\min c^t x$ s.t. $Ax = b$, $x \geq 0$, with $A \geq 0$, $b > 0$, and $\varphi(t) = t\log t - t + 1$, so that (14.3), (14.4) become

$$x^{k+1} = \operatorname{argmin}\ c^t x + \lambda_k \sum_{j=1}^n \left(x_j \log\left(\frac{x_j}{x_j^k}\right) + x_j^k - x_j \right) \tag{14.5}$$

$$\text{s.t. } Ax = b. \tag{14.6}$$

In order to solve the k-th subproblem (14.5)-(14.6) a method called MART is used. Let a^i $(1 \leq i \leq m)$ be the rows of A and a_{ij} $(1 \leq i \leq m,\ 1 \leq j \leq n)$ its entries. Observe that (14.5) is equivalent to

$$\min \hat{c}^t x + \sum_{j=1}^{n} x_j \log x_j \tag{14.7}$$

$$\text{s.t. } Ax = b \tag{14.8}$$

with $\hat{c}_j = \frac{c_j}{\lambda_k} - 1 - \log x_j^k$. MART, which is highly parallelizable, generates a sequence $\{x^\ell\}$ as follows. The rows of A are used in a predetermined order (say cyclically). Let $i(\ell)$ be the row to be used in iteration ℓ. Then the sequence $\{x^\ell\}$ is defined by

$$x_j^0 = e^{-(1+\hat{c}_j)} \tag{14.9}$$

$$x_j^{\ell+1} = x_j^\ell \left(\frac{b_{i(\ell)}}{\langle a^{i(\ell)}, x^\ell \rangle} \right)^{a_{i(\ell),j}}. \tag{14.10}$$

The sequence $\{x^\ell\}$ converges to a solution of (14.7)-(14.8) (see [8]) i.e. to the k+1-th iterate of the proximal point method with φ-divergences, as in (14.5). Usually a few iterations of MART are enough to generate an appropriate x^{k+1} (i.e. such that $\{x^k\}$ converges rapidly to a solution of $\min c^t x$ s.t. $Ax = b$, $x \geq 0$). We remark that λ_k, x^k and c appear only in the initialization step (14.9) of MART (through $\hat{c}$).

15. Approximate versions of the proximal point method with Bregman functions or φ-divergences.

The approximation we discuss here is different in spirit from the approximations presented in section 6. We do not discuss preservation of convergence of the methods when the k-th iterate is affected by an error ε_k, but rather iteration formulae which allow easier computation of the iterates, replacing subproblem (10.3) or (13.3) by simpler ones. The basic idea is to replace $f(x)$ in

$$x^{k+1} = \operatorname{argmin}\{f(x) + \lambda_k \delta(x, x^k)\} \tag{15.1}$$

where δ is either D_h or d_φ, by its linear approximation $\overline{f}_k(x) = f(x^k) + \nabla f(x^k)^t(x - x^k)$ so that (15.1) becomes, after discarding constant terms,

$$x^{k+1} = \operatorname{argmin}\{\nabla f(x^k)^t x + \lambda_k \delta(x, x^k)\}. \tag{15.2}$$

The difference between (15.1) and (15.2) is that in order to find x^{k+1} in (15.1) we have to solve, in the case of φ-divergences,

$$\varphi'\left(\frac{x_j}{x_j^k}\right) = -\frac{\nabla f(x)_j}{\lambda_k} \qquad (1 \le j \le n) \tag{15.3}$$

which is a system of n nonlinear equations in unknowns $x_1, \ldots, x_n$, while (15.2) reduces to

$$\varphi'\left(\frac{x_j}{x_j^k}\right) = -\frac{\nabla f(x^k)_j}{\lambda_k} \tag{15.4}$$

which is an uncoupled system. In fact, if φ' can be easily inverted (as in Examples 12.1, 12.2 and 12.3), (15.4) reduces to the closed formula

$$x_j^{k+1} = x_j^k (\varphi')^{-1}\left(-\frac{\nabla f(x^k)_j)}{\lambda_k}\right). \tag{15.5}$$

Similarly, in the case of Bregman functions, instead of solving the system

$$\nabla h(x) = \nabla h(x^k) - \frac{1}{\lambda_k}\nabla f(x) \tag{15.6}$$

in order to find x^{k+1}, we get, after linearizing f at x^k,

$$\nabla h(x) = \nabla h(x^k) - \frac{1}{\lambda_k}\nabla f(x^k) \tag{15.7}$$

which, if h is zone coercive so that ∇h is invertible, gives rise to the explicit formula

$$x^{k+1} = (\nabla h)^{-1}[\nabla h(x^k) - \frac{1}{\lambda_k}\nabla f(x^k)]. \tag{15.8}$$

We mention that, as zone coerciveness of h is necessary to have x^{k+1} well defined by (15.8), a similar condition, namely $\lim_{t\to\infty}\varphi'(t) = +\infty$, is necessary to have x^{k+1} well defined by (15.5). An adaptation to the method for the case of $\lim_{t\to\infty}\varphi'(t) < \infty$ can be found in [21].

These methods can also be seen as steepest descent methods with respect to $\delta(\cdot,\cdot)$, because the first order optimality condition of (15.2), namely

$$\nabla f(x^k) + \lambda_k \nabla_x \delta(\cdot, x^k) = 0 \tag{15.9}$$

with $\lambda_k > 0$, is also the first order optimality condition for

$$\min \nabla f(x^k)^t x$$

$$\text{s.t. } \delta(x, x^k) \leq \theta_k$$

for some θ_k. In fact, when $\delta = D_h$ with $h(x) = 1/2\,\|x\|^2$, (15.1) is just the usual steepest descent method with step $1/\lambda_k$. It should be clear that we cannot expect convergence of such scheme without further specifications, in the first place because $\{f(x^k)\}$ is not guaranteed to decrease; we only have $\overline{f}_k(x^{k+1}) \leq \overline{f}_k(x^k)$. As in the case of the usual steepest descent method, some linear search must be performed to ensure convergence. We have two basic alternatives for the search. One is to keep an exogenously given λ_k, call y^k the solution of (15.2) and take

$$x^{k+1} = (1-\alpha_k)x^k + \alpha_k y^k$$

where α_k is obtained through a linear search, i.e. we search in the interval between x^k and y^k. The other option is to determine λ_k with a search along the curve given by (15.5) or (15.8) with a variable λ, that is to say, in the case of (15.8) we take $\mu = 1/\lambda$ and define $x(\mu)$ as

$$x(\mu) = (\nabla h)^{-1}[\nabla h(x^k) - \mu \nabla f(x^k)] \tag{15.10}$$

and perform a linear search along $x(\mu)$ with μ in $[0, \tilde{\mu}]$ for some $\tilde{\mu} > 0$.

We have also choices with respect to the objective function of the linear (or curvilinear) search. The natural option is f, but it turns out to be the case that addition of a regularization term to the objective of the search makes it possible to obtain better convergence results.

It is also clear that, being generalizations of the usual steepest descent method, these algorithms will share its weaknesses, but there is a remarkable advantage: they will produce a minimizer of f on the convex set $\overline{S}$ (with Bregman functions) or $\mathbf{R}^n_{++}$ (with φ-divergences) at a cost just slightly higher (if ∇h or φ' are easily invertible) than the usual steepest descent method, which works only for unconstrained optimization.

The alternatives regarding both the objective and the line for the search mentioned above give rise to the following algorithms:

1. For $\min f(x)$ s.t. $x \in \overline{S}$ with $x(\mu)$ as in (15.10) and $x^0 \in S$,

   a) Algorithm BSD1: ($\tilde{\mu} > 0$ exogenously given)

$$\mu_k = \underset{\mu \in [0, \tilde{\mu}]}{\operatorname{argmin}} f(x(\mu)) \tag{15.11}$$

$$x^{k+1} = x(\mu_k). \tag{15.12}$$

   b) Algorithm BSD2: (μ_k exogenously given, satisfying $\mu_k \in [\hat{\mu}, \tilde{\mu}]$ for some $\tilde{\mu} \geq \hat{\mu} > 0$)

$$\alpha_k = \underset{\alpha \in [0,1]}{\operatorname{argmin}} f((1-\alpha)x^k + \alpha x(\mu_k)) \tag{15.13}$$

$$x^{k+1} = (1-\alpha_k)x^k + \alpha_k x(\mu_k). \tag{15.14}$$

   c) Algorithm BSD3: ($\tilde{\mu} > 0$ and η_k exogenously given, with $\eta_k \in [\hat{\eta}, \tilde{\eta}]$ for some $\tilde{\eta} \geq \hat{\eta} > 0$)

$$\mu_k = \underset{\mu \in [0, \tilde{\mu}]}{\operatorname{argmin}} \{f(x(\mu)) + \eta_k D_h(x(\mu), x^k)\} \tag{15.15}$$

$$x^{k+1} = x(\mu_k). \tag{15.16}$$

2. For $\min f(x)$ s.t. $x \geq 0$ with $x^0 > 0$, y^k equal to the right hand side of (15.5) and λ_k in (15.5) satisfying $\lambda_k \in [\hat{\lambda}, \tilde{\lambda}]$ for some $\tilde{\lambda} \geq \hat{\lambda} > 0$,

a) Algorithm DSD1:

$$\alpha_k = \operatorname*{argmin}_{\alpha \in [0,1]} f((1-\alpha)x^k + \alpha y^k) \tag{15.17}$$

$$x^{k+1} = (1-\alpha_k)x^k + \alpha_k y^k. \tag{15.18}$$

b) Algorithm DSD2: (η_k exogenously given, satisfying $\eta_k \leq \eta \lambda_k d_\varphi(x^k, y^k)$ for some $\eta \in (0,1]$)

$$\alpha_k = \operatorname*{argmin}_{\alpha \in [0,1]} \{f((1-\alpha)x^k + \alpha y^k) + \eta_k \alpha\} \tag{15.19}$$

$$x^{k+1} = (1-\alpha_k)x^k + \alpha_k y^k. \tag{15.20}$$

We discuss next the convergence results for the five algorithms presented above. BSD and DSD stand for Bregman and Divergence Steepest Descent respectively.

In the case of the exact proximal point methods discussed in previous sections, convexity of f was essential to ensure existence of the iterates, while for these approximate versions such existence is guaranteed by the linear search, and convergence results are available for nonconvex functions, as is the case with the usual steepest descent method. On the other hand, as is also the case for the usual steepest descent method, without the regularized search we only get *weak convergence*, in the following sense:

DEFINITION 15.1: A sequence $\{x^k\} \subset \mathbf{R}^n$ is weakly convergent to $V \subset \mathbf{R}^n$ if

i) $\{x^k\}$ is bounded,

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

iii) Every cluster point of $\{x^k\}$ belongs to V.

This notion of weak convergence to a subset of $\mathbf{R}^n$ is unrelated to the notion of weak convergence in Hilbert spaces discussed in section 4. We remark that in $\mathbf{R}^n$ strong and weak convergence, in the sense of section 4, are exactly the same, so that no confusion should arise: the concept in section 4 is used for infinite dimensional spaces and Definition 15.1 for finite dimensional ones.

In addition, for the algorithms with a nonregularized line search (BSD1, BSD2 and DSD1) we need a level set boundedness assumption on f to ensure boundedness of $\{x^k\}$, once again similarly to the situation for the usual steepest descent method.

The following results have been established in [**19**], [**21**].

THEOREM 15.1. *If f is differentiable, $\{x \in \mathbf{R}^n : f(x) \leq f(x^0)\}$ is bounded, h is zone coercive, $D_h(x,\cdot)$ is quasicovex for all $x \in S$ (only for BSD2), $\lim_{t\to\infty} \varphi'(t) = +\infty$ and the corresponding problems have solutions, then the sequences $\{x^k\}$ generated by BSD1, BSD2 or DSD1 satisfy*

i) $\{x^k\}$ *is bounded,*

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

iii) for every cluster point $\overline{x}$ *of* $\{x^k\}$ *it holds that* $\overline{x} \geq 0$, $\nabla f(\overline{x})^t \overline{x} = 0$ *(for DSD1),*

iv) for every $x \in \overline{S}$ *there exists a cluster point* $\overline{x}$ *of* $\{x^k\}$ *such that* $\langle \nabla f(\overline{x}), x - \overline{x}\rangle \geq 0$ *(for BSD1 and BSD2).*

Quasiconvexity of D_h in its second variable, required for BSD2 in order to prove item (ii) of Theorem 15.1, is a rather restrictive assumption. While all φ-divergences are convex in their second variable by Proposition 12.1(v), most Bregman distances are not. This assumption is not needed for BSD1 or BSD3.

THEOREM 15.2. *Under the hypotheses of Theorem 15.1, if f is convex then the sequence $\{x^k\}$ generated by BSD1, BSD2 or DSD1 is weakly convergent to the solution set of the corresponding problem.*

THEOREM 15.3. *If f is differentiable, h is zone coercive, $\lim_{t\to\infty} \varphi'(t) = +\infty$ and the corresponding problem has solutions, then*

i) the sequence $\{x^k\}$ generated by BSD3 satisfies items (i), (ii) and (iv) of Theorem 15.1,

ii) the sequence $\{x^k\}$ generated by DSD2 satisifes items (i), (ii) and (iii) of Theorem 15.1.

THEOREM 15.4. *Under the hypotheses of Theorem 15.3, if f is convex then the sequences $\{x^k\}$ generated by BSD3 or DSD2 converge to a solution of the corresponding problem.*

The proof of Theorem 15.4 uses the concept of quasi-Fejér convergence with respect to d_φ and Proposition 13.1 to obtain convergence of the whole sequence generated by DSD2, and the analogous versions with Bregman functions for the sequence generated by BSD3.

Of course, these algorithms are practical only when φ' or ∇h can be easily inverted. For $S = \mathbf{R}^n_{++}$, and φ as in Example 12.1 or h as in Example 9.2, we get

$$x(1/\lambda_k)_j = y^k_j = x^k_j \exp(-1/\lambda_k \nabla f(x^k)_j) \tag{15.21}$$

and the computational work of the algorithms reduces to the line searches. Examples of Bregman functions with easily invertible gradients can be found in [9] and [19] for the case of C being a ball, a box (as defined in section 9) or a polyhedron of the form $\{x \in \mathbf{R}^n : Ax \leq b\}$ with A nonsingular. In the case of the polyhedron, inversion of ∇h requires solution of linear systems with matrices A and A^t.

Bibliographical notes

References given in the text are mainly for results whose proofs have not been fully included. We acknowledge here the sources of the main ideas and results.

A modern exposition of the rather old concept of regularization discussed in section 1 can be found in [33]. [25] is an excellent survey on proximal point methods, with an approach somewhat different from ours, and includes comprehensive references on the subject. The material in section 2 is based upon [29], though our proofs follow a different approach. A very good introduction to monotone operators can be found in [7]. [28] is a basic reference for the proximal point method for monotone operators. More recent results on the subject can be found in [26]. The final step in the proof of Theorem 4.4 has been taken from [29], which is also the basis of sections 5 and 6. The material of section 7 has been drawn from [3], [4] and [30]. The concept of Bregman functions originates in [5]. The definition of Bregman functions used here is taken from [13], and the notions of zone and boundary coerciveness were introduced in [18]. The proximal point method with Bregman distances appeared for the first time in [15], but only for h as in Example 9.2. The method for a general h was first presented in [10], with a convergence analysis covering only solutions in the interior of the zone. A convergence analysis including the case of solutions in the boundary, using zone coercive Bregman functions, can be found in [11], from where we have taken the proof of Theorem 10.1, excepting for Step 1, where we use boundary instead of zone coerciveness, following [18], which contains also the material on the characterization of the limit point for linear and quadratic programming presented at the end of section 10. The exponential multipliers method of section 10 is presented in [4]. An extensive study of variational inequality problems can be found in [24]. Proximal

point methods with Bregman functions for finding zeroes of monotone operators were first discussed in [14], and then extended to variational inequality problems in [6], from where we have taken the material of section 11. The concept of paramonotonicity was introduced in [9]. The notion of φ-divergence originates in [12] and has been extended and developed in [32]. The convergence analysis of the proximal point method with φ-divergences can be found in [20] for the case of λ_k bounded away from 0 and in [22] for the general case. The notion of quasi-Fejér convergence appeared for the first time in [16] and was further developed in [20], from where we have taken the proof of Proposition 13.1. The idea of using proximal point methods with φ-divergences for variational inequalities was presented for the first time in [2]. The results on this subject discussed at the end of section 14 appear in [6]. Linearity of the convergence rate of proximal point methods with Bregman functions or φ-divergences for minimization of strongly convex functions has been proved in [23]. Similar results for monotone operators whose inverse is Lipschitz continuous at 0 are from [6]. The convergence rate of the proximal point method applied to linear programming problems is analyzed in [22] for the case of φ-divergences and in [6] for the case of Bregman functions. The approximate versions of the proximal point methods discussed in section 15 have been presented in [21] for the case of φ-divergences and in [19] for the case of Bregman functions.

REFERENCES

[1] Adler, I., Monteiro, R.D.C. Limiting behavior of the affine scaling continuous trajectories for linear programming. *Mathematical Programming* **5** (1991) 29-51.

[2] Auslender, A., Haddou, M. An interior-proximal method for convex linearly constrained problems and its extension to variational inequalities (to be published).

[3] Avriel, M. *Nonlinear Programming, Analysis and Methods*. Prentice Hall, New Jersey (1976).

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

[5] Bregman, L. 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** (1967) 200-217.

[6] Burachik, R. *Generalized proximal point methods for the variational inequality problem.* PhD Thesis. Instituto de Matemática Pura e Aplicada. Rio de Janeiro (1995).

[7] Brézis, H. *Operateurs monotones maximaux et semigroups de contractions dans les espaces de Hilbert*. Université de Paris-CNRS, Paris (1971).

[8] Censor, Y., De Pierro, A., Elfving, T., Herman, G., Iusem, A. On iterative methods for linearly constrained entropy maximization. In *Numerical Analysis and Mathematical Modelling. Banach Center Publication Series* **24** (1990) 145-163.

[9] Censor, Y., Iusem, A., Zenios, S. An interior point method with Bregman functions for the variational inequality problem with paramonotone operators (to be published).

[10] Censor, Y., Zenios, S. 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. Information-type measures of difference of probability distributions and indirect observations. *Studia Scientiae Mathematicae Hungaricae* **2** (1967) 299-318.

[13] De Pierro, A., Iusem, A. A relaxed version of Bregman's method for convex programming. *Journal of Optimization Theory and Applications* **51** (1986) 421-440.

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

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

[16] Ermol'ev, Yu.M. On the method of generalized stochastic gradients and quasi-Fejér sequences. *Cybernetics* **5** (1969) 208-220.

[17] Hoffman, A.J. On approximate solutions of systems of linear equations. *Journal of Research of the National Bureau of Standards* **49** (1952) 263-265.

[18] Iusem, A. On some properties of generalized proximal point methods for quadratic and linear programming (to be published in *Journal of Optimization Theory and Applications*).

[19] Iusem, A. Steepest descent methods with generalized distances for constrained optimization (to be published).

[20] Iusem, A., Svaiter, B., Teboulle, M. Entropy-like proximal methods in convex programming (to be published in *Mathematics of Operations Research*).

[21] Iusem, A., Svaiter, B., Teboulle, M. Multiplicative interior gradient methods for minimization over the nonnegative orthant (to be published).

[22] Iusem, A., Teboulle, M. Convergence rate analysis of nonquadratic proximal and augmented Lagrangian methods for convex and linear programming (to be published in *Mathematics of Operations Research*).

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

[24] Kinderlehrer, D., Stampacchia, G. *An Introduction to Variational Inequalities and their Applications*. Academic Press, New York (1980).

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

[26] Martinet, B. Perturbations des méthodes d'optimisation. *RAIRO. Analyse numérique* **12** (1978) 153-171.

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

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

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

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

[31] Taylor, A.E., Lay, D.C. *Introduction to Functional Analysis*. J. Wiley, New York (1980).

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

[33] Thikonov, A., Arsenin, V. *Solution of ill-posed problems*. Winston (1977).

[34] Zenios, S. On the fine grain decomposition of multicommodity transportation problems. *SIAM Journal on Optimization* **1** (1991) 401-423.

Impresso na Gráfica do
impa
pelo Sistema Xerox /1090