Annals of Operations Research 105, 77–98, 2001

Proximity Function Minimization Using Multiple Bregman Projections, with Applications to Split Feasibility and Kullback–Leibler Distance Minimization

CHARLES BYRNE Charles_Byrne@uml.edu
Department of Mathematical Sciences, University of Massachusetts at Lowell, Lowell, MA 01854, USA

YAIR CENSOR yair@math.haifa.ac.il
Department of Mathematics, University of Haifa, Mt. Carmel, Haifa 31905, Israel

Abstract. Problems in signal detection and image recovery can sometimes be formulated as a *convex feasibility problem* (CFP) of finding a vector in the intersection of a finite family of closed convex sets. Algorithms for this purpose typically employ orthogonal or generalized projections onto the individual convex sets. The simultaneous multiprojection algorithm of Censor and Elfving for solving the CFP, in which different generalized projections may be used at the same time, has been shown to converge for the case of nonempty intersection; still open is the question of its convergence when the intersection of the closed convex sets is empty.

Motivated by the geometric alternating minimization approach of Csiszár and Tusnády and the product space formulation of Pierra, we derive a new simultaneous multiprojection algorithm that employs generalized projections of Bregman to solve the convex feasibility problem or, in the inconsistent case, to minimize a proximity function that measures the average distance from a point to all convex sets. We assume that the Bregman distances involved are jointly convex, so that the proximity function itself is convex. When the intersection of the convex sets is empty, but the closure of the proximity function has a unique global minimizer, the sequence of iterates converges to this unique minimizer. Special cases of this algorithm include the "Expectation Maximization Maximum Likelihood" (EMML) method in emission tomography and a new convergence result for an algorithm that solves the split feasibility problem.

Keywords: Bregman projections, convex feasibility problem, product space, Kullback–Leibler distance, proximity function

1. Introduction

Let C_i, $i = 1, 2, \ldots, I$, be closed convex sets in the J-dimensional Euclidean space $\mathbb{R}^J$ and let C be their intersection. In many applications such convex sets represent constraints that we wish to impose on the solution and the algorithms employ projections onto these individual sets. For examples see Youla [53], Combettes [30], Byrne [16] and the recent books by Stark and Yang [50] and Censor and Zenios [29].

Typically, the projections of a point onto the individual sets C_i are more easily calculated than the projection onto the intersection C, therefore iterative methods whereby

the latter can be obtained from repeated use of the former are desirable. There are three cases to be considered:

(1) the intersection C is nonempty, but "small" in the sense that all members of C are quite similar;

(2) the intersection C is nonempty and "large", that is, the members of C are quite varied; and

(3) the set C is empty, meaning that the constraints we impose are mutually contradictory.

When we say that the members of C are "quite similar" or "quite varied", we mean that the real-world objects that they represent (e.g., the images in an image reconstruction task) are "similar" or "varied" according to some criteria appropriate for the task.

Case (1) usually occurs if I is large and/or the individual sets C_i are "small". In this case an algorithm that simply solves the *convex feasibility problem* (CFP), that is, one that finds some member of C, is useful. Case (2) occurs if there are few convex sets and/or they all are quite "large". In this case just obtaining some member of C may not be helpful; we want to get a member of C near to some prior estimate of the solution. The orthogonal projection onto C, or a generalized projection of the type to be discussed here, might be more helpful in this case; see, e.g., Dykstra [34,35], Censor and Reich [28], Bregman, Censor and Reich [6] and references therein. A more general approach is to optimize a cost function, such as entropy, over the set C. Recent related work on this topic is in Byrne [17]. Case (3) is dealt with by finding a point that is, in some sense, close to all the individual sets C_i. One way to achieve this is to set up a *proximity function* that measures the average distance to all the convex sets and then to minimize it. Case (3) is our main focus in the present paper.

These issues can be considered in a general context, involving Bregman distances and projections. Let S be an open convex subset of $\mathbb{R}^J$ and f a Bregman function from the closure $\overline{S}$ of S into $\mathbb{R}$; see, e.g., Censor and Lent [25], Censor and Zenios [29, chapter 2] and the appendix at the end of this paper. For a Bregman function $f(x)$, the Bregman distance D_f is defined by

$$D_f(z, x) \triangleq f(z) - f(x) - \langle \nabla f(x), z - x \rangle, \tag{1.1}$$

where $\langle \cdot, \cdot \rangle$ is the standard inner product in $\mathbb{R}^J$ and $\nabla f(x)$ is the gradient of f at x. When the function f has the form $f(x) = \sum_{j=1}^{J} g_j(x_j)$, with the g_j scalar Bregman functions, we say that f and the associated $D_f(z, x)$ are *separable*. With $g_j(t) = t^2$, for all j, the function $f(x) = \sum_{j=1}^{J} g_j(x_j) = \sum_{j=1}^{J} x_j^2$ is a separable Bregman function and $D_f(z, x)$ is the squared Euclidean distance between z and x.

For each i, denote by $P_{C_i}^f(x)$ the *Bregman projection* of $x \in S$ onto the set C_i with respect to the *Bregman function* f; that is, for any $x \in S$ we have $D_f(P_{C_i}^f(x), x) \leqslant D_f(z, x)$, for all $z \in C_i \cap \overline{S}$. If $C \triangleq \bigcap_{i=1}^{I} C_i$ is nonempty then the sequential iterative algorithm of successive projections, whose iterative step is given by $x^{k+1} = P_{C_{i(k)}}^f(x^k)$,

converges to a member of C. This was shown by Bregman [5] for the *cyclic control* defined by $i(k) = k \pmod I + 1$, for all $k \geqslant 0$, and by Censor and Reich [27] and Bauschke and Borwein [4] for the more general *repetitive control*. If the set C is empty then this scheme does not converge. In such a case it has been shown by Gubin, Polyak and Raik [37] that, for orthogonal projections in Hilbert space, the sequential iterative scheme exhibits *cyclic convergence,* i.e., convergence of the cyclic subsequences.

In this paper we investigate iterative methods of the simultaneous type. In the past such methods were proposed with arithmetic averaging for orthogonal projections, see, e.g., Auslender [2], Aharoni and Censor [1], Bauschke and Borwein [3], Butnariu and Censor [8,9], Censor [19,20], Combettes [30,31], Iusem and De Pierro [40], Kiwiel [41] and references therein. See also the recent book by Butnariu, Censor and Reich [11]. Recently, Censor and Elfving [21] proposed and studied a simultaneous projections algorithm for the convex feasibility problem that employs Bregman projections. However, the averaging of the simultaneous projections there is not arithmetic, but depends on the choice of the Bregman function (or functions).

Byrne and Censor studied in [18] simultaneous methods with arithmetic averaging for Bregman projections that are not necessarily orthogonal. Such a possibility was mentioned, in passing, by Censor and Herman [23, section 4.4], and was recently studied for the special case of entropic projections in Butnariu, Censor and Reich [10]; the results in [10] deal only with the consistent case $C \neq \emptyset$. The focus in [18] was on the behavior, in the inconsistent case $C = \emptyset$, of simultaneous methods with arithmetic averaging of Bregman projections based on a Bregman distance D_f that is both separable and jointly convex. Recent work by Butnariu, Iusem and Burachik [12] on stochastic convex feasibility problems contains a similar proximity function minimization algorithm and notes the importance of joint convexity of the distance.

In contrast with these efforts, we are concerned here with proximity functions $F(x)$ of the multiprojection type; that is, functions of the form

$$F(x) = \sum_{i=1}^{I} D_{f_i}\left(P_{C_i}^{f_i}(x), x\right), \tag{1.2}$$

where the D_{f_i} are Bregman distances derived from possibly distinct, possibly nonseparable Bregman functions f_i having zones S_{f_i}. The function F is defined, for all x in the open convex set $U = \bigcap_{i=1}^{I} S_{f_i}$, which is assumed nonempty. We shall assume that each distance $D_{f_i}(x, z)$ is jointly convex, i.e., is convex with respect to the vector formed by concatenating the vectors x and z. The usefulness of joint convexity has been noted in several recent publications, see, e.g., Butnariu, Iusem and Burachik [12], and examples of jointly convex distances are well-known. The joint convexity of these distances implies that $F(x)$ is convex, as well.

Our main use of convex analysis will be to apply the closure operation to the proximity function $F(x)$ in order to provide finite-valued extension to at least part of the boundary of U. In the standard presentation of Bregman functions and distances the

zone S is an open convex set with closure $\overline{S}$. The Bregman distance $D_f(x, z)$ is defined for $x \in \overline{S}$ and $z \in S$ and the Bregman projections are defined for $x \in S$. In Byrne and Censor [18] the definition of the distance $D_f(x, z)$, the Bregman projections and, thereby, the proximity function $F(x)$ are extended to include $x \in \overline{S}$, but only for the separable case. Such extension permits the treatment of the fairly common case in which the proximity function has no minimizer within S, but does have a minimizer when extended to $\overline{S}$. Similar extensions appear in Kiwiel [42] and in Censor and Zenios [29, section 6.8]. Our approach here is to extend only $F(x)$, not the Bregman projections themselves, using the concept of closure of a proper convex function, as discussed in Rockafellar [48, section 7].

We first extend the proximity function $F(x)$ of (1.2) to all of $\mathbb{R}^J$ by defining $F(x) = +\infty$, for all x not in U. The *closure* of the function F is the function $\operatorname{cl} F$ defined, for all x in $\mathbb{R}^J$, by

$$\operatorname{cl} F(x) = \liminf_{y \to x} F(y). \tag{1.3}$$

The functions F and $\operatorname{cl} F$ agree on U but $\operatorname{cl} F$ can differ from F by taking finite values at certain points on the boundary of U at which F takes the value $+\infty$. We shall prove convergence of our iterative method whenever $\operatorname{cl} F$ has a unique minimizer or when the set $C \cap \overline{U}$ is nonempty. In the next section we show a motivating example of the closure of a proximity function when we consider the likelihood maximization problem for independent Poisson random variables.

In designing our iterative algorithm we are influenced by the reformulation of the CFP in a product space, as suggested by Pierra [47], and the concepts of *Bregman distances* and *Bregman projections* as introduced by Censor and Lent [25] based on the work of Bregman in [5] and studied extensively, see, e.g., [29] and references therein. The third ingredient used here is the framework of alternating minimization of a functional of two vectors, proposed by Csiszár and Tusnády [32]. The first two of these concepts were used by Censor and Elfving [21], but, because they were concerned only with the consistent case, they used Bregman's successive projections approach, not the alternating minimization method of [32]. The proximity function minimization algorithm developed by us in [18] can also be recast in terms of Pierra's product space and the alternating minimization approach of Csiszár and Tusnády, but neither of these notions was explicitly used there. In recent work Eggermont and LaRiccia [36] make use of alternating minimization and prove that jointly convex Bregman distances enjoy the "four-point property" of [32]. As we shall see, this is an important aid in the present work.

We draw the reader's attention to the "Note added in proof", at the end of the paper, in which we add some important remarks about the validity or validation of the technical assumptions A1, A2 and A3 made in our analysis.

2. A motivating example: The expectation maximization maximum likelihood algorithm

The "Expectation Maximization Maximum Likelihood" (EMML) algorithm, as it is used in emission tomography (see, e.g., Vardi, Shepp and Kaufman [52], Lange and Carson [44], Tanaka [51], Byrne [13–15]) is a special case of the more general EM algorithm of Dempster, Laird and Rubin [33] for computing maximum likelihood estimators, see also McLachlan and Krishnan [46]. The EMML algorithm in this case provides a nonnegative minimizer of the Kullback–Leibler distance, which is a function of the type given in (1.2) that can be extended continuously to its closure.

Shannon's entropy function maps the nonnegative orthant $\mathbb{R}^J_+$ into $\mathbb{R}$ according to

$$\operatorname{ent} x \triangleq -\sum_{j=1}^{J} x_j \log x_j, \tag{2.1}$$

where "log" denotes the natural logarithms and, by definition, $0 \log 0 = 0$. Its negative, $f(x) \triangleq -\operatorname{ent} x$, is a Bregman function and the Bregman distance associated with it is the Kullback–Leibler (KL) distance (see Kullback and Leibler [43], see also [29, example 2.1.2 and lemma 2.1.3]), given by

$$D_f(x, z) = KL(x, z) = \sum_{j=1}^{J} \left(x_j \log\left(\frac{x_j}{z_j}\right) + z_j - x_j \right). \tag{2.2}$$

For positive scalars a, b, define $KL(a, b) = a \log(a/b) + b - a$, $KL(0, b) = b$ and $KL(a, 0) = +\infty$. For a given positive vector $y \in \mathbb{R}^I$ and a given nonnegative matrix $A = (a_{ij}) \in \mathbb{R}^{I \times J}$, all of whose column-sums are equal to one, and with no zero rows, denote by a^i the ith column of the transpose matrix A^{T} (so that $a^i_j = a_{ij}$) and consider the distance

$$KL(y, Ax) \triangleq D_f(y, Ax) = \sum_{i=1}^{I} \left(y_i \log \frac{y_i}{\langle a^i, x \rangle} + \langle a^i, x \rangle - y_i \right). \tag{2.3}$$

We see that the function $KL(y, Ax)$ is a proper convex function and can be extended continuously to a function that takes finite values at all points x of the boundary of the positive orthant for which the vector Ax has only positive entries. This extension is its closure function. If there is no nonnegative vector x such that $y = Ax$, then the minimum of the closure of $KL(y, Ax)$ over the nonnegative orthant occurs on the boundary of the nonnegative orthant, not in the interior.

Define the sets C_i as

$$C_i \triangleq \left\{ x \in \mathbb{R}^J \mid x \geqslant 0,\ \langle a^i, x \rangle = y_i \right\}, \tag{2.4}$$

and let $w_j^i \triangleq a_{ij}$, for all $1 \leqslant i \leqslant I$ and $1 \leqslant j \leqslant J$. The Bregman functions $f_i(x)$ are defined, for $i = 1, 2, \ldots, I$, as

$$f_i(x) \triangleq \sum_{j=1}^{J} a_{ij}(x_j \log x_j - x_j), \tag{2.5}$$

and the Bregman projection $P_{C_i}^{f_i}(x)$ of a point $x \in \mathbb{R}_+^J$ onto C_i, is a member of C_i which minimizes the distance

$$D_{f_i}(z, x) = \sum_{j=1}^{J} w_j^i KL(z_j, x_j), \tag{2.6}$$

over all $z \in C_i \cap \mathbb{R}_+^J$. It is not difficult to verify that, in this case, $P_{C_i}^{f_i}$ has the explicit form

$$\left(P_{C_i}^{f_i}(x)\right)_j = x_j \frac{y_i}{\langle a^i, x\rangle}, \quad 1 \leqslant j \leqslant J. \tag{2.7}$$

If $w_j^i = 0$, for some values of j, then there will be other members of C_i that also minimize the distance given by (2.6).

It is important to note that if there is an index j for which $x_j = 0$ but $z_j \neq 0$ then $KL(z, x) = +\infty$. Therefore, when we seek the Bregman projection of x onto a closed convex set C_i we must allow for the possibility that the Bregman distance from x to each member of C_i is infinite and then we do not define the Bregman projection of x onto this set. In our case, however, we see from (2.7) that $(P_{C_i}^{f_i}(x))_j = 0$ if and only if $x_j = 0$, so the Bregman distance from x to such C_i is always finite and the Bregman projection is always defined.

The proximity function F of (1.2) is defined, in this case, as

$$\begin{aligned} F(x) &\triangleq \sum_{i=1}^{I} D_{f_i}\left(P_{C_i}^{f_i}(x), x\right) = \sum_{i=1}^{I}\sum_{j=1}^{J} a_{ij} KL\left(\left(P_{C_i}^{f_i}(x)\right)_j, x_j\right) \\ &= \sum_{i=1}^{I}\sum_{j=1}^{J} a_{ij} KL\left(x_j \frac{y_i}{\langle a^i, x\rangle}, x_j\right) = KL(y, Ax). \end{aligned} \tag{2.8}$$

The iterative step of the EMML algorithm is given by

$$x_j^{k+1} = \sum_{i=1}^{I} w_j^i \left(P_{C_i}^{f_i}\left(x^k\right)\right)_j = \sum_{i=1}^{I} a_{ij} \frac{x_j^k y_i}{\langle a^i, x^k\rangle} = x_j^k \sum_{i=1}^{I} \frac{a_{ij} y_i}{\langle a^i, x^k\rangle}, \tag{2.9}$$

for all $1 \leqslant j \leqslant J$. It is instructive to note that, for $k = 1, 2, \ldots$, we have $\sum_{j=1}^{J} x_j^k = \sum_{i=1}^{I} y_i$ and, therefore, the sequence $\{x^k\}$ is bounded. The $F(x)$ of (2.8) clearly has nonnegative minimizers and the following result holds (see Iusem [38,39], Vardi, Shepp and Kaufman [52]).

Theorem 2.1. Let the entries of x^0 be positive. Any sequence $\{x^k\}_{k\geqslant 0}$, generated by the EMML algorithm (2.9), converges to a nonnegative vector that minimizes $KL(y, Ax)$.

In the inconsistent case $\{x \in \mathbb{R}^J \mid x \geqslant 0,\ Ax = y\} = \emptyset$ the nonnegative minimizer of $KL(y, Ax)$ is almost always unique, regardless of the values of I and J.

Definition 2.1. We say that a matrix $A = (a_{ij}) \in \mathbb{R}^{I\times J}$ has the full rank property (FRP) if A and every submatrix obtained from A by deleting columns have full rank.

The following result can be found in Byrne [13, proposition 1].

Theorem 2.2. If A has the FRP and if $y = Ax$ has no nonnegative solutions then there is a subset $L \subseteq \{1, 2, \ldots, J\}$, having at most $I - 1$ elements, such that, for all nonnegative minimizers $\hat{x} \geqslant 0$ of $KL(y, Ax)$, $\hat{x}_j > 0$ only if $j \in L$. Consequently, there can be only one such $\hat{x}$.

According to this theorem, the minimizer of the proximity function exists only if the function $F(x)$ is extended, via the closure operation, to the boundary of the region within which it is originally defined.

In [13,15] we compared the EMML algorithm to the closely related "Simultaneous Multiplicative Algebraic Reconstruction Technique" (SMART). The iterative step of the SMART algorithm is given by

$$x_j^{k+1} = x_j^k \exp\left(\sum_{i=1}^{I} a_{ij} \log\left(\frac{y_i}{\langle a^i, x^k\rangle}\right)\right), \tag{2.10}$$

for all $j = 1, 2, \ldots, J$. The SMART minimizes the function $KL(Ax, y)$ over the nonnegative orthant. While in the consistent case, $C \neq \emptyset$, the EMML algorithm converges to some member of C, the SMART, in contrast, converges to that member of C for which the cross-entropy (Kullback–Leibler) distance to the initial vector, $KL(x, x^0)$, is minimized. When the entries of the initial vector x^0 are all equal, the SMART converges to the solution for which the Shannon entropy, ent x, is maximized.

There is no loss of generality in considering here only systems of linear equations $Ax = y$ in which all entries of the matrix $A = (a_{ij})$ are nonnegative. This is so because given an arbitrary matrix and rescaling if necessary, we may assume that, for each j, the column sum $\sum_{i=1}^{I} a_{ij}$ is nonzero. Now redefine A and x without changing the notation as follows: replace a_{kj} by $a_{kj}/\sum_{i=1}^{I} a_{ij}$ and x_j by $x_j \sum_{i=1}^{I} a_{ij}$. This leaves the product Ax unchanged but the new A has all its column sums equal to one. The equality $Ax = y$ still holds, but now we know that $y_+ \triangleq \sum_{i=1}^{I} y_i = \sum_{j=1}^{J} x_j \triangleq x_+$. Let E be the matrix whose entries are all one and let $\gamma \geqslant 0$ be a large enough scalar so that $A^{\text{new}} = A + \gamma E$ has all nonnegative entries. Then $A^{\text{new}}x = Ax + (\gamma x_+)e$, where e is the vector whose entries are all one. So the new system of equations to solve is $A^{\text{new}}x = y + (\gamma y_+)e = y^{\text{new}}$.

3. Preliminaries: Csiszár and Tusnády's alternating minimization algorithm and the lemma of Eggermont and LaRiccia

Our new fully simultaneous algorithm employs Bregman projections onto the convex sets $\{C_i\}_{i=1}^I$ in $\mathbb{R}^J$. As will be seen below, the main algorithmic difference between this algorithm and the multiprojections method of Censor and Elfving [21] (see also [29, section 5.9]) is that here we use alternating minimizations, instead of successive projections. For symmetric distances the two approaches coincide, as in the case of the split feasibility problem discussed later. So far, convergence of the multiprojections algorithm of Censor and Elfving has been shown only in the consistent case $C \neq \emptyset$; the convergence results in this paper apply to both the consistent and inconsistent situations. In the inconsistent case our algorithm becomes a minimization tool for the closure of the proximity function $F(x)$ defined originally for $x \in U = \bigcap_{i=1}^I S_{f_i}$ by (1.2).

As we saw in the previous section, in the case of the EMML algorithm, the proximity function $F(x)$ may not have a minimizer within the open set U. In this case the function $F(x)$ was easily extended, from the positive orthant to portions of the boundary, within which we are able to locate minimizers of $KL(y, Ax)$. Following this example, we shall consider, for the general case, minimizers of the closure of the proximity function F.

Csiszár and Tusnády describe in [32] an *alternating minimization algorithm* for solving a proximity function minimization problem. Their problem involves only two convex sets. However, using Pierra's [47] product space reformulation of the CFP, the alternating minimization algorithm can be applied to obtain simultaneous algorithms for the general CFP. The alternating minimization algorithm derives its strength from two important geometric properties, called the *Three Point Property* (3PP) and the *Four Point Property* (4PP) in [32], of which we also make use here.

Given closed convex subsets C_i, $i = 1, 2, \ldots, I$, with (possibly empty) intersection $C \triangleq \bigcap_{i=1}^I C_i$, we reformulate the CFP in a product space framework. Following Pierra [47], we let $\mathcal{V}$ be the product of I copies of the Euclidean space $\mathbb{R}^J$, so that a typical element $v = (v_1, v_2, \ldots, v_I)$ of $\mathcal{V}$ is such that $v_i \in \mathbb{R}^J$, $i = 1, 2, \ldots, I$. We define $\mathcal{C} = \prod_{i=1}^I C_i$ to be the product of all sets C_i, i.e., the subset of $\mathcal{V}$ consisting of all v such that $v_i \in C_i$, $i = 1, 2, \ldots, I$, and we let $\mathcal{D}$ be the ("diagonal") subspace of $\mathcal{V}$ consisting of all v such that $v_i = x$, $i = 1, 2, \ldots, I$, where $x \in \mathbb{R}^J$, and express this by writing $v = d(x)$. It is easy to verify that an element $d(x^*)$ belongs to $\mathcal{D} \cap \mathcal{C}$ if and only if $x^* \in \bigcap_{i=1}^I C_i$ and, therefore, finding a solution of the two-sets feasibility problem in $\mathcal{V}$ yields a solution of the original CFP in $\mathbb{R}^J$.

In [21] Censor and Elfving obtain an iterative algorithm for solving the CFP by performing successive Bregman projections onto $\mathcal{C}$ and $\mathcal{D}$ with respect to a Bregman distance in $\mathcal{V}$, given by,

$$D_\lambda(v, w) \triangleq \sum_{i=1}^I \lambda_i D_{f_i}(v_i, w_i), \tag{3.1}$$

where $\lambda = (\lambda_i) \in \mathbb{R}^I$ is a fixed vector such that all λ_i are positive and $\sum_{i=1}^I \lambda_i = 1$. A different distance $D_{\mathcal{F}}(v, w)$ between $v \in \overline{\mathcal{S}}$ and $w \in \mathcal{S}$, where $\mathcal{S} \triangleq \prod_{i=1}^I S_i$, is given by

$$D_{\mathcal{F}}(v, w) = \sum_{i=1}^{I} D_{f_i}(v_i, w_i), \tag{3.2}$$

and it can be shown, along the same lines of proof of [21, lemma 3.1] (or see [29, lemma 5.9.1]), that this is a Bregman distance in the product space $\mathcal{V}$, induced by the Bregman function $\mathcal{F}(v) \triangleq \sum_{i=1}^I f_i(v_i)$. With this distance at hand we propose to solve the CFP by a simultaneous iterative algorithm that minimizes $D_{\mathcal{F}}(\alpha, \beta)$, over $\alpha \in \mathcal{C}$, $\beta \in \mathcal{D}$. If the CFP has a solution, then the minimum value will be zero. This approach involves the alternating minimization method of Csiszár and Tusnády [32] which we present now in a slightly simplified version.

Suppose that $\mathcal{P}$ and $\mathcal{Q}$ are two convex sets in the n-dimensional Euclidean space $\mathbb{R}^n$. Let $\Theta(p, q)$ be a real-valued function defined for all $p \in \mathcal{P}$, $q \in \mathcal{Q}$.

Algorithm 3.1 (The alternating minimization algorithm).
Initialization: $q^0 \in \mathcal{Q}$ is arbitrary.
Iterative step: Given q^k find p^{k+1} by solving

$$p^{k+1} = \operatorname{argmin}\{\Theta(p, q^k) \mid p \in \mathcal{P}\}, \tag{3.3}$$

then calculate q^{k+1} by solving

$$q^{k+1} = \operatorname{argmin}\{\Theta(p^{k+1}, q) \mid q \in \mathcal{Q}\}. \tag{3.4}$$

Assuming that all the minima exist, the pair of sequences $\{p^k\}$, $\{q^k\}$ is obtained and we then have the following monotonicity result.

Lemma 3.1. If all the minima in (3.3) and (3.4) exist then, for any pair of sequences $\{p^k\}$, $\{q^k\}$, generated by algorithm 3.1, the sequence $\{\Theta(p^k, q^k)\}$ is decreasing.

Proof. For all $k \geqslant 0$, we have

$$\Theta(p^k, q^k) \geqslant \Theta(p^{k+1}, q^k) \geqslant \Theta(p^{k+1}, q^{k+1}). \tag{3.5}$$

□

To obtain further results Csiszár and Tusnády introduce two geometric axioms, the *Three Point Property* (3PP) and the *Four Point Property* (4PP), which we discuss now.

Definition 3.1 (The three point property). The function $\Theta(p, q)$ has the 3PP if there is a nonnegative-valued function $\Delta(p, p')$, defined for all $p, p' \in \mathcal{P}$, such that, for

every $p \in \mathcal{P}$ and for every pair of iterative sequences, generated by algorithm 3.1, the inequality

$$\Theta\big(p, q^k\big) \geqslant \Delta\big(p, p^{k+1}\big) + \Theta\big(p^{k+1}, q^k\big) \tag{3.6}$$

holds, for all $k \geqslant 0$.

In many applications $\Theta(p, q) \geqslant 0$ and $\Delta(p, p') = \Theta(p, p')$. As we shall see, this holds for the distance function defined in (3.2).

Lemma 3.2. If all minima in (3.3) and (3.4) exist, if $\Theta(p, q)$ has the 3PP and if the sequence $\{\Theta(p^k, q^k)\}$ is bounded below then the sequence $\{\Delta(p^k, p^{k+1})\}$ converges to zero.

Proof. Using the 3PP and the definitions of p^k and q^k, we have, for all $k \geqslant 0$,

$$\Theta\big(p^k, q^k\big) \geqslant \Delta\big(p^k, p^{k+1}\big) + \Theta\big(p^{k+1}, q^k\big) \geqslant \Delta\big(p^k, p^{k+1}\big) + \Theta\big(p^{k+1}, q^{k+1}\big). \tag{3.7}$$

Since $\{\Theta(p^k, q^k)\}$ is bounded below, the sequence $\{\Theta(p^k, q^k) - \Theta(p^{k+1}, q^{k+1})\}$ converges to zero and the result follows. □

Suppose now that there exist $\hat{p} \in \mathcal{P}$ and $\hat{q} \in \mathcal{Q}$ for which $\Theta(p, q)$ is minimized over all $p \in \mathcal{P}$ and $q \in \mathcal{Q}$. From the 3PP we have

$$\Theta\big(\hat{p}, q^k\big) \geqslant \Delta\big(\hat{p}, p^{k+1}\big) + \Theta\big(p^{k+1}, q^k\big), \tag{3.8}$$

and we also have

$$\Theta\big(\hat{p}, q^k\big) = \Theta\big(\hat{p}, q^k\big) - \Theta(\hat{p}, \hat{q}) + \Theta(\hat{p}, \hat{q}). \tag{3.9}$$

It follows then that

$$\Theta\big(\hat{p}, q^k\big) - \Theta(\hat{p}, \hat{q}) \geqslant \Delta\big(\hat{p}, p^{k+1}\big). \tag{3.10}$$

We would like to have the related inequality

$$\Delta\big(\hat{p}, p^{k+1}\big) \geqslant \Theta\big(\hat{p}, q^{k+1}\big) - \Theta(\hat{p}, \hat{q}), \tag{3.11}$$

in order to establish the *double inequality*

$$\Theta\big(\hat{p}, q^k\big) \geqslant \Delta\big(\hat{p}, p^{k+1}\big) + \Theta(\hat{p}, \hat{q}) \geqslant \Theta\big(\hat{p}, q^{k+1}\big), \tag{3.12}$$

from which it would follow that the sequences $\{\Theta(\hat{p}, q^k)\}$ and (by rewriting (3.12) with $k + 1$ instead of k) $\{\Delta(\hat{p}, p^k)\}$ are decreasing. The 4PP is precisely what we need to establish the second part of the double inequality (3.12).

Definition 3.2 (The four point property). The function $\Theta(p, q)$ has the 4PP if there is a nonnegative-valued function $\Delta(p, p')$, defined for all $p, p' \in \mathcal{P}$, such that, for any $p \in \mathcal{P}$ and $q \in \mathcal{Q}$ and for every pair of iterative sequences, generated by algorithm 3.1, the following inequality holds,

$$\Delta\big(p, p^k\big) + \Theta(p, q) \geqslant \Theta\big(p, q^k\big). \tag{3.13}$$

Special cases of the double inequality (3.12) have appeared in the literature, although it does not appear in [32] itself; see, e.g., Byrne [15], where it is used in the proof of convergence of the EMML algorithm, and also in Matúš [45], in connection with entropic projections. If there exist $\hat{p} \in \mathcal{P}$ and $\hat{q} \in \mathcal{Q}$ for which $\Theta(p, q)$ is minimized over all $p \in \mathcal{P}$ and $q \in \mathcal{Q}$, then we can conclude that the sequences $\{\Theta(\hat{p}, q^k)\}$ and $\{\Delta(\hat{p}, p^k)\}$ are decreasing, but without further assumptions we cannot conclude that $\{q^k\}$ is convergent.

We now apply the alternating minimization method of Csiszár and Tusnády and the results given above, for $\Theta(p, q)$ in $\mathbb{R}^n$, to the distance $D_{\mathcal{F}}$, defined by (3.2) in the product space $\mathcal{V}$. To do this we let $n = I \times J$, $\mathcal{P} = \overline{\mathcal{S}} \cap \mathcal{C}$ and $\mathcal{Q} = \mathcal{S}$ and identify $\Theta(p, q)$ with $D_{\mathcal{F}}(v, w)$ (and in doing so we also take the freedom to use interchangeably (p, q) and (v, w)). An assumption of "zone consistency" must be made that will guarantee that the sequences $\{p^k\}$ and $\{q^k\}$ remain in $\overline{\mathcal{S}}$ and $\mathcal{S}$, respectively, throughout the iterations. The 3PP then follows from a basic inequality in the theory of Bregman distances, as will be seen below.

In order to have the 4PP for $D_{\mathcal{F}}$ we shall assume that each of the Bregman distances $D_{f_i}(x, z)$ involved is *jointly convex*, that is, convex as a function of the concatenated vector $u = (x, z)$, so that $D_{\mathcal{F}}$ in (3.2) is also a jointly convex Bregman distance. We then invoke the following lemma, due to Eggermont and LaRiccia [36, lemma 2.11].

Lemma 3.3. A jointly convex Bregman distance $D_{\mathcal{F}}$ has the 4PP with $\Delta = D_{\mathcal{F}}$, that is, for any pair of sequences $\{p^k\}$, $\{q^k\}$, generated by algorithm 3.1, and any $p \in \mathcal{P}$ and $q \in \mathcal{Q}$, we have, for all $k \geqslant 0$,

$$D_{\mathcal{F}}\big(p, p^k\big) + D_{\mathcal{F}}(p, q) \geqslant D_{\mathcal{F}}\big(p, q^k\big). \tag{3.14}$$

Proof. By joint convexity we have the inequality

$$D_{\mathcal{F}}(p, q) \geqslant D_{\mathcal{F}}\big(p^k, q^k\big) + \big\langle \nabla_1 D_{\mathcal{F}}\big(p^k, q^k\big), p - p^k \big\rangle + \big\langle \nabla_2 D_{\mathcal{F}}\big(p^k, q^k\big), q - q^k \big\rangle, \tag{3.15}$$

where, for $i = 1, 2$, $\nabla_i D_{\mathcal{F}}(p, q)$ denotes the partial gradient of $D_{\mathcal{F}}$, with respect to the ith vector variable, evaluated at (p, q). Since q^k minimizes $D_{\mathcal{F}}(p^k, q)$ over $q \in \mathcal{Q}$, we have

$$\big\langle \nabla_2 D_{\mathcal{F}}\big(p^k, q^k\big), q - q^k \big\rangle \geqslant 0. \tag{3.16}$$

Using the definition of $D_{\mathcal{F}}$ (see (1.1)), we obtain

$$\big\langle \nabla_1 D_{\mathcal{F}}\big(p^k, q^k\big), p - p^k \big\rangle = \big\langle \nabla \mathcal{F}\big(p^k\big) - \nabla \mathcal{F}\big(q^k\big), p - p^k \big\rangle. \tag{3.17}$$

It follows then that

$$D_{\mathcal{F}}\big(p, q^k\big) - D_{\mathcal{F}}\big(p, p^k\big) = D_{\mathcal{F}}\big(p^k, q^k\big) + \big\langle \nabla_1 D_{\mathcal{F}}\big(p^k, q^k\big), p - p^k \big\rangle \tag{3.18}$$

$$\leqslant D_{\mathcal{F}}(p, q) - \big\langle \nabla_2 D_{\mathcal{F}}\big(p^k, q^k\big), q - q^k \big\rangle \leqslant D_{\mathcal{F}}(p, q), \tag{3.19}$$

from which the 4PP follows. □

Since the set $\mathcal{Q} = \mathcal{S}$ is not closed, it is unlikely that minimizers $\hat{p}$ and $\hat{q}$ exist. As we saw in our discussion of the EMML algorithm, it is usually necessary to extend the proximity function to a portion of the boundary of its original domain of definition. Because the distances D_{f_i} are jointly convex, the proximity function F of (1.2) that we are trying to minimize is a proper convex function and its closure cl F provides the necessary finite extension to (part of) the boundary.

4. The new algorithm and its convergence theory

We turn now to our new iterative algorithm.

Algorithm 4.1.
Initialization: $x^0 \in U$ is arbitrary.
Iterative step: Given x^k find, for all $i = 1, 2, \ldots, I$, the projections $P_{C_i}^{f_i}(x^k)$ and calculate x^{k+1} from

$$\sum_{i=1}^{I} \nabla^2 f_i\big(x^{k+1}\big)x^{k+1} = \sum_{i=1}^{I} \nabla^2 f_i\big(x^{k+1}\big) P_{C_i}^{f_i}\big(x^k\big), \tag{4.1}$$

where $\nabla^2 f_i(x^{k+1})$ denotes the Hessian matrix (of second partial derivatives) of the function f_i at x^{k+1}.

As we shall see below, each x^{k+1} is obtained in this algorithm by minimizing the function

$$F_k(x) \triangleq \sum_{i=1}^{I} D_{f_i}\big(P_{C_i}^{f_i}\big(x^k\big), x\big), \tag{4.2}$$

over all $x \in U$. For algorithm 4.1 to be well-defined we must, therefore, make the following assumptions.

Assumption A1 (Zone consistency). For every $i = 1, 2, \ldots, I$, if $x^k \in S_{f_i}$ then $P_{C_i}^{f_i}(x^k) \in S_{f_i}$.

Assumption A2. For every $k = 1, 2, \ldots,$ the function $F_k(x)$ has a unique minimizer within U.

Clearly, implementation of our algorithm requires solving a series of minimization problems involving Bregman functions and Bregman projections. For particular Bregman functions and convex sets we can solve these problems in closed-form, as the examples in this paper illustrate. However, in most cases approximate techniques, such as the "cyclic subgradient projection" method, see, e.g., Censor and Lent [26], are required.

The following propositions lead to our convergence theorem of algorithm 4.1. First we use the lemma of Eggermont and LaRiccia.

Proposition 4.1. In addition to assumptions A1 and A2, assume that each Bregman distance D_{f_i} is jointly convex. Let $\{x^k\}$ be any sequence generated by algorithm 4.1 and let $x, z \in U$ be arbitrary. Then we have the following inequality:

$$\sum_{i=1}^{I} D_{f_i}\bigl(P_{C_i}^{f_i}(x),\, P_{C_i}^{f_i}\bigl(x^k\bigr)\bigr) \geqslant \sum_{i=1}^{I} D_{f_i}\bigl(P_{C_i}^{f_i}(x),\, x^{k+1}\bigr) - \sum_{i=1}^{I} D_{f_i}\bigl(P_{C_i}^{f_i}(x),\, z\bigr). \tag{4.3}$$

Proof. In the product space $\mathcal{V}$ we construct from the distances D_{f_i} the Bregman distance $D_{\mathcal{F}}$ according to (3.2). The joint convexity of all D_{f_i}'s implies the joint convexity of $D_{\mathcal{F}}$ and lemma 3.3 applies. The vectors p, q, p^k and q^k of lemma 3.3 are all in the product space $\mathcal{V}$ now and we identify p_i with $P_{C_i}^{f_i}(x)$, q with $d(z)$, $(p^k)_i$ with $P_{C_i}^{f_i}(x^k)$, and q^k with $d(x^{k+1})$, respectively, and the result follows. □

Proposition 4.2. Under the assumptions of proposition 4.1, the sequence $\{F(x^k)\}$, where $\{x^k\}$ is any sequence generated by algorithm 4.1, is decreasing.

Proof. From the definitions of F, $F_k(x^k)$ and x^{k+1} in (1.2), (4.2) and (4.1), respectively, we have

$$F\bigl(x^k\bigr) = F_k\bigl(x^k\bigr) \geqslant F_k\bigl(x^{k+1}\bigr). \tag{4.4}$$

The last inequality holds because x^{k+1} is, by assumption A2, by straightforward differentiation of (4.2) and by (4.1), the minimizer of $F_k(x)$ over U. Next we use a well-known inequality in the theory of Bregman projections, see, e.g., [29, theorem 2.4.1], which we quote in the appendix as theorem 6.1 for the reader's convenience. Applying this inequality for each D_{f_i} separately (identifying x^{k+1}, $P_{C_i}^{f_i}$ and $P_{C_i}^{f_i}(x^k)$ here with y, P_Ω and z, in theorem 6.1, respectively) and summing up all inequalities over $i = 1, 2, \ldots, I$, we then obtain

$$F_k\bigl(x^{k+1}\bigr) \geqslant F\bigl(x^{k+1}\bigr) + \sum_{i=1}^{I} D_{f_i}\bigl(P_{C_i}^{f_i}\bigl(x^k\bigr),\, P_{C_i}^{f_i}\bigl(x^{k+1}\bigr)\bigr). \tag{4.5}$$

Combining (4.4), (4.5) and the nonnegativity of the D_{f_i}'s – the result follows. □

Proposition 4.3. Under the assumptions of proposition 4.1 we have, for any sequence $\{x^k\}$ generated by algorithm 4.1,

$$\lim_{k\to\infty} F\bigl(x^k\bigr) = \inf\bigl\{\operatorname{cl} F(x) \mid x \in \mathbb{R}^J\bigr\}. \tag{4.6}$$

Proof. Since $\{F(x^k)\}$ is a real nonnegative decreasing sequence it converges to a nonnegative limit, which we denote by ϕ. Clearly, we have $\phi \geqslant \inf\{\operatorname{cl} F(x) \mid x \in \mathbb{R}^J\}$; to prove equality we denote $\Gamma = \inf\{\operatorname{cl} F(x) \mid x \in \mathbb{R}^J\}$ and assume, by way of negation, that $\phi - \Gamma = \delta > 0$.

Select a $v \in U$ such that $F(x^k) - F(v) \geqslant \delta/2$, for all $k \geqslant 0$. This is always possible since there exists a sequence $\{u^l\}$ with $F(u^l) \to \Gamma$, as $l \to \infty$ and, without

loss of generality, we may assume that $F(u^l) \leqslant \Gamma + \delta/2$, for all $l \geqslant 0$. Therefore, $F(x^k) - F(u^l) \geqslant \delta/2$, for all $k \geqslant 0$ and all $l \geqslant 0$, and we can pick $v = u^l$ for some fixed $l \geqslant 0$.

Using again the basic inequality in the theory of Bregman projections that we used in the previous proposition, applying it for each D_{f_i} separately (this time, identifying x^k, $P_{C_i}^{f_i}$ and $P_{C_i}^{f_i}(v)$ here with y, P_Ω and z, in theorem 6.1, respectively) and summing up all inequalities over $i = 1, 2, \ldots, I$, we obtain, using also (1.2), that, for all $k \geqslant 0$,

$$\sum_{i=1}^{I} D_{f_i}\left(P_{C_i}^{f_i}(v), x^k\right) \geqslant \sum_{i=1}^{I} D_{f_i}\left(P_{C_i}^{f_i}(v), P_{C_i}^{f_i}\left(x^k\right)\right) + F\left(x^k\right). \tag{4.7}$$

In order to use the 4PP we apply lemma 3.3 in a way similar to how it was applied in the proof of proposition 4.1. But we now identify p, q, p^k and q^k of lemma 3.3 as follows: p_i with $P_{C_i}^{f_i}(v)$, q with $d(v)$, $(p^k)_i$ with $P_{C_i}^{f_i}(x^k)$, and q^k with $d(x^{k+1})$, respectively, and get

$$\sum_{i=1}^{I} D_{f_i}\left(P_{C_i}^{f_i}(v), P_{C_i}^{f_i}\left(x^k\right)\right) \geqslant \sum_{i=1}^{I} D_{f_i}\left(P_{C_i}^{f_i}(v), x^{k+1}\right) - \sum_{i=1}^{I} D_{f_i}\left(P_{C_i}^{f_i}(v), v\right). \tag{4.8}$$

From (4.7) and (4.8) we conclude that, for all $k \geqslant 0$,

$$\sum_{i=1}^{I} D_{f_i}\left(P_{C_i}^{f_i}(v), x^k\right) \geqslant \sum_{i=1}^{I} D_{f_i}\left(P_{C_i}^{f_i}(v), x^{k+1}\right) + \frac{\delta}{2}. \tag{4.9}$$

But since Bregman distances are always nonnegative and a decreasing sequence of nonnegative terms has successive differences that converge to zero, we get $\delta = 0$, which is a contradiction and the proof of the proposition is complete. □

Our convergence theorem for algorithm 4.1 now follows. Let $F(x)$ be defined for $x \in U$ by (1.2) and for other $x \in \mathbb{R}^J$ let it be equal to $+\infty$ and let the set of minimizers of cl F over $\mathbb{R}^J$ be denoted by Φ. We also keep the notation $\Gamma = \inf\{\text{cl}\, F(x) \mid x \in \mathbb{R}^J\}$.

There is a subtle point we must discuss before proceeding with the statement and proof of our convergence theorem. It may seem obvious that if cl $F(x) = 0$ for some x then x is in the intersection $C = \bigcap_{i=1}^{I} C_i$, so that C is not empty. We have been unable to prove this, however. The problem reduces to showing that $x^k \to x$ and $D_{f_i}(P_{C_i}^{f_i}(x^k), x^k) \to 0$ imply $P_{C_i}^{f_i}(x^k) \to x$, for each i. This is true for separable Bregman functions f_i, which are those most often employed in practice (see the remarks following B5 in the appendix), and may be true more generally. We therefore make the following assumption (see, however, also the "Note added in proof" at the end of the paper):

Assumption A3. If cl $F(x) = 0$ for some x then x is in $C \cap \overline{U}$.

Theorem 4.1. Let assumptions A1, A2 and A3 hold and assume that the distances D_{f_i} are jointly convex, for all $i = 1, 2, \ldots, I$. In addition, assume that the set Φ is nonempty. If $\operatorname{cl} F$ has a unique minimizer then any sequence $\{x^k\}$, generated by algorithm 4.1, converges to this minimizer. If Φ is not a singleton but $\Gamma = 0$, then the intersection C of the sets C_i is nonempty and $\{x^k\}$ converges to a solution of the CFP.

Proof. Let $\{x^k\}$ be any sequence generated by algorithm 4.1. Assume first that $\operatorname{cl} F$ has a unique minimizer. From proposition 4.2 there is a $\beta > 0$ such that $F(x^k) = \operatorname{cl} F(x^k) \leqslant \beta$, for all $k \geqslant 0$. We recall the useful result, see, e.g., Rockafellar [48, corollary 8.7.1], which states that if the level set $L_\alpha = \{x \mid G(x) \leqslant \alpha\}$ of a closed proper convex function G is nonempty and bounded for a single value of α, then it is bounded for every α. Therefore, if the set Φ of minimizers of $\operatorname{cl} F$ is nonempty and bounded then L_Γ is nonempty and bounded, for $G = \operatorname{cl} F$. Consequently, L_β is bounded, implying that $\{x^k\}$ is bounded. In particular, if $\operatorname{cl} F$ has a unique minimizer, then $\{x^k\}$ is bounded. Let x^* be any cluster point of $\{x^k\}$. By the lower semi-continuity (see, e.g., Rockafellar [48, section 7]) of $\operatorname{cl} F$ we know that $\operatorname{cl} F(x^*) = \Gamma$, so x^* is unique and thus $\lim_{k\to\infty} x^k = x^*$.

Now suppose that Φ contains more then one element but $\Gamma = 0$, so that, by assumption A3, the intersection $C = \bigcap_{i=1}^I C_i$ is nonempty. We take $\hat{x} \in C \cap \overline{U}$ so that $\operatorname{cl} F(\hat{x}) = 0$. For this case we show now:

(i) that any sequence $\{x^k\}$, generated by algorithm 4.1, is bounded;

(ii) that every cluster point of it x^* is in the set C; and

(iii) that the sequence $\{x^k\}$ converges to x^*.

(i) Using once more the basic inequality in the theory of Bregman projections that we used in previous propositions, applying it for each D_{f_i} separately (this time, identifying x^k, $P_{C_i}^{f_i}$ and $\hat{x}$ here with y, P_Ω and z, in theorem 6.1, respectively) and summing up all inequalities over $i = 1, 2, \ldots, I$, we obtain

$$\sum_{i=1}^I D_{f_i}\big(\hat{x}, x^k\big) \geqslant \sum_{i=1}^I D_{f_i}\big(P_{C_i}^{f_i}\big(x^k\big), x^k\big) + \sum_{i=1}^I D_{f_i}\big(\hat{x}, P_{C_i}^{f_i}\big(x^k\big)\big). \tag{4.10}$$

From the 4PP (3.14) we have

$$\sum_{i=1}^I D_{f_i}\big(\hat{x}, P_{C_i}^{f_i}\big(x^k\big)\big) \geqslant \sum_{i=1}^I D_{f_i}\big(\hat{x}, x^{k+1}\big) - \sum_{i=1}^I D_{f_i}(\hat{x}, z), \tag{4.11}$$

for any $z \in U$. This is obtained by identifying p with $d(\hat{x})$, q with $d(z)$, $(p^k)_i$ with $P_{C_i}^{f_i}$ and q^k with $d(x^k)$, in lemma 3.3 and here, respectively. Therefore, combining (4.10) and (4.11),

$$\sum_{i=1}^I D_{f_i}\big(\hat{x}, x^k\big) \geqslant \Gamma + \sum_{i=1}^I D_{f_i}\big(\hat{x}, x^{k+1}\big) - \sum_{i=1}^I D_{f_i}(\hat{x}, z). \tag{4.12}$$

Selecting now $z \in U$ close enough to $\hat{x}$ and using $\Gamma = 0$, it follows that, for all $k \geqslant 0$,

$$\sum_{i=1}^{I} D_{f_i}\left(\hat{x}, x^k\right) \geqslant \sum_{i=1}^{I} D_{f_i}\left(\hat{x}, x^{k+1}\right). \tag{4.13}$$

This shows that the sequence $\{\sum_{i=1}^{I} D_{f_i}(\hat{x}, x^k)\}$ is decreasing, thus bounded and, since the distance functions D_{f_i} are always nonnegative, each $\{D_{f_i}(\hat{x}, x^k)\}$ must be bounded. Therefore, it follows from condition B3 in the definition of Bregman functions (consult the appendix) that the sequence $\{x^k\}$ is bounded.

(ii) Let x^* be any cluster point of $\{x^k\}$. Since cl $F(x^*) = 0$ it follows from assumption A3 that x^* is in the intersection of the sets C_i.

(iii) Using now x^* in place of $\hat{x}$ in the calculations above, we conclude that $\{\sum_{i=1}^{I} D_{f_i}(x^*, x^k)\}$ is decreasing and nonnegative, thus convergent. Since x^* is a cluster point of $\{x^k\}$ we use condition B4 in the definition of Bregman functions (see the appendix) and obtain that for the subsequence of which x^* is a limit $\lim_{k\to\infty} \sum_{i=1}^{I} D_{f_i}(x^*, x^k) = 0$. Therefore, the entire sequence converges to zero and it follows from condition B5 in the definition of Bregman functions (see the appendix) that the sequence $\{x^k\}$ converges to x^*. □

A special case of algorithm 4.1 is the iterative method that we presented in [18]. There the functions f_i are of the form

$$f_i(x) = \sum_{j=1}^{J} w_i^j g_j(x_j), \tag{4.14}$$

where $\{w_j^i\}$ are nonnegative weights such that, for every $j = 1, 2, \ldots, J$, $\sum_{i=1}^{I} w_j^i = 1$, and each $g_j(x_j)$ is a scalar Bregman function with associated Bregman distance (prime denotes derivative):

$$d_j(x_j, z_j) = g_j(x_j) - g_j(z_j) - g_j'(z_j)(x_j - z_j). \tag{4.15}$$

Then each D_{f_i} has the form

$$D_{f_i}(x, z) = \sum_{j=1}^{J} w_j^i d_j(x_j, z_j). \tag{4.16}$$

Equation (4.1) then simplifies and becomes (double prime denotes second derivative):

$$g_j''\left(x_j^{k+1}\right) x_j^{k+1} \left(\sum_{i=1}^{I} w_j^i \right) = g_j''\left(x_j^{k+1}\right) \sum_{i=1}^{I} w_j^i \left(P_{C_i}^{f_i}\left(x^k\right)\right)_j, \tag{4.17}$$

for all $j = 1, 2, \ldots, J$, and algorithm 4.1 takes the following form.

Algorithm 4.2.
Initialization: $x^0 \in U$ is arbitrary.

Iterative step: Given x^k find, for all $i = 1, 2, \ldots, I$, the projections $P_{C_i}^{f_i}(x^k)$ and calculate x^{k+1} from

$$x_j^{k+1} = \sum_{i=1}^{I} w_j^i \left(P_{C_i}^{f_i}\left(x^k\right)\right)_j. \tag{4.18}$$

As noted in [18], special cases include Combettes' iterative algorithm for the Euclidean case [31] and the "Expectation Maximization Maximum Likelihood" (EMML) method, as it occurs in emission tomography, presented in section 2. Algorithm 4.2 was used by Censor, Gordon and Gordon in [22] to obtain a fast image reconstruction method for sparse problems.

5. The split feasibility problem

Next we show how to apply algorithm 4.1 to another interesting problem discussed in [21] by Censor and Elfving.

Problem 5.1 (The split feasibility problem). Given two closed convex sets C, Q in $\mathbb{R}^J$ and an invertible matrix A, find $x \in C$ such that $Ax \in Q$.

For the consistent case, in which there are such x that solve the problem, one can, in principle, use the sequential projection method, projecting orthogonally successively onto the two sets $A(C)$ and Q, where $A(C)$ is the image set of C under the mapping A. However, the set $A(C)$ may not be simple to describe and computing the orthogonal projection onto it may not be easy since this orthogonal projection is equivalent to an oblique projection onto C, followed by A (see [21, section 6.1]). Censor and Elfving were motivated to consider multiprojection algorithms by the desire to replace the orthogonal projection onto $A(C)$ by the orthogonal projection onto C.

The iterative step of their algorithm is the following

$$x^{k+1} = A^{-1}\left(I + AA^{\mathrm{T}}\right)^{-1}\left(AP_C\left(x^k\right) + AA^{\mathrm{T}}P_Q\left(Ax^k\right)\right), \tag{5.1}$$

where A^{-1} and A^{T} are the inverse and the transpose of A, respectively, and P_C and P_Q are the orthogonal projections onto C and Q, respectively. In the consistent case, it follows from [21] that any sequence $\{x^k\}$, generated by (5.1), converges to $x^* \in C$, such that $Ax^* \in Q$.

We put this algorithm into the framework discussed above and prove convergence for the inconsistent case. Let $f_1(x) = \langle Ax, Ax\rangle = ||Ax||^2$ and $f_2(x) = ||x||^2$, with associated Bregman distances $D_1(x, z) = ||x - z||^2_{A^{\mathrm{T}}A}$ and $D_2(x, z) = ||x - z||^2$, where $||x||_H = \langle x, Hx\rangle$ is the "ellipsoidal norm", for any given square symmetric positive definite matrix H.

Applying algorithm 4.1 yields the iterative procedure of (5.1) and now we conclude from theorem 4.1 that the iterative sequence converges in the inconsistent case to a minimizer of the proximity function

$$F(x) = \operatorname{cl} F(x) = D_1\big(P_{A(C)}^{f_1}(x), x\big) + D_2\big(P_Q^{f_2}(x), x\big),$$

whenever such minimizers exist.

6. Summary

There are a number of iterative methods involving generalized Bregman projections onto convex sets that can be used to solve the convex feasibility problem (CFP). Except for the simultaneous multiprojection method of Censor and Elfving [21] these methods employ a single Bregman distance, with respect to which the projections are defined. Typically, these algorithms converge to a member of the intersection of the convex sets, provided that intersection is nonempty. In this paper we have presented a simultaneous multiprojection algorithm for the CFP, involving several distinct Bregman distances. We assumed that these distances are jointly convex, so that the proximity function itself is convex. When the intersection of the convex sets is nonempty, this algorithm converges to a solution of the CFP. When the intersection of the convex sets is empty, the algorithm converges to a minimizer of the closure of a proximity function that measures the average distance to all convex sets, provided such a minimizer exists and is unique.

Note added in proof

The technical conditions appearing in assumptions A1–A3 are not expressed as conditions on the problem parameters directly and, thus, the convergence of the algorithms cannot be determined with certainty by examining these parameters before running the algorithms.

We add to this observation two notes. First, zone consistency, appearing in assumption A1, can be guaranteed if the functions f_i are assumed to be *Legendre functions*, see Bauschke and Borwein [4, theorem 3.14], where more details about the verifiability of zone consistency can be found. Secondly, in a forthcoming paper, by Butnariu, Byrne and Censor [7], we show that assumption A3 always holds.

Appendix: Bregman functions, distances and projections

Let S be a nonempty open convex set in $\mathbb{R}^J$ with closure $\overline{S}$. Let $f : \overline{S} \to \mathbb{R}$ be differentiable and define $D_f(x, z) : \overline{S} \times S \to \mathbb{R}$ by

$$D_f(x, z) = f(x) - f(z) - \big\langle \nabla f(z), x - z \big\rangle. \tag{6.1}$$

Following Censor, Iusem and Zenios [24], we say that f is a Bregman function with zone S if the following conditions are satisfied.

B1. f is continuous and strictly convex on $\overline{S}$;

B2. f is twice continuously differentiable on S and its Hessian matrix $\nabla^2 f(x)$ is positive-definite, for all $x \in S$;

B3. for any fixed $x \in \overline{S}$ the level sets $\{z \mid D_f(x, z) \leqslant \alpha\}$ are bounded;

B4. if $y^k \in S$ and $\{y^k\} \to y^*$ then $D_f(y^*, y^k) \to 0$;

B5. if $x^k \in \overline{S}$ and $y^k \in S$, with $\{x^k\}$ bounded, $\{y^k\} \to y^*$ and $D_f(x^k, y^k) \to 0$, then $\{x^k\} \to y^*$.

Remarks.

(i) It can be shown that, if the Bregman function f is separable, then the condition in B5 that $\{x^k\}$ be bounded is redundant.

(ii) As noted by Bauschke and Borwein [4, remark 4.2], conditions B1–B5 imply that for any fixed $z \in S$ the level sets $\{x \mid D_f(x, z) \leqslant \alpha\}$ are also bounded.

(iii) If f is a Bregman function then D_f is the Bregman distance associated with it. The set S is referred to as the *zone* of f.

(iv) Condition B2 is often replaced with the weaker condition that f be continuously differentiable on S.

(v) Solodov and Svaiter [49] showed recently that condition B5 is redundant.

Let C be a closed convex set in $\mathbb{R}^J$ and $z \in S$ a given point. The *Bregman projection* of z onto C is the point $P_C^f(z) \in C$ which minimizes $D_f(x, z)$ over all $x \in C \cap \overline{S}$. Bregman projections exist and are unique provided that the set C is closed and convex and that $C \cap \overline{S}$ is nonempty (see, e.g., [29, lemma 2.1.2].) Furthermore, we assume that $P_C^f(z) \in S$ whenever $z \in S$ (this is commonly called *zone consistency*.) The basic (and useful) inequality expressed in the next theorem then holds, see, e.g., [29, theorem 2.4.1].

Theorem 6.1. Let f be a Bregman function with zone S and let $\Omega \subseteq \mathbb{R}^J$ be a closed convex set such that $\Omega \cap \overline{S} \neq \emptyset$. Assume that $y \in S$, implies $P_\Omega(y) \in S$. Let $z \in \Omega \cap \overline{S}$, then for any $y \in S$ the inequality

$$D_f\big(P_\Omega(y), y\big) \leqslant D_f(z, y) - D_f\big(z, P_\Omega(y)\big), \tag{6.2}$$

holds.

In [4] Bauschke and Borwein introduced the class of *Bregman/Legendre functions* and demonstrated that this class provides a suitable framework within which to treat Bregman projections. The Bregman/Legendre functions include most of the common Bregman functions, but the two classes are not identical.

Acknowledgments

We thank our colleagues Paul Eggermont, Tommy Elfving and Simeon Reich for enlightening discussions on this research, and two anonymous referees for constructive comments which helped us revise this paper. The work of Y. Censor was partially supported by grants 293/97 and 592/00 of the Israel Science Foundation founded by The Israel Academy of Sciences and Humanities and by NIH grant HL-28438 at the Medical Image Processing Group (MIPG), Department of Radiology, University of Pennsylvania, Philadelphia, PA, USA. Part of this work was done during visits of Y. Censor at the Department of Mathematics of the University of Linköping in Sweden. The support and hospitality of Professor Åke Björck, head of the Numerical Analysis Group there, are gratefully acknowledged.

References

[1] R. Aharoni and Y. Censor, Block-iterative projection methods for parallel computation of solutions to convex feasibility problems, Linear Algebra and its Applications 120 (1989) 165–175.
[2] A. Auslender, *Optimisation: Méthodes Numériques* (Masson, Paris, France, 1976).
[3] H.H. Bauschke and J.M. Borwein, On projection algorithms for solving convex feasibility problems, SIAM Review 38 (1996) 367–426.
[4] H.H. Bauschke and J.M. Borwein, Legendre functions and the method of random Bregman projections, Journal of Convex Analysis 4 (1997) 27–67.
[5] L.M. Bregman, The relaxation method of finding the common point 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] L.M. Bregman, Y. Censor and S. Reich, Dykstra's algorithm as the nonlinear extension of Bregman's optimization method, Journal of Convex Analysis 6 (1999) 319–333.
[7] D. Butnariu, C. Byrne and Y. Censor, What is a Bregman function? Technical Report (August 2000).
[8] D. Butnariu and Y. Censor, On the behavior of a block-iterative projection method for solving convex feasibility problems, International Journal of Computer Mathematics 34 (1990) 79–94.
[9] D. Butnariu and Y. Censor, Strong convergence of almost simultaneous block-iterative projection methods in Hilbert spaces, Journal of Computational and Applied Mathematics 53 (1994) 33–42.
[10] D. Butnariu, Y. Censor and S. Reich, Iterative averaging of entropic projections for solving stochastic convex feasibility, Computational Optimization and Applications 8 (1997) 21–39.
[11] D. Butnariu, Y. Censor and S. Reich, eds., *Inherently Parallel Algorithms in Feasibility and Optimization and Their Applications* (Elsevier Science, Amsterdam, The Netherlands, 2001).
[12] D. Butnariu, A.N. Iusem and R.S. Burachik, Iterative methods for solving stochastic convex feasibility problems and applications, Computational Optimization and Applications 14 (2000) 269–307.
[13] C.L. Byrne, Iterative image reconstruction algorithms based on cross-entropy minimization, IEEE Transactions on Image Processing IP-2 (1993) 96–103.
[14] C.L. Byrne, Erratum and addendum to "Iterative image reconstruction algorithms based on cross-entropy minimization", IEEE Transactions on Image Processing IP-4 (1995) 225–226.
[15] C.L. Byrne, Iterative reconstruction algorithms based on cross-entropy minimization, in: *Image Models (and their Speech Model Cousins)*, eds. S.E. Levinson and L. Shepp, IMA Volumes in Mathematics and its Applications, Vol. 80 (Springer, New York, 1996) pp. 1–11.
[16] C.L. Byrne, Iterative projection onto convex sets using multiple Bregman distances, Inverse Problems 15 (1999) 1295–1313.

[17] C.L. Byrne, Block-iterative interior point optimization methods for image reconstruction from limited data, Inverse Problems 16 (2000) 1405–1419.
[18] C.L. Byrne and Y. Censor, Proximity function minimization for separable, jointly convex Bregman distances, with applications, Technical Report (February 1998).
[19] Y. Censor, Row-action methods for huge and sparse systems and their applications, SIAM Review 23 (1981) 444–464.
[20] Y. Censor, Parallel application of block-iterative methods in medical imaging and radiation therapy, Mathematical Programming 42 (1988) 307–325.
[21] Y. Censor and T. Elfving, A multiprojection algorithm using Bregman projections in a product space, Numerical Algorithms 8 (1994) 221–239.
[22] Y. Censor, D. Gordon and R. Gordon, Component averaging: an efficient iterative parallel algorithm for large and sparse unstructured problems, Parallel Computing 27 (2001) 777–808.
[23] Y. Censor and G.T. Herman, On some optimization techniques in image reconstruction from projections, Applied Numerical Mathematics 3 (1987) 365–391.
[24] Y. Censor, A.N. Iusem and S.A. Zenios, An interior point method with Bregman functions for the variational inequality problem with paramonotone operators, Mathematical Programming 81 (1998) 373–400.
[25] Y. Censor and A. Lent, An iterative row-action method for interval convex programming, Journal of Optimization Theory and Applications 34 (1981) 321–353.
[26] Y. Censor and A. Lent, Cyclic subgradient projections, Mathematical Programming 24 (1982) 233–235.
[27] Y. Censor and S. Reich, Iterations of paracontractions and firmly nonexpansive operators with applications to feasibility and optimization, Optimization 37 (1996) 323–339.
[28] Y. Censor and S. Reich, The Dykstra algorithm with Bregman projections, Communications in Applied Analysis 2 (1998) 407–419.
[29] Y. Censor and S.A. Zenios, *Parallel Optimization: Theory, Algorithms and Applications* (Oxford University Press, New York, NY, USA 1997).
[30] P.L. Combettes, The foundations of set theoretic estimation, Proceedings of the IEEE 81 (1993) 182–208.
[31] P.L. Combettes, Inconsistent signal feasibility: Least-squares solutions in a product space, IEEE Transactions on Signal Processing SP-42 (1994) 2955–2966.
[32] I. Csiszár and G. Tusnády, Information geometry and alternating minimization procedures, Statistics and Decisions, Supp. 1 (1984) 205–237.
[33] A.P. Dempster, N.M. Laird and D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, Journal of the Royal Statistical Society, Series B 37 (1977) 1–38.
[34] R. Dykstra, An algorithm for restricted least squares regression, Journal of the American Statistical Society 78 (1983) 837–842.
[35] R. Dykstra, An iterative procedure for obtaining I-projections onto the intersection of convex sets, The Annals of Probability 13 (1985) 975–984.
[36] P.P.B. Eggermont and V.N. LaRiccia, On EM-like algorithms with EM-like properties for smoothed minimum distance estimation, Technical Report DE 19716, Department of Mathematical Sciences, University of Delaware, Newark, USA (March 1998).
[37] L.G. Gubin, B.T. Polyak and E.V. Raik, The method of projections for finding the common point of convex sets, USSR Computational Mathematics and Mathematical Physics 7 (1967) 1–24.
[38] A.N. Iusem, Convergence analysis for a multiplicatively relaxed EM algorithm, Mathematical Methods in the Applied Sciences 14 (1991) 573–593.
[39] A.N. Iusem, A short convergence proof of the EM algorithm for a specific Poisson model, Revista Brasileira de Probabilidade e Estatistica 6 (1992) 57–67.
[40] A.N. Iusem and A.R. De Pierro, Convergence results for an accelerated nonlinear Cimmino algorithm, Numerische Mathematik 49 (1986) 367–378.

[41] K.C. Kiwiel, Block iterative surrogate projection methods for convex feasibility problems, Linear Algebra and its Applications 215 (1995) 225–259.
[42] K.C. Kiwiel, Free–steering relaxation methods for problems with strictly convex costs and linear constraints, Mathematics of Operations Research 22 (1997) 326–349.
[43] S. Kullback and R. Leibler, On information and sufficiency, Annals of Mathematical Statistics 22 (1951) 79–86.
[44] K. Lange and R. Carson, EM reconstruction algorithms for emission and transmission tomography, Journal of Computer Assisted Tomography 8 (1984) 306–316.
[45] F. Matúš, On iterated averages of I-projections, Technical Report (March 1997).
[46] G.J. McLachlan and T. Krishnan, *The EM Algorithm and Extensions* (Wiley, New York, NY, 1997).
[47] G. Pierra, Decomposition through formalization in a product space, Mathematical Programming 28 (1984) 96–115.
[48] R.T. Rockafellar, *Convex Analysis* (Princeton University Press, Princeton, NJ, 1970).
[49] M.V. Solodov and B.F. Svaiter, An inexact hybrid generalized proximal point algorithm and some new results on the theory of Bregman functions, Mathematics of Operations Research 25 (2000) 214–230.
[50] H. Stark and Y. Yang, *Vector Space Projections: A Numerical Approach to Signal and Image Processing, Neural Nets and Optics* (Wiley, New York, NY, 1998).
[51] E. Tanaka, A fast reconstruction algorithm for stationary positron emission tomography based on a modified EM algorithm, IEEE Transactions on Medical Imaging MI-6 (1987) 98–105.
[52] Y. Vardi, L.A. Shepp and L. Kaufman, A statistical model for positron emission tomography, Journal of the American Statistical Association 80 (1985) 8–20.
[53] D.C. Youla, Mathematical theory of image restoration by the method of convex projections, in: *Image Recovery: Theory and Applications*, ed. H. Stark (Academic Press, Orlando, FL, 1987) pp. 29–78.