# Proximity Function Minimization Using Multiple Bregman Projections, with Applications to Entropy Optimization and Kullback–Leibler Distance Minimization

Charles Byrne (byrnec@cs.uml.edu)
Department of Mathematical Sciences
University of Massachusetts at Lowell
Lowell, MA 01854, USA
and
Yair Censor (yair@mathcs2.haifa.ac.il)
Department of Mathematics
University of Haifa,
Mt. Carmel, Haifa 31905, Israel

June 7, 1999

**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 number of closed, convex sets. When the intersection is empty, one can minimize a proximity function that measures the average distance to all the closed convex sets. Algorithms for these purposes typically employ projections, not necessarily orthogonal, onto the individual convex sets. The multiprojection algorithm of Censor and Elfving provides a simultaneous method for solving the CFP, in which different generalized projections may be used at the same time. Convergence of this multiprojection algorithm follows, for the case of nonempty intersection, from Bregman's theorem on convergence of sequential projections via a product space formulation. An important application of their algorithm is to the *split feasibility problem.* Still open is the issue of convergence of such multiprojection algorithms when the set intersection is empty.
We use here the geometric alternating minimization approach of Csiszár and Tusnády to obtain new multiprojection algorithms for proximity function minimization that converge even in the infeasible case. Special cases of these algorithms include the "Expectation Maximization Maximum Likelihood" (EMML) method in emission tomography, the "Simultaneous Multiplicative Algebraic Reconstruction Technique" (SMART), new methods for image reconstruction

that impose pixel-by-pixel upper and lower bounds on the reconstructed image and the related constrained maximum likelihood algorithm of Vardi and Zhang for estimating mixing distributions in statistics.

# 1 Introduction

Let $C_i$, $i = 1, \dots, I$, be closed convex sets in the $J$–dimensional Euclidean space $R^J$ 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; see, e.g., Youla [58], Combettes [30]. 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 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 [35, 36], Censor and Reich [28], Bregman, Censor and Reich [6] and references therein.

Case (3) is dealt with by finding a point that is, in some sense, close to all the individual $C_i$. 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 this function. If we also wish to impose as a hard constraint that $x$ be a member of another closed convex set $\Omega$, then we minimize the proximity function subject to this additional restriction on $x$. 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 $R^J$ and $f$ a Bregman function from the closure $\overline{S}$ of $S$ into $R$; see, e.g., Censor and Lent [21] or Censor and Zenios [23, Chapter 2].

For a Bregman function $f(x)$, the generalized distance $D_f$ is given 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 $R^J$ 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* (see Appendix B at the end of this paper). With $g_j(t) = g(t) = t^2$, for each $j$, 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 *generalized projection* of $x \in S$ onto the set $C_i$ with respect to the *generalized distance* $D_f$; that is, for any $x \in S$ we have $D_f(P_{C_i}^f(x), x) \leq D_f(z,x)$, for all $z \in C_i \bigcap \overline{S}$. If $C \triangleq \bigcap_{i=1}^{I} C_i$ is nonempty then the sequential iterative algorithm of successive projections $x^{k+1} = P_{i(k)}^f(x^k)$ converges to a member of $C$. This was shown by Bregman [5] for the *cyclic control* $i(k) = k \pmod I + 1$, $k \geq 0$, by Censor and Reich [27] and by Bauschke and Borwein [3] 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 [40] 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 weighting for orthogonal projections, see, e.g., Aharoni and Censor [1], Bauschke and Borwein [2], Butnariu and Censor [7, 8], Censor [19, 20], Combettes [30, 31], Iusem and De Pierro [45], Kiwiel [46] and references therein. Recently, Censor and Elfving [24] proposed and studied a simultaneous projections algorithm for the convex feasibility problem that employs generalized projections of the Bregman type. However, the weighting of the simultaneous projections there is not arithmetic, but depends on the choice of the Bregman function (or functions).

Byrne and Censor [18] studied recently simultaneous methods with arithmetic weighting for generalized projections that are not necessarily orthogonal. Such a possibility was mentioned, in passing, by Censor and Herman [25, Section 4.4], and was recently studied for the special case of entropic projections in Butnariu, Censor and Reich [9]; the results in [9] deal only with the consistent case $C \neq \emptyset$. The focus in [18] was on the behavior of simultaneous methods with arithmetic averaging for generalized projections in the inconsistent case $C = \emptyset$. It was shown there that, if $D_f$ is separable and jointly convex, then such methods converge to a minimizer of a proximity function $F(x)$ that measures the average generalized distance of $x$ to the family $\{C_i\}_{i=1}^I$. Recent work by Butnariu, Iusem and Burachik [10] on stochastic convex feasibility problems contains a similar proximity function minimization algorithm and notes the importance of joint convexity of the distance.

In the standard presentation of Bregman functions and distances the *zone* $S$ is an open convex set with closure $\overline{S}$. The Bregman distance $D_f(z,x)$ is defined for $z \in \overline{S}$ and $x \in S$ and the Bregman projections $P_i(x)$ are defined for $x \in S$. In [18] the definition of the distance $D_f(z,x)$, the projection $P_i(x)$ and, thereby, the proximity function $F(x)$ are extended to include $x \in \overline{S}$. This 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 [47] and in [23, Section 6.8]. We adopt this approach in this paper as well, again restricting our discussion to separable Bregman functions.

In contrast with [18], we shall be concerned here with proximity functions of the multiprojection type, defined for $x \in \bigcap_{i=1}^I S_i$, by

$$F(x) = \sum_{i=1}^{I} D_{f_i}(P_i(x), x), \tag{1.2}$$

where, for $i = 1, 2, \ldots, I$, the $D_{f_i}(z,x)$ are generalized distances derived from separable Bregman functions $f_i$ and the $P_i = P_{C_i}^{f_i}$ are the associated Bregman projections onto the $C_i$. In addition, we shall consider the *second variable* projection of $x$ onto the $C_i$, that is, the member $\tilde{P}_i(x) = \tilde{P}_{C_i}^{f_i}(x)$ of $C_i$ for which the quantity $D_{f_i}(x,z)$ is minimized over all $z \in C_i \bigcap S$, provided that the minimum exists and is attained. The associated proximity function to be minimized in this case is $\tilde{F}(x)$ having the form

$$\tilde{F}(x) = \sum_{i=1}^{I} D_{f_i}(x, \tilde{P}_i(x)). \tag{1.3}$$

In what follows we shall make considerable use of three important tools of mathematical algorithm design. The first tool is the reformulation of the problem in a product space, as suggested by Pierra [52]. The second tool is the concept of generalized distances and their projections onto convex sets, as introduced by Bregman in [5] and studied extensively under the names *Bregman distances* and *Bregman projections* by Censor and co-authors and by others (see [23] and the references therein). The third tool is the framework of alternating minimization of a functional of two vector variables, as proposed by Csiszár and Tusnády [32]. The first two of these tools were used in the work of Censor and Elfving [24], but, because they were only concerned with the feasible case, they used Bregman's successive projections approach, instead of the alternating minimization method of [32]. The proximity function minimization algorithm developed in [18] can 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 [38] make use of alternating minimization and prove the useful result 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.

To keep this presentation within reasonable bounds and make it accessible and useful to a wide audience we adopt a tutorial style, deliberately sacrificing some mathematical rigour of the presentation at places in order to present the new algorithms themselves and their applications as clearly as possible.

# 2 Pierra's product space formulation

Given closed convex subsets $C_i$, $i = 1, \ldots, I$, with (possibly empty) intersection $C \triangleq \bigcap_{i=1}^{I} C_i$, we reformulate the CFP in a product space framework. Following Pierra [52] we let $\mathcal{V}$ be the product of $I$ copies of the Euclidean space $R^J$, so that a typical element $v = (v_1, v_2, \ldots, v_I)$ of $\mathcal{V}$ is such that $v_i \in R^J$, $i = 1, \ldots, I$. We define $\mathcal{C} = \Pi_{i=1}^{I} C_i$ to be the product of all sets $C_i$, i.e., the subspace of $\mathcal{V}$ consisting of all $v$ such that $v_i \in C_i$, $i = 1, \ldots, I$, and we let $\mathcal{D}$ be the ("diagonal") subspace of $\mathcal{V}$ consisting of all $v$ such that $v_i = x$, $i = 1, \ldots, I$, where $x \in R^J$, and express this by writing $v = d(x)$. Our goal is to find a member of $\mathcal{V}$ in $\mathcal{C} \bigcap \mathcal{D}$. It is easy to verify that an element $d(x^*)$ belongs to $\mathcal{D} \bigcap \mathcal{C}$, 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 $R^J$.

We shall have occasion later to consider the problem of minimizing a proximity function over all $x$ within a given closed convex set $\Omega$. For such problems we let $\mathcal{H}$ be the subspace of $\mathcal{V}$ consisting of all $v$ such that $v_i = x$, $i = 1, \ldots, I$, where $x \in \Omega$ and we write for this $v = h(x)$.

In [24] 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 generalized distance in $\mathcal{V}$, given by,

$$D_\lambda(v,w) \triangleq \sum_{i=1}^{I} \lambda_i D_{f_i}(v_i, w_i), \tag{2.1}$$

where $\lambda = (\lambda_i) \in R^I$ is a fixed vector such that all $\lambda_i$ are positive and $\sum_{i=1}^{I} \lambda_i = 1$ .

Here we construct a distance measure $D(v,w)$ between $v \in \overline{\mathcal{S}}$ and $w \in \mathcal{S}$, where $\mathcal{S} \triangleq \Pi_{i=1}^{I} S_i$, as follows:

$$D(v,w) = \sum_{i=1}^{I} D_{f_i}(v_i, w_i), \tag{2.2}$$

where $D_{f_i}$ is the Bregman distance associated with the Bregman function $f_i$ with zone $S_i$. With this distance at hand we attempt to solve the CFP by finding iterative algorithms that will minimize $D(\alpha, \beta)$, 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 [32], which we describe in the following section.

# 3 The alternating minimization method of Csiszár and Tusnády

In this section we present a slightly simplified version of the alternating minimization method of Csiszár and Tusnády [32]. Suppose that $\mathcal{P}$ and $\mathcal{Q}$ are two closed convex sets in the $n$-dimensional Euclidean space $R^n$. 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 method)**
**Initialization:** $q^0 \in \mathcal{Q}$ is arbitrary.
**Iterative Step:** Given $q^k$ find $p^{k+1}$ by solving

$$p^{k+1} = \operatorname{argmin}\{\Theta(p, q^k) \mid p \in \mathcal{P}\}, \tag{3.1}$$

then calculate $q^{k+1}$ by solving

$$q^{k+1} = \operatorname{argmin}\{\Theta(p^{k+1}, q) \mid q \in \mathcal{Q}\}. \tag{3.2}$$

Assuming that all the minima exist, the sequences $\{p^k\}$, $\{q^k\}$ are obtained. Define

$$F_k \triangleq \Theta(p^k, q^k). \tag{3.3}$$

We then have the following monotonicity result:

**Lemma 3.1** *The sequence $\{F_k\}$ is decreasing.*

**Proof:** We have

$$F_k = \Theta(p^k, q^k) \geq \Theta(p^{k+1}, q^k) \geq \Theta(p^{k+1}, q^{k+1}) = F_{k+1}. \tag{3.4}$$

∎

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 three-point property 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, defined by Algorithm 3.1, the following inequality holds:*

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

In many applications $\Theta(p,q) \geq 0$ and $\Delta(p,p') = \Theta(p,p')$. As we shall see, this holds for the distance measure defined in (2.2).

**Lemma 3.2** *If $\Theta(p,q)$ has the 3PP and if the sequence $\{F_k\}$ is bounded below (in particular, if $F_k \geq 0$) then the sequence $\{\Delta(p^k, p^{k+1})\}$ converges to zero.*

**Proof:** Using the 3PP and the definitions of the vectors $p^k$ and $q^k$, we have:

$$F_k = \Theta(p^k, q^k) \geq \Delta(p^k, p^{k+1}) + \Theta(p^{k+1}, q^k) \geq \Theta(p^{k+1}, q^{k+1}) = F_{k+1}. \tag{3.6}$$

Since $\{F_k\}$ is bounded below, the sequence $\{F_k - F_{k+1}\}$ 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

$$\Theta(\hat{p}, q^k) \geq \Delta(\hat{p}, p^{k+1}) + \Theta(p^{k+1}, q^k), \tag{3.7}$$

and we also have

$$\Theta(\hat{p}, q^k) = \Theta(\hat{p}, q^k) - \Theta(\hat{p}, \hat{q}) + \Theta(\hat{p}, \hat{q}). \tag{3.8}$$

It follows then that

$$\Theta(\hat{p}, q^k) - \Theta(\hat{p}, \hat{q}) \geq \Delta(\hat{p}, p^{k+1}). \tag{3.9}$$

We would like to have the related inequality

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

in order to establish the *double inequality*

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

from which it would follow that the sequences $\{\Theta(\hat{p}, q^k)\}$ and $\{\Delta(\hat{p}, p^k)\}$ are decreasing. The 4PP is precisely what we need to establish the second part of the double inequality (3.11).

**Definition 3.2 (The four-point property)** *The function $\Theta(p, q)$ has the four-point property 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, defined by Algorithm 3.1, the following inequality holds,*

$$\Delta(p, p^k) + \Theta(p, q) \geq \Theta(p, q^k). \tag{3.12}$$

Special cases of the double inequality (3.11) have appeared in the literature, although it does not appear in [32] itself; see, e.g., Byrne [13], where it is used in the proof of convergence of the EMML algorithm, and also in Matúš [50], in connection with entropic projections.

We now apply the alternating minimization method of Csiszár and Tusnády and the results given above for $\Theta(p, q)$ in $R^n$ to the distance measure $D$, defined by (2.2) in the product space $\mathcal{V}$. To do this we let $n = I \times J$ and identify $\Theta(p, q)$ with $D(v, w)$ (and in doing so we also take the freedom to use interchangeably $(p, q)$ and $(v, w)$). Then, of course, we must assume that either $\mathcal{P} \subseteq \overline{\mathcal{S}}$ and $\mathcal{Q} \subseteq \mathcal{S}$ or that $\mathcal{P}$ and $\mathcal{Q}$ have nonempty intersections with $\overline{\mathcal{S}}$ and $\mathcal{S}$, respectively. In the latter case 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 (see Assumption 8.1 in Appendix B). The 3PP then follows from a standard inequality in the theory of Bregman distances, i.e., inequality (8.5) in Appendix B. In order to have the 4PP for $D$ 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 $D$ in (2.2) is also a jointly-convex Bregman distance. We then invoke the following lemma, due to Eggermont and LaRiccia [38, Lemma 2.11]:

**Lemma 3.3** *A jointly-convex Bregman distance $D_f$ has the 4PP with $\Delta = D_f$, that is*

$$D_f(p, p^k) + D_f(p, q) \geq D_f(p, q^k). \tag{3.13}$$

**Proof:** By joint-convexity we have the inequality:

$$D_f(p,q) \geq D_f(p^k,q^k) + \left\langle \nabla_1 D_f(p^k,q^k), p-p^k \right\rangle + \left\langle \nabla_2 D_f(p^k,q^k), q-q^k \right\rangle, \quad (3.14)$$

where $\nabla_i D_f(p,q)$ denotes the partial gradient of $D_f$, with respect to the $i$th vector variable, evaluated at $(p,q)$. Since $q^k$ minimizes $D_f(p^k,q)$ over $q$, we have

$$\left\langle \nabla_2 D_f(p^k,q^k), q-q^k \right\rangle \geq 0. \quad (3.15)$$

Using the definition of $D_f$ (see (1.1)), we obtain

$$\left\langle \nabla_1 D_f(p^k,q^k), p-p^k \right\rangle = \left\langle \nabla f(p^k) - \nabla f(q^k), p-p^k \right\rangle. \quad (3.16)$$

It follows then that

$$\begin{aligned} D_f(p,q^k) - D_f(p,p^k) &= D_f(p^k,q^k) + \left\langle \nabla_1 D_f(p^k,q^k), p-p^k \right\rangle & (3.17) \\ &\leq D_f(p,q) - \left\langle \nabla_2 D_f(p^k,q^k), q-q^k \right\rangle \leq D_f(p,q), & (3.18) \end{aligned}$$

from which the 4PP follows. ∎

Next, we impose further restrictions that will enable us to prove convergence of the iterative sequences to a minimizing pair $(\hat{p},\hat{q})$. We assume that $D(v,w) = \Delta(v,w) \geq 0$ is defined for $v,w \in \mathcal{S} \subseteq \mathcal{V}$ and that it has the 3PP and the 4PP. In the examples considered later, these conditions hold. We also make the following assumptions on $D$.

**Assumption 3.1 (Bounded level sets)** For any fixed $v \in \mathcal{S}$ and $t \geq 0$, the set $\{w \,|\, D(v,w) \leq t\}$ is bounded. Likewise, for any fixed $w \in \mathcal{S}$ and $t \geq 0$, the set $\{v \,|\, D(v,w) \leq t\}$ is bounded.

**Assumption 3.2** The points $\hat{p}$ and $q^0$ are chosen so that $D(\hat{p},q^0)$ is finite.

**Assumption 3.3** If $\{D(p,p^k)\}$ converges to zero, for some $p$ and some bounded sequence $\{p^k\}$, then $\{p^k\}$ converges to $p$.

From the double inequality (3.11) we know that the sequence $\{D(\hat{p},q^k)\}$ is decreasing and from Assumption 3.1 it follows that $\{q^k\}$ is bounded, so we can extract a subsequence converging to $q^*$. Let $p^*$ minimize $D(p,q^*)$. Taking limits in the double inequality (3.11) we have

$$D(\hat{p},q^*) = D(\hat{p},\hat{q}) + D(\hat{p},p^*), \quad (3.19)$$

while, from the 3PP, we also have

$$D(\hat{p},q^*) \geq D(p^*,q^*) + D(\hat{p},p^*). \quad (3.20)$$

Since $\{D(\hat{p}, q^k)\}$ is decreasing, it follows from Assumption 3.2 that $D(\hat{p}, q^*)$ is finite, and so $D(p^*, q^*) \leq D(\hat{p}, \hat{q})$; the pair $(p^*, q^*)$, therefore, minimizes $D(p, q)$. We can replace $\hat{p}$ with $p^*$ in the double inequality, and conclude that the sequence $\{D(p^*, p^{k+1})\}$ converges to zero. From Assumption 3.3 we then have that $p^k$ converges to $p^*$.

We summarize these results in the following theorem:

**Theorem 3.1** *Let $D$ satisfy the 3PP, 4PP and Assumptions 3.1–3.3, listed above. Let $\hat{p}$ and $\hat{q}$ be such that $D(\hat{p}, \hat{q}) \leq D(p, q)$, for all $p$ and $q$. Then $\{p^k\}$ converges to $p^*$, $\{q^k\}$ converges to $q^*$ and the pair $(p^*, q^*)$ satisfies that $D(p^*, q^*) \leq D(p, q)$, for all $p$ and $q$.*

# 4 The main results

In this section we present our new fully simultaneous algorithms which employ extended Bregman projections onto the convex sets $\{C_i\}_{i=1}^I$ in $R^J$. The main algorithmic difference between these algorithms and the multiprojections algorithm of Censor and Elfving [24] (see also [23, Section 5.9]) is the fact that here we use alternating minimizations, instead of successive projections. For symmetric distances the two approaches coincide. The multiprojections algorithm of Censor and Elfving has been shown to converge, so far, only in the consistent case $\bigcap_{i=1}^I C_i \neq \emptyset$, whereas our convergence results apply to both the consistent and inconsistent situations. In the inconsistent case our algorithms become minimization tools for the proximity functions defined below, and a minimizer might occur on the boundary of the zone $S_i$ of the Bregman function $f_i$. This possibility is the driving force behind our construction of *extended separable Bregman functions* in Appendices A and B.

We assume that $D(v, w)$ is defined in the product space $\mathcal{V}$ by (2.2) and that the $D_{f_i}$'s are jointly-convex separable extended Bregman distances, as defined in Appendix B. As we reformulate our problem within the product space framework, in order to apply the alternating minimization technique, we find that we must decide which of sets $\mathcal{C}$ or $\mathcal{D}$ is to be identified with $\mathcal{P}$. Our algorithm will depend on this choice. In Subsection 4.1 we consider the case in which $\mathcal{C}$ is taken to be $\mathcal{P}$ while the other choice is taken up in Subsection 4.2. In the last two subsections we discuss the modifications that must be made when $\mathcal{D}$ is replaced by $\mathcal{H}$.

## 4.1 The case in which $\mathcal{P}$ is identified with $\mathcal{C}$ and $\mathcal{Q}$ is identified with $\mathcal{D}$

Here we let the set $\mathcal{C}$ in the product space $\mathcal{V}$ play the role of $\mathcal{P}$ and let $\mathcal{D} \subseteq \mathcal{V}$ be $\mathcal{Q}$ of Section 3. We assume that $\{f_i\}_{i=1}^I$ is a family of separable extended Bregman functions with zones $\{S_i\}_{i=1}^I$, as defined in Appendix B, and that $\overline{S_i} \bigcap C_i \neq \emptyset$, for all $1 \leq i \leq I$. As discussed in Appendix B, we denote the extended Bregman projection of $z$ onto $C_i$ with respect to $f_i$, defined for all $z \in \mathrm{dom} P_i$, by

$$P_i(z) \triangleq P_{C_i}^{f_i}(z). \tag{4.1}$$

The *proximity function of the family of sets* $\{C_i\}_{i=1}^I$ *with respect to the family of separable extended Bregman functions* $\{f_i\}_{i=1}^I$, given in (8.1), is defined as

$$F(x) \triangleq \sum_{i=1}^{I} D_i(P_i(x), x), \tag{4.2}$$

for all $x$ in $\mathrm{dom} F \triangleq \bigcap_{i=1}^I \mathrm{dom} P_i$.

**Algorithm 4.1**
**Initialization:** $x^0 \in \mathrm{dom} F$ is arbitrary.
**Iterative Step:** Given $x^k$ find, for all $i = 1, \ldots, I$, the projections $P_i(x^k)$ and calculate $x^{k+1}$ from

$$\sum_{i=1}^{I} \nabla^2 f_i(x^{k+1}) x^{k+1} = \sum_{i=1}^{I} \nabla^2 f_i(x^{k+1}) P_i(x^k), \tag{4.3}$$

where $\nabla^2 f_i(x^{k+1})$ denotes the Hessian matrix (of second partial derivatives) of the function $f_i$, evaluated at $x^{k+1}$.
Since the $f_i$ are separable, we can rewrite (4.3) as

$$x_j^{k+1} \sum_{i=1}^{I} g''_{ij}(x_j^{k+1}) = \sum_{i=1}^{I} g''_{ij}(x_j^{k+1})(P_i(x^k))_j. \tag{4.4}$$

In order to minimize the proximity function over all $x \in \mathrm{dom} F$, we set $q^0 = d(x^0)$ to initialize the application of Algorithm 3.1 in the product space. In the iterative step of Algorithm 3.1, given $q^k = d(x^k)$, we solve the minimization of (3.1) by letting $p^{k+1} = (P_1(x^k), P_2(x^k), \ldots, P_I(x^k))$. This is precisely the expression for the projection of $q^k$ onto $\mathcal{C}$ in the product space $\mathcal{V}$ according to the separable extended Bregman distance $D(p, q)$ defined by (2.2) with the family of separable extended Bregman functions defined in Appendix B. This follows from an argument similar to the one used in the proof of [24, Lemma 4.1] (also appearing in [23, Lemma 5.9.2]). From this $p^{k+1}$ we then calculate $q^{k+1}$ of (3.2). This minimization is realized by $q^{k+1} = d(x^{k+1})$ where $x^{k+1}$ is the solution of (4.3), as can be verified along similar lines to those of [24, Lemma 4.2] (also appearing in [23, Lemma 5.9.3]). Admittedly, the ability to actually solve (4.3) for $x^{k+1}$ in practice cannot always be guaranteed. We have made the additional assumption that, for all $1 \leq i \leq I$, the generalized distances $D_i(x, z)$ are *jointly-convex* with respect to both $x$ and $z$. This implies the joint convexity of $D$ in (2.2), as well as the convexity of $F$. From our analysis of extended Bregman functions and distances in Appendices A and B, we know that Assumptions 3.1–3.3, needed for proving convergence in the previous section, hold. We, therefore, conclude that the iterative procedure of Algorithm 4.1 converges to a minimizer of the proximity function $F(x)$ whenever it has minimizers.

A special case of this algorithm is the iterative method presented in [18]. There the functions composing each $f_i$ in (8.1) of Appendix B are of the form

$$g_{ij} = w_i^j g_j(x_j)$$

where the $w_j^i$ are nonnegative weights such that, for each $j$, $\sum_{i=1}^{I} w_j^i = 1$, and each $g_i(t)$ is an extended scalar Bregman function as in Definition 7.1 of Appendix A. Then each $D_i$ has the form (8.2) with $d_{ij}(x_j, z_j) = w_j^i d_j(x_j, z_j)$, for all $i$ and $j$. Equation (4.4) then simplifies and becomes

$$g_j''(x_j^{k+1}) x_j^{k+1} \left(\sum_{i=1}^{I} w_j^i\right) = g_j''(x_j^{k+1}) \sum_{i=1}^{I} w_j^i (P_i(x^k))_j, \tag{4.5}$$

so that, for all $j = 1, \ldots, J$,

$$x_j^{k+1} = \sum_{i=1}^{I} w_j^i (P_i(x^k))_j. \tag{4.6}$$

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. See, e.g., Vardi, Shepp and Kaufman [55] and also Section 6 below.

When, instead of the choices made at the beginning of this subsection, we make $\mathcal{C}$ the set $\mathcal{Q}$ and $\mathcal{D}$ the set $\mathcal{P}$ we get a different algorithm, as we discuss next.

## 4.2 The case in which $\mathcal{P}$ is identified with $\mathcal{D}$ and $\mathcal{Q}$ is identified with $\mathcal{C}$

Now we interchange the roles of the sets chosen in the beginning of the previous subsection and let the set $\mathcal{D}$ in the product space $\mathcal{V}$ play the role of $\mathcal{P}$ and let

$\mathcal{C} \subseteq \mathcal{V}$ be $\mathcal{Q}$ of Section 3. We again assume that $\{f_i\}_{i=1}^I$ is a family of separable extended Bregman functions with zones $\{S_i\}_{i=1}^I$, as defined in Appendix B, and that $\overline{S_i} \bigcap C_i \neq \emptyset$, for all $1 \leq i \leq I$. In contrast with the previous subsection, we now look at

$$\tilde{P}_i(x) \triangleq \tilde{P}_{C_i}^{f_i}(x), \tag{4.7}$$

the *second-variable* extended Bregman projection of $x$ onto $C_i$ with respect to $f_i$, defined for all $x \in \mathrm{dom}\tilde{P}_i$, where (compare with (8.4))

$$\mathrm{dom}\tilde{P}_i \triangleq \{x \in \overline{S}_i \mid D_i(x,z) < +\infty, \, \text{for some} \, z \in C_i \bigcap \overline{S}_i\}.$$

The *proximity function of the family of sets* $\{C_i\}_{i=1}^I$ *with respect to the family of separable extended Bregman functions* $\{f_i\}_{i=1}^I$, given in (8.1), is, in this case, defined as

$$\tilde{F}(x) \triangleq \sum_{i=1}^{I} D_i(x, \tilde{P}_i(x)), \tag{4.8}$$

for all $x$ in $\mathrm{dom}\tilde{F} \triangleq \bigcap_{i=1}^I \mathrm{dom}\tilde{P}_i$.

**Algorithm 4.2**
**Initialization:** $x^0 \in \mathrm{dom}\tilde{F}$ is arbitrary.
**Iterative Step:** Given $x^k$ find, for all $i = 1, \ldots, I$, the projections $\tilde{P}_i(x^k)$ and calculate $x^{k+1}$ from

$$\sum_{i=1}^{I} \nabla f_i(x^{k+1}) = \sum_{i=1}^{I} \nabla f_i(\tilde{P}_i(x^k)), \tag{4.9}$$

Since the $f_i$ are separable, we can rewrite (4.9), for $j = 1, \ldots, J$, as

$$\sum_{i=1}^{I} g'_{ij}(x_j^{k+1}) = \sum_{i=1}^{I} g'_{ij}((\tilde{P}_i(x^k))_j). \tag{4.10}$$

The justification of Algorithm 4.2 is done along lines similar, but not identical, to those of the previous subsection. In order to minimize the proximity function over all $z \in \mathrm{dom}\tilde{F}$, we set now $p^0 = d(x^0)$ to initialize the application of Algorithm 3.1 in the product space. Note that for this case we must apply Algorithm 3.1 by doing first (3.2) and then (3.1) in every iterative step. So, given $p^k = d(x^k)$, we solve the minimization of $\Theta(p^k, q) = D(p^k, q)$ by letting $q^k = (\tilde{P}_1(x^k), \tilde{P}_2(x^k), \ldots, \tilde{P}_I(x^k))$. Now this is precisely the expression for the second-variable projection of $p^k$ onto $\mathcal{C}$ in the product space $\mathcal{V}$ according to the separable extended Bregman distance $D(p,q)$ defined by (2.2) with the family of separable extended Bregman functions defined in Appendix B. Again, this follows from an argument similar to the one used in the proof of [24, Lemma 4.1] (also appearing in [23, Lemma 5.9.2]). From this $q^k$ we then calculate $p^{k+1}$ by doing the other minimization. This minimization is realized by $p^{k+1} = d(x^{k+1})$ where $x^{k+1}$ is the solution of (4.9), as can be verified along similar lines to those of [24, Lemma 4.2] (also appearing in [23, Lemma 5.9.3]). Once again we must admit that the ability to actually solve (4.9) for $x^{k+1}$ in practice cannot always be guaranteed. Having made, as before, the additional assumption that, for all $1 \le i \le I$, the generalized distances $D_i(x,z)$ are *jointly-convex* with respect to both $x$ and $z$. This implies the joint convexity of $D$ in (2.2), as well as the convexity of $\tilde{F}$. From our analysis of extended Bregman functions and distances in Appendices A and B, we know that Assumptions 3.1–3.3, needed for proving convergence in the previous section, hold. We, therefore, conclude that the iterative procedure of Algorithm 4.2 converges to a minimizer of the proximity function $\tilde{F}(x)$ whenever it has minimizers.

A special case of this algorithm is the "simultaneous multiplicative algebraic reconstruction technique" (SMART) presented in [11, 29] (see also Section 6 below).

### 4.3 The case in which $\mathcal{P}$ is identified with $\mathcal{C}$ and $\mathcal{Q}$ is identified with $\mathcal{H}$

In this and the next subsections we replace the "diagonal" set $\mathcal{D}$ with the "sub-diagonal" set

$$\mathcal{H} \triangleq \{v \in \mathcal{V} \mid v_i = x,\ i = 1, \ldots, I,\ x \in \Omega\},$$

where $\Omega \subseteq R^J$ is some given closed convex set, see Section 2. Changing again the roles of the sets $\mathcal{P}$ and $\mathcal{Q}$ chosen in the previous subsections, we let now the set $\mathcal{C}$ in the product space $\mathcal{V}$ play the role of $\mathcal{P}$ and let $\mathcal{H} \subseteq \mathcal{V}$ be $\mathcal{Q}$ of Section 3. We again assume that $\{f_i\}_{i=1}^I$ is a family of separable extended Bregman functions with zones $\{S_i\}_{i=1}^I$, as defined in Appendix B, and that $\overline{S_i} \bigcap C_i \ne \emptyset$, for all $1 \le i \le I$. We now look again at

$$P_i(x) \triangleq P_{C_i}^{f_i}(x), \tag{4.11}$$

the extended Bregman projection of $x$ onto $C_i$ with respect to $f_i$, defined for all $x \in \mathrm{dom} P_i$.

The *proximity function of the family of sets* $\{C_i\}_{i=1}^I$ *with respect to the family of separable extended Bregman functions* $\{f_i\}_{i=1}^I$, given in (8.1), is, in this case, defined as

$$F_\Omega(x) \triangleq \sum_{i=1}^{I} D_i(P_i(x), x), \tag{4.12}$$

for all $x$ in

$$\text{dom}F_\Omega \triangleq \Omega \bigcap \Big( \bigcap_{i=1}^{I} \text{dom}P_i \Big).$$

**Algorithm 4.3**
**Initialization:** $x^0 \in \text{dom}F_\Omega$ is arbitrary.
**Iterative Step:** Given $x^k$ find, for all $i = 1, \ldots, I$, the projections $P_i(x^k)$ and calculate $x^{k+1}$ from

$$x^{k+1} = \text{argmin}\,\{\sum_{i=1}^{I} D_i(P_i(x^k), x) \mid x \in \Omega\} \tag{4.13}$$

The justification of Algorithm 4.3 is again done along lines similar, but not identical, to those of the previous subsections. In order to minimize the proximity function of (4.12) over all $x \in \text{dom}F_\Omega$, we set now $q^0 = h(x^0)$ to initialize the application of Algorithm 3.1 in the product space. So, given $q^k = h(x^k)$, we first solve the minimization of $D(p, q^k)$ by letting $p^{k+1} = (P_1(x^k), P_2(x^k), \ldots, P_I(x^k))$. This is precisely the expression for the projection of $q^k$ onto $\mathcal{C}$ in the product space $\mathcal{V}$ according to the separable extended Bregman distance $D(p, q)$ defined by (2.2) with the family of separable extended Bregman functions defined in Appendix B. Again, this follows from an argument similar to the one used in the proof of [24, Lemma 4.1] (also appearing in [23, Lemma 5.9.2]). From this $p^{k+1}$ we then calculate $q^{k+1}$ by doing the other minimization. This minimization is realized by $q^{k+1} = h(x^{k+1})$ where $x^{k+1}$ is the solution of (4.13). Having made, as before, the additional assumption that, for all $1 \leq i \leq I$, the generalized distances $D_i(x, z)$ are *jointly-convex* with respect to both $x$ and $z$. This implies the joint convexity of $D$ in (2.2), as well as the convexity of $F_\Omega$. From our analysis of extended Bregman functions and distances in Appendices A and B, we know that Assumptions 3.1–3.3, needed for proving convergence in the previous section, hold. We, therefore, conclude that the iterative procedure of Algorithm 4.3 converges to a minimizer of the proximity function $F_\Omega(x)$ whenever it has minimizers.

A special case of this algorithm is the iterative method of Vardi and Zhang, presented in [56, 57], for maximum likelihood estimation of mixing probabilities. In that work the authors consider a random variable $Z$, whose values lie in the set

$\{1, 2, \ldots, I\}$, such that the probability that $Z$ takes on the value $i$ is the entry $g_i$ of the probability vector $g = (g_1, \ldots, g_I)^T$, given by

$$g_i = \sum_{j=1}^{J} f_j p_{ij}, \;\; i = 1, \ldots, I, \tag{4.14}$$

where the $p_{ij}$ are known nonnegative weights such that, for every $j$, $\sum_{i=1}^{I} p_{ij} = 1$ and $f = (f_1, \ldots, f_J)^T$ is an unknown probability vector. The goal is to estimate $f$ from $N$ independent random samples of the random variable $Z$. For $i = 1, \ldots, I$ let $y_i$ be the number of times the value $i$ occurs as the value of the random variable $Z$, out of the sample of $N$. The likelihood function for $f$ is then defined by

$$L(f) = \Pi_{i=1}^{I} (g_i)^{y_i}, \tag{4.15}$$

so that the log-likelihood function becomes

$$LL(f) = \sum_{i=1}^{I} y_i \log g_i. \tag{4.16}$$

Vardi and Zhang consider the problem of maximizing $LL(f)$ subject to the constraint that $f$ be a probability vector and that $0 \leq a_j \leq f_j \leq b_j$, for all $j$. To obtain their algorithm we give each $D_i$ the form

$$D_i(x, z) = \sum_{j=1}^{J} p_{ij} KL(x_j, z_j), \tag{4.17}$$

so that $d_{ij}(x_j, z_j) = p_{ij} KL(x_j, z_j)$, for all $i$ and $j$. For each $i$, we now have $C_i = \{x \mid \sum_{j=1}^{J} p_{ij} x_j = y_i\}$. We take $\Omega$ to be the closed convex set

$$\Omega \triangleq \{x \in R^J \mid 0 \leq a_j \leq x_j \leq b_j, \text{ for all } j = 1, \ldots, J, \text{ and } \sum_{j=1}^{J} x_j = 1\}.$$

As shown in [56], the iterative step can be calculated in closed-form using scaling and chopping. When we make $\mathcal{C}$ the set $\mathcal{Q}$ and $\mathcal{H}$ the set $\mathcal{P}$ we get a different algorithm, as we discuss next.

## 4.4 The case in which $\mathcal{P}$ is $\mathcal{H}$ and $\mathcal{Q}$ is $\mathcal{C}$

With $\mathcal{H}$ as $\mathcal{P}$ and $\mathcal{C}$ as $\mathcal{Q}$, each $q \in \mathcal{Q} = \mathcal{C}$ has the form $q = (q_i)$, with $q_i \in C_i \subseteq R^J$ and each $p$ has the form $p = h(x)$, for some $x \in \Omega$. We assume that $\overline{S_i} \cap C_i \neq \emptyset$, for all $1 \leq i \leq I$, and that $d_{ij}(x_j, z_j)$ is extended to $z_j$ on the boundary of $V_{ij}$ as discussed in Appendix A. Denote again $\tilde{P}_i(x) \triangleq \tilde{P}_{C_i}^{f_i}(x)$, the *second variable* projection of $x$ onto $C_i$

with respect to $f_i$, defined, for all $x$ for which there is $z$ in $C_i$ with $D_i(x,z) < +\infty$, as that member of $C_i$ for which the distance $D(x,z)$ is minimized, over all $z \in C_i \bigcap \overline{S}_i$. Now the proximity function becomes

$$\tilde{F}_\Omega(x) \triangleq \sum_{i=1}^{I} D_i(x, \tilde{P}_i(x)), \tag{4.18}$$

defined for all $x$ in

$$\text{dom}\tilde{F}_\Omega \triangleq \Omega \bigcap \Big( \bigcap_{i=1}^{I} \text{dom}\tilde{P}_i \Big).$$

**Algorithm 4.4**
**Initialization:** $x^0 \in \text{dom}\tilde{F}_\Omega$ is arbitrary.
**Iterative Step:** Given $x^k$ find, for all $i = 1, \ldots, I$, the projections $\tilde{P}_i(x^k)$ and calculate $x^{k+1}$ from

$$x^{k+1} = \text{argmin}\,\{\sum_{i=1}^{I} D_i(x, \tilde{P}_i(x^k)) \mid x \in \Omega\} \tag{4.19}$$

One more time, the Algorithm 4.4 is justified along lines similar, but not identical, to those of the previous subsections. In order to minimize the proximity function of (4.18) over all $x \in \text{dom}\tilde{F}_\Omega$, we set now $p^0 = h(x^0)$ to initialize the application of Algorithm 3.1 in the product space. So, given $p^k = h(x^k)$, we first solve the minimization of $D(p^k, q)$ by letting $q^k = (\tilde{P}_1(x^k), \tilde{P}_2(x^k), \ldots, \tilde{P}_I(x^k))$. This expression for the (second-variable) projection of $p^k$ onto $\mathcal{C}$ in the product space $\mathcal{V}$ according to the separable extended Bregman distance $D(p,q)$ defined by (2.2) with the family of separable extended Bregman functions defined in Appendix B. Again, this follows from an argument similar to the one used in the proof of [24, Lemma 4.1] (also appearing in [23, Lemma 5.9.2]). From this $q^k$ we then calculate $p^{k+1}$ by doing the other minimization. This minimization is realized by $p^{k+1} = h(x^{k+1})$ where $x^{k+1}$ is the solution of (4.19). Having made, as before, the additional assumption that, for all $1 \leq i \leq I$, the generalized distances $D_i(x,z)$ are *jointly-convex* with respect to both $x$ and $z$. This implies the joint convexity of $D$ in (2.2), as well as the convexity of $\tilde{F}_\Omega$. From our analysis of extended Bregman functions and distances in Appendices A and B, we know that Assumptions 3.1–3.3, needed for proving convergence in Section 3, hold. We, therefore, conclude that the iterative procedure of Algorithm 4.4 converges to a minimizer of the proximity function $\tilde{F}_\Omega(x)$ whenever it has minimizers.

In the next section we discuss the application of our results to the *split feasibility problem* considered by Censor and Elfving in [24].

# 5 The split feasibility problem

In [24] Censor and Elfving discuss what they call the *split feasibility problem* which is the following. Given closed convex sets $C, Q$ in $R^J$ and an invertible matrix $A$, find $x \in C$ such that $Ax \in Q$. For the consistent case, in which there are such $x$, one can, in principle, use the sequential projection method, projecting orthogonally alternatingly onto the two sets $A(C)$ and $Q$. 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 equivalnet to an oblique projection onto $C$, followed by $A$, see [24, 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

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

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 [24] that any sequence $\{x^k\}$, generated by this algorithm, converges to $x^\infty \in C$, such that $Ax^\infty \in Q$.

We can put this algorithm into the framework discussed above and prove convergence for the inconsistent case. Let $\mathcal{C}$ be the product of $C_1 = A(C)$ and $C_2 = Q$, $\mathcal{D}$ the diagonal subspace of V, as before. Let $f_1(x) = (Ax)^T Ax$ and $f_2(x) = x^T x$, with associated Bregman distances $D_1(x, z) = (1/2)||x - z||^2_{A^T A}$ and $D_2(x, z) = (1/2)||x - z||^2$, where $||x||_G = \langle x, Gx \rangle$. Since these distances are symmetric, the first variable projection and the second variable projection coincide. The iterative algorithm we obtain is that given in (5.1). But now we can conclude that the iterative sequence converges in the inconsistent case to a minimizer of the proximity function $F(x) = D_1(P^{f_1}_{A(C)} x, x) + D_2(P^{f_2}_Q x, x)$.

# 6 The ABSMART and ABEMML algorithms

In this section we consider two iterative algorithms, called the ABSMART and ABEMML algorithms, that can be derived as special cases of the algorithms discussed in Section 4. These algorithms are quite similar to the EMML and SMART algorithms, but incorporate lower and upper bounds $a = (a_j)$ and $b = (b_j)$ on the vector of unknowns $x$ by using the functions $KL(s - a_j, t - a_j)$ and $KL(b_j - s, b_j - t)$ instead of $KL(s, t)$.

## 6.1 The EMML and SMART algorithms

The "Expectation Maximization Maximum Likelihood" (EMML) algorithm, as it is used in emission tomography (see, e.g., Byrne [11, 12, 13], Lange and Carson [49], Tanaka [53], Vardi, Shepp and Kaufman [55]), 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 [51]. The EMML algorithm considered here provides a nonnegative minimizer of the Kullback-Leibler distance as we explain now.

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

$$\text{ent } x \triangleq -\sum_{j=1}^{J} x_j \log x_j \tag{6.1}$$

where "log" denotes the natural logarithms and, by definition, $0 \log 0 = 0$. Its negative, $f(x) \triangleq -$ ent $x$, is a Bregman function and the generalized distance associated with it is the Kullback-Leibler (KL) distance (see Kullback and Leibler [48], see also [23, Example 2.1.2 and Lemma 2.1.3]), given by

$$D_f(x,z) = KL(x,z) = \sum_{j=1}^{J} \Big( x_j \log\Big(\frac{x_j}{z_j}\Big) + z_j - x_j \Big). \tag{6.2}$$

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 $y \in R^I$ and a given nonnegative matrix $A = (a_{ij}) \in R^{I \times J}$ all of whose column-sums are equal to one, consider the distance

$$KL(y, Ax) \triangleq D_f(y, Ax) = \sum_{i=1}^{I} \Big( y_i \log \frac{y_i}{(Ax)_i} + (Ax)_i - y_i \Big). \tag{6.3}$$

Define the sets $C_i$ as

$$C_i \triangleq \{x \in R^J \mid x \geq 0, (Ax)_i = y_i\}, \tag{6.4}$$

and let $w^i_j \triangleq a_{ij}$, for all $1 \leq i \leq I$ and $1 \leq j \leq J$. The functions $g_j$ are defined as

$$g_j(x_j) \triangleq x_j \log x_j, \quad \text{for all} \quad 1 \leq j \leq J. \tag{6.5}$$

The generalized projection $P_i(x)$ of a point $x \in R^J_+$ onto $C_i$, is a member $z$ of $C_i$ which minimizes the distance

$$D_i(z,x) = \sum_{j=1}^{J} w^i_j KL(z_j, x_j). \tag{6.6}$$

It can be verified that, in this case, $P_i$ has the explicit form

$$(P_i(x))_j = x_j \frac{y_i}{(Ax)_i}, \qquad 1 \le j \le J. \tag{6.7}$$

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 (6.6).

It is important to note that if there is an index $j$ for which $x_j = 0$ but $z_j \neq 0$ then $KL(z, x) = +\infty$. When we seek the generalized projection of $x$ onto a closed convex set $C_i$ we must allow for the possibility that the generalized distance from $x$ to each member of $C_i$ is infinite and then we do not define the generalized projection of $x$ onto this set. In our case, however, we see from (6.7) that $(P_i(x))_j = 0$ if and only if $x_j = 0$, so the generalized distance from $x$ to such $C_i$ is always finite and the generalized projection is always defined.

The proximity function $F$ is defined, in this case, as

$$F(x) \triangleq \sum_{i=1}^{I} D_i(P_i(x), x) = \sum_{i=1}^{I}\sum_{j=1}^{J} a_{ij} KL((P_i(x))_j, x_j) \tag{6.8}$$

$$= \sum_{i=1}^{I}\sum_{j=1}^{J} a_{ij} KL\Big(x_j \frac{y_i}{(Ax)_i}, x_j\Big) = KL(y, Ax). \tag{6.9}$$

The iterative step of the EMML algorithm is given by

$$x_j^{k+1} = \sum_{i=1}^{I} w_j^i (P_i(x^k))_j = \sum_{i=1}^{I} a_{ij} \frac{x_j^k y_i}{(Ax^k)_i} = x_j^k \sum_{i=1}^{I} \frac{a_{ij} y_i}{(Ax^k)_i}, \tag{6.10}$$

for all $1 \le j \le J$. The $F(x)$ of (6.9) clearly has nonnegative minimizers and the following result holds (see Iusem [42, 43], Vardi, Shepp and Kaufman [55]).

**Theorem 6.1** *Any sequence $\{x^k\}_{k \ge 0}$, generated by the EMML algorithm, converges to a minimizer of $KL(y, Ax)$.*

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

**Definition 6.1** *We say that a matrix $A = (a_{ij}) \in R^{I \times J}$ 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 [11, Proposition 1].

**Theorem 6.2** *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 $\hat{x} \ge 0$ of $KL(y, Ax)$, $\hat{x}_j > 0$ only if $j \in L$. Consequently, there can be only one such $\hat{x}$.*

We note that, according to this theorem, the minimizer of the proximity function can be on the boundary of the region within which the function $f$ is defined. It is, therefore, necessary to define the proximity function for all boundary points for which the generalized projections are defined.

Turning now to the SMART algorithm, we note that the *second variable* projection $\tilde{P}_i(x)$, of a point $x \in R^J_+$ onto $C_i$, is a member $z$ of $C_i$ which minimizes the distance

$$D_i(x,z) = \sum_{j=1}^{J} w^i_j KL(x_j, z_j). \tag{6.11}$$

It can be verified that, in this case, $\tilde{P}_i$ has the same explicit form as above, that is,

$$(\tilde{P}_i(x))_j = x_j \frac{y_i}{(Ax)_i}, \qquad 1 \le j \le J. \tag{6.12}$$

Again, if $w^i_j = 0$ for some values of $j$ then there will be other members of $C_i$ that also minimize (6.11).

The proximity function $\tilde{F}$ in this case is

$$\tilde{F}(x) \triangleq \sum_{i=1}^{I} D_i(x, \tilde{P}_i(x)) = \sum_{i=1}^{I}\sum_{j=1}^{J} a_{ij} KL(x_j, (\tilde{P}_i x)_j) \tag{6.13}$$

$$= \sum_{i=1}^{I}\sum_{j=1}^{J} a_{ij} KL\Big(x_j, x_j \frac{y_i}{(Ax)_i}\ \Big) = KL(Ax, y). \tag{6.14}$$

The iterative step of the SMART algorithm is given by

$$x^{k+1}_j = \Pi^I_{i=1}((P_i(x^k))_j)^{w^i_j} = x^k_j \exp\Big(\sum_{i=1}^{I} a_{ij} \log \frac{y_i}{(Ax^k)_i}\Big), \tag{6.15}$$

for all $1 \le j \le J$. The $\tilde{F}(x)$ of (6.14) clearly has nonnegative minimizers and we have the following result (see [11]).

**Theorem 6.3** *Any sequence $\{x^k\}_{k \ge 0}$ generated by the SMART algorithm converges to the minimizer of $KL(Ax, y)$ for which $KL(x, x^0)$ is minimized.*

There is no loss of generality in considering here only systems of linear equations $Ax = y$ in which all entries of the matrix are nonnegative. For suppose that $A$ is an arbitrary (real) matrix $A = (a_{ij})$. Rescaling if necessary, we may assume that for each $j$ the column sum $\sum_i a_{ij}$ is nonzero. Now redefine $A$ and $x$ without changing the notation as follows: replace $a_{kj}$ with $\frac{a_{kj}}{\sum_i a_{ij}}$ and $x_j$ with $x_j \sum_i a_{ij}$. This leaves the product $Ax$ unchanged but the new $A$ has all its column sums equal to one. The system $Ax = y$ still holds, but now we know that $y_+ \triangleq \sum_i y_i = \sum_j x_j \triangleq x_+$. Let

$U$ be the matrix whose entries are all one and let $\gamma \geq 0$ be large enough so that $A^{new} = A + \gamma U$ has all nonnegative entries. Then $A^{new}x = Ax + (\gamma x_+)u$, where $u$ is the vector whose entries are all one. So the new system of equations to solve is $A^{new}x = y + (\gamma y_+)u = y^{new}$.

There are also block-iterative versions of the SMART and EMML algorithms, as well as of the ABEMML and ABSMART algorithms, given in the next subsections. These algorithms use only part of the data at each step of the iteration. See, e.g., Byrne [14, 15, 16, 17] and Censor and Segman [29] for further details.

Suppose that, instead of the nonnegativity constraints $x_j \geq 0$, we wish to impose the box constraints $a_j \leq x_j \leq b_j$, for $j = 1, \ldots, J$, for some given $a = (a_j)$ and $b = (b_j)$ which are prior lower and upper bounds on $x = (x_j)$. The ABEMML and ABSMART algorithms presented below converge to a solution of $y = Ax$ with $a \leq x \leq b$ and, in addition, the ABSMART algorithm minimizes the quantity $KL(x - a, x^0 - a) + KL(b - x, b - x^0)$ over these same $x$, provided that $a < x^0 < b$ and that there is a solution of $y = Ax$ with $a \leq x \leq b$. The negative of the quantity $KL(x - a, x^0 - a) + KL(b - x, b - x^0)$ is a generalization of the Fermi-Dirac generalized entropy, which is obtained by taking $a_j = 0$ and $b_j = 1$, for all $j = 1, \ldots, J$.

In both cases considered below we find that calculating Bregman projections onto the sets $C_i = \{x \,|\, y_i = Ax_i\}$ using the distance

$$D_i^{ab}(x, z) = \sum_{j=1}^{J} a_{ij}\Big(KL(x_j - a_j, z_j - a_j) + KL(b_j - x_j, b_j - z_j)\Big) \tag{6.16}$$

cannot be done in closed form, whereas closed-form projections onto the $C_i$ using either the distance

$$D_i^{a}(x, z) = \sum_{j=1}^{J} a_{ij} KL(x_j - a_j, z_j - a_j) \tag{6.17}$$

or the distance

$$D_i^{b}(x, z) = \sum_{j=1}^{J} a_{ij} KL(b_j - x_j, b_j - z_j), \tag{6.18}$$

are possible. We obtain our algorithms by considering duplicates of each of the $C_i$ and letting $D_i = D_i^a$, for $i = 1, \ldots, I$, and $D_i = D_{i-I}^b$, for $i = I + 1, \ldots, 2I$.

## 6.2 The ABSMART algorithm

The ABSMART algorithm is the fully simultaneous version of the ABMART algorithm studied in [16, Section 3]. We assume that $(Aa)_i < y_i < (Ab)_i$, for all $i$.

**Algorithm 6.1 (ABSMART)**
**Initialization:** $x^0 \in R^J$ such that $a_j < x_j^0 < b_j$, for all $j$, is arbitrary.
**Iterative Step:** Given $x^k$ find, for all $j = 1, \ldots, J$, the components of $x^{k+1}$ from

$$x_j^{k+1} = \alpha_j^k b_j + (1 - \alpha_j^k) a_j \tag{6.19}$$

with

$$\alpha_j^k = \frac{c_j^k \prod_{i=1}^I (d_i^k)^{a_{ij}}}{1 + c_j^k \prod_{i=1}^I (d_i^k)^{a_{ij}}}, \tag{6.20}$$

$$c_j^k = \frac{x_j^k - a_j}{b_j - x_j^k}, \tag{6.21}$$

and

$$d_i^k = \frac{(y_i - (Aa)_i)((Ab)_i - (Ax^k)_i)}{((Ab)_i - y_i)((Ax^k)_i - (Aa)_i)}. \tag{6.22}$$

All terms in (6.22) are positive. We see from (6.19) that each term of the iterative sequence $\{x_j^k\}$ is a convex combination of the $a_j$ and $b_j$; the iteration proceeds until convergence to a convex combination for which $y = Ax$, if such exists. If there is no such solution of $y = Ax$ then the algorithm will converge to an approximate solution satisfying the constraints, specifically, the limit is the unique vector satisfying $a \leq x \leq b$ for which the function $KL(Ax - Aa, y - Aa) + KL(Ab - Ax, Ab - y)$ is minimized.

## 6.3 The ABEMML algorithm

The ABEMML algorithm is the fully simultaneous version of the algorithm presented in [16, Section 5]. Here we also assume that $(Aa)_i < y_i < (Ab)_i$, for all $i$.

**Algorithm 6.2 (ABEMML)**
**Initialization:** $x^0 \in R^J$ such that $a_j < x_j^0 < b_j$, for all $j$, is arbitrary.
**Iterative Step:** Given $x^k$ find, for all $j = 1, \ldots, J$, the components of $x^{k+1}$ from

$$x_j^{k+1} = \frac{\alpha_j^k b_j + \beta_j^k a_j}{d_j^k}, \tag{6.23}$$

with

$$\alpha_j^k = (x_j^k - a_j) e_j^k, \tag{6.24}$$

$$\beta_j^k = (b_j - x_j^k) f_j^k, \tag{6.25}$$

$$e_j^k = \sum_i a_{ij} \Big( \frac{y_i - (Aa)_i}{(Ax^k)_i - (Aa)_i} \Big), \tag{6.26}$$

$$f_j^k = \sum_i a_{ij} \Big( \frac{(Ab)_i - y_i}{(Ab)_i - (Ax^k)_i} \Big), \tag{6.27}$$

and

$$d_j^k = \alpha_j^k + \beta_j^k. \tag{6.28}$$

We see from (6.23) that each term of the iterative sequence $\{x_j^k\}$ is a convex combination of the $a_j$ and $b_j$. The iterations proceed until convergence to a convex combination for which $y = Ax$, if such exists. If there is no such solution of $y = Ax$ then the algorithm will converge to an approximate solution satisfying the constraints, specifically, the limit is the unique vector satisfying $a \leq x \leq b$ for which the function $KL(y - Aa, Ax - Aa) + KL(Ab - y, Ab - Ax)$ is minimized.

# 7 Appendix A: Extended scalar Bregman functions and distances

In this appendix we present a class of functions that we call "extended scalar Bregman functions" and use them to construct "extended scalar Bregman distances" between two real numbers. The latter are then used, in Appendix B, to define separable extended Bregman distances between vectors. Some of the results presented here can be deduced also from the work of Kiwiel [47].

Let $V$ be a nonempty open convex subset of the real line $R$, with closure $\overline{V}$ and boundary $\mathrm{bd}V = \overline{V} \setminus V$, those points in $\overline{V}$ but not in $V$. Then, for any $a, b \in R$, there are four cases to consider:

Case 1: $V = \overline{V} = R$ thus $\mathrm{bd}V = \emptyset$;

Case 2: $V = (-\infty, b), \overline{V} = (-\infty, b]$ thus $\mathrm{bd}V = \{b\}$;

Case 3: $V = (a, +\infty), \overline{V} = [a, +\infty)$ thus $\mathrm{bd}V = \{a\}$;

Case 4: $V = (a, b), \overline{V} = [a, b]$ thus $\mathrm{bd}V = \{a, b\}$.

**Definition 7.1** *(Extended Scalar Bregman Distance). Let $V$ be a nonempty open convex subset of $R$ and let $g : \overline{V} \to R$ be a function with the following properties:*

**P1:** *$g$ is continuous on $\overline{V}$;*

**P2:** *$g$ is continuously differentiable at all points of $V$ (that is, $g'$ exists and is continuous on $V$);*

**P3:** *$g$ is strictly convex on $\overline{V}$.*

*Define also, for Cases 2 and 4,*

$$g'(b) = \lim_{t \nearrow b} g'(t) \triangleq \beta \leq +\infty, \tag{7.1}$$

*and, for Cases 3 and 4,*

$$g'(a) = \lim_{t \searrow a} g'(t) \triangleq \alpha \geq -\infty. \tag{7.2}$$

*For $s, t \in \overline{V}$ define an extended scalar Bregman distance with respect to $g$ by*

$$d_g(s,t) \triangleq \begin{cases} g(s) - g(t) - g'(t)(s-t), & \text{if } g'(t) \text{ is finite,} \\ 0, & \text{if } s = t \in \mathrm{bd}V \text{ and } g'(t) \text{ is infinite,} \\ +\infty, & \text{if } s \neq t, t \in \mathrm{bd}V \text{ and } g'(t) \text{ is infinite.} \end{cases} \tag{7.3}$$

*We further define, for any fixed $\theta > 0$, the partial level set*

$$L_2^g(s,\theta) \triangleq \{t \in \overline{V} \mid d_g(s,t) \leq \theta\}, \tag{7.4}$$

*and make the additional assumption*

**P4:** *For all $\theta > 0$ and all $s \in \overline{V}$, the level sets $L_2^g(s,\theta)$ are bounded.*

A function $g$ having all the properties as in Definition 7.1 will be called an *extended scalar Bregman function.* Iusem [44, Proposition 9.2] relates, under certain conditions, the properties (7.1)–(7.2) to a property he calls the "zone coerciveness" of a Bregman function.

**Proposition 7.1** *If $g$ is an extended scalar Bregman function then $g'$ is strictly increasing on $V$, i.e., for $u > t$ in $V$ we have $g'(u) > g'(t)$.*

**Proof:** From the strict convexity property **P3**, we have $g(u) > g(t) + g'(t)(u-t)$ and $g(t) > g(u) + g'(u)(t-u)$. Together these give $(g'(u) - g'(t))(u-t) > 0$. ∎

Since the derivative $g'$ is strictly increasing on $V$ it makes sense to define the (possibly infinite-valued) derivative of $g$ at the boundaries through the one-sided limits.

Let us denote, here and henceforth, the expression in the first row of (7.3) by $\delta_g(s,t)$, i.e., $\delta_g(s,t) \triangleq g(s) - g(t) - g'(t)(s-t)$. Then, from the strict convexity of $g$ we know that, if $g'(t)$ is finite, then $\delta_g(s,t) \geq 0$, for all $s \in \overline{V}$ and all $t \in V$, and $\delta_g(s,t) = 0$ if and only if $s = t$, see, e.g., Bazaraa, Sherali and Shetty [4, Theorem 3.3.3]. The extension in (7.3) preserves this property.

**Proposition 7.2** *If $g$ is an extended scalar Bregman function then, for all $t \in \overline{V}$, we have* $\lim_{u \to t} d_g(t,u) = 0$.

**Proof:** We prove this by showing that $\lim_{u\to t} g'(u)(u-t) = 0$. If $g'(t)$ is finite the result clearly holds. Consider the case in which $t = b$ and $g'(b) = +\infty$. Since the derivative of $g$ is positive for $u$ near $b$ we have $g(u) < g(b)$ and $g'(u) > 0$. Then $g(b) > g(u) + g'(u)(b-u)$, so that $g(b) - g(u) > g'(u)(b-u) > 0$. Since $(g(b) - g(u)) \to 0$, as $u \nearrow b$, we have $g'(u)(b-u) \to 0$, as $u \nearrow b$. The case $t = a$ follows in a similar manner. ∎

From Proposition 7.2, and under the same conditions, the next two corollaries hold.

**Corollary 7.1** *If $g$ is an extended scalar Bregman function and if, for any fixed $s \in \overline{V}$, we let $\eta(t) = d_g(s,t)$ and, for any fixed $t \in \overline{V}$, we let $\xi(s) = d_g(s,t)$, then both $\eta$ and $\xi$ are continuous on $\overline{V}$.*

**Corollary 7.2** *Let $g$ be an extended scalar Bregman function, let $s \in \overline{V}$ be fixed and let $\{t_k\}_{k\geq 0} \subseteq \overline{V}$ be a bounded sequence. If $d_g(s,t_k) \to 0$ then $t_k \to s$, as $k \to +\infty$.*

We present some examples of such functions $g$, taking as $V$ the largest set satisfying the conditions of Definition 7.1. In each of the next five examples $\delta_g(s,t)$ denotes the expression in the first row of (7.3).

**Example 7.1** If $g(t) = t^2$ then we have $\delta_g(s,t) = L^2(s,t) = (s-t)^2$, the square of the Euclidean distance and $V = R$.

**Example 7.2** If $g(t) = t \log t$ then $\delta_g(s,t) = KL(s,t) = s\log(s/t) + t - s$, the scalar Kullback-Leibler distance and $V = (0, +\infty)$.

**Example 7.3** If $g(t) = (t-a)\log(t-a)$ then $\delta_g(s,t) = (s-a)\log((s-a)/(t-a)) + t - s = KL(s-a, t-a)$ and $V = (a, +\infty)$.

**Example 7.4** If $g(t) = (b-t)\log(b-t)$ then $\delta_g(s,t) = (b-s)\log((b-s)/(b-t)) + s - t = KL(b-s, b-t)$ and $V = (-\infty, b)$.

**Example 7.5** If $g(t) = (t-a)\log(t-a) + (b-t)\log(b-t)$ then we have $\delta_g(s,t) = (s-a)\log((s-a)/(t-a)) + (b-s)\log((b-s)/(b-t)) = KL(s-a, t-a) + KL(b-s, b-t)$ and $V = (a, b)$.

For any fixed $\alpha > 0$ we define the other *partial level set* of $d_g(s,t)$ by

$$L_1^g(\alpha, t) \triangleq \{s \in \overline{V} \mid d_g(s,t) \leq \alpha\}. \tag{7.5}$$

**Proposition 7.3** *Ig $g$ is an extended scalar Bregman function then, for any $t \in \overline{V}$ and $\alpha > 0$, the partial level set $L_1^g(\alpha, t)$ is bounded.*

**Proof:** Clearly, if $V$ is bounded there is nothing to prove. Now, if $t$ is on the boundary of $V$ and $d_g(s,t) \leq \alpha$ then $t = s$, so it is obvious that $L_1^g(\alpha, t)$ is bounded in this case. If there are $\alpha > 0$ and $t \in V$ such that $L_1^g(\alpha, t)$ is not bounded, then there is a sequence $\{s_k\}$ whose absolute values $|s_k| \to +\infty$ and such that $g(s_k) - g'(t)s_k \leq \alpha$, for all $k \geq 0$. We must consider two cases: (1) $s_k \to +\infty$ and (2) $s_k \to -\infty$.

We discuss case (1) in detail; since case (2) is similar, we omit it. As $s_k \to +\infty$ there are three possibilities: (1a) $g(s_k) \to +\infty$; (1b) $g(s_k) \to -\infty$; or (1c) $g(s_k) \to r$, for some $r \in R$. In case (1a) if $g'(t) \leq 0$ then we are done. So suppose $g'(t) > 0$. If $g(s_k) - g'(t)s_k \leq \alpha$, for all $k \geq 0$, then for some real $\phi$ we have $g(s_k) \leq \alpha + g'(t)s_k < \phi + g'(t)s_k$, for all $k \geq 0$. Since $g$ is strictly convex, there is $u > t$ with $g(u) > g(t) + g'(t)(u-t)$ and $g'(u) > g'(t)$. The line $l$ tangent to the graph of $g$ at $(u, g(u))$ has slope $g'(u) > g'(t)$, so $l$ intersects the line $y = g'(t)x + \phi$ at some point $x = v$. Then $g(v) > g'(t)v + \phi$, since $(v, g(v))$ is above the line $l$. This contradicts $g(s_k) < \phi + g'(t)s_k$, for all $k \geq 0$, since $g(s)$ is increasing, as $s \to +\infty$. Cases (1b) and (1c) are similar and we omit the details. ∎

So, while the boundedness of $L_1^g(\alpha, t)$ follows from the strict convexity of $g$, Property **P4** of Definition 7.1 does not follow from our other assumptions about $g$ and $d_g(s,t)$. Indeed, we can construct a function $g$ on $V = (0, +\infty)$ with $g(t) \to -\infty$ and $g'(t) \to -\epsilon < 0$, as $t \to +\infty$. Then $d_g(s,t)$ remains bounded as $t \to +\infty$.

**Proposition 7.4** *Let $g$ be an extended scalar Bregman function. If $t < u < s$ then $d_g(s,t) \geq d_g(s,u) + d_g(u,t)$ and $d_g(t,s) \geq d_g(t,u) + d_g(u,s)$.*

**Proof:** Both inequalities readily follow from Definition 7.1 and the fact that the derivative $g'$ is strictly increasing. ∎

From Proposition 7.4, and under the same assumptions, we obtain the next two corollaries.

**Corollary 7.3** *Let $g$ be an extended scalar Bregman function. If $t_k \to t$ and $\{d_g(s_k, t_k)\}$ is bounded, then $\{s_k\}$ is bounded.*

**Proof:** Suppose not. We consider two cases: (1) $t_k \searrow t$ and $s_k \nearrow +\infty$; (2) $t_k \nearrow t$ and $s_k \nearrow +\infty$. For case (1) we have $d_g(s_k, t_k) \geq d_g(s_k, t_1) + d_g(t_1, t_k)$, so that $\{d_g(s_k, t_1)\}$ is bounded; it follows that $\{s_k\}$ is bounded. For case (2) we have $d_g(s_k, t_k) \geq d_g(s_k, t) + d_g(t, t_k)$, so that $\{d_g(s_k, t)\}$ is bounded; it follows that $\{s_k\}$ is bounded. The remaining cases are similar and we omit them. ∎

**Corollary 7.4** *Let $g$ be an extended scalar Bregman function. If $s_k \to s$ and $\{d_g(s_k, t_k)\}$ is bounded, then $\{t_k\}$ is bounded.*

**Proof:** The proof is similar to that of the previous corollary and we omit it. ∎

**Proposition 7.5** *Let $g$ be an extended scalar Bregman Function, let $s_k, t_k \in \overline{V}$, for all $k$, and suppose that $\{t_k\} \to t \in \overline{V}$, $\{s_k\}$ is bounded and $d_g(s_k, t_k) \to 0$. Then $s_k \to t$.*

**Proof:** Without loss of generality we may assume that $s_k \to s$. If $g'(t)$ is finite then the result follows from the strict convexity of $g$ and the continuity of $g'$. So assume that $g'(t)$ is infinite. From $d_g(s_k, t_k) = g(s_k) - g(t_k) - g'(t_k)(s_k - t_k) \to 0$ it follows that $g'(t_k)(s_k - t_k)$ remains finite. But we know that $|g'(t_k)| \to +\infty$, so $(s_k - t_k) \to 0$ and $s = t$ follows. ∎

**Remark 7.1** If $s_k \to s$ and $t_k \to t$ it need not follow that $d_g(s_k, t_k) \to d_g(s, t)$. The implication is true if $g'(t)$ is finite or if $s \neq t$. But, if we let, for example, $g(x) = x \log x$, $t = 0$, $t_k \to 0$ and $s_k = -1/(1 + \log t_k)$, we find that $d_g(s_k, t_k) \to +1$.

# 8 Appendix B: Separable extended Bregman distances and projections

This material has its origin in Bregman's paper [5], Censor and Lent [21], and further developments which appear in the works of Bauschke and Borwein [3], Censor and Zenios [22], Censor, Iusem and Zenios [26], De Pierro and Iusem [34], Eckstein [37], Iusem [41], Teboulle [54] and others.

We use now the extended scalar Bregman functions and distances, presented in Appendix A, as building bricks to construct *extended Bregman functions*, which are not necessarily scalar, in the following natural way. Let $V_{ij} \subseteq R$ be nonempty, open convex subsets of $R$, for all $i = 1, \ldots, I$, $j = 1, \ldots, J$ and let $g_{ij}$ be *extended scalar*

*Bregman functions* with domains $\overline{V}_{ij}$ and with associated *extended scalar Bregman distance* $d_{ij} \stackrel{\Delta}{=} d_{g_{ij}}$, as in Definition 7.1 of Appendix A. In these circumstances, each function

$$f_i(x) \stackrel{\Delta}{=} \sum_{j=1}^{J} g_{ij}(x_j) \tag{8.1}$$

is a *separable extended Bregman function* over the zone $S_i \stackrel{\Delta}{=} \prod_{j=1}^{J} V_{ij} \subseteq R^J$, and the generalized distance associated with each $f_i$ is, for any $(x, z) \in \overline{S_i} \times \overline{S_i}$,

$$D_i(x, z) \stackrel{\Delta}{=} D_{f_i}(x, z) = \sum_{j=1}^{J} d_{ij}(x_j, z_j). \tag{8.2}$$

For every $1 \leq j \leq J$, and $x_j,\ z_j \in \overline{V}_{ij}$, the expression $\delta_{g_{ij}}$ of the first row of Definition 7.1 is

$$\delta_{ij}(x_j, z_j) \stackrel{\Delta}{=} g_{ij}(x_j) - g_{ij}(z_j) - g'_{ij}(z_j)(x_j - z_j). \tag{8.3}$$

Observe that here we construct each $f_i$ from extended scalar Bregman functions over a closed domain (see Definition 7.1 of Appendix A). Therefore, the generalized distances $D_i(x, z)$ are defined for all $(x, z) \in \overline{S}_i \times \overline{S}_i$. This is in contrast to the standard theory of Bregman distances, which are defined only on $\overline{S}_i \times S_i$. See, e.g., [21, 22, 27, 28] or [23, Chapter 2] for the standard theory of Bregman functions and the notions of Bregman function, zone and generalized distance. Note also that, by construction, the zone $S_i$ for the separable case is a "box" in $R^J$. We then have that $\sum_{i=1}^{I} D_i(x, z) = +\infty$ if and only if there are indices $i, j$ such that $x_j \neq z_j$ and $g'_{ij}(z_j)$ is infinite.

Let $C_i$ be a nonempty closed convex subsets of $R^J$, for all $i = 1, \ldots, I$. For any given $z \in \overline{S}_i$ we have either

(1) $D_i(x, z) = +\infty$, for all $x \in C_i \bigcap \overline{S}_i$, or

(2) $D_i(x, z) < +\infty$, for some $x \in C_i \bigcap \overline{S}_i$.

In case (2) there is a unique element of $C_i \bigcap \overline{S}_i$, denoted $P_i(z)$, for which $D_i(x, z) \geq D_i(P_i(z), z)$ for all $x \in C_i \bigcap \overline{S}_i$. Therefore, we define

$$\mathrm{dom} P_i \stackrel{\Delta}{=} \{z \in \overline{S}_i \mid D_i(x, z) < +\infty, \text{ for some } x \in C_i \bigcap \overline{S}_i\}, \tag{8.4}$$

extending the applicability of the Bregman projection operator defined in Section 1. In oreder to propagate the property of case (2) along an iterative process which employs projections we need the following "zone consistency" type assumption about the projections $P_i$:

**Assumption 8.1:** If $z \in \text{dom}P_i$ then $P_i(z) \in \text{dom}P_i$, i.e., $D_i(x, P_i(z)) < +\infty$, for some $x \in C_i \bigcap \overline{S}_i$.

The following inequality plays an important role in our forthcoming analysis.

**Proposition 8.1** *Let $f_i$ be an extended separable Bregman function, as defined above, and let Assumption 8.1 hold. Then, for any $x \in C_i \bigcap \overline{S}_i$ and $z \in \text{dom}P_i$, we have*

$$D_i(x, z) \geq D_i(x, P_i(z)) + D_i(P_i(z), z). \tag{8.5}$$

**Proof:** This was proven originally by Bregman [5] for generalized Bregman distances over $\overline{S}_i \times S_i$; see also, e.g., [23, Theorem 2.4.1]. With the added assumption made above the same proof holds. ∎

The next proposition establishes a continuity result for Bregman projections which is, to the best of our knowledge, the first of its kind in the literature on Bregman functions.

**Proposition 8.2** *Under the assumptions of Proposition 8.1, the function $\pi_j^i$, defined for all $x \in \text{dom}P_i$, by $\pi_j^i(x) \triangleq (P_i(x))_j$ is continuous for all $j$.*

**Proof:** Let $x, x^k \in \text{dom}P_i$, and $x^k \to x$. Then $D_i(P_i(x), x^k) \geq D_i(P_i(x), P_i(x^k))$ follows from the inequality (8.5). We show that $\{D_i(P_i(x), x^k)\}$ is bounded. If not, then there is a $j$ such that $\{d_{ij}((P_i(x))_j, x_j^k)\}$ is unbounded. It follows that $\{g'_{ij}(x_j^k)((P_i(x))_j - x_j^k)\}$ is unbounded; therefore $g'_{ij}(x_j)$ is infinite. So $(P_i(x))_j = x_j$, since $D_i(P_i(x), x)$ is finite. Consequently, $\{d_{ij}(x_j, x_j^k)\}$ is unbounded; this cannot be, since it converges to zero. Therefore, $\{D_i(P_i(x), x^k)\}$ and thus also $\{D_i(P_i(x), P_i(x^k))\}$ are bounded. It follows that $\{P_i(x^k)\}$ is bounded. Let $\{P_i(x^{k_m})\}$ be a subsequence converging to some $c \in C_i$. Then, from Proposition 8.1, we have $D_i(P_i(x), x^{k_m}) \geq D_i(P_i(x), P_i(x^{k_m})) + D_i(P_i(x^{k_m}), x^{k_m})$. The convergence of $\{d_{ij}((P_i(x^{k_m}))_j, x_j^{k_m})\}$ to $d_{ij}(c_j, x_j)$ is clear, for all $j$ for which $g'_{ij}(x_j)$ is finite. Since $\{D_i(P_i(x^{k_m}), x^{k_m})\}$ is a bounded sequence it follows that, for those $j$ for which $g'_{ij}(x_j)$ is infinite, $c_j = x_j = (P_i(x))_j$. Finally, taking $k_m \to +\infty$ in $D_i(P_i(x), x^{k_m}) \geq D_i(P_i(x), P_i(x^{k_m})) + D_i(P_i(x^{k_m}), x^{k_m})$, we get $D_i(P_i(x), x) \geq D_i(P_i(x), c) + \limsup\{D_i(P_i(x^{k_m}), x^{k_m})\} \geq D_i(P_i(x), c) + D_i(c, x)$, from which we conclude that $c_j = (P_i(x))_j$. ∎

**Proposition 8.3** *Under the assumptions of Proposition 8.1, let $x \in \overline{S}_i$. Then $x \in \text{dom}P_i$ if and only if, for some sequence $x^k \in S_i$, $x^k \to x$, the sequence $\{D_i(P_i(x^k), x^k)\}$ is bounded.*

**Proof:** Let $x \in \text{dom}P_i$; then $D_i(P_i(x), x^k) \geq D_i(P_i(x), P_i(x^k)) + D_i(P_i(x^k), x^k)$. Since $\{D_i(P_i(x), x^k)\}$ converges to $D_i(P_i(x), x)$, the sequence $\{D_i(P_i(x^k), x^k)\}$ is bounded. Conversely, if the sequence $\{D_i(P_i(x^k), x^k)\}$ is bounded, then, since $x^k \to x$, we have that $\{P_i(x^k)\}$ is bounded. Select a convergent subsequence $\{P_i(x^{k_m})\} \to c \in C_i$. We show that $D_i(c, x) < +\infty$. Since $\{D_i(P_i(x^k), x^k)\}$ is bounded, it follows that, for all $j$ for which $g_j'(x_j) = +\infty$, we must have $\{(P_i(x^{k_m}) - x^{k_m})\} \to 0$. It follows that $c_j = x_j$ for all such $j$; therefore, $D_i(c, x) < +\infty$. ∎

The classical examples are

**Example 8.1** When $g_j(t) \stackrel{\Delta}{=} g(t) = \frac{1}{2}t^2$ we have that $f(x) = \sum_{j=1}^{J} g_j(x_j) = \frac{1}{2}\|x\|^2$, and, taking $f_i \equiv f$, we get

$$D_i(x, z) = D_f(x, z) = \frac{1}{2}\|x - z\|^2 \stackrel{\Delta}{=} L^2(x, z), \tag{8.6}$$

and $V_j = R$, for all $j = 1, \ldots, J$.

**Example 8.2** For all $j = 1, \ldots, J$, let

$$g_j(t) \stackrel{\Delta}{=} (t - a_j)\log(t - a_j) + (b_j - t)\log(b_j - t)$$

, for $t \in V_j \stackrel{\Delta}{=} (a_j, b_j)$, where $a_j < b_j$ are real numbers . Then

$$f(x) = \sum_{j=1}^{J} g_j(x_j) \sum_{j=1}^{J} (x_j - a_j)\log(x_j - a_j) + (b_j - x_j)\log(b_j - x_j)$$

and, taking $f_i \equiv f$, we get

$$D_i(x, z) = D_f(x, z) = \sum_{j=1}^{J} (x_j - a_j)\log(\frac{x_j - a_j}{z_j - a_j}) + (b_j - x_j)\log(\frac{b_j - x_j}{b_j - z_j}).$$

If $a_j = 0$ and $b_j = +\infty$ then $D_f(x, z)$ becomes the Kullback-Leibler distance (6.2).

Further results on Bregman projections and additional examples can be found, e.g., in Censor and Reich [27], De Pierro and Iusem [34], Eckstein [37], Teboulle [54].

# Acknowledgments

We thank our colleagues Paul Eggermont, Tommy Elfving and Simeon Reich for enlightening discussions on the research presented here. The work of Y. Censor was partially supported by 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] Aharoni, R. and Censor, Y. (1989) Block-iterative projection methods for parallel computation of solutions to convex feasibility problems, *Linear Algebra and Its Applications*, **120**:165–175.

[2] Bauschke, H.H. and Borwein, J.M. (1996) On projection algorithms for solving convex feasibility problems, *SIAM Review*, **38**: 367–426.

[3] Bauschke, H.H. and Borwein, J.M. (1997) Legendre functions and the method of random Bregman projections, *Journal of Convex Analysis*, **4**: 27–67.

[4] Bazaraa, M.S., Sherali, H.D. and Shetty, C.M. (1993) *Nonlinear Programming: Theory and Algorithms*, Second Edition, John Wiley and Sons, New York.

[5] Bregman, L.M. (1967) 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**: 200–217.

[6] Bregman, L.M., Censor, Y. and Reich, S. (1999) Dykstra's algorithm as the nonlinear extension of Bregman's optimization method, *Technical Report*, March, 1999.

[7] Butnariu, D. and Censor, Y. (1990). On the behavior of a block-iterative projection method for solving convex feasibility problems, *International Journal of Computer Mathematics*, **34**: 79–94.

[8] Butnariu, D. and Censor, Y. (1994) Strong convergence of almost simultaneous block-iterative projection methods in Hilbert spaces, *Journal of Computational and Applied Mathematics*, **53**: 33–42.

[9] Butnariu, D., Censor, Y. and Reich, S. (1997) Iterative averaging of entropic projections for solving stochastic convex feasibility, *Computational Optimization and Applications*, **8**: 21–39.

[10] Butnariu, D., Iusem, A.N. and Burachik, R.S. (1997) Iterative methods for solving stochastic convex feasibility problems and applications, *Computational Optimization and Applications*, to appear.

[11] Byrne, C.L. (1993) Iterative image reconstruction algorithms based on cross-entropy minimization, *IEEE Transactions on Image Processing*, **IP-2**: 96–103.

[12] Byrne, C.L. (1995) Erratum and addendum to "Iterative image reconstruction algorithms based on cross-entropy minimization", *IEEE Transactions on Image Processing*, **IP-4**: 225–226.

[13] Byrne, C.L. (1996) Iterative reconstruction algorithms based on cross-entropy minimization, in: *Image Models (and their Speech Model Cousins)*, (S.E. Levinson and L. Shepp, Editors), the IMA Volumes in Mathematics and its Applications, Volume 80, Springer-Verlag, New York, 1–11.

[14] Byrne, C.L. (1996) Block-iterative methods for image reconstruction from projections, *IEEE Transactions on Image Processing*, **IP-5**: 792–794.

[15] Byrne, C.L. (1997) Convergent block-iterative algorithms for image reconstruction from inconsistent data, *IEEE Transactions on Image Processing*, **IP-6**: 1296–1304.

[16] Byrne, C.L. (1997) Iterative algorithms for deblurring and deconvolution with constraints, *Inverse Problems*, **14**: 1455–1467.

[17] Byrne, C.L. (1998) Accelerating the EMML algorithm and related iterative algorithms by rescaled block-iterative (RBI) methods, *IEEE Transactions on Image Processing*, **IP-7**: 100–109.

[18] Byrne, C.L. and Censor, Y. (1998) Proximity function minimization for separable, jointly convex Bregman distances, with applications, *Technical Report*, February 1998.

[19] Censor, Y. (1981) Row-action methods for huge and sparse systems and their applications, *SIAM Review*, **23**: 444–464.

[20] Censor, Y. (1988) Parallel application of block-iterative methods in medical imaging and radiation therapy, *Mathematical Programming*, **42**: 307–325.

[21] Censor, Y. and Lent, A. (1981) An iterative row-action method for interval convex programming, *Journal of Optimization Theory and Applications*, **34**: 321–353.

[22] Censor, Y., and Zenios, S. A. (1992) Proximal minimization algorithm with $D$-functions, *Journal of Optimization Theory and Applications*, **73**: 451–464.

[23] Censor, Y. and Zenios, S.A. (1997) *Parallel Optimization: Theory, Algorithms and Applications*, Oxford University Press, New York.

[24] Censor, Y. and Elfving, T. (1994) A multiprojection algorithm using Bregman projections in a product space, *Numerical Algorithms*, **8**: 221–239.

[25] Censor, Y. and Herman, G.T. (1987) On some optimization techniques in image reconstruction from projections, *Applied Numerical Mathematics*, **3**: 365–391.

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

[27] Censor, Y. and Reich, S. (1996) Iterations of paracontractions and firmly nonexpansive operators with applications to feasibility and optimization, *Optimization*, **37**: 323–339.

[28] Censor, Y. and Reich, S. (1998) The Dykstra algorithm with Bregman projections, *Communications in Applied Analysis*, **2**: 407–419.

[29] Censor, Y. and Segman, J. (1987) On block-iterative entropy maximization, *Journal of Information and Optimization Sciences*, **8**: 275–291.

[30] Combettes, P.L. (1993) The foundations of set theoretic estimation, *Proceedings of The IEEE*, **81**: 182–208.

[31] Combettes, P.L. (1994) Inconsistent signal feasibility: Least-squares solutions in a product space, *IEEE Transactions on Signal Processing*, **SP-42**: 2955–2966.

[32] Csiszár, I. and Tusnády, G. (1984) Information geometry and alternating minimization procedures, *Statistics and Decisions* Supplement Issue No. 1, pp. 205–237.

[33] Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977) Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society, Series B*, **37**: 1–38.

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

[35] Dykstra, R. (1983) An algorithm for restricted least squares regression, *Journal of the American Statistical Society*, **78**: 837–842.

[36] Dykstra, R. (1985) An iterative procedure for obtaining *I*-projections onto the intersection of convex sets, *The Annals of Probability*, **13**: 975–984.

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

[38] Eggermont, P.P.B. and LaRiccia, V.N. (1998) On EM-like algorithms with EM-like properties for smoothed minimum distance estimation, *Technical Report*, Department of Mathematical Sciences, University of Delaware, Newark, DE 19716, USA, March 1998.

[39] Han, S-P. (1988) A successive projection method, *Mathematical Programming*, **40**: 1–14.

[40] Gubin, L.G., Polyak, B.T. and Raik, E.V. (1967) The method of projections for finding the common point of convex sets, *USSR Computational Mathematics and Mathematical Physics*, **7**: 1–24.

[41] Iusem, A.N. (1991) On dual convergence and the rate of primal convergence of Bregman's convex programming method, *SIAM Journal on Optimization*, **1**: 401–423.

[42] Iusem, A.N. (1991) Convergence analysis for a multiplicatively relaxed EM algorithm, *Mathematical Methods in the Applied Sciences*, **14**: 573–593.

[43] Iusem, A.N. (1992) A short convergence proof of the EM algorithm for a specific Poisson model, *Revista Brasileira de Probabilidade e Estatistica*, **6**: 57–67.

[44] Iusem, A.N. (1995) *Proximal point methods in optimization*, Published (in English) under the title: *"Métodos De Ponto Proximal Em Optimização"*, Instituto de Matemática Pura e Aplicada (IMPA), Rio de Janeiro, RJ, Brazil, 20º Colóquio Brasileiro De Matemática, July 24–28, 1995.

[45] Iusem, A.N. and De Pierro, A.R. (1986) Convergence results for an accelerated nonlinear Cimmino algorithm, *Numerische Mathematik*, **49**: 367–378.

[46] Kiwiel, K.C. (1995) Block iterative surrogate projection methods for convex feasibility problems, *Linear Algebra and Its Applications*, **215**: 225–259.

[47] Kiwiel, K.C. (1997) Free–steering relaxation methods for problems with strictly convex costs and linear constraints, *Mathematics of Operations Research*, **22**: 326–349.

[48] Kullback, S. and Leibler, R. (1951) On information and sufficiency, *Annals of Mathematical Statistics*, **22**: 79–86.

[49] Lange, K. and Carson, R. (1984) EM reconstruction algorithms for emission and transmission tomography, *Journal of Computer Assisted Tomography*, **8**: 306–316.

[50] Matúš, F. (1997) On iterated averages of $I$-projections *Technical Report*, March 1997.

[51] McLachlan, G.J. and Krishnan, T. (1997) *The EM Algorithm and Extensions*, John Wiley and Sons, New York.

[52] Pierra, G. (1984) Decomposition through formalization in a product space, *Mathematical Programming*, **28**: 96–115.

[53] Tanaka, E. (1987) A fast reconstruction algorithm for stationary positron emission tomography based on a modified EM algorithm, *IEEE Transactions on Medical Imaging*, **MI-6**: 98–105.

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

[55] Vardi, Y., Shepp, L.A. and Kaufman, L.A. (1985) A statistical model for positron emission tomography, *Journal of the American Statistical Association*, **80**: 8–20.

[56] Vardi, Y. and Zhang, C-H. (1998) Estimating mixing probabilities subject to lower and upper bound constraints with applications, *Technical Report 97-006*, Department of Statistics, Rutgers University, New Brunswick, NJ, USA.

[57] Vardi, Y. and Zhang, C-H. (1999) Reconstruction of binary images from limited projections via the EM algorithm, in: Herman, G.T. and Kuba A. (Editors) (1999) *Discrete Tomography: Foundations, Algorithms, and Applications*, Birkhäuser, Boston, MA, USA, to appear.

[58] Youla, D.C. (1987) Mathematical theory of image restoration by the method of convex projections, in: Stark, H. (Editor) (1987) *Image Recovery: Theory and Applications*, Academic Press, Orlando, FL, USA, pp. 29–78.