NUMERICAL ANALYSIS AND MATHEMATICAL MODELLING
BANACH CENTER PUBLICATIONS, VOLUME 24
PWN-POLISH SCIENTIFIC PUBLISHERS
WARSAW 1990

ON ITERATIVE METHODS FOR LINEARLY CONSTRAINED ENTROPY MAXIMIZATION

YAIR CENSOR

Department of Mathematics and Computer Science, University of Haifa, Haifa, Israel

ALVARO R. DE PIERRO

Instituto de Mathemática, Estatistica e Ciência da Computaçás, Universidade de Campinas, Campinas, Brazil

TOMMY ELFVING

Department of Mathematics, Linköping University, Linköping, Sweden

GABOR T. HERMAN

Department of Radiology, Hospital of the University of Pennsylvania, Philadelphia, Pennsylvania, USA

ALFREDO N. IUSEM

Instituto de Mathemática Pura e Aplicada (IMPA), Rio de Janeiro, Brazil

1. Introduction

The "$x \log x$" *entropy functional*, $\operatorname{ent} x$, maps the nonnegative orthant $\boldsymbol{R}^n_+$ of the n-dimensional Euclidean space $\boldsymbol{R}^n$ into $\boldsymbol{R}$ according to

$$\operatorname{ent} x = -\sum_{j=1}^{n} x_j \log x_j. \tag{1}$$

Here log denotes the natural logarithm and, by definition, $0 \log 0 = 0$.

The *entropy maximization problems* that are addressed are of the form:

(2) Find the $x \in \boldsymbol{R}^n_+$ which maximizes $\operatorname{ent} x$ subject to $x \in Q$.

Q is a set of linear constraints of one of the two following forms:

$$Q_1 = \{x \in \boldsymbol{R}^n \mid Ax = b\}, \tag{3}$$

$$Q_2 = \{x \in \boldsymbol{R}^n \mid Ax \leqslant b\}. \tag{4}$$

Accordingly, (2) is referred to as entropy maximization for *equalities* or for *inequalities*, respectively. In (3) and (4), A is a real $m \times n$ matrix, b is a real m-dimensional vector, and the inequality $Ax \leqslant b$ is interpreted componentwise.

Such linearly constrained entropy maximization problems arise in many fields including transportation planning, statistics, linear numerical analysis, chemistry, geometric programming, and image processing [1], [4], [9], [14], [15], [16], [23], [24], [27], [28]. Their use is rigorously founded in several areas [22], [26], while in other situations entropy maximization is used on an empirical basis. In image reconstruction from projections [4], [7], [12], [17], [18], [19], [20], [25], which is the source of our motivation to study entropy maximization, arguments in favour of maximum entropy imaging have been given, but typically they express mainly the conviction that the maximum entropy approach yields the solution which is most objective or maximally uncommitted with respect to missing information.

The common aspect of the algorithms discussed in this paper is their row-action nature in the sense of [3]. Accordingly, they are all iterative methods in which the matrix A and the vector b are used unchanged during the iterations. In a single iterative step access is required to *only one row* of the constraints system. In addition, only the immediate predecessor of the next iterate is needed at any given step. Algorithms for norm-minimization having such properties were found to perform well on sparse and large systems such as those arising in image reconstruction from projections [4], [5], [12], [17], [18], [19], [20]. Limited experimental experience with some of the algorithms for *entropy optimization* presented below is also available [7], [13], [19], [20], [25], but all these represent preliminary tests and more experimental work is needed to assess the practical value of these algorithms.

This paper is not an overall review of iterative methods for linearly constrained entropy maximization, we rather concentrate on a specific family of algorithms to which our attention and efforts were attracted recently. We report here on theoretical developments and analysis and therefore make no claims regarding advantages of these algorithms in practical applications.

Lamond and Stewart [24] observed that many independently discovered balancing methods used in transportation planning and other fields for solving (2) are in fact special cases of Bregman's method [2]. They noted, however, one exception: the algorithm MART (Multiplicative Algebraic Reconstruction Technique), which was first suggested as a reconstruction technique in [17] and whose convergence was proved in [9] and in [25].

For the special case that A is a zero-one matrix (all entries are equal either to zero or to one), MART coincides with Bregman's method, but the question

of how they relate to each other (if at all) in the general case remained open until now. In this paper we answer this question in two ways. First we demonstrate, in Section 3, that a MART iterative step is a secant approximation to a Bregman iterative step. An interesting point is that, in spite of this observation, MART converges precisely to the maximum entropy solution. Usually, using an approximation in an algorithm causes an inevitable deviation from the conceptual algorithm. But here the crudest approximation to a Bregman step has a "conceptual life" of its own. This relationship, which is easy to verify once it is discovered, motivated us in the construction of an algorithm for entropy maximization over linear inequalities, which preserves the overall structure of Bregman's method but uses in each iteration a "MART step" instead of a "Bregman step". It is also shown in Section 3 that one step of this algorithm can be regarded as one step of an underrelaxed version of Bregman's method. This proof reveals the precise connection between MART and Bregman's method for entropy maximization under linear constraints. In Section 4 we show the convergence to maximum entropy of MART for inequalities. Finally, we present our conclusions in Section 5.

2. Bregman's method and MART for entropy maximization

Let a^i be the transpose of the ith row of A. We use suffixes to denote components of vectors, as in a^i_j and b_i. The inner product in $\boldsymbol{R}^n$ is denoted by $\langle \cdot, \cdot \rangle$.

We assume without further restating and without loss of generality that for all i, $a^i_j \neq 0$ for at least one j.

The next assumption is physically justifiable in image reconstruction from projections and in some other fields, without it some of the theory described below is invalid.

Let i be fixed ($1 \leqslant i \leqslant m$) and let a denote a^i and b denote b_i. Assume that either

$$b > 0 \quad \text{and} \quad 1 \geqslant a_j \geqslant 0, \quad \text{for } i \leqslant j \leqslant n, \tag{5}$$

or

$$b < 0 \quad \text{and} \quad -1 \leqslant a_j \leqslant 0, \quad \text{for } 1 \leqslant j \leqslant n. \tag{6}$$

This assumption also applies to all what follows *without restatement*. Condition (6) is significant only for inequalities, for equalities it is equivalent to condition (5).

When dealing with row-action methods, in the kth iterative step only one row of the system of equalities or inequalities is used; $i(k)$ denotes the index of this row. The sequence $\{i(k)\}$ is the *control* of the algorithm. We say that the control is *almost cyclic* if, for some fixed integer r, $\{1, 2, \ldots, m\} \subset \{i(k), \ldots, i(k+r)\}$ for all k.

We discuss first entropy maximization for equalities. All algorithms discussed here use the same iteration scheme to derive x^{k+1} from x^k.

Iteration scheme for equalities.

$$x_j^{k+1} = x_j^k \exp(c_k a_j^{i(k)}), \quad j = 1, \ldots, n. \tag{7}$$

(Note that if all components of x^k are positive, then the same is true for x^{k+1}).

The method introduced by Bregman [2] can be used to maximize any functional which satisfies a certain set of conditions ([8]). An essential part of the method is that after performing (7) the $i(k)$th constraint has to be satisfied, i.e., that

$$\langle a^{i(k)}, x^{k+1} \rangle = b_{i(k)}. \tag{8}$$

For any x^k such that $x_j^k > 0$ for $j = 1, \ldots, n$ there is a unique choice of x^{k+1}, c_k such that (7) and (8) are simultaneously satisfied, as can be seen from the following where 0 denotes the zero vector.

LEMMA 1. *Let $a \in \mathbf{R}^n - \{0\}$ and $b \in \mathbf{R}$ be such that either* (5) *or* (6) *hold, and let $x \in \mathbf{R}^n$, $x > 0$. Define, for real c,*

$$\phi(c) := \sum_{j=1}^{n} a_j x_j \exp(c a_j) - b. \tag{9}$$

Then there exists a unique c such that $\phi(c) = 0$.

Proof. Immediate, considering monotonicity and continuity of ϕ. ∎

Define

$$\operatorname{sgn} b_{i(k)} := \begin{cases} 1, & \text{if } b_{i(k)} > 0, \\ -1, & \text{if } b_{i(k)} < 0. \end{cases} \tag{10}$$

In MART, c_k in (7) is explicitly defined by

$$c_k := \operatorname{sgn}(b_{i(k)}) \log \frac{b_{i(k)}}{\langle a^{i(k)}, x^k \rangle}. \tag{11}$$

This is well defined as long as $x^{(k)} > 0$ because of (5) and (6).

It has been shown ([8], [25]) that under appropriate conditions both these iterative methods converge to the entropy maximizing element of Q_1. We do not repeat the precise statements here.

If A is a matrix whose elements are either -1 or 0 or 1, then the two iteration schemes produce the same sequence $\{x^k\}$ starting from the same $x^0 > 0$. This is not true for a general A. A computational advantage of the MART sequence is that c_k is given explicitly by (11), rather than by solving a system of $n+1$ nonlinear equations (7), (8).

Underrelaxation has been considered, in somewhat different ways, in both methods. An *underrelaxed Bregman step for equalities* (with relaxation parame-

ter α_k, $0 < \alpha_k \leqslant 1$), chooses c_k so that

$$\langle a^{i(k)},\ x^{k+1}\rangle = \alpha_k b_{i(k)} + (1-\alpha_k)\langle a^{i(k)},\ x^k\rangle \tag{12}$$

is satisfied instead of (8). (That this is well defined follows again from Lemma 1). An *underrelaxed MART step for equalities* (with relaxation parameter λ_k, $0 < \lambda_k \leqslant 1$), chooses c_k by

$$c_k = \lambda_k \operatorname{sgn}(b_{i(k)}) \log \frac{b_{i(k)}}{\langle a^{i(k)},\ x^k\rangle}. \tag{13}$$

Convergence of algorithms based on these iterative steps is discussed in [11] and [25], respectively.

In entropy maximization for *inequalities*, the algorithms use in addition to the sequence $\{x^k\}$ of *primal* iterates a sequence $\{z^k\}$ of m-dimensional *dual* iterates.

Iteration scheme for inequalities.

$$x_j^{k+1} = x_j^k \exp(d_k a_j^{i(k)}), \quad j = 1, \ldots, n, \tag{14}$$

$$z^{k+1} = z^k - d_k e^{i(k)}, \tag{15}$$

with

$$d_k = \min\{z_{i(k)}^k, c_k\}, \tag{16}$$

where e^i denotes the ith column of the identity matrix of order m. This is the overall scheme of Bregman's method for inequalities ([2], [8], [11]). For the underrelaxed Bregman method c_k is chosen here so that (7) and (12) are satisfied simultaneously.

We proposed (without analysis and proof) in [6] to create a "MART for inequalities" algorithm that will use the scheme (14), (15), (16) but with c_k in the closed-form defined in (13).

One of the main results of this paper is to provide a convergence proof for this new algorithm.

Those who wish to implement underrelaxed MART for inequalities, should observe that (13)–(16) completely describe an iterative step of this algorithm. A choice of initial values, control, and relaxation parameters which guarantees convergence to maximum entropy is stated in Theorem 2 below.

3. The relationship between the two iteration methods

We first discuss the *not* relaxed versions of the iteration methods, i.e., formulas (7), (8), (11). MART has an advantage over Bregman's method, because the latter requires finding a root of a function (9) in each iterative step. Here we show that the MART step is a one-step approximation to this root-finding procedure.

Let a, b, and x be as in Lemma 1 and define the function ψ on $\boldsymbol{R}_+$ by

$$\psi(u) := \sum_{j=1}^{n} a_j x_j u^{|a_j|} - b. \tag{17}$$

From (9) we get that, for any positive u,

$$\psi(u) = \phi(\operatorname{sgn} b(\log u)) \tag{18}$$

and

$$\psi(0) = -b \neq 0. \tag{19}$$

Hence, by Lemma 1 and (18), ψ has a unique root, and it is positive.

Now consider the line through $(0, -b)$ and $(1, \psi(1))$, in the plane of the graph of $\psi(u)$. This secant to the graph intersects the u-axis at the point $\tilde{u}$ given by

$$\tilde{u} = \frac{b}{\langle a, x\rangle}. \tag{20}$$

This $\tilde{u}$ is an approximation to the root of ψ and hence, by (18),

$$\tilde{c} = (\operatorname{sgn} b) \log \frac{b}{\langle a, x\rangle} \tag{21}$$

is an approximation to the root of ϕ. Thus, the MART choice for c_k is an approximation by one step with the secant method to the Bregman choice of c_k. The arithmetic cost of computing $\psi(u)$ for values of u other than 0 or 1 is high and so further secant steps, to better approximate the root of ψ, would be much more expensive than the first step. This is particularly important when considering a very large and possibly sparse constraints system as occurs in image reconstruction from projections and some other fields. To better use Bregman's method one would be inclined to further iterate in an inner-loop to obtain a better estimate of the root of (9). MART, for either equalities or, as proposed here, for inequalities, allows for a closed-form formula to replace the root finding calculations. In spite of that formula being recognized as a crude approximation to the root of (9), the algorithm still converges *conceptually* to the desired point. In applications where it is practical to compare the two approaches in actual computation, this should be, of course, done. However, in very large and sparse situations MART is sometimes left competitionless because anything else is impractical.

Convergence properties of MART and Bregman's method for *equalities* have been proved independently of each other ([8], [25]). Convergence of the underrelaxed Bregman's method for *inequalities* was given in [11]. The observations made in this section motivate us to propose MART, rather than Bregman's method, for *inequalities*. The basic idea of the convergence proof we present here, of the resulting algorithm, is to show that an (underrelaxed) MART step coincides with an (underrelaxed) Bregman step for a suitable

choice of the relaxation parameter. However, the previously known results on the convergence of Bregman's method are not strong enough to provide a convergence proof for MART for *inequalities*. Therefore, we need first to strengthen those known results for Bregman's method.

The first step of this program is to show that an underrelaxed MART step is the same as an underrelaxed Bregman step, provided that the relaxation parameter is appropriately chosen.

Define

$$u_k := \frac{b_{i(k)}}{\langle a^{i(k)}, x^k \rangle} \tag{22}$$

and

$$\alpha_k := \begin{cases} \dfrac{1}{b_{i(k)}} \cdot \dfrac{u_k}{u_k - 1} \displaystyle\sum_{j=1}^{n} a_j^{i(k)} x_j^k (u_k^{\lambda_k |a_j^{i(k)}|} - 1), & \text{if } u_k \neq 1, \\ 1, & \text{if } u_k = 1, \end{cases} \tag{23}$$

where, for all k and for some fixed ε,

$$0 < \varepsilon \leqslant \lambda_k \leqslant 1. \tag{24}$$

LEMMA 2. *Assume that $x^k > 0$ and let x^{k+1} be defined by (7) with c_k given by (13) and λ_k satisfying (24). Then x^{k+1} satisfies (12) for α_k defined by (23). Furthermore, this α_k falls in the range $0 < \alpha_k \leqslant 1$.*

Proof. If $u_k = 1$, then $b_{i(k)} = \langle a^{i(k)}, x^k \rangle$, and (12) is satisfied, for any α_k, as long as $x^{k+1} = x^k$. To show that this is the case it is sufficient to show that $c_k = 0$, but that is an immediate consequence of (13).

If $u_k \neq 1$, then starting with the right hand side of (12), we get

$$\begin{aligned} \alpha_k b_{i(k)} + (1 - \alpha_k) \langle a^{i(k)}, x^k \rangle &= \langle a^{i(k)}, x^k \rangle + \sum_{j=1}^{n} a_j^{i(k)} x_j^k (u_k^{\lambda_k |a_j^{i(k)}|} - 1) \\ &= \sum_{j=1}^{n} a_j^{i(k)} \left(x_j^k \exp\left(\log(u_k) \lambda_k \operatorname{sgn}(b_{i(k)}) a_j^{i(k)} \right) \right) \\ &= \langle a^{i(k)}, x^{k+1} \rangle. \end{aligned}$$

Since $x^k > \theta$ and $a^{i(k)} \neq \theta$, (5) and (6) imply that $u_k > 0$. Define

$$\varrho := \varepsilon \times \min_{a_j^i \neq 0} |a_j^i|, \tag{25}$$

where ε is from (24) and the minimum is taken over all entries of A. Thus ϱ does not depend on k and

$$0 < \varrho \leqslant \lambda_k |a_j^{i(k)}| \leqslant 1, \quad \text{for } a_j^{i(k)} \neq 0. \tag{26}$$

(This is the only point in this paper where the fact that $|a_j^i| \leqslant 1$ is used.)

From (26) one can check the following by considering separately the cases $u_k < 1$ and $u_k > 1$. (If $u_k = 1$, then $\alpha_k = 1$ by definition).

$$(27)\quad 0 < \frac{u_k^{\varrho}-1}{u_k-1} = \frac{1}{b_{i(k)}} \cdot \frac{u_k}{u_k-1} \sum_{j=1}^{n} a_j^{i(k)} x_j^k (u_k^{\varrho}-1)$$

$$\leqslant \alpha_k \leqslant \frac{1}{b_{i(k)}} \cdot \frac{u_k}{u_k-1} \sum_{j=1}^{n} a_j^{i(k)} x_j^k (u_k-1) = 1. \blacksquare$$

In both the Bregman and the MART case the iteration scheme for *inequalities* (14)–(16) defines a unique sequence $\{x^k\}$ provided that initial values $x^0 > \theta$ and z^0, the control $\{i(k)\}$, and the sequence of relaxation parameters $\{\alpha_k\}$ or $\{\lambda_k\}$ are specified.

LEMMA 3. *If Bregman's method and MART are specified to have the same initial values (with $x^0 > \theta$) and control, and if the relaxation parameters in MART satisfy* (24) *and those in Bergman's method are chosen based on those in MART using* (23), *then the two algorithms produce the same sequence $\{x^k\}$. Furthermore, at each iterative step, $x^k > \theta$ and the c_k's chosen by the two methods are the same.*

Proof. By induction using Lemma 2. ∎

It may appear that Lemma 3 and proven convergence properties of Bregman's method for *inequalities* should provide a convergence proof of MART for *inequalities*. The difficulty is that existing proofs of Bergman's method for inequalities [11] require $\{\alpha_k\}$ to be bounded below by a positive real number.

The following example shows that the lower bound on α_k in Lemma 2 cannot be replaced by a positive number.

Consider the system

$$(28)\qquad \begin{bmatrix} 0 & 1 \\ 0 & -1 \\ 1 & 1 \\ -1 & -1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \leqslant \begin{bmatrix} 1 \\ -1 \\ 1 \\ -1 \\ 1 \end{bmatrix},$$

whose only feasible point is $x_1 = 0$, $x_2 = 1$. Substituting into (23) with $\lambda_k = 0.5$ for all k we get $\alpha_{5s} = \sqrt{x_1^{5s}}/(1+\sqrt{x_1^{5s}})$. So, if the sequence $\{x^k\}$ converges to a feasible poin at all, then $\alpha_{5s} \to 0$ as $s \to \infty$, and convergence of MART for *inequalities* cannot possibly be established based on existing convergence theorems of Bregman's method.

The following observation is essential to our proof of convergence of underrelaxed MART for *inequalities*.

LEMMA 4. *Let the conditions of Lemma* 3 *be satisfied and let* K *be any set of nonnegative integers. Suppose that* $\{u_k | k \in K\}$ *is bounded, where* u_k *is defined by* (22). *Then there exists an* ε_K *such that*

$$0 < \varepsilon_K \leqslant \alpha_k \leqslant 1, \tag{29}$$

for all k *in* K.

Proof. By Lemma 3, the conditions of Lemma 2 are satisfied. Let ϱ be defined by (25) and define a function β on $\boldsymbol{R}_+$ by

$$\beta(u) := \begin{cases} \dfrac{u^\varrho - 1}{u - 1}, & \text{if } u \neq 1, \\ \varrho, & \text{if } u = 1. \end{cases} \tag{30}$$

β is continuous on $\boldsymbol{R}_+$ and, due to (26), $\beta(u) > 0$. Also, from (23) and (27) we get

$$0 < \beta(u_k) \leqslant \alpha_k \leqslant 1. \tag{31}$$

Let B_K be such that $|u_k| \leqslant B_K$ if $k \in K$. Then, from the properties of β it follows that

$$0 < \min_{0 \leqslant u \leqslant B_K} \beta(u). \tag{32}$$

Finally, set ε_K equal to the right hand side of (32). ∎

4. Convergence of MART for inequalities

MART produces the same sequence of iterates as Bregman's method, provided that the relaxation parameters α_k are chosen in a certain way. Known results regarding Bregman's method ([11]) imply convergence provided that the α_k's are bounded away from zero. Unfortunately, this condition is not necessarily satisfied by the α_k's of Lemma 3. Instead, we have the weaker condition expressed in Lemma 4. In this section we show that for entropy maximization this weaker condition is sufficient for the convergence of Bergman's method and hence the convergence of MART. Bergman functions were defined in [8] as follows.

Let Λ be a subset of $\boldsymbol{R}^n$ and let f: $\Lambda \to \boldsymbol{R}$. Let S be a nonempty convex set such that $\bar{S} \subseteq \Lambda$ (the bar over S denotes closure). Assume that $f(x)$ has continuous first partial derivatives at any $x \in S$ and denote by $\nabla f(x)$ its gradient at x.

From f, construct the function $D: \bar{S} \times S \subseteq \boldsymbol{R}^{2n} \to \boldsymbol{R}$ by

$$D(x, y) := f(x) - f(y) - \langle \nabla f(y), x - y \rangle. \tag{33}$$

Define the *partial level sets* of D, for $\alpha \in \boldsymbol{R}$ by

$$L_1(y, \alpha) := \{x \in \bar{S} \mid D(x, y) \leqslant \alpha\}, \tag{34}$$

$$L_2(x, \alpha) := \{y \in S \mid D(x, y) \leqslant \alpha\}. \tag{35}$$

A function f: $\Lambda \subseteq \boldsymbol{R}^n \to \boldsymbol{R}$ is called a *Bregman function* if there exists a non-empty, open, convex set S (called the *zone* of f), such that $\bar{S} \subseteq \Lambda$ and

(i) $f(x)$ is continuously differentiable at every $x \in S$;

(ii) $f(x)$ is strictly convex on $\bar{S}$;

(iii) $f(x)$ is continuous on $\bar{S}$;

(iv) for every $\alpha \in \boldsymbol{R}$ the partial level sets $L_1(y, \alpha)$ and $L_2(x, \alpha)$ are bounded for every $y \in S$ and for every $x \in \bar{S}$ respectively;

(v) if $y^k \xrightarrow{k \to \infty} y^* \in \bar{S}$, then $D(y^*, y^k) \xrightarrow{k \to \infty} 0$;

(vi) if $D(x^k, y^k) \xrightarrow{k \to \infty} 0$, $y^k \xrightarrow{k \to \infty} y^* \in \bar{S}$ and $\{x^k\}$ is bounded, then $x^k \xrightarrow{k \to \infty} y^*$.

LEMMA 5. *The function* $-\operatorname{ent} x$ *is a Bregman function with* $\Lambda = \boldsymbol{R}^n_+$ *and zone* S_e *defined by*

$$S_e := \{x \mid x > 0\}. \tag{36}$$

Proof. (i) and (ii) are simple. Property (iii) is valid due to the convention $0 \log 0 = 0$. For the function $-\operatorname{ent} x$

$$D(x, y) = \sum_{j=1}^{n} x_j(\log x_j - \log y_j - 1) + \sum_{j=1}^{n} y_j. \tag{37}$$

Fixing y let any component of x go to $+\infty$. Then $D(x, y) \to +\infty$ as well and so, for any $\alpha \in \boldsymbol{R}$, $L_1(y, \alpha)$ is bounded. A similar argument shows that $L_2(x, \alpha)$ is bounded, proving (iv). Property (v) also follows from (37).

Assume now that the premises of property (vi) are satisfied. It is sufficient to show that any convergent subsequence of $\{x^k\}$ converges to y^*. Consider a general term of (37), namely t: $\boldsymbol{R}_+ \times (\boldsymbol{R}_+ - \{0\}) \to \boldsymbol{R}$ defined by

$$t(x, y) := x(\log x - \log y - 1) + y, \tag{38}$$

($x \geqslant 0$ and $y > 0$). For any fixed y

$$t(x, y) \geqslant 0 \tag{39}$$

for all $x \geqslant 0$, and

$$t(x, y) = 0 \quad \text{if and only if} \quad x = y. \tag{40}$$

Now consider a convergent subsequence $\{x^{k_s}\}$ of $\{x^k\}$ and assume that $x^{k_s} \to \bar{x}$.

$$D(x^{k_s}, y^{k_s}) = \sum_{j=1}^{n} t(x_j^{k_s}, y_j^{k_s}) \xrightarrow{s \to \infty} 0. \tag{41}$$

From (39) it follows, that for each j

$$t(x_j^{k_s}, y_j^{k_s}) \xrightarrow{s \to \infty} 0. \tag{42}$$

Noting that $x_j^{k_s} \to \bar{x}_j$ and $y_j^{k_s} \to y_j^*$, consider two cases. If $y_j^* > 0$, then (42) and (40) imply that $y_j^* = \bar{x}_j$. If $y_j^* = 0$, then (42) and (38) imply that $\bar{x}_j = 0$, and again $y_j^* = \bar{x}_j$. Hence $y^* = \bar{x}$. ∎

Let H be a hyperplane in $\boldsymbol{R}^n$ defined by $\langle a, x\rangle = b$ such that $H \cap \bar{S}$ is not empty and let $y \in S$. Define the *Bregman projection of y onto H with respect to f* as a point $x \in H$ which satisfies, for some $\lambda \in \boldsymbol{R}$,

$$\nabla f(x) = \nabla f(y) + \lambda a. \tag{43}$$

It is easily shown that if x exists then it is unique ([2]). In this case λ, which is called the *Bregman parameter associated with the projection of y onto H with respect to f*, is also unique.

LEMMA 6. *Let $x^k \in S_e$ and let x^{k+1} and c_k be defined by* (7) *and* (12) *with $0 \leqslant \alpha_k \leqslant 1$. Then x^{k+1} is the Bregman projection of x^k onto the hyperplane*

$$\langle a^{i(k)}, x\rangle = \alpha_k b_{i(k)} + (1-\alpha_k)\langle a^{i(k)}, x^k\rangle \tag{44}$$

with respect to $-\operatorname{ent} x$ and c_k is the associated Bregman parameter.

Proof. Substituting into (43) $-$ent for f, x^{k+1} for x, x^k for y, c_k for λ and $a^{i(k)}$ for a, we get (7). Substituting x^{k+1} into (44) yields (12). ∎

Let f be a Bregman function with zone S and let H be a hyperplane such that $H \cap \bar{S}$ is not empty. f is said to be *strongly zone consistent* with respect to H if, for every $y \in S$ and for every hyperplane H' which is parallel to H and lies between y and H, the Bregman projection of y onto H' is in S.

LEMMA 7. *The function $-\operatorname{ent} x$ is strongly zone consistent with respect to any hyperplane $\langle a, x\rangle = b$ for which either* (5) *or* (6) *is satisfied.*

Proof. Follows from Lemma 6 noting that in (7) if $x^k \in S$, then $x^{k+1} \in S$. ∎

The next theorem is the main result of [11] which extends to the underrelaxed case the earlier results of [2], [8].

THEOREM 1. *Let f: $\boldsymbol{R}^n \to \boldsymbol{R}$ be a Bregman function with zone S and let*

$$\langle a^i, x\rangle \leqslant b_i, \quad 1 \leqslant i \leqslant m, \tag{45}$$

be any set of inequalities such that

(i) *the set X of elements of $\bar{S}$ satisfying* (45) *is not empty;*

(ii) *f is strongly zone consistent with respect to each of the hyperplanes $H_i = \{x \mid \langle a^i, x\rangle = b_i\}$ $(1 \leqslant i \leqslant m)$.*

Furthermore, let $\{i(k)\}$ be an almost cyclic control and let $\{\alpha_k\}$ be a sequence of relaxation parameters satisfying

$$0 < \bar{\varepsilon} \leqslant \alpha_k \leqslant 1, \tag{46}$$

for all k. Define two sequences $x^k \in \boldsymbol{R}^n$ and $z^k \in \boldsymbol{R}^m$ as follows. $z^0 \in \boldsymbol{R}^m_+$ is arbitrary and x^0 satisfies, for $1 \leqslant j \leqslant n$,

$$[\nabla f(x^0)]_j = -\sum_{i=1}^{n} a_j^i z_i^0. \tag{47}$$

Furthermore, for $k \geq 0$,

$$Vf(x^{k+1}) = Vf(x^k) + d_k a^{i(k)}, \tag{48}$$

$$z^{k+1} = z^k - d_k e^{i(k)}, \tag{49}$$

with

$$d_k = \min\{z^k_{i(k)}, c_k\}, \tag{50}$$

where c_k *is the Bregman parameter associated with the projection of* x^k *onto the hyperplane defined by* (44) *with respect to* f. *Under these circumstances* $\{x^k\}$ *converges to an* x^* *which minimizes* f *over* X.

COROLLARY. *Let* Q_2 *be as in* (4), *such that* $Q_2 \cap \boldsymbol{R}^n_+ \neq \emptyset$. *Let* $\{i(k)\}$ *be an almost cyclic control and* $\{\alpha_k\}$ *be a sequence of relaxation parameters satisfying* (46) *for all* k. *Define sequences* $x^k \in \boldsymbol{R}^n$ *and* $z^k \in \boldsymbol{R}^m$ *as follows.* $z^0 \in \boldsymbol{R}^m_+$ *is arbitrary and, for* $1 \leq j \leq n$,

$$x^0_j = \exp\big((-A^T z^0)_j - 1\big). \tag{51}$$

Furthermore, for $k \geq 0$, *let* (14)–(16) *be satisfied, with* c_k *chosen so that* (7) *and* (12) *are simultaneously satisfied. Under these circumstances* $\{x^k\}$ *converges to the* x^* *which maximizes* $\operatorname{ent} x$ *over* $Q_2 \cap \boldsymbol{R}^n_+$.

Proof. By Lemma 5, $-\operatorname{ent} x$ is a Bregman function with zone S_e defined by (36). $\bar{S}_e = \boldsymbol{R}^n_+$. The assumption $Q_2 \cap \boldsymbol{R}^n_+ \neq \emptyset$ is condition (i) in Theorem 1, while condition (ii) follows from Lemma 7. The conditions on the control and on the sequence of relaxation parameters are the same as in Theorem 1. Similarly, $z^0 \in \boldsymbol{R}^m_+$ is arbitrary in both places and for the function $-\operatorname{ent} x$ (47) and (51) are equivalent. That the c_k's are the same follows from Lemma 6. By induction then the sequences produced in Theorem 1 and here coincide and the desired convergence follows. ∎

This corollary specializes the general result of [11] to the entropy function. We desire a similar result for MART.

THEOREM 2. *Let* Q_2 *be as in* (4) (*with the conditions* (5) *and* (6) *satisfied*), *such that* $Q_2 \cap \boldsymbol{R}^n_+ \neq \emptyset$. *Let* $\{i(k)\}$ *be an almost cyclic control and* $\{\lambda_k\}$ *be a sequence of relaxation parameters satisfying* (24) *for all* k. *Define sequences* $x^k \in \boldsymbol{R}^n$ *and* $z^k \in \boldsymbol{R}^m$ *as follows.* $z^0 \in \boldsymbol{R}^m_+$ *is arbitrary and* x^0 *is defined by* (51). *Furthermore, for* $k \geq 0$, *let* (14)–(16) *be satisfied with* c_k *chosen by* (13). *Under these circumstances* $\{x^k\}$ *converges to the* x^* *which maximizes* $\operatorname{ent} x$ *over* $Q_2 \cap \boldsymbol{R}^n_+$.

Proof. By Lemma 3, the sequences produced by the algorithms in Theorem 2 and Theorem 1 are the same as long as the α_k's of Theorem 1 are chosen based on the λ_k's of Theorem 2 using (23). Unfortunately, such α_k's do not necessarily satisfy (46), and so Theorem 1 is not immediately applicable. Studying the proof of Theorem 1 in [11], we see that (46) is used only twice (in

the proofs of Propositions 4.9 and 4.10). For the case of entropy maximization, the proof of [11] can be refined so that instead of (46) the weaker condition expressed in Lemma 4 is used. The details are laborious and are provided in the Appendix. The so altered Theorem 1, called Theorem 1′ in the Appendix, combined with Lemma 3 proves Theorem 2. ∎

It is possible to construct another version of the MART algorithm by introducing relaxation in a slightly different manner, namely by picking instead of (13)

$$c_k := \operatorname{sgn}(b_{i(k)}) \log \frac{b'_{i(k)}}{\langle a^{i(k)}, x^k \rangle} \tag{52}$$

where

$$b'_{i(k)} = \lambda_k b_{i(k)} + (1 - \lambda_k) \langle a^{i(k)}, x^k \rangle. \tag{53}$$

Convergence of this variant can be proved in a similar way but the details are not repeated here.

5. Conclusions

Bregman's method is quite general, although, as pointed out in [8], there are only few commonly used optimization criteria which are Bregman functions. MART is an algorithm especially designed for entropy maximization in large and sparse systems, which has a similar structure to Bregman's method when applied to entropy. However, it has the computational advantage of using an explicitly defined parameter where Bregman's method requires an inner loop to estimate the so-called Bregman parameter. Convergence of both methods has been independently proved for equality constraints. For inequality constraints, only Bregman's method has been previously proved to converge. Our study of the two methods under a unified framework allows the construction of MART for inequalities and its interpretation as an underrelaxed Bregman's method for inequalities. This in turn allows us to prove convergence of the MART under inequality constraints.

Recently, some work has been done concerning the behavior of row-action methods when applied to inconsistent or infeasible systems (see, e.g., [5], [10], [12], and [21]). It is worthwhile to study similar questions regarding the entropy maximization algorithms presented here.

Acknowledgements. This work was done while Y. Censor, A. De Pierro, and T. Elfving were visiting the Medical Image Processing Group (MIPG), Department of Radiology, Hospital of the University of Pennsylvania, Philadelphia, under support of NIH grant HL-28438. Further progress was made during visits of Y. Censor to the Department of Mathematics at the University of Linköping and the National Defence Research Institute (FOA3) in Linköping,

Sweden, and to the Instituto de Mathemática Pura e Aplicada (IMPA) in Rio de Janeiro, Brazil. We are grateful to Professors Åke Björck, Lindolpho de Carvalho Dias, Kurt Jörnsten, and Torleiv Orhaug for making these visits possible.

We gratefully acknowledge useful discussions and communications with Y.-H. Kuo and A. Lent at the early stages of this research.

We thank Ms. M. A. Blue for wordprocessing the manuscript.

Appendix

As explained in the proof of Theorem 2, an analog of Theorem 1 which while specializing to entropy relaxes condition (46) is required.

THEOREM 1'. *Assume that* $Q_2 \cap \boldsymbol{R}^n_+ \neq \emptyset$, *and let* $\{i(k)\}$ *be an almost cyclic control and let* $\{\alpha_k\}$ *be a sequence of relaxation parameters. Define two sequences* $\{x^k\}$, $\{z^k\}$ *as in Theorem* 1 *with* $f(x) = -\text{ent}\, x$. *If for any subset* K *of the set of nonnegative integers for which* $\{u_k \mid k \in K\}$ *is bounded, there exists an* ε_k *such that* $0 < \varepsilon_k \leqslant \alpha_k \leqslant 1$ *for all* $k \in K$, *then* $\{x^k\}$ *converges to* x^* *which solves* (2) *with* $Q = Q_2$.

The results in the rest of this Appendix establish the proof of this theorem. We prove Propositions 4.9 and 4.10 of [11], without using (46), but using Lemma 4 instead. Because of Lemma 3, all results of [11] which do not make use of (46) are applicable to the algorithm of Theorem 2, with α_k defined by (23). This fact is used repeatedly without further emphasis. We need the following preliminary result.

LEMMA A. *During the execution of the algorithm of Theorem* 2, *the* d_k *of* (16) *are bounded.*

Proof. Let

$$L(x^k, z^k) = f(x^k) + \langle z^k, Ax^k - b\rangle \tag{A1}$$

and

$$e_k := L(x^{k+1}, z^{k+1}) - L(x^k, z^k). \tag{A2}$$

(Note: what we defined as e_k above is denoted by d_k in [11].)

By Proposition 4.6 of [11] $\{L(x^k, z^k)\}$ is bounded above. By Corollary 4.1 of [11], $e_k \geqslant 0$, and so $\{e_k\}$ is bounded.

(A1) and the fact that

$$\nabla f(x^k) = -A^T z^k \tag{A3}$$

(this is Proposition 4.2 of [11]) imply that

$$L(x^k, z^k) = f(x^k) - \langle \nabla f(x^k), x^k\rangle - \langle z^k, b\rangle. \tag{A4}$$

By substituting this into (A2) and using (15) and $f(x) = -\text{ent}\,x$, we get that

$$\left\{\sum_{j=1}^{n}(x_j^k - x_j^{k+1}) + d_k b_{i(k)}\right\} \tag{A5}$$

is bounded. Since $\{x^k\}$ is also bounded (Corollary 4.3 of [11]), the desired result follows. ∎

PROPOSITION 4.9 of [11]. *If* $x^{k_s} \xrightarrow[s\to\infty]{} x^*$, *then* $x^* \in Q_2$. (*Here* $\{x^{k_s}\}$ *denotes an infinite subsequence of* $\{x^k\}$).

Proof. Suppose the result is false. Then there exists p, $1 \leqslant p \leqslant m$, such that

$$\langle a^p, x^*\rangle > b_p. \tag{A6}$$

Since the control is almost cyclic, there exists an integer r such that, for all s, $i(k_s + l_s) = p$ for some $l_s \in \{1, 2, \ldots, r\}$. By Proposition 4.8 of [11], $x^{k_s + l_s} \xrightarrow[s\to\infty]{} x^*$. Furthermore, by Lemma 2, it is possible to choose a subsequence of $\{k_s + l_s\}$, such that the α_k's associated with this sequence converge to an α, $0 \leqslant \alpha \leqslant 1$. For notational convenience, we denote this subsequence by $\{k_s\}$ as well. For this new k_s the following are true:

$$x^{k_s} \xrightarrow[s\to\infty]{} x^*, \tag{A7}$$

$$i(k_s) = p, \tag{A8}$$

$$\alpha_{k_s} \xrightarrow[s\to\infty]{} \alpha, \tag{A9}$$

where $0 \leqslant \alpha \leqslant 1$.

From (A6)–(A8) one gets that, for sufficiently large s,

$$\langle a^{i(k_s)}, x^{k_s}\rangle > b_{i(k_s)}. \tag{A10}$$

By ignoring the beginning of the sequence we may assume that (A10) holds for all s.

From (13) it follows that $c_{k_s} < 0$. Since $z^k \geqslant \theta$ for all k (Proposition 4.3 of [11]), we have from (16) that, for all s,

$$d_{k_s} = c_{k_s}. \tag{A11}$$

Hence, Lemma A together with (13), (22), and (24) implies that $\{u_{k_s}\}$ is bounded. By Lemma 4, this implies that there is an ε_p such that, for all s,

$$0 < \varepsilon_p \leqslant \alpha_{k_s} \leqslant 1. \tag{A12}$$

This, together with (A9) implies that

$$0 < \alpha. \tag{A13}$$

By Lemma 3, the x^{k_s+1} produced by MART is the same as that produced by Bregman's method, and so, from (12),

$$\langle a^{i(k_s)}, x^{k_s+1}\rangle = \alpha_{k_s} b_{i(k_s)} + (1-\alpha_{k_s})\langle a^{i(k_s)}, x^{k_s}\rangle. \tag{A14}$$

Observing (A8) and letting $s \to \infty$ we get (by using again Proposition 4.8 of [11])

$$\langle a^p, x^*\rangle = \alpha b_p + (1-\alpha)\langle a^p, x^*\rangle. \tag{A15}$$

This implies that $\langle a^p, x^*\rangle = b_p$, contradicting (A6). ∎

PROPOSITION 4.10 of [11]. *If* $x^{k_s} \xrightarrow[s\to\infty]{} x^*$, *then, for sufficiently large* s *and for all* $p \in I_1(x^*) := \{i \mid \langle a^i, x^*\rangle < b_i\}$, $z_p^{k_s+r+1} = 0$. (*Here again* r *is the constant in the definition of almost cyclic control.*)

Proof. Let $p \in I_1(x^*)$ and define

$$l_s := \max_{1\le l\le r} \{l \mid i(k_s+l) = p\}. \tag{A16}$$

By Proposition 4.8 in [11], $x^{k_s+l_s} \xrightarrow[s\to\infty]{} x^*$, and $x^{k_s+l_s+1} \xrightarrow[s\to\infty]{} x^*$.

The proof that, for sufficiently large s,

$$d_{k_s} + l_s = z_p^{k_s+l_s} \tag{A17}$$

is divided into two cases depending on whether or not $\langle a^p, x^*\rangle = 0$.

Suppose that

$$\langle a^p, x^*\rangle = \delta \neq 0. \tag{A18}$$

Then, for s sufficiently large,

$$|\langle a^p, x^{k_s+l_s}\rangle| > |\tfrac{1}{2}\delta| \tag{A19}$$

and so, by (22), $\{u_{k_s+l_s}\}$ is bounded. Then, by Lemma 4, there is an ε_p such that $0 < \varepsilon_p \leqslant \alpha_{k_s+l_s}$ holds for sufficiently large s. Now define

$$\varrho_p := \frac{\varepsilon_p}{3} \cdot \frac{b_p - \langle a^p, x^*\rangle}{\|a^p\|}. \tag{A20}$$

By definition of I_1, $\varrho_p > 0$. Hence, for sufficiently large s,

$$\|x^{k_s+l_s} - x^*\| < \varrho_p, \qquad \|x^{k_s+l_s+1} - x^*\| < \varrho_p. \tag{A21}$$

Now suppose that (A17) is false. Then $d_{k_s+l_s} = c_{k_s+l_s}$, and so from (12)

$$\langle a^p, x^{k_s+l_s+1}\rangle = \alpha_{k_s+l_s} b_p + (1-\alpha_{k_s+l_s})\langle a^p, x^{k_s+l_s}\rangle. \tag{A22}$$

It follows, noting (A21), that for sufficiently large s

$$\begin{aligned}\alpha_{k_s+l_s}(b_p - \langle a^p, x^{k_s+l_s}\rangle) &= \langle a^p, x^{k_s+l_s+1} - x^{k_s+l_s}\rangle \\ &\leqslant \|a^p\|\,\|x^{k_s+l_s+1} - x^{k_s+l_s}\| \leqslant 2\|a^p\|\varrho_p.\end{aligned} \tag{A23}$$

This implies that for sufficiently large s

$$\varrho_p \geqslant \frac{\alpha_{k_s+l_s}}{2} \cdot \frac{(b_p - \langle a^p, x^{k_s+l_s}\rangle)}{\|\alpha\|^p}. \tag{A24}$$

Considering the limit of the right hand side of (A24) as $s \to \infty$, and noting that $x^{k_s+l_s} \to x^*$ and that $\alpha_{k_s+l_s} \geqslant \varepsilon_p$ for sufficiently large s, we get from (A20) that $\varrho_p > \varrho_p$. This contradiction shows that (A17) must be true for sufficiently large s if $\langle a^p, x^*\rangle \neq 0$.

Suppose now that

$$\langle a^p, x^*\rangle = 0. \tag{A25}$$

Then $\langle a^p, x^{k_s+l_s}\rangle \xrightarrow[s\to\infty]{} 0$ and so, by (13) and (24), either

$$b_p > 0 \quad \text{and} \quad c_{k_s+l_s} \xrightarrow[s\to\infty]{} +\infty \tag{A26}$$

or

$$b_p < 0 \quad \text{and} \quad c_{k_s+l_s} \xrightarrow[s\to\infty]{} -\infty. \tag{A27}$$

The nonnegativity of z^k (Proposition 4.3 of [11]), (16) and Lemma A show that (A27) cannot happen. Lemma A, (16) and (A26) now imply that for sufficiently large s (A17) is true.

In either case, from (15),

$$z_p^{k_s+l_s+1} = 0. \tag{A28}$$

By the definition of l_s in (A16), the pth component of z is not changed during any step k for which $k_s+l_s+1 \leqslant k \leqslant k_s+r$. This and (A28) show that $z_p^{k_s+r+1} = 0$. ∎

Note that Proposition 4.10 of [11] does not have "for sufficiently large s" in its statement. However this is due to an oversight there and does not make any difference to the rest of the proof.

References

[1] M. Bacharach, *Biproportional Matrices and Input-Output Change*, Cambridge University Press, Cambridge 1970.

[2] L. M. Bregman, *The relaxation method of finding the common points of convex sets and its application to the solution of problems in convex programming*, USSR Comput. Math. and Math. Phys. 7 (1967), 200–217.

[3] Y. Censor, *Row-action methods for huge and sparse systems and their applications*, SIAM Rev. 23 (1981), 444–466.

[4] —, *Finite series expansion reconstruction methods*, Proc. IEEE 71 (1983) 409–419.

[5] Y. Censor, P. P. B. Eggermont and D. Gordon, *Strong underrelaxation in Kaczmarz's method for inconsistent systems*, Numer. Math. 41 (1983), 83–92.

[6] Y. Censor, T. Elfving and G. T. Herman, *Methods for entropy maximization with applications in image processing*, in P. Johansen and P. W. Becker (eds.), *Proc. 3rd Scand. Conf. on Image Analysis*, Chartwell-Bratt Ltd., Lund 1983, 296–300.

[7] Y. Censor, A. V. Lakshminarayanan and A. Lent, *Relaxational methods for large-scale entropy optimization problems with application in image reconstruction*, in P. C. C. Wang *et al.* (eds.), *Information Linkage Between Applied Mathematics and Industry*, Academic Press, New York 1979, 539–546.

[8] Y. Censor and A. Lent, *An iterative row-action method for interval convex programming*, J. Optim. Theory Appl. 34 (1981), 321–353.

[9] J. N. Darroch and D. Ratcliff, *Generalized iterative scaling for log-linear methods*, Ann. Math. Stat. 43 (1972), 1470–1480.

[10] A. R. De Pierro and A. N. Iusem, *A simultaneous projection method for linear inequalities*, Linear Algebra Appl. 64 (1985), 243–253.

[11] —, —, *A relaxed version of Bregman's method for convex programming*, J. Optim. Theory Appl. 51 (1986), 421–440.

[12] P. P. B. Eggermont, G. T. Herman and A. Lent, *Iterative algorithms for large partitioned linear systems with applications to image reconstruction*, Linear Algebra Appl. 40 (1981), 37–67.

[13] T. Elfving, *On some methods for entropy maximization and matrix scaling*, Linear Algebra Appl. 34 (1980), 331–339.

[14] S. Erlander, *Entropy in linear programs*, Math. Programming, 21 (1981), 137–151.

[15] B. R. Frieden, *Statistical models for the image restoration problem*, Comput. Graph. Im. Process. 12 (1980), 40–59.

[16] —, *Probability, Statistical Optics and Data Testing: A Problem Solving Approach*, Springer-Verlag, Berlin 1983.

[17] R. Gordon, R. Bender and G. T. Herman, *Algebraic reconstruction techniques (ART) for three-dimensional electron microscopy and X-ray photography*, J. Theoret. Biol. 29 (1970), 471–481.

[18] G. T. Herman, *Image Reconstruction from Projections: The Fundamentals of Computerized Tomography*, Academic Press, New York 1980.

[19] —, *Mathematical optimization versus practical performance: A case study based on the maximum entropy criterion in image reconstruction*, Math. Programming Stud. 20 (1982), 96–112.

[20] —, *Application of maximum entropy and Bayesian optimization methods to image reconstruction from projections*, in C. R. Smith and W. T. Grandy, Jr. (eds.), *Maximum-Entropy and Bayesian Methods in Inverse Problems*, D. Reidel, Dordrecht 1985, 319–338.

[21] A. N. Iusem and A. R. De Pierro, *On the set of weighted least squares solutions of systems of convex inequalities*, Comment. Math. Univ. Carolinae 25 (1984), 667–678.

[22] E. T. Jaynes, *On the rationale of maximum-entropy methods*, Proc. IEEE 70 (1982), 939–952.

[23] J. N. Kapur, *Twenty-five years of maximum entropy principle*, J. Math. Phys. Sci. 17 (1983), 103–156.

[24] B. Lamond and N. F. Stewart, *Bregman's balancing method*, Transportation Res. Part B 15 (1981), 239–248.

[25] A. Lent, *A convergent algorithm for maximum entropy image restoration with a medical X-ray application*, in R. Show (ed.), *Image Analysis and Evaluation*, Society of Photographic Scientists and Engineers (SPSE), Washington, D. C., 1977, 249–257.

[26] R. D. Levine and M. Tribus (eds.), *The Maximum Entropy Formalism*, The MIT Press, Cambridge, Massachusetts, 1978.

[27] B. Parlett and C. Reinsch, *Balancing a matrix for calculation of eigenvalues and eigenvectors*, Numer. Math. 13 (1969), 296–304.

[28] D. S. Wong, *Maximum likelihood, entropy maximization, and the geometric programming approaches to the calibration of trip distribution models*, Transportation Res. Part B 15 (1981), 329–343.

Presented to the Semester
Numerical Analysis and Mathematical Modelling
February 25 – May 29, 1987