Iterative Methods for Large Scale Convex Optimization

By

Thomas Katsekpor

10014581

A thesis submitted to the Department of Mathematics, University of Ghana, Legon, in partial fulfillment of the requirements for the Degree of Doctor of Philosophy

University of Ghana

Legon

July 2017

Declaration

This thesis was written under the supervision of Professor Alvaro Rodolfo De Pierro, Institute of Mathematical Sciences and Computer Sciences (ICMC), University of São Paulo at São Carlos, Brazil.

I hereby declare that except where due acknowledgment is made, this work has never been presented wholly or in part for an award of a degree in any university.

Student: Thomas Katsekpor.....

Supervisor: Professor Alvaro Rodolfo De Pierro.....

Supervisor: Dr. Margaret Mclntyre.....

Supervisor: Dr. Douglas Adu-Gyamfi.....

Abstract

This thesis presents a detailed description and analysis of Bregman's iterative method for convex programming with linear constraints. Row and block action methods for large scale problems are adopted for convex feasibility problems. This motivates Bregman type methods for optimization.

A new simultaneous version of the Bregman's method for the optimization of Bregman function subject to linear constraints is presented and an extension of the method and its application to solving convex optimization problems is also made.

Closed-form formulae are known for Bregman's method for the particular cases of entropy maximization like Shannon and Burg's entropies. The algorithms such as the Multiplicative Algebraic Reconstruction Technique (MART) and the related methods use closed-form formulae in their iterations. We present a generalization of these closed-form formulae of Bregman's method when the objective function variables are separated and analyze its convergence.

We also analyze the algorithm MART when the problem is inconsistent and give some convergence results.

Acknowledgments

First and foremost, I want to thank the almighty God for his continuous protection and guidance throughout the period of the thesis. I also want to thank my mum for her prayers during the period of the thesis.

I thank Professor Alvaro Rodolfo De-Pierro for introducing me to convex optimization and making me like it. In fact, I am indebted to him for being my advisor and for his support, concern, insight and encouragement. His patience has sustained my research no matter how slow my research would seem to progress.

I am also grateful to Dr. Margaret Mclntyre for creating the needed atmosphere for this work to begin. In fact, I owe her a debt of gratitude for her advice, support and concern and for the urgent manner she has addressed issues concerning the thesis.

I also want to thank Dr. Adu-Gyamfi for his advice and encouragement throughout the period of the thesis.

I am also grateful to Daniel Reem of Technion for the explanations of some key concepts of my thesis and making some important materials available to me for use. I thank him for his concern and support.

The financial support granted by the University of Ghana, Office of Research, Innovation and Development, during my studies is gratefully acknowledged .

Finally, I thank all my colleagues in the department of mathematics who have encouraged and assisted me to get this work done.

Contents

List of Notations

The following is a list of frequently used symbols in the thesis. The symbols are used for the same purpose throughout the thesis. The meanings of those symbols that have multiple meanings will be clear from the context.

$\mathbb{N}$ or $\mathbb{Z}^+$	the set of natural numbers or positive integers.
$\mathbb{R}$	the set of real numbers.
$\mathbb{R}^n$	the n-dimensional Euclidean plane.
$\mathbb{R}^n_+$	the non-negative orthant of the n-dimensional Euclidean plane.
$\mathbb{R}^n_{++}$	the positive orthant of the n-dimensional Euclidean plane.
$\mathcal{R}^n_-$	the extended n-dimensional vector space, adding $-\infty$ coordinates.
$\mathbb{R}^{m\times n}$	the space of all real $m \times n$ matrices.
$\bar{S}$ or cl$S$	the closure of the set S.
bd$S$ or ∂S	the boundary of the set S.
$B(S)$	the family of Bregman functions with zone S.
$\mathcal{C}$	the closure of a convex hull.
A	a real $m \times n$ matrix.
a^i	the ith row of the matrix A.
a^i_j	the entry on ith row and jth column of the matrix A.
$D_f(x, y)$	generalized distance between the vectors x and y.
$\diamondsuit$	denotes the end of definition, algorithm, remark and assumption.

$\square$	denotes the end of proof.
b	a real m-dimensional vector.
b_i	the i element of the vector b.
A^T, x^T	the transpose of the matrix A or the vector x.
$\text{Im}(A^T)$ or $R(A^T)$	the row space of the matrix A.
e^i	the ith standard basis vector in $\mathbb{R}^m$ or $\mathbb{R}^n$.
$\text{Int}S$	the interior of the set S.
$\text{epi}(f)$	the epigraph of the function f.
f^{-1}	the inverse of the function f.
$\|\cdot\|$	the 2-norm.
$P_C(x)$	Bregman projection of the point x onto the closed convex set C.
H_i	the ith hyperplane.
$\langle\cdot,\cdot\rangle$	inner product, i.e., $\langle x, y\rangle = x^T y$.
$P_H(x)$	Bregman projection of the point x onto the hyperplane H.
$\pi_H(x)$	the parameter associated with the projection of x onto the hyperplane H.
$\arg\min_C f(x)$	a minimizer of the function f over the closed convex set C.
$f'(x)$ or $\nabla f(x)$	the derivative of the function f with respect to x.
$\nabla_x f(x, y)$	the derivative of the function f with respect to x.
$X \setminus Y$	the set difference containing all elements in the set X and not in Y.
$\mathbf{1}$	the vector of ones in $\mathbb{R}^n$ or $\mathbb{R}^m$

Chapter 1

Introduction

This thesis is about iterative methods for solving optimization problems with linear constraints, where the objective function to be optimized belongs to a wide family of functions known as Bregman functions.

The main feature of this kind of iterative method is its capability of dealing with the constraints by blocks [21], making them especially suited for solving large-scale problems arising in various fields of applications such as image reconstruction from projections and image restoration [55, 50]. Usually, when the constraints are linear, the matrix describing the constraints is sparse, but all too often, no special structure pattern is detectable in it. In such cases, row or block-action methods are the main option.

One important example is the so called Algebraic Reconstruction Technique (ART) [67] method that computes the projection of the starting point onto the solution set of a linear system of equations. An extension of ART is the Hildreth's quadratic programming algorithm [58] that computes the projection of a given point onto a polyhedron, that is, the solution set of a system of linear inequalities.

Lev Bregman, in a famous paper [14] extended all the previous methods to a large family of functions that are optimized over linear constraints. The main feature of these Bregman's methods is that they essentially consist of a sequence of pro-

jections that generalizes the sequence of the standard orthogonal projections in Euclidean spaces.

In image reconstruction from projections, systems of equations are not only very large but sometimes underdetermined, i.e., when the system has fewer equations than unknowns. In the case of incomplete data [21], they are overdetermined, and the system has more equations than unknowns, and possibly inconsistent [49, 25]. In the underdetermined systems of equations, where we usually have more than one solution, a particular solution is chosen based on some criteria. A common approach, based on some physical considerations, is to choose a maximum entropy solution [65, 68].

For the overdetermined systems of equations, where usually the equations are inconsistent with no solution, algorithms are developed that converge to the weighted least squares solutions of the systems.

The chapters of the thesis are organized in the following manner:

Chapter 2: We study the block Bregman methods, which involve a sequence of Bregman projections onto separating hyperplanes, for solving the convex feasibility problem. Issues of convergence associated with the methods are discussed and addressed. We also extend or generalize the concept of relaxation for Bregman projections, first proposed in [41] and further extended to general convex set in [29].

Chapter 3: This chapter deals with the optimization of Bregman functions subject to linear inequality constraints using the simultaneous method.

Chapter 4: We develop a general closed-form formula for the iterative step in Bregman's algorithm for linearly constrained convex optimization of any Bregman function whose variables are separated. That is, we replace the computational burden involved in an inner loop calculation of the projection parameter by a closed form formula. We derive specific closed-form formulae for Burg's and Shannon's entropies and compare the results with the existing ones in the literature. General underrelaxed Bregman's algorithm for linear inequality constraints

is also proposed.

Chapter 5: We analyze the behaviour of MART algorithm. Here, all problems are assumed to be inconsistent. Strongly underrelaxed parameters, with some specific conditions, are therefore incorporated into the methods to enable them converge to the desired solutions.

Chapter 6: This chapter contains the conclusion and suggestions for future work.

In this introductory chapter, we describe not only the basic and the detailed information needed to understand the remaining chapters of the thesis, but also give its historical framework. The results in this chapter are not new.

In the following two sections, we describe the optimization problem to be solved, the optimality conditions and some basic results on convex duality theory. Sections 1.4 and 1.5 present the convex feasibility problem and the detailed description of the available iterative algorithms in the literature for solving it. The general and the popular one, in its sequential form, is known as Projections onto Convex Sets (POCS) and its particular cases for linear systems of equations are ART and Cimmino.

Section 1.6 is dedicated to the definitions of Bregman measures and the corresponding generalized projections. The generalization of the Pythagoras theorem for Bregman measures and the main properties of Bregman projections onto hyperplanes or half-spaces are also presented in this section.

Section 1.7 describes the known versions of the sequential Bregman's method for linear equality constraints as well as its relaxed version introduced in [41].

Section 1.7 also describes the sequential Bregman's method for linear inequality constraints and its particular case for the quadratic optimization problem known as the Hildreth's method.

When using Bregman's method to maximize entropy functions with linear constraints, it is possible to obtain simpler closed form formulae for the iterations.

The Multiplicative Algebraic Reconstruction Techniques (MART) and the Simultaneous MART (SMART) are two of these well known methods for Shannon's entropy maximization [23] that are generated as relaxed Bregman's methods using appropriate relaxation parameters. This is described in Section 1.8.

1.1 The convex optimization problem

We are concerned with iterative methods for solving the Convex Optimization Problem (COP) defined by

$$\begin{aligned} &\text{minimize} \quad f(x) \\ &\text{subject to} \quad x \in C = \bigcap_{i=1}^{m} C_i, \end{aligned} \tag{1.1}$$

where $f : D \subseteq \mathbb{R}^n \to \mathbb{R}$ is a real valued convex function and C is the constraint set, nonempty, closed and convex. In most practical situations, C is the intersection of other closed convex sets C_i; that is, $C = \cap_{i=1}^{m} C_i$, as in (1.1). The function f is called the *objective function* or *cost function*.

Usually the set C is specified by a set of inequality constraints $f_i : \mathbb{R}^n \to \mathbb{R}$ for $i = 1, \ldots, m$, or equality constraints $h_j : \mathbb{R}^n \to \mathbb{R}$ for $j = 1, \ldots, p$, or, a combination of both, i.e.,

$$C := \{x \in \mathbb{R}^n \mid f_i(x) \leq 0, \ \forall i = 1, \ldots, m \text{ and } h_j(x) = 0, \ \forall j = 1, \ldots, p\}. \tag{1.2}$$

In this case, the set of points for which the objective function and all the constraint functions are defined

$$\mathcal{D} = D \cap \bigcap_{i=1}^{m} \text{dom} f_i \cap \bigcap_{j=1}^{p} \text{dom} h_j$$

is called the domain of the optimization problem (1.1). A point $x \in \mathbb{R}^n$ is called feasible if it satisfies the constraints $f_i(x) \leq 0$, $i = 1, \ldots, m$, and $h_j(x) = 0$, $j = 1, \ldots, p$. The problem (1.1) is said to be feasible if there exists at least one feasible point, and infeasible otherwise. The set of all feasible points is called the feasible set or the constraint set C.

The problem is convex if f and the f_i's are convex and the h_i's are affine. It is referred to as a nonlinear convex optimization problem, if in addition, f is either nonlinear or the f_i's are nonlinear. It is a linear programming problem if f is linear and C is a polyhedron.

In this work, we assume that all of the functions are continuously differentiable on C and the optimization problems we discuss and analyze consist of linear and nonlinear constraints. In this regard, we restate problem (1.1) with the convex set C decomposed into linear and nonlinear constraints as stated on page 373 of [11] as follows.

$$\begin{aligned} &\text{minimize} && f(x) \\ &\text{subject to} && x \in X,\ f_i(x) \leq 0,\ i = 1, \ldots, \bar{m}, \\ & && \langle c_j, x \rangle = d_j,\ j = 1, \ldots, p,\ \langle a_i, x \rangle \leq b_i,\ i = \bar{m}+1, \ldots, m, \end{aligned} \tag{1.3}$$

where X is a nonempty subset of $\mathbb{R}^n$, $f_i : \mathbb{R}^n \to \mathbb{R}$ are nonlinear functions, $c_j,\ a_i$ are nonzero vectors in $\mathbb{R}^n$ and $d_j,\ b_i$ are real numbers.

In the next section, we present the general material on convex optimization problem which will be used in the sequel.

1.2 Optimality conditions

Here, we first look at the conditions for the existence of optimal solution as described in [11] and then examine some basic results of convex analysis needed for the solution of the convex optimization problem stated in Section 1.1.

1.2.1 Existence of an optimal solution

For the function $f : D \subseteq \mathbb{R}^n \to \mathbb{R}$, i.e., the set $\{f(x) \in \mathbb{R} \mid x \in D\}$, there are two possibilities:

(i) The set $\{f(x) \mid x \in D\}$ is bounded below. In this case, $\inf_{x \in D} f(x) \in \mathbb{R}$.

(ii) The set $\{f(x) \mid x \in D\}$ is unbounded below. In this case, $\inf_{x \in D} f(x) = -\infty$.

Existence of at least one global minimum is guaranteed if f is a continuous function and D is a nonempty compact subset of $\mathbb{R}^n$. This is the *Weierstrass theorem*. We consider the following definitions.

Definition 1.2.1. Let $f : D \subseteq \mathbb{R}^n \to \mathbb{R}$ be a function. Then

(i) $x^* \in D$ is a *minimum* of f over D if $f(x^*) = \inf_{x \in D} f(x)$. We call x^* a minimizing point or a *minimizer* or a *minimum* of f over D. We denote this by $x^* = \arg\min_{x \in D} f(x)$.

(ii) A subset D of $\mathbb{R}^n$ is called *convex* if $\alpha x + (1-\alpha)y \in D, \forall x, y \in D, \forall \alpha \in [0,1]$.

(iii) Let D be a convex subset of $\mathbb{R}^n$. A function $f : D \to \mathbb{R}$ is called *convex* if

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha) f(y), \ \forall x, y \in D, \ \forall \alpha \in [0,1]$$

and *strictly convex* if

$$f(\alpha x + (1-\alpha)y) < \alpha f(x) + (1-\alpha) f(y), \ \forall x, y \in D, \ x \neq y, \ \forall \alpha \in (0,1).$$

(iv) A vector x is said to be a *relative interior point* of the nonempty convex set D if $x \in D$ and there exists an open sphere S centered at x such that $S \cap \text{aff}D \subset D$, i.e., x is an interior point of D relative to aff$D$, where aff$D$ is the notation for the *affine hull* of D, defined as the intersection of all affine sets containing D. The set of all relative interior points of D is called the *relative interior* of D.

♢

Proposition 1.2.2. *Let $f : D \subseteq \mathbb{R}^n \to \mathbb{R}$ be a convex function over the convex set D.*

(i) A local minimum of f over D is also a global minimum over D. If in addition f is strictly convex, then there exists at most one global minimum of f.

(ii) If f is convex and the set D is open then $\nabla f(x^*) = 0$ is a necessary and sufficient condition for a vector $x^* \in D$ to be a global minimum of f over D.

The proof of this proposition can be found on page 14 of [10].

1.2.2 Least-squares problems

A least-squares problem is problem (1.1) with the objective function f given by

$$f(x) = \|Ax - b\|^2 = \sum_{i=1}^{m} (\langle a^i, x\rangle - b_i)^2,$$

and with no constraints. $A \in \mathbb{R}^{m\times n}$ with $m \geq n$, a^i is the ith row of the matrix A and $x \in \mathbb{R}^n$ is the optimization variable.

For overdetermined linear system of equations $Ax = b$, $A \in \mathbb{R}^{m\times n}$ with $m > n$ and $b \in \mathbb{R}^m$, we cannot solve for x for most b. Therefore, we find $x = x_{ls}^*$ that minimizes $\|Ax-b\|^2$ and x_{ls}^* is the least-squares (approximate) solution of $Ax = b$.

It must be noted that if the objective function of an optimization problem is quadratic and the associated quadratic form is positive semidefinite then it is a least-squares problem. While the basic least-squares problem has a simple fixed form, several standard techniques are used to increase its flexibility in applications. In a weighted least-squares, the weighted least-squares cost

$$\sum_{i=1}^{m} w_i(\langle a^i, x\rangle - b_i)^2,$$

where the w_i's are positive is minimized. The w_i's are the weights and are chosen to reflect differing levels of concern about the sizes of the terms $\langle a^i, x\rangle - b_i$, or simply to influence the solution.

1.3 Lagrangian duality

In this section, we define the Lagrange dual function and the duality gap and then state the Karush-Kuhn-Tucker conditions for the optimization problem (1.1).

1.3.1 The Lagrange dual function

The Lagrangian $L : D \times \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$ associated with problem (1.1) with the convex set C specified by (1.2) is given by

$$L(x, \lambda, \nu) = f(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{j=1}^{p} \nu_j h_j(x) \tag{1.4}$$

where λ_i and ν_j are the Lagrange multipliers associated with the ith inequality constraint $f_i(x) \leq 0$ and the jth equality constraint $h_j(x) = 0$ respectively; the vectors $(\lambda, \nu) \in \mathbb{R}^m \times \mathbb{R}^p$ are the dual variables or the Lagrange multiplier vectors. $f_i(x)$ and $h_j(x)$ are assumed to be convex for all i and j.
We define the (Lagrange) dual function as

$$g(\lambda, \nu) = \inf_x L(x, \lambda, \nu).$$

When L is unbounded below in x, we have $g(\lambda, \nu) = -\infty$. Since the dual function is the pointwise infimum of a family of affine functions of (λ, ν), it is always concave, even if problem (1.1) is not convex.

If the optimal value of problem (1.1) is p^* then for any $\lambda_i \geq 0$ for all i and any ν, we have

$$g(\lambda, \nu) \leq p^*. \tag{1.5}$$

This inequality is justified in Subsection 1.3.3 under complementary slackness. The (Lagrange) dual problem associated with problem (1.1) is given by

$$\begin{array}{ll} \text{maximize} & g(\lambda, \nu) \\ \text{subject to} & \lambda_i \geq 0 \text{ for } i = 1, \ldots, m. \end{array} \tag{1.6}$$

In this context, the original problem (1.1) is sometimes called the *primal problem*.

1.3.2 Weak duality

Suppose the optimal value of the dual problem is d^*. Then, by (1.5), it is the best lower bound on p^* that can be obtained from the Lagrange dual function. In particular, we have

$$d^* \leq p^*.$$

This is the weak duality and it holds even when d^* and p^* are infinite. For example, if the primal problem is unbounded below, so that $p^* = -\infty$, then we have $d^* = -\infty$, i.e., the Lagrange dual problem is infeasible. Conversely, if the dual problem is unbounded above, so that $d^* = \infty$, then we must have $p^* = \infty$, i.e., the primal problem is infeasible. The difference $p^* - d^*$ is the optimal duality gap

of the original problem. We have strong duality if $p^* = d^*$. It was demonstrated on page 226 of [12] that the strong duality holds for convex problems where the feasible set has nonempty interior, i.e., the Slater's condition holds.

The following result, the strong duality theorem for linear and nonlinear constraints stated and proved on page 373 of [11], shows that, under suitable convexity assumptions and under a constraint qualification, there is no duality gap between the primal and the dual optimal objective function values. We repeat this theorem and Assumption 6.4.3 therein as Assumption 1.3.1 for the purpose of easy reference.

Assumption 1.3.1. (**Linear and nonlinear constraints**) *The optimal value p^* of problem (1.3) is finite, and the following hold:*

(i) The set X is the intersection of a polyhedral set and a convex set D.

(ii) The functions $f : \mathbb{R}^n \to \mathbb{R}$ and $f_i : \mathbb{R}^n \to \mathbb{R}$ are convex over D.

(iii) There exists a feasible vector $\bar{x}$ such that $f_i(\bar{x}) < 0$ for all $i = 1, \ldots, \bar{m}$, i.e., the Slater's condition holds.

(iv) There exists a vector that satisfies the linear constraints (but not necessarily the constraints $f_i(x) \leq 0$, $i = 1, \ldots, \bar{m}$), and belongs to X and to the relative interior of D. ♢

Theorem 1.3.2. Strong duality theorem for linear and nonlinear constraints *Let Assumption 1.3.1 hold for problem (1.3). Then there is no duality gap and there exists at least one geometric multiplier.*

Note: A vector (λ^*, ν^*) is said to be a geometric multiplier for problem (1.3) if $\lambda^* \geq 0$ and $p^* = \inf_{x \in X} L(x, \lambda^*, \nu^*)$.

1.3.3 Complementary slackness

Suppose that the primal and the dual optimal values are attained and equal (so, in particular, strong duality holds). Let x^* be a primal optimal and (λ^*, ν^*) be a

dual optimal point. This means that

$$\begin{aligned} f(x^*) &= g(\lambda^*, \nu^*), \\ &= \inf_x \left(f(x) + \sum_{i=1}^m \lambda_i^* f_i(x) + \sum_{j=1}^p \nu_j^* h_j(x) \right), \text{ by the definition of } g \\ &\leq f(x^*) + \sum_{i=1}^m \lambda_i^* f_i(x^*) + \sum_{j=1}^p \nu_j^* h_j(x^*) \\ &\leq f(x^*). \end{aligned} \tag{1.7}$$

The last inequality follows from $\lambda_i^* \geq 0$ and $f_i(x^*) \leq 0$ for $i = 1, \ldots, m$, and $h_j(x^*) = 0$ for $j = 1, \ldots, p$. We conclude that the two inequalities in this chain hold with equality. This means that $\sum_{i=1}^m \lambda_i^* f_i(x^*) = 0$ and since each term in this sum is nonpositive, we have

$$\lambda_i^* f_i(x^*) = 0 \text{ for } i = 1, \ldots, m.$$

This condition is known as complementary slackness; it holds for any primal optimal x^* and any dual optimal (λ^*, ν^*) (when strong duality holds). We can express the complementary slackness condition as

$$\lambda_i^* > 0 \Rightarrow f_i(x^*) = 0$$

or equivalently,

$$f_i(x^*) < 0 \Rightarrow \lambda_i^* = 0.$$

1.3.4 KKT optimality conditions

Suppose the functions $f_0, \ldots, f_m$, $h_1, \ldots, h_p$ are differentiable and therefore have open domains. Let x^* and (λ^*, ν^*) be any primal and dual optimal points with zero duality gap. Now since x^* minimizes $L(x, \lambda^*, \nu^*)$ over x, $\nabla L(x^*, \lambda^*, \nu^*) = 0$. Therefore we have

$$\begin{aligned} \nabla f(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla f_i(x^*) + \sum_{j=1}^{p} \nu_j^* \nabla h_j(x^*) &= 0, \\ \lambda_i^* f_i(x^*) &= 0, \text{ for } i = 1, \ldots, m, \\ \lambda_i^* &\geq 0, \text{ for } i = 1, \ldots, m, \\ f_i(x^*) &\leq 0, \text{ for } i = 1, \ldots, m, \\ h_j(x^*) &= 0, \text{ for } j = 1, \ldots, p, \end{aligned} \tag{1.8}$$

which are called the Karush-Kuhn-Tucker (KKT) conditions. Thus, for any optimization problem with differentiable objective and constraint functions for which strong duality holds, any pair of primal and dual optimal points must satisfy the KKT conditions (1.8). When the primal problem is convex, the KKT conditions are also sufficient for the points to be primal and dual optimal [[12], page 244]. In other words, if f_i are convex and h_j are affine, and $\bar{x}, \bar{\lambda}, \bar{\nu}$ are any points that satisfy the KKT conditions then $\bar{x}, \bar{\lambda}, \bar{\nu}$ are primal and dual optimal, with zero duality gap.

1.4 The Convex Feasibility Problem (CFP)

When there is no objective function to be minimized, the problem of convex optimization (1.1) reduces to just finding a point in the intersection of the closed convex sets (1.9) and this is called the convex feasibility problem. That is, we find

$$x \in C = \bigcap_{i=1}^{m} C_i. \tag{1.9}$$

A common feature of all the algorithms used to solve (1.9) and (1.1) in the thesis is their row-action nature in the sense of [21]. These algorithms obey a specific control sequence. In this section therefore, we would want to define a control sequence of an algorithm and to describe two different types of this sequence used in the thesis, and then give the definition of a row action method.

Definition 1.4.1. A control sequence *A control sequence* $\{i(k)\}$ is a sequence

of indices according to which individual sets C_i or blocks that are groups of the sets C_i in the convex feasibility problem (1.9) may be chosen for the execution of an iterative algorithm.

(i) *Cyclic control*: A control sequence $\{i(k)\}$ is *cyclic* if $i(k) = k \bmod m + 1$, where m is the total number of sets in problem (1.9).

(ii) *Almost cyclic control*: A control sequence $\{i(k)\}$ is *almost cyclic* on $I := \{1, 2, \ldots, m\}$ if $i(k) \in I$ for all $k \geq 0$, and there exists a fixed integer $r \geq m$ (called *almost cyclicality constant*) such that, for all $k \geq 0$, $I \subset \{i(k), \ldots, i(k + r)\}$. ♢

Definition 1.4.2. A *row-action method* is an iterative procedure which requires, in each iterative step, only the current iterate and one row of the matrix or a group of rows, and performs no transformation on the matrix elements. ♢

The matrix mentioned in the last definition is either the matrix $A \in \mathbb{R}^{m\times n}$ if the constraints in (1.9) are linear or the Jacobian matrix of first partial derivatives if the constraints are nonlinear. Row-action method is frequently used in areas where the matrix describing the constraints is huge and sparse as observed in the field of image reconstruction from projections.

A row-action iteration has the functional form

$$x^{k+1} = \mathcal{P}^r{}_i(x^k, C_i), \tag{1.10}$$

where k is the iteration index, and $i = i(k)$ for $1 \leq i \leq m$ is the control index, specifying the row that is acted upon by the algorithmic operator $\mathcal{P}^r{}_i$. The algorithmic operator generates, in some specified manner, the new iterate x^{k+1} from the current iterate x^k and from information contained in C_i for $1 \leq i \leq m$. $\mathcal{P}^r{}_i$ may depend on additional parameters that vary from iteration to iteration, such as relaxation parameters, weights, etc.

The constraints in (1.9) may be decomposed into M groups of constraints called *blocks* by choosing a sequence of integers $\{m_t\}_{t=0}^{M}$ such that

$$0 = m_0 < m_1 < \ldots < m_{M-1} < m_M = m$$

and defining for each t, $1 \leq t \leq M$, the subset

$$I_t = \{m_{t-1} + 1, m_{t-1} + 2, \ldots, m_t\}.$$

This yields a partition of the set:

$$I = \{1, 2, \ldots, m\} = I_1 \cup I_2 \cup \cdots \cup I_M. \tag{1.11}$$

A *block-action iteration* then has the functional form

$$x^{k+1} = \mathcal{P}^b{}_t(x^k, \{C_i\}_{i \in I_t}), \tag{1.12}$$

where $t = t(k)$ is the control index, $1 \leq t \leq M$, specifying the block that is used when the algorithmic operator $\mathcal{P}^b{}_i$ generates x^{k+1} from x^k and from information contained in all constraints in (1.9) whose indices belong to I_t. $\mathcal{P}^b{}_i$ may also depend on additional parameters that vary from iteration to iteration.

The iterative methods we consider may therefore be classified as having one of the following four basic structures.

(i) *Sequential algorithms.* For this class of algorithms we define a control sequence $\{i(k)\}$ and the algorithm performs, in a strictly sequential manner, row-action iterations according to (1.10), from an appropriate initial point until a stopping rule is applied.

(ii) *Simultaneous algorithms.* Algorithms in this class first execute simultaneously row-action iterations on all rows or constraints

$$x^{k+1,i} = \mathcal{P}^r{}_i(x^k, C_i), \ i = 1, 2, \ldots, m$$

and the next iterate x^{k+1} is a convex combination of the iterates $x^{k+1,i}$.

(iii) *Sequential block-iterative algorithms.* Here, the constraints in (1.9) are decomposed into fixed blocks in a form (1.11), and a control sequence $\{t(k)\}$ over the set $\{1, 2, \ldots, M\}$ is defined. The algorithm performs sequentially, according to the control sequence, block iterations of the form (1.12).

(iv) *Simultaneous block-iterative algorithms.* In this case, block iterations are first performed using the same current iterate x^k, on all blocks simultaneously

$$x^{k+1,t} = \mathcal{P}^b{}_t(x^k, \{C_i\}_{i\in I_t}),\ t = 1, 2, \ldots, M.$$

The next iterate x^{k+1} is then the convex combination of the iterates $x^{k+1,t}$.

1.5 Projections onto Convex Sets (POCS)

When we are dealing with a problem with a huge number of constraint sets C_i for $i = 1, \ldots, m$, it is important to develop methods that deal with a few constraints at a time. The most popular of these methods for the CFP consists of orthogonal projection in a sequential manner onto the convex sets, that is, given an initial point $x^0 \in \mathbb{R}^n$, the algorithm is defined by the sequence

$$x^{k+1} = P_{C_{i(k)}}(x^k) \tag{1.13}$$

where $P_{C_{i(k)}}$ denotes the orthogonal projection onto the convex set $C_{i(k)}$ and $\{i(k)\}$ is a control sequence, usually cyclic along the set of integers $I := \{1, 2, \ldots, m\}$. The control sequence $\{i(k)\}$ can also be 'almost cyclic'.

Figure 1.1 illustrates the case of feasibility of the CFP where orthogonal projections are used in a sequential manner onto the intersection of two convex sets. Figure 1.2 illustrates the case of infeasibility of the CFP where orthogonal projections are used in a sequential manner onto two non-intersecting convex sets. In this case, the sequence generated may not converge to the projection of the starting point.

It is also possible to define a relaxed version of POCS method as

$$x^{k+1} = x^k + \alpha_k(P_{C_{i(k)}}(x^k) - x^k) \tag{1.14}$$

where the α_k's are positive relaxation parameters. It can be proven (see [54]) that the above sequence converges if the constraint set is nonempty, control sequence is cyclic and $\alpha_k \in (0, 2)$.

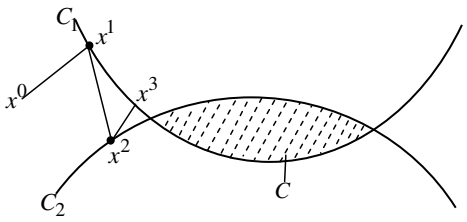

Figure 1.1: Feasibility: Orthogonal projection in a sequential manner onto two convex sets.

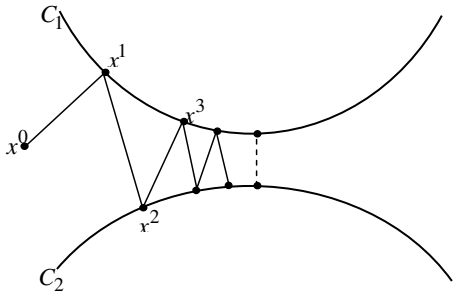

Figure 1.2: Infeasibility: Orthogonal projection in a sequential manner onto two convex sets.

Several of the methods we discuss in this thesis employ a sequence $\{\alpha_k\}$ of relaxation parameters. Loosely speaking, these parameters overdo or underdo the move prescribed in an iterative step. Relaxation parameters add an extra degree of freedom to the way a method might actually be implemented and have important consequences on the performance of the method in practice. The effects of relaxation parameters on the iterations and the technical conditions on them that guarantee convergence of an algorithm are discussed for each method, especially strongly underrelaxed methods in Chapter 5.

It should be noted that an iterative algorithm is said to be underrelaxed if the relaxation parameters incorporated are confined to the interval $(0, 1)$. It is however said to be overrelaxed if the relaxation parameters lie in the interval $(1, 2)$. In most convergence analysis, the relaxation parameters remain in the interval $[\epsilon, 2-\epsilon]$, for an arbitrarily small $\epsilon > 0$, in order to guarantee convergence, see, e.g Aharoni and Censor [5].

A simultaneous version of POCS method (SPOCS), suitable for parallel implementation, can also be defined as

$$x^{k+1} = x^k + \alpha_k \sum_{i=1}^{m} \lambda_i (P_{C_i}(x^k) - x^k) \tag{1.15}$$

where the λ_i's are positive real numbers such that $\sum_{i=1}^{m} \lambda_i = 1$. For this version, more general convergence results can be proven, including convergence to a least

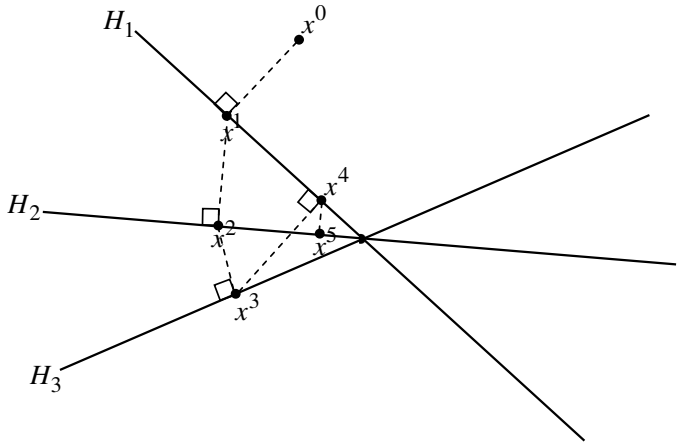

Figure 1.3: The iterative step of the algorithm ART with three equations and unity relaxation parameter.

squares solution in the infeasible case (see [40] for details) for system of linear equations.

1.5.1 ART and Cimino

When the problem is linear and the constraints are equalities, POCS method and its simultaneous version SPOCS method give rise to two well known algorithms in the literature: ART [67] and Cimmino [34]. Figure 1.3 illustrates the algorithm for ART with a system of three equations as in (1.16) with unity relaxation parameter. ART is very popular in image and signal processing [53], especially in Computerized Tomography [60, 55]. In this case, for a given matrix $A \in \mathbb{R}^{m\times n}$ with rows $a^i \neq 0$ for $i = 1, \ldots, m$ and $b \in \mathbb{R}^m$, problem (1.9) reduces to finding a solution of the linear system of equalities

$$Ax = b \tag{1.16}$$

and (1.13) becomes the algorithm ART or Kaczmarz's algorithm given by

$$x^{k+1} = x^k + \frac{b_i - \langle a^i, x^k\rangle}{\|a^i\|^2} a^i, \tag{1.17}$$

and from (1.14), its relaxed version is

$$x^{k+1} = x^k + \alpha_k \frac{b_i - \langle a^i, x^k\rangle}{\|a^i\|^2} a^i, \tag{1.18}$$

and with a cyclic control sequence along the set of integers $I := \{1, 2, \ldots, m\}$. $\langle \cdot, \cdot \rangle$ stands for the standard inner product in $\mathbb{R}^n$ and $\| \cdot \|$ for the Euclidean norm.

Similarly, from (1.15), the simultaneous version known as Cimmino's algorithm, is given by

$$x^{k+1} = x^k + \alpha_k \sum_{i=1}^{m} \lambda_i \frac{b_i - \langle a^i, x^k \rangle}{\|a^i\|^2} a^i. \tag{1.19}$$

As noted earlier, Cimmino's algorithm converges to a least squares solution in the case of infeasibility. This behaves like a gradient method applied to a square objective function [10, 11, 40].

If the equalities (1.16) are replaced with the inequalities

$$Ax \leq b, \tag{1.20}$$

then we have the relaxed method of Agmon, Motzkin, and Schoenberg (AMS) (see [1]) and its relaxed Cimmino's method respectively as

$$x^{k+1} = x^k + \alpha_k c_i^k a^i, \tag{1.21}$$

and

$$x^{k+1} = x^k + \alpha_k \sum_{i=1}^{m} \lambda_i c_i^k a^i, \tag{1.22}$$

where

$$c_i^k = \min \left\{ 0, \frac{b_i - \langle a^i, x^k \rangle}{\|a^i\|^2} \right\}$$

and with a cyclic control sequence.

Figure 1.4 illustrates the iterative step in Cimmino's algorithm using a system of three inequalities with unity relaxation parameter.

It is worth noting that, in general, a sequence of orthogonal projections onto closed convex sets does not converge to the orthogonal projection of the initial point onto the intersection. A simple counterexample is shown next.

Consider the half-spaces $C_1 = \{x \in \mathbb{R}^2 \mid x_2 \leq 0\}$ and $C_2 = \{x \in \mathbb{R}^2 \mid x_1 + x_2 \leq 0\}$ and let $C = C_1 \cap C_2$. If $x_0 = (2, 1)$ is the initial point of projection then, by simple calculation, the projection of x_0 onto C, i.e., the minimizer of the function $\frac{1}{2}\|x - x_0\|^2$ over C is $(\frac{1}{2}, -\frac{1}{2})$, i.e., $(\frac{1}{2}, -\frac{1}{2}) = \frac{1}{2} \arg\min\{\|x - x_0\|^2 \mid x \in C\}$. But

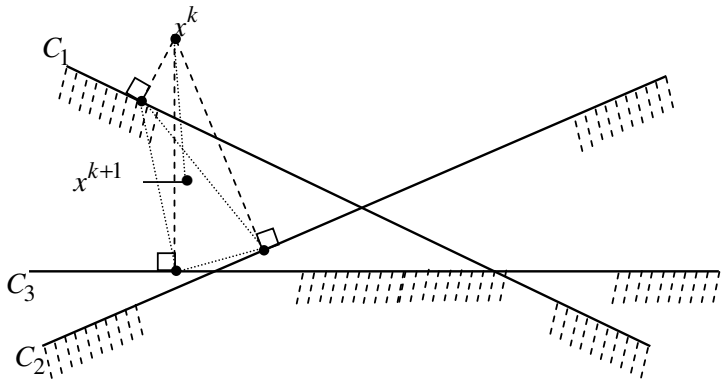

Figure 1.4: The iterative step of the simultaneous Cimmino's algorithm with three inequalities and unity relaxation parameter.

using Algorithm 1.13, with initial projection of x_0 onto C_1, the algorithm does not converge to $(\frac{1}{2}, -\frac{1}{2})$.

This counterexample motivates the development of Bregman measures and methods [14], where the objective function defines the measure to be used for the projections, and a clever duality scheme allows the preservation of information about the starting point and the function to be minimized. This is the subject of the next section.

1.6 Bregman measures and generalized projections

Let S be a nonempty, open, convex set, such that $\bar{S} \subseteq D$ with the function $f : D \subseteq \mathbb{R}^n \to \mathbb{R}$. Then S is called the zone of f and is sometimes defined as $S := \text{Int}(\text{dom} f)$.

Definition 1.6.1. Bregman Functions A function $f : D \subseteq \mathbb{R}^n \to \mathbb{R}$ with zone S is called a *Bregman function* if there exists a nonempty, open, convex set S, such that $\bar{S} \subseteq D$ and the following conditions hold:

(i) $f(x)$ has continuous first partial derivatives at every $x \in S$,

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

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

(iv) for every $\alpha \in \mathbb{R}$, the partial level sets $L_1^f(y,\alpha)$ and $L_2^f(x,\alpha)$ are bounded for every $y \in S$ and for every $x \in \bar{S}$, respectively, where $L_1^f(y,\alpha) := \left\{x \in \bar{S} \mid D_f(x,y) \leq \alpha\right\}$ and $L_2^f(x,\alpha) := \{y \in S \mid D_f(x,y) \leq \alpha\}$.

(v) If $y^k \in S$ for all $k \geq 0$ and $\lim_{k\to\infty} y^k = y^*$ then $\lim_{k\to\infty} D_f(y^*, y^k) = 0$,

(vi) If $y^k \in S$ and $x^k \in \bar{S}$ for all $k \geq 0$ and if $\lim_{k\to\infty} D_f(x^k, y^k) = 0$ and $\lim_{k\to\infty} y^k = y^*$, and $\left\{x^k\right\}$ is bounded, then $\lim_{k\to\infty} x^k = y^*$.

We denote the collection of all Bregman functions with zone S by $B(S)$. ♢

The function D_f, where $D_f : \bar{S} \times S \subseteq \mathbb{R}^{2n} \to \mathbb{R}$, is constructed from $f(x)$ in Definition 1.6.1 by

$$D_f(x,y) := f(x) - f(y) - \langle \nabla f(y), x - y\rangle. \tag{1.23}$$

This function is called the generalized distance function or the D-function.

$D_f(x,y)$ may be interpreted as the difference $f(x) - h(x)$, where $h(z)$ represents the hyperplane H given by

$$H = \{(z, h(z)) \mid h(z) = f(y) + \langle \nabla f(y), z - y\rangle\} \tag{1.24}$$

which is tangent to the epigraph of f at the point $(y, f(y))$ in $\mathbb{R}^{n+1}$.
Figure 1.5 shows the geometric interpretation of the generalized distance $D_f(x,y)$.

Remark 1.6.2. It is worth noting that the D-function is not necessarily symmetric, i.e., $D_f(x,y) \neq D_f(y,x)$, and in general, does not satisfy the triangle inequality. Therefore we will prefer to call D_f a Bregman measure instead of a distance or generalized distance function. However, the phrases 'Bregman distance' and 'Bregman divergence' are very common in the literature to denote D_f.
♢

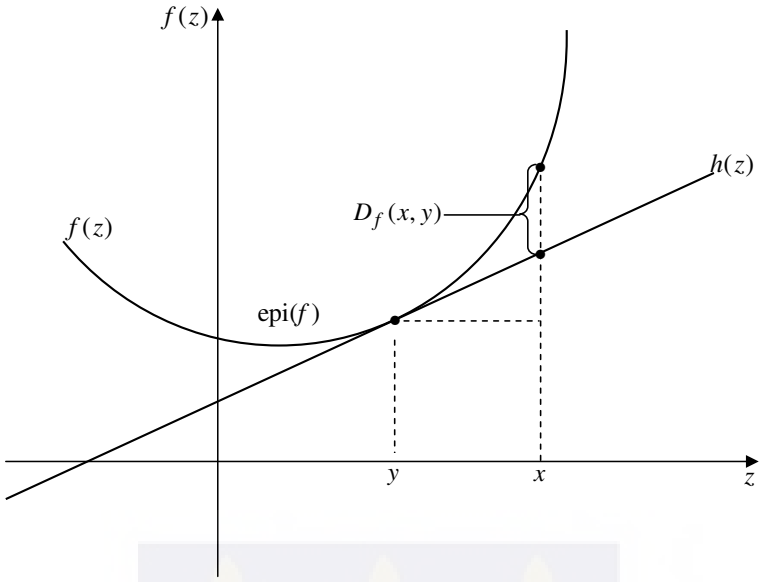

Figure 1.5: Geometric interpretation of the generalized distance.

Two important examples of Bregman measures are those defined by the functions

$$f(x) = \frac{1}{2}\|x\|^2 \text{ and } f(x) = -\sum_{j=1}^{n}(x_j \log x_j)$$

with zones $\mathbb{R}^n$ and $\mathbb{R}^n_+$ respectively.

The first one gives rise to the standard 2-norm,

$$D_f(x, y) = \frac{1}{2}\|x - y\|^2$$

which is symmetric. The second one, the Shannon entropy [68, 79], gives rise to the Kullback-Leibler information divergence [69],

$$D_f(x, y) = \sum_{j=1}^{n}\left(x_j \log \frac{x_j}{y_j} + y_j - x_j\right)$$

which is clearly not a distance function and not a metric.

Definition 1.6.3. Generalized projections Given a closed and nonempty convex set $C \subseteq \mathbb{R}^n$, $f \in B(S)$, and $y \in S$, a point $x^* \in C \cap \bar{S}$ for which

$$\min_{z \in C \cap \bar{S}} D_f(z, y) = D_f(x^*, y)$$

is denoted by $P_C(y)$ and is called a *generalized projection* or *Bregman projection* or simply a *projection* of a point y onto the set C. $\diamondsuit$

The next lemma shows the nonnegativity associated with the Bregman measure which is justified in the literature. This lemma is followed by Lemma 1.6.5 which guarantees the existence and uniqueness of generalized projections. Its proof can be found on page 32 of [32].

Lemma 1.6.4. *For every $f \in B(S)$, we have $D_f(x, y) \geq 0$ for all $x \in \bar{S}$ and for all $y \in S$, and $D_f(x, y) = 0$ if and only if $y = x$.*

Proof. This is an immediate consequence of the strict convexity of f and the definition of $D_f(x, y)$. □

Lemma 1.6.5. *If $f \in B(S)$, then for any closed convex set $C \subseteq \mathbb{R}^n$, such that $C \cap \bar{S} \neq \emptyset$, and for any $y \in S$, there exists a unique generalized projection $x^* = P_C(y)$.*

An important geometrical property about the Bregman measure and its projection, which is the basis for all the main convergence proofs, is briefly described next. Details of this description will be given in the next chapter.

1.6.1 A generalized Pythagoras theorem

First, we will prove the following.

Lemma 1.6.6. *If $f \in B(S)$ and D_f is a Bregman measure then for a given point $x \in S$ and its Bregman projection $P_C(x)$ onto a closed convex set C, the function defined by*

$$G(z) = D_f(z, x) - D_f(z, P_C(x))$$

is convex (as a matter of fact linear) for $z \in \bar{S} \cap C$.

Proof. Using the definition of D_f, we have

$$\begin{aligned} G(z) &= D_f(z, x) - D_f(z, P_C(x)) \\ &= f(z) - f(x) - \langle \nabla f(x), z - x \rangle - f(z) + f(P_C(x)) \\ &\quad + \langle \nabla f(P_C(x)), z - P_C(x) \rangle \\ &= f(P_C(x)) - f(x) - \langle \nabla f(x), z - x \rangle + \langle \nabla f(P_C(x)), z - P_C(x) \rangle \end{aligned}$$

which is linear in z. □

The most important property that underlines the application of Bregman projections in Bregman's method or algorithm for solving convex optimization problem is presented next. The property describes the geometry behind Bregman's method and makes it possible for the method to converge.

Theorem 1.6.7. *For a given $f \in B(S)$ and a closed convex set C, define $P_C(x)$ as the Bregman projection of x onto C. Then the following inequality holds for every $y \in C \cap \bar{S}$.*

$$D_f(P_C x, x) \leq D_f(y, x) - D_f(y, P_C x) \tag{1.25}$$

Proof. The proof of this theorem could be found in [14] as Lemma 1 on page 201. But because of its importance in the sequel, we will present the proof in detail in Chapter 2. □

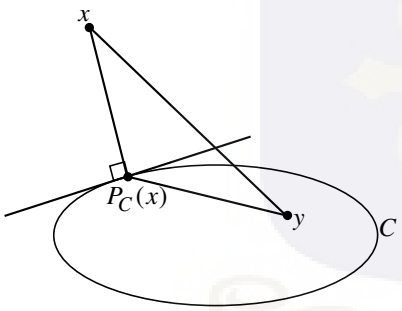

Figure 1.6: Geometric description of Theorem 1.6.7 when orthogonal projections are used.

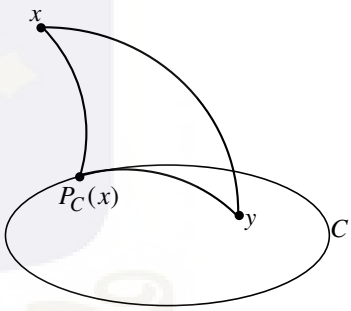

Figure 1.7: Geometric description of Theorem 1.6.7 when generalized projections are used.

Remark 1.6.8. It is worth nothing that when the function f is the half-squared 2-norm and the set C is a hyperplane, the inequality (2.3) becomes equality and the result becomes the old and well known Pythagoras theorem. A deeper analysis of this fact, together with the proof, will appear in Chapter 2. ♢

1.6.2 Generalized projections onto hyperplanes

A key role in the iterative projection methods for linear feasibility problems and for linearly constrained optimization problems is played by generalized projections onto hyperplanes. A hyperplane in $\mathbb{R}^n$ is a set of the form

$$H = \{x \in \mathbb{R}^n \mid \langle a, x\rangle = b\},$$

where $a \in \mathbb{R}^n$, $a \neq 0$, and $b \in \mathbb{R}$ are given. We will need the following definition.

Definition 1.6.9. Zone consistency

(i) A function $f \in B(S)$ is said to be *zone consistent* with respect to the convex set C if for every $y \in S$ we have $P_C(y) \in S$. That is, the generalized projection $P_C(y)$ of any point $y \in S$ onto the convex set C remains in S.

(ii) $f \in B(S)$ is said to be *strongly zone consistent* with respect to the hyperplane H and the point $y \in S$, if it is zone consistent with respect to H, and with respect to every other hyperplane H' which is parallel to H and lies between y and H. ♢

Two examples of strongly zone consistent Bregman functions with respect to the hyperplane $H = \{x \in \mathbb{R}^n \mid \langle a, x\rangle = b\}$ are the functions $f : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^n_+ \to \mathbb{R}$ given by

$$f(x) = \frac{1}{2}\|x\|^2 \text{ and } g(x) = -\sum_{j=1}^{n}(x_j \log x_j)$$

with zones $\mathbb{R}^n$ and $\mathbb{R}^n_+$ respectively

The following lemma characterizes generalized projections onto hyperplanes and its proof can be found on page 35 of [32].

Lemma 1.6.10. *Let $f \in B(S)$, $H = \{x \in \mathbb{R}^n \mid \langle a, x\rangle = b\}$ with $a \in \mathbb{R}^n \setminus \{0\}$, $b \in \mathbb{R}$, and assume that f is zone consistent with respect to H. Then, for any given $y \in S$, the system*

$$\nabla f(z) = \nabla f(y) + \lambda a, \tag{1.26}$$

$$\langle a, z\rangle = b, \tag{1.27}$$

determines uniquely the point z, which is the generalized projection of y onto H. For a fixed representation of H, i.e., for a fixed $a \in \mathbb{R}^n \setminus \{0\}$ and $b \in \mathbb{R}$, the system also determines uniquely the real number λ. In this case, the vector z and λ are the unknowns.

For a fixed representation of the hyperplane H, λ obtained from the system (1.26)-(1.27) is called the generalized projection parameter associated with the generalized projection of y onto H. We denote this by $\pi_H(y)$, i.e., $\lambda = \pi_H(y)$.

The next two results, Lemmas 1.6.11 and 1.6.12 whose proofs can be found on pages 36 and 37 of [32], give statements about the signs of the parameters associated with the generalized projections onto hyperplanes and the relationship between two such parameters if a point is projected onto two parallel hyperplanes.

Lemma 1.6.11. *Let $H = \{x \in \mathbb{R}^n \mid \langle a, x\rangle = b\}$ and $y \in S$. For any $f \in B(S)$ which is zone consistent with respect to H, the parameter λ associated with the generalized projection of y onto some particular representation of H, i.e., $a \neq 0$ and $b \in \mathbb{R}$ given, satisfies*

$$\lambda(b - \langle a, y\rangle) > 0, \quad \text{if} \quad y \notin H, \tag{1.28}$$

$$\lambda = 0, \quad \text{if} \quad y \in H. \tag{1.29}$$

Lemma 1.6.12. *Let $H_r = \{x \in \mathbb{R}^n \mid \langle a, x\rangle = b_r\}$ for $r = 1, 2$ be two parallel hyperplanes in $\mathbb{R}^n$ with $a \in \mathbb{R}^n \setminus \{0\}$, $b_r \in \mathbb{R}$, and let $x \in S$. Then for any $f \in B(S)$ which is zone consistent with respect to both hyperplanes,*

$$\pi_{H_1}(x) \leq \pi_{H_2}(x) \text{ if and only if } b_1 \leq b_2.$$

Lemma 1.6.13. *Let $H = \{x \in \mathbb{R}^n \mid \langle a, x\rangle = b\}$ with $a \neq 0$ be a hyperplane in $\mathbb{R}^n$ and let $y \in S$ and $P_H(y)$ be the Bregman projection of y onto H, $y \notin H$. If $b < \langle a, y\rangle$ then for any $f \in B(S)$ which is zone consistent with respect to the hyperplane, the projection parameter λ is negative.*

Proof. The proof follows from Lemma 1.6.11. If $b < \langle a, y\rangle$ then from (1.28), $\lambda < 0$. □

1.7 Bregman's method for linear constraints

Another common feature of our algorithms is their primal-dual nature. In view of Theorem 1.3.2, the primal-dual approach for constrained optimization problems aims at solving the dual unconstrained problem. This is done by an iterative scheme which alternates between the minimization of the Lagrangian and the application of a steepest ascent iteration [11] to the dual problem.

For simplicity, we first develop a primal-dual algorithm for the following linear equality constrained problem and then extend it to linear inequality constraints.

$$\min f(x) \tag{1.30}$$

$$\text{subject to } Ax = b, \tag{1.31}$$

$$x \in \mathbb{R}^n_+, \tag{1.32}$$

where $\mathbb{R}^n_+$ is the non-negative orthant of the n-dimensional Euclidean plane, A is an $m \times n$ matrix, $b \in \mathbb{R}^m$, $C = \{x \in \mathbb{R}^n_+ \mid Ax = b\} \neq \emptyset$ and f is a Bregman function zone consistent with respect to the hyperplane $H_i := \{x \in \mathbb{R}^n_+ \mid \langle a^i, x\rangle = b_i\}$, $a^i \neq 0$, for $i = 1, \ldots, m$.

The Lagrangian of (1.30)-(1.32) with respect to the equality constraints is

$$L(x, z) = f(x) + \langle z, Ax - b\rangle,$$

where $z \in \mathbb{R}^m_+$ is the dual vector of the Lagrange multipliers. The dual function $g : \mathbb{R}^m_+ \to \mathbb{R}$ is given by

$$g(z) = \min_{x \in \mathbb{R}^n_+} L(x, z)$$

and a necessary condition for the minimization of $L(x, z)$ is

$$\nabla_x L(x, z) = 0 \text{ for } x \in \mathbb{R}^n_{++}.$$

This implies that

$$\nabla f(x) = -A^T z. \tag{1.33}$$

From (1.33), for $x^k \in \mathbb{R}^n_{++}$ and $z^k \in \mathbb{R}^m_+$, $k \geq 0$, we generate an iterative primal-dual algorithm as follows:

$$\nabla f(x^k) = -A^T z^k. \tag{1.34}$$

To do this, the corrections to the dual vectors have to be prescribed. But from the Lagrange Duality Theorem 1.3.2,

$$\min_{x\in\mathbb{R}^n_+\cap H} f(x) = \max_{z\in\mathbb{R}^m_+} g(z).$$

Therefore the dual corrections should at least entail dual ascent, i.e., guarantee that the sequence $\{g(z^k)\}$ is increasing. This dual correction can be represented by

$$z^{k+1} = z^k + y^k \tag{1.35}$$

where $y^k \in \mathbb{R}^m$ is a dual correction vector. Therefore, using (1.34) and (1.35), we have

$$\nabla f(x^{k+1}) = -A^T z^{k+1} = -A^T z^k - A^T y^k = \nabla f(x^k) - A^T y^k. \tag{1.36}$$

If a decision is made to change only one component of z^k, i.e., the $i(k)$th component at each iteration, then we can write $y^k = \theta_k e^{i(k)}$. This enables us to write (1.35) in component form as

$$z_i^{k+1} = \begin{cases} z_i^k, \ i \neq i(k), \\ z_i^k + \theta_k, \ i = i(k), \end{cases} \tag{1.37}$$

and (1.36) in the form

$$\nabla f(x^{k+1}) = \nabla f(x^k) - \theta_k a^i \tag{1.38}$$

for $i \in I := \{1, 2, \ldots, m\}$.

Now if the projection parameter θ_k is calculated so that x^{k+1} satisfies

$$\langle a^i, x^{k+1}\rangle = b_i, \tag{1.39}$$

i.e., the next iterate x^{k+1} lies on the $i(k)$th hyperplane, we obtain Bregman's method for convex programming with linear equality constraints and with the almost cyclic control sequence $\{i(k)\}$.

The solution x^{k+1} of the system (1.38)-(1.39) is the generalized projection of the current primal iterate x^k onto the $i(k)$th hyperplane $H_{i(k)}$, as described in Subsection 1.6.2, and θ_k is the generalized projection parameter.

The primal-dual algorithm for linear inequality constraints also follows the same format and also calculates the parameter θ_k. However, before proceeding, it compares θ_k with the $i(k)$th component of the current dual vector z^k and uses the smaller of these two in the iterative step.

Figure 1.8 illustrates the geometric interpretation of all the possible cases in an iterative step of Bregman's primal-dual algorithm for linear inequality constraints. We state this linear inequality constrained problem and its Bregman's algorithm for solving it in the next subsection.

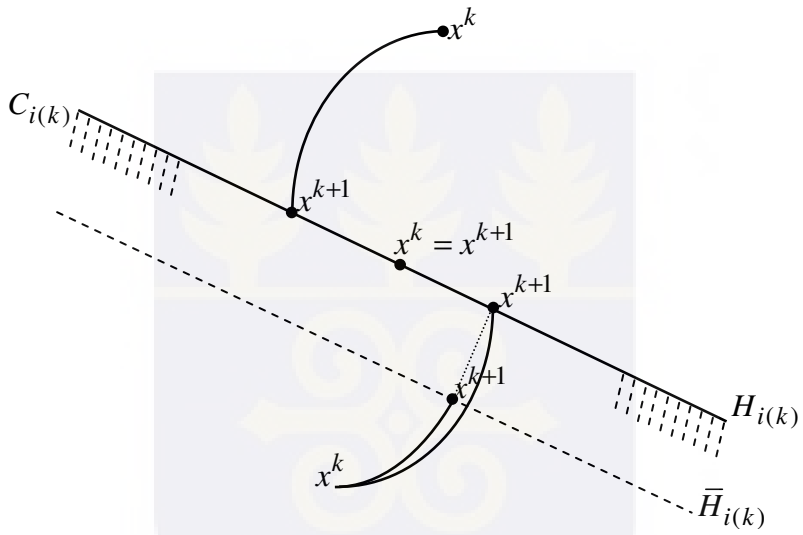

Figure 1.8: Geometric interpretation of all the possible cases of Bregman's algorithm for linear inequality constraints.

Since Bregman's algorithm is a row-action method, it takes at each iteration, a hyperplane H_i and projects onto it the current iterate according to the generalized distance constructed from the objective function $f(x)$, and computes the next iterate. That is, 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 sequence of the algorithm.

Next, we state Bregman's method for solving the linear equality constrained problem (1.30)-(1.32).

Algorithm 1.7.1. Bregman's algorithm for linear equalities

(i) **Initialization** $x^0 \in \text{Intdom} f$ $(\mathbb{R}^n_{++})$ is such that for an arbitrary $z^0 \in \mathbb{R}^m_+$,

$$\nabla f(x^0) = -A^T z^0. \tag{1.40}$$

(ii) **Iterative Step** Given x^k calculate x^{k+1} from the system

$$\nabla f(x^{k+1}) = \nabla f(x^k) + c_k a^{i(k)}, \tag{1.41}$$

$$\langle a^{i(k)}, x^{k+1} \rangle = b_{i(k)}, \tag{1.42}$$

where $c_k = \pi_{H_{i(k)}}(x^k)$. We assume that f is zone consistent with respect to the hyperplane $H_i := \{x \in \mathbb{R}^n_+ \mid \langle a^i, x \rangle = b_i\}$ for $i = 1, \ldots, m$ and that the representation of every hyperplane is fixed during the whole iteration process, so that the values of c_k are well defined.

(iii) **Control** The control sequence $\{i(k)\}$ is almost cyclic on the index set I. $\diamondsuit$

1.7.1 Bregman's method for linear inequality constraints

This subsection describes Bregman's method for linear inequality constraints and emphasizes its relationship with the application of Gauss-Seidel type methods to a dual problem.

Let $f \in B(S)$ and consider the problem

$$\min f(x), \tag{1.43}$$

$$\text{subject to } \langle a^i, x \rangle \leq b_i, \qquad i \in I := \{1, 2, \ldots, m\}, \tag{1.44}$$

$$x \in \bar{S}. \tag{1.45}$$

Let $H_i := \{x \in \mathbb{R}^n \mid \langle a^i, x \rangle = b_i\}$ and $C_i := \{x \in \mathbb{R}^n \mid \langle a^i, x \rangle \leq b_i\}$; denote also $C = \cap_{i=1}^m C_i \neq \emptyset$, and assume that $C \cap \bar{S} \neq \emptyset$. A is an $m \times n$ matrix whose ith row is a^i, and $b \in \mathbb{R}^m$. Assume that $a^i \neq 0$ for all $i \in I$ and that $f \in B(S)$ is strongly zone consistent with respect to every H_i.

Below is the Bregman's algorithm for solving the linear inequality constraints problem (1.43)-(1.45).

Algorithm 1.7.2. Bregman's algorithm for linear inequalities

(i) **Initialization** $x^0 \in S$ is such that for an arbitrary $z^0 \in \mathbb{R}^m_+$,

$$\nabla f(x^0) = -A^T z^0. \tag{1.46}$$

(ii) **Iterative Step** Given x^k and z^k, calculate x^{k+1} and z^{k+1} from

$$\begin{aligned} \nabla f(x^{k+1}) &= \nabla f(x^k) + c_k a^{i(k)}, & (1.47)\\ z^{k+1} &= z^k - c_k e^{i(k)}, & (1.48)\\ c_k &= \min\left(z^k_{i(k)}, \theta_k\right), & (1.49) \end{aligned}$$

where $\theta_k = \pi_{H_{i(k)}}(x^k)$.

(iii) **Control** The control sequence $\{i(k)\}$ is almost cyclic on the index set I. $\diamondsuit$

1.7.2 On relaxation

In [41], an important modification of Bregman's method was made by incorporating into the method a relaxation parameter. The underlying idea behind the relaxation strategy is to relax the constraints before computing the Bregman projections. More precisely, if at iteration k, the ith constraint

$$\langle a^i, x\rangle \leq b_i$$

is to be used, we substitute it for the relaxed constraint

$$\langle a^i, x\rangle \leq \alpha b_i + (1-\alpha)\langle a^i, x\rangle.$$

This simply means that instead of projecting onto the hyperplane

$$H_i = \{x \in \mathbb{R}^n \mid \langle a^i, x\rangle = b_i\},$$

we project onto the relaxed hyperplane

$$H_i(\alpha) = \{x \in \mathbb{R}^n \mid \langle a^i, x\rangle = \alpha b_i + (1-\alpha)\langle a^i, x^k\rangle\},$$

where the so called relaxation parameter (possibly also depending on k), α, lies in the interval $(0, 1]$. In Chapter 2, we elaborate further on this concept and its consequences to developing closed-form formulas in Chapter 4. Clearly, relaxation can also be applied to the last algorithm for inequality constraints.

1.8 Entropy maximization and closed formulas: MART, SMART and related methods

In this section, we describe existing results that simplify the application of Bregman's method to entropy maximization and its related problems.

The $x \log x$ entropy function, ent x, which maps the nonnegative orthant $\mathbb{R}^n_+$ into $\mathbb{R}$ according to

$$-\text{ent}(x) = \sum_{j=1}^{n} (x_j \log x_j),$$

where, by convention, $0 \log 0 = 0$, is a Bregman function with zone $\mathbb{R}^n_{++}$ (see page 33 of [32]). This function, known as Shannon's entropy, has its origins in information theory.

We now consider the following entropy optimization problem over linear equality constraints.

$$\min f(x), \tag{1.50}$$

$$\text{subject to } Ax = b, \tag{1.51}$$

$$x \in \mathbb{R}^n_+, \tag{1.52}$$

where $f(x) = \sum_{j=1}^{n} (x_j \log x_j)$.
Using Algorithm 1.7.2 with $f(x) = \sum_{j=1}^{n} (x_j \log x_j)$, we have the algorithm for solving (1.50)-(1.52) as

Algorithm 1.8.1. Bregman's method for equality constrained entropy optimization

(i) **Initialization** $x^0 \in \mathbb{R}^n_{++}$ is such that for an arbitrary $z^0 \in \mathbb{R}^m_+$

$$x^0_j = \exp\left(\left(-A^T z^0\right)_j - 1\right), \ j = 1, 2, \ldots, n.$$

(ii) **Iterative Step** Given x^k choose a control index $i(k)$ and solve the system

$$x^{k+1}_j = x^k_j \exp\left(c_k a^i_j\right), \ j = 1, 2, \ldots, n, \tag{1.53}$$

$$b_i = \langle a^i, x^{k+1} \rangle. \tag{1.54}$$

(iii) **Control** The sequence $\{i(k)\}$ is almost cyclic on $I = \{1, 2, \ldots, m\}$. ♢

We discuss Algorithm 1.8.1 and its relationship with the algorithm MART.

From Lemma 1.6.10, there exists a unique choice of x^{k+1} and c_k that satisfies the system (1.53)-(1.54). To proceed with the iteration, this system with $n+1$ equations needs to be solved for x^{k+1} and c_k.

An alternative algorithm for solving this same problem is MART which employs a closed-form formula for the iterative updates instead of solving a system of $n+1$ nonlinear equations. However, for the proof of convergence for MART, the following assumptions are made.

Assumption 1.8.2.

(i) Feasibility: $\{x \in \mathbb{R}^n \mid Ax = b\} \cap \mathbb{R}^n_+ \neq \emptyset$.

(ii) Non-negativity: $a^i_j \geq 0$, and $b_i > 0$ for all $i \in I$ and $j = 1, 2, \ldots, n$.

(iii) Normalization: $Ax = b$ is scaled so that for all $i \in I$ and $j = 1, 2, \ldots, n$, $a^i_j \leq 1$. ♢

With this assumption, we state the algorithm for MART:

Algorithm 1.8.3. Multiplicative Algebraic Reconstruction Technique

(i) **Initialization** $z^0 \in \mathbb{R}^m_+$ is arbitrary, and $x^0 \in \mathbb{R}^n_{++}$ is given by

$$1 + \log x^0_j = \left(-A^T z^0\right)_j, \ j = 1, 2, \ldots, n.$$

(ii) **Iterative Step**

$$x^{k+1}_j = x^k_j \left(\frac{b_i}{\langle a^i, x^k \rangle}\right)^{a^i_j}, \ j = 1, 2, \ldots, n. \tag{1.55}$$

(iii) **Control**: The sequence $\{i(k)\}$ is almost cyclic on I. ♢

The question we want to address briefly here, and in more detail in Chapter 4, is whether it is possible to replace the iterative step of the system (1.53)-(1.54)

in the Algorithm 1.8.1 with a closed-form formula as it is the case for orthogonal projections onto hyperplanes using Assumption 1.8.2?

The answer to this question will motivate the development of general closed-form formula for the iterative steps in Bregman's algorithm for linear constraints in Chapter 4.

To answer this question, observe that in order to use Algorithm 1.8.1, the system (1.53)-(1.54) has to be solved for c_k and x^{k+1} in each iterate. Doing this for the kth iterate, by eliminating x^{k+1} from this system, we have

$$\sum_{j=1}^{n} a_j^i x_j^k \exp\left(c_k a_j^i\right) - b_i = 0.$$

Let $\exp c_k = y_k$ and define the function $f_k : \mathbb{R}_+ \to \mathbb{R}$ by

$$f_k(y_k) = \sum_{j=1}^{n} a_j^i x_j^k \exp\left(c_k a_j^i\right) - b_i.$$

Then we need a positive root of f_k to determine c_k. Now, if the conditions or Assumption 1.8.2 that enable the convergence of MART are imposed on Algorithm 1.8.1 then $f_k(0) = -b_i < 0$, $\lim_{y_k \to \infty} f_k(y_k) = +\infty$, and for $y_k \geq 0$, we have the derivatives $f_k'(y_k) > 0$ and $f_k''(y_k) \leq 0$. Now, since $f_k(0) < 0$ and $\lim_{y_k \to \infty} f_k(y_k) = +\infty$, there exists $\widehat{y}_k > 0$ such that $f_k(\widehat{y}_k) > 0$. Also since f_k is continuous, the intermediate value theorem ensures that for some $y^* \in (0, \widehat{y}_k)$ we have $f_k(y_k^*) = 0$. Since $f_k'(y_k) > 0$ for all $y_k > 0$, it follows that f_k is strictly increasing on $(0, \infty)$. Hence it is one-to-one there, and therefore if $f_k(\widetilde{y}_k) = 0$ for some $\widetilde{y}_k > 0$ then ${y^*}_k = \widetilde{y}_k$ and so $y_k^* > 0$ is unique and $f_k(y_k^*) = 0$.

Now consider the line through the points $(0, -b_i)$ and $(1, f_k(1))$ in the plane of the graph of $f_k(y_k)$. This line or the secant line to the graph intersects the y_k-axis at the point $\bar{y}_k$ given by

$$\bar{y}_k = \frac{b_i}{\langle a^i, x^k \rangle}.$$

$\bar{y}_k$ is therefore considered as a secant approximation to the root y_k^* of f_k and $\exp \bar{c}_k = \bar{y}_k$. Hence

$$\bar{c}_k = \log \frac{b_i}{\langle a^i, x^k \rangle}.$$

When this approximate value $\bar{c}_k$ is substituted into (1.53) of Algorithm 1.8.1, we obtain the closed-form formula (1.55) of Algorithm 1.8.3 which is MART's iterative step. Therefore MART's iterative step is a secant approximation to Bregman's iterative step if the conditions that enable the convergence of MART hold. Details of these can be found in [23].

The question of whether the algorithms MART and SMART will converge when $\{x \in \mathbb{R}^n \mid Ax = b\} = \emptyset$ will be addressed in Chapter 5.

Chapter 2

The convex feasibility problem and block Bregman methods for equality constraints

Now, let us go back to the CFP (1.9) and its solution using a sequence of Bregman projections. In this chapter, we show how to generalize the method of relaxed Bregman projections onto closed convex sets, itself a generalization of POCS. As a consequence, we derive an application to convex but nonlinear sets of constraints and to linear equality constraints as well.

Before presenting the algorithm, we need to define a separating hyperplane.

Definition 2.0.4. Separating hyperplanes For a given point $\bar{x}$ and a closed, nonempty convex set C, define for $s \in \mathbb{R}^n \setminus \{0\}$ and $d \in \mathbb{R}$, $H^s = \{x \in \mathbb{R}^n \mid \langle s, x\rangle = d\}$. We say that H^s separates $\bar{x}$ and C, that is, it is a *separating hyperplane* for $\bar{x}$ and C, if $\langle s, x\rangle \leq d \, \forall x \in C$ and $\langle s, \bar{x}\rangle \geq d$. $\diamondsuit$

Definition 2.0.5. Supporting hyperplanes For a given closed and nonempty convex set C and a given point $\bar{x} \in \partial C$, the boundary of C, define for $s \in \mathbb{R}^n \setminus \{0\}$ the hyperplane, $H^s = \{x \in \mathbb{R}^n \mid \langle s, x - \bar{x}\rangle = 0\}$. We say that H^s *supports* C at $\bar{x}$ if either $\langle s, x - \bar{x}\rangle \geq 0$ for all $x \in C$, or $\langle s, x - \bar{x}\rangle \leq 0$ for all $x \in C$. $\diamondsuit$

It is clear from Proposition 11 in [24] that, for a given closed and nonempty

convex set C and a point x, if $P_C(x)$ is the orthogonal projection of x onto C, then $H = \{y \in \mathbb{R}^n \mid \langle y - P_C(x), x - P_C(x)\rangle = 0\}$ defines a tangent hyperplane at $P_C(x)$ which is a supporting and a separating hyperplane of $x \notin C$ and C.

The following result extends this concept to Bregman projections. If we consider $P_C(x)$ as a Bregman projection of x onto C then there exists a tangent hyperplane to C at the point $P_C(x)$, see [24] . This hyperplane will be related to our generalization of relaxation.

Our new general method for solving (1.9) will be defined as follows.

$$x^{k+1} = P^s_{C_{i(k)}}(x^k) \tag{2.1}$$

where $P^s_{C_{i(k)}}$ denotes the Bregman projection onto a separating hyperplane for the point x^k and the closed convex set C_i for $i = 1, \ldots, m$.

Next, we state a well-known result, the three-point lemma by Chen and Teboulle 1993, which is widely used in the analysis of generalized Bregman projection methods.

Lemma 2.0.6. *Let f be a Bregman function with zone S. Then for any $x, z \in S$ and $y \in \bar{S}$,*

$$D_f(y, x) = D_f(z, x) + D_f(y, z) + \langle \nabla f(x) - \nabla f(z), z - y\rangle. \tag{2.2}$$

The next theorem is Lemma 1 in [14]. We repeat the statement and its proof here because of its important to the work in this chapter.

Theorem 2.0.7. *Let f be a Bregman function with zone S and let C be a closed and nonempty convex set such that $C \cap \bar{S} \neq \emptyset$. Define $P_C(x)$ as the Bregman projection of x onto C and assume that for any $x \in S$, $P_C(x) \in S$. Then the following inequality is true*

$$D_f(P_C(x), x) \leq D_f(y, x) - D_f(y, P_C(x)) \tag{2.3}$$

for every $y \in C \cap \bar{S}$.

Proof. Using the convexity of the function G in Lemma 1.6.6 and the fact that $D_f(x,x) = 0$ for all $x \in S$, we have, for any $\alpha \in (0,1]$,

$$D_f(\alpha y + (1-\alpha)P_C(x), x) - D_f(\alpha y + (1-\alpha)P_C(x), P_C(x)) \leq \tag{2.4}$$

$$\alpha D_f(y,x) + (1-\alpha)D_f(P_C(x),x) - \alpha D_f(y, P_C(x)). \tag{2.5}$$

This implies that

$$D_f(y,x) - D_f(P_C(x),x) - D_f(y,P_C(x)) \geq \tag{2.6}$$

$$\frac{D_f(\alpha y + (1-\alpha)P_C(x),x) - D_f(P_C(x),x)}{\alpha} - \frac{D_f(\alpha y + (1-\alpha)P_C(x), P_C(x))}{\alpha}. \tag{2.7}$$

The first term on the right-hand side of (2.7) is non-negative $\forall \alpha \in (0,1]$ because $P_C(x)$ is a minimizer, and the second term tends to zero as α tends to zero because of the definition of the Bregman measure. That is,

$$\lim_{\alpha \to 0} \frac{D_f(\alpha y + (1-\alpha)P_C(x), P_C(x))}{\alpha}$$

$$= \lim_{\alpha \to 0} \frac{D_f(P_C(x) + \alpha(y - P_C(x)), P_C(x)) - D_f(P_C(x), P_C(x))}{\alpha}$$

$$= \langle \nabla_z D_f(z, P_C(x))|_{z=P_C(x)}, y - P_C(x)\rangle = 0,$$

since $D_f(x,y) = f(x) - f(y) - \langle \nabla f(y), x - y\rangle$ implies $\nabla_x D_f(x,y) = \nabla f(x) - \nabla f(y)$ and so $\nabla_x D_f(y,y) = 0$. This completes the proof. □

A deeper analysis of this result (Theorem 2.0.7) for inequalities, not considered by Bregman, gives us a better geometrical view of Bregman projections. In (2.4), the difference is not only convex but linear in the first variable and so

$$D_f(y,x) - D_f(P_C(x),x) - D_f(y,P_C(x)) = \tag{2.8}$$

$$\frac{D_f(\alpha y + (1-\alpha)P_C(x),x) - D_f(P_C(x),x)}{\alpha} \tag{2.9}$$

$$- \frac{D_f(\alpha y + (1-\alpha)P_C(x), P_C(x))}{\alpha}. \tag{2.10}$$

Using the definition of the Bregman measure to expand (2.9), we have

$$\frac{f(\alpha y + (1-\alpha)P_C(x)) - f(x) - \langle \nabla f(x), \alpha y + (1-\alpha)P_C(x) - x\rangle}{\alpha}$$

$$-\frac{[f(P_C(x)) - f(x) - \langle \nabla f(x), P_C(x) - x\rangle]}{\alpha}$$
$$= \frac{f(\alpha y + (1-\alpha)P_C(x)) - f(P_C(x)) - \alpha\langle \nabla f(x), y - P_C(x)\rangle}{\alpha} \tag{2.11}$$

and also for the expansion of (2.10), we have

$$\frac{f(\alpha y + (1-\alpha)P_C(x)) - f(P_C(x)) - \alpha\langle \nabla f(P_C(x)), y - P_C(x)\rangle}{\alpha}. \tag{2.12}$$

Therefore using (2.8), (2.9), (2.10), (2.11) and (2.12) , we have

$$D_f(y,x) - D_f(P_C(x),x) - D_f(y,P_C(x)) = \langle \nabla f(P_C(x)) - \nabla f(x), y - P_C(x)\rangle \tag{2.13}$$

for every $y \in C$, which, if $P_C(x) = z$, is the result of Lemma 2.0.6.

If f is $\frac{1}{2}\|\cdot\|^2$ then (2.13) gives, as expected, the cosine of the angle determined by the points x, $P_C(x)$ and y. When this is zero, the Pythagoras theorem for equalities is retrieved.

In the general case, if f is a Bregman function with zone S and C is a closed and nonempty convex set such that $C \cap \bar{S} \neq \emptyset$, and if f is zone consistent with respect to C then, for $x \notin C$, the equation

$$\langle \nabla f(x) - \nabla f(P_C(x)), y - P_C(x)\rangle = 0 \tag{2.14}$$

defines a hyperplane that is tangent to C at $P_C(x)$, where $P_C(x)$ is the Bregman projection of x onto C.

Remark 2.0.8. It should be noted that if $x \notin C$ then $x \neq P_C(x)$ and since f is a Bregman function assumed to be strictly convex on $\bar{S}$, ∇f is strictly monotone on S (see page 10 of [77]) and so $\nabla f(x) \neq \nabla f(P_C(x))$.

Alternatively, if f satisfies Assumption 2.2.2 so that ∇f is invertible from S onto $\mathbb{R}^n$ then ∇f is one-to-one and therefore $\nabla f(x) = \nabla f(P_C(x))$ implies $x = P_C(x)$, a contradiction to the assumption that $x \notin C$. Thus $\nabla f(x) \neq \nabla f(P_C(x))$ for all $x \in S$. ♢

Again, the next theorem which is Theorem 2.4.2 on page 43 of [32] is repeated here with its proof because of its important to the work of this section.

Theorem 2.0.9. *Under the assumption of Theorem 2.0.7, for any $x \in S$, $v \in C \cap \bar{S}$ is $P_C(x)$ if and only if*

$$\langle \nabla f(x) - \nabla f(v), y - v \rangle \leq 0 \tag{2.15}$$

for all $y \in C \cap \bar{S}$.

Proof. If $z = P_C(x)$ in Lemma 2.0.6 then (2.15) follows from (2.2) and (2.3). Conversely, if (2.15) holds then with $v = P_C(x)$ in Lemma 2.0.6, (2.2) reduces to the inequality

$$D_f(v, x) \leq D_f(y, x) \text{ for all } y \in C \cap \bar{S}$$

since D_f is nonnegative. Therefore $v = P_C(x)$.

□

This theorem says that a hyperplane H through the point $P_C(x)$, $x \notin C$, and which is perpendicular to $\nabla f(x) - \nabla f(P_C(x))$ supports the convex set C at the point $P_C(x)$. That is, the set C lies entirely on one side of H.

Definition 2.0.10. Generalized tangent hyperplane For a given closed and nonempty convex set C, and a given Bregman function f with zone S such that $C \cap \bar{S} \neq \emptyset$, define $P_C(x)$ as the Bregman projection of x onto C. If f is zone consistent with respect to C, then the *generalized tangent hyperplane at* $P_C(x)$ is the set

$$\{y \in \bar{S} \mid \langle y - P_C(x), \nabla f(x) - \nabla f(P_C(x)) \rangle = 0\} \tag{2.16}$$

if $x \notin C$. ♢

Now consider the minimization problem for a given $y \in S$, where S is a zone of the Bregman function f, and a closed and nonempty convex set C.

$$\min_{x \in C \cap \bar{S}} G(x) := f(x) - f(y) - \langle \nabla f(y), x - y \rangle. \tag{2.17}$$

The condition for the minimum $P_C(y)$, that is, the Bregman projection of y onto C, is given by

$$\langle -\nabla G(P_C(y)), x - P_C(y) \rangle = \langle \nabla f(y) - \nabla f(P_C(y)), x - P_C(y) \rangle \leq 0 \tag{2.18}$$

for all $x \in C \cap \bar{S}$, which describes the hyperplane from the Pythagorean equality. We note that, with $G(x) = f(x) - f(y) - \langle \nabla f(y), x - y \rangle$, $\nabla G(x) = \nabla f(x) - \nabla f(y)$ and so $\nabla G(P_C(y)) = \nabla f(P_C(y)) - \nabla f(y)$.
Thus $\langle \nabla G(P_C(y)), x - P_C(y) \rangle = \langle \nabla f(P_C(y)) - \nabla f(y), x - P_C(y) \rangle$ and (2.18) follows from Theorem 2.0.9.

2.1 An extension of relaxation

In this section, we extend the concept of relaxation for Bregman projections, first proposed in [41] and further extended in [29] to, not necessarily the linear case. As described in Subsection 1.7.2, for a given hyperplane

$$H = \{x \in \mathbb{R}^n \mid \langle a, x \rangle = b\},$$

a relaxed Bregman projection onto H with relaxation parameter $\alpha \in (0, 1]$ is defined as the Bregman projection onto the parallel hyperplane defined by

$$H = \{x \in \mathbb{R}^n \mid \langle a, x \rangle = \alpha b + (1 - \alpha)\langle a, x^k \rangle\}. \tag{2.19}$$

We generalize this concept to a general closed convex sets in a natural way in the next subsection.

2.1.1 Relaxed Bregman projections onto closed convex sets

For a given closed and nonempty convex set $C \subseteq \mathbb{R}^n$, a point $x \in \mathbb{R}^n$ and its Bregman projection $P_C(x)$ onto C, we define a relaxed Bregman projection onto C with parameter $\alpha \in (0, 1]$, $P_{H(\alpha,x)}(x)$, as the Bregman projection onto the hyperplane defined by

$$\begin{aligned} H(\alpha, x) \quad = \quad & \{y \in \mathbb{R}^n \mid \langle \nabla f(x) - \nabla f(P_C(x)), y \rangle = \\ & \alpha \langle \nabla f(x) - \nabla f(P_C(x)), P_C(x) \rangle + (1 - \alpha)\langle \nabla f(x) - \nabla f(P_C(x)), x \rangle\} \end{aligned} \tag{2.20}$$

if $x \notin C$. If $x \in C$ then the hyperplane is the entire space $\mathbb{R}^n$ and so $x, P_{H(\alpha,x)}(x) \in \mathbb{R}^n$, since in this case, the normal vector $\nabla f(x) - \nabla f(P_C(x)) = 0$. We show that

$H(\alpha, x)$ is a separating hyperplane for the point $x \notin C$ and C in the following proposition.

Proposition 2.1.1. *Suppose f is a Bregman function with zone S and let C be a closed and nonempty convex set such that $C \cap \bar{S} \neq \emptyset$. Define $P_C(x)$ as the Bregman projection of x onto C and assume that for any $x \in S$, $P_C(x) \in S$. Then, for $\alpha \in (0, 1]$ and $x \in S$, the hyperplane*

$$\begin{aligned} H(\alpha, x) &= \{y \in \mathbb{R}^n \mid \langle \nabla f(x) - \nabla f(P_C(x)), y\rangle = \\ &\quad \alpha\langle \nabla f(x) - \nabla f(P_C(x)), P_C(x)\rangle + (1-\alpha)\langle \nabla f(x) - \nabla f(P_C(x)), x\rangle\} \end{aligned}$$

separates the point $x \notin C$ from $C \cap \bar{S}$ for $x \in S$.

Proof. We first show that $H(1, x)$ is a separating and supporting hyperplane of C at the point $P_C(x)$ of C and then deduce that $H(\alpha, x)$ separates x from $C \cap \bar{S}$ for $\alpha \in (0, 1)$, since $H(\alpha, x)$ is parallel to $H(1, x)$ and lies between x and $H(1, x)$ for $\alpha \in (0, 1)$ and $x \in S$.

Now by definition,

$$\begin{aligned} H(1, x) &= \{y \in \mathbb{R}^n \mid \langle \nabla f(x) - \nabla f(P_C(x)), y\rangle = \langle \nabla f(x) - \nabla f(P_C(x)), P_C(x)\rangle\} \\ &= \{y \in \mathbb{R}^n \mid \langle \nabla f(x) - \nabla f(P_C(x)), y - P_C(x)\rangle = 0\} \end{aligned}$$

is the generalized tangent hyperplane in (2.16).

By Theorem 2.0.9,

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

for all $y \in C$ and so $C \subseteq \{y \in \mathbb{R}^n \mid \langle \nabla f(x) - \nabla f(P_C(x)), y - P_C(x)\rangle \leq 0\}$. Thus $C \cap \bar{S}$ lies on one side of $H(1, x)$, and since by definition, $H(\alpha, x)$ is parallel to $H(1, x)$ and lies between x and $H(1, x)$ for $\alpha \in (0, 1)$, $C \cap \bar{S}$ lies on one side of $H(\alpha, x)$.

On the other hand, if $x \in S$ but $x \notin C$ then

$$\begin{aligned} \langle \nabla f(x) - \nabla f(P_C(x)), x\rangle &= \alpha\langle \nabla f(x) - \nabla f(P_C(x)), P_C(x)\rangle \\ &\quad +(1-\alpha)\langle \nabla f(x) - \nabla f(P_C(x)), x\rangle \end{aligned}$$

implies

$$\langle \nabla f(x) - \nabla f(P_C(x)), \alpha(x - P_C(x)) \rangle = 0.$$

But, by the definition of D_f,

$$\begin{aligned} D_f(x,y) + D_f(y,x) &= f(x) - f(y) - \langle \nabla f(y), x - y \rangle \\ &\quad + f(y) - f(x) - \langle \nabla f(x), y - x \rangle \\ &= \langle \nabla f(x) - \nabla f(y), x - y \rangle. \end{aligned} \tag{2.21}$$

Therefore

$$\alpha \langle \nabla f(x) - \nabla f(P_C(x)), x - P_C(x) \rangle = \alpha(D_f(x, P_C(x)) + D_f(P_C(x), x)) > 0.$$

Hence, for $\alpha \in (0, 1]$ and $x \in S$, $H(\alpha, x)$ separates x from $C \cap \bar{S}$. □

2.1.2 The relationship with the Censor-Herman definition

A definition of underrelaxation of the Bregman projection onto a general closed convex set (not necessarily linear or half space) was given by Yair Censor and Gabor T. Herman in [29]. This definition includes as special cases the underrelaxed orthogonal projections and the underrelaxed Bregman projections onto linear constraints, i.e., hyperplanes and half-spaces, as given in [41]. Censor-Herman in [29] defines the underrelaxed Bregman projection of a point $x \in S$ onto a general closed convex set C, with respect to a Bregman function f and with a relaxation parameter $\lambda \in [0, 1]$, as a point $P_{C,\lambda}(x)$ that satisfies the equation

$$\nabla f(P_{C,\lambda}(x)) = (1 - \lambda)\nabla f(x) + \lambda \nabla f(P_C(x)), \tag{2.22}$$

where $P_C(x)$ is the Bregman projection of x onto C and S is the zone of f. In Proposition 1 in [29], it was shown that f is zone consistent with respect to C. This proposition also guarantees the invertibility of ∇f and hence the uniqueness of $P_{C,\lambda}$. In particular, for every $z \in S$, $P_{C,\lambda}(z) \in S$.

We prove in the following proposition that our new extended definition of underrelaxed Bregman projections onto a general closed convex sets contains that of Censor-Herman.

Proposition 2.1.2. *Let f be a Bregman function with zone $S = \mathrm{Int}(\mathrm{dom} f)$, and C be a general closed and nonempty convex set. Assume that f is strongly zone consistent with respect to the separating hyperplane $H(\alpha, x)$ for $x \in S$, $0 < \alpha \leq 1$, of C defined in (2.20). Then there exists $\lambda \in [0, 1]$ such that, for $x \in S$, $P_{C,\lambda}(x)$ of (2.22) is the underrelaxed Bregman projection $P_{H(\alpha,x)}(x)$, i.e., the Bregman projection of x onto the separating hyperplane $H(\alpha, x)$.*

Proof. We observe that if $x \in C \cap S$ then there is nothing to prove. Since, in this case, $\nabla f(x) - \nabla f(P_C(x)) = 0$ and the hyperplanes $H(1, x)$ and $H(\alpha, x)$ for $\alpha \in (0, 1)$ and $x \in S$ are the entire $\mathbb{R}^n$ and so, $P_C(x) = P_{H(1,x)}(x) = P_{H(\alpha,x)}(x) = x$. Thus

$$\nabla f(P_{H(\alpha,x)}(x)) = (1 - \lambda)\nabla f(x) + \lambda \nabla f(P_{H(1,x)}(x)),$$

holds for all $\lambda \in \mathbb{R}$ and in particular for $\lambda \in [0, 1]$. Therefore suppose that $x \notin C$. Then $P_{H(\alpha,x)}(x)$ is the relaxed Bregman projection of x onto $H(1, x)$ and by Lemma 1.6.10, it is uniquely determined with its projection parameter θ by the system

$$\begin{aligned} \nabla f(P_{H(\alpha,x)}(x)) &= \nabla f(x) + \theta(\nabla f(x) - \nabla f(P_C(x))) \quad (2.23) \\ \langle P_{H(\alpha,x)}(x), \nabla f(x) - \nabla f(P_C(x))\rangle &= \alpha\langle \nabla f(x) - \nabla f(P_C(x)), P_C(x)\rangle \\ &\quad + (1 - \alpha)\langle \nabla f(x) - \nabla f(P_C(x)), x\rangle. \end{aligned}$$

Similarly, the Bregman projection of x onto $H(1, x)$, $P_{H(1,x)}(x)$, is also uniquely determined with its projection parameter θ^* by the system

$$\begin{aligned} \nabla f(P_{H(1,x)}(x)) &= \nabla f(x) + \theta^*(\nabla f(x) - \nabla f(P_C(x))) \quad (2.24) \\ \langle P_{H(1,x)}(x), \nabla f(x) - \nabla f(P_C(x))\rangle &= \langle \nabla f(x) - \nabla f(P_C(x)), P_C(x)\rangle. \end{aligned}$$

Now, by Lemma 1.6.12, $\theta^* \leq \theta$, and since $\alpha \in (0, 1]$, i.e., $\alpha \neq 0$, $x \notin H(\alpha, x)$ and so by Lemma 1.6.11, $\theta^* \neq 0$. This means that $\theta^* \leq \theta$ and θ/θ^* is well defined. Therefore, from (2.23) and (2.24),

$$\begin{aligned} \nabla f(P_{H(\alpha,x)}(x)) &= \nabla f(x) + \theta(\nabla f(x) - \nabla f(P_C(x))) \\ &= \nabla f(x) + \frac{\theta}{\theta^*}(\nabla f(P_{H(1)}(x)) - \nabla f(x)) \\ &= \left(1 - \frac{\theta}{\theta^*}\right)\nabla f(x) + \frac{\theta}{\theta^*}\nabla f(P_{H(1,x)}(x)). \end{aligned}$$

Now since $x \notin C$ and the hyperplane $H(\alpha, x)$ separates C from x, $\nabla f(x) - \nabla f(P_C(x))$ is the outward normal to the hyperplane $H(\alpha, x)$ and so by Lemma 1.6.13 , the projection parameters of x onto $H(\alpha, x)$ for $\alpha \in (0, 1]$ are negative. Thus $\theta^* \leq \theta$ implies $1 \geq \frac{\theta}{\theta^*}$ and so if we define $\lambda = \frac{\theta}{\theta^*}$ then $0 \leq \lambda \leq 1$ and

$$\nabla f(P_{H(\alpha,x)}(x)) = (1-\lambda)\nabla f(x) + \lambda \nabla f(P_{H(1,x)}(x)),$$

which completes the proof. □

2.1.3 The relationship with the Aharoni-Berman-Censor definition

In [3], Aharoni, Berman and Censor associate the set $A_Q(x)$ with the closed convex set Q and a point $x \in \mathbb{R}^n$ as follows

$$A_Q(x) = \begin{cases} \{x\} \text{ if } x \in Q, \\ \{x + \lambda(P_H(x) - x) \mid H \in \mathcal{H}_{x,Q},\ \eta \leq \lambda \leq 2 - \eta\} \text{ if } x \notin Q, \end{cases} \tag{2.25}$$

where $0 \leq \eta \leq 1$ and λ is a relaxation parameter. Here $P_H(x)$ is the orthogonal projection of $x \notin Q$ onto a hyperplane H that separates the ball $B(x, \delta d(x, Q))$ with center x and radius $\delta d(x, Q)$ for $0 \leq \delta \leq 1$, from Q; $d(x, Q)$ is the Euclidean distance between x and Q. $\mathcal{H}_{x,Q}$ is the collection of all hyperplanes that separate $B(x, \delta d(x, Q))$ from Q. The set $A_Q(x)$ is then used to describe the iterative step in their (δ, η)-algorithm.

Next, we show that in the special case when $f = \frac{1}{2}\|\cdot\|^2$, method (2.1) coincides with the method by Aharoni-Berman-Censor in [3].

Proposition 2.1.3. *If $f = \frac{1}{2}\|\cdot\|^2$ then $H(\alpha, x) \in \mathcal{H}_{x,C}$ for a closed convex set C, $\alpha \in (0, 1]$, and $\{P_{H(\alpha,x)}(x)\} \in A_C(x)$, where $A_C(x)$ is as defined by (2.25) with Q replaced by C and $H(\alpha, x)$ is the separating hyperplane of C.*

Proof. Since, by the definition in [3], $\mathcal{H}_{x,C}$ is a collection of all hyperplanes that separate the ball $B(x, \delta d(x, C))$ from the closed convex set C for $0 \leq \delta \leq 1$ if $x \notin C$, by Proposition 2.1.1, $H(\alpha, x) \in \mathcal{H}_{x,C}$ for $\alpha \in (0, 1]$. Therefore $\{P_{H(\alpha,x)}(x)\} \in A_C(x)$. If $x \in C$ then, as observed in the beginning of the proof of Proposition 2.1.2, $P_{H(\alpha,x)}(x) = x$ and so $\{P_{H(\alpha,x)}(x)\} \in A_C(x)$. □

2.2 A general Bregman projection method

In this subsection, we propose a block-iterative algorithm with underrelaxed Bregman projections as defined in Subsection 2.1.1 for solving (1.9). By block-iterative, we mean that, at the kth iteration, the next iterate x^{k+1} is generated from the current iterate x^k by using a subset or a block of the family of the closed convex sets $\{C_i\}_{i=1}^m$ in the convex feasibility problem (1.9). Such a block-iterative scheme for a convex feasibility problem was first proposed in [5] by Aharoni and Censor.

We partition the integer interval $[1, m]$ into p disjoint blocks or intervals Ω_l, that is, $[1, m] = \cup_{l=1}^{p} \Omega_l$ and let $Q_l = \cap_{i\in\Omega_l} C_i$. Let f be a Bregman function with zone $S = \text{Int}(\text{dom} f)$, and assume that f is essentially smooth and strongly zone consistent with respect to the separating hyperplane $H_i(\alpha, x)$ of C_i such that $H_i(\alpha, x) \cap \bar{S} \neq \emptyset$. Then, our new general method for solving (1.9) will be defined as follows.

$$\nabla f(x^{k+1}) = \sum_{i\in\Omega_{l(k)}} \lambda_i \nabla f(P_{H_i(\alpha,x^k)}(x^k)), \tag{2.26}$$

where $P_{H_i(\alpha,x^k)}(x^k)$ denotes the Bregman projection of the current iterate x^k onto the hyperplane $H_i(\alpha, x^k)$ that separates x^k from the closed convex set $C_i \cap \bar{S}$, as defined in Subsection 2.1.1. λ_i's are positive numbers which are bounded away from zero, i.e., there exists $\epsilon > 0$ such that $\lambda_i \geq \epsilon$ for each i, and

$$\sum_{i\in\Omega_{l(k)}} \lambda_i = 1. \tag{2.27}$$

$\{l(k)\}$ is a cyclic control sequence over the index set $\{1, \ldots, p\}$.

Remark 2.2.1. It must be noted that the applicability of the algorithm defined by (2.26) depends on the ability to invert the gradient ∇f explicitly. If the Bregman function f is essentially smooth then ∇f is one-to-one mapping with continuous inverse $(\nabla f)^{-1}$, see [[78], Theorem 26.5]. We must also assume zone consistency; that is, all the gradients that appear in the iterative step are defined. ♢

Based on Remark 2.2.1, the Bregman function f used in the algorithm defined by (2.26) and the iterates produced by the algorithm must satisfy the following assumption.

Assumption 2.2.2. *$f \in B(S)$ is essentially smooth and zone consistent or strongly zone consistent with respect to every hyperplane and the iterate $x^k \in S$ for all $k \geq 0$.* ♢

2.2.1 A convergence theorem

We will establish the convergence of the sequence generated by (2.26) for $p = 1$. It however remains a conjecture that the sequence generated by the general block-iterative Bregman projection method (2.26) for $p > 1$ converges to the solution of (1.9).

To do this, we will use Propositions 2.2.4-2.2.6 below. The method of proof is closely related to the previous proofs given by Aharoni, Berman and Censor in [3], and Censor and Herman in [29].

The properties of the Bregman function f in relation to a closed and nonempty convex set C with zone $S = \text{Int}(\text{dom} f)$ such that $C \cap \bar{S} \neq \emptyset$ will be used repeatedly in the analysis leading to the proof of Theorem 2.2.9.

Properties 2.2.3.

(i) $\langle \nabla f(x) - \nabla f(P_C(x)), y - P_C(x) \rangle \leq 0$ for $y \in C \cap \bar{S}$ and $x \in S$.

(ii) $\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq 0$, $\forall x, y \in S$, i.e.,∇f is a monotone operator on S.

(iii) $\langle \nabla f(x) - \nabla f(y), x - y \rangle = D_f(x, y) + D_f(y, x)$ for all $x, y \in S$.

(iv) For every $\alpha \in \mathbb{R}$, 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 $y^k \in S$ for all $k \geq 0$, and $\lim_{k\to\infty} y^k = y^*$ then $\lim_{k\to\infty} D_f(y^*, y^k) = 0$.

(vi) If $y^k \in S$ and $x^k \in \bar{S}$ for all $k \geq 0$ and if $\lim_{k\to\infty} D_f(x^k, y^k) = 0$, and $\lim_{k\to\infty} y^k = y^*$, and $\{x^k\}$ is bounded, then $\lim_{k\to\infty} x^k = y^*$. ♢

Proof. The proof of (i) is the proof of Theorem 2.0.9. The proofs of (ii) and (iii) are part of the proof of Proposition 2.1.1, precisely (2.21). (iv) to (vi) are part of the definition of a Bregman function. □

Next, we prove the general Bregman measure descendant property for the algorithm defined by (2.26). This is known in the literature as the Fejér monotonicity.

Proposition 2.2.4. *Let f be a Bregman function with zone $S = \mathrm{Int}(\mathrm{dom} f)$, and assume that f is strongly zone consistent with respect to the separating hyperplane $H_i(\alpha, x)$ of C_i such that $H_i(\alpha, x) \cap \bar{S} \neq \emptyset$ for $\alpha \in (0, 1]$ and $x \in S$. Then the Bregman measure for the algorithm defined by (2.26) decreases towards $C \cap \bar{S}$ with respect to the second variable; that is,*

$$D_f(x, x^{k+1}) \leq D_f(x, x^k) \tag{2.28}$$

for every $x \in C \cap \bar{S}$. Moreover, the sequence $\{x^k\}$ is bounded.

Proof. By the definition of D_f, for every $x \in C$,

$$D_f(x, x^{k+1}) = f(x) - f(x^{k+1}) - \langle \nabla f(x^{k+1}), x - x^{k+1} \rangle.$$

Using the definition of the algorithm in (2.26) and condition (2.27), the last equality becomes

$$D_f(x, x^{k+1}) = \sum_{i=1}^{m} \lambda_i \left(f(x) - f(x^{k+1}) - \langle \nabla f(P_{H_i(\alpha, x^k)}(x^k)), x - x^{k+1} \rangle \right)$$

$$\begin{aligned} = \sum_{i=1}^{m} \lambda_i \big(& f(x) - f(P_{H_i(\alpha,x^k)}(x^k)) + f(P_{H_i(\alpha,x^k)}(x^k)) - f(x^{k+1}) \\ & - \langle \nabla f(P_{H_i(\alpha,x^k)}(x^k)), x - P_{H_i(\alpha,x^k)}(x^k) + P_{H_i(\alpha,x^k)}(x^k) - x^{k+1} \rangle \big) . \end{aligned}$$

Thus, using the definition of D_f again, we have

$$D_f(x, x^{k+1}) = \sum_{i=1}^{m} \lambda_i \left(D_f(x, P_{H_i(\alpha,x^k)}(x^k)) - D_f(x^{k+1}, P_{H_i(\alpha,x^k)}(x^k)) \right) .$$

Therefore, using (2.27), we have

$$D_f(x, x^k) - D_f(x, x^{k+1}) = \sum_{i=1}^{m} \lambda_i D_f(x^{k+1}, P_{H_i(\alpha, x^k)}(x^k)) \tag{2.29}$$

$$+ \sum_{i=1}^{m} \lambda_i \left(D_f(x, x^k) - D_f(x, P_{H_i(\alpha, x^k)}(x^k)) \right). \tag{2.30}$$

But, by Theorem 2.0.7,

$$D_f(P_{H_i(\alpha, x^k)}(x^k), x^k) \leq D_f(x, x^k) - D_f(x, P_{H_i(\alpha, x^k)}(x^k))$$

for every $x \in C \cap \bar{S}$ and so (2.29) becomes

$$D_f(x, x^k) - D_f(x, x^{k+1}) \geq \sum_{i=1}^{m} \lambda_i (D_f(x^{k+1}, P_{H_i(\alpha, x^k)}(x^k)) + D_f(P_{H_i(\alpha, x^k)}(x^k), x^k)).$$

Hence, since the weights λ_i are positive,

$$D_f(x, x^k) - D_f(x, x^{k+1}) \geq \sum_{i=1}^{m} \lambda_i D_f(P_{H_i(\alpha, x^k)}(x^k), x^k) \tag{2.31}$$

and (2.28) follows since D_f is nonnegative. In the literature, this means that the sequence $\{x^k\}$ is D_f-Fejér monotone with respect to $C \cap \bar{S}$ and implies that $\{x^k\}$ is bounded. This is because a repeated use of (2.28) gives $x^k \in L(x, \alpha_0)$ for all $k \geq 0$ with $\alpha_0 = D_f(x, x^0)$ and so $\{x^k\}$ is bounded by the definition of the Bregman function (see Properties 2.2.3(iv)). □

Several important consequences can be deduced from Proposition 2.2.4. We begin with the next corrollary.

Corollary 2.2.5. *The sequence generated by the algorithm defined by (2.26) is such that for every $y \in C$*

$$\lim_{k \to \infty} D_f(y, x^k) = \theta,$$

for some nonnegative θ and for $i \in I := \{1, \ldots, m\}$,

$$\lim_{k \to \infty} D_f(P_{H_i(\alpha, x^k)}(x^k), x^k) = 0. \tag{2.32}$$

Proof. By (2.28), for every $y \in C \cap \bar{S}$, the sequence $\{D_f(y, x^k)\}$ is monotonically decreasing, bounded below by zero and therefore convergent. Hence, there exists a nonnegative θ such that $\lim_{k\to\infty} D_f(y, x^k) = \theta$. Also, as observed in the proof of Proposition 2.2.4, the sequence $\{x^k\}$ is bounded and so the left hand side of (2.31) tends to zero. Therefore $\lim_{k\to\infty} D_f(P_{H_i(\alpha,x^k)}(x^k), x^k) = 0$ since the λ_i's are positive and bounded away from zero. □

Since the sequence generated by the algorithm defined by (2.26) is bounded, it has a limit point. In the next proposition, we show that if a limit point exists and it belongs to C then it is the limit of the sequence.

Proposition 2.2.6. *Any limit point $x^* \in C$ of the bounded sequence $\{x^k\}$ of the algorithm defined by (2.26) is the limit of the entire sequence.*

Proof. Suppose $x^* \in C$ is a limit point of $\{x^k\}$ and that x^{**} is another limit point of $\{x^k\}$. That is

$$\lim_{k\in N_1;k\to\infty} x^k = x^* \text{ and } \lim_{k\in N_2;k\to\infty} x^k = x^{**},$$

where N_1 and N_2 are two different infinite subsets of $N = \{0, 1, 2, \ldots, \}$. Then, since f is zone consistent with respect to the C_i's and $x^k \in S$ for all $k \geq 0$, $x^* \in \bar{S}$. Therefore, from (2.28), $\lim_{k\to\infty} D_f(x^*, x^k)$ exists. Applying Properties 2.2.3 (v) in the definition of Bregman function to the sequence $\{x^k\}_{k\geq 0, k\in N_1}$, we have $\lim_{k\in N_1;k\to\infty} D_f(x^*, x^k) = 0$ and so by (2.28) $\lim_{k\to\infty} D_f(x^*, x^k) = 0$ for all $k \geq 0$ which also holds for the sequence $\{x^k\}_{k\geq 0, k\in N_2}$. Therefore using Properties 2.2.3 (vi), we have $x^* = x^{**}$. □

Before the next step for the convergence proof, we need a condition that must be satisfied by the separating hyperplanes in order to guarantee convergence. This condition induces the following definition.

Definition 2.2.7. *For a given closed convex set C and a point $x \notin C$ we say that H is a δ separating hyperplane for C and x if H lies between C and $B(x, \delta d(x, P_C x))$, where $d(\cdot, \cdot)$ stands for the Euclidean distance and $\delta \in (0, 1)$ and $B(x, r)$ is the ball with center at x and radius r.* ♢

It is clear that this definition is equivalent to saying that for every $x \in \mathbb{R}^n$,

$$\|P_H(x) - x\| \geq \delta \|P_C(x) - x\|, \tag{2.33}$$

where $P_C(x)$ and $P_H(x)$ are the Bregman projections of x onto C and the separating hyperplane H respectively. In other words, if the Euclidean distance between x and its projection onto the hyperplane tends to zero, the same is valid for the Euclidean distance between x and its projection onto the associated convex set.

Therefore our assumption for the separating hyperplanes $H_i(\alpha, x)$ in (2.26) for $i = 1, \ldots, m$ will be

Assumption 2.2.8. *For the algorithm defined by (2.26), and for $i \in I := 1, \ldots, m$ and for all $x \in S$ which are not in C_i, the inequality (2.33) holds with $H_i(\alpha, x)$ defined in (2.20) with respect to C_i for $\alpha \in (0, 1)$.* ♢

Given that the algorithm defined by (2.26) satisfies Assumption 2.2.8, we can prove that every limit point of the sequence $\{x^k\}$ generated by (2.26) belong to C. The following theorem justifies this statement.

Theorem 2.2.9. *The whole sequence of the algorithm defined by (2.26) converges to a point in C.*

Proof. Using (2.32) and (2.33), we have

$$\lim_{k\to\infty} D_f(P_{C_i}(x^k), x^k) = 0, \ \forall i \in I := \{1, \ldots, m\}. \tag{2.34}$$

Suppose x^* is a limit point of the sequence $\{x^k\}$. The limit point x^* exists because $\{x^k\}$ is bounded by Proposition 2.2.4. This means that there exists a subsequence $\{x^{k_l}\}$ such that $\lim_{l\to\infty} x^{k_l} = x^*$, and using (2.34)

$$\lim_{l\to\infty} D_f(P_{C_i}(x^{k_l}), x^{k_l}) = 0 \text{ for all } i \in I. \tag{2.35}$$

Now, by Theorem 2.0.7,

$$D_f(x, P_{H_i(\alpha,x^k)}(x^k)) \leq D_f(x, x^k) - D_f(P_{H_i(\alpha,x^k)}(x^k), x^k)$$

for every $x \in C \cap \bar{S}$. But by (2.32) and (2.28), $\{D_f(P_{H_i(\alpha,x^k)}(x^k), x^k)\}$ and $\{D_f(x, x^k)\}$ are bounded for any $x \in C \cap \bar{S}$ and so $\{D_f(x, P_{H_i(\alpha,x^k)}(x^k))\}$ is

bounded. Therefore by Properties 2.2.3 (iv), $\{P_{H_i(\alpha,x^k)}(x^k)\}$ is bounded. The boundedness of $\{P_{C_i}(x^k)\}$ follows from Assumption 2.2.8. Therefore, using Properties 2.2.3 (vi), $\lim_{l\to\infty} P_{C_i}(x^{k_l}) = x^*$ for each $i \in I$. Hence $x^* \in C$. □

2.2.2 A general underrelaxed entropy projection method

In this section, we derive a general block iterative algorithm with underrelaxed entropy projections using projections onto separating hyperplanes. We derive this algorithm for the solution of a linear system of equations $Ax = b$.

The $x \log x$ entropy function is a Bregman function with zone $S = \text{Int}\mathbb{R}^n_+$, see page 33 of [32]. Therefore, by Lemma 1.6.10, the Bregman projection $P_{H_i(\alpha,x)}(x)$ of x onto $H(\alpha, x)$ satisfies the equation

$$\nabla f(P_{H_i(\alpha,x)}(x)) = \nabla f(x) + \theta(\nabla f(x) - \nabla f(P_{C_i}(x))), \tag{2.36}$$

where θ is the parameter associated with the projection of x onto $H(\alpha, x)$. For $f(x) = \sum_{j=1}^n (x_j \log x_j)$, the gradient of the jth component is $\nabla f(x)_j = 1 + \log x_j$. Thus using (2.36), we have

$$1 + \log(P_{H_i(\alpha,x^k)}(x^k))_j = 1 + \log x_j^k + \theta_i^k(\log x_j^k - \log(P_{C_i}(x^k))_j)$$

which simplifies to

$$\log(P_{H_i(\alpha,x^k)}(x^k))_j = \log x_j^k + \theta_i^k \log\left(\frac{x_j^k}{(P_{C_i}(x^k))_j}\right) \tag{2.37}$$

and by the definition of the general Bregman method given in (2.26) for $p = 1$,

$$\begin{aligned}
\nabla f(x^{k+1})_j &= \sum_{i=1}^m \lambda_i \nabla f(P_{H_i(\alpha,x^k)}(x^k))_j \\
1 + \log(x^{k+1})_j &= \sum_{i=1}^m \lambda_i (1 + \log(P_{H_i(\alpha,x^k)}(x^k))_j \\
\log(x^{k+1})_j &= \sum_{i=1}^m \lambda_i \log(P_{H_i(\alpha,x^k)}(x^k))_j \text{ for } j = 1, \ldots, n, \ k \geq 0.
\end{aligned}$$

Therefore, using (2.37), we have

$$\begin{aligned}\log x_j^{k+1} &= \sum_{i=1}^{m} \lambda_i \left(\log x_j^k + \theta_i^k \log \left(\frac{x_j^k}{(P_{C_i}(x^k))_j} \right) \right) \\ &= \log x_j^k + \sum_{i=1}^{m} \log \left(\frac{x_j^k}{(P_{C_i}(x^k))_j} \right)^{\lambda_i \theta_i^k}.\end{aligned}$$

Therefore, the iterative step becomes

$$x_j^{k+1} = x_j^k \prod_{i=1}^{m} \left(\frac{x_j^k}{(P_{C_i}(x^k))_j} \right)^{\lambda_i \theta_i^k}. \tag{2.38}$$

But $H_i = \{x \mid \langle a^i, x \rangle = b_i\}$, $a^i \neq 0$, for $i = 1, 2, \ldots, m$, and the jth component of the normal a_j^i is

$$\begin{aligned}a_j^i &= \nabla f(x^k)_j - \nabla f(P_{C_i}(x^k))_j = 1 + \log x_j^k - (1 + \log(P_{C_i}(x^k))_j) \\ &= \log \left(\frac{x_j^k}{(P_{C_i}(x^k))_j} \right).\end{aligned}$$

This implies that

$$\exp a_j^i = \frac{x_j^k}{(P_{C_i}(x^k))_j}.$$

Therefore, using (2.38), the iterative method becomes

$$x_j^{k+1} = x_j^k \prod_{i=1}^{m} (\exp a_j^i)^{\lambda_i \theta_i^k} = x_j^k \prod_{i=1}^{m} \exp \left(a_j^i \lambda_i \theta_i^k \right) \text{ for } j = 1, \ldots, n, \ k \geq 0$$

and if we replace the θ_i^k's with $d_i^k = \log \frac{b_i}{\langle a^i, x^k \rangle}$ for all i and $k \geq 0$ then the resulting formula resembles the iterative step formula of the block-iterative MART algorithm of Censor and Segman [31]. However if one replaces the θ_i^k's with $c_k = \alpha \log \frac{b_i}{\langle a^i, x^k \rangle}$, $0 < \alpha \leq 1$, for all i and $k \geq 0$, then the resulting formula resembles the iterative step formula of the underrelaxed MART algorithm in [23].

2.3 An application for general convex sets

When an exact Bregman projection onto a closed convex set is too costly to compute, we can use the algorithm defined by (2.26)when $p = 1$ and the separating

hyperplane defined by the approximation of the function that defines the convex set. That is, suppose that the convex set C_i is defined by

$$C_i = \{x \in \mathbb{R}^n \mid g_i(x) \leq 0\}, \tag{2.39}$$

where $g_i : \mathbb{R}^n \to \mathbb{R}$ is a differentiable convex function. We follow closely the definition of the approximating hyperplane on page 151 of [61], i.e., a separating hyperplane, $H_i(\alpha, x^k)$ for $\alpha \in (0,1)$ and $x^k \in S = \text{Int}(\text{dom}g)$, between the current iterate x^k and the current constraint C_i if $x^k \notin C_i$ as

$$H_i(\alpha, x^k) = \{x \in \mathbb{R}^n \mid \alpha g_i(x^k) + \nabla g_i(x^k)^T(x - x^k) = 0\}, \tag{2.40}$$

and $H_i(\alpha, x^k) = \mathbb{R}^n$ if $x^k \in C_i$. We assume $C = \cap_{i=1}^m C_i \neq \emptyset$.

For the algorithm defined by (2.26), $P_{H_i(\alpha,x^k)}(x^k)$ is the Bregman projection of x^k onto $H_i(\alpha, x^k)$. This means that for a function $f \in B(S)$ with $S = \text{Int}(\text{dom}f)$ which is strongly zone consistent with respect to the hyperplane H_i, the projection solves the equations given by

$$\nabla f(P_{H_i(\alpha,x^k)}(x^k)) = \nabla f(x^k) + \theta_i^k \nabla g_i(x^k), \tag{2.41}$$

$$\alpha g_i(x^k) + \nabla g_i(x^k)^T(P_{H_i(\alpha,x^k)}(x^k) - x^k) = 0 \tag{2.42}$$

for the projection parameter θ_i^k and $P_{H_i(\alpha,x^k)}(x^k)$.

Now from (2.41), we have

$$\langle \nabla f(P_{H_i(\alpha,x^k)}(x^k)) - \nabla f(x^k), P_{H_i(\alpha,x^k)}(x^k) - x^k \rangle = \theta_i^k \langle \nabla g_i(x^k), P_{H_i(\alpha,x^k)}(x^k) - x^k \rangle.$$

Therefore, using (2.42), we have

$$\langle \nabla f(P_{H_i(\alpha,x^k)}(x^k)) - \nabla f(x^k), P_{H_i(\alpha,x^k)}(x^k) - x^k \rangle = -\alpha \theta_i^k g_i(x^k).$$

Thus if $x^k \notin C_i$ or $g_i(x^k) > 0$ then

$$\theta_i^k = \frac{\langle \nabla f(P_{H_i(\alpha,x^k)}(x^k)) - \nabla f(x^k), P_{H_i(\alpha,x^k)}(x^k) - x^k \rangle}{-\alpha g_i(x^k)}$$

and (2.41) becomes

$$\nabla f(P_{H_i(\alpha,x^k)}(x^k)) = \nabla f(x^k) - \frac{\langle \nabla f(P_{H_i(\alpha,x^k)}(x^k)) - \nabla f(x^k), P_{H_i(\alpha,x^k)}(x^k) - x^k \rangle}{\alpha g_i(x^k)} \nabla g_i(x^k) \tag{2.43}$$

and if $x^k \in C_i$ or $g_i(x^k) \leq 0$ then $\nabla f(P_{H_i(\alpha,x^k)}(x^k)) = \nabla f(x^k)$.

Observe that if $\nabla g_i(x^k) = 0$ then g_i takes its minimal value at x^k, implying by the non-emptyness of C that $g_i(x^k) \leq 0$, so that $P_{H_i(\alpha,x^k)}(x^k) = x^k$ and so $\nabla f(P_{H_i(\alpha,x^k)}(x^k)) = \nabla f(x^k)$.

In order that the algorithm defined by (2.26) converges to $x^* \in C$ for the general convex set, we make the following assumption on the gradient of the function g_i for $i = 1, \ldots, m$. With this assumption, i.e., Assumption 2.3.1, Proposition 2.3.2 guarantees that $x^* \in C$ and ensures that Assumption 2.2.8 is satisfied.

Assumption 2.3.1. *For each of the functions $\{g_i\}$ used in (2.39), there exists $\epsilon > 0$ such that $\|\nabla g_i(x^k)\| \geq \epsilon$.*

Proposition 2.3.2. *On the basis of Assumption 2.3.1, $\lim_{k\to\infty} g_i(x^k) = 0$ and $\lim_{k\to\infty} x^k \in C_i$ for each $i \in \{1, \ldots, m\}$.*

Proof. If $\|\nabla g_i(x^k)\|$ is bounded away from zero as in Assumption 2.3.1 then (2.32) implies that $P_{H_i(\alpha,x^k)}(x^k) - x^k$ tends to zero as $k \to \infty$, and so (2.42) implies that $\alpha \lim_{k\to\infty} g_i(x^k) = 0$. Therefore, $\lim_{k\to\infty} x^k \in C_i$ for each i since the g_i's are continuous and $\alpha \in (0, 1)$. This further implies that $\lim_{k\to\infty}(P_{C_i}(x^k) - x^k) = 0$ which satisfies condition (2.33). □

Remark 2.3.3. It must be noted that, for the general convex sets, Assumption 2.3.1 is sufficient for the proof of Theorem 2.2.9. Assumption 2.2.8 is thus redundant.

The proof of Theorem 2.2.9 may therefore take the following simple form:

Proof. By Proposition 2.2.4, $\{x^k\}$ is bounded and so there exists a subsequence $\{x^{k_l}\}$ such that $\lim_{l\to\infty} x^{k_l} = x^*$. Therefore, by Proposition 2.3.2, $x^* \in C_i$ for each $i \in \{1, \ldots, m\}$. Thus $x^* \in C$ and hence by Proposition 2.2.6, x^* is the limit of the entire sequence. □

2.4 Linear equality constraints

In the case of linear equality constraints, given by a linear system of equations

$$Ax = b,$$

where $A \in \mathbb{R}^{m\times n}$, $x \in \mathbb{R}^n$ and $b \in \mathbb{R}^m$, if for some function $f \in B(S)$ which is strongly zone consistent with respect to the hyperplane $H_i := \{x \in \mathbb{R}^n \mid \langle a^i, x\rangle = b_i\}$, $a^i \neq 0$, for each $i \in \{1, \ldots, m\}$, the starting point of the algorithm defined by (2.26) is given by

$$\nabla f(x^0) = -A^T z^0$$

for a given $z^0 \in \mathbb{R}^m_+$ (observe that the obvious choice is $z^0 = 0$ and x^0 would be the unconstrained minimum of f) then we have the iterative formula given by

$$\nabla f(x^{k+1}) = \sum_{i=1}^{m} \lambda_i \nabla f(P_{H_i(\alpha,x^k)}(x^k))$$

where, from (1.34), $\nabla f(x^k) = -A^T z^k$ for $k \geq 0$ and so, in the limit, say x^*, $\nabla f(x^*) = -A^T z^*$, for some z^*. Therefore the Kuhn-Tucker conditions are satisfied for the optimization problem

$$\min f(x) \text{ subject to } Ax = b. \tag{2.44}$$

In this case, algorithm (2.26) solves (2.44) when $p = 1$.

2.5 A Conjecture for the strongly underrelaxed case

In the general case, where the constraints are not linear equalities, just the sequence of Bregman projections defined by algorithm (2.26) does not guarantee convergence to the solution of a minimization problem, say

$$\min f(x) \text{ subject to } \langle a^i, x\rangle \leq b_i$$

for the function $f \in B(S)$ with $S = \text{Intdom} f$ and $i = 1, \ldots, m$, but converges to the solution of only the convex feasibility problem, unless dual variables are

updated as in Chapter 3 for linear inequalities. So, what happens in the general case? Our conjecture is that in the purely sequential Bregman algorithm, when the relaxation parameters tend to zero, the sequence generated by algorithm (2.26) tends to the solution of the optimization problem $\lim_{x\in\mathbb{R}^n} D_f(Ax, b)$ in the general case.

Chapter 3

Block Bregman methods for inequality constraints

In this chapter, we present a new simultaneous version of the Bregman method for linear constraints with corresponding convergence results.

3.1 The problem

We recall the problem of Subsection 1.7.1,

$$\begin{aligned} &\min f(x), \\ &\text{subject to } \langle a^i, x\rangle \leq b_i, \qquad i \in I := \{1, 2, \ldots, m\}, \\ &x \in \bar{S}, \end{aligned}$$

where $H_i := \{x \in \mathbb{R}^n \mid \langle a^i, x\rangle = b_i\}$, $C_i := \{x \in \mathbb{R}^n \mid \langle a^i, x\rangle \leq b_i\}$; $C = \cap_{i=1}^m C_i$ and $C \cap \bar{S} \neq \emptyset$.

A is an $m \times n$ matrix whose ith row is a^i, and $b \in \mathbb{R}^m$, $a^i \neq 0$ for all $i \in I$. $S = \text{Int}(\text{dom} f)$ and $f \in B(S)$ is essentially smooth and strongly zone consistent with respect to every H_i.

3.1.1 Simultaneous under-relaxed Bregman's algorithm for linear inequality constraints

We present the simultaneous version of Algorithm 1.7.2 for solving problem (1.43)-(1.45).

Algorithm 3.1.1. Simultaneous under-relaxed Bregman's algorithm for linear inequalities

(i) **Initialization** $x^0 \in S$ is such that for an arbitrary $z^0 \in \mathbb{R}^m_+$,

$$\nabla f(x^0) = -A^T z^0. \tag{3.1}$$

(ii) **Iterative Step** Given x^k and z^k, calculate x^{k+1} and z^{k+1} from

$$\nabla f(x^{k+1}) = \nabla f(x^k) + \sum_{i \in I_{t(k)}} \lambda_i^k c_i^k a^i, \tag{3.2}$$

$$z^{k+1} = z^k - \sum_{i \in I_{t(k)}} \lambda_i^k c_i^k e^i \tag{3.3}$$

with

$$c_i^k = \begin{cases} \min\left(\dfrac{z_i^k}{\lambda_i^k}, \theta_i^k\right) & \text{if} \quad i \in I_{t(k)}, \\ 0 & \text{if} \quad i \notin I_{t(k)}, \end{cases} \tag{3.4}$$

where $\theta_i^k = \pi_{H(k)_i}(x^k)$, and there exists $\bar{\epsilon} > 0$ such that, for the positive weights λ_i^k with $\sum_{i \in I_{t(k)}} \lambda_i^k = 1$ for all $k \geq 0$, $\lambda_i^k \geq \bar{\epsilon}$ for $i \in I_{t(k)}$ and $\lambda_i^k = 0$ for $i \notin I_{t(k)}$.

(iii) $H(k)_i$ is parallel to $H_i = \{x \in \mathbb{R}^n \mid \langle a^i, x\rangle = b_i\}$ for all $i \in I_{t(k)}$ and $k \geq 0$. $H(k)_i := \{x \in \mathbb{R}^n \mid \langle a^i, x\rangle = \alpha^k b_i + (1-\alpha^k)\langle a^i, x^k\rangle\}$, where $\{\alpha^k\}$ such that $\epsilon \leq \alpha^k \leq 1$ for $\epsilon > 0$ is a sequence of relaxation parameters. We adopt the concept of relaxation highlighted in Subsection 1.7.2 and implemented in [41].

(iv) The sequence $\{t(k)\}_{k=0}^{\infty}$ is almost cyclic on the index set $\{1, 2, \ldots, M\}$, where M is the number of blocks. The index set $I := \{1, 2, \ldots, m\}$ of the m constraints has been partitioned into M nonempty disjoint blocks such that $I = I_1 \cup \ldots \cup I_M$. ♢

In order to justify the proof of Lemma 3.1.4 and the proofs of the propositions leading to the proof of the convergence theorem, Theorem 3.2.1, let P_i and Q_i be Bregman projections onto

$$C(k)_i = \{x \in \mathbb{R}^n \mid \langle a^i, x\rangle \leq \alpha^k b_i + (1-\alpha^k)\langle a^i, x^k\rangle\} \tag{3.5}$$

and $H(k)_i$ respectively so that for any function $f \in B(S)$ which is strongly zone consistent with respect to H_i for $i \in I_{t(k)}$ and $x^k \in S$ with $\theta_i^k = \pi_{H(k)_i}(x^k)$,

$$\nabla f(P_i x^k) = \nabla f(x^k) + \min\{0, \theta_i^k\} a^i, \tag{3.6}$$

$$\nabla f(Q_i x^k) = \nabla f(x^k) + \theta_i^k a^i. \tag{3.7}$$

Define

$$\nabla f(w_i^k) = \nabla f(x^k) + c_i^k a^i \text{ for } i \in I_{t(k)}. \tag{3.8}$$

The w_i^k's are 'modified relaxed Bregman projections' onto the H_i's (see page 233 of [62]) such that the next iterate x^{k+1} satisfies the equation $\nabla f(x^{k+1}) = \sum_{i\in I_{t(k)}} \lambda_i^k \nabla f(w_i^k)$.

This is because, multiplying (3.8) by λ_i^k for $i \in I_{t(k)}$ with $\sum_{i\in I_{t(k)}} \lambda_i^k = 1$ and summing over $i \in I_{t(k)}$, we have

$$\sum_{i\in I_{t(k)}} \lambda_i^k \nabla f(w_i^k) = \nabla f(x^k) + \sum_{i\in I_{t(k)}} \lambda_i^k c_i^k a^i.$$

Therefore, from iterative step (3.2),

$$\nabla f(x^{k+1}) = \sum_{i\in I_{t(k)}} \lambda_i^k \nabla f(w_i^k). \tag{3.9}$$

Since the applicability of Algorithm 3.1.1 depends on the ability to invert the gradient ∇f explicitly, we assume that the Bregman function f used in Algorithm 3.1.1 satisfies Assumption 2.2.2.

3.1.2 Preliminary results

The following lemma establishes the non-negativity of the dual variables.

Lemma 3.1.2. *For any $k \geq 0$, $z_i^k \geq 0$ for $i \in I_{t(k)}$.*

Proof. By initialization, $z^0 \in \mathbb{R}^m_+$. By induction, assume $z^k \in \mathbb{R}^m_+$. If $i \notin I_{t(k)}$ then $z_i^{k+1} = z_i^k \geq 0$ and if $i \in I_{t(k)}$ then using (3.4), $z_i^{k+1} = z_i^k - \lambda_i^k c_i^k \geq z_i^k - z_i^k = 0$. Hence $z^{k+1} \in \mathbb{R}^m_+$. □

Lemma 3.1.3. *$\nabla f(x^k) = -A^T z^k$ for any $k \geq 0$.*

Proof. We proceed by induction. By (3.1), the result is true when $k = 0$. Assume that the result is true for any $k \geq 0$. Then by (3.2) and (3.3),

$$\nabla f(x^{k+1}) = \nabla f(x^k) + \sum_{i \in I_{t(k)}} \lambda_i^k c_i^k a^i \tag{3.10}$$

$$= -A^T(z^k - \sum_{i \in I_{t(k)}} \lambda_i^k c_i^k e^i) = -A^T z^{k+1}. \tag{3.11}$$

□

The next lemma shows that the w_i^k is a Bregman projection of x^k onto $H(k)_i$ if $x^k \notin C(k)_i$ or a point in the segment between x^k and its projection onto $H(k)_i$ if $x^k \in C(k)_i$.

Lemma 3.1.4. *Let the function $f \in B(S)$ be strongly zone consistent with respect to the hyperplane H_i for all $i \in I_{t(k)}$ and $k \geq 0$, and let w_i^k satisfy (3.8). Then for the closed and nonempty convex set C with $C \cap \bar{S} \neq \emptyset$, the following statements hold if f satifies Assumption 2.2.2.*

(a) If $x^k \notin C(k)_i$ then $w_i^k = P_i x^k = Q_i x^k$ and $\theta_i^k = c_i^k < 0$; P_i and Q_i are as defined in (3.6) and (3.7).

(b) If $x^k \in C(k)_i$ then $0 \leq c_i^k \leq \theta_i^k$.

(c) $w_i^k \in C(k)_i$.

(d) $D_f(w_i^k, x^k) \leq D_f(y, x^k) - D_f(y, w_i^k)$ for all $y \in C \cap \bar{S}$.

Proof. (a) If $x^k \notin C(k)_i$ then by Lemma 1.6.11 $\theta_i^k < 0$, and since $z_i^k \geq 0$ by Lemma 3.1.2, $c_i^k = \theta_i^k$ by (3.4). Therefore, from the definitions of w_i^k, P_i and Q_i, we have $w_i^k = P_i x^k = Q_i x^k$.

(b) If $x^k \in C(k)_i$ then $\theta_i^k \geq 0$ and since $z_i^k \geq 0$, $0 \leq c_i^k \leq \theta_i^k$.

(c) Clearly $w_i^k \in C(k)_i$ if $c_i^k = \theta_i^k$. Thus, suppose $c_i^k \neq \theta_i^k$ or $c_i^k = z_i^k/\lambda_i^k$, i.e., $c_i^k < \theta_i^k$, and recall that θ_i^k, the parameter associated with the Bregman projection of x^k onto $H(k)_i$ (in what follows we denote this projection by $\bar{x}^{k+1}$) is obtained by solving the system

$$\nabla f(\bar{x}^{k+1}) = \nabla f(x^k) + \theta_i^k a^i, \tag{3.12}$$

$$\langle a^i, \bar{x}^{k+1} \rangle = \alpha^k b_i + (1 - \alpha^k)\langle a^i, x^k \rangle. \tag{3.13}$$

Now, using (3.8) and (3.12), we have

$$\nabla f(w_i^k) - \nabla f(\bar{x}^{k+1}) = (c_i^k - \theta_i^k)a^i$$

and so

$$\langle \nabla f(w_i^k) - \nabla f(\bar{x}^{k+1}), w_i^k - \bar{x}^{k+1} \rangle = \langle (c_i^k - \theta_i^k)a^i, w_i^k - \bar{x}^{k+1} \rangle.$$

But, by Properties 2.2.3 (iii),

$$\langle \nabla f(w_i^k) - \nabla f(\bar{x}^{k+1}), w_i^k - \bar{x}^{k+1} \rangle = D_f(w_i^k, \bar{x}^{k+1}) + D_f(\bar{x}^{k+1}, w_i^k) \geq 0$$

and since $c_i^k < \theta_i^k$, we have

$$\langle a^i, w_i^k - \bar{x}^{k+1} \rangle \leq 0.$$

Therefore

$$\langle a^i, w_i^k \rangle \leq \alpha^k b_i + (1 - \alpha^k)\langle a^i, x^k \rangle$$

since

$$\langle a^i, \bar{x}^{k+1} \rangle = \alpha^k b_i + (1 - \alpha^k)\langle a^i, x^k \rangle.$$

This means that $w_i^k \in C(k)_i$.

(d) The result follows from Theorem 2.0.7.

□

3.2 Convergence results

Next, is the convergence theorem for Algorithm 3.1.1.

Theorem 3.2.1. *Assume the following:*

(i) $f \in B(S)$,

(ii) f *is strongly zone consistent with respect to each* H_i, $i \in I_{t(k)}$,

(iii) $\{t(k)\}_{k=0}^{\infty}$ *is almost cyclic on* $\{1, 2, \ldots, M\}$ *with a constant of almost cyclicality* r,

(iv) $C \cap \bar{S} \neq \emptyset$.

Then any sequence $\{x^k\}$ *produced by Algorithm 3.1.1 converges to the point* x^*, *which is the solution of (1.43)-(1.45).*

To prove Theorem 3.2.1, we use Propositions 3.2.2 to 3.2.7.

Proposition 3.2.2. *If the assumptions of Theorem 3.2.1 hold and* w_i^k *satisfies (3.8), then*
$w_i^k = P_{\bar{H}(k)_i}(x^k)$ *and* $c_i^k = \pi_{\bar{H}(k)_i}(x^k)$ *where*

$$\bar{H}(k)_i = \{x \in \mathbb{R}^n \mid \langle a^i, x\rangle = \gamma^k b_i + (1-\gamma^k)\langle a^i, x^k\rangle\} \tag{3.14}$$

for some $\gamma^k \in \mathbb{R}$ *such that* $0 \leq \gamma^k \leq \alpha^k$ *for all* $k \geq 0$.

Proof. Note that w_i^k is the Bregman projection of x^k onto the hyperplane $\bar{H}(k)_i = \{\langle a^i, x\rangle = \langle a^i, w_i^k\rangle\}$, which is parallel to $H(k)_i$ and passes through w_i^k. It remains to demonstrate that the right-hand side $\langle a^i, w_i^k\rangle$ has the desired form as in (3.14). It is clear that if $c_i^k = \theta_i^k$ for $i \in I_{t(k)}$ then $\bar{H}(k)_i = H(k)_i$. This is because, by definition, $\bar{H}(k)_i$ is parallel to $H(k)_i$ and lies between x^k and $H(k)_i$, and θ_i^k is the parameter associated with the Bregman projection of x^k onto $H(k)_i$. Therefore $\bar{H}(k)_i$ and $H(k)_i$ coincide if $c_i^k = \theta_i^k$. Hence $\gamma^k = \alpha^k$.

On the other hand, if $c_i^k \neq \theta_i^k$ for $i \in I_{t(k)}$ then, by the definition of c_i^k, $0 \leq \frac{z_i^k}{\lambda_i^k} = c_i^k < \theta_i^k$.

Now consider the following which is possible by Properties 2.2.3 (iii) and (3.8):

$$\begin{aligned} 0 \leq D_f\left(w_i^k, x^k\right) + D_f\left(x^k, w_i^k\right) &= \langle \nabla f\left(w_i^k\right) - \nabla f\left(x^k\right), w_i^k - x^k \rangle, \\ &= \langle c_i^k a^i, w_i^k - x^k \rangle. \end{aligned} \tag{3.15}$$

Therefore if $c_i^k = 0$ for $i \in I_{t(k)}$ then $D_f\left(x^k, w_i^k\right) = 0$ and $D_f\left(w_i^k, x^k\right) = 0$ imply $w_i^k = x^k$, i.e.,$\langle a^i, w_i^k \rangle = \langle a^i, x^k \rangle$ and so we may take $\gamma^k = 0$.

Finally if $0 < c_i^k < \theta_i^k$ then $0 < \pi_{\bar{H}(k)_i}(x^k) < \pi_{H(k)_i}(x^k)$. Hence, by Lemma 1.6.12,

$$\langle a^i, x^k \rangle < \langle a^i, w_i^k \rangle < \alpha^k b_i + (1 - \alpha^k)\langle a^i, x^k \rangle \tag{3.16}$$

for $i \in I_{t(k)}$ and so there is $\gamma^k \in (0, \alpha^k)$ as desired. For more insight, see page 427 of [41] □

Proposition 3.2.3. *If the assumptions of Theorem 3.2.1 hold, λ_i^k and c_i^k are as defined in Algorithm 3.1.1 and w_i^k satisfies (3.8) then for $x^k \in S$,*

$$\sum_{i \in I_{t(k)}} \lambda_i^k \left\{ D_f(x^k, w_i^k) - D_f(x^k, x^{k+1}) \right\}$$

and

$$D_f\left(x^{k+1}, x^k\right) + \sum_{i \in I_{t(k)}} \lambda_i^k c_i^k \left(b_i - \langle a^i, x^{k+1} \rangle\right)$$

are non-negative.

Proof. Using Lemma 2.0.6 with $y = x^k$, $x = w_i^k$ and $z = x^{k+1}$, we have

$$D_f(x^k, w_i^k) = D_f(x^k, x^{k+1}) + D_f(x^{k+1}, w_i^k) - \langle \nabla f(w_i^k) - \nabla f(x^{k+1}), x^k - x^{k+1} \rangle.$$

Now, multiplying the last equation by λ_i^k for $i \in I_{t(k)}$, $\sum_{i \in I_{t(k)}} \lambda_i^k = 1$, and summing over $i \in I_{t(k)}$, we have

$$\sum_{i \in I_{t(k)}} \lambda_i^k D_f(x^k, w_i^k) = D_f(x^k, x^{k+1}) + \sum_{i \in I_{t(k)}} \lambda_i^k D_f(x^{k+1}, w_i^k)$$

since, by (3.9),

$$\begin{aligned}\sum_{i\in I_{t(k)}} \lambda_i^k \langle \nabla f(w_i^k) - \nabla f(x^{k+1}), x^k - x^{k+1}\rangle &= \langle \nabla f(x^{k+1}) - \nabla f(x^{k+1}), x^k - x^{k+1}\rangle \\ &= 0.\end{aligned}$$

Thus

$$\sum_{i\in I_{t(k)}} \lambda_i^k \left\{ D_f(x^k, w_i^k) - D_f(x^k, x^{k+1}) \right\} = \sum_{i\in I_{t(k)}} \lambda_i^k D_f(x^{k+1}, w_i^k) \geq 0. \tag{3.17}$$

Now let

$$\bar{d}_k = D_f\left(x^{k+1}, x^k\right) + \sum_{i\in I_{t(k)}} \lambda_i^k c_i^k \left(b_i - \langle a^i, x^{k+1}\rangle\right). \tag{3.18}$$

Then, using Properties 2.2.3 and (3.2), we have

$$\begin{aligned}\bar{d}_k + D_f\left(x^k, x^{k+1}\right) &= D_f\left(x^k, x^{k+1}\right) + D_f\left(x^{k+1}, x^k\right) \\ &\quad + \sum_{i\in I_{t(k)}} \lambda_i^k c_i^k \left(b_i - \langle a^i, x^{k+1}\rangle\right) \\ &= \langle \nabla f\left(x^{k+1}\right) - \nabla f\left(x^k\right), x^{k+1} - x^k\rangle \\ &\quad + \sum_{i\in I_{t(k)}} \lambda_i^k c_i^k \left(b_i - \langle a^i, x^{k+1}\rangle\right) \\ &= \langle \sum_{i\in I_{t(k)}} \lambda_i^k c_i^k a^i, x^{k+1} - x^k\rangle + \sum_{i\in I_{t(k)}} \lambda_i^k c_i^k \left(b_i - \langle a^i, x^{k+1}\rangle\right) \\ &= \sum_{i\in I_{t(k)}} \lambda_i^k c_i^k \left(b_i - \langle a^i, x^k\rangle\right).\end{aligned}$$

But, by Proposition 3.2.2, $\langle a^i, w_i^k\rangle = \gamma^k b_i + (1-\gamma^k)\langle a^i, x^k\rangle$ implies

$$\langle a^i, w_i^k - x^k\rangle = \gamma^k \left(b_i - \langle a^i, x^k\rangle\right). \tag{3.19}$$

Therefore, from (3.15),

$$0 \leq D_f\left(w_i^k, x^k\right) + D_f\left(x^k, w_i^k\right) = c_i^k \gamma^k \left(b_i - \langle a^i, x^k\rangle\right). \tag{3.20}$$

Therefore, using (3.20), if $\gamma^k = 0$ then $w_i^k = x^k$, and by (3.8), $w_i^k = x^k$ implies $c_i^k = 0$. Therefore by (3.2), $\nabla f(x^{k+1}) = \nabla f(x^k)$ implies $x^{k+1} = x^k$ if f satisfies Assumption 2.2.2. Hence

$$\bar{d}_k = -D_f\left(x^k, x^{k+1}\right) + \sum_{i\in I_{t(k)}} \lambda_i^k c_i^k \left(b_i - \langle a^i, x^k\rangle\right) = 0 \text{ if } \gamma^k = 0.$$

Next we consider when $\gamma^k > 0$ for $k \geq 0$. Using (3.20), we have

$$\bar{d}_k = \sum_{i \in I_{t(k)}} \lambda_i^k \left\{ -D_f\left(x^k, x^{k+1}\right) + \frac{D_f\left(w_i^k, x^k\right) + D_f\left(x^k, w_i^k\right)}{\gamma^k} \right\}$$

and since γ^k is less or equal to one, we have

$$\bar{d}_k \geq \sum_{i \in I_{t(k)}} \lambda_i^k \left\{ D_f\left(w_i^k, x^k\right) + D_f\left(x^k, w_i^k\right) - D_f\left(x^k, x^{k+1}\right) \right\}.$$

Therefore, using (3.17),

$$\bar{d}_k \geq \sum_{i \in I_{t(k)}} \lambda_i^k \left\{ D_f\left(w_i^k, x^k\right) + D_f\left(x^{k+1}, w_i^k\right) \right\}. \tag{3.21}$$

Thus

$$\bar{d}_k \geq 0 \text{ for all } k \geq 0.$$

□

Proposition 3.2.4. *If the Lagrangian of the minimization problem in (1.43)-(1.45) is $L(x,z) = f(x) + \langle z, Ax - b\rangle$ then, for any sequences $\{x^k\}$ and $\{z^k\}$ produced by Algorithm 3.1.1,*

(i) the sequence $\left\{L\left(x^k, z^k\right)\right\}$ is nondecreasing and $\lim_{k\to\infty} L\left(x^k, z^k\right)$ exists,

(ii) $\lim_{k\to\infty} D_f\left(w_i^k, x^k\right) = 0$ and $\lim_{k\to\infty} D_f\left(x^{k+1}, w_i^k\right) = 0$ for each $i \in I_{t(k)}$,

(iii) $\{x^k\}$ is bounded.

Proof. (i) Define

$$d_k = L\left(x^{k+1}, z^{k+1}\right) - L\left(x^k, z^k\right). \tag{3.22}$$

Then, from the definition of L,

$$\begin{aligned} d_k &= f(x^{k+1}) + \langle z^{k+1}, Ax^{k+1} - b\rangle - \left(f(x^k) + \langle z^k, Ax^k - b\rangle\right) \\ &= f(x^{k+1}) - f(x^k) + \langle z^{k+1}, Ax^{k+1}\rangle - \langle z^{k+1}, b\rangle - \langle z^k, Ax^k\rangle + \langle z^k, b\rangle. \end{aligned}$$

But $\langle z^k, Ax^k\rangle = \langle A^T z^k, x^k\rangle = -\langle \nabla f(x^k), x^k\rangle$.

Therefore

$$\begin{aligned} d_k &= f(x^{k+1}) - f(x^k) - \langle \nabla f(x^k), x^{k+1} - x^k\rangle + \langle \nabla f(x^k), x^{k+1} - x^k\rangle \\ &\quad -\langle z^{k+1} - z^k, b\rangle - \langle \nabla f(x^{k+1}), x^{k+1}\rangle + \langle \nabla f(x^k), x^k\rangle \\ &= D_f\left(x^{k+1}, x^k\right) - \langle z^{k+1} - z^k, b\rangle + \langle \nabla f(x^k) - \nabla f(x^{k+1}), x^{k+1}\rangle \\ &= D_f\left(x^{k+1}, x^k\right) - \langle \sum_{i\in I_{t(k)}} \lambda_i^k c_i^k a^i, x^{k+1}\rangle + \langle \sum_{i\in I_{t(k)}} \lambda_i^k c_i^k e^i, b\rangle \\ &= D_f\left(x^{k+1}, x^k\right) - \langle \sum_{i\in I_{t(k)}} \lambda_i^k c_i^k a^i, x^{k+1}\rangle + \sum_{i\in I_{t(k)}} \lambda_i^k c_i^k b_i \\ &= D_f\left(x^{k+1}, x^k\right) + \sum_{i\in I_{t(k)}} \lambda_i^k c_i^k \left(b_i - \langle a^i, x^{k+1}\rangle\right). \end{aligned}$$

Hence, by Proposition 3.2.3 and (3.18), $d_k = \bar{d}_k \geq 0$ and so $\left\{L(x^k, z^k)\right\}$ is non-decreasing.

We now prove the existence of $\lim_{k\to\infty} L\left(x^k, z^k\right)$ by showing that $\left\{L\left(x^k, z^k\right)\right\}$ is bounded from above on $C \cap \bar{S}$ for all $k \geq 0$.

To do this, we choose $z \in C \cap \bar{S}$ and consider:

$$\begin{aligned} D_f(z, x^k) &= f(z) - f(x^k) - \langle \nabla f(x^k), z - x^k\rangle \\ &= f(z) - f(x^k) + \langle A^T z^k, z - x^k\rangle \\ &= f(z) - f(x^k) + \langle z^k, Az\rangle - \langle z^k, Ax^k\rangle \\ &\leq f(z) - f(x^k) + \langle z^k, b - Ax^k\rangle = f(z) - L\left(x^k, z^k\right). \end{aligned}$$

Therefore

$$L\left(x^k, z^k\right) \leq f(z) - D_f(z, x^k) \leq f(z).$$

Hence $\left\{L\left(x^k, z^k\right)\right\}$ is bounded from above and so the limit exists.

(ii) Since $\lim_{k\to\infty} L\left(x^k, z^k\right)$ exists, by (3.22), $\lim_{k\to\infty} d_k = \lim_{k\to\infty} \bar{d}_k = 0$. Therefore, from (3.21),

$$\lim_{k\to\infty} \sum_{i\in I_{t(k)}} \lambda_i^k \left\{D_f\left(w_i^k, x^k\right) + D_f\left(x^{k+1}, w_i^k\right)\right\} = 0 \tag{3.23}$$

and since $\lambda_i^k \geq \bar{\epsilon} > 0$ for $i \in I_{t(k)}$ and $k \geq 0$, we have

$$\lim_{k\to\infty} D_f\left(w_i^k, x^k\right) = 0 \text{ and } \lim_{k\to\infty} D_f\left(x^{k+1}, w_i^k\right) = 0$$

for each $i \in I_{t(k)}$ and $k \geq 0$.

(iii) Applying $d_k = L\left(x^{k+1}, z^{k+1}\right) - L\left(x^k, z^k\right) \geq 0$ for all $k \geq 0$ recursively to the expression $D_f(z, x^k) \leq f(z) - L\left(x^k, z^k\right)$, we have

$$D_f(z, x^k) \leq f(z) - L\left(x^0, z^0\right) = \alpha,$$

where $\alpha \in \mathbb{R}$. Therefore, from the boundedness of the partial level set $L_2^f(z, \alpha)$, Properties 2.2.3 (iv), we have that $\left\{x^k\right\}$ is bounded.

□

Proposition 3.2.5. *Assume that* $\lim_{j\to\infty} x^{k_j} = x^*$. *Fix* q, *a positive integer, and take a sequence* $\{l_j\}$ *with* $l_j \in \{1, 2, \ldots, q\}$. *Then* $\lim_{j\to\infty} w_i^{k_j+l} = x^*$ *for* $i \in I_{t(k_j)}$ *and* $\lim_{j\to\infty} x^{k_j+l_j} = x^*$.

Proof. Consider first the q sequences $\left\{x^{k_j+l}\right\}_{j=0}^{\infty}$ with $1 \leq l \leq q$. Since $\left\{x^{k_j+l}\right\}_{j=0}^{\infty}$ are sub-sequences of the bounded sequence $\{x^k\}_{k=0}^{\infty}$, they are bounded, and from Proposition 3.2.4 (ii), for all $s \in \{0, 1, \ldots, q\}$, $i \in I_{t(k_j)}$ and $j \geq 0$,

$$\lim_{j\to\infty} D_f\left(w_i^{k_j+s}, x^{k_j+s}\right) = 0 \text{ and } \lim_{j\to\infty} D_f\left(x^{k_j+s+1}, w_i^{k_j+s}\right) = 0. \tag{3.24}$$

From Proposition 3.2.4 (ii) and (iii), $\{D_f(w_i^k, x^k)\}$ is bounded and $\{D_f(x, x^k)\}$ for each $x \in C \cap \bar{S}$ is bounded . Therefore by Lemma 3.1.4 (d), $\{D_f(x, w_i^k)\}$ is bounded for $x \in C \cap \bar{S}$. This means that, by Properties 2.2.3 (iv), $\{w_i^k\}$ is bounded. Therefore, using the first equation on the left hand side of (3.24) and applying recursively Properties 2.2.3 (vi), we have, for all l such that $0 \leq l \leq q$ and $i \in I_{t(k_j)}$,

$$\lim_{j\to\infty} w_i^{k_j+l} = x^*. \tag{3.25}$$

Hence, using the second equation on the right hand side of (3.24) and applying recursively Properties 2.2.3 (vi), we have that, for all l such that $0 \leq l \leq q$,

$$\lim_{j\to\infty} x^{k_j+l} = x^*.$$

Thus, interlacing these $q+1$ sequences, we can form the sequences

$$\left\{w_i^{k_1}, w_i^{k_1+1}, \ldots, w_i^{k_1+q}, w_i^{k_2}, w_i^{k_2+1}, \ldots, w_i^{k_2+q}, \ldots, w_i^{k_j}, w_i^{k_j+1} \ldots, w_i^{k_j+q}, \ldots\right\}$$

and

$$\left\{x^{k_1}, x^{k_1+1}, \ldots, x^{k_1+q}, x^{k_2}, x^{k_2+1}, \ldots, x^{k_2+q}, \ldots, x^{k_j}, x^{k_j+1} \ldots, x^{k_j+q}, \ldots\right\}$$

both converging to x^* with $\{w_i^{k_j+l_j}\}$ and $\{x^{k_j+l_j}\}$ as their respective sub-sequences.

□

Proposition 3.2.6. *All limit points of the sequence $\{x^k\}$ produced by Algorithm 3.1.1 belong to $C \cap \bar{S}$.*

Proof. Let $p \in \{1, 2, \ldots, M\}$ and $l_j \in \{1, 2, \ldots, r\}$, where r is the constant of almost cyclicality of $\{t(k_j)\}_{j=0}^{\infty}$, such that $t(k_j + l_j) = p$. Then, with $q = r$ as in Proposition 3.2.5 and with $i \in I_{t(k_j+l_j)} = I_p$,

$$\lim_{j\to\infty} w_i^{k_j+l_j} = x^* \text{ and } \lim_{j\to\infty} x^{k_j+l_j} = x^*$$

hold and so we can extract sub-sequences $\{x^{s_j}\}_{j=0}^{\infty}$ of $\left\{x^{k_j+l_j}\right\}_{j=0}^{\infty}$ such that $x^{s_j} \to x^*$, $\gamma^{s_j} \to \gamma$, $\alpha^{s_j} \to \alpha \geq \epsilon$, and by Lemma 3.1.4 (c) and Proposition 3.2.2, $x^* \in C(s_j)_i$ as $j \to \infty$. Therefore, it follows that for $i \in I_{t(s_j)}$,

$$\langle a^i, x^*\rangle = \gamma b_i + (1-\gamma)\langle a^i, x^*\rangle \text{ or } \gamma\left(\langle a^i, x^*\rangle - b_i\right) = 0. \tag{3.26}$$

Therefore if $\gamma \neq 0$ then $\langle a^i, x^*\rangle = b_i$ and if $\gamma^{s_j} \to \gamma = 0$ then (3.26) is inconclusive, and moreover $\gamma \neq \alpha$ since $\alpha^{s_j} \to \alpha \geq \epsilon > 0$, and so by Proposition 3.2.2, precisely (3.16), we have

$$\langle a^i, x^*\rangle \leq \alpha b_i + (1-\alpha)\langle a^i, x^*\rangle \text{ or } 0 \leq \alpha\left(b_i - \langle a^i, x^*\rangle\right)$$

for $i \in I_p$. Therefore $\langle a^i, x^*\rangle \leq b_i$ for $i \in I_p$. Hence, since p is an arbitrary index, we have $Ax^* \leq b$ and so $x^* \in C$. □

Proposition 3.2.7. *For $x \in C$, define $I_1^{t(k)}(x) := \{i \in I_{t(k)} : \langle a^i, x\rangle < b_i\}$ and $I_2^{t(k)}(x) := \{i \in I_{t(k)} : \langle a^i, x\rangle = b_i\}$ for all $k \geq 0$ and assume that $\lim_{j\to\infty} x^{k_j} = x^*$. Then $z_i^{k_j+r+1} = 0$ for all $i \in I_1^{t(k_j)}(x^*)$ and $j \geq 0$, where r is the constant of almost cyclicality.*

Proof. Let

$$\delta = \frac{\epsilon}{4} \min_{i \in I_1^{t(k_j)}(x^*)} \left\{ \frac{(b_i - \langle a^i, x^* \rangle)}{\|a^i\|} \right\} > 0.$$

By Proposition 3.2.5, there exists a natural number J such that

$$\|w_i^{k_j+l} - x^*\| < \delta \text{ for all } l \in \{0, \ldots, r+1\}, \text{ and } i \in I_1^{(t(k_j)+l)}(x^*) \text{ for all } j \geq J. \tag{3.27}$$

Define $l_j = \max_{0 \leq l \leq r} \{l : t(k_j + l) = p\}$. The existence of l_j is guaranteed by almost cyclicality of the control sequence. Let $s_j = k_j + l_j$ and assume that $c_i^{s_j} = \theta_i^{s_j}$. Then

$$\langle a^i, w_i^{s_j} \rangle = \alpha^{s_j} b_i + (1 - \alpha^{s_j}) \langle a^i, x^{s_j} \rangle$$

i.e.,

$$\langle a^i, w_i^{s_j} - x^{s_j} \rangle = \alpha^{s_j} \left(b_i - \langle a^i, x^{s_j} \rangle \right).$$

Therefore

$$\alpha_i^{s_j} \left(b_i - \langle a^i, x^* \rangle \right) = \langle a^i, w_i^{s_j} - x^{s_j} \rangle + \alpha^{s_j} \langle a^i, x^{s_j} - x^* \rangle.$$

Thus, using (3.27) and the fact that $0 < \epsilon \leq \alpha^{s_j} \leq 1$, we have the following inequalities for all $j \geq J$ and $i \in I_1^{t(s_j)}(x^*)$:

$$\begin{aligned}
\epsilon \left(b_i - \langle a^i, x^* \rangle \right) &\leq \langle a^i, w_i^{s_j} - x^{s_j} \rangle + \alpha^{s_j} \langle a^i, x^{s_j} - x^* \rangle \\
&\leq \|a^i\| \left(\|w_i^{s_j} - x^{s_j}\| + \alpha^{s_j} \|x^{s_j} - x^*\| \right) \\
&\leq \|a^i\| \left(\|w_i^{s_j} - x^*\| + \|x^{s_j} - x^*\| + \alpha^{s_j} \|x^{s_j} - x^*\| \right) \\
&< \|a^i\| (2 + \alpha^{s_j}) \delta \leq 3\delta \|a^i\|.
\end{aligned}$$

Hence we have the contradiction $3\delta > \epsilon \{(b_i - \langle a^i, x^* \rangle)/\|a^i\|\} \geq 4\delta$ for all $i \in I_1^{t(s_j)}(x^*)$. It therefore follows that $c_i^{s_j} \neq \theta_i^{s_j}$ and so $c_i^{s_j} = \frac{z_i^{s_j}}{\lambda_i^{s_j}}$ implies $z_i^{s_j+1} = 0$ for $i \in I_1^{t(s_j)}(x^*) = I_1^p(x^*)$. By the definition of l_j, the index p is not used in iteration $k_j + l$ for $l_j < l \leq r$ and so $z_i^{s_j+l}$ for $i \in I_1^p(x^*)$ remains unaffected. We conclude that $z_i^{s_j+r+1} = 0$ for $i \in I_1^p(x^*)$ with $p \in \{1, \ldots, M\}$. □

Now the proof of Theorem 3.2.1 follows:

Proof. Since $z_i^{k_j} = 0$ for $i \in I_1^{t(k_j)}(x^*)$ whiles $\langle a^i, x^* \rangle = b_i$ for $i \in I_2^{t(k_j)}(x^*)$, and $I_{t(k_j)} = I_1^{t(k_j)} \cup I_2^{t(k_j)}$ for all $j \geq 0$, we have

$$\begin{aligned} \langle z^{k_j}, Ax^{k_j} - b \rangle &= \langle z^{k_j}, Ax^{k_j} - Ax^* \rangle \\ &= \langle A^T z^{k_j}, x^{k_j} - x^* \rangle \end{aligned}$$

But $A^T z^{k_j} = -\nabla f(x^{k_j})$ and so

$$\begin{aligned} \langle z^{k_j}, Ax^{k_j} - b \rangle &= -\langle \nabla f(x^{k_j}), x^{k_j} - x^* \rangle \\ &= D_f(x^*, x^{k_j}) - f(x^*) + f(x^{k_j}) \to 0 \text{ as } j \to \infty. \end{aligned}$$

Therefore by the continuity of f on $\bar{S}$, we have

$$\lim_{j \to \infty} L(x^{k_j}, z^{k_j}) = \lim_{j \to \infty} \left(f(x^{k_j}) + \langle z^{k_j}, Ax^{k_j} - b \rangle \right) = f(x^*).$$

But $L(x^{k_j}, z^{k_j}) \leq f(z) - D_f(z, x^{k_j}) \leq f(z)$ for all $z \in C \cap \bar{S}$. Therefore $f(x^*) \leq f(z)$ for any $z \in C \cap \bar{S}$. This shows that x^* is a solution for (1.43)-(1.45), since by Proposition 3.2.6, $x^* \in C$. This solution is unique because of the strict convexity of f. □

Chapter 4

Closed form formulas for separated variables optimization

From the computational point of view, the difficult part of any Bregman's algorithm lies in the projection operation at each iterative step. For n-dimensional iterates, the system of equations to solve at each iterative step usually consists of $n+1$ equations, n of which are usually nonlinear. This means that if the system has to be solved numerically in each iteration, then the computational burden might reduce the efficiency of the algorithm. Numerical errors in the calculation of the projections may also cause the practical algorithm to deviate from the conceptual one.

To reduce the computational burden, we develop a closed-form formula for the iterative step in Bregman's algorithm for the optimization of any Bregman function over linear constraints. That is, we replace the computational burden involved in an inner loop calculation of the projection parameter by a closed-form formula.

In [23] and [22], closed-form formulas were derived for the iterative steps for the maximization of Burg's and Shannon's entropy. It was also established in [23] that MART step is indeed a secant approximation to Bregman's iterative step for entropy maximization and this is the motivation for the work in this chapter.

4.1 Analysis of Bregman's algorithm for optimization of variable separable functions

We consider the minimization of the Bregman function $f : \mathbb{R}^n \to \mathbb{R}$ given by $f(x) = \sum_{j=1}^{n} g_j(x_j)$ with zone $S = \text{Int}(\text{dom} f)$. We assume that $g_j'' > 0$ everywhere and that f is essentially smooth, twice continuously differentiable and zone consistent with respect to the hyperplane $H_i = \{x \in \mathbb{R}^n \mid \langle a^i, x\rangle = b_i\}$ for $i = 1, \ldots, m$ and $\cap_{i=1}^{m} H_i \neq \emptyset$.

We begin with the problem

$$\min f(x) \text{ subject to } Ax = b, \tag{4.1}$$

where A is an $m \times n$ real matrix; x and b are n and m dimensional vectors respectively. $a^i \neq 0$ for $i = 1, \ldots, m$ is the transpose of the ith row of A and b_i the ith component of b.

If x^k is the current iterate then, according to Lemma 1.6.10, the following equations determine uniquely the next iterate x^{k+1} and the projection parameter c_k in the Bregman's algorithm for solving (4.1) at kth iterative step.

$$\begin{aligned} \nabla f(x^{k+1}) &= \nabla f(x^k) + c_k a^i, \\ \langle a^i, x^{k+1}\rangle &= b_i. \end{aligned}$$

This implies

$$g_j'(x_j^{k+1}) = g_j'(x_j^k) + c_k a_j^i, \tag{4.2}$$

$$\langle a^i, x^{k+1}\rangle = b_i. \tag{4.3}$$

We assume that the function $f \in B(S)$ satisfies Assumption 2.2.2 since the applicability of the algorithm defined by (4.2) and (4.3) depends on the ability to invert the gradient ∇f explicitly.

Now eliminating x^{k+1} from these two equations, we have

$$\sum_{j=1}^{n} a_j^i G_j^{-1}\left(G_j(x_j^k) + c a_j^i\right) = b_i, \text{ where } c_k = c, \quad g_j' = G_j \tag{4.4}$$

and the inverse G_j^{-1} exists based on the assumption that f is essentially smooth and zone consistent with respect to the hyperplane $H_i = \{x \in \mathbb{R}^n \mid \langle a^i, x\rangle = b_i\}$.

We define the function $h : \mathbb{R} \to \mathbb{R}$ by

$$h(c) = \sum_{j=1}^{n} a_j^i G_j^{-1}\left(G_j(x_j^k) + ca_j^i\right) - b_i. \tag{4.5}$$

We find the approximate root of (4.5) using the root of a secant line to $h(c)$.
To verify whether $h(c)$ really has a root, we consider the following features:

$$h'(c) = \sum_{j=1}^{n} (a_j^i)^2 (G_j^{-1})'\left(G_j(x_j^k) + ca_j^i\right)$$

and

$$(G_j^{-1})'\left(G_j(x_j^k) + ca_j^i\right) = \frac{1}{G_j'\left(G_j^{-1}\left(G_j(x_j^k) + ca_j^i\right)\right)} = \frac{1}{g''\left(G_j^{-1}(G_j\left(x_j^k\right) + ca_j^i)\right)}$$

which is positive since g_j is a strictly convex and twice continuously differentiable function. Therefore $h'(c) > 0$ for all $c \in \mathbb{R}$. This means that h is monotonically increasing on its domain, and by Lemma 1.6.10, the system (4.2)-(4.3) also determines uniquely the parameter c. Thus $h(c)$ has unique real root.

Interestingly, these features of $h(c)$ remain unchanged with relaxation, i.e., if we substitute b_i in (4.5) for $\alpha_i^k b_i + (1 - \alpha_i^k)\langle a^i, x^k\rangle$, where the relaxation parameter α_i^k is such that $0 < \alpha_i^k < 1$ for all $k \geq 0$ and $i = 1, \ldots, m$.

To find an approximate value for the root of $h(c) = 0$, we first find a secant line to the curve $h(c)$ as follows.
The linear approximation of $G_j^{-1}\left(G_j(x_j^k) + ca_j^i\right)$ is given by

$$G_j^{-1}\left(G_j(x_j^k) + ca_j^i\right) \approx G_j^{-1}(G_j(x_j^k)) + ca_j^i(G_j^{-1})'(G_j(x_j^k)) = x_j^k + c\frac{a_j^i}{g_j''(x_j^k)}. \tag{4.6}$$

We substitute (4.6) into (4.5) and the resulting expression becomes the secant line $\hat{f}(c)$ to the curve $h(c)$. That is

$$\hat{f}(c) = \sum_{j=1}^{n} a_j^i \left(x_j^k + c\frac{a_j^i}{g_j''(x_j^k)}\right) - b_i = \langle a^i, x^k\rangle - b_i + c\sum_{j=1}^{n} \frac{(a_j^i)^2}{g_j''(x_j^k)}$$

and for the c-intercept, $\bar{c}$, we have

$$\bar{c} = \frac{b_i - \langle a^i, x^k\rangle}{\sum_{j=1}^{n} \frac{(a_j^i)^2}{g_j''(x_j^k)}}$$

and with underrelaxation (i.e., replacing b_i with $\alpha_i^k b_i + (1-\alpha_i^k)\langle a^i, x^k\rangle$), we have

$$\bar{c} = \alpha_i^k \frac{b_i - \langle a^i, x^k\rangle}{\sum_{j=1}^n \frac{(a_j^i)^2}{g_j''(x_j^k)}}. \tag{4.7}$$

Therefore using the last equation and (4.2), the closed-form formula in the iterative step in underrelaxed Bregman's algorithm for solving (4.1) is

$$x_j^{k+1} = G_j^{-1}\left(g_j'(x_j^k) + a_j^i \alpha_i^k \frac{b_i - \langle a^i, x^k\rangle}{\sum_{j=1}^n \frac{(a_j^i)^2}{g_j''(x_j^k)}}\right) \tag{4.8}$$

for $k \geq 0$, $j = 1, \ldots, n$ and the control sequence $\{i(k)\}$ is almost cyclic on $\{1, 2, \ldots, m\}$.

We therefore propose the following Bregman's algorithm which employs closed-form formula for the iterative updates for solving (4.1).

4.1.1 Bregman's algorithm for linear equalities using closed-form formula

Algorithm 4.1.1. Bregman's algorithm for linear equalities using closed-form formula

(i) **Initialization** $x^0 \in \text{Intdom} f$ is such that for an arbitrary $z^0 \in \mathbb{R}_+^m$,

$$\nabla f(x^0) = -A^T z^0.$$

(ii) **Iterative Step**

$$x_j^{k+1} = G_j^{-1}\left(g_j'(x_j^k) + a_j^i \alpha_i^k \frac{b_i - \langle a^i, x^k\rangle}{\sum_{j=1}^n \frac{(a_j^i)^2}{g_j''(x_j^k)}}\right)$$

The sequence $\{i(k)\}$ is almost cyclic on the index set $\{1, 2, \ldots, m\}$ and $\{\alpha_i^k\}$ is a sequence of relaxation parameters such that $\epsilon \leq \alpha_i^k \leq 1$ for a fixed $\epsilon > 0$.

$\diamondsuit$

4.1.2 General underrelaxed Bregman's algorithm for linear inequalities

The algorithm for linear inequality constraints also calculates the projection parameter $\bar{c}$ in (4.7). However, before proceeding, it compares it with the ith component of the current dual vector z^k and uses the smaller of the two in the iteration. We therefore propose the following general Bregman algorithm for the minimization of any Bregman function over inequality constraints with the projection parameter given in a closed-form in (4.7). That is, a general underrelaxed Bregman's algorithm for solving the problem

$$\min f(x) \text{ subject to } Ax \leq b,\ x \in \bar{S} \tag{4.9}$$

where A and b are as defined in (4.1).

Algorithm 4.1.2. General underrelaxed Bregman's algorithm for linear inequalities

(i) **Initialization** $x^0 \in \text{Intdom} f$ is such that for an arbitrary $z^0 \in \mathbb{R}^m_+$,

$$\nabla f(x^0) = -A^T z^0.$$

(ii) **Iterative Step** Given x^k and z^k, calculate x^{k+1} and z^{k+1} from

$$\begin{aligned} \nabla f(x^{k+1}) &= \nabla f(x^k) + c_i^k a^i, \\ z^{k+1} &= z^k - c_i^k e^i \end{aligned}$$

with

$$c_i^k = \min\left(z_i^k, \theta_i^k\right)$$

where

$$\theta_i^k = \alpha_i^k \frac{b_i - \langle a^i, x^k \rangle}{\sum_{j=1}^n (a_j^i)^2 / g_j''(x_j^k)}.$$

(iii) $H(k)_i = \{x \in \mathbb{R}^n \mid \langle a^i, x \rangle = \alpha_i^k b_i + (1 - \alpha_i^k)\langle a^i, x^k \rangle\}$.

(iv) The sequence $\{i(k)\}$ is almost cyclic on the index set $\{1, 2, \ldots, m\}$

and $\{\alpha_i^k\}$ is a sequence of relaxation parameters such that $\epsilon \leq \alpha_i^k \leq 1$ for a fixed $\epsilon > 0$. We assume that the problem is feasible and f is strongly zone consistent with respect to every H_i. $\diamondsuit$

Note: By analyzing Lemma 4 and its proof in [23] and the proof of Theorem 1 which is the main result of [82], we conjecture that algorithms 4.1.1 and 4.1.2 converge to the desired solutions of problems (4.1) and (4.9).

We will now use (4.8) to derive estimates for closed-form formulas for the iterative steps in Bregman's algorithm for the following functions:

4.1.3 The half-squared Euclidean norm

As noted in Chapter 1, the function $f = \frac{1}{2}\|\cdot\|^2$, which leads to orthogonal projections onto hyperplanes, always leads to a closed-form formula for the iterative step. That is the Kaczmarz's algorithm for solving the system $Ax = b$. When $g_j(x_j^k) = \frac{1}{2}(x_j^k)^2$ in (4.8), we recover exactly the closed-form formula for Kaczmarz's algorithm.

Thus, for $g_j(x_j^k) = \frac{1}{2}(x_j^k)^2$, we have

$$g_j'(x_j^k) = x_j^k \ , \ g_j''(x_j^k) = 1 \text{ and } G_j^{-1}(x_j^k) = x_j^k \tag{4.10}$$

and the substitution of (4.10) into (4.8) gives

$$x_j^{k+1} = x_j^k + a_j^i \alpha_i^k \frac{b_i - \langle a^i, x^k \rangle}{\sum_{j=1}^n (a_j^i)^2} = x_j^k + \alpha_i^k \frac{b_i - \langle a^i, x^k \rangle}{\|a^i\|^2} a_j^i$$

or

$$x^{k+1} = x^k + \alpha_i^k \frac{b_i - \langle a^i, x^k \rangle}{\|a^i\|^2} a^i$$

for all $k \geq 0$.

4.1.4 The negative Shannon entropy

We use (4.8) to derive a closed-form formula for the case when $g_j(x_j^k) = x_j^k \ln x_j^k$ for $x_j^k > 0$ for each j and establish its relation with MART. Now, for $g_j(x_j^k) = x_j^k \ln x_j^k$,

we have

$$g_j'(x_j^k) = 1 + \ln x_j^k \ , \ g_j''(x_j^k) = \frac{1}{x_j^k} \text{ and } G_j^{-1}(x_j^k) = \exp(x_j^k - 1) \tag{4.11}$$

and the substitution of (4.11) into (4.8) gives

$$\begin{aligned} x_j^{k+1} &= G_j^{-1}\left(1 + \ln x_j^k + a_j^i \alpha_i^k \frac{b_i - \langle a^i, x^k\rangle}{\sum_{j=1}^n (a_j^i)^2 x_j^k}\right) \\ &= \exp\left(\ln x_j^k + a_j^i \alpha_i^k \frac{b_i - \langle a^i, x^k\rangle}{\sum_{j=1}^n (a_j^i)^2 x_j^k}\right) \\ &= x_j^k \exp\left(a_j^i \alpha_i^k \frac{b_i - \langle a^i, x^k\rangle}{\sum_{j=1}^n (a_j^i)^2 x_j^k}\right) \\ &= x_j^k \exp \frac{a_j^i \alpha_i^k \langle a^i, x^k\rangle}{\sum_{j=1}^n (a_j^i)^2 x_j^k}\left(\frac{b_i}{\langle a^i, x^k\rangle} - 1\right). \end{aligned} \tag{4.12}$$

For the convergence of MART or Bregman's algorithm for entropy maximization over equality constraints, the following assumptions are made:

(i) $a_j^i > 0$ for $i = 1, \ldots, m$, and $j = 1, \ldots, n$.

(ii) $Ax = b$ is scaled so that $a_j^i \leq 1$ for $i = 1, \ldots, m$ and $j = 1, \ldots, n$.

Using these assumptions, $0 < a_j^i \leq 1$ implies $0 < (a_j^i)^2 x_j^k \leq a_j^i x_j^k$ and so

$$0 \leq \sum_{j=1}^n (a_j^i)^2 x_j^k \leq \langle a^i, x^k\rangle \text{ or } \frac{1}{\sum_{j=1}^n (a_j^i)^2 x_j^k} \geq \frac{1}{\langle a^i, x^k\rangle}. \tag{4.13}$$

Therefore, for the case $\frac{b_i}{\langle a^i, x^k\rangle} > 1$, multiplying the last inequality by $\alpha_i^k(b_i - \langle a^i, x^k\rangle)$, we have

$$a_j^i \alpha_i^k \frac{b_i - \langle a^i, x^k\rangle}{\sum_{j=1}^n (a_j^i)^2 x_j^k} \geq a_j^i \alpha_i^k \frac{b_i - \langle a^i, x^k\rangle}{\langle a^i, x^k\rangle} = a_j^i \alpha_i^k \left(\frac{b_i}{\langle a^i, x^k\rangle} - 1\right).$$

Now, using the inequality $\ln y \leq y - 1$ for $y > 1$, we have

$$a_j^i \alpha_i^k \left(\frac{b_i}{\langle a^i, x^k\rangle} - 1\right) \geq a_j^i \alpha_i^k \ln\left(\frac{b_i}{\langle a^i, x^k\rangle}\right) = \ln\left(\frac{b_i}{\langle a^i, x^k\rangle}\right)^{a_j^i \alpha_i^k}.$$

Using the last inequality and (4.12), we have

$$\begin{aligned} x_j^{k+1} &\geq \exp\left(\ln x_j^k + \ln\left(\frac{b_i}{\langle a^i, x^k\rangle}\right)^{a_j^i \alpha_i^k}\right) \\ &= x_j^k \left(\frac{b_i}{\langle a^i, x^k\rangle}\right)^{a_j^i \alpha_i^k}. \end{aligned}$$

This means that for the case $\frac{b_i}{\langle a^i, x^k\rangle} > 1$, our formula produces iterates that majorize those generated by MART.

Now for the case $0 < \frac{b_i}{\langle a^i, x^k\rangle} < 1$, multiplying (4.13) by $\alpha_i^k(b_i - \langle a^i, x^k\rangle)$, we have

$$a_j^i \alpha_i^k \frac{b_i - \langle a^i, x^k\rangle}{\sum_{j=1}^n (a_j^i)^2 x_j^k} \leq a_j^i \alpha_i^k \frac{b_i - \langle a^i, x^k\rangle}{\langle a^i, x^k\rangle} = \frac{a_j^i b_i \alpha_i^k}{\langle a^i, x^k\rangle}\left(1 - \frac{\langle a^i, x^k\rangle}{b_i}\right)$$

and using the inequality $1 - \frac{1}{y} \leq \ln y$ for $0 < y \leq 1$, we have

$$a_j^i \alpha_i^k \frac{b_i - \langle a^i, x^k\rangle}{\sum_{j=1}^n (a_j^i)^2 x_j^k} \leq \frac{a_j^i b_i \alpha_i^k}{\langle a^i, x^k\rangle} \ln\left(\frac{b_i}{\langle a^i, x^k\rangle}\right) = \alpha_i^k \frac{b_i}{\langle a^i, x^k\rangle} \ln\left(\frac{b_i}{\langle a^i, x^k\rangle}\right)^{a_j^i}.$$

If we define a sequence of relaxation $\{\lambda_i^k\}$ such that $\lambda_i^k = \alpha_i^k \frac{b_i}{\langle a^i, x^k\rangle}$ then (4.12) becomes

$$\begin{aligned} x_j^{k+1} &\leq \exp\left(\ln x_j^k + \lambda_i^k \ln\left(\frac{b_i}{\langle a^i, x^k\rangle}\right)^{a_j^i}\right) \\ &= x_j^k \left(\frac{b_i}{\langle a^i, x^k\rangle}\right)^{a_j^i \lambda_i^k}. \end{aligned}$$

This means that for the case $0 < \frac{b_i}{\langle a^i, x^k\rangle} < 1$, our formula produces iterates that are majorized by those of MART.

However, in order to compare the efficiency of the two algorithms, we ought to do some numerical experimentation on both MART and Algorithm 4.1.2.

4.1.5 The negative Burg's entropy

The Burg's entropy, $B(x)$, also known as '$\log x$-entropy' is defined on $\mathbb{R}^n_{++}$ by

$$B(x) = \sum_{j=1}^n \log x_j.$$

Burg's entropy was first proposed in [17] and has since then provoked a controversy regarding the question of which entropy functional should be used in different situations. This question was discussed in [46] and recently also in [66]. It must be noted that the negative Burg's entropy is not a Bregman function because it becomes singular on the boundary of its zone, i.e., when x_j^k tends to zero for even only one j, then $B(x)$ tends to ∞, demonstrating an essential discontinuity.

In this subsection, we use (4.8) to derive a closed-form formula for the minimization of the negative Burg's entropy, i.e., $f(x) = -B(x)$ with zone $S = \text{Int}\mathbb{R}^n_+$ and $g_j(x^k_j) = -\ln x^k_j$ for $x^k_j > 0$ and for each j. We will compare our method with the method in [22].

Now for $g_j(x^k_j) = -\ln x^k_j$, we have

$$g'_j(x^k_j) = -\frac{1}{x^k_j}\ ,\ g''_j(x^k_j) = \frac{1}{(x^k_j)^2} \text{ and } G_j^{-1}(x^k_j) = -\frac{1}{x^k_j}. \tag{4.14}$$

The substitution of (4.14) into (4.8) gives

$$\begin{aligned} x^{k+1}_j &= G_j^{-1}\left(-\frac{1}{x^k_j} + a^i_j\alpha^k_i\frac{b_i - \langle a^i, x^k\rangle}{\sum_{j=1}^n (a^i_j x^k_j)^2}\right) \\ &= \frac{x^k_j\sum_{j=1}^n (a^i_j x^k_j)^2}{\sum_{j=1}^n (a^i_j x^k_j)^2 + (a^i_j\alpha^k_i x^k_j)(\langle a^i, x^k\rangle - b_i)} \\ &= \frac{x^k_j}{1-u^k_j},\ \text{where } u^k_j = \frac{a^i_j\alpha^k_i x^k_j(b_i - \langle a^i, x^k\rangle)}{\sum_{j=1}^n (a^i_j x^k_j)^2} \neq 1. \end{aligned} \tag{4.15}$$

In [22], using underrelaxation and with the condition $a^i_j b_i \geq 0$ imposed on the elements of the matrix A and the vector b, the following expression was derived for the projection parameter c_k.

$$\varphi(a^i, x^k, b_i) = \begin{cases} \lambda_k\left(1 - \frac{\langle a^i, x^k\rangle}{b_i}\right) t_k \text{ if } b_i > 0, \\ \lambda_k\left(1 - \frac{\langle a^i, x^k\rangle}{b_i}\right) r_k \text{ if } b_i < 0, \end{cases} \tag{4.16}$$

where $0 < \lambda_k < 1$, and

$$r_k = \max\left\{\frac{1}{a^i_j x^k_j} \mid 1 \leq j \leq n,\ a^i_j < 0\right\}$$

and if $a^i_j \geq 0$ for all j then $r_k = -\infty$;

$$t_k = \min\left\{\frac{1}{a^i_j x^k_j} \mid 1 \leq j \leq n,\ a^i_j > 0\right\}$$

and if $a^i_j \leq 0$ for all j then $t_k = \infty$.

But our projection parameter c_k with relaxation is estimated here as

$$c_k = \alpha^k_i\frac{b_i - \langle a^i, x^k\rangle}{\sum_{j=1}^n (a^i_j x^k_j)^2} = \frac{b_i\alpha^k_i}{\sum_{j=1}^n (a^i_j x^k_j)^2}\left(1 - \frac{\langle a^i, x^k\rangle}{b_i}\right). \tag{4.17}$$

We compare (4.17) with (4.16) in [22] and derive any relation between the two estimates.

By the definition of t_k in [22], $a_j^i x_j^k \leq \frac{1}{t_k}$ if $a^i > 0$, and $t_k = \infty$ if $a_j^i \leq 0$ for all j. This means that, $(a_j^i x_j^k)^2 \leq \frac{a_j^i x_j^k}{t_k}$ or $\sum_{j=1}^n (a_j^i x_j^k)^2 \leq \frac{\langle a^i, x^k \rangle}{t_k}$.

Therefore

$$\frac{b_i}{\sum_{j=1}^n (a_j^i x_j^k)^2} \geq \frac{t_k b_i}{\langle a^i, x^k \rangle} \text{ for } b_i > 0 \text{ and } a_j^i \neq 0$$

since $a_j^i b_i > 0$. Thus, using (4.17) for $\frac{b_i}{\langle a^i, x^k \rangle} < 1$, we have

$$c_k \leq \frac{b_i}{\langle a^i, x^k \rangle} \alpha_i^k \left(1 - \frac{\langle a^i, x^k \rangle}{b_i}\right) t_k \tag{4.18}$$

and for $\frac{b_i}{\langle a^i, x^k \rangle} > 1$, we have

$$c_k \geq \frac{b_i}{\langle a^i, x^k \rangle} \alpha_i^k \left(1 - \frac{\langle a^i, x^k \rangle}{b_i}\right) t_k. \tag{4.19}$$

Similarly, by the definition of r_k in [22], $a_j^i x_j^k \geq \frac{1}{r_k}$ if $a^i < 0$, and $r_k = -\infty$ if $a_j^i \geq 0$ for all j. This means that, $(a_j^i x_j^k)^2 \leq \frac{a_j^i x_j^k}{r_k}$ or $\sum_{j=1}^n (a_j^i x_j^k)^2 \leq \frac{\langle a^i, x^k \rangle}{r_k}$.

Therefore

$$\frac{b_i}{\sum_{j=1}^n (a_j^i x_j^k)^2} \leq \frac{r_k b_i}{\langle a^i, x^k \rangle} \text{ for } b_i < 0$$

since $a_j^i b_i \geq 0$. Thus, using (4.17) for $\frac{b_i}{\langle a^i, x^k \rangle} < 1$, we have

$$c_k \geq \frac{b_i}{\langle a^i, x^k \rangle} \alpha_i^k \left(1 - \frac{\langle a^i, x^k \rangle}{b_i}\right) r_k \tag{4.20}$$

and for $\frac{b_i}{\langle a^i, x^k \rangle} > 1$, we have

$$c_k \leq \frac{b_i}{\langle a^i, x^k \rangle} \alpha_i^k \left(1 - \frac{\langle a^i, x^k \rangle}{b_i}\right) r_k. \tag{4.21}$$

Chapter 5

Analysis of inconsistent problems

5.1 Introduction

As stated in the introduction, ART is a very well known iterative algorithm for image processing and reconstruction [51] that aims at solving a linear system of equations

$$Ax = b, \tag{5.1}$$

where $A = (a^i_j)$ is the $m \times n$ nonnegative projection matrix, x is the image vector with components x_j and b is the m-dimensional vector of projection data with ith element b_i ($b_i \geq 0$). For a given starting point x^0, ART's iteration is defined by

$$x^{k+1} = x^k + \lambda_k \frac{b_i - \langle a^i, x^k \rangle}{\|a^i\|^2} a^i. \tag{5.2}$$

$a^i \neq 0$ is the ith column of the transpose A^T and λ_k is a positive relaxation parameter that lies in the open interval $(0, 2)$. The sequence $\{i(k)\}$ determines the ordering (usually cyclic) in which the matrix rows are selected [21].

In the consistent case, when the system (5.1) has solutions, (5.2) converges to the solution for which $\|x - x^0\|$ is minimized. If the system (5.1) has no solution, ordering is cyclic and the relaxation parameter is fixed, it has been proven [25] that (5.2) generates m fixed points, that, when λ tends to zero, (5.2) approaches a weighted least squares solution, that is, the limit solves the problem

$$\min_{x \geq 0} \sum_{i=1}^{m} \left(\frac{b_i - \langle a^i, x \rangle}{\|a^i\|} \right)^2. \tag{5.3}$$

In [24, 81], it was also proven that the sequence (5.2) itself converges to the solution of (5.3) provided that the relaxation parameters satisfy the conditions

$$\lambda_k \underset{k\to\infty}{\longrightarrow} 0 \tag{5.4}$$

and

$$\sum_{k=0}^{\infty} \lambda_k = +\infty. \tag{5.5}$$

Block versions of ART share the same properties mentioned above and the fully parallel version (a single block), known as Cimmino's method can be shown to converge to the solution in the feasible case and to a least squares solution in the infeasible case for a fixed relaxation parameter in (0, 2) (see [40]).

The question that follows is that, is it true that for other Bregman methods or algorithms, the under-relaxed version of the sequences generated by the algorithms in the inconsistent case converges to the solution of an optimization problem, as it is the case in ART, if the relaxation parameters satisfy certain condition? Also can the fully simultaneous version of the sequences generated by other Bregman methods in the inconsistent case converge to the solution of an optimization problem? In this chapter, we give an answer for the particular case of MART and its simultaneous version SMART, already analyzed in [18].

One of the first iterative algorithms in image reconstruction for underdetermined problems (limited number of views) is MART [51]. This is also a very well known nonlinear iterative algorithm for transmission computed tomography (CT) with very attractive properties. It converges to a maximum entropy solution of the linear CT equations [35, 71, 37] and it confines the reconstruction to the convex hull of the object [8, 33].

Both ART and MART were introduced by Gordon, Bender and Herman. But MART is limited to non-negative systems for which non-negative solutions are sought. That is, MART finds non-negative solutions to the system (5.1) provided (5.1) is non-negative, i.e., a_j^i's are non-negative and b_i's are positive.

In the under-determined case, both algorithms find the solution closest to the

starting point, in the two-norm or weighted two-norm sense for ART, and in the cross-entropy sense for MART. Thus, both algorithms can be viewed as solving optimization problem.

MART (now we consider the cyclic version, $i(k) = k(\text{mod } m)$) is defined by the following sequence: given a positive starting point x^0, and, in general, $x^k = x^{(k,0)}$, and $x^{k+1} = x^{(k,m)}$, then, for all $i = 1, \ldots, m$ and, $j = 1, \ldots, n$ such that $a^i_j > 0$ and $x^k_j > 0$,

$$x^{(k,i)}_j = x^{(k,i-1)}_j \left(\frac{b_i}{\langle a^i, x^{(k,i-1)} \rangle} \right)^{\lambda_k a^i_j}. \tag{5.6}$$

Like ART, in the consistent case, when the system (5.1) has solutions, (5.6) converges to the solution for which $K(x, x^0)$ is minimized, where the function K is as defined in (5.8) below.

Next, we prove that the properties of MART are similar to those of ART when the problem is inconsistent. We prove that, when properly underrelaxed as in (5.4) and (5.5), the algorithm converges to a solution of the problem

$$\min_{x \geq 0} L(x) := \sum_{i=1}^{m} \left(\langle a^i, x \rangle \ln \langle a^i, x \rangle - \langle a^i, x \rangle (\ln b_i + 1) \right) \tag{5.7}$$

which is the projection of the data vector b onto the range of A with respect to the Kullback-Leibler distance [69]. The Kullback-Leibler distance, or relative entropy of the vector x in $\mathbb{R}^n_+$ with respect to the vector y also in $\mathbb{R}^n_{++}$ is defined as

$$K(x, y) = \sum_{j=1}^{n} \left(x_j \ln \frac{x_j}{y_j} + y_j - x_j \right). \tag{5.8}$$

So, (5.7) is equivalent to

$$\min_{x \geq 0} K(Ax, b), \tag{5.9}$$

since, using (5.8) with the assumption that $a^i_j \geq 0$ and $b_i > 0$, we have $Ax \in \mathbb{R}^m_+$ and $b \in \mathbb{R}^m_{++}$ and so

$$K(Ax, b) = \sum_{i=1}^{m} (\langle a^i, x \rangle \ln \langle a^i, x \rangle - \langle a^i, x \rangle (\ln b_i + 1) + b_i), \tag{5.10}$$

which has the same minimizer as (5.7) since b_i is a positive constant for each i. The solution of (5.9) is not necessarily unique. We will prove that MART chooses

the one minimizing $K(x, x^0)$, where x^0 is the (positive) starting point. It gives the maximum entropy if $x^0 = \mathbf{1}$. We will confine our results to row action MART described above, for the sake of simplicity. Proofs for more general block versions, like those in [18], are similar.

5.2 Convergence results

5.2.1 Boundedness

Taking into account (5.4) for k large enough, we have that $\forall\, i, j$

$$\lambda_k a_j^i \;\leq\; 1. \tag{5.11}$$

Define

$$c \;=\; \min_{a_j^i \neq 0}\{a_j^i\} \tag{5.12}$$

and

$$\bar{b} \;=\; \max_i\, b_i. \tag{5.13}$$

Then, using (5.11) and the fact that $x_j^k > 0$ for each j and $k \geq 0$, we obtain the following inequalities for $a_j^i > 0$ (when $a_j^i = 0$ the iteration remains unmodified) and $\frac{b_i}{\langle a^i, x^k\rangle} > 1$.

$$\begin{aligned} 0 \;<\; x_j^{k+1} = x_j^k \left(\frac{b_i}{\langle a^i, x^k\rangle}\right)^{\lambda_k a_j^i} &\leq x_j^k \frac{b_i}{\langle a^i, x^k\rangle} \\ &\leq \frac{x_j^k a_j^i b_i}{c\sum_l a_l^i x_l^k} \leq \frac{b_i}{c\left(1 + \sum_{l\neq j} \frac{a_l^i x_l^k}{a_j^i x_j^k}\right)} \leq \frac{\bar{b}}{c}. \end{aligned} \tag{5.14}$$

The inequalities in (5.14) are as a result of (5.12) and (5.13).

Now suppose that, for some k, $x_j^k > \frac{\bar{b}}{c}$, then the inequality (5.14) is valid if $\frac{b_i}{\langle a^i, x^k\rangle} > 1$, and so $\frac{\bar{b}}{c} < x_j^k < x_j^k \frac{b_i}{\langle a^i, x^k\rangle} \leq \frac{\bar{b}}{c}$ which is a contradiction. Hence $x_j^k \leq \frac{\bar{b}}{c}$. Also if $\frac{b_i}{\langle a^i, x^k\rangle} \leq 1$, then from (5.6), $x_j^{k+1} \;\leq\; x_j^k$. Using induction on k, it can be easily proved that $x_j^k \;\leq\; x_j^{k_0}$ for some given k_0. So, the sequence generated by (5.6) is uniformly bounded (independently of λ).

5.2.2 Change of variables

Now, let us change variables in (5.6). Taking logarithms and using the notation $e^y = (e^{y_1}, \ldots, e^{y_n})^T$ for $y = (y_1, \ldots, y_n)^T$ and $y_j^{(k,i)} = \ln\left(x_j^{(k,i)}\right)$, we obtain the iteration

$$y_j^{(k,i)} = y_j^{(k,i-1)} + \lambda_k a_j^i \ln\left(\frac{b_i}{\langle a^i, e^{y^{(k,i-1)}}\rangle}\right) \text{ for } j = 1, \ldots, n. \tag{5.15}$$

Alternatively, we may consider $K(x_j^{(k,i)}, 1)$ and use $\nabla K(x_j^{(k,i)}, 1)$ for $y_j^{(k,i)}$ since $K(x_j^{(k,i)}, 1) = x_j^{(k,i)} \ln x_j^{(k,i)} + 1 - x_j^{(k,i)}$ and $\nabla K(x_j^{(k,i)}, 1) = \ln x_j^{(k,i)} = y_j^{(k,i)}$. (5.15) therefore takes the form

$$\nabla K(x_j^{(k,i)}, 1) = \nabla K(x_j^{(k,i-1)}, 1) + \lambda_k a_j^i \ln\left(\frac{b_i}{\langle a^i, x^{(k,i-1)}\rangle}\right) \text{ for } j = 1, \ldots, n. \tag{5.16}$$

Now, summing (5.15) over i for $j = 1, \ldots, n$, we have

$$y_j^{(k,m)} - y_j^{(k,0)} = \lambda_k \sum_{i=1}^{m} a_j^i \ln\left(\frac{b_i}{\langle a^i, e^{y^{(k,i-1)}}\rangle}\right). \tag{5.17}$$

(5.15) and (5.17) motivate the definition of the following objective function

$$\hat{L}(y) := \sum_{i=1}^{m} (\langle a^i, e^y\rangle \ln\langle a^i, e^y\rangle - \langle a^i, e^y\rangle(\ln b_i + 1)). \tag{5.18}$$

Equivalently, we write

$$\hat{L}(y) = \sum_{i=1}^{m} \hat{L}_i(y), \tag{5.19}$$

where

$$\hat{L}_i(y) = \langle a^i, e^y\rangle \ln\langle a^i, e^y\rangle - \langle a^i, e^y\rangle(\ln b_i + 1). \tag{5.20}$$

In order to simplify our notation, we will consider the extended n-dimensional vector space, adding possibly $-\infty$ coordinates, that we will denote by $\mathcal{R}_-^n$. There is a one-to-one correspondence between this set and $\mathbb{R}_+^n$ via the logarithm of the coordinates.

A trivial observation using (5.7) and (5.19) is that,

$$\min_{x \geq 0} L(x) = \min_{y \in \mathbb{R}^n} \hat{L}(y) \tag{5.21}$$

and for $x = e^y$,

$$e^{-y_j}\nabla\hat{L}(y)_j = \nabla L(x)_j. \tag{5.22}$$

It is easy to show that, if $y^* \in \mathcal{R}^n_-$ is a minimum, and $x^* = e^{y^*}$, then, for $i = 1, \ldots, m$, using (5.22),

$$\nabla\hat{L}(y^*)_j \;=\; x^*_j\nabla L(x^*)_j \;=\; 0. \tag{5.23}$$

Iteration (5.15) becomes

$$y_j^{(k,i)} = y_j^{(k,i-1)} - \lambda_k e^{-y_j{}^{(k,i-1)}}\nabla\hat{L}_i(y^{(k,i-1)})_j, \tag{5.24}$$

for $j = 1, \ldots, n$. This is because, from (5.20), we have

$$\begin{aligned}\nabla\hat{L}_i(y)_j &= \langle a^i, e^y\rangle\frac{a^i_j e^{y_j}}{\langle a^i, e^y\rangle} + a^i_j e^{y_j}\ln\langle a^i, e^y\rangle - a^i_j e^{y_j}\ln(b_i+1)\\ &= a^i_j e^{y_j}\ln\langle a^i, e^y\rangle - a^i_j e^{y_j}\ln b_i = -a^i_j e^{y_j}\ln\left(\frac{b_i}{\langle a^i, e^y\rangle}\right).\end{aligned}$$

(5.24) means that, for k large enough, $\hat{L}_i$ is decreased at each iteration (see Section 1.2 of [10]).

To simplify notation, we define the diagonal matrix

$$D(y) \;=\; \mathrm{diag}(e^{-y_1}, \ldots, e^{-y_n}), \tag{5.25}$$

and substitute D^k for $D(y^k)$, and in general, we substitute D for $D(y)$, whenever this is clear. Using this notation, (5.24) becomes

$$y^{(k,i)} = y^{(k,i-1)} - \lambda_k D(y^{(k,i-1)})\nabla\hat{L}_i(y^{(k,i-1)}) \tag{5.26}$$

and since in Subsection 5.2.1, $\{x^{(k,i)}\}$ is bounded, $\{y^{(k,i)}\}$ is bounded from above.

Lemma 5.2.1. *The differences between sub-iterates in (5.24) and major iterates in (5.26) tend to zero, that is, for $i = 1, \ldots, m$,*

$$y^{(k,i)} - y^{(k,i-1)} \underset{k\to\infty}{\longrightarrow} 0 \tag{5.27}$$

and

$$y^{k+1} - y^k \underset{k\to\infty}{\longrightarrow} 0. \tag{5.28}$$

Also, for $i = 1, \ldots, m$,

$$\frac{x_j^{(k,i)}}{x_j^{(k,i-1)}} \longrightarrow_{k\to\infty} 1. \tag{5.29}$$

Proof. The sequence $\{x^k\}$ is bounded and so $\{\langle a^i, e^{y^{(k,i-1)}}\rangle\}$ is bounded and

$$\left\{\frac{b_i}{\langle a^i, e^{y^{(k,i-1)}}\rangle}\right\} \tag{5.30}$$

is bounded away from zero for each i. If $\{\langle a^i, e^{y^{(k,i-1)}}\rangle\}$ is bounded away from zero then the factors multiplying the relaxation parameter in (5.15) and (5.17) are bounded and (5.27) and (5.28) hold. If not, then $\langle a^i, e^{y^{(k,i-1)}}\rangle > 0$ for $i = 1, \ldots, m$ and for all $k \geq 0$. This means that (5.30) is not bounded from above and so $\frac{b_i}{\langle a^i, e^{y^{(k,i-1)}}\rangle}$ may tend to $+\infty$ as $k \to \infty$. But, from (5.17), for $j = 1, \ldots, n$, we have

$$y_j^{k+1} = y_j^0 + \sum_{l=0}^{k} \lambda_l \sum_{i=1}^{m} a_j^i \ln\left(\frac{b_i}{\langle a^i, e^{y^{(l,i-1)}}\rangle}\right) \tag{5.31}$$

since $y_j^{k+1} = y_j^{(k,m)}$ and $y_j^k = y_j^{(k,0)}$.

Thus the series on the right hand side of (5.31) may diverge to $+\infty$ due to (5.5) and the fact that (5.30) is not bounded from above. This is a contradiction since $\{y_j^{k+1}\}$ is bounded from above for each j and so $\{\langle a^i, e^{y^{(k,i-1)}}\rangle\}$ must be bounded away from zero. Therefore $\{\langle a^i, e^{y^{(k,i-1)}}\rangle\}$ is bounded and bounded away from zero. Hence (5.29) follows, since $y_j^{(k,i)} - y_j^{(k,i-1)} \longrightarrow 0$ implies $\ln x_j^{(k,i)} - \ln x_j^{(k,i-1)} \longrightarrow 0$ and so $\frac{x_j^{(k,i)}}{x_j^{(k,i-1)}} \longrightarrow 1$, bearing in mind that ln is continuous. □

Another immediate consequence of the rationale in the last lemma is the following corollary. Now, let the set $\mathcal{C}$ be the closure of the convex hull of the iterates $(y^{(k,i)})$ in $\mathcal{R}_-^n$.

Corollary 5.2.2. $\|D(y)\nabla\hat{L}_l(y)\|$ *is bounded in* $\mathcal{C}$.

Proof. Considering only the jth component of $D(y)\nabla\hat{L}_l(y)$, we have

$$\begin{aligned} D(y)_j\nabla\hat{L}_l(y)_j &= e^{-y_j}\nabla\hat{L}_l(y)_j \\ &= e^{-y_j}\left(-a_j^i e^{y_j}\ln\left(\frac{b_i}{\langle a^i, e^{y^{(k,i)}}\rangle}\right)\right) \\ &= -a_j^i \ln\left(\frac{b_i}{\langle a^i, e^{y^{(k,i)}}\rangle}\right) \end{aligned} \tag{5.32}$$

and from the last lemma, $\{\langle a^i, e^{y^{(k,i)}}\rangle\}$ is bounded and bounded away from zero and so $\ln\left(\frac{b_i}{\langle a^i, e^{y^{(k,i)}}\rangle}\right)$ is bounded since ln is continuous and $\mathcal{C}$, being the closure of the convex hull of the bounded iterates $(y^{(k,i)})$ in $\mathcal{R}^n_-$, is bounded. □

The next lemma shows that the sequence is essentially asymptotically decreasing.

Lemma 5.2.3. *If there is an accumulation point $y^* \in \mathcal{R}^n_-$, and a positive number γ such that*

$$\|\nabla \hat{L}(y)\| > \gamma \tag{5.33}$$

for y in an open set Ω in $\mathcal{R}^n_-$ containing y^*, then for k large enough and $y^k \in \Omega$, $\hat{L}(y^{k+1}) \leq \hat{L}(y^k)$. In particular, if (5.33) is valid for every y, then the whole sequence decreases for k large enough.

Proof. Using (5.26) for a complete iteration, we have

$$y^{k+1} = y^k - \lambda_k \sum_{i=1}^{m} (D^{(k,i-1)} \nabla \hat{L}_i(y^{(k,i-1)})) \tag{5.34}$$

and using (5.20), we have

$$D^k \nabla \hat{L}(y^k) = \sum_{i=1}^{m} D^k \nabla \hat{L}_i(y^k). \tag{5.35}$$

Define the algorithm's direction as

$$d^k = \sum_{i=1}^{m} D^{(k,i-1)} \nabla \hat{L}_i(y^{(k,i-1)})$$

so that

$$y^{k+1} = y^k - \lambda_k d^k.$$

Now, using (5.34) and (5.35), for k large enough, we have

$$\begin{aligned} y^{k+1} &= y^k - \lambda_k D^k \nabla \hat{L}(y^k) - \lambda_k \sum_{i=1}^{m} (D^{(k,i-1)} \nabla \hat{L}_i(y^{(k,i-1)}) - D^k \nabla \hat{L}_i(y^k)) \\ &= y^k - \lambda_k D^k \nabla \hat{L}(y^k) + O({\lambda_k}^2), \end{aligned} \tag{5.36}$$

where the last equality is obtained by Lipschitz continuity and boundedness of $D(y)\nabla\hat{L}_i(y)$ on $\mathcal{C}$ by Corollary 5.2.2. That is, for some positive $W \in \mathbb{R}$, we have

$$\left\|\sum_{i=1}^{m}\lambda_k(D^{(k,i-1)}\nabla\hat{L}_i(y^{(k,i-1)}) - D^k\nabla\hat{L}_i(y^k))\right\| \leq \lambda_k W\sum_{i=1}^{m}\left\|y^{(k,i-1)} - y^k\right\|$$

$$= \lambda_k W\sum_{i=1}^{m}\left\|y^{(k,i-1)} - y^{(k,0)}\right\| = \lambda_k W\sum_{i=2}^{m}\left\|\sum_{l=1}^{i-1}(y^{(k,l)} - y^{(k,l-1)})\right\|$$

$$\leq (\lambda_k)^2 W\sum_{i=2}^{m}\sum_{l=1}^{i-1}\left\|D^{(k,l-1)}\nabla\hat{L}_l(y^{(k,l-1)})\right\| \leq (\lambda_k)^2 W\sum_{i=2}^{m}\sum_{l=1}^{i-1}M$$

where

$$M = \max_{y^{(k,l)}\in\mathcal{C}}\left\{\|D^{(k,l-1)}\nabla\hat{L}_l(y^{(k,l-1)})\|\right\}.$$

Therefore

$$\left\|\sum_{i=1}^{m}\lambda_k(D^{(k,i-1)}\nabla\hat{L}_i(y^{(k,i-1)}) - D^k\nabla\hat{L}_i(y^k))\right\| \leq (\lambda_k)^2 m^2 W M$$

and so for k large enough,

$$-\lambda_k\sum_{i=1}^{m}(D^{(k,i-1)}\nabla\hat{L}_i(y^{(k,i-1)}) - D^k\nabla\hat{L}_i(y^k)) = O({\lambda_k}^2).$$

Now consider the objective function $\hat{L}(y^k)$. Since $\nabla\hat{L}_i(y)$ is Lipschitz continuous on $\mathcal{C}$, for some $L_0 > 0$, we have from [[77], precisely (15) on p.6],

$$\left\|\hat{L}(y^{k+1}) - \hat{L}(y^k) - (y^{k+1} - y^k)^T\nabla\hat{L}(y^k)\right\| \leq \frac{L_0}{2}\|y^{k+1} - y^k\|^2.$$

Therefore, from (5.28), for k large enough, we have

$$\hat{L}(y^{k+1}) = \hat{L}(y^k) + (y^{k+1} - y^k)^T\nabla\hat{L}(y^k) + O(\|y^{k+1} - y^k\|^2). \tag{5.37}$$

Now, using (5.36) for k large enough and the fact that $D^k\nabla\hat{L}(y^k)$ is bounded , we have

$$\begin{aligned}(y^{k+1} - y^k)^T\nabla\hat{L}(y^k) &= (-\lambda_k D^k\nabla\hat{L}(y^k) + O({\lambda_k}^2))^T\nabla\hat{L}(y^k)\\ &= -\lambda_k(\nabla\hat{L}(y^k))^T D^k\nabla\hat{L}(y^k) + O({\lambda_k}^2)\nabla\hat{L}(y^k)\\ &= -\lambda_k(\nabla\hat{L}(y^k))^T D^k\nabla\hat{L}(y^k) + O({\lambda_k}^2)\end{aligned}$$

and

$$\begin{aligned}\|y^{k+1}-y^k\|^2 &= (-\lambda_k D^k\nabla\hat{L}(y^k)+O({\lambda_k}^2))^T(-\lambda_k D^k\nabla\hat{L}(y^k)+O({\lambda_k}^2))\\ &= {\lambda_k}^2(D^k\nabla\hat{L}(y^k))^T D^k\nabla\hat{L}(y^k)+O({\lambda_k}^2)\lambda_k D^k\nabla\hat{L}(y^k)\\ &\quad +O({\lambda_k}^2)O({\lambda_k}^2)\\ &= O({\lambda_k}^2)+O({\lambda_k}^2)+O({\lambda_k}^2)=O({\lambda_k}^2),\end{aligned}$$

due to (5.4). Therefore from (5.37), for k large enough, we have

$$\hat{L}(y^{k+1})=\hat{L}(y^k)-\lambda_k(\nabla\hat{L}(y^k))^T D^k\nabla\hat{L}(y^k)+O({\lambda_k}^2) \tag{5.38}$$

and the result holds since the factor multiplying λ_k is bounded away from zero. This is because D^k is bounded away from zero, since in the proof of Lemma 5.2.1, x^k or e^{y^k} is bounded away from zero. Thus, using (5.33), $(\nabla\hat{L}(y^k))^T D^k\nabla\hat{L}(y^k)$ is bounded away from zero. □

5.2.3 Limit points

In the next proposition, we further assume that in addition to conditions (5.4) and (5.5), the relaxation parameter λ_k is chosen such that $\sum_{k=1}^{\infty}(\lambda_k)^2<\infty$. In this case, $\sum_{k=1}^{\infty}O(\lambda_k)^2<\infty$.

Proposition 5.2.4. *If $\sum_{k=1}^{\infty}(\lambda_k)^2<\infty$ then there is a limit point x^* of the sequence (5.6) such that*

$$x_j^*\nabla L(x^*)_j = 0. \tag{5.39}$$

Equivalently

$$\nabla\hat{L}(y^*) = 0, \tag{5.40}$$

where $y^*\in\mathcal{R}_-^n$ and $y_j^*=\ln x_j^*$ for $j=1,\ldots,n$.

Proof. If there is no such a limit point then using (5.22),

$$D^{-1}(y^k)D(y^k)\nabla L(x^k)=D(y^k)\nabla\hat{L}(y^k)$$

is bounded away from zero and there exists a real positive number η such that

$$(\nabla\hat{L}(y^k))^T D^k\nabla\hat{L}(y^k)>\eta>0.$$

Thus, by Lemma 5.2.3, $\hat{L}(y^{k+1}) \leq \hat{L}(y^k)$ for every $k \geq k_0$, and, for every integer $l > k_0$, we have

$$\hat{L}(y^{l+1}) - \hat{L}(y^{k_0}) = \sum_{k=k_0}^{l} (\hat{L}(y^{k+1}) - \hat{L}(y^k)) \leq -\eta \sum_{k=k_0}^{l} \lambda_k + \sum_{k=k_0}^{l} O({\lambda_k}^2). \quad (5.41)$$

But, since by hypothesis, $\lim_{l\to\infty} \sum_{k=k_0}^{l} O({\lambda_k}^2)$ is finite, the right hand side of (5.41) tends to $-\infty$ as l tends to infinity due to (5.5) whiles the left hand side is bounded which is a contradiction. □

Theorem 5.2.5. *Every limit point of the sequence generated by (5.6) satisfies (5.39).*

Proof. From Proposition 5.2.4, we know that such a limit point, say x^* or y^*, exists. Suppose that there is another limit point $x^{**} = \lim_{j\to\infty} x^{k_j}$ ($y^{**} = \lim_{j\to\infty} y^{k_j} = \lim_{j\to\infty} \ln x^{k_j}$, such that $x^{**}\nabla L(x^{**}) \neq 0$ or $\nabla L(y^{**}) \neq 0$). Then, it is clear from (5.23) that, $\lim_{j\to\infty} \nabla\hat{L}(y^{k_j}) \neq 0$, that is, $\|\nabla\hat{L}(y^{k_j})\| > \gamma$ for some positive γ.

Now let LP be the set of limit points of the sequence $\{y^k\}$ such that $\nabla L(y) = 0$ or $x\nabla L(x) = 0$. Then LP is closed and bounded since the sequence is bounded. For a given positive ϵ, define the set $LP_\epsilon := \{y \mid \text{distance}(y, LP) \leq \epsilon\}$, and assume that the limit point $y^{**} \notin LP_\epsilon$ and that $\text{distance}(y^{**}, LP_\epsilon) = \delta > 0$. Then $\|\nabla\hat{L}(y^{k_j})\| > \gamma$ for $y^{k_j} \in \Omega$, where Ω is some open neighbourhood of y^{**} in $\mathcal{R}^n_-$ and a complement of LP_ϵ in $\mathcal{R}^n_-$.

Now considering the whole sequence $\{y^k\}$, by Lemma 5.2.3, $\{\hat{L}(y^k)\}$ decreases for k large enough and $y^k \in \Omega$, i.e., there exists a constant k_0 such that $\hat{L}(y^{k+1}) \leq \hat{L}(y^k)$ for $k \geq k_0$ if $y^k \notin LP_\epsilon$ or $y^k \in \Omega$. But, since there is a subsequence converging to $y^* \in LP$, there exists $k_1 > k_0$ such that $y^{k_1} \in LP_\epsilon$ and $\hat{L}(y^*) < \hat{L}(y^k)$. This is because if $\hat{L}(y^*) \leq \hat{L}(y^k)$ for k large enough and $y^k \in LP_\epsilon$ then

(i) there could not exist a subsequence converging to y^{**} this is because the distance between iterates tends to zero, i.e., $y^{k+1} - y^k \longrightarrow 0$,

(ii) the sequence $\{\hat{L}(y^k)\}$ decreases outside LP_ϵ and the distance between y^{**} and LP_ϵ is positive.

Thus, since $y^{k_1} \in LP_\epsilon$, if by continuity $LP_\epsilon \longrightarrow LP$ as $\epsilon \longrightarrow 0$ then y^{k_1} tends to say $\bar{y}$ as ϵ tends to zero (or a convergent subsequence for ϵ tending to zero). Therefore $\bar{y} \in LP$, $\nabla \hat{L}(\bar{y}) = 0$ and $\hat{L}(\bar{y}) \geq \hat{L}(y^*)$ which is a contradiction.

□

5.2.4 Convergence of the whole sequence

The next proposition is trivial, reflecting the fact that the solution set is the intersection of a linear system with the nonnegative orthant. The proof can be found in [19].

Proposition 5.2.6. *If x^* solves (5.7) then every other solution x (nonnegative) solves the linear system $Ax = Ax^*$.*

Theorem 5.2.7. *The whole sequence generated by (5.6) converges to a maximum entropy (relative to the starting point) solution of (5.7).*

Proof. From (5.16), it is clear that

$$\nabla K(x^k, \mathbf{1}) = \nabla K(x^{k-1}, \mathbf{1}) + A^T z^{k-1} \tag{5.42}$$

for some vector $z^{k-1} \in \mathbb{R}^m$ with $z_i^{k-1} = \lambda_k \ln\left(\frac{b_i}{\langle a^i, x^{(k,i-1)}\rangle}\right)$ and $\mathbf{1}$ stands for the vector of ones. Therefore iterating, we have

$$\nabla K(x^k, \mathbf{1}) - \nabla K(x^0, \mathbf{1}) = \nabla K(x^k, x^0) = A^T \sum_{s=1}^{k} z^{k-s} \tag{5.43}$$

and taking limits, we have

$$\nabla K(x^*, x^0) = 0 \in \mathrm{Im}(A^T) \tag{5.44}$$

since, in this case, for $j = 1, \ldots, n$,

$$\lim_{k\to\infty} \sum_{s=1}^{k} \sum_{i=1}^{m} a_j^i z_i^{k-s} = \lim_{k\to\infty} \sum_{s=1}^{k} \lambda_s \sum_{i=1}^{m} a_j^i \ln\left(\frac{b_i}{\langle a^i, x^{(s,i-1)}\rangle}\right) = 0$$

in view of (5.5) and the fact that x^* is a maximum entropy solution of (5.7). We must therefore have

$$\lim_{s\to\infty}\sum_{i=1}^{m} a_j^i \ln\left(\frac{b_i}{\langle a^i, x^{(s,i-1)}\rangle}\right) = \sum_{i=1}^{m} a_j^i \ln\left(\frac{b_i}{\langle a^i, x^*\rangle}\right) = 0.$$

But, considering Proposition 5.2.6, (5.44) completes the Kuhn-Tucker conditions for the problem

$$\min_{x\geq 0} K(x, x^0) \text{ subject to } x = \arg\min_{x\geq 0} L(\hat{x}). \tag{5.45}$$

The problem (5.45) has a unique solution (entropy is a strictly convex function with linear constraints); so, the whole sequence converges. □

5.3 On SMART

In this section, we deduce the Simultaneous Multiplicative Algebraic Reconstruction Technique, better known as SMART (see [18]), as a particular case of the majorizing functions technique, introduced by De Pierro in [15, 39], applied to the function defined by (5.7). Writing $\langle a^i, x\rangle = \sum_{j=1}^{n}\left(\frac{a_j^i x_j^k}{\langle a^i, x^k\rangle}(\frac{x_j}{x_j^k})\langle a^i, x^k\rangle\right)$ and noting that $\sum_{j=1}^{n}\frac{a_j^i x_j^k}{\langle a^i, x^k\rangle} = 1$, and using convexity of the $x\ln x$ functional, we have the following inequality,

$$L(x) \leq \sum_{i=1}^{m}\sum_{j=1}^{n}\left(\frac{a_j^i x_j^k}{\langle a^i, x^k\rangle}\cdot\frac{x_j}{x_j^k}\langle a^i, x^k\rangle \ln\frac{x_j}{x_j^k}\langle a^i, x^k\rangle - (1+\ln b_i)\langle a^i, x\rangle\right), \tag{5.46}$$

i.e.,

$$L(x) \leq \varphi(x, x^k) = \sum_{i=1}^{m}\sum_{j=1}^{n}\left(a_j^i x_j \ln\frac{x_j}{x_j^k}\langle a^i, x^k\rangle - (1+\ln b_i)\langle a^i, x\rangle\right). \tag{5.47}$$

The function $\varphi(x, x^k)$ is strictly convex with separated variables and has the following properties:

(i) $\varphi(x^k, x^k) = L(x^k)$,

(ii) $\nabla_x\varphi(x^k, x^k) = \nabla L(x^k)$.

Essentially this means that the new function majorizes the one that we want to minimize and coincides with its gradient at the current iterate x^k.

The majorizing algorithm is then defined iteratively as:

$$x^{k+1} = \arg\min_{x\geq 0} \varphi(x, x^k). \tag{5.48}$$

In order to compute the next iterate, we differentiate the variable separable function $\varphi(x, x^k)$ and have

$$\frac{\partial \varphi(x, x^k)}{\partial x_j} = \sum_{i=1}^{m} a_j^i \left(-\ln b_i + \ln x_j \frac{\langle a^i, x^k \rangle}{x_j^k} \right) = \ln \prod_{i=1}^{m} \left(\frac{x_j}{x_j^k} \frac{\langle a^i, x^k \rangle}{b_i} \right)^{a_j^i} = 0. \tag{5.49}$$

From this equation, we have

$$(x^{k+1})^{\sum_i a_j^i} = (x^k)^{\sum_i a_j^i} \prod_{i=1}^{m} \left(\frac{b_i}{\langle a^i, x^k \rangle} \right)^{a_j^i}. \tag{5.50}$$

When the row sums, $\sum_i a_j^i$, are one, we recover SMART. Following this approach, a convergence proof of SMART can be obtained, but this has been done in a more general setting by C. Byrne in [18].

Chapter 6

New results, conclusion and future work

6.1 New results and conclusion

We presented new results on Bregman iterative methods. In Chapter 2, a new approach to Bregman projection methods for the convex feasibility problem that generalizes the concept of relaxation introduced in [41, 29] was presented. This generalization allowed us to define approximate Bregman's projection method for general not necessarily linear convex feasibility problems. As a consequence, we derived an application to convex but nonlinear sets of constraints and to linear equality constraints as well.

In Chapter 3, we discussed block type methods for linear inequality constrained problems and the corresponding convergence proofs emphasizing on the dual approach. To the best of our knowledge, the simultaneous version of the Bregman's algorithmic scheme for solving linearly constrained optimization problem formulated in Chapter 3 does not exist anywhere in the literature. Therefore our simultaneous algorithmic scheme is a new method for minimizing any Bregman function over linear constraints. A convergence proof of this algorithm was also given.

Again, for the first time, a general closed-form formula for the iterative step in

Bregman's algorithm for the optimization of any Bregman function with separated variables was presented in Chapter 4. The secant approximation approach used in [23] to establish the relationship between MART and Entropy maximization over linear equalities was the motivation.

In Chapter 5, we analyzed the behaviour of Bregman type algorithm when the problem is inconsistent. We have proven that adequately underrelaxed MART for equalities converges to a minimum solution of Kullback Leibler distance.

6.2 Future work

In future, we will illustrate the conjecture outlined at the end of Chapter 2 by some numerical examples and prove it. This will enable us construct an algorithm using a sequence of Bregman projections in the general case, i.e., where the constraints are not only linear equalities, to minimize the Bregman function involved. We will then examine the specific case where the Bregman function is quadratic for norm minimization.

In Chapter 3, we produced a simultaneous Bregman's algorithm that minimizes Bregman function over polyhedron. As part of our future work, we will consider the case where the polyhedron is empty and find an approximate solution for the minimization problem. This will be a generalization of the result in [40], i.e., the simultaneous Hildreth method where the function $\frac{1}{2}\|\cdot\|^2$ was minimized over the empty polyhedron.

We will also consider a generalization of the work in [73], as a consequence of a result in [2]. We will do this by proving the following statements:

(i) Any sequence of orthogonal projections onto a finite family of polyhedral convex sets is bounded.

(ii) Any sequence of Bregman projections (including the orthogonal projections) onto a finite family of affine sets is bounded.

(iii) Any sequence of Bregman projections onto a finite family of polyhedral

convex sets is bounded.

The Algorithmic schemes 4.1.1 and 4.1.2 proposed in Chapter 4 are without convergence results. To establish the practicability of these algorithmic schemes, we shall first prove the convergence results of the schemes and demonstrate numerically the capabilities of these schemes for a particular Bregman function, i.e., the negative Shannon entropy as defined by the iterative steps (4.12) for solving large scale problems in image reconstruction from projections. We will go ahead and compare numerically the performance of the algorithm MART and the algorithm for entropy maximization over linear equalities as defined in (4.12). It will also enable us to compare the efficiencies of our proposed algorithms for maximization of Shannon entropy with that of the existing similar algorithms in the literature [22].

Finally, we will work on a MATLAB package with implementation of MART and SMART algorithms for discretization of inverse problems. In this regard, we shall provide a few simplified 2D and 3D tomography test problems from x-ray CT and then undertake a series of experiments to determine the importance of these algorithms with underrelaxation to image quality.

Bibliography

[1] Agmon S. 1954, The Relaxation Method for Linear Inequalities, *Canad. J. Math*, Vol. 6, pp. 382-392.

[2] Aharoni R., Duchet P. and Wajnryb B. 1984, Successive Projections on Hyperplanes, *Journal of Mathematical Analysis and Applications*, Vol. 103, pp. 134-138.

[3] Aharoni R., Breman A. and Censor Y 1983, An Interior Points Algorithm for the Convex Feasibility Problem, *Advances in Applied Mathematics*, Vol. 4, pp. 479-489.

[4] Aharoni R., Censor Y., De Pierro A. R. and Zaknoon M. 2004, On the convergence of sequential projection algorithms for the inconsistent convex feasibility problem, submitted.

[5] Aharoni R. and Censor Y. 1989, Block-iterative projection methods for parallel computation of solutions to convex feasibility problems, *Linear Algebra and its Applications*, Vol. 120, pp. 165-175.

[6] Avriel M. 1976, Nonlinear Programming: Analysis and Methods, *Prentice Hall, Englewood Cliffs, NJ.*

[7] Bauschke H. H. and Combettes P. L. 2011, Convex Analysis and Monotone Operator Theory in Hilbert Spaces, *Canadian Mathematical Society*, Address: Halifax, Nova Scotia, Canada.

[8] Beyer W. A. and Ulam S. M. 1968, Note on the visual hull of a set, *J. Comb.*

Theory, Vol. 4, 240-245.

[9] Bertsekas D. P. and Tsitsiklis J. N. 1996, Neuro-Dynamic Programming, *Athena Scientific, Bel-mont, MA.*, address: Athena Scientific, Belmont Mass. U.S.A

[10] Bertsekas D. P. 2003, Nonlinear Programming *Athena Scientific, Belmont, Massachusetts*, address: Athena Scientific, Belmont Mass. U.S.A

[11] Bertsekas D. P. 2003, Convex Analysis and Optimization, *Athena Scientific, Belmont, Massachusetts.*, address: Athena Scientific, Belmont Mass. U.S.A

[12] Boyd S. and Vandenberghe L. 2009, Convex Optimization, *Cambridge University Press*, Address: The Edinburgh Building, Cambridge, UK.

[13] Bregman L. M., Censor Y. and Reich S. 1999, Dykstra's Algorithim as the Nonlinear Extension of Bregman's Optimization Method, *J. Convex Analysis*, Vol. 6, pp. 319-333.

[14] 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, *U.S.S.R. 'Com-put.' Math. and Math. Phys,*, Vol. 7, pp. 200-217.

[15] Browne J. A. and De Pierro A. R. 1996, A row-action alternative to the EM algorithm for maximizing likelihoods in emission tomography, *IEEE Transactions on Medical Imaging*, Vol. 15, pp. 687-699.

[16] Burg J. P. 1967, Maximum Entropy Spectral Analysis, *Proceedings 37th Annual Meeting of the Society of Exploration Geophysicists, Oklahoma City, OK.*

[17] Burg J. P. 1967, Maximum Entropy Spectral Analysis, *Proceedings 37th Annual Meeting of the Society of Exploration Geophysicists*, Vol. 3

[18] Byrne C. 1996, Block iterative methods for image reconstruction from projections, *IEEE Trans. Image Process.* IP-5, 792-794.

[19] Byrne C. 1994, Iterative Reconstruction Algorithms Based on Cross-Entropy Minimization, in Image Models (and their Speech Model Cousins), S. E. Levinson and L. Shepp, Editors, IMA Volumes in Mathematics and its Applications, Vol. 80, pp. 1-11 New York: Springer-Verlag.

[20] Censor Y. 1979, Row-action methods for huge and sparse systems and their applications, *Tech. Rep. MIPG 28, Medical Image Processing Group, Dept. of Computer Science, State Univ. of New York*.

[21] Censor Y. 1981, "Row-action methods for huge and sparse systems and their applications", SIAM Rev., Vol. 23, pp. 444-466.

[22] Censor Y., De Pierro A. R. and Iusem A. N. 1991, Optimization of Burg's Entropy over Linear Constraints, *Applied Numerical Mathematics*, Vol. 7, pp. 151-165.

[23] Censor Y., De Pierro A. R., Elfving T., Herman G.T, and Iusem A. N. 1990, On Iterative Methods for Linearly Constrained Entropy Maximization, *Numerical Analysis and Mathematical Modelling, Banach Center Publications*, Vol. 24, pp. 145-163.

[24] Censor Y., De Pierro A. R. and Zaknoon M. 2004, Steered Sequential Projections for the Inconsistent Convex Feasibility Problem, *Transportation Res.*, Vol. 59, pp. 385-405.

[25] Censor Y., Eggermont P. P. B. and Gordon D. 1983, Strong underrelaxation in Kaczmarz's method for inconsistent systems, *Numer. Math.* Vol. 41, pp. 83-92.

[26] Censor Y., Elfving T. and Herman G.T. 1977, Maximum-Entropy and Bayesian Spectral Analysis and Estimation Problems, *Special-purpose algorithms for linearly constrained entropy maximization, in: C.R. Smith and G.J. Erickson, eds.* , pp.241-254.

[27] Censor Y., Elfving T. and Herman G.T. 1983, Methods for Entropy Maxi-

mization with Applications in Image Processing, in: P. Johansen and P.W. Becker, eds., *Proceedings Third Scandinavian Conference on Image Analysis* , pp. 296-300.

[28] Censor Y. and Herman G.T. 1979, Row-generation methods for feasibility and optimization problems involving sparse matrices and their applications, *Sparse Matrix Proceedings 1978, I.S. Duff and G. W. Stewart, eds., Society for Industrial and Applied Mathematics, Philadelphia*, Vol. 4, pp. 197-219.

[29] Censor Y. and Herman G. T. 2002, Block-Iterative Algorithms with Under-relaxed Bregman Projections, *SIAM Journal on Optimization*, Vol. 13, pp. 283-297.

[30] Censor Y. and Lent A. 1981, An Iterative Row-action Method for interval convex programming, *Journal of optimization Theory and Applications*, Vol. 34, pp. 321-353.

[31] Censor Y. and Segmen J. 1987, On block-iterative Entropy Maximization *Journal of Information and Optimization Sciences*, Vol. 8, pp. 275-291.

[32] Censor Y. and Zenios S. A. 1997, Parallel Optimization; Theory, Algorithms, and Applications, *Oxford University Press, Inc.*, address: New York, USA.

[33] Chaudhuri B. B. and Rosenfeld A. 1998, On the computation of the digital convex hull and circular hull of a digital region, *Pattern Recognition*, Vol. 31, 2007-16.

[34] Cimmino G. 1938, Calcoho approssimato per le soluzioni dei sistemi di equazioni lineari, *La Ricerca Scientifica, Roma*, Vol. XVI, Ser. II. Anno IX, pp. 326-333.

[35] Csiszàr I. and Tusnády G. 1984, Information geometry and alternating minimization procedures', in Statistics and Decisions, Supplement Issue No. 1, R. Oldenbourg Verlag, pp. 205-237.

[36] Daniel J. W. 1971, The Approximate Minimization of Functionals, *Prentice-*

Hall, Address:Englewood Cliffs, NJ.

[37] Darroch J. N. and Ratcliff D. 1972, Generalized iterative scaling for log-linear models, *Annals of Mathematical Statistics*, Vol. 43, 1470-80.

[38] Dempster A. P., Laird N. M. and Rubin D. D. 1977, Maximum likelihood from incomplete data via the EM algorithm, *J. Roy. Stat. Soc.*, Series B, Vol. 39, pp. 1-38.

[39] De Pierro A. R. 1995, A modified expectation maximization algorithm for penalized likelihood estimation in emission tomography, *IEEE Trans. Med. Imaging*, Vol. 14, 1, pp. 132-137.

[40] De Pierro A. R. and Iusem A. N. 1985, "A simultaneous projection method for linear inequalities", *Linear Algebra and its Applications*, Vol. 64, 243-253.

[41] De Pierro A. R. and Iusem A. N. 1986, A Relaxed Version of Bregman's Method for Convex Programming, *Optimization Theory and its Applications*, Vol. 5, pp. 421-440.

[42] De Pierro A. R. and Iusem A. N. 1988, A Finitely Convergent "'Row-Action"' Method for the Convex Feasibility Problem, *Applied Mathematics and Optimization*, Vol. 17, pp. 225-235.

[43] De Pierro A. R. and Iusem A. N. 1989, On the Asymptotic Behavior of Some Alternate Smoothing Series Expansion Iterative Methods, *Applied Numerical Mathematics*, Vol. 29, pp. 421-440.

[44] De Pierro A. R. and Iusem A. N. 1990, A Finitely Convergent "'Row-Action"'Method for the Convex Feasibility Problem, *Mathematical Programming*, Vol. 47, pp. 37-51.

[45] De Pierro A. R. and Iusem A. N. 1991, On The Convergence of Han's Method for Convex Programming with Quadratic Objective, *Mathematical Programming*, Vol. 52, pp. 265-284.

[46] D'Addario L. R. 1977, Maximum-Entropy Imaging, *Theory and philosophy, in: R. Shaw, ed.,Image Analysis and Evaluation (Society of Photographic Scientists and Engineers(SPSE))*, pp. 221-225.

[47] Elfving T. 1980, One Some Methods for Entropy Maximization and Matrix Scaling, *Linear Algebra and its Applications*, Vol. 34, pp. 329-343.

[48] Erlander S. 1981, Entropy in Linear Programmes, *Mathematical Programming*, Vol. 21, pp. 137-151.

[49] Eggermont P.P.B., Herman G.T. and Lent A. 1981, Iterative Algorithms for Large Partitioned Linear Systems, with Applications to Image Reconstruction, *Linear Algebra Appl.*, Vol. 40, pp. 37-67.

[50] Frieden B.R. 1980, Statistical Models for the Image Restoration Problem, *Comput. Graph Image Process*, Vol. 12, pp. 40-59.

[51] Gordon R., Bender R. and Herman G. T. 1970, Algebraic Reconstruction Techniques (ART) for three-dimensional electron microscopy and x-ray photography, *J. Theor. Biol.*, Vol. 29, 471-81.

[52] Gordon R. and Herman G. T. 1974, Three dimensional reconstruction from projections: a review of algorithms, *Int. Rev. Cytol.*, Vol. 38, 111-51.

[53] Gordon R, Bender R. and Herman G. T. 1970, Algebraic reconstruction techniques (ART) for three dimensional electron microscopy and x-ray photography, *Journal of Theoretical Biology*, Vol. 29, pp. 471-481.

[54] Gubin L., Polyak B. and Raik E. 1967, The Method of Projections for Finding the Common Point of Convex Sets, *USSR Computational Mathematics and Mathematical Physics*, Vol. 7, pp. 1-24.

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

[56] Herman G.T. and Lent A. 1978, A Family of Iterative Quadratic Optimiza-

tion Algorithms for Pairs of Inequalities, with Applications in Diagnostic Radiology, *Mathematical Programming Study*, Vol. 9, pp. 15-29.

[57] Herman G. T. and Lent A. 1976, Iterative Reconstruction Algorithms, *Computers in Biology and Medicine*, Vol. 6, pp. 273-294.

[58] Hildreth C. 1957, Quadratic Programming Procedure, *Michigan Agricultural Experiment Station*, Issue No. 361, Vol. 4, pp. 79-85.

[59] Hoffman A. J. 1952, Convex Analysis and Monotone Operator Theory in Hilbert Spaces, *J. Res. Nat. Bur. Standards*, A Vol. 49, pp. 263-265.

[60] Hounsfield G. N. 1973, Computerized transverse axial scanning tomography, Part I: Description of the system, *British Journal of Radiology*, Vol. 46, pp. 1016-1022.

[61] Iusem A. N. and Svaiter B. F. 1994, A Row-Action Method for Convex Programming *Mathematical Programming*, Vol. 64, pp. 149-171.

[62] Iusem A. N. and De Pierro A. R. 1987, A Simultaneous Iterative Method for Computing Projections on Polyhedra, *Siam J. Control and Optimization*, Vol. 25, pp. 231-242.

[63] Iusem A. N. and De Pierro A. R. 1990, On The Convergence Properties of Hildreth's Quadratic Programming Algorithm, *Discrete Mathematics*, Vol. 47, pp. 37-51.

[64] Iusem A. N. and De Pierro A. R. 1991, Optimization of Burg's Entropy over Linear Constraints, *Applied Numerical Mathematics*, Vol. 46, pp. 265-284.

[65] Jaynes E. T. 1982, On the rational of maximum entropy methods, *Proceedings of the IEEE*, Vol. 70, pp. 939-952.

[66] Johnson R. and Shore J. E. 1984, Which is the better Entropy Expression for Speech Processing:-SlogS or logS, *IEEE Trans. Acoust. Speech Signal Process*, Vol. 32, pp. 129-136.

[67] Kaczmarz S. 1937, Angenaherte Auflosung von Systemen linearer Gleichungen, *Bull. Acad. Polon. Sci. Lett.*, Vol. 35, pp. 355-357.

[68] Kapur J. N. 1983, A Comparative Assessment of Various Measures of Entropy, *J. Information Optimization Sciences*, Vol. 4, pp. 207-232.

[69] Kullback S. and Leibler R., 1951, "On information and sufficiency", *Ann. Math. Stat.* Vol. 22, 1, pp. 79-86.

[70] Lamond B. and Stewart N. P. 1981, Statistical Models for the Image Restoration Problem, *Transportation Res.*, Vol. 15B, pp. 239-248.

[71] Lent A. 1977, A convergent algorithm for maximum entropy image restoration with a medical x-ray application", In: Shaw, *Image Analysis and Evaluation*, pp. 249-257, Washington, D.C., Society of Photographic Scientists and Engineers.

[72] Lent A. and Censor Y. 1980, Extensions of Hildreth's Row-Action Method For Quadratic Programming, *Siam J. Control and Optimization*, Vol. 18, pp. 444-454.

[73] Meshulam R. 1996, On Products of Projections, *Discrete Mathematics*, Vol. 154, pp. 307-310.

[74] Nashed M. Z. 1981, Continuous and semicontinuous analogues of iterative methods of Cimmino and Kaczmarz with applications to inverse Radon transform, Mathematical Aspects of Computerized Tomography, *Lectures Notes in Medical Informatics, Springer Verlag Berlin*, Editors:Herman G. T. and Natterer F. Vol. 8, 160-178.

[75] Nowak R. D. 2003, Distributed EM algorithms for density estimation and clustering in sensor networks, *IEEE Trans. Signal Process.*, Vol. 51, pp. 2245-2253.

[76] Ostrowski A. M. 1973, *Solution of Equations in Euclidean and Banach Spaces*, New York, Academic Press.

[77] Polyak B. T. 1987, Introduction to Optimization, *New York: Optimization Software.*

[78] Rockafellar R. T. 1970, Convex Analysis, *Princeton University Press.*, address: Princeton, New Jersey, USA

[79] Shannon C. E. 1948, A Mathematical Theory of Communication, *Bell System Technical Journal*, Vol. 27, pp. 379-423.

[80] Solodov M. V. 2003, Incremental gradient algorithms with stepsizes bounded away from zero, *"'Com-put."' Optim. Appl.*, Vol. 11, pp. 23-35.

[81] Trummer M. R. 1984, Short Communications / Kurze Mitteilungen, A Note on the ART of Relaxation, *Computing by Springer-Verlag*, Vol. 33, pp. 349-352.

[82] -,-. 1986, A relaxed version of Bregman's method for convex programming, *J. Optim. Theory and Application*, Vol. 51, 421–440.