Asymmetric Bregman Forward-Backward Splitting with Projection Correction

Max Nilsson

LUND UNIVERSITY

Department of Automatic Control

MSc Thesis
TFRT-6187
ISSN 0280-5316

Department of Automatic Control
Lund University
Box 118
SE-221 00 LUND
Sweden

Printed in Sweden by Tryckeriet i E-huset
Lund 2022

Abstract

This thesis examines first-order Bregman algorithms in a primal and a primal-dual setting. The Bregman gradient descent algorithm is introduced from a majorization-minimization perspective and as a generalization of the gradient descent algorithm. Concepts such as relative smoothness and Legendreness are defined and are shown to be natural restrictions in order to show convergence results.

A special case of the NOFOB algorithm, proposed by Giselsson in 2021, with a Bregman setting is defined, which we call the Bregman NOFOB algorithm. This algorithm works in a primal-dual setting and consists of a nonlinear forward-backward splitting step followed by a projection correction. Both of these components are discussed with respect to the Bregman setting. The Bregman NOFOB framework unifies multiple algorithms, one of which is the celebrated Bregman Chambolle-Pock method. It also allows us to define novel Bregman primal-dual algorithms. Under certain assumptions on the solution set of the problem and on the projection steps, we show that the Bregman NOFOB method converges in duality gap.

This Bregman NOFOB algorithm with asymmetric kernel is compared with the Wolfe-Atwood (WA) algorithm on the D-optimal design optimization problem. As confirmed by the theory of this thesis - in the primal case - the two algorithms both converge by sequence and by function value, with sublinear (Bregman NOFOB) and linear (WA) rates. In the primal-dual case we experimentally show that the projection step sizes satisfy the duality gap convergence of the Bregman NOFOB algorithm. Indeed, this duality gap convergence is also verified experimentally. No comparison with the WA algorithm is made in the primal-dual case, since it is restricted to the primal setting.

Contents

1

Introduction

1.1 Preliminary Example

Consider the minimization problem

$$\underset{x_1,x_2\geq 0}{\text{minimize}} \;\; x_1 + x_2.$$

If we let $f : \mathbb{R}^2 \to \mathbb{R}$ be defined by

$$f(x) = x_1 + x_2$$

for all $x \in \mathbb{R}^2$ then we can rewrite our problem as

$$\underset{x\in\mathbb{R}^2}{\text{minimize}} \;\; f(x) + \iota_{\mathbb{R}^2_+}(x),$$

see Section 1.4 for a definition of the indicator function $\iota_{\mathbb{R}^2_+}$. It is immediate that the problem has the unique solution $x^\star = 0$. The function f is continuously differentiable and convex. Furthermore, the function f is β-smooth (see again Section 1.4 for the definition of β-smoothness) for each $\beta \geq 0$. Naively, one could apply the gradient descent algorithm to f in hope of solving the minimization problem. The update step of the gradient descent method is given by

$$x_{k+1} = x_k - \gamma \nabla f(x_k)$$

for some step-size $\gamma > 0$. But we will quickly run into trouble. As soon as one of the coordinates of x_k becomes negative, then we fall outside the effective domain of $f + \iota_{\mathbb{R}^2_+}$ and the function values of $f + \iota_{\mathbb{R}^2_+}$ become infinite.

We must let the constraints of the problem be reflected in the chosen algorithm in some way. One common algorithmic choice is the proximal gradient method. The proximal gradient method consists of a forward and a backward step. The update step of the proximal gradient method can be written as (see [18, Section 2.7])

$$x_{k+1} = [I + \gamma \partial \iota_{\mathbb{R}^2_+}]^{-1}[I - \gamma \nabla f] x_k.$$

The forward step is the part

$$x_k^+ \coloneqq [I - \gamma \nabla f] x_k = x_k - \gamma \nabla f(x_k)$$

and corresponds to the update step of the gradient descent method. The backward step is the part

$$[I + \gamma \partial \iota_{\mathbb{R}^2_+}]^{-1} x_k^+.$$

In this case, the backward step projects x_k^+ onto the set $\mathbb{R}^2_+$. Therefore, the proximal gradient method has reflected the constraints of the problem by a backward step.

In contrast, the Bregman gradient descent method incorporates the constraints in the forward step of its algorithm. The Bregman gradient descent method will be the main focus of Chapter 2, but we will here briefly introduce its main idea.

The reason that we cannot guarantee convergence of the gradient method in our problem is that $f + \iota_{\mathbb{R}^2_+}$ is no longer β-smooth for any $\beta > 0$. In fact, there exists no quadratic majorizers of $f + \iota_{\mathbb{R}^2_+}$. The Bregman gradient descent method asks the question if we can instead find some other type of majorizer and from that majorizer define an update step analogous to the gradient descent method. The type of function which defines this majorizer is what we will call a Bregman function and we will often denote it by h. In this particular case, the Bregman function

$$h(x) = \begin{cases} -\log(x_1) - \log(x_2) & \text{if } \ x \in \mathbb{R}^2_{++} \\ \infty & \text{otherwise} \end{cases} \tag{1.1}$$

suits our problem, but more on that in Chapter 2. Moreover, in Chapter 2 we will derive the update step of the Bregman gradient descent method (see (2.13)). Note that the effective domain of h equals that of $f + \iota_{\mathbb{R}^2_+}$. This should indicate that we have captured some of the problem geometry with our algorithm.

See Figure 1.1 of how the first five iterations of the proximal gradient method compares with the first five iterations of the Bregman gradient descent method.

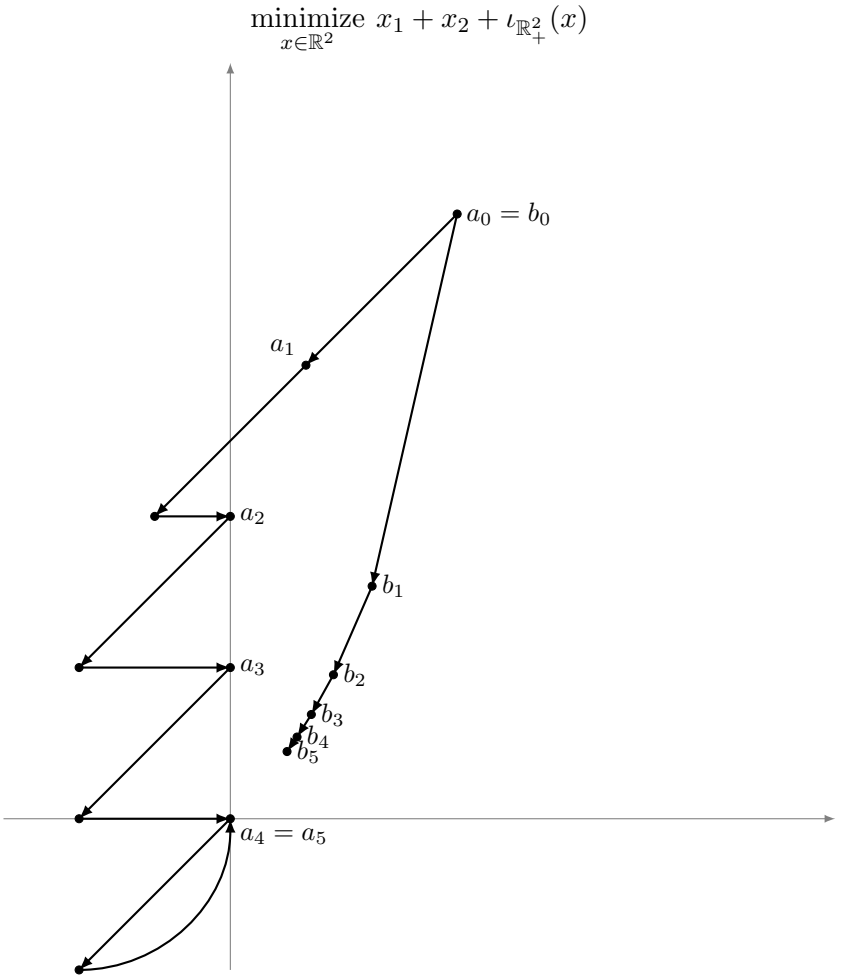


Figure 1.1 Comparing the first five iterations of the proximal gradient descent $\{a_k\}_{k=0}^{\infty}$ with the Bregman gradient descent $\{b_k\}_{k=0}^{\infty}$ on the objective function $f(x) = x_1 + x_2$ with the added constraint $x \in \mathbb{R}^2_+$. The Bregman function $h(x) = -\sum_{i=1}^{2} \log(x_i)$ was used in the Bregman gradient descent algorithm. In this example, both algorithms are implemented with a constant step-size.

1.2 Background

First-order Bregman type methods date back to classical treatments, such as [15], under the name of mirror descent. Since its introduction, it has inspired a great number of researchers and laid the groundwork for multiple algorithms. See for example [6] for a treatment of random Bregman projections

and [7] for Bregman analysis in the context of monotone operator theory in a general Banach space. Another example is a nonlinear proximal point algorithm based on Bregman functions in [12].

Despite - or perhaps because of - the longevity of the research of Bregman functions, it is still an active theoretical and applied research area. A new descent Lemma, with respect to Bregman distances, was developed in [4] which led to a natural derivation of the proximal-gradient scheme with Bregman distances with applications including solving the Poisson inverse problem. Relative smoothness with respect to Bregman functions, and its connection with convergence analysis, was introduced in [14], with interesting applications such as solving the classical D-optimal design problem. A stochastic first-order Bregman method was examined in [3] in the context of overparameterized nonlinear models.

First-order Bregman type methods have recently had applications in a primal-dual problem setting. In [10] the Chambolle-Pock method was introduced and extended in [11] to a Bregman variant. A drawback of this non-linear primal-dual algorithm is that it imposes some strong restrictions on the utilized Bregman functions. For instance, the two Bregman functions in the algorithm are both assumed to be 1-strongly convex. This restriction excludes for example the Bregman function defined in (1.1).

We will in Section 3.3 propose a new method with inspiration from the NOFOB algorithm [13] that tackles these limitations. We will call this method the Bregman NOFOB algorithm (see Algorithm 1 in Section 3.3). In particular, we will see that the Bregman Chambolle-Pock method is a special case of the Bregman NOFOB algorithm. We will in this thesis take the first steps towards a complete convergence analysis of the Bregman NOFOB algorithm.

1.3 Outline

In Chapter 2 we will introduce the Bregman function from an intuitive standpoint. We will in Section 2.1 follow the steps of the gradient descent convergence analysis in a more general setting, with an undefined function D_h called the Bregman distance. We will there see what properties we want our undefined function D_h to have in order for the Bregman gradient descent algorithm to converge. In Section 2.2 we properly define the Bregman function h and its associated distance D_h. We will also show what restriction we need to put on the Bregman function in order for the convergence analysis to hold. In Section 2.2 we introduce the symmetry measure α_h of a Bregman function h and how it relates to the step-sizes of the Bregman gradient descent. Lastly, in Section 2.4 we will extend the Bregman gradient descent algorithm to include a proximal part.

In Chapter 3 we will change the setting to a primal-dual one. The algo-

rithm called nonlinear forward-backward splitting with projection correction (NOFOB), introduced in [13], will be treated in a Bregman setting. The two steps of the algorithm: the forward-backward and the projection correction step, will be discussed in Section 3.1 and 3.2 respectively. In Section 3.3 these steps will be combined and form the novel Bregman NOFOB algorithm. In a primal-dual problem setting, it will be shown that the Bregman NOFOB algorithm generalizes the Bregman Chambolle-Pock algorithm [11]. Furthermore, the Bregman NOFOB algorithm has a special case with asymmetric updates, which yields a flexible choice of Bregman function.

This special case with asymmetric updates will in Chapter 4 be applied to the D-optimal design problem. Section 4.1 introduces the D-optimal design problem from a statistical and geometrical viewpoint. In Section 4.2 we compare the Bregman NOFOB algorithm with the Wolfe-Atwood algorithm [19] on some numerical experiments.

1.4 Notation and Definitions

In this section, we will collect the most common notation and definitions used in the upcoming chapters.

We will let $\mathbb{R}$ denote the real numbers and $\mathbb{R}^n_+$ be the subset of $\mathbb{R}^n$ which contains vectors with only nonnegative components

$$\mathbb{R}^n_+ \coloneqq \{x \in \mathbb{R}^n \mid x_i \geq 0 \text{ for all } i = 1, \ldots, n\}.$$

The set $\mathbb{R}^n_{++}$ is the subset of $\mathbb{R}^n$ which contains vectors with only strictly positive components. The subsets $\mathbb{R}^n_-$ and $\mathbb{R}^n_{--}$ are defined analogously. If $x, y \in \mathbb{R}^n$ we let

$$[x, y] \coloneqq \{z \in \mathbb{R}^n \mid z = \theta x + (1 - \theta)y \text{ for some } \theta \in [0, 1]\}.$$

Furthermore, we define $\overline{\mathbb{R}} \coloneqq \mathbb{R} \cup \{\infty\}$. We will let $\langle \cdot, \cdot \rangle$ denote the ordinary inner product on $\mathbb{R}^n$, defined by

$$\langle x, y \rangle = x^T y = \sum_{i=1}^{n} x_i y_i.$$

If A and B are two sets, we write $A \subset B$ if A is a subset of B. The power set of A is denoted by 2^A. The interior, relative interior, closure, and boundary of A are denoted by $\operatorname{int}(A)$, $\operatorname{relint}(A)$, $\overline{A}$, and $\operatorname{bd}(A)$ respectively. Recall that a function $f : A \to B$ is called a homeomorphism if f is bijective and both f and f^{-1} are continuous (with respect to some topologies defined on A and B). We will let $I : A \to A$ equal the identity function and the set

A will be clear from the context. The indicator function of a set $S \subset \mathbb{R}^n$ is denoted by $\iota_S : \mathbb{R}^n \to \overline{\mathbb{R}}$ and defined as

$$\iota_S(x) = \begin{cases} 0, & x \in S \\ \infty, & x \notin S. \end{cases}$$

Let $f : \mathbb{R}^n \to \overline{\mathbb{R}}$. We define the effective domain of f as

$$\operatorname{dom} f \coloneqq \{x \in \mathbb{R}^n \mid f(x) < \infty\}.$$

Recall that the function f is proper and closed if $\operatorname{dom} f \neq \emptyset$ and its epigraph $\{(x, r) \in \mathbb{R}^n \times \mathbb{R} \mid f(x) \leq r\}$ is closed, respectively. As usual, $\partial f : \mathbb{R}^n \to 2^{\mathbb{R}^n}$ denotes the subdifferential of f defined by

$$\partial f(x) = \{s \in \mathbb{R}^n \mid f(y) \geq f(x) + \langle s, y - x \rangle \text{ for all } y \in \mathbb{R}^n\}$$

for each $x \in \mathbb{R}^n$. The conjugate function $f^* : \mathbb{R}^n \to \overline{\mathbb{R}}$ is defined by

$$f^*(s) = \sup_{x \in \mathbb{R}^n} [\langle s, x \rangle - f(x)]$$

for each $s \in \mathbb{R}^n$. The level set of f at $\xi \in \mathbb{R}$ is defined by

$$\operatorname{lev}_\xi f \coloneqq \{x \in \mathbb{R}^n \mid f(x) \leq \xi\}.$$

A differentiable function f with $\operatorname{dom} f = \mathbb{R}^n$ is called β-smooth, for some $\beta \geq 0$, if

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\beta}{2} \|y - x\|_2^2$$
$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle - \frac{\beta}{2} \|y - x\|_2^2$$

hold for all $x, y \in \mathbb{R}^n$.

Let the sequence $\{x_k\}_{k=0}^\infty$, with $x_k \in \mathbb{R}^n$, be defined by the iterates from an algorithm, which seeks to minimize an objective function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$. Then we say that the algorithm converges by sequence if $\{x_k\}_{k=0}^\infty$ converges to some solution $x^\star$ of the minimization problem. Furthermore, $\{x_k\}_{k=0}^\infty$ converges by (function) value if $\{f(x_k)\}_{k=0}^\infty$ converges to the optimum value denoted by $f^\star$. Note that a sequence need not converge by sequence even if it converges by value.

Let $g : \mathbb{R}^n \to \mathbb{R}$ be a function with minimizing value $g^\star$. We define

$$\operatorname*{Argmin}_{x \in \mathbb{R}^n} g(x) = \{x \in \mathbb{R}^n \mid g(x) = g^\star\}.$$

The proximal operator of a function $\phi : \mathbb{R}^n \to \overline{\mathbb{R}}$, denoted by $\text{prox}_\phi : \mathbb{R}^n \to 2^{\mathbb{R}^n}$, is defined as

$$\text{prox}_\phi(z) \coloneqq \underset{x \in \mathbb{R}^n}{\text{Argmin}} \left[\phi(x) + \frac{1}{2} \|x - z\|_2^2 \right].$$

If ϕ is closed convex, then $\text{prox}_p \, hi$ is a single-valued mapping from $\mathbb{R}^n$ to $\mathbb{R}^n$.

An operator $M : \mathbb{R}^n \to 2^{\mathbb{R}^n}$ is monotone if for all $x, y \in \mathbb{R}^n$ and $u \in Mx$ and $v \in My$ if the following holds

$$\langle u - v, x - y \rangle \geq 0.$$

The zero-set of M is defined by

$$\text{Zer}[M] \coloneqq \{x \in \mathbb{R}^n \mid 0 \in Mx\}$$

and the fixed-point set of M by

$$\text{Fix}[M] \coloneqq \{x \in \mathbb{R}^n \mid x \in Mx\}.$$

A single-valued operator $C : \mathbb{R}^n \to \mathbb{R}^n$ is called $\frac{1}{\beta}$-cocoercive if for all $x, y \in \mathbb{R}^n$ the following holds

$$\langle Cx - Cy, x - y \rangle \geq \frac{1}{\beta} \|Cx - Cy\|_2^2 .$$

Let $A, B \in \mathbb{R}^{m \times n}$. The Hadamard product of A and B is denoted by $A \circ B \in \mathbb{R}^{m \times n}$ and is given by $(A \circ B)_{ij} = A_{ij} B_{ij}$. Recall that a matrix $C \in \mathbb{R}^{n \times n}$ is called skew-symmetric if $C^T = -C$. Let $u \in \mathbb{R}^n$ then $\text{Diag}(u) \in \mathbb{R}^{n \times n}$ denotes the diagonal matrix with u as its diagonal. Conversely, if $U \in \mathbb{R}^{n \times n}$ then $\text{diag}(U) \in \mathbb{R}^n$ is the diagonal vector of U. If a matrix $X \in \mathbb{R}^{n \times n}$ is symmetric and positive definite we write $X \succ 0$. The set $\mathbb{S}_{++}^n$ is the set of $n \times n$ positive definite (and symmetric) matrices

$$\mathbb{S}_{++}^n \coloneqq \{X \in \mathbb{R}^{n \times n} \mid X \succ 0\}.$$

We will as usual equip $\mathbb{S}_{++}^n$ with the standard inner product $\langle X, Y \rangle = \text{tr}(XY)$.

2

Bregman Primal Method

2.1 The Bregman Distance and Bregman Primal Method

Let $f : \mathbb{R}^n \to \mathbb{R}$ be some differentiable function and consider the standard primal minimization problem

$$\underset{x \in \mathbb{R}^n}{\text{minimize}}\ f(x). \tag{2.1}$$

Suppose further that f is β-smooth, which implies that for every x, y in $\mathbb{R}^n$ we have a following inequality

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\beta}{2}||y - x||_2^2. \tag{2.2}$$

The following update step is the well-known gradient method update step

$$x_{k+1} = x_k - \gamma_k \nabla f(x_k)$$

with a positive step-size γ_k. This update step gives rise to an algorithm which we hope will converge to some solution $x^\star$ of (2.1). In fact, the convergence to some minimizer $x^\star$ occurs under certain stronger conditions on f and on the step size γ_k. One such pair of conditions is that f is convex and that γ_k lies in the interval $[\varepsilon, \frac{2}{\beta} - \varepsilon]$ for all k, for some positive ε.

Now suppose we have a bounded step-size $0 < \gamma_k < \frac{1}{\beta}$. In this case, the gradient method update step can be viewed as an update step of a certain *majorization-minimization* algorithm. There exists some function $g : \mathbb{R}^n \to \mathbb{R}$ with $g \geq f$ such that

$$x_{k+1} = \underset{y \in \mathbb{R}^n}{\text{argmin}}\ g(y).$$

Take $g(y)$ to equal the right hand side of (2.2) and insert $x = x_k$. Minimizing g then turns into a quadratic programming problem:

$$\underset{y \in \mathbb{R}^n}{\text{argmin}}\ g(y) = \underset{y \in \mathbb{R}^n}{\text{argmin}}\ \langle \nabla f(x_k), y \rangle + \frac{\beta}{2}||y - x_k||_2^2.$$

The right hand side is a minimization of a strongly convex and differentiable function and by the first order optimality condition has a minimum x_{k+1} satisfying

$$0 = \nabla f(x_k) + \beta(x_{k+1} - x_k) \implies x_{k+1} = x_k - \frac{1}{\beta}\nabla f(x_k).$$

By a simple modification of g and using that $\beta < \frac{1}{\gamma_k}$ we can by an analogous argument arrive at the general gradient method update step.

One part to generalize in the previous discussion is the part played by the distance function $D : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ given by

$$D(y, x) = \frac{1}{2}\|y - x\|_2^2. \tag{2.3}$$

Is there any way to generalize this convergence theory for some different distance $D_h : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ depending on some function $h : \mathbb{R}^n \to \overline{\mathbb{R}}$? The answer is *yes* and the function h is what we later will call a Bregman function. How this function h determines D_h and what properties h should have, will be described in Section 2.2.

Assume from now on that f is convex and assume that we have found a Bregman distance D_h such that the descent condition

$$f(y) \leq f(x) + \langle \nabla f(x), y - x\rangle + \beta D_h(y, x) \tag{2.4}$$

holds. Intuitively, we want D_h to capture some of the geometry of f.

We can then define a majorization-minimization algorithm for any step-size $\gamma_k \in (0, \frac{1}{\beta})$ as

$$x_{k+1} = \operatorname*{argmin}_{y \in \mathbb{R}^n} f(x_k) + \langle \nabla f(x_k), y - x_k\rangle + \frac{1}{\gamma_k} D_h(y, x_k). \tag{2.5}$$

If we for the moment we will suppose that such a minimization problem is well defined, this update step gives rise to an algorithm, that we will call The Bregman Primal method. Let us now prove its convergence.

Convergence of The Bregman Primal Method

The descent condition and the algorithm update imply that the algorithm iterates satisfy

$$f(x_{k+1}) \leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k\rangle + \beta D_h(x_{k+1}, x_k).$$

Here we use that

$$\langle \nabla f(x_k), x_{k+1} - x_k\rangle = \langle \nabla f(x_k), z - x_k\rangle + \langle -\nabla f(x_k), z - x_{k+1}\rangle.$$

Continuing the standard convergence argument, we bound this last term. In the normal gradient descent scheme, when $D_h(y,x) = \frac{1}{2}\|y-x\|_2^2$, we use the fact that

$$\langle -\nabla f(x_k), z - x_{k+1}\rangle = \frac{1}{\gamma_k}\langle x_{k+1} - x_k, z - x_{k+1}\rangle$$

and the identity

$$\langle x-y, z-x\rangle = \frac{1}{2}\left(\|z-y\|_2^2 - \|z-x\|_2^2 - \|x-y\|_2^2\right) \tag{2.6}$$

for all $x, y, z \in \mathbb{R}^n$ to get that

$$\langle -\nabla f(x_k), z - x_{k+1}\rangle \leq \frac{1}{2\gamma_k}\left(\|z-x_k\|_2^2 - \|z-x_{k+1}\|_2^2 - \|x_{k+1}-x_k\|_2^2\right)$$

and proceed from there.

In the general case the equality in (2.6) can be relaxed to an inequality which we will from now on assume holds:

$$\langle -\nabla f(x_k), z - x_{k+1}\rangle \leq \frac{1}{\gamma_k}(D_h(z,x_k) - D_h(z,x_{k+1}) - D_h(x_{k+1},x_k)) \tag{2.7}$$

for all $z \in \mathbb{R}^n$ such that $D_h(z,x_k) \neq \infty$. The convexity of f gives that

$$f(x_k) + \langle \nabla f(x_k), x^\star - x_k\rangle \leq f(x^\star)$$

when combined with (2.7) gives that

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k\rangle + \beta D_h(x_{k+1},x_k) \\ &\leq f(x_k) + \langle \nabla f(x_k), x^\star - x_k\rangle \\ &\quad + \frac{1}{\gamma_k}\left(D_h(x^\star,x_k) - D_h(x^\star,x_{k+1})\right) + \left(\beta - \frac{1}{\gamma_k}\right) D_h(x_{k+1},x_k) \\ &\leq f(x^\star) + \frac{1}{\gamma_k}\left(D_h(x^\star,x_k) - D_h(x^\star,x_{k+1})\right) + \left(\beta - \frac{1}{\gamma_k}\right) D_h(x_{k+1},x_k) \end{aligned}$$

where we in the second inequality use $z = x^\star$ in (2.7) and in the third inequality use the convexity assumption on f. We get the inequality:

$$\begin{aligned} D_h(x^\star,x_{k+1}) \leq D_h(x^\star,x_k) &+ (\beta\gamma_k - 1)D_h(x_{k+1},x_k) \\ &- \gamma_k(f(x_{k+1}) - f(x^\star)). \end{aligned} \tag{2.8}$$

This inequality is a Lyapunov inequality if we further assume that $D_h \geq 0$.

Suppose now that $D_h(x,x) = 0$, which is a natural condition on some function D_h satisfying the descent condition (2.4). Without using the convexity of f and considering $z = x_k$ in (2.7) we can instead arrive at the weaker Lyapunov inequality

$$f(x_{k+1}) \leq f(x_k) - \frac{1}{\gamma_k} D_h(x_k, x_{k+1}) - \left(\frac{1}{\gamma_k} - \beta\right) D_h(x_{k+1}, x_k)$$

$$\implies f(x_{k+1}) - f(x_k) \leq -\frac{1}{\gamma_k} \underbrace{D_h(x_k, x_{k+1})}_{\geq 0} - \underbrace{\left(\frac{1}{\gamma_k} - \beta\right)}_{>0} \underbrace{D_h(x_{k+1}, x_k)}_{\geq 0} \leq 0.$$

This implies that $\{f(x_k)\}_{k=0}^{\infty}$ is a nonincreasing sequence. From this weaker Lyapunov inequality we can extract

$$\frac{1 - \beta\gamma_k}{\gamma_k} D_h(x_{k+1}, x_k) \leq f(x_k) - f(x_{k+1}). \tag{2.9}$$

In order to arrive at suitable telescoping sum, we need a lower bound of $(1 - \beta\gamma_k)/\gamma_k$ that is independent of k. Suppose that the step-sizes γ_k are not only strictly upper bounded by $1/\beta$ but uniformly bounded in the sense that

$$\gamma_k \leq \frac{1 - \varepsilon}{\beta} \iff 1 - \beta\gamma_k \geq \varepsilon$$

for some $\varepsilon \in (0, 1)$. Then we have that

$$\frac{1 - \beta\gamma_k}{\gamma_k} \geq \frac{\beta\varepsilon}{1 - \varepsilon} > 0.$$

Now we can telescope (2.9) and arrive at

$$\frac{\beta\varepsilon}{1 - \varepsilon} \sum_{k=0}^{n} D_h(x_{k+1}, x_k) \leq f(x_0) - f(x_{n+1}) \leq f(x_0) - f(x^\star) < \infty.$$

This implies that $\{D_h(x_{k+1}, x_k)\}_{k=0}^{\infty}$ is summable and $\lim_{k\to\infty} D_h(x_{k+1}, x_k) = 0$.

Let us return to the first Lyapunov inequality (2.8). Since $\beta\gamma_k - 1 < 0$, telescoping summation gives that

$$\sum_{k=0}^{n} \gamma_k (f(x_{k+1}) - f(x^\star)) \leq D_h(x^\star, x_0) - D_h(x^\star, x_{n+1}) \leq D_h(x^\star, x_0). \tag{2.10}$$

Since $\{f(x_k)\}_{k=0}^{\infty}$ is nonincreasining we get that

$$\sum_{k=0}^{n} \gamma_k (f(x_{n+1}) - f(x_{k+1})) \leq 0.$$

Adding this to equation (2.10) we finally get

$$f(x_{n+1}) - f(x^\star) \leq \frac{D_h(x^\star, x_0)}{\sum_{k=0}^{n} \gamma_k}.$$

If $\sum_{k=0}^{\infty} \gamma_k = \infty$, we have function value convergence. For instance, if the step-size is constant or if it is uniformly bounded below by some positive number, we have sublinear $\mathcal{O}(1/k)$ convergence.

Also notice how (2.8) implies the important property:

$$0 \leq D_h(x^\star, x_{k+1}) \leq D_h(x^\star, x_k) \tag{2.11}$$

which implies that $\{D_h(x^\star, x_k)\}_{k=0}^{\infty}$ converges, but not necessarily to zero. It is a quite subtle question whether or not these convergence properties imply sequence convergence, i.e., $\lim_{k\to\infty} x_k$ converges to some solution of the problem. This question will be discussed in the following section.

2.2 Bregman Functions

Let us now state the general definition of a Bregman distance.

Definition 2.2.1. Let $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ be proper, closed, convex, and differentiable on the non-empty set $\operatorname{int}(\operatorname{dom} h)$, then its corresponding **Bregman Distance** $D_h : \mathbb{R}^n \times \operatorname{int}(\operatorname{dom} h) \to \overline{\mathbb{R}}$, is given by

$$D_h(x, y) = h(x) - h(y) - \langle \nabla h(y), x - y \rangle.$$

Note that if $h(x) = \frac{1}{2}\|x\|_2^2$ then $D_h(x, y) = \frac{1}{2}\|x - y\|_2^2$, which is the function from (2.3) we wanted to generalize. Also, saying that D_h a distance is in many ways misguided. For instance it neither satisfies the triangle inequality or respects symmetry about its arguments.

We note that $D_h(\cdot, y)$ is differentiable on $\operatorname{int}(\operatorname{dom} h)$ with gradient

$$\nabla D_h(\cdot, y) = \nabla h(\cdot) - \nabla h(y).$$

Now let us move on with a formal discussion of Bregman functions. We begin by restating the properties we want a Bregman function D_h to have, according to the discussion of the previous section, in order to get a well-defined and convergent Bregman primal method. The following properties should hold for all $x \in \mathbb{R}^n$, $y \in \operatorname{int}(\operatorname{dom} h)$, $z \in \operatorname{dom} h$, and a given objective function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$, which is assumed to be convex. Furthermore, f is assumed to be differentiable on the non-empty set $\operatorname{int}(\operatorname{dom} f)$. Since h is supposed to capture the geometry of f it is natural to assume some properties of their effective domains. We will from now on assume that $\operatorname{dom} h \subset \operatorname{dom} f$.

With this inclusion of the effective domains in mind, we find it natural to suppose that the set

$$\operatorname{argmin}\{f(x) \mid x \in \overline{\operatorname{dom} h}\}$$

is non-empty. Lastly, we will also suppose that the descent condition (2.4) holds for some $\beta > 0$ and that the step-size satisfies $0 < \gamma_k < \frac{1}{\beta}$.

These are the properties that we will prove to hold:

P1. $D_h(y, y) = 0$.

P2. $D_h(x, y) \geq 0$.

P3. $f(x) \leq f(y) + \langle \nabla f(y), x - y \rangle + \beta D_h(x, y)$.

P4. For some fixed y the function $x \mapsto f(y) + \langle \nabla f(y), x - y \rangle + \frac{1}{\gamma_k} D_h(x, y)$, has a unique minimum for all k.

P5. $x_k \in \operatorname{int}(\operatorname{dom} h)$ for all $k \geq 1$ given that $x_0 \in \operatorname{int}(\operatorname{dom} h)$.

P6. $\langle -\gamma_k \nabla f(x_k), z - x_{k+1} \rangle \leq D_h(z, x_k) - D_h(z, x_{k+1}) - D_h(x_{k+1}, x_k)$.

P7. $\lim_{k \to \infty} x_k$ exists and the limit point is a solution of the problem (2.1).

P1, P2, and P3

Some of the basic properties from the previous section that we want from a Bregman distance to uphold can now be stated.

Proposition 2.2.2. *Let $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ be proper, closed, convex, and differentiable on the non-empty set $\operatorname{int}(\operatorname{dom} h)$. Then its corresponding Bregman distance D_h satisfies **P1** and **P2**.*

Proof. We see that

$$D_h(y, y) = h(y) - h(y) - \langle \nabla h(y), y - y \rangle = 0.$$

From the fact that h is convex we get the first order condition

$$h(x) \geq h(y) + \langle \nabla h(y), x - y \rangle$$

for all $x \in \mathbb{R}^n$ and $y \in \operatorname{int}(\operatorname{dom} h)$. Therefore, $D_h(x, y) \geq 0$. ■

The property **P3** is something that does not hold in general, since we have no real constraints on f. In fact, this upper bound of f is captured by the following definition, which we will from now on assume.

Definition 2.2.3. Let $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ and D_h be a Bregman distance with corresponding function $h : \mathbb{R}^n \to \overline{\mathbb{R}}$. We say that f is **β-relatively smooth** with respect to h if for any $x \in \mathbb{R}^n$ and $y \in \operatorname{int}(\operatorname{dom} h)$ there is a $\beta \in \mathbb{R}_+$ for which

$$f(x) \leq f(y) + \langle \nabla f(y), x - y \rangle + \beta D_h(x, y).$$

Remark. After expanding $D_h(x, y)$, we get that the condition for relative smoothness is equivalent with

$$\beta h(y) - f(y) + \langle \beta \nabla h(y) - \nabla f(y), x - y \rangle \leq \beta h(x) - f(x).$$

Therefore, f being β-relatively smooth with respect to h is equivalent with the function $\beta h - f$ being convex. Also notice that if f is β-relatively smoothness with respect to $h = \frac{1}{2} \|\cdot\|_2^2$ then f is β-smooth in the nominal sense.

P4 and P5

Proving properties **P4**, **P5**, and **P7** turns out to be a much more subtle matter. As we will see, **P6** follows by the standard Three-Point Property once we have shown in **P4** and **P5** that the iterations $\{x_k\}_{k=0}^{\infty}$ are well-defined. These properties described in **P4**, **P5**, and **P7** are closely related to the concepts of *essential smoothness*, *essential strict convexity*, and *Legendre*.

Definition 2.2.4. Let $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ be proper, closed, and convex. Then h is **Essentially Smooth** if it is differentiable on the non-empty set $\operatorname{int}(\operatorname{dom} h)$ and if for all convergent sequences $\{x_k\}_{k=0}^{\infty}$ in $\operatorname{int}(\operatorname{dom} h)$ with limit point in $\operatorname{bd}(\operatorname{dom} h)$, we have that

$$\lim_{k \to \infty} \|\nabla h(x_k)\|_2 = \infty.$$

The following proposition more than justifies the somewhat perhaps non-intuitive of essential smoothness.

Proposition 2.2.5 ([17, Theorem 26.1])**.** *Let $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ be proper, closed, and convex with $\operatorname{int}(\operatorname{dom} h) \neq \emptyset$. Then h is essentially smooth if and only if*

$$\partial h(x) = \begin{cases} \{\nabla h(x)\} & \text{if } x \in \operatorname{int}(\operatorname{dom} h) \\ \emptyset & \text{otherwise.} \end{cases}$$

Example 2.2.6. The function $h_1 : \mathbb{R} \to \overline{\mathbb{R}}$ given by

$$h_1(x) = \begin{cases} \frac{1}{2} x^2 & \text{if } x \geq 0 \\ \infty & \text{otherwise.} \end{cases}$$

is proper, closed, convex, and differentiable on $\operatorname{int}(\operatorname{dom} h) = (0, \infty)$ but not essentially smooth. For example, one can see that $\partial h_1(0) = \{0\}$ and use Proposition 2.2.5.

From only essential smoothness of h we can show the following weaker statement of **P5**. For this purpose, let us the define $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ for some fixed y in $\operatorname{int}(\operatorname{dom} h)$ by

$$g(x) = f(y) + \langle \nabla f(y), x - y \rangle + \frac{1}{\gamma_k} D_h(x, y). \tag{2.12}$$

Lemma 2.2.7. *Let $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ be proper, closed, convex, and essentially smooth. Suppose that g in (2.12) is minimized for some $x^\star \in \mathbb{R}^n$. Then*

$$x^\star \in \operatorname{int}(\operatorname{dom} h).$$

Proof. By Fermat's rule we get the optimality condition

$$\begin{aligned} & 0 \in \partial g(x^\star) \\ \iff\ & 0 \in \gamma_k \nabla f(x) + \partial h(x^\star) - \nabla h(x) \\ \iff\ & \nabla h(x) - \gamma_k \nabla f(x) \in \partial h(x^\star). \end{aligned}$$

In the first equivalence, we have used that a constraint qualification trivially holds. The last inclusion can only be true if $\partial h(x^\star)$ is non-empty. By Proposition 2.2.5 we get that $x^\star$ must lie in $\operatorname{int}(\operatorname{dom} h)$. ■

Definition 2.2.8. Let $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ be proper, closed, and convex. Then h is **Essentially Strictly Convex** if it is strictly convex on every convex subset of $\operatorname{dom} \partial h$.

Remark. Essential strict convexity is a weaker condition than strict convexity. In other words, there exists functions h that are essentially strictly convex but are not strictly convex on the convex set $\operatorname{dom} h$. The following example analyses such a function.

Example 2.2.9 ([17, Example in 26])**.** Consider the function $h_2 : \mathbb{R}^2 \to \overline{\mathbb{R}}$ given by

$$h_2(x, y) = \begin{cases} \frac{y^2}{2x} - 2\sqrt{y} & \text{if } x > 0 \text{ and } y \geq 0 \\ 0 & \text{if } x = y = 0 \\ \infty & \text{otherwise.} \end{cases}$$

It is proper and convex. Since it is zero at the origin it is also closed. The function is clearly not strictly convex on $\operatorname{dom} h$ since it is zero along the positive x-axis. If one considers the partial derivative

$$\frac{\partial}{\partial y} h_2(x, y) = \frac{y}{x} - \frac{1}{\sqrt{y}}$$

for $x, y > 0$ then it is evident that for any fixed $x > 0$:

$$\frac{\partial}{\partial y} h_2(x, y) \to -\infty \text{ as } y \to 0.$$

With a similar analysis for the origin we can conclude that $\operatorname{dom} \partial h_2 = \mathbb{R}_{++}$. We can compute the Hessian of h_2 on $\mathbb{R}_{++}$ as

$$\nabla^2 h_2(x, y) = \begin{bmatrix} \frac{y^2}{x^3} & -\frac{y}{x^2} \\ -\frac{y}{x^2} & \frac{1}{x} + \frac{1}{2y^{3/2}} \end{bmatrix}$$

which has determinant $\frac{\sqrt{y}}{2x^3} > 0$ and so the Hessian is positive definite for all $x, y > 0$. Therefore h_2 is not strictly convex on $\operatorname{dom} h$ but is strictly convex on $\operatorname{dom} \partial h$ and essentially strictly convex.

By Proposition 2.2.5 we have also shown that h_2 is essentially smooth.

Example 2.2.10. The function h_1 from example 2.2.6 has conjugate function

$$h_1^*(\mu) = \begin{cases} \frac{1}{2}\mu^2 & \text{if } \mu \geq 0 \\ 0 & \text{otherwise,} \end{cases}$$

which is not essentially strictly convex. We have here an example of a function h_1 which is not essentially smooth and has a conjugate that is not essentially strictly convex. This is in fact is not a coincidence.

Proposition 2.2.11 ([17, Theorem 26.3])**.** *Let $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ be proper, closed, and convex. Then h is essentially strictly convex if and only if h^* is essentially smooth.*

Now let us return to our main objective of proving existence and uniqueness of potential minimizers of g given in (2.12). We start with the uniqueness, which is where we need the extra assumption of essentially strict convexity on the Bregman function. This pair of assumptions are so common they have been given a name.

Definition 2.2.12. Let $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ be proper, closed, and convex. Then h is **Legendre** if it is essentially smooth and essentially strictly convex.

See Figure 2.1 for some typical examples of Legendre Bregman functions.

Lemma 2.2.13. *Let $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ be proper, closed, convex, and Legendre. Suppose that g in (2.12) is minimized for some $x^\star \in \mathbb{R}^n$, then it is the unique minimizer of g.*

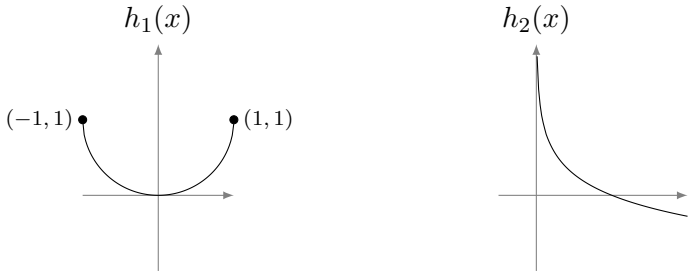


Figure 2.1 Two typical examples of Legendre Bregman functions. The function h_1 has a bounded and closed effective domain $[-1,1]$. The function h_2 has an unbounded and open effective domain $(0,\infty)$. Explicitly, the functions are defined by $h_1(x) = 1 - \sqrt{1-x^2}$ and $h_2(x) = -\log(x)$.

Proof. Suppose that g is also minimized by some $y^\star \in \mathbb{R}^n$. By Lemma 2.2.7 we have that both $x^\star, y^\star \in \operatorname{int}(\operatorname{dom} h)$. In general, we have the following inclusion [17, Theorem 23.4]

$$\operatorname{relint}(\operatorname{dom} h) \subset \operatorname{dom} \partial h.$$

In our case, we have the standing assumption that $\operatorname{int}(\operatorname{dom} h) \neq \emptyset$ which implies that [8, Fact 6.14(iii)]

$$\operatorname{relint}(\operatorname{dom} h) = \operatorname{int}(\operatorname{dom} h)$$

and so $x^\star, y^\star \in \operatorname{dom} \partial h$. The fact that h is strictly convex on the convex subset $[x^\star, y^\star] \subset \operatorname{dom} \partial h$ implies that $x^\star = y^\star$. ■

To prove **P4** and **P5**, it remains to show that a minimizer of g exists. There are multiple different requirements on the problem or its associated Bregman function that would imply this existence property. In many cases, when we are working with a fixed problem and Bregman function, we can show it directly in that special case.

It is quite striking that with all the demands we have put upon our problem, such as Legendreness of the Bregman function, it is still not enough to guarantee that g has a minimizer. There are many roads to choose from here; each giving seemingly different constraints on seemingly different parts of the problem. We choose a constraint on the solution set of the problem, namely that

$$S = \operatorname{argmin}\{f(x) \mid x \in \overline{\operatorname{dom} h}\}$$

is non-empty and compact.

Recall the definition of a level set of a function, for some $\xi \in \mathbb{R}$ we define

$$\operatorname{lev}_\xi f = \{x \in \mathbb{R}^n \mid f(x) \leq \xi\}.$$

Proposition 2.2.14. *Suppose that*

$$S = \operatorname{argmin}\{f(x) \mid x \in \overline{\operatorname{dom} h}\}$$

is non-empty and compact and that the function f is β-relatively smooth with respect to the Bregman function h, with $\operatorname{dom} h \subset \operatorname{dom} f$. Let $\gamma_k \in (0, \frac{1}{\beta})$ and $y \in \operatorname{int}(\operatorname{dom} h)$ then the function

$$g(x) = f(y) + \langle \nabla f(y), x - y \rangle + \frac{1}{\gamma_k} D_h(x, y)$$

has a minimizer over $\operatorname{dom} h$.

Proof. Note that the set S can be written as

$$S = \operatorname{argmin}\{(f + \iota_{\overline{\operatorname{dom} h}})(x) \mid x \in \mathbb{R}^n\}.$$

Let $x^\star \in S$ and $\xi = (f + \iota_{\overline{\operatorname{dom} h}})(x^\star)$. Then the level set $\operatorname{lev}_\xi(f + \iota_{\overline{\operatorname{dom} h}}) = S$ is non-empty and bounded. Since $f + \iota_{\overline{\operatorname{dom} h}}$ is proper, closed, and convex and we have found one level set S which is bounded, [8, Propositions 11.12, 11.13] gives that all level sets of $f + \iota_{\overline{\operatorname{dom} h}}$ are bounded. Let $x \in \operatorname{dom} h$. Since $\operatorname{dom} g = \operatorname{dom} h$ we have that $g(x) < \infty$ and $\operatorname{lev}_{g(x)}(f + \iota_{\overline{\operatorname{dom} h}})$ is bounded. Now we get that

$$\begin{aligned} & f(x) \leq g(x) \\ \implies & (f + \iota_{\overline{\operatorname{dom} h}})(x) \leq g(x) + \iota_{\overline{\operatorname{dom} h}}(x) = g(x) \\ \implies & x \in \operatorname{lev}_{g(x)} g \subset \operatorname{lev}_{g(x)}(f + \iota_{\overline{\operatorname{dom} h}}) \end{aligned}$$

and so

$$\operatorname{lev}_{g(x)} g = \operatorname{dom} h \cap \operatorname{lev}_{g(x)} g$$

is non-empty and bounded. By [8, Theorem 11.10] we have that g has a minimizer over $\operatorname{dom} h$.

■

Theorem 2.2.15. *Let $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ be proper, closed, convex and Legendre. Suppose that*

$$S = \operatorname{argmin}\{f(x) \mid x \in \overline{\operatorname{dom} h}\}$$

*is non-empty and compact and that the function f is β-relatively smooth with respect to h. Then **P4** and **P5** hold.*

Proof. By our assumptions, Proposition 2.2.14 gives that a minimizer $x^\star \in \operatorname{dom} h$ exists to the function g given in **P4**. By Lemma 2.2.13 $x^\star$ is unique and so **P4** holds. By Lemma 2.2.7, $x_1 \in \operatorname{int}(\operatorname{dom} h)$ if $x_0 \in \operatorname{int}(\operatorname{dom} h)$ and so by induction over k we have that **P5** holds and the update step in the Bregman primal method is well-defined. ■

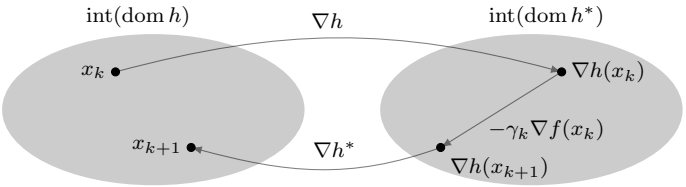


Figure 2.2 A representation of the update step of the Bregman gradient descent.

P6 and P7

From the proof of Lemma 2.2.7 we can now see that the iteration update rule of the Bregman primal method can be written as

$$\nabla h(x_{k+1}) = \nabla h(x_k) - \gamma_k \nabla f(x_k). \tag{2.13}$$

See Figure 2.2 for a visualization.

This update step is by our theory well-defined. If h is Legendre then we know from Proposition 2.2.11 that both h and h^* are essentially smooth which by Proposition 2.2.5 gives that ∂h is a bijective single-valued map from $\operatorname{int}(\operatorname{dom} h)$ to $\operatorname{int}(\operatorname{dom} h^*)$ with $(\partial h)^{-1} = \partial h^*$. If h is furthermore continuously differentiable, which is often common in practice, then ∇h is a homeomorphism between the two open sets $\operatorname{int}(\operatorname{dom} h)$ and $\operatorname{int}(\operatorname{dom} h^*)$. This implies that the update rule can be made explicit:

$$x_{k+1} = \nabla h^{-1}[\nabla h(x_k) - \gamma_k \nabla f(x_k)].$$

If we compare this update rule to the standard case when $h(x) = \frac{1}{2}\|x\|_2^2$ we fall back onto the standard gradient descent method

$$x_{k+1} = x_k - \gamma_k \nabla f(x_k).$$

We can now rewrite **P6** as for all $z \in \operatorname{dom} h$ then

$$\langle \nabla h(x_{k+1}) - \nabla h(x_k), z - x_{k+1} \rangle \leq D_h(z, x_k) - D_h(z, x_{k+1}) - D_h(x_{k+1}, x_k)$$

should hold. This follows directly from the three points identity of Bregman functions with the fact that $x_k \in \operatorname{int}(\operatorname{dom} h)$.

Proposition 2.2.16 (Three Points Identity)**.** *Let $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ be proper, closed, and convex. Then for all $x, y \in \operatorname{int}(\operatorname{dom} h)$ and $z \in \operatorname{dom} h$ we have that*

$$\langle \nabla h(x) - \nabla h(y), z - x \rangle = D_h(z, y) - D_h(z, x) - D_h(x, y).$$

Proof. Evaluating the right hand side reveals that all pure $h(\cdot)$ terms cancel and all that remains is

$$\begin{aligned} D_h(z,y) - D_h(z,x) - D_h(x,y) &= \langle -\nabla h(y), z-y\rangle \\ &\quad + \langle \nabla h(x), z-x\rangle + \langle \nabla h(y), x-y\rangle \\ &= \langle -\nabla h(y), z-x\rangle + \langle \nabla h(x), z-x\rangle \\ &= \langle \nabla h(x) - \nabla h(y), z-x\rangle. \end{aligned}$$

■

Now let us turn to the final property **P7**, which is that concerning the sequential convergence of $\{x_k\}_{k=0}^{\infty}$ to some solution of the problem. We must further demand some properties on our Bregman distance which have not been captured yet. This is summarized in the following Theorem 2.2.17.

Recall from the previous section (2.11) that if $z \in S$ then

$$0 \leq D_h(z, x_{k+1}) \leq D_h(z, x_k)$$

holds for all k. We also derived the following function value convergence

$$\lim_{k\to\infty} f(x_k) = f(z).$$

Theorem 2.2.17. *Let $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ be proper, closed, convex, and Legendre. Suppose that*

$$S = \operatorname{argmin}\{f(x) \mid x \in \overline{\operatorname{dom} h}\} = \operatorname{argmin}\{f(x) \mid x \in \operatorname{dom} h\}$$

is non-empty and compact and that the function f is β-relatively smooth with respect to h. Suppose further that for all $\xi \in \mathbb{R}$ and $x \in \operatorname{dom} h$ the level set $\operatorname{lev}_\xi D_h(x,\cdot)$ is bounded and that for all sequences $\{y_k\}_{k=0}^{\infty}$ with $y_k \in \operatorname{int}(\operatorname{dom} h)$ for all k and all $y \in \operatorname{dom} h$ we have the following equivalence

$$\lim_{k\to\infty} y_k = y \iff \lim_{k\to\infty} D_h(y, y_k) = 0.$$

Then the sequence $\{x_k\}_{k=0}^{\infty}$ converges to some solution $x^\star \in S$.

Proof. Let $z \in S$. Then we know that the sequence $\{D(z,x_k)\}_{k=0}^{\infty}$ converges and is nonincreasing which implies that for $\xi = D(z,x_0)$ we have that

$$x_k \in \operatorname{lev}_\xi D_h(z,\cdot) \text{ for all } k = 0, 1, \ldots$$

and $\{x_k\}_{k=0}^{\infty}$ is bounded. Therefore there exists some subsequence $\{x_{k_n}\}_{n=0}^{\infty}$ of $\{x_k\}_{k=0}^{\infty}$ which converges to some $x^\star \in \overline{\operatorname{dom} h}$. Furthermore, we know that

$x^\star$ must lie in S and solve the problem, which implies that $x^\star \in \operatorname{dom} h$. Since $\lim_{n\to\infty} x_{k_n} = x^\star$ we have that

$$\lim_{n\to\infty} D_h(x^\star, x_{k_n}) = 0.$$

Now let $\varepsilon > 0$ and there is some $N > 0$ such that $n \geq N$ implies that

$$D_h(x^\star, x_{k_n}) < \varepsilon.$$

Then for any $k \geq k_N$ we must also have that $D_h(x^\star, x_k) < \varepsilon$ and so

$$\lim_{k\to\infty} D_h(x^\star, x_k) = 0.$$

From our hypothesis, we know that this must imply convergence:

$$\lim_{k\to\infty} x_k = x^\star.$$

■

There are two properties of importance in the hypothesis of this Theorem 2.2.17. First is that the level sets $\operatorname{lev}_\xi D_h(x, \cdot)$ are bounded and second is that

$$\lim_{k\to\infty} y_k = y \iff \lim_{k\to\infty} D_h(y, y_k) = 0$$

holds. Giving conditions that in turn would imply these two properties is quite a complicated matter. For more details we refer to [6]. We will give an instant of such constraints that are in particular relevant to the later sections of this thesis.

Proposition 2.2.18. *Let $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ be proper, closed, convex, and Legendre. Suppose that both $\operatorname{dom} h$ and $\operatorname{dom} h^*$ are open. Then for all $\xi \in \mathbb{R}$ and $x \in \operatorname{dom} h$ the level set $\operatorname{lev}_\xi D_h(x, \cdot)$ is bounded and for all sequences $\{y_k\}_{k=0}^\infty$ with $y_k \in \operatorname{int}(\operatorname{dom} h)$ for all k and all $y \in \operatorname{dom} h$ we have the following equivalence*

$$\lim_{k\to\infty} y_k = y \iff \lim_{k\to\infty} D_h(y, y_k) = 0.$$

Proof. The level sets $\operatorname{lev}_\xi D_h(x, \cdot)$ are bounded by [6, Corollary 3.11]. The right and left implications in the equivalence above follow from [6, Proposition 3.2] and [6, Theorem 3.9] respectively. ■

Remark. It is interesting how we have yet to use in our analysis the fact that $\{D_h(x_{k+1}, x_k)\}_{k=0}^\infty$ is summable. This property will enter the picture in the next section, where we will relax the upper bound of the step-sizes γ_k to

be larger than $\frac{1}{\beta}$. We know from the standard gradient descent scheme that we might relax up to $\gamma_k \in (0, (2-\varepsilon)/\beta]$ for some $\varepsilon \in (0,2)$. We will see that if the iterates are well-defined (**P4** and **P5**) then an associated relaxation of the Bregman primal method holds. In fact, we can let $\gamma_k \in (0, (1+\alpha-\varepsilon)/\beta]$ for some $\varepsilon \in (0, 1+\alpha)$ where the α is bounded inside $[0,1]$ and depends on the *symmetry* of D_h.

2.3 Symmetry of the Bregman Distance

Let us first define the Symmetry Coefficient of a Bregman function.

Definition 2.3.1. Let $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ be proper, closed, convex, and Legendre. The **Symmetry Coefficient** α_h is the value

$$\alpha_h = \inf\left\{ \frac{D_h(x,y)}{D_h(y,x)} \mid x, y \in \operatorname{int}(\operatorname{dom} h) \text{ and } x \neq y \right\}.$$

Note that $D_h \geq 0$, $\operatorname{int}(\operatorname{dom} h) \neq \emptyset$ by essential smoothness, and by the essential strict convexity of h we can have $D_h(x,y) = 0$ only if $x = y$ and so α_h is always well-defined. Also see that if for all $x, y \in \operatorname{int}(\operatorname{dom} h)$ we have that

$$\frac{D_h(x,y)}{D_h(y,x)} > 1 \implies \frac{D_h(y,x)}{D_h(x,y)} < 1$$

which gives that $\alpha_h \in [0,1]$ for any Bregman function h. If $h = \frac{1}{2}\|\cdot\|_2^2$ then $D_h(x,y) = \frac{1}{2}\|x-y\|_2^2$ is entirely symmetric and so

$$D_h(x,y) = D_h(y,x) \implies \alpha_h = 1.$$

It is natural to ask if there exists more examples of Bregman functions h such that $\alpha_h = 1$. The answer is *yes*. If $h : \mathbb{R}^n \to \mathbb{R}$ is a strictly convex quadratic

$$h(x) = \frac{1}{2}\langle x, Ax\rangle + \langle b, x\rangle + c$$

for any positive definite matrix $A \in \mathbb{S}^n_{++}$, $b \in \mathbb{R}^n$ and $c \in \mathbb{R}$ then

$$D_h(x,y) = \frac{1}{2}\|x-y\|_A^2 \implies \alpha_h = 1.$$

Let us now the question: are there any strictly convex Bregman functions h other than the strictly convex quadratics such that $\alpha_h = 1$. The answer is *no* and it has been previously mentioned in for example [4]. It has been proven in [5] in a weaker form, where h was assumed to be a one-dimensional function $h : \mathbb{R} \to \mathbb{R}$ and twice differentiable. To the best of our knowledge, the general proof given below has not been covered before in this manner.

Proposition 2.3.2. *Let $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ be proper, closed, convex, and Legendre. Then $\alpha_h = 1$ if and only if h is a strictly convex quadratic function with $\operatorname{dom} h = \mathbb{R}^n$.*

Proof. From the previous remark, it remains to show the ($\implies$) direction. Therefore, suppose that $\alpha_h = 1$ holds from which we directly get that

$$D_h(x, y) = D_h(y, x) \tag{2.14}$$

for all $x, y \in \operatorname{int}(\operatorname{dom} h)$.

We claim that without loss of generality we can assume that $0 \in \operatorname{int}(\operatorname{dom} h)$, $h(0) = 0$ and $\nabla h(0) = 0$. Since there exists some $x_0 \in \operatorname{int}(\operatorname{dom} h)$ we can consider the translation $h \circ T_{x_0}$ where $T_{x_0}(x) = x - x_0$. It is easy to see that $h \circ T_{x_0}$ is also proper, closed, convex, and Legendre, with $0 \in \operatorname{int}(\operatorname{dom} h \circ T_{x_0})$. Importantly also $\alpha_{h \circ T_{x_0}} = \alpha_h$.

Now suppose that $0 \in \operatorname{int}(\operatorname{dom} h)$ and set $c = h(0)$ and $b = \nabla h(0)$. Consider a linear perturbation of h given by $\hat{h}(x) = h(x) - \langle b, x\rangle - c$. Again we see that $\hat{h}$ is proper, closed, convex, and Legendre with $\operatorname{int}(\operatorname{dom} \hat{h}) = \operatorname{int}(\operatorname{dom} h)$. Furthermore, $D_{\hat{h}} = D_h$ and so $\alpha_{\hat{h}} = \alpha_h$.

From now on, suppose therefore that $0 \in \operatorname{int}(\operatorname{dom} h)$, $h(0) = 0$ and $\nabla h(0) = 0$. We show that ∇h is a linear map on $\operatorname{int}(\operatorname{dom} h)$:

$$\nabla h(x) = Ax$$

for some matrix $A \in \mathbb{R}^{n \times n}$. This implies by essential smoothness that $\operatorname{int}(\operatorname{dom} h) = \mathbb{R}^n$ and so $\operatorname{dom} h = \mathbb{R}^n$. By essential strict convexity we get that $A \in \mathbb{S}^n_{++}$ and we are done.

Take $y = 0$ in (2.14) to get

$$\begin{aligned} h(x) &= -h(x) - \langle \nabla h(x), -x\rangle \\ \implies h(x) &= \frac{1}{2}\langle \nabla h(x), x\rangle \end{aligned}$$

which must hold for any $x \in \operatorname{int}(\operatorname{dom} h)$. Return to (2.14) for any $x, y \in \operatorname{int}(\operatorname{dom} h)$ with this in mind to get

$$\begin{aligned} h(x) - h(y) - \langle \nabla h(y), x - y\rangle &= h(y) - h(x) - \langle \nabla h(x), y - x\rangle \\ 2h(x) - \langle \nabla h(x), x\rangle + \langle \nabla h(x), y\rangle &= 2h(y) - \langle \nabla h(y), y\rangle + \langle \nabla h(y), x\rangle \\ \implies \langle \nabla h(x), y\rangle &= \langle \nabla h(y), x\rangle. \end{aligned}$$

Let C be the closed set $C = \operatorname{bd}(\operatorname{dom} h)$. We now show that C must be the empty set. Assume, in hope of a contradiction, that $C \neq \emptyset$ then the number

$$r := \inf_{x \in C} \|x\|_\infty$$

exists, since the right hand side is an infimum of a non-empty set of non-negative real numbers. Furthermore, $r > 0$ since $0 \in \operatorname{int}(\operatorname{dom} h)$ and C is closed. Again by C being closed and non-empty, we get that there exists some $x_r \in C$ such that $\|x_r\|_\infty = r$. Define $\mathcal{U}$ to be the open ball with center 0 and radius $r/2$ with respect to the infinity norm:

$$\mathcal{U} = \left\{ x \in \mathbb{R}^n \mid \|x\|_\infty < \frac{r}{2} \right\}.$$

Now let $x_1, x_2 \in \mathcal{U}$ which gives that $x_1 + x_2 \in \operatorname{int}(\operatorname{dom} h)$ and

$$\begin{aligned} \langle \nabla h(x_1 + x_2), y \rangle &= \langle \nabla h(y), x_1 + x_2 \rangle \\ &= \langle \nabla h(y), x_1 \rangle + \langle \nabla h(y), x_2 \rangle \\ &= \langle \nabla h(x_1), y \rangle + \langle \nabla h(x_2), y \rangle \end{aligned}$$

holds for all $y \in \operatorname{int}(\operatorname{dom} h)$. In particular it holds for any scaled standard basis vector $y = \lambda_i e_i$ for small enough $\lambda_i \in (0, 1)$. Then we have that

$$\lambda_i \langle \nabla h(x_1 + x_2), e_i \rangle = \lambda_i (\langle \nabla h(x_1), e_i \rangle + \langle \nabla h(x_2), e_i \rangle)$$

which implies that

$$\nabla h(x_1 + x_2) = \nabla h(x_1) + \nabla h(x_2).$$

Now consider the convergent sequence $\{x_k\}_{k=0}^\infty$ given by $x_k = \mu_k x_r$ where $\mu_k = 1 - 1/(k+1)$. It follows that for all $k \geq 0$ we have that $x_k/2 \in \mathcal{U}$ and so

$$\nabla h(x_k) = \nabla h\left(\frac{x_k}{2} + \frac{x_k}{2}\right) = 2\nabla h\left(\frac{x_k}{2}\right).$$

By essential smoothness of h and the fact that $\lim_{k\to\infty} x_k = x_r \in \operatorname{bd}(\operatorname{dom} h)$ we have that

$$\lim_{k\to\infty} \left\| \nabla h\left(\frac{x_k}{2}\right) \right\|_2 = \infty$$

which contradicts the definition of r. Therefore, $C = \emptyset$ and

$$\operatorname{int}(\operatorname{dom} h) = \operatorname{dom} h = \mathbb{R}^n.$$

Similar to the method above, we can show that for all $x, y \in \mathbb{R}^n$ and $c \in \mathbb{R}$:

$$\langle \nabla h(cx), y \rangle = c \langle \nabla h(x), y \rangle \implies \nabla h(cx) = c \nabla h(x).$$

With the fact that additivity of ∇h also holds, ∇h must be linear: $\nabla h(x) = Ax$ for some $A \in \mathbb{R}^{n \times n}$. By essential strict convexity we must have that A is positive definite and we are done. ■

Remark. If one were to skip the requirement of h being essentially smooth, then it is of course easy to construct other Bregman functions h such that $\alpha_h = 1$. Take for example $h : \mathbb{R} \to \mathbb{R}$ with

$$h(x) = \frac{1}{2}x^2 + \iota_{[-1,1]}(x).$$

An interesting observation is the equation

$$h(x) = \frac{1}{2}\langle \nabla h(x), x \rangle$$

which we arrived at in the proof of Proposition 2.3.2. This is related to the famous Euler's homogeneous function theorem. If h is assumed to be continuously differentiable with $\operatorname{dom} h = \mathbb{R}^n$, then this equation is solved precisely by functions h which are positively homogeneous of degree 2. In other words, the functions such that

$$h(cx) = c^2 h(x)$$

holds for each $x \in \mathbb{R}^n$ and $c > 0$. It is easy to construct such a function h which is not a strictly convex quadratic. Consider for example the proper, closed, convex, and Legendre $h : \mathbb{R}^2 \to \mathbb{R}$ given by

$$h(x_1, x_2) = \frac{1}{4}\sqrt{x_1^4 + x_2^4},$$

which is positively homogeneous of degree 2, but does not contradict Proposition 2.3.2. In fact, after some laborious calculations, one can check that $D_h(x, y) = D_h(y, x)$ holds only when

$$(x_1 y_1 - x_2 y_2)(x_1 y_2 - x_2 y_1) = 0,$$

which equals the union of two closed cones in $\mathbb{R}^4$.

Now let us use the symmetry coefficient α_h to perhaps relax the step-size requirement we previously have used. Let us from now on suppose that the update step is well-defined for step-sizes

$$\gamma_k \in \left(0, \frac{1 + \alpha - \varepsilon}{\beta}\right]$$

for some $\varepsilon \in (0, 1 + \alpha)$. First let us show that the sequence $\{f(x_k)\}_{k=0}^{\infty}$ is nonincreasining. Recall from Section 2.1 how we ended up with the inequality

$$f(x_{k+1}) - f(x_k) \leq -\frac{1}{\gamma_k} \underbrace{D_h(x_k, x_{k+1})}_{\geq 0} - \underbrace{\left(\frac{1}{\gamma_k} - \beta\right)}_{>0} \underbrace{D_h(x_{k+1}, x_k)}_{\geq 0} \leq 0.$$

Now this inequality still hold, but not for the same reason. Notice that we can no longer claim that $\gamma_k^{-1} - \beta > 0$. Instead we can use the symmetry of the terms $D_h(x_k, x_{k+1})$ and $D_h(x_{k+1}, x_k)$ with relation to the symmetry coefficient. We have that

$$\alpha D_h(x_{k+1}, x_k) \leq D_h(x_k, x_{k+1})$$

and so

$$\begin{aligned}\gamma_k(f(x_{k+1}) - f(x_k)) &\leq -D_h(x_k, x_{k+1}) - (1-\beta)D_h(x_{k+1}, x_k) \\ &\leq -\underbrace{(\alpha + 1 - \gamma_k\beta)}_{>\varepsilon} D_h(x_{k+1}, x_k) \leq 0.\end{aligned}$$

We have shown now that even in this relaxed case, $\{f(x_k)\}_{k=0}^{\infty}$ is nonincreasining.

With similar modifications, see [4], one can show that $\{D_h(x_{k+1}, x_k)\}_{k=0}^{\infty}$ is summable and that we get a similar function value convergence:

$$f(x_{n+1}) - f(x^\star) \leq \frac{D_h(x^\star, x_0) + \alpha W}{\sum_{k=0}^{n} \gamma_k}$$

where $W = \sum_{k=0}^{\infty} D_h(x_{k+1}, x_k) < \infty$.

2.4 The Proximal Gradient Bregman Method

Similar to how the standard gradient descent method can be generalized to a proximal-gradient method, we will see how a similar treatment can be applied to our Bregman gradient method. Fortunately, most of our previous proofs and arguments hold directly - or by some modification - in this new setting.

Let $\phi : \mathbb{R}^n \to \overline{\mathbb{R}}$ be some proper, closed, and convex function. Note that we do not require ϕ to be differentiable. We consider the primal minimization problem

$$\underset{x \in \mathbb{R}^n}{\text{minimize}}\ f(x) + \phi(x). \tag{2.15}$$

In the normal gradient descent method, where f is assumed to be β-relatively smooth to the Bregman function $h = \frac{1}{2}\|\cdot\|_2^2$, we would extend our optimization method to the *proximal gradient descent* with update step

$$x_{k+1} = \text{prox}_{\gamma_k \phi}[x_k - \gamma_k \nabla f(x_k)].$$

For a suitably small step-size $\gamma_k \in (0, 1/\beta]$ this update step can be seen as a majorization-minimization algorithm. We will now follow the same reasoning and arguments to develop a proximal gradient Bregman algorithm.

Let $x_k \in \operatorname{int}(\operatorname{dom} h)$ and define similar as before $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ to be

$$g(x) = \phi(x) + f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{\gamma_k} D_h(x, x_k).$$

If all our previous hypotheses hold concerning the differentiable function f, the Bregman function h, the step-size γ_k, and how they relate to each other, then first of all $g \geq f + \phi$. We also know that the forward gradient step $x_k^+ \in \operatorname{int}(\operatorname{dom} h)$ is well-defined and given by (see Lemma 2.2.7)

$$x_k^+ = \nabla h^*(\nabla h(x_k) - \gamma_k \nabla f(x_k)).$$

Now consider the expression $D_h(x, x_k^+)$. After expanding this we get the expression

$$\begin{aligned} D_h(x, x_k^+) = h(x) &- h(\nabla h^*(\nabla h(x_k) - \gamma_k \nabla f(x_k))) \\ &- \langle \underbrace{\nabla h \circ \nabla h^*}_{=I}(\nabla h(x_k) - \gamma_k \nabla f(x_k)), x - \nabla h^*(\nabla h(x_k) - \gamma_k \nabla f(x_k)) \rangle. \end{aligned}$$

If we remove the terms that are independent of x, we get the expression

$$h(x) - \langle \nabla h(x_k) - \gamma_k \nabla f(x_k), x \rangle.$$

Lastly, after multiplying this expression by a constant γ_k^{-1} and adding some terms independent of x we arrive at

$$\begin{aligned} &f(x_k) + \langle \nabla f(x_k), x \rangle + \frac{1}{\gamma_k}[h(x) - h(x_k) + \langle \nabla h(x_k), x - x_k \rangle] \\ &= f(x_k) + \langle \nabla f(x_k), x \rangle + \frac{1}{\gamma_k} D_h(x, x_k). \end{aligned}$$

What we end up with is that minimizing g is equivalent with minimizing the function $x \mapsto \phi(x) + \frac{1}{\gamma_k} D_h(x, x_k^+)$. We notice that it makes sense to assume at least that $\operatorname{dom} \phi \cap \operatorname{dom} h \neq \emptyset$. It is now natural to define a new type of a proximal operator, which depends on the Bregman function h. We will let $\operatorname{prox}^h_{\gamma_k \phi} : \operatorname{int}(\operatorname{dom} h) \to 2^{\mathbb{R}^n}$ be given by

$$\operatorname{prox}^h_{\gamma_k \phi}(y) = \operatorname*{Argmin}_{x \in \mathbb{R}^n} \left[\phi(x) + \frac{1}{\gamma_k} D_h(x, y) \right].$$

This operator is at least single-valued when for instance the level sets of ϕ are bounded (which naturally holds for example when ϕ is an indicator function of a bounded set or a non-differentiable norm regularizer), by [8, Corollary 11.16], since $D_h(\cdot, y) \geq 0$ always holds. The Bregman prox-operator is furthermore single-valued if h is assumed to be strictly convex. Recall that

strict convexity is a stronger condition than essential strict convexity by Example 2.2.9.

It can very well happen that the single unique minimizer given by the Bregman prox-operator lies on $\operatorname{dom} h \setminus \operatorname{int}(\operatorname{dom} h)$ which would make the next iteration step in our algorithm invalid. Consider for example a Legendre Bregman function $h : \mathbb{R} \to \overline{\mathbb{R}}$ with $\operatorname{dom} h = [-1, 1]$, $h(-1) = h(1) = 1$, $\nabla h(0) = 0$, and $\nabla h(x) \to \infty$ if $x \to -1^+$ or $x \to 1^-$. If for some initial value $x_0 \in (-1, 1)$ we have that $x_0^+ = 0$ we would be in trouble if $\phi(x) = |x - 1| - C$ for large enough $C > 0$. Then $\operatorname{prox}^h_{\gamma_k \phi}(x_0) = 1 \notin \operatorname{int}(\operatorname{dom} h)$. Furthermore, $x_1 = 1$ might not even be a minimizer of the objective function $f + \phi$. Note that this type of problem will never occur if we assume that $\operatorname{dom} \phi \subset \operatorname{int}(\operatorname{dom} h)$.

With small modifications of our discussion, function value convergence of $\{\phi(x_k) + f(x_k)\}_{k=0}^{\infty}$ and sequence convergence of $\{x_k\}_{k=0}^{\infty}$ can be extended in this prox-setting. See [4] for the complete story.

We will end off this chapter by showing how the update step of the proximal Bregman method can be described by operators. This will lead us to the framework of monotone inclusion problems, which is a theory that lets use analyse and unify many optimization algorithms effectively. One of these algorithms is the NOFOB algorithm [13], which the next chapter will concern (of which the proximal Bregman method is a special case).

We have seen that the update step is given by

$$x_{k+1} = \operatorname{prox}^h_{\gamma_k \phi}(x_k^+)$$

where the forward step x_k^+ satisfies

$$\nabla h(x_k^+) = \nabla h(x_k) - \gamma_k \nabla f(x_k).$$

Since we have assumed before the relevant constraint qualifications, we can freely distribute the subdifferental operator. We arrive at the optimality condition

$$\begin{aligned} & 0 \in \partial \phi(x_{k+1}) + \frac{1}{\gamma_k} \nabla h(x_{k+1}) - \frac{1}{\gamma_k} \nabla h(x_k^+) \\ \iff & \\ & x_{k+1} \in [\nabla h + \gamma_k \partial \phi]^{-1} \nabla h(x^+) \\ & \quad = [\nabla h + \gamma_k \partial \phi]^{-1} [\nabla h - \gamma_k \nabla f] x_k. \end{aligned}$$

By our theory, the x_{k+1} is uniquely determined:

$$x_{k+1} = [\nabla h + \gamma_k \partial \phi]^{-1} [\nabla h - \gamma_k \nabla f] x_k,$$

which exactly generalizes the proximal gradient update step when $\nabla h = I$.

3

Bregman Primal-Dual Methods

In this chapter our goal is to introduce the NOFOB algorithm [13] and show how our Bregman theory can be viewed in this framework. The NOFOB algorithm consists of two steps. First a forward-backward map and then a projection onto a hyperplane. The first two sections of this chapter will deal with each of these steps respectively, in a Bregman theory setting. In the third section, we will introduce a primal-dual problem and the NOFOB algorithm in a Bregman setting.

3.1 The forward-backward Step

We will begin by describing the general NOFOB setting. Consider the inclusion problem

$$0 \in [A + C]x$$

for some operators A, C. In order to coincide with previous results it is appropriate to suppose that $A : \mathbb{R}^n \to 2^{\mathbb{R}^n}$ is maximally monotone and $C : \mathbb{R}^n \to \mathbb{R}^n$ is $\frac{1}{\beta}$-cocoercive. We wish to find some $x \in \mathbb{R}^n$ that satisfies

$$x \in \mathrm{Zer}[A + C] \neq \emptyset.$$

The forward-backward step is defined by an operator $T_k : \mathbb{R}^n \to 2^{\mathbb{R}^n}$ which in turn is defined by some operator $M_k : \mathbb{R}^n \to \mathbb{R}^n$ by

$$T_k = [M_k + A]^{-1}[M_k - C].$$

We then define the forward-backward step by

$$x_{k+1} \in T_k x_k$$

and the initial x_0 can be chosen arbitrary. If we for the moment assume that this update step is well defined, we can see that

$$\begin{aligned} x \in \operatorname{Fix} T_k &\iff x \in T_k x \\ &\iff x \in [M_k + A]^{-1}[M_k - C]x \\ &\iff M_k x - Cx \in M_k x + Ax \\ &\iff 0 \in [A + C]x \iff x \in \operatorname{Zer}[A + C]. \end{aligned}$$

Which means that $\operatorname{Fix} T_k = \operatorname{Zer}[A + C]$.

At this point it is very tempting to compare this update step with the previously seen proximal gradient update step

$$x_{k+1} = \left[\frac{1}{\gamma_k}\nabla h + \partial\phi\right]^{-1}\left[\frac{1}{\gamma_k}\nabla h - \nabla f\right].$$

With this line of thought we would let A represent the proximal, proper, closed, convex, and non-smooth $\partial\phi$ and C represent the proper, closed, convex, and smooth ∇f. Furthermore, M_k would represent the Bregman part $\gamma_k^{-1}\nabla h$. But there are multiple problems with this approach. First let us describe the problems and then how we can remedy them.

First of all, setting $C = \nabla f$ makes no sense, since we have previously assumed $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ to allow $\operatorname{dom} f \neq \mathbb{R}^n$. Therefore, ∇f is certainly not in general $\frac{1}{\beta}$-cocoercive or β-Lipschitz continuous, since f may not be β-smooth. This is not a fault in our chosen problem setting, but a strength. It allows our method to be more flexible. We have introduced theory of f being β-relatively smooth with respect to a Bregman function h which we want to represent in this NOFOB setting.

But the problems do not end there. Similarly to the previous issue, we cannot simply set $M_k = \gamma_k^{-1}\nabla h$ because of discrepancies with its domain $\operatorname{dom} h$ which might not equal the entire $\mathbb{R}^n$. A deeper issue arises when examining what properties M_k should have in order to guarantee that the forward-backward step is well-defined and the convergence of the complete NOFOB algorithm. In [13], it is described that M_k should at least be 1-strongly monotone and we have nowhere assumed any strong convexity of h before.

Our remedy is the following. Let A instead represent the whole sum

$$A = \partial\phi + \nabla f,$$

which is certainly maximally monotone. This forces $C = 0$ which is certainly cocoercive. The forward part of T_k is now only an explicit M_k step, which inspires us to set

$$M_k = \frac{1}{\gamma_k}\nabla h - \nabla f.$$

Note that M_k is now maximally monotone, which follows directly from f being β-relatively smooth with respect to h and $\gamma_k \in (0, \frac{1}{\beta})$. It also follows that indeed

$$T_k = [\nabla h + \gamma_k \partial\phi]^{-1}[\nabla h - \gamma_k \nabla f].$$

From the theory of the previous chapter, if h is assumed further to be Legendre, ϕ has appropriate domain, and $x_k \in \operatorname{int}(\operatorname{dom} h)$, then the domain of T_k contains $\operatorname{int}(\operatorname{dom} h)$ and is single valued on $\operatorname{int}(\operatorname{dom} h)$ with $T_k x_k \in \operatorname{int}(\operatorname{dom} h)$.

3.2 The Bregman Projection

Let $x_k \in \operatorname{int}(\operatorname{dom} h)$ and $\hat{x}_k = T_k x_k$. Consider the half-space $H_k \subset \mathbb{R}^n$ defined by x_k and $\hat{x}_k$ given by

$$H_k = \{z \in \mathbb{R}^n \mid \langle M_k x_k - M_k \hat{x}_k, z - \hat{x}_k \rangle \leq 0\}. \tag{3.1}$$

There are some basic properties that we can extract from this definition. First of all $\hat{x}_k \in \operatorname{bd}(H_k)$ follows directly. If M_k is strictly monotone, which for instance happens when $\gamma_k \in (0, \frac{1}{\beta})$ we get that $x_k \notin H_k$ if and only if $x_k \neq \hat{x}_k$, which happens if and only if $x_k \notin \operatorname{Zer}[A + C] = \operatorname{Zer}[A]$.

Suppose that $x^\star \in \operatorname{Zer}[A]$. Can we relate this condition to H_k? First, since

$$\begin{aligned} \hat{x}_k &\in [M_k + A]^{-1} M_k x_k \\ M_k x_k - M_k \hat{x}_k &\in A\hat{x}_k, \end{aligned}$$

and the fact that $0 \in Ax^\star$ we can use that A is monotone with respect to $x^\star$ and x_k to get

$$\langle 0 - (M_k x_k - M_k \hat{x}_k), x^\star - \hat{x}_k \rangle \geq 0.$$

This is equivalent with $x^\star \in H_k$. We summarize our findings in the following proposition.

Proposition 3.2.1. *Suppose that $x_k \in \operatorname{int}(\operatorname{dom} h)$, M_k is strictly monotone and that $\hat{x}_k = T_k x_k$ is well-defined. Then $\hat{x}_k \in H_k$ and if $x_k \notin \operatorname{Zer}[A]$ then the hyper-plane $\operatorname{bd}(H_k)$ strictly separates $\operatorname{Zer}[A]$ and the singleton $\{x_k\}$. If $x_k \in \operatorname{Zer}[A]$ then also $x_k \in H_k$.*

We can now fully write out the second step of the NOFOB algorithm. It is of the form of a projection of x_k onto H_k with respect to the geometry induced by h:

$$x_{k+1} = \Pi^h_{H_k}(x_k) \coloneqq \operatorname*{argmin}_{x \in H_k} D_h(x, x_k).$$

There are some non-subtle and subtle details that we need to face. First of all, we need the minimizing point of $D_h(x, x_k)$ in H_k to exist and be unique. Furthermore, we must demand that $x_{k+1} \in \operatorname{int}(\operatorname{dom} h)$ in order to have a well-defined next step of the algorithm.

One might fear that these three desired properties regarding this particular Bregman projection will require new constraints on h. Beautifully, this is not the case. Every such desired property follows from the update steps of the algorithm and the Legendreness of h.

Theorem 3.2.2. *Suppose that $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ is proper, closed, convex, and Legendre. Let $y \in \operatorname{int}(\operatorname{dom} h)$. If C is some closed convex set satisfying the constraint qualification $C \cap \operatorname{int}(\operatorname{dom} h) \neq \emptyset$ then*

$$\Pi_C^h(y) = \operatorname*{argmin}_{x \in C} D_h(x, y)$$

is well-defined, single valued, and contained in $\operatorname{int}(\operatorname{dom} h)$.

Proof. We begin by showing that such a minimizer of $D_h(\cdot, y)$ exists over C. Since C is closed and convex with $C \cap \operatorname{dom} f \neq \emptyset$ then by [8, Proposition 11.15] such a minimizer exists if the level set $\operatorname{lev}_\xi D_h(\cdot, y)$ is bounded for all $\xi \in \mathbb{R}$. This is not the type of level sets that have appeared in the previous chapter, which were of the type $\operatorname{lev}_\xi D_h(x, \cdot)$. Since

$$\begin{aligned} D_h(x, y) &= h(x) - h(y) - \langle \nabla h(y), x - y \rangle \\ &= h(x) - \langle \nabla h(y), x \rangle - \underbrace{h(y) + \langle \nabla h(y), y \rangle}_{\text{independent of } x} \end{aligned}$$

we can instead show that the level sets $\operatorname{lev}_\xi [h(\cdot) - \langle \nabla h(y), \cdot \rangle]$ are bounded. Recall that the map ∇h is a homeomorphism between the open sets $\operatorname{int}(\operatorname{dom} h)$ and $\operatorname{int}(\operatorname{dom} h^*)$ which implies that $\nabla h(y) \in \operatorname{int}(\operatorname{dom} h^*)$. By [8, Theorem 14.17] we have precisely that the level sets $\operatorname{lev}_\xi [h(\cdot) - \langle \nabla h(y), \cdot \rangle]$ are bounded. Therefore, $D_h(\cdot, y)$ has a minimizer over C.

Now let us show that the set $K = \operatorname{Argmin}_{x \in C} D_h(x, y)$ is contained in $\operatorname{int}(\operatorname{dom} h)$. Suppose, in hope of a contradiction, that there exists some $x \in K \setminus \operatorname{int}(\operatorname{dom} h)$. It is clear that x must be an element of $\operatorname{dom} h$, otherwise $D_h(x, y)$ would be infinite. This implies that $x \in \operatorname{bd}(h) \cap \operatorname{dom} h$. By the constraint qualification, we can also find some $z \in C \cap \operatorname{int}(\operatorname{dom} h)$. The line segment $[x, z] \subset C$ and furthermore $[x, z] \subset K$. Let the function $l : [0, 1] \to K$ trace out this line segment:

$$l(t) = tx + (1 - t)z$$

and consider the function $d : [0, 1] \to \mathbb{R}$ given by $d(t) = D_h(l(t), y)$. By the chain rule, d is differentiable on $(0, 1)$ with derivative

$$d'(t) = \langle \nabla h \circ l(t), x - z \rangle - \langle \nabla h(y), x - z \rangle.$$

The limit as $t \to 0^+$ of the first of these two terms is the directional derivative of h at the point x with respect to the direction $x - z$. Since h is essentially smooth, we have that $\partial h(x) = \emptyset$ and it follows that

$$\lim_{t \to 0^+} \langle \nabla h \circ l(t), x - z \rangle = -\infty \implies \lim_{t \to 0^+} d'(t) = -\infty.$$

This is a clear contradiction to the fact that $[x, z] \subset K$, since by continuity we can find some small enough $t \in (0, 1)$ such that $d(t) < d(0)$. Therefore, $K \subset \operatorname{int}(\operatorname{dom} h)$.

Now it is straightforward to see that K must be a singleton. In fact, we have that

$$K \subset \operatorname{int}(\operatorname{dom} h) = \operatorname{dom} \partial h$$

and by essential strict convexity of h, h is strictly convex on K. ■

Note that in the setting of the NOFOB algorithm, the constraint qualification $H_k \cap \operatorname{int}(\operatorname{dom} h) \neq \emptyset$ follows for free. By Proposition 3.2.1 we have that $\hat{x}_k \in H_k$ and we also know that $\hat{x}_k \in \operatorname{int}(\operatorname{dom} h)$.

Furthermore, if $x_k \notin \operatorname{Zer}[A]$ then the projection x_{k+1} lies on the boundary of H_k. The argument is simple. Suppose that $x_k \in \operatorname{int}(H_k)$. Since $\operatorname{int}(H_k)$ is open, the necessary first-order condition of optimality says that

$$\nabla D_h(\cdot, x_k) \mid_{x_{k+1}} = \nabla h(x_{k+1}) - \nabla h(x_k) = 0.$$

From h being Legendre we know that ∇h is injective and so $x_{k+1} = x_k$. This is a contradiction by Proposition 3.2.1. We summarize this in the following proposition.

Proposition 3.2.3. *Suppose that $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ is proper, closed, convex, and Legendre. Suppose that $x_k \in \operatorname{int}(\operatorname{dom} h)$, M_k is strictly convex, and $\hat{x}_k = T_k x_k$ is well-defined. Let H_k be defined as in (3.1). Then $\Pi^h_{H_k}(x_k)$ is well-defined, single-valued, contained in* $\operatorname{int}(\operatorname{dom} h)$*, and satisfies*

$$\Pi^h_{H_k}(x_k) = \Pi^h_{\operatorname{bd} H_k}(x_k).$$

Note that the optimization problem given by $\Pi^h_{\operatorname{bd} H_k}(x_k)$ has the following Lagrange optimality condition

$$\begin{cases} \nabla h(x_{k+1}) &= \nabla h(x_k) - \lambda_k (M_k x_k - M_k \hat{x}_k) \\ 0 &= \langle M_k x_k - M_k \hat{x}_k, x_{k+1} - \hat{x}_k \rangle \end{cases} \tag{3.2}$$

solved by some unique Lagrange multiplier $\lambda_k \in \mathbb{R}$. These unique values λ_k will be called the **projection steps** of the coming Bregman NOFOB algorithm (see Section 3.3).

Remark. Recall the primal proximal Bregman setting from the previous chapter with backward step

$$x_{k+1} = \operatorname*{argmin}_{x \in \mathbb{R}^n} \left[\phi(x) + \frac{1}{\gamma_k} D_h(x, x_k^+) \right].$$

Consider the special case when ϕ is some proximal indicator function $\phi = \iota_C$ for some convex set C. Then this backward step reduces to a standard Bregman projection

$$x_{k+1} = \Pi_C^h(x_k^+).$$

The theory of this section can thus be utilized, if we impose the natural constraint qualification of C and h, to show that the backward step is well-defined, single valued. Also $x_{k+1} \in \operatorname{int}(\operatorname{dom} h)$ and so the algorithm may continue in the next update step.

3.3 The Bregman NOFOB Algorithm

Now we will combine a forward-backward step of a possibly nonlinear kernel $M_k : \mathbb{R}^p \to \mathbb{R}^p$ followed by a projection correction onto a half-space H_k, separating $\operatorname{Zer}[A]$ and the singleton $\{x_k\}$, with respect to some Bregman function $h : \mathbb{R}^p \to \overline{\mathbb{R}}$. In particular, we will from now focus on solving monotone inclusion problems $0 \in Az$ where $A : \mathbb{R}^p \to 2^{\mathbb{R}^p}$ has a particular form:

$$A = \partial F + K.$$

Here $F : \mathbb{R}^p \to \overline{\mathbb{R}}$ is some proper, closed, and convex function with subdifferential $\partial F : \mathbb{R}^p \to 2^{\mathbb{R}^p}$ and $K : \mathbb{R}^p \to \mathbb{R}^p$ is some linear skew-symmetric operator. Soon we will see why this form of A naturally occurs in a primal-dual setting, see (3.3).

We propose a kernel $M_k : \mathbb{R}^p \to 2^{\mathbb{R}^p}$ of a special form:

$$M_k = \partial \psi_k + \tilde{K}.$$

The function $\psi_k : \mathbb{R}^p \to \overline{\mathbb{R}}$ is assumed to be proper, closed, convex, and differentiable on the nonempty set $\operatorname{int}(\operatorname{dom} \psi)$. In some sense the part of ψ_k will capture a possible Bregman structure of the inclusion problem. On the other hand, the operator $\tilde{K} : \mathbb{R}^p \to \mathbb{R}^p$ will capture the linear skew-symmetric part. In fact, $\tilde{K}$ will be assumed to be linear skew-symmetric and take the form

$$\tilde{K} = K_M - K.$$

Once more $K_M : \mathbb{R}^p \to \mathbb{R}^p$ is assumed to be linear skew-symmetric. Note that for $z_k \in \operatorname{int}(\operatorname{dom} M_k)$ we have that the set $M_k z_k$ is a singleton and (with some abuse of notation) equals

$$M_k z_k = \nabla \psi_k(z_k) + K_M z_k - K z_k.$$

Algorithm 1: Bregman NOFOB

Let : $M_k = \partial\psi_k + \tilde{K}$, Legendre Bregman function h
Input : $z_0 \in \operatorname{int}(\operatorname{dom} M_k)$
for $k = 0, 1, \ldots$ **do**
  $\hat{z}_k \coloneqq [M_k + A]^{-1} M_k z_k = [\nabla\psi_k + K_M + \partial F]^{-1}[\nabla\psi_k + K_M - K]z_k$
  $H_k \coloneqq \{z \in \mathbb{R}^p \mid \langle M_k z_k - M_k \hat{z}_k, z - \hat{z}_k\rangle \leq 0\}$
  $z_{k+1} \coloneqq \Pi^h_{H_k}(z_k)$
end

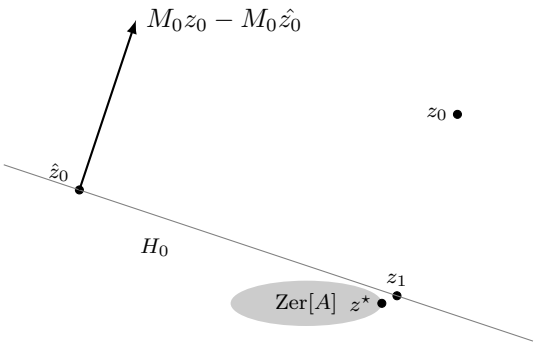


Figure 3.1 A representation of the first iteration of the Bregman NOFOB algorithm.

See Figure 3.1 for a graphical representation of the first iteration of the Bregman NOFOB algorithm.

The following standard proposition describes how many desirable properties of ψ_k carry over to its associated kernel M_k irrespective of how $\tilde{K}$ behaves.

Proposition 3.3.1. *Let $M_k = \partial\psi_k + \tilde{K}$ where $\tilde{K}$ is linear skew-symmetric. Then*

(i) If $\partial\psi_k$ is monotone then M_k is monotone.

(ii) If $\partial\psi_k$ is strictly monotone then M_k is strictly monotone.

(iii) If $\partial\psi_k$ is maximally monotone then M_k is maximally monotone.

Proof. Let $u \in M_k x$, $v \in M_k y$ then

$$\begin{aligned}\langle u - v, x - y\rangle &= \langle (u - \tilde{K}x) - (v - \tilde{K}y), x - y\rangle + \langle \tilde{K}(x-)y, x - y\rangle \\ &\geq \langle \tilde{K}x - \tilde{K}y, x - y\rangle = 0.\end{aligned}$$

The first inequality follows from $\partial\psi_k$ being monotone and is replaced by a strict inequality in case of $\partial\psi_k$ being strictly monotone. The second equality follows from $\tilde{K}$ being skew-symmetric. Therefore, both (i) and (ii) follows. The maximal part of (iii) is immediate. ■

Let us consider the standard primal optimization problem on composite form

$$\underset{x\in\mathbb{R}^n}{\text{minimize}}\, f(x) + g(Lx)$$

with $f : \mathbb{R}^n \to \overline{\mathbb{R}}$, $L \in \mathbb{R}^{m\times n}$ and $g : \mathbb{R}^m \to \overline{\mathbb{R}}$. If both f and g are proper, closed, and convex and the constraint qualification $L\,\mathrm{dom}\, f \cap \mathrm{int}(\mathrm{dom}\, g) \neq \emptyset$ holds then strong duality holds between the above primal problem and the Fenchel-Rockafellar dual problem

$$\underset{y\in\mathbb{R}^m}{\text{minimize}} -f^*(-L^T y) - g^*(y).$$

Let $p = m + n$. The Lagrangian $\mathbf{L} : \mathbb{R}^p \to \mathbb{R} \cup \{\pm\infty\}$, which generates the primal problem, is given by

$$\mathbf{L}(x, y) = f(x) - g^*(y) + \langle y, Lx\rangle.$$

The Lagrangian has the saddle subdifferential (see for reference [17, Example 2.2.3])

$$A \coloneqq \partial\mathbf{L} \coloneqq \begin{bmatrix} \partial_x \mathbf{L} \\ \partial_y - \mathbf{L} \end{bmatrix} = \underbrace{\begin{bmatrix} \partial f \\ \partial g^* \end{bmatrix}}_{\partial F} + \underbrace{\begin{bmatrix} 0 & L^T \\ -L & 0 \end{bmatrix}}_{K} \tag{3.3}$$

which fits our setting. Here we have implicitly defined F as $F = (f, g^*)$. We will find a solution $z^\star = (x^\star, y^\star) \in \mathbb{R}^d$ to the inclusion problem

$$0 \in Az = \partial\mathbf{L}z$$

by applying the Bregman NOFOB algorithm to this primal-dual setting. In other words, the point $z^\star$ is a saddle point of $\mathbf{L}$.

We will from now on restrict the type of kernel M_k that we will consider. First of all, the linear skew-symmetric operator K_M will be of the form

$$K_M = \begin{bmatrix} 0 & L_M^T \\ -L_M & 0 \end{bmatrix},$$

with $L_M \in \mathbb{R}^{m\times n}$. Furthermore, we will restrict the function ψ_k to the kind which satisfies for some functions $\psi_{xx}, \psi_{yx} : \mathbb{R}^n \to \mathbb{R}$ and $\psi_{xy}, \psi_{yy} : \mathbb{R}^m \to \mathbb{R}$:

$$\partial\psi_k = \begin{bmatrix} \partial\psi_{xx} & \partial\psi_{xy} \\ \partial\psi_{yx} & \partial\psi_{yy} \end{bmatrix}.$$

This implies that

$$\nabla\psi_k \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \nabla\psi_{xx}(x) + \nabla\psi_{xy}(y) \\ \nabla\psi_{yx}(x) + \nabla\psi_{yy}(y) \end{bmatrix}$$

holds for all $(x, y) \in \operatorname{int}(\operatorname{dom}\psi)$. Note that we have dropped the iteration index k on the right hand sides for the sake of readability.

Duality-Gap Convergence

Before we show our convergence results, we define a partial duality gap function $\mathcal{G}_{z^\star} : \mathbb{R}^p \to \mathbb{R} \cup \{\pm\infty\}$ for some fixed $z^\star = (x^\star, y^\star) \in \mathbb{R}^p$ by

$$\mathcal{G}_{z^\star}(x, y) = \mathbf{L}(x, y^\star) - \mathbf{L}(x^\star, y). \tag{3.4}$$

It is an elementary result that $\mathcal{G}_{z^\star}(z) \geq 0$ for all $z \in \mathbb{R}^p$ if $z^\star$ is a saddle point of $\mathbf{L}$, see for example [18].

Let the operator $T_k : \mathbb{R}^p \to 2^{\mathbb{R}^p}$ be defined by

$$T_k = [M_k + A]^{-1} M_k$$

and $h : \mathbb{R}^p \to \overline{\mathbb{R}}$ the Legendre Bregman function referenced in Algorithm 1. We will from now in this section assume that M_k is strictly convex and that T_k is non-empty, single-valued, and maps $\operatorname{int}(\operatorname{dom} h)$ in $\operatorname{int}(\operatorname{dom} h)$. By Proposition 3.2.3, the Bregman projection in Algorithm1 is well-defined, single-valued, and contained in $\operatorname{int}(\operatorname{dom} h)$. Therefore, if Algorithm 1 is well-defined up to iteration k then it is also well-defined at iteration $k + 1$.

The question remains of what properties of M_k are needed in order for this assumption on T_k to hold. The answer to this question is outside the scope of this thesis.

Lemma 3.3.2. *Let $\{z_k\}_{k=0}^{\infty}$ and $\{\hat{z}_k\}_{k=0}^{\infty}$ be the iterates given by Algorithm 1. Let $z = (x, y) \in \mathbb{R}^p$ be arbitrary. Then we have that*

$$0 \geq \mathcal{G}_z(\hat{z}_k) + \langle M_k z_k - M_k \hat{z}_k, z - \hat{z}_k \rangle.$$

Proof. The algorithmic forward-backward step is given by $M_k z_k - M_k \hat{z}_k \in A\hat{z}_k$. If we expand this expression we get that

$$\begin{cases} \partial f(\hat{x}_k) \ni \nabla\psi_{xx} x_k + \nabla\psi_{xy} y_k - (\nabla\psi_{xx}\hat{x}_k + \nabla\psi_{xy}\hat{y}_k) + L_M^T(y_k - \hat{y}_k) - L^T y_k \\ \partial g^*(\hat{y}_k) \ni \nabla\psi_{yx} x_k + \nabla\psi_{yy} y_k - (\nabla\psi_{yx}\hat{x}_k + \nabla\psi_{yy}\hat{y}_k) - L_M(x_k - \hat{x}_k) + L x_k. \end{cases}$$

Let us define the two vectors v_k and w_k by

$$\begin{cases} v_k = \nabla\psi_{xx} x_k + \nabla\psi_{xy} y_k - (\nabla\psi_{xx}\hat{x}_k + \nabla\psi_{xy}\hat{y}_k) + L_M^T(y_k - \hat{y}_k) \\ w_k = \nabla\psi_{yx} x_k + \nabla\psi_{yy} y_k - (\nabla\psi_{yx}\hat{x}_k + \nabla\psi_{yy}\hat{y}_k) - L_M(x_k - \hat{x}_k) \end{cases}$$

so that $v_k - L^T y_k \in \partial f(\hat{x}_k)$ and $w_k + Lx_k \in \partial g^*(\hat{y}_k)$. By the definition of the subdifferential operator we have that

$$\begin{cases} 0 \geq f(\hat{x}_k) - f(x) + \langle v_k - L^T y_k, x - \hat{x}_k \rangle \\ 0 \geq g^*(\hat{y}_k) - g^*(y) + \langle w_k + Lx_k, y - \hat{y}_k \rangle. \end{cases}$$

If we add these two inequalities we get

$$\begin{aligned} 0 &\geq [f(\hat{x}_k) - g^*(y) + \langle y, L\hat{x}_k \rangle] - [f(x) - g^*(\hat{y}_k) + \langle L^T \hat{y}_k, x \rangle] \\ &\quad + \langle v_k - L^T(y_k - \hat{y}_k), x - \hat{x}_k \rangle + \langle w_k + L(x_k - \hat{x}_k), y - \hat{y}_k \rangle \\ &\quad + \underbrace{\langle L^T \hat{y}_k, \hat{x}_k \rangle - \langle \hat{y}_k, L\hat{x}_k \rangle}_{=0} \\ &= \mathcal{G}_z(\hat{z}_k) + \langle M_k z_k - M_k \hat{z}_k, z - \hat{z}_k \rangle. \end{aligned}$$

■

Proposition 3.3.3. *Let $\{z_k\}_{k=0}^{\infty}$ and $\{\hat{z}_k\}_{k=0}^{\infty}$ be the iterates given by Algorithm 1. Suppose that $\operatorname{dom} h \cap \operatorname{Zer}[A] \neq \emptyset$. If the sequence of projection step lengths $\{\lambda_k\}_{k=0}^{\infty}$ (defined by (3.2)) is uniformly bounded from below by a positive number, i.e.,*

$$\liminf_{k \to \infty} \lambda_k = \varepsilon > 0$$

then the sequence $\{\hat{z}_k\}_{k=0}^{\infty}$ converges in duality gap, i.e.,

$$\lim_{k \to \infty} \mathcal{G}_{z^\star}(\hat{z}_k) = 0$$

holds for all $z^\star \in \operatorname{dom} h \cap \operatorname{Zer}[A]$.

Proof. Let $z^\star \in \operatorname{dom} h \cap \operatorname{Zer}[A]$. Recall that $\mathcal{G}_{z^\star}(\hat{z}_k) \geq 0$. By the three points identity of Bregman functions (Proposition 2.7) we have that

$$\begin{aligned} D_h(z^\star, z_{k+1}) &= D_h(z^\star, z_k) - D_h(z_{k+1}, z_k) - \langle \nabla h(z_{k+1}) - \nabla h(z_k), z^\star - z_{k+1} \rangle \\ &= D_h(z^\star, z_k) - D_h(z_{k+1}, z_k) - \lambda_k \langle M_k \hat{z}_k - M_k z_k, z^\star - z_{k+1} \rangle \\ &= D_h(z^\star, z_k) - D_h(z_{k+1}, z_k) - \lambda_k \langle M_k \hat{z}_k - M_k z_k, z^\star - \hat{z}_k + \hat{z}_k - z_{k+1} \rangle \\ &= D_h(z^\star, z_k) - D_h(z_{k+1}, z_k) - \lambda_k \langle M_k \hat{z}_k - M_k z_k, z^\star - \hat{z}_k \rangle \\ &\leq D_h(z^\star, z_k) - D_h(z_{k+1}, z_k) - \lambda_k \mathcal{G}_{z^\star}(\hat{z}_k). \end{aligned}$$

The second equality is the definition of the projection step from (3.2). The fourth equality follows from Proposition 3.2.3 and that $z_{k+1} \in \operatorname{bd}(H_k)$. The inequality is Lemma 3.3.2.

Now we telescope the inequality above from $k = 0, \dots, N$ and get that

$$0 \leq \sum_{k=0}^{N} \lambda_k \mathcal{G}_{z^\star}(\hat{z}_k) \leq D_h(z^\star, z_0) - D_h(z^\star, z_{N+1}) - \sum_{k=0}^{N} D_h(z_{k+1}, z_k)$$

$$\implies 0 \leq \varepsilon \sum_{k=0}^{N} \mathcal{G}_{z^\star}(\hat{z}_k) \leq D_h(z^\star, z_0) \implies \lim_{k\to\infty} \mathcal{G}_{z^\star}(\hat{z}_k) = 0.$$

■

Special Cases

Here we will give some examples of how to choose $M_k = \partial\psi_k + K_M - K$ and how these choices relate to other algorithms.

Standard Setting: $\psi_k = h_k$ and $K_M = K$.

Let ψ_k equal some Legendre Bregman function h_k for each k and $K_M = K$. The kernel takes the form

$$M_k = \begin{bmatrix} \partial\psi_{xx} & \partial\psi_{xy} \\ \partial\psi_{yx} & \partial\psi_{yy} \end{bmatrix}.$$

We get the forward-backward step

$$\hat{z}_k = [\nabla\psi_k + K + \partial F]^{-1} \nabla\psi_k(z_k).$$

If we expand the forward-backward step we get

$$\begin{cases} \partial f(\hat{x}_k) \ni \nabla\psi_{xx} x_k + \nabla\psi_{xy} y_k - (\nabla\psi_{xx}\hat{x}_k + \nabla\psi_{xy}\hat{y}_k) - L^T\hat{y}_k \\ \partial g^*(\hat{y}_k) \ni \nabla\psi_{yx} x_k + \nabla\psi_{xy} y_k - (\nabla\psi_{yx}\hat{x}_k + \nabla\psi_{yy}\hat{y}_k) + L\hat{x}_k. \end{cases} \tag{3.5}$$

The Bregman projection has optimality condition (see (3.2))

$$\begin{cases} \nabla h_k(z_{k+1}) & = \nabla h_k(z_k) - \lambda_k(\nabla h_k(z_k) - \nabla h_k(\hat{z}_k)) \\ 0 & = \langle \nabla h(z_k) - \nabla h(\hat{z}_k), z_{k+1} - \hat{z}_k \rangle \end{cases}.$$

The unique solution of the Bregman projection can in this case be determined directly. In fact, the system of equations are satisfied by $\lambda_k = 1$ and $z_{k+1} = \hat{z}_k$. In other words, the Bregman projection step is redundant and

$$z_{k+1} = [\nabla\psi_k + A]^{-1} \nabla\psi_k(z_k).$$

Therefore, the Bregman NOFOB update step reduces to a non-linear resolvent with kernel $\nabla\psi_k$.

Bregman Chambolle-Pock: $\psi_k = h_k$, $K_M = K$, $\nabla\psi_{xy} = -L^T$ and $\nabla\psi_{yx} = -L$.

The Chambolle-Pock primal-dual method was introduced in [10]. Later it was extended into the Bregman Chambolle-Pock method in [11]. We will see that the Bregman Chambolle-Pock method is a special case of the standard setting of the Bregman NOFOB algorithm. If we fix $\nabla\psi_{xy} = -L^T$ and $\nabla\psi_{yx} = -L$, which are the gradients of the function $(x, y) \mapsto -\langle Lx, y\rangle$ with respect to x and y respectively, then the kernel takes the form

$$M_k = \begin{bmatrix} \partial\psi_{xx} & -L^T \\ -L & \partial\psi_{yy} \end{bmatrix}.$$

Using (3.5) we expand the forward-backward step and get

$$\begin{cases} \partial f(\hat{x}_k) \ni \nabla\psi_{xx}x_k - \nabla\psi_{xx}\hat{x}_k - L^T y_k \\ \partial g^*(\hat{y}_k) \ni \nabla\psi_{yy}y_k - \nabla\psi_{yy}\hat{y}_k - L(x_k - 2\hat{x}_k). \end{cases} \tag{3.6}$$

Since this is a special case of the Standard Setting the Bregman projection step is yet again redundant. This is exactly the Bregman Chambolle-Pock algorithm considered in [11] with the added restriction that the gradients of the Bregman functions ψ_{xx} and ψ_{yy} can be written as

$$\begin{cases} \partial\psi_{xx} = \sigma^{-1}I + \partial\hat{\psi}_{xx} \\ \partial\psi_{yy} = \tau^{-1}I + \partial\hat{\psi}_{yy} \end{cases}$$

for some Bregman functions $\hat{\psi}_{xx}, \hat{\psi}_{yy}$. Then we can form

$$M_k = \begin{bmatrix} \partial\hat{\psi}_{xx} & 0 \\ 0 & \partial\hat{\psi}_{yy} \end{bmatrix} + \begin{bmatrix} \sigma^{-1}I & -L^T \\ -L & \tau^{-1}I \end{bmatrix}$$

where the second operator is strictly monotone if $\sigma\tau < \frac{1}{\|L\|^2}$. This is exactly the convergence requirement of Theorem 1 in [11] which guarantees convergence of $\{z_k\}_{k=0}^{\infty}$ in duality gap. Note that the original Chambolle-Pock algorithm in [10] is retrieved from the case when $\hat{\psi}_{xx} = \hat{\psi}_{yy} = 0$.

Asymmetric Updates: $\psi_{xy} = \psi_{yx} = 0$ and $K_M = 0$.

This novel algorithm will be the main focus during Chapter 4 and will be implemented on a nontrivial problem, see Section (4.2). In this case we get the kernel

$$M_k = \nabla\psi - K = \begin{bmatrix} \nabla\psi_{xx} & 0 \\ 0 & \nabla\psi_{yy} \end{bmatrix} - \begin{bmatrix} 0 & L^T \\ -L & 0 \end{bmatrix}$$

with expanded forward-backward step

$$\begin{cases} \partial f(\hat{x}_k) \ni \nabla\psi_{xx}x_k - \nabla\psi_{xx}\hat{x}_k - L^T y_k \\ \partial g^*(\hat{y}_k) \ni \nabla\psi_{yy}y_k - \nabla\psi_{yy}\hat{y}_k + Lx_k. \end{cases} \tag{3.7}$$

Comparing this with the Bregman Chambolle-Pock (3.6) we see that M_k is no longer symmetrical, which implies that the Bregman projection step is no longer trivial in the sense that $z_{k+1} = \hat{z}_k$ no longer holds. On the other hand, the expanded form shows that the coupling between $\hat{x}_k$ and $\hat{y}_k$ is removed, i.e., (3.7) can be implemented in parallel.

By Proposition 3.3.1 we have that M_k is here monotone if both ψ_{xx} and ψ_{yy} are convex. This contrasts the strong constraints on ψ_{xx} and ψ_{yy} imposed by the Bregman Chambolle-Pock algorithm. Note that we are yet to provide the connection between ψ and some Bregman function. In fact, we have great flexibility in this choice. Suppose for instance that $f = \hat{f} + \phi$ where $\hat{f}, \phi : \mathbb{R}^n \to \overline{\mathbb{R}}$ are proper, closed, and convex. Suppose also that $\hat{f}$ is differentiable on $\operatorname{int}(\operatorname{dom}\hat{f})$. We can then use the theory of Chapter 2 and utilize relative smoothness. To that end, suppose that $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ is β-relatively smooth with respect to $\hat{f}$. We can let $\psi_{xx} = \beta h - \hat{f}$ which is then convex. The first part of (3.7) becomes

$$\begin{aligned} \beta\nabla h(\hat{x}_k) + \partial\phi(\hat{x}_k) &\ni \beta\nabla h(x_k) - \nabla\hat{f}(x_k) - L^T y_k \\ &\implies \hat{x}_k = \operatorname{prox}^h_{\beta^{-1}\phi}(x_k^+) \end{aligned}$$

where

$$x_k^+ = \nabla h^*(\beta\nabla h(x_k) - \nabla\hat{f}(x_k) - L^T y_k).$$

Note that in this example we have not yet mentioned any properties of ψ_{yy} and g^*. This means that a similar approach, as for ψ_{xx} and f, may be implemented there. In some ways, we have decomposed a primal-dual problem into two independent Bregman type methods.

Also note that if the dual part of the primal-dual problem is ignored, i.e., that $m = 0$ and $L = 0$, then this algorithm reduces to the proximal Bregman gradient from Section 2.4.

4

D-optimal Design

In this chapter, we will begin by introducing two seemingly different optimization problems. The first problem is a geometrical one, which finds the minimum-volume ellipsoid that covers some reasonable finite subset of $\mathbb{R}^n$. The second problem has its roots in statistical optimality design and has for example applications within machine learning. We will show that these problems are dual to one another, with no duality gap. We refer to [19] for a full treatment.

The second section will connect the theory from Chapter 2 with the algorithms of Chapter 3 and solve this aforementioned optimization problem. Furthermore, it will be solved in a primal-dual setting, with an added 1-norm $\delta \|Lx - b\|_1$ regularizer for chosen $L \in \mathbb{R}^{m\times d}, b \in \mathbb{R}^d$, and $\delta \in \mathbb{R}_+$.

4.1 D-optimal Design and Minimum-Volume Ellipsoids

Consider some set of points $\{x_i\}_{i=1}^m$ where $x_i \in \mathbb{R}^n$ and construct the data matrix

$$X = \begin{bmatrix} 1 & x_1^T \\ \vdots & \vdots \\ 1 & x_m^T \end{bmatrix} \in \mathbb{R}^{m\times(n+1)}.$$

We will consider the supervised learning task of regression. To that end, let $\{y_i\}_{i=1}^m$ where $y_i \in \mathbb{R}$ be corresponding labels and we seek to find some model $\mathbf{m} : \mathbb{R}^n \to \mathbb{R}$ such that

$$y_i \approx \mathbf{m}(x_i).$$

In the method of least squares, we consider an affine model $\mathbf{m}_\theta(x) = \langle \theta, (1, x)\rangle$ where $\theta \in \mathbb{R}^{n+1}$. Note that the bias of the model is incorporated by the leading 1 in the $(1, x)$ vector. If we incorporate noise into our model and suppose appropriate independence, we get the standard regression

$$Y = X\theta + \varepsilon,$$

where ε is a random variable with normal distribution $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$ and $\sigma \geq 0$. In the least squares problem setting we want to solve the minimization problem

$$\underset{\theta \in \mathbb{R}^{n+1}}{\text{minimize}} \ \frac{1}{2} \|X\theta - Y\|_2^2 ,$$

where

$$Y = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix}.$$

Let the columns of X be given by $X_i \in \mathbb{R}^m$ for $i = 1, \ldots, (n+1)$ which we will call the features of X. In order for our least squares problem to have a unique solution, we suppose that X has full column rank. This is equivalent with the features being linearly independent and $m > n$. In that case, the matrix $X^T X$ is invertible and the optimal $\theta^\star$ is given explicitly by

$$\theta^\star = (X^T X)^{-1} X^T Y.$$

Let $\Delta_{m-1} \subset \mathbb{R}^m$ be the ordinary $(m-1)$-dimensional simplex

$$\Delta_{m-1} = \left\{ u \in \mathbb{R}_+^m \mid \sum_{i=1}^m u_i = 1 \right\}.$$

Let $u \in \Delta_{m-1}$ and consider the Diagonal matrix $U = \mathrm{Diag}(u) \in \mathbb{R}^{m \times m}$. If we let $\mathcal{X} = \cup_{i=1}^m \{x_i\} \subset \mathbb{R}^n$ then we can associate for each such $u \in \Delta_{m-1}$ a probability measure $\xi : \mathcal{X} \to \mathbb{R}_+$ given by $\xi(x_i) = u_i$. In optimal design literature, such as [1], the measure ξ is called a **design** and from such a design we can define an **information matrix** $M(\xi) \in \mathbb{R}^{(n+1)\times(n+1)}$ by

$$M(\xi) = \int_{x \in \mathcal{X}} x x^T \mathrm{d}\xi(x).$$

In our case, where $\mathcal{X}$ is finite, we simply get that

$$M(\xi) = X^T U X$$

and we will from now on denote the right hand side by $M(U)$, for any $U = \mathrm{Diag}(u)$ such that $u \in \Delta_{m-1}$. In optimal design, a measure ξ is called ***D*-optimal** if it solves the following optimization problem

$$\underset{\xi}{\text{minimize}} \ [\det M(\xi)^{-1}].$$

Minimizing $\xi \mapsto \det M(\xi)^{-1}$ is the same as minimizing $\xi \mapsto \log \det M(\xi)^{-1}$ and so we arrive at the D-optimal design problem:

$$\underset{u \in \Delta_{m-1}}{\text{minimize}} \ -\log(\det X^T U X) \tag{4.1}$$

where as before $U = \text{Diag}(u)$.

The information matrix also appears in the weighted least squares model. Let us return to the linear regression model from before. If we transform the model by

$$Y' = \sqrt{U}Y, \quad X' = \sqrt{U}X, \quad \varepsilon' = \sqrt{U}\varepsilon,$$

and assume that X' still has full column rank. This is for example true when $u \in \mathbb{R}^m_{++}$. We once again get the standard regression

$$Y' = X'\theta + \varepsilon'$$

with $\varepsilon' \sim \mathcal{N}(0, \sigma^2 U)$ and unique solution

$$\begin{aligned}\theta^\star &= ((X')^T X')^{-1}(X')^T Y' \\ &= (X^T U X)^{-1} X^T U Y.\end{aligned}$$

We have abused some notation above. Indeed, Y is a random variable and $\theta^\star$ is a point estimator of the random variable

$$\begin{aligned}\Theta^\star &= ((X')^T X')^{-1}(X')^T Y' \\ &= (X^T U X)^{-1} X^T U (X\theta^\star + \varepsilon) \\ &= \theta^\star + (X^T U X)^{-1} X^T U \varepsilon \\ \implies \mathbb{E}[\Theta^\star] &= \theta^\star.\end{aligned}$$

Therefore, the point estimate $\theta^\star$ is unbiased and has variance

$$\begin{aligned}\text{Var}[\Theta^\star] &= \mathbb{E}[\Theta^\star - \theta^\star][\Theta^\star - \theta^\star]^T \\ &= \mathbb{E}[(X^T U X)^{-1} X^T U \varepsilon][(X^T U X)^{-1} X^T U \varepsilon]^T \\ &= (X^T U X)^{-1} X^T U \mathbb{E}[\varepsilon\varepsilon^T] U X (X^T U X)^{-1} \\ &= \sigma^2 Z^T Z\end{aligned}$$

where

$$Z = \sqrt{U}X[M(U)]^{-1}.$$

If we supposed that the variance of the error of the original relationship $Y \approx X\theta$ varied between observations, in the form of $\varepsilon \sim \mathcal{N}(0, \sigma^2 U^{-1})$, we end up with

$$\begin{aligned}\text{Var}[\Theta^\star] &= (X^T U X)^{-1} X^T \sqrt{U} \mathbb{E}[\varepsilon'(\varepsilon')^T] \sqrt{U} X (X^T U X)^{-1} \\ &= \sigma^2 [M(U)]^{-1}.\end{aligned}$$

This could for instance occur when each observation x_i is not a single measurement, but is in fact an average of $N_i > 0$ measurements each with the same variance σ^2. This would correspond to the error variances

$$\text{Var}[\varepsilon_i] = \frac{\sigma^2}{N_i}$$

and weight matrix $U = \mathrm{Diag}(N_1, \ldots, N_m)$.

Minimum-Volume Ellipsoids

Now let us turn to something seemingly different: the problem of finding the minimum-volume ellipsoid containing the set of points $\{x_i\}_{i=1}^m$ where each $x_i \in \mathbb{R}^n$. We will begin by describing how one can find the minimum-volume ellipsoid centered around the origin which includes the set $\mathcal{X} = \cup_{i=1}^m \{x_i\}$. For that reason, it is natural to assume that the span of $\mathcal{X}$ equals $\mathbb{R}^n$. Otherwise, the minimum-volume ellipsoid would be of volume 0 and lie in some subspace of $\mathbb{R}^n$ homeomorphic to $\mathbb{R}^p$ where $p < n$.

Let $\tilde{X} \in \mathbb{R}^{m \times n}$ be defined by

$$\tilde{X} = \begin{bmatrix} x_1^T \\ \vdots \\ x_m^T \end{bmatrix} \in \mathbb{R}^{m \times n},$$

which is assumed to have rank n which implies that $m \geq n$. Let B_n be the closed unit sphere in $\mathbb{R}^n$ with respect to the 2-norm. Let $L : \mathbb{R}^n \to \mathbb{R}^n$ be some invertible map and consider the image set $L^{-1}B_n$. We have that $y \in L^{-1}B_n$ if and only if $Ly \in B_n$ and

$$\begin{aligned} \|Ly\|_2^2 &\leq 1 \\ y^T L^T L y &\leq 1 \\ y^T H y &\leq 1 \end{aligned}$$

for some positive definite $H = L^T L \in \mathbb{S}_{++}^n$. Conversely, starting with a positive definite map $H \in \mathbb{S}_{++}^n$ we can by a Cholesky factorization obtain an invertible map $L : \mathbb{R}^n \to \mathbb{R}^n$ such that $y^T H y \leq 1$ implies that $y \in L^{-1}B_n$. What we have shown is that the ellipsoids centered around the origin in $\mathbb{R}^n$ are precisely described by positive definite matrices and are of the form

$$\mathcal{E}(H) = \{y \in \mathbb{R}^n \mid y^T H y \leq 1\}.$$

The volume of $\mathcal{E}(H)$ is now given directly by

$$\mathrm{vol}(\mathcal{E}(H)) = \det(L^{-1})\mathrm{vol}(B_n) = (\det H)^{-1/2}\mathrm{vol}(B_n).$$

Therefore, the task of minimizing the volume of $\mathcal{E}(H)$, where $\mathcal{E}(H)$ is some ellipsoid centered around the origin and covering $\mathcal{X}$, is equivalent with maximizing the determinant of H. We now arrive at the minimum-volume enclosing ellipsoid problem:

$$\underset{H \in C \cap \mathbb{S}_{++}^n}{\text{minimize}} \; -\log(\det H),$$

where C is the convex set

$$C = \{H \in \mathbb{R}^{n\times n} \mid x_i^T H x_i \leq n \text{ holds for all } i = 1, \ldots, m\}.$$

Note that the change of $x_i^T H x_i \leq n$ is nothing but a scaling of H and does not change the geometrical arguments above in any substantial way.

In [19, Theorem 2.2] it was shown that the two problems

$$\underset{u\in\Delta_{m-1}}{\text{maximize}} \ \log(\det \tilde{X}^T U \tilde{X}) \quad \text{and} \quad \underset{H\in C\cap\mathbb{S}^n_{++}}{\text{minimize}} \ -\log(\det H)$$

are dual problems to each other and if X has full column rank, then strong duality holds. Furthermore, by [19, Proposition 2.5], given solutions $U^\star = \text{Diag}(u), u \in \Delta_{m-1}$ and $H^\star \in \mathbb{R}^{n\times n}$ of the two problems, respectively, they are related by

$$[\tilde{X}^T U^\star \tilde{X}]^{-1} = H^\star$$

with $H^\star$ being unique.

Lastly, let us consider the problem of finding the minimum-volume ellipsoid including the set $\mathcal{X}$, without any restriction on the center of the ellipsoid. In order to avoid degenerate ellipsoid solutions of volume 0, we suppose that the affine hull of $\mathcal{X}$ equals $\mathbb{R}^n$. This means that for any $x \in \mathbb{R}^n$ there exists λ_i such that $x = \sum_{i=1}^m \lambda_i x_i$ where $\sum_{i=1}^m \lambda_i = 1$. This is equivalent to the span of the set $\{(1, x_i)\}_{i=1}^m$ containing the subset $\{1\} \times \mathbb{R}^n \subset \mathbb{R}^{n+1}$. This is again equivalent to the span of the set $\{(1, x_i)\}_{i=1}^m$ equaling the entire $\mathbb{R}^{n+1}$ and X from (4.1) has full column rank and $m \geq n + 1$.

As it turns out, by [19, Theorem 2.10], if we instead solve the D-optimal design problem (4.1), then the $n \times n$ lower right submatrix of $H^\star = [X^T U^\star X]^{-1}$, call it $\tilde{H}$, determines this minimum-volume ellipsoid. In fact, it is given by the ellipsoid

$$\mathcal{E}(\tilde{H}, \bar{x}) = \{x \in \mathbb{R}^n \mid (x - \bar{x})^T \tilde{H}(x - \bar{x}) \leq n\}.$$

The center $\bar{x}$ is given by the $(n \times 1)$ trailing part of $(\tilde{X})^T \text{diag}(U^\star)$.

The Wolfe-Atwood Algorithm

Here we will review the Wolfe-Atwood (WA) algorithm, its rate of convergence and its time complexity. The WA algorithm is specifically designed to solve the D-optimal design problem. It is based on a coordinate-ascent framework. The ascent part comes from this particular implementation which instead maximizes and solves the just as difficult problem

$$\underset{u\in\Delta_{m-1}}{\text{maximize}} \ \log(\det X^T U X).$$

We will in this thesis not go into detail on the correctness of the algorithm, but it is included for completeness (see Algorithm 2). For a full treatment,

Algorithm 2: The WA Algorithm

Let : $X \in \mathbb{R}^{m\times n}$ with full column rank,
$f(u) \coloneqq \log(\det X^T \operatorname{Diag}(u)X)$
Input : $u_0 \in \Delta_{m-1}$
Compute: $\omega_0 \coloneqq \nabla f(u_0)$
for $k = 0, 1, \ldots$ **do**
  $i \coloneqq \underset{r=1,\ldots,m}{\operatorname{argmax}} [\omega_k]_r$
  $j \coloneqq \underset{r=1,\ldots,m:[u_k]_r>0}{\operatorname{argmin}} [\omega_k]_r$
  $\varepsilon_+ \coloneqq ([\omega_k]_i - n)/n$
  $\varepsilon_- \coloneqq (n - [\omega_k]_j)/n$
  if $\varepsilon_+ > \varepsilon_-$ **then**
    $\lambda^\star \coloneqq ([\omega_k]_i - n)/((n-1)[\omega_k]_i)$
    $u_{k+1} \coloneqq (1+\lambda^\star)^{-1}(u_k + \lambda^\star e_i)$
  else
    $\lambda^\star \coloneqq ([\omega_k]_j - n)/((n-1)[\omega_k]_j)$
    $\lambda \coloneqq \max\{-[u_k]_j, \lambda^\star\}$
    $u_{k+1} \coloneqq (1+\lambda)^{-1}(u_k + \lambda e_i)$
  end
  Compute: ω_{k+1}
end

see [19] and [2]. In [19, Chapter 3] it is shown that the WA algorithm satisfies local linear convergence.

The computation step $\omega_0 \coloneqq \nabla f(u_0)$ in Algorithm 2 is computationally expensive. As we will see in the next Section, see (4.2), this involves inverting the $(n \times n)$ matrix $X^T \operatorname{Diag}(u)X$, which in of itself might be numerically demanding for large enough n. Even worse, this is repeated at every iteration when computing ω_{k+1}. In order to avoid this expensive $\mathcal{O}(n^3)$ complexity at every iteration, we can approximate each ω_{k+1} by a Cholesky factorization $LL^T = X^T \operatorname{Diag}(u)X$, see [19] for the details. The Cholesky factorization is generally still $\mathcal{O}(n^3)$. We solve this issue by performing rank-1 updates of the Cholesky factorization at each step. After this optimization, we arrive at an algorithm with complexity $\mathcal{O}(mn)$ at each iteration.

4.2 NOFOB Bregman Method for D-optimal Design

From now on, let $f : \mathbb{R}^m \to \overline{\mathbb{R}}$ denote the objective function of the D-optimal Design Problem

$$f(u) = -\log(\det X^T \operatorname{Diag}(u)X)$$

for some matrix $X \in \mathbb{R}^{m\times n}$ with full column rank and $m > n$. Of course, we interpret that if the determinant of $X^T \operatorname{Diag}(u)X$ is non-positive then $f(u) = \infty$. Let $\phi : \mathbb{R}^m \to \overline{\mathbb{R}}$ equal the indicator function $\iota_{\Delta_{m-1}}$. Let $g : \mathbb{R}^d \to \mathbb{R}$ be defined by

$$g(y) = \delta \left\| y - b \right\|_1$$

where $b \in \mathbb{R}^d$ and $\delta \in \mathbb{R}_+$ are assumed to be fixed. Furthermore, let the matrix $L \in \mathbb{R}^{m\times d}$ be fixed. We will from now on mainly consider the following regularized D-optimal design optimization problem

$$\underset{u\in\mathbb{R}^m}{\text{minimize}}\, f(u) + \phi(u) + g(Lu).$$

Let us begin by analyzing f and finding a suitable Bregman function relatively smooth to it.

First we quickly derive the standard expressions for the gradient and the Hessian of f. Consider the function $\hat{f} : \mathbb{R}^{m\times m} \to \overline{\mathbb{R}}$ defined by

$$\hat{f}(A) = -\log(\det A).$$

It is a standard result (see for example [9]) that for any small $\Delta A \in \mathbb{R}^{m\times m}$ such that $A + \Delta A \in \mathbb{S}^m_{++}$ it follows that

$$\hat{f}(A + \Delta A) \approx \hat{f}(A) + \operatorname{tr}(-A^{-1}\Delta A)$$

from which it follows that $\nabla \hat{f}(A) = -A^{-1}$. Consider the composite function $\hat{f}_X : \mathbb{R}^{m\times m} \to \overline{\mathbb{R}}$ given by

$$\hat{f}_X(A) = \hat{f}(X^T A X).$$

From the chain rule, we have that

$$\nabla \hat{f}_X(A) = -X(X^T A X)^{-1} X^T.$$

In order to compute $\nabla f(u)$ we consider only points A and directions ΔA which are diagonal matrices: $A = \operatorname{Diag}(u)$ and $\Delta A = \operatorname{Diag}(v)$ for $u, v \in \mathbb{R}^m_{++}$. Let the matrix $C \in \mathbb{R}^{m\times m}$ be given by

$$C = X(X^T \operatorname{Diag}(u) X)^{-1} X^T.$$

Note that if we assume that X has full column rank and $u \in \mathbb{R}^m_{++}$, then C is positive semi-definite. It is then a standard check to see that

$$\operatorname{tr}(-C\Delta A) = \langle \operatorname{diag}(-C), v\rangle,$$

where $\langle \cdot, \cdot \rangle$ is the normal inner product of $\mathbb{R}^m$. Moreover, this implies that

$$\nabla f(u) = \operatorname{diag}(-C). \tag{4.2}$$

We proceed similarly with the Hessian and use the standard result

$$\hat{f}_X(A + \Delta A) \approx \hat{f}_X(A) + \operatorname{tr}(-C\Delta A) + \frac{1}{2}\operatorname{tr}(C\Delta A C\Delta A).$$

The quadratic term has the following identity when $\Delta A = \operatorname{Diag}(v)$ for $v \in \mathbb{R}^m_{++}$:

$$\operatorname{tr}(C\Delta A C\Delta A) = v^T(C \circ C)v$$

and so

$$\nabla^2 f(u) = C \circ C$$

which by the Schur product theorem is positive semi-definite, since C is positive semi-definite. We have shown that f is convex on its domain and that $\mathbb{R}^m_{++} \subset \operatorname{dom} f$. Now let us determine some appropriate Bregman function $h : \mathbb{R}^m \to \overline{\mathbb{R}}$ with $\operatorname{dom} h = \mathbb{R}^m_{++}$. When considering the case of X equaling some identity matrix, we get that $C = U^{-1}$ and that $\nabla^2 f(u) = U^{-2}$ which puts some restrictions on the Bregman function. Namely, we must have that

$$U^{-2} \preceq \nabla h^2(u).$$

We claim that $\nabla^2 h(u) = U^{-2}$ is a valid choice for which f is 1-relatively smooth. We have the natural choice of h being the logarithmic barrier function

$$h(u) = -\sum_{i=1}^{m} \log u_i.$$

It can be verified that h is Legendre and that it has conjugate function

$$h^*(s) = -m - \sum_{i=1}^{m} \log(-s_i)$$

with domain $\operatorname{dom} h^* = \mathbb{R}^m_{--}$. Hence both h and h^* have open domain and the conclusion of Proposition 2.2.18 holds.

First we will show that

$$C \preceq U^{-1} \iff P := \sqrt{U}C\sqrt{U} \preceq I.$$

Note that P is symmetric and $P^2 = P$ holds. It follows that the mapping P is idempotent and only has eigenvalues that equal 0 or 1. Therefore, $C \preceq U^{-1}$ holds.

By the Schur product theorem once more we have that both $C \circ (U^{-1} - C)$ and $(U^{-1} - C) \circ U^{-1}$ are positive definite, which is just to say that

$$\nabla^2 f(u) = C \circ C \preceq C \circ U^{-1} \preceq U^{-1} \circ U^{-1} = \nabla^2 h(u).$$

We have shown that f is 1-relatively smooth with respect to h on $\operatorname{dom} h = \mathbb{R}^m_{++}$.

Let us compute the symmetry coefficient α_h. Consider the two points $\mathbb{1}_m, (x, \mathbb{1}_{m-1}) \in \mathbb{R}^m_{++}$ where $\mathbb{1}_m = (1, 1, \ldots, 1) \in \mathbb{R}^m$ is the vector consisting of m 1s. We then have that

$$\begin{aligned} D_h(\mathbb{1}_m, (x, \mathbb{1}_{m-1})) &= \frac{1}{x}(x \log x - x + 1) \\ D_h((x, \mathbb{1}_{m-1}), \mathbb{1}_m) &= -\log x + x - 1 \\ \implies \lim_{x\to\infty} \frac{D_h(\mathbb{1}_m, (x, \mathbb{1}_{m-1}))}{D_h((x, \mathbb{1}_{m-1}), \mathbb{1}_m)} &= \lim_{x\to\infty} \frac{x \log x - x + 1}{x^2 - x - x \log x} = 0 \geq \alpha_h. \end{aligned}$$

Since also $\alpha_h \geq 0$ we have that $\alpha_h = 0$.

Recall the function $g : \mathbb{R}^d \to \mathbb{R}$ given by

$$g(y) = \delta \|y - b\|_1 .$$

A property of g is that its conjugate is proximal (or Bregman proximal with respect to the Bregman function $h(\cdot) = \frac{1}{2}\|\cdot\|_2^2$). This is a desired property in our primal-dual setting, since g^* will appear and serve the dual part of the algorithm. Evaluating $\operatorname{prox}_{\gamma_k g^*}(y)$ is a standard procedure and treated in for instance [16]. In fact, it takes the easily computed form of

$$\operatorname{prox}_{\gamma_k g^*}(y) = \left[\frac{\delta(y_i - \gamma_k b_i)}{\max\{\delta, |y_i - \gamma_k b_i|\}}\right]_i .$$

We are now ready to fully write out how the regularized D-optimal design problem can be solved with asymmetric Bregman forward-backward splitting with projection correction. We will solve the primal-dual inclusion problem of the form

$$0 \in Az$$

for $z = (x, y) \in \mathbb{R}^m \times \mathbb{R}^d$ where

$$A = \begin{bmatrix} \partial(f + \iota_{\Delta_{m-1}}) & 0 \\ 0 & \partial g^* \end{bmatrix} + \begin{bmatrix} 0 & L^T \\ -L & 0 \end{bmatrix}.$$

The full operator $M_k : \mathbb{R}^m_{++} \times \mathbb{R}^d \to \mathbb{R}$ will be chosen as

$$M_k = \begin{bmatrix} \beta_k \nabla h - \nabla f & 0 \\ 0 & \frac{1}{\gamma_k} I \end{bmatrix} - \begin{bmatrix} 0 & L^T \\ -L & 0 \end{bmatrix}.$$

Note that for all step-sizes $\beta_k > 1$ and $\gamma_k > 0$ the operator M_k is monotone. The forward-backward step for any $z_k = (u_k, y_k) \in \mathbb{R}^m_{++} \times \mathbb{R}^d$ is given by

$$\begin{aligned}
\hat{z}_k &= [M_k + A]^{-1} M_k z \\
&= \begin{bmatrix} \beta_k \nabla h + \partial \iota_{\Delta_{m-1}} & 0 \\ 0 & \frac{1}{\gamma_k} I + \partial g^* \end{bmatrix}^{-1} \begin{bmatrix} \beta_k \nabla h(u_k) - \nabla f(u_k) - L^T y_k \\ L u_k + \frac{1}{\gamma_k} y_k \end{bmatrix} \\
&= \begin{bmatrix} \nabla h + \frac{1}{\beta_k} \partial \iota_{\Delta_{m-1}} \\ I + \gamma_k \partial g^* \end{bmatrix}^{-1} \begin{bmatrix} \nabla h(u_k) - \frac{1}{\beta_k}(\nabla f(u_k) - L^T y_k) \\ \gamma_k L u_k + y_k \end{bmatrix} \\
&\iff \begin{cases} 0 \in \frac{1}{\beta_k} \partial \iota_{\Delta_{m-1}}(\hat{u}_k) + \nabla h(\hat{u}_k) - (\nabla h(u_k) - \frac{1}{\beta_k}(\nabla f(u_k) - L^T y_k)) \\ \hat{y}_k = \operatorname{prox}_{\gamma_k g^*}(\gamma_k L u_k + y_k) \end{cases} \\
&\iff \begin{cases} \hat{u}_k = \underset{u \in \Delta_{m-1}}{\operatorname{argmin}} \; h(u) + \langle c, u \rangle, \text{ with } c = -(\nabla h(u_k) - \frac{1}{\beta_k}(\nabla f(u_k) - L^T y_k)) \\ \hat{y}_k = \operatorname{prox}_{\gamma_k g^*}(\gamma_k L u_k + y_k). \end{cases}
\end{aligned}$$

The subproblem

$$\hat{u}_k = \underset{u \in \Delta_{m-1}}{\operatorname{argmin}} \; h(u) + \langle c, u \rangle$$

does not have a closed form solution, but can be solved by a 1-dimensional line search. To see this, note an optimality condition of this subproblem is

$$c - \operatorname{diag}(U^{-1}) + \theta \mathbb{1}_m = 0$$

and so

$$u_i = \frac{1}{c_i + \theta} > 0 \implies \theta_i > -c_i.$$

We end up determining the θ such that

$$\sum_{i=1}^m \frac{1}{c_i + \theta} - 1 = 0, \text{ on } \theta \in (-\min\{c_i\}_{i=1}^m, \infty).$$

The function

$$d(\theta) = \sum_{i=1}^m \frac{1}{c_i + \theta} - 1$$

is strictly decreasing and $\lim_{\theta \to (-\min\{c_i\})^-} d(\theta) = \infty$ and $\lim_{\theta \to \infty} d(\theta) = -1$ and so a unique zero of d exists on the chosen interval. In our implementation, we solved this one dimensional problem with Newton's method with initial guess $\theta_0 = -\min\{c_i\} + 1$.

Let us now move on to the second step of the algorithm, which is the Bregman projection part. Let $h' : \mathbb{R}^m \times \mathbb{R}^d \to \overline{\mathbb{R}}$ be given by

$$h'(u, y) = \beta_k h(u) + \frac{1}{\gamma_k} \|y\|_2^2 .$$

The Bregman projection step is

$$\hat{z}_{k+1} = \underset{z \in H_k}{\operatorname{argmin}} \, D_{h'}(z, z_k) \tag{4.3}$$

where the hyperplane H_k is given by

$$H_k = \{z \in \mathbb{R}^m \times \mathbb{R}^d \mid \langle M_k z_k - M_k \hat{z}_k, z - \hat{z}_k \rangle = 0\}.$$

Note how H_k is now defined as a hyperplane instead of a half-space, as in (3.1). By Proposition 3.2.3 this difference does not change the algorithm. The optimality condition from the Lagrangian function of this subproblem (4.3) is given by

$$\begin{aligned} \nabla h'(z_{k+1}) - \nabla h'(z_k) - \lambda_k (M_k z_k - M_k \hat{z}_k) &= 0 \\ \langle M_k z_k - M_k \hat{z}_k, z - \hat{z}_k \rangle &= 0 \end{aligned}$$

for some $\lambda_k \in \mathbb{R}$. We can apply similar methods as we did when determining bounds on θ above. The first m components of this equation give some restrictions. Specifically, we need

$$[\nabla h(x_k)]_i - \lambda_k [M_k z_k - M_k \hat{z}_k]_i < 0 \text{ for all } i = 1, \ldots, m$$

in order for the optimality condition to make sense. This leads us to restrict λ_k to some interval (a_k, b_k) where the zero of the following function is unique:

$$\lambda_k \mapsto \langle M_k z_k - M_k \hat{z}_k, (\nabla h')^{-1}(\nabla h'(z_k) - \lambda_k (M_k z_k - M_k \hat{z}_k)) - \hat{z}_k \rangle.$$

Precomputing the appropriate vectors above leads to a problem which can be efficiently solved by some 1-dimensional method. In our implementation we used a bisection method.

4.3 Comparing Bregman NOFOB with the WA algorithm

Let us now consider five different instances of D-optimal design, represented by Figures 4.1 to 4.5. Throughout all of the following examples, the step-sizes $\beta_k = 1.1$ and $\gamma_k = 1$ were used. For more information about each instance, we refer to the appropriate figure caption.

In Figure 4.1 we see a visual representation of the D-optimal design problem over $\mathbb{R}^3$. In this and the next case, we have chosen $d = 0$. The dual part of the primal-dual algorithm is ignored in this simplified setting. The reason for this is that we cannot apply the WA algorithm in the primal-dual setting. In this small toy example, it is evident that the solution set $\operatorname{Zer}[A]$ lies in the boundary of the relative interior of Δ_2. Both the Bregman NOFOB and the WA algorithm converge to a solution, but with seemingly different speed

and qualitative trajectory. Note how the iterates in Figure 4.1a contrast the jagged nature of the WA iterates in Figure 4.1b. This is to be expected since the WA algorithm is of a coordinate-descent type.

In Figure 4.2 we see an instance with the same dimensions as in Figure 4.1 but with some different behaviour. For one, the solution set is infinite and at least parts of it lies in the relative interior of Δ_2. Furthermore, both the Bregman NOFOB (Figure 4.2a) and the WA algorithms (Figure 4.2b) converge to points in the relative interior of Δ_2. Since these points are different, it is not possible to compare algorithm convergence in this case w.r.t one particular solution $u^\star$.

In Figure 4.3, we have increased the dimensions to $m = 100$ and $n = 50$ in the same D-optimal design problem."Due to the higher dimension, we here compare function value suboptimality and distance to solution vs iteration in Figure 4.3. The point $u^\star$ was simply defined as the iterate from either of the two algorithms (during the 10^4 iterations) which achieved the lowest value $f(u^\star) := f^\star$. With the discussion in mind from the last paragraph, one is right to feel suspicious about Figure 4.3a and 4.3b. These plots are simply included to strongly indicate sequential convergence of the two algorithms. Furthermore, they show what kind of convergence can be expected. We have previously shown that the Bregman NOFOB (in this primal setting) has a sublinear $\mathcal{O}(1/k)$ convergence. We have also stated that the WA algorithm has been shown to achieve local linear convergence. Both of these results are exemplified here, in sequential and function value form, see Figures 4.3c and 4.3d.

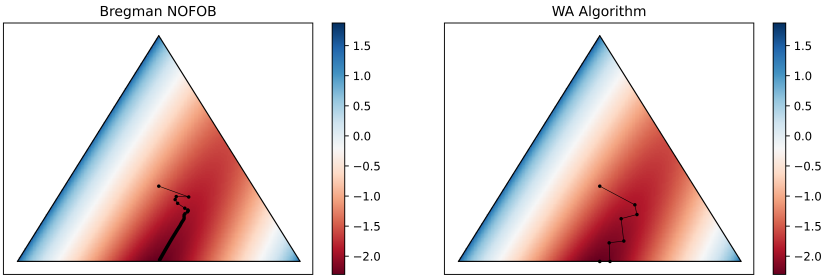
Recall that the WA algorithm, with rank-1 updates to the Cholesky factorization, achieves a $\mathcal{O}(mn)$ computational complexity per iteration. The Bregman NOFOB algorithm suffers greatly from the problem of evaluating $\nabla f(\cdot)$ multiple times each iteration. At a minimum, we need one evaluation of $\nabla f(u_k)$ during the forward-backward step and one evaluation of $\nabla f(\hat{u}_k)$ during the Bregman projection step, because of the inverse in the gradient expression (4.2). This leads to a substantially worse $\mathcal{O}(n^3)$ computational complexity per iteration. It would be very useful and interesting if there was a way to approximate these gradients in the Bregman NOFOB case as it was done in the WA algorithm.

Now let us increase the dimensionality of the problem and enter the primal-dual setting. In this case the WA algorithm cannot be applied. But the Bregman NOFOB method converges, with similar sublinear sequential and function value convergence as in Figure 4.3. Figure 4.4 shows two related plots during the first 10^4 iterates. Figure 4.4b shows the magnitude of the projection step-sizes λ_k. As we have hoped, given the theory of Chapter 3.3, they are uniformly bounded below by some positive number

$$\liminf_{k\to\infty} \lambda_k > 0.$$

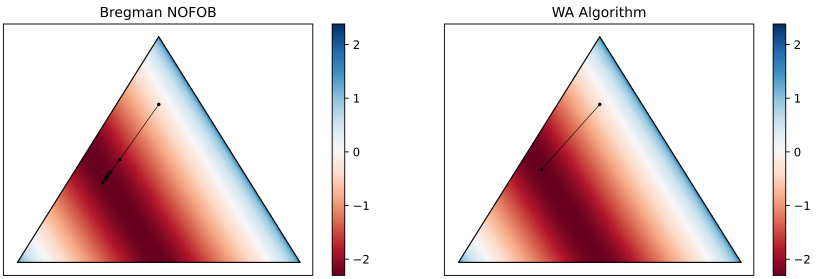
By Proposition 3.3.3, the duality gap given in (3.4) will converge to 0. This fact is demonstrated in Figure 4.4a.

Lastly, Figure 4.5 connects the primal D-optimal design problem to its dual minimum-volume ellipsoid problem. Figure 4.5a and 4.5b show two examples of this problem with 10 and 50 data-points x_i respectively. At the four iterate time points, of the Bregman NOFOB algorithm, $k = 0, 1, 10$, and 1000, we recovered the primal variable and plotted its corresponding ellipsoid. As expected, when dealing with more data-points, see Figure 4.5b, more iterations are needed. It is clearly visible that the 1000th ellipsoid does not yet contain all data. Indeed, it seems like that if the initial ellipsoid does not cover its data, then the ellipsoid iterates probably never will. This kind of behaviour is common in dual problems when recovering the primal iterates from the dual iterates.



(a) The Bregman NOFOB Algorithm. **(b)** The WA Algorithm.

Figure 4.1 The first $N = 1000$ iterations of the Bregman NOFOB and the WA algorithm on a simple instance of the D-optimal design problem. The dimensions for this problem are $m = 3$, $n = 2$, and $d = 0$. Each individual element of the matrix $X \in \mathbb{R}^{3\times 2}$ was sampled independently from a uniform distribution $\mathcal{U}_{[0,1)}$. The initial value u_0 was chosen as $u_0 = \frac{1}{3}(1, 1, 1)$. The black dots are plotted by the sequence $\{u_k\}_{k=0}^{N}$. The triangular region should be graphically interpreted as a 2-dimensional representation of $\Delta_2 \in \mathbb{R}^3$. The three corners correspond to the points $(1, 0, 0), (0, 1, 0), (0, 0, 1) \in \mathbb{R}^3$ and u_0 corresponds to the center of the triangle. The function values plotted in the heat-map are in fact $u \mapsto -\log(f(u) - f^\star)$ in order for the heat-map to be more readable.

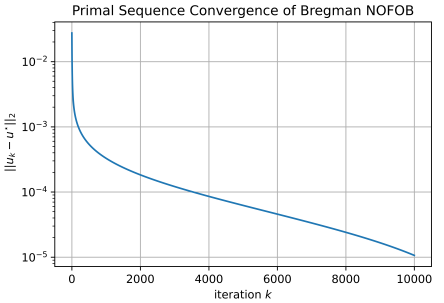


(a) The Bregman NOFOB Algorithm. **(b)** The WA Algorithm.

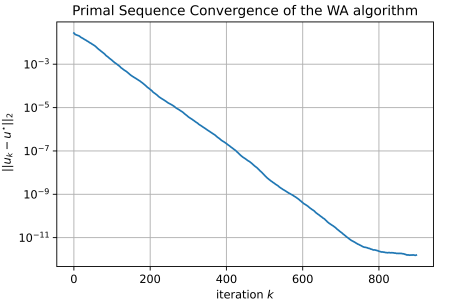
Figure 4.2 The first 30 iterations of the Bregman NOFOB and the WA algorithm on a simple instance of the D-optimal design problem. The dimensions for this problem are $m = 3$, $n = 2$, and $d = 0$. The matrix X was explicitly chosen as

$$X = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}.$$

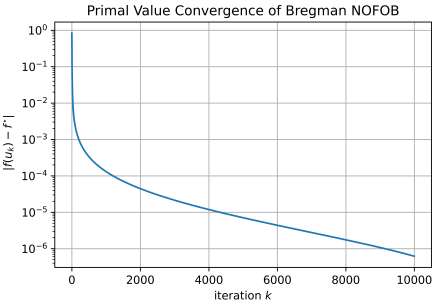
The initial value u_0 was chosen as $u_0 = (0.15, 0.15, 0.7)$.



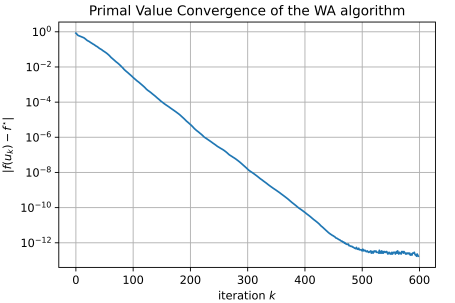
(a) Sequential convergence of Bregman NOFOB.



(b) Sequential convergence of the WA algorithm.

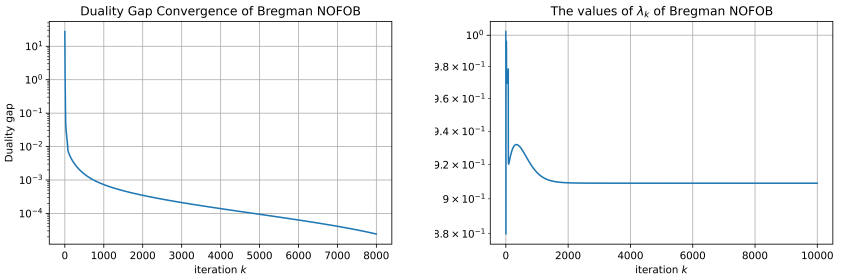


(c) Value convergence of Bregman NOFOB.



(d) Value convergence of the WA algorithm.

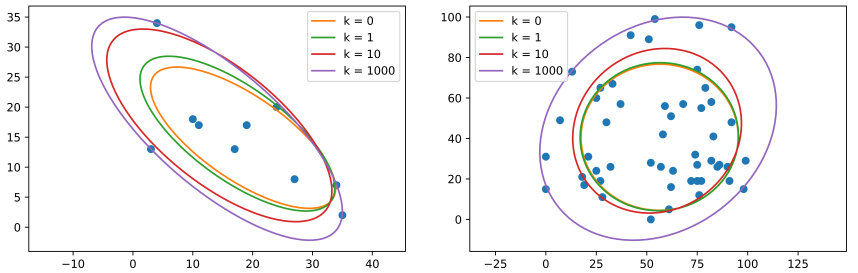
Figure 4.3 The first 10^4 iterations of the Bregman NOFOB and the WA algorithm on a more complex instance of the D-optimal design problem. The dimensions of this problem were $m = 100$, $n = 50$, and $d = 0$. The matrix X and the initial u_0 was chosen as in Figure 4.1. Note that the plots are in log-linear scale and that the scales of the plots differ.



(a) The duality gap iterations (defined in (3.4)) of Bregman NOFOB.

(b) The projection steps λ_k of Bregman NOFOB.

Figure 4.4 The first 10^4 iterations of the Bregman NOFOB on a more complex instance of the D-optimal design problem. The dimensions for this problem are $m = 100$, $n = 50$, and $d = 100$. The matrix X and the initial u_0 was chosen as in Figure 4.1. The vector b was chosen in a similarly random manner and we fixed $\delta = 1$.



(a) Iterates of minimum-volume ellipsoids with 10 data points.

(b) Iterates of minimum-volume ellipsoids with 50 data points.

Figure 4.5 The first 1000 iterations of the Bregman NOFOB algorithm on some simple instance of the D-optimal design problem, visualized in the dual-value setting. At four specific iterations we draw the corresponding ellipsoids described by the theory of Section 4.1.

Bibliography

[1] C. L. Atwood. "Optimal and efficient designs of experiments". *The Annals of Mathematical Statistics* (1969), pp. 1570–1602.

[2] C. L. Atwood. "Sequences converging to d-optimal designs of experiments". *The Annals of Statistics* (1973), pp. 342–352.

[3] N. Azizan, S. Lale, and B. Hassibi. "Stochastic mirror descent on overparameterized nonlinear models". *IEEE Transactions on Neural Networks and Learning Systems* (2021).

[4] H. H. Bauschke, J. Bolte, and M. Teboulle. "A descent lemma beyond lipschitz gradient continuity: first-order methods revisited and applications". *Mathematics of Operations Research* **42**:2 (2017), pp. 330–348.

[5] H. H. Bauschke and J. M. Borwein. "Joint and separate convexity of the bregman distance". In: *Studies in Computational Mathematics*. Vol. 8. Elsevier, 2001, pp. 23–36.

[6] H. H. Bauschke, J. M. Borwein, et al. "Legendre functions and the method of random bregman projections". *Journal of convex analysis* **4**:1 (1997), pp. 27–67.

[7] H. H. Bauschke, J. M. Borwein, and P. L. Combettes. "Bregman monotone optimization algorithms". *SIAM Journal on control and optimization* **42**:2 (2003), pp. 596–636.

[8] H. H. Bauschke, P. L. Combettes, et al. *Convex analysis and monotone operator theory in Hilbert spaces*. Vol. 408. Springer, 2011.

[9] S. Boyd, S. P. Boyd, and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[10] A. Chambolle and T. Pock. "A first-order primal-dual algorithm for convex problems with applications to imaging". *Journal of mathematical imaging and vision* **40**:1 (2011), pp. 120–145.

[11] A. Chambolle and T. Pock. "On the ergodic convergence rates of a first-order primal–dual algorithm". *Mathematical Programming* **159**:1 (2016), pp. 253–287.

[12] J. Eckstein. "Nonlinear proximal point algorithms using bregman functions, with applications to convex programming". *Mathematics of Operations Research* **18**:1 (1993), pp. 202–226.

[13] P. Giselsson. "Nonlinear forward-backward splitting with projection correction". *SIAM Journal on Optimization* **31**:3 (2021), pp. 2199–2226.

[14] H. Lu, R. M. Freund, and Y. Nesterov. "Relatively smooth convex optimization by first-order methods, and applications". *SIAM Journal on Optimization* **28**:1 (2018), pp. 333–354.

[15] A. S. Nemirovskij and D. B. Yudin. "Problem complexity and method efficiency in optimization" (1983).

[16] N. Parikh, S. Boyd, et al. "Proximal algorithms". *Foundations and trends® in Optimization* **1**:3 (2014), pp. 127–239.

[17] R. T. Rockafellar. *Convex analysis*. Vol. 18. Princeton university press, 1970.

[18] E. K. Ryu and W. Yin. *Large-Scale Convex Optimization: Algorithms & Analyses via Monotone Operators*. Cambridge University Press, 2022.

[19] M. J. Todd. *Minimum-volume ellipsoids: Theory and algorithms*. SIAM, 2016.

Lund University
Department of Automatic Control
Box 118
SE-221 00 Lund Sweden

Document name
MASTER'S THESIS

Date of issue
December 2022

Document Number
TFRT-6187

Author(s)
Max Nilsson

Supervisor
Pontus Giselsson, Dept. of Automatic Control, Lund University, Sweden
Emma Tegling, Dept. of Automatic Control, Lund University, Sweden (examiner)

Title and subtitle
Asymmetric Bregman Forward-Backward Splitting with Projection Correction

Abstract

This thesis examines first-order Bregman algorithms in a primal and a primal-dual setting. The Bregman gradient descent algorithm is introduced from a majorization-minimization perspective and as a generalization of the gradient descent algorithm. Concepts such as relative smoothness and Legendreness are defined and are shown to be natural restrictions in order to show convergence results.

A special case of the NOFOB algorithm, proposed by Giselsson in 2021, with a Bregman setting is defined, which we call the Bregman NOFOB algorithm. This algorithm works in a primal-dual setting and consists of a nonlinear forward-backward splitting step followed by a projection correction. Both of these components are discussed with respect to the Bregman setting. The Bregman NOFOB framework unifies multiple algorithms, one of which is the celebrated Bregman Chambolle-Pock method. It also allows us to define novel Bregman primal-dual algorithms. Under certain assumptions on the solution set of the problem and on the projection steps, we show that the Bregman NOFOB method converges in duality gap.

This Bregman NOFOB algorithm with asymmetric kernel is compared with theWolfe-Atwood (WA) algorithm on the D-optimal design optimization problem. As confirmed by the theory of this thesis – in the primal case - the two algorithms both converge by sequence and by function value, with sublinear (Bregman NOFOB) and linear (WA) rates. In the primal-dual case we experimentally show that the projection step sizes satisfy the duality gap convergence of the Bregman NOFOB algorithm. Indeed, this duality gap convergence is also verified experimentally. No comparison with the WA algorithm is made in the primal-dual case, since it is restricted to the primal setting.

Keywords

Classification system and/or index terms (if any)

Supplementary bibliographical information

ISSN and key title
0280-5316

ISBN

Language
English

Number of pages
1-65

Recipient's notes

Security classification

http://www.control.lth.se/publications/