ALTERNATING BREGMAN PROJECTIONS AND CONVERGENCE OF THE EM ALGORITHM

DOMINIKUS NOLL

Abstract. We investigate convergence of alternating Bregman projections between non-convex sets and prove convergence to a point in the intersection, or to points realizing a gap between the two sets. The speed of convergence is generally sub-linear, but may be linear under transversality. We apply our analysis to prove convergence of versions of the expectation maximization algorithm for non-convex parameter sets.

Keywords. Alternating Bregman projections · definable sets · o-minimal structures · EM algorithm · *em*-algorithm · Kullback-Leibler divergence · information geometry

AMS Classification 65K05 · 49J52 · 62D10 · 62B11

1. Introduction

Given a finite dimensional euclidean space $\mathbb{R}^d$ with inner product $\langle\cdot,\cdot\rangle$ and induced norm $\|\cdot\|$, and a convex function $f:\mathbb{R}^d\to\mathbb{R}\cup\{+\infty\}$ of Legendre type [53, Sect. 26], [8], the *Bregman distance* or *Bregman divergence* associated with f is defined as

$$D(x,y)=\begin{cases} f(x)-f(y)-\langle\nabla f(y),x-y\rangle & \text{if } y\in\text{int}(\text{dom}f)=:G\\ +\infty & \text{otherwise}\end{cases} \tag{1.1}$$

Given a closed subset $B\subset\text{dom}f$, the left Bregman distance to B, and the left Bregman projection onto B, are defined through

$$\overleftarrow{D}_B(a)=\min\{D(b',a):b'\in B\},\quad \overleftarrow{P}_B(a)=\text{argmin}\{D(b',a):b'\in B\}, \tag{1.2}$$

where the operator $\overleftarrow{P}_B$ is set-valued. Analogously, the right Bregman distance to, and projector onto, a closed set $A\subset G$ are

$$\vec{D}_A(b)=\min\{D(b,a'):a'\in A\},\quad \vec{P}_A(b)=\text{argmin}\{D(b,a'):a'\in A\}. \tag{1.3}$$

We consider sequences $a_n\in A$, $b_n\in B$ generated by alternating left-right Bregman projections as

$$b_n\in\overleftarrow{P}_B(a_n),\ a_{n+1}\in\vec{P}_A(b_n),\ n\in\mathbb{N},$$

and seek conditions under which these converge to a point in the intersection, $a_n,b_n\to x^*\in A\cap B$, or in the infeasible case $A\cap B=\emptyset$, to points $a_n\to a^*\in A$, $b_n\to b^*\in B$ minimizing the Bregman distance between A and B. We use the notation,

$$a_n\xrightarrow{l}b_n\xrightarrow{r}a_{n+1}\xrightarrow{l}b_{n+1},$$

and also the index-free form,

$$a\xrightarrow{l}b\xrightarrow{r}a^+\xrightarrow{l}b^+, \tag{1.4}$$

with $b\in\overleftarrow{P}_B(a)$, $a^+\in\vec{P}_A(b)$, $b^+\in\overleftarrow{P}_B(a^+)$, referring to these as *building blocks* of the alternating sequence. Note that (1.4) gives decrease of the distance in the sense that

$$D(b^+,a^+)\le D(b,a^+)\le D(b,a). \tag{1.5}$$

Alternating Bregman projections were first proposed in [23] as iterated left projections $a\xrightarrow{l}b\xrightarrow{l}a^+\xrightarrow{l}b^+$ between closed convex sets A,B. In the convex case substantial literature on Bregman projections is available, see e.g., [8, 9, 10, 11, 16, 25, 26, 28, 32, 31, 39]. Work addressing the non-convex case is scarce and includes for instance [18], even though many practical applications use non-convex alternating Bregman procedures without proper convergence certificate.

A strong motivation for the setup (1.4) is that it covers instances of the Expectation Maximization algorithm (EM algorithm), where the Bregman distance specializes to the *Kullback-Leibler divergence*. This link was established in [32, 31], where the authors introduce the *em-algorithm*, known to coincide

with the EM algorithm in the majority of cases [2]. In [32] convergence results for (1.4) with A, B convex are obtained. Convergence for certain non-convex instances of (1.4) appear already in [60]. Here we prove convergence for *definable* non-convex parameter sets, a hypothesis satisfied in practice.

Since alternating Bregman projections include euclidean alternating projections (AP), it is worth checking what is known in that case, as this gives an idea of what we may expect to achieve. Local convergence of non-convex AP was proved in [45] for *transversal intersections*, with [44, 12, 13] expanding on that idea, and here one expects linear convergence near $A \cap B$. *Tangential intersections* were considered in [52, 51], and then convergence drops to sub-linear speed. Convergence for tangential intersections is obtained under the *angle condition*, a geometric form of the *Kurdyka-Łojasiewicz inequality*, controlling the mutual position of the two sets near $x^* \in A \cap B$, and this extends also to the infeasible case; [52, 51, 61]. Presently we derive similar geometric notions for alternating Bregman projections, including both the feasible and the infeasible case.

Along with geometry, convergence requires as second ingredient a weak form of *regularity* in at least one of the sets, or less bindingly in [52, 51], a mild form of regularity which one of the sets has with regard to the other. Even in the euclidean case [14, 15, 52] regularity hypotheses cannot be avoided. Here we rely on the *three-point-inequality*, which in the non-convex setting made its first appearance in [52]. We obtain a suitable extension to Bregman alternating projections, and for the purpose of justification, we show how the three point inequality can be derived from conditions on the reach of A, B.

As demonstrated in [52, 51] for AP, angle condition and regularity are truly versatile when allowed to break the symmetry between A, B. By that we mean that under (1.4) regularity of B with regard to A does not have the same effect as regularity of A with regard to B. Since Bregman alternating projections are by themselves already asymmetric, half the theory applies to lr-building blocks $a \xrightarrow{l} b \xrightarrow{r} a^+$, the other half to rl-blocks $b \xrightarrow{r} a^+ \xrightarrow{l} b^+$. Fortunately, a *duality principle* allows to pass from one to the other, leading to a more unified convergence theory.

We obtain applications to the EM algorithm for discrete distributions, and for exponential families. In general only a sub-linear convergence rate $O(k^{-\rho})$ for some $\rho \in (0, \infty)$ can be affirmed. Our non-smooth geometric approach has elements in common with *information geometry*, a field combining Riemannian geometry, statistics and probability [2, 3].

The structure of the paper is as follows. Section 2 is preparatory, as is Section 3, where the classical notion of reach is extended to the Bregman setting. Section 4 prepares the infeasible case, and in Section 5 the geometric angle condition is derived from the Kurdyka-Łojasiewicz inequality. We also examine how the angle condition relates to tangential and transversal intersections. Section 6 extends the three-point-inequality from [52, 51] to the Bregman setting. Convergence is obtained in Section 7, dual convergence in Section 8. The worst case speed of convergence is considered in Section 9. In Section 10 we obtain sufficient conditions for the three point inequality from properties based on Bregman reach. The EM algorithm is discussed in Section 11. Terminology generally follows [54, 49], and [8, 18, 53] are useful for properties of Bregman distances and projectors.

2. Preparation

In this section we specify standing hypotheses, recall known results on Bregman projections, and discuss notions of reach. Throughout we assume that f is of Legendre type [53, 8], of class C^2 in the interior $G = \mathrm{int}(\mathrm{dom} f)$ of its domain, and satisfies $\nabla^2 f(x) \succ 0$ for every $x \in G$.

2.1. **Well-posedness of the alternating sequence.** Since $f - \langle \nabla f(a), \cdot \rangle$ is coercive for $a \in G$ by [53, Cor. 14.2.2] or [8, Thm. 3.7], existence of the left projection $\tilde{P}_B(a)$ of $a \in G$ in the sense that $\emptyset \neq \tilde{P}_B(a) \subset B \cap \mathrm{dom} f$ is assured as soon as $B \cap \mathrm{dom} f \neq \emptyset$ is closed in $\mathrm{dom} f$. This does not even require B to be closed.

On the other hand, existence of the right projection $\vec{P}_A(b)$ of $b \in \mathrm{dom} f$ needs in the first place $A \subset G$, but since $D(b, \cdot)$ is not coercive in general, we assume in the alternative that $A \subset G$ is closed bounded to get $\emptyset \neq \vec{P}_A(b) \subset A$. With these hypotheses alternating sequences are well-defined.

2.2. **Interiority.** The situation seems even less binding than for iterated left projections [23, 8], where in order to continue left projecting from $y^+ \in \tilde{P}_B(y)$, one needs $y^+ \in G$, a quest known as

interiority. While at first sight we do not need this here, we have different, yet compelling, reasons why we also want interiority $a_k, b_k \in G$ of the alternating sequence.

We call B *interiority preserving* if $B \cap \mathrm{dom} f \neq \emptyset$ is closed in $\mathrm{dom} f$ and $\overleftarrow{P}_B(a) \subset B \cap G$ for all $a \in G$. It follows from [8, Thm. 3.12] that a closed convex B is interiority preserving iff it satisfies the constraint qualification $B \cap G \neq \emptyset$. For non-convex B, the constraint qualification is necessary but no longer sufficient, a counterexample being given in Section 12.

When B is interiority preserving, then a_k, b_k stay in $G = \mathrm{int}(\mathrm{dom} f)$, which has the benefit to enable duality, given in the next section. However, to make full use of this, we also need accumulation points a^* of the a_k and b^* of the b_k to stay in G. Due to closedness of $A \subset G$, this is clear for the a^*.

Let b^* be an accumulation point of the b_k. By boundedness of A we may pass to subsequences $b_k \to b^*$, $a_k \to a^*$, $b_k \in \overleftarrow{P}_B(a_k)$. Suppose $b_k \to b^* \notin \mathrm{dom} f$. By coercivity we have $f(b_k) - \langle \nabla f(a^*), b_k \rangle \to \infty$, hence also $D(b_k, a_k) = f(b_k) - f(a_k) - \langle \nabla f(a^*), b_k - a_k \rangle + \langle \nabla f(a^*) - \nabla f(a_k), b_k - a_k \rangle \to \infty$, which is absurd because by (1.5) the sequence $D(b_k, a_k)$ is bounded. Therefore $b_k \to b^* \in \mathrm{dom} f$. But then $b^* \in \overleftarrow{P}_B(a^*)$, hence $b^* \in G$ since B is interiority preserving and $a^* \in G$. The agreeable consequence is that the alternating sequence stays in a compact subset of G if $A \subset G$ is closed bounded and B is interiority preserving.

For any such a_k, b_k we may now find a closed bounded subset B' of B such that $B' \subset G$ and $\overleftarrow{P}_B(a_k) \subset B'$, so that a_k, b_k remain alternating between A and B' with $\overleftarrow{P}_B(a_k) = \overleftarrow{P}_{B'}(a_k)$. This means the assumption $A \subset G$ closed bounded, B interiority preserving, can without loss of generality be replaced by the hypothesis $A, B \subset G$ closed bounded, which we adopt during the following sections.

We sketch one construction of B' when B is bounded. Find a ball $B(z, r)$ containing A, B in its interior, and let $G_0 = G \cap \mathrm{int} B(z, r)$. Then $G_1 = \mathrm{cl}\, G_0$ is a bounded closed convex body containing B, and containing A in its interior. Now use a standard approximation of G_1 by polytopes P_n from within [37], i.e., $P_n \subset G_0 \subset G_1 \subset (1 + \frac{1}{n}) P_n$. Let K be the set of all b_k and all their accumulation points, then K is compact, and by the above contained in G_0, hence $\mathrm{dist}(K, \partial G_0) > 0$. Therefore some $P = P_n$ contains K in its interior. Let $B' = B \cap P$, then a_k, b_k are alternating between A, B', and all accumulation points of the b_k are also in B'. B' is convex when B is, and is definable when B is (see Section 2.6). Moreover, B' is closed, because $B \cap \mathrm{dom} f = C \cap \mathrm{dom} f$ for a closed set C, hence $B \cap P = B \cap \mathrm{dom} f \cap P = C \cap \mathrm{dom} f \cap P = C \cap P$ is closed.

2.3. **Legendre duality.** The conjugate f^* of a function of Legendre type f is again of Legendre type [53, 8], associated with $G^* = \mathrm{int}(\mathrm{dom} f^*)$, so along with the Bregman distance D generated by f we also consider the distance D^* generated by f^* (cf. [53, Thm. 26.5], [8]). For $x, y \in G = \mathrm{int}(\mathrm{dom} f)$ it is known (cf. [8, Thm. 3.7]) that

$$D(x, y) = D^*(\nabla f(y), \nabla f(x)), \tag{2.1}$$

and this gives a link between left and right projections

$$\vec{P}_A = \nabla f^* \circ \overleftarrow{P}^*_{\nabla f(A)} \circ \nabla f, \tag{2.2}$$

obtained in [18, Prop. 7.1]. Here $\vec{P}^*_{\nabla f(B)}$, $\overleftarrow{P}^*_{\nabla f(A)}$ stand for projections with regard to f^*, D^*. Swapping f and f^*, we can also derive the formula

$$\overleftarrow{P}_B = \nabla f^* \circ \vec{P}^*_{\nabla f(B)} \circ \nabla f,$$

so that $\vec{P}_A \circ \overleftarrow{P}_B = \nabla f^* \circ \overleftarrow{P}^*_{\nabla f(A)} \circ \vec{P}^*_{\nabla f(B)} \circ \nabla f = \nabla f^* \circ \overleftarrow{P}^*_{B^*} \circ \vec{P}^*_{A^*} \circ \nabla f$, using $\nabla f^* \circ \nabla f = id$. Iterating this, we see that every alternating sequence (1.4) has a mirror sequence in dual space. More precisely, with

$$a^* = \nabla f(b),\ b^* = \nabla f(a^+),\ a^{+*} = \nabla f(b^+), \quad A^* = \nabla f(B),\ B^* = \nabla f(A),$$

we transform rl-building blocks $b \xrightarrow{r} a^+ \xrightarrow{l} b^+$ between $A, B \subset G$ into lr-blocks $a^* \xrightarrow{l*} b^* \xrightarrow{r*} a^{+*}$ between $A^*, B^* \subset G^*$. Commutativity of the following diagrams is referred to as the *duality principle*.

$$\begin{array}{ccccc} b & \xrightarrow{r} & a^+ & \xrightarrow{l} & b^+ \\ {\scriptstyle\nabla f}\downarrow & & \downarrow & & \downarrow \\ a^* & \xrightarrow{l*} & b^* & \xrightarrow{r*} & a^{+*} \end{array} \qquad \begin{array}{ccccc} a & \xrightarrow{l} & b & \xrightarrow{r} & a^+ \\ {\scriptstyle\nabla f}\downarrow & & \downarrow & & \downarrow \\ b^* & \xrightarrow{r*} & a^{+*} & \xrightarrow{l*} & b^{+*} \end{array}$$

It allows us to concentrate e.g. on results for rl-building blocks, obtaining those for lr-building blocks with minor effort. This will be used repeatedly.

2.4. **Norm bounds.** A consequence of non-degeneracy $\nabla^2 f(x) \succeq \epsilon I \succ 0$ of the Hessian on any compact $K \subset G$ is that we have an estimate of the form

$$m^2\|x-y\|^2 \leq D(x,y) \leq M^2\|x-y\|^2,\ x,y \in K, \tag{2.3}$$

with m, M depending only on f and K. Since Legendre functions satisfy $(\nabla f)^{-1} = \nabla f^*$, we also have an estimate of the form

$$l\|x-y\| \leq \|\nabla f(x) - \nabla f(y)\| \leq L\|x-y\|,\ x,y \in K, \tag{2.4}$$

where L is the Lipschitz constant of ∇f on K, while l is one over the Lipschitz constant of $(\nabla f)^{-1} = \nabla f^*$ on K. These two imply

$$m^2L^{-2}\|\nabla f(x) - \nabla f(y)\|^2 \leq D(x,y) \leq M^2 l^{-2}\|\nabla f(x) - \nabla f(y)\|^2 \tag{2.5}$$

on any such K. Since by our preprocessing in Section 2.2 we arranged $A, B \subset K \subset G$, these norm bounds will become useful in convergence analysis.

2.5. **Uniform second-order differentiability.** The fact that f is class C^2 on G assures that it has a uniform second-order Taylor-Young expansion on every compact $K \subset G$. More precisely, given $\epsilon > 0$, there exists $\delta > 0$ such that for all $x_0 \in K$ and all $x \in G$, $\|x - x_0\| < \delta$ implies $\big|f(x) - f(x_0) - \langle \nabla f(x_0), x - x_0\rangle - \frac{1}{2}\langle \nabla^2 f(x_0)(x-x_0), x-x_0\rangle\big| \leq \epsilon\|x-x_0\|^2$. For the notion of uniform differentiability see [55].

2.6. **Kurdyka-Łojasiewicz inequality.** The following definition is essential for our convergence theory.

Definition 2.1. (Kurdyka-Łojasiewicz inequality). A lower semi-continuous function $F : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ has the KŁ-property at $\bar{x} \in \mathrm{dom}(\partial F)$ if there exists $\eta > 0$, a neighborhood U of $\bar{x}$, and a continuous concave function $\phi : [0, \eta) \to \mathbb{R}_+$, called *de-singularizing function*, such that

i. $\phi(0) = 0$,
ii. ϕ is of class C^1 on $(0, \eta)$,
iii. $\phi'(s) > 0$ for $s \in (0, \eta)$,
iv. For all $x \in U \cap \{x : F(\bar{x}) < F(x) < F(\bar{x}) + \eta\}$ the KŁ-inequality

$$\phi'(F(x) - F(\bar{x}))\mathrm{dist}(0, \partial F(x)) \geq 1 \tag{2.6}$$

is satisfied, where ∂F is the limiting subdifferential.

Remark 2.2. We say that F satisfies the Łojasiewicz inequality when the de-singularizing function is $\phi'(s) = s^{-\theta}$ for some $\theta \in [\frac{1}{2}, 1)$, which means $\phi(s) = \frac{s^{1-\theta}}{1-\theta}$.

Remark 2.3. Information on convergence via the KŁ-property give e.g. [1, 4, 5, 21, 50, 6]. It is well-known that definability in an o-minimal structure [33, 34, 35], for short *definability*, implies the KŁ-inequality. See [43], and for non-smooth F, [22, Thm. 11], where it is shown that ϕ may be chosen concave. We shall have occasion to use the small o-minimal structure $\mathbb{R}_{an}$ of globally sub-analytic sets, but also large o-minimal structures containing at least $\mathbb{R}_{an,\exp}$, allowing exponential and logarithm. See in particular [33, 19, 30, 56, 59, 48].

3. Bregman reach

This section is still preparatory, but some results are of independent interest, in particular those concerning reach in the Bregman context, as well as concepts like Bregman geodesics encountered in information geometry [3].

3.1. **Bregman balls.** Bregman distances lead to two types of Bregman balls

$$\overleftarrow{\mathcal{B}}(a,r) = \{x \in \mathbb{R}^n : D(x,a) \le \tfrac{1}{2}r^2\}, \quad \vec{\mathcal{B}}(b,r) = \{x \in \mathbb{R}^n : D(b,x) \le \tfrac{1}{2}r^2\},$$

which generalize euclidean balls $B(x,r)$ in a natural way. Suppose $\overleftarrow{\mathcal{B}}(a,r) \subset G$, and let $b \in \partial\overleftarrow{\mathcal{B}}(a,r)$ be a point on the boundary, then there exists a euclidean ball entirely contained in $\overleftarrow{\mathcal{B}}(a,r)$, which touches the boundary $\partial\overleftarrow{\mathcal{B}}(a,r)$ at b from within $\overleftarrow{\mathcal{B}}(a,r)$. We ask how the radius of this euclidean ball is related to r.

Lemma 3.1. *Let $\overleftarrow{\mathcal{B}}(a,r) \subset G$, and define*

$$\overline{\kappa} := \max_{x \in \partial\overleftarrow{\mathcal{B}}(a,r)} \frac{\lambda_{\max}(\nabla^2 f(x))}{\|\nabla f(x) - \nabla f(a)\|}, \quad \underline{\kappa} := \min_{x \in \partial\overleftarrow{\mathcal{B}}(a,r)} \frac{\lambda_{\min}(\nabla^2 f(x))}{\|\nabla f(x) - \nabla f(a)\|}. \tag{3.1}$$

Then the euclidean ball with radius $1/\overline{\kappa}$ rolls freely inside the left Bregman ball $\overleftarrow{\mathcal{B}}(a,r)$. In addition, for every $b \in \partial\overleftarrow{\mathcal{B}}(a,r)$, this Bregman ball is contained in the euclidean ball with radius $1/\underline{\kappa}$ which has the same tangent hyperplane as $\overleftarrow{\mathcal{B}}(a,r)$ at b and has its center on the same side of the tangent hyperplane as a.

Proof. The boundary $\partial\overleftarrow{\mathcal{B}}(a,r)$ of $\overleftarrow{\mathcal{B}}(a,r)$ is the smooth surface implicitly given by the equation $F(x) = D(x,a) - \frac{1}{2}r^2 = 0$. The normal curvature of the surface at $x \in \partial\overleftarrow{\mathcal{B}}(b,r)$ in unit tangential direction v is therefore

$$\kappa_n(x,v) = \frac{\langle v, \nabla^2 F(x) v\rangle}{\|\nabla F(x)\|} = \frac{\langle v, \nabla^2 f(x) v\rangle}{\|\nabla f(x) - \nabla f(a)\|}, \tag{3.2}$$

so that $\underline{\kappa} \le \kappa_n(x,v) \le \overline{\kappa}$ for all $x \in \partial\overleftarrow{\mathcal{B}}(a,r)$ and all unit v from (3.1). By the Blaschke rolling theorem [20, §24, IV, p. 118] the euclidean ball with radius $\min_{x,\|v\|=1} 1/\kappa_n$ rolls freely inside the convex body $\overleftarrow{\mathcal{B}}(b,r)$, hence so does the smaller ball with radius $1/\overline{\kappa}$.

Conversely, the tangent hyperplane to $\overleftarrow{\mathcal{B}}(a,r)$ at b being $H = \{x : \langle \nabla f(a) - \nabla f(b), x - b\rangle = 0\}$, the euclidean ball $B(z, 1/\underline{\kappa})$ with $z = b + (1/\underline{\kappa})(\nabla f(a) - \nabla f(b))/\|\nabla f(a) - \nabla f(b)\|$ has H as tangent hyperplane at b, and has it center z on the same side of H as a. Since by (3.1) the normal curvature $\underline{\kappa}$ of the euclidean ball is everywhere smaller than the normal curvature $\kappa_n(x,v)$ of the Bregman ball, we get $\overleftarrow{\mathcal{B}}(a,r) \subset B(z, 1/\underline{\kappa})$, again by Blaschke's rolling theorem. □

On any compact subset of the interior of the domain of f the eigenvalues of the Hessians $\nabla^2 f(x)$ are bounded below and above by constants $0 < \lambda \le \Lambda < \infty$. Using (2.3) and (2.4), this gives $\overline{\kappa} \le \Lambda l^{-1}\|x-a\|^{-1} \le \Lambda l^{-1} M D(x,a)^{-1/2} = \Lambda l^{-1} M\sqrt{2} r^{-1}$, and $\underline{\kappa} \ge \lambda L^{-1}\|x-a\|^{-1} \ge \lambda m L^{-1} D(x,a)^{-1/2} = \lambda m L^{-1}\sqrt{2} r^{-1}$. We have proved the following

Proposition 3.2. *For every compact subset K of the interior of* $\mathrm{dom} f$ *there exist constants $0 < \underline{c} \le \overline{c}$ such that the following is true: If $b \in \partial\overleftarrow{\mathcal{B}}(a,r)$ with $\overleftarrow{\mathcal{B}}(a,r) \subset K$, then a euclidean ball of radius $\underline{c}r$ is contained in $\overleftarrow{\mathcal{B}}(a,r)$ and touches $\partial\overleftarrow{\mathcal{B}}(a,r)$ at b from within, and $\overleftarrow{\mathcal{B}}(a,r)$ is contained in a euclidean ball with radius $\overline{c}r$ which touches $\overleftarrow{\mathcal{B}}(a,r)$ at b from outside.*

We next consider right Bregman balls. Since these need not be convex, we need an extension of Blaschke's rolling theorem.

Lemma 3.3. *Let f be of class $C^{2,1}$ on G, and suppose $\vec{\mathcal{B}}(b,r) \subset G$. Let $1/r_0$ be a Lipschitz constant of $n(x) = \frac{\nabla^2 f(x)(x-b)}{\|\nabla^2 f(x)(x-b)\|}$ on $\partial\vec{\mathcal{B}}(b,r)$. Then a euclidean ball of radius r_0 rolls freely inside $\vec{\mathcal{B}}(b,r)$.*

Proof. Note that $\partial\vec{\mathcal{B}}(b,r)$ is the $d-1$-dimensional $C^{1,1}$-sub-manifold of $\mathbb{R}^d$ given implicitly by the equation $G(x) = f(b) - f(x) - \langle \nabla f(x), b - x\rangle - \frac{1}{2}r^2 = 0$. The outer unit normal at a point x on the boundary is therefore $n(x) = \frac{\nabla^2 f(x)(x-b)}{\|\nabla^2 f(x)(x-b)\|}$. Since $\partial\vec{\mathcal{B}}(b,r)$ is compact and $\nabla^2 f \succ 0$, $\nabla^2 f(x)$ is bounded and bounded away from 0 on the boundary. Moreover, $b \notin \partial\vec{\mathcal{B}}(b,r)$, hence the denominator $\|\nabla^2 f(x)(x-b)\|$ stays bounded away from 0. Finally, since by hypothesis $\nabla^2 f$ is locally Lipschitz, the unit normal $n(x)$ is Lipschitz on the compact $\partial\vec{\mathcal{B}}(b,r)$. Assuming $1/r_0$ is a Lipschitz constant, it follows with the extension [58, Thm. 1 (v)] of Blaschke's rolling theorem that a ball of radius r_0 rolls freely inside $\vec{\mathcal{B}}(b,r)$. □

Using (2.3) and $0 < \lambda \le \lambda_i \le \Lambda < \infty$ for the eigenvalues λ_i of $\nabla^2 f(x)$ on $\partial\vec{\mathcal{B}}(b,r)$, we may again show that r_0 is proportional to r in the rough sense of Proposition 3.2. We leave the details to the reader.

3.2. **Bregman geodesics.** Let $b^+ \in \overleftarrow{P}_B(a^+)$ with $a^+ \notin B$, and define $a_\lambda = (\nabla f)^{-1}(\lambda\nabla f(a^+) + (1-\lambda)\nabla f(b^+))$. Then $\overleftarrow{P}_B(a_\lambda) = b^+$ for $0 \le \lambda < 1$; cf. [18, Prop. 3.2]. We call the curve a_λ the left Bregman perpendicular to B at b^+ from $a^+ \in G\backslash B$. The direction $d = \nabla^2 f(b^+)^{-1}(\nabla f(a^+) - \nabla f(b^+))$ is tangent to the curve a_λ at $a_\lambda|_{\lambda=0} = b^+$; see also [18]. In addition, d is normal to the tangent hyperplane $H = \{b : \langle \nabla f(a^+) - \nabla f(b^+), b - b^+\rangle = 0\}$ to B at b^+ in the euclidean geometry $\|x\|^2_{b^+} = \langle x, \nabla^2 f(b^+)x\rangle = \langle x, x\rangle_{b^+}$. The curve a_λ consists of those points which satisfy $\overleftarrow{P}_H(a_\lambda) = b^+$ and are on the same side of H as a^+. Perpendiculars a_λ are also known as left Bregman geodesics. Note that a non-smooth $b^+ \in B$ may be left projected on from different points lying on different geodesics. Those may then be distinguished by their tangents d at b^+.

We next investigate right Bregman geodesics. As we shall see, this may be based on the dual formula (2.2).

Lemma 3.4. *Let $a^+ \in \vec{P}_A(b)$ with $b \notin A$, and define $b_\lambda = \lambda b + (1-\lambda)a^+$. Then $\vec{P}_A(b_\lambda) = a^+$ for $0 \le \lambda < 1$.*

Proof. We put $a^{*+} = \nabla f(b)$, so that $b = \nabla f^*(a^{*+})$, also $b^{*+} = \nabla f(a^+)$ and $B^* = \nabla f(A)$. Then $b^{*+} = \nabla f(a^+) \in \nabla f \circ \vec{P}_A(\nabla f^*(a^{*+})) = \overleftarrow{P}^*_{\nabla f(A)}(a^{*+}) = \overleftarrow{P}^*_{B^*}(a^{*+})$ by (2.2). This means we are in the situation of the left Bregman projection above, with f replaced by f^* and B replaced by B^*. The left Bregman perpendicular to B^* at b^{*+}, being $a^*_\lambda = \nabla f(\lambda\nabla f^*(a^{*+}) + (1-\lambda)\nabla f^*(b^{*+}))$, is defined for $\lambda \in [0,1]$, and for $0 \le \lambda < 1$, $\overleftarrow{P}^*_{B^*}(a^*_\lambda) = b^{*+}$ is single-valued. Reading the dual formula (2.2) backward means $\vec{P}_A(b_\lambda) = a^+$ single valued for $0 \le \lambda < 1$. □

We call b_λ the right Bregman perpendicular to A at $a^+ \in \vec{P}_A(b)$ from $b \notin A$, or right geodesic. Right geodesics are straight lines, while left geodesics are curved. Let $D(b,a^+) = \frac{1}{2}r^2$, then the tangent hyperplane to $\vec{\mathcal{B}}(b,r)$ at $a^+ \in \partial\vec{\mathcal{B}}(b,r)$ is $H = \{z : \langle \nabla^2 f(a^+)(b-a^+), z - a^+\rangle = 0\}$. The right Bregman perpendicular to A at a^+ in direction $d = b - a^+$ includes, and for convex right Bregman balls consists of, those points b_λ which satisfy $\vec{P}_H(b_\lambda) = a^+$ and lie on the same side of H as b. The geodesic is also the normal to H in the euclidean geometry induced by the scalar product $\langle x, \nabla^2 f(a^+)y\rangle$.

Remark 3.5. In the case of the Kullback-Leibler divergence, left and right perpendiculars are known as m- and e-geodesics; cf. [2, 3].

3.3. **Normal cones.** We recall the following properties of the left and right Bregman projectors.

Lemma 3.6. *Let $a \xrightarrow{l} b \xrightarrow{r} a^+ \xrightarrow{l} b^+$ be a building block. Then*

(i) $\nabla f(a^+) - \nabla f(b^+) \in N^p_B(b^+)$.
(ii) $\nabla^2 f(a^+)(b-a^+) \in \widehat{N}_A(a^+)$.
(iii) *When f is of class $C^{2,1}$, then $\nabla^2 f(a^+)(b-a^+) \in N^p_A(a^+)$.*

Proof. For (i) see [18, Prop. 3.3]. This can also be derived directly from Lemma 3.1. Place a euclidean ball $B(z, \underline{c}r)$ such that it is contained in $\overleftarrow{\mathcal{B}}(a^+, r)$ and touches at b^+ from within, where $D(b^+, a^+) = \frac{1}{2}r^2$. Then $b^+ \in P_B(z)$ with $z \in b^+ + \mathbb{R}_+(\nabla f(a^+) - \nabla f(b^+))$.

Concerning (ii), from $a^+ \in \vec{P}_A(b)$ we derive $a^+ \in \partial\vec{\mathcal{B}}(b,r)$ with $\frac{1}{2}r^2 = D(b,a^+)$. The tangent hyperplane to the regular surface $\partial\vec{\mathcal{B}}(b,r)$ at a^+ has normal $n(a^+) = \nabla^2 f(a^+)(b-a^+)$. This means $n(a^+)$ is a Fréchet normal to the sublevel set $S = \{x : -D(b,x) \le -\frac{1}{2}r^2\}$ at a^+, and since $a^+ \in A \subset S$, $n(a^+)$ is also a Fréchet normal to A at a^+.

Part (ii) may also be derived via duality. The building block $a \xrightarrow{l} b \xrightarrow{r} a^+$ is the image under ∇f^* of a building block $b^* \xrightarrow{r*} a^{*+} \xrightarrow{l*} b^{*+}$. For the latter $\nabla f^*(a^{*+}) - \nabla f^*(b^{*+}) \in N^p_{B^*}(b^{*+})$ by part (i). Using $N^p_{B^*}(b^{*+}) \subset \widehat{N}_{B^*}(b^{*+})$ and the chain rule [54, Thm. 10.6, Ex. 6.7] gives $\nabla^2 f(a^+)(b-a^+) \in \widehat{N}_A(a^+)$.

Finally, concerning (iii), when f is class $C^{2,1}$, we may by Lemma 3.3 place a euclidean ball $B(z, r_0)$ such that it has a^+ on its boundary and is contained in $\vec{\mathcal{B}}(b, r)$, where $D(b, a^+) = \frac{1}{2}r^2$. Since $B(z, r_0)$ and $\vec{\mathcal{B}}(b, r)$ share the tangent hyperplane $H = \{x : \langle \nabla^2 f(a^+)(b - a^+), x - a^+\rangle = 0\}$ at a^+, the center z must lie on the normal $a^+ + \mathbb{R}_+ d$, $d = \nabla^2 f(a^+)(b - a^+)$, which is therefore proximal. □ □

3.4. **Bregman reach.** A point $b^+ \in B$ is *projected on* if there exists $c \notin B$ with $b^+ \in P_B(c)$. Then $d := c - b^+ \in N_B^p(b^+) \neq \{0\}$, and the reach $R(b^+, d)$ of B at b^+ in direction d is the radius r of the largest ball $B(c, r)$ with center $c = b^+ + rd/\|d\|$ having no point of B in its interior (cf. [36]). It is possible that $R(b^+, d) = \infty$, when the largest ball becomes a half-space.

We extend this classical notion of reach [36] to the Bregman setting. The results in Section 3.1 show that if b^+ is projected on, then it is also *left Bregman projected on*, i.e., $b^+ \in \overleftarrow{P}_B(a^+)$ for some $a^+ \notin B$.

Definition 3.7. Let $b^+ \in \overleftarrow{P}_B(a^+)$ be left projected on from some $a^+ \in G \setminus B$, and let a_λ be the associated left Bregman geodesic. The left Bregman reach $\overleftarrow{R}(b^+, d)$ of B at $b^+ \in B$ in direction $d = \nabla^2 f(b^+)^{-1}(\nabla f(a^+) - \nabla f(b^+))$ is the largest left Bregman radius r_λ for which the Bregman ball $\overleftarrow{\mathcal{B}}(a_\lambda, r_\lambda)$ with $b^+ \in \partial\overleftarrow{\mathcal{B}}(b_\lambda, r_\lambda)$ has no points of B in its interior.

Remark 3.8. As in the euclidean case the Bregman reach may be infinite, which is when $\bigcup_{\lambda>0} \overleftarrow{\mathcal{B}}(a_\lambda, r_\lambda)$ contains no point of B except b^+. The definition reproduces the classical definition of reach in the euclidean case.

Definition 3.9. (Left Bregman reach). The set B has left Bregman reach at least $r > 0$ at $b^* \in B$, written $\overleftarrow{R}(b^*) \geq r$, if there exists a neighborhood U of b^* such that for every point $b^+ \in \overleftarrow{P}_B(a^+) \cap U$ left projected on from some $a^+ \in G \setminus B$, we have $\overleftarrow{R}(b^+, d) \geq r$ for the associated $d = \nabla^2 f(b^+)^{-1}(\nabla f(a^+) - \nabla f(b^+))$.

Definition 3.10. (Right Bregman reach). The right Bregman reach $\vec{R}(a^+, d)$ of A at $a^+ \in A$ in direction $d = b - a^+ \neq 0$ with $a^+ \in \vec{P}_A(b)$, is the radius r_λ of the largest right Bregman ball $\vec{\mathcal{B}}(b_\lambda, r_\lambda)$ with $a^+ \in \partial\vec{\mathcal{B}}(b_\lambda, r_\lambda)$ having no points of A in its interior. A has right Bregman reach at least $r > 0$ at $a^* \in A$, written $\vec{R}(a^*) \geq r$, if $\vec{R}(a^+, d) \geq r$ on a neighborhood U of a^*.

Remark 3.11. Suppose $D(b, a_\lambda) = D(b^+, a_\lambda) = \frac{1}{2}r_\lambda^2$, then $D^*(b_\lambda^*, a^*) = D^*(b_\lambda^*, a^{*+})$ by duality, as can be seen from $D^*(\nabla f(a_\lambda), \nabla f(b)) = D^*(\nabla f(a_\lambda), \nabla f(b^+)) = \frac{1}{2}r_\lambda^2$. This shows $\overleftarrow{R}(b^+, d) = \vec{R}(a^{*+}, d_*)$ with $d = \nabla^2 f(b^+)^{-1}(\nabla f(a^+) - \nabla f(b^+))$ and $d_* = b^* - a^{*+}$. In other words, Bregman reach is amenable to duality.

Remark 3.12. As a consequence of the results in Section 3.1, we now see that if $K \subset G$ is compact, then any set $B \subset K$ with left Bregman reach $r > 0$ has classical reach at least $\underline{c}r > 0$ and at most $\overline{c}r$, where $\underline{c}, \overline{c}$ depends only on K, and inversely, if B has classical reach r, then it has left Bregman reach at most $\underline{c}^{-1}r$, and at least $\overline{c}^{-1}r$. Right Bregman reach and classical reach are also proportional in this rough sense when f is class $C^{2,1}$.

3.5. **1-coercivity of f.** Consider a left perpendicular a_λ at $b^+ \in \overleftarrow{P}_B(a^+)$ from $a^+ \notin B$. Let $H = \{x : \langle \nabla f(b^+) - \nabla f(a^+), x - b^+\rangle = 0\}$ be the tangent hyperplane to $\partial\overleftarrow{\mathcal{B}}(a^+, r)$ at b^+, where $D(b^+, a^+) = \frac{1}{2}r^2$, H_+ the half space containing a^+, H_{++} the open half space. Suppose a_λ is defined for $0 \leq \lambda < \lambda_\infty$, where $\lambda_\infty \in (1, \infty]$, $D(b^+, a_\lambda) = \frac{1}{2}r_\lambda^2$. We ask whether the balls $\overleftarrow{\mathcal{B}}(a_\lambda, r_\lambda)$, $0 \leq \lambda < \lambda_\infty$, fill the open half space $H_{++} \cap \mathrm{int}(\mathrm{dom} f)$.

Proposition 3.13. *The following are equivalent:*

(i) *For some nonempty compact $B \subset G = \mathrm{int}(\mathrm{dom} f)$, and every $a^+ \notin B$, $b^+ \in \overleftarrow{P}_B(a^+)$ with perpendicular a_λ, the balls $\overleftarrow{\mathcal{B}}(a_\lambda, r_\lambda)$, $0 \leq \lambda < \lambda_\infty$ fill the half space $H_{++} \cap \mathrm{int}(\mathrm{dom} f)$.*

(ii) *f is 1-coercive.*

When these are satisfied, then (i) *holds for* every *such B, and we have $\lambda_\infty = \infty$ in* every *perpendicular curve a_λ.*

Proof. 1) Assume first that f is 1-coercive. Then by [8, Prop. 2.16] f^* is defined everywhere, hence so is ∇f^*, hence a_λ is defined for all $\lambda \geq 0$. Now let $b \in H_{++} \cap \mathrm{int}(\mathrm{dom} f)$. That means $\langle \nabla f(a^+) - \nabla f(b^+), b - b^+ \rangle > 0$. By the cosine theorem for Bregman distances and the definition of a_λ we have

$$D(b, a_\lambda) = D(b, b^+) + D(b^+, a_\lambda) - \lambda \langle \nabla f(a^+) - \nabla f(b^+), b - b^+ \rangle$$

hence

$$\begin{aligned} \frac{D(b, a_\lambda) - D(b^+, a_\lambda)}{\lambda} &= \frac{D(b, b^+)}{\lambda} - \langle \nabla f(a^+) - \nabla f(b^+), b - b^+ \rangle \\ &\to -\langle \nabla f(a^+) - \nabla f(b^+), b - b^+ \rangle < 0 \end{aligned}$$

as $\lambda \to \infty$, so that $D(b, a_\lambda) < D(b^+, a_\lambda)$ for λ large enough, which gives $b \in \tilde{\mathcal{B}}(a_\lambda, r_\lambda)$.

2) Conversely, suppose $H_{++} \cap \mathrm{int}(\mathrm{dom} f) \subset \bigcup \{\tilde{\mathcal{B}}(a_\lambda, r_\lambda) : 0 \leq \lambda < \lambda_\infty\}$. The normal curvature of $\partial \tilde{\mathcal{B}}(a_\lambda, r_\lambda)$ at $b^+ \in B$ in unit tangential direction v being

$$\kappa_n(\lambda) = \frac{\langle v, \nabla^2 f(b^+) v \rangle}{\|\nabla f(b^+) - \nabla f(a_\lambda)\|},$$

we see that $\kappa_n(\lambda)$ must tend to 0 as $\lambda \to \lambda_\infty$. Indeed, by convexity left Bregman balls touching H at b^+ have to get arbitrarily flat at b^+ as λ increases in order to contain points $b \in H_{++}$ a fixed distance away from b^+, while arbitrarily close to H. But that means the denominator $\|\nabla f(b^+) - \nabla f(a_\lambda)\|$ must tend to infinity, since the numerator is bounded below by $\lambda_{\min}(\nabla^2 f(b^+)) > 0$. That forces $\|\nabla f(a_\lambda)\| \to \infty$ as $\lambda \to \lambda_\infty$ for every perpendicular a_λ to B at any $b^+ \in B$ left projected on from some $a^+ \notin B$.

Now suppose that contrary to what is claimed 1-coercivity fails. By [53, Lemma 26.7] this means there exist a_k with $\|a_k\| \to \infty$ such that $\|\nabla f(a_k)\| \leq K < \infty$. Using compactness of $B \subset G$, find $b_k \in \tilde{P}_B(a_k)$, and let $a_{k,\lambda}$ be the left perpendicular to B at b_k from a_k. Then we can find points $\bar{a}_k$ on the perpendicular, having $b_k = \tilde{P}_B(\bar{a}_k)$, such that $\|\bar{a}_k\| \leq K' < \infty$, and at the same time $\|\bar{a}_k - b_k\| \geq \epsilon > 0$ for all k (the latter using $\|a_k - b_k\| \to \infty$, allowing to choose $\bar{a}_k$ in between bounded while away from b_k). Passing to subsequences, we may assume $b_k \to b^*$, $\bar{a}_k \to a^*$ with $b^* \in \tilde{P}_B(a^*)$ and $\|a^* - b^*\| \geq \epsilon$, while $\nabla f(a_k) \to v \notin \mathrm{dom} \nabla f$.

From $\nabla f(a_{k,\lambda}) = \nabla f(b_k) + \lambda(\nabla f(a_k) - \nabla f(b_k))$ and boundedness of the $\nabla f(a_k)$ follows $\|\nabla f(a_{k,\lambda})\| \propto \lambda$ uniformly over k. Since $\|\nabla f(a_{k,\lambda})\| \to \infty$ for fixed k, $a_{k,\lambda}$ must be defined for all $\lambda > 0$. Now parametrize the same perpendicular curve as $\nabla f(\bar{a}_{k,\mu}) = \nabla f(b_k) + \mu(\nabla f(\bar{a}_k) - \nabla f(b_k))$, then again $\nabla f(\bar{a}_{k,\mu}) \propto \mu$ uniformly over k. With the same argument as above, $\bar{a}_{k,\mu}$ must be defined for all $\mu > 0$.

Now for every k and $\lambda > 0$ there exists $\mu_k(\lambda)$ such that $a_{k,\lambda} = \bar{a}_{k,\mu_k(\lambda)}$. But the relation between the two is given by

$$\lambda(\nabla f(a_k) - \nabla f(b_k)) = \mu(\nabla f(\bar{a}_k) - \nabla f(b_k)), \tag{3.3}$$

hence by boundedness of the $\nabla f(a_k)$ we have $r\lambda \leq \mu_k(\lambda) \leq r'\lambda$ for all k with certain r, r' not depending on k.

Since $\bar{a}_k$ is between a_k and b_k, we have $\mu > 1$ when $\lambda = 1$. We argue that this implies $\mu_k(\lambda) > \lambda$ for almost all k. Indeed, if for some k there is a moment, where $\mu_k(\lambda) = \lambda$, then $\nabla f(a_k) = \nabla f(\bar{a}_k)$, and that could happen only a finite number of times, because $\nabla f(\bar{a}_k) \to \nabla f(a^*)$, while $\nabla f(a_k) \to v \notin \mathrm{dom} \nabla f$. Hence we can assume $\mu_k(\lambda) > \lambda$ for all k. From (3.3) we now get

$$\left(1 - \frac{\lambda}{\mu_k(\lambda)}\right) \nabla f(b_k) + \frac{\lambda}{\mu_k(\lambda)} \nabla f(a_k) = \nabla f(\bar{a}_k),$$

a convex combination, and passing to the limit ($k \to \infty$) in a subsequence $\mu_k(\lambda) \to \mu^* \geq \lambda$,

$$\left(1 - \frac{\lambda}{\mu^*}\right) \nabla f(b^*) + \frac{\lambda}{\mu^*} v = \nabla f(a^*),$$

proving $\nabla f(a^*) \in [\nabla f(b^*), v]$. But the $\bar{a}_k$ are independent of v, so we can arrange the same estimate for other points a^* on the limit perpendicular a^*_λ generated by $b^* \in \tilde{P}_B(a^*)$. But that means $\nabla f(a^*_\lambda) \in [\nabla f(b^*), v]$ for the entire perpendicular, contradicting $\|\nabla f(a^*_\lambda)\| \to \infty$. □

4. Gaps between sets

Let a_k, b_k be a Bregman alternating sequence. Since $D(b_k, a_k) \leq D(b_{k-1}, a_k) \leq D(b_{k-1}, a_{k-1})$, monotone convergence gives $D(b_k, a_k) \to \frac{1}{2}r^{*2}$, and $D(b_{k-1}, a_k) \to \frac{1}{2}r^{*2}$ for some $r^* \geq 0$. Now let $k \in N$ be an infinite subsequence of $\mathbb{N}$ such that $b_{k-1} \to b^*$, $a_k \to a^*$, $b_k \to \hat{b}$. As $a_k \in \vec{P}_A(b_{k-1})$, we have $a^* \in \vec{P}_A(b^*)$. Similarly, as $b_k \in \tilde{P}_B(a_k)$, we have $\hat{b} \in \tilde{P}_B(a^*)$. But $D(b^*, a^*) = D(\hat{b}, a^*) = \frac{1}{2}r^{*2}$, hence $b^* \in \tilde{P}_B(a^*)$, too. So we have found a pair $(b^*, a^*) \in B \times A$ with $a^* \in \vec{P}_A(b^*)$, $b^* \in \tilde{P}_B(a^*)$. We write $b^* \sim a^*$ for such pairs.

Let A^*, B^* be the sets of accumulation points of the a_k, b_k. The above argument shows that for every $a^* \in A^*$ there exists $b^* \in B^*$ such that $b^* \sim a^*$, and for every $b^* \in B^*$ there exists $a^* \in A^*$ with $b^* \sim a^*$. We call $K^* = \{(b^*, a^*) \in B^* \times A^* : b^* \sim a^*\}$ *the gap* of the alternating sequence. The case $r^* = 0$ is not excluded, where of course $b^* \sim a^*$ implies $b^* = a^* \in A \cap B$.

Abstracting from the sequence a_k, b_k, we call a pair (K^*, r^*) a *gap* between A and B if $K^* \subset B \times A$ is compact with $b^* \sim a^*$ for all $(b^*, a^*) \in K^*$ and $D(b^*, a^*) = \frac{1}{2}r^{*2}$. Note that every $x^* \in A \cap B$ gives rise to a zero gap $(\{(x^*, x^*)\}, 0)$.

Lemma 4.1. *Let $F(x, y) = i_B(x) + D(x, y) + i_A(y)$, and suppose $(b^*, a^*) \in K^*$, then $(0, 0) \in \partial F(b^*, a^*)$.*

Proof. Consider a building block $b \xrightarrow{r} a^+ \xrightarrow{l} b^+$, then $n_B(b^+) = \nabla f(a^+) - \nabla f(b^+) \in N_B(b^+)$ and $n_A(a^+) = \nabla^2 f(a^+)(b - a^+) \in N_A(a^+)$ by Lemma 3.6, hence

$$(\lambda n_B(b^+) + \nabla f(b^+) - \nabla f(a^+), \mu n_A(a^+) + \nabla^2 f(a^+)(a^+ - b^+)) \in \partial F(b^+, a^+) \tag{4.1}$$

for all $\lambda, \mu \geq 0$. Since $b^* \sim a^*$, we may choose $b = b^+ = b^*$ and $a^+ = a^*$ in the building block, which gives $(0, 0) \in \partial F(b^*, a^*)$ on putting $\lambda = \mu = 1$. □

Proposition 4.2. *Suppose A, B, f are definable. Then there is only a finite number of gap values $0 \leq r_1 < r_2 < \cdots < r_N$. There exist $\eta_i > 0$, $i = 1, \ldots, N$, such that every alternating sequence a_k, b_k which satisfies $\frac{1}{2}r_i^2 \leq D(b_k, a_k) < \frac{1}{2}r_i^2 + \eta_i$ must have value convergence $D(b_k, a_k) \to \frac{1}{2}r_i^2$, $D(b_{k-1}, a_k) \to \frac{1}{2}r_i^2$.*

Proof. This follows with [43, Prop. 2]. □

5. Angle condition

In this section we introduce the angle condition and show that it is a geometric form of the Kurdyka-Łojasiewicz inequality.

Definition 5.1. (Angle condition). Let $\sigma : (0, \infty) \to (0, \infty)$ be increasing. The set B satisfies the *rl*-angle condition with constant γ and shrinking function σ with respect to A at a gap pair $b^* \sim a^*$ with gap value r^*, if there exists a neighborhood W of (b^*, a^*) and $\eta > 0$ such that

$$\frac{1 - \cos\alpha}{\sigma(\tilde{D}_B(a^+) - \frac{1}{2}r^{*2})} \geq \gamma \tag{5.1}$$

for every building block $b \xrightarrow{r} a^+ \xrightarrow{l} b^+$ with $(b^+, a^+) \in W$, $\frac{1}{2}r^{*2} \leq D(b^+, a^+) < \frac{1}{2}r^{*2} + \eta$, where $\alpha = \sphericalangle(b - a^+, b^+ - a^+)$.

Remark 5.2. A standard compactness argument shows that if (K^*, r^*) is a gap such that the angle condition holds at every $(b^*, a^*) \in K^*$, then it holds for all building blocks in a neighborhood of K^* with the same σ, η, γ.

We now show that the angle condition can be understood as a geometric form of the KŁ-inequality.

Proposition 5.3. *Let (K^*, r^*) be a gap between A and B. Suppose $F(x, y) = i_B(x) + D(x, y) + i_A(y)$ satisfies the KŁ-condition at every $(b^*, a^*) \in K^*$. Then the angle condition is satisfied in the following form. There exists a neighborhood W of K^*, a de-singularizing function ϕ, and constants $\gamma, \eta > 0$ such that every building block $b \xrightarrow{r} a^+ \xrightarrow{l} b^+$ with $(b^+, a^+) \in W$ and $\frac{1}{2}r^{*2} \leq D(b^+, a^+) < \frac{1}{2}r^{*2} + \eta$ satisfies*

$$\phi'(\tilde{D}_B(a^+) - \tfrac{1}{2}r^{*2})^2 \tilde{D}_B(a^+)(1 - \cos\alpha) \geq \gamma, \tag{5.2}$$

where $\alpha = \sphericalangle(b - a^+, b^+ - a^+)$.

Proof. K^* being compact, and F having constant value $\frac{1}{2}r^{*2}$ on K^*, the KŁ-inequality is satisfied as follows: There exists a bounded neighborhood W of K^*, a de-singularizing function ϕ, and constants $\gamma, \eta > 0$ such that

$$\phi'(F(b^+, a^+) - \tfrac{1}{2}r^{*2})\mathrm{dist}_{|\cdot|}((0,0), \partial F(b^+, a^+)) \geq \gamma$$

for all $(b^+, a^+) \in W$ with $\frac{1}{2}r^{*2} \leq F(b^+, a^+) < \frac{1}{2}r^{*2} + \eta$, and where $|(x, y)| = \|x\| + \|y\|$.

Now consider building blocks $b \xrightarrow{r} a^+ \xrightarrow{l} b^+$ with $(b^+, a^+) \in W$ and $\frac{1}{2}r^{*2} \leq D(b^+, a^+) < \frac{1}{2}r^{*2}+\eta$. Since

$$(\lambda n_B(b^+) + \nabla f(b^+) - \nabla f(a^+), \mu n_A(a^+) + \nabla^2 f(a^+)(a^+ - b^+)) \in \partial F(b^+, a^+)$$

for all $n_B(b^+) \in N_B(b^+)$ and $n_A(a^+) \in N_A(a^+)$, Lemma 3.6 gives

$$\phi'(F(b^+, a^+) - \tfrac{1}{2}r^{*2})\,(\|\lambda(\nabla f(a^+) - \nabla f(b^+)) + \nabla f(b^+) - \nabla f(a^+))\| + \\ \|\mu\nabla^2 f(a^+)(b - a^+) + \nabla^2 f(a^+)(a^+ - b^+)\|) \geq \gamma$$

for all $\lambda, \mu \geq 0$. Choosing $\lambda = 1$, gives

$$\phi'(F(b^+, a^+) - \tfrac{1}{2}r^{*2})\|\mu\nabla^2 f(a^+)(b - a^+) + \nabla^2 f(a^+)(a^+ - b^+)\| \geq \gamma$$

for all $\mu \geq 0$. Now due to boundedness of W the eigenvalues of all Hessians $\nabla^2 f(a^+)$ with $(b^+, a^+) \in W$ are bounded above by a constants Λ. Therefore $\|\mu\nabla^2 f(a^+)(b - a^+) + \nabla^2 f(a^+)(a^+ - b^+)\| \leq \Lambda\|\mu(b - a^+) + a^+ - b^+\|$, so that

$$\phi'(F(b^+, a^+) - \tfrac{1}{2}r^{*2})\|\mu(b - a^+) + a^+ - b^+\| \geq \Lambda^{-1}\gamma$$

for all $\mu \geq 0$. Passing to the infimum over $\mu \geq 0$ implies

$$\phi'(F(b^+, a^+) - \tfrac{1}{2}r^{*2})\sin\alpha\|a^+ - b^+\| \geq \Lambda^{-1}\gamma,$$

for angles $\alpha = \sphericalangle(b - a^+, b^+ - a^+)$ less than $90°$, while for angles larger than $90°$ the minimum is attained at $\phi'(F(b^+, a^+) - \frac{1}{2}r^{*2})\|a^+ - b^+\|$. As the statement is clear in the latter case, we continue with $\alpha < 90°$. Here using $\|a^+ - b^+\| \leq MD(b^+, a^+)^{1/2}$, taking squares gives

$$\phi'(\tilde{D}_B(a^+) - \tfrac{1}{2}r^{*2})^2\tilde{D}_B(a^+)\sin^2\alpha \geq \gamma^2/M^2\Lambda^2$$

and then (using $1 - \cos\alpha \geq \frac{1}{2}\sin^2\alpha$):

$$\phi'(\tilde{D}_B(a^+) - \tfrac{1}{2}r^{*2})^2\tilde{D}_B(a^+)(1 - \cos\alpha) \geq \gamma^2/2M^2\Lambda^2 =: \gamma',$$

where γ' depends only on K^* and the bounds (2.3) associated with it. This proves the claim. □

Remark 5.4. Clearly (5.2) gives the *rl*-angle condition (5.1) with angle shrinking function $\sigma(s) = \phi'(s)^{-2}s^{-1}$ when $r^* = 0$. When $r^* > 0$, then the term $\tilde{D}_B(a^+)$ stays bounded away from 0 and hence does not contribute to the singularity. We can then shuffle it to the right, obtaining yet another γ'' depending only on K^* and r^*, now with $\sigma(s) = \phi'(s)^{-2}$. During the following we will always use the angle condition with angle shrinking functions generated by de-singularizing functions ϕ.

Corollary 5.5. *Let A, B, f be definable. Then there exists a unique de-singularizing function ϕ which works for each of the finitely many gap values. The angle shrinking function is $\sigma(s) = \phi'(s)^{-2}$ for gaps $r_i > 0$, and $\sigma(s) = \phi'(s)^{-2}s^{-1}$ for r_1 in case $r_1 = 0$.*

5.1. **Tangential and transversal intersection.** It is clear that the mutual geometric position of A, B near a point $\bar{x} \in A \cap B$ must be decisive for the speed of convergence of alternating Bregman projections near $\bar{x}$. We distinguish tangential and transversal intersection, expecting transversality to give linear speed, while tangential intersection should force a slowdown to sub-linear speed. We now give an interpretation of these using the angle condition.

Definition 5.6. (Transversality). We say that B intersects A *rl-transversally* at $\bar{x} \in A \cap B$, if there exists a neighborhood U of $\bar{x}$ and $\underline{\alpha} > 0$ such that for every building block $b \xrightarrow{r} a^+ \xrightarrow{l} b^+$ in U the angle $\alpha = \sphericalangle(b - a^+, b^+ - a^+)$ is larger than $\underline{\alpha}$. Otherwise we say that B intersects A *rl-tangentially* at $\bar{x}$.

This notion is weaker than classical transversality, and yet guarantees linear convergence of the alternating method, as will indeed be proved in Corollary 9.4. One classical notion of transversality in non-smooth calculus is the following:

Proposition 5.7. *Suppose $N_B(\bar{x}) \cap (-N_A(\bar{x})) = \{0\}$. Then B intersects A rl-transversally at $\bar{x} \in A \cap B$.*

Proof. Assume on the contrary that there exist building blocks $b_{k-1} \xrightarrow{r} a_k \xrightarrow{l} b_k$ with $b_{k-1}, a_k, b_k \to \bar{x}$ as $k \to \infty$ such that $\alpha_k = \sphericalangle(b_{k-1} - a_k, b_k - a_k) \to 0$. Let $u_k = (b_{k-1} - a_k)/\|b_{k-1} - a_k\|$, $v_k = (b_k - a_k)/\|b_k - a_k\|$, and $w_k = (\nabla f(a_k) - \nabla f(b_k))/\|a_k - b_k\|$.

Recall $n_A(a_k) = \nabla^2 f(a_k)(b_{k-1} - a_k) \in N_A(a_k)$ and $n_B(b_k) = \nabla f(a_k) - \nabla f(b_k) \in N_B(b_k)$ by Lemma 3.6. Hence $\nabla^2 f(a_k)u_k \in N_A(a_k)$ and $w_k \in N_B(b_k)$. Select an infinite subsequence $k \in K$ of $\mathbb{N}$ such that $u_k \to u \neq 0$, $v_k \to v \neq 0$, $w_k \to w \neq 0$, the latter due to (2.4). Then $\nabla^2 f(a_k)u_k \to \nabla^2 f(\bar{x})u \in N_A(\bar{x})$, and $\nabla^2 f(a_k)v_k \to \nabla^2 f(\bar{x})v$, $w_k \to w \in N_B(\bar{x})$. But $\sphericalangle(u_k, v_k) \to 0$ implies $\sphericalangle(u, v) = 0$, and since u, v are unit vectors, we have $u = v$. Since $\nabla^2 f(\bar{x})u \in N_A(\bar{x})$, this gives $\nabla^2 f(\bar{x})v \in N_A(\bar{x})$.

Taylor-Young expansion of ∇f at the a_k gives $\nabla f(a_k) - \nabla f(b_k) = \nabla^2 f(a_k)(a_k - b_k) + o(\|a_k - b_k\|)$, where the o-term may be made uniform on a bounded neighborhood of $\bar{x}$, see Section 2.5. Therefore

$$w_k = \frac{\nabla f(a_k) - \nabla f(b_k)}{\|a_k - b_k\|} = \nabla^2 f(a_k)(-v_k) + o(1) \to \nabla^2 f(\bar{x})(-v) \in -N_A(\bar{x}).$$

Since $w_k \to w \in N_B(\bar{x})$, we have $w \in N_B(\bar{x}) \cap (-N_A(\bar{x}))$, a contradiction. □

Remark 5.8. In [12, 13] the authors propose a refinement in the euclidean case, where $N_B(\bar{x})$ is replaced by a restricted normal cone $N_B^A(\bar{x})$, which only considers projections on B stemming from A, and similarly for $N_A^B(\bar{x})$. This could be extended to left and right Bregman projections. We do not pursue this natural idea further. The proof in [52, Prop. 1] may be adapted to the Bregman setting.

When B intersects A rl-tangentially at $\bar{x}$, then there are rl-building blocks for which the angle $\alpha = \sphericalangle(b - a^+, b^+ - a^+)$ shrinks to 0 as these $b \xrightarrow{r} a^+ \xrightarrow{l} b^+$ approach $\bar{x}$. Following [52, 51], we now relate this to the angle condition in the zero gap case $r^* = 0$. Note that for $r^* = 0$ condition (5.1) means

$$\frac{1 - \cos\alpha}{\sigma(\tilde{D}_B(a^+))} \geq \gamma \tag{5.3}$$

for every building block $b \xrightarrow{r} a^+ \xrightarrow{l} b^+$ in U, where $\alpha = \sphericalangle(b - a^+, b^+ - a^+)$. Therefore when the angle α tends to 0, meaning $1 - \cos\alpha \to 0$, then the speed with which it is allowed to do so is controlled by the angle shrinking function σ, which when obtained from the KŁ-inequality is $\sigma(s) = \phi'(s)^{-2}s^{-1}$ for a de-singularizing function ϕ. This was introduced in [52, 51] for the euclidean case with shrinking functions $\sigma(s) = s^\theta$ for $\theta \in [\frac{1}{2}, 1)$.

When (5.3) holds but $\sigma(\tilde{D}_B(a^+))$ does *not* shrink to 0, then α must also stay away from 0, which is precisely when B intersects A rl-transversally at $\bar{x}$. Hence this case is covered by the angle condition.

Definition 5.9. (Dual transversality). We say that A intersects B *lr-transversally* at $\bar{x} \in A \cap B$ if there exists a neighborhood U of $\bar{x}$ and $\underline{\alpha} > 0$ such that for every building block $a \xrightarrow{l} b \xrightarrow{r} a^+$ in U the angle $\alpha = \sphericalangle(\nabla f(a) - \nabla f(b), \nabla f(a^+) - \nabla f(b))$ is larger than $\underline{\alpha}$. Otherwise A intersects B lr-tangentially at $\bar{x}$.

As a first application of duality we have

Proposition 5.10. *Suppose $N_B(\bar{x}) \cap (-N_A(\bar{x})) = \{0\}$. Then A intersects B lr-transversally at $\bar{x} \in A \cap B$.*

Proof. Let $A^* = \nabla f(B)$, $B^* = \nabla f(A)$, then $\nabla^2 f(\bar{x})N_{B^*}(\nabla f(\bar{x})) = N_A(\bar{x})$ by the chain rule [54, Thm. 10.6, Ex. 6.7], applied to the C^1-diffeomorphism ∇f. Similarly $\nabla^2 f(\bar{x})N_{A^*}(\nabla f(\bar{x})) = N_B(\bar{x})$. Since $\nabla^2 f(\bar{x}) \succ 0$, the hypothesis implies $N_{A^*}(\nabla f(\bar{x})) \cap (-N_{B^*}(\nabla f(\bar{x}))) = \{0\}$. Therefore by the previous proposition B^* intersects A^* rl-transversally at $\nabla f(\bar{x}) \in A^* \cap B^*$. For building blocks $b^* \xrightarrow{r*} a^{*+} \xrightarrow{l} b^{*+}$ close to $\nabla f(\bar{x})$ this means $\sphericalangle(b^* - a^{*+}, b^{*+} - a^{*+}) \geq \underline{\alpha} > 0$. But rl-building blocks $a \xrightarrow{l} b \xrightarrow{r} a^+$ in the neighborhood of $\bar{x}$ are mapped by ∇f to lr-building block $b^* \xrightarrow{r*} a^{*+} \xrightarrow{l} b^{*+}$ in the neighborhood of $\nabla f(\bar{x})$ and this gives directly $\alpha = \sphericalangle(\nabla f(a) - \nabla f(b), \nabla f(a^+) - \nabla f(b)) \geq \underline{\alpha}$. □

This calls for a dual angle condition, which we will obtain shortly.

5.2. **Duality for the KŁ-inequality.** We show that the KŁ-inequality is directly amenable to duality.

Lemma 5.11. *Let $F(x,y) = i_B(x) + D(x,y) + i_A(y)$ and put $F_*(u,v) = i_{B^*}(u) + D^*(u,v) + i_{A^*}(v)$, where $A^* = \nabla f(B)$, $B^* = \nabla f(A)$. Suppose F satisfies a KŁ-inequality at $(\bar{x},\bar{y})$, then F_* satisfies a KŁ-inequality at $(\bar{u},\bar{v}) = (\nabla f(\bar{y}), \nabla f(\bar{x}))$ with the same de-singularizing function.*

Proof. We have $F(x,y) = i_{B\times A}(x,y) + D(x,y)$, hence $\partial F(x,y) = N_{B\times A}(x,y) + \partial D(x,y) = N_B(x) \times N_A(y) + \partial D(x,y)$, D being jointly differentiable, and using [54, Prop. 6.41]. Now $D(x,y) = D^*(\nabla f(y), \nabla f(x))$, and $i_A(y) = i_{\nabla f(A)}(\nabla f(y)) = i_{B^*}(\nabla f(y))$, $i_B(x) = i_{A^*}(\nabla f(x))$. Hence

$$F(x,y) = F_*(\Phi(x,y))$$

under the isomorphism $\Phi(x,y) = (\nabla f(y), \nabla f(x))$, where $F_*(u,v) = i_{B^*\times A^*}(u,v) + D^*(u,v)$. Then $\partial F(x,y) = \Phi'(x,y)^T \partial F_*(\Phi(x,y))$ by the chain rule [54, Thm. 10.6]. Suppose $(0,0) \in \partial F(\bar{x},\bar{y})$, then as F satisfies the KŁ-condition at $(\bar{x},\bar{y})$, we have

$$\phi'(F(x,y) - F(\bar{x},\bar{y}))\mathrm{dist}((0,0), \partial F(x,y)) \geq \gamma$$

for some $\gamma > 0$, some $\eta > 0$, a neighborhood W of $(\bar{x},\bar{y})$, and every $(x,y) \in W$ satisfying $F(\bar{x},\bar{y}) \leq F(x,y) \leq F(\bar{x},\bar{y}) + \eta$. Now let $u = \nabla f(y)$, $v = \nabla f(x)$, $\bar{u} = \nabla f(\bar{y})$, $\bar{v} = \nabla f(\bar{x})$. Then $(0,0) \in \partial F_*(\nabla f(\bar{y}), \nabla f(\bar{x}))$, because $\Phi'(\bar{x},\bar{y})^T$ is a linear isomorphism. Let $V = \Phi(W)$, then V is a neighborhood of $(\bar{u},\bar{v})$. Let $(u,v) \in V$, $(u,v) = \Phi(x,y)$ with $(x,y) \in W$. Let $h \in \partial F_*(u,v)$, then $h = \Phi'(x,y)^{-T} g$ for $g \in \partial F(x,y)$. But then

$$\begin{aligned}\phi'(F_*(u,v) - F_*(\bar{u},\bar{v}))\|h\| &= \phi'(F(x,y) - F(\bar{x},\bar{y}))\|\Phi'(x,y)^{-T} g\| \\ &\geq \phi'(F(x,y) - F(\bar{x},\bar{y}))k\|g\| \geq k\gamma,\end{aligned}$$

where we use $\|\Phi'(x,y)^T\| \leq k^{-1}$ for all $(x,y) \in W$. Here $\Phi'(x,y) = \mathrm{diag}(\nabla^2 f(y), \nabla^2 f(x)) = \Phi'(x,y)^T$, and we can without loss choose $W = W_1 \times W_2$ such that $\nabla^2 f(y)$ is bounded away from 0 on the neighborhood W_2 of $\bar{y}$, $\nabla^2 f(x)$ bounded away from 0 on the neighborhood W_1 of $\bar{x}$. This means, the KŁ-inequality for F_* is satisfied with $\gamma' = k\gamma$ and the same ϕ and η. □

5.3. **Duality for the angle condition.** We now show that duality leads the way to the correct dual angle condition. Observe that $\sigma(\vec{D}_B(a^+) - \frac{1}{2}r^{*2}) = \sigma(\vec{D}_{A^*}(b^*) - \frac{1}{2}r^{*2})$ due to (2.1) and (2.2). Also $\alpha = \sphericalangle(b - a^+, b^+ - a^+)$ is the same as $\alpha = \sphericalangle(\nabla f^*(a^*) - \nabla f^*(b^{*+}), \nabla f^*(a^{*+}) - \nabla f^*(b^{*+}))$. This means the correct definition is as follows:

Definition 5.12. (Dual angle condition). Let $\sigma : (0,\infty) \to (0,\infty)$ be monotonically increasing. The set A satisfies the *lr*-angle condition with constant $\gamma > 0$ and shrinking function σ with respect to B at a gap pair $b^* \sim a^*$, if there exists a neighborhood W of (b^*, a^*) and $\gamma, \eta > 0$ such that

$$\frac{1 - \cos\alpha}{\sigma(\vec{D}_A(b) - \frac{1}{2}r^{*2})} \geq \gamma \tag{5.4}$$

for every *lr*-building block $a \xrightarrow{l} b \xrightarrow{r} a^+$ with $(b, a^+) \in W$ and $\frac{1}{2}r^{*2} \leq \vec{D}_A(b) < \frac{1}{2}r^{*2} + \eta$, where $\alpha = \sphericalangle(\nabla f(a) - \nabla f(b), \nabla f(a^+) - \nabla f(b))$.

This can also be extended to gaps (K^*, r^*) using duality. Again we will content ourselves with angle shrinking functions of the form $\sigma(s) = \phi'(s)^{-2}s^{-1}$, respectively, $\sigma(s) = \phi'(s)^{-2}$ for a de-singularizing ϕ.

5.4. **Local projections.** The angle condition has an advantage over the KŁ-condition. We say that $a_k \in \vec{P}_A(b_k)$ is a local projection when the minimum (1.3) is local, while still guaranteeing decrease $D(b_k, a_{k-1}) > D(b_k, a_k)$. Consider a bounded alternating sequence which is local in this sense, in symbols $b_{k-1} \overset{r}{\dashrightarrow} a_k \xrightarrow{l} b_k$, and let A^*, B^* be the set of accumulation points of the a_k, b_k. Define $A^s = \{a_k : k \in \mathbb{N}\} \cup A^*$, $B^s = \{b_k : k \in \mathbb{N}\} \cup B^*$. Then a_k, b_k is a standard alternating sequence $b_{k-1} \xrightarrow{r} a_k \xrightarrow{l} b_k$ between the closed sets A^s, B^s, as points in A which previously might have made $\vec{P}_A$ local have been removed.

Now suppose A, B, f are definable, then the KŁ-inequality holds everywhere. Here we use the fact that the proof of Proposition 5.3 is still valid, since $\partial F(b^+, a^+)$ is not altered by $\vec{P}_A$ being

local. Hence the angle condition holds everywhere. But the angle condition is expressed in terms of building blocks, hence it is untouched when we restrict to the smaller sets A^s, B^s. This is significant, because A^s, B^s have no reason to be definable. Nor is there a good argument in favor of $F^s(x,y) = i_{B^s}(x) + D(x,y) + i_{A^s}(y)$ still having the KŁ-property, because the normal cones N_{A^s}, N_{B^s} are larger than N_A, N_B. This thread will be picked up again in Corollary 7.5.

6. Three point inequality and duality

In this section we discuss the second fundamental ingredient of our convergence theory, referred to as the three-point inequality in [52]. In the convex setting, a related notion has been discussed in [32]. We extend this now to Bregman projections.

Definition 6.1. The building block $b \xrightarrow{r} a^+ \xrightarrow{l} b^+$ satisfies the *rl*-three-point inequality with constant $\ell \in (0,1]$ if

$$D(b,a^+) \geq D(b^+,a^+) + \ell D(b,b^+). \tag{6.1}$$

The building block $a \xrightarrow{l} b \xrightarrow{r} a^+$ satisfies the *lr*-three-point inequality with $\ell \in (0,1]$ if

$$D(b,a) \geq D(b,a^+) + \ell D(a^+,a). \tag{6.2}$$

Remark 6.2. Putting $b^* = \nabla f(a^+)$, $a^* = \nabla f(b)$, $a^{*+} = \nabla f(b^+)$, we see from the dual relationship $D(x,y) = D^*(\nabla f(y), \nabla f(x))$ that (6.1) transforms into

$$D^*(b^*,a^*) \geq D^*(b^*,a^{*+}) + \ell D^*(a^{*+},a^*),$$

which is (6.2) for building blocks $a^* \xrightarrow{l*} b^* \xrightarrow{r*} a^{*+}$ alternating between $B^* = \nabla f(A)$ and $A^* = \nabla f(B)$ due to (2.2). In other words, the three-point inequality is directly amenable to duality.

In the euclidean case [52] *lr*- and *rl*-variants coincide. Now we define:

Definition 6.3. Let $a^* \in A$, $b^* \in B$, $b^* \sim a^*$. We say that the *rl*-three-point inequality holds at (b^*,a^*) if there exists $\ell \in (0,1)$ and $\delta > 0$ such that every *rl*-building block $b \xrightarrow{r} a^+ \xrightarrow{l} b^+$ with $b^+ \in B(b^*,\delta)$, $a^+ \in B(a^*,\delta)$ satisfies the *rl*-three-point inequality with constant ℓ.

The definition for the *lr*-case is analogous.

Proposition 6.4. *Suppose B is convex. Then the rl-three-point inequality holds for all A. Suppose $\nabla f(A)$ is convex. Then the lr-three-point inequality holds for all B.*

Proof. By the Bregman law of cosines $D(b,a^+) = D(b^+,a^+) + D(b,b^+) - \langle \nabla f(a^+) - \nabla f(b^+), b - b^+ \rangle$. Now $\nabla f(a^+) - \nabla f(b^+) \in N_B(b^+)$, hence by convexity of B the term $-\langle \nabla f(a^+) - \nabla f(b^+), b - b^+ \rangle$ is positive for $b \in B$, and we get $D(b,a^+) \geq D(b^+,a^+) + D(b,b^+)$, which is (6.1) with $\ell = 1$. For the second statement, by duality we have to show that building blocks $b^* \xrightarrow{r*} a^{*+} \xrightarrow{l*} b^{*+}$ satisfy the *rl*-three-point inequality. This follows from the first part, as now $B^* = \nabla f(A)$ is convex, while $A^* = \nabla f(B)$. □

Proposition 6.5. *Let (K^*,r^*) be a gap between A and B, and suppose that at every $(b^*,a^*) \in K^*$ a three-point inequality for rl-building blocks is satisfied. Then there exists $0 < \ell < 1$ and a neighborhood W of K^* such that the three point inequality holds with the same $\ell \in (0,1]$ for all building blocks $b \xrightarrow{r} a^+ \xrightarrow{l} b^+$ satisfying $(b^+,a^+) \in W$.*

Proof. This is a compactness argument. By hypothesis every $(b^*,a^*) \in K^*$ has a neighborhood $B(b^*,\delta_{b^*,a^*}) \times B(a^*,\delta_{b^*,a^*})$ such that *rl*-building blocks with $b^+ \in B(b^*,\delta_{b^*,a^*})$, $a^+ \in B(a^*,\delta_{b^*,a^*})$ satisfy the three-point inequality with $\ell_{b^*,a^*} \in (0,1)$. Now $K^* \subset \bigcup\{B(b^*,\delta_{b^*,a^*}/2) \times B(a^*,\delta_{b^*,a^*}/2) : (b^*,a^*) \in K^*\}$, and by compactness of K^* there exist finitely many pairs $(b_i^*,a_i^*) \in K^*$ such that $K^* \subset \bigcup\{B(b^*,\delta_{b_i^*,a_i^*}/2) \times B(a^*,\delta_{b_i^*,a_i^*}/2) : i = 1,\dots,m\} =: W$. Now let $\ell = \min\{\ell_{b_i^*,a_i^*} : i = 1,\dots,m\}$, and let $b \xrightarrow{r} a^+ \xrightarrow{l} b^+$ be a building block with $(b^+,a^+) \in W$. Then $(b^+,a^+) \in B(b^*,\delta_{b_i^*,a_i^*}/2) \times B(a^*,\delta_{b_i^*,a_i^*}/2)$ for some i. Therefore this building block satisfies the three-point inequality with $\ell_{b_i^*,a_i^*}$, hence also with ℓ. □

Interestingly, the three point inequality for building blocks $b \xrightarrow{r} a^+ \xrightarrow{l} b^+$ approaching a gap forces b, b^+ to get close.

Lemma 6.6. *Let a_k, b_k be a Bregman alternating sequence with gap (K^*, r^*). Suppose the rl-three point inequality holds in a neighborhood of K^*. Then $b_{k-1} - b_k \to 0$.*

Proof. Select an infinite subsequence $k \in N$ such that $b_{k-1} \to b^*$, $a_k \to a^*$, $b_k \to \hat{b}$. Suppose $\hat{b} \neq b^*$. We have $D(b_{k-1}, a_k) \to \frac{1}{2}r^{*2}$, $D(b_k, a_k) \to \frac{1}{2}r^{*2}$, $D(b_{k-1}, b_k) \to D(b^*, \hat{b}) > 0$. Therefore $D(b_{k-1}, a_k) \geq D(b_k, a_k) + \ell D(b_{k-1}, b_k) \to \frac{1}{2}r^{*2} + \ell D(b^*, \hat{b}) > \frac{1}{2}r^{*2}$, a contradiction. □

The analogue of Proposition 6.5 for lr-building blocks is obtained by duality, and the same goes for the compactness argument extending it from gap pairs to gaps. We skip the details. Sufficient conditions for the three-point inequality will be discussed in Section 10.

7. Convergence

We start with a global convergence result.

Theorem 7.1. (Global convergence). *Let a_k, b_k be a Bregman alternating sequence with gap (K^*, r^*). Suppose the rl-angle condition and rl-three-point inequality are satisfied at every $(b^*, a^*) \in K^*$. Then the sequence $b_k \in \tilde{P}_B \circ \vec{P}_A(b_{k-1})$ converges.*

Proof. 1) Since $D(b_{k-1}, a_{k-1}) \leq D(b_{k-1}, a_k) \leq D(b_k, a_k)$, we have $D(b_k, a_k) \to \frac{1}{2}r^{*2}$ and also $D(b_{k-1}, a_k) \to \frac{1}{2}r^{*2}$ by monotone convergence. Due to boundedness of the a_k, b_k the set K^* of accumulation point pairs (b^*, a^*) of the alternating sequence satisfying $b^* \sim a^*$ is compact, so by the usual compactness argument we find a neighborhood W of K^* on which the angle condition holds with the same γ, η and de-singularizing function ϕ, respectively, the same angle shrinking function $\sigma(s) = \phi'(s)^{-2}$ or $\sigma(s) = \phi'(s)^{-2}s^{-1}$, simultaneously for all lr-building blocks in W. Shrinking W further, and using Proposition 6.5, we may in addition arrange that the three point inequality is satisfied with ℓ for all lr-building blocks in W. Since there are only finitely many iterates outside W, we may without loss assume

$$\phi'(\tilde{D}_B(a_k) - \tfrac{1}{2}r^{*2})^2 \tilde{D}_B(a_k)(1 - \cos\alpha_k) \geq \gamma \tag{7.1}$$

for all k, where $\alpha_k = \sphericalangle(b_{k-1} - a_k, b_k - a_k)$. With the same argument we can also assume that the three-point inequality holds for all k with the same constant $\ell \in (0, 1)$, hence we also have the four point inequality

$$D(b_k, a_k) \geq D(b_{k+1}, a_{k+1}) + \ell D(b_k, b_{k+1}). \tag{7.2}$$

2) Due to our standing assumption $A, B \subset G$, we may further assume that (2.3) holds on W with suitable constants m, M. Now

$$\begin{aligned}
m^{-2}D(b_{k-1}, b_k) &\geq \|b_{k-1} - b_k\|^2 \\
&= \|b_{k-1} - a_k\|^2 + \|a_k - b_k\|^2 - 2\|b_{k-1} - a_k\|\|a_k - b_k\|\cos\alpha_k \\
&= (\|b_{k-1} - a_k\| - \|a_k - b_k\|)^2 + 2\|b_{k-1} - a_k\|\|a_k - b_k\|(1 - \cos\alpha_k) \\
&\geq 2\|b_{k-1} - a_k\|\|a_k - b_k\|(1 - \cos\alpha_k) \\
&\geq 2M^{-1}D(b_{k-1}, a_k)^{1/2}\|a_k - b_k\|(1 - \cos\alpha_k) \\
&\geq 2\ M^{-1}D(b_k, a_k)^{1/2}\|a_k - b_k\|(1 - \cos\alpha_k) \\
&\geq 2M^{-2}D(b_k, a_k)(1 - \cos\alpha_k) \\
&\geq 2M^{-2}\gamma\phi'(\tilde{D}_B(a_k) - \tfrac{1}{2}r^{*2})^{-2}.
\end{aligned} \tag{7.3}$$

Here lines 1,5,7 use (2.3), lines 2-4 concern the cosine theorem, line 6 uses $D(b_k, a_k) \leq D(b_{k-1}, a_k)$, and the last line uses the angle condition (7.1). We re-write this as

$$D(b_{k-1}, b_k)^{1/2} \geq \frac{m\sqrt{2\gamma}}{M}\phi'(\tilde{D}_B(a_k) - \tfrac{1}{2}r^{*2})^{-1}. \tag{7.4}$$

3) Concavity of the de-singularizing function ϕ now implies

$$\begin{aligned}\phi(\tilde{D}_B(a_k)-\tfrac{1}{2}r^{*2})-\phi(\tilde{D}_B(a_{k+1})-\tfrac{1}{2}r^{*2}) &\geq \phi'(\tilde{D}_B(a_k)-\tfrac{1}{2}r^{*2})\left[\tilde{D}_B(a_k)-\tfrac{1}{2}r^{*2}-(\tilde{D}_B(a_{k+1})-\tfrac{1}{2}r^{*2})\right]\\ &= \phi'(\tilde{D}_B(a_k)-\tfrac{1}{2}r^{*2})\left[D(b_k,a_k)-D(b_{k+1},a_{k+1})\right]\\ &\geq \phi'(\tilde{D}_B(a_k)-\tfrac{1}{2}r^{*2})\ell D(b_k,b_{k+1})\\ &\geq m\sqrt{2\gamma}M^{-1}\ell\frac{D(b_k,b_{k+1})}{D(b_{k-1},b_k)^{1/2}}.\end{aligned}$$

Here the third line uses the four point inequality (7.2), while the last line uses (7.4). Setting $C = M/m\sqrt{2\gamma}\ell$, we have

$$C\left[\phi(\tilde{D}_B(a_k)-\tfrac{1}{2}r^{*2})-\phi(\tilde{D}_B(a_{k+1})-\tfrac{1}{2}r^{*2})\right]D(b_{k-1},b_k)^{1/2}\geq D(b_k,b_{k+1}).$$

Since $a^2\leq bc$ implies $a\leq\frac{1}{2}b+\frac{1}{2}c$ for positive a,b,c, we deduce

$$D(b_k,b_{k+1})^{1/2}\leq\frac{1}{2}D(b_{k-1},b_k)^{1/2}+\frac{C}{2}\left[\phi(\tilde{D}_B(a_k)-\tfrac{1}{2}r^{*2})-\phi(\tilde{D}_B(a_{k+1})-\tfrac{1}{2}r^{*2})\right]. \tag{7.5}$$

Summing this from $k=1$ to n gives

$$\sum_{k=1}^{n}D(b_k,b_{k+1})^{1/2}\leq\frac{1}{2}\sum_{k=1}^{n}D(b_{k-1},b_k)^{1/2}+\frac{C}{2}\left[\phi(\tilde{D}_B(a_1)-\tfrac{1}{2}r^{*2})-\phi(\tilde{D}_B(a_{n+1})-\tfrac{1}{2}r^{*2})\right].$$

Re-arranging, and multiplying by 2, we obtain

$$\begin{aligned}\sum_{k=1}^{n}D(b_k,b_{k+1})^{1/2} &\leq D(b_0,b_1)^{1/2}+C\left[\phi(\tilde{D}_B(a_1)-\tfrac{1}{2}r^{*2})-\phi(\tilde{D}_B(a_{n+1})-\tfrac{1}{2}r^{*2})\right]-D(b_n,b_{n+1})^{1/2}\\ &\leq D(b_0,b_1)^{1/2}+C\phi(\tilde{D}_B(a_1)-\tfrac{1}{2}r^{*2}).\end{aligned}$$

Using (2.3), this implies

$$\sum_{k=1}^{n}\|b_k-b_{k+1}\|\leq m^{-1}M\|b_0-b_1\|+m^{-1}C\phi(\tilde{D}_B(a_1)-\tfrac{1}{2}r^{*2}).$$

This proves convergence of the series $\sum_{k=1}^{\infty}\|b_k-b_{k+1}\|$, hence the sequence b_k is Cauchy and converges to some $b^*\in B$. □

Remark 7.2. For $r^*>0$ the proof does not assure convergence of the a_k, but all accumulation points a^* of the a_k satisfy $D(b^*,a^*)=\frac{1}{2}r^{*2}$ and $b^*\sim a^*$, i.e., lie on the boundary of $\vec{\mathcal{B}}(b^*,r^*)$, and the gap has the form $K^*=\{b^*\}\times A^*$. In the feasible case $r^*=0$, convergence of the a_k is guaranteed, but here we have an even stronger result:

Theorem 7.3. (Convergence by attraction). *Let $x^*\in A\cap B$, and suppose rl-angle condition and rl-three point estimate are satisfied at x^*. Then there exists a neighborhood V of x^* such that every Bregman alternating sequence which enters V converges to some point in the intersection.*

Proof. 1) By Proposition 5.3 the angle condition (5.1) holds on a neighborhood U of $x^*\in G$ with the shrinking function $\sigma(s)=1/s\phi'(s)^2$ and a constant γ. In addition, U may be chosen such that every building block $b\xrightarrow{r}a^+\xrightarrow{l}b^+$ in U satisfies the three point estimate (6.1) with $0<\ell<1$. From the three-point inequality we immediately obtain the following four-point-inequality

$$D(b,a)\geq D(b^+,a^+)+\ell D(b,b^+) \tag{7.6}$$

for building blocks $a\xrightarrow{l}b\xrightarrow{r}a^+\xrightarrow{l}b^+$ with $b,a^+,b^+\in U$.

Since the alternating sequence including accumulation points is contained in the interior of $\mathrm{dom}\,f$, we may assume that (2.3) with constants m,M is satisfied in a neighborhood of the set of iterates. In particular, we assume that it is satisfied on the neighborhood U of x^*. Let $U=B(x^*,\epsilon)$ without loss. Now define

$$C=M/m\sqrt{2\gamma}\ell.$$

2) Choose $\delta > 0$ such that the following are satisfied:

$$\begin{aligned}(m^{-5}M^5 + m^{-4}M^4 + m^{-3}M^3 + m^{-2}M^2 + m^{-1}M + 1)\delta < \epsilon/2 \\ (m^{-1} + m^{-2}M + m^{-3}M^2 + m^{-4}M^3)C\phi(\xi) < \epsilon/2\end{aligned} \tag{7.7}$$

for all $|\xi| < M^2\delta^2$. The latter is possible due to $\phi(0) = 0$ and continuity of ϕ at 0. Now let $V = B(x^*, \delta)$. We claim that if the alternating sequence $a_k \xrightarrow{l} b_k \xrightarrow{r} a_{k+1} \xrightarrow{l} b_{k+1}$ enters V, then it converges to a point $b^\sharp \in A \cap B$. Relabeling the sequence, we may assume that $b_0 \in V$. The case where the a_k reach V first is treated analogously.

3) We shall prove by induction that for every $k \geq 1$,

$$b_0, a_1, b_1, \dots, a_k, b_k, a_{k+1}, b_{k+1} \in U \tag{7.8}$$

and

$$\sum_{j=1}^{k} D(b_j, b_{j+1})^{1/2} \leq \frac{1}{2}\sum_{j=1}^{k} D(b_{j-1}, b_j)^{1/2} + \frac{C}{2}\left[\phi(\tilde{D}_B(a_1)) - \phi(\tilde{D}_B(a_{k+1}))\right]. \tag{7.9}$$

Let us first prove (7.8) for $k = 1$. This means we have to show $a_1, b_1, a_2, b_2 \in U$. We have $D(b_0, a_1) \leq D(b_0, x^*)$ due to $x^* \in A$, hence $m\|b_0 - a_1\| \leq D(b_0, a_1)^{1/2} \leq D(b_0, x^*)^{1/2} \leq M\|b_0 - x^*\|$, giving $\|b_0 - a_1\| \leq m^{-1}M\delta$. Then $\|a_1 - x^*\| \leq \|a_1 - b_0\| + \|b_0 - x^*\| \leq (m^{-1}M + 1)\delta < \epsilon$ using (7.7), which is the first claim.

Now $D(b_1, a_1) \leq D(b_0, a_1)$, hence $m\|b_1 - a_1\| \leq M\|b_0 - a_1\| \leq m^{-1}M^2\delta$. Then $\|b_1 - a_1\| \leq M^2m^{-2}\delta$, giving $\|b_1 - x^*\| \leq \|b_1 - a_1\| + \|a_1 - x^*\| \leq (m^{-2}M^2 + m^{-1}M + 1)\delta < \epsilon$, again using (7.7). This is the second statement in $(7.8)_1$.

Next $D(b_1, a_2) \leq D(b_1, a_1)$, which gives $m\|b_1 - a_2\| \leq M\|b_1 - a_1\| \leq m^{-2}M^3\delta$, hence $\|b_1 - a_2\| \leq m^{-3}M^3\delta$. Then $\|a_2 - x^*\| \leq \|a_2 - b_1\| + \|b_1 - x^*\| \leq (m^{-3}M^3 + m^{-2}M^2 + m^{-1}M + 1)\delta < \epsilon$ from (7.7), which is the third statement in $(7.8)_1$.

Finally, from $D(b_2, a_2) \leq D(b_2, a_1)$ we get $m\|b_2 - a_2\| \leq M\|b_2 - a_1\|$, hence $\|b_2 - a_2\| \leq m^{-4}M^4\delta$, so that $\|b_2 - x^*\| \leq \|b_2 - a_2\| + \|a_2 - x^*\| < (m^{-4}M^4 + m^{-3}M^3 + m^{-2}M^2 + m^{-1}M + 1)\delta < \epsilon$, once again via (7.7). That proves $a_1, b_1, a_2, b_2 \in U$.

4) Before proving $(7.9)_1$, let us first do the induction step. Suppose $(7.8)_{k-1}$ and $(7.9)_{k-1}$ are satisfied. We have to prove $(7.8)_k$ and $(7.9)_k$. We first check (7.8) at k. By $(7.8)_{k-1}$ we know that $b_0, a_1, b_1, \dots, a_k, b_k \in U$, so it remains to prove $a_{k+1}, b_{k+1} \in U$. Now observe that $(7.9)_{k-1}$ implies

$$\begin{aligned}\sum_{j=1}^{k-1} D(b_j, b_{j+1})^{1/2} &\leq D(b_0, b_1)^{1/2} + C\left[\phi(\tilde{D}_B(a_1)) - \phi(\tilde{D}_B(a_k))\right] - D(b_{k-1}, b_k)^{1/2} \\ &\leq D(b_0, b_1)^{1/2} + C\phi(\tilde{D}_B(a_1)).\end{aligned}$$

Using (2.3) this implies

$$\sum_{j=1}^{k-1} \|b_{j-1} - b_j\| \leq m^{-1}M\|b_0 - b_1\| + m^{-1}C\phi(\tilde{D}_B(a_1)).$$

Using this, we have

$$\begin{aligned}\|b_k - x^*\| \leq \|b_k - b_1\| + \|b_1 - x^*\| &\leq \sum_{j=1}^{k-1} \|b_j - b_{j+1}\| + \|b_1 - x^*\| \\ &\leq m^{-1}M\|b_0 - b_1\| + m^{-1}C\phi(\tilde{D}_B(a_1)) + \|b_1 - x^*\| \\ &\leq m^{-1}M\|b_0 - a_1\| + m^{-1}M\|a_1 - b_1\| + m^{-1}C\phi(\tilde{D}_B(a_1)) + (m^{-2}M^2 + m^{-1}M + 1)\delta \\ &< m^{-2}M^2\delta + m^{-3}M^3\delta + m^{-1}C\phi(\tilde{D}_B(a_1)) + (m^{-2}M^2 + m^{-1}M + 1)\delta \\ &= (m^{-3}M^3 + 2m^{-2}M^2 + m^{-1}M + 1)\delta + m^{-1}C\phi(\tilde{D}_B(a_1)).\end{aligned}$$

Now $D(b_k, a_{k+1}) \leq D(b_k, x^*)$, hence $\|b_k - a_{k+1}\| \leq m^{-1}M\|b_k - x^*\| \leq (m^{-4}M^4 + \dots + m^{-1}M)\delta + m^{-2}MC\phi(\tilde{D}_B(a_1))$. Then $\|a_{k+1} - x^*\| \leq \|a_{k+1} - b_k\| + \|b_k - x^*\| \leq (m^{-4}M^4 + \dots + 1)\delta + (m^{-1} + m^{-2}M)C\phi(\tilde{D}_B(a_1)) < \epsilon/2 + \epsilon/2 = \epsilon$, where we use (7.7), being allowed to do so due to $\tilde{D}_B(a_1) <$

$M^2\delta^2$. Namely, $\tilde{D}_B(a_1) = D(b_1, a_1) \leq D(b_0, a_1) \leq D(b_0, x^*) \leq M^2\|b_0 - x^*\|^2 \leq M^2\delta^2$, which bounds the term $(m^{-1} + m^{-2}M)C\phi(\tilde{D}_B(a_1))$ according to the second condition in (7.7).

Finally, $D(b_{k+1}, a_{k+1}) \leq D(b_k, a_{k+1})$, so $\|b_{k+1} - a_{k+1}\| \leq m^{-1}M\|b_k - a_{k+1}\| \leq (m^{-5}M^5 + \cdots + m^{-1}M)\delta + m^{-3}M^2C\phi(\tilde{D}_B(a_1))$, which gives $\|b_{k+1} - x^*\| \leq \|b_{k+1} - a_{k+1}\| + \|a_{k+1} - x^*\| \leq (m^{-5}M^5 + \cdots + 1)\delta + (m^{-1} + m^{-2}M + m^{-3}M^2)C\phi(\tilde{D}_B(a_1)) < \epsilon/2 + \epsilon/2 = \epsilon$, using both estimates in (7.7). That proves $(7.8)_k$.

6) Now let us prove $(7.9)_k$. From the angle condition (5.1) with $\sigma(s) = \phi'(s)^{-2}s^{-1}$ we have

$$\phi'(\tilde{D}_B(a_k))^2\tilde{D}_B(a_k)(1 - \cos\alpha_k) \geq \gamma, \tag{7.10}$$

where $\alpha_k = \sphericalangle(b_{k-1} - a_k, b_k - a_k)$. We also have the four-point estimate

$$D(b_{k+1}, a_{k+1}) + \ell D(b_k, b_{k+1}) \leq D(b_k, a_k) \tag{7.11}$$

because $D(b_k, a_{k+1}) \leq D(b_k, a_k)$. Now we invoke precisely the same estimation as used in the proof of Theorem 7.1, which via (7.1) and (7.2) led to (7.4). Here we use instead (7.10) and (7.11) to derive (using $r^* = 0$):

$$D(b_{k-1}, b_k)^{1/2} \geq \frac{m\sqrt{2\gamma}}{M}\phi'(\tilde{D}_B(a_k))^{-1}. \tag{7.12}$$

Now by concavity of ϕ, the four point estimate (7.11), and (7.12), we get

$$\begin{aligned}
\phi(\tilde{D}_B(a_k)) - \phi(\tilde{D}_B(a_{k+1})) &\geq \phi'(\tilde{D}_B(a_k))\left(D(b_k, a_k) - D(b_{k+1}, a_{k+1})\right) \\
&\geq \phi'(\tilde{D}_B(a_k))\ell D(b_k, b_{k+1}) \\
&\geq m\sqrt{2\gamma}M^{-1}\ell\frac{D(b_k, b_{k+1})}{D(b_{k-1}, b_k)^{1/2}}.
\end{aligned}$$

With $C = M/m\sqrt{2\gamma}\ell$ this becomes

$$C\left[\phi(\tilde{D}_B(a_k)) - \phi(\tilde{D}_B(a_{k+1}))\right]D(b_{k-1}, b_k)^{1/2} \geq D(b_k, b_{k+1}).$$

Since $a^2 \leq bc$ implies $a \leq \frac{1}{2}b + \frac{1}{2}c$ for positive a, b, c, we deduce

$$D(b_k, b_{k+1})^{1/2} \leq \frac{1}{2}D(b_{k-1}, b_k)^{1/2} + \frac{C}{2}\left[\phi(\tilde{D}_B(a_k)) - \phi(\tilde{D}_B(a_{k+1}))\right]. \tag{7.13}$$

By the induction hypothesis $(7.9)_{k-1}$ we have

$$\sum_{j=1}^{k-1} D(b_j, b_{j+1})^{1/2} \leq \frac{1}{2}\sum_{j=1}^{k-1} D(b_{j-1}, b_j)^{1/2} + \frac{C}{2}\left[\phi(\tilde{D}_B(a_1)) - \phi(\tilde{D}_B(a_k))\right].$$

Adding this and (7.13) gives $(7.9)_k$ at stage k.

7) It remains to prove (7.9) at $k = 1$. This can be done by following the same steps as in 6) with $k = 1$.

8) Having proved $(7.9)_k$ for all k, we see that the series $\sum_{k=1}^{\infty} D(b_{k-1}, b_k)^{1/2}$ converges, and that all iterates stay in U. Hence via (2.3) the series $\sum_{k=1}^{\infty}\|b_{k-1} - b_k\|$ converges as well, hence the b_k form a Cauchy sequence, which converges to some $b^\sharp \in B \cap U$.

9) Convergence of the a_k is obtained as follows. We have $D(b_k, a_{k+1}) \leq D(b_k, b^\sharp)$ due to $b^\sharp \in A \cap B$ and $a_{k+1} \in \vec{P}_A(b_k)$, hence $\|a_{k+1} - b_k\| \leq m^{-1}M\|b_k - b^\sharp\|$, and then $\|a_{k+1} - b^\sharp\| \leq \|a_{k+1} - b_k\| + \|b_k - b^\sharp\| \leq (1 + m^{-1}M)\|b_k - b^\sharp\|$. □

Corollary 7.4. *Suppose A, B, f are definable and B is prox-regular at $x^* \in A \cap B$. Then there exists a neighborhood V of x^* such that every Bregman alternating sequence $a_k \xrightarrow{l} b_k \xrightarrow{r} a_{k+1}$ which reaches V converges to a point $b^\sharp \in A \cap B$.*

We pick up the thread of local projections from Section 5.4.

Corollary 7.5. *Consider a local Bregman alternating sequence $b_{k-1} \overset{r}{\dashrightarrow} a_k \xrightarrow{l} b_k$ with gap K^*. Suppose A, B, f are definable. Let the rl-three point inequality be satisfied on K^*. Then the sequence b_k converges.*

Proof. Since a_k, b_k is a standard Bregman alternating sequence between the sets A^s, B^s, we know from Section 5.4 that the angle condition, which holds throughout A, B, remains in place between A^s and B^s. Let K^* be the gap generated by the a_k, b_k, then K^* is also the gap of a_k, b_k when the latter is considered alternating between A^s and B^s. By hypothesis the rl-three point inequality holds on K^*, and since it is also expressed in terms of building blocks, it remains true for a_k, b_k alternating between A^s, B^s. Altogether, by Theorem 7.1, the sequence b_k converges. □

This is important from a practical point of view when A is not convex, as we then may want to solve (1.3) with a local NLP-solver, starting with the last a as initial guess, thereby assuring descent (1.5).

8. dual convergence

We continue to consider the alternating sequence under the form

$$a \xrightarrow{l} b \xrightarrow{r} a^+ \xrightarrow{l} b^+$$

where $D(b^+, a^+) \leq D(b, a^+) \leq D(b, a)$. However, we now reverse the roles of A and B, i.e., we assume that the dual angle condition (5.4) and the dual three point estimate (6.2) are satisfied, now for lr-building blocks $a \xrightarrow{l} b \xrightarrow{r} a^+$.

Theorem 8.1. *Let a_k, b_k be a Bregman alternating sequence with gap (K^*, r^*). Suppose the lr-angle condition and lr-three point inequality are satisfied at every pair $(b^*, a^*) \in K^*$. Then the sequence $a_k = \vec{P}_A \circ \tilde{P}_B(a_{k-1})$ converges.*

Proof. We use duality to obtain the mirror sequence a_k^*, b_k^* of the a_k, b_k. Then (K^*, r^*) is mapped into the dual gap $(\nabla f(K^*), r^*)$, and by amenability of the angle condition and the three-point inequality, the dual sequence now satisfies the rl-angle condition and rl-three-point inequality at $\nabla f(K^*)$. Therefore $b_k^* \in \tilde{P}^*_{B^*} \circ \vec{P}^*_{A^*}(b^*_{k-1})$ converges by Theorem 7.1. Mapping this back under ∇f^* yields convergence of $a_k = \vec{P}_A \circ \tilde{P}_B(a_{k-1})$. □

It is again possible to obtain convergence by attraction using duality.

Corollary 8.2. *Let $\bar{x} \in A \cap B$ and suppose lr-angle condition and lr-three-point inequality are satisfied at $\bar{x}$. Then there exists a neighborhood V of $\bar{x}$ such that every Bregman alternating sequence which enters V converges to some point in the intersection.*

Naturally, we can also address the case of local projections $a \dashrightarrow b \xrightarrow{r} a^+$ via duality. We skip the details.

9. Speed of convergence

We consider the case of the Łojasiewicz inequality, where the de-singularizing function is $\phi(s) = s^{1-\theta}$ for some $\theta \in [\frac{1}{2}, 1)$. In that case, worst case convergence rates can be obtained.

Corollary 9.1. *Under the hypotheses of Theorem 7.1, suppose the de-singularizing function is of the form $\phi'(s) = s^{-\theta}$ for some $\theta \in (\frac{1}{2}, 1)$. Then the speed of convergence of the sequence $b_k = \tilde{P}_B \circ \vec{P}_A(b_{k-1})$ is $\|b_k - b^*\| = O(k^{-\rho})$ with $\rho = \frac{1-\theta}{2\theta-1} \in (0, \infty)$. When $\theta = \frac{1}{2}$ the speed is R-linear. In the feasible case, the a_k converge to $b^* \in A \cap B$ with the same speed.*

Proof. Summing (7.5) from $k = N$ to $k = K$ gives

$$\begin{aligned} -\frac{1}{2} D(b_{N-1}, b_N)^{1/2} + \frac{1}{2} \sum_{k=N}^{K-1} D(b_k, b_{k+1})^{1/2} + D(b_K, b_{K+1})^{1/2} \\ \leq \frac{C}{2} \left[\phi(\tilde{D}_B(a_N) - \tfrac{1}{2} r^{*2}) - \phi(\tilde{D}_B(a_{K+1}) - \tfrac{1}{2} r^{*2}) \right]. \end{aligned} \tag{9.1}$$

Passing to the limit $K \to \infty$ gives

$$-\frac{1}{2} D(b_{N-1}, b_N)^{1/2} + \frac{1}{2} \sum_{k=N}^{\infty} D(b_k, b_{k+1})^{1/2} \leq \frac{C}{2} \phi(\tilde{D}_B(a_N) - \tfrac{1}{2} r^{*2}).$$

Introducing $S_N = \sum_{k=N}^{\infty} D(b_k, b_{k+1})^{1/2}$, this reads

$$-\frac{1}{2}(S_{N-1} - S_N) + \frac{1}{2}S_N \leq \frac{C}{2}\phi(\tilde{D}_B(a_N) - \tfrac{1}{2}r^{*2}).$$

On the other hand, (7.4) gives

$$\phi'(\tilde{D}_B(a_N) - \tfrac{1}{2}r^{*2})^{-1} \leq \frac{M}{m\sqrt{2\gamma}} D(b_{N-1}, b_N)^{1/2} = \frac{M}{m\sqrt{2\gamma}}(S_{N-1} - S_N).$$

Now by hypothesis we have $\phi'(s) = s^{-\theta}$ for $\theta \in [\frac{1}{2}, 1)$, hence $\phi(s) = (1-\theta)^{-1}s^{1-\theta}$. Therefore $\left[\phi'(s)^{-1}\right]^{\frac{1-\theta}{\theta}} = s^{1-\theta} = (1-\theta)\phi(s)$. Hence $\phi(\tilde{D}_B(a_N) - \frac{1}{2}r^{*2}) \leq (1-\theta)^{-1}[\phi'(\tilde{D}_B(a_N) - \frac{1}{2}r^{*2})^{-1}]^{\frac{1-\theta}{\theta}} \leq (1-\theta)^{-1}\left(\frac{M}{m\sqrt{2\gamma}}\right)^{\frac{1-\theta}{\theta}} (S_{N-1} - S_N)^{\frac{1-\theta}{\theta}}$. Substituting this gives

$$\frac{1}{2}S_N \leq C'(S_{N-1} - S_N)^{\frac{1-\theta}{\theta}} + \frac{1}{2}(S_{N-1} - S_N) \tag{9.2}$$

with $C' = \frac{C}{2}(1-\theta)^{-1}\left(\frac{M}{m\sqrt{2\gamma}}\right)^{\frac{1-\theta}{\theta}}$. Now for $\theta > \frac{1}{2}$ we have $\frac{1-\theta}{\theta} < 1$, so that the first term on the right hand side dominates the second term. Therefore there exists another constant C'' such that

$$S_N^{\frac{\theta}{1-\theta}} \leq C''(S_{N-1} - S_N).$$

From here we follow precisely the argument in [52, Cor. 4 (24) ff], where this leads to an estimate of the form

$$S_N \leq C''' N^{-\frac{1-\theta}{2\theta-1}}$$

for another constant C'''. Using (2.3), this shows $\widetilde{S}_N := \sum_{k=N}^{\infty} \|b_k - b_{k+1}\| \leq m^{-1}S_N \leq m^{-1}C'''N^{-\frac{1-\theta}{2\theta-1}}$, and since $\|b_N - b^*\| \leq \widetilde{S}_N$ by the triangle inequality, we get the desired estimate $\|b_k - b^*\| = O(k^{-\rho})$ with $\rho = \frac{1-\theta}{2\theta-1}$.

Now consider the case $\theta = \frac{1}{2}$, then (9.2) turns into

$$\frac{1}{2}S_N \leq C'(S_{N-1} - S_N) + \frac{1}{2}(S_{N-1} - S_N),$$

hence

$$S_N \leq \frac{1+2C'}{2+2C'}S_{N-1},$$

which gives Q-linear speed $S_N \to 0$, hence R-linear speed $\widetilde{S}_N \to 0$, and then also R-linear speed $\|b_N - b^*\| \to 0$.

Finally, in the feasible case we get the same speed of convergence for the a_k from part 9) of the proof of Theorem 7.3. □

Remark 9.2. This may be compared to [40], where for a non-convex version of the EM algorithm for exponential families a global estimate $D(b_k, b_{k+1})^{1/2} = O(k^{-1/2})$ is obtained without use of the KŁ-inequality, with D the Kullback-Leibler divergence. See also Example 12.1 in Section 12, and Section 11.

Remark 9.3. Naturally, via duality, the same speed of convergence is obtained for Theorem 8.1 if $\phi(s) = s^{1-\theta}$.

Corollary 9.4. (Linear convergence). *Suppose B intersects A rl-transversally at $\bar{x} \in A \cap B$, and the rl-three-point inequality is satisfied at $\bar{x}$. Then there exists a neighborhood V of $\bar{x}$ such that every Bregman alternating sequence which enters V converges to some point in the intersection with R-linear speed.*

Proof. The neighborhood V of $\bar{x}$ may be chosen such that the *rl*-three-point inequality holds on V. By hypothesis we may also assure that the numerator $1 - \cos\alpha$ in (5.1) stays bounded away from 0 in V. Therefore we can allow a constant as angle shrinking function σ. Now $\sigma(s) = \phi'(s)^{-2}s^{-1}$ gives constant σ as soon as $\phi'(s) \propto s^{-1/2}$, so the de-singularizing function is $\phi(s) = s^{1/2}$. Then by Corollary 9.1 convergence is R-linear near $\bar{x}$. □

The dual version of this result is also true.

10. Sufficient conditions for the three-point inequality

In this chapter we derive the three point inequality from conditions on the reach of the sets A, B. Bregman reach had been introduced in Section 3.4, and extends the classical notion of reach [36]. However, there are other ways to extend the classical notion of reach to the Bregman setting, each with advantages and inconveniences.

10.1. **Bregman reach larger than gap.** We start discussing left Bregman reach $\overleftarrow{R}(b^+, d)$, which we match with the distance between A and B measured by $D(b^+, a^+)^{1/2}$.

Proposition 10.1. *Let $b^* \sim a^*$ with gap value $r^* \geq 0$ and suppose the left Bregman reach at $b^* \in B$ is at least $r > r^*$. Suppose f is 1-coercive. Then there exist $\delta > 0$ and $0 < \ell < 1$ such that the three point inequality holds with ℓ for every building block $b \xrightarrow{r} a^+ \xrightarrow{l} b^+$ with $(b^+, a^+) \in B(b^*, \delta) \times B(a^*, \delta)$.*

Proof. 1) Recall that the three point inequality holds trivially for every $\ell \in (0, 1]$ if a building block satisfies $\langle \nabla f(a^+) - \nabla f(b^+), b - b^+ \rangle \leq 0$. We therefore assume $\langle \nabla f(a^+) - \nabla f(b^+), b - b^+ \rangle > 0$ throughout. Geometrically, this means that b, a^+ are strictly on the same side of the tangent hyperplane H at b^+.

2) We have $b^+ \in \overleftarrow{P}_B(a^+)$ with $a^+ \in A$, and the Bregman perpendicular a_λ to B at b^+ in direction $d = \nabla^2 f(b^+)^{-1}(\nabla f(a^+) - \nabla f(b^+))$ satisfies

$$\nabla f(a_\lambda) - \nabla f(b^+) = \lambda(\nabla f(a^+) - \nabla f(b^+)). \tag{10.1}$$

Now consider b strictly on the same side of the tangent hyperplane as a^+. We claim that there exists a parameter $\lambda(b) > 1$ for which the point $a_{\lambda(b)}$ on the geodesic gives equality $D(b, a_{\lambda(b)}) = D(b^+, a_{\lambda(b)})$. Indeed, the cosine theorem for Bregman distances in tandem with (10.1) gives

$$D(b, a_\lambda) = D(b, b^+) + D(b^+, a_\lambda) - \lambda \langle \nabla f(a^+) - \nabla f(b^+), b - b^+ \rangle, \tag{10.2}$$

hence

$$\begin{aligned} \frac{D(b, a_\lambda) - D(b^+, a_\lambda)}{\lambda} &= \frac{D(b, b^+)}{\lambda} - \langle \nabla f(a^+) - \nabla f(b^+), b - b^+ \rangle \\ &\to -\langle \nabla f(a^+) - \nabla f(b^+), b - b^+ \rangle < 0 \end{aligned}$$

as $\lambda \to \infty$, so that eventually $D(b^+, a_\lambda) > D(b, a_\lambda)$. Here we use the fact that due to 1-coercivity of f the a_λ are defined for all $\lambda \geq 0$, so that we may pass to the limit, and we use the fact that the limit term is negative. Since at $a_\lambda|_{\lambda=1} = a^+$ we have $D(b, a_1) > D(b^+, a_1)$, the intermediate value theorem gives $\lambda = \lambda(b) \in (1, \infty)$ with equality.

3) Differentiating (10.1) with respect to λ gives $\nabla^2 f(a_\lambda) \frac{d}{d\lambda} a_\lambda = \nabla f(a^+) - \nabla f(b^+)$. Therefore $\frac{d}{d\lambda} D(b^+, a_\lambda) = -\langle \frac{d}{d\lambda} a_\lambda, \nabla^2 f(a_\lambda)(b^+ - a_\lambda) \rangle = -\langle \nabla^2 f(a_\lambda) \frac{d}{d\lambda} a_\lambda, b^+ - a_\lambda \rangle = \langle \nabla f(a^+) - \nabla f(b^+), a_\lambda - b^+ \rangle > 0$, as all a_λ are on the same side of the tangent hyperplane as a^+. This means $\lambda \mapsto D(b^+, a_\lambda)$ is strictly increasing, hence the intermediate value $\lambda(b)$ found above is unique.

4) From $D(b, a_{\lambda(b)}) = D(b^+, a_{\lambda(b)})$ we get $D(b, b^+) = \lambda(b) \langle \nabla f(a^+) - \nabla f(b^+), b - b^+ \rangle$ using (10.2), hence the three point inequality holds with $\ell(b) = 1 - \lambda(b)^{-1}$ for every building block $b \xrightarrow{r} a^+ \xrightarrow{l} b^+$ with b strictly on the same side of the hyperplane H as a^+.

What remains to be shown is that there is one global ℓ which works for all these building blocks, or put differently, that the $\ell(b) = 1 - \lambda(b)^{-1}$ stay bounded away from 0 even when their building blocks approach the gap.

From the construction we have $b, b^+ \in \partial \overleftarrow{\mathcal{B}}(a_{\lambda(b)}, r_{\lambda(b)})$, so by the definition of the left reach $r_{\lambda(b)} \geq \overleftarrow{R}(b^+, d) \geq r > r^*$ for all such $(b^+, a^+) \in W$. Now assume contrary to what is claimed that there exist building blocks $b_{k-1} \xrightarrow{r} a_k \xrightarrow{l} b_k$ with $a_k \to a^*$, $b_k \to b^*$, such that $\lambda(b_{k-1}) \to 1$. Then $\nabla f(a_{\lambda(b_{k-1})}) = \nabla f(b_k) + \lambda(b_{k-1})(\nabla f(a_k) - \nabla f(b_k)) \to \nabla f(a^*)$. Since ∇f is an diffeomorphism from $\mathrm{int}(\mathrm{dom}\ f)$ onto $\mathrm{int}(\mathrm{dom} f^*)$, this implies $a_{\lambda(b_{k-1})} \to a^*$. Then $D(b_k, a_{\lambda(b_{k-1})}) \to D(b^*, a^*) = \frac{1}{2} r^{*2}$, hence $r_{\lambda(b_{k-1})} \to r^*$, a contradiction with the above. □

Remark 10.2. In view of Proposition 3.13 it is unlikely that the result still holds without 1-coercivity of f. This is why we consider alternative ways to define reach via $\widetilde{R}$ in the following sections.

10.2. **Mobile reach larger than gap.** We use a different notion of reach based on what we call a mobile euclidean norm. This requires measuring the gap between the sets differently.

With every building block $b \xrightarrow{r} a^+ \xrightarrow{l} b^+$ we associate the euclidean norm $\|x\|_{b^+}^2 = \langle x, \nabla^2 f(b^+)x\rangle = \langle x, x\rangle_{b^+}$. Due to $\nabla^2 f(b^+) \succeq \epsilon > 0$ for b^+ in a compact subset K of the interior of $\mathrm{dom} f$, we have an estimate of the form

$$m\|x\|_{b^+} \leq \|x\| \leq M\|x\|_{b^+}, \tag{10.3}$$

with m, M depending only on K, and for every $b^+ \in K$. The rationale of this norm stems from the fact that second-order Taylor-Young expansion of f at b^+ gives

$$D(b, b^+) = \tfrac{1}{2}\|b - b^+\|_{b^+}^2 + o(\|b - b^+\|^2), \quad D(b^+, a^+) = \tfrac{1}{2}\|b^+ - a^+\|_{b^+}^2 + o(\|b^+ - a^+\|^2).$$

Here the little-o terms may be made uniform on any compact set of b^+, see Section 2.5.

Let us fix some more terminology.

Definition 10.3. The proximal normal cone to B with regard to $\|\cdot\|_{b^+}$ is N_B^{p,b^+}. The orthogonal projector with regard to $\|\cdot\|_{b^+}$ is $P_B^{b^+}$, the angle is $\sphericalangle_{b^+}$, and $\|\cdot\|_{b^+}$-balls are $B_{b^+}(x, r)$. The reach of B at $b^+ \in B$ in direction $d \in N_B^{p,b^+}(b^+) \setminus \{0\}$ with regard to $\|\cdot\|_{b^+}$ is $\widetilde{R}(b^+, d)$.

This allows now the following

Definition 10.4. (Mobile reach). We say that B has mobile reach at least $\widetilde{R} > 0$ at $b^* \in B$, noted $\widetilde{R}(b^*) \geq \widetilde{R}$, if there exists a neighborhood U of b^* such that for every $b^+ \in P_B^{b^+}(a^+) \cap U$ for some $a^+ \notin B$ we have $\widetilde{R}(b^+, d) \geq \widetilde{R}$, where $d = \nabla^2 f(b^+)^{-1}(\nabla f(a^+) - \nabla f(b^+))$.

Remark 10.5. As seen in Sections 3.1 and 3.4, B has positive reach at $b^* \in B$ iff it has positive left Bregman reach at b^*. Using (10.3), this is now equivalent to having positive mobile reach.

Proposition 10.6. *Let $b^* \sim a^*$ with $\|\nabla f(b^*) - \nabla f(a^*)\|_{b^*} = \rho_*$. Suppose B has mobile reach at least $\widetilde{R} > \rho_*$ at b^*. Then there exist $\delta > 0$ and $0 < \ell < 1$ such that every building block $b \xrightarrow{r} a^+ \xrightarrow{l} b^+$ with $(b^+, a^+) \in B(b^*, \delta) \times B(a^*, \delta)$ and $\|b - b^+\| < \delta$ satisfies the rl-three-point inequality with ℓ.*

Proof. Fix $\ell \in (0, 1)$ such that $(1-\ell)^{-1}\rho_* < \widetilde{R}$. Then find $\epsilon > 0$ such that $(1+\epsilon)(1-\ell)^{-1}\rho_* < \widetilde{R}$. Writing $\widetilde{R} = (1+\epsilon)(1-\ell)^{-1}\widetilde{\rho}$ therefore means $\rho_* < \widetilde{\rho}$. Let U be a neighborhood of b^* as in Definition 10.4. Shrink U further until $(1+\epsilon)D(b, b^+) \geq \frac{1}{2}\|b - b^+\|_{b^+}^2$ for all $b, b^+ \in U$, using uniform second order Taylor-Young expansion at b^+ on U in tandem with (2.3). Ready to shrink U even further, combine it with a neighborhood V of a^* such that $(b^+, a^+) \in U \times V$ implies $\|\nabla f(b^+) - \nabla f(a^+)\|_{b^+} < \widetilde{\rho}$. This is possible, because $\rho_* < \widetilde{\rho}$, and since the norm $\|\cdot\|_{b^+}$ depends continuously on b^+.

Now let $b \xrightarrow{r} a^+ \xrightarrow{l} b^+$ be a building block with $b, b^+ \in U$ and $a^+ \in V$, and put $\bar{a} := b^+ + \nabla^2 f(b^+)^{-1}(\nabla f(a^+) - \nabla f(b^+))$. Then

$$\begin{aligned}\langle \nabla f(a^+) - \nabla f(b^+), b - b^+\rangle &= \langle \nabla^2 f(b^+)^{-1}(\nabla f(a^+) - \nabla f(b^+)), b - b^+\rangle_{b^+} \\ &= \langle \bar{a} - b^+, b - b^+\rangle_{b^+} \\ &= \|\bar{a} - b^+\|_{b^+}\|b - b^+\|_{b^+}\cos\beta,\end{aligned}$$

where $\beta = \sphericalangle_{b^+}(\bar{a} - b^+, b - b^+)$ is the angle in the euclidean geometry of $\|\cdot\|_{b^+}$. Now if $\cos\beta \leq 0$, the three-point inequality is trivially satisfied, so we may assume $\cos\beta > 0$. Then $\beta < 90°$, hence there exists a point $\hat{a}$ on the proximal normal to B at b^+ in the $\|\cdot\|_{b^+}$-geometry such that the triangle $b, \hat{a}, b^+$ is equilateral with two angles β at the corners b, b^+, and two edges of the same $\|\cdot\|_{b^+}$-length R joining b, b^+ to $\hat{a}$. Hence $\hat{a} = b^+ + Rd$, where $d = (\bar{a} - b^+)/\|\bar{a} - b^+\|_{b^+}$ and $R = \|b - \hat{a}\|_{b^+}$. Then $\|b - b^+\|_{b^+} = 2R\cos\beta$. But the ball $B_{b^+}(\hat{a}, R)$ contains b, b^+, hence $R \geq \widetilde{R}(b^+, d) \geq \widetilde{R} = (1+\epsilon)(1-\ell)^{-1}\widetilde{\rho}$. We deduce

$$\|b - b^+\|_{b^+} \geq 2(1+\epsilon)(1-\ell)^{-1}\widetilde{\rho}\cos\beta \geq 2(1+\epsilon)(1-\ell)^{-1}\|\nabla f(b^+) - \nabla f(a^+)\|_{b^+}\cos\beta.$$

Hence

$$\begin{aligned}(1-\ell)D(b, b^+) &\geq (1-\ell)(1+\epsilon)^{-1}\tfrac{1}{2}\|b - b^+\|_{b^+}^2 \\ &\geq (1-\ell)(1+\epsilon)^{-1}(1+\epsilon)(1-\ell)^{-1}\|\nabla f(b^+) - \nabla f(a^+)\|_{b^+}\|b - b^+\|_{b^+}\cos\beta \\ &= \langle \nabla f(a^+) - \nabla f(b^+), b - b^+\rangle.\end{aligned}$$

That proves the *rl*-three point inequality. □

Remark 10.7. While Proposition 10.1 needs 1-coercivity of f, which is not required here, we now need a nearness condition on b, b^+ to bring in the uniform Taylor-Young estimate, using that f is of class C^2. This seems acceptable in view of Lemma 6.6. We will see the consequences right below.

Theorem 10.8. (Global convergence). *Let a_k, b_k be a Bregman alternating sequence with gap (K^*, r^*). Suppose the rl-angle condition is satisfied at every $(b^*, a^*) \in K^*$. Then the sequence $b_k \in \bar{P}_B \circ \vec{P}_A(b_{k-1})$ converges under any of the following conditions:*

(i) *B has left Bregman reach $\bar{R}(b^*) > r^*$ for every $b^* \sim a^*$ in K^*, and f is 1-coercive.*
(ii) *B has mobile reach $\widetilde{R}(b^*) > \rho_* = \|\nabla f(b^*) - \nabla f(a^*)\|_{b^*}$ for all $b^* \sim a^*$ in K^*, and the sequence satisfies $b_{k-1} - b_k \to 0$.*

Proof. All we need is assure the *rl*-three point inequality for building blocks $b \xrightarrow{r} a^+ \xrightarrow{l} b^+$, as the rest is like in Theorem 7.1. We use Proposition 10.1 for the left Bregman reach case (i), and Proposition 10.6 in case (ii). For the latter, by compactness we have $\widetilde{R} > \max\{\|\nabla f(b^*) - \nabla f(a^*)\|_{b^*} : b^* \sim a^*\}$, so that the three point inequality holds for all gap pairs with the same ℓ. □

10.3. **Slowly vanishing reach.** We have seen that positive reach at some $b^* \in B$, or prox-regularity of B at b^*, could be expressed in four equivalent fashions, using $R, \bar{R}$, $\vec{R}$, and mobile reach $\widetilde{R}$. Unfortunately, exact quantitative relations among those four can only be obtained in the rough proportional sense of Section 3.1. Better quantitative results can be obtained for zero gaps. For those we may even allow B to have *slowly vanishing reach* at $b^* \in B$, a notion introduced in [52, 51] in the euclidean setting. As this increases the chances to establish the three-point inequality, we adapt this to the Bregman context. For an explanation of what is meant by vanishing reach see also Example 12.4.

The following preparatory result conveys the fact that asymptotically as $r \to 0$, left Bregman balls $\bar{\mathcal{B}}(a^+, r)$ with $b^+ \in \partial\bar{\mathcal{B}}(a^+, r)$ more and more resemble balls $B_{b^+}(\bar{a}, \|\bar{a} - b^+\|_{b^+})$ in the euclidean geometry $\langle x, y\rangle_{b^+} = \langle \nabla^2 f(b^+)x, y\rangle$, where $\bar{a} = b^+ + \nabla^2 f(b^+)^{-1}(\nabla f(a^+) - \nabla f(b^+))$.

Lemma 10.9. *Let $x^* \in A \cap B$. Then*

$$\liminf_{b^+ \in \bar{P}_B(a^+), A \ni a^+ \to x^*} \widetilde{R}(b^+, d) > 0 \text{ iff } \liminf_{b^+ \in \bar{P}_B(a^+), A \ni a^+ \to x^*} \bar{R}(b^+, d) > 0.$$

When both tend to zero, we have

$$\lim_{b^+ \in \bar{P}_B(a^+), A \ni a^+ \to x^*} \frac{\bar{R}(b^+, d)}{\widetilde{R}(b^+, d)} = 1. \tag{10.4}$$

Proof. Note that $\widetilde{R}(b^+, d)$ stays away from 0 iff $\bar{R}(b^+, d)$ stays away from 0 by the results of Section 3.1 in tandem with (10.3). Now suppose both reach terms shrink to 0. Let $\bar{R}(b^+, d)$ with $d = \nabla^2 f(b^+)^{-1}(\nabla f(a^+) - \nabla f(b^+))$ be realized at $\lambda > 1$ and radius $r_\lambda > 0$, with $\bar{\mathcal{B}}(a_\lambda, r_\lambda)$ the largest left Bregman ball having b^+ on its boundary and no point of B in its interior.

Working in the $\|\cdot\|_{b^+}$-geometry, uniform second-order Taylor-Young expansion of f reads

$$f(x + h) = f(x) + \langle \nabla^2 f(b^+)^{-1} \nabla f(x), h\rangle_{b^+} + \tfrac{1}{2}\langle h, \nabla^2 f(b^+)^{-1} \nabla^2 f(x) h\rangle_{b^+} + o(\|h\|^2),$$

hence the normal curvature of the left Bregman ball $\bar{\mathcal{B}}(a_\lambda, r_\lambda)$ at $x \in \partial\bar{\mathcal{B}}(a_\lambda, r_\lambda)$ in unit tangential direction v in the $\|\cdot\|_{b^+}$-geometry is

$$\kappa_{n,b^+}(x, v) = \frac{\langle v, \nabla^2 f(b^+)^{-1} \nabla^2 f(x) v\rangle_{b^+}}{\|\nabla^2 f(b^+)^{-1}(\nabla f(a_\lambda) - \nabla f(x))\|_{b^+}}.$$

Taylor-Young expansion gives $\nabla^2 f(b^+)^{-1}(\nabla f(a_\lambda) - \nabla f(x)) = a_\lambda - x + o(\|a_\lambda - b^+\|) + o(\|x - b^+\|)$. Since we assume $r_\lambda \to 0$ as the building block approaches x^*, we have $a_\lambda \to x^*$ and $x \to x^*$ for $x \in \partial\bar{\mathcal{B}}(a_\lambda, r_\lambda)$. Hence on a sufficiently small neighborhood U of x^*, $(1 - \epsilon)\|a_\lambda - x\|_{b^+} \le \|\nabla^2 f(b^+)^{-1}(\nabla f(a_\lambda) - \nabla f(x))\|_{b^+} \le (1 + \epsilon)\|a_\lambda - x\|_{b^+}$ for $x, a_\lambda, b^+ \in U$, using again that the little-o terms in the Taylor-Young expansion may be made uniform on a compact set of b^+ (Section 2.5).

Shrinking U further, we may arrange $\nabla^2 f(b^+)^{-1}\nabla^2 f(x) = I_d + E$ with $\|E\|_{b^+} \le \epsilon$ for $x, b^+ \in U$. Then

$$\frac{1-\epsilon}{(1+\epsilon)\|a_\lambda - x\|_{b^+}} \le \kappa_{n,b^+}(x,v) \le \frac{1+\epsilon}{(1-\epsilon)\|a_\lambda - x\|_{b^+}}$$

using $\|v\|_{b^+} = 1$. This means the constants of Proposition 3.2, applied in the $\|\cdot\|_{b^+}$-geometry, are $\underline{c} = \frac{1-\epsilon}{1+\epsilon}$ and $\bar{c} = \frac{1+\epsilon}{1-\epsilon}$. Since the $\|\cdot\|_{b^+}$-euclidean ball with radius $\bar{c}r_\lambda$ contains the Bregman ball and touches it from outside at b^+, this ball contains also b, hence its radius is larger than the $\|\cdot\|_{b^+}$-reach at b^+: $\frac{1+\epsilon}{1-\epsilon}\check{R}(b^+,d) = \frac{1+\epsilon}{1-\epsilon}r_\lambda \ge \widetilde{R}(b^+,d)$. On the other hand, the smaller $\|\cdot\|_{b^+}$-ball with radius $\underline{c}r_\lambda$ is contained in the Bregman ball and touches it at b^+ from inside, hence contains no points of B in its interior, whence $\frac{1-\epsilon}{1+\epsilon}r_\lambda \le \widetilde{R}(b^+,d)$. This shows that numerator and denominator in (10.4) agree asymptotically, which proves equality 1 in the limit. □

Definition 10.10. (Slowly vanishing reach). Let $x^* \in A \cap B$. We say that B has slowly vanishing reach with rate τ at x^* with regard to A if

$$\tau := \limsup_{b^+ \in \check{P}_B(a^+), A \ni a^+ \to x^*} \frac{D(b^+,a^+)^{1/2}}{\widetilde{R}(b^+,d)} < \frac{1}{\sqrt{2}}. \tag{10.5}$$

Replacing $\widetilde{R}(b^+,d)$ by $\check{R}(b^+,d)$ gives the same value τ by Lemma 10.9, hence an equivalent definition of slowly vanishing reach. We now see that $\widetilde{R}$ has the advantage over $\check{R}$ that it gives the same information at points $x^* \in A \cap B$, while based on a euclidean norm, the inconvenience being that it is a mobile one.

10.4. **Three point inequality from vanishing reach.** In this section we show that the three-point inequality holds in the vicinity of a zero gap $r^* = 0$ even when we allow the reach to shrink to zero, provided it shrinks slightly slower than the distance between the sets.

Proposition 10.11. *Let $x^* \in A \cap B$, and suppose B has slowly vanishing reach at x^* with rate $\tau < 1/\sqrt{2}$ with regard to A. Then there exist a neighborhood U of x^* such that every building block $b \xrightarrow{r} a^+ \xrightarrow{l} b^+$ with $b, a^+, b^+ \in U$ satisfies the rl-three-point inequality with ℓ as long as ℓ satisfies $\ell < 1 - \sqrt{2}\tau$.*

Proof. 1) The cosine theorem for Bregman distances gives

$$D(b,a^+) = D(b^+,a^+) + D(b,b^+) - \langle \nabla f(a^+) - \nabla f(b^+), b - b^+ \rangle.$$

Therefore the rl-three-point inequality holds with $0 < \ell < 1$ iff

$$(1-\ell)D(b,b^+) \ge \langle \nabla f(a^+) - \nabla f(b^+), b - b^+ \rangle.$$

This holds regardless of ℓ when $\langle \nabla f(a^+) - \nabla f(b^+), b - b^+ \rangle \le 0$. We therefore concentrate on the case $\langle \nabla f(a^+) - \nabla f(b^+), b - b^+ \rangle > 0$.

2) Let ℓ be as in the statement, and choose τ' such that $\tau < \tau' < \frac{1-\ell}{\sqrt{2}}$. Then choose $\epsilon > 0$ such that $\tau'(1+\epsilon)^4 < \frac{1-\ell}{\sqrt{2}}$. By the definition of τ we can find a neighborhood W of x^* such that

$$\frac{D(b^+,a^+)^{1/2}}{\widetilde{R}(b^+,d)} \le \tau' \tag{10.6}$$

for all $a^+ \xrightarrow{r} b^+$ with $b^+, a^+ \in W$.

We put $\bar{a} = b^+ + \nabla^2 f(b^+)^{-1}(\nabla f(a^+) - \nabla f(b^+))$. Shrinking W further such that b, b^+ are forced sufficiently close, using uniform Taylor-Young expansion $D(b,b^+) = \frac{1}{2}\|b - b^+\|^2_{b^+} + o(\|b-b^+\|^2)$ in tandem with (2.3), we may arrange the following:

$$\begin{aligned} &\text{(a)} \quad \tfrac{1}{2}\|b - b^+\|^2_{b^+} \le (1+\epsilon)^2 D(b,b^+), \\ &\text{(b)} \quad \tfrac{1}{2}\|b^+ - a^+\|^2_{b^+} \le (1+\epsilon)^2 D(b^+,a^+) \\ &\text{(c)} \quad \|b^+ - \bar{a}\|_{b^+} \le (1+\epsilon)\|b^+ - a^+\|_{b^+} \end{aligned} \tag{10.7}$$

for all building blocks with $a^+, b^+, \bar{a}, b \in W$. Let the angle $\beta = \sphericalangle_{b^+}(b - b^+, \bar{a} - b^+)$ be taken with regard to the $\|\cdot\|_{b^+}$-geometry. Then,

$$\begin{aligned}
\langle \nabla f(a^+) - \nabla f(b^+), b - b^+ \rangle &= \langle \nabla^2 f(b^+)^{-1}(\nabla f(a^+) - \nabla f(b^+)), b - b^+ \rangle_{b^+} \\
&= \|\nabla^2 f(b^+)^{-1}(\nabla f(a^+) - \nabla f(b^+))\|_{b^+} \|b - b^+\|_{b^+} \cos\beta \\
&= \|\bar{a} - b^+\|_{b^+} \|b - b^+\|_{b^+} \cos\beta \\
&\leq (1+\epsilon)\|b^+ - a^+\|_{b^+} \|b - b^+\|_{b^+} \cos\beta \\
&\leq (1+\epsilon)^2 \|b^+ - a^+\|_{b^+} \sqrt{2} D(b, b^+)^{1/2} \cos\beta \\
&\leq (1+\epsilon)^3 \sqrt{2} D(b^+, a^+)^{1/2} \sqrt{2} D(b, b^+)^{1/2} \cos\beta \\
&\leq (1+\epsilon)^3 \tau' 2 \widetilde{R}(b^+, d) D(b, b^+)^{1/2} \cos\beta.
\end{aligned} \tag{10.8}$$

These estimates use $\cos\beta \geq 0$ throughout, which holds due to part 1). Moreover, line four uses (10.7) (c), line five uses (10.7) (a), line six uses (10.7) (b), and the last line uses (10.6).

3) Now recall that $\bar{a} - b^+$ is a proximal normal to the set B at b^+ with regard to the $\|\cdot\|_{b^+}$-geometry, with $d = (\bar{a} - b^+)/\|\bar{a} - b^+\|_{b^+}$ the corresponding unit proximal normal. Since $\beta = \sphericalangle_{b^+}(b - b^+, \bar{a} - b^+) < 90°$ according to part 1), we can choose $R > 0$ such that the point $\hat{a} = b^+ + Rd$ on the proximal normal satisfies $\|\hat{a} - b^+\|_{b^+} = \|\hat{a} - b\|_{b^+} = R$, so that $b, \hat{a}, b^+$ form an equilateral triangle with two angles β adjacent to the side b, b^+ of length $\|b - b^+\|_{b^+}$, and two sides of equal length R joining $\hat{a}$. Therefore $\frac{1}{2}\|b - b^+\|_{b^+} = R\cos\beta$ by the perpendicular bisector theorem.

Now the $\|\cdot\|_{b^+}$-euclidean ball with center $\hat{a}$ and radius R contains the points $b, b^+ \in B$ on its boundary. By definition of the reach of B at b^+ with regard to the norm $\|\cdot\|_{b^+}$, and since the ball in question has its center on the $\|\cdot\|_{b^+}$-normal $b^+ + \mathbb{R}_+ d$, this means that R must be at least as large as the $\|\cdot\|_{b^+}$-reach $\widetilde{R}(b^+, d)$ in that direction. We derive

$$\|b - b^+\|_{b^+} = 2R\cos\beta \geq 2\widetilde{R}(b^+, d)\cos\beta. \tag{10.9}$$

Plugging this into (10.8) gives

$$\begin{aligned}
\langle \nabla f(a^+) - \nabla f(b^+), b - b^+ \rangle &\leq (1+\epsilon)^3 \tau' \|b - b^+\|_{b^+} D(b, b^+)^{1/2} \\
&\leq (1+\epsilon)^4 \tau' \sqrt{2} D(b, b^+) \\
&< (1 - \ell) D(b, b^+)
\end{aligned}$$

by the choice of ϵ and τ'. Therefore by the cosine theorem for Bregman distances

$$\begin{aligned}
D(b, a^+) &= D(b^+, a^+) + D(b, b^+) - \langle \nabla f(a^+) - \nabla f(b^+), b - b^+ \rangle \\
&\geq D(b^+, a^+) + D(b, b^+) - (1 - \ell) D(b, b^+) \\
&= D(b^+, a^+) + \ell D(b, b^+).
\end{aligned}$$

□

Remark 10.12. We see from the argument in part 3), and also from a similar one in Proposition 10.6, that the technique in part 2) of the proof of Proposition 10.1 looks like a Bregman version of the euclidean perpendicular bisector theorem.

10.5. **Convexity.** As we had seen in Proposition 6.4, convexity of B, or $\nabla f(A)$, makes things easier, but surprisingly, convexity of A doesn't seem to help. This discrepancy was also observed in [18, Thm. 7.3], where the right Bregman Chebyshev condition of A was shown to imply convexity of $\nabla f(A)$, not of A. For short, convexity is not amenable to duality. Can we still get something when A is convex?

Proposition 10.13. *Suppose A is closed bounded contained in* $\mathrm{int}(\mathrm{dom} f)$ *and has positive reach. Suppose f is of class $C^{2,1}$. Then there exists $r > 0$ such that the lr-three-point inequality is satisfied at all gaps with gap value $r^* \leq r$.*

Proof. Since A has positive reach and ∇f is a $C^{1,1}$-diffeomorphism, the image $B^* = \nabla f(A)$ has also positive reach. Since positive reach is equivalent to positive mobile reach, we have $\widetilde{R}(b^{*+}) \geq \widetilde{R}$ for some $\widetilde{R} > 0$ for the mobile reach in dual space and all $b^{*+} \in B^*$. Therefore, applying Proposition

10.6 in dual space, we get the rl-three-point inequality for dual building blocks $b^* \xrightarrow{r*} a^{*+} \xrightarrow{l*} b^{*+}$ near gaps with dual gap distance $\rho_* = \|\nabla f^*(b^{*+}) - \nabla f^*(a^{*+})\|_{b^{*+}} < \widetilde{R}$.

Going backwards in the dual formula (2.2), this means every lr-building block $a \xrightarrow{l} b \xrightarrow{r} a^+$ satisfies the lr-three-point inequality as soon as $\rho_* = \|\nabla f^*(b^{*+}) - \nabla f^*(a^{*+})\|_{b^{*+}} = \|a^+ - b\|_{b^{*+}} < \widetilde{R}$. Now $\|u\|^2_{b^{*+}} = \langle \nabla^2 f^*(b^{*+})u, u\rangle = \langle \nabla^2 f(a^+)^{-1}u, u\rangle$. Hence in primal space the sufficient condition reads $\rho_* = \langle \nabla^2 f(a^+)^{-1}(a^+ - b), a^+ - b\rangle^{1/2} < \widetilde{R}$. Since $\nabla^2 f(a^+)^{-1} \succeq \epsilon > 0$ on A, we have an estimate of the form $m' D(b, a^+)^{1/2} \leq \|a^+ - b\|_{b^{*+}}$ similar to (10.3) combined with (2.3) with the same m' for A, B, so that the lr-three-point inequality now holds for building blocks with $D(b, a^+)^{1/2} < m'^{-1}\widetilde{R}$, hence it holds for gaps with $r^* \leq r := m'^{-1}\widetilde{R}$. □

This shows that it is prox-regularity, or positive reach, which is amenable to duality, not convexity. However, the distortion caused by ∇f makes it hard to relate the reach of a set A in primal space to the reach of its ∇f-image B^* in dual space. Quantifying r could at best be achieved in specific situations.

Corollary 10.14. *Let $\bar{x} \in A \cap B$ and suppose A is prox-regular at $\bar{x}$. Let A, B, f be definable, and suppose f is of class $C^{2,1}$. Then there exists a neighborhood V of $\bar{x}$ such that every Bregman alternating sequence which enters V converges to a point in $A \cap B$.*

This should be compared with Corollary 7.4.

11. EM algorithm

We apply our convergence theory to the EM algorithm via the variant in [32], known as the *em*-algorithm. Criteria for the two to coincide are given in [2, Thm. 4, Ex. 10, §7.2]. As observed in [60], or [47, Sect 3.6], even when convergent, EM iterates may go to local minima or saddle points of the likelihood, which is not surprising in a non-convex setting. In this work, we focus on convergence of the iterates, which contains value convergence first discussed in [60], see also [40]. As literature on convergence of the EM algorithm is vast, we confine ourselves to a few pointers [32, 31, 2, 3, 57, 24, 26, 28, 29, 38, 40, 60]. En excellent survey up to 2008 is [47].

11.1. **Exposition of the method.** The *em*-algorithm requires a complete data space X, an incomplete data space Y, a measurable mapping $T : X \to Y$, a family of probability measures P on X and their images P^T under T on Y. Assuming $P \ll \mu$ for a σ-finite base measure on X, and $P^T \ll \mu^T$, we have densities $p = \frac{dP}{d\mu}$ and $p^T = \frac{dP^T}{d\mu^T}$. We further require an empirical distribution $\hat{P}$ on Y with $\hat{P} \ll \mu^T$, $\hat{p} = \frac{d\hat{P}}{d\mu^T}$, which we use to define the data set on X as $\mathcal{D} = \{P : P^T = \hat{P}\}$, respectively, $D = \{p : p^T = \hat{p}\}$. Finally, we need a statistical model $\mathcal{M}$ of distributions $Q \ll \mu$ on X, $M = \{q : Q \in \mathcal{M}\}$, and our goal is to minimize the Kullback-Leibler information distance [41] between D and M, given by

$$K(p||q) = \int_X p(x) \log \frac{p(x)}{q(x)} d\mu(x).$$

Then the *em*-algorithm is the following alternating procedure:

$$\begin{aligned} e\text{-step} \qquad & p \in \operatorname*{argmin}_{p' \in D} K(p'||q) \\ m\text{-step} \qquad & q^+ \in \operatorname*{argmin}_{q' \in M} K(p||q') \end{aligned}$$

The *m*-step is similar to the M step of the EM algorithm and represents maximum likelihood estimation in complete data space. On the other hand, the *e*-step may differ from the E step, as shown in [2, 6.3], and according to [2, Thm. 4], both agree iff the conditional expectation with respect to a candidate distribution of the missing data, given the observed data, is an affine function of observed data.

It turns out that the *e*-step is explicit. As follows from [42, Thm. 4.1], we have $K(p||q) \geq K(p^T||q^T)$, with equality iff $\frac{p(x)}{q(x)} = \frac{p^T(T(x))}{q^T(T(x))}$ μ-a.e. Applying this to the data set D shows that $\tilde{P}_D(q) = p$ is realized by setting $p(x) = \hat{p}(T(x))\frac{q(x)}{q^T(T(x))}$ μ-a.e.

11.2. **Discrete measures.** In the first place, let us assume $X = I, Y = J$ finite, with μ the counting measure. Then distributions on I are vectors $p = (p_i)_{i\in I}$ with $p_i \geq 0$, $\sum_{i\in I} p_i = 1$, and similarly, $\hat{p} = (\hat{p}_j)_{j\in J}$ with $\hat{p}_j \geq 0$, $\sum_{j\in J} \hat{p}_j = 1$. The Kullback-Leibler divergence on $\mathbb{R}^I_+$

$$K(p||q) = \sum_{i\in I} p_i \log \frac{p_i}{q_i} - p_i + q_i,$$

is now the Bregman divergence generated by $f(x) = \sum_{i\in I} x_i \log x_i - x_i$, making the algorithm amenable to our convergence theory. We assume $\hat{p}_j > 0$ for all $j \in J$ and observe that the data set is $D = \Big\{p \geq 0 : \sum_{i\in I} p_i = 1, \sum_{T(i)=j} p_i = \hat{p}_j \ \forall j \in J\Big\}$.

Theorem 11.1. *Let $M \subset \mathbb{R}^I_{++}$ be a closed definable set of statistical model distributions. Let p^k, q^k be a sequence generated by the em-algorithm. Then the p^k converge to a distribution $p^* \in D$. Every accumulation point $q^* \in M$ of the q^k satisfies*

$$p^*_i = \hat{p}_j \frac{q^*_i}{\sum_{T(i')=j} q^*_{i'}}, \quad i \in T^{-1}(j), j \in J. \tag{11.1}$$

When M is definable in $\mathbb{R}_{an}$, then the speed of convergence is no worse than $\|p^k - p^*\| = O(k^{-\rho})$ for some $\rho \in (0, \infty)$.

Proof. We have $p \in \tilde{P}_D(q)$ and $q^+ \in \vec{P}_M(p)$, hence $M = A$ and $D = B$ in our general scheme. Since D is convex and not entirely contained in the boundary of $\mathbb{R}^I_+$, it is interiority preserving, hence can be pre-processed as in Section 2.2. Therefore lr-building blocks $q \xrightarrow[e]{l} p \xrightarrow[m]{r} q^+$ satisfy the three-point inequality by Proposition 6.4. Moreover, D is semi-algebraic, $K(p||q)$ is definable because log is definable in $\mathbb{R}_{an,\exp}$, and M is definable by hypothesis. Therefore the lr-angle condition holds, and convergence of the p^k follows with Theorem 8.1.

Next observe that the e-step can be made explicit due to the structure of D and the observation made above. We have

$$p_i = \hat{p}_j \frac{q_i}{\sum_{T(i')=j} q_{i'}}, \quad i \in T^{-1}(j), j \in J$$

for $q \xrightarrow{l} p$, and we may clearly pass to the limit in this expression, which gives (11.1).

Finally, assume M is definable in $\mathbb{R}_{an}$. Since D is semi-algebraic, hence definable in $\mathbb{R}_{an}$, it remains to show definability of $K(p||q)$ in $\mathbb{R}_{an}$. Now observe that by $M \subset (0, 1]^I$ and closedness of M the q^k_i stay bounded away from 0, i.e. $q^k_i \geq \epsilon > 0$ for all i, k. Since $\hat{p}_j > 0$ for all $j \in J$, formula (11.1) shows that all p^k_i also stay uniformly bounded away from 0, say $p^k_i \geq \delta > 0$ for all i, k. Bearing in mind that $p^k_i \leq 1$, the logarithms $\log p_i$ in $K(p||q)$ can a priori be restricted to the interval $[\delta, 1]$. But the restriction $\log \restriction [\delta, 1]$ is globally sub-analytic, cf. [34, 35]. In consequence the KŁ-inequality becomes a Łojasiewicz inequality with de-singularizing function $\phi'(s) = s^{-\theta}$ for some $\theta \in [\frac{1}{2}, 1)$. Therefore Corollary 9.1 gives the claimed rate of convergence. □

Remark 11.2. In the feasible case $D \cap M$ has a neighborhood of attraction W such that any sequence p^k, q^k entering W converges to some $p^* \in D \cap M$ with the same rate $\|q^k - p^*\| = O(k^{-\rho})$, $\|p^k - p^*\| = O(k^{-\rho})$. Recall from Corollary 9.1 that ρ can be expressed in terms of the Lyapunov exponent θ of F, hence in some cases an even more specific rate may be obtainable.

Convergence of the q^k with the same rate also occurs for gaps > 0 when the left Bregman reach of $A = M$ at the q^k is larger than the gap value, as $\tilde{P}_M$ is then locally Lipschitz.

Note that (11.1) says $p^k = \mathbb{E}_{q^k}(p|\hat{p})$, i.e., p^k is the conditional expectation of p given $\hat{p}$ with regard to the probability distribution q^k. In the limit we then have $p^* = \mathbb{E}_{q^*}(p|\hat{p})$ for all accumulation points q^* of the q^k. Since definability is not a severe restriction, this is quite useful in practice.

11.3. **Exponential family.** We consider an exponential family of densities with respect to a base measure dx

$$p_\theta(x) = \exp(\langle\theta, t(x)\rangle - f(\theta) + k(x)), \tag{11.2}$$

where $t(x)$ is the sufficient statistic, θ the natural parameter varying in a parameter set Θ, $f(\theta)$ the log-normalizer function, and $\exp k(x)$ the carrier measure density.

Lemma 11.3. *The Kullback-Leibler divergence of two distributions $p_\theta(x)$ and $p_{\theta'}(x)$ belonging to the same exponential family is $K(p_{\theta'}||p_\theta) = D_f(\theta, \theta')$, where D_f is the Bregman divergence induced by the log-normalizer f.*

Proof. From (11.2), and since $\int_X p_\theta(x)dx = 1$, we have

$$f(\theta) = \log \int_X \exp\{\langle \theta, t(x)\rangle + k(x)\}dx.$$

Differentiation with respect to θ gives

$$\nabla f(\theta) = \int_X t(x) \exp\{\langle \theta, t(x)\rangle + k(x)\}dx \Big/ \int_X \exp\{\langle \theta, t(x)\rangle + k(x)\}dx.$$

Now $\exp f(\theta) = \int_X \exp\{\langle \theta, t(x)\rangle + k(x)\}dx$, hence $\nabla f(\theta) = \int_X t(x) \exp\{\langle \theta, t(x)\rangle - f(\theta) + k(x)\}dx = \int_X t(x)p_\theta(x)dx = E_\theta[t(x)]$, the expectation of the random variable $t(x)$ with respect to the distribution $p_\theta dx$ (see also [47, (1.57)]). Then

$$\begin{aligned} K(p_{\theta'}||p_\theta) &= \int p_{\theta'}(x) \log \frac{p_{\theta'}(x)}{p_\theta(x)}\, dx \\ &= \int_X p_{\theta'}(x) \left(f(\theta) - f(\theta') + \langle \theta' - \theta, t(x)\rangle\right) dx \\ &= \int_X p_{\theta'}(x) \left(D_f(\theta, \theta') + \langle \theta - \theta', \nabla f(\theta')\rangle + \langle \theta' - \theta, t(x)\rangle\right) dx \qquad (11.3) \\ &= D_f(\theta, \theta') + \int_X p_{\theta'}(x) \langle \theta - \theta', \nabla f(\theta') - t(x)\rangle dx \\ &= D_f(\theta, \theta') + \langle \theta - \theta', \nabla f(\theta') - E_{\theta'}[t(x)]\rangle \\ &= D_f(\theta, \theta'). \end{aligned}$$

□

In order to connect with our general set-up, we have to make sure that f is Legendre. We have the

Definition 11.4. The exponential family is called *steep* if its log-normalizer f is of Legendre type.

An exponential family is *regular* if the natural parameter space Θ is open. It is known that regular exponential families are steep, but the steep class is larger; cf. [24, 7]. We are now ready to concretize the *em*-algorithm for exponential families, specifying model set $\mathcal{M} = \{p_\theta : \theta \in M\}$ and data set $\mathcal{D} = \{p_\theta : \theta \in D\}$ by their parameter representatives $M, D \subset \Theta$.

Algorithm *em*-algorithm for exponential family

- **Step 1** (*e*-`step`)**.** Given current model density p_θ, $\theta \in M$, complete data with the help of the sample via $\theta' \in \vec{P}_D(\theta)$. Obtain completed data density $p_{\theta'}$, $\theta' \in D$.
- **Step 2** (*m*-`step`)**.** Given complete data density $p_{\theta'}$, $\theta' \in D$, improve parameter estimate by maximum likelihood in complete data space via $\theta \in \overleftarrow{P}_M(\theta')$. Back to step 1.

We observe that due to (11.3) left and right projections have been swapped, and the algorithm has now the form $\theta' \xrightarrow[m]{l} \theta \xrightarrow[e]{r} \theta'^{+} \xrightarrow[m]{l} \theta^{+}$, which matches (1.4) with $M = B$, $D = A$, $a = \theta'$, $b = \theta$, $a^+ = \theta'^+$, $b^+ = \theta^+$.

It remains to say a bit more about the data set. Following [2], we consider the expectation parameter $\eta(\theta) = E_\theta[t(x)]$, which by Lemma 11.3 is $\eta = \nabla f(\theta)$, with inverse $\theta = \nabla f^*(\eta)$. When the sufficient statistic is of the form $t(x) = (t_1(x), t_2(x))$ for observed $y = t_1(x)$ and hidden $z = t_2(x)$, then

$$D = \{\theta \in G : E_\theta[t_1(x)] = \hat{y}\}, \quad \mathcal{D} = \{p_\theta : \theta \in D\}, \qquad (11.4)$$

where $\hat{y}$ is the available sample. We can partition $\theta = (\theta_1, \theta_2)$ accordingly, so that $\langle \theta, \eta\rangle = \langle \theta_1, \eta_1\rangle + \langle \theta_2, \eta_2\rangle$, then $\nabla f(D) = \{(\eta_1, \eta_2) \in G^* : \eta_1 = \hat{y}, \eta_2 \text{ free}\}$, which in expectation coordinates is an affine subspace L^*, intersected with $\mathrm{dom}\nabla f^* = G^*$.

Theorem 11.5. *Suppose the log-normalizer $f(\theta)$ of the steep exponential family and the model parameter set M are definable. Suppose $M \subset G$ is closed bounded and D has the structure* (11.4)*. Then the em-algorithm $\theta'^{+} = \vec{P}_D \circ \overleftarrow{P}_M(\theta')$ converges.*

Proof. 1) By hypothesis M is closed bounded and contained in G, so its image $A^* := \nabla f(M)$ is closed bounded and contained in G^*. Since $D \subset G$, we have $B^* := \nabla f(D) \subset G^*$, and by the above $B^* = L^* \cap G^*$. This means $B^* = \nabla f(D)$ is convex, while not necessarily closed.

2) Put $B_1^* = L^* \cap \operatorname{cl} G^*$, then B_1^* is closed and satisfies the constraint qualification $B_1^* \cap G^* = B^* \neq \emptyset$. Therefore B_1^* is interiority preserving (Section 2.2, [8, Thm. 3.12]), hence $\overleftarrow{P}^*_{B_1^*}(y^*)$ is defined for $y^* \in G^*$ and we have $\overleftarrow{P}^*_{B_1^*}(y^*) \subset B_1^* \cap G^* = B^*$. This little detour shows that $\overleftarrow{P}^*_{B^*}(y^*) \neq \emptyset$ is well-defined for $y^* \in G^*$.

3) By compactness of $A^* \subset G^*$ dual right projections on A^* are well defined, and by 2) dual left projections from A^* are also well-defined and, moreover, go to B^*. Now consider $\theta' \in D$, then $\theta \in \overleftarrow{P}_M(\theta')$ is well defined by compactness of $M \subset G$, i.e. $\theta' \xrightarrow{l} \theta$ is well defined in primal space. Hence $\nabla f(\theta') \xrightarrow{r*} \nabla f(\theta)$ is well defined in G^*. But by what we had just seen we can now continue left projecting from $\nabla f(\theta) \in A^*$ into $B^* \subset G^*$, hence the dual rl-building block $\nabla f(\theta') \xrightarrow{r*} \nabla f(\theta) \xrightarrow{l*} \nabla f(\theta'^{+})$ is well defined and lies in G^*. By duality backward, the primal lr-building block $\theta' \xrightarrow{l} \theta \xrightarrow{r} \theta'^{+}$ is now also well defined and lies in G. Iterating this, the entire primal sequence is well-defined and lies in G, and its mirror sequence lies in G^*. It remains to show that the same holds for the accumulation points of these sequences.

4) As B_1^* is interiority preserving, we may apply Section 2.2 to the dual alternating sequence, which means there exists a closed bounded subset C^* of $B_1^* \cap G^* = B^*$ such that left projections of iterates from A^* go to C^*. In consequence the dual sequence alternates between C^* and A^*, now including accumulation points. Mapping this back via ∇f^* means the primal sequence alternates between $M = \nabla f^*(A^*)$ and the compact $C := \nabla f^*(C^*) \subset D$, and accumulation points belong to M, respectively, C. This makes the situation amenable to our convergence theory.

5) For that it remains to establish the angle condition. Since f, M are definable by hypothesis, it remains to check definability of C. Now observe that G as the interior of the domain of f is definable. Definability of ∇f also follows from definability of f. Therefore $G^* = \nabla f(G)$ as the image of a definable set under a definable diffeomorphism is definable (see [30]). Hence $B^* = L^* \cap G^*$ is definable, L^* being algebraic. Now recall that the construction in Section 2.2 gives a definable C^* because B^* is definable, and hence $C = \nabla f^*(C^*)$ is also definable, using that $\nabla f^* = (\nabla f)^{-1}$ is also a definable diffeomorphism. This means the lr-angle condition is satisfied.

Now all the hypotheses of Theorem 8.1 are satisfied, which gives convergence of the primal lr-sequence $\theta'^{+} = \vec{P}_D \circ \overleftarrow{P}_M(\theta')$. □

Remark 11.6. 1) The argument hinges on M being bounded, which is not always true in practice, but some boundedness hypothesis is required (see e.g. [60, (6)]), because in the infeasible case even euclidean alternating projections between unbounded convex sets may escape to infinity, and without convexity, this may happen even in the feasible case. When the sequence θ_k, θ'_k is bounded and the θ'_k stay away from ∂G, we may always select a closed bounded subset $M_0 \subset G$ of the model set M such that a_k, b_k remain alternating between D, M_0, and then in a second step use the treatment of Section 2.2 to get a bounded $D_0 \subset G$.

2) The result holds more generally when $y = t_1(x)$ is affine in $z = t_2(x)$, $y = Az + b$. Let $A \in \mathbb{R}^{n \times m}$ of maximal rank $m < n$, find Q invertible $n \times n$ with $A = [\widetilde{A}\ 0]\, Q$ and $\widetilde{A}$ regular of size $m \times m$, and make the change of coordinates $\tilde{z} = \operatorname{diag}(\widetilde{A}, I_{n-m})Qz =: Tz$, then $\tilde{z} = (\tilde{y}, \tilde{v})$ with $\tilde{y} = \pi_1(\tilde{z}) = Az$ and $y = \tilde{y} + b$, which reduces the affine case to (11.4).

3) An intriguing question is whether EM, respectively *em*, may fail to converge and generate a continuum of accumulation points. A natural place to look are euclidean alternating projections, as those arise when estimating the mean of a Gaussian with known variance. Counterexamples for AP with a continuum of accumulation points have been given in [14, 15], but do not apply to EM, because in these examples the structure of the set D playing the role of the data set is too exotic. We sketch a possible counterexample in Section 12.

4) Inspecting lists of exponential families, one finds that log-normalizers f often feature terms $\log \theta_i$ for components of the natural parameter θ. All cases we are aware of are governed by the

o-minimal structure $\mathbb{R}_{an,\exp}$ unifying globally sub-analytic sets with exponential and logarithm; cf. [34, 33, 59]. Yet there is interest to arrange model parameters $\theta \in M$ such that in these $\log \theta_i$-terms the θ_i may *a priori* be bounded on some $[\underline{\theta}, \overline{\theta}]$ with $0 < \underline{\theta}, \overline{\theta} < \infty$. Namely, by [34, 35], this has the benefit that f will be definable in $\mathbb{R}_{an}$. Since D is by default definable in $\mathbb{R}_{an}$, things depend on M. When M is also definable in $\mathbb{R}_{an}$, we get de-singularizing functions $\phi(s) = s^\alpha$ for some $\alpha \in [\frac{1}{2}, 1)$, which by Corollary 9.1 allows to quantify the speed of convergence to some $O(k^{-\rho})$.

5) Instances of sublinear speed of EM are mentioned e.g. in [47, p. 102], even though the general understanding seems to be that EM should converge linearly. However, it may be extremely hard to predict linear convergence a priori. For instance, even in the feasible case $p^* \in D \cap M$, and despite the convenient structure of D, we would have to show that D, M intersect transversally at p^* prior to coming to know p^*. This is possible only in very specific situations. Realistically, we should therefore only claim a rate $O(k^{-\rho})$. That may be predicted credibly, as definability of D, M, f is usually easy to check. A fair chance to prove transversality might be the case (11.4) when M has a simple structure.

11.4. **dSPECT imaging.** We end with an application in dynamic SPECT (dSPECT) imaging [17]. Voxels $i \in I$ have unknown activity x_{ik} varying in time t_k, $k \in K$. Camera bin $j \in J$ receives y_{jk} counts at time t_k at angular position α_k, where $\mathbb{E}(y_{jk}) = \sum_{i\in I} c_{ijk}x_{ik}$, and the known coefficients c_{ijk} reflect camera and collimator geometry. Complete data z are activities z_{ijk} emitted from voxel i at time t_k to camera bin j in position α_k, $\mathbb{E}(z_{ijk}) = c_{ijk}x_{ik}$. The x_{ik} are the unknown parameters. Two laws have been discussed in [17], Poisson, and Gaussians with known variance. The data set is $D = \{z : \sum_{i\in I} z_{ijk} = y_{jk} \forall j \in J, k \in K\}$. Since the problem is underdetermined, prior information on the time behavior of the x_{ik} is added. In [17] the authors use $x_{ik} = a_i e^{-\lambda_i t_k} + b_i e^{-\mu_i t_k} + d_i$, and this defines a model $M_e = \{v : v_{ijk} = c_{ijk}x_{ik}(a_i, b_i, d_i, \lambda_i, \mu_i)\}$ with $5|I|$ parameters. The variant in [46] uses a Prony model $x_{ik} = \alpha_i x_{i,k-2} + \beta_i x_{i,k-1} + \gamma_i$ with $M_p = \{v : v_{ijk} = c_{ijk}x_{ik}(\alpha_i, \beta_i, \gamma_i, x_i^0, x_i^1)\}$, where the $5|I|$ parameters are $\alpha_i, \beta_i, \gamma_i$ and two initial values x_i^0, x_i^1 per voxel. Clearly M_p is definable in $\mathbb{R}_{an}$, M_e in $\mathbb{R}_{an,\exp}$, so that convergence of the E step sequence for both cases is assured, giving convergence of the corresponding x. In the feasible case the M sequence converges as well. For the Prony model M_p the rate is $O(k^{-\rho})$, for M_e this holds when the exponentials can be restricted to a compact interval. For the Gaussian case convergence could be obtained from [51, 52], while for the Poisson model convergence is now for the first time established here.

12. Examples

Example 12.1. We consider the Kullback-Leibler divergence in $\mathbb{R}^2$, $K(x||y) = \sum_{i=1}^2 x_i \ln(x_i/y_i) - x_i + y_i$, with $0 \ln 0 = 0$. Take $\bar{y} = (1, 1)$, $\bar{x} = (1, 0)$ then $K(\bar{x}||\bar{y}) = K((1, 0)||(1, 1)) = 1 = \frac{1}{2}r^2$ with $r = \sqrt{2}$. The Bregman ball $\bar{\mathcal{B}}(\bar{y}, \sqrt{2})$ contains $(1, 0)$, but no other point $(z, 0)$, $z \neq 1$, because $K((z, 0)||(1, 1)) = z \log z - z + 2 \overset{!}{=} 1$ has only the solution $z = 1$, while $z \log z - z + 2 > 1$ for $z \neq 1$. Hence the line $x_2 = 0$ is tangent (a support hyperplane) to $\bar{\mathcal{B}}(\bar{y}, \sqrt{2})$ at $(1, 0)$.

We now squeeze a curve B in between $x_2 = 0$ and $\partial\bar{\mathcal{B}}(\bar{y}, \sqrt{2})$ in such a way that $B \cap \{x_2 = 0\} = \{\bar{x}\} = B \cap \bar{\mathcal{B}}(\bar{y}, \sqrt{2})$. Then $B \cap G \neq \emptyset$, but $\bar{P}_B(\bar{y}) = \bar{x} \notin G = \mathbb{R}^2_{++}$. Make the ansatz $B = \{(z, f(z)) : 1 - \epsilon \leq z \leq 1 + \epsilon\}$ with $f(1) = 1$, then $K((z, f(z))||(1, 1)) = z \log z - z + f(z) \log f(z) - f(z) + 2 > 1$ when we arrange $f(z)$ such that $f(z) \log f(z) - f(z) \leq \frac{1}{2}[-1 - z \log z + z]$ on $(1 - \epsilon, 1 + \epsilon)$.

Example 12.2. Failure of convergence of euclidean alternating projections with at least one of A, B non-convex are given e.g. in [14, 15, 52, 51]. In those cases a_k, b_k are bounded with $a_{k-1} - a_k \to 0$, $b_{k-1} - b_k \to 0$, but fail to converge because either angle condition or regularity fail, while the other holds, producing a continuum of accumulation points. For pictures see [14, 15].

Example 12.3. We sketch an example of failure of convergence of a bounded alternating sequence with a continuum of accumulation points in the feasible case $A \cap B \neq \emptyset$, where one of the sets is affine. This could be re-organized to give EM for estimating the mean of a gaussian with known variance with a non-convex curved parameter set.

Let $A = \{(x, 0) : x \in \mathbb{R}^n\}$ and $B = \{(x, f(x)) : x \in \mathbb{R}^n\}$ the graph of a C^1-function $f : \mathbb{R}^n \to \mathbb{R}$ with $f(x) > 0$ on $|x| < 1$, $f(x) = 0$ on $|x| = 1$. Take euclidean alternating projections, then $P_A(x_k, f(x_k)) = (x_k, 0)$, while $(x_k, f(x_k)) \in P_B(x_{k-1}, 0)$ iff $x_{k-1} = x_k + f(x_k)\nabla f(x_k)$. This method follows steepest ascent backwards.

Taking $n = 2$, we let B the graph of the mexican hat function [1] on $x_1^2 + x_2^2 \leq 1$, or likewise, its epigraph. Similar to the argument given for steepest descent with infinitesimal steps in [1], AP with infinitesimal steps will also follow the valley of the hat downward, endlessly circling around and approaching the boundary curve $x_1^2 + x_2^2 = 1$, where $f = 0$. Since the stepsize $f(x_k)$ goes in fact to 0, this argument is plausible. What is amiss for convergence is the angle condition, the graph of the mexican hat failing KŁ, while regularity is guaranteed since A is a plane. For a picture see [1].

Example 12.4. We explain vanishing reach in the euclidean case. Let $B = \{(x, |x|^{3/2}) : x \in \mathbb{R}\}$, then B has vanishing reach at the origin in direction $d = (0, 1)$. As shown in [51], the radius R_x of the largest ball touching B at $b = (x, |x|^{3/2})$ from above is of the order $R_x = O(|x|^{1/2})$ as $x \to 0$. In particular, the point $(0, 0) \in B$ cannot be projected on from above, while all other $(x, |x|^{3/2})$ can. For more details on this example see [52].

Example 12.5. In [2, Ex. 10] the author presents the case of a curved exponential family where EM and *em*-algorithms converge to different limit points, even though these agree asymptotically for large sample sizes. The *em* algorithm converges to a point in $D \cap M$, while EM converges to a local minimum.

References

[1] Absil, P.A., Mahony, R., Andrews, B.: Convergence of the iterates of descent methods for analytic cost functions. *SIAM J. Optim.*, 16(2), 531-547 (2005).

[2] Sh.-I. Amari. Information geometry of the EM- and em-algorithms for neural networks. *Neural Networks*, 8(9):1995,1379-1408.

[3] Sh.-I. Amari. *Information Geometry and its Applications*. Springer Applied Math. Sci. 194, 2016.

[4] H. Attouch, J. Bolte. On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. *Math. Programming*, 116(1-2, Ser. B), 5-16 (2009).

[5] H. Attouch, J. Bolte, P. Redont, A. Soubeyran. Proximal alternating minimization and projection methods for nonconvex problems: an approach based on the Kurdyka-Łojasiewicz inequality. *Mathematics of Operations Research*, 35(2):2010, 438–457.

[6] H. Attouch, J. Bolte, B.F. Svaiter. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods. *Math. Programming* 137:2013,91-129.

[7] A. Banerjee, S. Merugu, I.S. Dhillon, J. Ghosh. Clustering with Bregman divergences. *J. Mach. Learn. Res.* 6(2005),1705-1749.

[8] H.H. Bauschke, J.M. Borwein. Legendre functions and the method of random Bregman projections. *J. Convex Anal.* 4(1):1997,27-67.

[9] H.H. Bauschke, J.M. Borwein. Joint and separate convexity of the Bregman distance. *Stud. Comp. Math.* 8:2001,23-36.

[10] H.H. Bauschke, P.L. Combettes. Iterating Bregman retractions. *SIOPT* 13:2003,1159-1173.

[11] H.H. Bauschke, P.L. Combettes, D. Noll. Joint minimization with alternating Bregman proximity operators. *Pacific J. Optim.* 2:2006,401-424.

[12] H.H. Bauschke, D.R. Luke, H.M. Phan, X. Wang. Restricted normal cones and the method of alternating projections: theory. *Set-Valued and Variational Analysis*, 21:2013, 431 – 473.

[13] H.H. Bauschke, D.R. Luke, H.M. Phan, X. Wang. Restricted normal cones and the method of alternating projections: applications. *Set-Valued and Variational Analysis,* 21:2013, 475–501.

[14] H.H. Bauschke, D. Noll. On cluster points of alternating projections. *Serdica Math. Journal*, 39:2013, 355 – 364.

[15] H.H. Bauschke, D. Noll. On the local convergence of the Douglas-Rachford algorithm. *Archiv der Math.*, 102(6):2014, 589–600.

[16] H.H. Bauschke, D. Noll. The method of forward projections. *J. Nonlin. Convex Anal.* 3:2002,191-205.

[17] H.H. Bauschke, D. Noll, A. Celler, J.M. Borwein. An EM algorithm for dynamic SPECT. *IEEE Transactions on Medical Imaging*, 18(3):1999,252-261.

[18] H.H. Bauschke, X. Wang, J. Ye, X. Yuan. Bregman distances and Chebyshev sets. *J. Approx. Theory* 159:2009,3-25.

[19] E. Bierstone, P. Milman. Semianalytic and subanalytic sets. *IHES Publ. Math.*, 67:1988, 5-42.

[20] W. Blaschke. *Kreis und Kugel*. Leipzig, von Veit & Comp., 1916.

[21] J. Bolte, A. Daniilidis, O. Ley, L. Mazet. Characterizations of Łojasiewicz inequalities: subgradient flows, talweg, convexity. *Trans. Amer. math. Soc.* 362(6):2009,3319-3363.

[22] J. Bolte, A. Daniilidis, A.S. Lewis, M. Shiota. Clarke subgradients of stratifiable functions. *SIAM J. Optim.* 18:2007,556-572.

[23] L. Bregman. A relaxation method of finding a common point of convex sets and its application to problems of optimization. *Dokl. Akad. Nauk SSSR*, 171(5):1966,1019-1022. (English translation: *Soviet Math. Dokl.* 7, 1966, 1578-1581).

[24] L. D. Brown. Fundamentals of statistical exponential families with applications in statistical decision theory. *IMS Lecture Notes Monogr. Ser.*, 9: 284pp. (1986).
[25] D. Butnariu, A.N. Iusem. *Totally Convex Functions for Fixed Point Computation in Infinite Dimensional Optimization*. Kluwer, Dordrecht (2000).
[26] C. Byrne, Y. Censor. Proximity function minimization using multiple Bregman projections, with applications to split feasibility and Kullback-Leibler distance minimization, *Ann. Oper. Res.*, 105:2001,77–98.
[27] Y. Censor, T. Elfving. A multiprojection algorithm using Bregman projections in a product space. *Num. Alg.* 8:1994,221-239.
[28] Y. Censor, S.A. Zenios. *Parallel Optimization*. Oxford University Press, 1997.
[29] Y. Censor, M. Zaknoon. Algorithms and convergence results of projection methods for inconsistent feasibility problems. *Pure Appl. Func. Anal.* 3:2018,565-586.
[30] M. Coste. An introduction to o-minimal geometry. *RAAD Notes, Institut de Recherche de Rennes*, 1999.
[31] I. Csiszár. I-divergence geometry of probability distributions and minimization problems, *Ann. Probability*, 3:1975,146-158.
[32] I. Csiszár, G. Tusnády. Information geometry and alternating minimization procedures. *Statistics and Decisions*, 1(1):1984,205-237.
[33] L. van den Dries. *Tame Topology and O-minimal Structures*. Camb. Univ. Press, 1998.
[34] L. van den Dries, C. Miller. Geometric categories and o-minimal structures. *Duke Math. J.* 85:1996,487-540.
[35] L. van den Dries, C. Miller. On the real exponential field with restricted analytic functions, *Isr. J. Math.* 85, No. 1-3, 19-56 (1994).
[36] H. Federer. Hausdorff measure and Lebesgue area. *Proc. Nat. Acad. Sci. USA*, 37:1951,90-94.
[37] H. Hadwiger: *Altes und Neues über konvexe Körper*. Elemente der Mathematik vom höheren Standpunkt aus, vol. III, Springer 1955,
[38] H. Hino, S. Akaho, N. Murata. Geometry of EM and related iterative algorithms. `arXiv:2209.01301v2 [stat.ML] 12 Nov 2022`
[39] A.N. Iusem. A short convergence proof of the EM algorithm for a specific Poisson model. *Brazilian Journal of Probability and Statistics* 6(1): 1992,
[40] F. Kunstner, R. Kumar, M. Schmidt. Homeomorphic-invariance of EM: Non-asymptotic convergence in KL divergence for exponential families with mirror descent. Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130. `arXiv:2011.01170`
[41] S. Kullback. *Information Theory and Statistics*. Dover Publications, 1968.
[42] S. Kullback, R.A. Leibler. On information and sufficiency. *The Ann. Math. Stat.* 22(1):1951,79–86.
[43] K. Kurdyka. On gradients of functions definable in o-minimal structures. *Ann. Inst. Fourier*, 48:1998,769-783.
[44] A.S. Lewis, R. Luke, J. Malick. Local linear convergence for alternating and averaged non convex projections. *Found. Comp. Math.* 9:2009, 485–513.
[45] A.S. Lewis, J. Malick. Alternating projections on manifolds. *Math. Oper. Res.* 33:2008, 216–234.
[46] J. Maeght, S.P. Boyd, D. Noll. Dynamic emission tomography - regularization and inversion. *Can. Math. Soc. Conf. Proc.* 27, 2000, 211–234.
[47] G.J McLachlan, T. Krishnan. *The EM algorithm and extensions*. Wiley Series in Prob. and Stat. 2nd ed. 2008.
[48] C. Miller. Exponentiation is hard to avoid. *Proc. Amer. Math. Soc* 122(1):1994,257-259.
[49] B.S. Mordukhovich. *Variational Analysis and Generalized Differentiation*. Springer, New York, 2006.
[50] D. Noll. Convergence of nonsmooth descent methods using the Kurdyka-Łojasiewicz inequality. *J. Optim. Theory Appl.* 160(2):2014,553-572.
[51] D. Noll. Alternating projections with applications to Gerchberg-Saxton error reduction. *Set-Valued and Variational Analysis*, 29:2021,771-802.
[52] D. Noll, A. Rondepierre. On local convergence of the method of alternating projections. *Foundations of Computational Mathematics*, vol. 16, no. 2, 2016, pp. 425-455.
[53] R.T. Rockafellar. *Convex Analysis*. Princeton University Press, NJ, 1970.
[54] R.T. Rockafellar, R.J.B. Wets. *Variational Analysis*. Grundlehren der mathematischen Wissenschaften, Springer 317:2009.
[55] S. Rolewicz. On uniform differentiability. *Bull. Polish Acad. Sci. Math.* 56(3-4):2008,231-237.
[56] M. Shiota. *Geometry of Subanalytic and Semialgebraic Sets*. Birkhäuser Verlag 1997.
[57] P. Tseng. An analysis of the EM algorithm and entropy-like proximal point methods. Math. Oper. Res. 29(1):2004,27-44.
[58] G. Walther. On a generalization of Blaschke's rolling theorem and the smoothing of surfaces. *Math. Meth. Appl. Sci.* 22:1999,301–316.
[59] A. Wilkie. Model completeness results for expansions of the ordered field of real numbers by restricted Pfaffian functions and the exponential function. *J. Amer. Math. Soc.* 9:1996,1051–1094.
[60] C.F.J. Wu. On the convergence properties of the EM algorithm. *Annals of Statistics* 11:1983,95-103.
[61] Z. Zhu, X. Li. Convergence analysis of alternating projection method for nonconvex sets. `arXiv:1802.03889v2 [math.OC] 22 Jul 2019`

Institut de Mathématiques, Université de Toulouse, France
Email address: `dominikus.noll@math.univ-toulouse.fr`