Gradient descent in a generalised Bregman distance framework

Martin Benning[1], Marta M. Betcke[2], Matthias J. Ehrhardt[1], and Carola-Bibiane Schönlieb[1]

University of Cambridge, Dept. of Applied Mathematics & Theoretical Physics, Wilberfore Road, Cambridge CB3 0WA, United Kingdom
{mb941, me404, cbs31}@cam.ac.uk

University College London, Dept. of Computer Science, Gower Street, London WC1E 6BT, United Kingdom
m.betcke@ucl.ac.uk

1 Introduction

In this work we study a generalisation of classical gradient descent that has become known in the literature as the so-called linearised Bregman iteration [8, 7], and – as the key novelty of this publication – apply it to minimise smooth but not necessarily convex objectives $E : \mathcal{U} \to \mathbb{R}$ over a Banach space $\mathcal{U}$. For this generalisation we want to consider proper, lower semi-continuous (l.s.c.), convex but not necessarily smooth functionals $J : \mathcal{U} \to \mathbb{R} \cup \{\infty\}$, and consider their generalised Bregman distances

$$D_J^p(u, v) = J(u) - J(v) - \langle p, u - v \rangle ,$$

for $u, v \in \mathcal{U}$ and $p \in \partial J(v)$, where $\partial J(v)$ denotes the subdifferential of J. Note that in case J is smooth we omit p in the notation of the Bregman distance, as the subdifferential is single-valued in this case. We further assume that there exists a proper, l.s.c., convex and not necessary smooth functional $F : \mathcal{U} \to \mathbb{R} \cup \{\infty\}$ such that the functional $G := F - E$ is also convex. This will imply $D_G^{q-\nabla E(v)}(u, v) \geq 0$ for all $u, v \in \mathrm{dom}(G)$ and $q \in \partial F(v)$, since $q - \nabla E$ is the gradient of G. Hence, the convexity of G yields the descent estimate

$$E(u) \leq E(v) + \langle \nabla E(v), u - v \rangle + D_F^q(u, v) , \tag{1}$$

for all $u, v \in \mathrm{dom}(F)$ and $q \in \partial F(v)$. We want to emphasise that in case of $F(u) = \frac{L}{2}\|u\|_2^2$ (for some constant $L > 0$) (1) reduces to the classical Lipschitz estimate; this generalisation has also been discovered in [2] simultaneously to this work (without the generalisation of Bregman distances to non-smooth functionals, though).

2 Linearised Bregman iteration applied to non-convex problems

The linearised Bregman iteration that we are going to study in this work is defined as

$$u^{k+1} = \operatorname*{arg\,min}_{u \in \mathrm{dom}(J)} \left\{ \tau^k \langle u - u^k, \nabla E(u^k) \rangle + D_J^{p^k}(u, u^k) \right\} , \tag{2a}$$

$$p^{k+1} = p^k - \tau^k \nabla E(u^k) \, , \tag{2b}$$

for $k \in \mathbb{N}$, some $u^0 \in \mathcal{U}$ and $p^0 \in \partial J(u^0)$. Here $J : \mathcal{U} \to \mathbb{R} \cup \{\infty\}$ is not only proper, l.s.c. and convex, but also chosen such that the overall functional in (2a) is coercive and strictly convex and thus, its minimiser well-defined and unique.

We want to highlight that this model has been studied for several scenarios in which E is the convex functional $E(u) = \frac{1}{2}\|Ku - f\|_2^2$, for data f and linear and bounded operators K (cf. [8, 7]), for more general convex functionals E and smooth J in [6, 3], as well as for the non-convex functional $E(u) = \frac{1}{2}\|K(u) - f\|_2^2$ for data f and a smooth but non-linear operator K in [1]. However, to our knowledge this is the first work that studies (2) for general smooth but not necessarily convex functionals E.

3 A sufficient decrease property

We want to show that together with the descent estimate (1) we can guarantee a sufficient decrease property of the iterates (2) in terms of the symmetric Bregman distance. The symmetric Bregman distance $D_J^{\mathrm{symm}}(u, v)$ (cf. [5]) is simply defined as $D_J^{\mathrm{symm}}(u, v) = D_J^q(u, v) + D_J^p(v, u) = \langle u - v, p - q \rangle$ for all $u, v \in \mathrm{dom}(J)$, $p \in \partial J(u)$ and $q \in \partial J(v)$.

Lemma 1 (Sufficient decrease property)**.** *Let $E : \mathcal{U} \to \mathbb{R}$ be a l.s.c. and smooth functional that is bounded from below and for which a proper, l.s.c. and convex functional $F : \mathcal{U} \to \mathbb{R} \cup \{\infty\}$ exists such that $G := F - E$ is also convex. Further, let $J : \mathcal{U} \to \mathbb{R} \cup \{\infty\}$ be a proper, l.s.c. and convex functional such that* (2a) *is well defined and unique. Further we choose τ^k such that the estimate*

$$\rho D_J^{symm}(u^{k+1}, u^k) \leq \frac{1}{\tau^k} D_J^{symm}(u^{k+1}, u^k) - D_F^{q^k}(u^{k+1}, u^k) \tag{3}$$

holds true, for all $k \in \mathbb{N}$, $q^k \in \partial F(u^k)$ and a fixed constant $0 < \rho < \infty$. Then the iterates of the linearised Bregman iteration (2) *satisfy the descent estimate*

$$E(u^{k+1}) + \rho D_J^{symm}(u^{k+1}, u^k) \leq E(u^k) \, . \tag{4}$$

In addition, we observe

$$\lim_{k \to \infty} D_J^{symm}(u^{k+1}, u^k) = 0 \, .$$

Proof. First of all, we easily see that update (2b), i.e.

$$\tau^k \nabla E(u^k) + (p^{k+1} - p^k) = 0 \, ,$$

is simply the optimality condition of (2a), for $p^{k+1} \in \partial J(u^{k+1})$. Taking a dual product of (2b) with $u^{k+1} - u^k$ yields

$$\langle \nabla E(u^k), u^{k+1} - u^k \rangle = -\frac{1}{\tau^k} D_J^{\mathrm{symm}}(u^{k+1}, u^k) \, . \tag{5}$$

Due to (1) we can further estimate

$$E(u^{k+1}) \leq E(u^k) + \langle u^{k+1} - u^k, \nabla E(u^k) \rangle + D_F^{q^k}(u^{k+1}, u^k)\,,$$

for $q^k \in \partial F(u^k)$. Together with (5) we therefore obtain

$$E(u^{k+1}) + \frac{1}{\tau^k} D_J^{\mathrm{symm}}(u^{k+1}, u^k) - D_F^{q^k}(u^{k+1}, u^k) \leq E(u^k)\,.$$

Using (3) then allows us to conclude

$$0 \leq \rho D_J^{\mathrm{symm}}(u^{k+1}, u^k) \leq E(u^k) - E(u^{k+1})\,;$$

hence, summing up over all N iterates and telescoping yields

$$\begin{aligned}\sum_{k=0}^{N} \rho D_J^{\mathrm{symm}}(u^{k+1}, u^k) &\leq \sum_{k=0}^{N} E(u^k) - E(u^{k+1})\,,\\ &= E(u^0) - E(u^{N+1})\,,\\ &\leq E(u^0) - \overline{E} < \infty\,,\end{aligned}$$

where $\overline{E}$ denotes the lower bound of E. Taking the limit $N \to \infty$ then implies

$$\sum_{k=0}^{\infty} \rho D_J^{\mathrm{symm}}(u^{k+1}, u^k) < \infty\,,$$

and thus, we have $\lim_{k\to\infty} D_J^{\mathrm{symm}}(u^{k+1}, u^k) = 0$ due to $\rho > 0$. □

Remark 1. *We want to emphasise that Lemma 1 together with the duality $D_J^{symm}(u^{k+1}, u^k) = D_{J^*}^{symm}(p^{k+1}, p^k)$, for $p^{k+1} \in \partial J(u^{k+1})$ and $p^k \in \partial J(u^k)$, further implies*

$$\lim_{k\to\infty} D_{J^*}^{symm}(p^{k+1}, p^k) = 0\,,$$

and hence, a sufficient decrease property holds also for the dual iterates. Here $J^* : \mathcal{U}^* \to \mathbb{R} \cup \{\infty\}$ denotes the Fenchel conjugate of J, and $\mathcal{U}^*$ is the dual space of $\mathcal{U}$.

4 A global convergence statement

For the following part we assume that both J and J^* are strongly convex w.r.t. the $\mathcal{U}$- respectively the $\mathcal{U}^*$-norm, i.e. there exist constants $\gamma > 0$ and $\delta > 0$ such that

$$\gamma \|u - v\|_{\mathcal{U}}^2 \leq D_J^{\mathrm{symm}}(u, v) \qquad \text{and} \qquad \delta \|p - q\|_{\mathcal{U}^*}^2 \leq D_{J^*}^{\mathrm{symm}}(p, q) \tag{6}$$

hold true for all $u, v \in \mathcal{U}$ and $p, q \in \mathcal{U}^*$. From Lemma 1 and (6) we readily obtain

$$\rho_1 \|u^{k+1} - u^k\|_{\mathcal{U}}^2 \leq E(u^k) - E(u^{k+1})\,, \tag{7}$$

for $\rho_1 := \gamma/\rho$, which implies $\lim_{k\to\infty} \|u^{k+1} - u^k\|_{\mathcal{U}} = 0$.

We follow [4] and establish a global convergence result by proving that the dual norm of the gradient is bounded by the iterates gap in addition to the already proven descent result (7). Together with a generalised Kurdyka-Łojasiewicz property we will be able to prove a global convergence statement for (2).

Given (6), we obtain the necessary iterates gap in the corresponding Banach space norm as an upper bound for the gradient in the dual Banach space norm, as follows.

Lemma 2 (Gradient bound)**.** *Let the same assumptions hold true as in Lemma 1, and let* (6) *be fulfilled. Then the iterates* (2) *satisfy*

$$\|\nabla E(u^k)\|_{\mathcal{U}^*} \leq \rho_2 \|u^{k+1} - u^k\|_{\mathcal{U}}\,, \tag{8}$$

for $\rho_2 := 1/(\delta\overline{\tau})$ *and* $\overline{\tau} := \inf_k \tau^k$*.*

Proof. As pointed out in Remark 1, we have the duality $D^{\mathrm{symm}}_{J^*}(p^{k+1}, p^k) = D^{\mathrm{symm}}_{J}(u^{k+1}, u^k)$ for the symmetric Bregman distances. Together with the duality estimate $\langle u, p\rangle \leq \|u\|_{\mathcal{U}}\|p\|_{\mathcal{U}^*}$ we therefore obtain

$$D^{\mathrm{symm}}_{J^*}(p^{k+1}, p^k) = \langle p^{k+1} - p^k, u^{k+1} - u^k\rangle \leq \|u^{k+1} - u^k\|_{\mathcal{U}}\|p^{k+1} - p^k\|_{\mathcal{U}^*}\,.$$

Hence, using (2b) yields

$$\frac{D^{\mathrm{symm}}_{J^*}(p^k - \tau^k \nabla E(u^k), p^k)}{\tau^k \|\nabla E(u^k)\|_{\mathcal{U}^*}} \leq \|u^{k+1} - u^k\|_{\mathcal{U}}\,.$$

Together with the δ-strong convexity (6) and $\rho_2 := 1/(\delta\overline{\tau})$ we get (8). □

Remark 2. *Note that we have to ensure* $\overline{\tau} > 0$ *in order to ensure* $\rho_2 < \infty$*. Due to* (3) *we can ensure this as long as* $D^{q^k}_F(u^{k+1}, u^k)$ *is bounded from above for all* $k \in \mathbb{N}$*.*

Before we can establish a global convergence result, we have to restrict the functionals E to the following class of functionals satisfying a generalised Kurdyka-Łojasiewicz property.

Definition 1 (Generalised Kurdyka-Łojasiewicz (KL) property)**.** *We assume for* $\eta > 0$ *that* $\varphi : [0, \eta[\to \mathbb{R}_{>0}$ *is a function that is continuous at zero and satisfies* $\varphi(0) = 0$*,* $\varphi \in C^1(]0, \eta[)$*. Let further* $E : \mathcal{U} \to \mathbb{R}$ *be a proper, l.s.c. and smooth functional.*

1. *The functional E fulfils the (generalised) KL property at a point* $\overline{u} \in \mathcal{U}$ *if there exists* $\eta \in]0, \infty]$*, a neighbourhood U of* $\overline{u}$ *and a function* φ *satisfying the conditions above, such that for all*

$$u \in U \cap \{u \mid E(\overline{u}) < E(u) < E(\overline{u}) + \eta\}$$

we observe

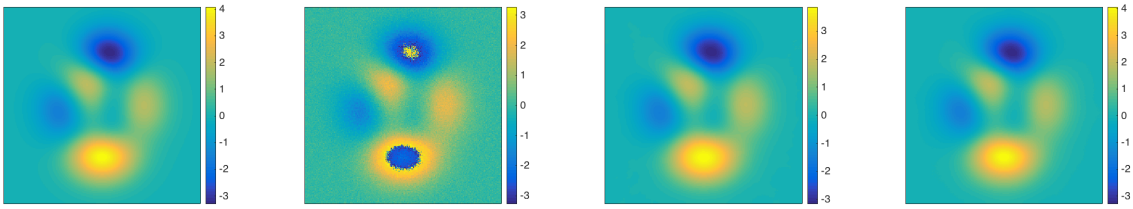
$$\varphi'(E(u) - E(\overline{u}))\|\nabla E(u)\|_{\mathcal{U}^*} \geq 1\,. \tag{9}$$

2. *If E satisfies the (generalised) KL property for all arguments in* $\mathcal{U}$*, E is called a (generalised) KL functional.*

Together with the previous results the generalised KL condition (9) allows to establish the following global convergence result.

Theorem 1 (Global convergence)**.** *Let the Banach space* $\mathcal{U}$ *be the dual of a separable normed space. Suppose that E is coercive, sequentially weak*-continuous and a KL function in the sense of Definition 1. Then the sequences* $\{u^k\}_{k\in\mathbb{N}}$ *and* $\{p^k\}_{k\in\mathbb{N}}$ *generated by* (2) *each have a strongly convergent subsequence with limits* $\hat{u}$ *and* $\hat{p}$*, with* $\nabla E(\hat{u}) = 0$ *and* $\hat{p} \in \partial J(\hat{u})$*. If* $\dim(\mathcal{U}) < \infty$*, then the convergence holds true for the entire sequences.*

Proof. The proof utilises (4), (8) and (9) to derive the statement. Due to page restrictions, the full length proof will be published separately in an extended version of this manuscript. □



(a) Ground truth (b) Gradient descent (c) $R(u) = \frac{1}{2}\|\nabla u\|^2_{L^2}$ (d) $R(u) = \|Cu\|_{\ell^1}$

Figure 1: A phase unwrapping example. Figure 1(a) shows the unknown, noise-free, ground truth signal. Figure 1(b) shows the result of classical gradient descent computation. Figure 1(c) visualises the solution of model 2.) with $\alpha = 1000$. Figure 1(d) shows the solution of model 3.) with $\alpha = 50$. All reconstructions have been computed from zero initialisations and were stopped according to the same discrepancy principle.

5 Phase unwrapping as a toy example

We want to conclude this paper with a numerical toy example for which we consider to minimise $E(u) := \frac{1}{2}\|K(u) - f\|^2_{L^2(\Omega;\mathbb{R}^2)}$ for $K(u) = (\cos(u), \sin(u))^T$, and choose $F(u) = \frac{L}{2}\|u\|^2_{L^2(\Omega)}$ with $L = 1$. We will minimise E via (2) with $J(u) := \frac{1}{2}\|u\|^2_{L^2(\Omega)} + \alpha R(u)$, for a positive scalar $\alpha > 0$ and three different choices of R: 1.) $R(u) = 0$, 2.) $R(u) = \frac{1}{2}\|\nabla u\|^2_{L^2(\Omega;\mathbb{R}^2)}$, and 3.) $R(u) = \|Cu\|_{\ell^1}$, where C denotes the two-dimensional discrete Cosine transform. The first case simply corresponds to classical gradient descent, case 2.) is gradient descent in a Hilbert space metric and 3.) corresponds to gradient descent in a non-smooth Bregman distance setting that does not correspond to a metric. Note that the question, whether E and J satisfy all conditions that are necessary for global convergence, will be omitted due to the page limit, but addressed in an extended version of this manuscript in the future. We do want to mention, though, that it is easy to see that J in 3.) does not meet the requirement (7); this, however, can be corrected via a smoothing of the ℓ^1-norm, for instance via a Huberised ℓ^1-norm.

In order to consider numerical examples, we discretise the above scenarios in a straight forward fashion. Input data f is created by applying the non-linear operator K to a multiple of the built-in MATLAB© signal 'peaks' (see Figure 1(a)) and additive normal distributed noise with mean zero and standard deviation $\sigma = 0.15$. Due to noise in the data, the iteration (2) is stopped as soon as $E(u^k) \leq \sigma^2 m/2$ is satisfied. Here m denotes the number of discrete samples. Reconstruction results for zero initialisations and the choice $\tau^k = 1.5$ for all $k \in \mathbb{N}$ can be found in Figure 1(b), 1(c) and 1(d). We want to emphasise that this example is just a toy example to demonstrate the impact of different choices of J; there are certainly much better unwrapping strategies, particularly for the unwrapping of smooth signals.

Code statement: The corresponding MATLAB© code can be downloaded at `https://doi.org/10.17863/CAM.6714`.

6 Conclusions & Outlook

We have presented a short convergence analysis of the linearised Bregman iteration for the minimisation of general smooth but non-convex functionals. We have proven a sufficient decrease property, and confirmed that the dual norm of the gradient is bounded by the

primal iterates under additional strong convexity assumptions of the convex functional that builds the basis for the Bregman iteration. Under a generalised KL condition, we have stated a global convergence result that we are going to refine in detail in a future release. We have concluded with a numerical toy example of phase unwrapping for three different Bregman distances. In a future work we are going to analyse the linearised Bregman iteration and its convergence behaviour in more detail and in a more generalised setting, and are going to investigate different Bregman distance choices as well as different numerical applications.

Acknowledgment

MB acknowledges support from the Leverhulme Trust Early Career project 'Learning from mistakes: a supervised feedback-loop for imaging applications' and the Isaac Newton Trust. MMB acknowledges support from the Engineering and Physical Sciences Research Council (EPSRC) 'EP/K009745/1'. MJE and CBS acknowledge support from the Leverhulme Trust project 'Breaking the non-convexity barrier'. CBS further acknowledges support from EPSRC grant 'EP/M00483X/1', EPSRC centre 'EP/N014588/1' and the Cantab Capital Institute for the Mathematics of Information. All authors acknowledge support from CHiPS (Horizon 2020 RISE project grant) that made this contribution possible.

References

[1] Markus Bachmayr and Martin Burger. Iterative total variation methods for nonlinear inverse problems. *Inverse Problems*, 25(10):26, 2009.

[2] Heinz Bauschke, Jérôme Bolte, and Marc Teboulle. A descent lemma beyond Lipschitz gradient continuity: first-order methods revisited and applications. *Mathematics of Operations Research (to appear.)*.

[3] Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

[4] Jérôme Bolte, Shoham Sabach, and Marc Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, 146(1-2):459–494, 2014.

[5] Martin Burger, Elena Resmerita, and Lin He. Error estimation for Bregman iterations and inverse scale space methods in image restoration. *Computing*, 81(2-3):109–135, 2007.

[6] Arkadi Nemirovski and David Borisovich Yudin. Problem complexity and method efficiency in optimization. 1982.

[7] Wotao Yin. Analysis and generalizations of the linearized Bregman method. *SIAM Journal on Imaging Sciences*, 3(4):856–877, 2010.

[8] Wotao Yin, Stanley Osher, Donald Goldfarb, and Jerome Darbon. Bregman iterative algorithms for ℓ_1-minimization with applications to compressed sensing. *SIAM Journal on Imaging Sciences*, 1(1):143–168, 2008.