CHOOSE YOUR PATH WISELY: GRADIENT DESCENT IN A BREGMAN DISTANCE FRAMEWORK*

MARTIN BENNING†, MARTA M. BETCKE‡, MATTHIAS J. EHRHARDT†, AND CAROLA-BIBIANE SCHÖNLIEB†

Abstract. We propose an extension of a special form of gradient descent — in the literature known as linearised Bregman iteration — to a larger class of non-convex functionals. We replace the classical (squared) two norm metric in the gradient descent setting with a generalised Bregman distance, based on a proper, convex and lower semi-continuous functional. The proposed algorithm is a generalisation of numerous well-known optimisation methods. Its global convergence is proven for functions that satisfy the Kurdyka-Łojasiewicz property. Examples illustrate that for suitable choices of Bregman distances this method — in contrast to traditional gradient descent — allows iterating along regular solution-paths. The effectiveness of the linearised Bregman iteration in combination with early stopping is illustrated for the applications of parallel magnetic resonance imaging, blind deconvolution as well as image classification.

Key words. Nonconvex Optimisation, Nonsmooth Optimisation, Gradient Descent, Bregman Iteration, Linearised Bregman Iteration, Parallel MRI, Blind Deconvolution, Deep Learning

AMS subject classifications. 49M37, 65K05, 65K10, 90C26, 90C30

1. Introduction.

Non-convex optimisation methods are indispensable mathematical tools for a large variety of applications [57]. For differentiable objectives, first-order methods such as gradient descent have proven to be useful tools in all kinds of scenarios. Throughout the last decade, however, there has been an increasing interest in first-order methods for non-convex and non-smooth objectives. These methods range from forward-backward, respectively proximal-type, schemes [1, 2, 3, 17, 18], over linearised proximal schemes [71, 15, 72, 56], to inertial methods [58, 62], primal-dual algorithms [70, 48, 53, 12], scaled gradient projection methods [63] and non-smooth Gauß-Newton extensions [32, 59].

In this paper, we follow a different approach of incorporating non-smoothness into first-order methods for non-convex problems. We present a direct generalisation of gradient descent, first introduced in [10], where the usual squared two-norm metric that penalises the gap of two subsequent iterates is being replaced by a potentially non-smooth distance term. This distance term is given in form of a generalised Bregman distance [19], where the underlying functional is proper, lower semi-continuous and convex, but not necessarily smooth. If the underlying functional is a Legendre function (see [66, Section 26] and [6]), the proposed generalisation coincides with the recently proposed Bregman proximal gradient [16]. In the more general case, the proposed method is a generalisation of the so-called linearised Bregman iteration [30, 74, 22, 21] to non-convex data fidelities.

The use of non-smooth Bregman distances for the penalisation of the iterates gap

*Submitted to the editors DATE.

Funding: This work was funded by the Leverhulme Trust Early Career Fellowship 'Learning from mistakes: a supervised feedback-loop for imaging applications', the Isaac Newton Trust, the Engineering and Physical Sciences Research Council (EPSRC) 'EP/K009745/1', the Leverhulme Trust project 'Breaking the non-convexity barrier', the EPSRC grant 'EP/M00483X/1', the EPSRC centre 'EP/N014588/1', the Cantab Capital Institute for the Mathematics of Information and CHiPS (Horizon 2020 RISE project grant).

†Department of Applied Mathematics and Theoretical Physics, University of Cambridge, UK (mb941@cam.ac.uk, m.j.ehrhardt@damtp.cam.ac.uk, cbs31@cam.ac.uk).

‡Department of Computer Science, University College London, UK (m.betcke@ucl.ac.uk).

allows for adaptive solution paths of the gradient descent. Replacing the squared two-norm for instance with a squared two-norm plus the Bregman distance w.r.t. a scaled one-norm or total variation leads to solution paths of different regularity. This control of the solution path can, for instance, be used to slowly evolve from very regular solutions to less regular solutions, thus creating more stable solution paths. This is in stark contrast to many of the non-smooth, non-convex first-order approaches mentioned above, where the methods are often initialised with random inputs that become more regular as we move along the solution path.

Our main contributions of this paper are the generalisation of the linearised Bregman iteration to non-convex functionals in a Banach space setting, a detailed convergence analysis of the proposed method as well as the presentation of numerical results that demonstrate that the use of solution paths of different regularity can lead to superior solutions.

The outline of the paper is as follows. We first recall key concepts of convex and non-convex analysis that are needed throughout the paper in Section 2. Subsequently, we define the extension of the linearised Bregman iteration for non-convex functionals in Section 3. Then, motivated by the informal convergence recipe of Bolte et al. [15, Section 3.2] we show a sufficient decrease property as well as a subgradient lower bound of the iterates gap for the proposed method in Section 4. In Section 5 we then prove a global convergence result in finite dimensions, which concludes the theoretical part. We conclude with the modelling of the applications of parallel Magnetic Resonance Imaging (MRI), blind deconvolution and image classification in Section 6, followed by corresponding numerical results in Section 7 as well as conclusions and outlook in Section 8.

2. Mathematical preliminaries. We briefly summarise some concepts of convex and non-convex analysis that are of importance for the remainder of this paper. We deal with general Banach spaces, denoted by $\mathcal{U}$, which are the dual of some Banach space $\mathcal{U}^*$. We frequently use functionals that are proper, lower semi-continuous and convex, and therefore define the following set of functionals:

$$\Gamma_0 := \{J : \mathcal{U} \to \mathbb{R} \cup \{\infty\} \,|\, J \text{ is proper, lower semi-continuous and convex}\} \,.$$

Convex, subdifferentiable functionals are of major importance throughout the entire paper. We therefore want to recall the definition of subgradients and the subdifferential of a convex functional.

DEFINITION 2.1 (Subdifferential). *Let $J \in \Gamma_0$. The functional J is called* subdifferentiable *at $u \in \mathcal{U}$, if there exists an element $p \in \mathcal{U}^*$ such that*

$$J(v) \geq J(u) + \langle p, v - u \rangle$$

holds, for all $v \in \mathcal{U}$. Furthermore, we call p a subgradient *at position u. The collection of all subgradients at position u, i.e.*

$$\partial J(u) := \{p \in \mathcal{U}^* \,|\, J(v) \geq J(u) + \langle p, v - u \rangle \,, \forall v \in \mathcal{U}\} \,,$$

is called subdifferential *of J at u.*

Bregman distances, introduced by Lev Bregman in 1967 (see [19]), play a vital role in the definition as well as in the convergence analysis of the linearised Bregman iteration for non-convex functionals. We recall its generalised variant for subdifferentiable functionals.

DEFINITION 2.2 (Bregman distance). *Let $J \in \Gamma_0$. Then the* generalised Bregman distance *D_J for a particular subgradient $q \in \partial J(v)$ is defined as*

$$D_J^q(u,v) := J(u) - J(v) - \langle q, u - v \rangle , \tag{1}$$

for $v \in \mathrm{dom}(J)$ and all $u \in \mathrm{dom}(J)$. Here $\mathrm{dom}(J)$ denotes the effective domain of J.

REMARK 1. *We omit the subgradient in the notation of the Bregman distance in case the subdifferential is single-valued and thus we have $q = \nabla J(v)$.*

Bregman distances are closely linked to the definition of strong convexity.

DEFINITION 2.3 (Strong convexity). *Let $J \in \Gamma_0$. Then J is said to be* γ-strongly convex *w.r.t. the $\mathcal{U}$-norm if there exists a constant γ such that*

$$D_J^q(u,v) \geq \frac{\gamma}{2}\|u - v\|_{\mathcal{U}}^2 \tag{2}$$

holds true for all $u, v \in \mathrm{dom}(J)$ and $q \in \partial J(v)$.

Bregman distances are not symmetric in general; however, they satisfy a dual symmetry $D_J^q(u,v) = D_{J^*}^u(q,p)$ for arguments $u, v \in \mathrm{dom}(J)$ and subgradients $p \in \partial J(u)$ and $q \in \partial J(v)$. Symmetry can nevertheless be achieved by simply adding two Bregman distances with interchanged arguments. The name symmetric Bregman distance goes back to [20].

DEFINITION 2.4 (Symmetric Bregman distance). *Let $J \in \Gamma_0$. Then the* symmetric generalised Bregman distance *$D_J^{symm}(u,v)$ is defined as*

$$D_J^{symm}(u,v) := D_J^q(u,v) + D_J^p(v,u) = \langle p - q, u - v \rangle ,$$

for $u, v \in dom(J)$ with $p \in \partial J(u)$ and $q \in \partial J(v)$.

REMARK 2. *Note that due to (2) we immediately observe*

$$\gamma\|u - v\|_{\mathcal{U}}^2 \leq D_J^{symm}(u,v) \tag{3}$$

in case J is γ-strongly convex.

Another concept that we exploit is Lipschitz-continuity, however not in a classical Hilbert-space setting but also in a Banach space setting. Generalised Lipschitz-continuity is then defined as follows.

DEFINITION 2.5 (Generalised Lipschitz continuity). *An operator $G : \mathcal{U} \to \mathcal{U}^*$ is said to be* (globally) Lipschitz-continuous *if there exists a constant $L > 0$ such that*

$$\|G(u) - G(v)\|_{\mathcal{U}^*} \leq L\|u - v\|_{\mathcal{U}} \tag{4}$$

is satisfied for all $u, v \in \mathcal{U}$.

Due to the importance of Lipschitz-continuous gradients, we define the following class of continuously differentiable functionals with Lipschitz-continuous gradient:

DEFINITION 2.6 (Smoothness). *A functional $J : \mathcal{U} \to \mathbb{R}$ is called* L-smooth *if it is Fréchet-differentiable and its derivative $\nabla J : \mathcal{U} \to \mathcal{U}^*$ is L-Lipschitz-continuous. The set of all L-smooth functionals is therefore denoted by $\mathcal{S}_L$ with*

$$\mathcal{S}_L := \left\{ J : \mathcal{U} \to \mathbb{R} \;\middle|\; \begin{array}{c} J \text{ is continuously Fréchet-differentiable} \\ \nabla J \text{ is } L\text{-Lipschitz-continuous} \end{array} \right\} .$$

Similar to classical Lipschitz continuity, generalised Lipschitz continuity of the gradient of a non-convex functional also implies the convexity of the difference of a multiple of a strongly convex functional and this non-convex functional.

LEMMA 2.7. *Let $H \in \Gamma_0$ be γ-strongly convex with convex domain $\mathrm{dom}(H)$, and let $E \in \mathcal{S}_L$, i.e.*

$$\|\nabla E(u) - \nabla E(v)\|_{\mathcal{U}^*} \leq L\|u - v\|_{\mathcal{U}}\,,$$

for all $u, v \in \mathcal{U}$. Then Lipschitz continuity of the gradient ∇E implies convexity of the functional $G(u) := \frac{L}{\gamma}H(u) - E(u)$ for all $u \in \mathrm{dom}(H)$.

Proof. Lipschitz-continuity of the gradient of E and γ-strong convexity of H imply

$$\|\nabla E(u) - \nabla E(v)\|_{\mathcal{U}^*}\|u - v\|_{\mathcal{U}} \leq L\,\|u - v\|_{\mathcal{U}}^2 \leq \frac{L}{\gamma}D_H^{\mathrm{symm}}(u, v)\,.$$

With the standard duality estimate $\langle u, p\rangle \leq \|u\|_{\mathcal{U}}\|p\|_{\mathcal{U}^*}$ and Definition 4 we obtain

$$\langle \nabla E(u) - \nabla E(v), u - v\rangle \leq \frac{L}{\gamma}D_H^{\mathrm{symm}}(u, v) = \frac{L}{\gamma}\langle p - q, u - v\rangle\,, \tag{5}$$

for all $p \in \partial H(u)$ and $q \in \partial H(v)$. By subtracting $\langle \nabla E(u) - \nabla E(v), u - v\rangle$ on both sides, estimate (5) therefore implies

$$0 \leq \left\langle \frac{L}{\gamma}p - \nabla E(u) - \left(\frac{L}{\gamma}q - \nabla E(v)\right), u - v\right\rangle = D_G^{\mathrm{symm}}(u, v)\,, \tag{6}$$

since $\nabla G(u) = \frac{L}{\gamma}\nabla H(u) - \nabla E(u)$ for all $u \in \mathrm{dom}(H)$. Now $D_G^{\mathrm{symm}}(u, v) \geq 0$, for all $u, v \in \mathrm{dom}(H)$ and all $p \in \partial H(u)$ and $q \in \partial H(v)$, already implies $D_G(u, v) \geq 0$ for all $u, v \in \mathrm{dom}(H)$. Since $\mathrm{dom}(H)$ is convex, $u, v \in \mathrm{dom}(H)$ implies $w = u + \lambda(v - u) \in \mathrm{dom}(H)$, for $\lambda \in [0, 1]$. Consequently, $D_G(u, w) \geq 0$ and $D_G(v, w) \geq 0$ is satisfied. Adding these two inequalities yields the convex combination $G(w) \leq \lambda G(v) + (1 - \lambda)G(u)$ for all $u, v \in \mathrm{dom}(H)$. Hence, G is convex. ∎

REMARK 3. *Note that the previous lemma also holds true for functionals where the domain is a convex subset of the Banach spaces. We also want to highlight that convexity of $G := \frac{L}{\gamma}H - E$ implies $D_G^{\frac{L}{\gamma}\nabla H(u) - \nabla E(u)}(u, v) \geq 0$. This leads to the very useful estimate*

$$E(u) \leq E(v) + \langle \nabla E(v), u - v\rangle + \frac{L}{\gamma}D_H(u, v)\,. \tag{7}$$

We denote the class of functionals that are γ-strongly convex and δ-smooth as $\Phi_{\gamma,\delta}$, i.e.

$$\Phi_{\gamma,\delta} := \{J \in \mathcal{S}_\delta \mid J \text{ is } \gamma\text{-strongly convex}\}\,.$$

We conclude this section by defining a proximal mapping based on functionals of the classes $\Phi_{\gamma,\delta}$ and Γ_0. Bregman proximal mappings based on the definition of the generalised Bregman distance (see for instance [24, 68, 7, 29]) are natural candidates. However, we follow the definition in [39] instead, as it allows us to be slightly less restrictive on the assumptions for the underlying functionals.

DEFINITION 2.8 (Proximal mapping). *Let $H \in \Phi_{\gamma,\delta}$ and let $R \in \Gamma_0$. Then we define the* proximal mapping *as the operator $(\nabla H + \alpha\partial R)^{-1} : \mathcal{U}^* \to \mathrm{dom}(R)$ with*

$$(\nabla H + \alpha\partial R)^{-1}(f) := \operatorname*{arg\,min}_{u\in\mathrm{dom}(R)} \left\{H(u) - \langle u, f\rangle + \alpha R(u)\right\} ,$$

for all arguments $f \in \mathcal{U}^*$ and a non-negative parameter $\alpha \geq 0$.

Now we have all the necessary mathematical tools to discuss the linearised Bregman iteration for the solution of non-convex optimisation problems. In the following section, we introduce the method whilst discussing its convergence properties in the subsequent section.

3. Linearised Bregman iteration for non-convex problems. We are interested in the minimisation of functionals $E : \mathcal{U} \to \mathbb{R} \cup \{\infty\}$ of the form

$$E := F + S \,, \tag{P}$$

for $F \in \mathcal{S}_L$ and $S \in \Gamma_0$. We want to emphasise that F does not necessarily have to be convex.

In order for the minimisation of E to make sense, we have to set some assumptions for this functional first. We define the set of critical points of E as

$$\mathrm{crit}(E) := \{u \in \mathrm{dom}(E) \mid -\nabla F(u) \in \partial S(u)\} \,. \tag{8}$$

Note that for $S \equiv 0$ the set $\mathrm{crit}(E)$ reduces to the set $\mathrm{crit}(F) := \{u \in \mathrm{dom}(F) \mid \nabla F(u) = 0\}$ of critical points of F. From now on we assume $E \in \Theta_L$ for Θ_L being defined as

$$\Theta_L := \left\{ \begin{array}{c|c} E := F + S & E \text{ is coercive} \vee E \text{ has bounded level sets} \\ F \in \mathcal{S}_L & \inf_u E(u) > -\infty \\ S \in \Gamma_0 & \mathrm{crit}(E) \neq \emptyset \end{array} \right\} .$$

We want to minimise E iteratively in a way that allows us to follow solution paths of different regularity. This regularity will be induced by an additional functional $J \in \Gamma_0$. Precisely, we approach the minimisation of (P) via the linearised Bregman iteration

$$u^{k+1} = \operatorname*{arg\,min}_{u\in\mathcal{U}} \left\{\tau^k \left(\langle \nabla F(u^k), u - u^k\rangle + S(u)\right) + D_J^{p^k}(u, u^k)\right\} , \tag{9a}$$

$$s^{k+1} \in \partial S(u^{k+1}) \,, \tag{9b}$$

$$p^{k+1} = p^k - \tau^k \left(\nabla F(u^k) + s^{k+1}\right) , \tag{9c}$$

for $k \in \{0, \ldots, N\}$, a sequence of positive parameters $\{\tau^k\}_{k\in\{1,\ldots,N\}}$ and initial values u^0 and p^0 with $p^0 \in \partial J(u^0)$. Note that (9c) is simply the optimality condition of (9a). If J is Fréchet-differentiable, ∂J is single-valued and we do not have to compute (9c) as we not need to pick a specific element from the set. However, if ∂J is multivalued, (9c) guarantees $p^k \in \partial J(u^k)$ for all $k \in \{0, \ldots, N\}$, and that we maintain a subgradient descent. This general form of linearised Bregman iteration for the minimisation of non-convex functionals is summed up in Algorithm 1.

REMARK 4. *For $J(u) = \frac{1}{2}\|u\|^2_{L^2(\Omega)}$, for some domain $\Omega \subset \mathbb{R}^n$, and $S \equiv 0$, (9) (and therefore also Algorithm 1) reduce to classical gradient descent. Hence, the linearised Bregman iteration is indeed a generalisation of gradient descent.*

Algorithm 1 Generalised linearised Bregman iteration for minimising (P)

Initialise N, $\{\tau^k\}_{k\in\{0,\dots,N-1\}}$, u^0 and $p^0 \in \partial J(u^0)$
for $k = 0, 1, \dots, N-1$ **do**
  Compute $u^{k+1} = \arg\min_{u\in\mathcal{U}} \left\{ \tau^k \left(\langle u - u^k, \nabla F(u^k) \rangle + S(u) \right) + D_J^{p^k}(u, u^k) \right\}$
  Pick $s^{k+1} \in \partial S(u^{k+1})$
  Compute $p^{k+1} = p^k - \tau^k \left(\nabla F(u^k) + s^{k+1} \right)$
end for
return u^N, p^N

Based on what has become known as the Bregman iteration [24, 68, 33, 42, 60], the linearised Bregman iteration has initially been proposed in [30] for the computation of sparse solutions of underdetermined linear systems of equations. It has been extensively studied in this context (cf. [74, 22, 21]) and also in the context of the minimisation of more general convex functionals (see [73]). It has further been analysed in the context of non-linear inverse problems in [4]. In [10] the linearised Bregman iteration has been studied in the context of minimising general smooth but non-convex functionals. Algorithm 1 allows us to pursue different solution paths for the minimisation of E over $\mathcal{U}$ depending on the choice of J. Note that we can also reformulate (9a) as follows:

$$u^{k+1} = \arg\min_{u\in\mathcal{U}} \left\{ \tau^k \left(\left\langle \nabla F(u^k) - \frac{1}{\tau^k} p^k, u - u^k \right\rangle + S(u) \right) + J(u) \right\} . \tag{10}$$

In order to ensure that a solution of Update (10) (respectively (9a)) exists, we choose J such that $J(u) + \tau^k(S(u) + \langle u^*, u\rangle)$ is coercive for all $u^* \in \mathcal{U}^*$. In particular, we choose J to be of the form $J = H + \tau^k R$, where H and R are functionals that satisfy $H \in \Phi_{\gamma,\delta}$ and $R \in \Gamma_0$. For this particular choice the iterates (9) read as

$$\begin{aligned} u^{k+1} &= \arg\min_{u\in\mathcal{U}} \left\{ \tau^k \left(\langle \nabla F(u^k), u - u^k \rangle + S(u) + D_R^{q^k}(u, u^k) \right) + D_H(u, u^k) \right\} , \\ &= \left(\nabla H + \tau^k \partial(R + S) \right)^{-1} \left(\nabla H(u^k) + \tau^k \left(q^k - \nabla F(u^k) \right) \right) , && \text{(11a)} \\ s^{k+1} &\in S(u^{k+1}) && \text{(11b)} \\ q^{k+1} &= q^k - \frac{1}{\tau^k} \left(\nabla H(u^{k+1}) - \nabla H(u^k) + \tau^k \left(\nabla F(u^k) + s^{k+1} \right) \right) , && \text{(11c)} \end{aligned}$$

for $q^k \in \partial R(u^k)$. Note that (11c) can be written as

$$q^k = q^0 - \sum_{n=0}^{k-1} \left[\frac{1}{\tau^n} (\nabla H(u^{n+1}) - \nabla H(u^n)) \right] - \sum_{n=0}^{k-1} \left[\nabla F(u^n) + s^{n+1} \right] , \tag{12}$$

and hence, for constant stepsize $\tau^k = \tau$ (11a) can be written as

$$u^k = \left(\nabla H + \tau \partial(R+S) \right)^{-1} \left(\nabla H(u^0) + \tau q^0 - \tau \left(\sum_{n=0}^{k-1} \nabla F(u^n) + \sum_{n=1}^{k-1} s^n \right) \right) . \tag{13}$$

Equations (11) are summarised in Algorithm 2.

REMARK 5. *We want to emphasise that Algorithm 2 is a generalisation of several well-established and some new algorithms: for the choices*

Algorithm 2 Specialised linearised Bregman iteration for minimising (P)

Initialise N, $\{\tau^k\}_{k\in\{0,\dots,N-1\}}$, u^0 and $q^0 \in \partial R(u^0)$
for $k = 0, 1, \dots, N-1$ **do**
  Get $u^{k+1} = \left(\nabla H + \tau^k \partial(R+S)\right)^{-1}\left(\nabla H(u^k) + \tau^k\left(q^k - \nabla F(u^k)\right)\right)$
  Pick $s^{k+1} \in \partial S(u^{k+1})$
  Compute $q^{k+1} = q^k - \frac{1}{\tau^k}\left(\nabla H(u^{k+1}) - \nabla H(u^k) + \tau^k\left(\nabla F(u^k) + s^{k+1}\right)\right)$
end for
return u^N, q^N

1. $H = \frac{1}{2}\|\cdot\|_2^2$, $R \equiv 0$ *and* $S = \chi_C(\cdot)$, *where* χ_C *denotes the characteristic function over a convex set* C, *we obtain the* projected gradient descent method *[37, 38, 13]*

$$u^{k+1} = proj_C(u^k - \tau^k \nabla F(u^k))\,.$$

2. $H = \frac{1}{2}\|\cdot\|_2^2$ *and* $R \equiv 0$ *we obtain the* proximal gradient descent method, *respectively the* forward-backward-splitting *[50]*

$$u^{k+1} = (I + \tau^k \partial S)^{-1}(u^k - \tau^k \nabla F(u^k))\,.$$

3. $R \equiv 0$ *we obtain the* Bregman Proximal Gradient *[16]*

$$u^{k+1} = \left(\nabla H + \tau^k \partial S\right)^{-1}\left(\nabla H(u^k) - \tau^k \nabla F(u^k)\right)\,.$$

For convex F *and the choices* $H(u) = \sum_{j=1}^n u_j \log(u_j) - u_j$ *and* $S = \chi_C$ *being the characteristic function over the simplex constraint* C, *we particularly derive the* (entropic) mirror descent *[55, 8]*

$$u^{k+1} = \operatorname*{arg\,min}_{u\in C}\left\{\sum_{j=1}^n [u_j \log(u_j) - u_j] - \tau^k \langle u, \nabla F(u^k)\rangle\right\}\,,$$

for which we obtain the component-wise update

$$u_i^{k+1} = \frac{u_i^k \exp\left(-\tau^k \nabla F(u^k)_i\right)}{\sum_{j=1}^n u_j^k \exp\left(-\tau^k \nabla F(u^k)_j\right)}\,,$$

for all $i \in \{1, \dots, n\}$.

4. $H = \frac{1}{2}\|\cdot\|_{\mathcal{U}}^2$, $S \equiv 0$ *and* $F = \frac{1}{2}\|K(\cdot) - f^\delta\|_{\mathcal{H}}^2$, *where* $K : \mathcal{U} \to \mathcal{H}$ *denotes a nonlinear operator (mapping from Hilbert space* $\mathcal{U}$ *to Hilbert space* $\mathcal{H}$*) and* $f^\delta \in \mathcal{H}$ *an element in the Hilbert space* $\mathcal{H}$, *we obtain the* Landweber-type iteration *[4]*

$$u^{k+1} = \operatorname*{arg\,min}_{u\in \mathrm{dom}(K)}\left\{\frac{1}{2}\left\|u - \left(u^k + \tau^k\left(q^k - \left(K'(u^k)^*\left(K(u^k) - f^\delta\right)\right)\right)\right)\right\|_{\mathcal{U}}^2 + \tau^k R(u)\right\},$$

$$q^{k+1} = q^k - \frac{1}{\tau^k}\left(u^{k+1} - u^k + \tau^k K'(u^k)^*\left(K(u^k) - f^\delta\right)\right)\,.$$

5. $S = \lambda R$ *for* $\lambda > 0$ *we obtain the* regularised linearised Bregman iteration

$$u^{k+1} = \left(\nabla H + \tau^k(1+\lambda)\partial R\right)^{-1}\left(\nabla H(u^k) + \tau^k\left(q^k - \nabla F(u^k)\right)\right)\,,$$

$$q^{k+1} = \frac{1}{1+\lambda}\left(q^k - \frac{1}{\tau^k}\left(\nabla H(u^{k+1}) - \nabla H(u^k) + \tau^k \nabla F(u^k)\right)\right)\,.$$

In the following we prove decrease properties of Algorithm 2 before we prove global convergence in a finite dimensional setting.

4. Properties of Algorithm 2. We begin our convergence analysis of Algorithm 2 by showing a sufficient decrease property of a surrogate for the energy E defined in (P) and a bound of the subgradients of that surrogate functional by the norm of the iterates gap. In order to do so, we first define the following surrogate functional for E.

DEFINITION 4.1 (Surrogate objective). *Assume $F \in \mathcal{S}_L$, $H \in \Phi_{\gamma,\delta}$ and $R \in \Gamma_0$. For an element $v \in \mathcal{U}$ and a corresponding subgradient $q \in \partial R(v)$ we define a surrogate objective $E^q : \mathcal{U} \times \mathcal{U} \to \mathbb{R} \cup \{\infty\}$ with*

$$E^q(u, v) := F(u) + S(u) + D_R^q(u, v)\,. \tag{14}$$

In addition we will make use of the concise notation $E^k(u) := E^{q^k}(u, u^k)$.

With the following theorem we prove a sufficient decrease property of the surrogate energy (14) for subsequent iterates.

THEOREM 4.2 (Sufficient decrease property). *Assume $E \in \Theta_L$, $R \in \Gamma_0$ and $H \in \Phi_{\gamma,\delta}$. Further, we assume that the stepsize τ^k satisfies the condition*

$$0 < \tau^k \leq \frac{\gamma C^{k+1}}{L + C^{k+1}\rho} \qquad \text{for} \qquad C^{k+1} := \frac{D_H^{symm}(u^{k+1}, u^k)}{D_H(u^{k+1}, u^k)}\,, \tag{15}$$

Then the iterates (11) satisfy the descent estimate

$$\begin{aligned} &E^k(u^{k+1}) + \rho D_H^{symm}(u^{k+1}, u^k) \\ &\quad + D_R^{q^{k+1}}(u^k, u^{k+1}) + D_R^{q^{k-1}}(u^k, u^{k-1}) \\ &\leq E^{k-1}(u^k)\,. \end{aligned} \tag{16}$$

In addition, we observe

$$\lim_{k\to\infty} D_H^{symm}(u^{k+1}, u^k) = 0 \quad \text{as well as} \quad \lim_{k\to\infty} D_R^{symm}(u^{k+1}, u^k) = 0\,.$$

Proof. First of all, we compute

$$\tau^k \left(\nabla F(u^k) + s^{k+1} + q^{k+1} - q^k\right) + \nabla H(u^{k+1}) - \nabla H(u^k) = 0$$

as the optimality condition of (11a), which is also the rearranged update formula (11c) as mentioned earlier (for $q^{k+1} \in \partial R(u^{k+1})$). Taking the dual product with $u^{k+1} - u^k$ therefore yields

$$\begin{aligned} -\langle \nabla F(u^k), u^{k+1} - u^k\rangle = \frac{1}{\tau^k} D_H^{\mathrm{symm}}(u^{k+1}, u^k) + D_R^{\mathrm{symm}}(u^{k+1}, u^k) \\ - \langle s^{k+1}, u^k - u^{k+1}\rangle\,. \end{aligned} \tag{17}$$

Due to the Lipschitz-continuity of the gradient of F we can use (7) and further estimate

$$F(u^{k+1}) \leq F(u^k) + \langle \nabla F(u^k), u^{k+1} - u^k\rangle + \frac{L}{\gamma} D_H(u^{k+1}, u^k)\,.$$

Together with (17) we therefore obtain the estimate

$$\begin{aligned} &F(u^{k+1}) - \langle s^{k+1}, u^k - u^{k+1}\rangle + D_R^{\mathrm{symm}}(u^{k+1}, u^k) \\ &+ \frac{1}{\tau^k} D_H^{\mathrm{symm}}(u^{k+1}, u^k) - \frac{L}{\gamma} D_H(u^{k+1}, u^k) \\ \leq\ & F(u^k)\,. \end{aligned}$$

Using the convexity estimate $S(u^{k+1}) - S(u^k) \leq -\langle s^{k+1}, u^k - u^{k+1}\rangle$, the stepsize bound (15) and adding $D_R^{q^{k-1}}(u^k, u^{k-1})$ to both sides of the inequality then allows us to conclude

$$\begin{aligned} &E^k(u^{k+1}) + D_R^{q^{k+1}}(u^k, u^{k+1}) + D_R^{q^{k-1}}(u^k, u^{k-1}) + \rho D_H^{\mathrm{symm}}(u^{k+1}, u^k) \\ \leq\ & E^{k-1}(u^k)\,. \end{aligned}$$

We have verified (16). Moreover, as $E(u^{k+1}) = E^k(u^{k+1}) - D_R^{q_k}(u^{k+1}, u^k)$ and $E(u^k) = E^{k-1}(u^k) - D_R^{q_{k-1}}(u^k, u^{k-1})$ we observe

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

hence, summing up over all N iterates and telescoping yields

$$\begin{aligned} \sum_{k=0}^{N} \left[\rho D_H^{\mathrm{symm}}(u^{k+1}, u^k) + D_R^{\mathrm{symm}}(u^{k+1}, u^k)\right] &\leq \sum_{k=0}^{N} E(u^k) - E(u^{k+1})\,, \\ &= E(u^0) - E(u^{N+1})\,, \\ &\leq E(u^0) - \inf_u E(u) < \infty\,. \end{aligned}$$

Taking the limit $N \to \infty$ therefore implies

$$\sum_{k=0}^{\infty} \left[\rho D_H^{\mathrm{symm}}(u^{k+1}, u^k) + D_R^{\mathrm{symm}}(u^{k+1}, u^k)\right] < \infty\,,$$

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

REMARK 6. *At first glance it looks as if we require a backtracking approach to ensure (15). However, we observe $1 \leq C^k \leq const < \infty$ for all $k \in \mathbb{N}$ and can therefore obtain an estimate for each τ^k that does not depend on the upcoming iterate.*

REMARK 7. *Since H is γ-strongly convex w.r.t. the $\mathcal{U}$-norm, Theorem 4.2 automatically implies*

$$\begin{aligned} &E^k(u^{k+1}) + \rho_1 \|u^{k+1} - u^k\|_{\mathcal{U}}^2 \\ &+ D_R^{q^{k+1}}(u^k, u^{k+1}) + D_R^{q^{k-1}}(u^k, u^{k-1}) \\ \leq\ & E^{k-1}(u^k)\,, \end{aligned}$$

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

REMARK 8. *As Theorem 4.2 implies the monotonic decrease $E^k(u^{k+1}) \leq E^{k-1}(u^k)$, we already know that the sequence $\{E^{k-1}(u^k)\}_{k\in\mathbb{N}}$ is bounded.*

It is worth mentioning that the name sufficient decrease can be misleading in the context of Algorithm 2 as it is not unusual for specific choices of R that the functional value of E does not change for several iterations.

We conclude this section with a bound for the subgradients of the surrogate energy at the iterates computed with Algorithm 2.

THEOREM 4.3 (A subgradient lower bound for the iterates gap). *Let the same assumptions hold true as in Theorem 4.2. Then the iterates (11) satisfy*

$$\|\nabla F(u^k) + s^k + q^k - q^{k-1}\|_{\mathcal{U}^*} \leq \rho_2 \|u^k - u^{k-1}\|_{\mathcal{U}}\,, \tag{18}$$

for $s^k \in \partial S(u^k)$, $q^k \in \partial R(u^k)$, $q^{k-1} \in \partial R(u^{k-1})$, $\rho_2 := \left(L + \delta/\tau^{min}\right)$ and $k \in \mathbb{N}$.

Proof. From (11c) we observe

$$\begin{aligned}
&\left\|\nabla F(u^k) + s^k + q^k - q^{k-1}\right\|_{\mathcal{U}^*} \\
&= \left\|\nabla F(u^k) - \nabla F(u^{k-1}) + \frac{1}{\tau^{k-1}}\left(\nabla H(u^{k-1}) - \nabla H(u^k)\right)\right\|_{\mathcal{U}^*}\,, \\
&\leq \left(L + \frac{\delta}{\tau^{\min}}\right)\|u^k - u^{k-1}\|_{\mathcal{U}}\,,
\end{aligned}$$

where we have made use of the Lipschitz-continuity of the gradient of F as well as the δ-Lipschitz continuity of ∇H. ∎

REMARK 9. *We want to point out that the generalised Lipschitz-continuity as defined in Definition 2.5 is not necessary if $R \equiv 0$ and $S \equiv 0$. In that case it is easy to see that we can obtain the estimate*

$$\|\nabla F(u^k)\|_{\mathcal{U}^*} \leq \frac{\delta}{\tau^{min}}\|u^{k+1} - u^k\|_{\mathcal{U}}$$

instead of (18) (see also [10]), without the use of Lipschitz-continuity. For the sufficient decrease Theorem 4.2 it is already enough to choose H such that $G := cH - F$ is convex for all arguments and a fixed constant c. This observation has already been made and exploited in [5, 10, 16].

5. A global convergence result for Algorithm 2 in finite dimensions. To conclude our convergence analysis we prove global convergence of Algorithm 2 with the help of the Kurdyka-Łojasiewicz (KL) property [51, 46]. We therefore assume a finite dimensional setting $\mathcal{U} = \mathbb{R}^n$ for the remainder of this section. Note that the finite dimensional setting still enables us to use different norms and their dual norms. For the definition of the KL property we follow [15] and define a distance between sub-sets and elements of $\mathbb{R}^n$ first.

DEFINITION 5.1. *Let $\Omega \subset \mathbb{R}^n$ and $u \in \mathbb{R}^n$. We define the distance from Ω to u as*

$$\operatorname{dist}(u, \Omega) := \begin{cases} \inf\{\|v - u\| \,|\, v \in \Omega\} & \Omega \neq \emptyset \\ \infty & \Omega = \emptyset \end{cases}.$$

Here $\|\cdot\|$ denotes the Euclidean norm.

Next, we recall the definition of the KL property based on the distance measure defined in Definition 5.1.

DEFINITION 5.2 (Kurdyka-Łojasierwicz property). *Let* dist *be defined as in Definition 5.1, and let $E \in \Theta_L$. Further, we assume there exists $\eta \in (0, \infty]$ such that $\varphi : [0, \eta) \to \mathbb{R}_{>0}$ is a concave function that satisfies $\varphi(0) = 0$, $\varphi \in C^1((0, \eta))$, is continuous at 0, and satisfies $\varphi'(s) > 0$ for all $s \in (0, \eta)$.*

- *Then the functional E is said to have the* Kurdyka-Łojasierwicz (KL) property *at $\overline{u} \in dom(\partial E) := \{u \in \mathbb{R}^n \,|\, \partial E(u) \neq \emptyset\}$ if there exists a neighbourhood N of $\overline{u}$ such that for all $u \in N \cap \{u \in \mathbb{R}^n \,|\, E(\overline{u}) < E(u) < E(\overline{u}) + \eta\}$ the inequality*

$$\varphi'\left(E(u) - E(\overline{u})\right) \operatorname{dist}(0, \partial E(u)) \geq 1 \tag{KL}$$

  holds.
- *If E satisfies the KL property at each point of* $\operatorname{dom}(\partial E)$*, E is called a* KL function.

Due to the equivalence of norms in finite dimensions, we can easily replace the Euclidean norm in (KL) with any other norm and obtain the same estimate by simply multiplying the function φ with the inverse of the appropriate norm-equivalence constant.

In order to apply the KL property, we have to verify some properties of the set of limit points. Let $\{u^k\}_{k\in\mathbb{N}}$ and $\{q^k\}_{k\in\mathbb{N}}$ be two sequences generated by Algorithm 2 from starting points u^0 and q^0 with $q^0 \in \partial R(u^0)$. The set of limit points is defined as

$$\omega(u^0) := \left\{ \overline{u} \in \mathbb{R}^n \;\middle|\; \text{there exists an increasing sequence of integers } \{k_j\}_{j\in\mathbb{N}} \text{ such that } \lim_{j\to\infty} u^{k_j} = \overline{u} \right\}.$$

First we show global convergence of Algorithm 2 without further assumptions on R. We then see from a toy example that we cannot guarantee convergence to a critical point of (P) without additional assumptions on R. A sufficient assumption is boundedness of the subgradients. This additional requirement will allow us to show convergence of Algorithm 2 to a critical point of (P).

For the remainder of this section we assume $\nabla F(u^k) \in \{\nabla F(u^k)\} + \partial S(u^k)$, for $\{u^k\}_{k\in\mathbb{N}}$ generated by Algorithm 2, and in particular $\nabla F(\overline{u}) \in \{\nabla F(\overline{u})\} + \partial S(\overline{u})$, for $\overline{u} \in \omega(u^0)$. This is obviously satisfied for $S \equiv 0$, but also for characteristic functions S over some convex domain such that the iterates and limiting points always lie within that set.

5.1. Global convergence of Algorithm 2. The following lemma guarantees that for a sequence converging to a limit point we also know that the surrogate objective converges to the objective evaluated at this limit point.

LEMMA 5.3. *Suppose $F \in \Theta_L$, $R \in \Gamma_0$ and $H \in \Phi_{\gamma,\delta}$, and let $\overline{u} \in \omega(u^0)$. Then we already know*

$$\lim_{k\to\infty} E^{k-1}(u^k) = F(\overline{u}). \tag{19}$$

Proof. Since $\overline{u}$ is a limit point of $\{u^k\}_{k\in\mathbb{N}}$ we know that there exists a subsequence

$\{u^{k_j}\}_{j\in\mathbb{N}}$ with $\lim_{j\to\infty} u^{k_j} = \overline{u}$. Hence, we immediately obtain

$$\begin{aligned}\lim_{j\to\infty} E^{k_j-1}(u^{k_j}) &= \lim_{j\to\infty} \left\{ F(u^{k_j}) + D_R^{q^{k_j-1}}(u^{k_j}, u^{k_j-1}) \right\}, \\ &= F(\overline{u}) = E(\overline{u}),\end{aligned}$$

due to the continuity of F and $\lim_{q\to\infty} D_R^{q^{k_j-1}}(u^{k_j}, u^{k_j-1}) = 0$ as a result of Theorem 4.2. Since $E^{k-1}(u^k)$ is also monotonically decreasing according to Remark 8 and therefore bounded, we can further conclude (19) for all $k \in \mathbb{N}$ as a consequence of the monotone convergence theorem. ∎

In addition to Lemma 5.3 the following lemma states that $\omega(u^0)$ is a non-empty, compact and connected set, and that the objective E is constant on that set.

LEMMA 5.4. *The set $\omega(u^0)$ is a non-empty, compact and connected set, the objective E is constant on $\omega(u^0)$ and we have $\lim_{k\to\infty} \operatorname{dist}(u^k, \omega(u^0)) = 0$.*

Proof. This follows from [15, Lemma 5]. ∎

Before we prove our first global convergence result, we recall one more result from [15] that is necessary for successfully carrying out the convergence proof.

LEMMA 5.5 ([15, Lemma 6]). *Let Ω be a compact set and let $E \in \Theta_L$, and suppose that E is constant on Ω and satisfies (KL) at each point in Ω. Then there exist $\varepsilon > 0$, $\eta > 0$ and $\varphi \in C^1((0,\eta))$ that satisfies the same conditions as in Definition 5.2, such that for all $\overline{u} \in \Omega$ and all u in*

$$\{u \in \mathbb{R}^n \mid \operatorname{dist}(u,\Omega) < \varepsilon\} \cap \{u \in \mathbb{R}^n \mid E(\overline{u}) < E(u) < E(\overline{u}) + \eta\} \tag{20}$$

condition (KL) is satisfied.

Now we have all the ingredients to show the following global convergence result for the finite dimensional version of Algorithm 2.

THEOREM 5.6 (Finite length property). *Suppose that $E^q(u,v)$ as a function of u is a KL function in the sense of Definition 5.2 for any fixed $v \in \mathbb{R}^n$ and $q \in \partial R(v)$, and assume $R \in \Gamma_0$ and $H \in \Phi_{\gamma,\delta}$. Let $\{u^k\}_{k\in\mathbb{N}}$ and $\{q^k\}_{k\in\mathbb{N}}$ be sequences generated by Algorithm (11). Then the sequence $\{u^k\}_{k\in\mathbb{N}}$ has finite length, i.e.*

$$\sum_{k=0}^{\infty} \|u^{k+1} - u^k\| < \infty. \tag{21}$$

Proof. We basically follow the steps of the proof of [15, Theorem 1] with minor modifications. Nevertheless, we present the proof in order to keep the paper self-contained.

The sequence $\{u^k\}_{k\in\mathbb{N}}$ is bounded, which follows from the assumption $F \in \Theta_L$ and the monotonic decrease. Thus, we know that there exists a convergent subsequence $\{u^{k_j}\}_{j\in\mathbb{N}}$ and $\overline{u} \in \mathbb{R}^n$ with

$$\lim_{j\to\infty} u^{k_j} = \overline{u}.$$

As a consequence of Lemma 5.3 we further know that $\lim_{k\to\infty} E^{k-1}(u^k) = E(\overline{u})$. If there exists an index $l \in \mathbb{N}$ with $E^{l-1}(u^l) = E(\overline{u})$ the results follow trivially. If there does not exist such an index, we observe that for any $\eta > 0$ there exists an index k_1 such that

$$E(\overline{u}) < E^{k-1}(u^k) < E(\overline{u}) + \eta$$

for all $k > k_1$, due to (19) and (16). In addition, for any $\varepsilon > 0$ there exists an index k_2 with

$$\operatorname{dist}(u^k, \omega(u^0)) < \varepsilon$$

for all $k > k_2$, due to Lemma 5.4. Hence, if we choose $l := \max(k_1, k_2)$, we know that u^k is in the set (20) for all $k > l$.

By Lemma 5.4, $\omega(u^0)$ satisfies all the assumptions of Lemma 5.5 and we have

$$1 \leq \varphi'(E^{k-1}(u^k) - E(\overline{u}))\ \operatorname{dist}(0, \partial E^{k-1}(u^k)) \tag{22}$$

for all $k > l$. This inequality makes sense due to $E^{k-1}(u^k) > E(\overline{u})$ for all k. From the concavity of φ we know that

$$\varphi'(x) \leq \frac{\varphi(x) - \varphi(y)}{x - y}$$

holds for all $x, y \in [0, \eta), x > y$, which we will use for the specific choices of $x = E^{k-1}(u^k) - E(\overline{u})$ and $y = E^k(u^{k+1}) - E(\overline{u})$. Combining the latter with Remark 7 and abbreviating

$$\varphi^k := \varphi(E^{k-1}(u^k) - E(\overline{u}))$$

yields

$$\varphi'(E^{k-1}(u^k) - E(\overline{u})) \leq \frac{\varphi^k - \varphi^{k+1}}{E^{k-1}(u^k) - E^k(u^{k+1})} \leq \frac{\varphi^k - \varphi^{k+1}}{\rho_1 \|u^{k+1} - u^k\|^2}\,.$$

Here we have made use of the trivial estimate $\rho_1\|u^{k+1} - u^k\|^2 \leq \rho_1\|u^{k+1} - u^k\|^2 + D_R^{q^{k+1}}(u^k, u^{k+1}) + D_R^{q^{k-1}}(u^k, u^{k-1})$. Inserting this result and the subgradient bound (18) into the KL inequality (22) yields

$$\|u^{k+1} - u^k\|^2 \leq \frac{\rho_2}{\rho_1}(\varphi^k - \varphi^{k+1})\|u^k - u^{k-1}\|\,.$$

Taking the square root, multiplying by 2 and using Young's inequality of the form $2\sqrt{ab} \leq a + b$ then yields

$$2\|u^{k+1} - u^k\| \leq \frac{\rho_2}{\rho_1}(\varphi^k - \varphi^{k+1}) + \|u^k - u^{k-1}\|\,.$$

Subtracting $\|u^{k+1} - u^k\|$ and summing from $k = 1, \ldots, N$ leads to

$$\begin{aligned}\sum_{k=1}^{N} \|u^{k+1} - u^k\| &\leq \frac{\rho_2}{\rho_1}(\varphi^1 - \varphi^{N+1}) + \|u^1 - u^0\| - \|u^{N+1} - u^N\| \\ &\leq \frac{\rho_2}{\rho_1}\varphi^1 + \|u^1 - u^0\| < \infty\,,\end{aligned}$$

and hence, we obtain the finite length property by taking the limit $N \to \infty$. ∎

Corollary 5.7 (Convergence). *Under the same assumptions as Theorem 5.6, the sequence $\{u^k\}_{k\in\mathbb{N}}$ converges.*

Proof. The finite length property Theorem 5.6 implies $\sum_{k=l}^{\infty} \|u^{k+1} - u^k\| \to 0$ for $l \to \infty$. Thus, for any $s \geq r \geq l$ we have

$$\|u^s - u^r\| = \left\|\sum_{k=r}^{s-1} u^{k+1} - u^k\right\| \leq \sum_{k=r}^{s-1} \|u^{k+1} - u^k\| \leq \sum_{k=l}^{\infty} \|u^{k+1} - u^k\|\,.$$

This shows that $\{u^k\}_{k\in\mathbb{N}}$ is a Cauchy sequence and, thus, is convergent. ∎

It is important to point out that Theorem 5.6 does imply convergence of Algorithm 2, but not necessarily convergence to a critical point of E, respectively F. The following remark will show that there is a simple counterexample for which Algorithm 2 does not converge to a critical point of (P).

REMARK 10. *Let* $F(u) = (u+1)^2/2$, $H(u) = u^2/2$, *and* $R(u) = \chi_{\geq 0}(u)$ *with*

$$\chi_{\geq 0}(u) := \begin{cases} 0 & u \geq 0 \\ \infty & u < 0 \end{cases},$$

be functions on $\mathcal{U} = \mathbb{R}$. *It is obvious that the only critical point of* $E = F$ *is* $\hat{u} = -1$. *However, Algorithm 2 can never converge to that point but will converge to* $\overline{u} = 0$ *due to the choice of* R. *This can be seen for instance for the choices* $u^0 > 0$, $q^0 = 0$ *and* $\tau^k = 1$. *Then the subsequent iterates are* $u^k = 0$ *and* $q^k = u^0 - k$, *thus,* $u^k \to 0$ *and* $q^k \to -\infty$.

For convex, quadratic fidelity terms (such as F in the example above) it is sufficient to satisfy a source condition of the form $\partial R(\hat{u}) \neq \emptyset$ (which in Remark 10 is clearly violated) in order to guarantee boundedness of the subgradients, see for instance [35]. For general non-convex terms F it is not straight forward to adapt the concept of source conditions, which is why we are simply going to assume boundedness of the subgradients instead. In the next subsection we will verify that this is enough to guarantee global convergence to a critical point.

5.2. Global convergence of Algorithm 2 to a critical point. For the remainder of this section we assume $R \in \Xi$, with Ξ being defined as

$$\Xi := \{ R \in \Gamma_0 \mid \exists C(u) \in \mathbb{R},\, \|q\| \leq C(u),\, q \in \partial R(u),\, \forall u \in \mathbb{R}^n \}\,.$$

Boundedness is not a very restrictive requirement as it is for instance satisfied for the large class of Lipschitz-continuous functions.

PROPOSITION 5.8. *Let* $R \in \Gamma_0$ *be a proper, convex and Lipschitz continuous function in the sense of* (4), *i.e.*

$$|R(u) - R(v)| \leq L\|u - v\|$$

is true for some constant $L > 0$ *and all arguments* $u, v \in \mathbb{R}^n$. *Then any subgradient* $q \in \partial R(v)$ *for any argument* $v \in \mathbb{R}^n$ *is bounded. Hence,* $R \in \Xi$.

Proof. From the convexity of R we automatically observe

$$|\langle q, u - v\rangle| \leq |R(u) - R(v)| \leq L\|u - v\|$$

for $q \in \partial R(v)$. If we choose $u = v + q$ we obtain $\|q\| \leq L$. Hence, $R \in \Xi$. ∎

The difference to the previous section is that we are now also able to show that $\omega(u^0) \subset \mathrm{crit}(E)$, with the help of the following lemma.

LEMMA 5.9. *Suppose* $E \in \Theta_L$, $R \in \Xi$, *and* $H \in \Phi_{\gamma,\delta}$. *Then* $\omega(u^0) \subset \mathrm{crit}(E)$.

Proof. Let $u^* \in \omega(u^0)$, which means $\lim_{k\to\infty} u^k = u^*$. Assume that $\nabla E(u^*) = \nabla F(u^*) + s^* = \nabla F(u^*) \neq 0$, $0 = s^* \in \partial S(u^*)$. It follows from the subgradient update (12) and the reverse triangle inequality $\|a + \sum_i a_i\| \geq \|a\| - \sum_i \|a_i\|$ that

$$\|q^k\| \geq \left\| \sum_{n=0}^{k-1} \nabla F(u^*) \right\| - \|q^0\| - \frac{1}{\tau^k}\|\nabla H(u^k) - \nabla H(u^0)\| - \sum_{n=0}^{k-1} \|\nabla F(u^n) - \nabla F(u^*)\|\,.$$

As $u^k \to u^*$, there exists $k_1, k_2 \in \mathbb{N}$ such that for all $k \geq k_1$ $\|\nabla H(u^k) - \nabla H(u^*)\| \leq 1$ and for all $k \geq k_2$ $\|\nabla F(u^k) - \nabla F(u^*)\| \leq \|\nabla F(u^*)\|/2$ holds, and therefore for all $k \geq \max(k_1, k_2)$ we have

$$\begin{aligned}
&-\frac{1}{\tau^k}\|\nabla H(u^k) - \nabla H(u^0)\| - \sum_{n=0}^{k-1} \|\nabla F(u^n) - \nabla F(u^*)\| \\
&\geq -\frac{1}{\tau^{\min}}\|\nabla H(u^k) - \nabla H(u^*)\| - \frac{1}{\tau^{\min}}\|\nabla H(u^*) - \nabla H(u^0)\|) \\
&\quad - \sum_{n=0}^{k_2-1} \|\nabla F(u^n) - \nabla F(u^*)\| - \sum_{n=k_2}^{k-1} \|\nabla F(u^n) - \nabla F(u^*)\| \\
&\geq -k\|\nabla F(u^*)\|/2 + \text{const}
\end{aligned}$$

with a constant independent of k. Combining these two estimates yields a contradiction to the boundedness of q^k

$$\|q^k\| \geq \left\| \sum_{n=0}^{k-1} \nabla F(u^*) \right\| - k\|\nabla F(u^*)\|/2 + \text{const} = k\|\nabla F(u^*)\|/2 + \text{const}\,.$$

Thus $\nabla F(u^*) = 0$ which means $u^* \in \operatorname{crit}(F) = \operatorname{crit}(E)$. ∎

Now we can finally prove that – under the additional assumption of boundedness of the subgradients – Algorithm 2 converges to a critical point.

THEOREM 5.10 (Global convergence to a critical point). *Suppose that $E^q(\cdot, v)$ is a KL function in the sense of Definition 5.2 for any fixed $v \in \mathbb{R}^n$ and $q \in \partial R(v)$, and assume $R \in \Gamma_0$, and $H \in \Phi_{\gamma,\delta}$. Let $\{u^k\}_{k\in\mathbb{N}}$ and $\{q^k\}_{k\in\mathbb{N}}$ be sequences generated by (11) that are both assumed to be bounded. Then the sequence $\{u^k\}_{k\in\mathbb{N}}$ converges to a critical point $\hat{u} \in \mathbb{R}^n$ of E as defined in (P).*

Proof. This result now follows in the same fashion as in the proof of Theorem 5.6, but with Lemma 5.9 concluding that the limit point is already a critical point of E. ∎

This concludes the theoretical analysis of Algorithm 2. In the following two sections we are going to discuss three applications, their mathematical modelling in the context of Algorithm 2 and their numerical results.

6. Applications. We demonstrate the capabilities of the linearised Bregman iteration by using it to approximately minimise several non-convex minimisation problems. We say approximately, as we do not exactly minimise the corresponding objective functionals, but rather compute iteratively regularised solutions to the associated inverse problems via early stopping of the iteration.

6.1. Parallel Magnetic Resonance Imaging. In (standard) Magnetic Resonance Imaging (MRI) the goal is to recover the spin-proton density from sub-sampled Fourier measurements that were obtained with a single radio-frequency (RF) coil. In parallel MRI, multiple RF coils are used for taking measurements, thus allowing to recover the spin-proton density from more measurements compared to the standard case. This, however, comes at the cost of having to model the sensitivities of the individual RF coils w.r.t. the measured material. We basically follow the mathematical modelling of [64, 69] and describe the recovery of the spin-proton density and the RF

coil sensitivities as the minimisation of the following energy functional:

$$E(u, b_1, \dots, b_s) := \frac{1}{2} \sum_{j=1}^{s} \|\mathcal{S}(\mathcal{F}((K(u, b_1, \dots, b_s))_j)) - f_j\|_2^2 + S(u, b_1, \dots, b_s)\,. \tag{23}$$

Here $\mathcal{F} \in \mathbb{C}^{n\times n}$ is the (discrete) Fourier transform, $\mathcal{S} \in \{0,1\}^{m\times n}$ is a sub-sampling operator, K is the non-linear operator $K(u, b_1, \dots, b_s) = (ub_1, ub_2, \dots, ub_s)^T$, u denotes the spin-proton density, $b_1, b_2, \dots, b_s$ the s coil sensitivities, $f_1, \dots, f_s$ the corresponding sub-sampled k-space data and $S(u, b_1, \dots, b_s) := \chi_{\|\cdot\|_\infty \le c_0}(u) + \sum_{j=1}^{s} \chi_{\|\cdot\|_\infty \le c_j}(b_j)$ is a sum of characteristic functions

$$\chi_{\|\cdot\|_\infty \le c_j}(x) := \begin{cases} 0 & \|x\|_\infty \le c_j \\ \infty & \|x\|_\infty > c_j \end{cases}$$

for constants $c_0, c_1, \dots, c_s$. Since $\mathbb{C}$ has the same topology as $\mathbb{R} \times \mathbb{R}$, we can formally treat all variables as variables in $\mathbb{R}^{2n}$.

Algorithm 3 Linearised Bregman iteration for parallel MRI reconstruction

Initialise $\{f_j\}_{j\in\{1,\dots,s\}}$, $\{\alpha_j\}_{j\in\{0,\dots,s\}}$, $\{w_l\}_{l\in\{1,\dots,n\}}$ $\tau^0 > 0$, u^0, $q^0 \in \partial \mathrm{TV}(u^0)$, $\{b_j^0\}_{j\in\{1,\dots,s\}}$ and $\{r_j^0\}_{j\in\{1,\dots,s\}}$ with $r_j^0 \in \partial R_2(b_j^0)$ for all $j \in \{1, \dots, s\}$

while (24) is not violated **do**

$v_0^{k+1} = \sum_{j=1}^{s} \overline{b_j^k} \mathcal{F}^{-1} \mathcal{S}^T \left(\mathcal{S}\mathcal{F}(u^k b_j^k) - f_j\right)$

$u^{k+1} = \arg\min_{\|u\|_\infty \le c_0} \left\{ \frac{1}{2} \left\| u - \left(u^k + \tau^k \left(\alpha q^k - v_0^{k+1}\right)\right)\right\|_2^2 + \alpha \tau^k \,\mathrm{TV}(u) \right\}$

$q^{k+1} = q^k - \frac{1}{\alpha\tau} \left(u^{k+1} - u^k + \tau^k v_0^{k+1}\right)$

for $j \in \{1, \dots, s\}$ **do**

$v_j^{k+1} = \overline{u^k} \mathcal{F}^{-1} \mathcal{S}^T \left(\mathcal{S}\mathcal{F}(u^k b_j^k) - f_j\right)$

$b_j^{k+1} = \arg\min_{\|b_j\|_\infty \le c_j} \left\{ \frac{1}{2} \left\| u - \left(u^k + \tau^k \left(\alpha r_j^k - v_j^{k+1}\right)\right)\right\|_2^2 + \tau^k \alpha_j \sum_{l=1}^{n} w_l |(C b_j)_l| \right\}$

$r_j^{k+1} = r_j^k - \frac{1}{\alpha\tau^k} \left(b_j^{k+1} - b_j^k + \tau^k v_j^{k+1}\right)$

end for

$k = k + 1$

end while

return $u^N, \{b_j^N\}_{j\in\{1,\dots,s\}}, q^N, \{r_j^N\}_{j\in\{1,\dots,s\}}$

The inverse problem of parallel MRI has been subject in numerous research publications [65, 44, 12]. We follow a different methodology here and apply Algorithm 2 to approximately minimise (23) with the following configuration. We choose H to be

$$H(u, b_1, \dots, b_s) = \frac{1}{2} \left(\|u\|_2^2 + \sum_{j=1}^{s} \|b_j\|_2^2 \right),$$

and the function R to be

$$R(u, b_1, \dots, b_s) = R_1(u) + \sum_{j=1}^{s} R_2(b_j)\,,$$

with

$$R_1(u) = \alpha_0 \mathrm{TV}(u) = \alpha_0 \| \|\nabla u\|_2 \|_1 \text{ and } R_2(b_j) = \alpha_j \sum_{l=1}^{n} w_l \left| (C\, b_j)_l \right|,\ \forall j \in \{1, \dots, s\}\,.$$

Here ∇ denotes a discrete finite forward difference approximation of the gradient, C denotes the discrete two-dimensional cosine transform, $\{w_l\}_{l\in\{1,\dots,n\}}$ is a set of weighting-coefficients and $\alpha_0,\dots,\alpha_{s+1}$ are positive scaling parameters. Note that $R \in \Xi$ and hence, the subgradients of Algorithm 2 will remain bounded. Iterating Algorithm 2 leads to unstable minimisers of (23) in case the k-space data $f_1,\dots,f_s$ are noisy, which is why we are going to apply Morozov's discrepancy principle [54] as a stopping criterion to stop the iteration early (see also [60, 36, 52], and [67, 4, 41] in the context of nonlinear inverse problems), i.e. we stop the iteration as soon as

$$E(u, b_1, \dots, b_s) \leq \eta \tag{24}$$

is satisfied, for some $\eta > 0$. Usually η depends on the variance of the normal-distributed noise.

Due to the boundedness assumption we know that the differentiable part of (23) has a Lipschitz-continuous gradient. For the particular choices of E, H, R and S Algorithm 2 reads as Algorithm 3.

6.2. Blind deconvolution. In blind (image) deconvolution the goal is to recover an unknown image u from a blurred and noisy image f. Assuming that the degradation is the same for each pixel, the problem of blind deconvolution can be modelled as the minimisation of the energy $E(u,h) := F(u,h) + S(u,h)$ with

$$F(u,h) := \frac{1}{2}\|u * h - f\|_2^2\,,$$

where $*$ denotes the two-dimensional, discrete convolution, $f \in \mathbb{R}^n$ is the blurry (and noisy) image, $h \in \mathbb{R}^r$ is the convolution kernel (with $r \leq n$) and $u \in \mathbb{R}^n$ the image without blur.

Blind deconvolution is extensively discussed in the literature, e.g. [45, 27, 23] and the references therein. We make the assumptions that the blur-free image u is bounded and has low total variation and that the kernel h satisfies a simplex constraint, i.e. all entries are non-negative and sum up to one. In order to do so, we choose $S(u,h) := \chi_{\|\cdot\|_2\leq 1}(u) + \chi_C(h)$ with

$$\chi_{\|\cdot\|_2\leq 1}(x) := \begin{cases} 0 & \|x\|_2 \leq 1 \\ \infty & \|x\|_2 > 1 \end{cases} \quad \text{and} \quad \chi_C(h) := \begin{cases} 0 & x \in C \\ \infty & x \notin C \end{cases},$$

for the simplex constraint set

$$C = \left\{ h \in \mathbb{R}^r \;\middle|\; \sum_{j=1}^{r} h_j = 1,\ h_j \geq 0,\ \forall j \in \{1,\dots,r\} \right\}.$$

The assumption of low total variation can for instance be motivated by [28]. We want to point out however, that minimising E with some additional total variation regularisation does often not lead to visually satisfactory results. The reason for this is that blurry images naturally have a lower total variation compared to their sharp counterparts. As a consequence, we obtain reconstructions with trivial Dirac delta convolution kernels.

We want to avoid such results and therefore apply Algorithm 2 instead, which allows us to choose a smooth solution path with increasing total variation that yields non-trivial reconstructions if the chosen solution-path is regular enough. We therefore

apply Algorithm 2 with $H : \mathbb{R}^n \times \mathbb{R}^r \to \mathbb{R}$ chosen to be the quadratic two norm for the image and the negative entropy for the convolution kernel, i.e.

$$H(u,h) = \frac{1}{2}\|u\|_2^2 + \sum_{j=1}^{r} L(h_j)\,,$$

where $L : \mathbb{R} \to \mathbb{R}$ is defined as

$$L(x) = \begin{cases} x\log(x) - x & x > 0 \\ 0 & x \leq 0 \end{cases}.$$

We know that H is 1-strongly convex in the first argument w.r.t. the 2-norm and 1-strongly convex in the second argument w.r.t. the 1-norm for $h \in (0,1)^r$, see for instance [8]. It is straightforward to see that the strong convexity in the second argument can be extended to the closed domain $[0,1]$. We want to point out that $\partial_h H$ is not Lipschitz continuous around 0, which in theory could lead to vanishing stepsizes τ^k. With a non-zero initialisation this, however, is unlikely to happen as we are going to see in the results section. In addition, our theoretical framework in principle requires $H(\cdot, h)$ to be strongly convex on the entire Banach space $\mathbb{R}^r$. However, it is straight forward to adapt all arguments for interiors of convex subsets. We conclude the modelling aspect by defining $R : \mathbb{R}^n \times \mathbb{R}^r \to \mathbb{R}$ as

$$R(u,h) = \alpha \mathrm{TV}(u)\,,$$

for $\alpha > 0$.

As in the previous example, we have $R \in \Xi$. These particular choices of H, R and S result in the following proximal mapping for h^{k+1}:

$$h^{k+1} = \operatorname*{arg\,min}_{h \in C} \left\{ \sum_{j=1}^{r} L(h_j) - \sum_{j=1}^{r} h_j \left(\log(h_j^k) - \tau^k \left(u^k *^T (u^k * h^k - f) \right)_j \right) \right\},$$

where $*^T$ denotes the transpose operation of the discrete convolution operator. This minimisation problem has the closed-form solution (see [8])

$$h_j^{k+1} = \frac{h_j^k \exp\left(-\tau^k \left(u^k *^T (u^k * h^k - f) \right)_j \right)}{\sum_{i=1}^{r} \left[h_i^k \exp\left(-\tau^k \left(u^k *^T (u^k * h^k - f) \right)_i \right) \right]}\,.$$

Hence, Algorithm 2 reads as Algorithm 4 in the context of blind deconvolution.

The boundedness assumptions on both u and h ensure Lipschitz-continuity of the gradient of F, such that we can ensure convergence of Algorithm 4.

6.3. Classification. The last application that we want to discuss is the classification of images. Given a set $D \in \mathbb{R}^{s \times r}$ of r training images (with s pixel each) in column vector form, we want to train a neural network to classify those images. We do so by learning the parameters $(A_1, \ldots, A_l)$ of the l-layer neural network

$$\rho(x) := \rho_1(A_1 \rho_2(A_2 \ldots \rho_l(A_l x)) \ldots)$$

in a supervised fashion. Here the parameters $A_j \in \mathbb{R}^{m_j \times n_j}$ are matrices of different size, and the functions $\{\rho_j\}_{j=1}^{l}$ are so-called activity functions of the neural net.

Algorithm 4 Linearised Bregman iteration for blind deconvolution

Initialise f, $\{\tau^k\}_{k\in\mathbb{N}}$, u^0 with $\|u^0\|_2 \leq 1$, $q^0 \in \alpha\partial\mathrm{TV}(u^0)$ and $h^0 \in C$
while (24) is not violated **do**
  $v^{k+1} = (u^k * h^k - f) *^T h^k$
  $u^{k+1} = \arg\min_{\|u\|_2\leq 1} \left\{ \frac{1}{2} \left\| u - \left(u^k + \tau^k \left(q^k - v^{k+1}\right)\right)\right\|_2^2 + \alpha\tau^k \,\mathrm{TV}(u) \right\}$
  $q^{k+1} = q^k - \frac{1}{\tau^k}\left(u^{k+1} - u^k + \tau^k \left((u^k * h^k - f) *^T h^k\right)\right)$
  for $j \in \{1, \ldots, r\}$ **do**
    $h_j^{k+1} = \frac{h_j^k \exp\left(-\tau^k \left(u^k *^T (u^k * h^k - f)\right)_j\right)}{\sum_{i=1}^r h_i^k \exp\left(-\tau^k (u^k *^T (u^k * h^k - f))_i\right)}$
  end for
  $k = k + 1$
end while
return u^N, h^N, q^N

Typical choices for activity functions are max- and min-functions, also known as rectifier. However, due to their non-differentiability it is common to approximate them with either the pointwise smooth-max-function, i.e.

$$\rho_j(x, c, \beta) := \frac{x\exp(\beta x) + c\exp(\beta c)}{\exp(\beta x) + \exp(\beta c)},$$

for $x \in \mathbb{R}$ and constants $\beta, \in \mathbb{R}$, or the soft-max-function, i.e.

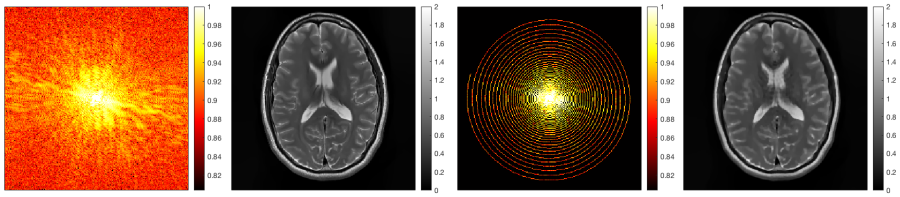
$$\rho_j(x)_i = \frac{\exp(x_i)}{\sum_{l=1}^m \exp(x_l)},$$

for $x \in \mathbb{R}^m$. The latter has the advantage that the function output automatically satisfies the simplex constraint. Defining a nonlinear operator $K(A_1, A_2, \ldots, A_l) := \rho(D) = \rho_1(A_1\rho_2(A_2 \ldots \rho_l(A_l D)) \ldots)$ for given matrix D and a given label matrix $Y \in \mathbb{R}^{m_1\times r}$, we aim to minimise

$$E(A_1, A_2, \ldots, A_l) := \mathcal{D}(K(A_1, A_2, \ldots, A_l), Y), \tag{25}$$

where $\mathcal{D} : \mathbb{R}^{m_1\times r} \times \mathbb{R}^{m_1\times r} \to \mathbb{R}$ denotes a functional that measures the distance between its arguments in some sense. Potential choices for $\mathcal{D}$ are the squared Frobenius norm $\mathcal{D}(X, Y) = \frac{1}{2}\|X - Y\|_{\mathrm{Fro}}^2$, the shifted Kullback-Leibler divergence $\mathcal{D}_\varepsilon(X, Y) = \sum_{i=1}^{m_1}\sum_{j=1}^{r}\left[(X_{ij} + \varepsilon)\log\left(\frac{X_{ij}+\varepsilon}{Y_{ij}+\varepsilon}\right) + Y_{ij} - X_{ij}\right]$, or the symmetrised shifted Kullback-Leibler divergence $\mathcal{D}_\varepsilon(X, Y) = \sum_{i=1}^{m_1}\sum_{j=1}^{r}\log\left(\frac{X_{ij}+\varepsilon}{Y_{ij}+\varepsilon}\right)(X_{ij} - Y_{ij})$.

As before, we aim to minimise (25) with Algorithm 2. This time we choose $H(A_1, \ldots, A_l) = \frac{1}{2}\sum_{j=1}^l \|A_j\|_{\mathrm{Fro}}^2$ and $R(A_1, \ldots, A_l) = \sum_{j=1}^l \alpha_j\|A_j\|_*$. Here $\{\alpha_j\}_j^l$ is a set of positive scaling parameters, and $\|X\|_* := \sum_{i=1}^{\mathrm{rank}(X)} \sigma_i$ is the one norm of the singular values $\{\sigma_i\}_{i=1}^{\mathrm{rank}(X)}$ of the argument X, also known as the nuclear norm. The rationale behind this choice for R is that we can create solution paths where the rank of the individual matrices are steadily increasing. This way we control the number of effective parameters and do not fit all parameters right from the start. Technically, we also have to add constraints to keep the parameters bounded, similarly to the previous two applications. However, we will discover in practice that boundedness is already enforced by the nuclear norm regularisation throughout the early iterations of Algorithm 2.



(a) Fully sampled (b) Recon. from 1a (c) Subsampled (d) Recon. from 1c

FIGURE 1. *Figure 1a shows a log-plot of the modulus of the fully sampled k-space data of the first coil taken from [44]. Figure 1b shows the reconstruction of the spin proton density from the data visualised in Figure 1a via Algorithm 3. In Figure 1c we see roughly 25 % of the k-space data visualised in Figure 1a, sampled on a spiral on a cartesian grid [11]. Figure 1d shows the reconstruction of the spin proton density from this subsampled k-space data with Algorithm 3.*

7. Numerical Results. We demonstrate the particular properties and idiosyncrasies of Algorithm 2 by computing several numerical solutions to the problems described in Section 6. All results have been computed with MATLAB R2017b. The code for the following examples is available at https://doi.org/10.17863/CAM.16931 and can be used under the Creative Commons Attribution (CC BY) license.

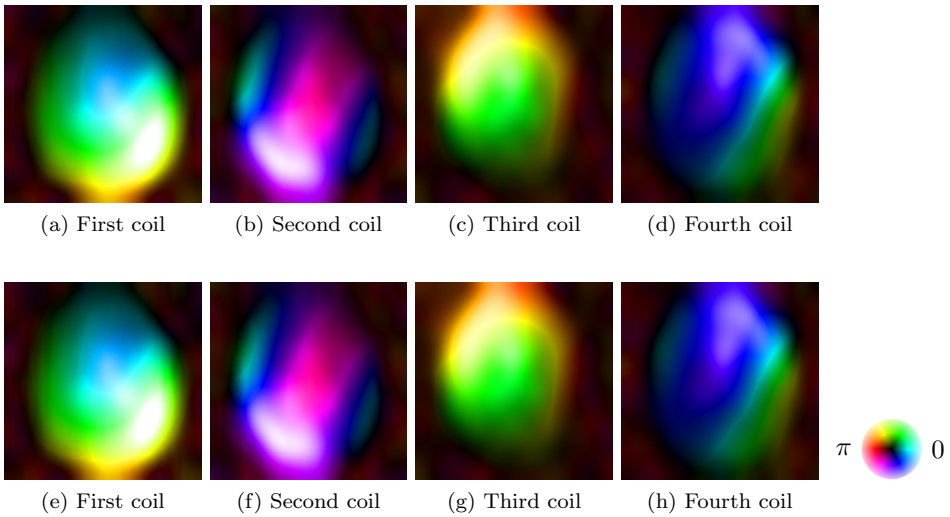
Since we do not use explicit Lipschitz constants, we employ a naïve backtracking strategy for the variable stepsize $\{\tau^k\}_{k\in\mathbb{N}}$. We start with an initial stepsize $\tau^0 > 0$ and check after each iteration whether $E(u^{k+1}) \leq E(u^k) + \varepsilon$ is satisfied. Here, $\varepsilon > 0$ is a small constant that accounts for numerical rounding errors that may cause $E(u^{k+1}) > E(u^k)$ when $E(u^{k+1}) \approx E(u^k)$. If the decrease is satisfied, we set $\tau^{k+1} = \tau^k$; otherwise we set $\tau^{k+1} = (3\tau^k)/4$ and backtrack again until we get a decrease.

7.1. Parallel MRI. We compute parallel MRI reconstructions from real k-space data. We use data from a T2-weighted TSE scan of a transaxial slice of a brain acquired with a four-channel head-coil in [43]. A reconstruction from fully sampled data is taken as a ground truth. The spiral sub-sampling is simulated by point-wise multiplication of the k-space data with the spiral pattern visualised in Figure 1c. We initialise with $u^0 = 2 \times \mathbf{1}^{65536\times 1}$ and $b_j^0 = \mathbf{1}^{65536\times 1}$ for $j \in \{1, \dots, 4\}$, and compute a $q^0 \in \partial R(u^0)$.

With the parameters $\alpha_j = 1$ for all $j \in \{0, \dots, s\}$, $\tau^0 = 1/2$, $w_1 = w_2 = w_{\sqrt{n}+1} = w_{\sqrt{n}+2} = 10^{-6}$ and $w_l = 5$ for $l \in \{1, \dots, n\} \setminus \{1, 2, \sqrt{n}+1, \sqrt{n}+2\}$, and $\eta = 3.45$ we obtain the spin proton density reconstruction visualised in Figure 1b, as well as the coil sensitivity reconstructions in Figure 2a - 2d. In Figure 1d and Figure 2e - 2h we show the results of the reconstructions from sub-sampled data using the sub-sampling scheme in Figure 1c.

7.2. Blind deconvolution. To simulate blurring of a gray-scale image $f_{\text{orig}} \in \mathbb{R}^{424\times 640}$ we subtract its mean, normalise it and subsequently blur f_{orig} with a motion-blur filter $h \in \mathbb{R}^{9\times 31}$. The filter was obtained with the MATLAB©-command `fspecial('motion', 30, 15)`, and we assume periodic boundary conditions for the blurring process. Subsequently we add normally distributed noise with mean zero and standard deviation $\sigma = 10^{-4}$ to obtain a blurry and noisy image f with ground truth f_{orig}. Both f_{orig} and f, as well as h are visualised in Figure 3.

We use f as our input image for Algorithm 4. We initialise Algorithm 4 with $u^0 = 0$ and $q^0 = 0$. We choose $h_0 = 1/(r^2) \times \mathbf{1}_{r\times r}$ for $r = 35$ to ensure that h_0 satisfies the simplex constraint. We set $\tau^0 = 2$ and pick $\alpha \in \{10^{-1}, 10^{-3}\}$. We then iterate



(a) First coil (b) Second coil (c) Third coil (d) Fourth coil

(e) First coil (f) Second coil (g) Third coil (h) Fourth coil

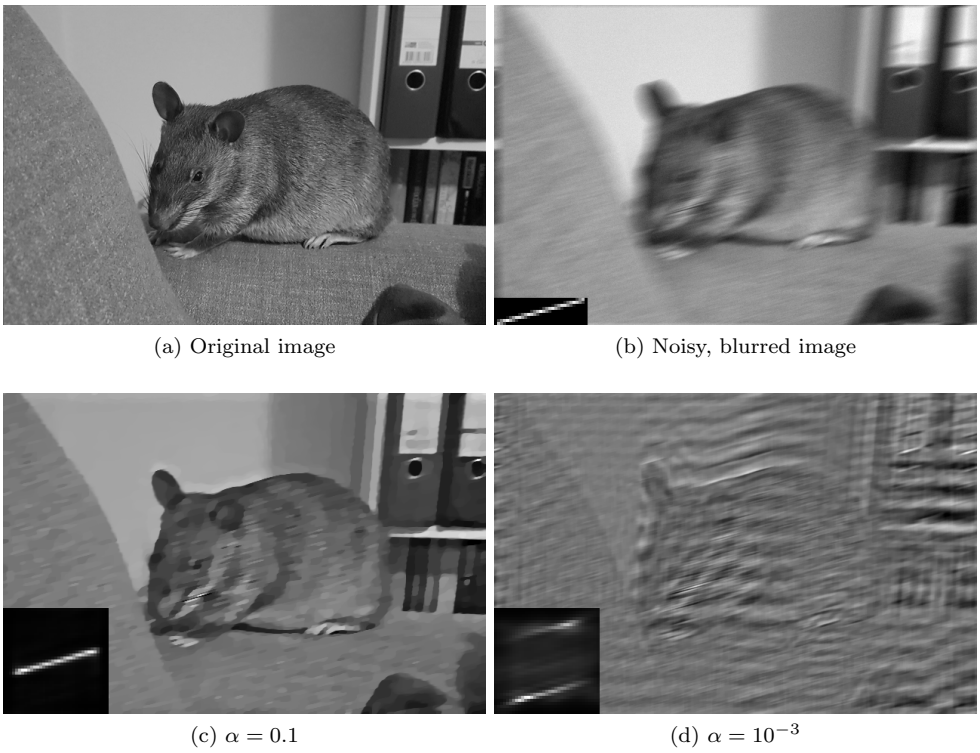
FIGURE 2. *Figure 2a - 2d show the reconstructions of the coil sensitivities from the fully sampled data. Figure 2e - 2h show the reconstructions of the same quantities from the sub-sampled data.*

Algorithm 4 until the discrepancy principle is violated for $\eta = (1.2\sigma^2)/(2\sqrt{424 \times 640})$. The inner total variation sub-problem is solved with the primal-dual hybrid gradient method [75, 61, 34, 25, 26]. The results are visualised in Figure 3.

7.3. Classification. We test the proposed framework for the classification of images of hand-written digits. We use the well-known MNIST dataset [47] as the basis for our classification. Ten example images of each class are visualised in Figure 4a. We pick 50000 images from the training dataset to create our training data matrix D, and use the remaining 10000 for cross validation. We model our classifier as a two-level neural network as described in Section 6.3. We choose the original rectifier activation functions for the networks' architecture. We overcome the non-differentiability by defining the derivatives to be zero at the non-differentiable points. This is consistent with the smooth-max approximation of the rectifier for $\beta \to \infty$. We choose E to be the squared Frobenius norm and set the scaling parameters to $\alpha_1 = \alpha_2 = 0.2$. The stepsize τ^0 is initialised with $\tau^0 = 10^{-3}$. Subsequently, we run Algorithm 2 for 10000 iterations. The prediction results of the classifier and the rank of the trained matrices are visualised in Figure 4.

8. Conclusions & Outlook. We have presented a generalisation of gradient descent that allows the incorporation of non-smooth Bregman distances, and therefore can also be seen as an extension of the linearised Bregman iteration to non-convex functionals. We have shown that the proposed method satisfies a sufficient decrease property and that the computed subgradients are bounded by the gap of the primal iterates. In finite dimensions, we have proven a global convergence result, where the limit is guaranteed to be a critical point of the energy if the subgradients are also bounded. The numerical experiments suggest that the solution path of the proposed method together with early stopping can be designed to obtain solutions superior to those attained with conventional variational regularisation methods.

There are several natural directions that can be explored from here. One could



(a) Original image (b) Noisy, blurred image

(c) $\alpha = 0.1$ (d) $\alpha = 10^{-3}$

FIGURE 3. *Figure 3a shows an image of Pixel the Gambian pouched rat, courtesy of Monique Boddington. Figure 3b shows a motion-blurred version of that image, together with some added normal distributed noise. The corresponding convolution kernel is depicted in the bottom left corner. Figure 3c visualises the reconstruction of the image and the convolution kernel with Algorithm 4 for the choice $\alpha = 10^{-1}$. Figure 3d show the reconstructions of the same quantities for the choice $\alpha = 10^{-3}$. We clearly see that a larger choice of α results in a regular solution, whereas a smaller α will mimic traditional gradient descent with almost no additional regularity of the reconstruction.*

study a linearised block coordinate variant of the proposed method, which would be similar in analysis to [71, 15]. In the wake of [58, 62], a generalisation of the proposed method to include inertial terms (or even multi-step inertial terms as in [49]) seems to be an intuitive approach for accelerating the method. Another direction that can be explored is the direction of non-smooth quasi-Newton extensions similar to [9]. Motivated by applications in deep learning, one could also follow up on incremental or stochastic variants of the proposed algorithm (cf. [40, 31, 14]).

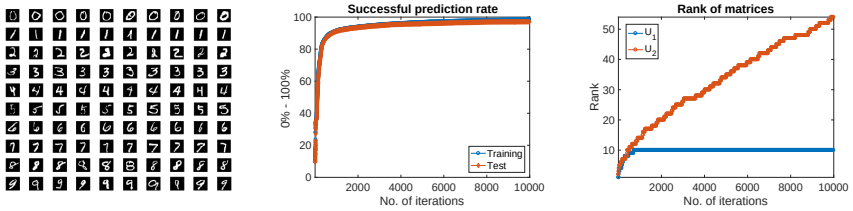
REFERENCES

[1] H. ATTOUCH AND J. BOLTE, *On the convergence of the proximal algorithm for nonsmooth functions involving analytic features*, Mathematical Programming, 116 (2009), pp. 5–16.

[2] H. ATTOUCH, J. BOLTE, P. REDONT, AND A. SOUBEYRAN, *Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the Kurdyka-łojasiewicz inequality*, Mathematics of Operations Research, 35 (2010), pp. 438–457.

[3] H. ATTOUCH, J. BOLTE, AND B. F. SVAITER, *Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized Gauss–Seidel methods*, Mathematical Programming, 137 (2013), pp. 91–129.

[4] M. BACHMAYR AND M. BURGER, *Iterative total variation methods for nonlinear inverse*



(a) MNIST [47] (b) Prediction (c) Rank

FIGURE 4. *Figure 4a shows ten randomly chosen images of each digit from the MNIST training data [47]. Figure 4b shows the successful prediction rate of the classifier throughout the iteration both for the training and the test data. Figure 4c shows the rank of the two matrices U_1 and U_2 that are reconstructed. It becomes evident that the rank is monotonically increasing throughout the course of the iteration, allowing the model to fit only a reduced no. of effective parameters at a time.*

problems, Inverse Problems, 25 (2009), p. 26, https://doi.org/10.1088/0266-5611/25/10/105004.

[5] H. H. Bauschke, J. Bolte, and M. Teboulle, *A descent lemma beyond Lipschitz gradient continuity: first-order methods revisited and applications*, Mathematics of Operations Research, 42 (2016), pp. 330–348.

[6] H. H. Bauschke, J. M. Borwein, and P. L. Combettes, *Essential smoothness, essential strict convexity, and Legendre functions in Banach spaces*, Communications in Contemporary Mathematics, 3 (2001), pp. 615–647.

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

[8] A. Beck and M. Teboulle, *Mirror descent and nonlinear projected subgradient methods for convex optimization*, Operations Research Letters, 31 (2003), pp. 167–175.

[9] S. Becker and J. Fadili, *A quasi-newton proximal splitting method*, in Advances in Neural Information Processing Systems, 2012, pp. 2618–2626.

[10] M. Benning, M. M. Betcke, M. J. Ehrhardt, and C.-B. Schönlieb, *Gradient descent in a generalised Bregman distance framework*, in Geometric Numerical Integration and its Applications, G. R. W. Quispel, P. Bader, D. I. McLaren, and D. Tagami, eds., vol. 74, MI Lecture Notes series of Kyushu University, April 2017, pp. 40–45, http://www.imi.kyushu-u.ac.jp/eng/files/imipublishattachment/file/math_58ec341a238fe.pdf.

[11] M. Benning, L. Gladden, D. Holland, C.-B. Schönlieb, and T. Valkonen, *Phase reconstruction from velocity-encoded MRI measurements–a survey of sparsity-promoting variational approaches*, Journal of Magnetic Resonance, 238 (2014), pp. 26–43.

[12] M. Benning, F. Knoll, C.-B. Schönlieb, and T. Valkonen, *Preconditioned ADMM with nonlinear operator constraint*, in IFIP Conference on System Modeling and Optimization, Springer, 2015, pp. 117–126.

[13] D. Bertsekas, *On the Goldstein-Levitin-Polyak gradient projection method*, IEEE Transactions on automatic control, 21 (1976), pp. 174–184.

[14] D. P. Bertsekas, *Incremental gradient, subgradient, and proximal methods for convex optimization: A survey*, in Optimization for Machine Learning, S. Sra, S. and Nowozin, S. and Wright, ed., MIT Press, 2011, pp. 85–120.

[15] J. Bolte, S. Sabach, and M. Teboulle, *Proximal alternating linearized minimization for nonconvex and nonsmooth problems*, Mathematical Programming, 146 (2014), pp. 459–494.

[16] J. Bolte, S. Sabach, M. Teboulle, and Y. Vaisbourd, *First order methods beyond convexity and Lipschitz gradient continuity with applications to quadratic inverse problems*, arXiv preprint arXiv:1706.06461, (2017).

[17] S. Bonettini, I. Loris, F. Porta, and M. Prato, *Variable metric inexact line-search based methods for nonsmooth optimization*, SIAM Journal on Optimization, 26 (2016), pp. 891–921, https://doi.org/http://dx.doi.org/10.1137/15M1019325, http://arxiv.org/abs/1506.00385.

[18] S. Bonettini, I. Loris, F. Porta, M. Prato, and S. Rebegoldi, *On the convergence of a linesearch based proximal-gradient method for nonconvex optimization*, Inverse Problems, (2017), https://doi.org/http://dx.doi.org/10.1088/1361-6420/aa5bfd, https://arxiv.

org/abs/1605.03791. Accepted.
[19] L. M. Bregman, *The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming*, USSR computational mathematics and mathematical physics, 7 (1967), pp. 200–217.
[20] M. Burger, E. Resmerita, and L. He, *Error estimation for Bregman iterations and inverse scale space methods in image restoration*, Computing, 81 (2007), pp. 109–135.
[21] J.-F. Cai, S. Osher, and Z. Shen, *Convergence of the linearized Bregman iteration for ℓ^1-norm minimization*, Mathematics of Computation, 78 (2009), pp. 2127–2136.
[22] J.-F. Cai, S. Osher, and Z. Shen, *Linearized Bregman iterations for compressed sensing*, Mathematics of Computation, 78 (2009), pp. 1515–1536.
[23] P. Campisi and K. Egiazarian, *Blind image deconvolution: theory and applications*, CRC press, 2016.
[24] Y. Censor and S. A. Zenios, *Proximal minimization algorithm with d-functions*, Journal of Optimization Theory and Applications, 73 (1992), pp. 451–464.
[25] A. Chambolle and T. Pock, *A first-order primal-dual algorithm for convex problems with applications to imaging*, Journal of mathematical imaging and vision, 40 (2011), pp. 120–145.
[26] A. Chambolle and T. Pock, *An introduction to continuous optimization for imaging*, Acta Numerica, 25 (2016), pp. 161–319.
[27] T. F. Chan and J. Shen, *Image Processing and Analysis: Variational, PDE, Wavelet, and Stochastic Methods*, Other titles in applied mathematics, Society for Industrial and Applied Mathematics, 2005.
[28] T. F. Chan and C.-K. Wong, *Total variation blind deconvolution*, IEEE Transactions on Image Processing, 7 (1998), pp. 370–375.
[29] P. L. Combettes and N. N. Reyes, *Moreau's decomposition in Banach spaces*, Mathematical Programming, 139 (2013), pp. 103–114.
[30] J. Darbon and S. Osher, *Fast discrete optimization for sparse approximations and deconvolutions*, (2007).
[31] A. Defazio, F. Bach, and S. Lacoste-Julien, *Saga: A fast incremental gradient method with support for non-strongly convex composite objectives*, Nips, (2014), pp. 1–12, https://arxiv.org/abs/arXiv:1407.0202v2.
[32] D. Drusvyatskiy, A. D. Ioffe, and A. S. Lewis, *Nonsmooth optimization using Taylor-like models: error bounds, convergence, and termination criteria*, arXiv preprint arXiv:1610.03446, (2016).
[33] J. Eckstein, *Nonlinear proximal point algorithms using Bregman functions, with applications to convex programming*, Mathematics of Operations Research, 18 (1993), pp. 202–226.
[34] E. Esser, X. Zhang, and T. F. Chan, *A general framework for a class of first order primal-dual algorithms for convex optimization in imaging science*, SIAM Journal on Imaging Sciences, 3 (2010), pp. 1015–1046.
[35] K. Frick and O. Scherzer, *Regularization of ill-posed linear equations by the non-stationary augmented lagrangian method*, The Journal of Integral Equations and Applications, (2010), pp. 217–257.
[36] G. Garrigos, L. Rosasco, and S. Villa, *Iterative regularization via dual diagonal descent*, arXiv preprint arXiv:1610.02170, (2016).
[37] A. A. Goldstein, *Convex programming in Hilbert space*, Bulletin of the American Mathematical Society, 70 (1964), pp. 709–710.
[38] A. A. Goldstein, *Constructive real analysis*, tech. report, Washington Univ. Seattle Dept. of Mathematics, 1967.
[39] T. Hohage and C. Homann, *A generalization of the chambolle-pock algorithm to banach spaces with applications to inverse problems*, arXiv preprint arXiv:1412.0126, (2014).
[40] R. Johnson and T. Zhang, *Accelerating Stochastic Gradient Descent using Predictive Variance Reduction*, Nips, 1 (2013), pp. 315–323.
[41] B. Kaltenbacher, F. Schöpfer, and T. Schuster, *Iterative methods for nonlinear ill-posed problems in banach spaces: convergence and applications to parameter identification problems*, Inverse Problems, 25 (2009), p. 065003.
[42] K. C. Kiwiel, *Proximal minimization methods with generalized Bregman functions*, SIAM journal on control and optimization, 35 (1997), pp. 1142–1168.
[43] F. Knoll, K. Bredies, T. Pock, and R. Stollberger, *MRI raw data: T2 weighted TSE scan of a healthy volunteer (4 channel head coil)*, Dec. 2010, https://doi.org/10.5281/zenodo.800525, https://doi.org/10.5281/zenodo.800525.
[44] F. Knoll, C. Clason, K. Bredies, M. Uecker, and R. Stollberger, *Parallel imaging with nonlinear reconstruction using variational penalties*, Magnetic Resonance in Medicine, 67

(2012), pp. 34–41.
[45] D. Kundur and D. Hatzinakos, *Blind image deconvolution*, IEEE Signal Processing Magazine, 13 (1996), p. 43, https://doi.org/10.1109/79.489268.
[46] K. Kurdyka, *On gradients of functions definable in o-minimal structures*, in Annales de l'institut Fourier, vol. 48, Chartres: L'Institut, 1950-, 1998, pp. 769–784.
[47] Y. LeCun, C. Cortes, and C. J. Burges, *MNIST handwritten digit database*, AT&T Labs [Online], 2 (2010), http://yann.lecun.com/exdb/mnist.
[48] G. Li and T. K. Pong, *Global convergence of splitting methods for nonconvex composite optimization*, SIAM Journal on Optimization, 25 (2015), pp. 2434–2460.
[49] J. Liang, J. Fadili, and G. Peyré, *A multi-step inertial forward-backward splitting method for non-convex optimization*, in Advances in Neural Information Processing Systems, 2016, pp. 4035–4043.
[50] P.-L. Lions and B. Mercier, *Splitting algorithms for the sum of two nonlinear operators*, SIAM Journal on Numerical Analysis, 16 (1979), pp. 964–979.
[51] S. Lojasiewicz, *Une propriété topologique des sous-ensembles analytiques réels*, Les équations aux dérivées partielles, 117 (1963), pp. 87–89.
[52] S. Matet, L. Rosasco, S. Villa, and B. L. Vu, *Don't relax: early stopping for convex regularization*, arXiv preprint arXiv:1707.05422, (2017).
[53] M. Moeller, M. Benning, C. Schönlieb, and D. Cremers, *Variational depth from focus reconstruction*, IEEE Transactions on Image Processing, 24 (2015), pp. 5369–5378.
[54] V. A. Morozov, *Methods for solving incorrectly posed problems*, Springer Science & Business Media, 2012.
[55] A. Nemirovskii, D. B. Yudin, and E. R. Dawson, *Problem complexity and method efficiency in optimization*, (1983).
[56] M. Nikolova and P. Tan, *Alternating proximal gradient descent for nonconvex regularised problems with multiconvex coupling terms*, (2017).
[57] J. Nocedal and S. J. Wright, *Numerical optimization 2nd*, 2006.
[58] P. Ochs, Y. Chen, T. Brox, and T. Pock, *ipiano: Inertial proximal algorithm for nonconvex optimization*, SIAM Journal on Imaging Sciences, 7 (2014), pp. 1388–1419.
[59] P. Ochs, J. Fadili, and T. Brox, *Non-smooth non-convex Bregman minimization: Unification and new algorithms*, arXiv preprint arXiv:1707.02278, (2017).
[60] S. Osher, M. Burger, D. Goldfarb, J. Xu, and W. Yin, *An iterative regularization method for total variation-based image restoration*, Multiscale Modeling & Simulation, 4 (2005), pp. 460–489.
[61] T. Pock, D. Cremers, H. Bischof, and A. Chambolle, *An algorithm for minimizing the Mumford-Shah functional*, in Computer Vision, 2009 IEEE 12th International Conference on, IEEE, 2009, pp. 1133–1140.
[62] T. Pock and S. Sabach, *Inertial proximal alternating linearized minimization (iPALM) for nonconvex and nonsmooth problems*, SIAM Journal on Imaging Sciences, 9 (2016), pp. 1756–1787.
[63] M. Prato, S. Bonettini, I. Loris, F. Porta, and S. Rebegoldi, *On the constrained minimization of smooth Kurdyka-Łojasiewicz functions with the scaled gradient projection method*, Journal of Physics: Conference Series, 756 (2016), p. 012001, https://doi.org/http://dx.doi.org/10.1088/1742-6596/756/1/012001.
[64] K. P. Pruessmann, M. Weiger, M. B. Scheidegger, and P. Boesiger, *SENSE: Sensitivity Encoding for Fast MRI*, Magnetic Resonance in Medicine, 42 (1999), pp. 952–62.
[65] S. Ramani and J. A. Fessler, *Parallel MR image reconstruction using augmented lagrangian methods*, IEEE Transactions on Medical Imaging, 30 (2011), pp. 694–706.
[66] R. T. Rockafellar, *Convex Analysis*, vol. 28, Princeton University Press, 1970.
[67] F. Schöpfer, A. K. Louis, and T. Schuster, *Nonlinear iterative methods for linear ill-posed problems in Banach spaces*, Inverse Problems, 22 (2006), p. 311.
[68] M. Teboulle, *Entropic proximal mappings with applications to nonlinear programming*, Mathematics of Operations Research, 17 (1992), pp. 670–690.
[69] M. Uecker, T. Hohage, K. T. Block, and J. Frahm, *Image reconstruction by regularized nonlinear inversion-joint estimation of coil sensitivities and image content*, Magnetic Resonance in Medicine, 60 (2008), pp. 674–682.
[70] T. Valkonen, *A primal–dual hybrid gradient method for nonlinear operators with applications to MRI*, Inverse Problems, 30 (2014), p. 055012.
[71] Y. Xu and W. Yin, *A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion*, SIAM Journal on imaging sciences, 6 (2013), pp. 1758–1789.
[72] Y. Xu and W. Yin, *A globally convergent algorithm for nonconvex optimization based on block*

coordinate update, Journal of Scientific Computing, (2017), pp. 1–35.

[73] W. YIN, *Analysis and generalizations of the linearized Bregman method*, SIAM Journal on Imaging Sciences, 3 (2010), pp. 856–877.

[74] W. YIN, S. OSHER, D. GOLDFARB, AND J. DARBON, *Bregman iterative algorithms for ℓ_1-minimization with applications to compressed sensing*, SIAM Journal on Imaging sciences, 1 (2008), pp. 143–168.

[75] M. ZHU AND T. CHAN, *An efficient primal-dual hybrid gradient algorithm for total variation image restoration*, UCLA CAM Report, 34 (2008).