Bregman Methods for Large-Scale Optimisation with Applications in Imaging

Martin Benning and Erlend Skaldehaug Riis

Contents

M. Benning (✉)
The School of Mathematical Sciences, Queen Mary University of London, London, UK
e-mail: m.benning@qmul.ac.uk

E. S. Riis
The Department of Applied Mathematics and Theoretical Physics, Cambridge, UK
e-mail: erlend.s.riis@gmail.com

K. Chen et al. (eds.), *Handbook of Mathematical Models and Algorithms in Computer Vision and Imaging*, https://doi.org/10.1007/978-3-030-03009-4_62-1

Abstract

In this chapter we review recent developments in the research of Bregman methods, with particular focus on their potential use for large-scale applications. We give an overview on several families of Bregman algorithms and discuss modifications such as accelerated Bregman methods, incremental and stochastic variants, and coordinate descent-type methods. We conclude this chapter with numerical examples in image and video decomposition, image denoising, and dimensionality reduction with auto-encoders.

Keywords

Optimisation · Bregman proximal methods · Bregman iterations · Inverse problems · Nesterov acceleration · Mirror descent · Kaczmarz method · Coordinate descent · Itoh-Abe method · Alternating direction method of multipliers · Primal-dual hybrid gradient · Robust principal components analysis · Deep learning · Image denoising

Introduction

Bregman methods have a long history in mathematical research areas such as optimisation, inverse and ill-posed problems, statistical learning theory, and machine learning. In this review, we mainly focus on the areas of optimisation and inverse and ill-posed problems and the application of popular Bregman methods to potentially large-scale problems. Following Lev Bregman's seminal work in 1967 (Bregman 1967), it was not before the work of Censor and Lent (1981) in 1981 that the use of Bregman methods has slowly but steadily been popularised in the area of mathematical optimisation, shortly followed by the advent of the mirror descent algorithm (Nemirovsky and Yudin 1983). Bregman proximal methods, which we discuss in greater detail in the following section, were first introduced by Censor and Zenios in their seminal work in 1992 (Censor and Zenios 1992), shortly followed by Teboulle (1992), Teboulle and Chen (1993) and Eckstein (1993). Bregman methods have been extensively studied since, see, for example, Bauschke et al. (2003) and references therein, and many notable extensions were developed, with one of the most popular ones in the context of inverse and ill-posed problems being the so-called Bregman iteration (Osher et al. 2005), which is based on a generalised Bregman distance notion (Kiwiel 1997b). Bregman iterations have been shown to possess favourable regularisation properties over traditional linear iterative regularisation methods, especially in the context of imaging and image processing applications, and therefore gained a lot of attention in those research fields. We refer to Osher et al. (2005), Burger (2016) and Benning and Burger (2018) for an overview on Bregman iterations.

The goal of this chapter is to provide a non-exhaustive overview over some recent developments in the adaptation of Bregman methods to handle poten-

tially large-scale problems. These extensions range from simple linearisations to accelerated versions of Bregman methods, incremental and stochastic adaptations, and coordinate descent variants to Bregman extensions of popular primal-dual frameworks. The chapter is therefore structured as follows. In section "Bregman Proximal Methods" we give an overview over Bregman proximal methods and some notable extensions. In section "Accelerated Bregman Methods" we discuss accelerations of the linearised Bregman iteration, before we focus on incremental and stochastic variants in section "Incremental and Stochastic Bregman Proximal Methods". Subsequently, we discuss coordinate descent-type Bregman methods in section "Bregman Coordinate Descent Methods" and saddle-point formulations of Bregman algorithms in section "Saddle-Point Methods". We present several application examples in section "Applications" before concluding this chapter with section "Conclusions and Outlook".

Bregman Proximal Methods

The Bregman proximal method or Bregman proximal algorithm is defined as the following iterative procedure. Starting with an initial value $x^0 \in \mathbb{R}^n$, we compute

$$x^{k+1} = \arg\min_{x \in \mathbb{R}^n} \left\{ F(x) + D_R(x, x^k) \right\}, \tag{1}$$

for $k \in \mathbb{N}$. Here $F : \mathbb{R}^n \to \mathbb{R}$ is a function that we wish to minimise via (1). We assume that F is bounded from below and that both F and R satisfy conditions that guarantee existence and uniqueness of the solution of (1), without discussing them in greater detail. The term $D_R(x, y)$ denotes the Bregman distance w.r.t. a convex and continuously differentiable function $R : \mathbb{R}^n \to \mathbb{R}$, which is defined as

$$D_R(x, y) = R(x) - R(y) - \langle \nabla R(y), x - y \rangle, \tag{2}$$

for all $x, y \in \mathbb{R}^n$, see Bregman (1967) and Censor and Lent (1981). In the following example, we recall a few relevant examples of Bregman distances.

Example 1 (Bregman distances). For a symmetric, positive semi-definite matrix $Q \in \mathbb{R}^{n \times n}$ and the function $R(x) := \frac{1}{2}\langle Qx, x \rangle$, we observe

$$D_R(x, y) = \frac{1}{2}\langle Q(x - y), x - y \rangle.$$

Special cases include the squared Euclidean distance if Q is the identity matrix and the squared Mahalanobis distance (cf. Mahalanobis 1936) if Q is a covariance matrix.

The generalised Kullback-Leibler divergence, i.e.

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

can be obtained by choosing R as the (shifted, negative) Boltzmann-Shannon entropy, i.e. $R(x) := \sum_{j=1}^{n} \left[x_j \log(x_j) - x_j\right]$. Other notable examples include the Itakura–Saito distance (cf. Itakura 1968) and the Hellinger distance (cf. Hellinger 1909).

Note that $D_R(x, y) \geq 0$ is guaranteed for all $x, y \in \mathbb{R}^n$ due to the convexity of R. Before we are briefly going to discuss how this Bregman framework unifies implicit and explicit gradient methods in the following section, we want to recall some basic and well-known properties of (1).

Corollary 1. *Let $F : \mathbb{R}^n \to \mathbb{R}$ and $R : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable functions, where R is also convex, and suppose for some $\overline{x} \in \mathbb{R}^n$ that x^* is defined as*

$$x^* := \arg\min_{x \in \mathbb{R}^n} \left\{ F(x) + D_R(x, \overline{x}) \right\}. \tag{3}$$

Then, the following identity holds:

$$F(x^*) + D_F(x, x^*) + D_R(x, x^*) + D_R(x^*, \overline{x}) = F(x) + D_R(x, \overline{x}). \tag{4}$$

Corollary 1 can easily be verified by computing the optimality condition of (3), subsequent computation of the inner product of the optimality condition with $x^* - x$, and the use of the three-point identity for Bregman distances, first proven in Chen and Teboulle (1993, Lemma 3.1). Corollary 1 allows us to verify the following convergence result of the Bregman method with convergence rate $1/k$ for convex functions F.

Theorem 1. *Let $F : \mathbb{R}^n \to \mathbb{R}$ and $R : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable and convex functions. Suppose $\hat{x}$ is a global minimiser of F that exists. Then, for any x^0, the iterates (1) satisfy*

$$F(x^k) - F(\hat{x}) \leq \frac{D_R(\hat{x}, x^0) - D_R(\hat{x}, x^k)}{k},$$

for $k \in \mathbb{N}$.

Proof. Applying Corollary 1 for $x^* = x^{k+1}$, $\overline{x} = x^k$, and $x = \hat{x}$ yields

$$F(x^{k+1}) + D_F(\hat{x}, x^{k+1}) + D_R(\hat{x}, x^{k+1}) + D_R(x^{k+1}, x^k) = F(\hat{x}) + D_R(\hat{x}, x^k),$$

which implies

$$F(x^{k+1}) - F(\hat{x}) \leq D_R(\hat{x}, x^k) - D_R(\hat{x}, x^{k+1}),$$

due to the convexity of F and R. Summing up this inequality from $k = 0, \ldots, K-1$ leads to

$$\sum_{k=0}^{K-1} F(x^{k+1}) - K\, F(\hat{x}) \leq D_R(\hat{x}, x^0) - D_R(\hat{x}, x^K).$$

Applying Corollary 1 again – but this time for $x^* = x^{k+1}$, $\overline{x} = x^k$ and $x = x^k$ – leaves us with

$$F(x^{k+1}) + D_F(x^k, x^{k+1}) + D_R(x^k, x^{k+1}) + D_R(x^{k+1}, x^k) = F(x^k) + \underbrace{D_R(x^k, x^k)}_{=0},$$

which in return implies $F(x^{k+1}) \leq F(x^k)$ due to the convexity of F and R (which is also an immediate consequence of the variational formulation of the Bregman method). Hence, we observe $K\, F(x^K) \leq \sum_{k=0}^{K-1} F(x^{k+1})$, which concludes the proof.

Remark 1. Note that the conditions on F and R in Theorem 1 alone do not necessarily guarantee uniqueness or even existence of x^{k+1} in (1). However, if the solution exists and is unique and computable, then Theorem 1 applies.

Let us now turn our attention to implicit and explicit gradient methods and how they can both be formulated as special cases of (1).

A Unified Framework for Implicit and Explicit Gradient Methods

While it is common in numerical analysis to distinguish between implicit and explicit methods, a feature of the Bregman framework is that it covers both types of methods. This can be seen by considering (1), i.e.

$$x^{k+1} = \arg\min_{x \in \mathbb{R}^n} \left\{ F(x) + D_J(x, x^k) \right\}, \tag{5}$$

for the special choice of $J : \mathbb{R}^n \to \mathbb{R}$ with

$$J(x) := \begin{cases} R(x) & \text{implicit} \\ \frac{1}{\tau} R(x) - F(x) & \text{explicit} \end{cases}. \tag{6}$$

Evaluating the Bregman distance w.r.t. J turns (5) into

$$x^{k+1} = \arg\min_{x\in\mathbb{R}^n} \begin{Bmatrix} F(x) + D_R(x, x^k) & \text{implicit} \\ F(x^k) + \langle \nabla F(x^k), x - x^k \rangle + \frac{1}{\tau} D_R(x, x^k) & \text{explicit} \end{Bmatrix};$$

Hence, we can construct Bregman methods that are either implicit or explicit w.r.t. ∇F. Whenever we use J as the notation of our function throughout this manuscript, we implicitly refer to J as defined in (6). Whenever we use R, we refer to a function R that is not of the form $\frac{1}{\tau}R - F$. Note that we rediscover the traditional gradient descent algorithm for the choice $R(x) = \frac{1}{2}\|x\|^2$ as a special case of the explicit formulation. Furthermore, note that the explicit formulation

$$x^{k+1} = \arg\min_{x\in\mathbb{R}^n} \left\{ F(x^k) + \langle \nabla F(x^k), x - x^k \rangle + \frac{1}{\tau} D_R(x, x^k) \right\} \tag{7}$$

is also known as mirror descent (Ben-Tal et al. 2001; Beck and Teboulle 2003; Juditsky et al. 2011), Bregman gradient method (Teboulle 2018), or recently also as NoLips (Bauschke et al. 2017). In order to guarantee convergence of (5), one usually has to guarantee convexity of J. In the explicit setting, this implies that τ and R have to be chosen to ensure convexity of $\frac{1}{\tau}R - F$ or equivalently that F is $1/\tau$-smooth if R is also a quadratic function. The latter condition has basically been proposed in Bauschke et al. (2017) and further discussed in Benning et al. (2017a,b) and Bolte et al. (2018). It has also been shown that if the step size τ is chosen such that c R - F is convex, for a some constant $c > 0$ and a function F, the estimate $0 < \tau \leq \Big(\big(1 + \gamma(R)\big) - \delta\Big)/c$ is sufficient to guarantee convergence under mild assumptions that are outlined in detail in Bauschke et al. (2017). Here $\gamma(R)$ denotes the symmetry coefficient defined as

$$\gamma(R) := \inf \left\{ D_R(x, y)/D_R(y, x) \;\middle|\; (x, y) \in (\text{int dom } R)^2 \backslash \{x, y \,|\, x{=}y\} \right\} \in [0, 1],$$

and δ is a constant that satisfies $\delta \in (0, 1 + \gamma(R))$. In the following section, we want to review the special case of Bregman gradient methods where F is the sum of two functions.

Bregman Proximal Gradient Method

An interesting, special case frequently considered in the literature is the case where F is a sum of two functions L and S, i.e. the Bregman method reads

$$x^{k+1} = \arg\min_{x\in\mathbb{R}^n} \left\{ L(x) + S(x) + D_J(x, x^k) \right\}, \tag{8}$$

where we assume that $L : \mathbb{R}^n \to \mathbb{R}$ is a continuously differentiable function. The function $S : \mathbb{R}^n \to \overline{\mathbb{R}}$ on the other hand is proper, lower semi-continuous (l.s.c.) and convex, for $\overline{\mathbb{R}} := \mathbb{R} \cup \{\infty\}$. If we choose $J(x) := \frac{1}{2\tau}\|x\|^2 - L(x)$ in the spirit of (6), then (8) reads

$$\begin{aligned} x^{k+1} &= \arg\min_{x\in\mathbb{R}^n} \left\{ \frac{1}{2} \left\| x - \left(x^k - \tau\, \nabla L(x^k)\right) \right\|^2 + \tau S(x) \right\}, \\ &=: (I + \tau S)^{-1} \left(x^k - \tau\, \nabla L(x^k)\right), \end{aligned}$$

where $(I + \tau S)^{-1} : \mathbb{R}^n \to \mathbb{R}^n$ is known as the proximal map or resolvent, see, for instance, (Parikh et al. 2014). This is the classical proximal gradient method, also known as forward backward splitting (Lions and Mercier 1979). More general proximal gradient methods can be derived for different choices of J and S, for example, the entropic mirror descent algorithm (Nemirovsky and Yudin 1983; Beck and Teboulle 2003; Beck 2017; Doan et al. 2018), i.e.

$$x_j^{k+1} = \frac{x_j^k \exp\left(-\tau(\nabla L(x^k))_j\right)}{\sum_{j=1}^n x_j^k \exp\left(-\tau(\nabla L(x^k))_j\right)},$$

for $j \in \{1, \ldots, n\}$, the difference of the negative Boltzmann Shannon entropy as defined in Example 1 and the function L, i.e. $J(x) := \frac{1}{\tau}\sum_{j=1}^n \left[x_j \log(x_j) - x_j\right] - L(x)$ with the convention $0\log(0) \equiv 0$, and the characteristic function

$$S(x) := \begin{cases} 0 & x \in \Sigma \\ +\infty & x \notin \Sigma \end{cases},$$

over the simplex constraint

$$\Sigma := \left\{ x \in \mathbb{R}^n \;\middle|\; x_j \geq 0,\ \forall j \in \{1, \ldots, n\},\ \sum_{j=1}^n x_j = 1 \right\}.$$

We also mention *variable metric proximal gradient methods*, an important class of algorithms which may be viewed as an instance of Bregman proximal gradient methods where the Bregman function J_k is iteration-dependent. Denoting by $(A_k)_{k\in\mathbb{N}}$ a sequence of symmetric positive definite matrices, which act as preconditioners, we define $J_k(x) := \frac{1}{2\tau_k}\langle x, A_k x\rangle - L(x)$. Note that if $S \equiv 0$, $A_k = \nabla^2 L(x^k)$, and $\tau_k = 1$, then one recovers the Newton method for L

$$x^{k+1} = x^k - (\nabla^2 L(x^k))^{-1} \nabla L(x^k).$$

More generally when $S \not\equiv 0$, one may choose A_k to be an approximation to the Hessian of L at x^k, so as to incorporate elements of quasi-Newton methods to the proximal gradient scheme. These schemes were studied by Bonnans et al. (1995) and later studied for non-convex objective functions (Chouzenoux et al. 2014; Frankel et al. 2015), Hilbert spaces (Combettes and Vũ 2014), and extensions to inertial methods (Bonettini et al. 2018), to mention a few examples.

In the next section, we focus on extensions of the Bregman proximal methods to convex but nonsmooth functions.

Bregman Iteration

A very important generalisation of (1), first proposed in Osher et al. (2005), allows us to also use convex but nonsmooth functions J as defined in (6) instead of convex and continuously differentiable functions J. Suppose we are given a proper, l.s.c. and convex function $J : \mathbb{R}^n \to \overline{\mathbb{R}}$. Then its subdifferential, defined as

$$\partial J(y) := \left\{ p \in \mathbb{R}^n \;\middle|\; J(x) - J(y) \geq \langle p, x - y \rangle,\ \forall x \in \mathbb{R}^n \right\},$$

is non-empty. It therefore makes sense to extend the definition (2) to a generalised Bregman distance (Kiwiel 1997a) for subdifferentiable functions, i.e.

$$D_J^p(x, y) = J(x) - J(y) - \langle p, x - y \rangle,$$

for $p \in \partial J(y)$. A generalisation of (1), commonly known as Bregman iteration, can then be defined as

$$x^{k+1} = \arg\min_{x \in \mathbb{R}^n} \left\{ F(x) + D_J^{p^k}(x, x^k) \right\}, \tag{9a}$$

$$p^{k+1} = p^k - \nabla F(x^{k+1}), \tag{9b}$$

for initial values $x^0 \in \mathbb{R}^n$ and $p^0 \in \partial J(x^0)$. Note that Corollary 1 and Theorem 1 also apply to Bregman iterations (cf. Benning and Burger 2018, Corollary 6.5), as those statements did not utilise any potential differentiability of J. Furthermore, note that the explicit variant of the Bregman iteration is known as the linearised Bregman iteration and has extensively been studied in Yin et al. (2008), Cai et al. (2009a,b,c), and Yin (2010).

Linearised Bregman Iteration as Gradient Descent

With the particular choice $J(x) = \frac{1}{2\tau}\|x\|^2 + \frac{1}{\tau}R(x) - F(x)$, the Bregman iteration (9) turns into the linearised Bregman iteration, which reads

$$x^{k+1} = \arg\min_{x\in\mathbb{R}^n} \left\{ F(x^k)+\langle \nabla F(x^k), x-x^k\rangle + \frac{1}{2\tau}\|x-x^k\|^2 + \frac{1}{\tau} D_R^{q^k}(x, x^k) \right\},$$

$$= (I+\partial R)^{-1}\left(x^k+q^k-\tau\nabla F(x^k)\right), \tag{10a}$$

$$q^{k+1} = q^k - \left(x^{k+1}-x^k+\tau\nabla F(x^k)\right), \tag{10b}$$

where $(I+\partial R)^{-1}$ denotes the proximal mapping w.r.t. the function R and $q^k \in \partial R(x^k)$ the subgradient of R at x^k that is iteratively defined via (10b) and some initial value $q^0 \in \partial R(x^0)$. Suppose we assume that $(x^k+q^k)/\tau - \nabla F(x^k)$ is in the range of some matrix $A \in \mathbb{R}^{m\times n}$ and that we therefore can substitute $\tau A^\top b^k := x^k + q^k - \tau\nabla F(x^k)$. Then (10) can be written as

$$x^{k+1} = (I+\partial R)^{-1}(\tau A^\top b^k), \tag{11a}$$

$$A^\top b^{k+1} = A^\top b^k - \nabla F(x^{k+1}). \tag{11b}$$

In the following, we want to focus on the special case $F(x) = \frac{1}{2}\|Ax-b^\delta\|^2$ with $\nabla F(x) = A^\top(Ax-b^\delta)$ for a matrix $A \in \mathbb{R}^{m\times n}$, for which (11) simplifies to

$$x^{k+1} = (I+\partial R)^{-1}(\tau A^\top b^k), \tag{12a}$$

$$b^{k+1} = b^k - \left(Ax^{k+1}-b^\delta\right), \tag{12b}$$

with initial value $b^0 = b^\delta$, given the assumption that the initial values of the original formulation were $x^0 = 0$ and $p^0 = 0$. Note that we can also write (12) as

$$b^{k+1} = b^k - \left(A(I+\partial R)^{-1}\left(\tau A^\top b^k\right) - b^\delta\right). \tag{13}$$

Hence, if we can identify an energy G_τ for which we can associate its gradient ∇G_τ with $A(I+\partial R)^{-1}\left(\tau A^\top \cdot\right) - b^\delta$, we can consider the linearised Bregman iteration a gradient descent method applied to this specific energy. In Yin (2010) and Huang et al. (2013), this energy has been identified as

$$G_\tau(b) := \frac{\tau}{2}\|A^\top b\|^2 - \langle b, b^\delta\rangle - \frac{1}{\tau}\tilde{R}(\tau A^\top b),$$

where $\tilde{R}$ denotes the Moreau-Yosida regularisation of R (cf. Moreau 1965; Yosida 1964), i.e.

$$\tilde{R}(z) := \inf_{x\in\mathbb{R}^n}\left\{ R(x) + \frac{1}{2}\|x-z\|^2 \right\}.$$

Since the gradient of the Moreau-Yosida regularisation of R reads $\nabla \tilde{R}(z) = z - (I + \partial R)^{-1}(z)$ (see, for instance, Attouch et al. 2014, Proposition 17.2.1), we easily verify

$$\nabla G_\tau(b) = A(I + \partial R)^{-1}(\tau A^\top b) - b^\delta.$$

As a consequence, (13) is equivalent to

$$b^{k+1} = b^k - \nabla G_\tau(b^k),$$

and the linearised Bregman iteration for $F(x) = \frac{1}{2}\|Ax - b^\delta\|^2$ reduces to a gradient descent method. This equivalence will be useful when studying acceleration methods.

Bregman Iterations as Iterative Regularisation Methods

Bregman iterations are not only useful for solving optimisation problems but are also extremely important in the context of solving inverse and ill-posed problems. The reason for this is that Bregman iterations can be used as iterative regularisation methods. If we consider the deterministic linear inverse problem

$$Ax^\dagger = b^\dagger, \tag{14}$$

for a given matrix $A \in \mathbb{R}^{m\times n}$, the aim of solving this inverse problem is to approximate $x^\dagger$ in (14), for given A and data b^δ with $\|b^\dagger - b^\delta\| \leq \delta$. Here, δ is a known, positive bound on the error of the measured data b^δ and the data $b^\dagger$ that satisfies (14).

Suppose we consider a convex function F that depends on A and b^δ, which we will denote as F_{b^δ}. It then can easily be shown that the iterates of (9) satisfy

$$D_J^{p^{k+1}}(x^\dagger, x^{k+1}) < D_J^{p^k}(x^\dagger, x^k),$$

for all indices $k \leq k^*(\delta)$ that satisfy Morozov's discrepancy principle (Morozov 1966), i.e.

$$F_{b^\delta}(x^{k^*(\delta)}) \leq \eta\delta < F_{b^\delta}(x^k),$$

for a parameter $\eta \geq 1$, see Osher et al. (2005) and Burger et al. (2007). Note that for $\eta > 1$ it can be guaranteed that $k^*(\delta)$ is finite. With the additional regularity assumption that $x^\dagger$ satisfies the so-called range condition (Benning and Burger 2018, Definition 5.8), i.e.

$$x^\dagger \in \arg\min_{x\in\mathbb{R}^n} \left\{ F_g(x) + R(x) \right\},$$

for some data $g \in \mathbb{R}^m$, one can prove the error estimate

$$D_J^{p^k}(x^\dagger, x^k) \leq \frac{\|w\|^2}{2k} + \delta\|w\| + \delta^2 k,$$

for the special case $F_{b^\delta}(x) := \frac{1}{2}\|Ax - b^\delta\|^2$, see Burger et al. (2007, Theorem 4.3). Here, w is defined as $w := g - Ax^\dagger \in \mathbb{R}^m$, which satisfies the source condition $A^*w \in \partial J(x^\dagger)$, cf. (Chavent and Kunisch 1997; Burger and Osher 2004). If $k^*(\delta)$ is of order $1/\delta$, we therefore observe

$$D_J^{p^{k^*(\delta)}}(x^\dagger, x^{k^*(\delta)}) = \mathcal{O}(\delta);$$

Hence, $x^{k^*(\delta)}$ converges to $x^\dagger$ in terms of the Bregman distances if δ converges to zero.

For more details on how to use Bregman iterations in the context of (linear) inverse problems, we refer the reader to Osher et al. (2005), Resmerita and Scherzer (2006), Schuster et al. (2012), Burger (2016), and Benning and Burger (2018). For the remainder of this paper, we want to discuss modifications of Bregman iterations and Bregman proximal methods that are suitable to large-scale optimisation and inverse problems.

Inverse Scale Space Flows

In what follows, we describe the *inverse scale space* (ISS) flow, a system of differential equations which can be derived as the continuous time limit of the Bregman iterations. For a Bregman function $J : \mathbb{R}^n \to \overline{\mathbb{R}}$ and objective function $F : \mathbb{R}^n \to \mathbb{R}$, this flow is given by

$$\dot{p}(t) = -\nabla F(x(t)), \quad p(t) \in \partial J(x(t)). \tag{15}$$

It is straightforward to verify that Bregman iterations (9b) and linearised Bregman iterations (10) can be derived, respectively, as the forward and backward Euler discretisation of (15).

The term *inverse scale space flow* was coined by Scherzer and Groetsch (2001) in 2001. In addition to its connection to Bregman schemes, the ISS flow itself is an active topic of research. Initially studied by Burger et al. (2006, 2007, 2013), and Burger (2016), it has found applications in nonlinear spectral analysis by Burger et al. (2016), Gilboa et al. (2016), and Schmidt et al. (2018).

The ISS flow itself has largely been studied in the context of scale space methods and data filtering, where the objective functions generally take the more specific forms $\|x - b^\dagger\|^2/2$ or $\|Ax - b^\dagger\|^2/2$. We mention some papers that address questions regarding the existence and uniqueness results for solutions to (15). Burger et al. (2007) proved existence, uniqueness, and certain regularity properties of the solution

to the flow when J is the total variation seminorm. These results were extended by Frick and Scherzer (2007) to all convex, proper, lower semicontinuous functions J, while in Burger et al. (2013), Burger et al. characterise the solution to the flow explicitly for the case $J = \|\cdot\|_1$. We note that while these studies do not assume strict convexity of J, strong convexity is ensured for F by the $\|\cdot\|^2$ term in F (restricted to the range of the linear operator A), so that the iterations (and flow) are still well-defined.

By supposing that J were twice continuously differentiable and μ-convex for some $\mu > 0$ (i.e. strongly convex with parameter μ, see Hiriart-Urruty and Lemaréchal 1993), we can provide an additional interpretation of the ISS flow, rewriting (15) as

$$\dot{x}(t) = -(\nabla^2 J(x(t)))^{-1} \nabla F(x(t)). \tag{16}$$

With this formulation, one can interpret the Hessian of $J(x(t))$ as a preconditioner for the flow. Furthermore, by using the chain rule, we derive an energy dissipation law for the system

$$\frac{d}{dt} F(x(t)) = \left\langle \dot{x}(t), \nabla F(x(t)) \right\rangle = -\left\langle \dot{x}(t), \nabla^2 J(x(t)) \dot{x}(t) \right\rangle \leq -\mu \|\dot{x}(t)\|^2,$$

where the final inequality follows from μ-convexity of J. Furthermore, observe that if $J = F$, (16) reduces to a continuous-time variant of Newton's method. One may tie this back to the variable metric proximal gradient methods, which were designed to incorporate quasi-Newton preconditioning to proximal gradient methods.

In section "The Bregman Itoh–Abe Method", we describe the Bregman Itoh–Abe (BIA) method (Benning et al. 2020), an iterative system derived by applying structure-preserving methods from numerical integration to the flow. Thus the ISS flow provides an alternative way to consider variational formulations for formulating Bregman schemes.

Accelerated Bregman Methods

Not only when dealing with large-scale problems, reducing the number of iterations is an important goal to achieve when designing an algorithm. In Theorem 1 we have seen that the Bregman proximal method (1) has a convergence rate of order $1/k$. In the wake of Nesterov (1983), many acceleration strategies have been developed for first-order optimisation methods that aim at minimising convex functions. As we focus on Bregman methods, we want to highlight the following adaptation of Nesterov (1983), first developed in Huang et al. (2013) for quadratic functions F. There, the authors consider the linearised Bregman iteration, i.e. (9) for the choice $J(x) = \frac{1}{2\tau}\|x\|^2 + \frac{1}{\tau}R(x) - F(x)$, as shown in (10). We have seen that (10) can be formulated as the gradient descent (13) for the special case $F(x) = \frac{1}{2}\|Ax - b^\delta\|^2$. The authors in Huang et al. (2013) have applied the idea of Nesterov acceleration to

formulation (13), which reads

$$b^{k+1} = (1+\beta_k)b^k - \beta_k b^{k-1} - \nabla G_\tau((1+\beta_k)b^k - \beta_k b^{k-1}), \tag{17}$$

where $\{\beta_k\}_{k\in\mathbb{N}}$ is a sequence of positive scalars. Applying $\tau A^\top$ to both sides of the equation and substituting $\tau A^\top b^k = x^k + q^k - \tau A^\top(Ax^k + b^\delta)$ then yields the equivalent formulation

$$x^{k+1} = \arg\min_{x\in\mathbb{R}^n} \left\{ F(x) + (1+\beta_k) D_J^{p^k}(x, x^k) - \beta_k D_J^{p^{k-1}}(x, x^{k-1}) \right\}, \tag{18a}$$

$$p^{k+1} = (1+\beta_k)p^k - \beta_k p^{k-1} - \nabla F(x^{k+1}), \tag{18b}$$

for $J(x) = \frac{1}{2\tau}\|x\|^2 + \frac{1}{\tau}R(x) - F(x)$, $F(x) = \frac{1}{2}\|Ax - b^\delta\|^2$, $p^k = \frac{1}{\tau}(x^k + q^k) - \nabla F(x^k) \in \partial J(x^k)$, and $q^k \in \partial R(x^k)$ for all $k \in \mathbb{N}$.

Remark 2. We want to emphasise that the equivalence between (17) and (18) does not hold for arbitrary functions F as we have exploited the linearity of ∇F by making use of $\nabla F((1+\beta_k)x^k - \beta_k x^{k-1}) = (1+\beta_k)\nabla F(x^k) - \beta_k \nabla F(x^{k-1})$.

Note that (17) can also be written in less compact form as

$$x^{k+1} = (I + \partial R)^{-1}(z^k), \tag{19a}$$

$$y^{k+1} = z^k - \tau \nabla F(x^{k+1}), \tag{19b}$$

$$z^{k+1} = \left(1 + \beta_{k+1}\right) y^{k+1} - \beta_{k+1} y^k, \tag{19c}$$

if we substitute $y^k = \tau A^\top b^k$. Following the same approach as in Chambolle and Dossal (2015), (19) can also be written as

$$x^{k+1} = (I + \partial R)^{-1}(z^k), \tag{20a}$$

$$y^{k+1} = z^k - \tau \nabla F(x^{k+1}), \tag{20b}$$

$$z^{k+1} = \left(1 - \frac{1}{t_{k+1}}\right) y^{k+1} + \frac{1}{t_{k+1}} u^{k+1}, \tag{20c}$$

$$u^{k+1} = y^k + t_{k+1}(y^{k+1} - y^k). \tag{20d}$$

for $\beta_k := (t_k - 1)/t_{k+1}$ and a sequence $\{t_k\}_{k\in\mathbb{N}}$ of positive parameters.

An open problem which has attracted interest in recent years concerns whether accelerated versions of Bregman (proximal) gradient methods with generic, strongly convex Bregman distances are possible (Teboulle 2018). In a recent work by Dragomir et al. (2019), this question is partly answered in the negative, concluding

that for Bregman distances, based on smooth functions R or functions R that satisfy that $\frac{1}{\tau}R - F$ is convex, the $\mathcal{O}(1/k)$ convergence rate is optimal for first-order methods that use previous gradient and Bregman proximal evaluations. However, for more restrictive function classes, faster convergence rates can be achieved, as has been shown in Hanzely et al. (2018) and Gutman and Peña (2018).

Acceleration strategies such as Nesterov acceleration have also been analysed in the context of iterative regularisation strategies (e.g. (9) combined with early stopping as described in section "Bregman Iterations as Iterative Regularisation Methods"), see, for instance, Matet et al. (2017), Neubauer (2017), Garrigos et al. (2018), and Calatroni et al. (2019).

Incremental and Stochastic Bregman Proximal Methods

Many large-scale problems, in particular in machine learning, involve the minimisation of functions of the form

$$F(x) := \frac{1}{m}\sum_{i=1}^{m} f_i(x). \tag{21}$$

In other words, the objective function is a sum of m individual functions. If m happens to be extremely large, computing the gradient of F can be computationally extremely expensive, rendering the application of traditional methods such as (1) or (18) computationally infeasible. Feasible alternatives are methods that make use of gradients that are only based on a subset $B \subset \{1, \dots, m\}$ of all indices. Such methods include incremental gradient methods (Bertsekas et al. 2011a) and stochastic gradient methods (Robbins and Monro 1951). If we assume that F in (21) is of the form

$$F(x) = L(x) + S(x) = \frac{1}{m}\sum_{i=1}^{m} \ell_i(x) + \frac{1}{m}\sum_{i=1}^{m} s_i(x), \tag{22}$$

an incremental version of the Bregman proximal gradient as in (8) can be formulated as

$$x^k = \arg\min_{x\in\mathbb{R}^n} \left\{ \ell_{i(k)}(x) + s_{i(k)}(x) + D_{J_k}(x, x^{k-1}) \right\}. \tag{23}$$

Here $i : \mathbb{N} \to \{1, \dots, m\}$ denotes the index function $i(x) := x$ modulo m, although other cycle orderings are certainly possible as well. A special case of (23) is the classical incremental proximal gradient method (Bertsekas et al. 2011b)

$$x^k = \left(I + \tau_k \partial s_{i(k)}\right)^{-1} \left(x^{k-1} - \tau_k \nabla \ell_{i(k)}(x^{k-1}) \right)$$

for the choice of $J_k(x) = \frac{1}{2\tau_k}\|x\|^2 - \ell_{i(k)}(x)$. If we further pick $s_i \equiv 0$ for all i, we obtain the classical incremental gradient descent (Widrow and Hoff 1960; Bertsekas et al. 2011a), i.e.

$$\begin{aligned} x^k &= x^{k-1} - \tau_k \nabla \ell_{i(k)}(x^{k-1}), \\ &= x^{k-1} - \tau_k \nabla f_{i(k)}(x^{k-1}), \end{aligned} \tag{24}$$

as a special case.

In the following sections, we discuss extensions of stochastic gradient descent (SGD) and Kaczmarz methods in the Bregman framework, before highlighting the connection between single cycles of incremental Bregman proximal methods and deep neural network architectures.

Stochastic Mirror Descent

Stochastic gradient descent generalises naturally to the Bregman proximal setting with the *stochastic mirror descent* (SMD) method (recall that mirror descent is equivalent to the Bregman gradient or linearised Bregman iteration). SMD is one of the most popular families of methods for stochastic optimisation, and the method is defined as Nemirovski et al. (2009)

$$x^{k+1} = \arg\min_{x \in \mathbb{R}^n} \{\tau_k \langle \nabla f_{i(k)}(x^k), x\rangle + D_J^{p^k}(x, x^k)\}. \tag{25}$$

As in the setting of incremental descent methods, $i(k) \in \{1, \ldots, n\}$ represents a sequence of indices, which in the setting of SMD are typically randomised.

SMD was originally introduced by Nemirovsky and Yudin (1983), while subsequent, significant contributions include Nemirovski et al. (2009), Nesterov (2009), and Xiao (2010). The framework and its convergence analysis were further extended by Duchi et al. (2012) to cases where the samples from the distribution are not assumed to be independent.

Similar to SGD, the SMD algorithms are suitable for large-scale optimisation and online learning settings, yet furthermore they come with the added benefits of Bregman iterations of exploiting structures in the data. Because of this, SMD is one of the most widely used family of methods for large-scale stochastic optimisation (Azizan and Hassibi 2018; Zhou et al. 2017).

In the aforementioned works on SMD, the Bregman function J is assumed to be differentiable. In contrast, the use of nonsmooth Bregman functions, e.g. that invoke the ℓ^1-norm, is significant in the context of Bregman iterations and sparse signal processing. In the following section, we cover a Bregman method for sparse reconstruction of linear systems which can be seen as an instance of SMD, using the nonsmooth Bregman function $J(x) = \|x\|^2/2 + \lambda\|x\|_1$.

The Sparse Kaczmarz Method

The Kaczmarz method is a scheme for solving quadratic problems of the form $\min_x \langle x, Ax\rangle/2 - \langle b, x\rangle$. The method was originally introduced by Kaczmarz (1937) and later by Gordon et al. (1970) under the name *algebraic reconstruction technique*. In this section, we review the extension of *Kaczmarz methods* to *sparse Kaczmarz methods* (Lorenz et al. 2014b) and their block variants. The motivation for sparse Kaczmarz methods is to find sparse solutions to linear problems $Ax = b$ via the problem formulation

$$\min_{x\in\mathbb{R}^n} \left\{ \frac{1}{2}\|x\|^2 + \lambda\|x\|_1 \ : Ax = b \right\}. \tag{26}$$

We first briefly review the original Kaczmarz method. For $x^0 = 0$, time steps $\tau_k > 0$, and a sequence of indices $(i(k))_{k\in\mathbb{N}}$, the (randomised) Kaczmarz method is given by

$$x^{k+1} = x^k - \tau_k(\langle a_{i(k)}, x^k\rangle - b_{i(k)})a_{i(k)}. \tag{27}$$

Here $a_{i(k)}$ denotes the i^{th} row vector of A. If $i(k)$ comprise a subset of indices, then the block-variant of the Kaczmarz method is given by

$$x^{k+1} = x^k - \tau_k a_{i(k)}^{\dagger}(a_{i(k)}x^k - b_{i(k)}),$$

where $a_{i(k)}$ denotes the submatrix formed by the row vectors of A indexed by $i(k)$ and $a_{i(k)}^{\dagger}$ denotes the Moore-Penrose pseudo-inverse of $a_{i(k)}$. The iterates of the randomised Kaczmarz methods converge linearly to a solution of $Ax = b$ (Gower and Richtárik 2015).

Lorenz et al. (2014b) proposed a sparse Kaczmarz method as follows. Given starting points $x^0 = z^0 = 0$, the updates are given by

$$\begin{aligned} z^{k+1} &= z^k - \tau_k(\langle a_{i(k)}, x^k\rangle - b_{i(k)})a_{i(k)}, \\ x^{k+1} &= S_\lambda(z^{k+1}). \end{aligned} \tag{28}$$

Here S_λ denotes the soft-thresholding operator with threshold λ. The iterates $(x^k)_{k\in\mathbb{N}}$ converge linearly to a solution of (26) (Schöpfer and Lorenz 2019, Theorem 3.2).

A block variant of the sparse Kaczmarz method was proposed in Lorenz et al. (2014b). For blocks of rows of A denoted by sets of indices $i(k)$, it consists of the updates

$$\begin{aligned} z^{k+1} &= z^k - \tau_k a_{i(k)}^{\top}(a_{i(k)}x^k - b_{i(k)}), \\ x^{k+1} &= S_\lambda(z^{k+1}). \end{aligned} \tag{29}$$

Note that this uses the transpose $a_{i(k)}^{\top}$, unlike the standard block Kaczmarz method which uses the pseudo-inverse $a_{i(k)}^{\dagger}$. This too converges to a solution of (26) (Lorenz et al. 2014a, Corollary 2.9).

The sparse (block-)Kaczmarz method (29) has connections to two aforementioned Bregman schemes. First, one may verify that it corresponds to the SMD method (25) for $J(x) = \|x\|^2/2 + \lambda\|x\|_1$ and $F(x) = \sum_{i=1}^{n} |\langle a_i, x\rangle - b_i|^2$. Second, if one takes the entire matrix A as each block, then one recovers the linearised Bregman method for the same J (Lorenz et al. 2014b).

As with the general SMD method, the sparse Kaczmarz method is particularly suitable in online reconstruction settings, where the rows of the linear system A and/or data entries b are not all available instantly but successively are made available over time. We refer the reader to Lorenz et al. (2014b) for numerical examples which include the application of online compressed sensing.

Deep Neural Networks

We can generalise the incremental Bregman proximal gradient (23) by including an additional, potentially nonlinear projection $H_k : \mathbb{R}^{n_{k-1}} \to \mathbb{R}^{n_k}$, to obtain

$$x^k = \arg\min_{x \in \mathbb{R}^{n_k}} \left\{ \ell_k(x) + s_k(x) + D_{J_k}(x, H_k(x^{k-1})) \right\}, \tag{30}$$

for a sequence of dimensions $\{n_k\}_{k=1}^{l}$ with $n_k \in \mathbb{N}$ for all $k = 1, \ldots, l$. We are interested in a single cycle of this incremental Bregman proximal method only, which is why we have simplified the indexing notation from $i(k)$ to k throughout this subsection. In the following, we want to demonstrate how certain deep neural network architectures are special cases of (30). This connection was first investigated in the context of variational networks by Kobler et al. (2017), in the context of Bregman methods by Benning and Burger (2018), and in the context of proximal gradient methods by Frerix et al. (2017), Combettes and Pesquet (2018), and Bertocchi et al. (2019). Gradient-based learning with Bregman algorithms has also been studied in the context of image segmentation by Ochs et al. in (2015), and Bregman distances are used to analyse regularisation strategies based on neural networks (Li et al. 2020). With the following example, we want to demonstrate how a class of feedforward neural networks coincides with (30).

Example 2 (Feedforward neural network with ReLU activation function). In this example we want to demonstrate how basic feedforward neural networks can be interpreted as variants of Algorithm (30). If we, for instance, choose $\{\ell_k\}_{k=1}^{l}$ to be of the form

$$\ell_k(x) := \frac{1}{2} \left\langle (I - M_k)x - 2\, b_k, x \right\rangle,$$

for quadratic matrices $\{M_k\}_{k=1}^l$ and vectors $\{b_k\}_{k=1}^l$ with $M_k \in \mathbb{R}^{n_k \times n_k}$ and $b_k \in \mathbb{R}^{n_k}$, which has the gradient

$$\nabla \ell_k(x) = \left(I - \frac{1}{2}\left(M_k + M_k^\top \right) \right) x - b_k,$$

and if we choose $\{s_k\}_{k=1}^l$ of the form

$$s_k(x) := \chi_{\geq 0}(x) = \begin{cases} 0 & \forall j : x_j \geq 0 \\ \infty & \exists j : x_j < 0 \end{cases}$$

for all $k \in \{1, \ldots, l\}$, then we easily verify that for the choice $J_k(x) = \|x\|^2/2 - \ell_k(x)$ the update

$$x^k = \max\left(0,\ A_k(x^{k-1}) + b_k\right),$$

with $A_k := \frac{1}{2}(M_k + M_k^T) \circ H_k$ is the unique solution of (30). Hence, we can consider this l-layer feedforward neural network with rectified linear units (ReLU) as activation functions (Nair and Hinton 2010) as a special case of the modified incremental Bregman gradient method (30) if we further guarantee that x^0 is chosen to be the input of the network.

Many other neural network architectures can be recovered in similar fashion to Example 2, where different activation functions can be recovered as proximal mappings for different choices of functions s_k, such as in Combettes and Pesquet (2018), and Bertocchi et al. (2019). For a recent overview of machine learning algorithms in the context of inverse problems, we refer to Arridge et al. (2019).

Bregman Incremental Aggregated Gradient

Two particularly interesting instances of incremental Bregman proximal methods are the *incremental aggregated gradient* (IAG) method (Blatt et al. 2007) and its stochastic counterpart *stochastic averaged gradient* (SAG) (Schmidt et al. 2017). For the sake of brevity, we focus on the incremental version in this paper. The IAG method reads

$$x^{k+1} = x^k - \frac{\tau_k}{m} g^k, \tag{31a}$$

$$g^{k+1} = g^k - \nabla f_{i(k+1)}(x^{k+1-m}) + \nabla f_{i(k+1)}(x^{k+1}). \tag{31b}$$

Here $\{\tau_k\}_{k\in\mathbb{N}}$ is a sequence of positive scalars and $i : \mathbb{N} \to \{1, \dots, m\}$ is defined as in "A Unified Framework for Implicit and Explicit Gradient Methods". Please also note that m arbitrary points $x^{1-m}, x^{2-m}, \dots, x^0$ have to be chosen as initialisation. It is easy to see and has also been pointed out in Blatt et al. (2007) that (31) can be rewritten as

$$x^{k+1} = x^k - \frac{\tau_k}{m} \sum_{l=0}^{m-1} \nabla f_{i(k-l)}(x^{k-l}), \tag{32}$$

for $k \geq m$. Note that this is equivalent to the following characterisation in terms of Bregman distances, in analogy to the explicit gradient descent characterisation in section "A Unified Framework for Implicit and Explicit Gradient Methods": if we rewrite (21) to $F(x) = \sum_{l=0}^{m-1} f_{i(k-l)}(x)$ for any $k \in \mathbb{N}$ and suppose we consider a Bregman method of the form

$$x^{k+1} = \arg\min_{x\in\mathbb{R}^n} \left\{ F(x) + \frac{1}{2\tau_k}\|x - x^k\|^2 - \frac{1}{m}\sum_{l=0}^{m-1} D_{f_{i(k-l)}}(x, x^{k-l}) \right\}, \tag{33a}$$

$$= \arg\min_{x\in\mathbb{R}^n} \left\{ \frac{1}{m}\sum_{l=0}^{m-1} \left[f_{i(k-l)}(x^{k-l}) + \langle \nabla f_{i(k-l)}(x^{k-l}), x - x^{k-l} \rangle \right] + \frac{1}{2\tau_k}\|x - x^k\|^2 \right\}, \tag{33b}$$

then it becomes evident from computing the optimality condition of (33a) that the update (33b) is equivalent to (32) and hence (31) for $k \geq m$. Note that we can rewrite (33a) to

$$x^{k+1} = \arg\min_{x\in\mathbb{R}^n} \left\{ F(x) - \frac{1}{m}\sum_{l=1}^{m-1} D_{f_{i(k-l)}}(x, x^{k-l}) + D_{J_k}(x, x^k) \right\}, \tag{34}$$

for $J_k(x) := \frac{1}{2\tau_k}\|x\|^2 - \frac{1}{m} f_{i(k)}(x)$. The notable difference to the conventional IAG method is that we can replace the Bregman distance $D_{J_k}(x, x^k)$ in (34) with more generic Bregman distances. As in section "A Unified Framework for Implicit and Explicit Gradient Methods", we can for example choose $J_k(x) = \frac{1}{2\tau_k}\|x\|^2 + \frac{1}{\tau_k} R(x) - \frac{1}{m} f_{i(k)}(x)$ and therefore derive incremental Bregman iterations of the form

$$x^{k+1} = (I + \partial R)^{-1}\left(x^k + q^k - \frac{\tau_k}{m} g^k \right)$$

$$q^{k+1} = q^k - \left(x^{k+1} - x^k + \frac{\tau_k}{m} g^k \right),$$

$$g^{k+1} = g^k - \nabla f_{i(k+1)}(x^{k+1-m}) + \nabla f_{i(k+1)}(x^{k+1}),$$

where $q^k \in \partial R(x^k)$ for all k. Hence, substituting $y^k = x^k + q^k - \frac{\tau_k}{m} g^k$ yields the equivalent formulation

$$\begin{aligned} x^{k+1} &= (I + \partial R)^{-1}\left(y^k\right), \\ g^{k+1} &= g^k - \nabla f_{i(k+1)}(x^{k+1-m}) + \nabla f_{i(k+1)}(x^{k+1}), \\ y^{k+1} &= y^k - \frac{\tau_{k+1}}{m} g^{k+1}. \end{aligned}$$

If F is of the form (22), where $s_i = s$ for some (convex) function $s : \mathbb{R}^n \to \overline{\mathbb{R}}$ for all indices $i \in \{1, \ldots, m\}$ and if we choose $J_k(x) = \frac{1}{m\,\tau_k} R(x) - \frac{1}{m} \ell_{i(k)}(x)$ for continuously differentiable R, we recover the proximal-like incremental aggregated gradient (PLIAG) method, recently proposed in Zhang et al. (2017), which reads

$$x^{k+1} = \arg\min_{x \in \mathbb{R}^n} \left\{ s(x) + \sum_{l=0}^{m-1} \left[\ell_{i(k-l)}(x^{k-l}) + \langle \nabla \ell_{i(k-l)}(x^{k-l}), x - x^{k-l} \rangle \right] + \frac{1}{\tau_k} D_R(x, x^k) \right\}.$$

Needless to say, many different IAG or SAG methods can be derived for different choices of $\{J_k\}_{k=1}^m$. Choosing J_k such that convergence of the above algorithms is guaranteed is a delicate issue and involves carefully chosen assumptions, cf. Zhang et al. (2017, Section 2.3). Convergence guarantees for J_k as defined above with an arbitrary (proper, convex, and l.s.c.) function R which is an open problem. Having considered incremental variants of Bregman proximal algorithms, we now want to review coordinate descent adaptations of this algorithm in the following section.

Bregman Coordinate Descent Methods

In the previous section, we have reviewed Bregman adaptations of popular algorithms for minimising objective functions that are sums of individual objective functions that occur in numerous large-scale applications, such as empirical risk minimisation in machine learning.

In this section, we want to focus on Bregman adaptations of algorithms that aim to minimise multi-variable functions $F : \mathbb{R}^n \to \mathbb{R}$ by minimising the objective with respect to one variable at a time. If we consider (1) for example, a simple coordinate descent adaption is

$$x_i^{k+1} = \arg\min_{x \in \mathbb{R}} \left\{ F(x_1^{k+1}, x_2^{k+1}, \ldots, x_{i-1}^{k+1}, x, x_{i+1}^k, \ldots, x_n^k) + D_{J_i}(x, x_i^k) \right\},$$

See, for example, Hua and Yamashita (2016), Corona et al. (2019a,b), Ahookhosh et al. (2019), Benning et al. (2020), and Gao et al. (2020). In the following, we want to give a brief overview on Bregman coordinate descent-type methods, with particular emphasis on an Itoh-Abe discrete gradient-based method, and also highlight their connections to traditional coordinate descent algorithms (and their Bregman adaptations) such as successive over-relaxation (SOR).

The Bregman Itoh–Abe Method

The Bregman Itoh–Abe (BIA) method (Benning et al. 2020) is a particular form for coordinate descent, derived by applying the discrete gradient method to the ISS flow (15). Discrete gradients are methods from geometric numerical integration for solving differential equations while preserving geometric structures – for details on geometric numerical integration, see, e.g. Hairer et al. (2006) and McLachlan and Quispel (2001) – and have found several applications to optimisation, e.g. Benning et al. (2020), Grimm et al. (2017), Ehrhardt et al. (2018), Riis et al. (2018), and Ringholm et al. (2018) due to their ability to preserve energy dissipation laws.

A discrete gradient is an approximation to a gradient that must satisfy two properties as follows.

Definition (Discrete gradient). Let F be a continuously differentiable function. A *discrete gradient* is a continuous map $\overline{\nabla} F : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ such that for all $x, y \in \mathbb{R}^n$,

$$\langle \overline{\nabla} F(x, y), y - x \rangle = F(y) - F(x) \quad \text{(Mean value)}, \tag{35}$$

$$\lim_{y \to x} \overline{\nabla} F(x, y) = \nabla F(x) \quad \text{(Consistency)}. \tag{36}$$

Given a choice of $\overline{\nabla} F$, starting points x^0, $p^0 \in \partial J(x^0)$, and time steps $(\tau_k)_{k \in \mathbb{N}}$, the Bregman discrete gradient scheme is defined as

$$p^{k+1} = p^k - \tau_k \overline{\nabla} F(x^k, x^{k+1}), \qquad p^{k+1} \in \partial J(x^{k+1}). \tag{37}$$

As with the other Bregman schemes, this is a discretisation of (15). Furthermore, the following dissipation property is an immediate consequence of the definition of discrete gradients.

Remark 3. When $J(x) = \|x\|^2/2$, then the ISS flow reduces to the Euclidean gradient flow, and we refer to the corresponding BIA method simply as the Itoh–Abe (IA) method.

Proposition. *Suppose J is μ-convex and that (x^{k+1}, p^{k+1}) solves the update (35) given (x^k, p^k) and time step $\tau_k > 0$. Then*

$$F(x^{k+1}) - F(x^k) = -\frac{1}{\tau_k} D_J^{symm}(x^k, x^{k+1}) \leq -\frac{\mu}{\tau_k}\|x^k - x^{k+1}\|^2, \tag{38}$$

where $D_J^{symm}(x, y)$ *is the* symmetrised Bregman distance *defined as*

$$D_J^{symm}(x, y) := D_J^p(x, y) + D_J^q(y, x) = \langle p - q, y - x\rangle \quad \textit{for} \quad p \in \partial J(y),\ q \in \partial J(x).$$

Proof. By (35) and (37) respectively, we have

$$F(x^{k+1}) - F(x^k) = \langle \overline{\nabla} F(x^k, x^{k+1}), x^{k+1} - x^k\rangle = -\frac{1}{\tau_k}\langle p^{k+1} - p^k, x^{k+1} - x^k\rangle.$$

The result then follows from monotonicity of convex functions, see, e.g. Hiriart-Urruty and Lemaréchal (1993, Theorem 6.1.2).

While there are various discrete gradients (see, e.g. McLachlan et al. 1999), the *Itoh–Abe discrete gradient* (Itoh and Abe 1988) (also known as the coordinate increment discrete gradient) is of particular interest in optimisation as it is derivative-free and can be implemented for nonsmooth functions. It is defined as

$$\overline{\nabla} F(x, y) = \begin{pmatrix} \frac{F(y_1, x_2, \ldots, x_n) - F(x)}{y_1 - x_1} \\ \frac{F(y_1, y_2, x_3, \ldots, x_n) - F(y_1, x_2, \ldots, x_n)}{y_2 - x_2} \\ \vdots \\ \frac{F(y) - F(y_1, \ldots, y_{n-1}, x_n)}{y_n - x_n} \end{pmatrix}, \tag{39}$$

where $0/0$ is interpreted as $\partial_i F(x)$.

The BIA method is derived by plugging in the Itoh–Abe discrete gradient for $\overline{\nabla} F$ in (37). Provided that J is separable in the coordinates, i.e. $J(x) = \sum_{i=1}^n J_i(x_i)$, for $J_i : \mathbb{R} \to \overline{\mathbb{R}}$, then this method reduces to sequential updates along the coordinates. Specifically, it can be written as

$$\begin{aligned} p_i^{k+1} &= p_i^k - \tau_{k,i} \frac{F(y^{k,i}) - F(y^{k,i-1})}{x_i^{k+1} - x_i^k}, \qquad p_i^{k+1} \in \partial J_i(y_i^{k,i}), \\ y^{k,i} &= [x_1^{k+1}, \ldots, x_i^{k+1}, x_{i+1}^k, \ldots, x_n^k], \qquad i = 1, \ldots, n. \end{aligned} \tag{40}$$

In addition to having a derivative-free formulation, the BIA method has convergence guarantees for a large group of objective functions. In particular, if the Bregman function J is nonsmooth and strongly convex, and if F is locally Lipschitz continuous with a regularity assumption (see Benning et al. 2020 for details), the BIA scheme converges to a set of *Clarke stationary points* (Benning et al. 2020, Theorem 4.5). Clarke stationarity refers to the optimality criteria $0 \in \partial^C F(x)$, where $\partial^C F(x)$ denotes the *Clarke subdifferential* of F at x (Clarke 1990).

This scheme comes with the cost that the updates (40) are in general implicit. However, for the cases

$$J(x) = \frac{1}{2}\|x\|^2, \qquad J(x) = \frac{1}{2}\|x\|^2 + \lambda\|x\|_1,$$
$$F(x) = \frac{1}{2}\|Ax - b^\delta\|^2, \qquad F(x) = \frac{1}{2}\|Ax - b^\delta\|^2 + \gamma\|x\|_1,$$

the updates are explicit (Benning et al. 2020).

In section "Student-t Regularised Image Denoising", we present an example of a nonsmooth, nonconvex image denoising model, previously considered in Benning et al. (2020), for which one can significantly speed up convergence by exploiting sparsity in the residual $x^* - x^\delta$.

Equivalencies of Certain Bregman Coordinate Descent Methods

In what follows, we briefly discuss and draw connections between various approaches to coordinate descent methods using Bregman distances. This builds on the observation by Miyatake et al. (2018) that the Itoh–Abe method applied to quadratic functions $F(x) = \langle x, Ax\rangle/2 - \langle b, x\rangle$ is equivalent to the Gauss–Seidel and successive-over-relaxation (SOR) methods (Young 1971).

The explicit coordinate descent method (Beck and Tetruashvili 2013; Wright 2015) for minimising F is given by

$$\begin{aligned} y^{k,0} &= x^k \\ y^{k,i} &= y^{k,i-1} - \overline{\tau}_i [\nabla F(y^{k,i-1})]_i e^i, \\ x^{k+1} &= y^{k,n}, \end{aligned} \tag{41}$$

where $\overline{\tau}_i > 0$ is the time step and e^i denotes the i^{th} basis vector. As mentioned in Wright (2015), the SOR method is also equivalent to the coordinate descent method with F as above and the time steps scaled coordinate-wise by $1/A_{i,i}$. Hence, in this setting, the Itoh–Abe discrete gradient method is equivalent not only to SOR methods but to explicit coordinate descent.

Furthermore, these equivalencies extend to discretisations of the inverse scale space flow for certain quadratic objective functions and certain forms of Bregman functions J. Consider a quadratic function $F(x) = \langle x, Ax\rangle/2 - \langle b, x\rangle$ where A is symmetric and positive definite, and denote by B the diagonal matrix for which $A_{i,i} = B_{i,i}$ for each i. Given a scaling parameter $\omega > 0$ and the Bregman function

$$J(x) = \frac{1}{2\omega}\langle x, Bx\rangle + \lambda\|x\|_1, \tag{42}$$

The Itoh–Abe method yields a sparse SOR scheme as detailed in Benning et al. (2020). We may compare this to a *Bregman linearised coordinate descent* scheme

$$\begin{aligned} y^{k,0} &= x^k, \quad p^k \in \partial J(x^k), \\ z_i &= \arg\min_y [\nabla F(y^{k,i-1})]_i \cdot y + D_J^{p^k}(y^{k,i-1}, y^{k,i-1} + ye^i), \\ y^{k,i} &= y^{k,i-1} + z_i e^i, \\ x^{k+1} &= y^{k,n}, \end{aligned}$$

where J is given by (42) for some $\omega = \omega_E \in (0, 2)$. One can verify that these schemes are equivalent if one sets $\omega_E = \frac{1}{1/\omega + 1/2}$. We furthermore mention that these equivalencies also hold if we were to consider (implicit) Bregman iterations rather than linearised ones.

Remark 4. It is worth noting at this stage that while the Kaczmarz method (27) is closely related to SOR (Oswald and Zhou 2015), this connection does not carry over to the BIA method versus the sparse Kaczmarz method.

Saddle-Point Methods

Many problems in imaging (Chambolle and Pock 2016a) and machine learning (Goldstein et al. 2015; Adler and Öktem 2018) can be formulated as minimisation problems of the form

$$\min_{x \in \mathbb{R}^n, z \in \mathbb{R}^m} G(x) + F(z) \quad \text{subject to} \quad K(x, z) = c. \tag{43}$$

Here $G : \mathbb{R}^n \to \overline{\mathbb{R}}$ and $F : \mathbb{R}^m \to \overline{\mathbb{R}}$ are proper and lower semi-continuous and usually also convex functions, the operator $K : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^s$ is a bounded, and usually linear operator and $c \in \mathbb{R}^s$ are a vector. A classical linear example for K is

$$K(x, z) = Ax + Bz,$$

where $A \in \mathbb{R}^{s \times n}$ and $B \in \mathbb{R}^{s \times m}$ are matrices (Boyd et al. 2011).

In terms of optimisation, the equality constraint can be incorporated with the help of a Lagrange multiplier $y \in \mathbb{R}^s$. We can then re-formulate (43) as finding a saddle point of an augmented Lagrange function, i.e. we solve

$$\min_{x \in \mathbb{R}^n, z \in \mathbb{R}^m} \max_{y \in \mathbb{R}^s} \mathscr{L}_\delta(x, z; y)$$

for the augmented Lagrangian

$$\mathscr{L}_\delta(x, z; y) := G(x) + F(z) + \langle y, K(x, z) - c \rangle + \frac{1}{2\delta} \|K(x, z) - c\|^2, \tag{44}$$

where $\delta > 0$ is a positive scalar. For the special case $K(x, z) = Ax - z$ and $c \equiv 0$, one can replace $F(Ax)$ with its convex conjugate and formulate the alternative saddle-point problem

$$\min_{x\in\mathbb{R}^n} \max_{y\in\mathbb{R}^m} G(x) + \langle Ax, y\rangle - F^*(y), \tag{45}$$

where the convex conjugate or Fenchel conjugate F^* of F is defined as

$$F^*(y) := \sup_{x\in\mathbb{R}^n} \langle x, y\rangle - F(x).$$

We want to emphasise that extensions for nonconvex functions (Li and Pong 2015; Moeller et al. 2015; Möllenhoff et al. 2015) and extensions for nonlinear operators A (Valkonen 2014; Benning et al. 2015; Clason and Valkonen 2017) or nonlinear replacements of the dual product (Clason et al. 2019) exist. In the following, we review Bregman algorithms for the numerical computation of solutions of those saddle-point formulations.

Alternating Direction Method of Multipliers

The alternating direction method of multipliers (ADMM), (Gabay 1983), is a coordinate descent method applied to the augmented Lagrangian functional (44). The augmented Lagrangian is furthermore modified to also include appropriate penalisation terms, so that we compute

$$x^{k+1} = \arg\min_{x\in\mathbb{R}^n} \mathscr{L}_\delta(x, z^k; \mu^k) + D_{J_x}(x, x^k), \tag{46a}$$

$$z^{k+1} = \arg\min_{z\in\mathbb{R}^m} \mathscr{L}_\delta(x^{k+1}, z; y^k) + D_{J_z}(z, z^k), \tag{46b}$$

$$y^{k+1} = \arg\max_{y\in\mathbb{R}^m} \mathscr{L}_\delta(x^{k+1}, z^{k+1}; y) - D_{J_y}(y, y^k), \tag{46c}$$

in an alternating fashion. To our knowledge, the first adaptation of ADMM to more general Bregman functions was proposed in Wang and Banerjee (2014). In the setting discussed here, the functions J_x, J_z, and J_y are convex and continuously differentiable functions. In the most basic scenario, we choose $K(x, z) = Ax + Bz$, J_x, and J_y as the zero functions, i.e. $J_x(x) = 0$ and $J_z(z) = 0$ for all $x \in \mathbb{R}^n$ $z \in \mathbb{R}^m$, while J_y is chosen to be a positive multiple of the squared Euclidean norm $J_y(y) := \frac{1}{2\tau}\|y\|^2$. Then (46) reduces to the classical ADMM setting (cf. Boyd et al. 2011)

$$x^{k+1} = \left(A^\top A + \delta\,\partial G\right)^{-1}\left(A^\top\left(c - (Bz^k + \delta y^k)\right)\right),$$

$$z^{k+1} = \left(B^\top B + \delta\,\partial F\right)^{-1}\left(B^\top\left(c - (Ax^{k+1} + \delta y^k)\right)\right),$$

$$y^{k+1} = y^k + \tau \left(A x^{k+1} + B z^{k+1} - c \right).$$

Depending on the choices of J_x, J_z, and J_y, many other useful variants are possible, such as

$$\begin{aligned}
x^{k+1} &= (I + \tau_x \delta \, \partial G)^{-1} \left(x^k - \tau_x A^\top \left(A x^k + B z^k + \delta y^k - c \right) \right), \\
z^{k+1} &= \left(I + \tau_z \delta \, \partial F \right)^{-1} \left(z^k - \tau_z B^\top \left(A x^{k+1} + B z^k + \delta y^k - c \right) \right), \\
y^{k+1} &= y^k + \tau_y \left(A x^{k+1} + B z^{k+1} - c \right),
\end{aligned}$$

for the choices $J_x(x) = \frac{1}{2\delta\,\tau_x}\|x\|^2 - \frac{1}{2\delta}\|Ax\|^2$, $J_z(z) = \frac{1}{2\delta\,\tau_z}\|z\|^2 - \frac{1}{2\delta}\|Bz\|^2$, and $J_y(y) = \frac{1}{2\tau_y}\|y\|^2$, which is fully explicit with respect to the operators A and B. Moreover, J_x is convex for $0 < \tau_x < \|A\|^2$, while J_z is convex for $0 < \tau_z < \|B\|^2$. A unified Bregman framework for primal-dual algorithms is discussed in greater detail in Zhang et al. (2011).

Primal-Dual Hybrid Gradient Method

In this section we focus on the special saddle-point formulation (45). It is straightforward to verify that for convex G and F a saddle point $(\hat{x}, \hat{y})^\top$ is characterised by the optimality system

$$0 \in \partial G(\hat{x}) + A^\top \hat{y}, \tag{47a}$$

$$0 \in \partial F^*(\hat{y}) - A\hat{x}. \tag{47b}$$

It is sensible and has indeed been suggested in Chambolle and Pock (2016b), and Hohage and Homann (2014) to solve this nonlinear inclusion problem with a fixed point algorithm of the form

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix} \in \begin{pmatrix} \partial G(x^{k+1}) + A^\top y^{k+1} \\ \partial F^*(y^{k+1}) - A x^{k+1} \end{pmatrix} + \partial J(x^{k+1}, y^{k+1}) - \partial J(x^k, y^k). \tag{48}$$

Here ∂J denotes the subdifferential of some convex function $J : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$. For the choice

$$J(x,y) := \frac{1}{2}\left\|\begin{pmatrix}x\\y\end{pmatrix}\right\|_M^2 \quad \text{with} \quad \left\|\begin{pmatrix}x\\y\end{pmatrix}\right\|_M := \sqrt{\left\langle M\begin{pmatrix}x\\y\end{pmatrix}, \begin{pmatrix}x\\y\end{pmatrix}\right\rangle}$$
$$\text{and} \quad M := \begin{pmatrix}\frac{1}{\tau}I & -A^\top \\ -A & \frac{1}{\sigma}I\end{pmatrix},$$

and $\tau\sigma\|A\|^2 < 1$, we obtain the conventional primal-dual hybrid gradient (PDHG) method (with relaxation parameter set to one) as proposed and discussed in Zhu and Chan (2008), Pock et al. (2009), Esser et al. (2010), and Chambolle and Pock (2011, 2016a), which reads

$$x^{k+1} = (I + \tau\partial G)^{-1}\left(x^k - \tau A^\top y^k\right), \tag{49a}$$
$$y^{k+1} = \left(I + \sigma\partial F^*\right)^{-1}\left(y^k + \sigma A(2x^{k+1} - x^k)\right). \tag{49b}$$

Note that we can reformulate (48) to

$$\begin{pmatrix}0\\0\end{pmatrix} \in \begin{pmatrix}\partial G(x^{k+1}) - \partial G(\hat{x}) + A^\top(y^{k+1} - \hat{y})\\ \partial F^*(y^{k+1}) - \partial F^*(\hat{y}) - A(x^{k+1} - \hat{x})\end{pmatrix}$$
$$+ \partial J(x^{k+1}, y^{k+1}) - \partial J(\hat{x}, \hat{y}) - \left(\partial J(x^k, y^k) - \partial J(\hat{x}, \hat{y})\right), \tag{50}$$

if we add the optimality system (47) to (48), for a saddle point $(\hat{x}, \hat{y})^\top$. Taking a dual product of

$$\begin{pmatrix}\partial G(x^{k+1}) - \partial G(\hat{x}) + A^\top(y^{k+1} - \hat{y})\\ \partial F^*(y^{k+1}) - \partial F^*(\hat{y}) - A(x^{k+1} - \hat{x})\end{pmatrix}$$

with $(x^{k+1} - \hat{x},\, y^{k+1} - \hat{y})^\top$ therefore yields

$$\left\langle\begin{pmatrix}\partial G(x^{k+1}) - \partial G(\hat{x}) + A^\top(y^{k+1} - \hat{y})\\ \partial F^*(y^{k+1}) - \partial F^*(\hat{y}) - A(x^{k+1} - \hat{x})\end{pmatrix}, \begin{pmatrix}x^{k+1} - \hat{x}\\ y^{k+1} - \hat{y}\end{pmatrix}\right\rangle$$
$$= D_G^{\text{symm}}(x^{k+1}, \hat{x}) + D_{F^*}^{\text{symm}}(y^{k+1}, \hat{y}) \geq 0.$$

Here $D_J^{\text{symm}}(x,y)$ denotes the symmetric Bregman distance $D_J^{\text{symm}}(x,y) = D_J^q(x,y) + D_J^p(y,x) = \langle p - q, x - y\rangle$, for subgradients $p \in \partial J(x)$ and $q \in \partial J(y)$, which is also known as Jeffreys–Bregman divergence and closely related to other symmetrisations such as Jensen–Bregman divergences (Nielsen and Boltz 2011) and Burbea Rao distances (Burbea and Rao 1982b,a). As an immediate consequence, we observe

$$0 \geq \left\langle \partial J(x^{k+1}, y^{k+1}) - \partial J(x^k, y^k), \begin{pmatrix} x^{k+1} - \hat{x} \\ y^{k+1} - \hat{y} \end{pmatrix} \right\rangle$$
$$= D_J\left(\begin{pmatrix} \hat{x} \\ \hat{y} \end{pmatrix}, \begin{pmatrix} x^{k+1} \\ y^{k+1} \end{pmatrix}\right) - D_J\left(\begin{pmatrix} \hat{x} \\ \hat{y} \end{pmatrix}, \begin{pmatrix} x^{k} \\ y^{k} \end{pmatrix}\right) + D_J\left(\begin{pmatrix} x^{k+1} \\ y^{k+1} \end{pmatrix}, \begin{pmatrix} x^{k} \\ y^{k} \end{pmatrix}\right),$$

where we have made use of the three-point identity for Bregman distances (Chen and Teboulle 1993). Thus, we can conclude

$$D_J\left(\begin{pmatrix} \hat{x} \\ \hat{y} \end{pmatrix}, \begin{pmatrix} x^{k+1} \\ y^{k+1} \end{pmatrix}\right) + D_J\left(\begin{pmatrix} x^{k+1} \\ y^{k+1} \end{pmatrix}, \begin{pmatrix} x^{k} \\ y^{k} \end{pmatrix}\right) \leq D_J\left(\begin{pmatrix} \hat{x} \\ \hat{y} \end{pmatrix}, \begin{pmatrix} x^{k} \\ y^{k} \end{pmatrix}\right)$$

for all iterates. Consequently, the iterates are bounded in the Bregman distance setting with respect to J. Summing up the dual product of (48) with $(x^{k+1} - \hat{x}, y^{k+1} - \hat{y})^\top$ therefore yields

$$\sum_{k=0}^{N} \left[D_G^{\text{symm}}(x^{k+1}, \hat{x}) + D_{F^*}^{\text{symm}}(y^{k+1}, \hat{y}) \right] + \sum_{k=0}^{N} D_J\left(\begin{pmatrix} x^{k+1} \\ y^{k+1} \end{pmatrix}, \begin{pmatrix} x^{k} \\ y^{k} \end{pmatrix}\right)$$
$$= \sum_{k=0}^{N} \left[D_J\left(\begin{pmatrix} \hat{x} \\ \hat{y} \end{pmatrix}, \begin{pmatrix} x^{k} \\ y^{k} \end{pmatrix}\right) - D_J\left(\begin{pmatrix} \hat{x} \\ \hat{y} \end{pmatrix}, \begin{pmatrix} x^{k+1} \\ y^{k+1} \end{pmatrix}\right) \right] \leq D_J\left(\begin{pmatrix} \hat{x} \\ \hat{y} \end{pmatrix}, \begin{pmatrix} x^{0} \\ y^{0} \end{pmatrix}\right)$$
$$< +\infty.$$

Hence, we can conclude $D_G^{\text{symm}}(x^N, \hat{x}) \to 0$, $D_{F^*}^{\text{symm}}(y^N, \hat{y}) \to 0$, and $D_J\left(\begin{pmatrix} x^N & y^N \end{pmatrix}^\top, \begin{pmatrix} x^k & y^k \end{pmatrix}^\top\right) \to 0$ for $N \to \infty$. If G and F^* are at least convex and if J is strongly convex with respect to some norm, one can further guarantee convergence of the corresponding iterates in norm to a saddle-point (x, y) solution of (45) with standard arguments. For more details, analysis, and extensions of PDHG methods, we refer the reader to Chambolle and Pock (2016a).

Applications

In the following we want to show applications for some of the Bregman algorithms discussed in this review chapter. We want to emphasise that none of the applications shown are really large-scale applications. The idea of this section is rather to demonstrate that the algorithms are applicable to a wide range of different problems, offering the potential to enhance actual large-scale problems. We focus on three combinations of applications and algorithms: robust principal component analysis

via the accelerated linearised Bregman iteration, deep learning with an incremental proximal Bregman architecture, and image denoising via the Bregman Itoh–Abe method.

Robust Principal Component Analysis

Robust principal component analysis is an extension of principal component analysis first proposed in Candès et al. (2011). The key idea is to decompose a matrix $X \in \mathbb{R}^{m\times n}$ into a low-rank matrix $L \in \mathbb{R}^{m\times n}$ and a sparse matrix $S \in \mathbb{R}^{m\times n}$ by solving the optimisation problem

$$\min_{L,S} \alpha_1 \|L\|_* + \alpha_2 \|S\|_1 \qquad \text{subject to} \qquad X = L + S. \tag{51}$$

Here $\|S\|_1$ is the one norm of the matrix S, i.e. $\|S\|_1 = \sum_{i=1}^{m}\sum_{j=1}^{n} |s_{ij}|$, while $\|L\|_*$ denotes the nuclear norm of L, which is the one norm of the singular values of L, i.e. $\|L\|_* = \sum_{j=1}^{\min(n,m)} \sigma_j$, for $L = U\Sigma V^*$ with $\Sigma_{ij} = \begin{cases} \sigma_j & i = j \\ 0 & i \neq j \end{cases}$ and U and V being orthogonal. There are numerous strategies for solving (51) numerically (Bouwmans et al. 2018); we focus on using the accelerated linearised Bregman iteration as discussed in section "Accelerated Bregman Methods". For this we use formulation (12) of the linearised Bregman iteration, respectively (19), in the accelerated case. We choose $A = \begin{pmatrix} I & I \end{pmatrix}^\top$, $b^\delta = X$, and $R = \alpha_1\|\cdot\|_* + \alpha_2\|\cdot\|_1$ and therefore obtain

$$\begin{aligned} L^{k+1} &= \left(I + \alpha_1 \partial\|\cdot\|_*\right)^{-1}\left(\tau X^k\right), \\ S^{k+1} &= \left(I + \alpha_2 \partial\|\cdot\|_1\right)^{-1}\left(\tau X^k\right), \\ X^{k+1} &= X^k - \left(L^{k+1} + S^{k+1} - X\right), \end{aligned}$$

in the case of (12), respectively

$$\begin{aligned} L^{k+1} &= \left(I + \alpha_1 \partial\|\cdot\|_*\right)^{-1}\left(\tau Y^k\right), \\ S^{k+1} &= \left(I + \alpha_2 \partial\|\cdot\|_1\right)^{-1}\left(\tau Y^k\right), \\ X^{k+1} &= Y^k - \left(L^{k+1} + S^{k+1} - X\right), \\ Y^{k+1} &= (1+\beta_{k+1})X^{k+1} - \beta_{k+1}X^k, \end{aligned}$$

(a) (b) (c) (d)

Fig. 1 From left to right: the first image of the Yale B faces database, its approximation which is the sum of a low-rank and a sparse matrix, the low-rank matrix, and the sparse matrix. (**a**) Original (**b**) Approximation (**c**) Low-rank part (**d**) Sparse part

in the case of (17), for $X^0 := X$. We choose the parameters to be $\tau = 1/\|A\|^2 = 1/2$, $\alpha_1 = 10\sqrt{\max(m,n)}$, $\alpha_2 = 10$, and $\beta_k = (k-1)/(k+3)$ for $k \geq 1$. Note that the latter automatically implies $Y^0 = X$. We run the algorithm on two test datasets; inspired by Brunton and Kutz (2019), the first one is the Yale Faces B dataset (Lee et al. 2005), and the second one is a video sequence of a Cornell box with a moving shadow, from (Benning et al. 2007). Figure 1 shows the first image of the Yale B faces database, its approximation, and its decomposition into a low-rank and a sparse part.

The more important aspect in terms of this review paper is certainly the comparison between the linearised Bregman iteration and its accelerated counterpart. A log-scale plot of the decrease of the loss function $\frac{1}{2}\|L + S - X\|_{\mathrm{F}}^2$, where $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm, over the course of the iterations of the two algorithms is visualised in Fig. 2. The plot is an empirical validation that (18) converges at rate $\mathcal{O}(1/k^2)$ as opposed to the $\mathcal{O}(1/k)$ rate of its non-accelerated counterpart.

In Fig. 3 we see the 1st, 50th, 100th, and 150th frame of the original Cornell box video sequence from Benning et al. (2007), together with a low-rank approximation and a sparse component computed with the accelerated linearised Bregman iteration.

Deep Learning

Ever since Alexnet entered the scene in 2012 (Krizhevsky et al. 2012), thwarting then state-of-the-art image classification approaches in terms of accuracy in the process, deep neural networks (DNNs) have been central to research in computer vision and imaging. In this section, we merely want to support the analogy between incremental Bregman proximal methods and DNNs as shown in section "Deep

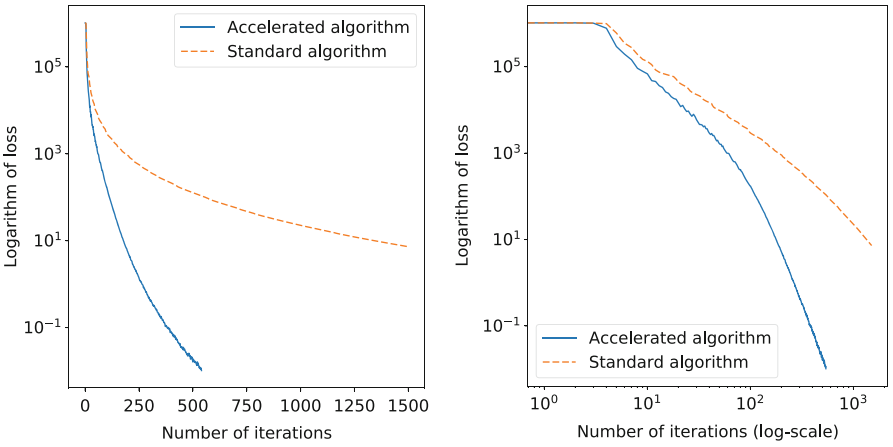

Fig. 2 This is an empirical validation of the different convergence rates of the linearised Bregman iteration and its accelerated counterpart (with regular scaling of the iterations on the left-hand side and a logarithmic scaling on the right-hand side)

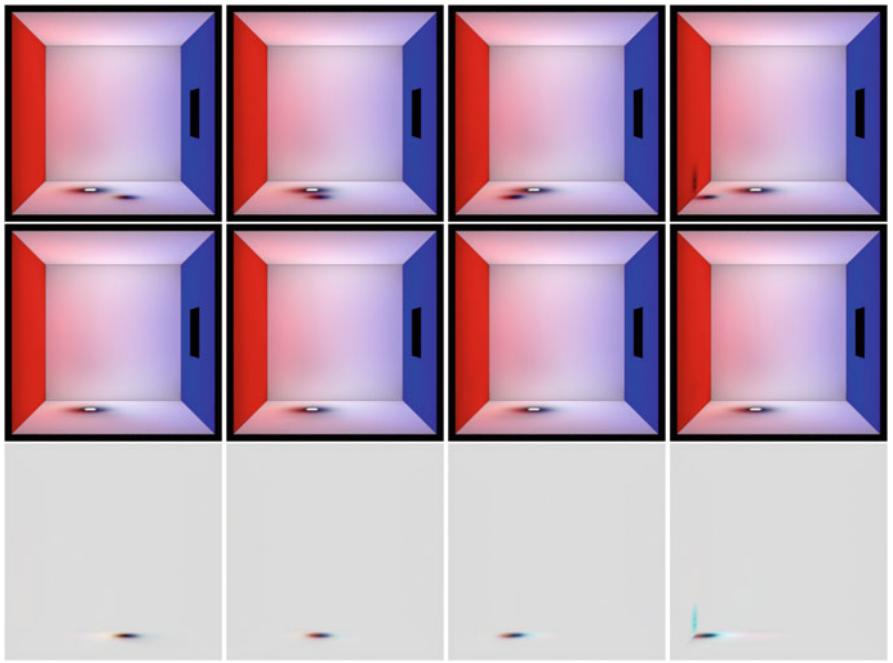

Fig. 3 First row: the 1st, 50th, 100th, and 150th frame of the original video sequence from Benning et al. (2007). Second row: the same frames of the computed low-rank part. Third row: the same frames of the computed sparse part

Neural Networks" with a practical example, rather than engaging in a discussion of when and why DNNs based on (30) should be used or what advantages or shortcomings they possess compared to other neural network architectures. For a comprehensive overview over developments in deep learning, we refer the reader to Goodfellow et al. (2016).

In this example, we set up a DNN-based auto-encoder for dimensionality reduction and compare it to classical dimensionality reduction via singular value decomposition. The auto-encoder is of the form

$$\begin{aligned} x^k &= \left(I + \partial\|\cdot\|_1\right)^{-1}\left(A_k x^{k-1} + b_k\right), \\ &= S_1\left(A_k x^{k-1} + b_k\right), \end{aligned}$$

for $k \in \{1,2,3,4\}$ and $x^0 = x$, where x denotes the input of the network, $A_k := \frac{1}{2}(M_k + M_k^\top) \circ H_k$ for matrices $M_k \in \mathbb{R}^{m_k \times m_k}$ dimensions $m_1 = 196$, $m_2 = 49$, $m_3 = 196$, and $m_4 = 784$, and where H_1 and H_2 are two-dimensional average pooling operators with window size 2×2 and H_3 and H_4 are nearest-neighbour interpolation operators that upscale by a factor of two. The vectors $\{b_k\}_{k=1}^4$ are bias vectors of dimensions $\{m_k\}_{k=1}^4$, and the operator S_1 is the soft-shrinkage operator as described in section "The Sparse Kaczmarz Method". Please note that this auto-encoder architecture is of the form (30) and represents a parametrised mapping Φ_Θ from $\mathbb{R}^{784}$ to $\mathbb{R}^{784}$, where $\Theta = (\{M_k\}_{k=1}^4, \{b_k\}_{k=1}^4)$ denotes the collection of parameters. We train the auto-encoder by minimising the empirical risk based on the mean-squared error for a set of samples $\{x_i\}_{i=1}^s$, $s = 60000$, via stochastic gradient descent (which is the randomised version of (24)), i.e. we approximately estimate optimal parameters $\hat{\Theta}$ via

$$\hat{\Theta} = \arg\min_\Theta \frac{1}{2s}\sum_{i=1}^{s}\left(\Phi_\Theta(x_i) - x_i\right)^2.$$

We emphasise that the soft-thresholding activation function S_1 leaves Φ_Θ as not differentiable, which is why the application of (24) is technically a stochastic subgradient method. We train the auto-encoder with the help of PyTorch for a fixed number of epochs (500) and fixed step size $\tau = 2$ with batch size 100 on the MNIST training dataset (LeCun et al. 1998). In Fig. 4, we visualised several samples and the corresponding transformed outputs of the auto-encoder. In Fig. 5, we have visualised random images from the same dataset in comparison to their truncated singular value decomposition reconstructions where all but the first 49 singular values are cut off. As to be expected, nonlinear dimensionality reduction can outperform linear dimensionality reduction, achieving visually superior results for the same subspace dimensionalty.

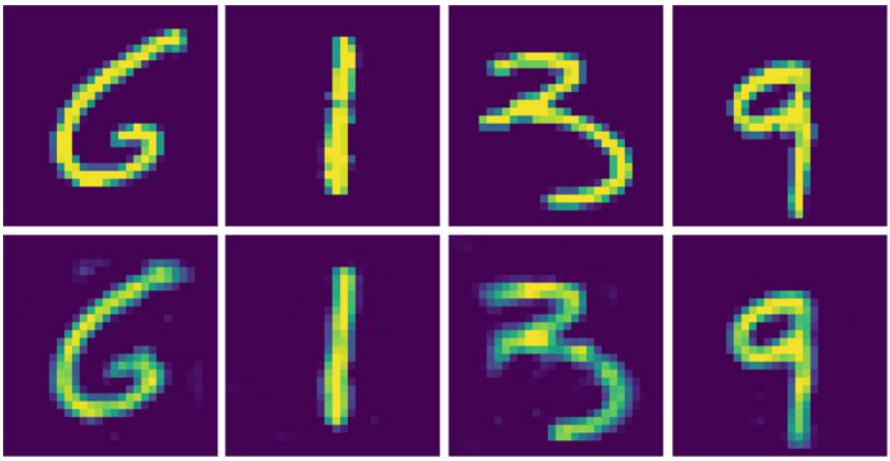

Fig. 4 Top row: random samples from the MNIST dataset. Bottom row: the corresponding approximations with the trained auto-encoder

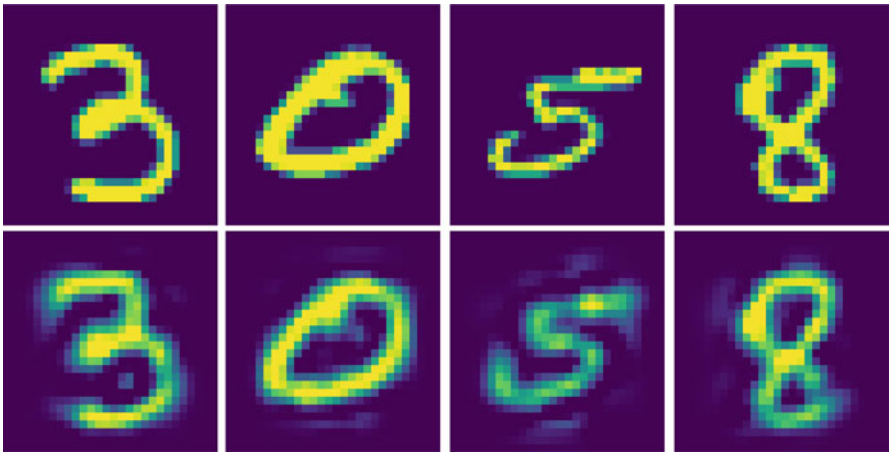

Fig. 5 Top row: random samples from the MNIST dataset. Bottom row: the corresponding approximations with the first 49 singular vectors

Student-t Regularised Image Denoising

In what follows, we apply BIA methods for solving a nonsmooth, nonconvex image denoising model, previously presented in Ochs et al. (2014). A priori knowledge of the noise distribution allows the use of Bregman functions $J(x)$ that exploit sparsity structures of the problem. As we will see, this yields significantly improved convergence rates in comparison with the default Itoh–Abe scheme (i.e. $J(x) =$

$\|x\|^2/2$). The application of the BIA method for this example was previously presented in Benning et al. (2020).

The objective function is given by

$$F : \mathbb{R}^n \to \mathbb{R}, \quad F(x) := \sum_{i=1}^{N} \varphi_i \Phi(K_i x) + \|x - x^\delta\|_1. \tag{52}$$

Here $\{K_i\}_{i=1}^N$ is a collection of linear filters, $(\varphi_i)_{i=1}^N \subset [0, \infty)$ are coefficients, $\Phi : \mathbb{R}^n \to \mathbb{R}$ is the nonconvex function based on the student-t distribution, defined as

$$\Phi(x) := \sum_{j=1}^{n} \psi(x_i), \quad \psi(x) := \log(1 + x^2),$$

and x^δ is an image corrupted by impulse noise (salt and pepper noise).

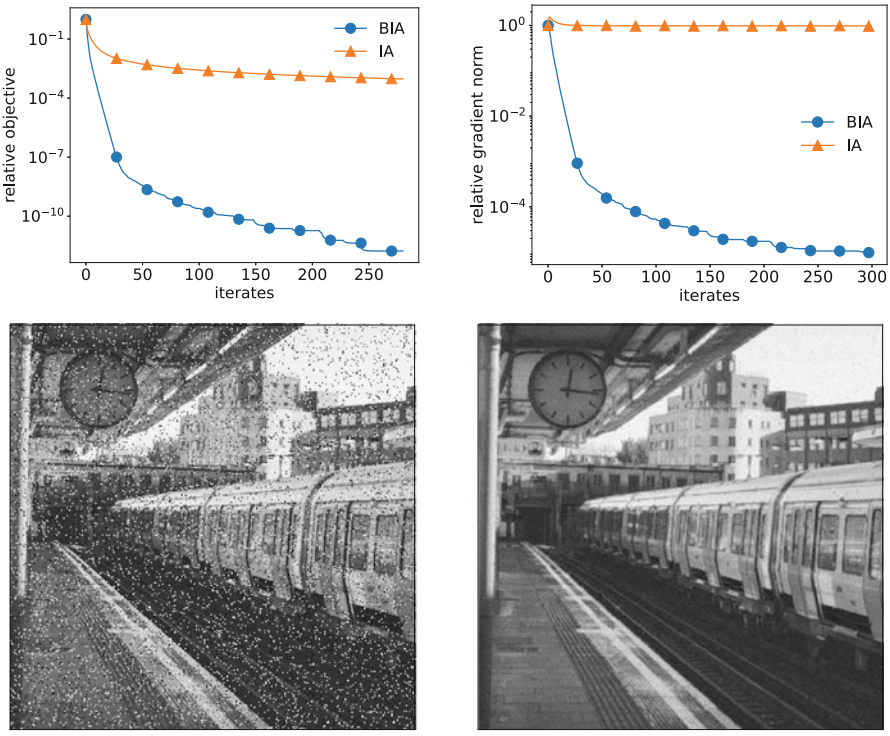

Fig. 6 Comparison of BIA and IA methods, for student-t regularised image denoising. First: convergence rate for relative objective. Second: convergence rate for relative gradient norm. Third: input data. Fourth: reconstruction

As impulse noise only affects a fraction of pixels, we use the data fidelity term $x \mapsto \|x - x^\delta\|_1$ to promote sparsity of $x^* - x^\delta$ for $x^* \in \arg\min F(x)$. As linear filters, we consider the simple case of finite difference approximations to first-order derivatives of x. We note that by applying a gradient flow to this regularisation function, we observe a similarity to Perona–Malik diffusion (Perona and Malik 1990).

For the BIA method, we consider the Bregman function

$$J(x) := \frac{1}{2}\|x\|^2 + \gamma\|x - x^\delta\|_1,$$

to account for the sparsity of the residual $x^* - x^\delta$ and compare the method to the regular Itoh–Abe discrete gradient method (abbreviated to IA).

We set the starting point $x^0 = x^\delta$ and the parameters to $\tau_k = 1$ for all k, $\gamma = 0.5$, and $\varphi_i = 2$, $i = 1, 2$. For the impulse noise, we use a noise density of 10%. In the case where x_i^{k+1} is not set to x_i^δ, we use the scalar root solver *scipy.optimize.brenth* on Python. Otherwise, the updates are in closed form.

See Fig. 6 for numerical results. By gradient norm, we mean $\mathrm{dist}(\partial^C F(x^k), 0)$.

Conclusions and Outlook

In this review paper, we gave a selective overview on a range of topics concerning adaptations of Bregman algorithms suited for large-scale problems in imaging. In particular, we discussed Nesterov accelerations of the Bregman (proximal) gradient or linearised Bregman iteration, incremental variants of Bregman methods, and coordinate descent-type Bregman algorithms with a particular focus on a Bregman Itoh–Abe scheme.

Despite the variety of numerous adaptations, a lot of research on Bregman algorithms is yet to be done. We conclude this chapter by discussing some open problems as well as ongoing directions of research.

Examples of open problems are adaptations for nonconvex objectives (following recent advances in papers such as Ahookhosh et al. 2019), extensions to nonlinear inverse problems (Bachmayr and Burger 2009) or inverse problems with non-quadratic data fidelity terms (Benning and Burger 2011) and the closer analysis and numerical realisation of neural network architectures inspired by Bregman algorithms. We also want to emphasise that Bregman variants of incremental or stochastic variants of ADMM or the PDHG method in the spirit of Ouyang et al. (2013) and Chambolle et al. (2018) are still open problems.

Another important topic of ongoing research is to understand the scope for and limitations of accelerated Bregman methods, as stated by Teboulle (2018). Dragomir et al. (2019) point out the open problem of whether accelerated Bregman methods are possible if one makes further assumptions on the objective and Bregman functions or by allowing access to second-order information. Another interesting approach is to consider ODEs – see, e.g. Krichene et al. (2015) in which Krichene et

al. investigate accelerating mirror descent via the ODE interpretation of Nesterov's acceleration (Su et al. 2016).

Going from optimisation to sampling, some recent papers consider methods for sampling of distributions which incorporate elements of mirror descent in the underlying dynamics. Hsieh et al. (2018) propose a framework for sampling from constrained distributions, termed *mirrored Langevin dynamics.* In a similar vein, Zhang et al. (2020) propose a Mirror Langevin Monte Carlo algorithm, to improve the smoothness and convexity properties for the distribution.

Acknowledgments MB thanks Queen Mary University of London for their support. ESR acknowledges support from the London Mathematical Society.

References

Adler, J., Öktem, O.: Learned primal-dual reconstruction. IEEE Trans. Med. Imaging **37**(6), 1322–1332 (2018)

Ahookhosh, M., Hien, L.T.K., Gillis, N., Patrinos, P.: Multi-block bregman proximal alternating linearized minimization and its application to sparse orthogonal nonnegative matrix factorization. arXiv preprint arXiv:1908.01402 (2019)

Arridge, S., Maass, P., Öktem, O., Schönlieb, C.-B.: Solving inverse problems using data-driven models. Acta Numerica **28**, 1–174 (2019)

Attouch, H., Buttazzo, G., Michaille, G.: Variational analysis in Sobolev and BV spaces: applications to PDEs and optimization. SIAM (2014)

Azizan, N., Hassibi, B.: Stochastic gradient/mirror descent: Minimax optimality and implicit regularization. arXiv preprint arXiv:1806.00952 (2018)

Bachmayr, M., Burger, M.: Iterative total variation schemes for nonlinear inverse problems. Inverse Prob. **25**(10), 105004 (2009)

Bauschke, H.H., Bolte, J., Teboulle, M.: A descent lemma beyond lipschitz gradient continuity: first-order methods revisited and applications. Math. Oper. Res. **42**(2), 330–348 (2017)

Bauschke, H.H., Borwein, J.M., Combettes, P.L.: Bregman monotone optimization algorithms. SIAM J. Control. Optim. **42**(2), 596–636 (2003)

Beck, A.: First-Order Methods in Optimization, Vol. 25. SIAM (2017)

Beck, A., Teboulle, M.: Mirror descent and nonlinear projected subgradient methods for convex optimization. Oper. Res. Lett. **31**(3), 167–175 (2003)

Beck, A., Tetruashvili, L.: On the convergence of block coordinate descent type methods. SIAM J. Optim. **23**(4), 2037–2060 (2013)

Ben-Tal, A., Margalit, T., Nemirovski, A.: The ordered subsets mirror descent optimization method with applications to tomography. SIAM J. Optim. **12**(1), 79–108 (2001)

Benning, M., Betcke, M., Ehrhardt, M., Schönlieb, C.-B.: Gradient descent in a generalised bregman distance framework. In: Geometric Numerical Integration and its Applications, Vol. 74, pp. 40–45. MI Lecture Notes series of Kyushu University (2017)

Benning, M., Betcke, M.M., Ehrhardt, M.J., Schönlieb, C.-B.: Choose your path wisely: gradient descent in a bregman distance framework. SIAM Journal on Imaging Sciences (SIIMS). arXiv preprint arXiv:1712.04045 (2017)

Benning, M., Burger, M.: Error estimates for general fidelities. Electron. Trans. Numer. Anal. **38**(44–68), 77 (2011)

Benning, M., Burger, M.: Modern regularization methods for inverse problems. Acta Numerica **27**, 1–111 (2018)

Benning, M., Knoll, F., Schönlieb, C.-B., Valkonen, T.: Preconditioned admm with nonlinear operator constraint. In: IFIP Conference on System Modeling and Optimization, pp. 117–126. Springer (2015)

Benning, M., Lee, E., Pao, H., Yacoubou-Djima, K., Wittman, T., Anderson, J.: Statistical filtering of global illumination for computer graphics. IPAM Research in Industrial Projects for Students (RIPS) Report (2007)
Benning, M., Riis, E.S., Schönlieb, C.-B.: Bregman Itoh–Abe methods for sparse optimisation. In print: J. Math. Imaging Vision (2020)
Bertocchi, C., Chouzenoux, E., Corbineau, M.-C., Pesquet, J.-C., Prato, M.: Deep unfolding of a proximal interior point method for image restoration. Inverse Prob. **36**, 034005 (2019)
Bertsekas, D.P.: Incremental gradient, subgradient, and proximal methods for convex optimization: A survey. Optim. Mach. Learn. **2010**(1–38), 3 (2011)
Bertsekas, D.P.: Incremental proximal methods for large scale convex optimization. Math. Program. **129**(2), 163 (2011)
Blatt, D., Hero, A.O., Gauchman, H.: A convergent incremental gradient method with a constant step size. SIAM J. Optim. **18**(1), 29–51 (2007)
Bolte, J., Sabach, S., Teboulle, M., Vaisbourd, Y.: First order methods beyond convexity and lipschitz gradient continuity with applications to quadratic inverse problems. SIAM J. Optim. **28**(3), 2131–2151 (2018)
Bonettini, S., Rebegoldi, S., Ruggiero, V.: Inertial variable metric techniques for the inexact forward–backward algorithm. SIAM J. Sci. Comput. **40**(5), A3180–A3210 (2018)
Bonnans, J.F., Gilbert, J.C., Lemaréchal, C., Sagastizábal, C.A.: A family of variable metric proximal methods. Math. Program. **68**(1–3), 15–47 (1995)
Bouwmans, T., Javed, S., Zhang, H., Lin, Z., Otazo, R.: On the applications of robust pca in image and video processing. Proc. IEEE **106**(8), 1427–1457 (2018)
Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J., et al.: Distributed optimization and statistical learning via the alternating direction method of multipliers. Found. Trends® Mach. Learn. **3**(1), 1–122 (2011)
Bregman, L.M.: The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. USSR Comput. Math. Math. Phys. **7**(3), 200–217 (1967)
Brunton, S.L., Kutz, J.N.: Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control. Cambridge University Press (2019)
Burbea, J., Rao, C.: On the convexity of higher order jensen differences based on entropy functions (corresp.). IEEE Trans. Inf. Theory **28**(6), 961–963 (1982)
Burbea, J., Rao, C.: On the convexity of some divergence measures based on entropy functions. IEEE Trans. Inf. Theory **28**(3), 489–495 (1982)
Burger, M.: Bregman distances in inverse problems and partial differential equations. In: Advances in Mathematical Modeling, Optimization and Optimal Control, pp. 3–33. Springer (2016)
Burger, M., Frick, K., Osher, S., Scherzer, O.: Inverse total variation flow. Multiscale Model. Simul. **6**(2), 366–395 (2007)
Burger, M., Gilboa, G., Moeller, M., Eckardt, L., Cremers, D.: Spectral decompositions using one-homogeneous functionals. SIAM J. Imag. Sci. **9**(3), 1374–1408 (2016)
Burger, M., Gilboa, G., Osher, S., Xu, J.: Nonlinear inverse scale space methods. Commun. Math. Sci. **4**(1), 179–212 (2006)
Burger, M., Moeller, M., Benning, M., Osher, S.: An adaptive inverse scale space method for compressed sensing. Math. Comput. **82**(281), 269–299 (2013)
Burger, M., Osher, S.: Convergence rates of convex variational regularization. Inverse Prob. **20**(5), 1411 (2004)
Burger, M., Resmerita, E., He, L.: Error estimation for bregman iterations and inverse scale space methods in image restoration. Computing **81**(2–3), 109–135 (2007)
Cai, J.-F., Osher, S., Shen, Z.: Convergence of the linearized bregman iteration for ℓ^1-norm minimization. Math. Comput. **78**(268), 2127–2136 (2009)
Cai, J.-F., Osher, S., Shen, Z.: Linearized bregman iterations for compressed sensing. Math. Comput. **78**(267), 1515–1536 (2009)
Cai, J.-F., Osher, S., Shen, Z.: Linearized bregman iterations for frame-based image deblurring. SIAM J. Imag. Sci. **2**(1), 226–252 (2009)

Calatroni, L., Garrigos, G., Rosasco, L., Villa, S.: Accelerated iterative regularization via dual diagonal descent. arXiv preprint arXiv:1912.12153 (2019)
Candès, E.J., Li, X., Ma, Y., Wright, J.: Robust principal component analysis? J. ACM **58**(3), 11 (2011)
Censor, Y., Lent, A.: An iterative row-action method for interval convex programming. J. Optim. Theory Appl. **34**(3), 321–353 (1981)
Censor, Y., Stavros Zenios, A.: Proximal minimization algorithm with d-functions. J. Optim. Theory Appl. **73**(3), 451–464 (1992)
Chambolle, A., Dossal, C.: On the convergence of the iterates of the "fast iterative shrinkage/thresholding algorithm. J. Optim. Theory Appl. **166**(3), 968–982 (2015)
Chambolle, A., Ehrhardt, M.J., Richtárik, P., Carola-Schonlieb, B.: Stochastic primal-dual hybrid gradient algorithm with arbitrary sampling and imaging applications. SIAM J. Optim. **28**(4), 2783–2808 (2018)
Chambolle, A., Pock, T.: A first-order primal-dual algorithm for convex problems with applications to imaging. J. Math. Imaging Vision **40**(1), 120–145 (2011)
Chambolle, A., Pock, T.: An introduction to continuous optimization for imaging. Acta Numerica **25**, 161–319 (2016)
Chambolle, A., Pock, T.: On the ergodic convergence rates of a first-order primal–dual algorithm. Math. Prog. **159**(1–2), 253–287 (2016)
Chavent, G., Kunisch, K.: Regularization of linear least squares problems by total bounded variation. ESAIM Control Optim. Calc. Var. **2**, 359–376 (1997)
Chen, G., Teboulle, M.: Convergence analysis of a proximal-like minimization algorithm using bregman functions. SIAM J. Optim. **3**(3), 538–543 (1993)
Chouzenoux, E., Pesquet, J.-C., Repetti, A.: Variable metric forward–backward algorithm for minimizing the sum of a differentiable function and a convex function. J. Optim. Theory Appl. **162**(1), 107–132 (2014)
Clarke, F.H.: Optimization and Nonsmooth Analysis. Classics in Applied Mathematics, 1st edn. SIAM, Philadelphia (1990)
Clason, C., Mazurenko, S., Valkonen, T.: Acceleration and global convergence of a first-order primal-dual method for nonconvex problems. SIAM J. Optim. **29**(1), 933–963 (2019)
Clason, C., Valkonen, T.: Primal-dual extragradient methods for nonlinear nonsmooth pde-constrained optimization. SIAM J. Optim. **27**(3), 1314–1339 (2017)
Combettes, P.L., Pesquet, J.-C.: Deep neural network structures solving variational inequalities. arXiv preprint arXiv:1808.07526 (2018)
Combettes, P.L., Vũ, B.C.: Variable metric forward–backward splitting with applications to monotone inclusions in duality. Optimization **63**(9), 1289–1318 (2014)
Corona, V., Benning, M., Ehrhardt, M.J., Gladden, L.F., Mair, R., Reci, A., Sederman, A.J., Reichelt, S., Schönlieb, C.-B.: Enhancing joint reconstruction and segmentation with non-convex bregman iteration. Inverse Prob. **35**(5), 055001 (2019)
Corona, V., Benning, M., Gladden, L.F., Reci, A., Sederman, A.J., Schoenlieb, C.-B.: Joint phase reconstruction and magnitude segmentation from velocity-encoded mri data. arXiv preprint arXiv:1908.05285 (2019)
Doan, T.T., Bose, S., Nguyen, D.H., Beck, C.L.: Convergence of the iterates in mirror descent methods. IEEE Control Syst. Lett. **3**(1), 114–119 (2018)
Dragomir, R.-A., Taylor, A., d'Aspremont, A., Bolte, J.: Optimal complexity and certification of bregman first-order methods. arXiv preprint arXiv:1911.08510 (2019)
Duchi, J.C., Agarwal, A., Johansson, M., Jordan, M.I.: Ergodic mirror descent. SIAM J. Optim. **22**(4), 1549–1578 (2012)
Eckstein, J.: Nonlinear proximal point algorithms using bregman functions, with applications to convex programming. Math. Oper. Res. **18**(1), 202–226 (1993)
Ehrhardt, M.J., Riis, E.S., Ringholm, T., Schönlieb, C.-B.: A geometric integration approach to smooth optimisation: Foundations of the discrete gradient method. ArXiv e-prints (2018)
Esser, E., Zhang, X., Chan, T.F.: A general framework for a class of first order primal-dual algorithms for convex optimization in imaging science. SIAM J. Imag. Sci. **3**(4), 1015–1046 (2010)

Frankel, P., Garrigos, G., Peypouquet, J.: Splitting methods with variable metric for Kurdyka–łojasiewicz functions and general convergence rates. J. Optim. Theory Appl. **165**(3), 874–900 (2015)
Frerix, T., Möllenhoff, T., Moeller, M., Cremers, D.: Proximal backpropagation. arXiv preprint arXiv:1706.04638 (2017)
Frick, K., Scherzer, O.: Convex inverse scale spaces. In: International Conference on Scale Space and Variational Methods in Computer Vision, pp. 313–325. Springer (2007)
Gabay, D.: Chapter ix applications of the method of multipliers to variational inequalities. In: Studies in Mathematics and Its Applications, Vol. 15, pp. 299–331. Elsevier (1983)
Gao, T., Lu, S., Liu, J., Chu, C.: Randomized bregman coordinate descent methods for non-Lipschitz optimization. arXiv preprint arXiv:2001.05202 (2020)
Garrigos, G., Rosasco, L., Villa, S.: Iterative regularization via dual diagonal descent. J. Math. Imaging Vision **60**(2), 189–215 (2018)
Gilboa, G., Moeller, M., Burger, M.: Nonlinear spectral analysis via one-homogeneous functionals: Overview and future prospects. J. Math. Imaging Vision **56**(2), 300–319 (2016)
Goldstein, T., Li, M., Yuan, X.: Adaptive primal-dual splitting methods for statistical learning and image processing. In: Advances in Neural Information Processing Systems, pp. 2089–2097 (2015)
Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press (2016)
Gordon, R., Bender, R., Herman, G.T.: Algebraic reconstruction techniques (ART) for three-dimensional electron microscopy and X-ray photography. J. Theor. Biol. **29**(3), 471–481 (1970)
Gower, R.M., Richtárik, P.: Randomized iterative methods for linear systems. SIAM J. Matrix Anal. Appl. **36**(4), 1660–1690 (2015)
Grimm, V., McLachlan, R.I., McLaren, D.I., Quispel, G.R.W., Schönlieb, C.-B.: Discrete gradient methods for solving variational image regularisation models. J. Phys. A **50**(29), 295201 (2017)
Gutman, D.H., Peña, J.F.: A unified framework for bregman proximal methods: subgradient, gradient, and accelerated gradient schemes. arXiv preprint arXiv:1812.10198 (2018)
Hairer, E., Lubich, C., Wanner, G.: Geometric Numerical Integration: Structure-Preserving Algorithms for Ordinary Differential Equations, Vol. 31, 2nd edn. Springer Science & Business Media, Berlin (2006)
Hanzely, F., Richtarik, P., Xiao, L.: Accelerated bregman proximal gradient methods for relatively smooth convex optimization. arXiv preprint arXiv:1808.03045 (2018)
Hellinger, E.: Neue begründung der theorie quadratischer formen von unendlichvielen veränderlichen. Journal für die reine und angewandte Mathematik (Crelles Journal) **1909**(136), 210–271 (1909)
Hiriart-Urruty, J.-B., Lemaréchal, C.: Convex analysis and minimization algorithms I: Fundamentals, volume 305 of Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathemati- cal Sciences], 2nd edn. Springer, Berlin (1993)
Hohage, T., Homann, C.: A generalization of the chambolle-pock algorithm to banach spaces with applications to inverse problems. arXiv preprint arXiv:1412.0126 (2014)
Hsieh, Y.-P., Kavis, A., Rolland, P., Cevher, V.: Mirrored Langevin dynamics. In: Advances in Neural Information Processing Systems, pp. 2878–2887 (2018)
Hua, X., Yamashita, N.: Block coordinate proximal gradient methods with variable bregman functions for nonsmooth separable optimization. Math. Program. **160**(1–2), 1–32 (2016)
Huang, B., Ma, S., Goldfarb, D.: Accelerated linearized bregman method. J. Sci. Comput. **54**(2–3), 428–453 (2013)
Itakura, F.: Analysis synthesis telephony based on the maximum likelihood method. In: The 6th International Congress on Acoustics, 1968, pp. 280–292 (1968)
Itoh, T., Abe, K.: Hamiltonian-conserving discrete canonical equations based on variational difference quotients. J. Comput. Phys. **76**(1), 85–102 (1988)
Juditsky, A., Nemirovski, A., et al.: First order methods for nonsmooth convex large-scale optimization, I: General purpose methods. Optim. Mach. Learn. 121–148 (2011). https://doi.org/10.7551/mitpress/8996.003.0007

Kaczmarz, M.S.: Angenäherte Auflösung von Systemen linearer Gleichungen. Bulletin International de l'Académie Polonaise des Sciences et des Lettres. Classe des Sciences Mathématiques et Naturelles. Série A, Sciences Mathématiques **35**, 355–357 (1937)
Kiwiel, K.C.: Free-steering relaxation methods for problems with strictly convex costs and linear constraints. Math. Oper. Res. **22**(2), 326–349 (1997)
Kiwiel, K.C.: Proximal minimization methods with generalized bregman functions. SIAM J. Control. Optim. **35**(4), 1142–1168 (1997)
Kobler, E., Klatzer, T., Hammernik, K., Pock, T.: Variational networks: connecting variational methods and deep learning. In: German Conference on Pattern Recognition, pp. 281–293. Springer (2017)
Krichene, W., Bayen, A., Bartlett, P.L.: Accelerated mirror descent in continuous and discrete time. In: Advances in Neural Information Processing Systems, pp. 2845–2853 (2015)
Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
LeCun, Y., Cortes, C., Burges, C.J.C.: The mnist database of handwritten digits (1998). http://yann.lecun.com/exdb/mnist 10:34 (1998)
Lee, K.-C., Ho, J., Kriegman, D.J.: Acquiring linear subspaces for face recognition under variable lighting. IEEE Trans. Pattern Anal. Mach. Intell. **27**(5), 684–698 (2005)
Li, G., Pong, T.K.: Global convergence of splitting methods for nonconvex composite optimization. SIAM J. Optim. **25**(4), 2434–2460 (2015)
Li, H., Schwab, J., Antholzer, S., Haltmeier, M.: Nett: Solving inverse problems with deep neural networks. Inverse Prob. **36**, 065005 (2020)
Lions, P.-L., Mercier, B.: Splitting algorithms for the sum of two nonlinear operators. SIAM J. Numer. Anal. **16**(6), 964–979 (1979)
Lorenz, D.A., Schöpfer, F., Wenger, S.: The linearized Bregman method via split feasibility problems: Analysis and generalizations. SIAM J. Imag. Sci. **7**(2), 1237–1262 (2014)
Lorenz, D.A., Wenger, S., Schöpfer, F., Magnor, M.: A sparse Kaczmarz solver and a linearized Bregman method for online compressed sensing. arXiv e-prints (2014)
Prasanta, P.C.: On the generalized distance in statistics. National Institute of Science of India (1936)
Matet, S., Rosasco, L., Villa, S., Vu, B.L.: Don't relax: Early stopping for convex regularization. arXiv preprint arXiv:1707.05422 (2017)
McLachlan, R.I., Quispel, G.R.W.: Six lectures on the geometric integration of ODEs, pp. 155–210. London Mathematical Society Lecture Note Series. Cambridge University Press, Cambridge (2001)
McLachlan, R.I., Quispel, G.R.W., Robidoux, N.: Geometric integration using discrete gradients. Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci. **357**(1754), 1021–1045 (1999)
Miyatake, Y., Sogabe, T., Zhang, S.-L.: On the equivalence between SOR-type methods for linear systems and the discrete gradient methods for gradient systems. J. Comput. Appl. Math. **342**, 58–69 (2018)
Moeller, M., Benning, M., Schönlieb, C., Cremers, D.: Variational depth from focus reconstruction. IEEE Trans. Image Process. **24**(12), 5369–5378 (2015)
Möllenhoff, T., Strekalovskiy, E., Moeller, M., Cremers, D.: The primal-dual hybrid gradient method for semiconvex splittings. SIAM J. Imag. Sci. **8**(2), 827–857 (2015)
Moreau, J.-J.: Proximité et dualité dans un espace hilbertien. Bulletin de la Société mathématique de France **93**, 273–299 (1965)
Morozov, V.A.: Regularization of incorrectly posed problems and the choice of regularization parameter. USSR Comput. Math. Math. Phys. **6**(1), 242–251 (1966)
Nair, V., Hinton, G.E.: Rectified linear units improve restricted boltzmann machines. In: Proceedings of the 27th International Conference on Machine Learning (ICML-10), pp. 807–814 (2010)
Nemirovski, A., Juditsky, A., Lan, G., Shapiro, A.: Robust stochastic approximation approach to stochastic programming. SIAM J. Optim. **19**(4), 1574–1609 (2009)
Nemirovsky, A.S., Yudin, D.B.: Problem complexity and method efficiency in optimization (1983)

Nesterov, Y.: A method for unconstrained convex minimization problem with the rate of convergence $\mathcal{O}(1/k^2)$. In: Doklady AN USSR, Vol. 269, pp. 543–547 (1983)
Nesterov, Y.: Primal-dual subgradient methods for convex problems. Math. Program. **120**(1), 221–259 (2009)
Neubauer, A.: On nesterov acceleration for landweber iteration of linear ill-posed problems. J. Inverse Ill-posed Prob. **25**(3), 381–390 (2017)
Nielsen, F., Boltz, S.: The burbea-rao and bhattacharyya centroids. IEEE Trans. Inf. Theory **57**(8), 5455–5466 (2011)
Ochs, P., Chen, Y., Brox, T., Pock, T.: iPiano: Inertial proximal algorithm for nonconvex optimization. SIAM J. Imag. Sci. **7**(2), 1388–1419 (2014)
Ochs, P., Ranftl, R., Brox, T., Pock, T.: Bilevel optimization with nonsmooth lower level problems. In: International Conference on Scale Space and Variational Methods in Computer Vision, pp. 654–665. Springer (2015)
Osher, S., Burger, M., Goldfarb, D., Xu, J., Yin, W.: An iterative regularization method for total variation-based image restoration. Multiscale Model. Simul. **4**(2), 460–489 (2005)
Oswald, P., Zhou, W.: Convergence analysis for Kaczmarz-type methods in a Hilbert space framework. Linear Algebra Appl. **478**, 131–161 (2015)
Ouyang, H., He, N., Tran, L., Gray, A.: Stochastic alternating direction method of multipliers. In: International Conference on Machine Learning, pp. 80–88 (2013)
Parikh, N., Boyd, S., et al.: Proximal algorithms. Found. Trends® Optim. **1**(3), 127–239 (2014)
Perona, P., Malik, J.: Scale-space and edge detection using anisotropic diffusion. IEEE Trans. Pattern Anal. Mach. Intell. **12**(7), 629–639 (1990)
Pock, T., Cremers, D., Bischof, H., Chambolle, A.: An algorithm for minimizing the mumford-shah functional. In: 2009 IEEE 12th International Conference on Computer Vision, pp. 1133–1140. IEEE (2009)
Resmerita, E., Scherzer, O.: Error estimates for non-quadratic regularization and the relation to enhancement. Inverse Prob. **22**(3), 801 (2006)
Riis, E.S., Ehrhardt, M.J., Quispel, G.R.W., Schönlieb, C.-B.: A geometric integration approach to nonsmooth, nonconvex optimisation. Foundations of Computational Mathematics (FOCM). ArXiv e-prints (2018)
Ringholm, T., Lazić, J., Schönlieb, C.-B.: Variational image regularization with Euler's elastica using a discrete gradient scheme. SIAM J. Imag. Sci. **11**(4), 2665–2691 (2018)
Robbins, H., Monro, S.: A stochastic approximation method. Ann. Math. Stat. **22**, 400–407 (1951)
Scherzer, O., Groetsch, C.: Inverse scale space theory for inverse problems. In: International Conference on Scale-Space Theories in Computer Vision, pp. 317–325. Springer (2001)
Marie Schmidt, F., Benning, M., Schönlieb, C.-B.: Inverse scale space decomposition. Inverse Prob. **34**(4), 179–212 (2018)
Schmidt, M., Le Roux, N., Bach, F.: Minimizing finite sums with the stochastic average gradient. Math. Program. **162**(1–2), 83–112 (2017)
Schöpfer, F., Lorenz, D.A.: Linear convergence of the randomized sparse Kaczmarz method. Math. Program. **173**(1), 509–536 (2019)
Schuster, T., Kaltenbacher, B., Hofmann, B., Kazimierski, K.S.: Regularization methods in Banach spaces, Vol. 10. Walter de Gruyter (2012)
Su, W., Boyd, S., Candes, E.J.: A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights. J. Mach. Learn. Res. **17**(153), 1–43 (2016)
Teboulle, M.: Entropic proximal mappings with applications to nonlinear programming. Math. Oper. Res. **17**(3), 670–690 (1992)
Teboulle, M.: A simplified view of first order methods for optimization. Math. Program. **170**(1), 67–96 (2018)
Teboulle, M., Chen, G.: Convergence analysis of a proximal-like minimization algorithm using bregman function. SIAM J. Optim. **3**(3), 538–543 (1993)
Valkonen, T.: A primal–dual hybrid gradient method for nonlinear operators with applications to mri. Inverse Prob. **30**(5), 055012 (2014)

Wang, H., Banerjee, A.: Bregman alternating direction method of multipliers. In: Advances in Neural Information Processing Systems, pp. 2816–2824 (2014)
Widrow, B., Hoff, M.E.: Adaptive switching circuits. Technical report, Stanford Univ Ca Stanford Electronics Labs (1960)
Wright, S.J.: Coordinate descent algorithms. Math. Program. **1**(151), 3–34 (2015)
Xiao, L.: Dual averaging methods for regularized stochastic learning and online optimization. J. Mach. Learn. Res. **11**, 2543–2596 (2010)
Yin, W.: Analysis and generalizations of the linearized bregman method. SIAM J. Imag. Sci. **3**(4), 856–877 (2010)
Yin, W., Osher, S., Goldfarb, D., Darbon, J.: Bregman iterative algorithms for \ell_1-minimization with applications to compressed sensing. SIAM J. Imag. Sci. **1**(1), 143–168 (2008)
Yosida, K.: Functional Analysis. Springer (1964)
Young, D.M.: Iterative Solution of Large Linear Systems. Computer Science and Applied Mathematics, 1st edn. Academic Press, Inc., Orlando (1971)
Zhang, H., Dai, Y.-H., Guo, L., Peng, W.: Proximal-like incremental aggregated gradient method with linear convergence under bregman distance growth conditions. arXiv preprint arXiv:1711.01136 (2017)
Zhang, K.S., Peyré, G., Fadili, J., Pereyra, M.: Wasserstein control of mirror Langevin Monte Carlo. arXiv e-prints (2020)
Zhang, X., Burger, M., Osher, S.: A unified primal-dual algorithm framework based on bregman iteration. J. Sci. Comput. **46**(1), 20–46 (2011)
Zhou, Z., Mertikopoulos, P., Bambos, N., Boyd, S., Glynn, P.W.: Stochastic mirror descent in variationally coherent optimization problems. In: Advances in Neural Information Processing Systems, pp. 7040–7049 (2017)
Zhu, M., Chan, T.: An efficient primal-dual hybrid gradient algorithm for total variation image restoration. UCLA CAM Report 34 (2008)