B-SMART: Bregman-Based First-Order Algorithms for Non-Negative Compressed Sensing Problems

Stefania Petra*, Christoph Schnörr, Florian Becker, and Frank Lenzen

IPA & HCI, Heidelberg University,
Speyerer Str. 6, 69115 Heidelberg, Germany
{petra,schnoerr,becker}@math.uni-heidelberg.de
frank.lenzen@iwr.uni-heidelberg.de
http://ipa.iwr.uni-heidelberg.de
http://hci.iwr.uni-heidelberg.de

Abstract. We introduce and study Bregman functions as objectives for non-negative sparse compressed sensing problems together with a related first-order iterative scheme employing non-quadratic proximal terms. This scheme yields closed-form multiplicative updates and handles constraints implicitly. Its analysis does not rely on global Lipschitz continuity in contrast to established state-of-the-art gradient-based methods, hence it is attractive for dealing with very large systems. Convergence and a $O(k^{-1})$ rate are proved. We also introduce an iterative two-step extension of the update scheme that accelerates convergence. Comparative numerical experiments for non-negativity and box constraints provide evidence for a $O(k^{-2})$ rate and reveal competitive and also superior performance.

Key words: multiplicative algebraic reconstruction, compressed sensing, underdetermined systems of nonnegative linear equations, convergence rates, limited angle tomography

1 Introduction

Overview. Since the advent of Compressed Sensing [8, 12] it is well-known that the sparsest solution of an underdetermined system of equations can be found via ℓ_1-minimization under adequate conditions. In many interesting applications the vector x^* to be recovered is nonnegative or even binary. Recent results [24, 17, 13, 21] show that under appropriate conditions, a sparse nonnegative (or binary) solution is also the *unique* solution of

$$Ax = b, \quad x \in X, \tag{1}$$

with $X = \mathbb{R}^n_+$ or $X = [0,1]^n$, and thus recovery reduces to a simpler feasibility problem. As a consequence, this may lead to alternatives superior to ℓ_1-minimization since *any* objective function subject to the constraints (1) can recover the sparse solution. On the other hand, (1) becomes infeasible when noise is present, and we have to allow for a distance of Ax^* to b.

In this paper we suggest and study the approach

$$x^* = \operatorname{argmin}_{x \in X} f(x), \qquad f(x) := B_\phi(Ax, b), \tag{2}$$

with B_ϕ an appropriate Bregman distance induced by ϕ. In the case of the Euclidean distance $B_\phi(x,y) = \frac{1}{2}\|x-y\|_2^2$ it is shown in [22] that recovery of nonnegative sparse solutions via nonnegative least-squares is stable and outperforms ℓ_1-regularization when combined with thresholding.

* gratefully acknowledges support by the State Ministry of Baden-Württemberg for Sciences, Research and the Arts.

Other choices for B_ϕ can be more adequate, however, if the noise is non-Gaussian, like e.g. Poisson noise in tomographic applications, or when the data b and the sensor matrix A are nonnegative.

For particular sparse (nonnegative) images and tomographic projection matrices A the *Simultaneous Multiplicative Algebraic Reconstruction Technique (SMART)* recently proved to be quite efficient by returning meaningful solutions after few iterations [1]. It applies only to the specific but important case of systems with nonnegative b and A. SMART has been invented and re-invented several times in the field of medical imaging. Convergence was proved in [7]. For consistent projection equations (1), it returns the feasible point in $\{Ax = b, x \geq 0\}$ that minimizes the cross-entropy distance $KL(x, x^0)$ to the initial vector x^0. When all entries of x^0 are all equal SMART converges to the maximizer of the Shannon entropy.

In a nutshell, past studies showed that SMART:

1. is adequate for ill-conditioned problems and huge problem sizes,
2. converges provably,
3. performs at each iteration only a single multiplication with A and $A^\top$, and
4. returns meaningful solution after few iterations.

Contribution and Organization. Motivated by the specific case of SMART (section 2.1), we introduce in section 2.2 an iterative scheme for the general case (2) based on a linearized objective and a related Bregman-based proximal term, that enables closed-form multiplicative updates and handles the constraints implicitly. We prove convergence and the convergence rate $O(k^{-1})$ in section 2.3.

Our approach may be understood as a blend of (i) optimal gradient-based schemes based on a linearized objective and upper bound surrogates through quadratic proximation, and (ii) fully nonlinear Bregman-based proximal iterations studied in [14]. While each step of the latter scheme is as costly as the original objective, the former schemes depend on the Lipschitz constant of the gradient of the objective that can be very large in large-scale nonnegative problems like 3D algebraic tomography. Our approach and the analyis do not require global Lipschitz continuity.

In section 2.4 we specifically consider Bregman distances induced by the Shannon entropy and by the Fermi-Dirac entropy and the corresponding multiplicative updates, to deal with nonnegativity or box constraints. Connections of the resulting objectives to nonnegative least-squares and ℓ_1-regression, that substantiate our approach more formally, are outlined in section 2.5.

While proving a $O(k^{-2})$ convergence rate is beyond the scope of the present conference contribution, we suggest two algorithmic extensions called F(AST)-SMART in section 3, aking to a Bregman-based versions of established first-order optimal schemes [20, 5]. Competitive numerical experiments discussed in section 4 illustrate the discussion above and support our claims.

Notation. We set $[n] = \{1, \dots, n\}$ for $n \in \mathbb{N}$. $\langle \cdot, \cdot \rangle$ denotes the Euclidean inner product and $\| \cdot \| = \| \cdot \|_2 = \langle \cdot, \cdot \rangle^{1/2}$ the corresponding norm. $\mathbb{1} = (1, \dots, 1)^\top$, that is $\|x\|_1 = \langle \mathbb{1}, x \rangle$ for $x \in \mathbb{R}^n_+$. Vectors are enumerated with superscripts x^i, and vector and matrix components with subscripts x_i, A_{ij}, while matrix rows and columns we denote by $A_{i,\bullet}$ and $A_{\bullet,j}$ respectively. Vector inequalities $x \geq y$ and $\log x, \exp x$ etc., are understood component-wise. By x_+ we denote $\mathbb{1}^\top x$. $\Delta_n = \{x \geq 0 \colon \|x\|_1 = 1\} \subset \mathbb{R}^n_+$ denotes the probability simplex. $KL(x, y)$ denotes the Kullback-Leibler distance of two nonnegative vectors, see Appendix.

2 B-SMART

2.1 Motivation: The SMART Iteration

It is well known [7] that the *Simultaneous Multiplicative Algebraic Reconstruction Technique (SMART)* minimizes $f(x) = KL(Ax, b)$ over the positive orthant, provided that $A \geq 0$, $(A_{\bullet,j})_+ >$

$0, j \in [n]$ and $b > 0$. This corresponds to (2) with φ being the negative entropy (21). For a positive iterate $x^k \in \mathbb{R}^n_{++}$ the SMART iteration reads

$$x_j^{k+1} = x_j^k \prod_{i=1}^m \left(\frac{b_i}{\langle A_{i,\bullet}, x^k \rangle} \right)^{t_k A_{ij}}, \quad j \in [n]. \tag{3}$$

Here t_k is a relaxation parameter, with $t_k \leq \min_j \{(A_{\bullet,j})_+\}$. We observe that algorithm (3) employs at each step the minimization of the linearized objective f plus a "prox"-like term of the form

$$x^{k+1} = \operatorname{argmin}_{x \in \mathbb{R}^n_+} f(x^k) + \nabla f(x^k)^\top (x - x^k) + \frac{1}{t_k} KL(x, x^k), \tag{4}$$

with arbitrary starting vector $x^0 > 0$. This implies

$$\log(x^{k+1}) = \log x^k - t_k A^\top (\log Ax^k - \log b), \tag{5}$$

since for every $x^k > 0$, $x^{k+1} > 0$ holds as well. This is exactly the SMART iteration with relaxation parameter t_k.

Remark 1. We note that the above algorithm (4) is closely related to the gradient descent method

$$x^{k+1} = \operatorname{argmin}_x f(x^k) + \nabla f(x^k)^\top (x - x^k) + \frac{1}{2t_k} \|x - x^k\|^2, \tag{6}$$

better known as $x^{k+1} = x^k - t_k \nabla f(x^k)$. For convex LC^1 functions there exist precise bounds for the value of t_k depending on the Lipschitz constant of the gradient of f. Moreover, convergence rates are well understood and optimal gradient methods have been established [18, 5, 19, 23]. Our objective function f however is only locally Lipschitz-continuous, due to differentiability, and non-differentiable on the boundary of $\mathbb{R}^n_+$, where sparse solutions occur.

2.2 A Nonlinear Projected Gradient Method

In this section we derive convergence rates for the iteration (4) by considering a general minimization scheme for problems of the form (2).

Let $\varphi : X \to \mathbb{R}$ and $\phi : Y \to \mathbb{R}$ be convex and continuously differentiable on $\operatorname{int}(X)$ and $\operatorname{int}(Y)$ respectively, with $A(X) \subset Y$. Further define the distance-like functions $B_\varphi : X \times \operatorname{int}(X) \to \mathbb{R}$ and $B_\phi : Y \times \operatorname{int}(Y) \to \mathbb{R}$ by

$$B_\varphi(x, y) = \varphi(x) - \varphi(y) - \langle x - y, \nabla \varphi(y) \rangle \tag{7}$$

and

$$B_\phi(x, y) = \phi(x) - \phi(y) - \langle x - y, \nabla \phi(y) \rangle. \tag{8}$$

We assume $A(X) \subset Y$ and $b \in \operatorname{int}(Y)$, and define $f : X \to \mathbb{R}$ by

$$f(x) = B_\phi(Ax, b). \tag{9}$$

Choosing an appropriate constant $c > 0$, we apply with $\nabla_x B_\phi(Ax, b) = A^\top(\nabla \phi(Ax) - \nabla \phi(b))$ the iteration

$$x^{k+1} = \operatorname{argmin}_{x \in X} f(x^k) + \langle \nabla f(x^k), x - x^k \rangle + \frac{c}{t_k} B_\varphi(x, x^k) \tag{10}$$

$$= \operatorname{argmin}_{x \in X} f(x^k) + \langle \nabla \phi(Ax^k) - \nabla \phi(b), Ax - Ax^k \rangle + \frac{c}{t_k} B_\varphi(x, x^k). \tag{11}$$

We will see that under an appropriate assumption the r.h.s. of (10) is an upper bound of f.

2.3 Convergence and Convergence Rates

Iteration (10) is exactly the nonlinear projected gradient method from [4], except for the fact that due to the particular form of the objective function, only relaxed conditions of f are required. In fact, we can replace the Lipschitz-condition in [4] by Assumption A, part (b), below.
Assumption A:

(a) X is a closed and convex set with nonempty interior;
(b) We have $B_\phi(Ax, Ay) \leq cB_\varphi(x, y)$ for all $x, y \in X$;
(c) The set of optimal solutions $X^* := \operatorname{argmin}_{x\in X} f(x)$ is nonempty.

The following results will turn out to be useful in the sequel.

Lemma 1 ([10, Lem 3.1]). *Let $S \subset \mathbb{R}^n$ be an open set with closure $\overline{S}$, and let $\psi : \overline{S} \to R$ be continuously differentiable on S. Then for any three points $a, b \in S$ and $c \in \overline{S}$ the following identity holds*

$$B_\psi(c, a) + B_\psi(a, b) - B_\psi(c, b) = \langle \nabla\psi(b) - \nabla\psi(a), c - a\rangle.$$

Theorem 1 ([3, Thm. 3.12]). *Suppose φ is closed proper convex and differentiable on* $\operatorname{int}(\operatorname{dom}\varphi)$*, X is closed convex with $X \cap \operatorname{int}(\operatorname{dom}\varphi) \neq \emptyset$, and $y \in \operatorname{int}(\operatorname{dom}\varphi)$. If φ is Legendre, then the Bregman projection $\overline{x}$ of y is unique and contained in* $\operatorname{int}(\operatorname{dom}\varphi)$*,*

$$\operatorname*{argmin}_{x\in X\cap\operatorname{dom}\varphi} B_\varphi(x, y) = \{\overline{x}\}, \qquad \overline{x} \in \operatorname{int}(\operatorname{dom}\varphi). \tag{12}$$

Remark 2. It is easy to see that assertion (12) also holds for the case

$$\operatorname*{argmin}_{x\in X\cap\operatorname{dom}\varphi} \big\{B_\varphi(x, y) + \langle l, x\rangle\big\} = \{z\}, \qquad z \in \operatorname{int}(\operatorname{dom}\varphi), \tag{13}$$

with $l \in \mathbb{R}^n$ arbitrary and $\|l\| \leq \infty$.

Our main result is stated next.

Theorem 2. *Under* **Assumption A** *above, for the sequence $\{x_k\}_{k\leq\kappa}$ generated by* (10) *with starting point $x^0 \in \operatorname{int}(X)$ and $t_k = t \leq 1$, one has:*

(a) Iteration (10) *is well defined.*
(b) For every κ,

$$\min_{0\leq k\leq\kappa} f(x^k) - \min_X f(x) \leq \frac{cB_\varphi(x^*, x^0)}{t\kappa}\ . \tag{14}$$

(c) The sequence $\{f(x_k)\}_{k\leq\kappa}$ is decreasing. In particular, the method converges.

Proof. Statement (a) follows by Remark 2.
(b) Let x^* be the optimal solution. The optimality conditions for (10) imply

$$\langle x - x^{k+1}, t_k\nabla f(x^k) + c\left(\nabla\varphi(x^{k+1}) - \nabla\varphi(x^k)\right)\rangle \geq 0, \quad x \in X.$$

In particular, for $x = x^*$ we get

$$\langle x^* - x^{k+1}, c\left(\nabla\varphi(x^k) - \nabla\varphi(x^{k+1})\right) - t_k\nabla f(x^k)\rangle \leq 0, \quad x \in X. \tag{15}$$

Since f is convex

$$0 \leq t_k(f(x^k) - f(x^*)) \;\leq\; t_k\langle x^k - x^*, \nabla f(x^k)\rangle \tag{16}$$
$$= \langle x^* - x^{k+1}, c\left(\nabla\varphi(x^k) - \nabla\varphi(x^{k+1})\right) - t_k\nabla f(x^k)\rangle \tag{17}$$
$$+ c\langle x^* - x^{k+1}, \nabla\varphi(x^{k+1}) - \nabla\varphi(x^k)\rangle + \langle x^k - x^{k+1}, t_k\nabla f(x^k)\rangle \tag{18}$$
$$:= s_1 + cs_2 + s_3. \tag{19}$$

By equation (15) $s_1 \leq 0$ holds, and by Lemma 1 we have

$$s_2 := \langle x^* - x^{k+1}, \nabla\varphi(x^{k+1}) - \nabla\varphi(x^k)\rangle = B_\varphi(x^*, x^k) - B_\varphi(x^*, x^{k+1}) - B_\varphi(x^{k+1}, x^k).$$

Furthermore

$$s_3 = \langle x^k - x^{k+1}, t_k\nabla f(x^k)\rangle \;=\; t_k\langle\nabla\phi(Ax^k) - \nabla\phi(b), Ax^k - Ax^{k+1}\rangle$$
$$\overset{\text{Lem. 1}}{=} t_k(B_\phi(Ax^{k+1}, Ax^k) + B_\phi(Ax^k, b) - B_\phi(Ax^{k+1}, b)).$$

Summarizing

$$t_k(f(x^k) - f(x^*)) \leq cB_\varphi(x^*, x^k) - cB_\varphi(x^*, x^{k+1})$$
$$+ \underbrace{t_kB_\phi(Ax^{k+1}, Ax^k) - cB_\varphi(x^{k+1}, x^k)}_{\leq 0, \text{Ass.}(b)} + \underbrace{t_kB_\phi(Ax^k, b)}_{=t_kf(x^k)} - t_kB_\phi(Ax^{k+1}, b)) \;,$$

gives

$$t_k(f(x^{k+1}) - f(x^*)) \leq cB_\varphi(x^*, x^k) - cB_\varphi(x^*, x^{k+1}).$$

Summing over k yields

$$\min_{0\leq k\leq\kappa} f(x^{k+1}) - f(x^*) \leq \frac{cB_\varphi(x^*, x^0) - cB_\varphi(x^*, x^{\kappa+1})}{t(\kappa+1)} \leq \frac{cB_\varphi(x^*, x^0)}{t(\kappa+1)}.$$

(c) By Lemma 1 we have

$$\langle\nabla\phi(Ax^k) - \nabla\phi(b), Ax - Ax^k\rangle = B_\phi(Ax, b) - B_\phi(Ax, Ax^k) - B_\phi(Ax^k, b)$$
$$= f(x) - f(x^k) - B_\phi(Ax, Ax^k) \;.$$

Thus

$$x^{k+1} = \text{argmin}_{x\in X} f(x) + \underbrace{\frac{c}{t_k}B_\varphi(x, x^k) - B_\phi(Ax, Ax^k)}_{:=f_k(x)}, \tag{20}$$

where $f_k(x) \geq 0$ due to Assumption A., part (b) and $f_k(x^k) = 0$. Consequently, algorithm (10) minimizes an upper bound on f, in analogy to the classical gradient method. Now,

$$f(x^{k+1}) + f_k(x^{k+1}) \leq f(x^k) + f_k(x^k) = f(x^k)$$

follows and

$$f(x^k) - f(x^{k+1}) \geq f_k(x^{k+1}) \geq 0 \;.$$

Hence, the sequence $\{f(x^k)\}_k$ is decreasing and bounded from below by 0. Statement (c) then follows by standard arguments. □

2.4 Application: Multiplicative Updates

It is well-known that multiplicative updates as e.g. employed by the exponential gradient method [4, 15], typically lead to faster convergence if the solution x^* of the optimization problem is sparse. As discussed in Section 2.1 the choice

$$\varphi_1(x) = \langle x, \log x \rangle, \quad x \in \mathbb{R}^n_+ \tag{21}$$

and $\phi_1(x) = \langle x, \log(x) \rangle, x \in \mathbb{R}^m_+$, leads to the update rule (3) of SMART, since $B_{\varphi_1}(x, y) = KL(x, y)$ and $f(x) = B_{\phi_1}(Ax, b) = KL(Ax, y)$. For this particular choice we obtain $c_1 = 1$, for matrices A with columns that sum up to one, compare Appendix, Prop. 2.

To include an upper bound on feasible points x, often known in applications (e.g. $x \in [0, 1]^n$), we additionally consider the generalization of the Fermi-Dirac entropy

$$\varphi_2(x) = \langle x - l, \log(x - l) \rangle + \langle u - x, \log(u - x) \rangle, \quad x \in X = [l, u],\ l < u. \tag{22}$$

A simple computation shows $B_{\varphi_2}(x, y) = KL(x - l, y - l) + KL(u - x, u - y)$. With B_{φ_2} and $f(x) = B_{\phi_2}(x, y) = KL(Ax, b)$, we obtain again $c_2 = 1$, compare Appendix, Prop. 3. This choice leads to the following algorithm that we call **bounded-SMART**

$$\frac{(x^{k+1} - l)_j}{(u - x^{k+1})_j} = \frac{(x^k - l)_j}{(u - x^k)_j} \prod_{i=1}^{m} \left(\frac{b_i}{\langle A_{i,\bullet}, x^k \rangle} \right)^{t_k A_{ij}}. \tag{23}$$

Proposition 1 below provides convergence rates for the multiplicative updates (3) and (23). The proof of the following preparatory Lemma is given in the Appendix.

Lemma 2. *For a minimizer $x^* \in X^*$ and some arbitrary starting point $x^0 \in \operatorname{int} X$ with $x^0_{\min} := \min_{i \in [n]} x^0_i$, we have*

$$B_{\varphi_i}(x^*, x^0) \ \leq \ \begin{cases} R(\log R - \log x^0_{\min} - 1) + \|x^0\|_1, & i = 1,\ X = \mathbb{R}^n_+ \\ \log n, & i = 1,\ X = \Delta_n \\ 2R(R - \log x^0_{\min}), & i = 2,\ X = [l, u] \end{cases} \tag{24}$$

for some sufficiently large $R > 0$ such that $\|x^*\|_1 \leq R$. In the case of $\varphi = \varphi_2$ and $X = [l, u]$, we have $R = \|u - l\|_1$.

Proposition 1. *Algorithms* (3) *and* (23) *converge for $t_k = 1$ with rate*

$$\min_{0 \leq k \leq \kappa} f(x^k) - f(x^*) \leq \frac{c(R)}{\kappa},$$

with $c(R)$ given by (24)*, for any $x^0 \in \operatorname{int} X$ and all $\kappa \geq 0$.*

Proof. Propositions 2 and 3 in the Appendix establish $c = 1$ in both cases, in the context of Assumption A, part (b). Parameter t_k can be fixed to 1. Together with parts (a) and (c), Theorem 2 then yields the assertion. □

2.5 B-SMART: An Alternative to Nonnegative Least Squares and ℓ_1–Regression

We briefly relate our approach to the more established objective functions

$$\min_{x \geq 0} \|Ax - b\|^2 \qquad \text{and} \qquad \min_{x \geq 0} \|Ax - b\|_1. \tag{25}$$

The nonnegative least-squares approach on the l.h.s. corresponds to the special case $\varphi_3(x) = \phi_3(x) = \frac{1}{2}\|x\|^2$, cf. (7), (8). Iteration (10) reads (up to a constant)

$$x^{k+1} = \operatorname{argmin}_{x\in X} \langle Ax^k - b, A(x - x^k)\rangle + c\|x - x^k\|^2, \tag{26}$$

with $c = \|A\|_2^2 = \lambda_{\max}(A^\top A)$ for Assumption A, part (b), to hold. While directly tackling the normal equations corresponding to the l.h.s. of (25) is known to be ill-conditioned, the surrogate (right-most term) in (26) provides only a poor approximation of the objective. The large weight c entails only small steps, in addition to the need to take non-smooth projections onto X into account. By contrast, $c = 1$ suffices for both cases (21) and (22), and the feasible set X is taken implicitly into account by closed-form iterative updates.

Adopting a probabilistic viewpoint, **non-negative least-squares** may be critized because the residuals $(Ax - b)_i^2$ do *not* follow a Gaussian distribution. Rather than rectifying this for specific applications (e.g. by a Poisson model in connection with tomography), our approach is additionally motivated by recent results of compressed sensing for non-negative sensing matrices, corresponding to sparse expander graphs with constant column sums (cf., e.g., [6]): $\mathbb{1}^\top A = d\mathbb{1}$, $d > 0$. As a consequence, we have $\|d^{-1}Ax\|_1 = \|x\|_1 = \|b\|_1$ for consistent systems $Ax = b$, that is $x, b \in \Delta_n$, up to a common scale factor. This suggests to adopt the distance $KL(Ax, b)$ in the inconsistent case (noisy measurements b), that is more natural for comparing points in the simplex Δ_n. Applying Jensen's inequality, we then get

$$\begin{aligned} KL(Ax, b) &\le \log\Big(\sum_i \frac{(Ax)_i^2}{b_i}\Big) = \log\Big(\sum_i \frac{(Ax)_i^2}{b_i} + \underbrace{\|b\|_1 - 2\|Ax\|_1 + 1}_{=0}\Big) \\ &= \log\Big(\sum_i \frac{(Ax - b)_i^2}{b_i} + 1\Big) = \log\Big(1 + \big\langle Ax - b, \operatorname{Diag}(b)^{-1}(Ax - b)\big\rangle\Big). \end{aligned} \tag{27}$$

This relates in view of non-negative least-squares our objective to (the logarithm of) a *scaled* squared Euclidean objective, which is known as the χ^2-distance that provides a first-order expansion of the KL-distance at b [11].

The ℓ_1**-regression objective** in (25), suggested e.g. by [9], may be considered as total variation distance $d_{\mathrm{TV}}(Ax, b) = \sum_i |(Ax - b)_i|$, again from the viewpoint of discrete probability distributions. Our objective upper-bounds this distance, $\frac{1}{2}KL \ge d_{\mathrm{TV}}^2$, as shown in [16], hence minimizes the total variation as well. On the other hand, unlike the *residuals* $(Ax - b)_i$ are known to be sparse (cf. [9]) (rather than x), considering the KL distance seems more appropriate.

Summing up, there are good reasons to consider and study (2) as objective for a range of non-negative compressed sensing scenarios.

3 F-SMART: Towards an Optimal Nonlinear Proj. Gradient Method

Above we showed that SMART and its bounded version converge with rate $O(k^{-1})$. We believe that it should be possible to design an "optimal" entropic gradient method in the sense of [20] with rate $O(k^{-2})$. An elaboration is beyond the scope of the present conference contribution. We therefore confine ourselves to specifying below the algorithm and to providing empirical evidence supporting our conjecture in Section 4.

Similar to Alg. 1 in [23], we suggest the following iteration called **F(ast)-SMART 1**,

$$y^k = (1 - \theta_k)x^k + \theta_k z^k \tag{28a}$$

$$z^{k+1} = \operatorname{argmin}_{x\in X} \langle \nabla f(y^k), x - y^k\rangle + c\,\theta_k B_\varphi(x, z^k) \tag{28b}$$

$$x^{k+1} = (1 - \theta_k)x^k + \theta_k z^{k+1}, \tag{28c}$$

where $x^0 = z^0 \in \operatorname{int}(\operatorname{dom}\varphi)$ and $\theta_k \in (0,1]$ satisfies

$$\frac{1-\theta_{k+1}}{\theta_{k+1}^2} \leq \frac{1}{\theta_k^2}. \tag{29}$$

Additionally, similar to FISTA [5], we suggest the following scheme called **F(ast)-SMART 2**,

$$x^k = \operatorname{argmin}_{x\in X} B_\varphi(x, y^{k-1}) + \langle \nabla f(y^{k-1}), x - y^{k-1}\rangle \tag{30a}$$

$$v^k = \Pi_X\Big(x^{k-1} + \frac{1}{\theta_k}(x^k - x^{k-1})\Big) \tag{30b}$$

$$y^k = (1-\theta_{k+1})x^k + \theta_{k+1}v^k, \tag{30c}$$

where $x^0 = y^0 = v^0 \in \operatorname{int}(\operatorname{dom}\varphi)$ and θ_k satisfies again (29).

Numerical evidence for convergence and the rate of both F-SMART variants is provided in the next section.

4 Experiments and Discussion

In this section we illustrate the performance of BSMART (10) compared to FISTA [5]. BSMART includes the *SMART* scheme for $\varphi_1(x)$ (21) and *b(ounded)-SMART* for $\varphi_2(x)$ (22) as special cases. In the following SMART (3), FSMART1 (28) and FSMART2 (30) will minimize $f(x) = KL(Ax, b)$ over $X = R_+^n$, while b(ounded)-SMART, b(ounded)-FSMART1 and b(ounded)-FSMART2 minimizes $f(x) = KL(Ax, b)$ over $X = [0,1]^n$. Cf. the discussion of Eq. (27), FISTA will be applied to $f(x) = 0.5\|Ax-b\|^2$ and $f(x) = 0.5\langle Ax-b, \operatorname{Diag}(b)^{-1}(Ax-b)\rangle$ subject to both $X = \mathbb{R}_+^n$ and $X = [0,1]^n$. Matrix A will be scaled so that the every column sums up to one.

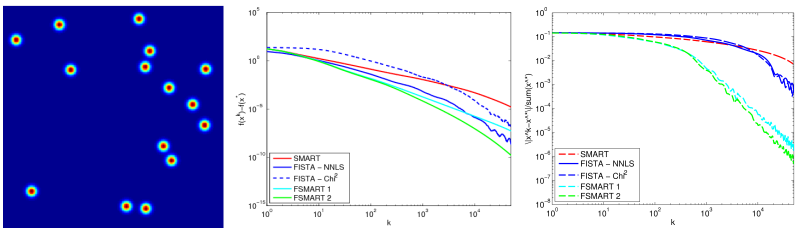

Fig. 1. The first test image consists of 15 particles at random positions (left). Comparison of function value errors $f(x^k) - f(x^*)$ for all algorithms (middle). While BSMART is competitive, the relative error decays faster for FSMART1 and FSMART1 (right).

Test Case 1: Here we consider an infeasible ill-conditioned problem inspired by a real-world application [1]. The original sparse image I, see Fig. 1 left, consists of 15 Gaussian blobs (particles) at random positions in a square. The measurement vector $b \in \mathbb{R}^{200}$ is computed by integrating the particle image exactly along 50×4 lines arranged in 4 fan beams (angles $45^o, 15^o, -15^o, -45^o$). Image I is discretized in 66×66 Gaussian basis functions positioned on a regular grid. The matrix entries A_{ij} equal the line integral of every basis function along every line, thus $A \in \mathbb{R}^{200\times4356}$ and $A \geq 0$. After scaling $\mathbb{1}^\top A = \mathbb{1}$, and $L_{\chi^2} = 1004.8$, $L_{\text{NNLS}} \approx 53.6$. We underline that *no* nonnegative solution exists which satisfies the constraints $Ax = b$. Additionally we added

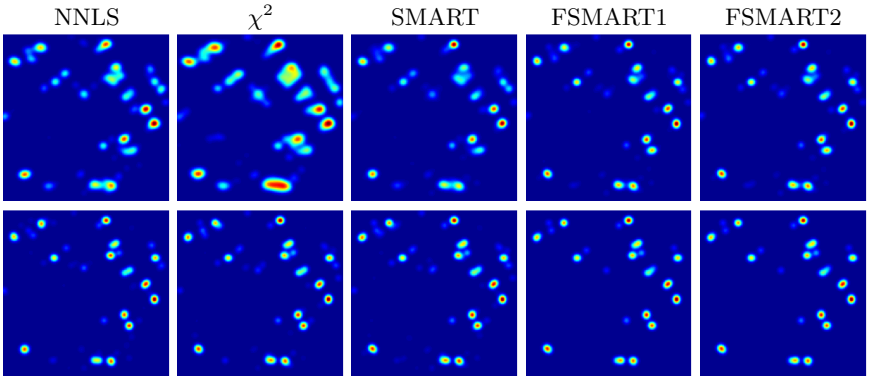

Fig. 2. Reconstructions of 15 particles at random positions at iteration 100 (top row) and at the final iteration (bottom row). The reconstruction is accurate after a (significantly) smaller number of iterations in the case of the KL objective that copes better with an ill-conditioned matrix A.

uniform (non-Gaussian) noise to b. The parameters for FISTA, FSMART1 and FSMART2 are chosen as $\theta_k = 1$, $\theta_{k+1} = 0.5(\sqrt{\theta_{k+1}^4 + 4\theta_{k+1}^2} - \theta_{k+1}^2)$ and satisfy (29), according to [23]. The function value at iteration k of all algorithms is depicted in Fig. 1. The function value for FSMART2 is lower than for FISTA, which is explained by the high values of L_{χ^2} and L_{NNLS}. The decay of $f(x^k) - f(x^*)$ for both FSMART1 and FSMART2 suggests a $O(k^{-2})$ rate, consistently with FISTA, see Fig. 1, middle. The solutions x^* for the three problems considered, $\min_{x\in\mathbb{R}^n_+} KL(Ax, b)$, $\min_{x\in\mathbb{R}^n_+} 0.5\langle Ax - b, \text{Diag}(b)^{-1}(Ax - b)\rangle$ and $\min_{x\in\mathbb{R}^n_+} 0.5\|Ax - b\|^2$, are not known, but we computed an accurate solution via an interior point solver for the KKT conditions. Iteration 100 and the final one are described in Fig. 2. *The reconstructions produced by SMART, FSMART1 and FSMART2 are of better quality even if only few iterations are performed.*

These preliminary computational results indicate that BSMART is sometimes even faster then the proven predicted theoretical rate and FSMART is a promising extension with a high potential for designing fast algorithms for nonnegative data.

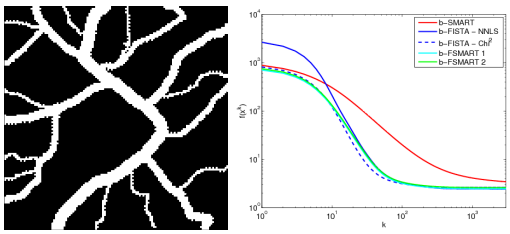

Fig. 3. Original 256×256 binary test image from [2] (left). Comparison of function values for all algorithms and $X = [0, 1]^n$ (right). Again FSMART1 and FSMART2 exhibit an $O(k^{-2})$ rate.

Test Case 2: The second 256×256 test image [2] is a vascular system containing larger and smaller vessels, see Fig. 3 (left). We consider 20 projecting directions, although the uncorrupted image binary image is determined by 18 projections and is unique in $[0, 1]^n$. Here $A \in \mathbb{R}^{7240\times 65536}$,

$L_{\chi^2} = 5.0948$, $L_{\mathrm{NNLS}} = 12.2308$ and $X = [0,1]^n$. Thus A is better conditioned than the previous one. To vector b we add again 5% uniform (nongaussian) noise. This results in an infeasible problem. Due to the low Lipschitz constant L_{χ^2} we expect a similar behavior of FISTA and FSMART, which is exactly what happens, see Fig. 3 (right) for the decrease of the function values.

Adding the additional information that the image entries are in $[0,1]^n$ leads to a fairly good reconstruction in Fig. 4 within the first iterations. This can be improved by thresholding.

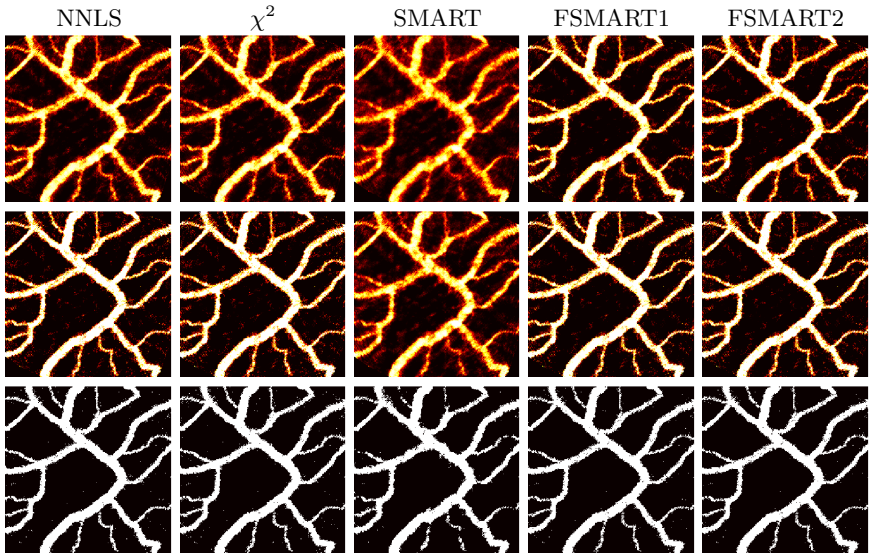

Fig. 4. Reconstructions after 50 iterations (top row) and after 100 iterations (middle row). By replacing at iteration 100 all values above a globally determined threshold with one and the others with zero, we obtain similar results for all algorithms with slightly better and faster reconstructions for FSMART and FISTA.

5 Conclusion and Further Work

This paper advocates Bregman functions as objectives for constrained nonnegative compressed sensing problems, together with a corresponding non-quadratic proximation scheme that only requires first-order gradient evaluations of the objective. The attractive properties of this approach concerning both mathematical and algorithmic aspects deserve further study. Our future work therefore will take a closer look on the pros and cons in connection with other established objectives in the field of compressed sensing, as initiated in section 2.5. Furthermore, in view of established optimal first-order methods with $O(k^{-2})$ convergence rate, we will study from a more general mathematical viewpoint surrogate objectives based on non-quadratic proximation that lead to efficient two-step iterations with multiplicative updates, with a focus on the resulting convergence rates.

Appendix

Properties of the Kullback-Leibler distance. For positive scalars a, b, define $KL(a,b) = a\log(a/b) + b - a$, $KL(0,b) = b$ and $KL(a,0) = +\infty$. The Kullback-Leibler distance can be extended to nonnegative vectors

$$KL(x,y) := \sum_{j=1}^{n} \left(x_j \log\left(\frac{x_j}{y_j}\right) + y_j - x_i \right). \tag{31}$$

It is well known that for all $x, y \geq 0$, we have $KL(x,y) \geq 0$ and $KL(x,y) = 0$ iff $x = y$. Furthermore, by Jensen's inequality, we have (see, e.g., [11, Thm. 2.7.1])

$$\sum_{i=1}^{n} x_i \log\frac{x_i}{y_i} \geq \Big(\sum_{i=1}^{n} x_i\Big) \log\frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} y_i}, \qquad \forall x, y \in \mathbb{R}^n_+. \tag{32}$$

Proposition 2. *For $A \geq 0$ with $1^\top A = 1^\top$, we have*

$$KL(Ax, Ay) \leq KL(x,y), \qquad \forall x, y \in \mathbb{R}^n_+. \tag{33}$$

Proof. We compute

$$\begin{aligned}
KL(x,y) &= \sum_{j=1}^{n} \Big(x_j \log\frac{x_j}{y_j} + y_j - x_i \Big) = \sum_{j=1}^{n} \underbrace{\sum_{i=1}^{m} A_{ij}}_{=1} \Big(x_j \log\frac{x_j}{y_j} + y_j - x_i \Big) \\
&= \sum_{i=1}^{m} \Big(\sum_{j=1}^{n} A_{ij} x_j \log\frac{A_{ij}x_j}{A_{ij}y_j} + \sum_{j=1}^{n} A_{ij} y_j - \sum_{j=1}^{n} A_{ij} x_j \Big) \\
&\overset{\text{Eq. (32)}}{\geq} \sum_{i=1}^{m} \Big[\Big(\sum_{j=1}^{n} A_{ij} x_j \Big) \log\frac{\sum_{j=1}^{n} A_{ij} x_j}{\sum_{j=1}^{n} A_{ij} y_j} + \sum_{j=1}^{n} A_{ij} y_j - \sum_{j=1}^{n} A_{ij} x_j \Big] \\
&= KL(Ax, Ay).
\end{aligned}$$

Lemma 3. *For any $x, y \geq 0$, with $x \geq t$ and $y \geq t$, we have $KL(x-t, y-t) \geq KL(x,y)$.*

Proof. Let $g(t) = KL(x-t, x-t)$. Then $g'(t) = \frac{x-t}{y-t} - 1 - \log\left(\frac{x-t}{y-t}\right) \geq 0$. Thus $g(t) \geq g(0)$.

This immediately implies

Proposition 3. *For $A \geq 0$ with $1^\top A = 1^\top$ and $x, y \in [l, u]$, we have*

$$\begin{aligned}
KL(Ax, Ay) &\leq KL(Ax - Al, Ay - Al) + KL(Au - Ax, Au - Ay) &(34)\\
&\leq KL(x - l, y - l) + KL(u - x, u - y)\ . &(35)
\end{aligned}$$

Proof of Lemma 2

Proof. In the case $X = \mathbb{R}^n_+$, we may assume $\|x\|_1 \leq R$ for some sufficiently large $R > 0$, due to Assumption A, part (c). Hence

$$\begin{aligned}
B_{\varphi_1}(x^*, x^0) &= \sum_i \big(x_i^* \log\frac{x_i^*}{x_i^0} + x_i^0 - x_i^* \big) = \|x^*\|_1 \sum_i \frac{x_i^*}{\|x^*\|_1} \log\frac{x_i^*}{x_i^0} + \sum_i (x_i^0 - x_i^*) \\
&= \|x^*\|_1 \sum_i \frac{x_i^*}{\|x^*\|_1} \big(\log\frac{x_i^*}{\|x^*\|_1} + \log\|x^*\|_1 - \log x^0_{\min} \big) + \sum_i (x_i^0 - x_i^*) \\
&\leq \|x^*\|_1 (\log\|x^*\|_1 - \log x^0_{\min} - 1) + \|x^0\|_1 \leq R(\log R - \log x^0_{\min} - 1) + \|x^0\|_1
\end{aligned}$$

In the case $X = \Delta_n$, we have $R = 1$ and may choose $x^0 = n^{-1}\mathbb{1}$. In the case $X = [l, u]$, the last two summands in (31) cancel. A similar computation then yields

$$\begin{aligned} B_{\varphi_2}(x^*, x^0) &\leq \|x^* - l\|_1(\log \|x^* - l\|_1 - \log x^0_{\min}) + \|u - x^*\|_1(\log \|u - x^*\|_1 - \log x^0_{\min}) \\ &\leq 2\|u - l\|_1(\log \|u - l\|_1 - \log x^0_{\min}). \end{aligned}$$

□

References

1. C. Atkinson and J. Soria. An efficient simultaneous reconstruction technique for tomographic particle image velocimetry. *Experiments in Fluids*, 47(4):553–568, 2009.
2. K. J. Batenburg. A network flow algorithm for reconstructing binary images from discrete x-rays. *J. Math. Imaging Vis.*, 27(2):175–191, February 2007.
3. H.H. Bauschke and J.M. Borwein. Legendre Functions and the Method of Random Bregman Projections. *J. Convex Analysis*, 4(1):27–67, 1997.
4. A. Beck and M. Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, May 2003.
5. A. Beck and M. Teboulle. A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems. *SIAM J. Img. Sci.*, 2:183–202, March 2009.
6. R. Berinde, A.C. Gilbert, P. Indyk, H. Karloff, and M.J. Strauss. Combining Geometry and Combinatorics: A Unified Approach to Sparse Signal Recovery. *CoRR*, 2008. Preprint arXiv:0804.4666.
7. C.L. Byrne. Iterative image reconstruction algorithms based on cross-entropy minimization. *IEEE Transactions on Image Processing*, pages 96–103, 1993.
8. E. Candès. Compressive sampling. In *Int. Congress of Math.*, volume 3, Madrid, Spain, 2006.
9. E. Candes, M. Rudelson, T. Tao, and R. Vershynin. Error Correcting via Linear Programming. In *46th Ann. IEEE Symp. Found. Computer Science (FOCS'05)*, pages 295–308, 2005.
10. G. Chen and M. Teboulle. Convergence analysis of a proximal-like minimization algorithm using Bregman functions. *SIAM Journal on Optimization*, 3(3):538–543, 1993.
11. T.M. Cover and J.A. Thomas. *Elements of Information Theory*. John Wiley & Sons, 1991.
12. D. Donoho. Compressed Sensing. *IEEE Trans. Information Theory*, 52:1289–1306, 2006.
13. D.L. Donoho and J. Tanner. Counting the Faces of Randomly-Projected Hypercubes and Orthants, with Applications. *Discrete & Computational Geometry*, 43(3):522–541, 2010.
14. J. Eckstein. Nonlinear Proximal Point Algorithms using Bregman Functions, with Applications to Convex Programming. *Math. Oper. Res.*, 18(1):202–226, 1993.
15. J. Kivinen and M. Warmuth. Exponentiated Gradient versus Gradient Descent for Linear Predictors. *Inform. Comput.* , 132:1–63, 1997.
16. S. Kullback. A Lower Bound for Discrimination Information in Terms of Variation. *IEEE Trans. Inf. Theory*, 13(1):126–127, 1967.
17. O.L. Mangasarian and B. Recht. Probability of Unique Integer Solution to a System of Linear Equations. *European Journal of Operational Research*, 214(1):27–30, 2011.
18. Y. Nesterov. A method of solving a convex programming problem with convergence rate O (1/k2). *Soviet Mathematics Doklady*, 27(2):372–376, 1983.
19. Y. Nesterov. Smooth minimization of non-smooth functions. *Math. Program.*, 103(1):127–152, 2005.
20. Y. E. Nesterov and A. S. Nemirovski. *Interior-Point Polynomial Algorithms in Convex Programming (Studies in Applied and Numerical Mathematics)*. Society for Industrial Mathematics, 1994.
21. S. Petra, C. Schnörr, and A. Schröder. Critical Parameter Values and Reconstruction Properties of Discrete Tomography: Application to Experimental Fluid Dynamics. *Fundamenta Informaticae*, 2013. to appear, and arXiv:1209.4316, 2012.
22. M. Slawski and M. Hein. Non-negative least squares for sparse recovery in the presence of noise. In *In Proc. SPARS*, 2011.
23. P. Tseng. On accelerated proximal gradient methods for convex-concave optimization., 2008. submitted to SIAM J. Control Optim.
24. M. Wang, W. Xu, and A. Tang. A Unique "Nonnegative" Solution to an Underdetermined System: From Vectors to Matrices. *IEEE Transactions on Signal Processing*, 59(3):1007–1016, 2011.