Linear convergence of the Randomized Sparse Kaczmarz Method

Frank Schöpfer* D

Dirk A. Lorenz[†]

October 11, 2016

Abstract

The randomized version of the Kaczmarz method for the solution of linear systems is known to converge linearly in expectation. In this work we extend this result and show that the recently proposed Randomized Sparse Kaczmarz method for recovery of sparse solutions, as well as many variants, also converges linearly in expectation. The result is achieved in the framework of split feasibility problems and their solution by randomized Bregman projections with respect to strongly convex functions. To obtain the expected convergence rates we prove extensions of error bounds for projections. The convergence result is shown to hold in more general settings involving smooth convex functions, piecewise linear-quadratic functions and also the regularized nuclear norm, which is used in the area of low rank matrix problems. Numerical experiments indicate that the Randomized Sparse Kaczmarz method provides advantages over both the non-randomized and the non-sparse Kaczmarz methods for the solution of over- and under-determined linear systems.

Keywords: randomized Kaczmarz method, linear convergence, Bregman projections, sparse solutions, split feasibility problem, error bounds **AMS classification:** 65F10, 68W20, 90C25

1 Introduction

In this paper we analyse a randomized variant of the recently proposed Sparse Kaczmarz method to recover sparse solutions of linear systems. Let $A \in \mathbb{R}^{m \times n}$ be a matrix with rows $a_i^T \in \mathbb{R}^n$ and $b \in \mathbb{R}^m$ be such that the linear system

^{*}Institut für Mathematik, Carl von Ossietzky Universität Oldenburg, 26111 Oldenburg, Germany, frank.schoepfer@uni-oldenburg.de

[†]Institute for Analysis and Algebra, TU Braunschweig, 38092 Braunschweig, Germany, d.lorenz@tu-braunschweig.de, fon +49-531-391-7423, fax +49-531-391-7414. The work of D.L. was partially supported by the National Science Foundation under Grant DMS-1127914 to the Statistical and Applied Mathematical Sciences Institute. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Ax = b is consistent. For the standard Kaczmarz method [24] one goes through the indices of the rows cyclically, and projects a given iterate onto the solution space of this row. For i = mod(k - 1, m) + 1 the method iterates

$$x_{k+1} = x_k - \frac{\langle a_i, x_k \rangle - b_i}{\|a_i\|_2^2} \cdot a_i.$$
(1)

It is known that the method converges to the minimum norm solution \hat{x} of Ax = b when it is initialized with $x_0 = 0$, but the speed of convergence is not simple to quantify, and especially, depends on the ordering of the rows, see e.g. [20]. The situation changes if one considers a randomization such that in each step one chooses a row of the system at random. In the seminal paper [41] it has been shown that a choice of row i with probability $||a_i||_2^2/||A||_F^2$ leads to a linear convergence rate in expectation,

$$\mathbb{E}\left[\|x_{k+1} - \hat{x}\|_{2}^{2}\right] \leq \left(1 - \frac{\sigma_{\min}^{2}}{\|A\|_{F}^{2}}\right) \cdot \mathbb{E}\left[\|x_{k} - \hat{x}\|_{2}^{2}\right],$$

where $||A||_F^2$ is the Frobenius norm and σ_{\min} denotes the smallest positive singular value of A. Since then similar results have been obtained for randomized Block Kaczmarz methods and systems of equalities and inequalities, see [9, 26, 31] and connections to stochastic gradient descent have been drawn [30].

In [27, 28] a variant of the Kaczmarz method has been proposed that produces sparse solutions. This *Sparse Kaczmarz method* uses two variables and reads as

$$x_{k+1}^{*} = x_{k}^{*} - \frac{\langle a_{i}, x_{k} \rangle - b_{i}}{\|a_{i}\|_{2}^{2}} \cdot a_{i}$$

$$x_{k+1} = S_{\lambda}(x_{k+1}^{*})$$
(2)

with $\lambda > 0$ and the soft shrinkage function $S_{\lambda}(x) = \max\{|x| - \lambda, 0\} \cdot \operatorname{sign}(x)$. It has been shown in [27] that the iterates x_k converge to the solution of the regularized Basis Pursuit problem,

$$\min_{x \in \mathbb{R}^n} \lambda \|x\|_1 + \frac{1}{2} \|x\|_2^2 \quad s.t. \quad Ax = b,$$
(3)

see e.g. [15, 18, 21], and also [38] for explicit values of $\lambda > 0$ that guarantee exact recovery of sparse solutions. But no convergence rate has been given. In [33] sublinear convergence rates have been obtained for the *Randomized Sparse Kaczmarz method* by identifying the iteration as a randomized coordinate gradient descent method applied to the unconstrained dual of (3), see also [32, 42]. However, linear convergence could only be obtained by smoothing the objective function in (3), which results in an iteration that is slightly different from (2), and need not solve (3). Here we will show that the Randomized Sparse Kaczmarz method in fact converges linearly in expectation without smoothing. We use the theoretical framework developed in [27], which treats the Sparse Kaczmarz method as a special case of so-called Bregman projections for split feasibility problems. Using this flexible framework we will show (sub-)linear convergence rates for a broad range of problems. Especially, linear rates are also obtained for randomized iterations of the form

$$X_{k+1}^* = X_k^* - \frac{\langle A_i, X_k \rangle - b_i}{\|A_i\|_F^2} \cdot A_i$$

$$X_{k+1} = S_\lambda(X_{k+1}^*)$$
(4)

to solve the *regularized nuclear norm* optimization problem in the area of low rank matrix problems,

$$\min_{X \in \mathbb{R}^{n_1 \times n_2}} \lambda \|X\|_* + \frac{1}{2} \|X\|_F^2 \quad s.t. \quad \langle A_i, X \rangle = b_i, \ i = 1, \dots, m,$$
(5)

where $\langle A, X \rangle = \text{trace}(A^T \cdot X)$ for two matrices $A, X \in \mathbb{R}^{n_1 \times n_2}$, and $S_{\lambda}(X)$ denotes the singular value thresholding operator, see eg. [14, 25, 34, 43].

In the next section we recall the basic properties of Bregman projections. In section 3 we prove some error bounds which are crucial for the convergence analysis of the method of randomized Bregman projections in section 4. The special case of the Randomized Sparse Kaczmarz method is treated in section 5. In the last section we report some numerical results illustrating the performance of the Sparse Kaczmarz method with and without randomization, and also its benefit for sparsity problems compared to the standard Kaczmarz method, even in the case of overdetermined systems.

2 Basic notions

We recall some well known concepts and properties of convex functions, see [37], and state basic assumption that will be used throughout the paper.

Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex. Since f is assumed to be finite everywhere, it is also continuous. By $\partial f(x)$ we denote the subdifferential of f at $x \in \mathbb{R}^n$,

$$\partial f(x) = \left\{ x^* \in \mathbb{R}^n \, | \, f(y) \ge f(x) + \langle x^* \, , \, y - x \rangle \quad \text{for all } y \in \mathbb{R}^n \right\},$$

which is nonempty, compact and convex. Furthermore for all R > 0 we have

$$\sup_{x \in B_R, \, x^* \in \partial f(x)} \|x^*\|_2 < \infty \quad , \quad \text{where} \quad B_R := \{x \in \mathbb{R}^n \mid \|x\|_2 \le R\}.$$

Definition 2.1. The convex function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be α -strongly convex for some $\alpha > 0$, if for all $x, y \in \mathbb{R}^n$ and $x^* \in \partial f(x)$ we have

$$f(y) \ge f(x) + \langle x^*, y - x \rangle + \frac{\alpha}{2} \cdot ||y - x||_2^2.$$

The convex conjugate function of f is $f^* : \mathbb{R}^n \to \mathbb{R}$,

$$f^*(x^*) = \sup_{x \in \mathbb{R}^n} \langle x^*, x \rangle - f(x)$$

Theorem 2.2. If $f : \mathbb{R}^n \to \mathbb{R}$ is α -strongly convex then the conjugate function f^* is differentiable with a $1/\alpha$ -Lipschitz-continuous gradient, i.e.

$$\|\nabla f^*(x^*) - \nabla f^*(y^*)\|_2 \le \frac{1}{\alpha} \cdot \|x^* - y^*\|_2 \text{ for all } x^*, y^* \in \mathbb{R}^n.$$

Definition 2.3. A convex function $f : \mathbb{R}^n \to \mathbb{R}$ is called *piecewise linear-quadratic* if there are finitely many polyhedral sets $F_i \subset \mathbb{R}^n$, $i \in I := \{1, \ldots, p\}$, whose union equals \mathbb{R}^n , and relative to each of which f(x) is given by a convex linear-quadratic function

$$f(x) = \frac{1}{2} \cdot \langle x, A_i x \rangle + \langle a_i, x \rangle + \alpha_i \quad , \quad x \in F_i \,,$$

with symmetric positive-semidefinite matrices $A_i \in \mathbb{R}^{n \times n}$, vectors $a_i \in \mathbb{R}^n$ and $\alpha_i \in \mathbb{R}$. For $x \in \mathbb{R}^n$ we define $I_f(x) := \{i \in I \mid x \in F_i\}$ and $F_x := \bigcap_{i \in I_f(x)} F_i$.

Note that each F_x is polyhedral and there are only finitely many different sets F_x .

Theorem 2.4. If $f : \mathbb{R}^n \to \mathbb{R}$ is convex piecewise linear-quadratic then f^* is also convex piecewise linear-quadratic, and for all $x \in \mathbb{R}^n$ we have

$$\partial f(x) = \operatorname{conv}\{A_i x + a_i \,|\, i \in I_f(x)\}$$

2.1 Bregman distance

The concept of Bregman distance and projections goes back to Bregman [8] and has been successfully used in optimization, see e.g. [2, 4, 10, 13, 40]. The definitions and results in this and the next subsection are taken from [27].

Definition 2.5. Let $f : \mathbb{R}^n \to \mathbb{R}$ be strongly convex. The *Bregman distance* $D_f^{x^*}(x,y)$ between $x, y \in \mathbb{R}^n$ with respect to f and a subgradient $x^* \in \partial f(x)$ is defined as

$$D_f^{x^*}(x,y) := f(y) - f(x) - \langle x^*\,,\,y-x
angle = f^*(x^*) - \langle x^*\,,\,y
angle + f(y)$$
 .

If f is differentiable then we have $\partial f(x) = \{\nabla f(x)\}$ and hence we simply write $D_f(x, y) = D_f^{x^*}(x, y)$.

Note that for $f(x) = \frac{1}{2} ||x||_2^2$ we just have $D_f(x, y) = \frac{1}{2} ||x - y||_2^2$. In general D_f is not a distance function in the usual sense, as it need neither be symmetric, nor does it have to obey a (quasi-)triangle inequality. Nevertheless it has some distance-like properties which we state in the following lemma.

Lemma 2.6. Let $f : \mathbb{R}^n \to \mathbb{R}$ be α -strongly convex. For all $x, y \in \mathbb{R}^n$ and $x^* \in \partial f(x), y^* \in \partial f(y)$ we have

$$\frac{\alpha}{2} \|x - y\|_2^2 \le D_f^{x^*}(x, y) \le \langle x^* - y^*, x - y \rangle \le \|x^* - y^*\|_2 \cdot \|x - y\|_2$$

and hence

$$D_f^{x^*}(x,y) = 0 \quad \Leftrightarrow \quad x = y.$$

For sequences x_k and $x_k^* \in \partial f(x_k)$ boundedness of $D_f^{x_k^*}(x_k, y)$ implies boundedness of both x_k and x_k^* . If f has a L-Lipschitz-continuous gradient then we also have $D_f(x, y) \leq \frac{L}{2} \cdot ||x - y||_2^2$.

2.2 Bregman projections

Definition 2.7. Let $f : \mathbb{R}^n \to \mathbb{R}$ be strongly convex, and $C \subset \mathbb{R}^n$ be a nonempty closed convex set. The *Bregman projection* of x onto C with respect to f and $x^* \in \partial f(x)$ is the unique point $\prod_{C}^{x^*}(x) \in C$ such that

$$D_f^{x^*}(x, \Pi_C^{x^*}(x)) = \min_{y \in C} D_f^{x^*}(x, y) =: \operatorname{dist}_f^{x^*}(x, C)^2.$$

For differentiable f we simply write $\Pi_C(x)$ and $\operatorname{dist}_f(x, C)$.

The notation for the Bregman projection does not capture its dependence on the function f, which, however, will always be clear from the context. Note that for $f(x) = \frac{1}{2} ||x||_2^2$ the Bregman projection is just the orthogonal projection onto C. To distinguish this case we denote the orthogonal projection by $P_C(x)$. We point out that in this case dist $_f(x, C)^2$ and the usual dist $(x, C)^2$ differ by a factor of 2, but we prefer this slight inconsistency to incorporating the factor into the definition of dist_f. The Bregman projection can also be characterized by a variational inequality.

Lemma 2.8 ([27, Lemma 2.2]). Let $f : \mathbb{R}^n \to \mathbb{R}$ be strongly convex. Then a point $\hat{x} \in C$ is the Bregman projection of x onto C with respect to f and $x^* \in \partial f(x)$ iff there is some $\hat{x}^* \in \partial f(\hat{x})$ such that one of the following equivalent conditions is fulfilled

$$\langle \hat{x}^* - x^*, y - \hat{x} \rangle \ge 0 \quad \text{for all} \quad y \in C$$

 $D_f^{\hat{x}^*}(\hat{x}, y) \le D_f^{x^*}(x, y) - D_f^{x^*}(x, \hat{x}) \quad \text{for all} \quad y \in C$

We call any such \hat{x}^* an admissible subgradient for $\hat{x} = \prod_{C}^{x^*}(x)$.

Bregman projections onto affine subspaces and half-spaces can be computed efficiently.

Definition 2.9. Let $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, $u \in \mathbb{R}^n$ and $\beta \in \mathbb{R}$. By L(A, b) we denote the *affine subspace*

$$L(A,b) := \{ x \in \mathbb{R}^n \, | \, Ax = b \},\$$

by $H(u,\beta)$ the hyperplane

$$H(u,\beta) := \{ x \in \mathbb{R}^n \, | \, \langle u \,, \, x \rangle = \beta \}$$

and by $H_{\leq}(u,\beta)$ the half-space

$$H_{<}(u,\beta) := \{ x \in \mathbb{R}^n \, | \, \langle u, x \rangle \le \beta \}.$$

Lemma 2.10 ([27, Lemma 2.4]). Let $f : \mathbb{R}^n \to \mathbb{R}$ be α -strongly convex.

(a) The Bregman projection of $x \in \mathbb{R}^n$ onto $L(A, b) \neq \emptyset$ is

$$\hat{x} := \Pi_{L(A,b)}^{x^*}(x) = \nabla f^*(x^* - A^T \hat{w})$$

where $\hat{w} \in \mathbb{R}^m$ is a solution of

$$\min_{w \in \mathbb{R}^m} f^*(x^* - A^T w) + \langle w, b \rangle$$

Moreover, an admissible subgradient for \hat{x} is $\hat{x}^* := x^* - A^T \hat{w}$. If A has full row rank then for all $y \in L(A, b)$ we have

$$D_f^{\hat{x}^*}(\hat{x}, y) \le D_f^{x^*}(x, y) - \frac{\alpha}{2} \cdot \|(AA^T)^{-\frac{1}{2}}(Ax - b)\|_2^2.$$

(b) The Bregman projection of $x \in \mathbb{R}^n$ onto $H(u, \beta)$ with $u \neq 0$ is

$$\hat{x} := \Pi_{H(u,\beta)}^{x^*}(x) = \nabla f^*(x^* - \hat{t} \cdot u),$$

where $\hat{t} \in \mathbb{R}$ is a solution of

$$\min_{t\in\mathbb{R}}f^*(x^*-t\cdot u)+t\cdot\beta.$$

Moreover, an admissible subgradient for \hat{x} is $\hat{x}^* := x^* - \hat{t} \cdot u$ and for all $y \in H(u, \beta)$ we have

$$D_f^{\hat{x}^*}(\hat{x}, y) \le D_f^{x^*}(x, y) - \frac{\alpha}{2} \cdot \frac{(\langle u, x \rangle - \beta)^2}{\|u\|_2^2}$$

If $x \notin H_{\leq}(u,\beta)$ then we necessarily have $\hat{t} > 0$, $\Pi_{H_{\leq}(u,\beta)}^{x^*}(x) = \hat{x}$ and the above inequality holds for all $y \in H_{\leq}(u,\beta)$.

3 Bounded linear regularity and error bounds

As in [3] for the case of metric projections, we will establish convergence rates with Bregman projections under the assumption of bounded linear regularity. By $\operatorname{rint}(C)$ we denote the relative interior of a subset $C \subset \mathbb{R}^n$.

Definition 3.1. Let $C_1, \ldots C_r \subset \mathbb{R}^n$ be closed convex sets with nonempty intersection $C := \bigcap_{i=1}^r C_i$.

(a) The collection $\{C_1, \ldots, C_r\}$ is called *boundedly linearly regular*, if for every R > 0 there exists $\gamma > 0$ such that for all $x \in B_R$ we have

$$\operatorname{dist}(x, C)^2 \le \gamma \cdot \sum_{i=1}^r \operatorname{dist}(x, C_i)^2,$$

and it is called *linearly regular*, if such an estimate holds globally for all $x \in \mathbb{R}^n$.

(b) The collection $\{C_1, \ldots, C_r\}$ satisfies the standard constraint qualification, if there exists $q \in \{0, \ldots, r\}$ such that C_{q+1}, \ldots, C_r are polyhedral and

$$\bigcap_{i=1}^{q} \operatorname{rint}(C_i) \cap \bigcap_{i=q+1}^{r} C_i \neq \emptyset.$$

Theorem 3.2 (Corollary 3 and 6 in [6]). If the collection $\{C_1, \ldots, C_r\}$ satisfies the standard constraint qualification then it is boundedly linearly regular. And if C is also bounded, then $\{C_1, \ldots, C_r\}$ is linearly regular.

By Lemma 2.6, and since $\operatorname{dist}_{f}^{x^*}(x, C)^2 \leq D_{f}^{x^*}(x, P_C(x))$, we can immediately bound the Bregman distance by the metric distance.

Lemma 3.3. Let $f : \mathbb{R}^n \to \mathbb{R}$ be strongly convex.

(a) For all $x \in \mathbb{R}^n$, $x^* \in \partial f(x)$ and $y^* \in \partial f(P_C(x))$ we have

$$\operatorname{dist}_{f}^{x^{*}}(x,C)^{2} \leq ||x^{*} - y^{*}||_{2} \cdot \operatorname{dist}(x,C)$$

(b) If f has a L-Lipschitz-continuous gradient then we have for all $x \in \mathbb{R}^n$

$$\operatorname{dist}_f(x,C)^2 \leq \frac{L}{2} \cdot \operatorname{dist}(x,C)^2$$

In general, it is not obvious how to extend the second (and better) estimate to non-differentiable functions f, because we lack an inequality like $||x^* - y^*||_2 \le L \cdot ||x - y||_2$. However, we can achieve the better estimate for convex piecewise linear-quadratic f. The result is based on the following lemma, which exploits the fact that the subgradients on the sets F_x are closely related, cf. Definition 2.3.

Lemma 3.4. Let $f : \mathbb{R}^n \to \mathbb{R}$ be strongly convex piecewise linear-quadratic and $C \subset \mathbb{R}^n$ be closed convex. Then for all R > 0 there exists L > 0 such that for all $x \in B_R$ and $x^* \in \partial f(x)$ we have

$$\operatorname{dist}_{f}^{x^{*}}(x,C)^{2} \leq \begin{cases} L \cdot \operatorname{dist}(x,C)^{2} & , F_{x} \cap C = \emptyset \\ L \cdot \operatorname{dist}(x,F_{x} \cap C)^{2} & , F_{x} \cap C \neq \emptyset \end{cases}$$

Proof. Since B_R is compact we have $dist(B_R \cap F_x, C) > 0$ for all $x \in B_R$ with $F_x \cap C = \emptyset$. Since there are only finitely many different sets F_x it follows that

$$d := \min\{\operatorname{dist}(B_R \cap F_x, C) \mid x \in B_R \text{ with } F_x \cap C = \emptyset\} > 0.$$

Furthermore there is a constant c > 0 such that $||x^* - y^*||_2 \le c$ for all $x \in B_R$, $x^* \in \partial f(x)$ and $y^* \in \partial f(P_C(x))$. Let $x \in B_R$ and $x^* \in \partial f(x)$. By Theorem 2.4 there are $\lambda_i \in [0, 1]$ with $\sum_{i \in I_f(x)} \lambda_i = 1$ such that

$$x^* = \sum_{i \in I_f(x)} \lambda_i \cdot (A_i x + a_i).$$

In case $F_x \cap C = \emptyset$ we have dist $(x, C) \ge d$, and hence by Lemma 3.3 we get

$$\operatorname{dist}_{f}^{x^{*}}(x,C)^{2} \leq \|x^{*} - y^{*}\|_{2} \cdot \operatorname{dist}(x,C) \leq \frac{c}{d} \cdot \operatorname{dist}(x,C)^{2}.$$

In case $F_x \cap C \neq \emptyset$ we set $\hat{x} := P_{F_x \cap C}(x)$. Since $\hat{x} \in F_x$ we have $I_f(x) \subset I_f(\hat{x})$, and therefore we can choose the following subgradient of f at \hat{x} ,

$$\hat{x}^* := \sum_{i \in I_f(x)} \lambda_i \cdot (A_i \hat{x} + a_i)$$

with the same λ_i as for x^* . We set $L_f := \max\{||A_i||_2 \mid i \in I\}$ and estimate

$$\langle x^* - \hat{x}^*, x - \hat{x} \rangle = \sum_{i \in I_f(x)} \lambda_i \cdot \langle A_i(x - \hat{x}), x - \hat{x} \rangle \le L_f \cdot ||x - \hat{x}||_2^2,$$

which yields $\operatorname{dist}_{f}^{x^{*}}(x,C)^{2} \leq \langle x^{*} - \hat{x}^{*}, x - \hat{x} \rangle \leq L_{f} \cdot \operatorname{dist}(x, F_{x} \cap C)^{2}.$

Now we can prove the main theorem of this section.

Theorem 3.5. Let $f : \mathbb{R}^n \to \mathbb{R}$ be strongly convex piecewise linear-quadratic, and let $C \subset \mathbb{R}^n$ be closed convex such that the collections $\{F_x, C\}$ are boundedly linearly regular for all $x \in \mathbb{R}^n$ with $F_x \cap C \neq \emptyset$. Then for all R > 0 there exists L > 0 such that for all $x \in B_R$ and $x^* \in \partial f(x)$ we have

$$\operatorname{dist}_{f}^{x^{*}}(x,C)^{2} \leq L \cdot \operatorname{dist}(x,C)^{2}.$$

Proof. The assertion immediately follows from Lemma 3.4 and Definition 3.1, because $dist(x, F_x) = 0$.

Remark 3.6. If C is polyhedral then by Theorem 3.2 all collections $\{F_x, C\}$ are boundedly linearly regular.

For the split feasibility problem we also need the following generalization of Hoffmann's error bound [23] to possibly non-polyhedral sets, which are defined by convex constraints in the range $\mathcal{R}(A)$ of a matrix A.

Lemma 3.7. Let the convex set $C \subset \mathbb{R}^n$ have the form $C = \{x \in \mathbb{R}^n | Ax \in Q\}$ with $A \in \mathbb{R}^{m \times n}$ and $Q \subset \mathbb{R}^m$ closed convex such that the collection $\{Q, \mathcal{R}(A)\}$ is boundedly linearly regular. Then for every R > 0 there exists $\gamma > 0$ such that for all $x \in B_R$ we have

$$\operatorname{dist}(x, C) \le \gamma \cdot \operatorname{dist}(Ax, Q).$$

Proof. In case A = 0 (and $0 \in Q$) we have $C = \mathbb{R}^n$ and hence the assertion holds trivially. Otherwise let $\sigma_{min} > 0$ be the smallest positive singular value of A, and let R > 0. Since $\{Q, \mathcal{R}(A)\}$ is boundedly linearly regular, there exists $\gamma > 0$ such that for all $x \in B_R$ we have

$$\operatorname{dist}(Ax, Q \cap \mathcal{R}(A)) \leq \gamma \cdot \operatorname{dist}(Ax, Q)$$

To $x \in B_R$ we find some $\hat{x} \in C$ such that $A\hat{x} = P_{Q \cap \mathcal{R}(A)}(Ax)$. Since $\hat{x} + \mathcal{N}(A) \subset C$ for the nullspace $\mathcal{N}(A)$ of A we get

$$dist(x, C) \leq ||x - P_{\hat{x} + \mathcal{N}(A)}(x)||_2 = ||(x - \hat{x}) - P_{\mathcal{N}(A)}(x - \hat{x})||_2$$

$$\leq \frac{1}{\sigma_{min}} \cdot ||Ax - A\hat{x}||_2 = \frac{1}{\sigma_{min}} \cdot dist (Ax, Q \cap \mathcal{R}(A))$$

$$\leq \frac{\gamma}{\sigma_{min}} \cdot dist(Ax, Q),$$

from which the assertion follows.

Note that for polyhedral sets Q the collection $\{Q, \mathcal{R}(A)\}$ is always boundedly linearly regular. Moreover in this case the classical result of Hoffmann holds globally for all $x \in \mathbb{R}^n$, cf. [23]. For non-polyhedral sets Q the assertion holds if $\operatorname{rint}(Q) \cap \mathcal{R}(A) \neq \emptyset$, cf. Theorem 3.2. Indeed, if this condition is not fulfilled, the assertion cannot be guaranteed in general, as the following counterexample demonstrates: For $Q = \{x \in \mathbb{R}^2 \mid ||x - (0, 1)^T||_2 \leq 1\}$ and $A = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$ we have $Q \cap \mathcal{R}(A) = \{0\}, C = \{0\} \times \mathbb{R}$ and hence for $x_1 > 0$ we get

$$\frac{\operatorname{dist}(A(x_1,0)^T,Q)}{\operatorname{dist}((x_1,0)^T,C)} = \frac{\sqrt{1+x_1^2}-1}{x_1} = \frac{x_1}{\sqrt{1+x_1^2}+1} \longrightarrow 0 \quad \text{for} \quad x_1 \searrow 0 \,.$$

Finally we concentrate on feasible linearly constrained optimization problems,

$$\min_{x \in \mathbb{R}^n} f(x) \quad \text{s.t.} \quad Ax = b \tag{6}$$

like in (3) or (5). If the objective function f is strongly convex then (6) has a unique solution \hat{x} which fulfills $\partial f(\hat{x}) \cap \mathcal{R}(A^T) \neq \emptyset$, and hence coincides with the Bregman projection $\Pi_{L(A,b)}^{x^*}(x)$ with respect to f for all $x \in \mathbb{R}^n$ with $x^* \in \partial f(x) \cap \mathcal{R}(A^T) \neq \emptyset$, cf. Lemma 2.10 (a). As a consequence for all such x, x^* we have $\operatorname{dist}_f^{x^*}(x, L(A, b))^2 = D_f^{x^*}(x, \hat{x})$. Our next aim is an error bound of the form $D_f^{x^*}(x, \hat{x}) \leq \gamma \cdot ||Ax - b||_2^2$. For piecewise linear-quadratic or differentiable f this immediately follows from Lemma 3.5 and 3.3 (b) and Hoffmann's error bound. But we will also achieve this result under weaker assumtions. To clarify these assumtions we need the concept of calmness of a set-valued mapping [37].

Definition 3.8. A set-valued mapping $S : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ is *calm* at $\hat{x} \in \mathbb{R}^n$ if $S(\hat{x}) \neq \emptyset$ and there are constants $\epsilon, L > 0$ such that

$$S(x) \subset S(\hat{x}) + L \cdot ||x - \hat{x}||_2 \cdot B_1 \quad , \quad ||x - \hat{x}||_2 \le \epsilon.$$

Example 3.9. (a) Any polyhedral multifunction, i.e. a set-valued mapping whose graph is the union of finitely many polyhedral sets, is calm at each $\hat{x} \in \mathbb{R}^n$. In particular this holds for the subdifferential mapping $\partial f(x)$ of a convex piecewise linear-quadratic function $f : \mathbb{R}^n \to \mathbb{R}$, see Proposition 1 in [36].

(b) Let $\sigma(X) \in \mathbb{R}^m$ denote the vector of singular values of $X \in \mathbb{R}^{n_1 \times n_2}$ (with $m = \min\{n_1, n_2\}$), and let $h : \mathbb{R}^m \to \mathbb{R}$ be a convex piecewise linearquadratic function which is absolutely symmetric, i.e. $h(x_1, \ldots, x_m) = h(|x_{\pi(1)}|, \ldots, |x_{\pi(m)}|)$ for any permutation π of the indices. Then the subdifferential mapping of $f(X) := h(\sigma(X))$ is calm at each $\hat{X} \in \mathbb{R}^{n_1 \times n_2}$. In particular this holds for the nuclear norm $||X||_* := ||\sigma(X)||_1$, the spectral norm $||X||_2 := ||\sigma(X)||_{\infty}$ and $f(X) = \lambda \cdot ||X||_* + \frac{1}{2} \cdot ||X||_F^2$. Furthermore the subdifferential mapping of

$$f(X_1, X_2) = \frac{1}{2} \cdot \|X_1\|_F^2 + \lambda_1 \cdot \|X_1\|_* + \frac{1}{2} \cdot \|X_2\|_F^2 + \lambda_2 \cdot \|X_2\|_1$$

is calm at each $(\hat{X}_1, \hat{X}_2) \in \mathbb{R}^{n_1 \times n_2} \times \mathbb{R}^{n_1 \times n_2}$, where $||X||_1$ denotes the 1-norm of all entries of a matrix X, see Example 2.10 in [39].

Now we can reformulate Theorem 2.12 in [39] to fit the present context.

Theorem 3.10. Consider the linearly constrained optimization problem (6) with $A \in \mathbb{R}^{m \times n}$, $b \in \mathcal{R}(A)$, and strongly convex $f : \mathbb{R}^n \to \mathbb{R}$. Let $x_0 \in \mathbb{R}^n$ and $x_0^* \in \partial f(x_0) \cap \mathcal{R}(A^T)$ be given. If the subdifferential mapping of f is calm at the unique solution \hat{x} of (6) and if the collection $\{\partial f(\hat{x}), \mathcal{R}(A^T)\}$ is linearly regular, then there exists $\gamma > 0$ such that for all $x \in \mathbb{R}^n$ and $x^* \in \partial f(x) \cap \mathcal{R}(A^T)$ with $D_f^{x^*}(x, \hat{x}) \leq D_f^{x_0^*}(x_0, \hat{x})$ we have

$$\operatorname{dist}_{f}^{x^{*}}(x, L(A, b))^{2} = D_{f}^{x^{*}}(x, \hat{x}) \leq \gamma \cdot \|Ax - b\|_{2}^{2}.$$

Proof. To obtain the error bound we apply the results of [39] to the objective function $g(y) = f^*(A^T y) - \langle b, y \rangle$ of the unconstrained dual

$$\min_{y \in \mathbb{R}^m} f^*(A^T y) - \langle b, y \rangle,$$

which relates to the Bregman distance in the following way by setting $x^* = A^T y$, $x = \nabla f^*(x^*)$ and observing that $\langle b, y \rangle = \langle x^*, \hat{x} \rangle$,

$$D_{f}^{x^{*}}(x,\hat{x}) = g(y) - g_{min}.$$

It follows from Theorem 2.12 in [39] that the function g is restricted strongly convex on all of its level sets. Hence, by Lemma 2.2 in [39], there exists $\gamma > 0$ such that for all $x \in \mathbb{R}^n$ and $x^* \in \partial f(x) \cap \mathcal{R}(A^T)$ with $D_f^{x^*}(x, \hat{x}) \leq D_f^{x^*_0}(x_0, \hat{x})$ we have

$$D_f^{x^*}(x, \hat{x}) = g(y) - g_{min} \le \gamma \cdot \|\nabla g(y)\|_2^2 = \gamma \cdot \|Ax - b\|_2^2.$$

4 Randomized Bregman Projections for SFP

The convex feasibility problem (CFP) is to find a common point of finitely many closed convex sets $C_i \subset \mathbb{R}^n$, $i \in I := \{1, \ldots, m\}$, with nonempty intersection,

find
$$x \in C := \bigcap_{i \in I} C_i$$
. (7)

A simple and widely known idea to solve (7) is to project successively onto the individual sets C_i and we refer to [3] for an excellent introduction. By now there is a vast literature on CFPs and projection algorithms for their solution, see e.g. [4, 5, 8, 12, 17, 44]. These projection algorithms are most efficient if the projections onto the individual sets are relatively cheap. Here we concentrate on a special instance of the CFP, also called *split feasibility problem* (SFP) [11, 13, 16, 40], where some or all of the sets C_i arise by imposing convex constraints $Q_i \subset \mathbb{R}^{m_i}$ in the range of a matrix $A_i \in \mathbb{R}^{m_i \times n}$,

$$C_i = \{ x \in \mathbb{R}^n \, | \, A_i x \in Q_i \} \,. \tag{8}$$

In general projections onto such sets can be prohibitively expensive and it is often preferable to use projections onto suitable enclosing halfspaces. The following lemma shows a construction of such an enclosing halfspace, see [27].

Lemma 4.1. Let $Q \subset \mathbb{R}^m$ be a nonempty closed convex set and $A \in \mathbb{R}^{m \times n}$. Assume that $\tilde{x} \notin C = \{x \in \mathbb{R}^n \mid Ax \in Q\}$ and set

$$w := A\tilde{x} - P_Q(A\tilde{x})$$
 and $\beta := \langle A^T w, \tilde{x} \rangle - \|w\|_2^2$

Then it holds that $A^T w \neq 0$, $\tilde{x} \notin H_{\leq}(A^T w, \beta)$ and $C \subset H_{\leq}(A^T w, \beta)$. In other words, the hyperplane $H(A^T w, \beta)$ separates \tilde{x} from C.

To solve a split feasibility problem one can proceed as follows: Let $I_Q \subset I$ be the subset of all indices *i* belonging to sets of the form (8), and denote by $I_C := I \setminus I_Q$ the set of the remaining indices. Encounter the different constraints C_i successively and project the current iterate onto C_i in case $i \in I_C$, or onto an enclosing halfspace according to Lemma 4.1 and Lemma 2.10 (b) in case $i \in I_Q$, see Algorithm 1. In [27] convergence of the iterates to a solution of (7) was shown for Bregman projections with respect to nondifferentiable functions, and for quite general control sequences $i : \mathbb{N} \to I$. The only requirement was that $(i(k))_{k \in \mathbb{N}}$ encounters each index in I infinitely often.¹ However, no assertion was made about convergence rates. Here we follow [1, 9, 19, 26, 29, 31, 35, 41, 45] and show that a randomized version of the algorithm converges in expectation to a solution of (7) with an expected (sub-)linear convergence rate.

Theorem 4.2. Let $f : \mathbb{R}^n \to \mathbb{R}$ be α -strongly convex. Consider the SFP (7) under the assumption that the collections $\{C_1, \ldots, C_r\}$ and $\{Q_i, \mathcal{R}(A_i)\}$ for each $i \in I_Q$ are boundedly linearly regular. Then for any starting points $x_0 \in \mathbb{R}^n$ and $x_0^* \in \partial f(x_0)$ the iterates x_k and x_k^* of Algorithm 1 remain bounded, the Bregman distances to C decrease monotonically,

$$\operatorname{dist}_{f}^{x_{k+1}^{*}}(x_{k+1}, C) \leq \operatorname{dist}_{f}^{x_{k}^{*}}(x_{k}, C),$$

¹Because very general control sequences $i : \mathbb{N} \to I$ besides simple cyclic control fulfill this requirement, the corresponding method was also called *method of random Bregman projections* in [4]. But such control sequences are not necessarily stochastic objects, in contrast to the situation in the present work. Hence we use the word *randomized* in Algorithm 1 instead of *random* to distinguish between the cases.

Algorithm 1 Randomized Bregman projections for split feasibility problems (RBPSFP)

Input: starting points $x_0 \in \mathbb{R}^n$, $x_0^* \in \partial f(x_0)$ and probabilities $p_i > 0, i \in I$ **Output:** a solution of (7)

1: initialize k = 0

2: repeat

- 3: choose an index $i_k = i \in I$ at random with probability $p_i > 0$
- if $i_k \in I_C$ then 4:
- update $x_{k+1} = \prod_{C_{i_k}}^{x_k^*}(x_k)$ together with an admissible subgradient 5: $x_{k+1}^* \in \partial f(x_{k+1})$, cf. Lemma 2.8 else if $i_k \in I_Q$ then
- 6:
- 7:
- set $w_k \in I_Q$ under set $w_k = A_{i_k} x_k P_{Q_{i_k}} (A_{i_k} x_k)$ and $\beta_k = \langle A_{i_k}^T w_k, x_k \rangle \|w_k\|_2^2$ update $x_{k+1} = \prod_{H \leq (A_{i_k}^T w_k, \beta_k)}^{x_k^*} (x_k)$ with $x_{k+1}^* \in \partial f(x_{k+1})$ as in 8: Lemma 2.10 (b)

end if 9:

- increment k = k + 110:
- 11: **until** a stopping criterion is satisfied

and converge in expectation to zero, where the expectation is taken with respect to the probability distribution $p_i > 0$, $i \in I$. The expected rate of convergence is at least sublinear: There is a constant c > 0 such that

$$\mathbb{E}\left[\operatorname{dist}(x_k, C)\right] \le \frac{c}{\sqrt{k}}$$

Proof. At first we consider the case $i_k \in I_C$. By Lemma 2.6 we have

$$D_f^{x_k^*}(x_k, x_{k+1}) \ge \frac{\alpha}{2} \cdot \|x_k - x_{k+1}\|_2^2 \ge \frac{\alpha}{2} \cdot \operatorname{dist}(x_k, C_{i_k})^2,$$

and together with Lemma 2.8 we can estimate for all $x \in C$

$$D_f^{x_{k+1}^*}(x_{k+1}, x) \le D_f^{x_k^*}(x_k, x) - \frac{\alpha}{2} \cdot \operatorname{dist}(x_k, C_{i_k})^2.$$
(9)

Now we consider the case $i_k \in I_Q$. By Lemma 4.1 we have $C \subset H_{\leq}(A_{i_k}^T w_k, \beta_k)$, and together with Lemma 2.10 (b) we can estimate for all $x \in C$

$$D_{f}^{x_{k+1}^{*}}(x_{k+1},x) \le D_{f}^{x_{k}^{*}}(x_{k},x) - \frac{\alpha}{2 \cdot \|A_{i_{k}}\|_{2}^{2}} \cdot \|A_{i_{k}}x_{k} - P_{Q_{i_{k}}}(A_{i_{k}}x_{k})\|_{2}^{2}.$$
 (10)

We fix some $x \in C$ and conclude from (9), (10) and Lemma 2.6 that both x_k and x_k^* remain bounded. Hence by Lemma 3.7 and the bounded linear regularity of all $\{Q_i, \mathcal{R}(A_i)\}, i \in I_Q$, there exist $\gamma_i > 0$ such that for all k we have

$$\operatorname{dist}(x_k, C_i) \le \gamma_i \cdot \|A_{i_k} x_k - P_{Q_{i_k}}(A_{i_k} x_k)\|_2.$$

Inserting this estimate into (10) we get

$$D_f^{x_{k+1}^*}(x_{k+1}, x) \le D_f^{x_k^*}(x_k, x) - \frac{\gamma_i^2 \cdot \alpha}{2 \cdot \|A_{i_k}\|_2^2} \cdot \operatorname{dist}(x_k, C_{i_k})^2.$$

Together with (9) this implies that the Bregman distances decrease monotonically, and that there is a constant c > 0 such that

$$\operatorname{dist}_{f}^{x_{k+1}^{*}}(x_{k+1}, C)^{2} \leq \operatorname{dist}_{f}^{x_{k}^{*}}(x_{k}, C)^{2} - c \cdot \operatorname{dist}(x_{k}, C_{i_{k}})^{2}.$$
(11)

For the moment we fix the values of the indices i_0, \ldots, i_{k-1} and consider only i_k as a random variable with values in I. Taking the expectation on both sides of (11) conditional to the values of the indices i_0, \ldots, i_{k-1} yields

$$\mathbb{E}\left[\operatorname{dist}_{f}^{x_{k+1}^{*}}(x_{k+1},C)^{2} \mid i_{0},\ldots,i_{k-1}\right] \leq \operatorname{dist}_{f}^{x_{k}^{*}}(x_{k},C)^{2} - \sum_{i \in I} p_{i} \cdot c \cdot \operatorname{dist}(x_{k},C_{i})^{2}.$$

By boundedness of x_k and bounded linear regularity of the collection $\{C_1, \ldots, C_m\}$ there is $\gamma > 0$ such that for all k we have

$$\mathbb{E}\left[\operatorname{dist}_{f}^{x_{k+1}^{*}}(x_{k+1},C)^{2} \middle| i_{0},\ldots,i_{k-1}\right] \leq \operatorname{dist}_{f}^{x_{k}^{*}}(x_{k},C)^{2} - \gamma \cdot \operatorname{dist}(x_{k},C)^{2}.$$
 (12)

Furthermore, by Lemma 3.3 (a) there is L > 0 such that for all k we have $\operatorname{dist}_{f}^{x^{*}}(x_{k}, C)^{4} \leq L \cdot \operatorname{dist}(x_{k}, C)^{2}$, and hence we get

$$\mathbb{E}\left[\operatorname{dist}_{f}^{x_{k+1}^{*}}(x_{k+1},C)^{2} \middle| i_{0},\ldots,i_{k-1}\right] \leq \operatorname{dist}_{f}^{x_{k}^{*}}(x_{k},C)^{2} - \frac{\gamma}{L} \cdot \operatorname{dist}_{f}^{x_{k}^{*}}(x_{k},C)^{4}.$$

Now we consider all indices i_0, \ldots, i_k as random variables with values in I, and take the full expectation on both sides,

$$\mathbb{E}\left[\operatorname{dist}_{f}^{x_{k+1}^{*}}(x_{k+1},C)^{2}\right] \leq \mathbb{E}\left[\operatorname{dist}_{f}^{x_{k}^{*}}(x_{k},C)^{2}\right] - \frac{\gamma}{L} \cdot \mathbb{E}\left[\operatorname{dist}_{f}^{x_{k}^{*}}(x_{k},C)^{4}\right]$$
$$\leq \mathbb{E}\left[\operatorname{dist}_{f}^{x_{k}^{*}}(x_{k},C)^{2}\right] - \frac{\gamma}{L} \cdot \left(\mathbb{E}\left[\operatorname{dist}_{f}^{x_{k}^{*}}(x_{k},C)^{2}\right]\right)^{2}.$$

We set $d_k := \mathbb{E}\left[\operatorname{dist}_f^{x_k^*}(x_k, C)^2\right]$. Then we have $d_{k+1} \leq d_k - \frac{\gamma}{L}d_k^2$. We observe that d_k is decreasing and by rearranging the inequality to

$$\frac{1}{d_{k+1}} \ge \frac{1}{d_k} + \frac{\gamma}{L} \frac{d_k}{d_{k+1}} \ge \frac{1}{d_k} + \frac{\gamma}{L}$$

we obtain $\frac{1}{d_{k+1}} \geq \frac{1}{d_0} + \frac{\gamma}{L}(k+1)$, and we conclude $d_k \leq \frac{Ld_0}{L+\gamma d_0 \cdot k}$ as desired. The expected sublinear convergence rates for $\operatorname{dist}(x_k, C)$ now follow from the estimate $\mathbb{E}\left[\operatorname{dist}(x_k, C)\right] \leq \sqrt{\frac{2}{\alpha}} \cdot \mathbb{E}\left[\operatorname{dist}_{f}^{x_k^*}(x_k, C)\right]$, cf. Lemma 2.6. **Remark 4.3.** According to Lemma 2.10 (b) the computation of the Bregman projection $x_{k+1} = \prod_{H \leq (A_{i_k}^T w_k, \beta_k)}^{x_k^*}(x_k)$ onto the halfspace $H_{\leq}(A_{i_k}^T w_k, \beta_k)$ in step 8 of Algorithm 1 amounts to an exact linesearch. In practice, this is feasible only in special cases, e.g. for $f(x) = ||x||_2^2$ or $f(x) = \lambda \cdot ||x||_1 + \frac{1}{2} ||x||_2^2$. But the assertions of Theorem 4.2 and the next two theorems remain true for inexact linesearches as well, cf. [27]. In particular, we may choose

$$t_k := \alpha \cdot \frac{\|w_k\|_2^2}{\|A_{i_k}^T w_k\|_2^2} \quad , \quad x_{k+1}^* := x_k^* - t_k \cdot A_{i_k}^T w_k \quad , \quad x_{k+1} = \nabla f^*(x_{k+1}^*) \, .$$

For piecewise linear-quadratic or differentiable f the expected rate of convergence is even linear.

Theorem 4.4. If f is piecewise linear-quadratic or has a Lipschitz-continuous gradient, then under the assumptions of Theorem 4.2 the expected rate of convergence is linear: There are constants $q \in (0, 1)$ and c > 0 such that

$$\mathbb{E}\left[\operatorname{dist}_{f}^{x_{k+1}^{*}}(x_{k+1},C)^{2}\right] \leq q \cdot \mathbb{E}\left[\operatorname{dist}_{f}^{x_{k}^{*}}(x_{k},C)^{2}\right],$$

and hence

$$\mathbb{E}\left[\operatorname{dist}(x_k, C)\right] \le c \cdot q^{\frac{k}{2}}.$$

Proof. By Theorem 3.5 and Lemma 3.3 (b) respectively, there is L > 0 such that for all k we have $\operatorname{dist}_{f}^{x_{k}^{*}}(x_{k}, C)^{2} \leq L \cdot \operatorname{dist}(x_{k}, C)^{2}$. Hence, using this in (12) in the proof of Theorem 4.2 we get

$$\mathbb{E}\left[\operatorname{dist}_{f}^{x_{k+1}^{*}}(x_{k+1},C)^{2}\right] \leq \left(1 - \frac{\gamma}{L}\right) \cdot \mathbb{E}\left[\operatorname{dist}_{f}^{x_{k}^{*}}(x_{k},C)^{2}\right],$$

from which the linear convergence rates follow.

Finally we turn to linearly constrained optimization problems.

Theorem 4.5. Consider the linearly constrained optimization problem (6) under the assumptions of Theorem 3.10. Let I_1, \ldots, I_r be a covering of $\{1, \ldots, m\}$ (not necessarily disjoint), denote by A_i the matrix consisting of the rows of A indexed by I_i , and let b_i denote the vector consisting of the entries of b indexed by I_i . The constraints $A_i x = b_i$ may be considered both as constraints with $i \in I_C$, cf. Lemma 2.10 (a), or with $i \in I_Q$ and $Q_i = \{b_i\}$. If the initial values are chosen as $x_0^* \in \mathcal{R}(A^T)$ and $x_0 = \nabla f^*(x_0^*)$ then the iterates of Algorithm 1 converge in expectation to the solution \hat{x} of (6). The expected rate of convergence is linear: There are constants $q \in (0, 1)$ and c > 0 such that

$$\mathbb{E}\left[D_f^{x_{k+1}^*}(x_{k+1},\hat{x})\right] \le q \cdot \mathbb{E}\left[D_f^{x_k^*}(x_k,\hat{x})\right]$$

and hence

$$\mathbb{E}\left[\left\|x_k - \hat{x}\right\|\right] \le c \cdot q^{\frac{\kappa}{2}}$$

Proof. Since $x_0^* \in \mathcal{R}(A^T)$ and the updates are of the form $x_k^* = x_{k-1}^* - A^T v_k$ for some $v_k \in \mathbb{R}^m$, we inductively get $x_k^* \in \mathcal{R}(A^T)$ for all $k \ge 0$. Hence the assertion follows from Theorem 3.10 as in the proofs of Theorem 4.2 and 4.4.

5 Linear convergence of the Randomized Sparse Kaczmarz method

Here we show how to apply Theorem 4.5 to obtain linear convergence of the Randomized Sparse Kaczmarz method. As illustrated in [28], the Sparse Kaczmarz method (2) can be considered as a special case of Algorithm 1 applied to the regularized Basis Pursuit problem (3). The objective function

$$f(x) = \lambda \|x\|_1 + \frac{1}{2} \|x\|_2^2 \tag{13}$$

is 1-strongly convex and also piecewise linear-quadratic with $\nabla f^*(x^*) = S_{\lambda}(x^*)$. We formulate the constraint Ax = b with sets $Q_i = \{b_i\}$ and mappings $A_i = a_i^T$ with the rows a_i^T of $A, i \in \{1, \ldots, m\}$. Step 7 in Algorithm 1 then reads as

$$w_k = \langle a_{i_k}, x_k \rangle - b_{i_k} , \beta_k = \langle a_{i_k} w_k, x_k \rangle - |w_k|^2.$$

According to Lemma 2.10, the Bregman projection $x_{k+1} = \prod_{H(A_{i_k}^T w_k, \beta_k)}^{x_k^*}(x_k)$ in Step 8 can be computed as

$$x_{k+1} = \nabla f^*(x_k^* - t_k \cdot a_{i_k} \cdot w_k) = S_\lambda \left(x_k^* - t_k \cdot \left(\langle a_{i_k}, x_k \rangle - b_{i_k} \right) \cdot a_{i_k} \right)$$

with an appropriate stepsize t_k . Now we use the inexact stepsize according to Remark 4.3 with $\alpha = 1$, namely

$$t_k = \frac{|w_k|^2}{\|a_{i_k}w_k\|_2^2} = \frac{1}{\|a_{i_k}\|_2^2}.$$

Hence, we do not need the quantity β_k to perform the iteration, and the full step reads as

$$x_{k+1}^* = x_k^* - \frac{\langle a_{i_k}, x_k \rangle - b_{i_k}}{\|a_{i_k}\|_2^2} \cdot a_{i_k} \quad , \quad x_{k+1} = S_\lambda(x_{k+1}^*) \cdot a_{i_k}$$

We recover the Randomized Sparse Kaczmarz method, which we state here as Algorithm 2.

As already noted in [27], it is also possible to perform an exact linesearch for the Sparse Kaczmarz method. To do so, in each step one has to solve the one-dimensional problem

$$t_k = \operatorname*{argmin}_{t \in \mathbb{R}} f^*(x_k^* - t \cdot a_{i_k}) + t \cdot b_{i_k} \tag{14}$$

which can be done in reasonable time since f^* is piecewise linear-quadratic, see [27, Section 2.5.2]. This results in the Exact-Step Randomized Sparse Kaczmarz (ERSK) method, stated as Algorithm 3. Note that ERSK can also be derived by directly considering the constraints as $C_i = H(a_i, b_i)$ and performing exact Bregman projections onto C_i .

As a consequence of Theorem 4.5 we can conclude the following:

Algorithm 2 Randomized Sparse Kaczmarz method (RSK)

Input: starting points $x_0 = x_0^* = 0 \in \mathbb{R}^n$, matrix $A \in \mathbb{R}^{m \times n}$, vector $b \in \mathbb{R}^m$ such that Ax = b is consistent, and probabilities $p_i > 0, i \in \{1, \ldots, m\}$

Output: the solution of $\min_{x \in \mathbb{R}^n} \lambda ||x||_1 + \frac{1}{2} ||x||_2^2$ s.t. Ax = b

- 1: initialize k = 0
- 2: repeat
- 3: choose an index $i_k = i \in \{1, ..., m\}$ at random with probability $p_i > 0$
- 4: set $a_{i_k}^T$ to the i_k -th row of A
- 5: update $x_{k+1}^* = x_k^* \frac{\langle a_{i_k}, x_k \rangle b_{i_k}}{\|a_{i_k}\|_2^2} \cdot a_{i_k}$
- 6: update $x_{k+1} = S_{\lambda}(x_{k+1}^*)$
- 7: increment k = k + 1
- 8: until a stopping criterion is satisfied

Algorithm 3 Exact-Step Randomized Sparse Kaczmarz method (ERSK)

Input: starting points $x_0 = x_0^* = 0 \in \mathbb{R}^n$, matrix $A \in \mathbb{R}^{m \times n}$, vector $b \in \mathbb{R}^m$ such that Ax = b is consistent, and probabilities $p_i > 0, i \in \{1, \ldots, m\}$

Output: the solution of $\min_{x \in \mathbb{R}^n} \lambda ||x||_1 + \frac{1}{2} ||x||_2^2$ s.t. Ax = b

- 1: initialize k = 0
- 2: repeat
- 3: choose an index $i_k = i \in \{1, ..., m\}$ at random with probability $p_i > 0$
- 4: set $a_{i_k}^T$ to the *i*-th row of A
- 5: calculate $t_k = \operatorname{argmin}_{t \in \mathbb{R}} f^*(x_k^* t \cdot a_{i_k}) + t \cdot b_{i_k}$
- 6: update $x_{k+1}^* = x_k^* t_k \cdot a_{i_k}$
- 7: update $x_{k+1} = S_{\lambda}(x_{k+1}^*)$
- 8: increment k = k + 1
- 9: until a stopping criterion is satisfied

Corollary 5.1. Let $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$ be in the range of A and let $\lambda > 0$. Then both the RSK method from Algorithm 2 and the ERSK method from Algorithm 3 converge in expectation to the unique solution \hat{x} of

$$\min_{x \in \mathbb{R}^n} \lambda \|x\|_1 + \frac{1}{2} \|x\|_2^2 \quad s.t. \quad Ax = b$$

at a linear rate, i.e. in both cases there exist $q \in (0,1)$ and c > 0 such that

$$\mathbb{E}\left[\left\|x_k - \hat{x}\right\|\right] \le c \cdot q^{\frac{k}{2}}.$$

Expected linear convergence for a randomized and smoothed Sparse Kaczmarz method was also shown in [33]. There the objective function (13) was replaced by

$$f_{\epsilon}(x) = \lambda \cdot r_{\epsilon}(x) + \frac{1}{2} \|x\|_2^2 \tag{15}$$

with $\epsilon > 0$ and $r_{\epsilon}(x)$ beeing the Moreau envelope of $||x||_1$,

$$r_{\epsilon}(x) = \sum_{i=1}^{n} \begin{cases} |x_i| - \frac{\epsilon}{2} & , |x_i| > \epsilon \\ \frac{x_i^2}{2\epsilon} & , |x_i| \le \epsilon \end{cases}$$

The function f_{ϵ} is 1-strongly convex and has a Lipschitz-continuous gradient. Hence linear convergence is also guaranteed by Theorem 3.10. But as shown above, Theorem 3.10 also allows us to prove this result without smoothing the objective function. Of course this also holds for the Randomized Block Sparse Kaczmarz method considered in [33] by applying Theorem 3.10 with a covering I_1, \ldots, I_r of $\{1, \ldots, m\}$.

6 Numerical examples

In two experiments we illustrate the impact of the Randomized Sparse Kaczmarz method versus the (non-sparse) Randomized Kaczmarz and the (nonrandomized) Sparse Kaczmarz method.

6.1 Sparse vs. non-sparse Randomized Kaczmarz

We constructed overdetermined linear systems with Gaussian matrices $A \in \mathbb{R}^{m \times n}$ for $m \geq n$, and sparse solutions $\hat{x} \in \mathbb{R}^n$ with corresponding right hand sides $b = A\hat{x} \in \mathbb{R}^m$ and also respective noisy right hand sides b^{δ} . We ran the usual Randomized Kaczmarz method (RK), the Randomized Sparse Kaczmarz method (RSK) (Algorithm 2), and the Exact-Step Randomized Sparse Kaczmarz method (ERSK) (Algorithm 3) on the problem. Note that, since with high probability the matrices A have full rank, in the case of no noise the solution \hat{x} is unique, and so all methods are expected to converge to the same solution \hat{x} .

Figure 1 shows the result for a five times overdetermined and consistent system without noise. Note that the usual RK performs consistently well over all trials, while the performance of RSK and ERSK differs drastically between different instances. As denoted by the quantiles, there are a few instances on which RSK and ERSK are remarkably fast, especially for the exact-step method, while for other instance they are rather slow. Also, the asymptotic linear rate of the medians is fastest for ERSK, and also RSK has a faster asymptotic rate than non-sparse RK.

Figures 2 and 3 show the results for noisy right hand sides. Figure 2 uses a two times overdetermined system with 10% relative noise, Figure 3 has the same noise level and a five times overdetermined system. All methods consistently stagnate at a residual level which is comparable to the noise level, however, ERSK achieves this faster than RSK which in turn is faster than RK. Regarding the reconstruction error, ERSK and RK achieve reconstructions with an error in the size of the noise level, while SRK achieves an even lower reconstruction error. The last effect is not explained by our theory. On an intuitive level one may argue that the Sparse Kaczmarz method obtains better reconstructions since it incorporates the sparsity of the solutions, but that the exact steps in the Sparse Kaczmarz method spoil this advantage by trying to fulfill all equations exactly, despite the noise. In fact, RSK with inexact stepsize may be seen as a kind of relaxed Kaczmarz method.



Figure 1: Experiment A: Comparison of Randomized Kaczmarz (black) Randomized Rparse Kaczmarz (red), and Exact-Step Randomized Sparse Kaczmarz (green), n = 200, m = 1000, sparsity s = 25, no noise. Left: Plots of relative residual ||Ax - b||/|b||, right: plots of error $||x - x^{\dagger}||/||x^{\dagger}||$. Thick line shows median over 60 trials, light area is between min and max, darker area indicate 25th and 75th quantile.



Figure 2: Experiment A: Comparison of Randomized Kaczmarz (black) Randomized Sparse Kaczmarz (red), and Exact-Step Randomized Sparse Kaczmarz (green), n = 200, m = 400, sparsity s = 25, 10% relative noise. Left: Plots of relative residual $||Ax - b^{\delta}||/||b^{\delta}||$, right: plots of error $||x - x^{\dagger}||/||x^{\dagger}||$. Thick line shows median over 60 trials, light area is between min and max, darker area indicate 25th and 75th quantile.



Figure 3: Experiment A: Comparison of Randomized Kaczmarz (black) Randomized Sparse Kaczmarz (red), and Exact-Step Randomized Sparse Kaczmarz (green), n = 200, m = 1000, sparsity s = 25, 10% relative noise. Left: Plots of relative residual $||Ax - b^{\delta}||/||b^{\delta}||$, right: plots of error $||x - x^{\dagger}||/||x^{\dagger}||$. Thick line shows median over 60 trials, light area is between min and max, darker area indicate 25th and 75th quantile.

6.2 Sparse cyclic vs. Randomized Sparse Kaczmarz

To investigate the impact of randomization within the Sparse Kaczmarz framework, we studied an academic tomography problem. We used the AIRtools toolbox [22] to create CT-measurement matrices of different sizes. We used fanbeam geometry throughout and worked with overdetermined systems, sparse solutions and noisefree right hand sides. We compared RSK with the cyclic version of the Sparse Kaczmarz method, where we process the rows of the linear system in their "natural" order. Figure 4 shows the result for a small problem with n = 100 pixels, and Figure 5 shows the result for a problem with n = 900pixels. In both cases the randomization shows improvements for the median as well as for the extreme cases.

7 Conclusion

Using error bounds and the theoretical framework of Bregman projections for split feasibility problems, we proved expected linear convergence for the Randomized Sparse Kaczmarz method. Numerical experiments confirm the linear convergence and demonstrate the benefit of using the method to recover sparse solutions of linear systems, even in the overdetermined case. However, we could not explicitly quantify the linear rate in terms of the problem data, as for the standard Randomized Kaczmarz method. The contraction constants q in Theorem 4.5 and Corollary 5.1 depend on quantities which are not easily accessible, like the constants L from Theorem 3.5 and γ from Theorem 3.10.

As demonstrated in [27] the presented framework also allows for numerous generalizations which we did not further pursue here. For example, in the presence of noise we could replace equality constraints $\langle a_i, x \rangle = b_i$ by inequalities



Figure 4: Experiment B: Sparse Kaczmarz (blue) vs. Sparse Randomized Kaczmarz (red), n = 100, m = 1164, sparsity s = 20. Left: Plots of relative residual ||Ax - b||/||b||, right: plots of error $||x - x^{\dagger}||/||x^{\dagger}||$. Thick line shows median over 40 trials, light area is between min and max, darker area indicate 25th and 75th quantile.



Figure 5: Experiment B: Sparse Kaczmarz (blue) vs. Sparse Randomized Kaczmarz (red), n = 900, m = 3660, sparsity s = 180. Left: Plots of relative residual ||Ax - b||/||b||, right: plots of error $||x - x^{\dagger}||/||x^{\dagger}||$. Thick line shows median over 40 trials, light area is between min and max, darker area indicate 25th and 75th quantile.

 $|\langle a_i, x \rangle - b_i| \leq \delta_i$ to reflect an error estimate for each measurement. Algorithms 2 and 3 would only have to be changed slightly by projecting onto the modified hyperplanes $H_{\leq}(a_i, b_i + \delta_i)$ or $H_{\leq}(-a_i, -b_i + \delta_i)$, and we still obtain linear convergence.

Let us remark that, motivated by the excellent performance of the Randomized Sparse Kaczmarz method, we also tried to solve the regularized nuclear norm problem (5) by applying a randomized Kaczmarz iteration of the form (4). Somewhat disappointingly, our preliminary numerical experiments indicated that this unduly increases the number of times we have to perform the expensive singular value thresholding. It would be interesting to know if the use of low-rank matrices A_i in (4) allows for more efficient updates of $S_\lambda(X_k^*)$ to compensate for this. A possible approach could be to use low-rank modifications of the singular value decomposition of the dual iterates $X_{k+1}^* = X_k^* - t_k \cdot A_i$ as shown in [7].

References

- A. Agaskar, C. Wang, and Y. M. Lu. Randomized Kaczmarz algorithms: Exact MSE analysis and optimal sampling probabilities. In *IEEE Global* Conference on Signal and Information Processing (GlobalSIP), 2015.
- [2] Y. Alber and D. Butnariu. Convergence of Bregman projection methods for solving consistent convex feasibility problems in reflexive Banach spaces. *Journal of Optimization Theory and Applications*, 92(1):33–61, 1997.
- [3] H. H. Bauschke and J. M. Borwein. On projection algorithms for solving convex feasibility problems. SIAM Review, 38(3):367–426, 1996.
- [4] H. H. Bauschke and J. M. Borwein. Legendre functions and the method of random Bregman projections. *Journal of Convex Analysis*, 4(1):27–67, 1997.
- [5] H. H. Bauschke, J. M. Borwein, and P. L. Combettes. Bregman monotone optimization algorithms. SIAM J. Control Optim., 42(2):596–636, 2003.
- [6] H. H. Bauschke, J. M. Borwein, and W. Li. Strong conical hull intersection property, bounded linear regularity, Jameson's property (G), and error bounds in convex optimization. *Mathematical Programming*, 86(1):135–160, 1999.
- [7] M. Brand. Fast low-rank modifications of the thin singular value decomposition. *Linear algebra and its applications*, 415(1):20–30, 2006.
- [8] L. M. Bregman. The relaxation method for finding common points of convex sets and its application to the solution of problems in convex programming. USSR Computational Mathematics and Mathematical Physics, 7:200-217, 1967.

- [9] J. Briskman and D. Needell. Block Kaczmarz method with inequalities. Journal of Mathematical Imaging and Vision, pages 1–12, 2014.
- [10] M. Burger. Bregman distances in inverse problems and partial differential equations. In Advances in Mathematical Modeling, Optimization and Optimal Control, pages 3–33. Springer, 2016.
- [11] C. Byrne. Iterative oblique projection onto convex sets and the split feasibility problem. *Inverse Problems*, 18:441–453, 2002.
- [12] C. Byrne. A unified treatment of some iterative algorithms in signal processing and image reconstruction. *Inverse Problems*, 20:103–120, 2004.
- [13] C. Byrne and Y. Censor. Proximity function minimization using multiple Bregman projections, with applications to split feasibility and Kullback-Leibler distance minimization. Annals of Operations Research, 105:77–98, 2001.
- [14] J.-F. Cai, E. J. Candès, and Z. Shen. A singular value thresholding algorithm for matrix completion. SIAM Journal on Optimization, 20(4):1956– 1982, 2010.
- [15] J.-F. Cai, S. Osher, and Z. Shen. Convergence of the linearized Bregman iteration for ℓ_1 -norm minimization. *Math. Comp.*, 78:2127–2136, 2009.
- [16] Y. Censor and T. Elfving. A multiprojection algorithm using Bregman projections in a product space. *Numer. Algorithms*, 8:221–239, 1994.
- [17] Y. Censor, T. Elfving, N. Kopf, and T. Bortfeld. The multiple-sets split feasibility problem and its applications for inverse problems. *Inverse Prob*lems, 21:2071–2084, 2005.
- [18] S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. SIAM Journal on Scientific Computing, 20(1):33–61, 1998.
- [19] X. Chen and A. M. Powell. Almost sure convergence of the Kaczmarz algorithm with random measurements. *Journal of Fourier Analysis and Applications*, 18(6):1195–1214, 2012.
- [20] F. Deutsch and H. Hundal. The rate of convergence for the method of alternating projections, ii. Journal of Mathematical Analysis and Applications, 205(2):381–405, 1997.
- [21] M. Elad. Sparse and redundant representations: from theory to applications in signal and image processing. Springer, 2010.
- [22] P.C. Hansen and M. Saxild-Hansen. AIR Tools A MATLAB package of algebraic iterative reconstruction methods. *Journal of Computational and Applied Mathematics*, 236(8):2167–2178, 2012.

- [23] A. J. Hoffman. On approximate solutions of systems of linear inequalities. Journal of Research of the National Bureau of Standards, 49(4):263–265, 1952.
- [24] S. Kaczmarz. Angenäherte Auflösung von Systemen linearer Gleichungen. Bull. Internat. Acad. Polon. Sci. Lettres A, pages 355–357, 1937.
- [25] M. J. Lai and W. Yin. Augmented ℓ_1 and nuclear-norm models with a globally linearly convergent algorithm. *SIAM J. Imaging Sci.*, 6(2):1059–1091, 2013.
- [26] D. Leventhal and A. S. Lewis. Randomized methods for linear constraints: convergence rates and conditioning. *Mathematics of Operations Research*, 35(3):641–654, 2010.
- [27] D. A. Lorenz, F. Schöpfer, and S. Wenger. The linearized Bregman method via split feasibility problems: Analysis and generalizations. SIAM J. Imaging Sciences, 7(2):1237–1262, 2014.
- [28] D. A. Lorenz, S. Wenger, F. Schöpfer, and M. Magnor. A sparse Kaczmarz solver and a linearized Bregman method for online compressed sensing. In 2014 IEEE International Conference on Image Processing (ICIP), pages 1347–1351. IEEE, 2014.
- [29] H. Mansour and O. Yilmaz. A fast randomized Kaczmarz algorithm for sparse solutions of consistent linear systems. arXiv preprint arXiv:1305.3803, 2013.
- [30] D. Needell, N. Srebro, and R. Ward. Stochastic gradient descent, weighted sampling, and the randomized Kaczmarz algorithm. *Mathematical Pro*gramming, 155(1):549–573, 2016.
- [31] D. Needell and J. A. Tropp. Paved with good intentions: Analysis of a randomized block Kaczmarz method. *Linear Algebra and its Applications*, 441:199–221, 2014.
- [32] Y. Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. SIAM Journal on Optimization, 22(2):341–362, 2012.
- [33] S. Petra. Randomized sparse block Kaczmarz as randomized dual blockcoordinate descent. Analele Stiintifice Ale Universitatii Ovidius Constanta-Seria Matematica, 23(3):129–149, 2015.
- [34] B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3):471–501, 2010.
- [35] P. Richtárik and M. Takáč. Iteration complexity of randomized blockcoordinate descent methods for minimizing a composite function. *Mathematical Programming*, 144(1-2):1–38, 2014.

- [36] S. M. Robinson. Some continuity properties of polyhedral multifunctions. Mathematical Programming Study, 14:206–214, 1981.
- [37] R. T. Rockafellar and R. J.-B. Wets. Variational Analysis. Springer, Berlin, 2009.
- [38] F. Schöpfer. Exact regularization of polyhedral norms. SIAM J. Optim., 22(4):1206–1223, 2012.
- [39] F. Schöpfer. Linear convergence of descent methods for the unconstrained minimization of restricted strongly convex functions. SIAM J. Optim., 26(3):1883–1911, 2016.
- [40] F. Schöpfer, T. Schuster, and A. K. Louis. An iterative regularization method for the solution of the split feasibility problem in Banach spaces. *Inverse Problems*, 24, 2008.
- [41] T. Strohmer and R. Vershynin. A randomized Kaczmarz algorithm with exponential convergence. *Journal of Fourier Analysis and Applications*, 15(2):262–278, 2009.
- [42] M. Wang and D. P. Bertsekas. Stochastic first-order methods with random constraint projection. SIAM Journal on Optimization, 26(1):681–717, 2016.
- [43] H. Zhang, J. F. Hui Cai, L. Cheng, and J. Zhu. Strongly convex programming for exact matrix completion and robust principal component analysis. *Inverse Problems and Imaging*, 6(2):357–372, 2012.
- [44] J. Zhao and Q. Yang. Several solution methods for the split feasibility problem. *Inverse Problems*, 21:1791–1799, 2005.
- [45] A. Zouzias and N. M. Freris. Randomized extended Kaczmarz for solving least squares. SIAM Journal on Matrix Analysis and Applications, 34(2):773–793, 2013.