CARL VON OSSIETZKY universität OLDENBURG

IfM INSTITUT FÜR MATHEMATIK

Linear convergence of the Randomized Sparse Kaczmarz Method *

Frank Schöpfer[1], Dirk A. Lorenz[2]

IfM Preprint No. 2018-01
September 2018

Institut für Mathematik
Carl von Ossietzky Universität Oldenburg
D-26111 Oldenburg, Germany

*Published in *method. Mathematical Programming, 2018.* DOI: 10.1007/s10107-017-1229-1
[1]frank.schoepfer@uni-oldenburg.de
[2]d.lorenz@tu-braunschweig.de

Noname manuscript No.
(will be inserted by the editor)

Linear convergence of the Randomized Sparse Kaczmarz Method

Frank Schöpfer · Dirk A. Lorenz

the date of receipt and acceptance should be inserted later

Abstract The randomized version of the Kaczmarz method for the solution of consistent linear systems is known to converge linearly in expectation. And even in the possibly inconsistent case, when only noisy data is given, the iterates are expected to reach an error threshold in the order of the noise-level with the same rate as in the noiseless case. In this work we show that the same also holds for the iterates of the recently proposed Randomized Sparse Kaczmarz method for recovery of sparse solutions. Furthermore we consider the more general setting of convex feasibility problems and their solution by the method of randomized Bregman projections. This is motivated by the observation that, similarly to the Kaczmarz method, the Sparse Kaczmarz method can also be interpreted as an iterative Bregman projection method to solve a convex feasibility problem. We obtain expected sublinear rates for Bregman projections with respect to a general strongly convex function. Moreover, even linear rates are expected for Bregman projections with respect to smooth or piecewise linear-quadratic functions, and also the regularized nuclear norm, which is used in the area of low rank matrix problems.

Keywords randomized Kaczmarz method, linear convergence, Bregman projections, sparse solutions, split feasibility problem, error bounds

Mathematics Subject Classification (2000) 65F10, 68W20, 90C25

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.

Frank Schöpfer
Institut für Mathematik, Carl von Ossietzky Universität Oldenburg, 26111 Oldenburg, Germany,
E-mail: frank.schoepfer@uni-oldenburg.de

Dirk A. Lorenz
Institute for Analysis and Algebra, TU Braunschweig, 38092 Braunschweig, Germany,
E-mail: d.lorenz@tu-braunschweig.de, Tel.: +49-531-391-7423, Fax: +49-531-391-7414

1 Introduction

In this paper we analyse a randomized variant of the recently proposed *Sparse Kaczmarz method* [28, 29] to recover sparse solutions of linear systems. Let $A \in \mathbb{R}^{m\times n}$ be a matrix with rows $0 \neq a_i^T \in \mathbb{R}^n$ and $b \in \mathbb{R}^m$ be such that the linear system $Ax = b$ is consistent. For the standard Kaczmarz method [25] one goes through the indices of the rows cyclically, and projects a given iterate onto the solution space of this row. For $i = \mathrm{mod}(k-1, m) + 1$ the method iterates

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

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. [21]. 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 [43] it has been shown that a choice of row i with probability $p_i = \|a_i\|_2^2/\|A\|_F^2$ leads to a linear convergence rate in expectation,

$$\mathbb{E}\left[\|x_{k+1} - \hat{x}\|_2^2\right] \leq \left(1 - \tfrac{1}{\kappa^2}\right) \cdot \mathbb{E}\left[\|x_k - \hat{x}\|_2^2\right]\,, \tag{2}$$

where $\kappa = \frac{\|A\|_F}{\sigma_{\min}(A)}$, $\|A\|_F$ is the Frobenius norm and $\sigma_{\min}(A)$ denotes the smallest positive singular value of A. The result was obtained for a consistent overdetermined system $Ax = b$ with a full rank matrix A. But even without the assumption of full rank, and in the possibly inconsistent case, when only a noisy right hand side b^δ is given with $\|b - b^\delta\|_2 \leq \delta$, the iterates are expected to reach an error threshold in the order of the noise-level with the same rate as in the noiseless case, cf. [31, 49],

$$\mathbb{E}\left[\|x_k - \hat{x}\|_2\right] \leq \left(1 - \tfrac{1}{\kappa^2}\right)^{\frac{k}{2}} \cdot \|\hat{x}\|_2 + \tfrac{\delta}{\sigma_{\min}(A)}\,.$$

Since then similar results have been achieved for randomized Block Kaczmarz methods and systems of equalities and inequalities, see [9, 27, 33], and connections to stochastic gradient descent have been drawn [32].

In [28, 29] a variant of the Kaczmarz method has been proposed that produces sparse solutions. This *Sparse Kaczmarz method* uses two variables and reads as

$$\begin{aligned} x_{k+1}^* &= x_k^* - \frac{\langle a_i\,,\, x_k\rangle - b_i}{\|a_i\|_2^2} \cdot a_i \\ x_{k+1} &= S_\lambda(x_{k+1}^*) \end{aligned} \tag{3}$$

with $\lambda > 0$ and the soft shrinkage function $S_\lambda(x) = \max\{|x| - \lambda, 0\} \cdot \mathrm{sign}(x)$. It has been shown in [28] that for a consistent system $Ax = b$ with an arbitrary matrix A the iterates x_k converge to the unique solution $\hat{x}$ of the *regularized Basis Pursuit problem*,

$$\min_{x\in\mathbb{R}^n} f(x) = \lambda\|x\|_1 + \tfrac{1}{2}\|x\|_2^2 \quad s.t. \quad Ax = b\,, \tag{4}$$

see e.g. [15, 19, 22], and also [40] for explicit values of $\lambda > 0$ that guarantee exact recovery of sparse solutions. But no convergence rate has been given. In [35], in the noiseless case, 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 objective function

$$g(y) = \tfrac{1}{2} \cdot \|S_\lambda(A^T y)\|_2^2 - \langle b\,,\, y\rangle \tag{5}$$

of the unconstrained dual of (4), see also [34, 44]. However, the rates given in [35] are in terms of the dual objective function g, and not of the primal iterates only, although, as mentioned there in the conclusions, the experimental results indicate that such rates also hold for the primal iterates. Furthermore, linear convergence could only be obtained by smoothing the primal objective function f in (4), which results in an iteration that is slightly different from (3), and need not solve (4).

Here we show that in the noiseless case the *Randomized Sparse Kaczmarz method* in fact converges linearly in expectation without smoothing. And in the noisy case, similarly to the Randomized Kaczmarz method, the iterates are expected to reach an error threshold in the order of the noise-level with the same rate as in the noiseless case. The proof is mainly based on two observations: First, linear rates can be achieved because g is *restricted strongly convex*, cf. [26, 41, 47]. Second, using the notion of *Bregman distance* with respect to f as in [28] easily allows us to express the rates in terms of the primal iterates only. Concretely, let $\operatorname{supp}(\hat{x}) = \{j \in \{1,\dots,n\} \,|\, \hat{x}_j \neq 0\}$, denote by A_J the matrix that is formed by the columns of A indexed by J, define

$$\tilde{\sigma}_{\min}(A) = \min\{\sigma_{\min}(A_J) \mid J \subset \{1,\dots,n\},\ A_J \neq 0\}, \tag{6}$$

and set $\tilde{\kappa} = \frac{\|A\|_F}{\tilde{\sigma}_{\min}(A)}$. In case $b \neq 0$ we also have $\hat{x} \neq 0$ and hence

$$|\hat{x}|_{\min} = \min\{|\hat{x}_j| \,|\, j \in \operatorname{supp}(\hat{x})\} > 0\,. \tag{7}$$

If row i is chosen with probability $p_i = \|a_i\|_2^2/\|A\|_F^2$, then the iterates of (3) fulfill

$$\mathbb{E}\left[\|x_k - \hat{x}\|_2\right] \leq \left(1 - \tfrac{1}{\tilde{\kappa}^2} \cdot \tfrac{|\hat{x}|_{\min}}{2|\hat{x}|_{\min}+4\lambda}\right)^{\frac{k}{2}} \cdot \sqrt{2\lambda\|\hat{x}\|_1 + \|\hat{x}\|_2^2} + \sqrt{\tfrac{2|\hat{x}|_{\min}+4\lambda}{|\hat{x}|_{\min}}} \cdot \tfrac{\delta}{\tilde{\sigma}_{\min}(A)}\,.$$

The values of $\tilde{\sigma}_{\min}(A)$ and $|\hat{x}|_{\min}$ are those used in [26] to quantify the linear convergence rate for the linearized Bregman method applied to (4).

Furthermore, we extend these results to a more general setting by using the theoretical framework developed in [28]. There the solution $\hat{x}$ of (4) is considered as a solution to the *convex feasibility problem* (CFP)

$$\text{find} \quad x \in C := \bigcap_{i=1}^{m} C_i \tag{8}$$

with the hyperplanes $C_i = \{x \in \mathbb{R}^n \,|\, \langle a_i\,,\, x\rangle = b_i\}$. The Sparse Kaczmarz method is then interpreted as an iterative projection method to solve (8),

where in each iteration, instead of orthogonal projections, *Bregman projections* with respect to f onto the sets C_i are employed. In this context the iteration with exact Bregman projections reads as

$$\begin{aligned} x_{k+1}^* &= x_k^* - t_k \cdot a_i \\ x_{k+1} &= S_\lambda(x_{k+1}^*)\,, \end{aligned}$$

where t_k is obtained by an exact linesearch procedure, and the choice $t_k = \frac{\langle a_i\,,\,x_k\rangle - b_i}{\|a_i\|_2^2}$ as in (3) corresponds to an inexact linesearch or relaxed Bregman projection. The CFP-framework is quite flexible and allows to include other convex constraints like inequalities $C_i = \{x \in \mathbb{R}^n \mid \langle a_i\,,\,x\rangle \leq b_i\}$. We show sublinear convergence rates in expectation for the *method of randomized Bregman projections* to solve a CFP, where in each iteration an index i is chosen with some probability $p_i > 0$, and a Bregman projection with respect to a general strongly convex function f onto C_i is employed. As in [3] for the case of orthogonal projections, these results are proven with *error bounds* that hold under the assumption of *bounded linear regularity* of the collection $\{C_1, \dots C_m\}$. Moreover, we derive sufficient conditions which ensure even linear convergence rates. Especially, based on the recent results of [41], we get linear rates for any *piecewise linear-quadratic* f, and also randomized iterations of the form

$$\begin{aligned} X_{k+1}^* &= X_k^* - \tfrac{\langle A_i\,,\,X_k\rangle - b_i}{\|A_i\|_F^2} \cdot A_i \\ X_{k+1} &= S_\lambda(X_{k+1}^*) \end{aligned} \tag{9}$$

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

$$\min_{X \in \mathbb{R}^{n_1 \times n_2}} f(X) = \lambda \|X\|_* + \tfrac{1}{2}\|X\|_F^2 \quad s.t. \quad \langle A_i\,,\,X\rangle = b_i\,,\; i = 1, \dots, m\,, \tag{10}$$

where $\langle A\,,\,X\rangle = \operatorname{trace}(A^T \cdot X)$ for two matrices $A, X \in \mathbb{R}^{n_1 \times n_2}$, and $S_\lambda(X)$ denotes the *singular value thresholding* operator, see e.g. [14, 26, 36, 46].

In the next section we recall the basic properties of Bregman distances and Bregman projections. The linear convergence rates for the Randomized Sparse Kaczmarz method are derived in section 3. In section 4 we prove the error bounds which are crucial for the convergence analysis in section 5, where we treat the general case of the solution of the CFP and related *split feasibility problems* (SFP) by the method of randomized Bregman projections. 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

At first we recall some well known concepts and properties of convex functions [39].

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 $x \in \mathbb{R}^n$,

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

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

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

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

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

When the concrete value of α is relevant we indicate this by saying that f is *α-strongly convex*.

Theorem 2.2 ([39, Proposition 12.60]) *If $f : \mathbb{R}^n \to \mathbb{R}$ is α-strongly convex then the conjugate function $f^*(x^*) := \sup_{x \in \mathbb{R}^n} \langle x^* \,,\, x\rangle - f(x)$ is differentiable with a $1/\alpha$-Lipschitz-continuous gradient, i.e.*

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

Example 2.3 The objective function $f(x) = \lambda\|x\|_1 + \frac{1}{2}\|x\|_2^2$ in (4) is 1-strongly convex with $\partial f(x) = \{x + \lambda \cdot s \,|\, s_j = \operatorname{sign}(x_j) \text{ if } x_j \neq 0, \text{ and } s_j \in [-1, 1] \text{ if } x_j = 0\}$, $f^*(x^*) = \frac{1}{2}\|S_\lambda(x^*)\|_2^2$ and $\nabla f^*(x^*) = S_\lambda(x^*)$, cf. [45].

2.1 Bregman distance

The concept of Bregman distance goes back to Bregman [8] and since then has successfully been used as a tool to analyse and design optimization algorithms, see e.g. [2, 4, 10, 13, 28, 42].

Definition 2.4 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

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

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

For $f(x) = \frac{1}{2}\|x\|_2^2$ we just have $D_f(x, y) = \frac{1}{2}\|x - y\|_2^2$. For the objective function in (4) a short reformulation yields the following.

Example 2.5 For $f(x) = \lambda\|x\|_1 + \frac{1}{2}\|x\|_2^2$ and any $x^* = x + \lambda \cdot s \in \partial f(x)$ we have

$$D_f^{x^*}(x, y) = \frac{1}{2}\|x - y\|_2^2 + \lambda \cdot (\|y\|_1 - \langle s \,,\, y\rangle)\,.$$

In the following lemma we state the key properties of the Bregman distance that are needed for the convergence analysis of the randomized methods. They immediately follow from the assumption of strong convexity, cf. [28].

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*

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

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 $\Pi_C^{x^*}(x) \in C$ such that

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

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 $\operatorname{dist}_f(x,C)^2$ and the usual $\operatorname{dist}(x,C)^2$ 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 ([28, 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*

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

$$D_f^{\hat{x}^*}(\hat{x}, y) \leq D_f^{x^*}(x,y) - D_f^{x^*}(x,\hat{x}) \quad \textit{for all} \quad y \in C\,.$$

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

The next lemma shows that Bregman projections onto affine subspaces and half-spaces can be computed by solving unconstrained optimization problems involving the differentiable conjugate function f^*.

Lemma 2.9 ([28, Lemma 2.4]) *Let $f : \mathbb{R}^n \to \mathbb{R}$ be α-strongly convex, $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, $u \in \mathbb{R}^n$ and $\beta \in \mathbb{R}$.*

(a) The Bregman projection of $x \in \mathbb{R}^n$ onto the affine subspace $L(A,b) := \{x \in \mathbb{R}^n \,|\, Ax = b\} \neq \emptyset$ *is*

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

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, $\hat{x}^* := x^* - A^T\hat{w}$ is an admissible subgradient for $\hat{x}$ according to Lemma 2.8. If A has full row rank then for all $y \in L(A,b)$ we have

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

(b) The Bregman projection of $x \in \mathbb{R}^n$ onto the hyperplane $H(u,\beta) := \{x \in \mathbb{R}^n \,|\, \langle u\,,\, x\rangle = \beta\}$ *with $u \neq 0$ is*

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

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

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

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

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

If x is not in the half-space $H_{\leq}(u,\beta) := \{x \in \mathbb{R}^n \,|\, \langle u\,,\, x\rangle \leq \beta\}$ *then we necessarily have $\hat{t} > 0$, $\Pi^{x^*}_{H_{\leq}(u,\beta)}(x) = \hat{x}$ and the above inequality holds for all $y \in H_{\leq}(u,\beta)$.*

3 Linear convergence of the Randomized Sparse Kaczmarz method

Here we show expected linear convergence for the Randomized Sparse Kaczmarz method to solve the regularized Basis Pursuit problem (4) with objective function

$$f(x) = \lambda\|x\|_1 + \tfrac{1}{2}\|x\|_2^2\,,$$

where we assume that $b \neq 0$ is in the range $\mathcal{R}(A)$ of A, and hence (4) has a unique solution $\hat{x} \neq 0$. Although linear convergence also follows from the general result in section 5, the short proof given here illustrates well the main ideas used to prove the general case, where most constants to quantify the rates

Algorithm 1 Randomized Sparse Kaczmarz method (RaSK)

Input: starting points $x_0 = x_0^* = 0 \in \mathbb{R}^n$, matrix $A \in \mathbb{R}^{m \times n}$ with rows $0 \neq a_i^T \in \mathbb{R}^n$, and vector $b \in \mathbb{R}^m$
Output: (approximate) 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 $i_k = i \in \{1, \dots, m\}$ at random with probability $p_i = \|a_i\|_2^2 / \|A\|_F^2$
4: update $x_{k+1}^* = x_k^* - \frac{\langle a_{i_k}, x_k \rangle - b_{i_k}}{\|a_{i_k}\|_2^2} \cdot a_{i_k}$
5: update $x_{k+1} = S_\lambda(x_{k+1}^*)$
6: increment $k = k + 1$
7: **until** a stopping criterion is satisfied

are only given implicitly. The Randomized Sparse Kaczmarz method (RaSK) with stepsize $t_k = \frac{\langle a_i, x_k \rangle - b_i}{\|a_i\|_2^2}$ is stated here as Algorithm 1.

The Exact-Step Randomized Sparse Kaczmarz (ERaSK) method with exact linesearch corresponding to a Bregman projection onto the hyperplane $H(a_i, b_i)$ is stated as Algorithm 2. An exact linesearch is indeed computationally feasible, since in this case the derivative of the one-dimensional objective function in line 4 of Algorithm 2 is piecewise linear, see [28, Section 2.5.2]. Computing the "kinks" to determine the linear pieces and then the optimal value can be done with at most $12n$ floating point operations and an $\mathcal{O}\big(n \cdot \ln(n)\big)$-sorting procedure. We provide a corresponding MATLAB code in the complementary material.

Algorithm 2 Exact-Step Randomized Sparse Kaczmarz method (ERaSK)

Input: starting points $x_0 = x_0^* = 0 \in \mathbb{R}^n$, matrix $A \in \mathbb{R}^{m \times n}$ with rows $0 \neq a_i^T \in \mathbb{R}^n$, and vector $b \in \mathbb{R}^m$
Output: (approximate) 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 $i_k = i \in \{1, \dots, m\}$ at random with probability $p_i = \|a_i\|_2^2 / \|A\|_F^2$
4: calculate $t_k = \operatorname{argmin}_{t \in \mathbb{R}} f^*(x_k^* - t \cdot a_{i_k}) + t \cdot b_{i_k}$
5: update $x_{k+1}^* = x_k^* - t_k \cdot a_{i_k}$
6: update $x_{k+1} = S_\lambda(x_{k+1}^*)$
7: increment $k = k + 1$
8: **until** a stopping criterion is satisfied

As noted in the introduction, linear convergence follows from the restricted strong convexity of the dual objective function (5), which yields the following error bound.

Lemma 3.1 *Let $\tilde{\sigma}_{\min}(A)$ and $|\hat{x}|_{\min}$ be as defined in (6) and (7), respectively. Then for all $x \in \mathbb{R}^n$ with $\partial f(x) \cap \mathcal{R}(A^T) \neq \emptyset$ and all $x^* = A^T y \in \partial f(x) \cap \mathcal{R}(A^T)$ we have*

$$D_f^{x^*}(x, \hat{x}) \leq \frac{1}{\tilde{\sigma}_{\min}^2(A)} \cdot \frac{|\hat{x}|_{\min} + 2\lambda}{|\hat{x}|_{\min}} \cdot \|Ax - b\|_2^2 .$$

Proof The proof is based on the results of [26]. There problem (4) is equivalently formulated with primal objective function $\tilde{f}(x) = \frac{1}{\lambda} \cdot f(x)$, and hence the dual objective function is $\tilde{g}(y) = \frac{\lambda}{2} \cdot \|S_1(A^T y)\|_2^2 - \langle b\,,\, y\rangle$. In Lemma 7 of [26] it was shown that $\tilde{g}$ is restricted strongly convex: Let $\tilde{Y}$ denote the set of minimizers of $\tilde{g}$. Then for all $y \in \mathbb{R}^m$ we have

$$\langle y - P_{\tilde{Y}}(y)\,,\, \nabla\tilde{g}(y)\rangle \geq \tilde{\sigma}_{\min}^2(A) \cdot \tfrac{\lambda\cdot|\hat{x}|_{\min}}{|\hat{x}|_{\min}+2\lambda} \cdot \|y - P_{\tilde{Y}}(y)\|_2^2\,.$$

We just have to reformulate this result for g in (5). Because of the relation $S_\lambda(\lambda \cdot A^T y) = \lambda \cdot S_1(A^T y)$ we have $\nabla g(\lambda \cdot y) = \nabla\tilde{g}(y)$. Hence the set of minimizers $\hat{Y}$ of g and $\tilde{Y}$ are related by $\hat{Y} = \lambda \cdot \tilde{Y}$, and we have $P_{\hat{Y}}(\lambda \cdot y) = \lambda \cdot P_{\tilde{Y}}(y)$. From this observation we immediately infer the estimate

$$\langle y - P_{\hat{Y}}(y)\,,\, \nabla g(y)\rangle \geq \tilde{\sigma}_{\min}^2(A) \cdot \tfrac{|\hat{x}|_{\min}}{|\hat{x}|_{\min}+2\lambda} \cdot \|y - P_{\hat{Y}}(y)\|_2^2\,.$$

For $x^* = A^T y \in \partial f(x) \cap \mathcal{R}(A^T)$ and $x = S_\lambda(x^*)$ this yields

$$\tilde{\sigma}_{\min}(A) \cdot \tfrac{|\hat{x}|_{\min}}{|\hat{x}|_{\min}+2\lambda} \cdot \|y - P_{\hat{Y}}(y)\|_2 \leq \|\nabla g(y)\|_2 = \|Ax - b\|_2\,.$$

Finally, with $\hat{x}^* := A^T P_{\hat{Y}}(y) \in \partial f(\hat{x})$ and Lemma 2.6 we can estimate

$$D_f^{x^*}(x, \hat{x}) \leq \langle x^* - \hat{x}^*\,,\, x - \hat{x}\rangle = \langle y - P_{\hat{Y}}(y)\,,\, Ax - b\rangle \leq \|y - P_{\hat{Y}}(y)\|_2 \cdot \|Ax - b\|_2\,,$$

from which the assertion follows. □

Now we can prove the main theorems of the article.

Theorem 3.2 (noiseless case) *The iterates x_k of both the RaSK method from Algorithm 1 and the ERaSK method from Algorithm 2 converge in expectation to the unique solution $\hat{x}$ of the regularized Basis Pursuit problem* (4) *with a linear rate, namely with $\tilde{\kappa} = \frac{\|A\|_F}{\tilde{\sigma}_{\min}(A)}$ and contraction factor*

$$q = 1 - \tfrac{1}{\tilde{\kappa}^2} \cdot \tfrac{1}{2} \cdot \tfrac{|\hat{x}|_{\min}}{|\hat{x}|_{\min}+2\lambda} \tag{11}$$

it holds that

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

and

$$\mathbb{E}\left[\|x_k - \hat{x}\|_2\right] \leq q^{\frac{k}{2}} \cdot \sqrt{2\lambda\|\hat{x}\|_1 + \|\hat{x}\|_2^2}\,,$$

where the expectation is taken with respect to the probability distribution $p_i = \|a_i\|_2^2/\|A\|_F^2$.

Proof By Theorem 2.8 of [28] the following estimate from Lemma 2.9 (b) holds for both the exact and the inexact stepsize,

$$D_f^{x_{k+1}^*}(x_{k+1}, \hat{x}) \leq D_f^{x_k^*}(x_k, \hat{x}) - \tfrac{1}{2} \cdot \tfrac{(\langle a_{i_k}\,, x_k\rangle - b_{i_k})^2}{\|a_{i_k}\|_2^2}\,. \tag{12}$$

For the moment we fix the values of the indices $i_0, \dots, i_{k-1}$ and consider only i_k as a random variable with values in $\{1, \dots, m\}$. Taking the expectation on both sides of (12) conditional to the values of the indices $i_0, \dots, i_{k-1}$ yields

$$\begin{aligned}\mathbb{E}\left[D_f^{x_{k+1}^*}(x_{k+1}, \hat{x}) \,\middle|\, i_0, \dots, i_{k-1}\right] &\leq D_f^{x_k^*}(x_k, \hat{x}) - \sum_{i=1}^{m} \frac{\|a_i\|_2^2}{\|A\|_F^2} \cdot \frac{1}{2} \cdot \frac{(\langle a_i\,, x_k\rangle - b_i)^2}{\|a_i\|_2^2} \\ &= D_f^{x_k^*}(x_k, \hat{x}) - \frac{1}{2} \cdot \frac{\|Ax_k - b\|_2^2}{\|A\|_F^2}\,. \end{aligned} \tag{13}$$

Together with Lemma 3.1 we get

$$\mathbb{E}\left[D_f^{x_{k+1}^*}(x_{k+1}, \hat{x}) \,\middle|\, i_0, \dots, i_{k-1}\right] \leq \left(1 - \frac{1}{\tilde{\kappa}^2} \cdot \frac{1}{2} \cdot \frac{|\hat{x}|_{\min}}{|\hat{x}|_{\min} + 2\lambda}\right) \cdot D_f^{x_k^*}(x_k, \hat{x})\,.$$

Now considering all indices $i_0, \dots, i_k$ as random variables with values in $\{1, \dots, m\}$, and taking the full expectation on both sides, yields the first estimate of the assertion. The second estimate then follows from Lemma 2.6 with $\alpha = 1$. □

Remark 3.3 (a) The contraction factor $q = 1 - \frac{1}{\tilde{\kappa}^2} \cdot \frac{1}{2} \cdot \frac{|\hat{x}|_{\min}}{|\hat{x}|_{\min} + 2\lambda}$ depends on $\hat{x}$. This reflects the fact that the dual objective is just restricted strongly convex for $\lambda > 0$. The dependence on $\hat{x}$ disappears in the case $\lambda = 0$ corresponding to the strongly convex $f(x) = \frac{1}{2}\|x\|_2^2$.

(b) By Example 2.5, if the iterates x_k are close enough to $\hat{x}$, such that the signs of the components of x_k and $\hat{x}$ coincide on $\operatorname{supp}(\hat{x})$, we actually have $D_f^{x_k^*}(x_k, \hat{x}) = \frac{1}{2}\|x_k - \hat{x}\|_2^2$. In this case we can remove the factor $\frac{1}{2}$ in (13) and consequently from q. Furthermore, if both $\hat{x}$ and x_k are sparse enough, instead of applying Lemma 3.1 to (13), we may use a *restricted isometry constant* [16], i.e. the smallest constant δ_r such that

$$(1 - \delta_r) \cdot \|x\|_2^2 \leq \|Ax\|_2^2 \leq (1 - \delta_r) \cdot \|x\|_2^2$$

for all $x \in \mathbb{R}^n$ with at most r nonzero components. This leads to the contraction factor $q = 1 - \frac{1-\delta_r}{\|A\|_F^2}$, which is possibly smaller.

(c) If A has full column rank, we have $\tilde{\sigma}_{\min}(A) = \sigma_{\min}(A)$. Hence for $\lambda = 0$ we recover the rate (2) for the standard Randomized Kaczmarz method.

Theorem 3.4 (noisy case) *Assume that instead of exact data $b \in \mathcal{R}(A)$ only a noisy right hand side $b^\delta \in \mathbb{R}^m$ with $\|b^\delta - b\|_2 \leq \delta$ is given. If the iterates x_k of the RaSK method from Algorithm 1 and the ERaSK method from Algorithm 2 are computed with b replaced by b^δ, then, with the same contraction factor q from* (11) *as in the noiseless case, we have*

(a) for the RaSK method:

$$\mathbb{E}\left[\|x_k - \hat{x}\|_2\right] \leq q^{\frac{k}{2}} \cdot \sqrt{2\lambda\|\hat{x}\|_1 + \|\hat{x}\|_2^2} + \sqrt{\frac{2|\hat{x}|_{\min} + 4\lambda}{|\hat{x}|_{\min}}} \cdot \frac{\delta}{\tilde{\sigma}_{\min}(A)}\,.$$

(b) for the ERaSK method with $\|A\|_{1,2} := \sqrt{\sum_{i=1}^m \|a_i\|_1^2}$:

$$\mathbb{E}\left[\|x_k - \hat{x}\|_2\right] \leq q^{\frac{k}{2}} \cdot \sqrt{2\lambda\|\hat{x}\|_1 + \|\hat{x}\|_2^2} + \sqrt{\frac{2|\hat{x}|_{\min} + 4\lambda}{|\hat{x}|_{\min}}} \cdot \frac{\delta}{\tilde{\sigma}_{\min}(A)} \cdot \sqrt{1 + \frac{4\|A\|_{1,2}}{\delta}}\,.$$

Proof As in [31] we make use of the simple but important observation that $x_k^\delta := \hat{x} + \frac{b_{i_k}^\delta - b_{i_k}}{\|a_{i_k}\|_2^2} \cdot a_{i_k} \in H(a_{i_k}, b_{i_k}^\delta)$. Similar to (12) this allows us to estimate

$$D_f^{x_{k+1}^*}(x_{k+1}, x_k^\delta) \leq D_f^{x_k^*}(x_k, x_k^\delta) - \tfrac{1}{2} \cdot \frac{(\langle a_{i_k}, x_k\rangle - b_{i_k}^\delta)^2}{\|a_{i_k}\|_2^2},$$

which after a short reformulation is equivalent to

$$D_f^{x_{k+1}^*}(x_{k+1}, \hat{x}) \leq D_f^{x_k^*}(x_k, \hat{x}) - \tfrac{1}{2} \cdot \frac{(\langle a_{i_k}, x_k\rangle - b_{i_k}^\delta)^2}{\|a_{i_k}\|_2^2} + \langle x_{k+1}^* - x_k^*, x_k^\delta - \hat{x}\rangle.$$

In the RaSK method we have $x_{k+1}^* - x_k^* = -\frac{\langle a_{i_k}, x_k\rangle - b_{i_k}^\delta}{\|a_{i_k}\|_2} \cdot a_{i_k}$ and thus

$$\begin{aligned}\langle x_{k+1}^* - x_k^*, x_k^\delta - \hat{x}\rangle &= \tfrac{b_{i_k}^\delta - b_{i_k}}{\|a_{i_k}\|_2^2} \cdot \langle x_{k+1}^* - x_k^*, a_{i_k}\rangle \\ &= \tfrac{(b_{i_k}^\delta - b_{i_k})^2}{\|a_{i_k}\|_2^2} - \tfrac{(b_{i_k}^\delta - b_{i_k})\cdot(\langle a_{i_k}, x_k\rangle - b_{i_k})}{\|a_{i_k}\|_2^2}.\end{aligned}$$

By rewriting

$$-\tfrac{1}{2}\cdot\tfrac{(\langle a_{i_k}, x_k\rangle - b_{i_k}^\delta)^2}{\|a_{i_k}\|_2^2} = -\tfrac{1}{2}\cdot\tfrac{(\langle a_{i_k}, x_k\rangle - b_{i_k})^2}{\|a_{i_k}\|_2^2} + \tfrac{(b_{i_k}^\delta - b_{i_k})\cdot(\langle a_{i_k}, x_k\rangle - b_{i_k})}{\|a_{i_k}\|_2^2} - \tfrac{1}{2}\cdot\tfrac{(b_{i_k}^\delta - b_{i_k})^2}{\|a_{i_k}\|_2^2}$$

we get

$$D_f^{x_{k+1}^*}(x_{k+1}, \hat{x}) \leq D_f^{x_k^*}(x_k, \hat{x}) - \tfrac{1}{2} \cdot \frac{(\langle a_{i_k}, x_k\rangle - b_{i_k})^2}{\|a_{i_k}\|_2^2} + \tfrac{1}{2} \cdot \frac{(b_{i_k}^\delta - b_{i_k})^2}{\|a_{i_k}\|_2^2}.$$

As in the proof for the noiseless case we get

$$\mathbb{E}\left[D_f^{x_{k+1}^*}(x_{k+1}, \hat{x})\right] \leq q \cdot \mathbb{E}\left[D_f^{x_k^*}(x_k, \hat{x})\right] + \tfrac{1}{2} \cdot \frac{\|b^\delta - b\|_2^2}{\|A\|_F^2}.$$

Inductively we infer that

$$\mathbb{E}\left[D_f^{x_k^*}(x_k, \hat{x})\right] \leq q^k \cdot (\lambda\|\hat{x}\|_1 + \tfrac{1}{2}\|\hat{x}\|_2^2) + \tfrac{1}{1-q} \cdot \tfrac{1}{2} \cdot \frac{\|b^\delta - b\|_2^2}{\|A\|_F^2}.$$

By Lemma 2.6 with $\alpha = 1$, and since $\sqrt{u+v} \leq \sqrt{u} + \sqrt{v}$, we get

$$\mathbb{E}\left[\|x_k - \hat{x}\|_2\right] \leq q^{\frac{k}{2}} \cdot \sqrt{2\lambda\|\hat{x}\|_1 + \|\hat{x}\|_2^2} + \sqrt{\tfrac{1}{1-q} \cdot \frac{\|b^\delta - b\|_2^2}{\|A\|_F^2}},$$

from which (a) follows. Now we turn to the ERaSK method. By Example 2.3 we have $x_{k+1}^* - x_k^* = (x_{k+1} - x_k) + (s_{k+1} - s_k)$ with $\|s_k\|_\infty, \|s_{k+1}\|_\infty \leq 1$. Since the exact linesearch guarantees $\langle x_{k+1}, a_{i_k}\rangle = b_{i_k}^\delta$, we get

$$\begin{aligned}\langle x_{k+1}^* - x_k^*, x_k^\delta - \hat{x}\rangle &= \tfrac{b_{i_k}^\delta - b_{i_k}}{\|a_{i_k}\|_2^2} \cdot \left(\langle x_{k+1} - x_k, a_{i_k}\rangle + \langle s_{k+1} - s_k, a_{i_k}\rangle\right) \\ &\leq \tfrac{(b_{i_k}^\delta - b_{i_k})^2}{\|a_{i_k}\|_2^2} - \tfrac{(b_{i_k}^\delta - b_{i_k})\cdot(\langle a_{i_k}, x_k\rangle - b_{i_k})}{\|a_{i_k}\|_2^2} + \tfrac{2|b_{i_k}^\delta - b_{i_k}|\cdot\|a_{i_k}\|_1}{\|a_{i_k}\|_2^2}.\end{aligned}$$

From this (b) follows analogously as for the RaSK method by using the estimate $\sum_{i=1}^m |b_i^\delta - b_i| \cdot \|a_i\|_1 \leq \|b^\delta - b\|_2 \cdot \|A\|_{1,2}$. □

Again we recover the result for the standard Randomized Kaczmarz method, because there the stepsize $t_k = \frac{\langle a_i\,,\,x_k\rangle - b_i}{\|a_i\|_2^2}$ in fact corresponds to an exact line-search. The error threshold for ERaSK is worse than the one for RaSK, which is also observed in our numerical experiments in Section 6. Please note that Theorem 3.4 tells us that RaSK and ERask are most useful for problems which are almost consistent and are only affected by moderately small noise. This is also the case for the standard Randomized Kaczmarz method, which aims at solving the constrained problem $\min_{x\in\mathbb{R}^n} \|x\|_2 \quad s.t. \quad Ax = b$. If one wishes to compute minimum 2-norm solutions to unconstrained least squares problems of the form $\min_{x\in\mathbb{R}^n} \|Ax - b\|_2$, then one has to modify the iterations of the standard Kaczmarz method appropriately, see [49]. We do not know yet how to modify RaSK and ERask so as to compute minimum 1-norm solutions to such least squares problems.

4 Bounded linear regularity and error bounds

In this section we derive some error bounds that we need for the analysis of the method of randomized Bregman projections to solve general convex feasibility problems. As in [3] for the case of orthogonal 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 $C \subset \mathbb{R}^n$, i.e. the interior of C relative to its affine hull.

Definition 4.1 Let $C_1, \dots C_m \subset \mathbb{R}^n$ be closed convex sets with nonempty intersection $C := \bigcap_{i=1}^m C_i$.

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

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

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

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

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

The standard constraint qualification generalizes the well known Slater condition for convex constraints, where $C_1, \dots, C_q$ are level sets of convex functions. By the next theorem it implies bounded regularity.

Theorem 4.2 (Corollary 3 and 6 in [6]) *If the collection $\{C_1, \dots C_m\}$ satisfies the standard constraint qualification then it is boundedly linearly regular. And if C is also bounded, then $\{C_1, \dots C_m\}$ is linearly regular.*

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

Lemma 4.3 *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 $y^* \in \partial f\big(P_C(x)\big)$ 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 $x \in \mathbb{R}^n$

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

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 $\|x^* - y^*\|_2 \leq L \cdot \|x - y\|_2$. However, we can achieve the better estimate for convex piecewise linear-quadratic f.

Definition 4.4 A convex function $f : \mathbb{R}^n \to \mathbb{R}$ is called *piecewise linear-quadratic* if there are finitely many polyhedral sets $F_i \subset \mathbb{R}^n$, $i \in I := \{1, \dots, p\}$, whose union equals $\mathbb{R}^n$, and relative to each of which $f(x)$ is given by a convex linear-quadratic function

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

with symmetric positive-semidefinite matrices $A_i \in \mathbb{R}^{n\times n}$, vectors $a_i \in \mathbb{R}^n$ and $\alpha_i \in \mathbb{R}$. Without loss of generality we may assume that all F_i have nonempty interior $\operatorname{int}(F_i)$ and that $\operatorname{int}(F_i) \cap \operatorname{int}(F_j) = \emptyset$ for $i \neq j$. For $x \in \mathbb{R}^n$ we define $I_f(x) := \{i \in I \,|\, x \in F_i\}$ 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. The next lemma characterizes the subdifferential of a convex piecewise linear-quadratic function.

Lemma 4.5 *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*

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

Proof The assertion about f^* follows from Theorem 11.14 in [39]. In its proof there is also given a characterisation of $\partial f(x)$, from which the above characterisation follows. But since the derivation would require some additional notation, and because we have not found this result explicitly stated elsewhere in the literature, we give a short proof here for convenience. The assertion is clear for $x \in \operatorname{int}(C_i)$ since f is differentiable on $\operatorname{int}(C_i)$ with $\nabla f(x) = A_i x + a_i$. By taking limits to a boundary point $x \in C_i$ for $i \in I_f(x)$, it follows that $A_i x + a_i \in \partial f(x)$ and thus $S_x := \operatorname{conv}\{A_i x + a_i \,|\, i \in I_f(x)\} \subset \partial f(x)$. Now suppose there exists some $x^* \in \partial f(x)$ such that $x^* \notin S_x$. Since S_x is closed

convex we can strictly separate x^* from S_x by a hyperplane, i.e. there are $u \in \mathbb{R}^n$ and $\beta \in \mathbb{R}$ such that

$$\langle u\,,\, x^*\rangle > \beta \geq \langle u\,,\, v\rangle \quad \text{for all } v \in \partial f(x)\,. \tag{14}$$

To x and u we find some $i \in I_f(x)$ such that $x + t \cdot u \in C_i$ for all $t > 0$ small enough. Since $A_i(x+t\cdot u)+a_i \in \partial f(x+t\cdot u)$ and the subdifferential mapping is monotone [39, Theorem 12.17], it follows that

$$0 \leq \langle A_i(x+t\cdot u)+a_i - x^*\,,\, (x+t\cdot u)-x\rangle = t^2\cdot\langle u\,,\, A_i u\rangle + t\cdot(\langle A_i x+a_i\,,\, u\rangle - \langle x^*\,,\, u\rangle)$$

We apply (14) to $v := A_i x + a_i \in \partial f(x)$ and get $0 < \langle u\,,\, x^*\rangle - \beta \leq t\cdot\langle u\,,\, A_i u\rangle$, which for $t \searrow 0$ leads to a contradiction. □

We also need the following lemma, which exploits the fact that the subgradients on the sets $F_x = \bigcap_{i\in I_f(x)} F_i$ are closely related.

Lemma 4.6 *Let $f : \mathbb{R}^n \to \mathbb{R}$ be strongly convex piecewise linear-quadratic and set $L_f := \max\{\|A_i\|_2 \,|\, i \in I\}$, cf. Definition 4.4. To $R > 0$ and a closed convex set $C \subset \mathbb{R}^n$ choose $c > 0$ such that $\|x^* - y^*\|_2 \leq c$ for all $x \in B_R$, $x^* \in \partial f(x)$ and $y^* \in \partial f\big(P_C(x)\big)$, and set*

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

Then for all $x \in B_R$ and $x^* \in \partial f(x)$ we have

$$\operatorname{dist}_f^{x^*}(x,C)^2 \leq \begin{cases} \frac{c}{d}\cdot\operatorname{dist}(x,C)^2 & , F_x \cap C = \emptyset \\ L_f\cdot\operatorname{dist}(x, F_x\cap C)^2 & , F_x \cap C \neq \emptyset \end{cases}.$$

Proof Since B_R is compact we have $\operatorname{dist}(B_R \cap F_x, C) > 0$ 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 indeed $d > 0$. Let $x \in B_R$ and $x^* \in \partial f(x)$. By Lemma 4.5 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 $\operatorname{dist}(x,C) \geq d$, and hence by Lemma 4.3 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^*. Hence we can estimate

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

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 4.7 *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 4.6 and Definition 4.1, because $\operatorname{dist}(x, F_x) = 0$. □

Remark 4.8 If C is polyhedral then by Theorem 4.2 all collections $\{F_x, C\}$ are boundedly linearly regular.

We also need the following generalization of Hoffmann's error bound [24] to possibly non-polyhedral sets, which are defined by convex constraints in the range $\mathcal{R}(A)$ of a matrix A.

Lemma 4.9 *Let the convex set $C \subset \mathbb{R}^n$ have the form $C = \{x \in \mathbb{R}^n \mid Ax \in Q\}$ 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*

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

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

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

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

$$\begin{aligned} \operatorname{dist}(x, C) &\leq \|x - P_{\hat{x}+\mathcal{N}(A)}(x)\|_2 = \|(x - \hat{x}) - P_{\mathcal{N}(A)}(x - \hat{x})\|_2 \\ &\leq \tfrac{1}{\sigma_{min}} \cdot \|Ax - A\hat{x}\|_2 = \tfrac{1}{\sigma_{min}} \cdot \operatorname{dist}\big(Ax, Q \cap \mathcal{R}(A)\big) \\ &\leq \tfrac{\gamma}{\sigma_{min}} \cdot \operatorname{dist}(Ax, Q) , \end{aligned}$$

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. [24]. For non-polyhedral sets Q the assertion holds if $\operatorname{rint}(Q) \cap \mathcal{R}(A) \neq \emptyset$, cf. Theorem 4.2. 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,

$$\min_{x \in \mathbb{R}^n} f(x) \quad \text{s.t.} \quad Ax = b \tag{15}$$

like in (4) or (10). If the objective function f is strongly convex then (15) 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 $\mathit{\Pi}^{x^*}_{L(A,b)}(x)$ 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 2.9 (a). As a consequence for all such x, x^* we have $\operatorname{dist}^{x^*}_f(x, L(A,b))^2 = D^{x^*}_f(x, \hat{x})$. Our next aim is an error bound of the form $D^{x^*}_f(x,\hat{x}) \leq \gamma \cdot \|Ax - b\|_2^2$. For piecewise linear-quadratic or differentiable f this immediately follows from Lemma 4.7 and 4.3 (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 [39].

Definition 4.10 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

$$S(x) \subset S(\hat{x}) + L \cdot \|x - \hat{x}\|_2 \cdot B_1 \quad \text{for any } x \text{ with} \quad \|x - \hat{x}\|_2 \leq \epsilon\,.$$

The following examples were given in [41].

Example 4.11 (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 [38].

(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 linear-quadratic function which is *absolutely symmetric*, i.e. $h(x_1, \ldots, x_m) = h\big(|x_{\pi(1)}|, \ldots, |x_{\pi(m)}|\big)$ for any permutation π of the indices. Then the subdifferential mapping of $f(X) := h\big(\sigma(X)\big)$ 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

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

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 [41].

Now we can reformulate Theorem 2.12 in [41] to fit the present context.

Theorem 4.12 *Consider the linearly constrained optimization problem* (15) *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* (15) *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* $D_f^{x^*}(x, \hat{x}) \le D_f^{x_0^*}(x_0, \hat{x})$ *we have*

$$\operatorname{dist}_f^{x^*}(x, L(A,b))^2 = D_f^{x^*}(x, \hat{x}) \le \gamma \cdot \|Ax - b\|_2^2 .$$

Proof To obtain the error bound we apply the results of [41] 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 [41] that the function g is *restricted strongly convex* on all of its level sets. Hence, by Lemma 2.2 in [41], 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 $D_f^{x^*}(x, \hat{x}) \le D_f^{x_0^*}(x_0, \hat{x})$ we have $D_f^{x^*}(x, \hat{x}) = g(y) - g_{min} \le \gamma \cdot \|\nabla g(y)\|_2^2 = \gamma \cdot \|Ax - b\|_2^2$. □

5 Randomized Bregman Projections for convex feasibility problems

We consider general convex feasibility problems (8) with finitely many closed convex sets $C_i \subset \mathbb{R}^n$ and nonempty intersection $C := \bigcap_{i=1}^m C_i$. As demonstrated in [28] this framework allows for numerous generalizations of (4). A widely known idea to solve a CFP is to project successively onto the individual sets C_i, see e.g. [3–5, 8, 12, 18, 48]. For efficiency it is essential that projections onto these sets can be computed relatively cheaply. But if some of the sets have the form

$$C_i = \{x \in \mathbb{R}^n \mid A_i x \in Q_i\} \quad , \quad i \in I_Q \subset \{1, \dots, m\} \tag{16}$$

with a closed convex set $Q_i \subset \mathbb{R}^{m_i}$ and matrix $A_i \in \mathbb{R}^{m_i \times n}$, then projecting onto such a set can be very expensive for large dimensions, and hence it is often preferable to project onto an enclosing halfspace as in the following lemma.

Lemma 5.1 (Lemma 2.6 in [28]) *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*

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

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

CFP's with some sets of the form (16) are also called *split feasibility problems*, see e.g. [11, 13, 17, 28, 42]. We analyse the convergence behaviour of a randomized projection algorithm to solve the CFP (8), stated as Algorithm 3. It uses Bregman projections onto C_i in case $i \notin I_Q$, and Bregman projections onto an enclosing halfspace according to Lemma 5.1 for sets of the form (16) in case $i \in I_Q$. In each iteration an index i is chosen with probability $p_i > 0$.

Algorithm 3 Randomized Bregman projections (RBP)

Input: data according to (8) and (16), starting points $x_0 \in \mathbb{R}^n$, $x_0^* \in \partial f(x_0)$ and probabilities $p_i > 0$, $i \in \{1, \dots, m\}$
Output: a solution of (8)
1: initialize $k = 0$
2: **repeat**
3: choose an index $i_k = i \in \{1, \dots, m\}$ at random with probability $p_i > 0$
4: **if** $i_k \notin I_Q$ **then**
5: update $x_{k+1} = \Pi_{C_{i_k}}^{x_k^*}(x_k)$ together with an admissible subgradient $x_{k+1}^* \in \partial f(x_{k+1})$, cf. Lemma 2.8
6: **else if** $i_k \in I_Q$ (cf. (16)) **then**
7: set $w_k = A_{i_k} x_k - P_{Q_{i_k}}(A_{i_k} x_k)$ and $\beta_k = \langle A_{i_k}^T w_k \,,\, x_k \rangle - \|w_k\|_2^2$
8: update $x_{k+1} = \Pi_{H_{\leq}(A_{i_k}^T w_k, \beta_k)}^{x_k^*}(x_k)$ with $x_{k+1}^* \in \partial f(x_{k+1})$ as in Lemma 2.9 (b)
9: **end if**
10: increment $k = k + 1$
11: **until** a stopping criterion is satisfied

In [28] convergence of the iterates to a solution of (8) was shown for Bregman projections with respect to nondifferentiable functions, and for quite general control sequences $i : \mathbb{N} \to \{1, \dots, m\}$. The only requirement was that $\big(i(k)\big)_{k \in \mathbb{N}}$ encounters each index in $\{1, \dots, m\}$ infinitely often.[1] However, no assertion was made about convergence rates. Here we follow [1, 9, 20, 27, 30, 33, 37, 43, 49] and show that the iterates of the randomized version Algorithm 3 converge in expectation to a solution of (8) with an expected (sub-)linear convergence rate.

Theorem 5.2 *Consider the CFP* (8) *where some sets may have the form* (16)*, and assume that the collections* $\{C_1, \dots, C_m\}$ *and* $\{Q_i, \mathcal{R}(A_i)\}$ *for each* $i \in I_Q$ *are boundedly linearly regular. Let* $f : \mathbb{R}^n \to \mathbb{R}$ *be* α*-strongly convex. 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 3 remain bounded, the Bregman distances to the intersection* C *decrease monotonically,*

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

[1] Because very general control sequences 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 3 instead of *random* to distinguish between the cases.

and converge in expectation to zero, where the expectation is taken with respect to the probability distribution $p_i > 0$, $i \in \{1, \dots, m\}$. The expected rate of convergence is at least sublinear: There is a constant $c > 0$ such that

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

Proof At first we consider the case $i_k \notin I_Q$. By Lemma 2.6 we have

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

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

$$D_f^{x_{k+1}^*}(x_{k+1}, x) \leq D_f^{x_k^*}(x_k, x) - \frac{\alpha}{2} \cdot \operatorname{dist}(x_k, C_{i_k})^2\,. \tag{17}$$

Now we consider the case $i_k \in I_Q$. By Lemma 5.1 we have $C \subset H_{\leq}(A_{i_k}^T w_k, \beta_k)$, and together with Lemma 2.9 (b) we can estimate for all $x \in C$

$$D_f^{x_{k+1}^*}(x_{k+1}, x) \leq D_f^{x_k^*}(x_k, x) - \frac{\alpha}{2 \cdot \|A_{i_k}\|_2^2} \cdot \|A_{i_k} x_k - P_{Q_{i_k}}\left(A_{i_k} x_k\right)\|_2^2\,. \tag{18}$$

We fix some $x \in C$ and conclude from (17), (18) and Lemma 2.6 that both x_k and x_k^* remain bounded. Hence by Lemma 4.9 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

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

Inserting this estimate into (18) we get

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

Together with (17) this implies that the Bregman distances decrease monotonically, and that there is a constant $c > 0$ such that

$$\operatorname{dist}_f^{x_{k+1}^*}(x_{k+1}, C)^2 \leq \operatorname{dist}_f^{x_k^*}(x_k, C)^2 - c \cdot \operatorname{dist}(x_k, C_{i_k})^2\,. \tag{19}$$

For the moment we fix the values of the indices $i_0, \dots, i_{k-1}$ and consider only i_k as a random variable with values in $\{1, \dots, m\}$. Taking the expectation on both sides of (19) conditional to the values of the indices $i_0, \dots, i_{k-1}$ yields

$$\mathbb{E}\left[\operatorname{dist}_f^{x_{k+1}^*}(x_{k+1}, C)^2 \,\middle|\, i_0, \dots, i_{k-1}\right] \leq \operatorname{dist}_f^{x_k^*}(x_k, C)^2 - \sum_{i=1}^{m} p_i \cdot c \cdot \operatorname{dist}(x_k, C_i)^2\,.$$

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

$$\mathbb{E}\left[\operatorname{dist}_f^{x_{k+1}^*}(x_{k+1}, C)^2 \,\middle|\, i_0, \dots, i_{k-1}\right] \leq \operatorname{dist}_f^{x_k^*}(x_k, C)^2 - \gamma \cdot \operatorname{dist}(x_k, C)^2\,. \tag{20}$$

Furthermore, by Lemma 4.3 (a) there is $L > 0$ such that for all k we have $\mathrm{dist}_f^{x^*}(x_k, C)^4 \leq L \cdot \mathrm{dist}(x_k, C)^2$, and hence we get

$$\mathbb{E}\left[\mathrm{dist}_f^{x_{k+1}^*}(x_{k+1}, C)^2 \,\middle|\, i_0, \ldots, i_{k-1}\right] \leq \mathrm{dist}_f^{x_k^*}(x_k, C)^2 - \tfrac{\gamma}{L} \cdot \mathrm{dist}_f^{x_k^*}(x_k, C)^4 .$$

Now we consider all indices $i_0, \ldots, i_k$ as random variables with values in $\{1, \ldots, m\}$, and take the full expectation on both sides,

$$\begin{aligned}\mathbb{E}\left[\mathrm{dist}_f^{x_{k+1}^*}(x_{k+1}, C)^2\right] &\leq \mathbb{E}\left[\mathrm{dist}_f^{x_k^*}(x_k, C)^2\right] - \tfrac{\gamma}{L} \cdot \mathbb{E}\left[\mathrm{dist}_f^{x_k^*}(x_k, C)^4\right] \\ &\leq \mathbb{E}\left[\mathrm{dist}_f^{x_k^*}(x_k, C)^2\right] - \tfrac{\gamma}{L} \cdot \left(\mathbb{E}\left[\mathrm{dist}_f^{x_k^*}(x_k, C)^2\right]\right)^2 .\end{aligned}$$

We set $d_k := \mathbb{E}\left[\mathrm{dist}_f^{x_k^*}(x_k, C)^2\right]$. 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

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

we obtain $\frac{1}{d_{k+1}} \geq \frac{1}{d_0} + \frac{\gamma}{L}(k+1)$, and we conclude $d_k \leq \frac{L d_0}{L + \gamma d_0 \cdot k}$ as desired. The expected sublinear convergence rates for $\mathrm{dist}(x_k, C)$ now follow from the estimate $\mathbb{E}\left[\mathrm{dist}(x_k, C)\right] \leq \sqrt{\frac{2}{\alpha}} \cdot \mathbb{E}\left[\mathrm{dist}_f^{x_k^*}(x_k, C)\right]$, cf. Lemma 2.6. □

Remark 5.3 According to Lemma 2.9 (b) the computation of the Bregman projection $x_{k+1} = \Pi^{x_k^*}_{H_\leq(A_{i_k}^T w_k, \beta_k)}(x_k)$ onto the halfspace $H_\leq(A_{i_k}^T w_k, \beta_k)$ in step 8 of Algorithm 3 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 5.2 and the next two theorems remain true for inexact linesearches as well, cf. [28]. In particular, we may choose

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

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

Theorem 5.4 *If f is piecewise linear-quadratic or has a Lipschitz-continuous gradient, then under the assumptions of Theorem 5.2 the expected rate of convergence is linear: There are constants $q \in (0, 1)$ and $c > 0$ such that*

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

and hence

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

Proof By Theorem 4.7 and Lemma 4.3 (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 (20) in the proof of Theorem 5.2 we get

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

from which the linear convergence rates follow. □

Finally we turn to linearly constrained optimization problems.

Theorem 5.5 *Consider the linearly constrained optimization problem* (15) *under the assumptions of Theorem 4.12. Let $I_1, \dots, I_r$ be a covering of $\{1, \dots, m\}$ (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. In Algorithm 3 we may choose to project directly onto the sets $C_i = \{x \in \mathbb{R}^n \mid A_i x = b_i\}$ according to Lemma 2.9 (a), or onto an enclosing halfspace according to Lemma 5.1 with $Q_i = \{b_i\}$. 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 3 converge in expectation to the solution $\hat{x}$ of* (15)*. The expected rate of convergence is linear: There are constants $q \in (0, 1)$ and $c > 0$ such that*

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

and hence

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

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 $k \geq 0$. Hence the assertion follows from Theorem 4.12 as in the proofs of Theorem 5.2 and 5.4. □

By Theorem 5.5 we get linear rates for randomized iterations of the form (9) to solve the regularized nuclear norm problem (10). Note that expected linear convergence for a randomized and smoothed Sparse Kaczmarz method to approximately solve the regularized Basis Pursuit problem (4) was also shown in [35]. There the objective function in (4) was replaced by

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

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

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

The function f_ϵ is 1-strongly convex and has a Lipschitz-continuous gradient. Hence linear convergence is also guaranteed by Theorem 5.5. But as shown in Section 3, this result holds without smoothing the objective function. Of course this also holds for the Randomized Block Sparse Kaczmarz method considered in [35] by applying Theorem 5.5 with a covering $I_1, \dots, I_r$ of $\{1, \dots, m\}$.

6 Numerical examples

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

6.1 Experiment A: 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 (RaSK) (Algorithm 1), and the Exact-Step Randomized Sparse Kaczmarz method (ERaSK) (Algorithm 2) 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 where the value $\lambda = 1$ was used for RaSK and ERaSK. Note that the usual RK performs consistently well over all trials, while the performance of RaSK and ERaSK differs drastically between different instances. As denoted by the quantiles, there are a few instances on which RaSK and ERaSK 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 ERaSK, and also RaSK has a faster asymptotic rate than non-sparse RK.

Figure 2 shows results for the underdetermined and consistent case with $\lambda = 3$ for RaSK and ERaSK. The ERaSK method takes advantage of the fact that the vectors $\hat{x}$ are very sparse. On the other hand, the RaSK method does not reduce the residual as fast as the RK method does. However, since the problem is underdetermined, the RK method does not converge to a sparse solution and hence, the error does not converge to zero.

Figures 3 and 4 show the results for noisy right hand sides, both with $\lambda = 1$ for RaSK and ERaSK. Figure 3 uses a two times overdetermined system with 10% relative noise, Figure 4 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, ERaSK achieves this faster than RaSK which in turn is faster than RK. Regarding the reconstruction error, ERaSK and RK achieve reconstructions with an error in the size of the noise level, while RaSK achieves an even lower reconstruction error. At least the smaller error of RaSK compared to ERaSK is explained by comparing the estimates in Theorem 3.4 (a) and (b): The constant term in estimate in (b) for ERaSK is worse than the one for RaSK by a factor of about $\sqrt{1 + 4\|A\|_{1,2}/\delta}$. On an intuitive level one may argue that the Sparse Kaczmarz method obtains better reconstructions since it incorporates the sparsity of the solutions, but

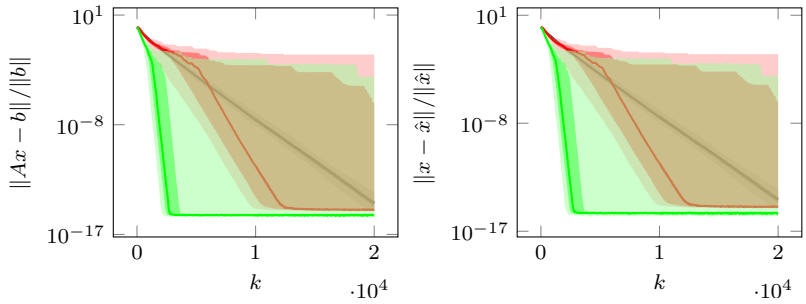

Fig. 1 Experiment A: Comparison of Randomized Kaczmarz (black) Randomized Sparse Kaczmarz (red), and Exact-Step Randomized Sparse Kaczmarz (green), $n = 200$, $m = 1000$, sparsity $s = 25$, no noise, $\lambda = 1$. Left: Plots of relative residual $\|Ax - b\|/\|b\|$, right: plots of error $\|x - \hat{x}\|/\|\hat{x}\|$.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

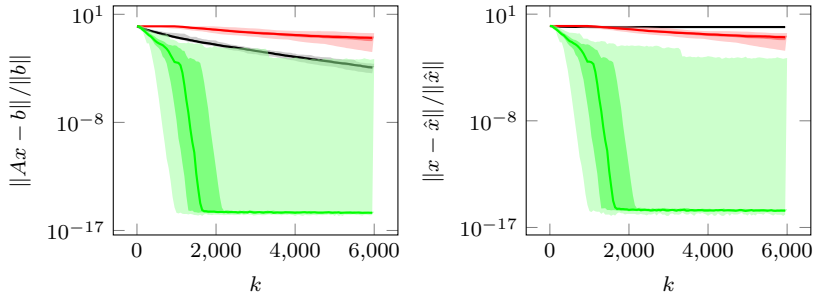

Fig. 2 Experiment A: Comparison of Randomized Kaczmarz (black) Randomized Sparse Kaczmarz (red), and Exact-Step Randomized Sparse Kaczmarz (green), $n = 600$, $m = 200$, sparsity $s = 10$, no noise, $\lambda = 3$. Left: Plots of relative residual $\|Ax - b\|/\|b\|$, right: plots of error $\|x - \hat{x}\|/\|\hat{x}\|$. Thick line shows median over 60 trials, light area is between min and max, darker area indicate 25th and 75th quantile.

that the exact steps in the Sparse Kaczmarz method spoil this advantage by trying to fullfill all equations exactly, despite the noise. In fact, RaSK with inexact stepsize may be seen as a kind of relaxed Kaczmarz method.

6.2 Experiment B: 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 [23] to create CT-measurement matrices of different sizes. We used fan-beam geometry throughout and worked with overdetermined systems, sparse solutions and noisefree right hand sides. We used $\lambda = 1$ and compared RaSK with the cyclic version of the Sparse Kaczmarz method, where we process the rows of the linear system in their "natural" order. Figure 5 shows the result for a small problem with $n = 100$ pixels, and Figure 6 shows the result for a problem with $n = 900$ pixels. In both cases the randomization shows improve-

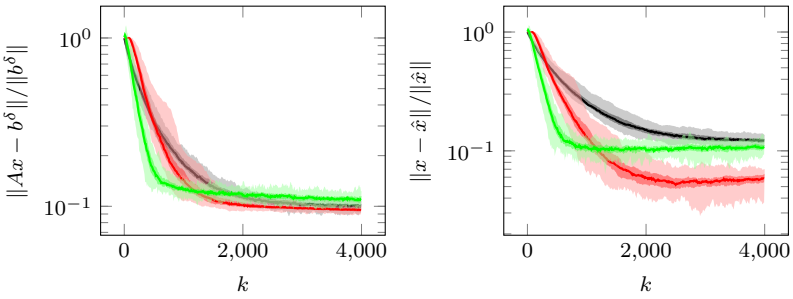

Fig. 3 Experiment A: Comparison of Randomized Kaczmarz (black) Randomized Sparse Kaczmarz (red), and Exact-Step Randomized Sparse Kaczmarz (green), $n = 200$, $m = 400$, sparsity $s = 25$, 10% relative noise, $\lambda = 1$. Left: Plots of relative residual $\|Ax - b^\delta\|/\|b^\delta\|$, right: plots of error $\|x - \hat{x}\|/\|\hat{x}\|$. Thick line shows median over 60 trials, light area is between min and max, darker area indicate 25th and 75th quantile.

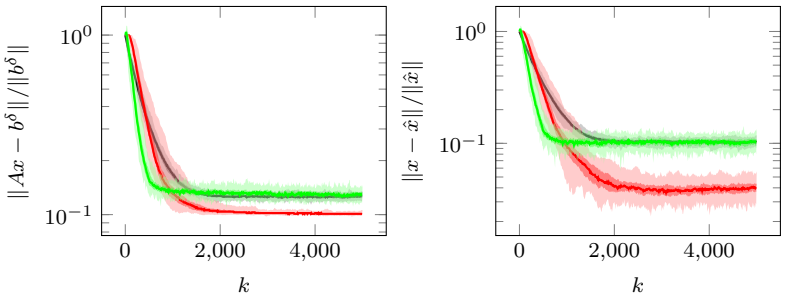

Fig. 4 Experiment A: Comparison of Randomized Kaczmarz (black) Randomized Sparse Kaczmarz (red), and Exact-Step Randomized Sparse Kaczmarz (green), $n = 200$, $m = 1000$, sparsity $s = 25$, 10% relative noise, $\lambda = 1$. Left: Plots of relative residual $\|Ax - b^\delta\|/\|b^\delta\|$, right: plots of error $\|x - \hat{x}\|/\|\hat{x}\|$. Thick line shows median over 60 trials, light area is between min and max, darker area indicate 25th and 75th quantile.

ments for the median as well as for the extreme cases.

7 Conclusion

We proved that the iterates of the Randomized Sparse Kaczmarz method are expected to converge linearly for consistent linear systems, and derived explicit estimates for the rates, cf. Theorem 3.2. Additionally, we show that in the noisy/inconsistent case, the iterates reach an error threshold in the order of the noise-level with the same rate as in the noiseless case. Numerical experiments confirm the theoretical results and demonstrate the benefit of using the method to recover sparse solutions of linear systems, even in the overdetermined case. We also obtained (sub-)linear convergence rates in expectation for the method of Randomized Bregman projections to solve general convex feasibility problems. Let us remark that, motivated by the excellent perfor-

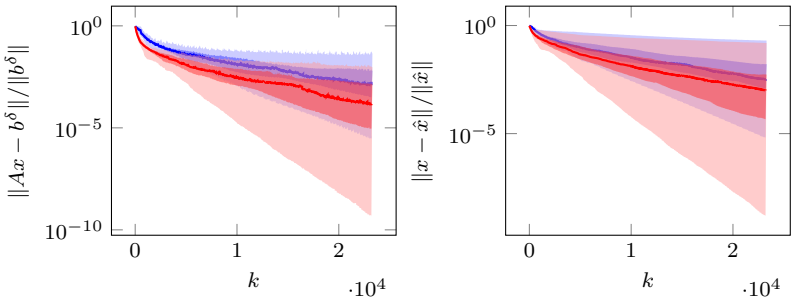

Fig. 5 Experiment B: Sparse Kaczmarz (blue) vs. Randomized Sparse Kaczmarz (red), $n = 100$, $m = 1164$, sparsity $s = 20$, $\lambda = 1$. Left: Plots of relative residual $\|Ax - b\|/\|b\|$, right: plots of error $\|x - \hat{x}\|/\|\hat{x}\|$. Thick line shows median over 40 trials, light area is between min and max, darker area indicate 25th and 75th quantile.

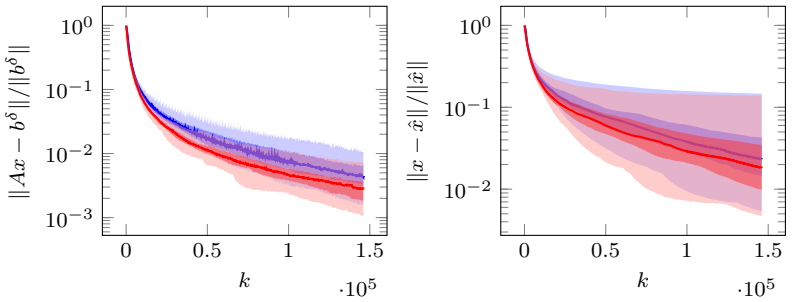

Fig. 6 Experiment B: Sparse Kaczmarz (blue) vs. Randomized Sparse Kaczmarz (red), $n = 900$, $m = 3660$, sparsity $s = 180$, $\lambda = 1$. Left: Plots of relative residual $\|Ax - b\|/\|b\|$, right: plots of error $\|x - \hat{x}\|/\|\hat{x}\|$. Thick line shows median over 40 trials, light area is between min and max, darker area indicate 25th and 75th quantile.

mance of the Randomized Sparse Kaczmarz method, we also tried to solve the regularized nuclear norm problem (10) by applying a randomized Kaczmarz iteration of the form (9). 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 (9) 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

1. 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 Journal on Control and Optimization*, 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. *Mathematics of Computation*, 78:2127–2136, 2009.
16. E. J. Candès and T. Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, 2005.
17. Y. Censor and T. Elfving. A multiprojection algorithm using Bregman projections in a product space. *Numerical Algorithms*, 8:221–239, 1994.
18. Y. Censor, T. Elfving, N. Kopf, and T. Bortfeld. The multiple-sets split feasibility problem and its applications for inverse problems. *Inverse Problems*, 21:2071–2084, 2005.

19. 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.
20. 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.
21. 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.
22. M. Elad. *Sparse and redundant representations: from theory to applications in signal and image processing*. Springer, 2010.
23. 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.
24. 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.
25. S. Kaczmarz. Angenäherte Auflösung von Systemen linearer Gleichungen. *Bulletin International de l'Academie Polonaise des Sciences et des Lettres*, pages 355–357, 1937.
26. M. J. Lai and W. Yin. Augmented ℓ_1 and nuclear-norm models with a globally linearly convergent algorithm. *SIAM Journal on Imaging Science*, 6(2):1059–1091, 2013.
27. D. Leventhal and A. S. Lewis. Randomized methods for linear constraints: convergence rates and conditioning. *Mathematics of Operations Research*, 35(3):641–654, 2010.
28. D. A. Lorenz, F. Schöpfer, and S. Wenger. The linearized Bregman method via split feasibility problems: Analysis and generalizations. *SIAM Journal on Imaging Sciences*, 7(2):1237–1262, 2014.
29. 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.
30. H. Mansour and O. Yilmaz. A fast randomized Kaczmarz algorithm for sparse solutions of consistent linear systems. *arXiv preprint arXiv:1305.3803*, 2013.
31. D. Needell. Randomized Kaczmarz solver for noisy linear systems. *BIT Numerical Mathematics*, 50(2):395–403, 2010.
32. D. Needell, N. Srebro, and R. Ward. Stochastic gradient descent, weighted sampling, and the randomized Kaczmarz algorithm. *Mathematical Programming*, 155(1):549–573, 2016.
33. 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.
34. Y. Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.

35. S. Petra. Randomized sparse block Kaczmarz as randomized dual block-coordinate descent. *Analele Stiintifice Ale Universitatii Ovidius Constanta-Seria Matematica*, 23(3):129–149, 2015.
36. 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.
37. P. Richtárik and M. Takáč. Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. *Mathematical Programming*, 144(1-2):1–38, 2014.
38. S. M. Robinson. Some continuity properties of polyhedral multifunctions. *Mathematical Programming Study*, 14:206–214, 1981.
39. R. T. Rockafellar and R. J.-B. Wets. *Variational Analysis*. Springer, Berlin, 2009.
40. F. Schöpfer. Exact regularization of polyhedral norms. *SIAM Journal on Optimization*, 22(4):1206–1223, 2012.
41. F. Schöpfer. Linear convergence of descent methods for the unconstrained minimization of restricted strongly convex functions. *SIAM Journal on Optimization*, 26(3):1883–1911, 2016.
42. 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.
43. T. Strohmer and R. Vershynin. A randomized Kaczmarz algorithm with exponential convergence. *Journal of Fourier Analysis and Applications*, 15(2):262–278, 2009.
44. M. Wang and D. P. Bertsekas. Stochastic first-order methods with random constraint projection. *SIAM Journal on Optimization*, 26(1):681–717, 2016.
45. W. Yin. Analysis and generalizations of the linearized Bregman method. *SIAM Journal on Imaging Science*, 3(4):856–877, 2010.
46. 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.
47. H. Zhang and L. Cheng. Restricted strong convexity and its applications to convergence analysis of gradient-type methods in convex optimization. *Optimization Letters*, pages 1–19, 2014.
48. J. Zhao and Q. Yang. Several solution methods for the split feasibility problem. *Inverse Problems*, 21:1791–1799, 2005.
49. 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.