manuscript No.
(will be inserted by the editor)

Extended Randomized Sparse Kaczmarz Method for Sparse Least Squares Solutions

Frank Schöpfer · Dirk A. Lorenz · Lionel N. Tondji · Maximilian Winkler

Abstract

MW: Add complex/block generalization?

The Extended Randomized Kaczmarz method is a well known iterative scheme which finds the Moore-Penrose inverse solution of a possibly inconsistent linear system and requires only one additional column of the system matrix in each iteration. Also, the Sparse Randomized Kaczmarz method has been shown to converge linearly to a sparse solution of a consistent linear system. Here, we combine both ideas and propose an Extended Sparse Randomized Kaczmarz method. We show linear expected convergence to a sparse least squares solution in the sense that an extended kind of the regularized basis pursuit problem is solved. Next, we generalize the additional step in the method and prove convergence to a more abstract optimization problem. We demonstrate numerically that our method can find sparse least squares solutions if the noise is concentrated in the complement of $\mathcal{R}(A)$ and that our generalization can handle impulsive noise.

Keywords randomized Kaczmarz method, sparse solutions, least squares, impulsive noise

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

The work of L.N.T. and D.L. has been supported by the ITN-ETN project TraDE-OPT funded by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 861137. This work represents only the author's view and the European Commission is not responsible for any use that may be made of the information it contains.

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

We consider the fundamental problem of approximating sparse solutions of large and possibly inconsistent linear systems

$$Ax = b$$

with matrix $A \in \mathbb{K}^{m\times n}$ and right hand side $b \in \mathbb{K}^m$, in the real case $\mathbb{K} = \mathbb{R}$ as well as in the complex case $\mathbb{K} = \mathbb{C}$. In particular, we have in mind situations where $A = M \cdot D$ is the product of a tall matrix $M \in \mathbb{K}^{m\times r}$ with $m > r$, and a matrix $D \in \mathbb{K}^{r\times n}$ with $r \leq n$, which acts as a basis or overcomplete dictionary that allows for a sparse representation of the solution, and where the given data b may be corrupted by noise and need not be contained in the range $\mathcal{R}(A)$ of A. This setting is somewhat more general than the usual one in the field of compressed sensing [7], where mostly flat matrices A with $m << n$ and full row rank are considered. It arises e.g. in geophysical sparsity-promoting imaging problems [28], where the system matrix is the product of a Curvelet transform matrix, which is suitable for a sparse representation of the solution, and a Jacobian, which corresponds to a linearized Born model, so that besides noisy measurement data there is also inconsistent due to a linearization error.

Here we set out to tackle such problems by solving combined optimization problems of the form

$$\begin{aligned} &\min_{x\in\mathbb{K}^n} f(x) \quad \text{s.t.} \quad Ax = \hat{y}, \\ &\text{where} \quad \hat{y} = \operatorname*{argmin}_{y\in\mathbb{K}^m} g^*(b-y) \quad \text{s.t.} \quad y \in \mathcal{R}(A) \end{aligned} \tag{1}$$

with sparsity promoting functions f and suitable data misfit functions g^*. For instance, it is known that the choice $f(x) = \lambda \cdot \|x\|_1 + \frac{1}{2} \cdot \|x\|_2^2$ favors sparse solutions for appropriate choices of $\lambda > 0$, see [4, 6, 20], where $\|x\|_1$ and $\|x\|_2$ denote the ℓ_1-norm and ℓ_2-norm of x, respectively. Similarly, by dividing the components of x into K groups $x = (x_1, \ldots, x_K)$ with $x_j \in \mathbb{K}^{n_j}$, the function $f(x) = \lambda \cdot \sum_{j=1}^K \|x_j\|_2 + \frac{1}{2} \cdot \|x\|_2^2$ favors group sparsity [23]. And in the related area of low rank matrix solutions [3, 18] we may choose $f(X) = \lambda \cdot \|X\|_* + \frac{1}{2} \cdot \|X\|_F^2$, where $\|X\|_*$ and $\|X\|_F$ denote the nuclear norm and Frobenius norm of a matrix X, respectively. Suitable data misfit functions are $g^*(b-y) = \frac{1}{2} \cdot \|b-y\|_2^2$ for least squares solutions, and ℓ_1-norm-like functions in situations where the data b is corrupted by impulsive noise, i.e. the case where only some components of the data are faulty, but with possibly large errors, see [25–27].

The linear system may be so large that full matrix operations are very expensive or even infeasible. Then it appears desirable to use iterative algorithms with low computational cost and storage per iteration that produce good approximate solutions of (1) after relatively few iterations. A celebrated example for the computation of minimum ℓ_2-norm solutions of consistent linear systems is the Kaczmarz method [11], also known as Algebraic Reconstruction Technique (ART), and its block and randomized variants [16] which started to get

popular due to the seminal paper [24]. In its most simple form for $\mathbb{K} = \mathbb{R}$, in each iteration a row vector a_i^T of A is chosen at random and the new iterate x_{k+1} is then computed as the orthogonal projection of x_k onto the solution hyperplane corresponding to the i-th equation $\langle a_i\,,\,x\rangle = b_i$, i.e.[1]

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

with initial value $x_0 = 0$. The Randomized Sparse Kaczmarz method [15, 17, 22] is a relatively new variant of the Kaczmarz method with the same low cost and storage requirements, and which has shown good performance in approximating sparse solutions of large consistent linear systems. It uses two variables x_k^* and x_k and reads as

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

with initial values $x_0 = x_0^* = 0$, and the soft shrinkage operator, which acts componentwise on a vector x as

$$\big(S_\lambda(x)\big)_j = \max\{|x_j| - \lambda, 0\} \cdot \operatorname{sign}(x_j)\,.$$

For consistent systems the iterates converge in expectation to the solution of the regularized *Basis Pursuit Problem*

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

However, for inconsistent systems the iterates need not converge, see [5] for a detailed study of this phenomenon. This behaviour is also well-known for the vanilla Kaczmarz method. As a remedy, in [8, 30] an Extended Randomized Kaczmarz method was proposed, which additionally uses one column $\tilde{a}_j$ of A in each step and finds the Moore-Penrose inverse solution, i.e. the least-squares solution with minimum ℓ^2-norm. Using an additional variable z_k with initial value $z_0 = b$, the iterates are computed as

$$\begin{aligned} z_{k+1} &= z_k - \tfrac{\langle \tilde{a}_j\,,\,z_k\rangle}{\|\tilde{a}_j\|_2^2} \cdot \tilde{a}_j\,, \\ x_{k+1} &= x_k - \tfrac{\langle a_i\,,\,x_k\rangle - b_i + z_{k+1,i}}{\|a_i\|_2^2} \cdot a_i\,. \end{aligned}$$

In this paper, we adopt this idea and propose the Generalized Extended Randomized Kaczmarz method to solve (1), see Algorithm 1. For example, to obtain sparse least squares solutions via

$$\begin{gathered} \min_{x\in\mathbb{R}^n} \lambda \cdot \|x\|_1 + \tfrac{1}{2} \cdot \|x\|_2^2 \quad \text{s.t.} \quad Ax = \hat{y}, \\ \text{where} \quad \hat{y} = \operatorname*{argmin}_{y\in\mathbb{R}^m} \tfrac{1}{2} \cdot \|b - y\|_2^2 \quad \text{s.t.} \quad y \in \mathcal{R}(A) \end{gathered}$$

[1] We use subscript indices for components of a vector, columns or rows of a matrix, and also as iteration indices. But the meaning should always be clear from the context.

the iteration reads as

$$\begin{aligned} z_{k+1} &= z_k - \tfrac{\langle \tilde{a}_j \,,\, z_k \rangle}{\|\tilde{a}_j\|_2^2} \cdot \tilde{a}_j \,, \\ x^*_{k+1} &= x^*_k - \tfrac{\langle a_i \,,\, x_k \rangle - b_i + z_{k+1,i}}{\|a_i\|_2^2} \cdot a_i \,, \\ x_{k+1} &= S_\lambda(x^*_{k+1}) \,, \end{aligned}$$

where $\tilde{a}_j$ is the j-th column of A. We also consider block versions, and prove expected convergence with rates under appropriate assumptions for general functions f and g^*. In particular, convergence in the complex case $\mathbb{K} = \mathbb{C}$ is shown by considering the iteration as a suitable block method in real variables.

2 Preliminaries

For $x, y \in \mathbb{R}^n$, we denote the standard inner product by $\langle x \,,\, y \rangle$ and for $p \in [1, +\infty[$ the ℓ_p norm by

$$\|x\|_p := \big(\sum_{i=1}^n |x_i|^p \big)^{\frac{1}{p}}.$$

For a nonempty closed convex set $E \subset \mathbb{R}^n$, we write its Euclidean projector as P_E and its distance function by

$$\operatorname{dist}(x, E) := \inf_{z \in E} \|x - z\|_2.$$

The Borel-σ-algebra on $\mathbb{R}^n$ will be denoted by $\mathcal{B}(\mathbb{R}^n)$. We will define all random elements on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Note that this is not a restriction, since such a one can always be obtained as a suitable product space, see [12]. We will refer to $(\mathcal{F}, \mathcal{B}(\mathbb{R}^n))$-measurable functions $f \colon \Omega \to \mathbb{R}^n$ as random variables. For a nonnegative or integrable random variable $f \colon \Omega \to \mathbb{R}$, by $\mathbb{E}[f]$ we denote its expectation w.r.t. the probability measure $\mathbb{P}$. By the abstract transformation formula [12, Theorem 4.10] the expectation of a composition $g \circ f$ w.r.t. $\mathbb{P}$ is equal to the expectation of g w.r.t. to the image measure $\mathbb{P} \circ f^{-1}$ (whenever the respective quantities are defined). To a sub-σ-algebra $\mathcal{G} \subset \mathcal{F}$, we associate the conditional expectation $\mathbb{E}[\cdot \mid \mathcal{G}]$, see [12, Chapter 8]. Also, we will consider the *Bochner spaces*

$$L^p(\Omega, \mathcal{G}, \mathbb{P}, \mathbb{R}^m) = \{X \colon \Omega \to \mathbb{R}^m \mid X \text{ is } \mathcal{G} \text{ -measurable and } \mathbb{E}[\|X\|_p^p] < \infty\},$$

which are Banach spaces when equipped with the norm $\|X\|_{L^p} := \mathbb{E}[\|X\|_p^p]^{\frac{1}{p}}$, see [10, p.21]. In particular for $p = 2$, we obtain Hilbert spaces with scalar product $\langle X, Y \rangle_{L^2} := \mathbb{E}[\langle X, Y \rangle]$.

2.1 Basic notions and Bregman distance

As in [22] we will analyze the convergence of the algorithms with the help of the Bregman distance [2] with respect to the objective function f. To this end we recall some well known concepts and properties of convex functions [19]. Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex and finite everywhere. Then f is continuous and its *subdifferential*

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

at any $x \in \mathbb{R}^n$ is nonempty, compact and convex. Throughout the paper we assume that f is even *strongly convex*, i.e. there is some $\alpha > 0$ such that for all $x, y \in \mathbb{R}^n$ and *subgradients* $x^* \in \partial f(x)$ we have

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

Then f is *coercive*, i.e.

$$\lim_{\|x\|_2 \to \infty} f(x) = \infty \,,$$

and its *conjugate function* $f^* : \mathbb{R}^n \to \mathbb{R}$ with

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

is also convex, finite everywhere and coercive. Additionally, f^* is differentiable with a *Lipschitz-continuous gradient* with constant $L_{f^*} = \frac{1}{\alpha}$, i.e. for all $x^*, y^* \in \mathbb{R}^n$ we have

$$\|\nabla f^*(x^*) - \nabla f^*(y^*)\|_2 \leq L_{f^*} \cdot \|x^* - y^*\|_2 \,,$$

which implies the estimate

$$f^*(y^*) \leq f^*(x^*) - \langle \nabla f^*(x^*) \,,\, y^* - x^*\rangle + \tfrac{L_{f^*}}{2} \cdot \|x^* - y^*\|_2^2 \,. \tag{2}$$

Example 2.1 (cf. [14, 29]) The sparsity promoting objective function

$$f(x) := \lambda \cdot \|x\|_1 + \tfrac{1}{2} \cdot \|x\|_2^2 \tag{3}$$

is strongly convex with constant $\alpha = 1$ for any $\lambda \geq 0$, its subdifferential is

$$\partial f(x) = \{x + \lambda \cdot s \,|\, s_j = \operatorname{sign}(x_j) \text{ if } x_j \neq 0, \text{ and } s_j \in [-1, 1] \text{ if } x_j = 0\} \,,$$

and its conjugate function can be computed with the soft shrinkage operator as

$$f^*(x^*) = \tfrac{1}{2} \cdot \|S_\lambda(x^*)\|_2^2 \quad \text{with} \quad \nabla f^*(x^*) = S_\lambda(x^*) \,.$$

Definition 2.2 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

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

Fenchel's equality states that $f(x) + f^*(x^*) = \langle x\,,\, x^*\rangle$ if $x^* \in \partial f(x)$ and implies that the Bregman distance can be written as

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

Example 2.3 (cf. [22]) For $f(x) = \frac{1}{2}\cdot\|x\|_2^2$ we just have $D_f^{x^*}(x,y) = \frac{1}{2}\|x-y\|_2^2$. For $f(x) = \lambda\cdot\|x\|_1 + \frac{1}{2}\cdot\|x\|_2^2$ and any $x^* = x + \lambda\cdot s \in \partial f(x)$ we have

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

The following inequalities are crucial for the convergence analysis of the randomized algorithms. They immediately follow from the definition of the Bregman distance and the assumption of strong convexity of f, cf. [14]. For all $x, y \in \mathbb{R}^n$ and $x^* \in \partial f(x)$, $y^* \in \partial f(y)$ we have

$$\frac{\alpha}{2}\|x-y\|_2^2 \leq D_f^{x^*}(x,y) \leq \langle x^* - y^*\,,\, x-y\rangle \leq \|x^*-y^*\|_2\cdot\|x-y\|_2 \qquad (4)$$

Note that if f is differentiable with a Lipschitz-continuous gradient, then we also have the (better) upper estimate $D_f^{x^*}(x,y) \leq L_f\cdot\|x-y\|_2^2$, but in general this need not be the case. The following example was also used in [17] as a smoothed version of (3).

Example 2.4 For $\varepsilon > 0$ the *Huber function* [9] is defined by

$$r_\varepsilon(x) := \sum_{j=1}^{n} \begin{cases} |x_j| - \frac{\varepsilon}{2} & ,|x_j| > \varepsilon \\ \frac{1}{2\cdot\varepsilon}\cdot x_j^2 & ,|x_j| \leq \varepsilon. \end{cases}.$$

Then for $\tau > 0$ the function

$$f(x) := r_\varepsilon(x) + \tfrac{\tau}{2}\cdot\|x\|_2^2,$$

is τ-strongly convex and has a $\left(\frac{1}{\varepsilon}+\tau\right)$-Lipschitz-continuous gradient with

$$\big(\nabla f(x)\big)_j = \left(\tfrac{1}{\max(\varepsilon,|x_j|)} + \tau\right)\cdot x_j\,.$$

2.2 Probability theory

The following lemma tells how a conditional expectation of a composed random variable simplifies if its arguments can be split into a 'measurable' and an 'independent' part, and will be used in our convergence analysis.

Lemma 2.5 *Let $\mathcal{G} \subseteq \mathcal{F}$ be a sub-σ-algebra and $(\mathcal{S},\mathcal{E})$ a measurable space. Consider functions with the following properties:*

(i) $X\colon \Omega \to \mathcal{S}$, $(\mathcal{G},\mathcal{E})$*-measurable,*
(ii) $I\colon \Omega \to \{1,\dots,k\}$, $\big(\mathcal{F},\mathcal{P}(\{1,\dots,k\})\big)$*-measurable and independent of* $\mathcal{G}$*,*
(iii) $\varphi\colon \{1,...,k\}\times\mathcal{S} \to \mathbb{R}$ *such that each function* $\varphi(i,\cdot)$ *is* $(\mathcal{E},\mathcal{B}(\mathbb{R}))$*-measurable and* $\varphi(i,X)\colon \Omega\to\mathbb{R}$ *is integrable.*

Then, also $\varphi(I,X)$ is integrable and

$$\mathbb{E}\big[\varphi(I,X) \mid \mathcal{G}\big] = \sum_{i=1}^{l} \mathbb{P}\big[\{I=i\}\big]\varphi(i,X).$$

Proof Since it holds

$$\varphi(I,X) = \sum_{i=1}^{l} 1_{\{I=i\}}\varphi(i,X)$$

with

$$1_A(\omega) := \begin{cases} 1 \ ,\omega \in A \\ 0 \ ,\omega \notin A, \end{cases} \qquad A \subset \Omega,$$

$\varphi(I,X)$ is integrable as a sum of integrable functions. Hence, its conditional expectation is defined and fulfils

$$\mathbb{E}\big[\varphi(I,X) \mid \mathcal{G}\big] = \sum_{i=1}^{l} \mathbb{E}\big[1_{\{I=i\}}\varphi(i,X) \mid \mathcal{G}\big].$$

Since X is $(\mathcal{G},\mathcal{E})$-measurable and $\varphi(i,\cdot)$ is $(\mathcal{E},\mathcal{B}(\mathbb{R}))$-measurable, $\varphi(i,X)$ is $(\mathcal{G},\mathcal{B}(\mathbb{R}))$-measurable. Hence, by [12, Theorem 8.14(iii)],

$$\mathbb{E}\big[1_{\{I=i\}}\varphi(i,X) \mid \mathcal{G}\big] = \mathbb{E}\big[1_{\{I=i\}}|\mathcal{G}\big] \cdot \varphi(i,X).$$

Since I and $\mathcal{G}$ are independent, also $1_{\{I=i\}} = 1_{\{\cdot=i\}} \circ I$ and $\mathcal{G}$ are independent. By [12, Theorem 8.14(vi)], this implies that

$$\mathbb{E}\big[1_{\{I=i\}} \mid \mathcal{G}\big] = \mathbb{E}[1_{\{I=i\}}] = \mathbb{P}[\{I=i\}]$$

and the assertion follows. □

3 Error bounds for linearly constrained optimization problems

Consider the feasible, convex and linearly constrained optimization problem

$$\min_{x\in\mathbb{R}^n} f(x) \quad \text{s.t.} \quad Ax = b \tag{5}$$

with a nonzero matrix $A \in \mathbb{R}^{m\times n}$, right hand side $b \in \mathcal{R}(A)$, and strongly convex objective function $f : \mathbb{R}^n \to \mathbb{R}$. This problem has a unique solution $\hat{x}$ which fulfills $\partial f(\hat{x}) \cap \mathcal{R}(A^T) \neq \emptyset$. To obtain convergence rates for the solution algorithms, we need to estimate the Bregman distance of the iterates to the solution $\hat{x}$ by *error bounds* of the form $D_f^{x^*}(x,\hat{x}) \leq \gamma \cdot \|Ax-b\|_2$ or $D_f^{x^*}(x,\hat{x}) \leq \gamma \cdot \|Ax-b\|_2^2$. We will see that such error bounds always hold if f has a Lipschitz-continuous gradient. But they also hold under weaker conditions. The following example was already proved in [22] (and here it also follows from Theorem 3.9).

Example 3.1 Let $\hat{x}$ be the unique solution of (5) with objective function $f(x) = \lambda \cdot \|x\|_1 + \frac{1}{2} \cdot \|x\|_2^2$. Then there exists $\gamma(\hat{x}) > 0$ such that for all $x \in \mathbb{R}^n$ and $x^* \in \partial f(x) \cap \mathcal{R}(A^T)$ we have

$$D_f^{x^*}(x, \hat{x}) \leq \gamma(\hat{x}) \cdot \|Ax - b\|_2^2 \,.$$

Based on the results of [13], an explicit expression of $\gamma(\hat{x})$ for $\hat{x} \neq 0$ was given in [22] as follows: Let A_J denote the submatrix that is formed by the columns of A indexed by $J \subset \{1, \ldots, n\}$, and set

$$\tilde{\sigma}_{\min}(A) := \min\{\sigma_{\min}(A_J) \,|\, J \subset \{1, \ldots, n\}, A_J \neq 0\} \,.$$

For $\hat{x} \neq 0$ we define $|\hat{x}|_{\min} = \min\{|\hat{x}_j| \,|\, \hat{x}_j \neq 0\}$. Then we have

$$\gamma(\hat{x}) = \frac{1}{\tilde{\sigma}_{\min}^2(A)} \cdot \frac{|\hat{x}|_{\min} + 2\lambda}{|\hat{x}|_{\min}} \,.$$

Moreover, for $\hat{x} = 0$ we may use $\gamma(0) = \frac{\sqrt{n}}{\sigma_{\min}^2(A)}$ (this can be shown with inequality (9) in the beginning of the proof of Lemma 3.8 below, but since this explicit expression is not so important here, we omit the details). Note that $\gamma(\hat{x})$ is quite discontinuous with respect to $\hat{x}$ and may become arbitrarily large, since $\lim_{\hat{x} \neq 0, |\hat{x}|_{\min} \to 0} \gamma(\hat{x}) = \infty$. We do not know whether these expressions for $\gamma(\hat{x})$ are the best possible.

To clarify the assumptions under which such error bounds hold for more general objective functions, we introduce the concepts of calmness [19] and linear regularity [1]. Let B_2 denote the closed unit ball of the 2-norm.

Definition 3.2 The (set-valued) subdifferential mapping $\partial f : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ is *calm* at $\hat{x}$ if there are constants $\varepsilon, L > 0$ such that

$$\partial f(x) \subset \partial f(\hat{x}) + L \cdot \|x - \hat{x}\|_2 \cdot B_2 \quad \text{for any } x \text{ with } \quad \|x - \hat{x}\|_2 \leq \varepsilon \,. \tag{6}$$

Note that calmness is a local growth condition similar to Lipschitz-continuity of a gradient mapping, but for fixed $\hat{x}$. Furthermore, this does not imply that for all $x^* \in \partial f(x)$ and all $\hat{x}^* \in \partial f(\hat{x})$ we have $\|x^* - \hat{x}^*\|_2 \leq L \cdot \|x - \hat{x}\|_2$, but only for some $\hat{x}^*$ which may depend on x^*. Of course, any Lipschitz-continuous gradient mapping is calm everywhere.

Example 3.3 (a) The subdifferential mapping of any convex piecewise linear-quadratic function $f : \mathbb{R}^n \to \mathbb{R}$ is calm everywhere. In particular, this holds for $f(x) = \lambda \cdot \|x\|_1 + \frac{1}{2} \cdot \|x\|_2^2$.

(b) For matrices $X \in \mathbb{R}^{n_1 \times n_2}$ the subdifferential mapping of $f(X) = \lambda \cdot \|X\|_* + \frac{1}{2} \cdot \|X\|_F^2$ is calm everywhere.

(c) The subdifferential mapping of

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

is calm everywhere with

$$\partial f(x) = \begin{cases} \lambda \cdot \frac{x}{\|x\|_2} + x, & x \neq 0 \\ \lambda B_2, & x = 0 \end{cases}$$

$$f^*(x^*) = \tfrac{1}{2} \cdot \|x^* - P_{\lambda \cdot B_2}(x^*)\|_2^2$$

$$\nabla f^*(x^*) = x^* - P_{\lambda \cdot B_2}(x^*) = \max\left\{0, 1 - \tfrac{\lambda}{\|x^*\|_2}\right\} \cdot x^*$$

(d) Divide the components of x into K groups $x = (x_1, \ldots, x_K)$ with $x_j \in \mathbb{R}^{n_j}$. Then the subdifferential mapping of $f(x) = \lambda \cdot \sum_{j=1}^K \|x_j\|_2 + \frac{1}{2} \cdot \|x\|_2^2$ is calm everywhere.

Proof (a) and (b) were already given in [21], and (d) is the group-version of (c). Here we show (c). For $\hat{x} \neq 0$ and $x \neq 0$ the function f is indeed differentiable with

$$\|\nabla f(x) - \nabla f(\hat{x})\|_2 \leq \left(1 + \tfrac{2\lambda}{\|\hat{x}\|_2}\right) \cdot \|x - \hat{x}\|_2 \,,$$

i.e. (6) holds with $L = 1 + \frac{2\lambda}{\|\hat{x}\|_2}$ for all x with $\|x - \hat{x}\|_2 < \frac{\|\hat{x}\|_2}{2}$. For $\hat{x} = 0$ and $x \neq 0$ we have $\nabla f(x) = \lambda \cdot \frac{x}{\|x\|_2} + x \in \partial f(\hat{x}) + \|x - \hat{x}\|_2 \cdot \frac{x - \hat{x}}{\|x - \hat{x}\|_2}$, i.e. (6) holds with $L = 1$ for all x since $\frac{x - \hat{x}}{\|x - \hat{x}\|_2} \in B_2$. □

Definition 3.4 Let $\partial f(x) \cap \mathcal{R}(A^T) \neq \emptyset$. Then the collection $\{\partial f(\hat{x}), \mathcal{R}(A^T)\}$ is *linearly regular*, if there is a constant $\gamma > 0$ such that for all $x^* \in \mathbb{R}^n$ we have

$$\operatorname{dist}\left(x^*, \partial f(\hat{x}) \cap \mathcal{R}(A^T)\right) \leq \gamma \cdot \left(\operatorname{dist}\left(x^*, \partial f(\hat{x})\right) + \operatorname{dist}\left(x^*, \mathcal{R}(A^T)\right)\right). \quad (7)$$

Obviously, if f is differentiable at $\hat{x}$, i.e. if $\partial f(\hat{x}) = \{\nabla f(\hat{x})\}$ is a singleton, then we have linear regularity.

Example 3.5 (cf. [1, 22]) The collection $\{\partial f(\hat{x}), \mathcal{R}(A^T)\}$ is linearly regular, if

(a) $\partial f(\hat{x})$ is polyhedral (which holds for piecewise linear-quadratic f in particular), or if
(b) $\operatorname{rint}\left(\partial f(\hat{x})\right) \cap \mathcal{R}(A^T) \neq \emptyset$, where $\operatorname{rint}\left(\partial f(\hat{x})\right)$ denotes the relative interior of $\partial f(\hat{x})$.

The condition in Example 3.5 (b) is a standard regularity assumption, similar to the Slater condition. In [22] local error bounds where sufficient to prove convergence, because all iterates were guaranteed to be bounded. In the present paper this need not be the case (we will in general only show boundedness in expectation). But here we will derive global error bounds under a global growth condition on the subdifferential mapping of f.

Definition 3.6 We say the subdifferential mapping of f *grows at most linearly*, if there exist $\eta, \rho \geq 0$ such that for all $x \in \mathbb{R}^n$ and $x^* \in \partial f(x)$ we have

$$\|x^*\|_2 \leq \eta \cdot \|x\|_2 + \rho \,. \quad (8)$$

Example 3.7 Any Lipschitz-continuous gradient mapping grows at most linearly. Furthermore, the subdifferential mappings of all functions in Example 3.3 grow at most linearly.

Lemma 3.8 *Let $\hat{x}$ be the unique solution of* (5)*. If the subdifferential mapping of f grows at most linearly, then there exists some constant $c > 0$ such that for all $x \in \mathbb{R}^n$ and $x^* \in \partial f(x) \cap \mathcal{R}(A^T)$ we have*

$$D_f^{x^*}(x, \hat{x}) \leq c \cdot \left(\sqrt{D_f^{x^*}(x, \hat{x})} + \|\hat{x}\|_2 + 1 \right) \cdot \|Ax - b\|_2 \,.$$

Proof To $\hat{x}$ there is some $\hat{x}^* \in \partial f(\hat{x}) \cap \mathcal{R}(A^T)$. We choose $u, \hat{u} \in \mathcal{N}(A^T)^\perp$ with $x^* = A^T u$ and $\hat{x}^* = A^T \hat{u}$, and by (4) we estimate

$$\begin{aligned} D_f^{x^*}(x, \hat{x}) &\leq \langle x^* - \hat{x}^* \,,\, x - \hat{x} \rangle = \langle A^T u - A^T \hat{u} \,,\, x - \hat{x} \rangle = \langle u - \hat{u} \,,\, Ax - b \rangle \\ &\leq \|u - \hat{u}\|_2 \cdot \|Ax - b\|_2 \leq \tfrac{1}{\sigma_{\min}^2(A)} \cdot \|A^T u - A^T \hat{u}\|_2 \cdot \|Ax - b\|_2 \\ &= \tfrac{1}{\sigma_{\min}^2(A)} \cdot \|x^* - \hat{x}^*\|_2 \cdot \|Ax - b\|_2 \,. \end{aligned} \tag{9}$$

It remains to estimate $\|x^* - \hat{x}^*\|_2$. The assumption of at most linear growth (8) together with (4) implies

$$\begin{aligned} \|x^* - \hat{x}^*\|_2 &\leq \eta \cdot (\|x\|_2 + \|\hat{x}\|_2) + 2\rho \\ &\leq \eta \cdot \|x - \hat{x}\|_2 + 2 \cdot (\eta \cdot \|\hat{x}\|_2 + \rho) \\ &\leq \eta \cdot \sqrt{\tfrac{2}{\alpha} \cdot D_f^{x^*}(x, \hat{x})} + 2 \cdot (\eta \cdot \|\hat{x}\|_2 + \rho) \,, \end{aligned}$$

from which the assertion follows. □

Now we can derive the global error bound.

Theorem 3.9 *Let $f : \mathbb{R}^n \to \mathbb{R}$ be strongly convex. If its subdifferential mapping grows at most linearly, is calm at the unique solution $\hat{x}$ of* (5)*, and if the collection $\{\partial f(\hat{x}), \mathcal{R}(A^T)\}$ is linearly regular, then there exists $\gamma(\hat{x}) > 0$ such that for all $x \in \mathbb{R}^n$ and $x^* \in \partial f(x) \cap \mathcal{R}(A^T)$ we have the global error bound*

$$D_f^{x^*}(x, \hat{x}) \leq \gamma(\hat{x}) \cdot \|Ax - b\|_2^2 \,. \tag{10}$$

In particular, this holds if f has a Lipschitz-continuous gradient.

Proof Let $\alpha > 0$ be the strong convexity constant, and let $\varepsilon, L > 0$ be as in (6) in the definition of calmness. At first we consider the case $D_f^{x^*}(x, \hat{x}) \leq \frac{\alpha}{2} \cdot \varepsilon^2$. Then by (4) we have $\|x - \hat{x}\|_2 \leq \varepsilon$, so that by (6) and (9) we get

$$\operatorname{dist}\left(x^*, \partial f(\hat{x})\right) \leq L \cdot \|x - \hat{x}\|_2 \leq L \cdot \sqrt{\tfrac{2}{\alpha} \cdot D_f^{x^*}(x, \hat{x})} \,. \tag{11}$$

Let $C := \partial f(\hat{x}) \cap \mathcal{R}(A^T)$. By choosing $\hat{x}^* := P_C(x^*)$ in (9) in the beginning of the proof of Lemma 3.8, we conclude that

$$D_f^{x^*}(x, \hat{x}) \leq \tfrac{1}{\sigma_{\min}^2(A)} \cdot \operatorname{dist}\left(x^*, C\right) \cdot \|Ax - b\|_2 \,.$$

Since $x^* \in \mathcal{R}(A^T)$, linear regularity (7) ensures that

$$\operatorname{dist}\left(x^*, C\right) \leq \gamma \cdot \operatorname{dist}\left(x^*, \partial f(\hat{x})\right).$$

Hence, together with (11) we get

$$D_f^{x^*}(x, \hat{x}) \leq \tfrac{1}{\sigma_{\min}^2(A)} \cdot L \cdot \gamma \cdot \sqrt{\tfrac{2}{\alpha} \cdot D_f^{x^*}(x, \hat{x})} \cdot \|Ax - b\|_2 ,$$

which implies (10). And in case $\frac{\alpha}{2} \cdot \varepsilon^2 < D_f^{x^*}(x, \hat{x})$ we apply Lemma 3.8 to get

$$\begin{aligned} D_f^{x^*}(x, \hat{x}) &\leq c \cdot \left(\sqrt{D_f^{x^*}(x, \hat{x})} + \|\hat{x}\|_2 + 1\right) \cdot \|Ax - b\|_2 \\ &\leq c \cdot \sqrt{D_f^{x^*}(x, \hat{x})} \cdot \left(1 + (\|\hat{x}\|_2 + 1) \cdot \sqrt{\tfrac{2}{\alpha}} \cdot \tfrac{1}{\varepsilon}\right) \cdot \|Ax - b\|_2 , \end{aligned}$$

which also implies (10). □

4 Convergence analysis of the GERK method

At first we consider the real case $\mathbb{K} = \mathbb{R}$ and prove expected convergence of the generalized extended randomized block Kaczmarz method (GERK) Algorithm 1 to the unique solution of (1) for suitable strongly convex functions f and g^*. We will derive convergence rates with the help of the following technical lemma.

Lemma 4.1 *Let $a, b > 0$, $q \in (0, 1)$, and $(d_k)_{k \geq 1}$ be a sequence with $d_k > 0$ and*

$$d_{k+1} \leq d_k - a \cdot d_k^2 + b \cdot q^k . \tag{12}$$

Then there exists some $c > 0$ such that $d_k \leq \frac{c}{k}$ for all $k \geq 1$.

Proof To $q \in (0, 1)$ we find some $c > 0$ such that for all $k \geq 1$ we have

$$\sqrt{\tfrac{2b}{a} \cdot q^k} + b \cdot q^k \leq \frac{c}{k+1} . \tag{13}$$

At first we assume that there are infinitely many indices k_j (in increasing order) for which $d_{k_j}^2 \leq \frac{2b}{a} \cdot q^{k_j}$. From (12) and (13) we infer that for these indices we have $d_{k_j} \leq \sqrt{\frac{2b}{a} \cdot q^{k_j}} \leq \frac{c}{k_j}$ and

$$d_{k_j+1} \leq d_{k_j} + b \cdot q^{k_j} \leq \sqrt{\tfrac{2b}{a} \cdot q^{k_j}} + b \cdot q^{k_j} \leq \frac{c}{k_j + 1} . \tag{14}$$

Furthermore, in case $k_{j+1} > k_j + 1$, for all $k = k_j + 1, \ldots, k_{j+1} - 1$ we have $b \cdot q^k < \frac{a}{2} \cdot d_k^2$, and thus (12) yields the recursion

$$d_{k+1} \leq d_k - a \cdot d_k^2 + b \cdot q^k \leq d_k - \tfrac{a}{2} \cdot d_k^2 . \tag{15}$$

It follows that $d_{k+1} \leq d_k$, and therefore division by d_k and d_{k+1} yields

$$\frac{1}{d_k} \leq \frac{1}{d_{k+1}} - \frac{a}{2} \cdot \frac{d_k}{d_{k+1}} \leq \frac{1}{d_{k+1}} - \frac{a}{2}\,,$$

which together with (14) implies

$$(k - k_j - 1) \cdot \frac{a}{2} \leq \sum_{i=k_j+1}^{k-1} \frac{1}{d_{i+1}} - \frac{1}{d_i} = \frac{1}{d_k} - \frac{1}{d_{k_j+1}} \leq \frac{1}{d_k} - \frac{k_j+1}{c}\,.$$

We conclude that $d_k \leq \frac{1}{\min\{\frac{a}{2}, \frac{1}{c}\}} \cdot \frac{1}{k}$ for all $k \geq 1$. In the remaining case that the index set $I = \{k \in \mathbb{N} \,|\, d_k^2 \leq \frac{2b}{a} \cdot q^k\}$ is finite or empty, the assertion follows from inequality (15) for $k \notin I$ with a similar conclusion. □

Algorithm 1 Generalized Extended Randomized Block Kaczmarz (GERK)

Input: starting points $x_0 = x_0^* = 0 \in \mathbb{R}^n$ and $z_0^* = b \in \mathbb{R}^m$, $z_0 = \nabla g^*(z_0^*)$, matrix $A \in \mathbb{R}^{m \times n}$ with M_r row-blocks $0 \neq A_i \in \mathbb{R}^{m_i \times n}$ and N_c column-blocks $0 \neq \tilde{A}_j \in \mathbb{R}^{m \times n_j}$
Output: (approximate) solution of
$\min_{x \in \mathbb{R}^n} f(x)$ s.t. $Ax = \hat{y}$, where $\hat{y} \in \operatorname{argmin}_{y \in \mathbb{R}^m} g^*(b - y)$ s.t. $y \in \mathcal{R}(A)$

initialize $k = 0$
repeat
  choose a column-block index $j_k = j \in \{1, \ldots, N_c\}$ at random with probability $\tilde{p}_j > 0$
  update $z_{k+1}^* = z_k^* - \tilde{t}_k \cdot \tilde{A}_{j_k} \tilde{A}_{j_k}^T z_k$ with stepsize $\tilde{t}_k = \frac{1}{L_{g^*} \cdot \|\tilde{A}_{j_k}\|_2^2}$
  update $z_{k+1} = \nabla g^*(z_{k+1}^*)$
  choose a row-block index $i_k = i \in \{1, \ldots, N_c\}$ at random with probability $p_i > 0$
  update $x_{k+1}^* = x_k^* - t_k \cdot A_{i_k}^T (A_{i_k} x_k - b_{i_k} + z_{k+1,i_k}^*)$ with stepsize $t_k = \frac{1}{L_{f^*} \cdot \|A_{i_k}\|_2^2}$
  update $x_{k+1} = \nabla f^*(x_{k+1}^*)$
  increment $k = k + 1$
until a stopping criterion is satisfied

Theorem 4.2 *Let $g^* : \mathbb{R}^m \to \mathbb{R}$ be strongly convex with a Lipschitz-continuous gradient. Then the iterates z_k^* of the GERK method from Algorithm 1 converge in expectation to $b - \hat{y}$, where $\hat{y} \in \mathcal{R}(A)$ is the unique solution of*

$$\min_{y \in \mathbb{R}^m} g^*(b - y) \quad s.t. \quad y \in \mathcal{R}(A)\,. \tag{16}$$

If the subdifferential mapping of the strongly convex function $f : \mathbb{R}^n \to \mathbb{R}$ grows at most linearly, then the iterates x_k converge in expectation to the corresponding unique solution $\hat{x}$ of

$$\min_{x \in \mathbb{R}^n} f(x) \quad s.t. \quad Ax = \hat{y}\,. \tag{17}$$

For some $q \in (0, 1)$ and $c > 0$ the expected rates of convergence are

$$\mathbb{E}\left[\|z_k^* - (b - \hat{y})\|_2^2\right] \leq c \cdot q^k\,, \tag{18}$$

and for all $k \geq 1$

$$\mathbb{E}\left[\|x_k - \hat{x}\|_2^2\right] \leq \frac{c}{k}\,. \tag{19}$$

Moreover, if a global error bound holds at $\hat{x}$*, then we even have*

$$\mathbb{E}\left[\|x_k - \hat{x}\|_2^2\right] \leq c \cdot (1+k) \cdot q^k\,. \tag{20}$$

Proof We split the proof into two parts. In the first part we show convergence of the iterates z_k^* with the help of the results in [22], and in the second part we show convergence of the iterates x_k.

Part 1: At first we note that the iterates z_k^* are independent from x_k and i_k, so that convergence of the z_k^* can be analyzed seperately. The assumptions on g^* imply that the conjugate $g = (g^*)^*$ is also strongly convex with a Lipschitz-continuous gradient. Hence, this also holds for the objective function $h(z) := g(z) - \langle b\,,\, z\rangle$ of the dual to (16),

$$\min_{z\in\mathbb{R}^m} h(z) = g(z) - \langle b\,,\, z\rangle \quad \text{s.t.} \quad A^T z = 0\,. \tag{21}$$

Set $\tilde{z}_k^* := z_k^* - b$. Then we have $\tilde{z}_0^* = 0$ and $\nabla h^*(\tilde{z}_k^*) = \nabla g^*(\tilde{z}_k^* + b) = \nabla g^*(z_k^*)$. Hence, the iteration can be written in the form

$$\tilde{z}_{k+1}^* = \tilde{z}_k^* - \tilde{t}_k \cdot \tilde{A}_{j_k} \tilde{A}_{j_k}^T z_k \quad , \quad z_{k+1} = \nabla h^*(\tilde{z}_{k+1}^*)$$

with initial value $\tilde{z}_0^* = 0$. By Theorem 5.5 in [22] the iterates z_k converge in expectation to the unique solution $\hat{z}$ of (21) with rate $\mathbb{E}\left[\|z_k - \hat{z}\|_2^2\right] \leq c \cdot q^k$. By duality and comparison of the optimality conditions of convex programs (cf. Example in [19]), the solution $\hat{y}$ of (16) and the solution $\hat{z}$ of (21) are related by $\nabla g(\hat{z}) = b - \hat{y}$. Expected convergence of the iterates z_k^* to $b - \hat{y}$ with rate (18) then follows from the estimate

$$\|z_k^* - (b - \hat{y})\|_2 = \|\nabla g(z_k) - \nabla g(\hat{z})\|_2 \leq L_g \cdot \|z_k - \hat{z}\|_2\,.$$

Part 2: Let $w_k := A_{i_k} x_k - b_{i_k} + z_{k+1,i_k}^*$. By Definition 2.2 of the Bregman distance, and since $A_{i_k}\hat{x} = \hat{y}_{i_k}$, we have

$$D_f^{x_{k+1}^*}(x_{k+1}, \hat{x}) = f^*\left(x_k^* - t_k \cdot A_{i_k}^T w_k\right) - \langle x_k^*\,,\, \hat{x}\rangle + t_k \cdot \langle w_k\,,\, \hat{y}_{i_k}\rangle + f(\hat{x})\,.$$

Using estimate (2) for f^* yields

$$D_f^{x_{k+1}^*}(x_{k+1}, \hat{x}) \leq D_f^{x_k^*}(x_k, \hat{x}) - t_k \cdot \langle w_k\,,\, A_{i_k} x_k - \hat{y}_{i_k}\rangle + \tfrac{L_{f^*}}{2} \cdot t_k^2 \cdot \|A_{i_k}^T w_k\|_2^2\,.$$

Since $t_k^2 = \frac{1}{L_{f^*}^2 \cdot \|A_{i_k}\|_2^4}$ and $\|A_{i_k}^T w_k\|_2^2 \leq \|A_{i_k}^T\|_2^2 \cdot \|w_k\|_2^2$, we get

$$D_f^{x_{k+1}^*}(x_{k+1}, \hat{x}) \leq D_f^{x_k^*}(x_k, \hat{x}) - t_k \cdot \langle w_k\,,\, A_{i_k} x_k - \hat{y}_{i_k}\rangle + \tfrac{t_k}{2} \cdot \|w_k\|_2^2\,.$$

We rewrite the last two summands as

$$\langle w_k\,,\, A_{i_k} x_k - \hat{y}_{i_k}\rangle = \|A_{i_k} x_k - \hat{y}_{i_k}\|_2^2 + \langle \hat{y}_{i_k} - b_{i_k} + z_{k+1,i_k}^*\,,\, A_{i_k} x_k - \hat{y}_{i_k}\rangle$$

and

$$\begin{aligned}\tfrac{1}{2}\cdot\|w_k\|_2^2 = \tfrac{1}{2}\cdot\|A_{i_k}x_k-\hat{y}_{i_k}\|_2^2 + \langle \hat{y}_{i_k}-b_{i_k}+z^*_{k+1,i_k}\,,\,A_{i_k}x_k-\hat{y}_{i_k}\rangle \\ + \tfrac{1}{2}\cdot\|\hat{y}_{i_k}-b_{i_k}+z^*_{k+1,i_k}\|_2^2\end{aligned}$$

to get

$$D_f^{x_{k+1}^*}(x_{k+1},\hat{x}) \leq D_f^{x_k^*}(x_k,\hat{x}) - \tfrac{t_k}{2}\cdot\|A_{i_k}x_k-\hat{y}_{i_k}\|_2^2 + \tfrac{t_k}{2}\cdot\|\hat{y}_{i_k}-b_{i_k}+z^*_{k+1,i_k}\|_2^2\,. \tag{22}$$

Set $c_1 := \min\limits_{i=1,\ldots,M_r} \frac{p_i}{2\cdot L_{f^*}\cdot\|A_i\|_2^2}$ and $c_2 := \max\limits_{i=1,\ldots,M_r} \frac{p_i}{2\cdot L_{f^*}\cdot\|A_i\|_2^2}$. Then we have $0 < c_1 \leq p_i\cdot\frac{t_k}{2} \leq c_2$ for all $i = 1,\ldots,M_r$, so that taking expectation yields the recursion

$$\mathbb{E}\left[D_f^{x_{k+1}^*}(x_{k+1},\hat{x})\right] \leq \mathbb{E}\left[D_f^{x_k^*}(x_k,\hat{x})\right] - c_1\cdot\mathbb{E}\left[\|Ax_k-\hat{y}\|_2^2\right] + c_2\cdot\mathbb{E}\left[\|\hat{y}-b+z^*_{k+1}\|_2^2\right]\,.$$

Using (18) we arrive at

$$\mathbb{E}\left[D_f^{x_{k+1}^*}(x_{k+1},\hat{x})\right] \leq \mathbb{E}\left[D_f^{x_k^*}(x_k,\hat{x})\right] - c_1\cdot\mathbb{E}\left[\|Ax_k-\hat{y}\|_2^2\right] + c_2\cdot c\cdot q^{k+1}\,. \tag{23}$$

This recursion implies boundedness of $\mathbb{E}\left[D_f^{x_k^*}(x_k,\hat{x})\right]$, because by the choice $x_0 = x_0^* = 0$ the initial Bregman distance $D_f^{x_0^*}(x_0,\hat{x}) = f(\hat{x})$ is finite. For ease of notation, in the following we use a generic constant $c > 0$ that is independent of the iteration index k and the random choices of the algorithm. By Lemma 3.8, the linear growth assumption on ∂f implies

$$\begin{aligned}\mathbb{E}\left[D_f^{x_k^*}(x_k,\hat{x})\right] &\leq c\cdot\mathbb{E}\left[\sqrt{D_f^{x_k^*}(x_k,\hat{x})}\cdot\|Ax_k-\hat{y}\|_2\right] + c\cdot\mathbb{E}\left[\|Ax_k-\hat{y}\|_2\right] \\ &\leq c\cdot\sqrt{\mathbb{E}\left[D_f^{x_k^*}(x_k,\hat{x})\right]}\cdot\sqrt{\mathbb{E}\left[\|Ax_k-\hat{y}\|_2^2\right]} + c\cdot\mathbb{E}\left[\|Ax_k-\hat{y}\|_2\right] \\ &\leq c\cdot\sqrt{\mathbb{E}\left[\|Ax_k-\hat{y}\|_2^2\right]}\,,\end{aligned}$$

which yields

$$\left(\mathbb{E}\left[D_f^{x_k^*}(x_k,\hat{x})\right]\right)^2 \leq c\cdot\mathbb{E}\left[\|Ax_k-\hat{y}\|_2^2\right]\,.$$

We insert this inequality into recursion (23) to get

$$\mathbb{E}\left[D_f^{x_{k+1}^*}(x_{k+1},\hat{x})\right] \leq \mathbb{E}\left[D_f^{x_k^*}(x_k,\hat{x})\right] - c\cdot\left(\mathbb{E}\left[D_f^{x_k^*}(x_k,\hat{x})\right]\right)^2 + c\cdot q^{k+1}\,.$$

The sublinear convergence rate (19) then follows from Lemma 4.1. Now we turn to the asymptotically better rate (20) under the stronger assumption that a global error bound of the form

$$D_f^{x_k^*}(x_k,\hat{x}) \leq \gamma\cdot\|Ax_k-\hat{y}\|_2^2$$

holds with some constant $\gamma > 0$. We set $q_1 := \max\{0, 1 - c_1 \cdot \gamma\}$. Then we have $q_1 \in [0, 1)$, and inserting the error bound into (23) we get

$$\mathbb{E}\left[D_f^{x_{k+1}^*}(x_{k+1}, \hat{x})\right] \leq q_1 \cdot \mathbb{E}\left[D_f^{x_k^*}(x_k, \hat{x})\right] + c_2 \cdot c \cdot q^{k+1}\,.$$

Finally, we set $\tilde{q} := \max\{q_1, q\}$ and conclude inductively that

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

from which the rate (20) follows by (4). □

Remark 4.3 According to [22], the stepsize $\tilde{t}_k$ for the z_k^*-update in line 4 of Algorithm 1 may also be chosen as

$$\tilde{t}_k = \tfrac{1}{L_{g^*}} \cdot \tfrac{\|\tilde{A}_{j_k}^T z_k\|_2^2}{\|\tilde{A}_{j_k} \tilde{A}_{j_k}^T z_k\|_2^2}$$

or determined by an exact linesearch. But so far we do not know whether we can also choose the stepsize t_k for the x_k^*-update in line 7 by an exact linesearch or as $t_k = \frac{1}{L_{f^*}} \cdot \frac{\|w_k\|_2^2}{\|A_{i_k}^T w_k\|_2^2}$ with $w_k := A_{i_k} x_k - b_{i_k} + z_{k+1,i_k}^*$. The main problem with this choice here seems to be that we only have a lower estimate $t_k \geq \frac{1}{L_{f^*} \cdot \|A_{i_k}\|_2^2}$, but after inequality (22) in the above proof we would also need a suitable upper estimate (note that w_k need not be contained in $\mathcal{R}(A_{i_k})$).

To apply Theorem 4.2 in the complex case $\mathbb{K} = \mathbb{C}$, we just split the variables into real and imaginary parts. In this way, a complex linear system $Ax = b$ can equivalently be written as a real linear system of the form

$$\begin{pmatrix} \Re(A) & , -\Im(A) \\ \Im(A) & , \Re(A) \end{pmatrix} \cdot \begin{pmatrix} \Re(x) \\ \Im(x) \end{pmatrix} = \begin{pmatrix} \Re(b) \\ \Im(b) \end{pmatrix}$$

and a vector update as in lines 4,7 of Algorithm 1 for a complex vector then corresponds to block updates of the real and imaginary parts. But we must take some care when we consider a function $f : \mathbb{C}^n \to \mathbb{R}$ in complex variables as a function $f : \mathbb{R}^{2n} \to \mathbb{R}$ in real variables. In particular, there is a notable subtlety regarding the sparsity promoting function $f(x) = \lambda \cdot \|x\|_1 + \frac{1}{2} \cdot \|x\|_2^2$ for complex vectors $x \in \mathbb{C}^n$. Considering it as a real function of the form

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

the gradient ∇f^* of the conjugate function would just be componentwise shrinkage of the vector $\big(\Re(x), \Im(x)\big)$, i.e. sparsity of the real and imaginary part is enforced seperately. On the one hand, this means that sparsity of the real vector $\big(\Re(x), \Im(x)\big)$ does not necessarily imply sparsity of the complex vector x. On the other hand, a global error bound is guaranteed to hold, cf. Examples 3.3 (a), 3.5 (a), and 3.7. A more suitable way to enforce sparsity of a complex vector seems to be to just use the complex ℓ_1-norm, i.e.

$$f\big(\Re(x), \Im(x)\big) = \lambda \cdot \sum_{j=1}^{n} \sqrt{\big(\Re(x_j)\big)^2 + \big(\Im(x_j)\big)^2} + \tfrac{1}{2} \cdot \big(\|\Re(x)\|_2^2 + \|\Im(x)\|_2^2\big)\,,$$

where, by Examples 3.3 (c) and (d), the gradient ∇f^* of the conjugate function amounts to componentwise shrinkage of the complex vector x,

$$\Big(\big(\nabla f^*(x)\big)_j \mathrel{\hat{=}} \Big) \qquad \big(S_\lambda(x)\big)_j = \max\{|x_j| - \lambda, 0\} \cdot \tfrac{x_j}{|x_j|} \quad , x \in \mathbb{C}^n \,, \tag{24}$$

i.e. sparsity of the real and imaginary part is enforced simultaneously. But since this is a special form of group sparsity, we can guarantee a global error bound, and hence the better rate (20), only under an additional regularity assumption as in Example 3.5 (b).

Remark 4.4 Algorithm 1 can also be directly implemented with complex number operations. We just have to replace the transposed matrices $\tilde{A}_{j_k}^T$ and $A_{i_k}^T$ in lines 4,7 by the complex adjoints $\overline{\tilde{A}}_{j_k}^T$ and $\overline{A}_{i_k}^T$, respectively. The updates in lines 5,8 must be performed by replacing the real gradient mappings ∇g^* and ∇f^* with the corresponding complex operators, e.g. using the complex shrinkage operator (24). Note that the expressions for the Huber function and its gradient in Example 2.4 are also meaningful for complex vectors x, and the corresponding real function is still strongly convex and has a Lipschitz-continuous gradient.

Example 4.5 Here are some concrete choices for the functions f and g^* that can be used in both the real and complex case $\mathbb{K} = \mathbb{R}$ or $\mathbb{K} = \mathbb{C}$, so that the assumptions in Theorem 4.2 are fulfilled. We indicate by (RA) if a regularity assumption as in Example 3.5 (b) is needed for f to ensure a global error bound and hence the better rate (20).

(a) **(Least squares)** $g^*(y) = \frac{1}{2} \cdot \|y\|_2^2$
(b) **(Impulsive noise)** $g^*(y) = r_\varepsilon(y) + \frac{\tau}{2} \cdot \|y\|_2^2$ with the Huber function r_ε
(c) **(Minimum 2-norm)** $f(x) = \frac{1}{2} \cdot \|x\|_2^2$
(d) **(Sparsity, (RA) needed only for $\mathbb{K} = \mathbb{C}$)** $f(x) = \lambda \cdot \|x\|_1 + \frac{1}{2} \cdot \|x\|_2^2$
(e) **(Group sparsity (RA))** $f(x) = \lambda \cdot \sum_{j=1}^K \|x_j\|_2 + \frac{1}{2} \cdot \|x\|_2^2$
(f) **(Low rank matrices (RA))** $f(X) = \lambda \|X\|_* + \frac{1}{2} \|X\|_F^2$

Note that instead of $f(x) = \lambda \cdot \|x\|_1 + \frac{1}{2} \cdot \|x\|_2^2$ we could also use $f(x) = r_\varepsilon(x) + \frac{\tau}{2} \cdot \|x\|_2^2$ as sparsity promoting function, as was done in [17]. But this requires tuning the two parameters ε,τ instead of only λ. On the contrary, so far we could not prove convergence for non-smooth data misfit functions g^*, so that we cannot use $g^*(y) = \lambda \cdot \|y\|_1 + \frac{1}{2} \cdot \|y\|_2^2$ for impulsive noise.

5 Numerical examples

In this part, we report numerical results of Algorithm 1 (GERK) for multiple settings and compare with the Sparse Randomized Kaczmarz method (SRK) and the Extended Randomized Kaczmarz method (REK) from [30]. All examples are run in `MATLAB 2019b` on a computer with 1,2 GHz Quad-Core Intel Core i7 processor and 4 cores at 1,2GHz and 16GM RAM.

We consider two kinds of experiments:

(i) In a first experiment, we want to find sparse solutions of the least-squares problem $\min \|Ax - b\|_2$ for the real and complex case. We use the function g^* from Example 4.5(a) and f from Example 4.5(d).
At first, we choose a similar example as in [8] by setting

$$\hat{b} = A\hat{x}, \quad \tilde{b} = \hat{b} + \eta_{\mathcal{R}(A)^\perp}$$

with $\eta_{\mathcal{R}(A)^\perp} = Nv \in \mathcal{R}(A)^\perp = \mathcal{N}(A^*)$, where the columns of N form an orthonormal basis of $\mathcal{N}(A^*)$ and v is a random vector uniformly distributed on a sphere $\partial\mathcal{B}_r(0)$ with $r = \alpha\|b\|_2$ and a factor α. Finally, we choose a vector $\hat{x} \in \mathbb{R}^n$ with sparsity

$$|\mathrm{supp}(\hat{x})| = \frac{\min(m,n)}{20}$$

and $\mathrm{rank}(A) = \min(m,n)/2$.
Note that we do not consider a matrix with full rank due to the following reason: If $m \leq n$ and A has full rank, it holds $\mathcal{R}(A) = \mathbb{R}^m$. If $m \geq n$, the matrix A^TA is invertible and the least-squares solution of the possibly inconsistent system $Ax = b$ is unique. In both cases, the sparse solution can be found either by the existing randomized sparse Kaczmarz method or by the existing randomized extended Kaczmarz method.
We choose 50 instances of a matrix $A = U\Sigma V^T$, where U and V are generated by applying the `orth` command to a random normal matrix, or a complex matrix with random normal real and imaginary part, and Σ is a diagonal matrix with $\mathrm{rank}(A)$ many nonzero values sampled from the uniform distribution on $[0.001, 100]$ at uniformly chosen random positions.

Algorithm	Figure 1	Figure 2
REK	500/500/500	500/500/500
SRK	57/82.5/126	98/130/159
ExSRK	25/26/45	25/25/29

Table 1 Sparsity of last iterates ($\#|x_{N,i}| > 10^{-5}$) in Figures 1 and 2

For systems with noise in $\mathrm{R(A)}^\perp$ and no noise in R(A), the ExSRK method seems to find sparse least-squares solutions, see Figures 1 and 2.

(ii) In a second experiment, we add impulsive noise and use Algorithm 1 with g^* from Example 4.5(b) and f from Example 4.5(d), see Figures 3 and 3. The data is generated as in (i) with the only difference that we choose the singular values in $[0.001, 10]$. We observe that, different to the ExSRK method, the specific GERK method is able to reconstruct the sparse vectors and gives the sparsest iterates. The examplary plot of the vector components suggests that, in case of the specific GERK method, the remaining

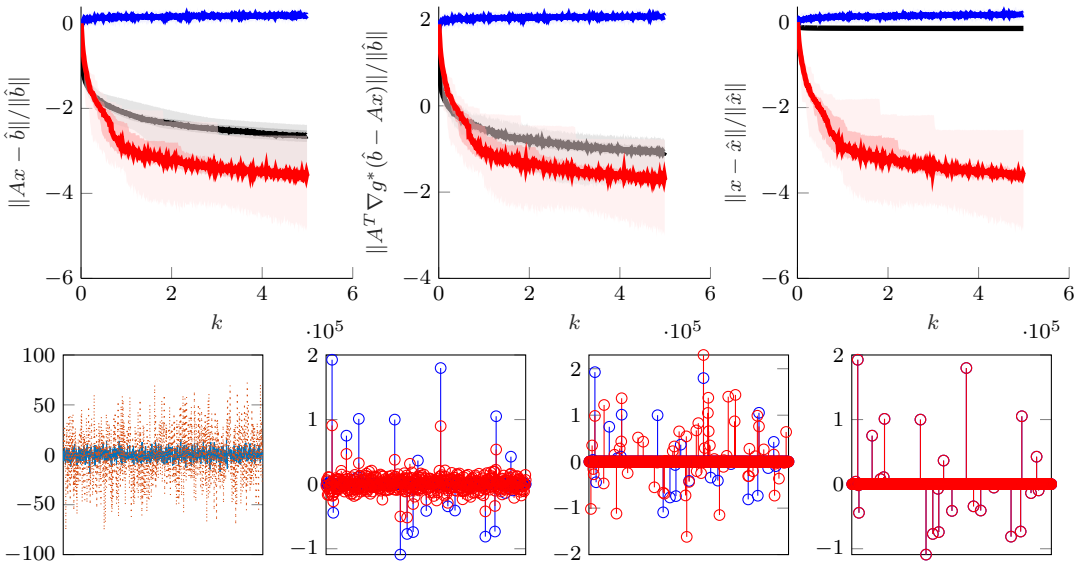

Fig. 1 A comparison of real randomized extended Kaczmarz (black), randomized sparse Kaczmarz (blue) and ExSRK method (red), $m = 1000, n = 500$, sparsity= 25, experiment (i) with $\alpha = 5$, $\lambda = 5$, uniform probabilities p. Upper figures, left: Plot of relative residual $\|Ax - b\|/\|b\|$, middle: Plot of gradient norm $\|A^T\nabla g^*(\hat{b} - Ax_k)\|/\|\hat{b}\|$, right: plot of error $\|x - \hat{x}\|$. Thick line shows median over all trials, light area is between min and max, darker area indicates 25th and 75th quantile. Lower figures, left: Plot of b (blue) and noisy b (red), right: Plot of $\hat{x}$ (blue) and last iterate x (red) of randomized extended Kaczmarz, randomized sparse Kaczmarz and ExSRK method

nonzero components might vanish after more steps. However, many iterations are needed. In this example, we ran the complex method for 10^7, the real method even for $2 \cdot 10^7$ iterations to get a good reconstruction.

Algorithm	Figure 3	Figure 4
REK	200/200/200	200/200/200
SRK	100/140.5/173	154/184/200
ExSRK	118/137.5/165	155/184/194
GERK	61/68.5/93	64/94/132

Table 2 Sparsity of last iterates ($\#|x_{N,i}| > 10^{-5}$) in Figures 1 and 2

In both experiments, there is no notable difference for $m < n$. We have observed that in experiment (ii), the reconstruction quality depends more on the condition of A than in experiment (i).

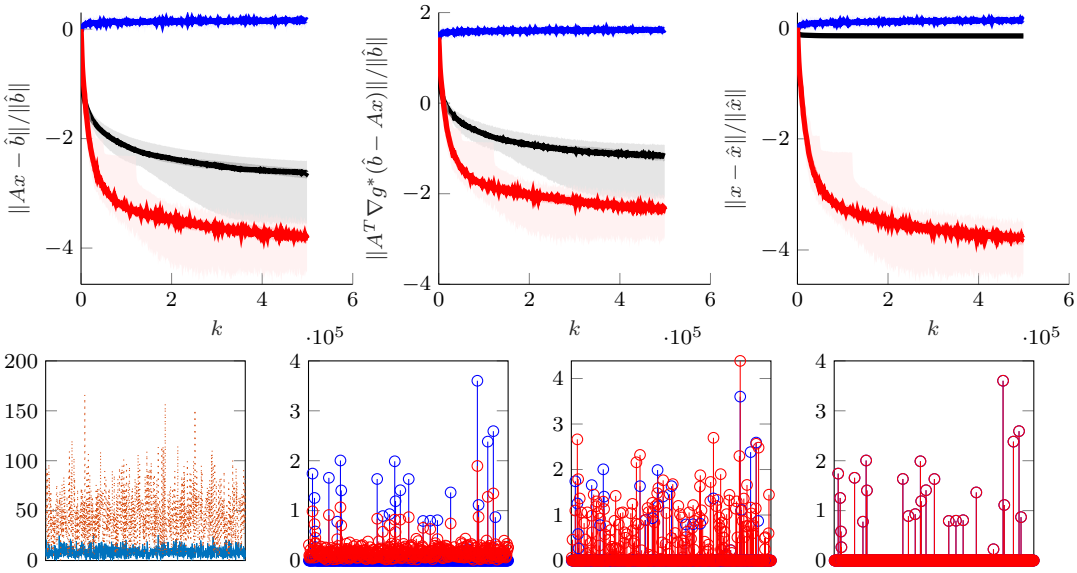

Fig. 2 Experiment from Figure 1 with complex A, b and $\hat{x}$ and the complex method, cf. Remark 4.4. In the lower figures, the absolute values are shown.

References

1. 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.
2. 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.
3. J.-F. Cai, E. J. Candès, and Z. Shen. A Singular Value Thresholding Algorithm for Matrix Completion. *ArXiv e-prints*, October 2008.
4. 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.
5. Emmanouil Daskalakis, Felix J Herrmann, and Rachel Kuske. Accelerating sparse recovery by reducing chatter. *SIAM Journal on Imaging Sciences*, 13(3):1211–1239, 2020.
6. D. L. Donoho. For most large underdetermined systems of linear equations the minimal l1-norm solution is also the sparsest solution. *Comm. Pure Appl. Math.*, 59(6):797–829, 2006.
7. David L Donoho. Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–1306, 2006.
8. Kui Du. Tight upper bounds for the convergence of the randomized extended Kaczmarz and Gauss–Seidel algorithms. *Numerical Linear Algebra*

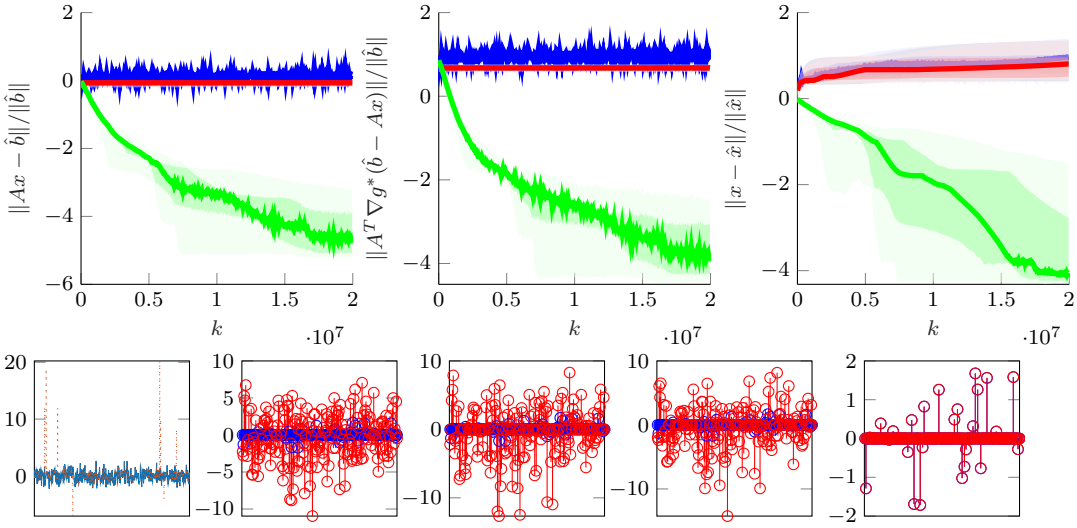

Fig. 3 A comparison of real randomized extended Kaczmarz (black), randomized sparse Kaczmarz (blue) and GERK method with Huber function $g*$ from Example 4.5(b) and f from Example 4.5(d) (red), $\epsilon = 10^{-4}$, $\tau = 10^{-3}$, $\lambda = 5$, uniform probabilities p, $m = 500, n = 200$, sparsity= 25, 10 repeats. Experiment (ii) with uniform noise from $[-2\|b\|_2, 2\|b\|_2]$ on 10 components. Upper figures, left: Plot of relative residual $\|Ax - b\|/\|b\|$, middle: Plot of gradient norm $\|A^T\nabla g^*(\hat{b} - Ax_k)\|/\|\hat{b}\|$, right: plot of error $\|x - \hat{x}\|$. Thick line shows median over all trials, light area is between min and max, darker area indicates 25th and 75th quantile. Lower figures, left: Plot of b (blue) and noisy b (red), right: Plot of $\hat{x}$ (blue) and last iterate x (red) of randomized extended Kaczmarz, randomized sparse Kaczmarz, ExSRK method and GERK method with Huber function

with Applications, 26(3):e2233, 2019.

9. Peter J Huber. Robust regression: asymptotics, conjectures and Monte Carlo. *The annals of statistics*, pages 799–821, 1973.
10. Tuomas Hytönen, Jan Van Neerven, Mark Veraar, and Lutz Weis. *Analysis in Banach spaces*, volume 12. Springer, 2016.
11. S. Kaczmarz. Angenäherte Auflösung von Systemen linearer Gleichungen. *Bull. Internat. Acad. Polon. Sci. Lettres A*, pages 355–357, 1937.
12. Achim Klenke. *Probability theory: a comprehensive course*. Springer Science & Business Media, 2013.
13. 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.
14. 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.
15. Dirk A Lorenz, Stephan Wenger, Frank Schöpfer, and Marcus 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.

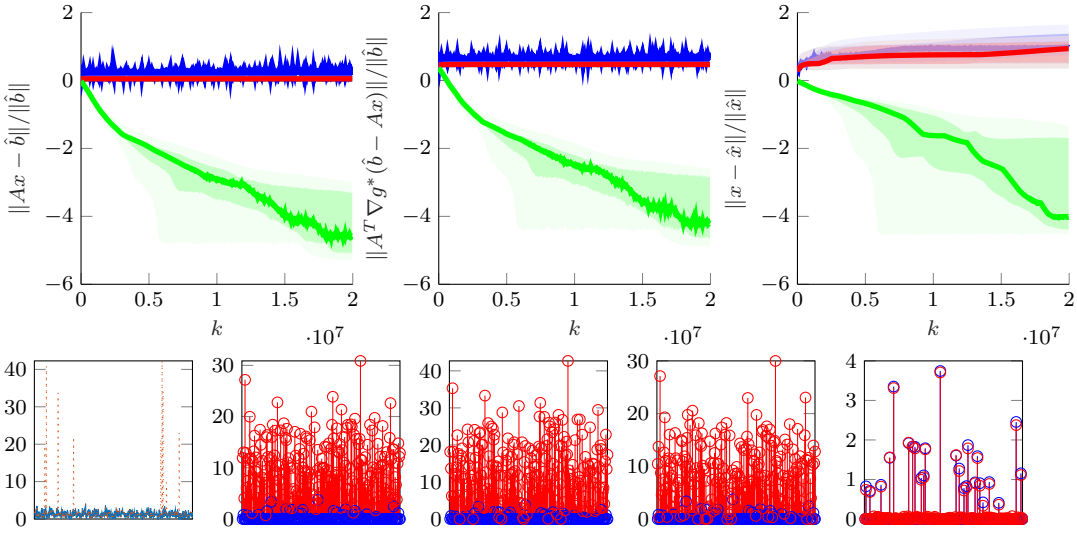

Fig. 4 Experiment from Figure 3 with complex A, b and $\hat{x}$ and the complex method, cf. Remark 4.4, with i.i.d. uniform noise in real and imaginary part. In the lower figures, the absolute values are shown.

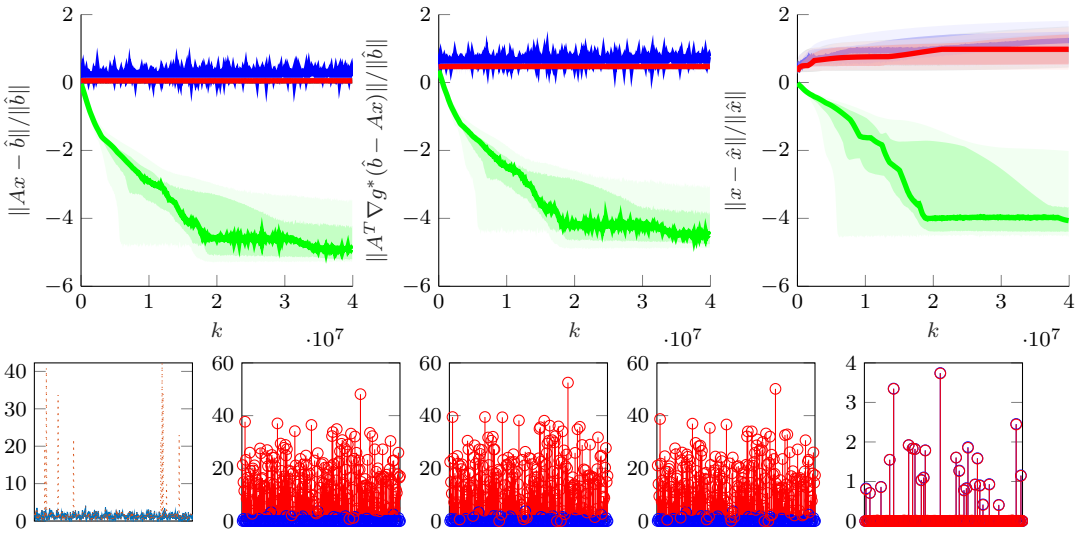

Fig. 5 Experiment from Figure 4 with $4 \cdot 10^7$ iterations

16. Deanna Needell and Joel A Tropp. Paved with good intentions: Analysis of a randomized block Kaczmarz method. *Linear Algebra and its Applications*, 441:199–221, 2014.
17. Stefania Petra. Randomized sparse block Kaczmarz as randomized dual block-coordinate descent. *Analele Stiintifice Ale Universitatii Ovidius*

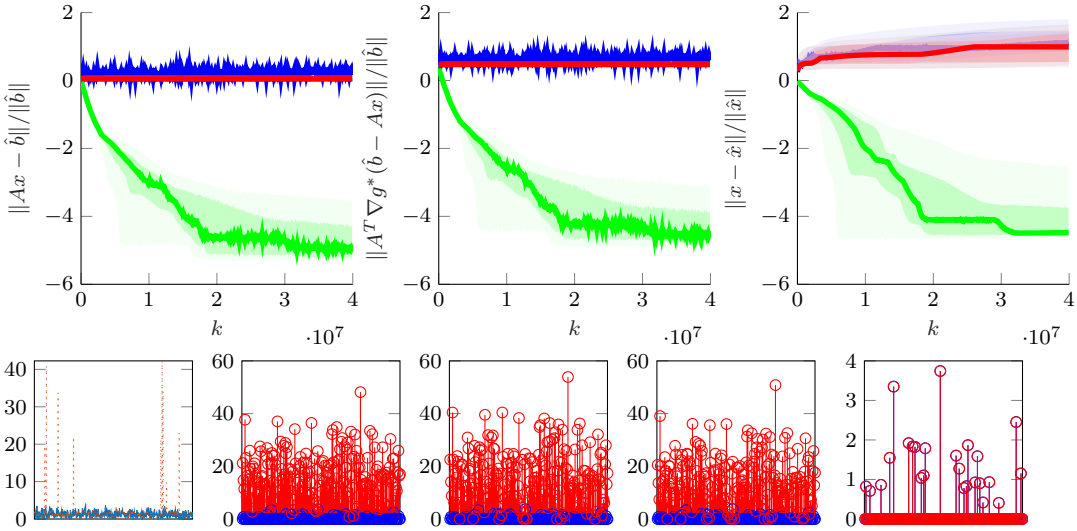

Fig. 6 Experiment from Figure 4 with $\lambda = 10$ and $4 \cdot 10^7$ iterations

Constanta-Seria Matematica, 23(3):129–149, 2015.

18. 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.
19. R. T. Rockafellar and R. J.-B. Wets. *Variational Analysis*. Springer, Berlin, 2009.
20. F. Schöpfer. Exact regularization of polyhedral norms. *SIAM J. Optim.*, 22(4):1206–1223, 2012.
21. 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.
22. Frank Schöpfer and Dirk A Lorenz. Linear convergence of the randomized sparse Kaczmarz method. *Mathematical Programming*, 173(1):509–536, 2019.
23. Mihailo Stojnic, Farzad Parvaresh, and Babak Hassibi. On the reconstruction of block-sparse signals with an optimal number of measurements. *IEEE Transactions on Signal Processing*, 57(8):3075–3085, 2009.
24. Thomas Strohmer and Roman Vershynin. A randomized Kaczmarz algorithm with exponential convergence. *Journal of Fourier Analysis and Applications*, 15(2):262–278, 2009.
25. Christoph Studer, Patrick Kuppinger, Graeme Pope, and Helmut Bolcskei. Recovery of sparsely corrupted signals. *IEEE Transactions on Information Theory*, 58(5):3115–3130, 2011.
26. Fei Wen, Peilin Liu, Yipeng Liu, Robert C Qiu, and Wenxian Yu. Robust sparse recovery in impulsive noise via ℓ_p-ℓ_1 optimization. *IEEE Transactions on Signal Processing*, 65(1):105–118, 2016.

27. Junfeng Yang and Yin Zhang. Alternating direction algorithms for ℓ_1-problems in compressive sensing. *SIAM journal on scientific computing*, 33(1):250–278, 2011.
28. Mengmeng Yang, Philipp Witte, Zhilong Fang, and Felix Herrmann. Time-domain sparsity-promoting least-squares migration with source estimation. In *2016 SEG International Exposition and Annual Meeting*. OnePetro, 2016.
29. W. Yin. Analysis and generalizations of the linearized Bregman method. *SIAM J. Imaging Sci.*, 3(4):856–877, 2010.
30. Anastasios Zouzias and Nikolaos M Freris. Randomized extended Kaczmarz for solving least squares. *SIAM Journal on Matrix Analysis and Applications*, 34(2):773–793, 2013.