

FASTER RANDOMIZED BLOCK KACZMARZ ALGORITHMS *

ION NECOARA[†]

Abstract. The Kaczmarz algorithm is a simple iterative scheme for solving consistent linear systems. At each step, the method projects the current iterate onto the solution space of a single constraint. Hence, it requires very low cost per iteration and storage, and it has a linear rate of convergence. Distributed implementations of Kaczmarz have become, in recent years, the *de facto* architectural choice for large-scale linear systems. Therefore, in this paper we develop a family of randomized block Kaczmarz algorithms that uses at each step a subset of the constraints and extrapolated stepsizes, and can be deployed on distributed computing units. Our approach is based on several new ideas and tools, including stochastic selection rule for the blocks of rows, stochastic conditioning of the linear system, and novel strategies for designing extrapolated stepsizes. We prove that randomized block Kaczmarz algorithm converges linearly in expectation, with a rate depending on the geometric properties of the matrix and its submatrices and on the size of the blocks. Our convergence analysis reveals that the algorithm is most effective when it is given a good sampling of the rows into well-conditioned blocks. Besides providing a general framework for the design and analysis of randomized block Kaczmarz methods, our results resolve an open problem in the literature related to the theoretical understanding of observed practical efficiency of extrapolated block Kaczmarz methods.

Key words. Consistent linear systems, Kaczmarz algorithm, random blocks of rows, expected linear convergence. L^AT_EX

AMS subject classifications. 15A06 , 90C20, 90C06.

1. Introduction. Given a real matrix $A \in \mathbb{R}^{m \times n}$ and a real vector $b \in \mathbb{R}^m$, in this paper we search for a solution of the linear system $Ax = b$:

$$(1.1) \quad \text{Find } x \quad \text{s.t.} \quad Ax = b.$$

We assume throughout the paper that the system is consistent, that is there exists a vector $x^* \in \mathbb{R}^n$ for which $Ax^* = b$. Let us denote the set of solutions by $\mathcal{X} = \{x \in \mathbb{R}^n : Ax = b\}$. Linear systems represent a modeling paradigm for solving many engineering and physics problems: partial differential equations [19], sensor networks [27], filtering [11], signal processing [7], computerized tomography [9], machine learning and optimal control [20]. In these applications it is usually sufficient to find a point which is not too far from the solution set \mathcal{X} . In particular, one chooses the error tolerance $\varepsilon > 0$ and aims to find a point x satisfying $\|x - \Pi_{\mathcal{X}}(x)\|^2 \leq \varepsilon$, where $\Pi_{\mathcal{X}}(\cdot) = \arg \min_{y \in \mathcal{X}} \|\cdot - y\|$ is the projection function onto solution set \mathcal{X} , and $\|\cdot\|$ is the standard Euclidean norm on \mathbb{R}^n . In the case when a randomized algorithm is used to find x , which renders x a random vector, one replace this condition with $\mathbf{E} [\|x - \Pi_{\mathcal{X}}(x)\|^2] \leq \varepsilon$, where $\mathbf{E} [\cdot]$ denotes the expectation with respect to the randomness of the algorithm.

1.1. Iterative methods. In practice, m and n are usually large so that iterative methods, e.g. the so-called row-action methods are preferred (in a row-action method only one block of rows of A is used in a certain iteration [2]). One of these methods is the iterative method of Kaczmarz [10, 23, 13]. In some situations, it is even more efficient than the conjugate gradient method, which is the most popular iterative algorithm for solving large linear systems [19]. In fact Kaczmarz algorithm was implemented by Hounsfield in the very first medical scanner [9]. At each step, the Kaczmarz algorithm projects the current iterate onto the solution space of a single row $a_{i_k}^T$ and then choose the next iterate along the line connecting the current iterate and the projection, leading to the following iterative process:

$$(1.2) \quad x^{k+1} = x^k - \alpha_k \frac{a_{i_k}^T x^k - b_{i_k}}{\|a_{i_k}\|^2} a_{i_k}.$$

Usually, the stepsize α_k is chosen in the interval $(0, 2)$. For $\alpha_k = 1$ we recover the basic Kaczmarz algorithm [10]. Note that this update rule requires low cost per iteration and storage of order $\mathcal{O}(n)$. In

*Submitted to the editors DATE.

Funding: This work was supported by the Executive Agency for Higher Education, Research and Innovation Funding (UEFIS-CDI), Romania, PNIII-P4-PCE-2016-0731, project ScaleFreeNet, no. 39/2017. The author thanks Yu. Nesterov and F. Glineur from Universite Catholique de Louvain for useful discussions on the Chebyshev-based Kaczmarz scheme.

[†]Department of Automatic Control and Systems Engineering, University Politehnica Bucharest, Splaiul Independentei 313, Bucharest, 060042, Romania (ion.necoara@acse.pub.ro).

contrast, in *block* Kaczmarz methods a subset of rows A_{J_k} are used at each iteration, with $J_k \subseteq [m]$ and $|J_k| > 1$. We usually distinguish two approaches. The first variant is simply a block generalization of basic Kaczmarz algorithm, that is, we project the current iterate onto the solution space of the *entire* block A_{J_k} and then choose the next iterate along the line connecting the current iterate and the projection:

$$(1.3) \quad x^{k+1} = x^k - \alpha_k A_{J_k}^\dagger (A_{J_k} x^k - b_{J_k}),$$

where $A_{J_k}^\dagger$ denotes the pseudoinverse of A_{J_k} . Usually, the stepsize α_k is chosen 1. This is the approach followed e.g. in [5, 8, 16, 22] and we refer to this iterative process as *block projection* Kaczmarz algorithm. The main drawback of (1.3) is that each iteration is expensive, since we need to apply the pseudoinverse to a vector, or equivalently, we must solve a least-squares problem at each iteration, having cost per iteration of order $\mathcal{O}(\tau^2 n)$, where $\tau = |J_k|$. Moreover, it is not adequate for distributed implementations. The second variant of block Kaczmarz avoids these issues, by projecting the current estimate onto *each individual* row that forms the block matrix A_{J_k} , and the resulting projections are averaged to form the next iterate. This leads to the following iteration:

$$(1.4) \quad x^{k+1} = x^k - \alpha_k \left(\sum_{i \in J_k} \omega_i \frac{a_i^T x^k - b_i}{\|a_i\|^2} a_i \right),$$

where the weights $\omega_i \in [0, 1]$ such that $\sum_{i \in J_k} \omega_i = 1$, and $\alpha_k \in (0, 2)$. Note that update (1.4) is very easy to implement on distributed computing units and it is comparable in terms of cost per iteration with the basic Kaczmarz update (1.2), i.e., of order $\mathcal{O}(\tau n)$. This is the scheme considered e.g. in [1, 2, 14, 21] and we also analyze it in this paper and refer to it as *block Kaczmarz algorithm*. Assuming $\alpha_k \in (0, 2)$, then the iterative process (1.2) is known to converge linearly [13, 23] (see also Section 3.3). Moreover, linear convergence results for the iteration (1.3), with particular stepsize $\alpha_k = 1$, were recently derived in [8, 16, 22]. However, we are not aware of any convergence rates depending on the size of the blocks $|J_k|$ and the geometric properties of the matrix A and its submatrices A_{J_k} for the iterative process (1.4).

1.2. Extrapolation. It is well known that the practical performance of block Kaczmarz method (1.4) can be enhanced, and often dramatically so, using *extrapolation*. This refers to the practice of moving *further* along the line connecting the last iterate and the average of the projections by using a stepsize $\alpha_k \geq 2$, see e.g. [1]. For example, since the iterative process (1.4) can be slow, in [14, 21] an extrapolated variant of (1.4) has been introduced with the following *adaptive* choice for the stepsize α_k :

$$(1.5) \quad \alpha_k = \frac{2 \sum_{i \in J_k} \bar{\omega}_i (a_i^T x^k - b_i)^2}{\| \sum_{i \in J_k} \bar{\omega}_i (a_i^T x^k - b_i) a_i \|^2},$$

where we use the notation $\bar{\omega}_i = \omega_i / \|a_i\|^2$ and, for convenience, we define $0/0 = 1$. From Jensen's inequality it follows that $\alpha_k \geq 2$. However, in numerical experiments, it has been observed that the extrapolation parameter α_k from (1.5) can be much larger than 2. Moreover, the sequence x^k generated by the iterative process (1.4) using the extrapolated adaptive stepsize α_k from (1.5) usually converges much faster than the same sequence x^k from (1.4) but generated with stepsize $\alpha_k \in (0, 2)$ [1, 2, 3, 14, 21]. However, despite more than 80 years of research on block Kaczmarz methods, the empirical success of extrapolation schemes is not supported by theory. That is, to the best of our knowledge, there is no theory explaining why these methods with $\alpha_k \geq 2$ require less iterations than their non-extrapolated variants $\alpha_k = 1$.

1.3. Rows importance. While selecting the index set $J \subseteq [m]$ uniformly random appears as the most natural choice, it is likely the case that some blocks of rows of A are more important than others. As an illustration, consider the scenario in which there exists $T \subset [m]$ such that $\mathcal{X} = \{x \in \mathbb{R}^n : A_T x = b_T\}$, where A_T denotes the block matrix of A whose rows are indexed in the set T . Clearly, the rows a_i for $i \in T$ are more important than the rows a_i for $i \notin T$. This is an extreme scenario: if T is known, one should simply remove the non-important rows from the representation to begin with. However, even if none of the rows can be removed, it is often the case that some (blocks of) rows are more important than others in the

sense that one should project on these more often. In fact, the operator theory shows that some sampling strategies of the blocks of rows are more effective than others, in terms of conditioning, see e.g. [16, 25]. We are not aware of any paper on block Kaczmarz method (1.4) that take importance of blocks of rows into consideration. An exception to this are some recent works [16, 8, 22], but on the block projection Kaczmarz algorithm (1.3) (i.e., [16, 8, 22] analyze rows importance for the method that projects the current estimate on the entire solution space of $A_J x = b_J$, as opposed to our algorithm (1.4), where we only project on the individual rows of the submatrix A_J and then average).

1.4. Outline. In Section 2 we summarize selected key contributions of this paper. In Section 3 we present some preliminary results for Kaczmarz algorithm. In Section 4 we define general random block Kaczmarz algorithms and derive new convergence rates. In Section 5 we present an acceleration of block Kaczmarz algorithm using Chebyshev-based stepsizes and derive the corresponding convergence rates.

1.5. Notation. For $x \in \mathbb{R}^n$, the standard Euclidean norm is denoted by $\|x\| = \sqrt{x^T x}$. For a positive integer m , let $[m] = \{1, 2, \dots, m\}$. By e_i we denote the i th column of the identity matrix $I_n \in \mathbb{R}^{n \times n}$. Let $A \in \mathbb{R}^{m \times n}$ be a matrix. By $\|A\|_F$, $\|A\|$, $\text{rank}(A)$, a_i^T , $\lambda_{\min}^{\text{nz}}(A)$ and $\lambda_{\max}(A)$ we denote its Frobenius norm, spectral norm, rank, i th row, the smallest non-zero eigenvalue, and the largest eigenvalue, respectively. For an index set $J \subset [m]$, by $A_J \in \mathbb{R}^{|J| \times n}$ we denote the matrix with the rows a_i^T for $i \in J$. The projection of a point x onto a closed convex set X is denoted by $\Pi_X(x) = \arg \min_z \{\|x - z\| : z \in X\}$. A matrix is called *normalized* if all its rows have the Euclidean norm equal to 1.

2. Contributions. In this section we briefly review our key contributions and results, leaving the theoretical details to the rest of the paper.

2.1. General framework. We develop a unified framework for studying extrapolation and rows importance questions for consistent linear systems, together with randomized block Kaczmarz methods for solving such systems of linear equalities. We define a probability space $([m], \mathcal{F}, \mathbf{P})$. By sampling $J \sim \mathbf{P}$, we are choosing a block of rows A_J from the matrix A . In this way we achieve two goals at the same time:

(i) First, this sampling defines a general *stochastic selection rule* which we shall use to design a *randomized block Kaczmarz method*, described in Section 2.2 below.

(ii) Second, the choice of probability measure is a natural way to assign *importance* to the blocks of A . Note that the probability \mathbf{P} is a *parameter* playing the dual role of controlling the representation of the solution set \mathcal{X} as an intersection of blocks of rows of matrix A , and defining the importance sampling procedure, which in turn defines the algorithm. For matrices with normalized rows (i.e. each row has norm 1), we have identified the following *stochastic conditioning* parameter:

$$(2.1) \quad \lambda_{\max}^{\text{block}} = \max_{J \sim \mathbf{P}} \lambda_{\max}(A_J^T A_J)$$

as the key quantity characterizing importance sampling. In particular, our analysis reveals that the most effective importance rule is the one that makes $\lambda_{\max}^{\text{block}}$ small, i.e. there is a sampling of the blocks of the rows into well-conditioned blocks. Moreover, the operator theory literature provides detailed information about the existence and construction of such good sampling (see Section 4.3).

2.2. Algorithms. We propose a block Kaczmarz algorithmic framework that uses a randomized scheme to choose a subset of the constraints at each iteration (see Sections 4 and 5):

Draw at each step a sample $J_k \sim \mathbf{P}$ and update:

$$(\text{RBK}) : \quad x^{k+1} = x^k - \alpha_k \left(\sum_{i \in J_k} \omega_i^k \frac{a_i^T x^k - b_i}{\|a_i\|^2} a_i \right),$$

where the weights satisfy $\omega_i^k \in [0, 1]$ such that $\sum_{i \in J_k} \omega_i^k = 1$. One important property of our algorithmic framework is the use of several *extrapolated* stepsizes α_k , that, in general, are much larger than the stepsize $\alpha_k \in (0, 2)$ usually used in the literature. More precisely, we analyze three choices for the stepsize α_k : (i) one depending on the geometric properties of the submatrices of A of the form $\mathcal{O}(1/\lambda_{\max}^{\text{block}})$; (ii) one

RBK algorithm	Convergence rates	Remarks
constant stepsize normalized A	$\mathbf{E} [\ x^k - x_k^*\ ^2] \leq \left(1 - \frac{\tau}{m} \frac{\lambda_{\min}^{\text{nz}}}{\lambda_{\max}^{\text{block}}}\right)^k \ x^0 - x_0^*\ ^2$	Theorem 4.1
adaptive stepsize normalized A	$\mathbf{E} [\ x^k - x_k^*\ ^2] \leq \left(1 - \frac{\tau}{m} \frac{\lambda_{\min}^{\text{nz}}}{\lambda_{\max}^{\text{block}}}\right)^k \ x^0 - x_0^*\ ^2$	Theorem 4.2
Chebyshev stepsize normalized $A \& \lambda_{\min} > 0$	$\ \mathbf{E} [x^k - x_k^*]\ ^2 \leq \left(1 - \sqrt{\frac{\lambda_{\min}}{\lambda_{\max}}}\right)^{2k} \ x^0 - x_0^*\ ^2$	Theorem 5.1

TABLE 1

The key convergence results obtained in this paper for algorithm RBK for the three choices of the extrapolated stepsize. Here, matrix A is normalized and λ_{\max} and $\lambda_{\min}(\lambda_{\min}^{\text{nz}})$ denote the largest and smallest (non-zero) eigenvalue of AA^T , respectively.

adaptive stepsize similar to (1.5); (iii) one stepsize using the roots of the Chebyshev polynomials. All three extrapolation procedures yield $\alpha_k \geq 2$ and hence they accelerate drastically the convergence of RBK algorithm. Another feature of our algorithm is that it allows to project in *parallel* onto several rows, thus providing flexibility in matching the implementation of the algorithm on the distributed architecture at hand. Moreover, RBK algorithm can be interpreted, for some particular choices of the weights and stepsize, as a *minibatch* stochastic gradient descent or *block* coordinate descent method applied to a specific optimization problem.

2.3. Convergence rates. To the best of our knowledge, convergence rates of Kaczmarz type methods were only previously derived for stepsizes belonging to the interval $(0, 2)$ [8, 16, 22, 23]. Moreover, the existing convergence estimates for block Kaczmarz algorithm (1.4) do not show any dependence on the size of the blocks $|J|$ or on the geometric properties of the block submatrices A_J [1, 2, 3, 14, 21]. On the other hand, our convergence analysis for the randomized block Kaczmarz (RBK) algorithm is one of the first proving an (expected) linear rate of convergence that is expressed explicitly in terms of the geometric properties of the matrix and its submatrices and of the size of the blocks. Moreover, our analysis allows to derive convergence estimates for all three choices of the extrapolated stepsize. *From our knowledge, this is the first time the randomized block Kaczmarz algorithm with extrapolation ($|J| > 1$ and $\alpha_k > 2$) is shown to have a better convergence rate than its basic variant (1.2) ($|J| = 1$ and $\alpha_k = 1$).* We have identified $\lambda_{\max}^{\text{block}}$ as the key quantity determining whether extrapolation helps or not, and how much (the smaller $\lambda_{\max}^{\text{block}}$, the more it helps). For example, for normalized matrices, RBK with the extrapolation rules (i)–(ii) has an expected linear rate for the square distance of the iterates to the optimal solution set of the form (see Table 1):

$$\mathcal{O} \left(\frac{m \lambda_{\max}^{\text{block}}}{\tau \lambda_{\min}^{\text{nz}}} \log \frac{1}{\varepsilon} \right),$$

where $\lambda_{\min}^{\text{nz}}$ denotes the smallest non-zero eigenvalue of AA^T . Thus, a convergence rate depending on the geometric properties of the matrix A and its submatrices A_J and on the size of the blocks $\tau = |J|$. When comparing RBK with basic Kaczmarz in terms of total computational cost to achieve an ε solution we get:

$$\mathcal{O} \left(\tau n \cdot \frac{m \lambda_{\max}^{\text{block}}}{\tau \lambda_{\min}^{\text{nz}}} \log \frac{1}{\varepsilon} \right) \quad \text{vs.} \quad \mathcal{O} \left(n \cdot \frac{m}{\lambda_{\min}^{\text{nz}}} \log \frac{1}{\varepsilon} \right).$$

Therefore, our convergence rate also explains *why* and *when* the randomized block Kaczmarz algorithm with the constant extrapolated stepsize (4.2) or adaptive extrapolated stepsize (1.5) works better compared to its basic counterpart. In particular, the analysis reveals that a distributed implementation of extrapolated RBK algorithm is most effective when the sampling of the blocks of rows yields a partition into well-conditioned blocks, that is, the stochastic conditioning parameter $\lambda_{\max}^{\text{block}}$ is small.

For the third choice of the extrapolated stepsize, depending on the roots of Chebyshev polynomials, and for normalized matrices having $\lambda_{\min} > 0$ we get a linear rate for the expected iterates of the form (see Table 1):

$$\mathcal{O} \left(\sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}} \log \frac{1}{\varepsilon} \right),$$

where $\lambda_{\min}(\lambda_{\max})$ denote the smallest (largest) eigenvalue of AA^T , respectively. Note that this convergence estimate is the same as for the conjugate gradient method and it is optimal for this class of iterative schemes, as the condition number of the matrix is square rooted.

3. Preliminaries. Note that the problem of finding a solution of the linear system $Ax = b$ can be posed as a quadratic optimization problem, the so-called linear least-square problem:

$$(3.1) \quad \min_{x \in \mathbb{R}^n} \frac{1}{2m} \|Ax - b\|^2 \quad \left(:= \frac{1}{2m} \sum_{i=1}^m (a_i^T x - b_i)^2 \right).$$

A more particular formulation is to find the least-norm solution of the linear system:

$$(3.2) \quad \min_{x \in \mathbb{R}^n} \frac{1}{2} \|x\|^2 \quad \text{s.t.} \quad Ax = b.$$

The dual of optimization problem (3.2) takes also the form of a quadratic program:

$$(3.3) \quad \min_{y \in \mathbb{R}^m} \frac{1}{2} \|A^T y\|^2 - b^T y,$$

where the primal variable x and the dual variable y are related through the relation $x = A^T y$. Let us define the primal and dual objective functions $f(x) = (1/2m) \|Ax - b\|^2$ and $g(y) = 1/2 \|A^T y\|^2 - b^T y$, respectively. Recall that the set of solutions is denoted $\mathcal{X} = \{x \in \mathbb{R}^n : Ax = b\}$ and for any given x we define its projection onto \mathcal{X} by $x^* = \Pi_{\mathcal{X}}(x)$.

3.1. Basic Kaczmarz algorithm. The Kaczmarz algorithm is an iterative scheme for solving the linear system $Ax = b$ that requires only $\mathcal{O}(n)$ cost per iteration and storage and has a linear rate of convergence. At each iteration k , the algorithm selects (cyclically, randomly) a row $i_k \in [m]$ of the linear system and does an orthogonal projection of the current estimate vector x^k onto the corresponding hyperplane $a_{i_k}^T x = b_{i_k}$:

$$\min_x \|x - x^k\|^2 \quad \text{s.t.} \quad a_{i_k}^T x = b_{i_k}.$$

Then, we choose the next iterate along the line connecting the current iterate and the projection. This leads to the following iteration for randomized/cyclic Kaczmarz algorithm [10, 23]:

Algorithm 3.1 (Algorithm Kaczmarz)

- 1: choose $x^0 \in \mathbb{R}^n$
 - 2: **for** $k \geq 0$ **do**
 - 3: choose an index $i_k \in [m]$ (random, cyclic) and update:
 - 4: $x^{k+1} = x^k - \alpha_k \frac{a_{i_k}^T x^k - b_{i_k}}{\|a_{i_k}\|^2} a_{i_k}$.
 - 5: **end for**
-

Usually, α_k is chosen constant in interval $(0, 2)$. For $\alpha_k = 1$ we recover basic Kaczmarz algorithm [10].

3.2. Interpretations. We can view randomized Kaczmarz algorithm, i.e. when i_k is chosen randomly, as an optimization method for solving a specific primal or dual optimization problem. More precisely, Kaczmarz algorithm is a particular case of:

SGD (Stochastic Gradient Descent): The randomized Kaczmarz (Algorithm 3.1) is equivalent to one step of the stochastic gradient descent method [17] applied to the finite sum problem (3.1). Specifically, a component function i_k , $f_{i_k}(x) = 1/2(a_{i_k}^T x - b_{i_k})^2$, is chosen randomly and a negative gradient step (having $\nabla f_{i_k}(x) = (a_{i_k}^T x - b_{i_k})a_{i_k}$) of this partial function in x^k with stepsize $\alpha_k/\|a_{i_k}\|^2$ is considered:

$$x^{k+1} = x^k - \frac{\alpha_k}{\|a_{i_k}\|^2} \nabla f_{i_k}(x^k).$$

RCD (Random Coordinate Descent): The randomized Kaczmarz (Algorithm 3.1) is equivalent to one step of randomized coordinate descent method [18] applied to the dual problem (3.3). Specifically, a negative gradient step in the random i_k th component of y (having the expression $\nabla_{i_k} g(y) = a_{i_k}^T A^T y - b_{i_k}$) with stepsize $\alpha_k / \|a_{i_k}\|^2$ is taken, yielding:

$$y^{k+1} = y^k - \frac{\alpha_k}{\|a_{i_k}\|^2} \nabla_{i_k} g(y^k) \cdot e_{i_k},$$

where e_i denotes the i th column of the identity matrix in $\mathbb{R}^{n \times n}$. We recover easily the iteration of Algorithm 3.1 by simply multiplying this update with A^T and using the relation between the primal and dual variables given by $x = A^T y$. Note that in both interpretations, we need to choose a specific stepsize, in order to prove convergence, see [17, 18].

3.3. Convergence properties. It is known that Algorithm 3.1 converges to the minimum norm solution 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 [4]. The situation changes if one considers a randomization such that in each step one chooses a row of the system matrix at random, according to a probability \mathbf{P} . In the seminal paper [23] it has been shown that sampling the rows of A with probability $\mathbf{P}(i = i_k) = \frac{\|a_{i_k}\|^2}{\|A\|_F^2}$ for all $i \in [m]$ and using constant stepsize $\alpha = 1$, we get a linear convergence rate in expectation of the form:

$$(3.4) \quad \mathbf{E} [\|x^k - x_k^*\|^2] \leq \left(1 - \frac{\lambda_{\min}^{\text{nz}}(A^T A)}{\|A\|_F^2}\right)^k \|x^0 - x_0^*\|^2,$$

where $\lambda_{\min}^{\text{nz}}(\cdot)$ denotes the minimum non-zero eigenvalue of a given matrix and $x_k^* = \Pi_{\mathcal{X}}(x^k)$. For completeness, let us derive this convergence rate. Considering the stepsize α_k constant in the interval $(0, 2)$ and using that $\langle x - x^*, (a_i^T x - b_i) a_i \rangle = (a_i^T x - b_i)^2$ for any x^* a solution of $Ax = b$, we get:

$$\begin{aligned} \|x^{k+1} - x^*\|^2 &= \|x^k - x^*\|^2 - 2\alpha \frac{(a_i^T x^k - b_i)^2}{\|a_i\|^2} + \alpha^2 \frac{(a_i^T x^k - b_i)^2}{\|a_i\|^2} \\ &= \|x^k - x^*\|^2 - \alpha(2 - \alpha) \frac{(a_i^T x^k - b_i)^2}{\|a_i\|^2}. \end{aligned}$$

Taking now the conditional expectation under the probability $\mathbf{P}(i = i_k) = \frac{\|a_{i_k}\|^2}{\|A\|_F^2}$, we get:

$$\mathbf{E}_i [\|x^{k+1} - x^*\|^2 | x^k] \leq \|x^k - x^*\|^2 - \frac{\alpha(2 - \alpha)}{\|A\|_F^2} \|Ax^k - b\|^2.$$

Further, it is well known from the Courant-Fischer theorem that for any matrix A we have $\|Ax\|^2 \geq \lambda_{\min}^{\text{nz}}(AA^T)\|x\|^2$ for all $x \in \text{range}(A^T)$. Moreover, we have that $x - \Pi_{\mathcal{X}}(x) \in \text{range}(A^T)$ for any x . In conclusion, if we denote $x_k^* = \Pi_{\mathcal{X}}(x^k)$, we get:

$$\|Ax^k - b\|^2 = \|A(x^k - x_k^*)\|^2 \geq \lambda_{\min}^{\text{nz}}(AA^T)\|x^k - x_k^*\|^2.$$

Using this inequality in the recurrence above and taking expectation over the entire history we get the following linear convergence rate in expectation:

$$(3.5) \quad \mathbf{E} [\|x^{k+1} - x_k^*\|^2] \leq \left(1 - \frac{\alpha(2 - \alpha)\lambda_{\min}^{\text{nz}}(AA^T)}{\|A\|_F^2}\right) \mathbf{E} [\|x^k - x_k^*\|^2].$$

For the optimal choice $\alpha^* = 1$ (i.e. $\alpha^* = \arg \max_{\alpha} \alpha(2 - \alpha)$) we get the simpler convergence estimate (3.4) derived in [23]. Note that for ill-conditioned problems, i.e. $\lambda_{\min}^{\text{nz}}(AA^T)$ small and $\|A\|_F$ large, this linear convergence is very slow using a constant stepsize $\alpha \in (0, 2)$. In the next sections we prove that block variants of randomized Kaczmarz (Algorithm 3.1) with properly chosen extrapolated stepsize α_k larger than 2 can substantially accelerate the convergence rate (3.5).

3.4. Preliminary probability results. Let J be a random set-valued map with values in $2^{[m]}$. Any realization $J \subseteq [m]$ of this random variable, referred to as sampling and having the same notation as the random variable, is characterized by the probability distribution $\mathbf{P}(J)$. We also define the probability with which an index $i \in [m]$ can be found in J as:

$$p_i = \mathbf{P}(i \in J).$$

Then, for any scalars θ_i , with $i \in [m]$, the following relation holds in expectation:

$$(3.6) \quad \mathbf{E}_J \left[\sum_{i \in J} \theta_i \right] = \sum_{J \subseteq [m]} \left(\sum_{i \in J} \theta_i \right) \mathbf{P}(J) = \sum_{i \in [m]} \theta_i \left(\sum_{J: i \in J} \mathbf{P}(J) \right) = \sum_{i \in [m]} p_i \theta_i.$$

The following examples for sampling blocks of rows of $A \in \mathbb{R}^{m \times n}$ will be used in our subsequent analysis. *Uniform sampling*: One natural choice is the *uniform* sampling of τ unique indexes of rows that make up J , i.e. $|J| = \tau$ for all samplings, with $1 \leq \tau \leq m$ fixed. For this choice of the random variable J , we observe that we have a total number of $\binom{m}{\tau}$ possible values that J can take. Thus, for the uniform sampling we have $\mathbf{P}(J) = 1/\binom{m}{\tau}$. We can also express p_i for the uniform sampling as:

$$(3.7) \quad p_i = \mathbf{P}(i \in J) = \sum_{J: i \in J} \mathbf{P}(J) = \frac{\binom{m-1}{\tau-1}}{\binom{m}{\tau}} = \frac{\tau}{m}.$$

Partition sampling: Another choice is the *partition* sampling, i.e. consider a partition of $[m]$ given by $\{J_1, \dots, J_\ell\}$, and then take $\mathbf{P}(J) = 1/\ell$ or $\mathbf{P}(J) = \|A_J\|_F^2/\|A\|_F^2$ for all $J \in \{J_1, \dots, J_\ell\}$. For example, for the first probability choice of the partition sampling, p_i is given by:

$$(3.8) \quad p_i = \frac{1}{\ell}.$$

In particular, if all the subsets in the partition have the same cardinality, i.e. $|J_l| = \tau$ for all $l \in [\ell]$, and A is normalized, then the two probabilities are the same and $\ell = m/\tau$. Hence, $p_i = \frac{\tau}{m}$. These preliminary results will help us in the convergence analysis of randomized block Kaczmarz algorithms we propose next.

4. Randomized block Kaczmarz algorithms. In this section we design new variants of randomized Kaczmarz, Algorithm (3.1), considering at each step a block of rows of the linear system $Ax = b$ and different choices for the stepsize. For all these methods we prove expected linear convergence rates. Note that block Kaczmarz methods have been also considered in other works, see e.g. [1, 2, 14, 21] and the references therein. Nevertheless, to our knowledge, this paper is the first one that provides an expected linear rate of convergence that depends explicitly on geometric properties of the system matrix A and its submatrices A_J . Moreover, the convergence estimates hold for several extrapolated stepsizes. In our Randomized Block Kaczmarz (RBK) algorithm, at each iteration, instead of projecting on only one hyperplane, we consider projections onto several hyperplanes and then take as a new direction a convex combination of these projections with some stepsize (see Algorithm 4.1). Here $J_k = \{i_k^1, \dots, i_k^{\tau_k}\} \subseteq [m]$ is the set of indexes corresponding

Algorithm 4.1 (Algorithm RBK)

- 1: choose $x^0 \in \mathbb{R}^n$, stepsize sequence $(\alpha_k)_{k \geq 0}$, and weights sequence $(\omega_k)_{k \geq 0}$
 - 2: **for** $k \geq 0$ **do**
 - 3: draw sample $J_k \sim \mathbf{P}$ and update:
 - 4: $x^{k+1} = x^k - \alpha_k \left(\sum_{i \in J_k} \omega_k^i \frac{a_i^T x^k - b_i}{\|a_i\|^2} a_i \right)$.
 - 5: **end for**
-

to the rows selected at iteration k of size $\tau_k \in [1, m]$ and \mathbf{P} denotes the probability distribution over the

collection of subsets of indexes of $[m]$. Moreover, the weights $\omega_k = (\omega_k^i)_{i \in J_k}$ are chosen positive and summing to 1. Thus, in our analysis we assume bounded weights satisfying $0 < \omega_{\min} \leq \omega_k^i \leq \omega_{\max} < 1$ for all $i \in J_k$ and $k \geq 0$. Two simple choices for the weights are e.g. $\omega_k^i = \|a_i\|^2 / \sum_{i \in J_k} \|a_i\|^2$ or $\omega_k^i = 1/\tau_k$ for all $k \geq 0$. In these two particular cases we get the following compact updates:

$$x^{k+1} = x^k - \alpha_k \frac{A_{J_k}^T (A_{J_k} x^k - b_{J_k})}{\|A_{J_k}\|_F^2} \quad \text{or} \quad x^{k+1} = x^k - \alpha_k \frac{A_{J_k}^T D_{J_k} (A_{J_k} x^k - b_{J_k})}{\tau_k},$$

respectively, where the diagonal matrix $D_J = \text{diag}(1/\|a_i\|^2, i \in J) \in \mathbb{R}^{\tau \times \tau}$. Several choices for the stepsize will be given in the next sections, based on over-relaxations (extrapolations), i.e. $\alpha_k > 2$. Similarly, as for Kaczmarz algorithm, RBK (Algorithm 4.1) can be interpreted as:

BSGD (Batch Stochastic Gradient Descent): One iteration of RBK algorithm can be viewed as one step of the batch stochastic gradient descent [17] applied to the finite sum problem (3.1) when the weights ω_k are chosen in a particular fashion. Specifically, if we choose the particular weights $\omega_k^i = \|a_i\|^2 / \sum_{i \in J_k} \|a_i\|^2$ and uniform probability, then we recover the batch stochastic gradient descent method with a certain choice of the stepsize:

$$x^{k+1} = x^k - \frac{\tau_k \alpha_k}{\sum_{i \in J_k} \|a_i\|^2} \left(\frac{1}{\tau_k} \sum_{i \in J_k} (a_i^T x^k - b_i) a_i \right).$$

RBCD (Randomized Block Coordinate Descent): One iteration of RBK algorithm can be viewed as one step of the block coordinate descent method [15, 18] applied to the dual problem (3.3) when the weights ω_k are chosen in a particular fashion. Specifically, if we choose the particular weights $\omega_k^i = \|a_i\|^2 / \sum_{i \in J_k} \|a_i\|^2$, then we recover the block coordinate descent method with a certain choice of the stepsize:

$$x^{k+1} = x^k - \frac{\alpha_k}{\sum_{i \in J_k} \|a_i\|^2} \left(\sum_{i \in J_k} (a_i^T x^k - b_i) a_i \right).$$

However, for general weights ω_k and stepsize α_k , RBK algorithm cannot be interpreted in these ways, thus our scheme is more general. In the following, we denote $x_k^* = \Pi_{\mathcal{X}}(x^k)$, that is the projection of x^k onto the solution set \mathcal{X} of the linear system $Ax = b$.

4.1. Randomized block Kaczmarz algorithm with constant stepsize. In this section we investigate the convergence rate of RBK algorithm for constant extrapolated stepsize $\alpha_k = \alpha > 2$ and weights $\omega_k^i = \omega_i$ for all k . Thus, the iteration of RBK (Algorithm 4.1) becomes in this case:

$$(4.1) \quad x^{k+1} = x^k - \alpha \left(\sum_{i \in J_k} \omega_i \frac{a_i^T x^k - b_i}{\|a_i\|^2} a_i \right).$$

The weights are chosen to satisfy $0 < \omega_{\min} \leq \omega_i \leq \omega_{\max} < 1$ for all i and sum to 1. Let us also define the following stochastic conditioning parameter depending on the geometric properties of the submatrices A_J :

$$\lambda_{\max}^{\text{block}} = \max_{J \sim \mathbf{P}} \lambda_{\max} \left(A_J^T \text{diag} \left(\frac{1}{\|a_i\|^2}, i \in J \right) A_J \right).$$

Then, we consider an extrapolated constant stepsize of the form:

$$(4.2) \quad 0 < \alpha < \frac{2\omega_{\min}}{\omega_{\max}^2 \lambda_{\max}^{\text{block}}}.$$

When we choose a random variable such that all the samplings satisfy $|J| = \tau$, with $\tau \in [1, m]$, then it is straightforward to see that $\lambda_{\max}^{\text{block}} < \tau$ provided that $\text{rank}(A_J) \geq 2$. Hence, in this case we use an over-relaxed (extrapolated) stepsize, since usually $2\omega_{\min}/\omega_{\max}^2 \lambda_{\max}^{\text{block}} > 2$. For example, for $\omega_i = 1/\tau$, we get

$2\tau/\lambda_{\max}^{\text{block}} > 2$. Using (3.6) we also define the positive semidefinite matrix W as:

$$W = \mathbf{E}_J \left[\sum_{i \in J} \frac{a_i a_i^T}{\|a_i\|^2} \right] = \sum_{i \in [m]} p_i \frac{a_i a_i^T}{\|a_i\|^2} = A^T \text{diag} \left(\frac{p_i}{\|a_i\|^2}, i \in [m] \right) A.$$

From our best knowledge, the choice (4.2) for the stepsize in the block Kaczmarz algorithm seems to be new. The next theorem proves the convergence rate of this algorithm which depends explicitly on the geometric properties of the system matrix A and its submatrices A_J .

THEOREM 4.1. *Let $\{x^k\}_{k \geq 0}$ be generated by RBK (Algorithm 4.1) with the particular update (4.1), i.e. the weights satisfy $0 < \omega_{\min} \leq \omega_i \leq \omega_{\max} < 1$ for all $i \in [m]$ and the stepsize $\alpha = \frac{(2-\delta)\omega_{\min}}{\omega_{\max}^2 \lambda_{\max}^{\text{block}}}$ for some $\delta \in (0, 1]$. Then, we have the following linear convergence rate in expectation:*

$$(4.3) \quad \mathbf{E} [\|x^k - x^*\|^2] \leq \left(1 - \frac{(2-\delta)\omega_{\min}^2 \lambda_{\min}^{nz}(W)}{\omega_{\max}^2 \lambda_{\max}^{\text{block}}} \right)^k \|x^0 - x^*\|^2.$$

Proof. Since we assume a consistent linear system, that is there is x^* such that $Ax^* = b$, we have:

$$\begin{aligned} \|x^{k+1} - x^*\|^2 &= \|x^k - x^* - \alpha \left(\sum_{i \in J_k} \omega_i \frac{a_i^T x^k - b_i}{\|a_i\|^2} a_i \right)\|^2 \\ &= \|x^k - x^* - \alpha \left(\sum_{i \in J_k} \omega_i \frac{a_i a_i^T}{\|a_i\|^2} (x^k - x^*) \right)\|^2 \\ &= \| \left(I_n - \alpha \left(\sum_{i \in J_k} \omega_i \frac{a_i a_i^T}{\|a_i\|^2} \right) \right) (x^k - x^*) \|^2 \\ &= (x^k - x^*)^T \left(I_n - 2\alpha \sum_{i \in J_k} \omega_i \frac{a_i a_i^T}{\|a_i\|^2} + \alpha^2 \left(\sum_{i \in J_k} \omega_i \frac{a_i a_i^T}{\|a_i\|^2} \right)^2 \right) (x^k - x^*). \end{aligned}$$

We need to take conditional expectation over J_k . However, for a general random sampling J we have from (3.6) that the expectation over the first sum from above yields the lower bound:

$$\begin{aligned} \mathbf{E}_J \left[\sum_{i \in J} \omega_i \frac{a_i a_i^T}{\|a_i\|^2} \right] &\succeq (\min_{i \in J} \omega_i) \mathbf{E}_J \left[\sum_{i \in J} \frac{a_i a_i^T}{\|a_i\|^2} \right] = \omega_{\min} \sum_{i \in [m]} p_i \frac{a_i a_i^T}{\|a_i\|^2} \\ &= \omega_{\min} A^T \text{diag} \left(\frac{p_i}{\|a_i\|^2}, i \in [m] \right) A = \omega_{\min} W. \end{aligned}$$

Thus, we obtained:

$$(4.4) \quad \mathbf{E}_J \left[\sum_{i \in J} \omega_i \frac{a_i a_i^T}{\|a_i\|^2} \right] \succeq \omega_{\min} W.$$

Moreover, using that for any $Q \succeq 0$ we have $Q^2 \preceq \lambda_{\max}(Q)Q$, the expectation over the second sum also

yields the following upper bound:

$$\begin{aligned}
\mathbf{E}_J \left[\left(\sum_{i \in J} \omega_i \frac{a_i a_i^T}{\|a_i\|^2} \right)^2 \right] &\preceq \mathbf{E}_J \left[\lambda_{\max} \left(\sum_{i \in J} \omega_i \frac{a_i a_i^T}{\|a_i\|^2} \right) \left(\sum_{i \in J} \omega_i \frac{a_i a_i^T}{\|a_i\|^2} \right) \right] \\
&\preceq (\max_{i \in J} \omega_i) \mathbf{E}_J \left[\lambda_{\max} \left(\sum_{i \in J} \frac{a_i a_i^T}{\|a_i\|^2} \right) \left(\sum_{i \in J} \omega_i \frac{a_i a_i^T}{\|a_i\|^2} \right) \right] \\
&\preceq (\max_{i \in J} \omega_i) \mathbf{E}_J \left[\lambda_{\max} \left(A_J^T \text{diag} \left(\frac{1}{\|a_i\|^2}, i \in J \right) A_J \right) \left(\sum_{i \in J} \omega_i \frac{a_i a_i^T}{\|a_i\|^2} \right) \right] \\
&\preceq (\max_{i \in J} \omega_i)^2 \lambda_{\max}^{\text{block}} \mathbf{E}_J \left[\sum_{i \in J} \frac{a_i a_i^T}{\|a_i\|^2} \right] = \omega_{\max}^2 \lambda_{\max}^{\text{block}} W,
\end{aligned}$$

where recall that $\lambda_{\max}^{\text{block}} = \max_{J \sim \mathbf{P}} \lambda_{\max} \left(A_J^T \text{diag} \left(\frac{1}{\|a_i\|^2}, i \in J \right) A_J \right)$. Therefore, taking conditional expectation w.r.t. the block J_k over entire history $\mathcal{F}_k = \{J_0, \dots, J_{k-1}\}$ in the recurrence above, we get:

$$\mathbf{E}_J [\|x^{k+1} - x^*\|^2 | \mathcal{F}_k] \leq (x^k - x^*)^T (I_n - 2\alpha\omega_{\min} W + \alpha^2 \omega_{\max}^2 \lambda_{\max}^{\text{block}} W) (x^k - x^*).$$

In order to ensure decrease we need $2\alpha\omega_{\min} - \alpha^2 \omega_{\max}^2 \lambda_{\max}^{\text{block}} \geq 0$, that is we get an extrapolated stepsize:

$$\alpha \leq \frac{2\omega_{\min}}{\omega_{\max}^2 \lambda_{\max}^{\text{block}}},$$

and the optimal stepsize is obtained by maximizing $2\alpha\omega_{\min} - \alpha^2 \omega_{\max}^2 \lambda_{\max}^{\text{block}}$ in α which leads to:

$$\alpha^* = \frac{\omega_{\min}}{\omega_{\max}^2 \lambda_{\max}^{\text{block}}}.$$

Hence, taking stepsize $\alpha = (2 - \delta)\omega_{\min}/\omega_{\max}^2 \lambda_{\max}^{\text{block}}$ for some $\delta \in (0, 1]$, we get:

$$\mathbf{E}_J [\|x^{k+1} - x^*\|^2 | \mathcal{F}_k] \leq (x^k - x^*)^T \left(I_n - (2 - \delta) \frac{\omega_{\min}^2}{\omega_{\max}^2 \lambda_{\max}^{\text{block}}} W \right) (x^k - x^*).$$

On the other hand, it is well-known from the Courant-Fischer theorem that for any matrix A we have $\|Ax\|^2 \geq \lambda_{\min}^{\text{nz}}(AA^T)\|x\|^2$ for all $x \in \text{range}(A^T)$. Moreover, we have that $x - \Pi_{\mathcal{X}}(x) \in \text{range}(A^T)$ for any x . In conclusion, using that $W = A^T D A$ with the diagonal matrix $D = \text{diag} \left(\frac{p_i}{\|a_i\|^2}, i \in [m] \right)$ invertible, we get that:

$$\begin{aligned}
(x^k - x^*)^T W (x^k - x^*) &= \|D^{1/2} A (x^k - x^*)\|^2 \geq \lambda_{\min}^{\text{nz}}(A^T D A) \|x^k - x^*\|^2 \\
&= \lambda_{\min}^{\text{nz}}(W) \|x^k - x^*\|^2.
\end{aligned}$$

Using this inequality in the recurrence above and taking expectation over the entire history we get:

$$\mathbf{E} [\|x^{k+1} - x_{k+1}^*\|^2] \leq \left(1 - (2 - \delta) \frac{\omega_{\min}^2 \lambda_{\min}^{\text{nz}}(W)}{\omega_{\max}^2 \lambda_{\max}^{\text{block}}} \right) \mathbf{E} [\|x^k - x_k^*\|^2],$$

which shows an expected linear convergence rate for RBK depending on the parameters $\lambda_{\min}^{\text{nz}}(W)$ and $\lambda_{\max}^{\text{block}}$ associated to the system matrix A and its submatrices A_J , respectively. \square

Now, let us consider the uniform and partition sampling examples of Section 3.4 where all the blocks sampling have the same size $|J| = \tau$. In this case we have $p_i = \frac{\tau}{m}$. Let us also consider the particular choices

$\delta = 1$, weights $\omega_i = 1/\tau$, and matrices A with normalized rows, i.e. $\|a_i\| = 1$ for all $i \in [m]$. Hence, $\|A\|_F^2 = m$. Then, our convergence rate (4.3) becomes:

$$(4.5) \quad \mathbf{E} [\|x^k - x_k^*\|^2] \leq \left(1 - \frac{\tau}{\lambda_{\max}^{\text{block}}} \frac{\lambda_{\min}^{\text{nz}}(A^T A)}{m}\right)^k \|x^0 - x_0^*\|^2.$$

Comparing with the convergence rate (3.4) of the basic Kaczmarz method (recall that for normalized matrices $\|A\|_F^2 = m$) we get an improvement $\frac{\tau}{\lambda_{\max}^{\text{block}}} > 1$, which shows that for RBK algorithm with the new extrapolated stepsize (4.2) we can get a speed-up even of order approximately τ compared to basic Kaczmarz algorithm on matrices with well-conditioned blocks (i.e. on matrices having $\lambda_{\max}^{\text{block}} \ll \tau$). Section 4.3 provides choices for the sampling that lead to a small stochastic conditioning parameter $\lambda_{\max}^{\text{block}}$.

4.2. Randomized block Kaczmarz algorithm with adaptive stepsize. Since the previous algorithm involves a stepsize depending on $\lambda_{\max}^{\text{block}}$, which may be difficult to compute in large-scale settings (i.e. when the random variable J is complicated and the number of rows m is large), in this section we design a randomized block Kaczmarz algorithm with an adaptive stepsize, which does not require the computation of $\lambda_{\max}^{\text{block}}$. More precisely, we consider a variant of RBK (Algorithm 4.1) with variable weights and an adaptive stepsize approximating online $\lambda_{\max}^{\text{block}}$. For simplicity of the notation let us define $\bar{\omega}_i^k = \frac{\omega_i^k}{\|a_i\|^2}$. Then, we consider the iteration of RBK (Algorithm 4.1) with an adaptive extrapolated stepsize of the form:

$$(4.6) \quad 0 < \alpha_k < 2L_k, \quad \text{where} \quad L_k = \begin{cases} \frac{\sum_{i \in J_k} \bar{\omega}_i^k (a_i^T x^k - b_i)^2}{\|\sum_{i \in J_k} \bar{\omega}_i^k (a_i^T x^k - b_i) a_i\|^2} & \text{if } a_i^T x^k - b_i \neq 0 \ \forall i \in J_k \\ \frac{1}{\lambda_{\max}(A_{J_k}^T \text{diag}(\bar{\omega}_i^k, i \in J_k) A_{J_k})} & \text{otherwise.} \end{cases}$$

Note that we do not need to compute L_k for the second case when implementing the algorithm. Recall that we consider weights satisfying $0 < \omega_{\min} \leq \omega_i^k \leq \omega_{\max} < 1$ for all k, i , and summing to 1. Hence, from Jensen's inequality we always have $L_k \geq 1$ and consequently $2L_k \geq 2$, i.e. we use extrapolation. Further, in our convergence analysis we take a stepsize of the form $\alpha_k = (2 - \delta)L_k$ for some $\delta \in (0, 1]$. Moreover, we denote $x_k^* = \Pi_{\mathcal{X}}(x^k)$, that is the projection of x^k onto the solution set of the linear system. It has been observed in practice that block Kaczmarz iteration with this adaptive choice for the stepsize has better performances than the same algorithm but with stepsize $\alpha_k \in (0, 2)$, see e.g. [1, 2, 3, 14, 21]. However, from our knowledge, there is no theory explaining *why* and *when* this adaptive method works. The next theorem proves the convergence rate of the adaptive algorithm depending explicitly on the geometric properties of the system matrix A and its submatrices A_J and answers to the question related to the theoretical understanding of observed practical efficiency of extrapolated block Kaczmarz methods.

THEOREM 4.2. *Let $\{x^k\}_{k \geq 0}$ be generated by RBK (Algorithm 4.1) with the adaptive stepsize $\alpha_k = (2 - \delta)L_k$ for some $\delta \in (0, 1]$ and the weights satisfying $0 < \omega_{\min} \leq \omega_i^k \leq \omega_{\max} < 1$ for all k, i . Then, we have the following linear convergence in expectation:*

$$(4.7) \quad \mathbf{E} [\|x^k - x_k^*\|^2] \leq \left(1 - \frac{\delta(2 - \delta)\omega_{\min}\lambda_{\min}^{\text{nz}}(W)}{\omega_{\max}\lambda_{\max}^{\text{block}}}\right)^k \|x^0 - x_0^*\|^2.$$

Proof. Using that $\langle x - x^*, (a_i^T x - b_i) a_i \rangle = (a_i^T x - b_i)^2$ in the update of RBK, we get:

$$\begin{aligned}
\|x^{k+1} - x_{k+1}^*\|^2 &= \left\| x^k - \alpha_k \left(\sum_{i \in J_k} \omega_k^i \frac{a_i^T x^k - b_i}{\|a_i\|^2} a_i \right) - x_{k+1}^* \right\|^2 \\
&= \|x^k - x_{k+1}^*\|^2 - 2\alpha_k \left(\sum_{i \in J_k} \omega_k^i \frac{(a_i^T x^k - b_i)^2}{\|a_i\|^2} \right) + \alpha_k^2 \left\| \sum_{i \in J_k} \omega_k^i \frac{(a_i^T x^k - b_i)^2}{\|a_i\|^2} a_i \right\|^2 \\
&= \|x^k - x^{k+1,*}\|^2 - 2(2-\delta) \frac{\left(\sum_{i \in J_k} \bar{\omega}_i^k (a_i^T x^k - b_i)^2 \right)^2}{\left\| \sum_{i \in J_k} \bar{\omega}_i^k (a_i^T x^k - b_i) a_i \right\|^2} + (2-\delta)^2 \frac{\left(\sum_{i \in J_k} \bar{\omega}_i^k (a_i^T x^k - b_i)^2 \right)^2}{\left\| \sum_{i \in J_k} \bar{\omega}_i^k (a_i^T x^k - b_i) a_i \right\|^2} \\
&= \|x^k - x^{k+1,*}\|^2 - \delta(2-\delta) \frac{\sum_{i \in J_k} \bar{\omega}_i^k (a_i^T x^k - b_i)^2}{\left\| \sum_{i \in J_k} \bar{\omega}_i^k (a_i^T x^k - b_i) a_i \right\|^2} \sum_{i \in J_k} \bar{\omega}_i^k (a_i^T x^k - b_i)^2 \\
&= \|x^k - x^{k+1,*}\|^2 - \delta(2-\delta) L_k \sum_{i \in J_k} \bar{\omega}_i^k (a_i^T x^k - b_i)^2.
\end{aligned}$$

Note that we get the same recurrence also for the trivial choice $L_k = 1/\lambda_{\max}(A_{J_k}^T \text{diag}(\bar{\omega}_i^k, i \in J_k) A_{J_k})$. Now let us bound L_k . For the nontrivial case, using that $\lambda_{\max}(MN) = \lambda_{\max}(NM)$ for any two matrices M and N of appropriate dimensions, we have:

$$\begin{aligned}
L_k &= \frac{(A_{J_k} x^k - b_{J_k})^T \text{diag}(\bar{\omega}_i^k, i \in J_k) (A_{J_k} x^k - b_{J_k})}{\|A_{J_k}^T \text{diag}(\bar{\omega}_i^k, i \in J_k) (A_{J_k} x^k - b_{J_k})\|^2} \\
&\geq \frac{1}{\lambda_{\max}(\text{diag}(\sqrt{\bar{\omega}_i^k}, i \in J_k) A_{J_k} A_{J_k}^T \text{diag}(\sqrt{\bar{\omega}_i^k}, i \in J_k))} \\
&= \frac{1}{\lambda_{\max}(A_{J_k}^T \text{diag}(\bar{\omega}_i^k, i \in J_k) A_{J_k})}.
\end{aligned}$$

This inequality holds trivially for the second choice (case) of L_k . Therefore, we can further bound L_k for all the cases as follows:

$$\begin{aligned}
L_k &\geq \frac{1}{\lambda_{\max}(A_{J_k}^T \text{diag}(\bar{\omega}_i^k, i \in J_k) A_{J_k})} \\
&\geq \frac{1}{(\max_{i \in J_k} \omega_i^k) \lambda_{\max}(A_{J_k}^T \text{diag}(1/\|a_i\|^2, i \in J_k) A_{J_k})} \\
&\geq \frac{1}{\omega_{\max} \max_{J \sim \mathbf{P}} \lambda_{\max}(A_J^T \text{diag}(1/\|a_i\|^2, i \in J) A_J)} \\
(4.8) \quad &= \frac{1}{\omega_{\max} \lambda_{\max}^{\text{block}}}.
\end{aligned}$$

Using this bound in the recurrence above we get:

$$\begin{aligned}
\|x^{k+1} - x_{k+1}^*\|^2 &\leq \|x^k - x_k^*\|^2 - \delta(2-\delta) \frac{1}{\omega_{\max} \lambda_{\max}^{\text{block}}} \sum_{i \in J_k} \omega_i^k \frac{(a_i^T x^k - b_i)^2}{\|a_i\|^2} \\
&\leq \|x^k - x_k^*\|^2 - \delta(2-\delta) \frac{\omega_{\min}}{\omega_{\max} \lambda_{\max}^{\text{block}}} \sum_{i \in J_k} \frac{(a_i^T x^k - b_i)^2}{\|a_i\|^2} \\
&= \|x^k - x_k^*\|^2 - \delta(2-\delta) \frac{\omega_{\min}}{\omega_{\max} \lambda_{\max}^{\text{block}}} (x^k - x_k^*)^T \sum_{i \in J_k} \frac{a_i a_i^T}{\|a_i\|^2} (x^k - x_k^*).
\end{aligned}$$

Taking now the conditional expectation and using again (3.6), we get:

$$\begin{aligned} \mathbf{E}_J [\|x^{k+1} - x_{k+1}^*\|^2 | \mathcal{F}_k] &\leq \|x^k - x_k^*\|^2 - \delta(2-\delta) \frac{\omega_{\min}}{\omega_{\max} \lambda_{\max}^{\text{block}}} (x^k - x_k^*)^T A^T \text{diag} \left(\frac{p_i}{\|a_i\|^2}, i \in [m] \right) A (x^k - x_k^*) \\ &= \|x^k - x_k^*\|^2 - \delta(2-\delta) \frac{\omega_{\min}}{\omega_{\max} \lambda_{\max}^{\text{block}}} (x^k - x_k^*)^T W (x^k - x_k^*) \end{aligned}$$

It is also known from the Courant-Fischer theorem that for any matrix A we have $\|Ax\|^2 \geq \lambda_{\min}^{\text{nz}}(AA^T)\|x\|^2$ for all $x \in \text{range}(A^T)$. Moreover, we have that $x - \Pi_{\mathcal{X}}(x) \in \text{range}(A^T)$ for any x . In conclusion, using that $W = A^T D A$, with the diagonal matrix $D = \text{diag} \left(\frac{p_i}{\|a_i\|^2}, i \in [m] \right)$ invertible, we get that:

$$\begin{aligned} (x^k - x_k^*)^T W (x^k - x_k^*) &= \|D^{1/2} A (x^k - x_k^*)\|^2 \geq \lambda_{\min}^{\text{nz}}(A^T D A) \|x^k - x_k^*\|^2 \\ &= \lambda_{\min}^{\text{nz}}(W) \|x^k - x_k^*\|^2. \end{aligned}$$

Using this inequality in the recurrence above and taking expectation over the entire history we get:

$$\mathbf{E} [\|x^{k+1} - x_{k+1}^*\|^2] \leq \left(1 - \delta(2-\delta) \frac{\omega_{\min}}{\omega_{\max} \lambda_{\max}^{\text{block}}} \lambda_{\min}^{\text{nz}}(W) \right) \mathbf{E} [\|x^k - x_k^*\|^2],$$

hence proving the statement of the theorem. \square

There is a tight connection between the constant stepsize (4.2) and the adaptive stepsize (4.6). Indeed, for simplicity let us consider uniform weights $\omega_i^k = 1/\tau$ and normalized matrices ($\|a_i\| = 1$ for all i, k). Then, from (4.8) we obtain:

$$L_k = \tau \frac{\|A_{J_k} x^k - b_{J_k}\|^2}{\|A_{J_k}^T (A_{J_k} x^k - b_{J_k})\|^2} \geq \tau \frac{1}{\lambda_{\max}(A_{J_k}^T A_{J_k})} \geq \tau \frac{1}{\lambda_{\max}^{\text{block}}}.$$

Hence, L_k represents an online approximation of $\tau/\lambda_{\max}^{\text{block}}$ and therefore:

$$(4.9) \quad \alpha_k = 2L_k \geq \alpha = 2 \frac{\omega_{\min}}{\omega_{\max}^2 \lambda_{\max}^{\text{block}}} = 2\tau \frac{1}{\lambda_{\max}^{\text{block}}}.$$

In conclusion, the adaptive stepsize (4.6) can be viewed as a practical online approximation of the constant extrapolated stepsize (4.2). Finally, let us simplify the convergence rate (4.7) for the uniform and partition sampling examples of Section 3.4 having all the blocks sampling the same size $|J| = \tau$. In this case we have $p_i = \frac{\tau}{m}$. Let us also consider the particular choices $\delta = 1$, weights $\omega_i = 1/\tau$, and normalized matrices A . Then, our convergence rate (4.7) becomes:

$$(4.10) \quad \mathbf{E} [\|x^k - x_k^*\|^2] \leq \left(1 - \frac{\tau}{\lambda_{\max}^{\text{block}}} \frac{\lambda_{\min}^{\text{nz}}(A^T A)}{m} \right)^k \|x^0 - x_0^*\|^2.$$

We observe that this convergence rate coincides with (4.5). However, the adaptive block Kaczmarz scheme has more chances to accelerate, since from (4.9) the variable stepsize is, in general, larger than the constant stepsize counterpart.

4.3. When block Kaczmarz works?. Comparing the convergence rates of RBK algorithm with the constant stepsize (4.2) and with the adaptive stepsize (4.6) given in (4.5) and (4.10), respectively, with the convergence rate of the basic Kaczmarz method given in (3.4), we obtain an improvement $\frac{\tau}{\lambda_{\max}^{\text{block}}} > 1$ for the block variants. Recall that the stochastic conditioning parameter $\lambda_{\max}^{\text{block}}$ is defined as:

$$\lambda_{\max}^{\text{block}} = \max_{J \sim \mathbf{P}} \lambda_{\max} \left(A_J^T \text{diag} \left(\frac{1}{\|a_i\|^2}, i \in J \right) A_J \right).$$

Therefore, we can get a speed-up even of order approximately τ for well conditioned matrices, i.e. for matrices having $\lambda_{\max}^{\text{block}} \ll \tau$. This shows that the probability \mathbf{P} plays a key role in defining the importance sampling procedure and consequently in the convergence behavior of RBK. Fortunately, the operator theory literature provides detailed information about the existence of such good probabilities defining the importance sampling. This is usually referred in the literature as *good paving* [16]. This section summarizes the main results from the literature on row paving and provides a technique for constructing a good paving. The idea is to find a random partition of the rows of the matrix A such that each subset has approximately equal size. Results on existence of good row pavings were derived e.g. in [26]:

LEMMA 4.3. *Let A be a normalized matrix with m rows and $\theta \in (0, 1)$. Then, there is a randomized partition $\{J_1, \dots, J_\ell\}$ of the rows indices with $\ell \geq \mathcal{O}(\|A\|^2 \log(1+m)/\theta^2)$ such that $\lambda_{\max}^{\text{block}} \leq 1 + \theta$.*

Although this is only an existential result, the literature describes several efficient algorithms for constructing good row pavings. For example, assume that κ is a permutation of the set $[m] = \{1, 2, \dots, m\}$, chosen uniformly at random. For each $i = 1 : \ell$, define the subsets:

$$J_i = \left\{ \kappa(l) : l = \lfloor (i-1) \frac{m}{\ell} \rfloor + 1, \dots, \lfloor \frac{m}{\ell} \rfloor \right\}.$$

It is clear that $\{J_1, \dots, J_\ell\}$ is a random partition of $[m]$ into ℓ blocks of approximately equal size. For every normalized matrix, such a random partition leads to a row paving whose $\lambda_{\max}^{\text{block}}$ is relatively small.

LEMMA 4.4. *Let A be a normalized matrix with m rows. Consider a randomized partition $\{J_1, \dots, J_\ell\}$ of the rows indices with $\ell \geq \|A\|^2$ subsets. Then, $\{J_1, \dots, J_\ell\}$ is a row paving with the upper bound $\lambda_{\max}^{\text{block}} \leq 6 \log(1+m)$ with probability at least $1 - m^{-1}$.*

A proof of this type of result appears in [25], see also [16]. By merging our theorems on the convergence of RBK algorithm with the previous result on the good paving, we obtain:

THEOREM 4.5. *Let A be a normalized matrix and $\{J_1, \dots, J_\ell\}$ be a random partition of the rows of A , as given by Lemma 4.4, such that $\tau = m/\ell$ is a positive integer. Under the assumptions of Theorems 4.1 and 4.2, the randomized block Kaczmarz method, Algorithm 4.1, with weights $\omega_i^k = 1/\tau = \ell/m$ for all i, k , and constant stepsize (4.2) or adaptive stepsize (4.6) with $\delta = 1$, admits the convergence estimate:*

$$(4.11) \quad \mathbf{E} [\|x^k - x_k^*\|^2] \leq \left(1 - \frac{\lambda_{\min}^{\text{nz}}(A^T A)}{6 \log(1+m) \|A\|^2} \right)^k \|x^0 - x_0^*\|^2.$$

In conclusion, our new convergence analysis shows when a block variant of Kaczmarz algorithm really works, i.e. we can choose a subset of rows $\tau > 1$ at each step, when $\lambda_{\max}^{\text{block}} \ll \tau$. Hence, a distributed implementation of the RBK algorithm is most effective when the probability distribution \mathbf{P} yields a partition of the rows into well-conditioned blocks. Otherwise, we can just apply the basic Kaczmarz algorithm with $\tau = 1$. Moreover, our analysis shows that the *optimal batchsize* is of order $\tau \sim m/\|A\|^2$. Assuming, for simplicity, that $\tau = m/\ell$ is a positive integer, from Lemma 4.4

$$\lambda_{\max}^{\text{block}} \leq 6 \log(1+m) \ll \tau = \frac{m}{\ell} \simeq \frac{m}{\|A\|^2}$$

holds with high probability, provided that the matrix A satisfies the following inequality

$$(4.12) \quad \|A\|^2 \ll \frac{m}{6 \log(1+m)}.$$

Recall that, for a normalized matrix A with m rows, the squared spectral norm $\|A\|^2$ attains its maximal value m when $\text{rank}(A) = 1$, i.e. its rows are identical. Therefore, the inequality (4.12) stipulates that the rows of A must exhibit a large amount of diversity in order for RBK algorithm with extrapolated stepsizes (4.2) or (4.6) to perform better than the basic Kaczmarz scheme. Note that convergence rates similar to (4.11) has been derived in [16] for the block projection Kaczmarz algorithm (1.3) with the particular stepsize $\alpha_k = 1$. However, RBK requires the computation of τ scalar products in \mathbb{R}^n at each iteration, so that its computational cost per iteration is $\mathcal{O}(\tau n)$, and thus cheaper than the one corresponding to block projection Kaczmarz (1.3) that requires solving a least-squares problem at each iteration in about $\mathcal{O}(\tau^2 n)$.

5. Randomized block Kaczmarz algorithm with Chebychev-based stepsize. Finally, we show that we can also choose extrapolated stepsizes in RBK (Algorithm 4.1) based on the roots of Chebyshev polynomials. For simplicity, we consider either the uniform or partition sampling of Section 3.4 having $|J| = \tau$. We also assume normalized matrices A and constant weights $\omega_k^i = 1/\tau$ for all k, i . Under these settings, for RBK algorithm with Chebyshev-based stepsize we derive linear or sublinear convergence estimates depending whether $\lambda_{\min}(AA^T) > 0$ or $\lambda_{\min}(AA^T) = 0$, respectively. Below we investigate these two cases.

5.1. Case 1: $\lambda_{\min}(AA^T) > 0$. We get the following linear convergence for this variant of RBK:

THEOREM 5.1. *Assume normalized matrix A such that $\lambda_{\min}(AA^T) > 0$. Let $\{x^k\}_{k \geq 0}$ be generated by RBK (Algorithm 4.1) with the uniform or partition sampling and the weights $\omega_k^i = 1/\tau$ for all k, i . Further, for a fixed number of iterations k the stepsizes $\{\alpha_j\}_{j=0}^{k-1}$ are depending on the roots of the Chebyshev polynomial of degree k (see Appendix) as follows:*

$$\alpha_j = \frac{2m}{(\lambda_{\max}(AA^T) + \lambda_{\min}(AA^T)) + (\lambda_{\max}(AA^T) - \lambda_{\min}(AA^T)) \cos\left(\frac{2\kappa(j)+1}{2k}\pi\right)},$$

where κ is a permutation of $[0:k-1]$. Then, we have the following linear convergence for expected iterates:

$$(5.1) \quad \|\mathbf{E}[x^k - x_k^*]\|^2 \leq \frac{4\lambda_{\max}(AA^T)}{\lambda_{\min}(AA^T)} \left(1 - \sqrt{\frac{\lambda_{\min}(AA^T)}{\lambda_{\max}(AA^T)}}\right)^{2k} \|x^0 - x_0^*\|^2.$$

Proof. For the iteration of RBK (Algorithm 4.1) we have for any solution $x^* \in \mathcal{X}$:

$$\begin{aligned} x^{k+1} - x^* &= x^k - x^* - \alpha_k \left(\sum_{i \in J_k} \omega_k^i \frac{a_i^T x^k - b_i}{\|a_i\|^2} a_i \right) \stackrel{\omega_k^i = 1/\tau, \|a_i\|=1}{=} x^k - x^* - \frac{\alpha_k}{\tau} \left(\sum_{i \in J_k} (a_i^T x^k - b_i) a_i \right) \\ &= x^k - x^* - \frac{\alpha_k}{\tau} \left(\sum_{i \in J_k} a_i a_i^T (x^k - x^*) \right) = \left(I_n - \frac{\alpha_k}{\tau} \left(\sum_{i \in J_k} a_i a_i^T \right) \right) (x^k - x^*). \end{aligned}$$

Taking conditional expectation and using (3.6) with $p_i = \tau/m$ for uniform or partition sampling, we get:

$$(5.2) \quad \begin{aligned} \mathbf{E}_J[x^{k+1} - x^* | \mathcal{F}_k] &= \left(I_n - \frac{\alpha_k}{\tau} \left(\sum_{i \in [m]} p_i a_i a_i^T \right) \right) (x^k - x^*) \\ &\stackrel{p_i = \tau/m}{=} \left(I_n - \frac{\alpha_k}{m} A^T A \right) (x^k - x^*). \end{aligned}$$

Multiplying from the left this recurrence with A we get:

$$\mathbf{E}_J[Ax^{k+1} - Ax^* | \mathcal{F}_k] = \left(A - \frac{\alpha_k}{m} AA^T A \right) (x^k - x^*) = \left(I_m - \frac{\alpha_k}{m} AA^T \right) (Ax^k - Ax^*)$$

or equivalently, using that $Ax^* = b$ and taking expectations over the entire history, we obtain:

$$\mathbf{E}[Ax^{k+1} - b] = \left(I_m - \frac{\alpha_k}{m} AA^T \right) \mathbf{E}[Ax^k - b].$$

Iterating this recurrence and defining the matrix $G = \frac{1}{m} AA^T \in \mathbb{R}^{m \times m}$ we obtain:

$$\mathbf{E}[Ax^k - b] = \prod_{j=0}^{k-1} \left(I_m - \alpha_j \frac{1}{m} AA^T \right) (Ax^0 - b) = \prod_{j=0}^{k-1} (I_m - \alpha_j G) (Ax^0 - b).$$

If we define the polynomial in the matrix G as $P_k(G) = \prod_{j=0}^{k-1} (I_m - \alpha_j G)$, then we can bound the norm of the expected residual by:

$$\|\mathbf{E}[Ax^k - b]\| = \|P_k(G)(Ax^0 - b)\| \leq \|P_k(G)\| \cdot \|Ax^0 - b\|.$$

Recall that we consider consistent linear system with $\lambda_{\min}(AA^T) > 0$. Then, from standard reasoning the spectrum of $G = \frac{1}{m}AA^T$ satisfies $\Lambda(G) \subset \mathbb{R}_{++}$. More precisely:

$$0 < \underbrace{\frac{1}{m}\lambda_{\min}(AA^T)}_{=\ell} \leq \lambda_i(G) \leq \underbrace{\frac{1}{m}\lambda_{\max}(AA^T)}_{=u} < \infty \quad \forall i = 1 : m.$$

Therefore, if we denote by λ_i the i th eigenvalue of G , we have the following bound:

$$\|\mathbf{E}[Ax^k - b]\| \leq \|P_k(G)\| \cdot \|Ax^0 - b\| \leq \max_{i=1:m} |P_k(\lambda_i)| \cdot \|Ax^0 - b\| \leq \max_{\lambda \in [\ell, u]} |P_k(\lambda)| \cdot \|Ax^0 - b\|.$$

In conclusion, we can choose the stepsizes α_j for $j = 0 : k-1$ such that $P_k(\lambda) = \prod_{j=0}^{k-1} (1 - \alpha_j \lambda)$ is the polynomial least deviating from zero on the interval $[\ell, u]$ and satisfying $P_k(0) = 1$. It is well known that this is the polynomial given in terms of a Chebyshev polynomial (see Appendix for a brief review of the main properties of Chebyshev polynomials):

$$P_k(\lambda) = T_k \left(\frac{2\lambda}{u-\ell} - \frac{u+\ell}{u-\ell} \right) / T_k \left(-\frac{u+\ell}{u-\ell} \right).$$

Then, we can guarantee the following linear convergence in expectation (see Lemma 5.3 in Appendix):

$$(5.3) \quad \|\mathbf{E}[Ax^k - b]\| \leq 2 \left(\frac{\sqrt{u} - \sqrt{\ell}}{\sqrt{u} + \sqrt{\ell}} \right)^k \|Ax^0 - b\| \leq 2 \left(1 - \sqrt{\frac{\lambda_{\min}(AA^T)}{\lambda_{\max}(AA^T)}} \right)^k \|Ax^0 - b\|.$$

The stepsizes α_j , for $j = 0 : k-1$, are chosen as the inverse roots of polynomial $P_k(\lambda)$ (see Appendix):

$$\begin{aligned} \alpha_j &= 2 / \left((u + \ell) + (u - \ell) \cos \left(\frac{2\kappa(j) + 1}{2k} \pi \right) \right) \\ &= 2m / \left((\lambda_{\max}(AA^T) + \lambda_{\min}(AA^T)) + (\lambda_{\max}(AA^T) - \lambda_{\min}(AA^T)) \cos \left(\frac{2\kappa(j) + 1}{2k} \pi \right) \right), \end{aligned}$$

where κ is some fixed permutation of $[0 : k-1]$. We can also derive convergence rates in $\mathbf{E}[x^k - x_k^*]$ using that $\mathbf{E}[x^k - x_k^*] \in \text{range}(A^T)$, and consequently from Courant-Fischer lemma and (5.3) we have:

$$\begin{aligned} \lambda_{\min}(AA^T) \|\mathbf{E}[x^k - x_k^*]\|^2 &\leq \|A\mathbf{E}[x^k - x_k^*]\|^2 = \|\mathbf{E}[Ax^k - b]\|^2 \\ &\leq 4 \left(1 - \sqrt{\frac{\lambda_{\min}(AA^T)}{\lambda_{\max}(AA^T)}} \right)^{2k} \|Ax^0 - b\|^2 \leq 4\lambda_{\max}(AA^T) \left(1 - \sqrt{\frac{\lambda_{\min}(AA^T)}{\lambda_{\max}(AA^T)}} \right)^{2k} \|x^0 - x_0^*\|^2. \end{aligned}$$

proving thus the linear convergence estimate of the theorem. \square

From Jensen's inequality we have $\|\mathbf{E}[\cdot]\| \leq \mathbf{E}[\|\cdot\|]$. In conclusion, $\|\mathbf{E}[\cdot]\|$ is a weaker criterion than $\mathbf{E}[\|\cdot\|]$. Note that convergence rates in the weaker criterion $\|\mathbf{E}[x^k - x_k^*]\|$ have been also given for another variant of Kaczmarz algorithm in [22] or for the random coordinate descent method in [24]. Moreover, the convergence rate from Theorem 5.1 is the same as for the conjugate gradient method and it is optimal for this class of iterative schemes. However, since this rate does not depend on the size of the blocks $|J|$, then we usually implement this accelerated variant of Kaczmarz by sampling single rows, that is, $|J| = 1$.

5.2. Case 2: $\lambda_{\min}(AA^T) = 0$. In this case we get sublinear convergence for this variant of RBK:

THEOREM 5.2. *Assume normalized matrix A such that $\lambda_{\min}(AA^T) = 0$. Let $\{x^k\}_{k \geq 0}$ be generated by RBK (Algorithm 4.1) with the uniform or partition sampling and the weights $\omega_k^i = 1/\tau$ for all k, i .*

Further, for a fixed number of iterations k the stepsizes $\{\alpha_j\}_{j=0}^{k-1}$ are depending on the roots of the Chebyshev polynomial of degree k as follows:

$$\alpha_j = \frac{m \left(1 - \cos \left(\frac{2k+1}{2(k+1)} \pi \right) \right)}{\lambda_{\max}(AA^T) \left(\cos \left(\frac{2\kappa(j)+1}{2(k+1)} \pi \right) - \cos \left(\frac{2k+1}{2(k+1)} \pi \right) \right)},$$

where κ is some permutation of $[0 : k - 1]$. Then, we have the following sublinear convergence for the residual of the normal system in expectation:

$$(5.4) \quad \|\mathbf{E} [A^T Ax^k - A^T b]\| = \|\mathbf{E} [Ax^k - b]\|_{(AA^T)} \leq \frac{\pi \lambda_{\max}(AA^T)}{2(k+1)^2} \|x^0 - x^*\|.$$

Proof. From (5.2) we also get the relation:

$$\mathbf{E} [x^k - x^*] = \prod_{j=0}^{k-1} \left(I_n - \alpha_j \frac{1}{m} A^T A \right) (x^0 - x^*).$$

Now, if we consider the normal system $A^T Ax = A^T b$, which coincides with $\nabla f(x) = 0$, we have:

$$\|\mathbf{E} [A^T Ax^k - A^T b]\| = \|\mathbf{E} [A^T A(x^k - x^*)]\| = \|A^T A \prod_{j=0}^{k-1} \left(I_n - \alpha_j \frac{1}{m} A^T A \right) (x^0 - x^*)\|,$$

where x^* denotes any solution of $Ax = b$ (recall that we consider consistent linear systems). If we define the matrix $G = \frac{1}{m} A^T A$ and the polynomial $Q_k(G) = G \prod_{j=0}^{k-1} (I_n - \alpha_j G)$, then we obtain the following bound for the residual of the normal system in expectation:

$$\|\mathbf{E} [A^T Ax^k - A^T b]\| = m \|Q_k(G)(x^0 - x^*)\| \leq m \|Q_k(G)\| \|x^0 - x^*\|.$$

Since we assume $\lambda_{\min}(AA^T) = \lambda_{\min}(A^T A) = 0$, then the spectrum of $G = \frac{1}{m} A^T A$ satisfies:

$$0 \leq \lambda_i(G) \leq \underbrace{\frac{1}{m} \lambda_{\max}(A^T A)}_{=u} < \infty \quad \forall i = 1 : m.$$

Therefore, if we denote by λ_i the i th eigenvalue of G , we have the following bound:

$$\begin{aligned} \|\mathbf{E} [A^T Ax^k - A^T b]\| &\leq m \|Q_k(G)\| \cdot \|x^0 - x^*\| \leq m \max_{i=1:m} |Q_k(\lambda_i)| \cdot \|x^0 - x^*\| \\ &\leq m \max_{\lambda \in [0, u]} |Q_k(\lambda)| \cdot \|x^0 - x^*\|. \end{aligned}$$

In conclusion, we can choose the stepsizes α_j for $j = 0 : k - 1$ such that $Q_k(\lambda) = \lambda \prod_{j=0}^{k-1} (1 - \alpha_j \lambda)$ of degree $k + 1$ is the polynomial least deviating from zero on the interval $[0, u]$ and satisfying $Q_k(0) = 0$ and $Q'_k(0) = 1$. We show below that this polynomial is also given in terms of a Chebyshev polynomial. Indeed, let us consider the closest root to -1 of the Chebyshev polynomial of degree $k + 1$ (i.e. T_{k+1}):

$$r_{k+1} = \cos \left(\frac{2k+1}{2(k+1)} \pi \right) = \cos \left(\pi - \frac{1}{2(k+1)} \pi \right).$$

Then, we define the polynomial:

$$Q_k(\lambda) = \frac{u}{1 - r_{k+1}} \frac{T_{k+1} \left(r_{k+1} + \frac{1-r_{k+1}}{u} \lambda \right)}{T'_{k+1}(r_{k+1})}.$$

Note that this polynomial satisfies the required properties: $\deg(Q_k) = k+1$, $Q_k(0) = \frac{u T_{k+1}(r_{k+1})}{(1-r_{k+1}) T'_{k+1}(r_{k+1})} = 0$ (recall that r_{k+1} is the $k+1$ root of T_{k+1}) and $Q'_k(0) = \frac{T'_{k+1}(r_{k+1})}{T'_{k+1}(r_{k+1})} = 1$. In conclusion, we get the following bound for this choice of $Q_k(\lambda)$:

$$m \max_{\lambda \in [0, u]} |Q_k(\lambda)| = m \max_{\lambda \in [0, u]} \left| \frac{u}{1-r_{k+1}} \frac{T_{k+1}\left(r_{k+1} + \frac{1-r_{k+1}}{u}\lambda\right)}{T'_{k+1}(r_{k+1})} \right| \leq m \frac{u}{|T'_{k+1}(r_{k+1})|} = \frac{\lambda_{\max}(A^T A)}{|T'_{k+1}(r_{k+1})|},$$

where in the inequality we used that $|T_{k+1}(x)| \leq 1$ for any $x \in [-1, 1]$ and that the root $r_{k+1} \leq 0$ (see Appendix). Further, since $T_{k+1}(\cos(\theta)) = \cos((k+1)\theta)$, if we differentiate we get $\sin(\theta)T'_{k+1}(\cos(\theta)) = (k+1)\sin((k+1)\theta)$. Now, for $r_{k+1} = \cos(\pi - \pi/(2k+2))$ we obtain:

$$|T'_{k+1}(r_{k+1})| = \frac{(k+1)|\sin((k+1)\pi - \pi/2)|}{|\sin(\pi - \pi/(2k+2))|} = \frac{k+1}{|\sin(\pi - \pi/(2k+2))|} = \frac{2(k+1)^2}{\pi},$$

for k sufficiently large (we used that $\sin(\pi - \theta) \sim \theta$ for θ small). In conclusion, we get the following sublinear convergence (using the notation $\|u\|_{(AA^T)} = \|A^T u\|$):

$$\|\mathbf{E}[A^T Ax^k - A^T b]\| = \|\mathbf{E}[Ax^k - b]\|_{(AA^T)} \leq \frac{\pi \lambda_{\max}(A^T A)}{2(k+1)^2} \|x^0 - x^*\|,$$

for k sufficiently large (i.e. for k such that $\sin(\pi - \pi/(2k+2)) \sim \pi/(2k+2)$). Finally, using that $\lambda_{\max}(A^T A) = \lambda_{\max}(AA^T)$ we get (5.4). The stepsizes α_j , for $j = 0 : k-1$, are chosen as the inverse roots of polynomial $Q_k(\lambda)$ (see Appendix):

$$\alpha_j = (1-r_{k+1})u^{-1}/\left(\cos\left(\frac{2\kappa(j)+1}{2(k+1)}\pi\right) - r_{k+1}\right) = \frac{m\left(1 - \cos\left(\frac{2k+1}{2(k+1)}\pi\right)\right)}{\lambda_{\max}(AA^T)\left(\cos\left(\frac{2\kappa(j)+1}{2(k+1)}\pi\right) - \cos\left(\frac{2k+1}{2(k+1)}\pi\right)\right)},$$

where κ is some fixed permutation of $[0 : k-1]$. □

Note that the RBK algorithm with Chebyshev-based stepsize belongs to the class of Chebyshev semi-iterative methods [6]. However, from our knowledge, this work is the first one that uses the properties of the Chebyshev polynomials in order to accelerated the convergence rate of randomized block Kaczmarz (RBK) algorithm. Other types of acceleration of Kaczmarz algorithm have been proposed e.g. in [7, 12, 22]. For example, in [22] two dependent steps of basic randomized Kaczmarz algorithm are taken, one from x^k and one from x^{k-1} , and then an affine combination of the results produces the next iterate x^{k+1} . For this scheme, [22] derives a similar convergence rate as in Theorem 5.1. In [12] Nesterov's accelerated random coordinate descent method from [18] is applied to the dual problem (3.3), leading in the primal space to an accelerated randomized Kaczmarz scheme with momentum. For this accelerated Kaczmarz scheme [12] derives the convergence rate $\mathbf{E}[\|x^k - x^*\|^2] \leq (1 - \sqrt{\lambda_{\min}(AA^T)/m})^k \|x^0 - x_0^*\|^2$. Although this rate is worse than (5.1) in terms of constants, it is given in the stronger criterion $\mathbf{E}[\|x^k - x^*\|^2]$. Remains an open problem whether Theorem 5.1 can be also given in the stronger criterion $\mathbf{E}[\|x^k - x^*\|^2]$.

REFERENCES

- [1] H. Bauschke, P. Combettes and S. Kruk, *Extrapolation algorithm for affine-convex feasibility problems*, Numerical Algorithms, 41(3):239–274, 2006.
- [2] Y. Censor, *Row action methods for huge sparse systems and their applications*, SIAM Review, 23: 444–466, 1981.
- [3] Y. Censor, W. Chen, P. Combettes, R. Davidi and G. Herman, *On the Effectiveness of Projection Methods for Convex Feasibility Problems with Linear Inequality Constraints*, Computational Optimization and Applications, 51(3): 1065–1088, 2012.
- [4] F. Deutsch and H. Hundal, *The rate of convergence for the method of alternating projections*, Journal of Mathematical Analysis and Applications, 205(2): 381–405, 1997.
- [5] T. Elfving, *Block-iterative methods for consistent and inconsistent linear equations*, Numer. Math., 35(1): 1–12, 1980.

- [6] G. Golub and R. Varga, *Chebyshev semi-iterative methods, successive over-relaxation methods and second-order Richardson iterative methods I, II*, Numer. Math., 3:147–168, 1961.
- [7] M. Hanke and W. Niethammer, *On the acceleration of Kaczmarz’s method for inconsistent linear systems*, Linear Algebra Applications, 130: 83–98, 1990.
- [8] R. Gower and P. Richtarik, *Randomized iterative methods for linear systems*, SIAM Journal on Matrix Analysis and Applications, 36(4): 1660–1690, 2015.
- [9] G. Hounsfield, *Computerized transverse axial scanning (tomography): Part I. Description of the system*, British Journal of Radiology, 46: 1016–1022, 1973.
- [10] S. Kaczmarz, *Angenäherte Auflösung von Systemen linearer Gleichungen*, Bull. Acad. Sci. Pologne, A35: 355–357, 1937.
- [11] U. Khan and J. Moura, *Distributed Kalman filters in sensor networks: bipartite fusion graphs*, Workshop on Statistical Signal Processing: 700–704, 2007.
- [12] J. Liu and S. Wright, *An accelerated randomized Kaczmarz algorithm*, Mathematics of Computation, 85: 153–178, 2016.
- [13] D. Leventhal and A. Lewis, *Randomized methods for linear constraints: convergence rates and conditioning*, Mathematics of Operations Research, 35(3): 641–654, 2010.
- [14] Y. Merzlyakov, *On a relaxation method of solving systems of linear inequalities*, USSR Comput. Math. Phys., 2: 504–510, 1963.
- [15] I. Necoara and D. Clipici, *Parallel random coordinate descent methods for composite minimization: convergence analysis and error bounds*, SIAM Journal on Optimization, 26(1): 197–226, 2016.
- [16] D. Needell and J. Tropp, *Paved with good intentions: Analysis of a randomized block Kaczmarz method*, Linear Algebra Applications, 441: 199–221, 2014.
- [17] A. Nemirovski, A. Juditsky, G. Lan and A. Shapiro, *Robust stochastic approximation approach to stochastic programming*, SIAM Journal on Optimization, 19(4): 1574–1609, 2009.
- [18] Y. Nesterov, *Efficiency of coordinate descent methods on huge-scale optimization problems*, SIAM Journal on Optimization, 22(2): 341–362, 2012.
- [19] M. Olshanskii and E. Tyrtyshnikov, *Iterative methods for linear systems: theory and applications*, SIAM, 2014.
- [20] A. Patrascu and I. Necoara, *Nonasymptotic convergence of stochastic proximal point algorithms for constrained convex optimization*, Journal of Machine Learning Research, 18(198): 1–42, 2018.
- [21] G. Pierra, *Decomposition through formalization in a product space*, Mathematical Programming, 28: 96–115, 1984.
- [22] P. Richtarik and M. Takac, *Stochastic reformulations of linear systems: algorithms and convergence theory*, arxiv, 2017.
- [23] T. Strohmer and R. Vershynin, *A randomized Kaczmarz algorithm with exponential convergence*, Journal of Fourier Analysis and Applications, 15(2): 262–278, 2009.
- [24] R. Sun and Y. Ye, *Worst-case complexity of cyclic coordinate descent: $O(n^2)$ gap with randomized version*, arxiv, 2017.
- [25] J. Tropp, *Improved analysis of the subsampled randomized Hadamard transform*, Advances in Adaptive Data Analysis, 3(1-2): 115–126, 2011.
- [26] J. Tropp, *Column subset selection, matrix factorization, and eigenvalue optimization*, Symposium on Discrete Algorithms, 978–986, 2009.
- [27] L. Xiao, S. Boyd and S. Lall, *A scheme for robust distributed sensor fusion based on average consensus*, International Symposium on Information Processing in Sensor Networks: 63–70, 2005.

Appendix (Chebyshev polynomials). In this section some properties of the Chebyshev polynomials are briefly reviewed. We refer to e.g. [19] for more details on Chebyshev polynomials. The Chebyshev polynomials $T_k(x)$, where $\deg(T_k) = k$ and $k \geq 0$, are defined by the recursive relation:

$$T_0(x) = 1, \quad T_1(x) = 1, \quad T_{k+1}(x) = 2xT_k(x) - T_{k-1}(x).$$

From the above recurrence we observe that the leading coefficient of $T_k(x)$ is 2^{k-1} , i.e. $T_k(x) = 2^{k-1}x^k + \text{lower powers of } x$. In particular, for $x \in [-1, 1]$, the Chebyshev polynomials can be written equivalently:

$$T_k(x) = \cos(k \arccos(x)).$$

The equivalence can be verified as follows using that $x = \cos(\theta)$:

$$\begin{aligned} T_k(x) &= 2x \cos((k-1) \arccos(x)) - \cos((k-2) \arccos(x)) = 2 \cos(\theta) \cos((k-1)\theta) - \cos((k-2)\theta) \\ &= \cos(k\theta) + \cos((k-2)\theta) - \cos((k-2)\theta) = \cos(k\theta) = \cos(k \arccos(x)). \end{aligned}$$

It follows that $T_k(1) = 1$. From this representation of $T_k(x)$ it also follows that:

$$\max_{x \in [-1, 1]} |T_k(x)| = 1.$$

Moreover, all the k roots of $T_k(x)$ are given by:

$$x_i = \cos\left(\frac{2i-1}{2k}\pi\right) \quad \text{for } i = 1 : k.$$

In conclusion, we get also the following representation for $T_k(x)$:

$$T_k(x) = 2^{k-1} \cdot \prod_{i=1}^k \left(x - \cos \left(\frac{2i-1}{2k}\pi \right) \right).$$

It is also easy to see the following interval transformation $[\ell, u] \rightarrow [-1, 1]$ through the relation:

$$-1 \leq \frac{2x}{u-\ell} - \frac{u+\ell}{u-\ell} \leq 1 \quad \text{for } \ell \leq x \leq u.$$

One important property of the Chebyshev polynomials is that $\frac{1}{2^{k-1}} T_k(x)$ has minimal deviation from 0 among all polynomials of degree k with leading coefficient 1 on $[-1, 1]$:

$$(5.5) \quad \max_{x \in [-1, 1]} \frac{1}{2^{k-1}} |T_k(x)| \leq \max_{x \in [-1, 1]} |P_k(x)| \quad \forall P_k(x) \text{ with leading coefficient 1 and } \deg(P_k) = k.$$

An immediate consequence of the above property valid for Chebyshev polynomials is the following lemma:

LEMMA 5.3. *Let $0 < \ell < u$ and $T_k^{(\ell,u)}(x) = T_k \left(\frac{2x}{u-\ell} - \frac{u+\ell}{u-\ell} \right)$. Then, the optimal value and the optimal polynomial P_k^* of the following optimization problem are:*

$$\min_{P_k(x): \deg(P_k)=k, P_k(0)=1} \max_{x \in [\ell, u]} |P_k(x)| = \frac{1}{T_k^{(\ell,u)}(0)} \leq 2 \left(\frac{\sqrt{u} - \sqrt{\ell}}{\sqrt{u} + \sqrt{\ell}} \right)^k \quad \text{and} \quad P_k^*(x) = \frac{T_k^{(\ell,u)}(x)}{T_k^{(\ell,u)}(0)}.$$