BIT Numerical Mathematics
https://doi.org/10.1007/s10543-020-00824-1

Randomized Kaczmarz with averaging

Jacob D. Moorman[1] · Thomas K. Tu[1] · Denali Molitor[1] · Deanna Needell[1]

Received: 7 February 2020 / Accepted: 22 July 2020

Abstract
The randomized Kaczmarz (RK) method is an iterative method for approximating the least-squares solution of large linear systems of equations. The standard RK method uses sequential updates, making parallel computation difficult. Here, we study a parallel version of RK where a weighted average of independent updates is used. We analyze the convergence of RK with averaging and demonstrate its performance empirically. We show that as the number of threads increases, the rate of convergence improves and the convergence horizon for inconsistent systems decreases.

Keywords Randomized Kaczmarz · Algebraic reconstruction technique · Parallel methods · Inconsistent linear systems

Mathematics Subject Classification 15A06 · 15B52 · 65F10 · 65F20 · 65Y20 · 68Q25 · 68W10 · 68W20 · 68W40

1 Introduction

In computed tomography, image processing, machine learning, and many other fields, a common problem is that of finding solutions to large linear systems of equations. Given $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$, we aim to find $x \in \mathbb{R}^n$ which solves the linear system of equations

$$\mathbf{A}x = b. \tag{1}$$

We will generally assume the system is overdetermined, with $m \gg n$. For simplicity, we assume throughout that $\mathbf{A}$ has full rank so that the solution is unique when it

Communicated by Rosemary Anne Renaut.

Molitor, Moorman, and Needell were partially supported by NSF CAREER DMS #1348721 and NSF BIGDATA #1740325. Moorman was additionally supported by NSF Grant DGE-1829071. Tu was supported by DARPA under agreement number FA8750-18-2-0066.

✉ Jacob D. Moorman
jacob@moorman.me

[1] Department of Mathematics, University of California, Los Angeles, Los Angeles, CA 90095-0001, USA

Published online: 11 August 2020

exists. However, this assumption can be relaxed by choosing the solution with least-norm when multiple solutions exist.

When a solution to Eq. (1) exists, we denote the solution by $x^\star$ and refer to the problem as *consistent*. Otherwise, the problem is *inconsistent*, and $x^\star$ instead denotes the *least-squares* solution

$$x^\star \overset{\text{def}}{=} \arg\min_{x\in\mathbb{R}^n} \frac{1}{2}\|b - \mathbf{A}x\|_2^2.$$

The least-squares solution can be equivalently written as $x^\star = \mathbf{A}^\dagger b$, where $\mathbf{A}^\dagger$ is the Moore-Penrose pseudoinverse of $\mathbf{A}$. We denote the least-squares *residual* as $r^\star \overset{\text{def}}{=} b - \mathbf{A}x^\star$, which is zero for consistent systems.

1.1 Randomized Kaczmarz

Randomized Kaczmarz (RK) is a popular iterative method for approximating the least-squares solution of large, overdetermined linear systems [16,29]. At each iteration, an equation is chosen at random from the system in Eq. (1) and the current iterate is projected onto the solution space of that equation. In a relaxed variant of RK, a step is taken in the direction of this projection with the size of the step depending on a relaxation parameter.

Let x^k be the kth iterate. We use $\mathbf{A}_i$ to denote the ith row of $\mathbf{A}$ and $\|\cdot\| \overset{\text{def}}{=} \|\cdot\|_2$. The *relaxed RK* update is given by

$$x^{k+1} = x^k - \alpha_k \frac{\mathbf{A}_{i_k}x^k - b_{i_k}}{\|\mathbf{A}_{i_k}\|^2}\mathbf{A}_{i_k}^\top, \tag{2}$$

where i_k is sampled from some fixed distribution $\mathcal{D}$ at each iteration and α_k are relaxation parameters [4]. Fixing the relaxation parameters $\alpha_k = 1$ for all iterations k leads to the standard RK method in which one projects the current iterate x^k onto the solution space of the chosen equation $\mathbf{A}_{i_k}x = b_{i_k}$ at each iteration [29]. Choosing relaxation parameters $\alpha_k \neq 1$ can be used to accelerate convergence or dampen the effect of noise in the linear system [4,13,14].

For consistent systems, RK converges exponentially in mean squared error (MSE) to the solution $x^\star$ [29], which when multiple solutions exist is the least-norm solution [19, 32]. For inconsistent systems, there exists at least one equation $\mathbf{A}_j x = b_j$ that is not satisfied by $x^\star$. As a result RK cannot converge for inconsistent systems, since it will occasionally project onto the solution space of such an equation. One can, however, guarantee exponential convergence in MSE to within a radius of the least-squares solution [21,23,32]. This radius is commonly referred to as the *convergence horizon*.

1.2 Randomized Kaczmarz with averaging

In order to take advantage of parallel computation and speed up the convergence of RK, we consider a simple extension of the RK method, where at each iteration multiple independent updates are computed in parallel and a weighted average of the updates is used. Specifically, we write the averaged RK update

$$x^{k+1} = x^k - \frac{1}{q} \sum_{i \in \tau_k} w_i \frac{\mathbf{A}_i x^k - b_i}{\|\mathbf{A}_i\|^2} \mathbf{A}_i^\top, \tag{3}$$

where τ_k is a random set of q row indices sampled *with replacement* and w_i represents the weight corresponding to the ith row. RK with averaging is detailed in Algorithm 1. If τ_k is a set of size one, i.e. $\tau_k = \{i_k\}$, and the weights are chosen as $w_i = 1$ for $i = 1, \ldots, m$, we recover the standard RK method.

Algorithm 1 Randomized Kaczmarz with Averaging

```
Input $\mathbf{A} \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, $x^0 \in \mathbb{R}^n$, weights $w \in \mathbb{R}^m$, number of maximum number of iterations $K$,
    distribution $\mathcal{D}$, number of threads $q$
for $k = 0, \ldots, K-1$ do
    $\tau_k \leftarrow q$ indices sampled from $\mathcal{D}$
    Compute $\delta \leftarrow \frac{1}{q} \sum_{i \in \tau_k} w_i \frac{\mathbf{A}_i x^k - b_i}{\|\mathbf{A}_i\|^2} \mathbf{A}_i^\top$ in parallel
    Update $x^{k+1} \leftarrow x^k - \delta$
Output $x^K$
```

1.3 Contributions

We derive a general convergence result for RK with averaging, and identify the conditions required for convergence to the least-squares solution. These conditions guide the choices of weights and probabilities of row selection, up to a relaxation parameter α. When $q = 1$ and appropriate weights and probabilities are chosen, we recover the standard convergence for RK [21,29,32].

For uniform weights and consistent systems, we relate RK with averaging to a more general parallel sketch-and-project method [27]. We also provide an estimate of the optimal choice for the relaxation parameter α, and compare to the estimated optimal relaxation parameter for the sketch-and-project method [27]. Through experiments, we show that our estimate lies closer to the observed result.

1.4 Organization

In Sect. 2, we give a general analysis of the convergence of RK with averaging and discuss the special case where the system is consistent. In Sect. 3, we discuss the special case where the weights are chosen to be uniform ($w_i = \alpha$ for all i) and derive the optimal relaxation parameter $\alpha^\star$ for consistent systems. In Sect. 4, we experimentally explore the effects of the number of threads q, the relaxation parameter α, the weights w_i, and the distribution $\mathcal{D}$ on the convergence properties of RK with averaging.

1.5 Related work

The Kaczmarz algorithm was originally proposed by Kaczmarz [16], though it was later independently developed by researchers in computed tomography as the Algebraic Reconstruction Technique [3,10]. The original Kaczmarz method cycles through rows in a fixed order; however, this is known to perform poorly for certain orders of the rows [12]. Other Kaczmarz variants [30] use deterministic methods to choose the rows, but their analysis is complicated and convergence results are somewhat unintuitive.

Some randomized control methods were proposed [15], but with no explicit proofs of convergence until Strohmer and Vershynin's 2009 paper [29], which proved that RK converges linearly in MSE, with a rate directly related to geometric properties of the matrix A. This proof was later extended to inconsistent systems [21], showing convergence within a convergence horizon of the least-squares solution.

RK is a well-studied method with many variants. We do not provide an exhaustive review of the related literature [5,7,17,24,32], but instead only remark on some closely related parallel extensions of RK.

Block Kaczmarz [1,6,8,23,31] randomly selects a block of rows from $\mathbf{A}$ at each iteration and computes its Moore-Penrose pseudoinverse. The pseudoinverse is then applied to the relevant portion of the current residual and added to the estimate, solving the least-squares problem only on the selected block of rows. Computing the pseudoinverse, however, is costly and difficult to parallelize.

The CARP algorithm [9] also distributes rows of $\mathbf{A}$ into blocks. However, instead of taking the pseudoinverse, the Kaczmarz method is then applied to the rows contained within each block. Multiple blocks are computed in parallel, and a component-averaging operator combines the approximations from each block. While CARP is shown to converge for consistent systems and to converge cyclically for inconsistent systems, no exponential convergence rate is given.

AsyRK [18] is an asynchronous parallel RK method that results from applying Hogwild! [26] to the least-squares objective. In AsyRK, each thread chooses a row $\mathbf{A}_i$ at random and updates a random coordinate within the support of that row $\mathbf{A}_i$ with a weighted RK update. AsyRK is shown to have exponential convergence, given conditions on the step size. Their analysis requires that $\mathbf{A}$ is sparse, while we do not make this restriction.

Recent work of Necoara [20] analyzes a slight generalization of Algorithm 1 under the name "randomized block Kaczmarz (RBK)". Rather than sampling indices i.i.d. as in Algorithm 1, RBK allows for more general sampling strategies such as sampling from a partition of the rows of $\mathbf{A}$. RBK was shown to converge exponentially in MSE to the solution of *consistent* systems of equations. The convergence rate of RBK shown by Necoara is dependent on the conditioning of the most ill-conditioned block of the partition when a partition is used and on the most ill-conditioned block of the entire matrix $\mathbf{A}$ when the indices are sampled i.i.d. as in Algorithm 1. Our analysis of Algorithm 1 does not depend on the most ill-conditioned block of the matrix $\mathbf{A}$ and applies to inconsistent systems as well as consistent systems.

RK falls under a more general class of methods often called sketch-and-project methods [11]. For a linear system $\mathbf{A}x = b$, sketch-and-project methods iteratively project the current iterate onto the solution space of a sketched subsystem $\mathbf{S}^\top \mathbf{A}x - \mathbf{S}^\top b$. In particular, RK is a sketch-and-project method with $\mathbf{S}^\top = \mathbf{I}_i$, where $\mathbf{I}_i$ is the ith row of the identity matrix. Other popular iterative methods such as coordinate descent can also be framed as sketch-and-project methods. In [27], the authors discuss a more general version of Algorithm 1 for sketch-and-project methods with averaging. Their analysis and discussion, however, focus on consistent systems and require uniform weights. We instead restrict our analysis to RK, but allow inconsistent systems and general weights w_i.

RK can also be interpreted as a subcase of stochastic gradient descent (SGD) [28] applied to the loss function [22]

$$F(x) = \sum_{i=1}^{m} f_i(x) = \sum_{i=1}^{m} \frac{1}{2}(\mathbf{A}_i x - b_i)^2.$$

In this context, RK with averaging can be seen as mini-batch SGD [2,25] with importance sampling, with the update

$$x^{k+1} = x^k - \frac{1}{q} \sum_{i \in \tau_k} \frac{w_i}{L_i} \nabla f_i(x),$$

where $L_i = \|\mathbf{A}_i\|^2$ is the Lipschitz constant of $\nabla f_i(x) = (\mathbf{A}_i x - b_i)\mathbf{A}_i^\top$.

2 Convergence of RK with averaging

For inconsistent systems, RK satisfies the error bound

$$\mathbb{E}_k \left[\|e^{k+1}\|^2 \right] \leq \left(1 - \frac{\sigma_{\min}^2(\mathbf{A})}{\|\mathbf{A}\|_F^2} \right) \|e^k\|^2 + \frac{\|r^\star\|^2}{\|\mathbf{A}\|_F^2}, \tag{4}$$

where $e^k \overset{\text{def}}{=} x^k - x^\star$ is the error of the kth iterate, $\sigma_{\min}(\mathbf{A})$ is the smallest *nonzero* singular value of $\mathbf{A}$, $\|\mathbf{A}\|_F^2 = \sum_{i,j} \mathbf{A}_{ij}^2$, $r^\star \overset{\text{def}}{=} b - \mathbf{A}x^\star$ is the least-squares residual, and $\mathbb{E}_k[\,\cdot\,] \overset{\text{def}}{=} \mathbb{E}\left[\,\cdot \mid \tau_{k-1}, \ldots, \tau_0\right]$ is the expectation conditioned on the samples from iterations $0, 1, \ldots, k-1$ with $\tau_k = \{i_k\}$ for RK [21,32]. Taking the full expectation on both sides of Eq. (4) and iterating the error bound yields

$$\mathbb{E}\left[\|e^k\|^2 \right] \leq \left(1 - \frac{\sigma_{\min}^2(\mathbf{A})}{\|\mathbf{A}\|_F^2} \right)^k \|e^0\|^2 + \frac{\|r^\star\|^2}{\sigma_{\min}^2(\mathbf{A})}.$$

For consistent systems the least-squares residual is $r^\star = 0$ and this bound guarantees exponential convergence in mean squared error (MSE) at a *convergence rate* of

$1 - \frac{\sigma_{\min}^2(\mathbf{A})}{\|\mathbf{A}\|_F^2}$ [29]. For inconsistent systems, this bound only guarantees exponential convergence in MSE to within a *convergence horizon* $\|r^\star\|^2/\sigma_{\min}^2(\mathbf{A})$.

We derive a convergence result for Algorithm 1 which is similar to Eq. (4) and leads to a better convergence rate and a smaller convergence horizon for inconsistent systems when using uniform weights. To analyze the convergence, we begin by finding the update to the error at each iteration. Subtracting the exact solution $x^\star$ from both sides of the update rule in Eq. (3) and using the fact that $\mathbf{A}_i e^k - r_i^\star = \mathbf{A}_i x^k - b_i$, we arrive at the error update

$$e^{k+1} = e^k - \frac{1}{q}\sum_{i \in \tau_k} w_i \frac{\mathbf{A}_i e^k - r_i^\star}{\|\mathbf{A}_i\|^2}\mathbf{A}_i^\top. \tag{5}$$

To simplify notation, we define the following matrices.

Definition 1 Define the weighted sampling matrix

$$\mathbf{M}_k \overset{\text{def}}{=} \frac{1}{q}\sum_{i \in \tau_k} w_i \frac{\mathbf{I}_i^\top \mathbf{I}_i}{\|\mathbf{A}_i\|^2},$$

where τ_k is a set of indices sampled independently from $\mathcal{D}$ with replacement and $\mathbf{I}$ is the identity matrix.

Using Definition 1, the error update from Eq. (5) can be rewritten as

$$e^{k+1} = (\mathbf{I} - \mathbf{A}^\top \mathbf{M}_k \mathbf{A})e^k + \mathbf{A}^\top \mathbf{M}_k r^\star. \tag{6}$$

Definition 2 Let $\mathbf{Diag}\,(d_1, d_2, \ldots, d_m)$ denote the diagonal matrix with $d_1, d_2, \ldots d_m$ on the diagonal. Define the normalization matrix

$$\mathbf{D} \overset{\text{def}}{=} \mathbf{Diag}\,(\|\mathbf{A}_1\|, \|\mathbf{A}_2\|, \ldots, \|\mathbf{A}_m\|)$$

so that the matrix $\mathbf{D}^{-1}\mathbf{A}$ has rows with unit norm, the probability matrix

$$\mathbf{P} \overset{\text{def}}{=} \mathbf{Diag}\,(p_1, p_2, \ldots, p_m)\,,$$

where $p_j = \mathbb{P}(i = j)$ with $i \sim \mathcal{D}$, and the weight matrix

$$\mathbf{W} \overset{\text{def}}{=} \mathbf{Diag}\,(w_1, w_2, \ldots, w_m)\,.$$

The convergence analysis additionally relies on the expectations given in Lemma 1, whose proof can be found in Appendix A.

Lemma 1 *Let the weighted sampling matrix* $\boldsymbol{M}_k$*, the normalization matrix* $\boldsymbol{D}$*, the probability matrix* $\boldsymbol{P}$*, and the weight matrix* $\mathbf{W}$ *be defined as in Definitions 1 and 2. Then*

$$\mathbb{E}_k\left[\boldsymbol{M}_k\right] = \boldsymbol{PWD}^{-2}$$

and

$$\mathbb{E}_k\left[\boldsymbol{M}_k^\top \boldsymbol{AA}^\top \boldsymbol{M}_k\right] = \frac{1}{q}\boldsymbol{PW}^2\boldsymbol{D}^{-2} + \left(1 - \frac{1}{q}\right)\boldsymbol{PWD}^{-2}\boldsymbol{AA}^\top\boldsymbol{PWD}^{-2}.$$

2.1 Coupling of weights and probabilities

Note that the weighted sampling matrix $\mathbf{M}_k$ is a sample average, with the number of samples being the number of threads q. Thus, as the number of threads q goes to infinity, we have

$$\mathbf{M}_k \overset{q\to\infty}{\longrightarrow} \mathbb{E}_{i\sim\mathcal{D}}\left[w_i \frac{\mathbf{I}_i^\top \mathbf{I}_i}{\|\mathbf{A}_i\|^2}\right] = \mathbf{PWD}^{-2}.$$

Therefore, as we take more and more threads, the averaged RK update of Eq. (3) approaches the deterministic update

$$x^{k+1} = (\mathbf{I} - \mathbf{A}^\top\mathbf{PWD}^{-2}\mathbf{A})x^k + \mathbf{A}^\top\mathbf{PWD}^{-2}b.$$

Likewise, the corresponding error update in Eq. (6) approaches the deterministic update

$$e^{k+1} = (\mathbf{I} - \mathbf{A}^\top\mathbf{PWD}^{-2}\mathbf{A})e^k + \mathbf{A}^\top\mathbf{PWD}^{-2}r^\star.$$

Since we want the error of the limiting averaged RK method to converge to zero, we should require that this limiting error update have the zero vector as a fixed point. Thus, we ask that

$$0 = \mathbf{A}^\top\mathbf{PWD}^{-2}r^\star$$

for any least-squares residual $r^\star$. This is guaranteed if $\mathbf{PWD}^{-2} = \beta\mathbf{I}$ for some scalar β. For convenience, we choose to express β as $\frac{\alpha}{\|\mathbf{A}\|_F^2}$ for some relaxation parameter α.

Assumption 1 The probability matrix $\mathbf{P}$ and weight matrix $\mathbf{W}$ are chosen to satisfy

$$\mathbf{PWD}^{-2} = \frac{\alpha}{\|\mathbf{A}\|_F^2}\mathbf{I}.$$

for some scalar relaxation parameter $\alpha > 0$.

2.2 General result

We now state a general convergence result for RK with averaging in Theorem 1. The proof is given in Appendix B. Theorem 1 in its general form is difficult to interpret, so we defer a detailed analysis to Sect. 3 in which the assumption of uniform weights ($\mathbf{W} = \alpha\mathbf{I}$) simplifies the bound significantly.

Theorem 1 *Let the weighted sampling matrix $\boldsymbol{M}_k$, the normalization matrix $\boldsymbol{D}$, the probability matrix $\boldsymbol{P}$, and the weight matrix $\mathbf{W}$ be defined as in Definitions 1 and 2. Suppose $\boldsymbol{P}$ and $\mathbf{W}$ are chosen such that $\boldsymbol{P}\mathbf{W}\boldsymbol{D}^{-2} = \frac{\alpha}{\|\boldsymbol{A}\|_F^2}\boldsymbol{I}$ for relaxation parameter $\alpha > 0$ (Assumption 1). Then the error at each iteration of Algorithm 1 satisfies*

$$\mathbb{E}_k\left[\|e^{k+1}\|^2\right] \leq \sigma_{\max}\left(\left(\boldsymbol{I} - \alpha\frac{\boldsymbol{A}^\top\boldsymbol{A}}{\|\boldsymbol{A}\|_F^2}\right)^2 - \frac{\alpha^2}{q}\left(\frac{\boldsymbol{A}^\top\boldsymbol{A}}{\|\boldsymbol{A}\|_F^2}\right)^2\right)\|e^k\|^2 + \frac{\alpha}{q}\frac{\|r^k\|_{\mathbf{W}}^2}{\|\boldsymbol{A}\|_F^2},$$

where $r^k \overset{def}{=} b - \boldsymbol{A}x^k$ is the residual of the kth iterate, $\|\cdot\|_{\mathbf{W}}^2 = \langle\cdot, \mathbf{W}\cdot\rangle$ and $\|\boldsymbol{A}\|_F^2 = \sum_{i,j}\boldsymbol{A}_{ij}^2$.

From Theorem 1, we see that the residual term decreases $\frac{\alpha}{q}\frac{\|r^k\|_{\mathbf{W}}^2}{\|\mathbf{A}\|_F^2} \to 0$ as the number of threads increases $q \to \infty$. Additionally, the convergence rate of the MSE $\mathbb{E}\left[\|e^k\|^2\right]$ approaches $\sigma_{\max}^2\left(\mathbf{I} - \alpha\frac{\mathbf{A}^\top\mathbf{A}}{\|\mathbf{A}\|_F^2}\right)$.

The recent work of Necoara [20] points out that

$$\|r^k\|_{\mathbf{W}}^2 \leq \left(\max_{i\in[m]} w_i\right)\|r^k\|^2.$$

Using this fact, the dependence of Theorem 1 on r^k can be loosened to a dependence on $r^\star$ since

$$\|r^k\|^2 = \|\mathbf{A}e^k\|^2 + \|r^\star\|^2.$$

In this way, results analogous to those in Sect. 3 may be derived without restricting to uniform weights ($\mathbf{W} = \alpha\mathbf{I}$).

2.3 Consistent systems

For consistent systems, Algorithm 1 converges to the solution $x^\star$ exponentially in MSE with the following guaranteed convergence rate.

Corollary 1 *Let the weighted sampling matrix $\boldsymbol{M}_k$, the normalization matrix $\boldsymbol{D}$, the probability matrix $\boldsymbol{P}$, and the weight matrix $\mathbf{W}$ be defined as in Definitions 1 and 2. Suppose that $\boldsymbol{P}$ and $\mathbf{W}$ are chosen such that $\boldsymbol{P}\mathbf{W}\boldsymbol{D}^{-2} = \frac{\alpha}{\|\mathbf{A}\|_F^2}\boldsymbol{I}$ for relaxation parameter $\alpha > 0$ (Assumption 1) and that the system of equations $\boldsymbol{A}x = b$ is consistent. Then the error at each iteration of Algorithm 1 satisfies*

$$\mathbb{E}_k\left[\|e^{k+1}\|^2\right] \leq \sigma_{\max}\left(\left(\boldsymbol{I} - \alpha\frac{\boldsymbol{A}^\top\boldsymbol{A}}{\|\boldsymbol{A}\|_F^2}\right)^2 + \frac{\boldsymbol{A}^\top}{\|\boldsymbol{A}\|_F}\left(\frac{\alpha}{q}\mathbf{W} - \frac{\alpha^2}{q}\frac{\boldsymbol{A}\boldsymbol{A}^\top}{\|\boldsymbol{A}\|_F^2}\right)\frac{\boldsymbol{A}}{\|\boldsymbol{A}\|_F}\right)\|e^k\|^2.$$

Corollary 1 can be derived from the proof of Theorem 1 with $r^\star = 0$.

3 Uniform weights

In this section, we simplify Theorem 1 under the additional assumption that the weights are uniform. That is, $w_i = \alpha$ for all i. Under this assumption, the convergence horizon can be determined explicitly, removing the dependence on the current residual r^k in Theorem 1. In this case, the dependence of the convergence rate and convergence horizon on the relaxation parameter α and number of threads q becomes clear.

When we assume that the weights are uniform, the update from Eq. (3) becomes

$$x^{k+1} = x^k - \frac{\alpha}{q} \sum_{i \in \tau_k} \frac{\mathbf{A}_i x^k - b_i}{\|\mathbf{A}_i\|^2} \mathbf{A}_i^\top,$$

where $i \in \tau_k$ are independent samples from $\mathcal{D}$ with $p_i = \frac{\|\mathbf{A}_i\|^2}{\|\mathbf{A}\|_F^2}$. Under these conditions, the convergence result of Theorem 1 can be simplified to remove the dependence on r^k. This simplification leads to the more interpretable error bound given in Corollary 2. In particular, increasing the number of threads q leads to both a faster convergence rate and smaller convergence horizon. If the relaxation parameter α is chosen as $\alpha = 1$ and a single row is selected at each iteration, i.e. $q = 1$, we arrive at the RK method [29]. Using a relaxation parameter α other than one results in the relaxed RK method [13,14].

Corollary 2 *Suppose the probabilities $p_i = \frac{\|\boldsymbol{A}_i\|^2}{\|\boldsymbol{A}\|_F^2}$ and the weights $w_i = \alpha$ for all i. Then the error at each iteration of Algorithm 1 satisfies*

$$\mathbb{E}_k\left[\|e^{k+1}\|^2\right] \le \sigma_{\max}\left(\left(\boldsymbol{I} - \alpha \frac{\boldsymbol{A}^\top \boldsymbol{A}}{\|\boldsymbol{A}\|_F^2}\right)^2 + \frac{\alpha^2}{q}\left(\boldsymbol{I} - \frac{\boldsymbol{A}^\top \boldsymbol{A}}{\|\boldsymbol{A}\|_F^2}\right)\frac{\boldsymbol{A}^\top \boldsymbol{A}}{\|\boldsymbol{A}\|_F^2}\right)\|e^k\|^2 + \frac{\alpha^2 \|r^\star\|^2}{q\|\boldsymbol{A}\|_F^2}.$$

The proof of Theorem 2 follows immediately from Theorem 1 and can be found in Appendix D.1.

Corollary 2 shows that the convergence horizon is proportional to $\frac{\alpha^2}{q}$, so smaller relaxation parameters α and larger number of threads q lead to a smaller convergence horizon. From the convergence rate term of Corollary 2, we see that increasing the relaxation parameter α improves the convergence rate of the algorithm up to a optimal relaxation parameter $\alpha^\star$ beyond which further increasing α leads to slower convergence rates. Increasing the number of threads q improves the convergence rate, asymptotically approaching an optimal rate as $q \to \infty$.

If a single row is chosen at each iteration, with weights $w_i = 1$ and probabilities $p_i = \frac{\|\mathbf{A}_i\|^2}{\|\mathbf{A}\|_F^2}$, then Algorithm 1 becomes the version of RK stated in [29]. In this case,

$$\|r^k\|_{\mathbf{W}}^2 = \|\mathbf{A}e^k\|^2 + \|r^\star\|^2. \tag{7}$$

Applying Corollary 2 leads to the following result, which recovers the error bound in Eq. (4).

Corollary 3 *Suppose the number of threads $q = 1$, the weights $w_i = 1$ for all i, and the probabilities $p_i = \frac{\|\boldsymbol{A}_i\|^2}{\|\boldsymbol{A}\|_F^2}$. Then the error at each iteration of Algorithm 1 satisfies*

$$\begin{aligned}\mathbb{E}_k\left[\|e^{k+1}\|^2\right] &\leq \sigma_{\max}\left(\boldsymbol{I} - \frac{\boldsymbol{A}^\top\boldsymbol{A}}{\|\boldsymbol{A}\|_F^2}\right)\|e^k\|^2 + \frac{\|r^\star\|^2}{\|\boldsymbol{A}\|_F^2}\\ &= \left(1 - \frac{\sigma_{\min}^2(\boldsymbol{A})}{\|\boldsymbol{A}\|_F^2}\right)\|e^k\|^2 + \frac{\|r^\star\|^2}{\|\boldsymbol{A}\|_F^2}.\end{aligned}$$

A proof of Corollary 3 is included in Appendix D.2.

3.1 Suggested relaxation parameter α for consistent systems with uniform weights

For consistent systems and using uniform weights,Algorithm 1 becomes a subcase of the parallel sketch-and-project method described by Richtárik and Takáč [27]. They suggest a choice for the relaxation parameter

$$\alpha^{\mathrm{RT}} = \frac{q}{1 + (q-1)\,\frac{\sigma_{\max}^2(\mathbf{A})}{\|\mathbf{A}\|_F^2}} \tag{8}$$

chosen to optimize their convergence guarantee

$$\mathbb{E}_k\left[\|e^{k+1}\|^2\right] \leq \left(1 - \alpha\left(2 - \frac{\alpha}{q}\left(1 + (q-1)\,\frac{\sigma_{\max}^2(\mathbf{A})}{\|\mathbf{A}\|_F^2}\right)\right)\frac{\sigma_{\min}^2(\mathbf{A})}{\|\mathbf{A}\|_F^2}\right)\|e^k\|^2. \tag{9}$$

Analogously, for consistent systems and using uniform weights, we can calculate the value of α to minimize the bound given in Corollary 2.

Theorem 2 *Suppose the probabilities $p_i = \frac{\|\boldsymbol{A}_i\|^2}{\|\boldsymbol{A}\|_F^2}$ and the weights $w_i = \alpha$. Suppose also that the system of equations $\boldsymbol{A}x = b$ is consistent. Then, the relaxation parameter α which yields the fastest convergence guarantee in Corollary 1 is*

$$\alpha^\star = \begin{cases} \frac{q}{1+(q-1)s_{\min}}, & s_{\max} - s_{\min} \leq \frac{1}{1-q}\\ \frac{2q}{1+(q-1)(s_{\min}+s_{\max})}, & s_{\max} - s_{\min} > \frac{1}{1-q}\end{cases} \tag{10}$$

where $s_{\min} = \frac{\sigma_{\min}^2(\boldsymbol{A})}{\|\boldsymbol{A}\|_F^2}$ and $s_{\max} = \frac{\sigma_{\max}^2(\boldsymbol{A})}{\|\boldsymbol{A}\|_F^2}$.

The proof of this result can be found in Appendix C.

When a single thread $q = 1$ is used, we see that our optimal relaxation parameter is $\alpha^\star = 1$. Whereas, when multiple threads $q > 1$ are used, we see

$$1 < \alpha^\star \leq q$$

since

$$0 < s_{\min} \leq s_{\max} \leq 1.$$

Additionally, by viewing the condition $s_{\max} - s_{\min} \leq \frac{1}{1-q}$ in terms of the number of threads, $q \leq 1 + \frac{1}{s_{\max}-s_{\min}}$, we see that for low numbers of threads the first form $\alpha^\star = \frac{q}{1+(q-1)s_{\min}}$ is used, while for high numbers of threads, the second form $\alpha^\star = \frac{2q}{1+(q-1)(s_{\min}+s_{\max})}$ is used.

Note that our relaxation parameter $\alpha^\star$ differs from the relaxation parameter α^{RT} suggested by Richtárik and Takáč [27], given in Eq. (8). This is due to the fact that our convergence rate guarantee is tighter, and thus we expect that our suggested relaxation parameter $\alpha^\star$ should be closer to the truly optimal value. We compare these two choices of the relaxation parameter α experimentally in Sect. 4.3 and show that our suggested relaxation parameter $\alpha^\star$ is indeed closer to the true optimal value, especially for large numbers of threads q.

4 Experiments

We present several experiments to demonstrate the convergence of Algorithm 1 under various conditions. In particular, we study the effects of the number of threads q, the relaxation parameter α, the weight matrix $\mathbf{W}$, and the probability matrix $\mathbf{P}$.

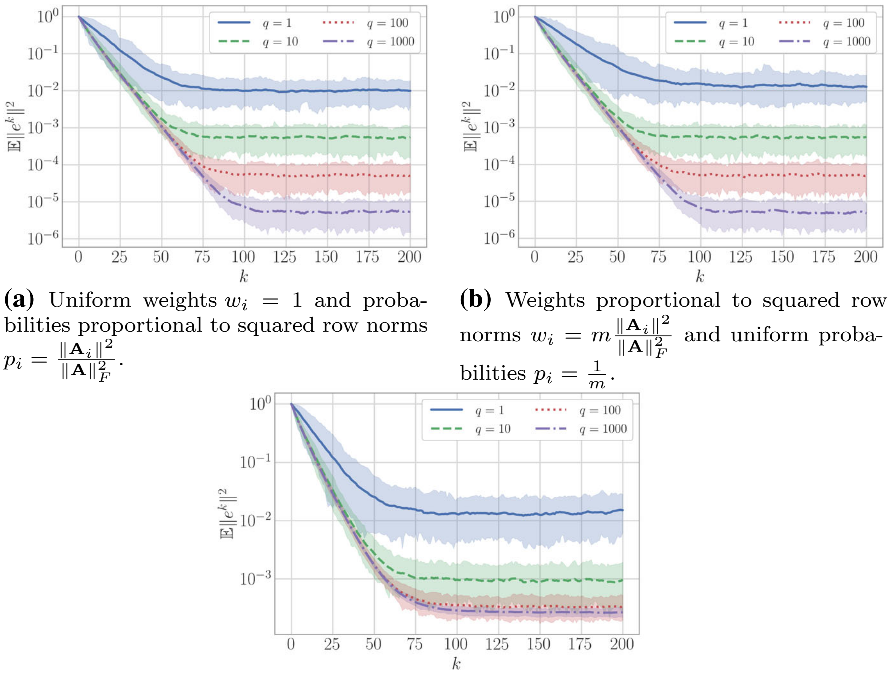
4.1 Procedure

For each experiment, we run 100 independent trials each starting with the initial iterate $x^0 = 0$ and average the squared error norms $\|e^k\|^2$ across the trials to estimate the MSE $\mathbb{E}\left[\|e^k\|^2\right]$. Shaded confidence intervals for the 5th and 95th percentiles are plotted when appropriate. For some plots, such as Fig. 3, outlier trials cause $\mathbb{E}\left[\|e^k\|^2\right]$ to lie outside of the confidence intervals. We sample $\mathbf{A}$ from 100×10 standard Gaussian matrices and least-squares solution $x^\star$ from 10-dimensional standard Gaussian vectors, normalized so that $\|x^\star\| = 1$. To form inconsistent systems, we generate the least-squares residual $r^\star$ as a Gaussian vector orthogonal to the range of $\mathbf{A}$, also normalized so that $\|r^\star\| = 1$. Finally, the right side b is computed as $r^\star + \mathbf{A}x^\star$.

4.2 The effect of the number of threads

In Fig. 1, we see the effects of the number of threads q on the approximation error of Algorithm 1 for different choices of the weight matrices $\mathbf{W}$ and probability matrices $\mathbf{P}$. In Fig. 1a, b $\mathbf{W}$ and $\mathbf{P}$ satisfy Assumption 1, while in Fig. 1c they do not.

In ,Figs. 1a, b as the number of threads q increases by a factor of ten, we see a corresponding decrease in the magnitude of the convergence horizon by approximately the same factor. This result corroborates what we expect based on Theorem 1 and Corollary 2. For Fig. 1c, we do not see the same consistent decrease in the magnitude of the convergence horizon. As q increases, for weight matrices $\mathbf{W}$ and probability



(a) Uniform weights $w_i = 1$ and probabilities proportional to squared row norms $p_i = \frac{\|\mathbf{A}_i\|^2}{\|\mathbf{A}\|_F^2}$.

(b) Weights proportional to squared row norms $w_i = m\frac{\|\mathbf{A}_i\|^2}{\|\mathbf{A}\|_F^2}$ and uniform probabilities $p_i = \frac{1}{m}$.

(c) Uniform weights $w_i = 1$ and uniform probabilities $p_i = \frac{1}{m}$.

Fig. 1 The effect of the number of threads on the MSE versus iteration for Algorithm 1 applied to inconsistent systems. The weights w_i and probabilities p_i in a and b satisfy Assumption 1, while in c they do not. Shaded regions are 5th and 95th percentiles, measured over 100 trials

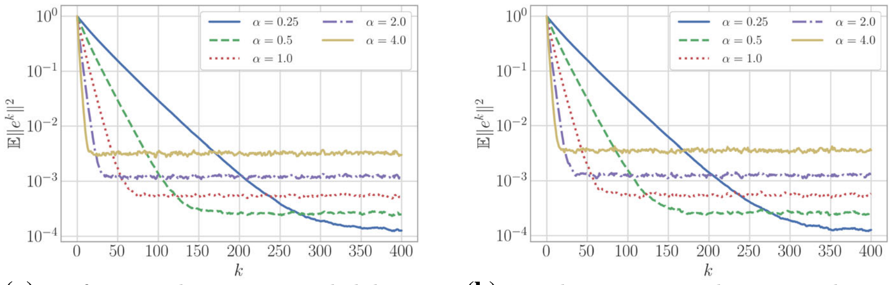
matrices $\mathbf{P}$ that do not satisfy Assumption 1, the iterates x^k approach a weighted least-squares solution instead of the desired least-squares solution $x^\star$ (see Sect. 2.1).

The rate of convergence in Fig. 1 also improves as the number of threads q increases. As q increases, we see diminishing returns in the convergence rate. We expect this behavior based on the dependence on $\frac{1}{q}$ in Theorem 1 and Corollary 2.

4.3 The effect of the relaxation parameter α

In Fig. 2, we observe the effect on the convergence rate and convergence horizon as we vary the relaxation parameter α. From Theorem 1, we expect that the convergence horizon increases with α and indeed observe this experimentally. The MSE $\mathbb{E}\left[\|e^k\|^2\right]$ behaves similarly as α varies for both sets of weights and probabilities considered, each of which satisfy Assumption 1.

For larger values of the relaxation parameter α, the convergence rate for Algorithm 1 eventually decreases and the method can ultimately diverge. This behavior can be seen in Fig. 3, which plots the estimated MSE after 100 iterations for consistent Gaussian systems, various α, and various numbers of threads q.



(a) Uniform weights $w_i = \alpha$, probabilities proportional to squared row norms $p_i = \frac{\|\mathbf{A}_i\|^2}{\|\mathbf{A}\|_F^2}$, and number of threads $q = 10$.

(b) Weights proportional to squared row norms $w_i = \alpha m \frac{\|\mathbf{A}_i\|^2}{\|\mathbf{A}\|_F^2}$, uniform probabilities $p_i = \frac{1}{m}$, and number of threads $q = 10$.

Fig. 2 The effect of the relaxation parameter α on the MSE versus iteration for Algorithm 1 applied to inconsistent systems

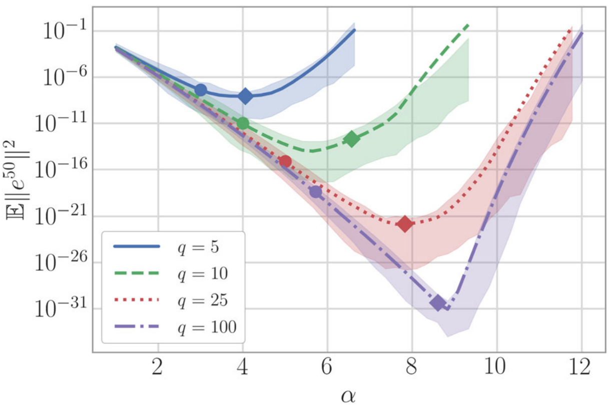


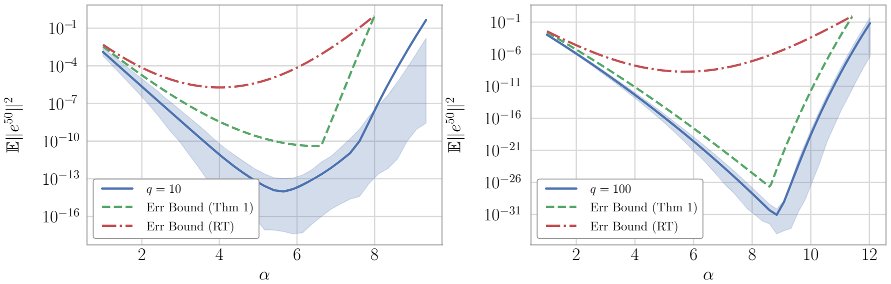
Fig. 3 Estimated MSE after 50 iterations of Algorithm 1 on consistent systems with weights $w_i = \alpha$, and probabilities $p_i = \frac{\|\mathbf{A}_i\|^2}{\|\mathbf{A}\|_F^2}$ for various choices of relaxation parameter α. Shaded regions are the 5th and 95th percentiles, measured over 100 trials. Diamond markers are estimates of the optimal relaxation parameter using Theorem 2, and circle markers are estimates using the formula from Richtárik and Takáč [27]

For each value of q, we plot two markers on the curve to show the estimated optimal values of α. The diamond markers are optimal values of $\alpha^\star$ computed using Theorem 2, and the circle markers are optimal values of α^{RT} using Eq. (8) from Richtárik and Takáč [27]. These values are also contained in Table 1. In terms of the number of iterations required, we find that the optimal value for α increases with q. Comparing the α^{RT} values from [27] with the α that minimize the curves in Fig. 3, we find that these values generally underestimate the optimal α that we observe experimentally. In comparison, the optimal $\alpha^\star$ calculated using Theorem 2 are much closer to the empirically optimal values of α, especially for high q.

We believe this is due to our bound being relatively tighter than Eq. (8). In Fig. 4a, b, we plot the error bounds produced by Eqs. (8) and (10) after 50 iterations for

Table 1 Calculated optimal relaxation parameters α for matrix $\mathbf{A}$ used in Fig. 3 for various numbers of threads q

	$q = 5$	$q = 10$	$q = 25$	$q = 100$
α^{RT} [Eq. (8)] [27]	3.00	4.00	5.00	5.72
$\alpha^{\star}$ (Theorem 2)	4.06	6.57	7.83	8.61



(a) $q = 10$. **(b)** $q = 100$.

Fig. 4 Estimated MSE after 50 iterations of Algorithm 1 on consistent systems for various choices of relaxation parameter α. Uniform weights $w_i = \alpha$ and probabilities proportional to squared row norms $p_i = \frac{\|\mathbf{A}_i\|^2}{\|\mathbf{A}\|_F^2}$. The first error bound is from Theorem 1, while the second is from Eq. 8 [27]

$q = 10$ and $q = 100$. We observe that as the number of threads increases, our bound approaches the empirical result.

5 Conclusion

We prove a general error bound for RK with averaging (Algorithm 1) in terms of the number of threads q and a relaxation parameter α. We find a natural coupling between the probabilities p_i and the weights w_i that leads to a reduced convergence horizon. We demonstrate that for uniform weights ($w_i = \alpha$ for all i), the rate of convergence and convergence horizon for Algorithm 1 improve both in theory and practice as the number of threads q increases. Based on the error bound, we also derive an optimal value for the relaxation parameter α which increases convergence speed, and compare with existing results.

A Proof of Lemma 1

Let $\mathbb{E}_i\left[\,\cdot\,\right]$ denote $\mathbb{E}_{i\sim\mathcal{D}}\left[\,\cdot\,\right]$. Expanding the definition of the weighted sampling matrix $\mathbf{M}_k$ as a weighted average of the i.i.d. sampling matrices $\frac{\mathbf{I}_i^\top \mathbf{I}_i}{\|\mathbf{A}_i\|^2}$, we see that

$$\mathbb{E}_k\left[\mathbf{M}_k\right] = \mathbb{E}_k\left[\frac{1}{q}\sum_{i\in\tau_k} w_i \frac{\mathbf{I}_i^\top \mathbf{I}_i}{\|\mathbf{A}_i\|^2}\right] = \mathbb{E}_i\left[w_i \frac{\mathbf{I}_i^\top \mathbf{I}_i}{\|\mathbf{A}_i\|^2}\right] = \sum_{i=1}^m p_i w_i \frac{\mathbf{I}_i^\top \mathbf{I}_i}{\|\mathbf{A}_i\|^2} = \mathbf{PWD}^{-2}.$$

Likewise, we can compute

$$\begin{aligned}
&\mathbb{E}_k\left[\mathbf{M}_k^\top \mathbf{A}\mathbf{A}^\top \mathbf{M}_k\right]\\
&= \mathbb{E}_k\left[\left(\frac{1}{q}\sum_{i\in\tau_k} w_i \frac{\mathbf{I}_i^\top \mathbf{A}_i}{\|\mathbf{A}_i\|^2}\right)\left(\frac{1}{q}\sum_{j\in\tau_k} w_j \frac{\mathbf{A}_j^\top \mathbf{I}_j}{\|\mathbf{A}_j\|^2}\right)\right]\\
&= \frac{1}{q}\mathbb{E}_i\left[\left(w_i \frac{\mathbf{I}_i^\top \mathbf{A}_i}{\|\mathbf{A}_i\|^2}\right)\left(w_i \frac{\mathbf{A}_i^\top \mathbf{I}_i}{\|\mathbf{A}_i\|^2}\right)\right] + (1-\frac{1}{q})\mathbb{E}_i\left[w_i \frac{\mathbf{I}_i^\top \mathbf{A}_i}{\|\mathbf{A}_i\|^2}\right]\mathbb{E}_i\left[w_i \frac{\mathbf{A}_i^\top \mathbf{I}_i}{\|\mathbf{A}_i\|^2}\right]\\
&= \frac{1}{q}\mathbb{E}_i\left[\left(w_i \frac{\mathbf{I}_i^\top \mathbf{A}_i}{\|\mathbf{A}_i\|^2}\right)\left(w_i \frac{\mathbf{A}_i^\top \mathbf{I}_i}{\|\mathbf{A}_i\|^2}\right)\right] + (1-\frac{1}{q})\mathbb{E}_i\left[w_i \frac{\mathbf{I}_i^\top \mathbf{I}_i}{\|\mathbf{A}_i\|^2}\right]\mathbf{A}\mathbf{A}^\top\mathbb{E}_i\left[w_i \frac{\mathbf{I}_i^\top \mathbf{I}_i}{\|\mathbf{A}_i\|^2}\right]\\
&= \frac{1}{q}\mathbb{E}_i\left[w_i^2 \frac{\mathbf{I}_i^\top \mathbf{I}_i}{\|\mathbf{A}_i\|^2}\right] + \left(1-\frac{1}{q}\right)\mathbf{PWD}^{-2}\mathbf{A}\mathbf{A}^\top\mathbf{PWD}^{-2}\\
&= \frac{1}{q}\mathbf{PW}^2\mathbf{D}^{-2} + \left(1-\frac{1}{q}\right)\mathbf{PWD}^{-2}\mathbf{A}\mathbf{A}^\top\mathbf{PWD}^{-2}
\end{aligned}$$

by separating the cases where $i = j$ from those where $i \neq j$ and utilizing the independence of the indices sampled in τ_k.

B Proof of Theorem 1

We now prove Theorem 1 starting from the error update in Eq. (6). Expanding the squared error norm,

$$\begin{aligned}
\|e^{k+1}\|^2 &= \|(\mathbf{I} - \mathbf{A}^\top\mathbf{M}_k\mathbf{A})e^k + \mathbf{A}^\top\mathbf{M}_k r^\star\|^2\\
&= \|(\mathbf{I} - \mathbf{A}^\top\mathbf{M}_k\mathbf{A})e^k\|^2 + 2\langle(\mathbf{I} - \mathbf{A}^\top\mathbf{M}_k\mathbf{A})e^k, \mathbf{A}^\top\mathbf{M}_k r^\star\rangle + \|\mathbf{A}^\top\mathbf{M}_k r^\star\|^2.
\end{aligned} \tag{11}$$

Under Assumption 1, the expectations in Lemma 1 simplify to

$$\mathbb{E}_k\left[\mathbf{M}_k\right] = \frac{\alpha}{\|\mathbf{A}\|_F^2}\mathbf{I}$$

and

$$\mathbb{E}_k\left[\mathbf{M}_k^\top \mathbf{A}\mathbf{A}^\top \mathbf{M}_k\right] = \frac{\alpha}{q}\frac{\mathbf{W}}{\|\mathbf{A}\|_F^2} + \alpha^2\left(1-\frac{1}{q}\right)\frac{\mathbf{A}\mathbf{A}^\top}{\|\mathbf{A}\|_F^4}.$$

Upon taking expections on both sides of Eq. (11), the middle term simplifies since

$$\mathbb{E}_k\left[\langle e^k, \mathbf{A}^\top \mathbf{M}_k r^\star\rangle\right] = \langle e^k, \mathbf{A}^\top \mathbb{E}_k\left[\mathbf{M}_k\right] r^\star\rangle = \langle e^k, \frac{\alpha}{\|\mathbf{A}\|_F^2}\mathbf{A}^\top r^\star\rangle = 0.$$

Thus,

$$\begin{aligned}&\mathbb{E}_k\left[\|e^{k+1}\|^2\right]\\ &= \underbrace{\mathbb{E}_k\left[\|(\mathbf{I}-\mathbf{A}^\top\mathbf{M}_k\mathbf{A})e^k\|^2\right]}_{\textcircled{1}} - \underbrace{2\mathbb{E}_k\left[\langle\mathbf{A}^\top\mathbf{M}_k\mathbf{A}e^k, \mathbf{A}^\top\mathbf{M}_k r^\star\rangle\right]}_{\textcircled{1}} + \underbrace{\mathbb{E}_k\left[\|\mathbf{A}^\top\mathbf{M}_k r^\star\|^2\right]}_{\textcircled{1}}.\end{aligned} \tag{12}$$

Making use of Lemma 1 to take the expectation in the first term in Eq. (12),

$$\begin{aligned}\textcircled{1} &= \mathbb{E}_k\left[\|(\mathbf{I}-\mathbf{A}^\top\mathbf{M}_k\mathbf{A})e^k\|^2\right]\\ &= \mathbb{E}_k\left[\left\langle e^k, (\mathbf{I}-\mathbf{A}^\top\mathbf{M}_k\mathbf{A})^\top(\mathbf{I}-\mathbf{A}^\top\mathbf{M}_k\mathbf{A})e^k\right\rangle\right]\\ &= \left\langle e^k, (\mathbf{I}-2\mathbf{A}^\top\mathbb{E}_k\left[\mathbf{M}_k\right]\mathbf{A}+\mathbf{A}^\top\mathbb{E}_k\left[\mathbf{M}_k^\top\mathbf{A}\mathbf{A}^\top\mathbf{M}_k\right]\mathbf{A})e^k\right\rangle\\ &\overset{lem.1}{=} \left\langle e^k, \left(\mathbf{I}-2\alpha\frac{\mathbf{A}^\top\mathbf{A}}{\|\mathbf{A}\|_F^2}+\frac{\alpha}{q}\frac{\mathbf{A}^\top\mathbf{W}\mathbf{A}}{\|\mathbf{A}\|_F^2}+\alpha^2\left(1-\frac{1}{q}\right)\left(\frac{\mathbf{A}^\top\mathbf{A}}{\|\mathbf{A}\|_F^2}\right)^2\right)e^k\right\rangle\\ &= \left\langle e^k, \left(\left(\mathbf{I}-\alpha\frac{\mathbf{A}^\top\mathbf{A}}{\|\mathbf{A}\|_F^2}\right)^2+\frac{\mathbf{A}^\top}{\|\mathbf{A}\|_F}\left(\frac{\alpha}{q}\mathbf{W}-\frac{\alpha^2}{q}\frac{\mathbf{A}\mathbf{A}^\top}{\|\mathbf{A}\|_F^2}\right)\frac{\mathbf{A}}{\|\mathbf{A}\|_F}\right)e^k\right\rangle.\end{aligned}$$

For the second term in Eq. 12,

$$\begin{aligned}\textcircled{2} &= 2\mathbb{E}_k\left[\langle\mathbf{A}^\top\mathbf{M}_k\mathbf{A}e^k, \mathbf{A}^\top\mathbf{M}_k r^\star\rangle\right]\\ &= 2\langle\mathbf{A}e^k, \mathbb{E}_k\left[\mathbf{M}_k^\top\mathbf{A}\mathbf{A}^\top\mathbf{M}_k\right]r^\star\rangle\\ &\overset{lem.1}{=} 2\langle\mathbf{A}e^k, \left(\frac{\alpha}{q}\mathbf{W}+\alpha^2\left(1-\frac{1}{q}\right)\mathbf{A}\mathbf{A}^\top\right)r^\star\rangle\\ &\overset{\mathbf{A}^\top r^\star=0}{=} 2\frac{\alpha}{q\|\mathbf{A}\|_F^2}\langle\mathbf{A}e^k, \mathbf{W}r^\star\rangle.\end{aligned}$$

Similarly, for the last term in Eq. (12),

$$\textcircled{3} = \mathbb{E}_k\left[\|\mathbf{A}^\top\mathbf{M}_k r^\star\|^2\right] = \frac{\alpha}{q}\frac{\|r^\star\|_{\mathbf{W}}^2}{\|\mathbf{A}\|_F^2}.$$

Combining these in Eq. (12),

$$
\begin{aligned}
\mathbb{E}\left[\|e^{k+1}\|^2\right] &= \left\langle e^k, \left(\mathbf{I} - \alpha \frac{\mathbf{A}^\top \mathbf{A}}{\|\mathbf{A}\|_F^2}\right)^2 e^k \right\rangle \\
&\quad + \left\langle e^k, \frac{\mathbf{A}^\top}{\|\mathbf{A}\|_F^2} \left(\frac{\alpha}{q}\mathbf{W} - \frac{\alpha^2}{q}\frac{\mathbf{A}\mathbf{A}^\top}{\|\mathbf{A}\|_F^2}\right)\mathbf{A}e^k \right\rangle - 2\frac{\alpha}{q}\frac{\langle \mathbf{A}e^k, \mathbf{W}r^\star\rangle}{\|\mathbf{A}\|_F^2} + \frac{\alpha}{q}\frac{\|r^\star\|_\mathbf{W}^2}{\|\mathbf{A}\|_F^2} \\
&= \left\langle e^k, \left(\left(\mathbf{I} - \alpha \frac{\mathbf{A}^\top \mathbf{A}}{\|\mathbf{A}\|_F^2}\right)^2 - \frac{\alpha^2}{q}\left(\frac{\mathbf{A}^\top \mathbf{A}}{\|\mathbf{A}\|_F^2}\right)^2\right) e^k \right\rangle + \frac{\alpha}{q}\frac{\|r^k\|_\mathbf{W}^2}{\|\mathbf{A}\|_F^2} \\
&\le \sigma_{\max}\left(\left(\mathbf{I} - \alpha \frac{\mathbf{A}^\top \mathbf{A}}{\|\mathbf{A}\|_F^2}\right)^2 - \frac{\alpha^2}{q}\left(\frac{\mathbf{A}^\top \mathbf{A}}{\|\mathbf{A}\|_F^2}\right)^2\right)\|e^k\|^2 + \frac{\alpha}{q}\frac{\|r^k\|_\mathbf{W}^2}{\|\mathbf{A}\|_F^2}.
\end{aligned}
$$

C Proof of Theorem 2

Proof We seek to optimize the convergence rate constant from Corollary 1 when using uniform weights $\mathbf{W} = \alpha \mathbf{I}$,

$$
\sigma_{\max}\left(\left(\mathbf{I} - \alpha \frac{\mathbf{A}^\top \mathbf{A}}{\|\mathbf{A}\|_F^2}\right)^2 + \frac{\alpha^2}{q}\left(\mathbf{I} - \frac{\mathbf{A}^\top \mathbf{A}}{\|\mathbf{A}\|_F^2}\right)\frac{\mathbf{A}^\top \mathbf{A}}{\|\mathbf{A}\|_F^2}\right) \tag{13}
$$

with respect to α. To do this, we first simplify from a matrix polynomial to a maximum over scalar polynomials in α with coefficients based on each singular value of $\mathbf{A}$. We then show that the maximum occurs when either the minimum or maximum singular value of $\mathbf{A}$ is used. Finally, we derive a condition for which singular value to use, and determine the optimal α that minimizes the maximum singular value.

Defining $\mathbf{Q}^\top \mathbf{\Sigma} \mathbf{Q} = \frac{\mathbf{A}^\top \mathbf{A}}{\|\mathbf{A}\|_F^2}$ as the eigendecomposition, and the polynomial

$$
p(\sigma) \overset{\text{def}}{=} 1 - 2\alpha\sigma + \alpha^2\left(\frac{\sigma}{q} + \left(1 - \frac{1}{q}\right)\sigma^2\right),
$$

the convergence rate constant from Eq. (13) can be written as $\sigma_{\max}\left(p\left(\frac{\mathbf{A}^\top \mathbf{A}}{\|\mathbf{A}\|_F^2}\right)\right)$. Since $p\left(\frac{\mathbf{A}^\top \mathbf{A}}{\|\mathbf{A}\|_F^2}\right)$ is a polynomial of a symmetric matrix, its singular vectors are the same as those of its argument, while its corresponding singular values are the polynomial p applied to the singular values of the original matrix. That is,

$$
p\left(\frac{\mathbf{A}^\top \mathbf{A}}{\|\mathbf{A}\|_F^2}\right) = p\left(\mathbf{Q}^\top \mathbf{\Sigma} \mathbf{Q}\right) = \mathbf{Q}^\top p\left(\mathbf{\Sigma}\right)\mathbf{Q}.
$$

Thus, the convergence rate constant can be written as

$$\sigma_{\max}\left(p\left(\frac{\mathbf{A}^\top\mathbf{A}}{\|\mathbf{A}\|_F^2}\right)\right) = \sigma_{\max}\left(p(\mathbf{\Sigma})\right).$$

Moreover, we can bound this extremal singular value by the maximum of the polynomial p over an interval containing the spectrum of $\mathbf{\Sigma}$

$$\sigma_{\max}\left(p\left(\mathbf{\Sigma}\right)\right) \leq \max\left|p\left(\sigma\right)\right| \quad \text{subject to} \quad \sigma \in [s_{\min}, s_{\max}].$$

Here, the singular values of $\mathbf{\Sigma}$ are bounded from below by $s_{\min} \overset{\text{def}}{=} \frac{\sigma_{\min}^2(\mathbf{A})}{\|\mathbf{A}\|_F^2}$ and above by $s_{\max} \overset{\text{def}}{=} \frac{\sigma_{\max}^2(\mathbf{A})}{\|\mathbf{A}\|_F^2}$ since $\mathbf{\Sigma}$ is the diagonal matrix of singular values of $\frac{\mathbf{A}^\top\mathbf{A}}{\|\mathbf{A}\|_F^2}$. Note that the polynomial can be factored as $p(\sigma) = (1-\sigma\alpha)^2 + \frac{\sigma\alpha^2}{q}(1-\sigma)$, and is positive for $\sigma \in [0, 1]$, which contains $[s_{\min}, s_{\max}]$. Also, since the coefficient of the σ^2 term of the polynomial p is $\alpha^2\left(1-\frac{1}{q}\right)$ which is greater than or equal to zero, the polynomial is convex in σ on the interval $[s_{\min}, s_{\max}]$. Thus, the maximum of p on the interval $[s_{\min}, s_{\max}]$ is attained at one of the two endpoints $s_{\min}, s_{\max}$ and we have the bound

$$\sigma_{\max}\left(p\left(\mathbf{\Sigma}\right)\right) = \max\left(p\left(s_{\min}\right), p\left(s_{\max}\right)\right).$$

To optimize this bound with respect to α, we first find conditions on α such that $p(s_{\min}) < p(s_{\max})$. If $s_{\max} = s_{\min}$, this obviously never holds; otherwise, $s_{\max} > s_{\min}$ and

$$p(s_{\min}) < p(s_{\max})$$
$$1 - 2\alpha s_{\min} + \alpha^2\left[\frac{s_{\min}}{q} + \left(1-\frac{1}{q}\right)s_{\min}^2\right] < 1 - 2\alpha s_{\max} + \alpha^2\left[\frac{s_{\max}}{q} + \left(1-\frac{1}{q}\right)s_{\max}^2\right]$$

Grouping like terms and cancelling, we get

$$\alpha\left(2-\frac{\alpha}{q}\right)(s_{\max}-s_{\min}) < \alpha^2\left(1-\frac{1}{q}\right)\left(s_{\max}^2 - s_{\min}^2\right)$$

Since $\frac{\alpha}{q} > 0$, we can divide it from both sides.

$$(2q-\alpha)(s_{\max}-s_{\min}) < \alpha(q-1)\left(s_{\max}^2 - s_{\min}^2\right)$$

Since $s_{\max} > s_{\min}$, we can divide both sides by $s_{\max} - s_{\min}$.

$$\begin{aligned} 2q - \alpha &< \alpha\,(q-1)\,(s_{\max}+s_{\min}) \\ 2q &< \alpha\,(1 + (q-1)\,(s_{\max}+s_{\min})) \end{aligned}$$

and since the number of threads $q \geq 1$, we can divide both sides by $1 + (q-1)\left(s_{\min} + s_{\max}\right)$ to get

$$\alpha > \frac{2q}{1 + (q-1)\left(s_{\min} + s_{\max}\right)} \overset{\text{def}}{=} \widehat{\alpha}.$$

Thus,

$$\sigma_{\max}\left(p\left(\mathbf{\Sigma}\right)\right) = \begin{cases} p\left(s_{\max}\right), & \alpha > \widehat{\alpha} \\ p\left(s_{\min}\right), & \alpha \leq \widehat{\alpha} \end{cases}$$

For the first term,

$$\begin{aligned} \frac{\partial}{\partial \alpha} p(s_{\max}) &= -2s_{\max} + 2\left(\frac{s_{\max}}{q} + \left(1 - \frac{1}{q}\right)s_{\max}^2\right)\alpha \\ &> -2s_{\max} + 2\left(\frac{s_{\max}}{q} + \left(1 - \frac{1}{q}\right)s_{\max}^2\right)\widehat{\alpha} \end{aligned}$$

since $\alpha > \widehat{\alpha}$ and the coefficient is positive. Factoring $\frac{2s_{\max}}{q}$ from the second term and substituting for $\widehat{\alpha}$, we get

$$\begin{aligned} &= -2s_{\max} + \frac{2s_{\max}}{q}\left(1 + (q-1)\,s_{\max}\right)\widehat{\alpha} \\ &= -2s_{\max} + \frac{2s_{\max}}{q}\left(1 + (q-1)\,s_{\max}\right)\frac{2q}{1 + (1-q)\left(s_{\min} + s_{\max}\right)} \\ &= -2s_{\max} + 2s_{\max}\frac{2\left(1 + (q-1)s_{\max}\right)}{1 + (q-1)(s_{\max} + s_{\min})} \\ &= 2s_{\max}\left[-1 + \frac{2\left(1 + (q-1)s_{\max}\right)}{1 + (q-1)(s_{\max} + s_{\min})}\right] \\ &= 2s_{\max}\left[\frac{1 + (q-1)(s_{\max} - s_{\min})}{1 + (q-1)(s_{\max} + s_{\min})}\right] \\ &> 0 \end{aligned}$$

since all terms in both numerator and denominator are positive. Thus, the function is monotonic increasing on $\alpha \in [\widehat{\alpha}, \infty)$, and the minimum is at the lower endpoint, i.e. $\alpha^{\star} = \widehat{\alpha}$.

Similarly, for the second term, $\alpha \leq \widehat{\alpha}$ and

$$\begin{aligned} \frac{\partial}{\partial \alpha} p(s_{\min}) &= -2s_{\min} + 2\left(\frac{s_{\min}}{q} + \left(1 - \frac{1}{q}\right)s_{\min}^2\right)\alpha \\ &\leq -2s_{\min} + 2\left(\frac{s_{\min}}{q} + \left(1 - \frac{1}{q}\right)s_{\min}^2\right)\widehat{\alpha} \\ &= 2s_{\min}\left[\frac{1 - (q-1)(s_{\max} - s_{\min})}{1 + (q-1)(s_{\max} + s_{\min})}\right] \end{aligned}$$

If

$$1-(q-1)(s_{\max}-s_{\min})<0, \tag{14}$$

this function is monotonic decreasing on $\alpha \in (-\infty, \alpha^\star]$, and the minimum is at the upper endpoint i.e. $\alpha = \alpha^\star$. Otherwise, since $p(s_{\min})$ is quadratic in α with positive leading coefficient, the minimum occurs at the critical point, so we set the derivative to 0 and solve for $\alpha^\star$

$$\begin{aligned}\frac{\partial}{\partial\alpha}p(s_{\min}) &= -2s_{\min}+2\left(\frac{s_{\min}}{q}+\left(1-\frac{1}{q}\right)s_{\min}^2\right)\alpha^\star\\ &= -2s_{\min}+\frac{2s_{\min}}{q}\left(1+(q-1)\,s_{\min}\right)\alpha^\star\\ &=0\\ \frac{2s_{\min}}{q}\left(1+(q-1)\,s_{\min}\right)\alpha^\star &= 2s_{\min}\\ \alpha^\star &= \frac{q}{1+(q-1)s_{\min}}\end{aligned}$$

D Corollary Proofs

We provide proofs for the corollaries of Sect. 2, which follow from Theorem 1.

D.1 Proof of Corollary 2

Suppose $p_i = \frac{\|\mathbf{A}_i\|^2}{\|\mathbf{A}\|_F^2}$ and $\mathbf{W} = \alpha\mathbf{I}$. From the proof of Theorem 1,

$$\mathbb{E}_k\left[\|e^{k+1}\|^2\right]=\left\langle e^k,\left(\left(\mathbf{I}-\alpha\frac{\mathbf{A}^\top\mathbf{A}}{\|\mathbf{A}\|_F^2}\right)^2-\frac{\alpha^2}{q}\left(\frac{\mathbf{A}^\top\mathbf{A}}{\|\mathbf{A}\|_F^2}\right)^2\right)e^k\right\rangle+\frac{\alpha}{q}\frac{\|r^k\|_{\mathbf{W}}^2}{\|\mathbf{A}\|_F^2}.$$

In this case, since $\mathbf{A}^\top r^\star = 0$, $\langle \mathbf{A}e^k, r^\star\rangle = 0$ and

$$\begin{aligned}\|r^k\|_{\mathbf{W}}^2 &= \alpha\|\mathbf{A}e^k\|^2+2\alpha\langle\mathbf{A}e^k, r^\star\rangle+\alpha\|r^\star\|^2\\ &=\alpha\langle e^k, \mathbf{A}^\top\mathbf{A}e^k\rangle+\alpha\|r^\star\|^2.\end{aligned}$$

Combining the inner products,

$$\mathbb{E}_k\left[\|e^{k+1}\|^2\right]=\left\langle e^k,\left(\left(\mathbf{I}-\alpha\frac{\mathbf{A}^\top\mathbf{A}}{\|\mathbf{A}\|_F^2}\right)^2+\frac{\alpha^2}{q}\left(\mathbf{I}-\frac{\mathbf{A}^\top\mathbf{A}}{\|\mathbf{A}\|_F^2}\right)\frac{\mathbf{A}^\top\mathbf{A}}{\|\mathbf{A}\|_F^2}\right)e^k\right\rangle+\frac{\alpha^2\|r^\star\|^2}{q\|\mathbf{A}\|_F^2}$$

$$\leq \sigma_{\max}\left(\left(\mathbf{I}-\alpha\frac{\mathbf{A}^\top\mathbf{A}}{\|\mathbf{A}\|_F^2}\right)^2+\frac{\alpha^2}{q}\left(\mathbf{I}-\frac{\mathbf{A}^\top\mathbf{A}}{\|\mathbf{A}\|_F^2}\right)\frac{\mathbf{A}^\top\mathbf{A}}{\|\mathbf{A}\|_F^2}\right)\|e^k\|^2+\frac{\alpha^2\|r^\star\|^2}{q\|\mathbf{A}\|_F^2}.$$

D.2 Proof of Corollary 3

Suppose $q=1$, $\mathbf{W}=\mathbf{I}$ and $p_i=\frac{\|\mathbf{A}_i\|^2}{\|\mathbf{A}\|_F^2}$.

$$\begin{aligned}\mathbb{E}_k\left[\|e^{k+1}\|^2\right]&\leq \sigma_{\max}\left(\mathbf{I}-\frac{\mathbf{A}^\top\mathbf{A}}{\|\mathbf{A}\|_F^2}\right)\|e^k\|^2+\frac{\|r^\star\|^2}{\|\mathbf{A}\|_F^2}\\&=\left(1-\frac{\sigma_{\min}^2(\mathbf{A})}{\|\mathbf{A}\|_F^2}\right)\|e^k\|^2+\frac{\|r^\star\|^2}{\|\mathbf{A}\|_F^2}.\end{aligned}$$

From the proof of Theorem 1,

$$\mathbb{E}_k\left[\|e^{k+1}\|^2\right]=\left\langle e^k,\left(\left(\mathbf{I}-\frac{\mathbf{A}^\top\mathbf{A}}{\|\mathbf{A}\|_F^2}\right)^2-\left(\frac{\mathbf{A}^\top\mathbf{A}}{\|\mathbf{A}\|_F^2}\right)^2\right)e^k\right\rangle+\frac{\|r^k\|^2}{\|\mathbf{A}\|_F^2}.$$

Decomposing r^k,

$$\begin{aligned}\|r^k\|^2&=\|\mathbf{A}e^k\|^2+\|r^\star\|^2\\&=\langle e^k,\mathbf{A}^\top\mathbf{A}e^k\rangle+\|r^\star\|^2.\end{aligned}$$

Combining the inner products,

$$\begin{aligned}\mathbb{E}_k\left[\|e^{k+1}\|^2\right]&=\left\langle e^k,\left(\left(\mathbf{I}-\frac{\mathbf{A}^\top\mathbf{A}}{\|\mathbf{A}\|_F^2}\right)^2-\left(\frac{\mathbf{A}^\top\mathbf{A}}{\|\mathbf{A}\|_F^2}\right)^2+\frac{\mathbf{A}^\top\mathbf{A}}{\|\mathbf{A}\|_F^2}\right)e^k\right\rangle+\frac{\|r^\star\|^2}{\|\mathbf{A}\|_F^2}\\&=\left\langle e^k,\left(\mathbf{I}-\frac{\mathbf{A}^\top\mathbf{A}}{\|\mathbf{A}\|_F^2}\right)e^k\right\rangle+\frac{\|r^\star\|^2}{\|\mathbf{A}\|_F^2}\\&\leq\sigma_{\max}\left(\mathbf{I}-\frac{\mathbf{A}^\top\mathbf{A}}{\|\mathbf{A}\|_F^2}\right)\|e^k\|^2+\frac{\|r^\star\|^2}{\|\mathbf{A}\|_F^2}\\&=\left(1-\frac{\sigma_{\min}^2(\mathbf{A})}{\|\mathbf{A}\|_F^2}\right)\|e^k\|^2+\frac{\|r^\star\|^2}{\|\mathbf{A}\|_F^2}.\end{aligned}$$

References

1. Aharoni, R., Censor, Y.: Block-iterative projection methods for parallel computation of solutions to convex feasibility problems. Linear Algebra Appl. **120**, 165–175 (1989)

2. Bottou, L.: Online algorithms and stochastic approximations. In: Online Learning and Neural Networks. Cambridge University Press, Cambridge (1998)
3. Charles, L.: Applied Iterative Methods. Ak Peters/CRC Press, Boca Raton (2007)
4. Cai, Y., Zhao, Y., Tang, Y.: Exponential convergence of a randomized Kaczmarz algorithm with relaxation. In: Gaol, F.L., Nguyen, Q.V. (eds.) Proceedings of the 2011 2nd International Congress on Computer Applications and Computational Science, pp. 467–473, Springer, Berlin (2012)
5. Chen, X., Powell, A.M.: Almost sure convergence of the Kaczmarz algorithm with random measurements. J. Fourier Anal. Appl. **18**(6), 1195–1214 (2012)
6. Eggermont, P.P.B., Herman, G.T., Lent, A.: Iterative algorithms for large partitioned linear systems, with applications to image reconstruction. Linear Algebra Appl. **40**, 37–67 (1981)
7. Eldar, Y.C., Needell, D.: Acceleration of randomized Kaczmarz method via the Johnson–Lindenstrauss lemma. Numer. Algorithms **58**(2), 163–177 (2011)
8. Elfving, T.: Block-iterative methods for consistent and inconsistent linear equations. Numer. Math. **35**(1), 1–12 (1980)
9. Gordon, D., Gordon, R.: Component-averaged row projections: a robust, block-parallel scheme for sparse linear systems. SIAM J. Sci. Comput. **27**(3), 1092–1117 (2005)
10. Gordon, R., Bender, R., Herman, G.T.: Algebraic reconstruction techniques (art) for three-dimensional electron microscopy and X-ray photography. J. Theor. Biol. **29**(3), 471–481 (1970)
11. Gower, R.M., Richtárik, P.: Randomized iterative methods for linear systems. SIAM J. Matrix Anal. Appl. **36**(4), 1660–1690 (2015)
12. Hamaker, C., Solmon, D.C.: The angles between the null spaces of X rays. J. Math. Anal. Appl. **62**(1), 1–23 (1978)
13. Hanke, M., Niethammer, W.: On the acceleration of Kaczmarz's method for inconsistent linear systems. Linear Algebra Appl. **130**, 83–98 (1990)
14. Hanke, M., Niethammer, W.: On the use of small relaxation parameters in Kaczmarz method. Z. Angew. Math. Mech. **70**(6), T575–T576 (1990)
15. Herman, G.T., Meyer, L.B.: Algebraic reconstruction techniques can be made computationally efficient (positron emission tomography application). IEEE Trans. Med. Imaging **12**(3), 600–609 (1993)
16. Kaczmarz, S.M.: Angenäherte auflösung von systemen linearer gleichungen. Bull. Int. Acad. Pol. Sci. Lett. Classe Sci. Math. Nat. Sér. A Sci. Math. **35**, 355–357 (1937)
17. Leventhal, D., Lewis, A.S.: Randomized methods for linear constraints: convergence rates and conditioning. Math. Oper. Res. **35**(3), 641–654 (2010)
18. Liu, J., Wright, S.J., Srikrishna, S.: An asynchronous parallel randomized Kaczmarz algorithm. arXiv:1401.4780 (2014)
19. Ma, A., Needell, D., Ramdas, A.: Convergence properties of the randomized extended Gauss–Seidel and Kaczmarz methods. SIAM J. Matrix Anal. A **36**(4), 1590–1604 (2015)
20. Necoara, I.: Faster randomized block Kaczmarz algorithms. SIAM J. Matrix Anal. Appl. **40**(4), 1425–1452 (2019)
21. Needell, D.: Randomized Kaczmarz solver for noisy linear systems. BIT Numer. Math. **50**(2), 395–403 (2010)
22. Needell, D., Srebro, N., Ward, R.: Stochastic gradient descent, weighted sampling, and the randomized Kaczmarz algorithm. Math. Program. **155**(1), 549–573 (2015)
23. Needell, D., Tropp, J.A.: Paved with good intentions: analysis of a randomized block Kaczmarz method. Linear Algebra Appl. **441**(August), 199–221 (2012)
24. Needell, D., Ward, R.: Two-subspace projection method for coherent overdetermined systems. J. Fourier Anal. Appl. **19**(2), 256–269 (2013)
25. Needell, D., Ward, R.: Batched stochastic gradient descent with weighted sampling. Approx. Theory XV San Antonio **2016**, 279–306 (2017)
26. Niu, F., Recht, B., Ré, C., Wright, S.J.: HOGWILD!: A lock-free approach to parallelizing stochastic gradient descent. In: Neural Information Processing Systems (2011)
27. Richtárik, P., Takáč, M.: Stochastic reformulations of linear systems: algorithms and convergence theory. SIAM J. Matrix Anal. Appl. **41**(2), 487–524 (2020)
28. Robbins, H., Monro, S.: A stochastic approximation method. Ann. Math. Stat. **22**(3), 400–407 (1951)
29. Strohmer, T., Vershynin, R.: A randomized Kaczmarz algorithm with exponential convergence. J. Fourier Anal. Appl. **15**(2), 262–278 (2009)
30. Jinchao, X., Zikatanov, L.: The method of alternating projections and the method of subspace corrections in hilbert space. J. Am. Math. Soc. **15**(3), 573–597 (2002)

31. Yangyang, X., Yin, W.: Block stochastic gradient iteration for convex and nonconvex optimization. SIAM J. Optim. **25**(3), 1686–1716 (2015)
32. Zouzias, A., Freris, N.M.: Randomized extended Kaczmarz for solving least squares. SIAM J. Matrix Anal. Appl. **34**(2), 773–793 (2013)