3-17-2014

Randomized Block Kaczmarz Method with Projection for Solving Least Squares

Deanna Needell
Claremont McKenna College

Ran Zhao

Anastasios Zouzias

Recommended Citation

Needell, D., Zhao, R., Zouzias, A., "Randomized Block Kaczmarz Method with Projection for Solving Least Squares", Submitted, arXiv preprint arXiv:1403.4192, 2014.

RANDOMIZED BLOCK KACZMARZ METHOD WITH PROJECTION FOR SOLVING LEAST SQUARES

DEANNA NEEDELL, RAN ZHAO AND ANASTASIOS ZOUZIAS

ABSTRACT. The Kaczmarz method is an iterative method for solving overcomplete linear systems of equations $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$. The randomized version of the Kaczmarz method put forth by Strohmer and Vershynin iteratively projects onto a randomly chosen solution space given by a single row of the matrix $\boldsymbol{A}$ and converges exponentially in expectation to the solution of a *consistent* system. In this paper we analyze two block versions of the method each with a randomized projection, that converge in expectation to the least squares solution of inconsistent systems. Our approach utilizes a paving of the matrix $\boldsymbol{A}$ to guarantee exponential convergence, and suggests that paving yields a significant improvement in performance in certain regimes. The proposed method is an extension of the block Kaczmarz method analyzed by Needell and Tropp and the Randomized Extended Kaczmarz method of Zouzias and Freris. The contribution is thus two-fold; unlike the standard Kaczmarz method, our methods converge to the least-squares solution of inconsistent systems, and by using appropriate blocks of the matrix this convergence can be significantly accelerated. Numerical experiments suggest that the proposed algorithm can indeed lead to advantages in practice.

1. INTRODUCTION

The Kaczmarz method [Kac37] is a popular iterative solver of overdetermined systems of linear equations $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$. Because of its simplicity and performance, the method and its derivatives are used in a range of applications from image reconstruction to digital signal processing [CFM+92, Nat01, SS87]. The method performs a series of orthogonal projections and iteratively converges to the solution of the system of equations. It is therefore computationally feasible even for very large and overdetermined systems.

Given a vector $\boldsymbol{b}$ and an $n \times d$ full rank matrix $\boldsymbol{A}$ with rows $\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots \boldsymbol{a}_n$, the algorithm begins with an initial estimate $\boldsymbol{x}_0$ and cyclically projects the estimation onto each of the solution spaces. This process can be described as follows:

$$\boldsymbol{x}_j = \boldsymbol{x}_{j-1} + \frac{\boldsymbol{b}[i] - \langle \boldsymbol{a}_i, \boldsymbol{x}_{j-1} \rangle}{\|\boldsymbol{a}_i\|_2^2} \boldsymbol{a}_i,$$

where $\boldsymbol{b}[i]$ denotes the ith coordinate of $\boldsymbol{b}$, $\boldsymbol{x}_j$ is the estimation in the jth iteration, $\|\cdot\|_2$ denotes the usual ℓ_2 vector norm, and $i = (j \mod n) + 1$ cycles through the rows of $\boldsymbol{A}$.

Since the method cycles through the rows of $\boldsymbol{A}$, the performance of the algorithm may depend heavily on the ordering of these rows. A poor ordering may lead to very slow convergence. To overcome this obstacle, one can select the row $\boldsymbol{a}_i$ at random to improve the convergence rate [HM93, Nat01]. Strohmer and Vershynin proposed and analyzed a method which selects a given row with probability proportional to its Euclidean norm [SV09, SV06]. They show that with this selection strategy, the randomized Kaczmarz method has an expected exponential convergence rate to the unique solution $\boldsymbol{x}_\star$:

D.N. is thankful to the Simons Foundation Collaboration grant. A.Z. has received funding from the European Research Council under the European Union's Seventh Framework Program (FP7/2007-2013) / ERC grant agreement n^o 259569.

$$\mathbb{E}\|\boldsymbol{x}_j - \boldsymbol{x}_\star\|_2^2 \le \left(1 - \frac{1}{R}\right)^j \|\boldsymbol{x}_0 - \boldsymbol{x}_\star\|_2^2, \tag{1}$$

where R is the scaled condition number, $R = \|\boldsymbol{A}^{-1}\|^2 \|\boldsymbol{A}\|_F^2$, $\|\cdot\|_F$ denotes the Frobenius norm, and $\|\boldsymbol{A}^{-1}\| \overset{\text{def}}{=} \inf\{M : M\|\boldsymbol{A}\boldsymbol{x}\|_2 \ge \|\boldsymbol{x}\|_2 \text{ for all } x\}$ is well-defined since $\boldsymbol{A}$ has full column rank. This convergence rate (1) is essentially independent of the number of rows of $\boldsymbol{A}$ and shows that for well-conditioned matrices, the randomized Kaczmarz method converges to the solution in just $O(p)$ iterations [SV09]. This yields an overall runtime of $O(d^2)$ which is much superior to others such as $O(nd^2)$ for Gaussian elimination. There are also cases where randomized Kaczmarz even outperforms the conjugate gradient method, see the discussion in [SV09] for details.

When the system is perturbed by noise or no longer consistent, $\boldsymbol{A}\boldsymbol{x}_\star + \boldsymbol{e} = \boldsymbol{b}$, the randomized Kaczmarz method still provides expected exponential convergence down to an error threshold [Nee10]. When the rows of $\boldsymbol{A}$ have unit norm, this result yields the following convergence bound:

$$\mathbb{E}\|\boldsymbol{x}_j - \boldsymbol{x}_\star\|_2 \le \left(1 - \frac{1}{R}\right)^j \|\boldsymbol{x}_0 - \boldsymbol{x}_\star\|_2 + \sqrt{R}\|\boldsymbol{e}\|_\infty. \tag{2}$$

This result is sharp, and shows that the randomized Kaczmarz method converges with a radius proportional the magnitude of the largest entry of the noise in the system. Since the iterates of the Kaczmarz method always lie in a single solution space, the method clearly will not converge to the least-squares solution of an inconsistent system.

1.1. **Randomized Extended Kaczmarz.** The bound (2) demonstrates that the randomized Kaczmarz method performs well when the noise in inconsistent systems is small. Zouzias and Freris introduced a variant of the method which utilizes a random projection to iteratively reduce the norm of the error [ZF12]. They show that the estimate of this *Randomized Extended Kaczmarz* (REK) method converges exponentially in expectation to the least squares solution of the system, breaking the radius barrier of the standard method. The algorithm maintains not only an estimate $\boldsymbol{x}_j$ to the solution but also an approximation $\boldsymbol{z}_j$ to the projection of $\boldsymbol{b}$ onto the range of $\boldsymbol{A}$:

$$\boldsymbol{x}_j = \boldsymbol{x}_{j-1} + \frac{\boldsymbol{b}[i] - \boldsymbol{z}_{j-1}[i] - \langle \boldsymbol{a}_i, \boldsymbol{x}_{j-1}\rangle}{\|\boldsymbol{a}_i\|_2^2}\boldsymbol{a}_i, \quad \boldsymbol{z}_j = \boldsymbol{z}_{j-1} - \frac{\langle \boldsymbol{a}^{(k)}, \boldsymbol{z}_{j-1}\rangle}{\|\boldsymbol{a}^{(k)}\|_2^2}\boldsymbol{a}^{(k)}, \tag{3}$$

where in iteration j, $\boldsymbol{a}_i$ and $\boldsymbol{a}^{(k)}$ is the row and column of $\boldsymbol{A}$, respectively, each chosen randomly with probability proportional to their Euclidean norms. In this setting, we no longer require that the matrix $\boldsymbol{A}$ be full rank, and ask for the least squares solution,

$$\boldsymbol{x}_{LS} \overset{\text{def}}{=} \underset{\boldsymbol{x}}{\operatorname{argmin}} \|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}\|_2 = \boldsymbol{A}^\dagger \boldsymbol{b},$$

where $\boldsymbol{A}^\dagger$ denotes the pseudo-inverse of $\boldsymbol{A}$. Zouzias and Freris showed that the REK method converges exponentially in expectation to the least squares solution [ZF12],

$$\mathbb{E}\|\boldsymbol{x}_j - \boldsymbol{x}_{LS}\|_2^2 \le \left(1 - \frac{1}{K^2(\boldsymbol{A})}\right)^{j/2} \left(\|\boldsymbol{x}_{LS}\|_2^2 + \frac{2\|\boldsymbol{b}\|_2^2}{\sigma_{\min}(\boldsymbol{A})}\right), \tag{4}$$

where $\sigma_{\min}(\boldsymbol{A})$ is the smallest non-zero singular value of $\boldsymbol{A}$ and $K(\boldsymbol{A}) = \frac{\|\boldsymbol{A}\|_F}{\sigma_{\min}(\boldsymbol{A})}$ denotes its scaled condition number.

1.2. **The block Kaczmarz method.** Recently, Needell and Tropp analyzed a block version of the simple randomized Kaczmarz method [NT13]. For simplicity and to avoid degenerate cases, we assume here that the rows of the matrix $\boldsymbol{A}$ all have unit ℓ_2 norm. We will refer to such matrices as *row-standardized* (and the transpose $\boldsymbol{A}^*$ as *column-standardized*). Like the traditional method, this version iteratively projects the current estimation onto the solution spaces. However, rather than using the solution space of a *single* equation, the block method projects onto the solution space of many equations simultaneously by selecting a block of rows rather than a single row. For a subset $\tau \subset \{1, 2, \ldots, n\}$, denote by $\boldsymbol{A}_\tau$ the submatrix of $\boldsymbol{A}$ whose rows are indexed by τ. We again begin with an arbitrary guess $\boldsymbol{x}_0$ for the solution of the system. Then for each iteration $j \geq 1$, select a block $\tau = \tau_j$ of rows. To obtain the next iterate, we project the current estimation onto the solution space of the equations listed in τ [NT13]:

$$\boldsymbol{x}_j = \boldsymbol{x}_{j-1} + (\boldsymbol{A}_{\boldsymbol{\tau}})^\dagger (\boldsymbol{b}_\tau - \boldsymbol{A}_\tau \boldsymbol{x}_{j-1}). \tag{5}$$

Here, the conditioning of the blocks $\boldsymbol{A}_{\boldsymbol{\tau}}$ plays a crucial row in the behavior of the method. Indeed, if each block is well-conditioned, its pseudoinverse can be applied efficiently using an iterative method such as CGLS [Bjö96]. To guarantee such properties, Needell and Tropp utilize a *paving* of the matrix $\boldsymbol{A}^*$.

Definition 1 (Column Paving)**.** *A* (p, α, β) column paving *of a* $d \times n$ *column-standardized matrix* $\boldsymbol{A}$ *is a partition* $T = \{\tau_1, \ldots, \tau_p\}$ *of the columns such that*

$$\alpha \leq \lambda_{\min}(\boldsymbol{A}_\tau^* \boldsymbol{A}_\tau) \quad \textit{and} \quad \lambda_{\max}(\boldsymbol{A}_\tau^* \boldsymbol{A}_\tau) \leq \beta \quad \textit{for each } \tau \in T,$$

where again we denote by $\boldsymbol{A}_\tau$ *the* $d \times |\tau|$ *submatrix of* $\boldsymbol{A}$*. We refer to the number* p *of blocks as the* size *of the paving, and the numbers* α *and* β *are called the* lower *and* upper paving bounds.

We refer to a column paving of $\boldsymbol{A}^*$ as a *row paving* of $\boldsymbol{A}$. We thus seek pavings of $\boldsymbol{A}$ with small number of blocks p and upper paving constant β. A surprising result shows that every column-standardized matrix admits such a column paving. Tropp proves the following result in [Tro09, Thm. 1.2], whose origins are due to Bourgain and Tzafriri [BT87, BT91] and Vershynin [Ver06].

Proposition 1 (Existence of Good Pavings)**.** *Fix a number* $\delta \in (0, 1)$ *and column-standardized matrix* $\boldsymbol{A}$ *with n columns. Then* $\boldsymbol{A}$ *admits a* (p, α, β) *column paving with*

$$p \leq C \cdot \delta^{-2} \|\boldsymbol{A}\|^2 \log(1 + n) \quad \textit{and} \quad 1 - \delta \leq \alpha \leq \beta \leq 1 + \delta,$$

where C denotes an absolute constant.

Proposition 1 shows the *existence* of paving, but the literature provides various efficient mechanisms for the construction of good pavings as well. In many cases one constructs such a paving simply by choosing a partition of an appropriate size *at random*. See [NT13] and the references therein for a thorough discussion of these types of results. Equipped with such a column paving of $\boldsymbol{A}^*$, the main result of [NT13] shows that the randomized block Kaczmarz algorithm (5) exhibits exponential convergence in expectation:

$$\mathbb{E}\|\boldsymbol{x}_j - \boldsymbol{x}_\star\|_2^2 \leq \left(1 - \frac{1}{C' \kappa^2(\boldsymbol{A}) \log n}\right)^j \|\boldsymbol{x}_0 - \boldsymbol{x}_\star\|_2^2 + \frac{3\|\boldsymbol{e}\|_2^2}{\sigma_{\min}^2(\boldsymbol{A})}, \tag{6}$$

where C' is an absolute constant and $\kappa(\boldsymbol{A}) = \frac{\sigma_{\max}(\boldsymbol{A})}{\sigma_{\min}(\boldsymbol{A})}$ denotes the condition number of $\boldsymbol{A}$.

Since each iteration of the block method utilizes multiple rows, one can compare the rate of (6) and (1) by considering convergence per epoch (one cycle through the rows of $\boldsymbol{A}$). From this analysis, one finds the bounds to be comparable. However, the block method can utilize fast matrix multiplies and efficient implementation, yielding dramatic improvements in computational time. See [NT13] for details and empirical results.

1.3. **Contribution.** The REK method breaks the so-called *convergence horizon* of standard Kaczmarz method, allowing convergence to the least-squares solution of inconsistent systems. The block Kaczmarz method on the other hand, allows for significant computational speedup and accelerated convergence to within a fixed radius of the least-squares solution. The main contribution of this paper analyzes a randomized block Kaczmarz method which also incorporates a blocked projection step, which provides accelerated convergence to the least-squares solution. In this case we need a column partition for the projection step and a row partition for the Kaczmarz step. We assume here that the matrix $\boldsymbol{A}$ is row-standardized, and will also utilize the column-standardized version of $\boldsymbol{A}$, denoted $\overline{\boldsymbol{A}}$, obtained by normalizing the columns of $\boldsymbol{A}$. We thus propose the following randomized block extended Kaczmarz method using double partitioning.

Algorithm 1 Randomized Double Block Kaczmarz Least Squares Solver

```
procedure $(\boldsymbol{A}, \boldsymbol{b}, T, \mathscr{T})$ ▷ $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, column-standardized $\overline{\boldsymbol{A}}$, $\boldsymbol{b} \in \mathbb{R}^m$, $T \in \mathbb{N}$, column partition $\mathscr{T}$ of $[n]$, row partition $\mathscr{S}$ of $[m]$
    Initialize $\mathbf{x}_0 = \mathbf{0}$ and $\mathbf{z}_0 = \boldsymbol{b}$
    Create column-standardized $\overline{\boldsymbol{A}}$
    for $k = 1, 2, \ldots, T$ do
        Pick $\tau_k \in \mathscr{T}$ and $\sigma_k \in \mathscr{S}$ uniformly at random
        Set $\mathbf{z}_k = \mathbf{z}_{k-1} - \overline{\boldsymbol{A}}_{\tau_k}(\overline{\boldsymbol{A}}_{\tau_k})^\dagger \mathbf{z}_{k-1}$    ▷ $\overline{\boldsymbol{A}}_{\tau_k}$: the $m \times |\tau_k|$ submatrix of $\overline{\boldsymbol{A}}$
        Update $\boldsymbol{x}_k = \boldsymbol{x}_{k-1} + (\boldsymbol{A}_{\sigma_k})^\dagger(\boldsymbol{b}_{\sigma_k} - (\boldsymbol{z}_k)_{\sigma_k} - \boldsymbol{A}_{\sigma_k}\boldsymbol{x}_{k-1})$    ▷ $\boldsymbol{A}_{\sigma_k}$: the $|\sigma_k| \times n$ submatrix of $\boldsymbol{A}$
    end for
    Output $\mathbf{x}_T$
end procedure
```

Our main result shows that this method offers both exponential convergence to the least-squares solution $\mathbf{x}_{\mathrm{LS}}$, and improved convergence speed due to the blocking of both the rows and columns.

Theorem 2. *Suppose Algorithm 1 is run with input* $\boldsymbol{A}$, $\boldsymbol{b}$, $T \in \mathbb{N}$, $(\overline{p}, \overline{\alpha}, \overline{\beta})$ *column paving* $\mathscr{T}$ *of* $\overline{\boldsymbol{A}}$, *and* (p, α, β) *column paving* $\mathscr{S}$ *of* $\boldsymbol{A}^*$, *both guaranteed by Proposition 1. Then the estimate vector* $\mathbf{x}_T$ *satisfies*

$$\mathbb{E}\,\|\boldsymbol{x}_T - \mathbf{x}_{LS}\|_2^2 \le \gamma^T \|\boldsymbol{x}_0 - \mathbf{x}_{LS}\|_2^2 + \left(\gamma^{\lfloor T/2 \rfloor} + \overline{\gamma}^{\lfloor T/2 \rfloor}\right) \frac{C \left\|\boldsymbol{b}_{\mathscr{R}(\boldsymbol{A})}\right\|_2^2}{(1-\gamma)},$$

where $\gamma = 1 - \frac{C}{\kappa^2(\boldsymbol{A})\log(1+n)}$, $\overline{\gamma} = 1 - \frac{C}{\kappa^2(\overline{\boldsymbol{A}}\log(1+d)}$*and* $\kappa(\boldsymbol{A}) = \frac{\sigma_{\max}(\boldsymbol{A})}{\sigma_{\min}(\boldsymbol{A})}$ *denotes the condition number of* $\boldsymbol{A}$.

1.4. **Organization.** In addition to Algorithm 1, we propose an additional variant of the block and extended Kaczmarz methods that serves both as motivation for the analysis of the main result, as well as a useful method in its own right. In Section 2 we first introduce a block coordinate descent method with exponential convergence to the least squares solution, and whose analysis illustrates the convergence of the projection $\boldsymbol{z}_k$ of Algorithm 1. We prove our main result, Theorem 2, in Section 3. In Section 4 we discuss implementation details and present experimental results for the various algorithms. We conclude with a discussion of related work and open directions in Section 5. The appendix includes proofs of intermediate results used along the way.

2. A Randomized Block Coordinate Descent Method

Utilizing the benefits of both the block variant and the randomized extension, we propose the following randomized block coordinate descent method for the inconsistent case. We assume here that the matrix $\boldsymbol{A}$ is *column-standardized*; its columns each have unit norm.

Algorithm 2 Randomized Block Least Squares Solver

```
procedure $(\boldsymbol{A}, \boldsymbol{b}, T, \mathscr{T})$        ▷ $\boldsymbol{A} \in \mathbb{R}^{m\times n}, \boldsymbol{b} \in \mathbb{R}^{m}$, $T \in \mathbb{N}$, column partition $\mathscr{T}$ of $[n]$
    Initialize $\mathbf{x}_0 = \mathbf{0}$ and $\mathbf{z}_0 = \boldsymbol{b}$
    for $k = 1, 2, \ldots, T$ do
        Pick $\tau_k \in \mathscr{T}$ uniformly at random
        Compute $\boldsymbol{a}_k = (\boldsymbol{A}_{\tau_k})^{\dagger}\mathbf{z}_{k-1}$        ▷ $\boldsymbol{A}_{\tau}$: the $m \times |\tau_k|$ submatrix of $\boldsymbol{A}$
        Update $(\mathbf{x}_k)_{\tau_k} = (\mathbf{x}_{k-1})_{\tau_k} + \boldsymbol{a}_k$
        Set $\mathbf{z}_k = \mathbf{z}_{k-1} - \boldsymbol{A}_{\tau_k}\boldsymbol{a}_k$
    end for
    Output $\mathbf{x}_T$
end procedure
```

2.1. **Analysis of Randomized Block Least Squares Solver.** In Algorithm 2 we see that the partition $\mathscr{T}$ plays an important role, both in terms of convergence and computation. To apply the pseudo-inverse of the block $\boldsymbol{A}_{\tau}$ efficiently, we would like to ensure each block is well-conditioned. To this end, we consider column pavings as in Definition 1, which guarantee the singular values of each block are controlled. The following lemma is motivated by Lemma 2.2 of [NT13], and shows that the iterates $\mathbf{z}_k$ converge exponentially to the projection of $\boldsymbol{b}$ onto the kernel of $\boldsymbol{A}^{*}$.

Lemma 3. *Let $\boldsymbol{b}$ be a fixed vector and $\mathscr{T}$ be a (p, α, β) column paving of $\boldsymbol{A}$. Assuming the notation of Algorithm 2, for every $k > 0$ it holds that*

$$\mathbb{E}\left\|\mathbf{z}_k - \boldsymbol{b}_{\mathscr{R}(\boldsymbol{A})^{\perp}}\right\|_2^2 \leq \left(1 - \frac{\sigma_{\min}^2(\boldsymbol{A})}{p\beta}\right)^{k} \left\|\boldsymbol{b}_{\mathscr{R}(\boldsymbol{A})}\right\|_2^2. \tag{7}$$

where $\boldsymbol{b}_{\mathscr{R}(\boldsymbol{A})^{\perp}} := (\mathbf{I} - \boldsymbol{A}\boldsymbol{A}^{\dagger})\boldsymbol{b}$.

Proof. Let $\boldsymbol{P}_{\tau_k} = \boldsymbol{A}_{\tau_k}(\boldsymbol{A}_{\tau_k})^{\dagger}$ and notice $\mathbf{z}_k = (\boldsymbol{I} - \boldsymbol{P}_{\tau_k})\mathbf{z}_{k-1}$. Define $\mathbf{e}_k = \mathbf{z}_k - \boldsymbol{b}_{\mathscr{R}(\boldsymbol{A})^{\perp}}$ for $k \geq 0$. Then,

$$\begin{aligned}
\mathbf{e}_k &= (\boldsymbol{I} - \boldsymbol{P}_{\tau_k})\mathbf{z}_{k-1} - \boldsymbol{b}_{\mathscr{R}(\boldsymbol{A})^{\perp}} \\
&= (\boldsymbol{I} - \boldsymbol{P}_{\tau_k})\mathbf{z}_{k-1} - (\boldsymbol{I} - \boldsymbol{P}_{\tau_k})\boldsymbol{b}_{\mathscr{R}(\boldsymbol{A})^{\perp}} \\
&= (\boldsymbol{I} - \boldsymbol{P}_{\tau_k})\mathbf{e}_{k-1},
\end{aligned}$$

where the first equality follows by the definition of $\mathbf{z}_k$, the second by orthogonality between $\boldsymbol{P}_{\tau_k}$ and $\boldsymbol{b}_{\mathscr{R}(\boldsymbol{A})^{\perp}}$, and the final equality by definition of $\mathbf{e}_{k-1}$. Next, we prove that

$$\mathbb{E}_{k-1}\left\|\mathbf{e}_k\right\|_2^2 \leq \left(1 - \frac{\sigma_{\min}^2(\boldsymbol{A})}{p\beta}\right)\left\|\mathbf{e}_{k-1}\right\|_2^2$$

where $\mathbb{E}_{k-1}$ is the expectation conditioned over the first $(k-1)$ iterations of the algorithm. By orthogonality $\left\|(\boldsymbol{I} - \boldsymbol{P}_{\tau_k})\mathbf{e}_{k-1}\right\|_2^2 = \left\|\mathbf{e}_{k-1}\right\|_2^2 - \left\|\boldsymbol{P}_{\tau_k}\mathbf{e}_{k-1}\right\|_2^2$, hence it suffices to lower bound $\mathbb{E}_{k-1}\left\|\boldsymbol{P}_{\tau_k}\mathbf{e}_{k-1}\right\|_2^2$. Let $\boldsymbol{A}_{\tau_k} := \boldsymbol{U}_{\tau_k}\boldsymbol{\Sigma}_{\tau_k}\boldsymbol{V}_{\tau_k}^{\top}$ be its truncated SVD decomposition of $\boldsymbol{A}_{\tau_k}$ where $\boldsymbol{\Sigma}_{\tau_k}$ is an $\mathbf{rank}\left(\boldsymbol{A}_{\tau_k}\right) \times \mathbf{rank}\left(\boldsymbol{A}_{\tau_k}\right)$

diagonal matrix containing the non-zero singular values of $\boldsymbol{A}_{\tau_k}$. Then,

$$\begin{aligned}
\mathbb{E}_{k-1}\left\|\boldsymbol{P}_{\tau_k}\mathbf{e}_{k-1}\right\|_2^2 &= \mathbb{E}_{k-1}\left\|\boldsymbol{U}_{\tau_k}^\top\mathbf{e}_{k-1}\right\|_2^2\\
&= \mathbb{E}_{k-1}\left\|\boldsymbol{\Sigma}_{\tau_k}^{-1\top}\boldsymbol{V}_{\tau_k}^\top\boldsymbol{A}_{\tau_k}^\top\mathbf{e}_{k-1}\right\|_2^2\\
&\geq \mathbb{E}_{k-1}\,\sigma_{\min}^2(\boldsymbol{\Sigma}_{\tau_k}^{-1}\boldsymbol{V}_{\tau_k}^\top)\left\|\boldsymbol{A}_{\tau_k}^\top\mathbf{e}_{k-1}\right\|_2^2\\
&= \mathbb{E}_{k-1}\,\frac{\left\|\boldsymbol{A}_{\tau_k}^\top\mathbf{e}_{k-1}\right\|_2^2}{\sigma_{\max}^2(\boldsymbol{\Sigma}_{\tau_k})}\\
&\geq \frac{1}{p\beta}\sum_{\tau_k\in\mathscr{T}}\left\|\boldsymbol{A}_{\tau_k}^\top\mathbf{e}_{k-1}\right\|_2^2\\
&= \frac{1}{p\beta}\left\|\boldsymbol{A}^\top\mathbf{e}_{k-1}\right\|_2^2\\
&\geq \frac{\sigma_{\min}^2(\boldsymbol{A})}{p\beta}\left\|\mathbf{e}_{k-1}\right\|_2^2,
\end{aligned}$$

where the first equality follows since $\boldsymbol{P}_{\tau_k}=\boldsymbol{U}_{\tau_k}\boldsymbol{U}_{\tau_k}^\top$ and dropping $\boldsymbol{U}_{\tau_k}$ using the unitary invariance property of the Euclidean norm, the second equality by replacing $\boldsymbol{U}_{\tau_k}^\top$ with $\boldsymbol{\Sigma}_{\tau_k}^{-1}\boldsymbol{V}_{\tau_k}^\top\boldsymbol{A}_{\tau_k}^\top$, the first equality follows since $\sigma_{\min}^2(\boldsymbol{\Sigma}_{\tau_k}^{-1}\boldsymbol{V}_{\tau_k}^\top)=1/\sigma_{\max}^2(\boldsymbol{\Sigma}_{\tau_k})=1/\sigma_{\max}^2(\boldsymbol{A}_{\tau_k})$, the second inequality follows by the paving assumption, and the final inequality follows since[1] $\mathbf{e}_k\in\mathscr{R}(\boldsymbol{A})$ for all $k\geq 0$. It follows that

$$\mathbb{E}_{k-1}\left\|\mathbf{e}_k\right\|_2^2=\left\|\mathbf{e}_{k-1}\right\|_2^2-\mathbb{E}_{k-1}\left\|\boldsymbol{P}_{\tau_k}\mathbf{e}_{k-1}\right\|_2^2\leq\left(1-\frac{\sigma_{\min}^2(\boldsymbol{A})}{p\beta}\right)^k\left\|\mathbf{e}_{k-1}\right\|_2^2. \tag{8}$$

Repeat the above inequality k times and notice that $\mathbf{e}_0=\boldsymbol{b}-\boldsymbol{b}_{\mathscr{R}(\boldsymbol{A})^\perp}=\boldsymbol{b}_{\mathscr{R}(\boldsymbol{A})}$ to conclude. □

To utilize this result, we aim to relate the iterates $\mathbf{z}_k$ to the estimation $\mathbf{x}_k$. The following claim quantifies precisely this relation.

Lemma 4. *For every $k\geq 0$, at the end of the k-th iteration, it holds that $\mathbf{z}_{k+1}=\boldsymbol{b}-\boldsymbol{A}\mathbf{x}_{k+1}$.*

Proof. By induction. For $k=0$, $(\mathbf{x}_1)_{\tau_0}=\boldsymbol{a}_0$ and $(\mathbf{x}_1)_{[n]\setminus\tau_0}=\mathbf{0}$ and moreover, $\mathbf{z}_1=\mathbf{z}_0-\boldsymbol{A}_{\tau_0}\boldsymbol{a}_0=\boldsymbol{b}-\boldsymbol{A}\mathbf{x}_1$. Assume that $\mathbf{z}_l=\boldsymbol{b}-\boldsymbol{A}\mathbf{x}_l$ is true for some $l>0$, we will show that it holds for $l+1$. For the sake of notation, denote $\boldsymbol{P}_l=\boldsymbol{A}_{\tau_l}{\boldsymbol{A}_{\tau_l}}^\dagger$. Then

$$\mathbf{z}_{l+1}=\mathbf{z}_l-\boldsymbol{P}_l\mathbf{z}_l=\boldsymbol{b}-\boldsymbol{A}\mathbf{x}_l-\boldsymbol{P}_l\mathbf{z}_l \tag{9}$$

the first equality follows by the definition of $\mathbf{z}_{l+1}$, the second equality follows by induction hypothesis. Now, it follows that

$$\begin{aligned}
\boldsymbol{A}\mathbf{x}_{l+1} &= \boldsymbol{A}_{\tau_l}(\mathbf{x}_{l+1})_{\tau_l}+\boldsymbol{A}_{[n]\setminus\tau_l}(\mathbf{x}_{l+1})_{[n]\setminus\tau_l}\\
&= \boldsymbol{A}_{\tau_l}(\mathbf{x}_l)_{\tau_l}+\boldsymbol{A}_{\tau_l}\boldsymbol{a}_l+\boldsymbol{A}_{[n]\setminus\tau_l}(\mathbf{x}_{l+1})_{[n]\setminus\tau_l}\\
&= \boldsymbol{A}_{\tau_l}(\mathbf{x}_l)_{\tau_l}+\boldsymbol{A}_{\tau_l}\boldsymbol{a}_l+\boldsymbol{A}_{[n]\setminus\tau_l}(\mathbf{x}_l)_{[n]\setminus\tau_l}\\
&= \boldsymbol{A}\mathbf{x}_l+\boldsymbol{A}_{\tau_l}\boldsymbol{a}_l.
\end{aligned}$$

the first equality follows by Step 6 of the algorithm (update on $\mathbf{x}$), the second equality because $(\mathbf{x}_{l+1})_{[n]\setminus\tau_l}=(\mathbf{x}_l)_{[n]\setminus\tau_l}$. Hence, $\boldsymbol{A}\mathbf{x}_l=\boldsymbol{A}\mathbf{x}_{l+1}-\boldsymbol{A}_{\tau_l}\boldsymbol{a}_l$. Now, the right hand side of Eqn. (9) can be rewritten as

$$\begin{aligned}
\boldsymbol{b}-\boldsymbol{A}\mathbf{x}_l-\boldsymbol{P}_l\mathbf{z}_l &= \boldsymbol{b}-\boldsymbol{A}\mathbf{x}_{l+1}+\boldsymbol{A}_{\tau_l}\boldsymbol{a}_l-\boldsymbol{P}_l\mathbf{z}_l\\
&= \boldsymbol{b}-\boldsymbol{A}\mathbf{x}_{l+1}.
\end{aligned}$$

[1] Indeed, $\mathbf{e}_0=\boldsymbol{b}_{\mathscr{R}(\boldsymbol{A})}\in\mathscr{R}(\boldsymbol{A})$ and it follows that $\mathbf{e}_k\in\mathscr{R}(\boldsymbol{A})$ for every $k\geq 0$ by the recursive definition of $\mathbf{e}_k$.

the last equality follows since $\boldsymbol{a}_l = (\boldsymbol{A}_{\tau_l})^\dagger \mathbf{z}_l$. Therefore, we conclude that $\mathbf{z}_{l+1} = \boldsymbol{b} - \boldsymbol{A}\mathbf{x}_{l+1}$. □

Combining these two lemmas yields the following result which shows convergence of the estimation to the least squares solution under the map $\boldsymbol{A}$.

Theorem 5. *Algorithm 2 with input* $\boldsymbol{A}$*,* $\boldsymbol{b}$*,* $T \in \mathbb{N}$*, and* (p, α, β) *column paving* $\mathscr{T}$*, outputs an estimate vector* $\mathbf{x}_T$ *that satisfies*

$$\mathbb{E}\,\|\boldsymbol{A}(\mathbf{x}_{LS} - \mathbf{x}_T)\|_2^2 \le \left(1 - \frac{\sigma^2_{\min}(\boldsymbol{A})}{p\beta}\right)^T \left\|\boldsymbol{b}_{\mathscr{R}(\boldsymbol{A})}\right\|_2^2.$$

Proof. We observe that

$$\boldsymbol{A}(\mathbf{x}_{\mathrm{LS}} - \mathbf{x}_{(k)}) = \boldsymbol{b}_{\mathscr{R}(\boldsymbol{A})} - \boldsymbol{A}\mathbf{x}_{(k)} = \boldsymbol{b} - \boldsymbol{A}\mathbf{x}_{(k)} - \boldsymbol{b}_{\mathscr{R}(\boldsymbol{A})^\perp} = \mathbf{z}_{(k)} - \boldsymbol{b}_{\mathscr{R}(\boldsymbol{A})^\perp}$$

where the first equality follows by $\boldsymbol{b}_{\mathscr{R}(\boldsymbol{A})} = \boldsymbol{A}\boldsymbol{A}^\dagger \boldsymbol{b} = \boldsymbol{A}\mathbf{x}_{\mathrm{LS}}$, the second by orthogonality $\boldsymbol{b} = \boldsymbol{b}_{\mathscr{R}(\boldsymbol{A})^\perp} + \boldsymbol{b}_{\mathscr{R}(\boldsymbol{A})}$ and the last equality from Lemma 4. Combined with inequality (7) this yields the desired result. □

When $\boldsymbol{A}$ has full column rank, we may bound the estimation error $\|\mathbf{x}_{\mathrm{LS}} - \mathbf{x}_T\|_2$ by $\frac{1}{\sigma_{\min}(\boldsymbol{A})}\,\|\boldsymbol{A}(\mathbf{x}_{\mathrm{LS}} - \mathbf{x}_T)\|_2$ which combined with the fact that $\left\|\boldsymbol{b}_{\mathscr{R}(\boldsymbol{A})}\right\|_2 \le \sigma_{\max}(\boldsymbol{A})\,\|\mathbf{x}_{\mathrm{LS}}\|_2$ implies the following corollary.

Corollary 6. *Algorithm 2 with full-rank* $\boldsymbol{A}$*,* $\boldsymbol{b}$*,* $T \in \mathbb{N}$*, and* (p, α, β) *column paving* $\mathscr{T}$*, outputs an estimate vector* $\mathbf{x}_T$ *that satisfies*

$$\mathbb{E}\,\|\mathbf{x}_{LS} - \mathbf{x}_T\|_2^2 \le \left(1 - \frac{\sigma^2_{\min}(\boldsymbol{A})}{p\beta}\right)^T \kappa^2(\boldsymbol{A})\,\|\mathbf{x}_{LS}\|_2^2.$$

2.2. **Comparison of Convergence Rates.** This bound improves upon that of the randomized block Kaczmarz method because it demonstrates exponential convergence to the least squares solution $\mathbf{x}_{\mathrm{LS}}$, whereas Eqn. (6) shows convergence only within a radius proportional to $\|\mathbf{e}\|_2^2$, which we call the *convergence horizon*. Algorithm 2 is able to break this barrier because it iteratively removes the component of $\boldsymbol{b}$ which is orthogonal to the range of $\boldsymbol{A}$. This of course is also true of the randomized Extended Kaczmarz method (3) as it also breaks this horizon barrier. To compare the rate of Eqn. (4) to that of Corollary 6, we consider two important scenarios.

First, consider the case when $\boldsymbol{A}$ is nearly square, and each submatrix $\boldsymbol{A}_\tau$ can be applied efficiently via a fast multiply. In this case, each iteration of Algorithm 2 incurs approximately the same computational cost as an iteration of the REK method. Thus, we may directly compare the convergence rates of Corollary 6 and (4) to find that Algorithm 2 is about $n/(p\beta)$ times faster than REK in this setting. Thus when n is much larger than $p\beta$, this can result in a significant speedup.

Alternatively, if the matrix $\boldsymbol{A}$ does not admit a fast multiply, it is fair to only compare the convergence rate per *epoch*, since each iteration of Algorithm 2 may require more computational cost than those of REK. Since an epoch of Algorithm 2 and REK consist of p and m iterations, respectively, we see that the rate of the former is proportional to $\sigma^2_{\min}(\boldsymbol{A})/\beta$ whereas that of REK is proportional to $\sigma^2_{\min}(\boldsymbol{A})$. We see in this case that these bounds suggest REK exhibits faster convergence. However, as observed in the randomized Block Kaczmarz method, the block methods still display faster convergence than their single counterparts because of implicit computational issues in the linear algebraic subroutines. See the discussion in [NT13] and the experimental results below for further details.

2.3. **Convergence of Least Squares Solver via Good Pavings.** From these results we see that the convergence of the solver will be controlled by the paving parameters. Equipped with a column paving as in Proposition 1, Theorem 5 and Corollary 6 imply the following result.

Corollary 7. *Suppose Algorithm 2 is run with input $\boldsymbol{A}$, $\boldsymbol{b}$, $T \in \mathbb{N}$, and (p, α, β) paving $\mathscr{T}$ guaranteed by Proposition 1 for some fixed constant δ. Then the estimate vector $\mathbf{x}_T$ satisfies*

$$\mathbb{E}\,\|\boldsymbol{A}(\mathbf{x}_{LS} - \mathbf{x}_T)\|_2^2 \le \left(1 - \frac{C'}{\kappa^2(\boldsymbol{A})\log(n)}\right)^T \|\boldsymbol{b}_{\mathscr{R}(\boldsymbol{A})}\|_2^2,$$

where $\kappa(\boldsymbol{A}) = \frac{\sigma_{\max}(\boldsymbol{A})}{\sigma_{\min}(\boldsymbol{A})}$ denotes the condition number of $\boldsymbol{A}$ and C' is an absolute constant.

If the matrix $\boldsymbol{A}$ is full rank then,

$$\mathbb{E}\,\|\mathbf{x}_{LS} - \mathbf{x}_T\|_2^2 \le \left(1 - \frac{C'}{\kappa^2(\boldsymbol{A})\log(n)}\right)^T \kappa^2(\boldsymbol{A})\,\|\mathbf{x}_{LS}\|_2^2.$$

Remark 1. *In considering the improvements offered by both the REK method and the block Kaczmarz method, one may ask whether it is advantageous to run a traditional REK projection step as in* (3) *along with a traditional block Kaczmarz update step as in* (5)*. However, empirically we have observed that such a combination actually leads to a degradation in performance and requires far more epochs to converge than the algorithms discussed above. We conjecture that it is important to run both the projection update and the Kaczmarz update "at the same speed"; if the Kaczmarz update utilizes many rows at once, so should the projection update, and vice versa.*

3. Analysis of Main Result

It is natural to ask whether one can consider both a row partition and a column partition in the Kaczmarz method, blocking both in the Kaczmarz update step and the projection step. Indeed, utilizing blocking in both steps yields Algorithm 1 above. Combining the theoretical approaches in [NT13, ZF12] we will prove the following result about the convergence of Algorithm 1. This result utilizes both a column paving and row paving, the latter taken as a column paving of $\boldsymbol{A}^*$.

Theorem 8. *Algorithm 1 with input $\boldsymbol{A}$, $\boldsymbol{b}$, $T \in \mathbb{N}$, $(\overline{p}, \overline{\alpha}, \overline{\beta})$ column paving $\mathscr{T}$ of $\overline{\boldsymbol{A}}$, and (p, α, β) column paving $\mathscr{S}$ of $\boldsymbol{A}^*$, outputs an estimate vector $\mathbf{x}_T$ that satisfies*

$$\mathbb{E}\,\|\boldsymbol{x}_T - \mathbf{x}_{LS}\|_2^2 \le \gamma^T \|\boldsymbol{x}_0 - \mathbf{x}_{LS}\|_2^2 + \left(\gamma^{\lfloor T/2 \rfloor} + \overline{\gamma}^{\lfloor T/2 \rfloor}\right) \frac{\|\boldsymbol{b}_{\mathscr{R}(\boldsymbol{A})}\|_2^2}{\alpha(1-\gamma)},$$

where $\gamma = 1 - \frac{\sigma^2_{\min}(\boldsymbol{A})}{p\beta}$ and $\overline{\gamma} = 1 - \frac{\sigma^2_{\min}(\overline{\boldsymbol{A}})}{\overline{p}\overline{\beta}}$.

In light of Proposition 1, Theorem 8 implies the main result, Theorem 2.

Proof of Theorem 8. Observe that Steps 5 and 6 of Algorithm 1 are identical with Algorithm 2, therefore Lemma 3 implies that

$$\mathbb{E}\,\left\|\mathbf{z}_k - \boldsymbol{b}_{\mathscr{R}(\boldsymbol{A})^\perp}\right\|_2^2 \le \left(1 - \frac{\sigma^2_{\min}(\overline{\boldsymbol{A}})}{\overline{p}\overline{\beta}}\right)^k \left\|\boldsymbol{b}_{\mathscr{R}(\boldsymbol{A})}\right\|_2^2. \tag{10}$$

Lemma 2.2 of [NT13] shows that for any vector $\boldsymbol{u}$,

$$\mathbb{E}\left\|\left(\mathbf{I} - (\boldsymbol{A}_v)^\dagger \boldsymbol{A}_v\right)\boldsymbol{u}\right\|_2^2 \le \left(1 - \frac{\sigma^2_{\min}(\boldsymbol{A})}{p\beta}\right)\|\boldsymbol{u}\|_2^2. \tag{11}$$

Since the range of $\mathbf{I} - (\boldsymbol{A}_v)^\dagger \boldsymbol{A}_v$ and $(\boldsymbol{A}_v)^\dagger$ are orthogonal, we thus have

$$\|\boldsymbol{x}_k - \mathbf{x}_{\mathrm{LS}}\|_2^2 = \left\|\left(\mathbf{I} - (\boldsymbol{A}_v)^\dagger \boldsymbol{A}_v\right)(\boldsymbol{x}_{k-1} - \mathbf{x}_{\mathrm{LS}})\right\|_2^2 + \left\|(\boldsymbol{A}_v)^\dagger((\boldsymbol{z}_k)_v - \boldsymbol{b}_v^\perp)\right\|_2^2. \tag{12}$$

Combining (11) with $\boldsymbol{u} = \boldsymbol{x}_{k-1} - \mathbf{x}_{\mathrm{LS}}$ along with (12), we have

$$\begin{aligned}
\mathbb{E}\,\|\boldsymbol{x}_k - \mathbf{x}_{\mathrm{LS}}\|_2^2 &\leq \left(1 - \frac{\sigma^2_{\min}(\boldsymbol{A})}{p\beta}\right)\mathbb{E}\,\|\boldsymbol{x}_{k-1} - \mathbf{x}_{\mathrm{LS}}\|_2^2 + \mathbb{E}\left\|(\boldsymbol{A}_v)^{\dagger}((\boldsymbol{z}_k)_v - \boldsymbol{b}_v^{\perp})\right\|_2^2 \\
&\leq \left(1 - \frac{\sigma^2_{\min}(\boldsymbol{A})}{p\beta}\right)\mathbb{E}\,\|\boldsymbol{x}_{k-1} - \mathbf{x}_{\mathrm{LS}}\|_2^2 + \frac{1}{\sigma^2_{\min}(\boldsymbol{A}_v)}\,\mathbb{E}\left\|(\boldsymbol{z}_k)_v - \boldsymbol{b}_v^{\perp}\right\|_2^2 \\
&\leq \left(1 - \frac{\sigma^2_{\min}(\boldsymbol{A})}{p\beta}\right)\mathbb{E}\,\|\boldsymbol{x}_{k-1} - \mathbf{x}_{\mathrm{LS}}\|_2^2 + \frac{1}{\alpha}\,\mathbb{E}\left\|\boldsymbol{z}_k - \boldsymbol{b}^{\perp}\right\|_2^2 && (13) \\
& && (14)
\end{aligned}$$

To apply this bound recursively, we will utilize an elementary lemma. It is essentially proved in [ZF12, Theorem 8] but for completeness we recall its proof in the appendix.

Lemma 9. *Suppose that for some $\gamma, \overline{\gamma} < 1$, the following bounds hold for all $\ell, k^* \geq 0$:*

$$\mathbb{E}\,\|\boldsymbol{x}_{k^*} - \mathbf{x}_{LS}\|_2^2 \leq \gamma\,\mathbb{E}\,\|\boldsymbol{x}_{k^*-1} - \mathbf{x}_{LS}\|_2^2 + r_{k^*} \quad \textit{and} \quad r_{k^*} \leq \overline{\gamma}^{k^*} B. \tag{15}$$

Then for any $T > 0$,

$$\mathbb{E}\,\|\boldsymbol{x}_T - \mathbf{x}_{LS}\|_2^2 \leq \gamma^T \|\boldsymbol{x}_0 - \mathbf{x}_{LS}\|_2^2 + \left(\gamma^{\lfloor T/2 \rfloor} + \overline{\gamma}^{\lfloor T/2 \rfloor}\right)\frac{B}{1-\gamma}.$$

Using $r_k = \frac{1}{\alpha}\mathbb{E}\left\|\boldsymbol{z}_k - \boldsymbol{b}^{\perp}\right\|_2^2$, $B = \frac{\|\boldsymbol{b}_{\mathscr{R}(\boldsymbol{A})}\|_2^2}{\alpha}$, $\gamma = 1 - \frac{\sigma^2_{\min}(\boldsymbol{A})}{p\beta}$, and $\overline{\gamma} = 1 - \frac{\sigma^2_{\min}(\overline{\boldsymbol{A}})}{\overline{p}\overline{\beta}}$, we see by (10) and (13) that the bounds (15) hold. Applying Lemma 9 completes the proof.

□

3.1. **Comparison of Convergence Rates.** The convergence rates displayed in Theorems 8 and 2 depend on the column-standardized $\overline{\boldsymbol{A}}$ version of the matrix $\boldsymbol{A}$. For this reason, it is difficult to make direct comparisons for arbitrary linear systems. In some cases, $\boldsymbol{A}$ and $\overline{\boldsymbol{A}}$ may have substantially different condition numbers, and paving both simultaneously may not lead to much improvement in convergence. However, there are also cases where column pavings appear naturally with row pavings. For example, if the matrix is positive semi-definite (or symmetric), then $\boldsymbol{A} = \overline{\boldsymbol{A}}$ in which case one gets the column paving for free from the row paving. In nice cases like this, the convergence bounds from Theorems 8 and 2 offer the same improvements as those discussed for Corollary 6. However, since Algorithm 1 utilizes blocking in two steps, we expect even more improvement in convergence due to implicit computational issues. Finally, we note that it is not necessary to actually compute the column standardized version of $\boldsymbol{A}$. Indeed, paving results analogous to those of Proposition 1 exist for matrices which are not standardized [Ver01]. For simplicity of presentation we only consider standardized versions.

4. Experimental Results

Here we present some experiments using simple examples to illustrate the benefits of block methods. We refer the reader to [ZF12, NT13] for more empirical results for both REK and block methods. In all experiments, one matrix is created and 40 trials of each method are run. In our first experiment, the matrix is a 300×100 matrix with standard normal entries, whose rows are then normalized to each have norm one, yielding a condition number of 3.7. The vector $\boldsymbol{x}$ is created to have independent standard normal entries, and the right hand side $\boldsymbol{b}$ is set to $\boldsymbol{Ax}$. We track the ℓ_2-error $\|\boldsymbol{x}_{LS} - \boldsymbol{x}_{\boldsymbol{k}}\|_2$ across each epoch[2] as well as the CPU time (measured in Matlab using the

[2] We refer to an epoch as the number of iterations that is equivalent to one cycle through m rows, even though rows and blocks are selected with replacement. Thus for REK, an epoch is m iterations, and for a block version with b blocks, one epoch is b iterations.

cputime command). In all experiments we considered a trial successful when the error reached 10^{-7}. The results for this case are presented in Figures 1 and 2. In all figures, a heavy line represents median performance, and the shaded region spans the minimum to the maximum value across all trials. As is demonstrated, even when the matrix does not have any natural block structure, the proposed algorithms outperform standard REK both in terms of epochs and CPU time.

Figure 3 shows similar plots, but in this case the system is no longer consistent. For these experiments, we used the same type and size of the matrix $\boldsymbol{A}$, but the right hand side vector $\boldsymbol{b}$ was generated as a Gaussian vector as well. We created $\boldsymbol{b}$ so that the residual norm $\|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x_{LS}}\|_2 = 0.5$. We then track the ℓ_2-error between the iterate $\boldsymbol{x_k}$ and the least-squares solution $\boldsymbol{x_{LS}}$ which we computed by $\boldsymbol{A}^{\dagger}\boldsymbol{b}$. The behavior, as predicted by our main results, is quite similar to the consistent case and thus breaks the convergence horizon of the standard Kaczmarz method.

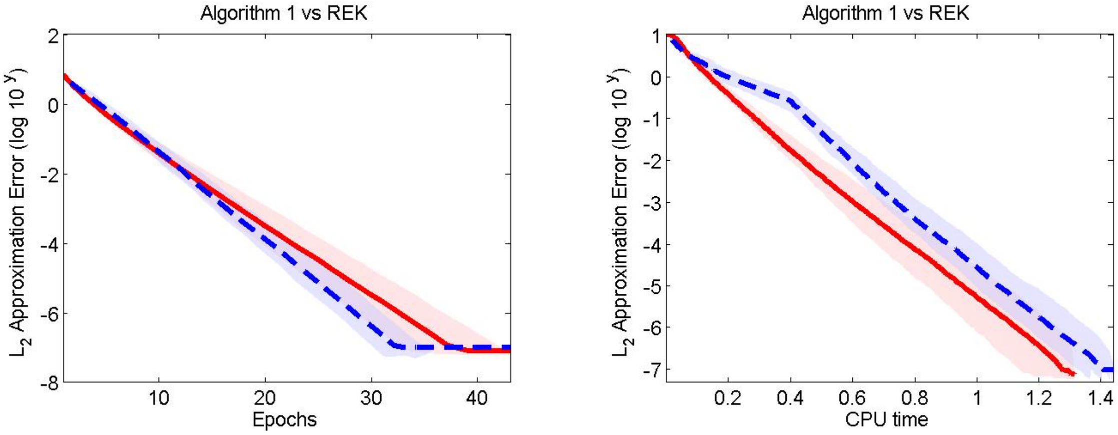

Figure 1 ℓ_2-norm error for REK (blue dashed) and Algorithm 1 (red) across epochs (left) and CPU time (right). Matrix is 300 × 100 Gaussian, system is consistent.

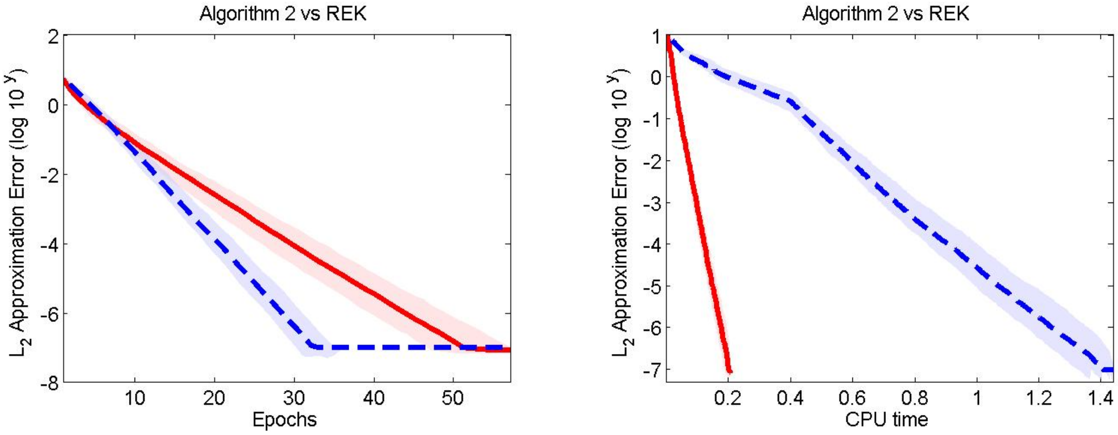

Figure 2 ℓ_2-norm error for REK (blue dashed) and Algorithm 2 (red) across epochs (left) and CPU time (right). Matrix is 300 × 100 Gaussian, system is consistent.

5. Related Work and Discussion

The Kaczmarz method was first introduced in the 1937 work of Kaczmarz himself [Kac37]. Since then, the method has been revitalized by researchers in computer tomography, under the name

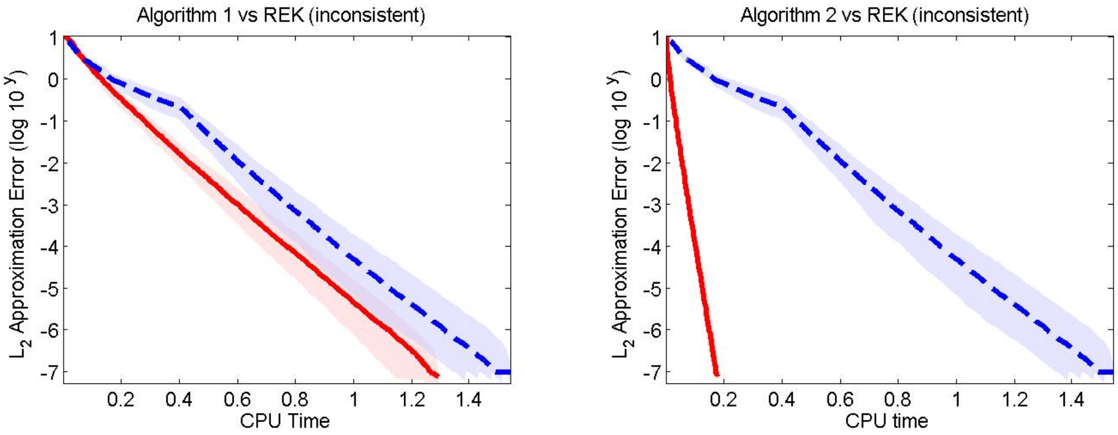

Figure 3 ℓ_2-norm error for REK (blue dashed) and Algorithm 2 (red) across epochs (left) and CPU time (right). Matrix is 300×100 Gaussian, system is inconsistent.

Algebraic Reconstruction Technique (ART) [GBH70, Byr08, Nat01, Her09]. Deterministic convergence results for the method often depend on properties of the matrix that are difficult to compute or analyze [Deu85, DH97, XZ02, Gal05]. Moreover, it has been well observed that random choice of row selection often speeds up the convergence [HS78, HM93, CFM+92, Nat01].

Recently, Strohmer and Vershynin [SV09] derived the first provable convergence rate of the Kaczmarz method, showing that when each row is selected with probability proportional to its norm the method exhibits the expected exponential convergence of (1). This work was extended to the inconsistent case in [Nee10], which shows exponential convergence to within some fixed radius of the least-squares solution. The almost-sure guarantees were recently derived by Chen and Powell [CP12]. To break the convergence barrier, relaxation parameters can be introduced, so that each iterate is over or under projected onto each solution space. Whitney and Meany prove that if the relaxation parameters tend to zero that the iterates converge to the least-squares solution [WM67]. Further results using relaxation have also been obtained, see for example [CEG83, Tan71, HN90, ZF12]. An alternative to relaxation parameters was recently proposed by Zouzias and Freris [ZF12] as the REK method described by (3). Rather than alter the projection step, motivated by ideas of Popa [Pop98] they introduce a secondary step which aims to reduce the residual.

The Kaczmarz method has been extended beyond linear systems as well. For example, Leventhal and Lewis [LL10] analyze the method for systems with polyhedral constraints, and Richtárik and Takáč [RT11] build on these results for general optimization problems.

Another important aspect of research in this area focuses on accelerating the convergence of the methods. Geometric brute force methods can be used [EN11], additional row directions may be added [PPKR12], or instead one can select *blocks* of rows rather than a single row in each iteration. The block version of the Kaczmarz method is originally due to work of Elfving [Elf80] and Eggermont et al. [EHL81]. Its convergence rates were recently studied in [NW13] and analyzed via pavings by Needell and Tropp [NT13]. The block Kaczmarz method is of course a special instance in a broader class of block projection algorithms, see for example [XZ02] for a more general analysis and [Byr08] for a presentation of other block variants.

To use block methods effectively, one needs to obtain a suitable partition of the rows (and/or columns). Popa constructs such partitions by creating orthogonal blocks [Pop99, Pop01, Pop04], whereas Needell and Tropp promote the use of row pavings to construct the partition [NT13].

Construction of pavings has been studied for quite some time now, and most early results rely on random selection. The guarantee of lower and upper paving bounds has been derived by Bourgain and Tzafriri [BT87] and Kashin and Tzafriri [KT94], respectively. Simultaneous guarantees were later derived by Bourgain and Tzafriri [BT91] with suboptimal dependence on the matrix norm. Recently, Spielman and Srivastava [SS12] and Youssef [You12b] provided simple proofs of the results from [BT87] and [KT94], respectively. Vershynin [Ver01] and Srivastava [Sri10] extend the paving results to general matrices with arbitrary row norms; see also [You12b, You12a]. Proposition 1 follows from the work of Vershynin [Ver06] and Tropp [Tro09], and is attributed to the seminal work of Bourgain and Tzafriri [BT87, BT91]. For particular classes of matrices, the paving can even be obtained from a random partition of the rows with high probability. This is proved by Tropp [Tro08a] using ideas from [BT91, Tro08b], and is refined in [CD12].

Appendix A. Proof of intermediate results

Here we include the proof of Lemma 9.

Proof of Lemma 9. Assume the bounds (15) hold. Applying the first bound in (15) recursively yields

$$\begin{aligned}
\mathbb{E}\,\|\boldsymbol{x}_{k^*} - \mathbf{x}_{\mathrm{LS}}\|_2^2 &\le \gamma^{k^*}\|\boldsymbol{x}_0 - \mathbf{x}_{\mathrm{LS}}\|_2^2 + \sum_{j=0}^{k^*-1}\gamma^{k^*-1-j}r_j \\
&\le \gamma^{k^*}\|\boldsymbol{x}_0 - \mathbf{x}_{\mathrm{LS}}\|_2^2 + \sum_{j=0}^{\infty}\gamma^j B \\
&\le \gamma^{k^*}\|\boldsymbol{x}_0 - \mathbf{x}_{\mathrm{LS}}\|_2^2 + \frac{B}{1-\gamma},
\end{aligned}$$

where the second inequality holds by the assumption that $r_k \le \gamma^k B \le B$, and the last by the properties of the geometric summation. Similarly, observe that for any k and k^* we have

$$\begin{aligned}
\mathbb{E}\,\|\boldsymbol{x}_{k+k^*} - \mathbf{x}_{\mathrm{LS}}\|_2^2 &\le \gamma^k\,\mathbb{E}\,\|\boldsymbol{x}_{k^*} - \mathbf{x}_{\mathrm{LS}}\|_2^2 + \sum_{j=0}^{k-1}\gamma^{k-1-j}r_{j+k^*} \\
&\le \gamma^k\,\mathbb{E}\,\|\boldsymbol{x}_{k^*} - \mathbf{x}_{\mathrm{LS}}\|_2^2 + \overline{\gamma}^{k^*}\sum_{j=0}^{\infty}\gamma^j B \\
&\le \gamma^k\,\mathbb{E}\,\|\boldsymbol{x}_{k^*} - \mathbf{x}_{\mathrm{LS}}\|_2^2 + \overline{\gamma}^{k^*}\frac{B}{1-\gamma}.
\end{aligned}$$

Now we choose k and k^* such that $T = k + k^*$ and $k = k^*$ if T is even, or $k = k^* + 1$ if T is odd. Combining the two inequalities above, we have

$$\begin{aligned}
\mathbb{E}\,\|\boldsymbol{x}_T - \mathbf{x}_{\mathrm{LS}}\|_2^2 &= \mathbb{E}\,\|\boldsymbol{x}_{k+k^*} - \mathbf{x}_{\mathrm{LS}}\|_2^2 \\
&\le \gamma^k\,\mathbb{E}\,\|\boldsymbol{x}_{k^*} - \mathbf{x}_{\mathrm{LS}}\|_2^2 + \overline{\gamma}^{k^*}\frac{B}{1-\gamma} \\
&\le \gamma^k\left(\gamma^{k^*}\|\boldsymbol{x}_0 - \mathbf{x}_{\mathrm{LS}}\|_2^2 + \frac{B}{1-\gamma}\right) + \overline{\gamma}^{k^*}\frac{B}{1-\gamma} \\
&= \gamma^{k+k^*}\|\boldsymbol{x}_0 - \mathbf{x}_{\mathrm{LS}}\|_2^2 + \left(\gamma^k + \overline{\gamma}^{k^*}\right)\frac{B}{1-\gamma} \\
&\le \gamma^T\|\boldsymbol{x}_0 - \mathbf{x}_{\mathrm{LS}}\|_2^2 + \left(\gamma^{\lfloor T/2\rfloor} + \overline{\gamma}^{\lfloor T/2\rfloor}\right)\frac{B}{1-\gamma}.
\end{aligned}$$

This completes the proof. □

References

[Bjö96] Å. Björck. *Numerical Methods for Least Squares Problems*. SIAM, Philadelphia, 1996.

[BT87] J. Bourgain and L. Tzafriri. Invertibility of "large" submatrices with applications to the geometry of Banach spaces and harmonic analysis. *Israel J. Math.*, 57(2):137–224, 1987.

[BT91] J. Bourgain and L. Tzafriri. On a problem of Kadison and Singer. *J. Reine Angew. Math.*, 420:1–43, 1991.

[Byr08] C. L. Byrne. *Applied iterative methods*. A K Peters Ltd., Wellesley, MA, 2008.

[CD12] S. Chrétien and S. Darses. Invertibility of random submatrices via tail decoupling and a matrix Chernoff inequality. *Statist. Probab. Lett.*, 82(7):1479–1487, 2012.

[CEG83] Y. Censor, P. P. B. Eggermont, and D. Gordon. Strong underrelaxation in kaczmarz's method for inconsistent systems. *Numerische Mathematik*, 41(1):83–92, 1983.

[CFM$^+$92] C. Cenker, H. G. Feichtinger, M. Mayer, H. Steier, and T. Strohmer. New variants of the POCS method using affine subspaces of finite codimension, with applications to irregular sampling. In *Proc. SPIE: Visual Communications and Image Processing*, pages 299–310, 1992.

[CP12] X. Chen and A. Powell. Almost sure convergence of the Kaczmarz algorithm with random measurements. *J. Fourier Anal. Appl.*, pages 1–20, 2012. 10.1007/s00041-012-9237-2.

[Deu85] F. Deutsch. Rate of convergence of the method of alternating projections. *Parametric optimization and approximation*, 76:96–107, 1985.

[DH97] F. Deutsch and H. Hundal. The rate of convergence for the method of alternating projections, ii. *J. Math. Anal. Appl.*, 205(2):381–405, 1997.

[EHL81] P. P. B. Eggermont, G. T. Herman, and A. Lent. Iterative algorithms for large partitioned linear systems, with applications to image reconstruction. *Linear Algebra Appl.*, 40:37–67, 1981.

[Elf80] T. Elfving. Block-iterative methods for consistent and inconsistent linear equations. *Numer. Math.*, 35(1):1–12, 1980.

[EN11] Y. C. Eldar and D. Needell. Acceleration of randomized Kaczmarz method via the Johnson-Lindenstrauss lemma. *Numer. Algorithms*, 58(2):163–177, 2011.

[Gal05] A. Galántai. On the rate of convergence of the alternating projection method in finite dimensional spaces. *J. Math. Anal. Appl.*, 310(1):30–44, 2005.

[GBH70] R. Gordon, R. Bender, and G. T. Herman. Algebraic reconstruction techniques (ART) for three-dimensional electron microscopy and X-ray photography. *J. Theoret. Biol.*, 29:471–481, 1970.

[Her09] G. T. Herman. *Fundamentals of computerized tomography: image reconstruction from projections*. Springer, 2009.

[HM93] G. Herman and L. Meyer. Algebraic reconstruction techniques can be made computationally efficient. *IEEE Trans. Medical Imaging*, 12(3):600–609, 1993.

[HN90] M. Hanke and W. Niethammer. On the acceleration of kaczmarz's method for inconsistent linear systems. *Linear Algebra Appl.*, 130:83–98, 1990.

[HS78] C. Hamaker and D. C. Solmon. The angles between the null spaces of X-rays. *J. Math. Anal. Appl.*, 62(1):1–23, 1978.

[Kac37] S. Kaczmarz. Angenäherte auflösung von systemen linearer gleichungen. *Bull. Internat. Acad. Polon.Sci. Lettres A*, pages 335–357, 1937.

[KT94] B. Kashin and L. Tzafriri. Some remarks on coordinate restriction of operators to coordinate subspaces. Insitute of Mathematics Preprint 12, Hebrew University, Jerusalem, 1993–1994.

[LL10] D. Leventhal and A. S. Lewis. Randomized methods for linear constraints: convergence rates and conditioning. *Math. Oper. Res.*, 35(3):641–654, 2010.

[Nat01] F. Natterer. *The mathematics of computerized tomography*, volume 32 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2001. Reprint of the 1986 original.

[Nee10] D. Needell. Randomized Kaczmarz solver for noisy linear systems. *BIT*, 50(2):395–403, 2010.

[NT13] D. Needell and J. A. Tropp. Paved with good intentions: Analysis of a randomized block kaczmarz method. *Linear Algebra Appl.*, pages 199–221, 2013.

[NW13] D. Needell and R. Ward. Two-subspace projection method for coherent overdetermined linear systems. *J. Fourier Anal. Appl.*, 19(2):256–269, 2013.

[Pop98] C. Popa. Extensions of block-projections methods with relaxation parameters to inconsistent and rank-deficient least-squares problems. *BIT*, 38(1):151–176, 1998.

[Pop99] C. Popa. Block-projections algorithms with blocks containing mutually orthogonal rows and columns. *BIT*, 39(2):323–338, 1999.

[Pop01] C. Popa. A fast Kaczmarz-Kovarik algorithm for consistent least-squares problems. *Korean J. Comput. Appl. Math.*, 8(1):9–26, 2001.

[Pop04] C. Popa. A Kaczmarz-Kovarik algorithm for symmetric ill-conditioned matrices. *An. Ştiinţ. Univ. Ovidius Constanţa Ser. Mat.*, 12(2):135–146, 2004.

[PPKR12] C. Popa, T. Preclik, H. Köstler, and U. Rüde. On KaczmarzŠs projection iteration as a direct solver for linear least squares problems. *Linear Algebra Appl.*, 436(2):389–404, 2012.

[RT11] P. Richtárik and M. Takáč. Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. Available at `arXiv:1107.2848`, Apr. 2011.

[Sri10] N. Srivastava. *Spectral sparsification and restricted invertibility*. Phd dissertation, Yale University, New Haven, CT, 2010.

[SS87] K. M. Sezan and H. Stark. Applications of convex projection theory to image recovery in tomography and related areas. In H. Stark, editor, *Image Recovery: Theory and application*, pages 415Ű–462. Acad. Press, 1987.

[SS12] D. A. Spielman and N. Srivastava. An elementary proof of the restricted invertibility theorem. *Israel J. Math.*, 190:83–91, 2012.

[SV06] T. Strohmer and R. Vershynin. A randomized solver for linear systems with exponential convergence. In *RANDOM 2006 (10th International Workshop on Randomization and Computation)*, number 4110 in Lecture Notes in Computer Science, pages 499–507. Springer, 2006.

[SV09] T. Strohmer and R. Vershynin. A randomized Kaczmarz algorithm with exponential convergence. *J. Fourier Anal. Appl.*, 15(2):262–278, 2009.

[Tan71] K. Tanabe. Projection method for solving a singular system of linear equations and its applications. *Numerische Mathematik*, 17(3):203–214, 1971.

[Tro08a] J. A. Tropp. Norms of random submatrices and sparse approximation. *C. R. Math. Acad. Sci. Paris*, 346(23-24):1271–1274, 2008.

[Tro08b] J. A. Tropp. The random paving property for uniformly bounded matrices. *Studia Math.*, 185(1):67–82, 2008.

[Tro09] J. A. Tropp. Column subset selection, matrix factorization, and eigenvalue optimization. In *Proceedings of the Twentieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 978–986, Philadelphia, PA, 2009. SIAM.

[Ver01] R. Vershynin. John's decompositions: selecting a large part. *Israel J. Math.*, 122:253–277, 2001.

[Ver06] R. Vershynin. Random sets of isomorphism of linear operators on Hilbert space. In *High dimensional probability*, volume 51 of *IMS Lecture Notes Monogr. Ser.*, pages 148–154. Inst. Math. Statist., Beachwood, OH, 2006.

[WM67] T. M. Whitney and R. K. Meany. Two algorithms related to the method of steepest descent. *SIAM J. Numer. Anal.*, 4(1):109–118, 1967.

[XZ02] J. Xu and L. Zikatanov. The method of alternating projections and the method of subspace corrections in Hilbert space. *J. Amer. Math. Soc.*, 15(3):573–597, 2002.

[You12a] P. Youssef. A note on column subset selection. Available at `arXiv:1212.0976`, Dec. 2012.

[You12b] P. Youssef. Restricted invertibility and the Banach–Mazur distance to the cube. Available at `arXiv:1206.0654`, June 2012.

[ZF12] A. Zouzias and N. M. Freris. Randomized extended Kaczmarz for solving least-squares. *SIAM J. Matrix Anal. A.*, 34(2):773–793, 2012.