6-28-2014

Block Kaczmarz Method with Inequalities

Jonathan Briskman

Deanna Needell
Claremont McKenna College

BLOCK KACZMARZ METHOD WITH INEQUALITIES

J. BRISKMAN AND D. NEEDELL

ABSTRACT. The randomized Kaczmarz method is an iterative algorithm that solves overdetermined systems of linear equations. Recently, the method was extended to systems of equalities and inequalities by Leventhal and Lewis. Even more recently, Needell and Tropp provided an analysis of a block version of the method for systems of linear equations. This paper considers the use of a block type method for systems of mixed equalities and inequalities, bridging these two bodies of work. We show that utilizing a matrix paving over the equalities of the system can lead to significantly improved convergence, and prove a linear convergence rate as in the standard block method. We also demonstrate that using blocks of inequalities offers similar improvement only when the system satisfies a certain geometric property. We support the theoretical analysis with several experimental results.

1. INTRODUCTION

The Kaczmarz method [Kac37] is an iterative algorithm for solving linear systems of equations. Each iteration projects from one constraint to the next, converging to the solution at a linear rate.The method has been applied to fields including image reconstruction, digital signal processing, and computer tomography [SS87, FCM$^+$92, Nat01, FS95]. Leventhal and Lewis modify the standard randomized Kaczmarz method to apply to systems of linear equalities and inequalities [LL10]. Unlike the standard algorithm which enforces a single constraint at each iteration, the block Kaczmarz approach recently analyzed by Needell and Tropp [NT14] enforces multiple constraints simultaneously and thus offers computational advantages. Here we demonstrate convergence for a system of linear equalities and inequalities by combining a randomized block Kaczmarz method for the equalities with a randomized Kaczmarz algorithm for the inequalities. These results indicate that the block Kaczmarz method can be used for a system of equalities and inequalities, and in some cases may quicken convergence. We also consider the case of utilizing blocking in both the equalities and inequalities, although this can be detrimental unless the geometry of the system meets certain conditions.

1.1. **Model and Notation.** We consider a linear system

$$Ax = b, \tag{1.1}$$

where A is a real (or complex) $n \times d$ matrix, typically with $n \gg d$.

The ℓ_p vector norm for $p \in [1, \infty]$ is denoted $\|\cdot\|_p$, while $\|\cdot\|$ is the spectral norm and $\|\cdot\|_F$ refers to the Frobenius norm. For an $n \times d$ matrix A, the singular values are arranged in decreasing order and we write

$$\sigma_{\max}(A) \overset{\text{def}}{=} \sigma_1(A) \ge \sigma_2(A) \ge \cdots \ge \sigma_d(A) \overset{\text{def}}{=} \sigma_{\min}(A).$$

We define the eigenvalues $\lambda_{\min}(A), \ldots, \lambda_{\max}(A)$ of a matrix analogously. For convenience we will assume that each row a_i of A has unit norm, $\|a_i\|_2 = 1$, and we call such matrices *standardized*.

We define the usual condition number $\kappa(A) \overset{\text{def}}{=} \sigma_{\max}(A)/\sigma_{\min}(A)$, and write the Moore-Penrose pseudoinverse of matrix A by $A^{\dagger}$. Recall that for a matrix A with full row rank, the pseudoinverse is obtained by $A^{\dagger} \overset{\text{def}}{=} A^*(AA^*)^{-1}$.

Now we consider a system of linear equalities and inequalities and denote by S its non-empty set of feasible solutions. We thus consider the matrix $\boldsymbol{A}$ whose rows can be arranged such that

$$\boldsymbol{A} = \left[\begin{array}{c} \boldsymbol{A}_= \\ \boldsymbol{A}_\le \end{array} \right], \tag{1.2}$$

and we will write $I_=$ and $I_\le$ to denote the row indices of $\boldsymbol{A}_=$ and $\boldsymbol{A}_\le$, respectively. Therefore, we ask that

$$\langle \boldsymbol{a}_i, \boldsymbol{x} \rangle \le \boldsymbol{b}_i \quad (i \in I_\le) \quad \text{and} \quad \langle \boldsymbol{a}_i, \boldsymbol{x} \rangle = \boldsymbol{b}_i \quad (i \in I_=) \tag{1.3}$$

We will assume that the set of rows $\{1, 2, ..., n\}$ is partitioned such that the first n_e rows correspond to equalities, and the remaining $n_i = n - n_e$ rows to inequalities. Thus $\boldsymbol{A}_=$ is an $n_e \times d$ matrix and $\boldsymbol{A}_\le$ is $n_i \times d$.

The error bound for this system of linear inequalities uses function $e : \mathbf{R}^n \to \mathbf{R}^n$ defined as in [LL10] by

$$e(y)_i = \left\{ \begin{array}{ll} y_i^+ & \text{for } i \in I_\le \\ y_i & \text{for } i \in I_= \end{array} \right.$$

where the positive part is defined as usual, $x^+ \stackrel{\text{def}}{=} \max(x, 0)$.

1.2. **Details of Kaczmarz.** The simple Kaczmarz method is an iterative algorithm that approximates a least-squares minimizer $\boldsymbol{x}_\star$ to the problem in (1.1). It takes an arbitrary initial approximation $\boldsymbol{x}_0$, and at each iteration j the current iterate is projected orthogonally onto the solution hyperplane $\{\langle \boldsymbol{a}_i, \boldsymbol{x} \rangle = \boldsymbol{b}_i\}$, using the update rule

$$\boldsymbol{x}_{j+1} = \boldsymbol{x}_j + \frac{\boldsymbol{b}_i - \langle \boldsymbol{a}_i, \boldsymbol{x}_j \rangle}{\|\boldsymbol{a}_i\|_2^2} \boldsymbol{a}_i \tag{1.4}$$

where $i = j \bmod n + 1$ [Kac37]. With an unfortunate ordering of the rows, this method as-is can produce very slow convergence. However, it has been well known that using randomized selection often eliminates this effect [HS78, HM93]. The randomized Kaczmarz method put forth by Strohmer and Vershynin [SV09] uses a random selection method for the selection of row i such that each row i is selected with probability proportional to $\|\boldsymbol{a}_i\|_2^2$. This randomization provides an algorithm that is both simple to analyze and enforce in many cases. In this paper we assume each row has unit norm, so each row is selected uniformly at random from {1,...,n} in the simple randomized Kaczmarz approach. Strohmer and Vershynin prove a linear rate of convergence for consistent systems that depends on the scaled condition number of $\boldsymbol{A}$, and not on the number of equations n [SV09],

$$\mathbb{E}\|\boldsymbol{x}_j - \boldsymbol{x}_\star\|_2^2 \le \left[1 - \frac{1}{K} \right]^j \|\boldsymbol{x}_0 - \boldsymbol{x}\|_2^2, \tag{1.5}$$

where $\boldsymbol{x}_\star$ is the solution to the consistent system (1.1) and $K = \|\boldsymbol{A}\|_F^2 / \sigma_{\min}^2(\boldsymbol{A})$ denotes the scaled condition number. Needell extended this work to the inconsistent case and proves linear convergence to the least-squares solution within some fixed radius [Nee10],

$$\mathbb{E}\|\boldsymbol{x}_j - \boldsymbol{x}_\star\|_2^2 \le \left[1 - \frac{1}{K} \right]^j \|\boldsymbol{x}_0 - \boldsymbol{x}\|_2^2 + K \|\boldsymbol{e}\|_\infty^2,$$

where $\boldsymbol{e} = \boldsymbol{A}\boldsymbol{x}_\star - \boldsymbol{b}$ denotes the residual vector. Because the Kaczmarz method projects directly onto each solution hyperplane, such a convergence radius is unavoidable without adding a relaxation parameter.

The randomized Kaczmarz method can be adapted to the case of a linear system of equalities and inequalities described in (1.3). Leventhal and Lewis [LL10] apply the Kaczmarz method to a consistent system of linear equalities and inequalities (here consistent simply means the feasible set S is non-empty). At each iteration j, the previous iterate only projects onto the solution hyperplane if the inequality is not

already satisfied. If the inequality is satisfied for row i selected at iteration j ($\boldsymbol{a}_i^T \boldsymbol{x} \le \boldsymbol{b}_i$), the approximation $\boldsymbol{x}_j$ is set as $\boldsymbol{x}_{j-1}$ [LL10]. The update rule for this algorithm is thus

$$\boldsymbol{x}_{j+1} = \boldsymbol{x}_j - \frac{e(\boldsymbol{a}_i^T \boldsymbol{x}_j - \boldsymbol{b}_i)}{\|\boldsymbol{a}_i\|_2^2} \boldsymbol{a}_i. \tag{1.6}$$

This algorithm converges linearly in expectation [LL10], with

$$\mathbb{E}\left[d(\boldsymbol{x}_j, S)^2 \mid \boldsymbol{x}_{j-1} \right] \le d(\boldsymbol{x}_{j-1}, S)^2 - \frac{\|e(\boldsymbol{A}\boldsymbol{x}_{j-1} - \boldsymbol{b})\|_2^2}{\|\boldsymbol{A}\|_F^2}.$$

In order to bound the right hand side of this expression, the authors rely on a lemma due to Hoffman [Hof52, LL10]. This result states that for any system (1.3) with non-empty solution set S, there exists a constant L independent of $\boldsymbol{b}$ such that for all $\boldsymbol{x}$,

$$d(\boldsymbol{x}, S) \le L \|e(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b})\|_2. \tag{1.7}$$

When $\boldsymbol{A}_= = \boldsymbol{A}$ is full column rank, the Hoffman constant is the inverse of the smallest singular value, $L = \sigma_{\min}^{-1}(\boldsymbol{A})$.

Using this their result becomes

$$\mathbb{E}\left[d(\boldsymbol{x}_j, S)^2 \right] \le \left[1 - \frac{1}{L^2 \|\boldsymbol{A}\|_F^2} \right]^j \cdot d(\boldsymbol{x}_0, S)^2, \tag{1.8}$$

which coincides with (1.5) for consistent systems of equalities.

1.3. **Block Kaczmarz.** A block variant of the randomized Kaczmarz method due to Elfving [Elf80] has been recently analyzed by Needell and Tropp [NT14] and can improve the convergence rate in certain cases. The block Kaczmarz method first partitions the rows $\{1, ..., n\}$ into m blocks, denoted $\tau_1, \ldots \tau_m$. Instead of selecting one row per iteration as done with the simple Kaczmarz method, the block Kaczmarz algorithm chooses a block uniformly at random at each iteration. Thus the block Kaczmarz method enforces multiple constraints simultaneously. At each iteration, the previous iterate $\boldsymbol{x}_{j-1}$ is projected onto the solution space to $\boldsymbol{A}_\tau \boldsymbol{x} = \boldsymbol{b}_\tau$, which enforces the set of equations in block τ [NT14]. $\boldsymbol{A}_\tau$ and $\boldsymbol{b}_\tau$ are written as the row submatrix of $\boldsymbol{A}$ and the subvector of $\boldsymbol{b}$ indexed by τ respectively, yielding an iterative rule of

$$\boldsymbol{x}_j = \boldsymbol{x}_{j-1} + (\boldsymbol{A}_\tau)^\dagger (\boldsymbol{b}_\tau - \boldsymbol{A}_\tau \boldsymbol{x}_{j-1}). \tag{1.9}$$

The pseudoinverse used in (1.9) returns the solution to the underdetermined least squares problem for a wide or square row submatrix $\boldsymbol{A}_\tau$.

Depending on the characteristics of the submatrix $\boldsymbol{A}_\tau$, the block method can provide better convergence than the simple method. If we assume that the submatrices $\boldsymbol{A}_\tau$ are well conditioned and especially if they admit a fast multiply, then the computational cost of the block algorithm (1.9) is similar to the cost of the simple update rule in (1.4) [NT14]. Since the convergence depends heavily on the conditioning of each submatrix, one seeks partitions of the rows into blocks for which each block is well-conditioned. The notion of a *row-paving* allows one to do precisely that.

Definition 1.1. We define an (m, β) row paving[1] of matrix $\boldsymbol{A}$ as a partition $T = \{\tau_1, ... \tau_m\}$ of the row indices such that

$$\lambda_{\max}(\boldsymbol{A}_\tau \boldsymbol{A}_\tau^*) \le \beta \quad \text{for each } \tau \in T.$$

---

[1]The standard definition of a row paving also includes a constant α which serves as a lower bound to the smallest singular value. We ignore that parameter here since it will not be utilized.

The *size* of the paving, or number of blocks, is m. The value of β is the upper paving bound, which controls the conditioning of the submatrices. Needell and Tropp [NT14] show that these parameters determine the performance of the algorithm, with convergence for a consistent system admitting an (m, β) paving given by

$$\mathbb{E}\|\boldsymbol{x}_j - \boldsymbol{x}_\star\|_2^2 \leq \left[1 - \frac{\sigma_{\min}^2(\boldsymbol{A})}{\beta m}\right]^j \|\boldsymbol{x}_0 - \boldsymbol{x}_\star\|_2^2. \tag{1.10}$$

Therefore the convergence rate depends on the size m and upper bound β; the algorithm's performance improves with low values of m and β, and large $\sigma_{\min}^2(\boldsymbol{A})$. The authors also prove convergence for inconsistent systems, with the same convergence rate and convergence radius which depends also on the minimum of all $\lambda_{\min}(\boldsymbol{A}_\tau \boldsymbol{A}_\tau^*)$, see [NT14] for details.

Surprisingly, every standardized matrix admits a good row paving. The following result is due to [Ver06, Tro09] which builds off the foundational work of [BT87, BT91].

Proposition 1.2 (Existence of Good Row Pavings)**.** *For any $\delta \in (0,1)$ and standardized $n \times d$ matrix $\boldsymbol{A}$, there is a row paving satisfying*

$$m \leq \mathrm{C} \cdot \delta^{-2} \|\boldsymbol{A}\|^2 \log(1 + n) \quad and \quad 1 - \delta \leq \beta \leq 1 + \delta.$$

where C *is an absolute constant.*

Although this is an existential result, there are constructive methods to obtain such pavings, and for certain classes of matrices, they can even be obtained by a random partitioning of the rows [Tro08, CD12, NT14].

With such a paving in tow, the convergence of (1.10) becomes

$$\mathbb{E}\|\boldsymbol{x}_j - \boldsymbol{x}_\star\|^2 \leq \left[1 - \frac{1}{\mathrm{C}\kappa^2(\boldsymbol{A})\log(1 + n)}\right]^j \|\boldsymbol{x}_0 - \boldsymbol{x}_\star\|_2^2$$

Although often comparable to the convergence rate for the simple method (1.5), numerical results confirm that the block method offers significant reduction in computation time due to the speed of matrix–vector multiplication (see e.g. [NT14]).

1.4. **Contribution.** This paper analyzes the system with matrix described in (1.2) using an algorithm with the block Kaczmarz approach for the equalities given by $\boldsymbol{A}_=$ and the simple method for the inequalities given by $\boldsymbol{A}_\leq$. A paving is created for $\boldsymbol{A}_=$, with the inequalities excluded. At each iteration, we select from $\boldsymbol{A}_=$ with a fixed probability p and from $\boldsymbol{A}_\leq$ with probability $1 - p$. In the former case, we select a block τ from paving T uniformly at random, and in the latter case we select a row i of $\boldsymbol{A}_\leq$ uniformly at random. In the case of a block of equalities being selected, the algorithm proceeds by updating $\boldsymbol{x}_j$ using (1.9). When an inequality row is selected, $\boldsymbol{x}_j$ is updated using the rule (1.6). We prove that this method yields linear convergence to the solution set S. We also include a discussion about paving both $\boldsymbol{A}_=$ and $\boldsymbol{A}_\leq$, which identifies a geometric property of the system which allows for improved convergence by utilizing two pavings. We show that when this property is not satisfied, utilizing both pavings can be detrimental to convergence.

1.5. **Organization.** Section 2 lays out our main result, Theorem 2.1, and provides a proof. We discuss blocking the full matrix in Section 3 and Section 4 explains numerical experiments and results. We conclude with discussion and related work in Section 5.

2. ANALYSIS OF THE BLOCK KACZMARZ ALGORITHM FOR A SYSTEM OF INEQUALITIES

In this section we analyze the convergence of the described method, which is detailed in Algorithm 2.1.

Notice that the probability of selecting a block of $A_=$ is $\frac{\beta m}{n_i + \beta m}$. This quantity corresponds to the relative size of $A_=$ in the system, where the size is measured in terms of the paving quantities βm. This value may be difficult to compute precisely, and the simpler threshold of n_e/n appears to also work well in practice. We provide no evidence that our selection of this threshold is most efficient, nor any more efficient than using one proportional to the number of equality rows n_e. We find that this algorithm yields linear convergence in expectation with a rate that only depends on the number of inequalities n_i, paving size m, and upper bound β.

Our main result is described in Theorem 2.1.

Theorem 2.1 (Convergence). *Let $A \in \mathbb{R}^{n \times d}$ and $b \in \mathbb{R}^n$ correspond to a system as in* (1.2) *with the first n_e rows being equalities and the remaining $n_i = n - n_e$ rows being inequalities. Let T be an (m, β) row paving of $A_=$. Let x_0 be an arbitrary initial estimate and S the non-empty feasible region. Then Algorithm 2.1 satisfies for each iteration $j = 1,2,3,...$,*

$$\mathbb{E}\left[d(x_j, S)^2\right] \leq \left(1 - \frac{1}{L^2(n_i + \beta m)}\right)^j \cdot d(x_0, S)^2,$$

where L is the Hoffman constant (1.7).

Remarks.
1. Note that when there are no block projections, no inequalities, or neither, Theorem 2.1 recovers the results of the standard randomized Kaczmarz for inequalities [LL10], the standard randomized block Kaczmarz method [NT14] or the standard randomized Kaczmarz method [SV09], respectively. We thus view this result as a completely generalized convergence bound.

Algorithm 2.1 Block Kaczmarz Method for a System of Inequalities

---

Input:
- Matrix A with dimension $n \times d$
- Right-hand side b with dimension n
- Number of rows representing equalities, n_e, and inequalities, $n_i = n - n_e$
- Partition $T = \{\tau_1, \ldots, \tau_m\}$ of the row indices $\{1, \ldots, n_e\}$
- Initial iterate x_0 with dimension d
- Convergence tolerance $\varepsilon > 0$

Output: An estimate $\hat{x}$ to the solution of the system (1.3)

$j \leftarrow 0$
repeat
    $j \leftarrow j + 1$
    Draw uniformly at random q from $[0, 1]$
    if $q \leq \frac{\beta m}{n_i + \beta m}$
        Choose a block τ uniformly at random from T
        $x_j \leftarrow x_{j-1} + (A_\tau)^\dagger(b_\tau - A_\tau x_{j-1})$ (Solve least-squares approximation)
    else
        Choose a row i uniformly at random from $\{n_e + 1, \ldots, n\}$
        $x_{j+1} \leftarrow x_j - \frac{e(a_i^T x_j - b_i)}{\|a_i\|_2^2} a_i$
until $\|e(Ax_j - b)\|_2^2 \leq \varepsilon^2$
$\hat{x} \leftarrow x_j$

2. If we let ρ_s and ρ_b be the convergence rates of the simple and block methods for mixed systems, respectively, then by (1.8) and Theorem 2.1,

$$\rho_s \geq \frac{1}{L^2 n} \quad \text{and} \quad \rho_b \geq \frac{1}{L^2(n_i + \beta m)}.$$

It is evident that our expected convergence rate will be faster per iteration than the simple method when $n_i + \beta m < n$. Since β can be chosen close to 1 and $m < n_e$ is then number of rows in $A_=$, this holds quite easily.

3. Since a single iteration using a block A_τ in general may cost more than an iteration utilizing a single row, it is more fair to compare per epoch, rather than per iteration. An epoch is typically the minimum number of iterations needed to visit each row of the matrix. When there are inequalities present that are already satisfied in a given iteration, that iteration may make no contribution and cost very little computationally. Thus the notion of epoch may be slightly skewed here, but if we ignore this subtlety the simple method will have approximately n iterations per epoch, compared to $n_i + m$ iterations per epoch with the block method. The approximate per epoch convergence rates can thus be compared as

$$n \cdot \rho_s \geq \frac{1}{L^2} \quad \text{and} \quad (n_i + m) \cdot \rho_b \geq \frac{n_i + m}{L^2(n_i + \beta m)}.$$

This result is similar to that found by Needell and Tropp [NT14], with the block convergence rate at best equal to that of the simple convergence rate when $\beta = 1$. However, as already noted, the block method is quite advantageous computationally.

Combining the paving result of Prop. 1.2 with Theorem 2.1 yields the following corollary.

Corollary 2.2. *Instate the assumptions and notation of Theorem 2.1 and let $A_=$ be equipped with an (m, β) row-paving as in Proposition 1.2. Then the iterates of Algorithm 2.1 satisfy*

$$\mathbb{E}\left[d(\boldsymbol{x}_j, S)^2\right] \leq \left(1 - \frac{1}{L^2(n_i + C\|A_=\|^2 \log(1+n))}\right)^j \cdot d(\boldsymbol{x}_0, S)^2,$$

for some constant C.

Proof of Theorem 2.1. Fix an iteration j of Algorithm 2.1. We proceed as in [NT14] and [LL10]. First, we suppose that $q \leq \frac{\beta m}{n_i + \beta m}$, so that a block τ of equalities is selected this iteration. Then writing P_S as the orthogonal projection onto S, we have $\boldsymbol{b}_\tau = A_\tau P_S \boldsymbol{x}_{j-1}$ since $P_S \boldsymbol{x}_{j-1} \in S$. We then have

$$\begin{aligned}
\boldsymbol{x}_j &= \boldsymbol{x}_{j-1} + A_\tau^\dagger(\boldsymbol{b}_\tau - A_\tau \boldsymbol{x}_{j-1}) \\
&= \boldsymbol{x}_{j-1} + A_\tau^\dagger(A_\tau P_S \boldsymbol{x}_{j-1} - A_\tau \boldsymbol{x}_{j-1}) \\
&= \boldsymbol{x}_{j-1} + A_\tau^\dagger A_\tau(P_S \boldsymbol{x}_{j-1} - \boldsymbol{x}_{j-1}).
\end{aligned}$$

Thus,

$$\begin{aligned}
\|\boldsymbol{x}_j - P_S \boldsymbol{x}_{j-1}\|_2^2 &= \|\boldsymbol{x}_{j-1} - P_S \boldsymbol{x}_{j-1} - A_\tau^\dagger A_\tau(\boldsymbol{x}_{j-1} - P_S \boldsymbol{x}_{j-1})\|_2^2 \\
&= \|(\mathbf{I} - A_\tau^\dagger A_\tau)(\boldsymbol{x}_{j-1} - P_S \boldsymbol{x}_{j-1})\|_2^2.
\end{aligned}$$

Taking expectation (over the choice of i, conditioned on previous choices), and using the fact that $A_\tau^\dagger A_\tau$ is an orthogonal projector, along with the properties of the paving yields

$$\begin{aligned}
\mathbb{E}\|\boldsymbol{x}_j - P_S \boldsymbol{x}_{j-1}\|_2^2 &\leq \mathbb{E}\|(I - A_\tau^\dagger A_\tau)(\boldsymbol{x}_{j-1} - P_S \boldsymbol{x}_{j-1})\|_2^2 \\
&= \|\boldsymbol{x}_{j-1} - P_S \boldsymbol{x}_{j-1}\|_2^2 - \mathbb{E}\|A_\tau^\dagger A_\tau(\boldsymbol{x}_{j-1} - P_S \boldsymbol{x}_{j-1})\|_2^2 \\
&\leq \|\boldsymbol{x}_{j-1} - P_S \boldsymbol{x}_{j-1}\|_2^2 - \frac{1}{\beta}\mathbb{E}\|A_\tau(\boldsymbol{x}_{j-1} - P_S \boldsymbol{x}_{j-1})\|_2^2.
\end{aligned}$$

Since $d(\boldsymbol{x}_{j-1}, S) = \|\boldsymbol{x}_{j-1} - P_S \boldsymbol{x}_{j-1}\|_2$ and $d(\boldsymbol{x}_j, S) \le \|\boldsymbol{x}_j - P_S \boldsymbol{x}_{j-1}\|_2$, this means that

$$
\begin{aligned}
\mathbb{E}\left[d(\boldsymbol{x}_j, S)^2\right] &\le d(\boldsymbol{x}_{j-1}, S)^2 - \frac{1}{\beta}\mathbb{E}\|A_{\boldsymbol{\tau}}(\boldsymbol{x}_{j-1} - P_S \boldsymbol{x}_{j-1})\|_2^2 \\
&= d(\boldsymbol{x}_{j-1}, S)^2 - \frac{1}{\beta m}\sum_{\tau \in T}\|A_\tau \boldsymbol{x}_{j-1} - \boldsymbol{b}_\tau\|_2^2 \\
&= d(\boldsymbol{x}_{j-1}, S)^2 - \frac{1}{\beta m}\sum_{i \in I_=}e(A_= \boldsymbol{x}_{j-1} - \boldsymbol{b}_=)_i^2.
\end{aligned}
\tag{2.1}
$$

Next suppose that instead $i \in I_\le$ is selected. Then since each row $\boldsymbol{a_i}$ has unit norm,

$$
\begin{aligned}
d(\boldsymbol{x}_j, S)^2 &\le \|\boldsymbol{x}_j - P_S \boldsymbol{x}_{j-1}\|_2^2 \\
&= \|\boldsymbol{x}_{j-1} - e(A\boldsymbol{x}_{j-1} - b)_i \boldsymbol{a_i} - P_S \boldsymbol{x}_{j-1}\|_2^2 \\
&= \|\boldsymbol{x}_{j-1} - P_S \boldsymbol{x}_{j-1}\|_2^2 + e(A\boldsymbol{x}_{j-1} - b)_i^2 - 2e(A\boldsymbol{x}_{j-1} - b)_i \langle \boldsymbol{a_i}, \boldsymbol{x}_{j-1} - P_S \boldsymbol{x}_{j-1} \rangle \\
&\le d(\boldsymbol{x}_{j-1}, S)^2 - e(A\boldsymbol{x}_{j-1} - b)_i^2,
\end{aligned}
$$

where the last line follows from the fact that $\langle \boldsymbol{a_i}, P_S \boldsymbol{x}_{j-1} \rangle \le b_i$ and $e(A\boldsymbol{x}_{j-1} - b)_i \ge 0$. Now taking expectation again we have

$$
\begin{aligned}
\mathbb{E}\left[d(\boldsymbol{x}_j, S)^2\right] &\le d(\boldsymbol{x}_{j-1}, S)^2 - \mathbb{E}(e(A\boldsymbol{x}_{j-1} - \boldsymbol{b})_i^2) \\
&= d(\boldsymbol{x}_{j-1}, S)^2 - \frac{1}{n_i}\sum_{i \in I_\le}e(A_\le \boldsymbol{x}_{j-1} - \boldsymbol{b}_\le)_i^2.
\end{aligned}
$$

Combining these results and letting $E_=$ and $E_\le$ denote the events that a block from T and a row from $I_\le$ is selected, respectively, we have

$$
\begin{aligned}
\mathbb{E}\left[d(\boldsymbol{x}_j, S)^2\right] &= p \cdot \mathbb{E}[d(\boldsymbol{x}_j, S)^2 | E_=] + (1 - p) \cdot \mathbb{E}[d(\boldsymbol{x}_j, S)^2 | E_\le] \\
&\le p\left[d(\boldsymbol{x}_{j-1}, S)^2 - \frac{1}{\beta m}\sum_{i \in I_=}e(A_= \boldsymbol{x}_{j-1} - \boldsymbol{b}_=)_i^2\right] + (1 - p)\left[d(\boldsymbol{x}_{j-1}, S)^2 - \frac{1}{n_i}\sum_{i \in I_\le}e(A_\le \boldsymbol{x}_{j-1} - \boldsymbol{b}_\le)_i^2\right] \\
&= d(\boldsymbol{x}_{j-1}, S)^2 - p \cdot \frac{1}{\beta m}\sum_{i \in I_=}e(A_= \boldsymbol{x}_{j-1} - \boldsymbol{b}_=)_i^2 - (1 - p) \cdot \frac{1}{n_i}\sum_{i \in I_\le}e(A_\le \boldsymbol{x}_{j-1} - \boldsymbol{b}_\le)_i^2.
\end{aligned}
$$

Since $p = \frac{\beta m}{n_i + \beta m}$, we have $\frac{1-p}{n_i} = \frac{1}{n_i + \beta m}$ and we can simplify

$$
\begin{aligned}
\mathbb{E}\left[d(\boldsymbol{x}_j, S)^2\right] &\le d(\boldsymbol{x}_{j-1}, S)^2 - \frac{1}{n_i + \beta m}\left[\sum_{i \in I_=}e(A_= \boldsymbol{x}_{j-1} - \boldsymbol{b}_=)_i^2 + \sum_{i \in I_\le}e(A_\le \boldsymbol{x}_{j-1} - \boldsymbol{b}_\le)_i^2\right] \\
&= d(\boldsymbol{x}_{j-1}, S)^2 - \frac{1}{n_i + \beta m}\|e(A\boldsymbol{x}_{j-1} - \boldsymbol{b})\|_2^2 \\
&\le d(\boldsymbol{x}_{j-1}, S)^2 - \frac{1}{L^2(n_i + \beta m)} \cdot d(\boldsymbol{x}_{j-1}, S)^2 \\
&= \left[1 - \frac{1}{L^2(n_i + \beta m)}\right]d(\boldsymbol{x}_{j-1}, S)^2,
\end{aligned}
$$

where we have utilized the Hoffman bound (1.7) in the second inequality.

Utilizing independence of the random selections and recursing on this relation yields the desired result. $\square$

3. A Discussion about Blocking Inequalities

It is natural to ask whether one can benefit by blocking both the equalities as above and also the inequalities, as described by Algorithm 3.2. Indeed, Section 4 will show dramatic improvements in computational time when the rows of $A_=$ are paved and block projections as in Algorithm 2.1 are used. So can one benefit even more by paving also the rows of $A_\leq$? The answer to this question heavily depends on the structure of the matrix A.

Algorithm 3.2 Double Block Kaczmarz Method for a System of Inequalities

---

Input:
- Matrix A with dimension $n \times d$
- Right-hand side b with dimension n
- Partition $T' = \{\tau'_1, \dots, \tau'_{m'}\}$ of the row indices $\{1, \dots, n_i\}$
- Partition $T = \{\tau_1, \dots, \tau_m\}$ of the row indices $\{1, \dots, n_e\}$
- Initial iterate x_0 with dimension d
- Convergence tolerance $\varepsilon > 0$

Output: An estimate $\hat{x}$ to the solution of the system (1.3)

$j \leftarrow 0$
repeat
    $j \leftarrow j + 1$
    Draw uniformly at random q from $[0, 1]$
    if $q \leq \frac{\beta m}{\beta' m' + \beta m}$
        Choose a block τ uniformly at random from T
        $x_j \leftarrow x_{j-1} + (A_\tau)^\dagger (b_\tau - A_\tau x_{j-1})$ (Solve least-squares approximation)
    else
        Choose a block τ' uniformly at random from T'
        Set $\sigma = \{i \in \tau' : \langle a_i, x_{j-1} \rangle > b_i\} \subset \tau'$ (Select unsatisfied subset)
 $x_j \leftarrow x_{j-1} + (A_\sigma)^\dagger (b_\sigma - A_\sigma x_{j-1})$ (Solve least-squares approximation)
until $\|e(Ax_j - b)\|_2^2 \leq \varepsilon^2$
$\hat{x} \leftarrow x_j$

If we only consider $A_=$, a block projection as in (1.9) enforces all the equations indexed by τ to be satisfied. This is of course desirable when the rows indexed by τ correspond to equalities. Also, if a single inequality corresponding to row i in $A_\leq$ is not satisfied and we perform a single projection as in (1.4), we are again enforcing that inequality to hold with equality. However, this improves the estimation since in this case we know the solution set S lies on the opposite side of the hyperplane $\{x : \langle x, a_i \rangle = b_i\}$ as the current estimation (see Figure 1 (a)). On the other hand, if we employ a block projection as in (1.9) to a set of inequalities indexed by τ which are not satisfied by the current estimation x_{j-1} then we enforce *all of them to hold with equality simultaneously*. Depending on the geometry of the involved rows, this may result in an improved estimation or actually one much farther from the solution set. Of course, one might alternatively want solve the convex program to project onto the intersection of the corresponding half-spaces, but we would like to maintain the efficiency and simplicity of the block Kaczmarz method.

As an illustrative example, Figure 1 (b) and (c) demonstrate two possible scenarios in two dimensions. Here, the solution space is a single point marked S, and we draw two hyperplanes H_{i_1} and H_{i_2} where $H_i = \{x : \langle a_i, x \rangle = b_i\}$. The yellow shaded regions denote areas where both inequalities hold true: $\{x : \langle a_{i_1}, x \rangle \leq b_{i_1}$ and $\langle a_{i_2}, x \rangle \leq b_{i_2}\}$. Notice that in (b), when the angle between $x_{j-1} - x_j$ and $s - x_j$ is obtuse,
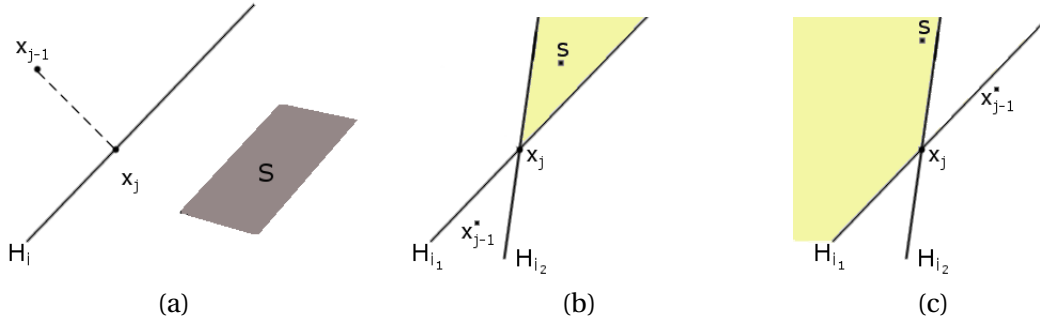
Figure 1 Possible geometries of the system. S denotes solution space (or solution point). Yellow shading denotes regions where inequalities i_1 and i_2 are both satisfied. (a) A single projection onto hyperplane $H_i = \{x : \langle a_i, x \rangle = b_i\}$ provides improved estimation. (b) Block projection onto intersection of hyperplanes also may provide improved estimation. (c) Block projection onto intersection of hyperplanes may provide improved estimation.

the orthogonal projection of estimation x_{j-1} onto their intersection is guaranteed to be closer to the solution set S. On the other hand, when that angle is acute we see exactly the opposite, as in (c). We can quantify this notion by the following definition.

Definition 3.1. For an $r \times d$ matrix A and $b \in \mathbb{R}^r$, for row i denote by $\tilde{H}_i$ and H_i the half-space $\tilde{H}_i = \{\langle a_i, x \rangle \leq b_i\}$ and hyperplane $H_i = \{\langle a_i, x \rangle = b_i\}$, respectively, and write P_S as the orthogonal projection onto a convex set S. An *obtuse* (m, β) *row paving* of the matrix A is an (m, β) row paving $T = \{\tau_1, \ldots, \tau_m\}$ that also satisfies the following. Let $\tau \in T$ and let $s \in \cap_{i \in \tau} \tilde{H}_i$, $w \in \cap_{i \in \tau} \tilde{H}_i^c$, and $z = P_{\cap_{i \in \tau} H_i} w$. Then

$$\langle w - z, s \rangle < 0.$$

In other words, the angle between $w - z$ and s (and thus $s - z$) is obtuse.

We will see that performing block projections on the inequalities in the system only makes sense when one can obtain an obtuse row paving. We will use $w = x_{j-1}$, $z = x_j$, and $s \in S$. Notice that if $i_1, i_2 \in \tau \in T$, then the partition used in the system depicted in Figure 1 (c) does not constitute an obtuse row paving.

We conduct two simple experiments to demonstrate the different behavior of the algorithm. In all cases the matrix A is a 300×100 matrix with standard normal entries, 100 rows correspond to inequalities, and b is generated so that the solution set S is non-empty. We measure the residual error which we define as $\|e(Ax_j - b)\|_2$. Figure 2 (a) shows the behavior of the block method with this matrix and a row paving obtained via a random row partition of 30 blocks (10 rows per block). This generation will create a matrix with paving that with very high probability is not an obtuse row paving. As Figure 2 demonstrates, the block method diverges in this case. However, as Figure 2 (c) shows, the simple Kaczmarz method succeeds in identifying a point in the solution space. Next, we create a matrix in the exact same way, and create the same random row paving. Then, however, we iterate through every block in the paving corresponding to inequalities and if two rows i and k in a block satisfy $\langle a_i, a_k \rangle > 0$, we replace row a_i with $-a_i$ and entry b_i with $-b_i$. This guarantees every block in the paving yields a geometry like that shown in Figure 1 (b), and gives an obtuse row paving. Note that of course this changes the solution space as well so one cannot employ this strategy in general. We then add positive values to the entries in b corresponding to inequalities to ensure the solution set S is non-empty. With this new system and paving, we again run the block method and see that the method now converges to a point in the solution set, as seen in Figure 2 (b).

With this definition we obtain the following result, whose proof can be found in the appendix.
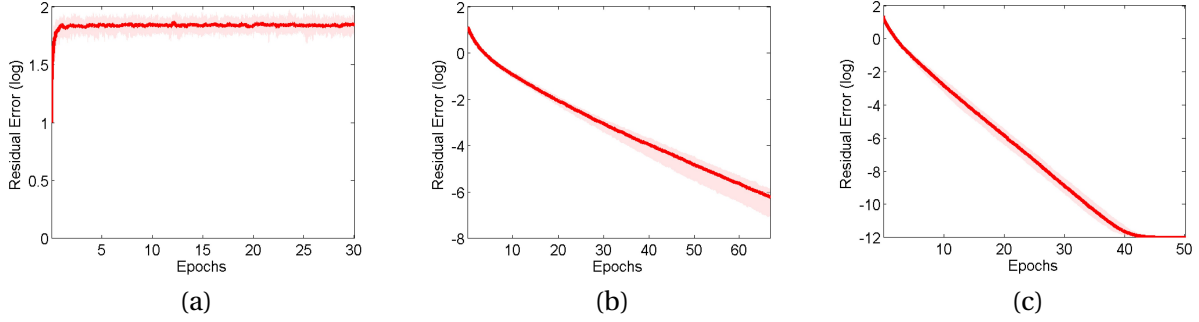
Figure 2 Residual error of the Kaczmarz Method per epoch: a) Median residual error of block method over 40 trials for matrix A not having an obtuse row paving, b) Median residual error of block method over 40 trials for matrix A using an obtuse row paving, c) Median residual error of simple method over 40 trials for same matrix as in a). Shaded region spans across minimum and maximum values over all trials and solid line denotes median value.

Theorem 3.2. *Let A satisfy the assumptions of Theorem 2.1 and in addition have an obtuse (m', β') row paving of $A_\leq$. Let $x_1, \ldots$ denote the iterates of Algorithm 3.2. Then using the notation of Theorem 2.1,*

$$\mathbb{E}[d(\boldsymbol{x}_j, S)^2] \leq \left[1 - \frac{1}{L^2(\beta' m' + \beta m)}\right]^j d(\boldsymbol{x}_0, S)^2.$$

Note that row pavings of standardized matrices can be obtained readily, often by random partitions [Tro11, Tro12, NT14], whereas obtuse row pavings may be much more challenging to obtain in general. Of course, by default the trivial paving which assigns each set τ to a single row always admits an obtuse row paving. We focus on Algorithm 2.1 which paves only $A_=$, and leave further analysis of Algorithm 3.2 and constructions of obtuse row pavings for future work.

4. EXPERIMENTS

We use MATLAB to run some experiments using random matrices to test the convergence of the block Kaczmarz method applied to a system of equalities and inequalities. In each experiment, we create a random 500 by 100 matrix A where each element is an independent standard normal random variable. Each entry is then divided by the norm of its row so that the matrix is standardized. The first 400 rows of matrix A compose $A_=$, and the remaining 100 rows are set as inequalities of $A_\leq$ in the method described by (1.3). The experiments are run using the following procedure. For each of 100 trials,

(1) Create matrix A in the manner described above.
(2) Create $\boldsymbol{x}_\star$ where each entry is selected independently from a standard normal distribution. Set $\boldsymbol{b} = A\boldsymbol{x}_\star$.
(3) Pave submatrix $A_=$ into 16 blocks with 25 equalities per block by a random partitioning of the rows.
(4) Set initial approximations $\boldsymbol{x}_0^{\text{block}} = \boldsymbol{x}_0^{\text{simp}} = A^* \boldsymbol{b}$.
(5) Draw q uniformly at random from $[0, 1]$.
  (a) If $q \leq \frac{n_e}{n}$, choose block $\{1, \ldots, m\}$ uniformly at random and update iterate $\boldsymbol{x}_j^{\text{block}}$ using (1.9). (Note that the threshold $\frac{n_e}{n}$ is different than that given in the main algorithm and theorem, but it is easier to calculate and seems to work fine in practice.)
  (b) Else, choose a row uniformly at random from $\{401, \ldots, 500\}$ and update iterate $\boldsymbol{x}_j^{\text{block}}$ using (1.6).

(c) Update iterate $\boldsymbol{x}_j^{\text{simp}}$ using (1.6).

For both the simple and block algorithms, the median, minimum, and maximum values of the residual $\|e(\boldsymbol{Ax}_j - \boldsymbol{b})\|_2^2$ of the 100 trials are recorded for each iteration j.

Figure 3 compares the performance of the block Kaczmarz method used in this paper and the standard Kaczmarz method described by Leventhal and Lewis [LL10]. The plot in Figure 3 (a) compares convergence per iteration. As the block Kaczmarz method enforces multiple equalities per iteration, it is unsurprising that it performs better in this experiment. Figure 3 (b) displays the convergence of the two methods per epoch. The block Kaczmarz algorithm has an epoch of $m + n_i$ iterations, and the standard Kaczmarz method has an epoch of size n. Here, to be fair we only count an iteration towards an epoch if the estimated solution $\boldsymbol{x}_j \neq \boldsymbol{x}_{j-1}$. Thus in the case where a chosen inequality is already satisfied for iteration j, this iteration does not count towards an epoch since no computation is being performed. We noticed, however, that whether or not we modified the count in this way, the behavior still produces results very similar to Figure 3. Once again the experiments yielded faster convergence with the block Kaczmarz approach. It is interesting to compare the results of Figure 3 (b) and those of Figure 2 (b) and (c). The per-epoch convergence of the methods and whether the block or standard appears faster varies slightly and depends on both the number of rows and columns. In general, the per-epoch convergence rates are reasonably comparable, as the analysis suggests. However, Figure 3 (c) compares the rate of convergence of the two algorithms by plotting the residual against the CPU time expended in the simulation. The ability to utilize efficient matrix–vector multiplication gives the method significantly improved convergence per second relative to the standard Kaczmarz algorithm.

(a) (b) (c)
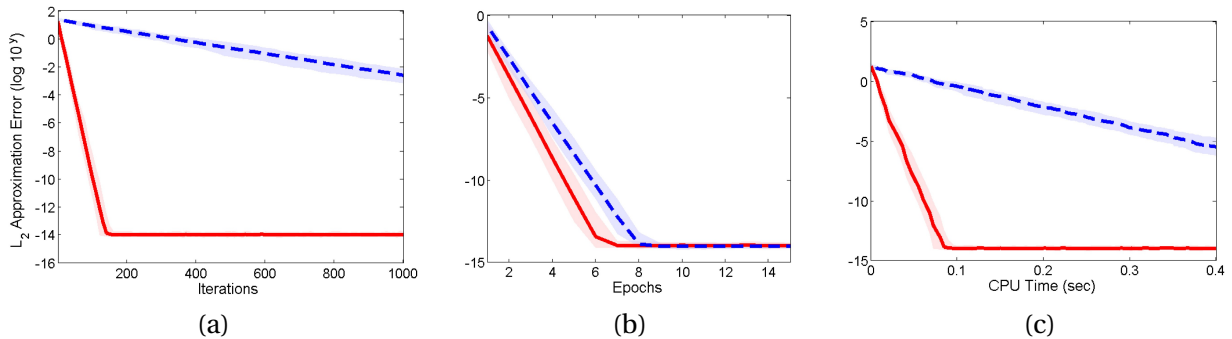
Figure 3 Residual error of the Block Kaczmarz Method (solid red) vs. Simple Kaczmarz Method (dashed blue) as a function of (a) Iterations, (b) Epochs, (c) CPU time. Shaded region spans from minimum to maximum value over 100 trials; lines denote the median value.

5. Conclusion and Related Work

The Kaczmarz algorithm was first proposed in [Kac37]. Kaczmarz demonstrated that the method converged to the solution of linear system $\boldsymbol{Ax} = \boldsymbol{b}$ for square, non-singular matrix $\boldsymbol{A}$. Since then, the method has been utilized in the context of computer tomography as the *Algebraic Reconstruction Technique* (ART) [GBH70, Byr08, Nat01, Her09]. Empirical results suggested that randomized selection offered improved convergence over the cyclic scheme [HS78, HM93]. Strohmer and Vershynin [SV09] were the first to prove an expected linear convergence rate using a randomized Kaczmarz algorithm with specific random control. This result was extended by Needell [Nee10] to apply to inconsistent systems, which shows a linear convergence rate to within a fixed radius around the least-squares solution. Almost-sure convergence guarantees were recently proved by Chen and Powell [CP12]. Zouzias and Freris [ZF13] analyze a modified version of the method in the inconsistent case, using a variant motivated by Popa [Pop99] to

reduce the residual and thereby converge to the least squares solution. Relaxation parameters can also be introduced to obtain convergence to the least squares solution, see e.g. [WM67, CEG83, Tan71, HN90], and partially weighted sampling can lead to a tradeoff between convergence rate and radius [NSW13]. Liu, Wright, and Sridhar [LWS14] discuss applying a parallelized variant of the randomized Kaczmarz method, demonstrating that the convergence rate can be increased almost linearly by bounding the number of processors by a multiple of the number of rows of A.

The block Kaczmarz updating method was introduced by Elfving [Elf80] as a special case of the more general framework by Eggermont et.al. [EHL81]. The notion of using blocking in projection methods is certainly not new, and there is a large amount of literature on these types of methods, see e.g. [XZ02, Byr08] and references therein. Needell and Tropp [NT14] provide the first analysis showing an expected linear convergence rate which depends on the properties of the matrix A and of the submatrices A_τ resulting from the paving, connecting pavings and the block Kaczmarz scheme. The use of specialized blocks appears elsewhere, in particular, the works of Popa use blocks with orthogonal rows that are beneficial for the block Kaczmarz method [Pop99, Pop01, Pop04]. Needell, Zhao, and Zouzias [NZZ14] expand on the results from [NT14] and [ZF13] to demonstrate convergence to the least-squares solution for an inconsistent system using the block Kaczmarz method. Again the block approach can yield faster convergence than the simple method.

The Kaczmarz method was first applied to a system of equalities and inequalities by Leventhal and Lewis [LL10], who also consider polynomial constraints with the method. They give a linear convergence rate to the feasible solution space S, using $\|A\|_F^2$ and the Hoffman constant [Hof52]. We apply the block Kaczmarz scheme to the system described in [LL10], combining their result with that of Needell and Tropp [NT14] to acquire a completely generalized result. We highlight several important complications which arise when attempting to apply the block scheme to inequalities. Nonetheless, whether a paving is used only partially or for the complete system, significant reduction in computational time can be achieved.

5.1. **Future Work.** There are many interesting open problems related to the block Kazcmarz method and linear systems with inequalities. It has been well observed in the literature that selecting rows (or blocks) *without replacement* rather than with replacement as in the theoretical results leads to faster a convergence rate empirically [RR12, NT14]. When selecting without replacement, independence between iterations vanishes, making a theoretical analysis more challenging. Secondly, it would be interesting to further investigate the use of obtuse row pavings. In systems with a large number of inequalities, the ability to pave the submatrix $A_\le$ with an obtuse row paving would lead to significantly faster convergence. In that case, one may like to identify a more general geometric property about the system that permits such pavings or an alternative formulation that offers convergence of the full block method.

APPENDIX A. PROOF OF THEOREM 3.2

Proof. Fix an iteration j of Algorithm 3.2. As in the proof of Theorem 2.1, if a block of equalities is selected this iteration, then we again have (2.1). So we next instead consider the case when a block of inequalities is selected, and call this block τ', and its pruned subset σ. Set $s = P_S x_{j-1}$, where again P_S denotes the orthogonal projection onto the solution set S. If we write $\tilde{H}_i = \{x : \langle a_i, x \rangle \le b_i\}$ and $H_i = \{x : \langle a_i, x \rangle = b_i\}$, then by their definitions we have

$$s \in \cap_{i \in \sigma} \tilde{H}_i, \quad x_{j-1} \in \cap_{i \in \sigma} \tilde{H}_i^c, \quad \text{and} \quad x_j = P_{\cap_{i \in \sigma} H_i} x_{j-1}.$$

Then since σ is part of an obtuse paving, the angle between $x_j - x_{j-1}$ and $s - x_{j-1}$ must be obtuse. There thus exists a point t on the line segment $L = \{\gamma x_{j-1} + (1 - \gamma)s : 0 \le \gamma \le 1\}$ such that $x_{j-1} - x_j$ and $t - x_j$ are orthogonal (see Figure 4).
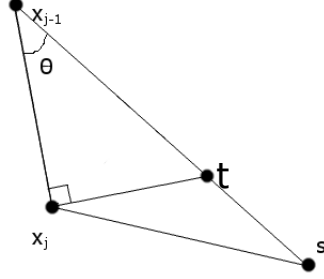
Figure 4 Geometry of system.

Now since $\boldsymbol{t} \in L$, we have $\|\boldsymbol{t} - \boldsymbol{x}_{j-1}\|_2 \le \|\boldsymbol{x}_{j-1} - \boldsymbol{s}\|_2$, and thus letting θ denote the angle between $\boldsymbol{x}_j - \boldsymbol{x}_{j-1}$ and $\boldsymbol{t} - \boldsymbol{x}_{j-1}$ (see Figure 4), we have

$$
\begin{aligned}
\|\boldsymbol{x}_j - \boldsymbol{x}_{j-1}\|_2 &\le \|\boldsymbol{x}_{j-1} - \boldsymbol{s}\|_2 \cdot \frac{\|\boldsymbol{x}_j - \boldsymbol{x}_{j-1}\|_2}{\|\boldsymbol{t} - \boldsymbol{x}_{j-1}\|_2} \\
&= \|\boldsymbol{x}_{j-1} - \boldsymbol{s}\|_2 \cdot \cos\theta \\
&= \frac{\|\boldsymbol{x}_j - \boldsymbol{x}_{j-1}\|_2 \cdot \|\boldsymbol{s} - \boldsymbol{x}_{j-1}\|_2 \cdot \cos\theta}{\|\boldsymbol{x}_j - \boldsymbol{x}_{j-1}\|_2} \\
&= \frac{\langle \boldsymbol{s} - \boldsymbol{x}_{j-1}, \boldsymbol{x}_j - \boldsymbol{x}_{j-1} \rangle}{\|\boldsymbol{x}_j - \boldsymbol{x}_{j-1}\|_2} \\
&= \frac{-\langle \boldsymbol{x}_{j-1} - \boldsymbol{s}, \boldsymbol{x}_j - \boldsymbol{x}_{j-1} \rangle}{\|\boldsymbol{x}_j - \boldsymbol{x}_{j-1}\|_2}.
\end{aligned}
$$

Thus, we have that

$$
\langle \boldsymbol{x}_{j-1} - \boldsymbol{s}, \boldsymbol{x}_j - \boldsymbol{x}_{j-1} \rangle \le -\|\boldsymbol{x}_j - \boldsymbol{x}_{j-1}\|_2^2.
$$

By the definition of $\boldsymbol{x}_j$, this means that

$$
\langle \boldsymbol{x}_{j-1} - \boldsymbol{s}, A_\sigma^\dagger(\boldsymbol{b}_\sigma - A_\sigma \boldsymbol{x}_{j-1}) \rangle \le -\|A_\sigma^\dagger(\boldsymbol{b}_\sigma - A_\sigma \boldsymbol{x}_{j-1})\|_2^2. \tag{A.1}
$$

Using this along with the paving properties we see that

$$
\begin{aligned}
\|\boldsymbol{x}_j - \boldsymbol{s}\|_2^2 &= \|\boldsymbol{x}_{j-1} - \boldsymbol{s} + A_\sigma^\dagger(\boldsymbol{b}_\sigma - A_\sigma \boldsymbol{x}_{j-1})\|_2^2 \\
&= \|\boldsymbol{x}_{j-1} - \boldsymbol{s}\|_2^2 + 2\langle \boldsymbol{x}_{j-1} - \boldsymbol{s}, A_\sigma^\dagger(\boldsymbol{b}_\sigma - A_\sigma \boldsymbol{x}_{j-1}) \rangle + \|A_\sigma^\dagger(\boldsymbol{b}_\sigma - A_\sigma \boldsymbol{x}_{j-1})\|_2^2 \\
&\le \|\boldsymbol{x}_{j-1} - \boldsymbol{s}\|_2^2 - \|A_\sigma^\dagger(\boldsymbol{b}_\sigma - A_\sigma \boldsymbol{x}_{j-1})\|_2^2 \\
&= d(\boldsymbol{x}_{j-1}, S)^2 - \|A_\sigma^\dagger(\boldsymbol{b}_\sigma - A_\sigma \boldsymbol{x}_{j-1})\|_2^2 \\
&\le d(\boldsymbol{x}_{j-1}, S)^2 - \frac{1}{\beta'}\|\boldsymbol{b}_\sigma - A_\sigma \boldsymbol{x}_{j-1}\|_2^2.
\end{aligned}
$$

Thus, taking expectation (over the choice of τ', conditioned on previous choices), yields

$$\mathbb{E}[d(\boldsymbol{x}_j, S)^2] \leq \mathbb{E}\|\boldsymbol{x}_j - \boldsymbol{s}\|_2^2$$

$$\leq d(\boldsymbol{x}_{j-1}, S)^2 - \frac{1}{\beta'}\mathbb{E}\|\boldsymbol{b}_\sigma - \boldsymbol{A}_\sigma \boldsymbol{x}_{j-1}\|_2^2$$

$$= d(\boldsymbol{x}_{j-1}, S)^2 - \frac{1}{\beta'}\mathbb{E}\|e(\boldsymbol{b}_{\tau'} - \boldsymbol{A}_{\tau'}\boldsymbol{x}_{j-1})\|_2^2$$

$$= d(\boldsymbol{x}_{j-1}, S)^2 - \frac{1}{m'\beta'}\sum_{\tau' \in T'}\|e(\boldsymbol{b}_{\tau'} - \boldsymbol{A}_{\tau'}\boldsymbol{x}_{j-1})\|_2^2$$

$$= d(\boldsymbol{x}_{j-1}, S)^2 - \frac{1}{m'\beta'}\|e(\boldsymbol{b}_\leq - \boldsymbol{A}_\leq \boldsymbol{x}_{j-1})\|_2^2.$$

Combining this with (2.1) and letting $E_=$ and $E_\leq$ denote the events that a block from T and a block from T' is selected, respectively, we have

$$\mathbb{E}\big[(d(\boldsymbol{x}_j, S)^2\big] = p \cdot \mathbb{E}[d(\boldsymbol{x}_j, S)^2 | E_=] + (1-p) \cdot \mathbb{E}[d(\boldsymbol{x}_j, S)^2 | E_\leq]$$

$$\leq p\left[d(\boldsymbol{x}_{j-1}, S)^2 - \frac{1}{\beta m}\sum_{i \in I_=} e(\boldsymbol{A}_= \boldsymbol{x}_{j-1} - \boldsymbol{b}_=)_i^2\right] + (1-p)\left[d(\boldsymbol{x}_{j-1}, S)^2 - \frac{1}{m'\beta'}\|e(\boldsymbol{b}_\leq - \boldsymbol{A}_\leq \boldsymbol{x}_{j-1})\|_2^2\right]$$

$$= d(\boldsymbol{x}_{j-1}, S)^2 - p \cdot \frac{1}{\beta m}\sum_{i \in I_=} e(\boldsymbol{A}_= \boldsymbol{x}_{j-1} - \boldsymbol{b}_=)_i^2 - (1-p) \cdot \frac{1}{m'\beta'}\|e(\boldsymbol{b}_\leq - \boldsymbol{A}_\leq \boldsymbol{x}_{j-1})\|_2^2$$

Since $p = \frac{\beta m}{\beta'm' + \beta m}$, we have $\frac{1-p}{\beta'm'} = \frac{1}{\beta'm' + \beta m}$ and we can simplify

$$\mathbb{E}\big[d(\boldsymbol{x}_j, S)^2\big] \leq d(\boldsymbol{x}_{j-1}, S)^2 - \frac{1}{\beta'm' + \beta m}\left[\sum_{i \in I_=} e(\boldsymbol{A}_= \boldsymbol{x}_{j-1} - \boldsymbol{b}_=)_i^2 + \|e(\boldsymbol{b}_\leq - \boldsymbol{A}_\leq \boldsymbol{x}_{j-1})\|_2^2\right]$$

$$= d(\boldsymbol{x}_{j-1}, S)^2 - \frac{1}{\beta'm' + \beta m}\|e(\boldsymbol{A}\boldsymbol{x}_{j-1} - \boldsymbol{b})\|_2^2$$

$$\leq d(\boldsymbol{x}_{j-1}, S)^2 - \frac{1}{L^2(\beta'm' + \beta m)} \cdot d(\boldsymbol{x}_{j-1}, S)^2$$

$$= \left[1 - \frac{1}{L^2(\beta'm' + \beta m)}\right]d(\boldsymbol{x}_{j-1}, S)^2,$$

where we have utilized the Hoffman bound (1.7) in the second inequality.

Iterating this relation along with independence of the random control completes the proof. □

REFERENCES

[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. JRMAA8; 46L05 (46L30 47B35 47D25); 1124564 (92j:46104); H. Halpern.

[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. *Numer. Math.*, 41(1):83–92, 1983.

[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.

[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.

[FCM$^+$92]   H. G. Feichtinger, C. Cenker, 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 *Applications in Optical Science and Engineering*, pages 299–310. International Society for Optics and Photonics, 1992.

[FS95]   H. G. Feichtinger and T. Strohmer. A kaczmarz-based approach to nonperiodic sampling on unions of rectangular lattices. In *SampTA '95: 1995 Workshop on Sampling Theory and Applications*, pages 32–37, Jurmala, Latvia, Sep. 1995.

[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 T. Med. 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.

[Hof52]   A. J. Hoffman. On approximate solutions of systems of linear inequalities. *J. Research Nat. Bur. Standards*, 49:263–265, 1952.

[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. Int. Acad. Polon. Sci. Lett. Ser. A*, pages 335–357, 1937.

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

[LWS14]   J. Liu, S. J. Wright, and S. Srikrishna. An asynchronous parallel randomized kaczmarz algorithm. Available at `arXiv: 1401.4780`, January 2014.

[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.

[NSW13]   D. Needell, N. Srebro, and R. Ward. Stochastic gradient descent and the randomized kaczmarz algorithm. Submitted, 2013.

[NT14]   D. Needell and J. A. Tropp. Paved with good intentions: Analysis of a randomized block Kaczmarz method. *Linear Algebra Appl.*, 441:199–221, 2014.

[NZZ14]   D. Needell, R. Zhao, and A. Zouzias. Randomized block kaczmarz method with projection for solving least squares. Submitted, 2014.

[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.

[RR12]   B. Recht and C. Ré. Beneath the valley of the noncommutative arithmetic–geometric mean inequality: Conjectures, case studies, and consequences. In *Proc. 25th Ann. Conf. Learning Theory*, Edinburgh, June 2012.

[SS87]   M. I. Sezan and H. Stark. Applications of convex projection theory to image recovery in tomography and related areas. *Image Recovery: Theory and Application*, pages 155–270, 1987.

[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. *Numer. Math.*, 17(3):203–214, 1971.

[Tro08]   J. A. Tropp. The random paving property for uniformly bounded matrices. *Studia Math.*, 185(1):67–82, 2008. 46B09 (15A52 46B20 60E15); 2379999 (2008k:46030); Sasha Sodin.

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

[Tro11]   J. A. Tropp. Improved analysis of the subsampled randomized hadamard transform. *Advances in Adaptive Data Analysis*, 3(01n02):115–126, 2011.

[Tro12]   J. A. Tropp. User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.*, 12(4):389–434, 2012.

[Ver06]   R. Vershynin. *Random sets of isomorphism of linear operators on Hilbert space*, volume 51 of *High dimensional prob-ability*, pages 148–154. Inst. Math. Statist, Beachwood, OH, 2006. 46B09 (46B07 46B20); 2387766 (2009h:46023); Dirk Werner.

[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.

[ZF13]   A. Zouzias and N. M. Freris. Randomized extended kaczmarz for solving least squares. *SIAM J. Matrix Anal. A.*, 34(2):773–793, 2013.