Greed Works: An Improved Analysis of Sampling Kaczmarz-Motzkin*

Jamie Haddock† and Anna Ma‡

Abstract. Stochastic iterative algorithms have gained recent interest in machine learning and signal processing for solving large-scale systems of equations, $A\mathbf{x} = \mathbf{b}$. One such example is the Randomized Kaczmarz (RK) algorithm, which acts only on single rows of the matrix A at a time. While RK randomly selects a row of A to work with, Motzkin's Method (MM) employs a greedy row selection. Connections between the two algorithms resulted in the Sampling Kaczmarz-Motzkin (SKM) algorithm which samples a random subset of β rows of A and then greedily selects the best row of the subset. Despite their variable computational costs, all three algorithms have been proven to have the same theoretical upper bound on the convergence rate. In this work, an improved analysis of the range of random (RK) to greedy (MM) methods is presented. This analysis improves upon previous known convergence bounds for SKM, capturing the benefit of partially greedy selection schemes. This work also further generalizes previous known results, removing the theoretical assumptions that β must be fixed at every iteration and that A must have normalized rows.

Key words. Kaczmarz method, iterative methods, greedy methods, randomization

AMS subject classifications. 65F10, 65F20

1. Introduction.

Large-scale systems of equations arise in many areas of data science, including in machine learning and as subroutines of several optimization methods [12]. We consider solving these large systems of linear equations, $A\mathbf{x} = \mathbf{b}$, where $A \in \mathbb{R}^{m\times n}$, $\mathbf{b} \in \mathbb{R}^m$, and $m \gg n$. Iterative methods which use a small portion of the data in each iteration are typically employed in this domain. These methods offer a small memory footprint and good convergence guarantees. The Kaczmarz method [36] is such an iterative method that consists of sequential orthogonal projections towards the solution set of a single equation (or subsystem). Given the system $A\mathbf{x} = \mathbf{b}$, the method computes iterates by projecting onto the hyperplane defined by the equation $\mathbf{a}_i^T\mathbf{x} = b_i$ where $\mathbf{a}_i^T$ is a selected row of the matrix A and b_i is the corresponding entry of $\mathbf{b}$. The iterates are recursively defined as

$$\mathbf{x}_{j+1} = \mathbf{x}_j + \frac{b_i - \mathbf{a}_{i_j}^T\mathbf{x}_i}{\|\mathbf{a}_{i_j}\|^2}\mathbf{a}_i.$$

We assume that $A\mathbf{x} = \mathbf{b}$ is consistent and $m > n$, but make no assumption on rank(A). We will use $\mathbf{r}_j := A\mathbf{x}_j - \mathbf{b}$ to represent the jth residual and $\mathbf{e}_j := \mathbf{x}_j - \mathbf{x}^*$ to represent the jth error term. We let $A^\dagger$ denote the Moore-Penrose pseudoinverse of the matrix A. Additionally,

*Submitted to the editors X.

Funding: This material was supported by the NSF DMS-1440140 while the authors were in residence at the Mathematical Science Research Institute in Berkeley, California, during the Fall 2017 semester. Both authors were partially supported by NSF CAREER DMS-1348721 and NSF BIGDATA 1740325. JH was partially funded by NSF DMS-1522158 and NSF DMS-1818969. AM was partially supported the U.S. Air Force Award FA9550-18-1-0031 led by Roman Vershynin.

†Department of Mathematics, University of California, Los Angeles, Los Angeles, CA (jhaddock@math.ucla.edu).

‡Department of Mathematics, University of California, Irvine, Irvine, CA (anna.ma@uci.edu).

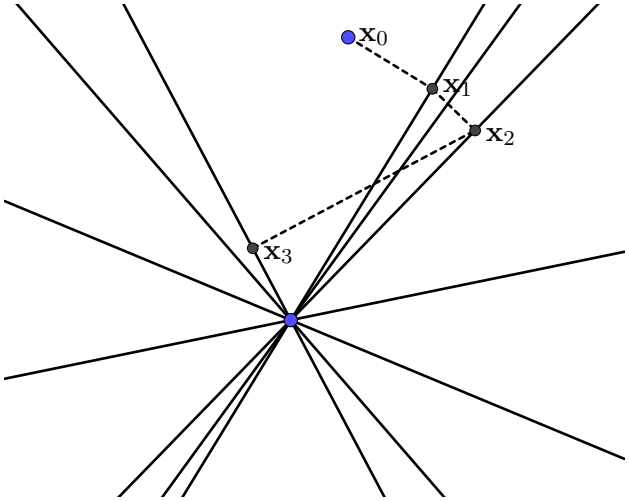


Figure 1.1: Several iterations of a Kaczmarz method. The iterate $\mathbf{x}_{j+1}$ is the orthogonal projection of $\mathbf{x}_j$ onto the solution set of the selected equation (represented by a line).

we let $\sigma_{\min}(A)$ be the smallest nonzero singular value of A and unless otherwise noted, we let $\|\cdot\|$ represent the Euclidean norm. We let $\|\cdot\|_F$ denote the Frobenius norm and $\|\cdot\|_\infty$ denote the ℓ^∞ norm. A visualization of several iterations of a Kaczmarz method are shown in Figure 1.1.

The Kaczmarz method was originally proposed in the late 30s [36] and rediscovered in the 1970's under the name *algebraic reconstruction technique (ART)* as an iterative method for reconstructing an image from a series of angular projections in computed tomography [27, 35]. This method has seen popularity among practitioners and researchers alike since the beginning of the digital age [14, 33], but saw a renewed surge of interest after the elegant convergence analysis of the *Randomized Kaczmarz (RK) method* in [60]. In [60], the authors showed that for a consistent system with unique solution, RK (with specified sampling distribution) converges at least linearly in expectation with the guarantee

$$\mathbb{E}\|\mathbf{e}_k\|^2 \leq \left(1 - \frac{\sigma_{\min}^2(A)}{\|A\|_F^2}\right)^k \|\mathbf{e}_0\|^2. \tag{1.1}$$

Many variants and extensions followed, including convergence analyses for inconsistent and random linear systems [49, 15], connections to other popular iterative algorithms [44, 51, 56, 57, 21], block approaches [52, 58], acceleration and parallelization strategies [22, 37, 47, 45], and techniques for reducing noise and corruption [68, 32].

Another popular Kaczmarz method extension is greedy (rather than randomized) row selection, which has been rediscovered several times in the literature as the "most violated constraint control" or the "maximal-residual control" [13, 54, 55]. This method was proposed in the 1950's as an iterative relaxation method for linear programming by Agmon, Motzkin, and Schoenberg under the name *Motzkin's relaxation method for linear inequalities (MM)* [48, 1]. In [1], the author showed that MM converges at least linearly (deterministically) with

the convergence rate of (1.1). The bodies of literature studying this greedy strategy have remained somewhat disjoint, with analyses for linear systems of equations in the numerical linear algebra community and analyses for linear systems of inequalities in the operations research and linear programming community [25, 26, 62, 4, 8, 9, 16]. There has been recent work in analyzing variants of this greedy strategy [20, 6, 7, 59]. In [59], the authors analyze MM on a system to which a Gaussian sketch has been applied. In [6, 7], the authors analyze variants of MM in which the equation selected in each iteration is chosen randomly amongst the set whose residual values are sufficiently near the maximal residual value. In [20], the authors provide a convergence analysis for a generalized version of MM in which the equation chosen in each iteration is that which has the maximal *weighted residual* value which are the residual values divided by the norm of the corresponding row of the measurement matrix. In [19], the authors illustrated the connection between MM and RK and proposed a family of algorithms that interpolate between the two, known as the *Sampling Kaczmarz-Motzkin (SKM) methods.*

The SKM methods operate by randomly sampling a subset of the system of equations, computing the residual of this subset, and projecting onto the equation corresponding to the largest magnitude entry of this sub-residual. The family of methods (parameterized by the size of the random sample of equations, β) interpolates between MM, which is SKM with $\beta = m$, and RK, which is SKM with $\beta = 1$. In [19], the authors prove that the SKM methods converge at least linearly in expectation with the convergence rate specified in (1.1). Meanwhile, the empirical convergence of this method is seen to depend upon β; however, increasing β also increases the computational cost of each iteration so the per iteration gain from larger sample size may be outweighed by the in-iteration cost. This is reminiscent of other methods which use sampled subsets of data in each iteration, such as the block projection methods [2, 52, 53].

Like SKM, the randomized block Kaczmarz (RBK) methods use a subset of rows $\tau \subset [m]$ to produce the next iterate; rather than forcing the next iterate to satisfy the single sampled equation as in RK, block iterates satisfy *all* the equations in the randomly sampled block. The $(k+1)$st RBK iteration is given by

$$\mathbf{x}_{k+1} = \mathbf{x}_k + (A_\tau)^\dagger(\mathbf{b}_\tau - A_\tau \mathbf{x}_k),$$

where A_τ and $\mathbf{b}_\tau$ represent the restriction onto the row indices in τ. In [52], the authors prove that on a system with a row-normalized measurement matrix and a well-conditioned row-paving RBK converges at least linearly in expectation with the guarantee

$$\mathbb{E}\|\mathbf{e}_k\|^2 \leq \left(1 - \frac{\sigma_{\min}^2(A)}{C\|A\|^2 \log(m+1)}\right)^k \|\mathbf{e}_0\|^2. \tag{1.2}$$

where C is an absolute constant and $\|A\|$ denotes the operator norm of the matrix. This can be a significant improvement over the convergence rate of (1.1) when $\|A\|_F^2 \gg \|A\|^2 \log(m+1)$. However, the cost per iteration scales with the size of the blocks. In [53], the authors generalize this result to inconsistent systems and show that, up to a convergence horizon, RBK converges to the least-squares solution.

In [29, 30], the authors introduce a framework of iterative methods known as the *sketch-and-project* methods. The sketch-and-project framework of methods produce each new iterate by projecting the previous iterate onto a sketch of the linear system; i.e., $\mathbf{x}_{k+1} =$

$\text{argmin}_{\mathbf{x}\in\mathbb{R}^n}\|\mathbf{x}-\mathbf{x}_k\|_B^2$ s.t. $S_k^\top A\mathbf{x} = S_k^\top \mathbf{b}$. Subsequent works proposed and analyzed variants with momentum [40], inexact variants [42], and adaptive variants [28]. This framework includes as special cases many forms of row- and column-action methods and second-order iterative least-squares methods [64]. Kaczmarz methods which iteratively project onto the solution spaces of subsets of rows in each iteration (like Block RK or SKM) can be interpreted and analyzed in this framework. Single row-action methods are recovered when the sketching matrices select a single row of the system. The SKM methods are recovered when the sketching matrices select a single row of the system and the choice of which sketch to use in each iteration is made in the same way as SKM (a randomized sample then a greedy selection based upon sketched residual). The results recovered from [29] for this interpretation of SKM coincide with (1.1).

In [50], the authors, inspired by the sketching framework in [29], construct a block-type method which iterates by projecting onto a Gaussian sketch of the equations. They show that this method converges at least linearly in expectation with the guarantee

$$\mathbb{E}\|\mathbf{e}_k\|^2 \leq \left(1-\left[\frac{\sqrt{s}\sigma_{\min}(A)}{9\sqrt{s}\|A\|+C\|A\|_F}\right]^2\right)^k \|\mathbf{e}_0\|^2. \tag{1.3}$$

where C is an absolute constant and s is the number of rows in the resulting sketched system. This result requires a Gaussian sketch which is a costly operation, however the authors suggest using a Gaussian sketch of only a subset of the equations. This result is most related to SKM and to our main result due to the presence of s, the size of the sketched system, in the bound.

2. Previous Results. This section focuses on the convergence behavior of the RK, MM, and SKM methods. Each of these projection methods is a special case of Algorithm 2.1 with a different *selection rule* (Line 4). In iteration j, RK uses the randomized selection rule that chooses $t_j = i$ with probability $\|\mathbf{a}_i\|_2^2/\|A\|_F^2$, MM uses the greedy selection rule $t_j = \arg\max_i |\mathbf{a}_i^\top \mathbf{x}_{j-1} - b_i|$, and SKM uses the hybrid selection rule that first samples a subset of β rows, τ_j, uniformly at random from all subsets of size β, $\tau_j \sim \text{unif}(\binom{[m]}{\beta})$, and then chooses $t_j = \arg\max_{i\in\tau_j} |\mathbf{a}_i^\top \mathbf{x}_{j-1} - b_i|$. As previously mentioned, RK and MM are special cases of the SKM method when the sample size $\beta = 1$ and $\beta = m$, respectively. Each of the methods converge linearly when the system is consistent with unique solution (RK and SKM converge linearly in expectation, MM converges linearly deterministically). In Table 2.1, we present the selection rules and convergence rates for RK, MM, and SKM. Note that under the assumption that A has been normalized so that $\|\mathbf{a}_i\|^2 = 1$, each of these upper bounds on the convergence rate is the same since $\|A\|_F^2 = m$. Thus, these results do not reveal any advantage the more computationally expensive methods (MM, SKM with $\beta \gg 1$) enjoy over RK. There are, in fact, pathological examples on which RK, MM, and SKM exhibit nearly the same behavior (e.g., consider the system defining two lines that intersect at one point in $\mathbb{R}^2$), so it is not possible to prove significantly different convergence rates without leveraging additional properties of the system.

In [31], the authors demonstrate that MM can converge faster than RK or SKM and that the convergence rate depends on the structure of the residual terms of the iterations,

Algorithm 2.1 Generic Kaczmarz Method

procedure KACZ($A, \mathbf{b}, \mathbf{x}_0$)
  $k = 1$
  repeat
    Choose $t_k \in [m]$ according to selection rule.
    $\mathbf{x}_k = \mathbf{x}_{k-1} - \frac{\mathbf{a}_{t_k}^T \mathbf{x}_{k-1} - b_{t_k}}{\|\mathbf{a}_{t_k}\|_2^2} \mathbf{a}_{t_k}$.
    $k = k + 1$
  until stopping criterion reached
  return $\mathbf{x}_k$
end procedure

	Selection Rule	Convergence Rate
RK [60]	$\mathbb{P}(t_j = i) = \frac{\|\mathbf{a}_i\|^2}{\|A\|_F^2}$	$\mathbb{E}\|\mathbf{e}_k\|^2 \leq (1 - \frac{\sigma_{\min}^2(A)}{\|A\|_F^2})^k \|\mathbf{e}_0\|^2$
SKM [19]	$\tau_j \sim \text{unif}(\binom{[m]}{\beta})$ $t_j = \arg\max_{i \in \tau_j} \|\mathbf{a}_i^\top \mathbf{x}_{j-1} - b_i\|$	$\mathbb{E}\|\mathbf{e}_k\|^2 \leq (1 - \frac{\sigma_{\min}^2(A)}{m})^k \|\mathbf{e}_0\|^2$
MM [1]	$t_j = \arg\max_i \|\mathbf{a}_i^\top \mathbf{x}_{j-1} - b_i\|$	$\|\mathbf{e}_k\|^2 \leq (1 - \frac{\sigma_{\min}^2(A)}{m})^k \|\mathbf{e}_0\|^2$

Table 2.1: The selection rules and convergence rates of RK, SKM, and MM. The presented results for MM and SKM assume that A has been normalized so that $\|\mathbf{a}_i\|^2 = 1$.

$\mathbf{r}_k = A\mathbf{x}_k - \mathbf{b}$. In particular, they prove that

$$\|\mathbf{e}_k\|^2 \leq \Pi_{j=0}^{k-1} \Big(1 - \frac{\sigma_{\min}^2(A)}{4\gamma_j}\Big) \|\mathbf{e}_0\|^2,$$

where γ_j is the *dynamic range* of the ith residual, $\gamma_j := \frac{\|\mathbf{r}_j\|^2}{\|\mathbf{r}_j\|_\infty^2}$. Our main contribution in this paper is to prove that the SKM methods can exhibit a similarly accelerated convergence rate and the advantage scales with the size of the sample, β. Again, this advantage depends upon the structure of the residuals of the iterations. We define here a generalization of the dynamic range used in [31]; our dynamic range is defined as

$$\gamma_j = \frac{\sum_{\tau \in \binom{[m]}{\beta}} \|A_\tau \mathbf{x}_{j-1} - \mathbf{b}_\tau\|^2}{\sum_{\tau \in \binom{[m]}{\beta}} \|A_\tau \mathbf{x}_{j-1} - \mathbf{b}_\tau\|_\infty^2}. \tag{2.1}$$

Now, we let $\mathbb{E}_{\tau_j}$ denote expectation with respect to the random sample τ_j conditioned upon the sampled τ_i for $i < j$, and $\mathbb{E}$ denote expectation with respect to all random samples τ_i for $1 \leq i \leq j$ where j is understood to be the last iteration in the context in which $\mathbb{E}$ is applied. We state our main result below in Corollary 2.1; this is a corollary of our generalized result which will be discussed and proven later.

Corollary 2.1. *Let A be normalized so $\|\mathbf{a}_i\| = 1$ for all rows $i = 1, ..., m$. Suppose the system of equations $A\mathbf{x} = \mathbf{b}$ is consistent, define $\mathbf{x}^* = A^\dagger \mathbf{b}$, and let $\mathbf{x}_0 \in$ range$(A^\top)$. Then SKM*

converges at least linearly in expectation and the bound on the rate depends on the dynamic range, γ_k of the random sample of β rows of A, τ_k. Precisely, in the kth iteration of SKM, we have

$$\mathbb{E}_{\tau_k}\|\mathbf{x}_k - \mathbf{x}^*\|^2 \leq \left(1 - \frac{\beta\sigma^2_{\min}(A)}{\gamma_k m}\right)\|\mathbf{x}_{k-1} - \mathbf{x}^*\|^2,$$

so applying expectation with respect to all iterations, we have

$$\mathbb{E}\|\mathbf{x}_k - \mathbf{x}^*\|^2 \leq \prod_{j=1}^{k}\left(1 - \frac{\beta\sigma^2_{\min}(A)}{\gamma_j m}\right)\|\mathbf{x}_0 - \mathbf{x}^*\|^2.$$

Corollary 2.1 shows that SKM experiences at least linear convergence where the contraction term is a product of terms that are less than one and dependent on the sub-sample size β. When $\beta = 1$, as in RK, $\gamma_k = 1$, so Corollary 2.1 recovers the upper bound for RK shown in [60]. However, when $\beta = m$ for MM, Corollary 2.1 offers an improved upper bound on the error over [31]; specifically

$$\|\mathbf{e_k}\|^2 \leq \left(1 - \frac{\sigma^2_{\min}(A)}{\gamma_k}\right)\|\mathbf{e}_{k-1}\|^2.$$

Our result illustrates that the progress made by an iteration of the SKM algorithm depends upon the dynamic range of the residual of that iteration. The dynamic range of each iteration, γ_j, satisfies

$$1 \leq \gamma_j \leq \beta.$$

Note that the upper bound, $\gamma_j = \beta$, is achieved by a constant residual where $|\mathbf{a}_i^T\mathbf{x}_j - b_i| = |\mathbf{a}_{i'}^T\mathbf{x}_j - b_{i'}|$ for all $i, i' \in [m]$, while the lower bound is achieved by the residual with one nonzero entry. As smaller γ_j provides a smaller upper bound on the new error $\mathbf{e}_j$, we consider the situation with one nonzero entry in the residual as the "best case" and the situation with a constant residual as the "worst case." We now compare our single iteration result in the best and worst cases to the previously known single iteration results of [1, 19, 31, 60]. These are summarized in Table 2.2; we present only the contraction terms α such that

$$\mathbb{E}_{\tau_k}\|\mathbf{e}_k\|^2 \leq \alpha\|\mathbf{e}_{k-1}\|^2,$$

for each upper bound in the case that A is normalized so that $\|\mathbf{a}_i\|^2 = 1$ for $i \in [m]$. In particular, note that the worst case residual provides the same upper bound rate as those of [60, 19, 1].

3. Main Results. Corollary 2.1 is a specialization of our general result to SKM with a fixed sample size β and systems that are row-normalized. Our general result requires neither row-normalization nor a static sample size. However, we must additionally generalize the SKM sampling distribution for systems that are not row-normalized. We now consider the general SKM method which samples β_k many rows of A in the kth iteration (according to probability distribution $p_{\mathbf{x}_{k-1}}$ defined in (3.1)) and projects onto the hyperplane associated to the largest magnitude entry of the sampled sub-residual.

	Best Case	Worst Case	Previous Best Case	Previous Worst Case
MM	$1-\sigma_{\min}^2(A)$	$1-\frac{\sigma_{\min}^2(A)}{m}$	$1-\frac{\sigma_{\min}^2(A)}{4}$ [31]	$1-\frac{\sigma_{\min}^2(A)}{m}$ [1, 19, 60]
SKM	$1-\frac{\beta\sigma_{\min}^2(A)}{m}$		$1-\frac{\sigma_{\min}^2(A)}{m}$	
RK	$1-\frac{\sigma_{\min}^2(A)}{m}$			

Table 2.2: Contraction terms α such that $\mathbb{E}_{\tau_k}\|\mathbf{e}_k\|^2 \leq \alpha\|\mathbf{e}_{k-1}\|^2$ for the best and worst case bounds of MM, SKM, and RK.

The generalized probability distribution over the subset of rows of A of size β_k is denoted $p_{\mathbf{x}} : \binom{[m]}{\beta_k} \to [0,1)$. The sampled subset of rows of A, $\tau_k \sim p_{\mathbf{x}}$ where

$$p_{\mathbf{x}}(\tau_k) = \frac{\|\mathbf{a}_{t(\tau_k,\mathbf{x})}\|^2}{\sum_{\tau\in\binom{[m]}{\beta_k}}\|\mathbf{a}_{t(\tau,\mathbf{x})}\|^2}, \tag{3.1}$$

and $t(\tau,\mathbf{x}) = \arg\max_{t\in\tau}(\mathbf{a}_t^\top\mathbf{x}-b_t)^2$. Thus, our generalized SKM method is Algorithm 2.1 with selection rule $\tau_j \sim p_{\mathbf{x}_{j-1}}$ and $t_j = t(\tau_j, \mathbf{x}_{j-1})$. Similar to the RK probability distribution of [60], the computation of (3.1) is utilized here simply to theoretically analyze the SKM algorithm without requiring normalized rows. This choice of sampling distribution conveniently simplifies the expected value computation in the proof of Theorem 3.1 by cancelling the numerator of the probability with the squared norm of the sampled row. We do not suggest that this probability distribution be implemented in a real world setting as it is computationally prohibitive.

One could instead implement a uniform distribution over rows or learn the distribution with probabilities proportional to the squared norms of the rows (as suggested in [60]). Neither of these is guaranteed to coincide with the distribution defined in (3.1), due to the dependence on the iterate $\mathbf{x}$. However, for many datasets where the row norms are (approximately) equal, the uniform distribution (approximately) coincides with (3.1). In particular, when the rows of A all have equal norm, as in the case of incidence matrices (see Section 4.2), then (3.1) reduces to the uniform distribution over samples of size β_k. Past works which analyze SKM [19, 47, 46] assume that the rows of A are normalized and that the probability distribution over the samples of size β is uniform. To the best of our knowledge, ours is the first work in this area to analyze an iterative projection method with an iteration dependent sampling distribution.

Our main result shows that the generalized SKM converges at least linearly in expectation with a bound that depends on the dynamic range of the sampled sub-residual, the size of the sample, and the minimum squared nonzero singular value of A, $\sigma_{\min}^2(A)$. In the event that there are multiple rows within the sub-residual which achieve $\max_{t\in\tau}(\mathbf{a}_t^\top\mathbf{x}-b_t)^2$, an arbitrary choice can be made amongst those rows and the main result will not be affected by this choice.

Theorem 3.1 provides theoretical convergence guarantees for the generalized SKM method. Whereas previous guarantees for SKM required normalized rows or fixed sample sizes β [19, 31, 47, 46], the guarantees presented here do not require either assumption. In addition, the contraction term of the generalized SKM method shows dependence on the dynamic range, another feature lacking in previous works. Following the statement of the theorem, we use

standard techniques in the Kaczmarz literature to prove our main result.

We additionally describe a simple generalization of the main result in the case that the samples are not made according to the generalized SKM distribution (3.1), but instead according to a distribution $\tilde{p}(\tau)$ whose probabilities are at least a constant factor of those in (3.1). We also remark on the special case in which rows have equal norm (and thus subsets τ are selected uniformly at random), the case where β_k is fixed, and the case in which $\beta_k = 1$ in order to make connections to previous results. Due to the dependence of the sampling distribution upon the current iterate, our main result is not easily iterable to provide the usual form of a Kaczmarz type result (e.g., $\mathbb{E}\|\mathbf{e}_k\|^2 \leq \alpha^k \|\mathbf{e}_0\|^2$) so we present the bound for only a single iteration. However, in the special cases we describe in Remarks 2 and 3 we are able to iterate the simplified expression due to the simplicity of the sampling distribution.

Theorem 3.1. *Suppose the system of equations $A\mathbf{x} = \mathbf{b}$ is consistent, define $\mathbf{x}^* = A^\dagger \mathbf{b}$, and let $\mathbf{x}_0 \in$ range$(A^\top)$. Then generalized SKM converges at least linearly in expectation and the bound on the rate depends on the dynamic range, γ_k of the random sample of β_k rows of A, τ_k. Precisely, in the kth iteration of generalized SKM, we have*

$$\mathbb{E}_{\tau_k}\|\mathbf{x}_k - \mathbf{x}^*\|^2 \leq \left(1 - \frac{\beta_k \binom{m}{\beta_k} \sigma_{\min}^2(A)}{\gamma_k m \sum_{\tau \in \binom{[m]}{\beta_k}} \|\mathbf{a}_{t(\tau, \mathbf{x}_{k-1})}\|^2}\right) \|\mathbf{x}_{k-1} - \mathbf{x}^*\|^2.$$

Proof. We begin by rewriting the generalized SKM iterate $\mathbf{x}_k$ and simplifying the resulting expression which yields

$$\begin{aligned}
\|\mathbf{x}_k - \mathbf{x}^*\|^2 &= \left\| \mathbf{x}_{k-1} - \frac{\mathbf{a}_{t(\tau_k, \mathbf{x}_{k-1})}^\top \mathbf{x}_{k-1} - b_{t(\tau_k, \mathbf{x}_{k-1})}}{\|\mathbf{a}_{t(\tau_k, \mathbf{x}_{k-1})}\|^2} \mathbf{a}_{t(\tau_k, \mathbf{x}_{k-1})} - \mathbf{x}^* \right\|^2 \\
&= \|\mathbf{x}_{k-1} - \mathbf{x}^*\|^2 - \frac{(\mathbf{a}_{t(\tau_k, \mathbf{x}_{k-1})}^\top \mathbf{x}_{k-1} - b_{t(\tau_k, \mathbf{x}_{k-1})})^2}{\|\mathbf{a}_{t(\tau_k, \mathbf{x}_{k-1})}\|^2} \\
&= \|\mathbf{x}_{k-1} - \mathbf{x}^*\|^2 - \frac{\|A_{\tau_k}\mathbf{x}_{k-1} - \mathbf{b}_{\tau_k}\|_\infty^2}{\|\mathbf{a}_{t(\tau_k, \mathbf{x}_{k-1})}\|^2},
\end{aligned}$$

where the first equation uses the definition of the generalized SKM iterate, and the second follows from the fact that $\mathbf{a}_{t(\tau_k, \mathbf{x}_{k-1})}^\top (\mathbf{x}_{k-1} - \mathbf{x}^*) = \mathbf{a}_{t(\tau_k, \mathbf{x}_{k-1})}^\top \mathbf{x}_{k-1} - b_{t(\tau_k, \mathbf{x}_{k-1})}$. Note that

$$\mathbf{a}_{t(\tau_k, \mathbf{x}_{k-1})}^\top \mathbf{x}^* = (AA^\dagger \mathbf{b})_{t(\tau_k, \mathbf{x}_{k-1})} = b_{t(\tau_k, \mathbf{x}_{k-1})}$$

since $\mathbf{b} \in$ range(A) and $AA^\dagger$ is the projector onto range(A).

Now, we take expectation of both sides (with respect to the sampled τ_k according to the

distribution (3.1)). This gives

$$
\begin{aligned}
\mathbb{E}_{\tau_k}\|\mathbf{x}_k-\mathbf{x}^*\|^2 &= \|\mathbf{x}_{k-1}-\mathbf{x}^*\|^2-\mathbb{E}_{\tau_k}\frac{\|A_{\tau_k}\mathbf{x}_{k-1}-\mathbf{b}_{\tau_k}\|_\infty^2}{\|\mathbf{a}_{t(\tau_k,\mathbf{x}_{k-1})}\|^2}\\
&= \|\mathbf{x}_{k-1}-\mathbf{x}^*\|^2-\sum_{\tau\in\binom{[m]}{\beta_k}} p_{\mathbf{x}_{k-1}}(\tau)\frac{\|A_\tau\mathbf{x}_{k-1}-\mathbf{b}_\tau\|_\infty^2}{\|\mathbf{a}_{t(\tau,\mathbf{x}_{k-1})}\|^2}\\
&= \|\mathbf{x}_{k-1}-\mathbf{x}^*\|^2-\sum_{\tau\in\binom{[m]}{\beta_k}} \frac{\|\mathbf{a}_{t(\tau,\mathbf{x}_{k-1})}\|^2}{\sum_{\pi\in\binom{[m]}{\beta_k}}\|\mathbf{a}_{t(\pi,\mathbf{x}_{k-1})}\|^2}\frac{\|A_\tau\mathbf{x}_{k-1}-\mathbf{b}_\tau\|_\infty^2}{\|\mathbf{a}_{t(\tau,\mathbf{x}_{k-1})}\|^2} \qquad (3.2)\\
&= \|\mathbf{x}_{k-1}-\mathbf{x}^*\|^2-\frac{1}{\sum_{\pi\in\binom{[m]}{\beta_k}}\|\mathbf{a}_{t(\pi,\mathbf{x}_{k-1})}\|^2}\sum_{\tau\in\binom{[m]}{\beta_k}}\|A_\tau\mathbf{x}_{k-1}-\mathbf{b}_\tau\|_\infty^2\\
&= \|\mathbf{x}_{k-1}-\mathbf{x}^*\|^2-\frac{1}{\gamma_k\sum_{\pi\in\binom{[m]}{\beta_k}}\|\mathbf{a}_{t(\pi,\mathbf{x}_{k-1})}\|^2}\sum_{\tau\in\binom{[m]}{\beta_k}}\|A_\tau\mathbf{x}_{k-1}-\mathbf{b}_\tau\|^2\\
&= \|\mathbf{x}_k-\mathbf{x}^*\|^2-\frac{\binom{m}{\beta_k}\beta_k}{\gamma_k m\sum_{\pi\in\binom{[m]}{\beta_k}}\|\mathbf{a}_{t(\pi,\mathbf{x}_{k-1})}\|^2}\|A\mathbf{x}_{k-1}-\mathbf{b}\|^2\\
&\le \left(1-\frac{\binom{m}{\beta_k}\beta_k\sigma_{\min}^2(A)}{\gamma_k m\sum_{\pi\in\binom{[m]}{\beta_k}}\|\mathbf{a}_{t(\pi,\mathbf{x}_{k-1})}\|^2}\right)\|\mathbf{x}_{k-1}-\mathbf{x}^*\|^2,
\end{aligned}
$$

where the last line follows from standard properties of singular values and the fact that $\mathbf{x}_{k-1}\in\text{range}(A^\top)$ (since $\mathbf{x}_0\in\text{range}(A^\top)$ and the SKM update preserves membership in $\text{range}(A^\top)$). This completes our proof. ∎

Now, we provide a corollary of the previous result which provides a bound on the expected error for the SKM algorithm which samples subsets of rows τ_k according to an alternate probability distribution $\tilde{p}(\tau)$ satisfying $\tilde{p}(\tau)\ge\epsilon p_{\mathbf{x}_{k-1}}(\tau)$ for all $\tau\in\binom{[m]}{\beta_k}$. In this case, we can exploit the relationship between probabilities to reuse the proof of Theorem 3.1. Provided that the probability distribution $\tilde{p}(\tau)$ is fixed between iterations we can iterate the bound unlike in Theorem 3.1. An application of Corollary 3.2 with the uniform distribution is given in Remark 1.

Corollary 3.2. *Suppose the system of equations $A\mathbf{x}=\mathbf{b}$ is consistent, define $\mathbf{x}^*=A^\dagger\mathbf{b}$, and let $\mathbf{x}_0\in\text{range}(A^\top)$. Suppose one runs SKM with $\tau_k\sim\tilde{p}(\tau)$ and $t_k=t(\tau_k,\mathbf{x}_{k-1})$ and that the probabilities used to sample, $\tilde{p}(\tau)$, are at least a constant factor of the probabilities (3.1); that is $\tilde{p}(\tau)\ge\epsilon p_{\mathbf{x}_{k-1}}(\tau)$ for all $\tau\in\binom{[m]}{\beta_k}$. Then we have that*

$$
\tilde{\mathbb{E}}_{\tau_k}\|\mathbf{x}_k-\mathbf{x}^*\|^2\le\left(1-\frac{\epsilon\binom{m}{\beta_k}\beta_k\sigma_{\min}^2(A)}{\gamma_k m\sum_{\pi\in\binom{[m]}{\beta_k}}\|\mathbf{a}_{t(\pi,\mathbf{x}_{k-1})}\|^2}\right)\|\mathbf{x}_{k-1}-\mathbf{x}^*\|^2 \qquad (3.3)
$$

where $\tilde{\mathbb{E}}_{\tau_k}$ denotes expectation taken with respect to the sampling of τ_k according to $\tilde{p}(\tau)$ and conditioned on the choices of τ_j for $j<k$. Furthermore, if $\tilde{p}(\tau)$ is constant between iterations

(so $\beta_j = \beta$ is constant) and independent of $\mathbf{x}_{k-1}$, we can iterate the previous result and have

$$\tilde{\mathbb{E}}\|\mathbf{x}_k - \mathbf{x}^*\|^2 \leq \prod_{j=1}^{k} \left(1 - \frac{\epsilon\binom{m}{\beta}\beta\sigma_{\min}^2(A)}{\gamma_j m \sum_{\pi\in\binom{[m]}{\beta}}\|\mathbf{a}_{t(\pi,\mathbf{x}_{j-1})}\|^2}\right)\|\mathbf{x}_0 - \mathbf{x}^*\|^2 \tag{3.4}$$

where $\tilde{\mathbb{E}}$ denotes expectation taken with respect to all samples of τ_j for $j = 1, ..., k$.

Proof. This proof is identical to that of Theorem 3.1 but where we first replace $\mathbb{E}_{\tau_k}$ with $\tilde{\mathbb{E}}_{\tau_k}$ and $p_{\mathbf{x}_{k-1}}(\tau)$ with $\tilde{p}(\tau)$. We replace the equation in (3.2) with an inequality and must add an ϵ to the numerator of the subtracted term in each line from (3.2) on. The iterated bound follows from recursively applying the bounds on the conditional expectations of each iteration. ∎

Remark 1. *(Uniform probability distribution) We consider the case of the uniform distribution over samples, i.e., $\tilde{p}(\tau) = 1/\binom{m}{\beta_k}$. We note that since*

$$\frac{\min_{i\in[m]}\|\mathbf{a}_i\|^2}{\binom{m}{\beta_k}\max_{i\in[m]}\|\mathbf{a}_i\|^2} \leq p_{\mathbf{x}_{k-1}}(\tau) \leq \frac{\max_{i\in[m]}\|\mathbf{a}_i\|^2}{\binom{m}{\beta_k}\min_{i\in[m]}\|\mathbf{a}_i\|^2}$$

when $min_{i\in[m]}\|\mathbf{a}_i\|^2 > 0$, we have

$$\tilde{p}(\tau) \geq \frac{\min_{i\in[m]}\|\mathbf{a}_i\|^2}{\max_{i\in[m]}\|\mathbf{a}_i\|^2} p_{\mathbf{x}_{k-1}}(\tau).$$

Thus, Corollary 3.2 holds for the uniform distribution with $\epsilon = \min_{i\in[m]}\|\mathbf{a}_i\|^2/\max_{i\in[m]}\|\mathbf{a}_i\|^2$. We note that this additionally provides a convergence analysis for RK of [60] in the case that the matrix A has unnormalized rows and the uniform distribution over rows is employed in sampling.

The next remarks make simplifying assumptions on the generalized SKM algorithm and our main result to provide better context for comparison with previous works.

Remark 2. *(Recovery of RK guarantees) If all of the rows of A have equal norm (not necessarily unit norm), then our result specializes to*

$$\mathbb{E}_{\tau_k}\|\mathbf{x}_k - \mathbf{x}^*\|^2 \leq \left(1 - \frac{\beta_k\sigma_{\min}^2(A)}{\gamma_k\|A\|_F^2}\right)\|\mathbf{x}_{k-1} - \mathbf{x}^*\|^2, \tag{3.5}$$

and we can iteratively apply this per-iteration guarantee to give a bound on the error in expectation with respect to all samples,

$$\mathbb{E}\|\mathbf{x}_k - \mathbf{x}^*\|^2 \leq \prod_{j=1}^{k}\left(1 - \frac{\beta_j\sigma_{\min}^2(A)}{\gamma_j\|A\|_F^2}\right)\|\mathbf{x}_0 - \mathbf{x}^*\|^2. \tag{3.6}$$

Additionally, when $\beta_k = 1$, the sampling distribution (3.1) and theoretical error upper bound (3.5) simplifies to the probability distribution and error guarantees of [60].

Remark 3. *(Improvement of MM guarantees) Corollary 2.1 is obtained from Theorem 3.1 when rows of A have unit norm and $\beta_k = \beta$. When $\beta_k = m$, then an improved convergence rate of*

$$\|\mathbf{x}_k - \mathbf{x}^*\|^2 \leq \left(1 - \frac{\sigma_{\min}^2(A)}{\gamma_k}\right)\|\mathbf{x}_{k-1} - \mathbf{x}^*\|^2 \leq \prod_{j=1}^{k}\left(1 - \frac{\sigma_{\min}^2(A)}{\gamma_j}\right)\|\mathbf{x}_0 - \mathbf{x}^*\|^2,$$

for MM over that shown in [31],

$$\|\mathbf{x}_k - \mathbf{x}^*\|^2 \leq \left(1 - \frac{\sigma_{\min}^2(A)}{4\gamma_k}\right)\|\mathbf{x}_{k-1} - \mathbf{x}^*\|^2 \leq \prod_{j=1}^{k}\left(1 - \frac{\sigma_{\min}^2(A)}{4\gamma_j}\right)\|\mathbf{x}_0 - \mathbf{x}^*\|^2,$$

is obtained.

Remark 4. *(Connection to Block RK) Note that this bound on the convergence rate of SKM additionally provides a bound on the convergence rate of a block Kaczmarz variant. This variant is distinct from the block Kaczmarz method considered in [52]. The analysis of [52] requires a pre-partitioned row paving, while the variant considered here allows the blocks to be sampled randomly and not pre-partitioned. Consider the block Kaczmarz variant which in each iteration selects a block of β_k rows of A, τ_k, and projects the previous iterate into the solution space of the entire block of β_k equations. This variant necessarily converges faster than SKM as it makes more progress in each iteration. In particular, note that $\{\mathbf{x}|A_{\tau_k}\mathbf{x} = \mathbf{b}_{\tau_k}\} \subset \{\mathbf{x}|\mathbf{a}_{t(\tau_k,\mathbf{y})}^T\mathbf{x} = b_{t(\tau_k,\mathbf{y})}\}$. Given iterate $\mathbf{x}_{k-1}$ and sample of rows τ_k, let $\mathbf{x}_k^{SKM}$ denote the iterate produced by SKM and $\mathbf{x}_k^{BK}$ denote the iterate produced by this block Kaczmarz variant. Note that $\mathbf{x}_k^{SKM}$ is the closest point to $\mathbf{x}_{k-1}$ on the hyperplane associated to equation $t(\tau_k, \mathbf{x}_{k-1})$ so, since $\mathbf{x}_k^{BK}$ also lies on this hyperplane, we have*

$$\|\mathbf{x}_k^{SKM} - \mathbf{x}_{k-1}\|^2 \leq \|\mathbf{x}_k^{BK} - \mathbf{x}_{k-1}\|^2.$$

Now, we note that by orthogonality of the projections, we have

$$\|\mathbf{x}_k^{SKM} - \mathbf{x}_{k-1}\|^2 + \|\mathbf{x}_k^{SKM} - \mathbf{x}^*\|^2 = \|\mathbf{x}_{k-1} - \mathbf{x}^*\|^2 = \|\mathbf{x}_k^{BK} - \mathbf{x}_{k-1}\|^2 + \|\mathbf{x}_k^{BK} - \mathbf{x}^*\|^2$$

so by the above inequality, we have

$$\|\mathbf{x}_k^{BK} - \mathbf{x}^*\|^2 \leq \|\mathbf{x}_k^{SKM} - \mathbf{x}^*\|^2.$$

A visualization of this situation is presented in Figure 3.1. Thus, the progress made by BK in any fixed iteration is at least as large as the progress made by SKM, so it must converge at least as quickly.

One may be assured that the contraction term in Theorem (3.1) is always strictly positive. We prove this simple fact in Proposition 3.3.

Proposition 3.3. *For any matrix A defining a consistent system with $\mathbf{x}^* = A^\dagger\mathbf{b}$ and $\mathbf{x}_{j-1} \in$ range$(A^\top)$, we have*

$$\gamma_j \geq \frac{\beta_j\binom{m}{\beta_j}\sigma_{\min}^2(A)}{m\sum_{\tau\in\binom{[m]}{\beta_j}}\|\mathbf{a}_{t(\tau,\mathbf{x}_{j-1})}\|^2}.$$

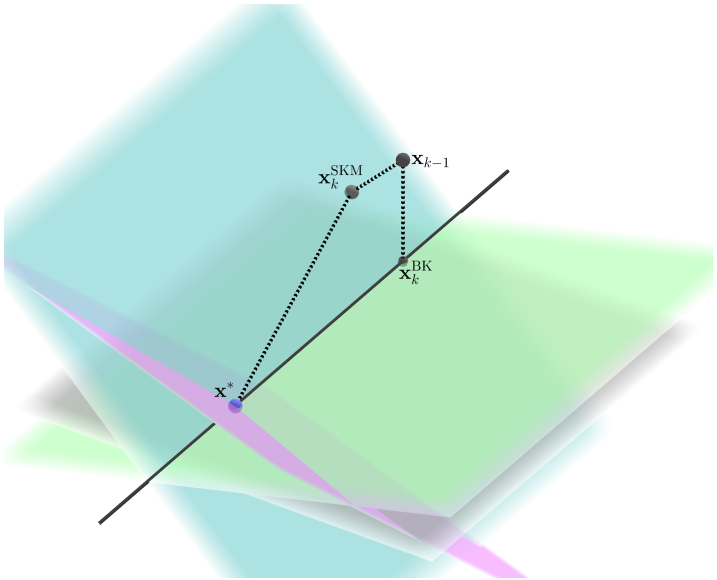


Figure 3.1: The SKM and BK iterates, $\mathbf{x}_k^{\mathrm{SKM}}$ and $\mathbf{x}_k^{\mathrm{BK}}$, generated by one iteration starting at $\mathbf{x}_{k-1}$ satisfy $\|\mathbf{x}_k^{\mathrm{BK}} - \mathbf{x}^*\|^2 \leq \|\mathbf{x}_k^{\mathrm{SKM}} - \mathbf{x}^*\|^2$.

Proof. Beginning with the definition of γ_j, we have

$$\begin{aligned}
\frac{\sum_{\tau_j \in \binom{[m]}{\beta_j}} \|A_{\tau_j}\mathbf{x}_{j-1} - \mathbf{b}_{\tau_j}\|^2}{\sum_{\tau_j \in \binom{[m]}{\beta_j}} \|A_{\tau_j}\mathbf{x}_{j-1} - \mathbf{b}_{\tau_j}\|_\infty^2} &= \frac{\frac{\beta_j}{m}\binom{m}{\beta_j}\|A(\mathbf{x}_{j-1} - \mathbf{x}^*)\|^2}{\sum_{\tau_j \in \binom{[m]}{\beta_j}} |\mathbf{a}_{t(\tau_j, \mathbf{x}_{j-1})}^\top (\mathbf{x}_{j-1} - \mathbf{x}^*)|^2} \\
&\geq \frac{\frac{\beta_j}{m}\binom{m}{\beta_j}\sigma_{\min}^2(A)\|\mathbf{x}_{j-1} - \mathbf{x}^*\|^2}{\sum_{\tau_j \in \binom{[m]}{\beta_j}} \|\mathbf{a}_{t(\tau_j, \mathbf{x}_{j-1})}\|^2 \|\mathbf{x}_{j-1} - \mathbf{x}^*\|^2} \\
&= \frac{\beta_j \binom{m}{\beta_j}\sigma_{\min}^2(A)}{m\sum_{\tau \in \binom{[m]}{\beta_j}} \|\mathbf{a}_{t(\tau, \mathbf{x}_{j-1})}\|^2},
\end{aligned}$$

where the inequality follows from properties of singular values and Cauchy-Schwartz. ∎

Because Theorem 3.1 shows that the contraction coefficient for generalized SKM is dependent on the dynamic range, the following section discusses bounds on the dynamic range for special types of linear systems.

4. Analysis of the Dynamic Range. Since the dynamic range plays an integral part in the convergence behavior for generalized SKM, the dynamic range is analyzed here for different types of specialized linear systems. Note that the dynamic range has also appeared in other works, although not under the guise of "dynamic range". For example, in [6] the authors proposed a Greedy Randomized Kaczmarz (GRK) algorithm that finds a subset of indices to randomly select the next row to project onto. The operation of finding this subset relies on a

ratio between the ℓ_∞ and ℓ_2 norms of the residual at the current iteration, essentially using a proxy of the dynamic range. In the next section, we analyze the dynamic range for random Gaussian linear systems and remark on the extension to other random linear systems. In the following section, we analyze the dynamic range for linear systems encoding average consensus problems on undirected graphs via the incidence matrix.

4.1. Gaussian Matrices. When entries of the measurement matrix A are drawn i.i.d. from a standard Gaussian distribution, it can be shown that the dynamic range is upper bounded by $\mathcal{O}(n\beta/\log\beta)$. The proof of the upper bound of γ_k is similar to Lemma 2 of [31], where the authors analyze the dynamic range for $\beta = m$. Here, we generalized the bound for varying samples sizes β_k.

Proposition 4.1. *Let $A \in \mathbb{R}^{m\times n}$ be a random Gaussian matrix with $a_{ij} \sim \mathcal{N}(0,\sigma^2)$. For each subset $\tau \in \binom{[m]}{\beta_k}$, let $I_\tau \subseteq \tau$ denote the set of rows in τ that are independent of $\mathbf{x}$ and note $|I_\tau| \leq \beta_k$. Assuming there is at least m' rows in $[m]$ which are independent of $\mathbf{x}$, the dynamic range can be upper bounded as:*

$$\gamma_j = \frac{\sum_{\tau\in\binom{[m]}{\beta}} \mathbb{E}_a\|A_\tau\mathbf{x}\|^2}{\sum_{\tau\in\binom{[m]}{\beta}} \mathbb{E}_a\|A_\tau\mathbf{x}\|_\infty^2} \leq \frac{\binom{m}{\beta_k}\left(\beta_k n + \sum_{i\in\tau\setminus I_\tau}\|\mathbf{a}_i\|^2/\sigma^2\right)}{\binom{m'}{\beta_k}\log(\beta_k)}. \tag{4.1}$$

Remark 5. *Note that the factor $\binom{m}{\beta_k}/\binom{m'}{\beta_k}$ is $\mathcal{O}(1)$ as $m\to\infty$ since*

$$\frac{\binom{m}{\beta_k}}{\binom{m'}{\beta_k}} = \frac{m!}{\beta_k!\,(m-\beta_k)!}\frac{\beta_k!\,(m-j-\beta_k)!}{(m-j)!} = \prod_{i=0}^{j}\frac{m-i}{m-\beta_k-i}.$$

Thus, we conclude that the expected dynamic range for any iteration k is $\mathcal{O}(n\beta_k/\log(\beta_k))$.

Proof. Without loss of generality, we let the solution to the system $\mathbf{x}^* = 0$ so that $\mathbf{b} = 0$. We are then interested in finding an upper bound on the dynamic range (2.1) in expectation. Here, the expectation is taken with respect to the random i.i.d. draws of the entries of A. To that end, we derive upper bounds and lower bounds on the numerator and denominator

of (2.1). Starting with the upper bound on the numerator we have

$$\begin{aligned}
\sum_{\tau\in\binom{[m]}{\beta_k}} \mathbb{E}_a\|A_\tau \mathbf{x}\|^2 &\leq \sum_{\tau\in\binom{[m]}{\beta}} \sum_{i\in\tau} \mathbb{E}_a\left(\|\mathbf{a}_i\|^2\|\mathbf{x}\|^2\right) \\
&= \sum_{\tau\in\binom{[m]}{\beta_k}} \left(\sum_{i\in I_\tau} \mathbb{E}_a\|\mathbf{a}_i\|^2\|\mathbf{x}\|^2 + \sum_{i\in\tau\setminus I_\tau} \|\mathbf{a}_i\|^2\|\mathbf{x}\|^2\right) \\
&= \sum_{\tau\in\binom{[m]}{\beta_k}} \left(\sum_{i\in I_\tau} n\sigma^2\|\mathbf{x}\|^2 + \sum_{i\in\tau\setminus I_\tau} \|\mathbf{a}_i\|^2\|\mathbf{x}\|^2\right) \\
&\leq \sum_{\tau\in\binom{[m]}{\beta_k}} \left(\beta_k n\sigma^2 + \sum_{i\in\tau\setminus I_\tau} \|\mathbf{a}_i\|^2\right)\|\mathbf{x}\|^2 \\
&= \binom{m}{\beta_k}\left(\beta_k n\sigma^2 + \sum_{i\in\tau\setminus I_\tau} \|\mathbf{a}_i\|^2\right)\|\mathbf{x}\|^2
\end{aligned}$$

where the first inequality follows from the Cauchy-Schwartz inequality and remaining computation uses the fact that $\mathbb{E}_a\|\mathbf{a}_i\|^2 = n\sigma^2$ and simplifies the expression. The lower bound follows from

$$\begin{aligned}
\sum_{\tau\in\binom{[m]}{\beta}} \mathbb{E}_a\|A_\tau \mathbf{x}\|_\infty^2 &= \sum_{\tau\in\binom{[m]}{\beta}} \mathbb{E}_a \max_{i\in\tau}\langle \mathbf{a}_i, \mathbf{x}\rangle^2 \geq \sum_{\tau\in\binom{[m]}{\beta}} \mathbb{E}_a \max_{i\in I_\tau}\langle \mathbf{a}_i, \mathbf{x}\rangle^2 \\
&\geq \sum_{\tau\in\binom{[m]}{\beta}} \left(\mathbb{E}_a \max_{i\in I_\tau}\langle \mathbf{a}_i, \mathbf{x}\rangle\right)^2 \geq \sum_{\substack{\tau\in\binom{[m]}{\beta} \\ |I_\tau|=\beta_k}} \left(\mathbb{E}_a \max_{i\in I_\tau}\langle \mathbf{a}_i, \mathbf{x}\rangle\right)^2 \\
&\geq \sum_{\substack{\tau\in\binom{[m]}{\beta} \\ |I_\tau|=\beta_k}} \sigma^2\|\mathbf{x}\|^2\log(\beta_k) \\
&\geq \binom{m'}{\beta_k}\sigma^2\|\mathbf{x}\|^2\log(\beta_k),
\end{aligned}$$

where the second to last inequality uses the fact that for i.i.d. Gaussian random variables $g_1, g_2, \ldots g_N \sim \mathcal{N}(0,\sigma^2)$, we have that $\mathbb{E}(\max_{i\in[N]} g_i) \gtrsim \sigma\sqrt{\log N}$ and that $\langle \mathbf{a}_i, \mathbf{x}\rangle \sim \mathcal{N}(0, \sigma^2\|\mathbf{x}\|^2)$. Therefore, we have

$$\begin{aligned}
\gamma_k &= \frac{\sum_{\tau\in\binom{[m]}{\beta}} \mathbb{E}_a\|A_\tau \mathbf{x}\|^2}{\sum_{\tau\in\binom{[m]}{\beta}} \mathbb{E}_a\|A_\tau \mathbf{x}\|_\infty^2} \\
&\leq \frac{\binom{m}{\beta_k}\left(\beta_k n\sigma^2 + \sum_{i\in\tau\setminus I_\tau}\|\mathbf{a}_i\|^2\right)}{\binom{m'}{\beta_k}\sigma^2\log(\beta_k)}.
\end{aligned}$$

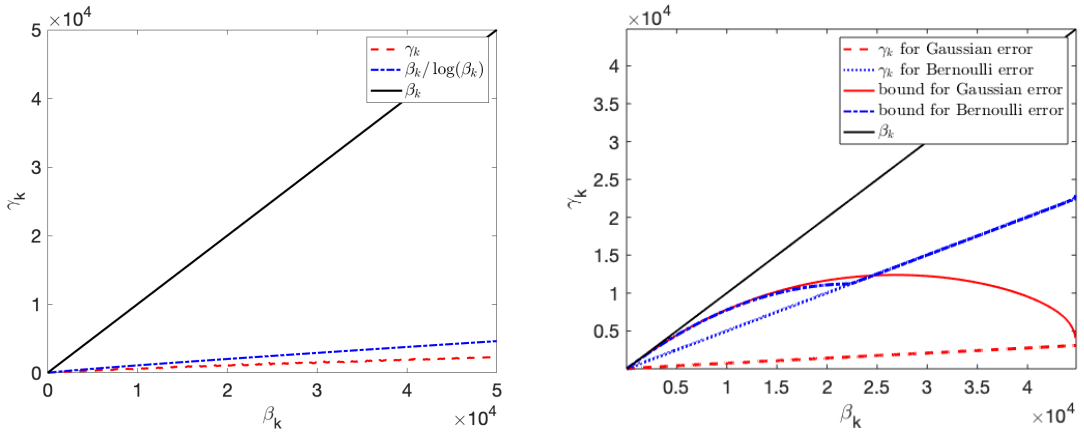


Figure 4.1: Dynamic ranges γ_k for various sample sizes β_k. Left: Gaussian matrix $A \in \mathbb{R}^{50000\times 500}$ and Gaussian error $\mathbf{e}_k \in \mathbb{R}^{500}$ (red). Conjectured bound is plotted in blue. Right: Incidence matrix $Q \in \mathbb{R}^{44850\times 300}$ for complete graph K_{300} with Gaussian error $\mathbf{e}_k \in \mathbb{R}^{300}$ (red) and a sparse error $\mathbf{e}_k$ with Bernoulli random variable entries (blue). Bounds are plotted in the same colors with different line styles.

Dividing all terms by σ^2 attains the desired result (4.1). ∎

We conjecture that the true bound is actually $\mathcal{O}(\beta_k/\log(\beta_k))$ and that the n is an artifact of our proof technique; throughout our experiments varying n (and for various m), we have not found any dependence of γ_k on n. For this reason, we have plotted γ_k and the corresponding conjectured bound in the left of Figure 4.1 for a Gaussian matrix of size 50000×500.

Remark 6. *To extend to other distributions, one can simply note that as the signal dimension n gets large, the Law of Large numbers can be invoked and a similar computation can be used to show an upper bound on the dynamic range of the system.*

With this estimate for γ_k, one can now ask for the optimal choice for β_k in terms of computational time. We can use Theorem 3.1 to estimate the expected number of SKM iterations required so that $\|\mathbf{x}_k - \mathbf{x}^*\|^2/\|\mathbf{x}_0 - \mathbf{x}^*\|^2 \leq \epsilon$ for $\epsilon < 1$; assuming that A is row normalized, we expect this relative error stopping threshold is reached for number of iterations

$$k \geq \frac{\log(\epsilon)}{\log\left(1 - \frac{\beta_k \sigma^2_{\min}(A)}{\gamma_k m}\right)}.$$

To estimate the optimal parameter β_k, we must have an accurate estimate for computational effort of each SKM iteration with sample size β. We first use an estimate for FLOPS required in each iteration of SKM as a proxy for computational time (which is notoriously difficult to estimate); the per-iteration flop requirement for each iteration of SKM is $\mathcal{O}(n\beta_k + n)$.

Therefore, the required FLOPS to reach relative error threshold ϵ is

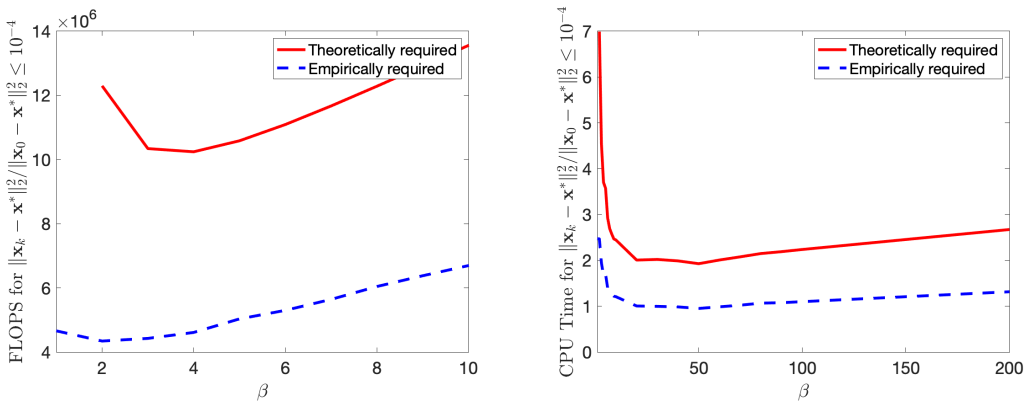
$$\frac{(n\beta_k + n)\log(\epsilon)}{\log\left(1 - \frac{\beta_k \sigma^2_{\min}(A)}{\gamma_k m}\right)} \approx \frac{(n\beta_k + n)\log(\epsilon)}{\log\left(1 - \frac{\log(\beta_k)\sigma^2_{\min}(A)}{nm}\right)},$$

where we have used the estimate $\gamma_k \approx n\beta_k/\log(\beta_k)$. This estimate is plotted for a variety of sample sizes β_k for a row-normalized Gaussian matrix $A \in \mathbb{R}^{50000\times 500}$ with $\epsilon = 10^{-4}$ (solid red line) in Figure 4.2a ; we additionally plot $(n\beta_k+n)k_{\text{emp}}$ where k_{emp} is the number of iterations empirically required to reach relative error stopping threshold $\epsilon = 10^{-4}$ (dashed blue line).

We next estimate computational time required per iteration as the average CPU time for a single iteration of an empirical trial of SKM, t_{β_k}. We estimate the required CPU time to reach relative error threshold ϵ as

$$\frac{t_{\beta_k}\log(\epsilon)}{\log\left(1 - \frac{\beta_k \sigma^2_{\min}(A)}{\gamma_k m}\right)} \approx \frac{t_{\beta_k}\log(\epsilon)}{\log\left(1 - \frac{\log(\beta_k)\sigma^2_{\min}(A)}{nm}\right)}.$$

This estimate is plotted for a variety of sample sizes β_k for a row-normalized Gaussian matrix $A \in \mathbb{R}^{50000\times 500}$ with $\epsilon = 10^{-4}$ (solid red line) in Figure 4.2b; we additionally plot the total CPU time empirically required to reach relative error stopping threshold $\epsilon = 10^{-4}$ (dashed blue line). We note that the estimated optimal choice for β_k differs given the estimates of per-iteration computational burden. This is to be expected as our estimate for FLOPS per SKM iteration ignores computational overhead and communication time that contribute to the CPU time.



(a) Estimate of theoretically required and empirically required FLOPS.

(b) Estimate of theoretically required and empirically required CPU time.

Figure 4.2: Comparison of theoretically and empirically required FLOPS and CPU time to reach given relative error stopping threshold and estimate of optimal β_k for Gaussian system defined by row-normalized $A \in \mathbb{R}^{50000\times 500}$, averaged over five independent trials.

This estimate of the optimal choice for β_k for a given system was dependent upon having an estimate for γ_k that is fixed between iterations and does not depend upon the current iterate. Gaussian systems are the only type where we have such an estimate; in the next section, we provide a bound for systems defined by incidence matrices, but this bound is iterate-dependent. When there is no iteration-consistent estimate for γ_k, one can attempt to choose the optimal β_k using an estimate for γ_k based on the sampled portion of the residual. We take this naive, empirical approach in Section 5 and call this selection strategy 'useDynRng'.

4.2. Incidence Matrices. In the previous subsection, we analyzed the dynamic range for systems with measurement matrices that are randomly generated. Deterministically generated measurement matrices are additionally of interest. In this subsection, we analyze the dynamic range associated to incidence matrices of undirected graphs, $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. The incidence matrix Q associated to an undirected graph is of size $|\mathcal{E}|\times|\mathcal{V}|$. For each edge, $(i, j) \in \mathcal{E}$ which connects vertex i to vertex j, the associated row of Q is all zeros with a one and negative one in the ith and jth entries. These types of matrices arise in one formulation of the *average consensus* problem as a system of linear equations.

The average consensus problem on a graph asks that all nodes on the graph learn the average value of initial, secret values held by each node using only local information; that is, each node i initially knows c_i and at solution they should all know $\frac{1}{|\mathcal{V}|}\sum_{i\in\mathcal{V}} c_i$ with communication only across edges. This problem models computation in many real life applications such as clock synchronization [23], localization without GPS [67], distributed data fusion in sensor networks [66], and load balancing [17]. Many analyses of (asynchronous and synchronous) distributed methods for this problem exploit its formulation as a system of linear equations. The problem over a directed graph may be formulated as a linear system using either the incidence matrix (described above) or the Laplacian matrix, $L = D - A$ where D is the diagonal matrix of node degrees and A is the adjacency matrix, or more generally as an *average consensus system* defined in [39].

The *gossip methods* that solve the average consensus problem are generalized by the Kaczmarz methods [43]. Early work making this connection focused on the formulation of the average consensus problem as a Laplacian system [69], but subsequent work generalized this connection to systems formulated more generally [39]. RK specializes to the *randomized gossip* method in which the pair of nodes which update are selected at random [65, 11]. The connection between gossip methods and other Kaczmarz variants have been observed; the connection to block methods was noted in [39], extended methods in [69], and accelerated methods in [41, 38]. This connection has also spawned new gossip methods; in [34] the authors propose a privacy preserving gossip method, and in [5], the authors propose an accelerated, decentralized gossip method. In [43], the authors summarize many of these advances and connections between the Kaczmarz literature and gossip literature. They first noted and exploited the fact that SKM specializes to a variant of *greedy gossip with eavesdropping (GGE)* in which the nodes are selected from amongst a random sample to maximize the update [63]. By utilizing the connection noted between GGE and SKM in [43] and with some adjustments to the method and theoretical setup here (namely sampling according to the connectivity of the network and redefining the dynamic range accordingly), the approach for proving our main convergence rate can be employed in proving a similar convergence rate for GGE.

Now, we consider the dynamic range for an incidence matrix. We can derive a simple bound on the dynamic range in each iteration that depends only upon the entries of the current error vector, $\mathbf{e}_k := \mathbf{x}_k - \mathbf{x}^*$. In particular,

$$\gamma_k = \frac{\sum_{\tau \in \binom{[m]}{\beta_k}} \|Q_\tau \mathbf{e}_k\|^2}{\|} \sum_{\tau \in \binom{[m]}{\beta_k}} \|Q_\tau \mathbf{e}_k\|_\infty^2 = \frac{\binom{m}{\beta_k} \frac{\beta_k}{m} \sum_{(i,j)\in\mathcal{E}} (e_k^{(i)} - e_k^{(j)})^2}{\sum_{\tau \in \binom{[m]}{\beta_k}} \max_{(i,j)\in\tau} (e_k^{(i)} - e_k^{(j)})^2} \tag{4.2}$$

$$\leq \frac{\beta_k (m - \beta_k + 1) \sum_{(i,j)\in\mathcal{E}} (e_k^{(i)} - e_k^{(j)})^2}{m \sum_{n=\beta_k}^{|\mathcal{E}|} (e_k^{(n_i)} - e_k^{(n_j)})^2},$$

where n_i and n_j denote the vertices connected by the nth smallest magnitude difference across an edge. This bound improves for iterates with a sufficient amount of variation in the coordinates. We have plotted γ_k and the corresponding bounds in the right of Figure 4.1. We calculate these values for the incidence matrix $Q \in \mathbb{R}^{44850\times 300}$ of the complete graph K_{300} in the cases when the error is a Gaussian vector (red) and a Bernoulli vector (blue).

Proposition 4.2. *If the right-hand-side vector associated to the system $Q\mathbf{x} = \mathbf{b}$ is $\mathbf{b} = \mathbf{0}$, as in the average consensus problem, then this bound on the dynamic range is easily computed from the current iterate,*

$$\gamma_k \leq \frac{\beta_k (m - \beta_k + 1) \sum_{(i,j)\in\mathcal{E}} (x_k^{(i)} - x_k^{(j)})^2}{m \sum_{n=\beta_k}^{|\mathcal{E}|} (x_k^{(n_i)} - x_k^{(n_j)})^2}.$$

Remark 7. *We note that this bound on the dynamic range holds for any incidence matrix Q, including those associated with directed graphs. In the case of directed graphs, however, additional assumptions must be made to ensure the well-posedness of the average consensus problem. Additionally, the Kaczmarz methods must be altered to ensure communication in only one direction along edges for directed graphs. Analyses of regular Kaczmarz methods, such as RK or SKM, do not apply to average consensus systems on directed graphs. We leave consideration of Kaczmarz type methods for this variant of the average consensus problem to future work.*

5. Experiments. In this section, we present simulated and real world experiments using SKM for varying sample sizes β. In the simulated experiments, we compare the theoretical convergence guarantees to the empirical performance of SKM, measured by approximation error $\|\mathbf{e}_j\|^2$, averaged over 20 random trials. The number of rows $m = 50000$ and number of columns $n = 500$ are fixed for all simulated experiments. The solution to the system is a vector $\mathbf{x}^* \in \mathbb{R}^n$ where each entry is drawn i.i.d. from a standard Gaussian distribution. In each experiment, the systems are consistent so that $\mathbf{b} = A\mathbf{x}^*$. The sample sizes considered for this experiment are $\beta = \{1, 100, 200, 500, 1000\}$. Unless otherwise stated, the rows of A are uniformly selected without replacement. The experiments presented in this section were are performed in MATLAB 2017b on a MacBook Pro 2015 with a 2.7 GHz Dual-Core Intel Core i5 and 8GB RAM. For practical reasons, we normalize the rows of A and utilize the bound shown in Corollary 2.1.

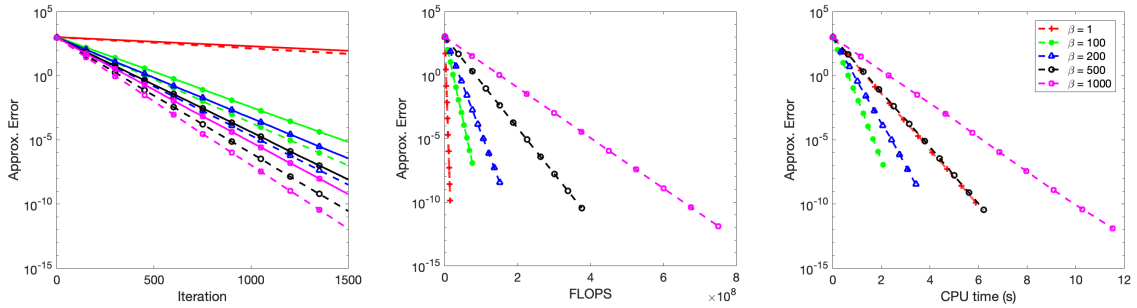


Figure 5.1: Comparison of SKM for various choices of fixed β values on linear system with entries of A drawn from i.i.d. from $\mathcal{N}(0,1)$. (left) Iteration vs Approximation Error with dashed lines representing average empirical performance of SKM and solid lines representing theoretical upper bounds for SKM. (middle) FLOPS vs Approximation Error. (right) CPU time vs Approximation Error

Figure 5.1 and Figure 5.2 show the results for Gaussian and Uniform random matrices A respectively. For Gaussian random matrices, each entry of A is drawn i.i.d. from a standard Gaussian distribution. For Uniform random matrices, entries of A are drawn i.i.d. uniformly from the interval $[0,1]$. In each figure, we plot along the horizontal axis the (left subplot) iteration, (middle subplot) FLOPS or floating point operations, and (right subplot) CPU time in seconds. The vertical axis for all plots indicate the average approximation error across random trials. Note that the left most subplot for both figures also contains a solid line, which indicates the theoretical upper bound of the algorithm provided by Corollary 2.1.

For linear systems with Gaussian random matrices, we see in Figure 5.1 that the convergence upper bound proven in this work closely matches the behavior of the SKM algorithm regardless of the choice of sample size β. To compare this result to previous works, note that when $\beta = 1$, the upper bound provided in Corollary 2.1 simply recovers the previous known upper bound for SKM with normalized rows, a bound which was completely independent of β. In other words, the solid red line is the comparative previous known SKM upper bound for all β.

Of course, choices of large sample sizes β come at a cost, which are captured in the middle and right most subplots of Figure 5.1. When measuring efficiency, it seems that $\beta = 1$ makes the most progress with minimal FLOPS while $\beta = 100$ is optimal amongst the tested sample sizes with respect to CPU time. This difference is typically explained by the programming and computer architecture (e.g., it may be more efficient to work on batches of rows as opposed to single rows at a time).

Figure 5.2 uses a uniform random matrix A instead of a Gaussian random matrix. While the algorithm efficiency with respect to FLOPS and CPU time have similar conclusions to those in the Gaussian measurement matrix case (as one would expect), the iteration vs approximation error plot now tells a different story. Unlike in the Gaussian case, the theoretical upper bound no longer closely tracks the approximation error of SKM. The looseness here comes from lower bounding the norm of $\|A\mathbf{x}\|_2^2$ with the magnitude of $\mathbf{x}$ times the smallest

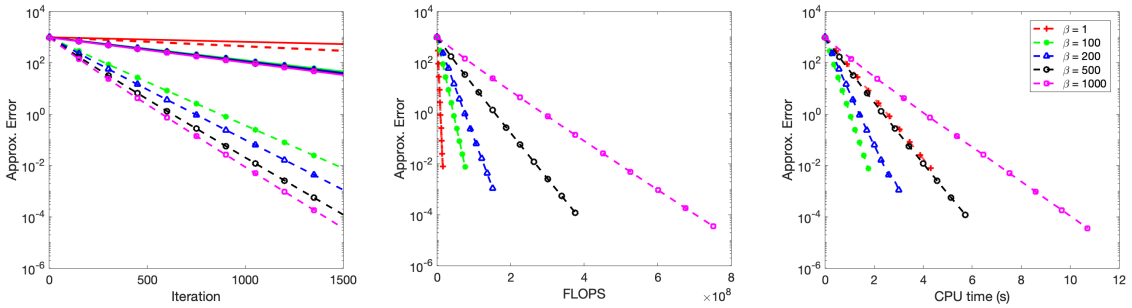


Figure 5.2: Comparison of SKM for various choices of fixed β values on linear system with entries of a A drawn from unif($[0, 1]$). (left) Iteration vs Approximation Error with dashed lines representing average empirical performance of SKM and solid lines representing theoretical upper bounds for SKM. (middle) FLOPS vs Approximation Error. (right) CPU time vs Approximation Error

nonzero singular value of A squared. Empirically, we have seen that this lower bound is tighter for Gaussian systems than Uniform systems. It should be noted that even though our theoretical bounds do not track the approximation error for SKM as tightly, they are still a slight improvement over the previous known bounds for SKM.

In addition to being an improvement over the previously known SKM bound, the convergence bound shown in this work enjoys the flexibility of being amenable to a dynamically selected sample size β_k. Figure 5.3 and Figure 5.4 show the empirical results from experiments where β_k is allowed to change at every iteration. In Figure 5.3 the measurement matrix A is again a random Gaussian matrix and in Figure 5.4 the measurement matrix entries are drawn i.i.d. from Unif($[0, 1]$). We consider three sampling regimes that change β_k at every iteration: 'useDynRng' which allocates β_k as a function of the dynamic range, 'slowInc' which increases β_k at every iteration until $\beta_k = m$, and finally 'rand' which uniformly at random selects a $\beta_k \in [m]$ at every iteration. More specifically, the 'useDynRng' uses the heuristic $\beta_k = \lceil \max(m, \frac{m\|A_{\tau_{k-1}}\mathbf{x}_{k-1} - \mathbf{b}_{\tau_{k-1}}\|_\infty}{n\|A_{\tau_{k-1}}\mathbf{x}_k - \mathbf{b}_{\tau_{k-1}}\|_2}) \rceil$. Note that this choice of β_k relies directly on the inverse of an *approximation* of the dynamic range γ_k computed without incurring additional computational cost, in a naive attempt to optimize the contraction term of the theoretical bound for SKM. Even though β_k changes at each iteration, we see that the theoretical guarantees proven in this work still track the progress of SKM. This indeed opens up new and interesting avenues of research including how one can compute an optimal β_k at every iteration. Since the focus of this work is the improvement of the convergence bound of SKM, we leave this for future work.

Figure 5.5 employs the upper bound on the dynamic range derived in Proposition 3.3 to approximate an upper bound for the error of SKM iterates when $\beta = 100$. Here, we compare the empirical performance of SKM with its previous known upper bound using the contraction term $1 - \frac{\sigma^2_{min}}{m}$ and $1 - \frac{\log(\beta)\sigma^2_{min}}{m}$. Note that we drop the factor of n apparent in Proposition 3.3 as we suspect it to be an artifact of the proof technique used and conjecture that the true

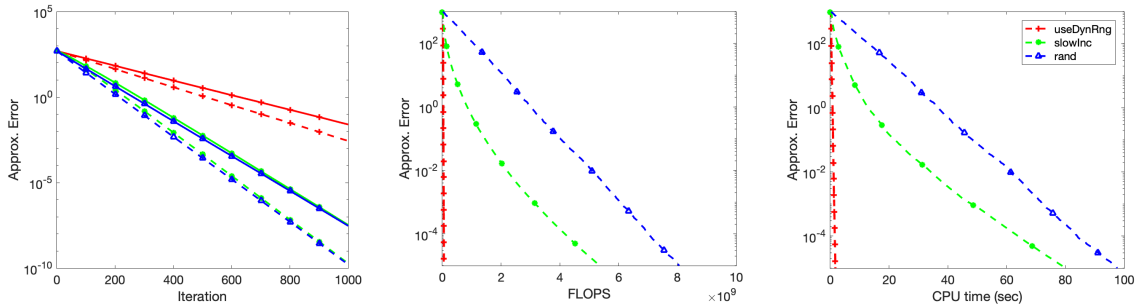


Figure 5.3: Comparison of SKM for various choices of dynamically selected β_k values on linear system with entries of A drawn from $\mathcal{N}(0,1)$.

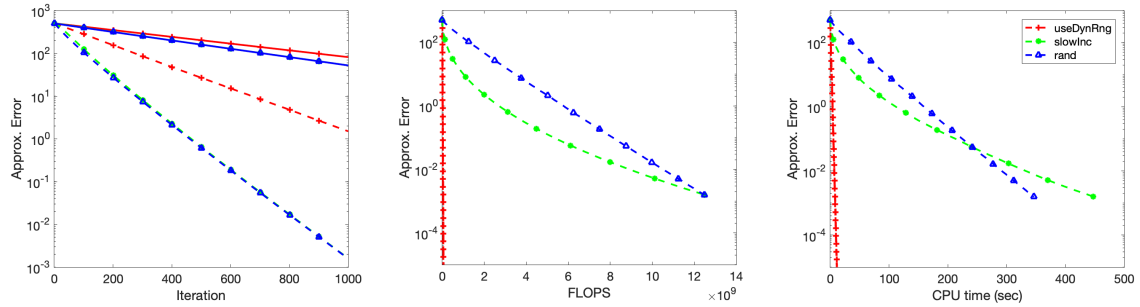


Figure 5.4: Comparison of SKM for various choices of dynamically selected β_k values on linear system with entries of A drawn from Unif$([0,1])$.

upper bound of the dynamic range is actually $\mathcal{O}(\beta/\log(\beta))$.

In both the Gaussian and Uniform synthetic experiments, the row norms of A are of similar magnitude on average and thus choosing β_k rows of the measurement matrix uniformly at random will behave similarly to the theoretically imposed probability distribution introduced in (3.1). In the next experiment, we consider a setting where the entries of the measurement matrix are $a_{ij} \sim \mathcal{N}(0, i/\sqrt{n})$ so that for each row, $\mathbb{E}\|\mathbf{a}_i\|^2 = i$. Since (3.1) is computationally impractical to implement, we will continue to select rows of A uniformly at random without replacement to evaluate the performance of SKM for various choices of β. The results of this experiment are presented in Figure 5.6. As in the previous synthetically generated experiments, $m = 50000$, $n = 500$, and the underlying signal x is a standard Gaussian random vector. Despite not sampling rows as imposed by (3.1), we see that SKM still converges with rates similar to those in Figure 5.1 and $\beta = 100$ outperforms the others with respect to CPU time.

Next we move on to evaluate the performance of SKM on incidence matrices of graphs. We start with the the AC systems discussed in Section 4.2. Here, the graph $\mathcal{G}$ is a complete graph K_{100} with corresponding incidence matrix $Q \in \{-1, 0, 1\}^{4950\times 100}$. The unknown underlying vector $\mathbf{x} \in \mathbb{R}^{100}$ is $\mathbf{x} = \hat{\mu}\mathbf{1}_{100}$ where $\mathbf{1}_{100}$ is a 100-dimensional vector of ones and $\hat{\mu}$ is the

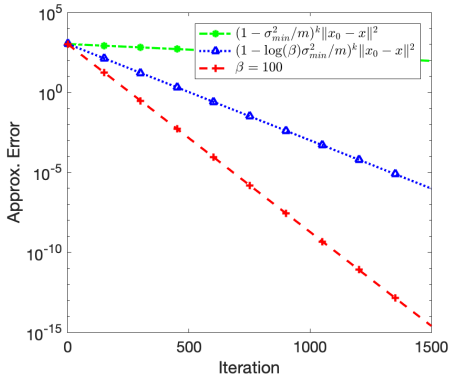


Figure 5.5: Comparison of SKM with $\beta = 100$ with previous known theoretical upper bounds and our upper bound with conjectured Gaussian system dynamic range γ_k.

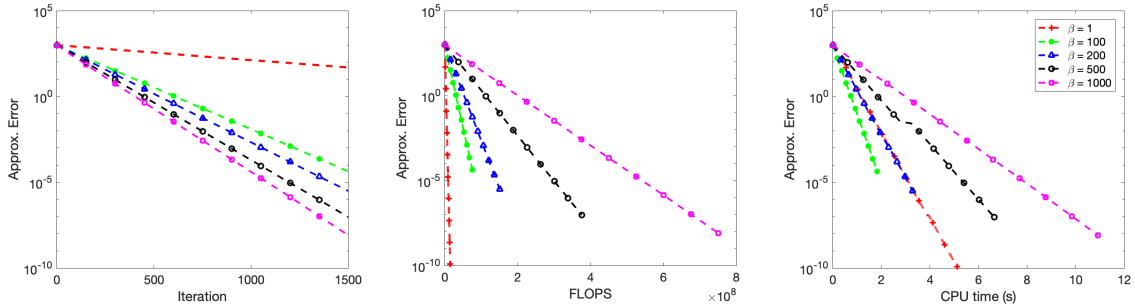


Figure 5.6: Comparison of SKM for various choices of fixed β values on linear system with entries a_{ij} drawn from $\mathcal{N}(0, i/\sqrt{n})$. (left) Iteration vs Approximation Error with dashed lines representing average empirical performance of SKM and solid lines representing theoretical upper bounds for SKM. (middle) FLOPS vs Approximation Error. (right) CPU time vs Approximation Error

empirical average of 100 random draws from a standard normal distribution. The results of this experiment are provided in Figure 5.7.

Figure 5.8 demonstrates the performance of SKM on a graph $\mathcal{G}$ which reflects a scale-free network, i.e., a graph whose degree distribution follows the power law. To create the graph, we employ the implementation of the Barabási-Albert (BA) model [3] with an initial graph of five vertices and ending with a graph of 300 vertices [24]. For more details on scale-free networks, see [3]. In Figure 5.8 we again observe exponential convergence in the mean approximation error and optimal performance with respect to CPU time when $\beta = 10$.

In Figures 5.9a and 5.9b we present the performance of SKM for GGE problems (See Section 4.2). In such problems, instead of randomly selecting a subset of β rows of the incidence matrix of a graph uniformly, we randomly select a node (column), collect all rows corresponding to neighbors of said node (rows corresponding to nonzero entries in that column), and

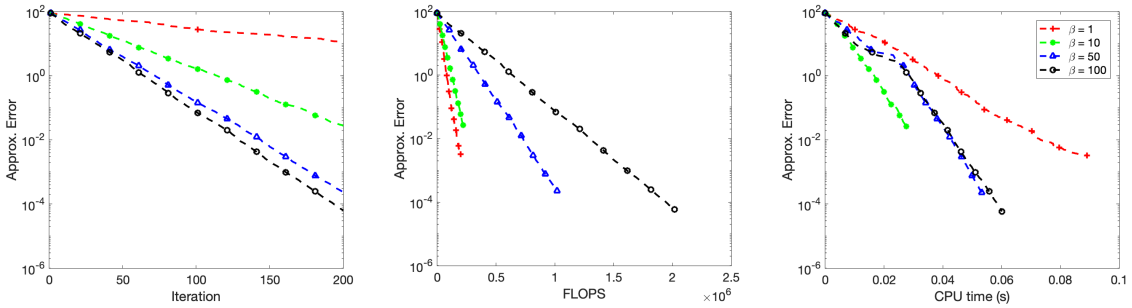


Figure 5.7: Comparison of SKM for various choices of fixed β on AC systems. (left) Iteration vs Approximation Error with dashed lines representing average empirical performance of SKM. (middle) FLOPS vs Approximation Error. (right) CPU time vs Approximation Error

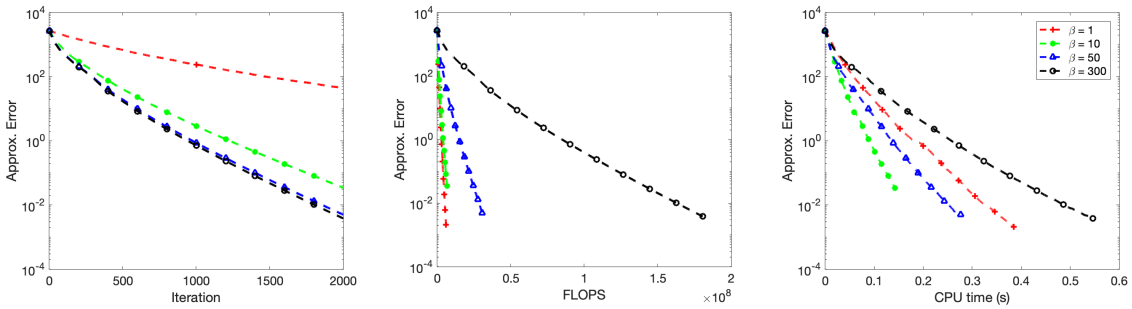
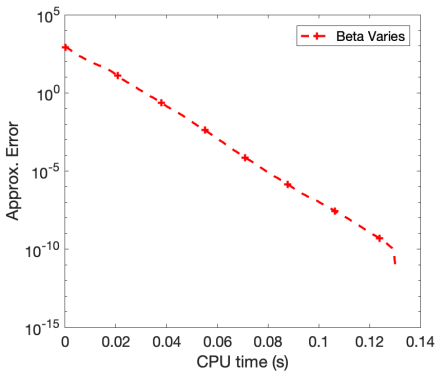


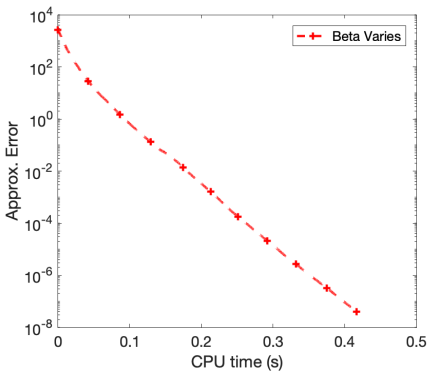
Figure 5.8: Comparison of SKM for various choices of fixed β on an incidence matrix of a scale-free network. (left) Iteration vs Approximation Error with dashed lines representing average empirical performance of SKM. (middle) FLOPS vs Approximation Error. (right) CPU time vs Approximation Error

then of those rows, select the row that returns the max absolute residual value to project on. Note that here, the size of the subset varies at each iteration, depending on the size of the neighborhood of the randomly selected node. Since this is significantly different from the standard SKM selection scheme, we refer to this method as GGE-SKM. Figure 5.9a presents the computational results for an incidence matrix of a complete graph (system set up as in Figure 5.7) while Figure 5.9b does the same for an incidence matrix of a BA model graph (system set up as in Figure 5.8).

Our last experiment demonstrates the performance of SKM with varying choices of β on real world data from the SuiteSparse Matrix Market [18]. For these experiments, we employ the 'Maragal_4', 'well1850', and 'ash958' matrices which are of dimension 1964×1034, 1850×712, and 958×292 respectively. It is useful to note that 'Maragal_4' and 'well1850' have low rank data structures. These matrices are used as the measurement matrix A and the underlying signal is arbitrarily chosen to be in the range of A^T. Here, we allow SKM to run for a maximum of 10^6 iterations or terminate once the approximation error has reached



(a) Performance of CPU time vs Approximation Error of GGE-SKM on an incidence matrix of a complete graph.



(b) Performance of CPU time vs Approximation Error of GGE-SKM on an incidence matrix of a scale-free network.

Figure 5.9: Performance of GGE-SKM on GGE problem with two different graphs.

	SKM $\beta = 1$			**SKM** $\beta = 10$		
	Error	FLOPS	CPU time	Error	FLOPS	CPU time
Maragal_4	2.45	2.07e+09	176.25	0.164	1.13e+10	460.88
well1850	7.67	1.42e+09	112.75	0.064	7.83e+09	238.59
ash958	1.06e-06	4.91e+06	0.323	1.00e-06	5.33e+06	0.126

Table 5.1: Performance of SKM using $\beta = 1$ and $\beta = 10$ on real world data from SuiteSparse Matrix Market.

the allotted error tolerance of 10^{-6}. The results of using $\beta \in \{1, 10, 50\}$ are presented in Table 5.1 and Table 5.2. We also included the performance of Conjugate Gradient Least Squares (GCLS) [10, 61] for comparison. Note that we do not claim to have optimized either implementation.

Table 5.1 and Table 5.2 demonstrates that for a fixed number of iterations, increasing β results in a lower approximation error. It also highlights the trade off between the subset size, the FLOP cost, and the CPU time. As we increase β, in general, both the FLOP and CPU time increase as well. One interesting observation is that for the 'ash958' experiment, the optimal choice of subset size is $\beta = 10$ with respect to CPU time. This is further motivation for future work in optimal β selection. Finally, as expected, CGLS outperforms the three choices of β. However, SKM can be more naturally implemented in distributed computing settings. We leave that direction as an avenue for future work as well.

6. Conclusion.

This work unifies the spectrum between the randomized Kaczmarz and a greedy variant of the Kaczmarz (Motzkin's Method) algorithm by improving the convergence bound of SKM, a hybrid randomized-greedy algorithm. We show that the behavior of SKM

	SKM $\beta = 50$			CGLS		
	Error	FLOPS	CPU time	Error	FLOPS	CPU time
Maragal_4	5.01e-02	5.27e+10	1593.4	3.97e-02	2.97e+07	5.41
well1850	2.49e-03	3.63e+10	1161	1.01e-06	4.06e+06	1.07
ash958	1.01e-06	1.44e+07	0.426	1.88e-06	41158	6.66e-3

Table 5.2: Performance of SKM using $\beta = 50$ and CGLS on real world data from SuiteSparse Matrix Market.

depends on the sample parameter β_k and the dynamic range of the linear system. This result improves upon previous work showing only the linear convergence of SKM. In presenting an improved convergence bound for SKM that highlights the impact of the sub-sample size β_k, we have opened up new and exciting avenues for SKM-type algorithms. Future directions of this work include finding optimal sample sizes for different types of linear systems and designing adaptive sample size selection schemes.

Acknowledgements. The authors would like to thank the manuscript referees for their thoughtful and detailed comments which significantly improved earlier versions of this work. The authors also thank Jacob Moorman, Liza Rebrova, Hanbaek Lyu, Deanna Needell, Jesús A. De Loera, and Roman Vershynin for useful conversations and suggestions.

REFERENCES

[1] S. Agmon, *The relaxation method for linear inequalities*, Canadian J. Math., 6 (1954), pp. 382–392.
[2] R. Aharoni and Y. Censor, *Block-iterative projection methods for parallel computation of solutions to convex feasibility problems*, Linear Algebra Appl., 120 (1989), pp. 165–175, https://doi.org/10.1016/0024-3795(89)90375-3, http://dx.doi.org/10.1016/0024-3795(89)90375-3.
[3] R. Albert and A.-L. Barabási, *Statistical mechanics of complex networks*, Rev. Mod. Phys., 74 (2002), p. 47.
[4] E. Amaldi and R. Hauser, *Boundedness theorems for the relaxation method*, Math. Oper. Res., 30 (2005), pp. 939–955, https://doi.org/10.1287/moor.1050.0164, http://dx.doi.org/10.1287/moor.1050.0164.
[5] N. S. Aybat and M. Gürbüzbalaban, *Decentralized computation of effective resistances and acceleration of consensus algorithms*, in IEEE Glob. Conf. Sig., IEEE, 2017, pp. 538–542.
[6] Z.-Z. Bai and W.-T. Wu, *On greedy randomized Kaczmarz method for solving large sparse linear systems*, SIAM J. Sci. Comput., 40 (2018), pp. A592–A606.
[7] Z.-Z. Bai and W.-T. Wu, *On relaxed greedy randomized Kaczmarz methods for solving large sparse linear systems*, Appl. Math. Lett., 83 (2018), pp. 21–26.
[8] U. Betke, *Relaxation, new combinatorial and polynomial algorithms for the linear feasibility problem*, Discrete Comput. Geom., 32 (2004), pp. 317–338, https://doi.org/10.1007/s00454-004-2878-4, http://dx.doi.org/10.1007/s00454-004-2878-4.
[9] U. Betke and P. Gritzmann, *Projection algorithms for linear programming*, Eur. J. Oper. Res., 60 (1992), pp. 287 – 295, https://doi.org/http://dx.doi.org/10.1016/0377-2217(92)90080-S, http://www.sciencedirect.com/science/article/pii/037722179290080S.
[10] Å. Björck, *Numerical methods for least squares problems*, SIAM, 1996.
[11] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah, *Randomized gossip algorithms*, IEEE ACM T. Network., 14 (2006), pp. 2508–2530.
[12] S. Boyd and L. Vandenberghe, *Convex optimization*, Cambridge university press, 2004.

[13] Y. Censor, *Row-action methods for huge and sparse systems and their applications*, SIAM Rev., 23 (1981), pp. 444–466, https://doi.org/10.1137/1023097, http://dx.doi.org/10.1137/1023097.
[14] Y. Censor, P. P. B. Eggermont, and D. Gordon, *Strong underrelaxation in Kaczmarz's method for inconsistent systems*, Numer. Math., 41 (1983), pp. 83–92.
[15] X. Chen and A. Powell, *Almost sure convergence of the Kaczmarz algorithm with random measurements*, J. Fourier Anal. Appl., (2012), pp. 1–20, http://dx.doi.org/10.1007/s00041-012-9237-2. 10.1007/s00041-012-9237-2.
[16] S. Chubanov, *A polynomial relaxation-type algorithm for linear programming*, Optimization Online, February, (2011).
[17] G. Cybenko, *Dynamic load balancing for distributed memory multiprocessors*, Journal of parallel and distributed computing, 7 (1989), pp. 279–301.
[18] T. A. Davis and Y. Hu, *The University of Florida sparse matrix collection*, ACM T. Math. Software, 38 (2011), pp. 1–25.
[19] J. A. De Loera, J. Haddock, and D. Needell, *A sampling Kaczmarz-Motzkin algorithm for linear feasibility*, SIAM J. Sci. Comput., 39 (2017), pp. S66–S87.
[20] K. Du and H. Gao, *A new theoretical estimate for the convergence rate of the maximal weighted residual Kaczmarz algorithm*, Numer. Math. - Theory Me., 12 (2019), pp. 627–639.
[21] B. Dumitrescu, *On the relation between the randomized extended Kaczmarz algorithm and coordinate descent*, BIT, (2014), pp. 1–11.
[22] Y. C. Eldar and D. Needell, *Acceleration of randomized Kaczmarz method via the Johnson-Lindenstrauss lemma*, Numer. Algorithms, 58 (2011), pp. 163–177, https://doi.org/10.1007/s11075-011-9451-z, http://dx.doi.org/10.1007/s11075-011-9451-z.
[23] N. M. Freris and A. Zouzias, *Fast distributed smoothing of relative measurements*, in 2012 IEEE 51st IEEE Conference on Decision and Control (CDC), IEEE, 2012, pp. 1411–1416.
[24] M. George, *B-A scale-free network generation and visualization*, 2020, https://www.mathworks.com/matlabcentral/fileexchange/11947-b-a-scale-free-network-generation-and-visualization.
[25] J.-L. Goffin, *The relaxation method for solving systems of linear inequalities*, Math. Oper. Res., 5 (1980), pp. 388–414, https://doi.org/10.1287/moor.5.3.388, http://dx.doi.org/10.1287/moor.5.3.388.
[26] J.-L. Goffin, *On the nonpolynomiality of the relaxation method for systems of linear inequalities*, Math. Program., 22 (1982), pp. 93–103, https://doi.org/10.1007/BF01581028, http://dx.doi.org/10.1007/BF01581028.
[27] 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 (1970), pp. 471–481.
[28] R. Gower, D. Molitor, J. Moorman, and D. Needell, *Adaptive sketch-and-project methods for solving linear systems*, arXiv preprint arXiv:1909.03604, (2019).
[29] R. M. Gower and P. Richtárik, *Randomized iterative methods for linear systems*, SIAM J. Matrix Anal. A., 36 (2015), pp. 1660–1690.
[30] R. M. Gower and P. Richtárik, *Stochastic dual ascent for solving linear systems*, arXiv preprint arXiv:1512.06890, (2015).
[31] J. Haddock and D. Needell, *On Motzkin's method for inconsistent linear systems*, BIT, 59 (2019), pp. 387–401.
[32] J. Haddock and D. Needell, *Randomized projection methods for linear systems with arbitrarily large sparse corruptions*, SIAM J. Sci. Comput., 41 (2019), pp. S19–S36.
[33] M. Hanke and W. Niethammer, *On the acceleration of Kaczmarz's method for inconsistent linear systems*, Linear Algebra Appl., 130 (1990), pp. 83–98.
[34] F. Hanzely, J. Konečnỳ, N. Loizou, P. Richtárik, and D. Grishchenko, *Privacy preserving randomized gossip algorithms*, arXiv preprint arXiv:1706.07636, (2017).
[35] G. Herman and L. Meyer, *Algebraic reconstruction techniques can be made computationally efficient*, IEEE T. Med. Imaging, 12 (1993), pp. 600–609.
[36] S. Kaczmarz, *Angenäherte auflösung von systemen linearer gleichungen*, Bull. Int. Acad. Polon. Sci. Lett. Ser. A, (1937), pp. 335–357.
[37] J. Liu, S. J. Wright, and S. Sridhar, *An asynchronous parallel randomized Kaczmarz algorithm*, arXiv preprint arXiv:1401.4780, (2014).
[38] N. Loizou, M. Rabbat, and P. Richtárik, *Provably accelerated randomized gossip algorithms*, in Int.

Conf. Acoust. Spee., IEEE, 2019, pp. 7505–7509.
[39] N. Loizou and P. Richtárik, *A new perspective on randomized gossip algorithms*, in IEEE Glob. Conf. Sig., IEEE, 2016, pp. 440–444.
[40] N. Loizou and P. Richtárik, *Momentum and stochastic momentum for stochastic gradient, Newton, proximal point and subspace descent methods*, arXiv preprint arXiv:1712.09677, (2017).
[41] N. Loizou and P. Richtárik, *Accelerated gossip via stochastic heavy ball method*, in 2018 56th Annual Allerton Conference on Communication, Control, and Computing (Allerton), IEEE, 2018, pp. 927–934.
[42] N. Loizou and P. Richtárik, *Convergence analysis of inexact randomized iterative methods*, arXiv preprint arXiv:1903.07971, (2019).
[43] N. Loizou and P. Richtárik, *Revisiting randomized gossip algorithms: General framework, convergence rates and novel block and accelerated protocols*, arXiv preprint arXiv:1905.08645, (2019).
[44] A. Ma, D. Needell, and A. Ramdas, *Convergence properties of the randomized extended Gauss–Seidel and Kaczmarz methods*, SIAM J. Matrix Anal. A., 36 (2015), pp. 1590–1604.
[45] J. D. Moorman, T. K. Tu, D. Molitor, and D. Needell, *Randomized Kaczmarz with averaging*, arXiv preprint arXiv:2002.04126, (2020).
[46] M. S. Morshed, M. S. Islam, et al., *On generalization and acceleration of randomized projection methods for linear feasibility problems*, arXiv preprint arXiv:2002.07321, (2020).
[47] M. S. Morshed, M. S. Islam, and M. Noor-E-Alam, *Accelerated sampling Kaczmarz Motzkin algorithm for the linear feasibility problem*, J. Global Optim., (2019), pp. 1–22.
[48] T. S. Motzkin and I. J. Schoenberg, *The relaxation method for linear inequalities*, Canadian J. Math., 6 (1954), pp. 393–404.
[49] D. Needell, *Randomized Kaczmarz solver for noisy linear systems*, BIT, 50 (2010), pp. 395–403, https://doi.org/10.1007/s10543-010-0265-5, http://dx.doi.org/10.1007/s10543-010-0265-5.
[50] D. Needell and E. Rebrova, *On block Gaussian sketching for iterative projections*, arXiv preprint arXiv:1905.08894, (2019).
[51] D. Needell, N. Srebro, and R. Ward, *Stochastic gradient descent and the randomized Kaczmarz algorithm*, Math. Program. A, 155 (2016), pp. 549–573.
[52] D. Needell and J. A. Tropp, *Paved with good intentions: Analysis of a randomized block Kaczmarz method*, Linear Algebra Appl., (2013).
[53] D. Needell, R. Zhao, and A. Zouzias, *Randomized block Kaczmarz method with projection for solving least squares*, Linear Algebra Appl., 484 (2015), pp. 322–343.
[54] J. Nutini, B. Sepehry, A. Virani, I. Laradji, M. Schmidt, and H. Koepke, *Convergence Rates for Greedy Kaczmarz Algorithms*, UAI, (2016).
[55] S. Petra and C. Popa, *Single projection Kaczmarz extended algorithms*, Numer. Algorithms, (2015), pp. 1–16, https://doi.org/10.1007/s11075-016-0118-7, https://arxiv.org/abs/1504.00231.
[56] C. Popa, *A fast Kaczmarz-Kovarik algorithm for consistent least-squares problems*, Korean J. Comput. Appl. Math., 8 (2001), pp. 9–26.
[57] C. Popa, *A Kaczmarz-Kovarik algorithm for symmetric ill-conditioned matrices*, An. Ştiinţ. Univ. Ovidius Constanţa Ser. Mat., 12 (2004), pp. 135–146.
[58] C. Popa, T. Preclik, H. Köstler, and U. Rüde, *On Kaczmarz's projection iteration as a direct solver for linear least squares problems*, Linear Algebra Appl., 436 (2012), pp. 389–404.
[59] E. Rebrova and D. Needell, *Sketching for Motzkin's iterative method for linear systems*, Proc. 50th Asilomar Conf. on Signals, Systems and Computers, (2019).
[60] T. Strohmer and R. Vershynin, *A randomized Kaczmarz algorithm with exponential convergence*, J. Fourier Anal. Appl., 15 (2009), pp. 262–278.
[61] Systems Optimization Laboratory, *CGLS: CG method for $Ax = b$ and least squares*, https://web.stanford.edu/group/SOL/software/cgls/.
[62] J. Telgen, *On relaxation methods for systems of linear inequalities*, Eur. J. Oper. Res., 9 (1982), pp. 184–189, https://doi.org/10.1016/0377-2217(82)90071-6, http://dx.doi.org/10.1016/0377-2217(82)90071-6.
[63] D. Ustebay, B. N. Oreshkin, M. J. Coates, and M. G. Rabbat, *Greedy gossip with eavesdropping*, IEEE T. Signal Proces., 58 (2010), pp. 3765–3776.
[64] H. Xiang and L. Zhang, *Randomized iterative methods with alternating projections*, arXiv preprint

arXiv:1708.09845, (2017).
[65] L. Xiao and S. Boyd, *Fast linear iterations for distributed averaging*, Syst. Control Lett., 53 (2004), pp. 65–78.
[66] L. Xiao, S. Boyd, and S. Lall, *A scheme for robust distributed sensor fusion based on average consensus*, in IPSN 2005. Fourth International Symposium on Information Processing in Sensor Networks, 2005., IEEE, 2005, pp. 63–70.
[67] J. Zhang, J. Cui, Z. Wang, Y. Ding, and Y. Xia, *Distributed joint cooperative self-localization and target tracking algorithm for mobile networks*, Sensors, 19 (2019), p. 3829.
[68] A. Zouzias and N. M. Freris, *Randomized extended Kaczmarz for solving least squares*, SIAM J. Matrix Anal. A., 34 (2013), pp. 773–793.
[69] A. Zouzias and N. M. Freris, *Randomized gossip algorithms for solving Laplacian systems*, in 2015 European Control Conference (ECC), IEEE, 2015, pp. 1920–1925.