SIAM J. MATRIX ANAL. APPL.
Vol. 44, No. 3, pp. 1349–1369

RANDOMIZED BLOCK ADAPTIVE LINEAR SYSTEM SOLVERS[*]

VIVAK PATEL[†], MOHAMMAD JAHANGOSHAHI[‡], AND D. ADRIAN MALDONADO[§]

Abstract. Randomized linear solvers randomly compress and solve a linear system with compelling theoretical convergence rates and computational complexities. However, such solvers suffer a substantial disconnect between their theoretical rates and actual efficiency in practice. Fortunately, these solvers are quite flexible and can be adapted to specific problems and computing environments to ensure high efficiency in practice, even at the cost of lower effectiveness (i.e., having a slower theoretical rate of convergence). While highly efficient adapted solvers can be readily designed by application experts, will such solvers still converge and at what rate? To answer this, we distill three general criteria for randomized adaptive solvers, which, as we show, will guarantee a worst-case exponential rate of convergence of the solver applied to consistent and inconsistent linear systems irrespective of whether such systems are overdetermined, underdetermined, or rank deficient. As a result, we enable application experts to design randomized adaptive solvers that achieve efficiency and can be verified for effectiveness using our theory. We demonstrate our theory on 26 solvers, nine of which are novel or novel block extensions of existing methods to the best of our knowledge.

Key words. block solvers, adaptive solvers, randomized solvers, linear systems, sketching

MSC codes. 15A06, 15B52, 65F10, 65F25, 65N75, 65Y05, 68W20, 68W40

DOI. 10.1137/22M1488715

1. Introduction.

Solving linear systems and solving least squares problems remain critical operations in scientific and engineering applications. As the size of systems or the sheer number of systems that need to be solved grows, faster and approximate linear solvers have become essential to scalability. Recently, randomized linear solvers have become of interest as they can compress the information in the original system in a problem-blind fashion, which can then be used to inexpensively and approximately solve the original system [36]. Moreover, by iterating on this procedure, randomized linear solvers will converge exponentially fast to the solution of the original system [26]. In fact, a rather simple randomized linear system solver was recently shown to achieve a *universal* exponential rate of convergence for any consistent linear system with high probability [33].

Despite such an incredible result, as we show through a salient example (see section 2), randomized linear solvers suffer a substantial disconnect between their convergence rate theory and actual efficiency in practice because they often violate simple computing principles (e.g., the locality principle [6]). Briefly, in the example in section 2, an "oracle" linear solver inspired by [33] is applied to a specific $10^7 \times 100$ system such that it only requires 100 arithmetic operations to find a solution with absolute error of 10^{-16}, yet it is *slower* than block Kaczmarz—which,

[*] Received by the editors April 5, 2022; accepted for publication (in revised form) by P. Drineas April 27, 2023; published electronically September 6, 2023.
https://doi.org/10.1137/22M1488715

Funding: The work of the authors was supported by University of Wisconsin–Madison WARF award AAD5914 and DOE grant DE-AC02-06CH11347.

[†] Department of Statistics, University of Wisconsin–Madison, Madison, WI 53706 USA (vivak.patel@wisc.edu).

[‡] Susquehanna International Group, Bala Cynwyd, PA 19004 USA (mjahangoshahi@uchicago.edu).

[§] Department of Mathematics and Computer Science, Argonne National Laboratories, Lemont, IL 60439 USA (maldonadod@anl.gov).

in theory, requires over 10^{10} arithmetic operations to find such a solution—because of access patterns that violate data locality. Unfortunately, nearly all variations of such linear solvers that exist [1, 18, 34, 2, 38, 11, 21, 22, 23, 3, 12, 19, 10, 29, 32] can be shown to suffer from this disconnect between their theoretical convergence rates and actual efficiency by specific choices of the linear system, software environment, or hardware.

A pessimistic view of these solvers would imply that they should be wholly abandoned. An alternative perspective would suggest a better prognosis: because of the adaptability of such solvers, they can be highly tailored to specific linear systems, software environments, and hardware to achieve high efficiency even at the expense of worse theoretical convergence rates. This latter view is the one adopted in this work.

A bevy of adapted methods can be designed and deployed by atomizing, composing, and customizing key components of randomized linear solvers.[1] Owing to the freedom of creating such solvers, understanding whether the efficient highly adapted method will still converge and at what cost to the rate (e.g., will the solver now converge subexponentially?) becomes integral to a practitioner's decision to implement the method.

To address this consideration, a handful of adaptive solvers were shown to retain exponential convergence by [10], but in a limited context: the set of projections must be finite, and the exactness assumption [30, Assumption 2] must be satisfied, which is generally difficult to verify in practice.[2] In our previous work [25, 24], adaptive solvers relying on vector operations were shown to retain exponential convergence. While our previous work accounted for a number of existing solvers (e.g., [34, 38, 11, 32, 1, 18, 3, 12]), adaptive solvers using high-efficiency block operations did not fall within our results. As block operations have been critical to achieving high efficiency in traditional factorizations (e.g., QR [8, Chap. 5]), in classical iterative methods (e.g., Krylov Iterations [31, Chap. 6]), in randomized factorization methods [15, sect. 16.2], and on GPUs [4], adaptive solvers using block operations must be shown to retain exponential convergence.

Therefore, in this work, we provide generic sufficient conditions that if satisfied by a randomized block adaptive solver (RBAS) will guarantee a worst-case (i.e., with probability one) exponential rate of convergence.[3] In particular, we provide these generic sufficient conditions and consequent worst-case exponential convergence rates in two contexts:

1. for row-action RBASs on consistent linear systems, which may be overdetermined, underdetermined, or rank deficient (see Corollaries 3.8 and 3.10), and
2. for column-action RBASs for linear least squares problems, which may be overdetermined, underdetermined, or rank deficient (see Corollaries 3.18 and 3.20).

We then show how to apply these results to 26 different solvers, nine of which—to the best of our knowledge—are either novel or novel block-operation extensions of existing methods. Thus, in this work, we give end-users the tools to design effective solvers for their specific problems and environments.

[1] We are implementing a software package to enable this approach. See https://github.com/numlinalg/RLinearAlgebra.jl.

[2] See subsection SM1.19 on how we can eliminate this assumption for an important class of methods.

[3] Other worst-case rates can be provided using similar ideas that we present herein, but we do not know of a context where such rates are useful.

The remainder of this work is organized as follows. In section 2, we demonstrate the disconnect between rates of convergence and efficiency. In section 3, we present the two archetype RBASs, provide examples for each, state and discuss the refined properties that such solvers satisfy, and state our convergence results for each type. In section 4, we provide a common formulation for the two types of RBASs, prove the convergence of these methods using this common formulation, and interlace numerical experiments that demonstrate key parts of the theory. In section 5, we show how to apply our convergence theory to a variety of existing and novel RBASs, and we provide numerical experiments where appropriate. In section 6, we conclude.

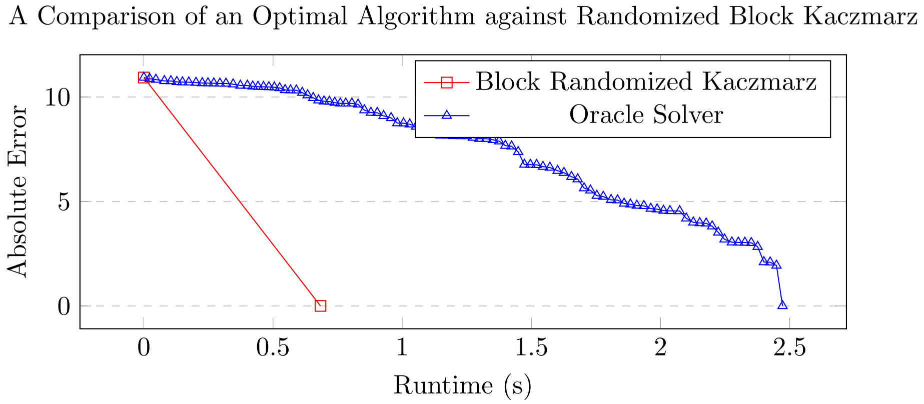
2. Counter example.

Here, we demonstrate that the theoretical convergence rates of randomized solvers can be quite disconnected from their actual efficiency in practice. Consider a consistent, linear system with $n = 10^7$ equations and $d = 100$ unknowns represented with double precision. Owing to the size of the system relative to the 4 gigabytes of memory available on an Intel i5 8th Generation CPU computer, the system is split into 0.5 gigabyte chunks, which contain at most 66,666 equations each.

Consider an "oracle" solver inspired by [33], which can randomly replace d equations in the original system in such a way that the coefficients of the resulting replaced d equations correspond to the rows of the $d \times d$ identity matrix and the system is still consistent. Then, with knowledge of the index of these d equations, the solver applies Kaczmarz to these rows to solve the system. As a result, the oracle solver requires d iterations and $\mathcal{O}(d)$ arithmetic operations. For this specific example, the oracle solver requires about 100 arithmetic operations.

Consider an alternative solver, the random block Kaczmarz solver, which will randomly choose a chunk from the system and perform a block updated to its iterate. In our example, a single block Kaczmarz update requires approximately 10^7 arithmetic operations, and, with an expected squared error rate of convergence of at least 0.993 [21, Theorem 1.2], it will required over 5,500 iterations and, correspondingly, over 5×10^{10} operations to achieve an expected absolute *squared* error of 10^{-16}.

Clearly, from a theoretical perspective, the "oracle" solver is substantially faster than the random block Kaczmarz solver as the former requires 10-fold fewer iterations and 10^8 fewer operations. However, when applied to the system, the "oracle" solver is trounced by random block Kacmzarz (see Figure 1). To understand this, the "oracle"



ht

FIG. 1. *A comparison runtime of the "oracle" algorithm against block Kaczmarz for the described system. The optimal algorithm achieves an absolute error of* 0 *in* 100 *iterations requiring* 2.47 *seconds. Block Kaczmarz achieves an absolute error of* 10^{-15} *in* 1 *iteration requiring* 0.68 *seconds.*

solver needs to read in a new chunk (in expectation and in reality) to access the equations that it has embedded, which is highly expensive as it violates data locality. On the other hand, the block Kaczmarz solver simply does what it can with the information that is given in a single chunk, which turns out to contain sufficient information for finding a high quality solution in one iteration. To summarize, these solvers behave very differently in their theoretical convergence rates and in practice, as this example shows.

This observation is motivation to adapt such solvers to ensure that they are efficient for specific problems and computing environments. In this work, we provide sufficient conditions that, if satisfied by an adapted solver, will be effective—that is, the solver will have a worst-case exponential rate of convergence.

3. Randomized block adaptive solvers.

Consider solving the consistent linear system

$$Ax = b, \tag{3.1}$$

or consider finding the least squares solution for a (possibly) inconsistent system by solving

$$\min_x \|Ax - b\|_2, \tag{3.2}$$

where $A \in \mathbb{R}^{n\times d}$, $x \in \mathbb{R}^d$, and $b \in \mathbb{R}^n$. We emphasize *we have not* required that $n < d$, $n > d$ or that A has full rank; in other words, we allow for underdetermined systems, overdetermined systems, and rank deficient linear systems. To solve these systems, we will consider two archetypes of RBAS methods: row-action RBAS methods for (3.1) and column-action RBAS methods for (3.1) and (3.2). We will define each variation below, provide examples, state the assumptions, and present the main convergence results.

3.1. Row-action RBASs.

For row-action methods, we will need to assume as follows.

Assumption 3.1. The system (3.1) is consistent. That is, the set $\mathcal{H} := \{x \in \mathbb{R}^d : Ax = b\}$ is nonempty.

With this assumption, we begin with an iterate $x_0 \in \mathbb{R}^d$ and some prior information, encapsulated by $\zeta_{-1} \in \mathfrak{Z}$, where $\mathfrak{Z}$ is finite in some sense (e.g., the product of a finite set and a finite dimensional linear space). We then generate a sequence of iterates, $\{x_k : k \in \mathbb{N}\}$, according to

$$x_{k+1} = x_k - A^\intercal W_k (W_k^\intercal A A^\intercal W_k)^\dagger W_k^\intercal (Ax_k - b), \tag{3.3}$$

where $\cdot^\dagger$ represents a pseudoinverse, and $\{W_k \in \mathbb{R}^{n\times n_k}\}$ are possibly random quantities (i.e., vectors or matrices) generated according to a possibly random, adaptive procedure, φ_R, which supplies

$$W_k, \zeta_k = \varphi_R(A, b, \{x_j : j \leq k\}, \{W_j : j < k\}, \{\zeta_j : j < k\}) \in \mathbb{R}^{n\times n_k} \times \mathfrak{Z}. \tag{3.4}$$

We make several comments about this procedure. First, n_k can be selected adaptively so long as it is known given the arguments of φ_R. Second, ζ_k contains information generated from previous iterations that may be essential to the operation of the adaptive procedure (see examples below). Third, we can change the inner product space as is done in [11] without issue (see subsection SM1.19). The next examples illustrate this formulation of row-action RBASs.

Example 3.2 (cyclic vector Kaczmarz). The cyclic vector Kaczmarz method cycles through the equations of $Ax = b$ (without reordering) and updates the current iterate by projecting it onto the hyperplane that solves the selected equation. To rephrase the cyclic vector Kaczmarz method in our framework, let $\{e_i : i = 1, \ldots, n\}$ denote the standard basis elements of $\mathbb{R}^n$. Moreover, let $\mathfrak{Z} = \{0\} \cup \mathbb{N}$, and $\zeta_{-1} = 0$. We then define φ_R to be

$$\varphi_R(A, b, \{x_j : j \le k\}, \{W_j : j < k\}, \{\zeta_j : j < k\}) = (e_{\text{rem}(\zeta_{k-1}, n)+1}, \zeta_{k-1} + 1). \tag{3.5}$$

With this choice of (W_k, ζ_k), we readily see that the described cyclic vector Kaczmarz method is equivalent to

$$x_{k+1} = x_k - A^{\intercal} e_{\text{rem}(\zeta_{k-1},n)+1} \frac{e^{\intercal}_{\text{rem}(\zeta_{k-1},n)+1}(Ax_k - b)}{\left\| A^{\intercal} e_{\text{rem}(\zeta_{k-1},n)+1} \right\|_2^2}, \tag{3.6}$$

which is exactly (3.3). We highlight that φ_R only depends on ζ_{k-1} and the number of equations in the linear system, which will be important in our discussion below.

Example 3.3 (random permutation block Kaczmarz). The random permutation block Kaczmarz method partitions the equations of $Ax = b$ (not necessarily equal partitions) into blocks of equations, generates a random permutation of the blocks, selects a block by cycling through the permutation, updates the current iterate by projecting it onto the hyperplane that solves all of the equations in the block, and, if the random permutation is exhausted, generates a new random permutation of the blocks.

To rephrase this method in our framework, let $\{E_i\}$ be matrices whose columns are generated by some partitioning of the identity matrix in $\mathbb{R}^{n \times n}$, and let $\epsilon = |\{E_i\}|$. Moreover, let $\mathfrak{Z}$ be product of the set of all permutations of $\{1, \ldots, \epsilon\}$ with the empty set, and $\{0\} \cup \mathbb{N}$. Let $\{Z_k : k + 1 \in \mathbb{N}\}$ be independent random permutations of $\{1, \ldots, \epsilon\}$. Let $\zeta_{-1} = (Z_0, 0)$. Then, we can define φ_R to be

$$\begin{aligned} &\varphi_R(A, b, \{x_j : j \le k\}, \{W_j : j < k\}, \{\zeta_j : j < k\}) \\ &\quad = \begin{cases} (E_{\zeta_{k-1}[1][\text{rem}(\zeta_{k-1}[2], \epsilon)+1]}, (\zeta_{k-1}[1], \zeta_{k-1}[2] + 1)), & \text{rem}(\zeta_{k-1}[2], \epsilon) < \epsilon - 1, \\ (E_{\zeta_{k-1}[1][\epsilon]}, (Z_{\text{div}(\zeta_{k-1}[2]+1, \epsilon)}, \zeta_{k-1}[2] + 1)), & \text{rem}(\zeta_{k-1}[2], \epsilon) = \epsilon - 1, \end{cases} \end{aligned} \tag{3.7}$$

where $\zeta_k[1]$ is the permutation component of ζ_k, $\zeta_k[1][j]$ is the jth element of the permutation, and $\zeta_k[2]$ is the iteration counter. With this choice of (W_k, ζ_k), it is easy to see that the random permutation block Kaczmarz method can be equivalently written as (3.3). We highlight that φ_R only depends on ζ_{k-1}, the partitioning of the identity matrix, and the size of the partition.

Example 3.4 (greedy block selection Kaczmarz). This method partitions the equations of $Ax = b$, computes the residual norm of each block at the given iteration, selects the block with the largest residual norm, and updates the current iterate by projecting it onto the hyperplane that solves all of the equations in the block.

To rephrase this method in our framework, let $\{E_i\}$ be matrices whose columns are generated by some partitioning of the $n \times n$ identity matrix, and let ϵ be the size of this set. Moreover, let $\mathfrak{Z} = \{\emptyset\}$, $\zeta_{-1} = \emptyset$, and let

$$\pi(k) = \underset{i=1,\ldots,\epsilon}{\operatorname{argmax}} \left\| E_i^\intercal (Ax_k - b) \right\|_2 . \tag{3.8}$$

Then, we can define φ_R to be

$$\varphi_R(A, b, \{x_j : j \leq k\}, \{W_j : j < k\}, \{\zeta_j : j < k\}) = (E_{\pi(k)}, \emptyset). \tag{3.9}$$

With this choice of (W_k, ζ_k), it is easy to see that this method is of the form (3.3). We emphasize that φ_R only depends on A, b, x_{k-1}, and the partitioning of the identity matrix.

One of the key properties that is apparent in the examples above is that they are *forgetful.* In other words, the choice of (W_k, ζ_k) only depends on some finite number of previous iterations. To state this formally, for all $j+1 \in \mathbb{N}$ and $k \in [1, j+1] \cap \mathbb{N}$, let

$$\mathcal{F}_k^j = \sigma(\zeta_{j-k}, x_{j-k+1}, W_{j-k+1}, \ldots, W_{j-1}, \zeta_{j-1}, x_j), \tag{3.10}$$

that is, the σ-algebra generated by the random variables indicated. Note that we take $\mathcal{F}_1^j = \sigma(\zeta_{j-1}, x_j)$ and $\mathcal{F}_0^j$ to be the trivial σ-algebra. Then, we can formalize this forgetfulness property as follows.

Definition 3.5 (Markovian). *A row-action RBAS is Markovian if there exists a finite $M \in \mathbb{N}$ such that for any measurable sets $\mathcal{W} \subset \mathbb{R}^{n \times n_k}$ and $\mathcal{Z} \subset \mathfrak{Z}$,*

$$\mathbb{P}\left[W_k \in \mathcal{W}, \zeta_k \in \mathcal{Z} \middle| \mathcal{F}_{k+1}^k\right] = \mathbb{P}\left[W_k \in \mathcal{W}, \zeta_k \in \mathcal{Z} \middle| \mathcal{F}_{\min\{M, k+1\}}^k\right]. \tag{3.11}$$

Remark 3.6. As discussed in [17, Chap. 3], a Markov process that depends on some extended period of information can be rewritten into a Markov process that depends on the most recent information only, which can be achieved by expanding the state of the Markov process. For a Markovian RBAS, we can do the same by adding this information in ζ_k, so long as we ensure that $\mathfrak{Z}$ is finite. Thus, the value of M in the preceding definition can always be taken as 1. We also note that if $\mathfrak{Z}$ is finite, then it cannot be used to store all previous iterates.

Another key property of the above examples is that either the iterate will be updated within some reasonable amount of time or the current iterate is the solution. For instance, in the random permutation block Kaczmarz method, if x_0 is not a solution, then within ϵ iterations from $k = 0$, we will find an $E_i^\intercal (Ax_0 - b) \neq 0$. As a result, x_0 will eventually be updated. We can generalize this property as follows.

Definition 3.7 (N, π-exploratory). *A row-action RBAS is N, π-exploratory for some $N \in \mathbb{N}$ and $\pi \in (0, 1]$ if*

$$\sup_{\substack{x_0 \in \mathbb{R}^d : x_0 \neq \mathcal{P}_{\mathcal{H}} x_0 \\ \zeta_{-1} \in \mathfrak{Z}}} \mathbb{P}\left[\bigcap_{j=0}^{N-1} \{\operatorname{col}(A^\intercal W_j) \perp x_0 - \mathcal{P}_{\mathcal{H}} x_0\} \middle| \mathcal{F}_1^0 \right] \leq 1 - \pi. \tag{3.12}$$

Here, we come to a bifurcation point in the theory of RBAS methods based on whether $\{\operatorname{col}(A^\intercal W_k)\}$ is a finite set or if it is an infinite set. In all of the examples above, $\{\operatorname{col}(A^\intercal W_k)\}$ belong to a finite set. In this case, we have the following result.

COROLLARY 3.8. *Let $A \in \mathbb{R}^{n\times d}$ and $b \in \mathbb{R}^n$, satisfying Assumption* 3.1. *Let $x_0 \in \mathbb{R}^d$ and $\zeta_{-1} \in \mathfrak{Z}$. Let $\{x_k : k \in \mathbb{N}\}$ be a sequence generated by* (3.3) *and* (3.4) *satisfying Definitions* 3.5 *and* 3.7 *for some $N \in \mathbb{N}$ and $\pi \in (0,1]$. If the elements of $\{\text{col}(A^\intercal W_k) : k+1 \in \mathbb{N}\}$ take value in a finite set, then either*

1. *there exists a stopping time τ with finite expectation such that $x_\tau = \mathcal{P}_\mathcal{H} x_0$, or*
2. *there exists a sequence of nonnegative stopping times $\{\tau_j : j+1 \in \mathbb{N}\}$ for which $\mathbb{E}[\tau_j] \leq j[(\text{rank}(A)-1)(N/\pi)+1]$, and there exist $\gamma \in (0,1)$ and a sequence of random variables $\{\gamma_j : j+1 \in \mathbb{N}\} \subset (0,\gamma]$ such that*

$$\mathbb{P}\left[\bigcap_{j=0}^{\infty}\left\{\left\|x_{\tau_j} - \mathcal{P}_\mathcal{H} x_0\right\|_2^2 \leq \left(\prod_{\ell=0}^{j-1}\gamma_\ell\right)\left\|x_0 - \mathcal{P}_\mathcal{H} x_0\right\|_2^2\right\}\right] = 1. \tag{3.13}$$

Comparing Corollary 3.8 to classical results about the convergence of cyclic Kaczmarz-type methods (see [5, Thm. 1]), we see that our result is a probabilistic analogue: rather than guaranteeing a certain amount of convergence within a fixed number of iterations, we offer a certain amount of convergence within a random number of iterations whose expectation is controlled by a regularly increasing value (i.e., $\mathbb{E}[\tau_j] \leq j[(\text{rank}(A)-1)(N/\pi)+1]$). Moreover, Corollary 3.8 includes the important possibility of the procedure terminating in a finite amount of time. Finally, we have a guaranteed worst-case rate (i.e., with probability one) of convergence for all such methods. Of course, this rate is pessimistic, but, given the generality of the methods (e.g., adaptive, deterministic, random) that fall within the scope of our result, it is quite surprising that such a bound can be found under such few, very general assumptions.

Now, the alternative case to $\{\text{col}(A^\intercal W_k)\}$ belonging to a finite set is that it belongs to an infinite set, for which the canonical example is the row-action analogue to Example 3.14.[4] Unfortunately, our strategy for proving Corollary 3.8 will break down for the infinite set case: in the proof of Corollary 3.8, we set γ to be the maximum over a finite set of elements that are all strictly less than one; however, if we attempt to use the same strategy for the infinite set case, we can find systems and methods such that the supremum over the same set produces a $\gamma = 1$ (an explicit example is constructed in subsection 4.5). Thus, rather than looking at the supremum, we can attempt to control the distribution of $\{\gamma_\ell : \ell+1 \in \mathbb{N}\}$. Surprisingly, we will only need to control the mean behavior of these random quantities rather than the entire distribution.

To state this notion of control, we will need some notation. First, for each $\ell+1 \in \mathbb{N}$, let

$$\chi_\ell = \begin{cases} 1, & x_{\ell+1} \neq x_\ell, \\ 0 & \text{otherwise}, \end{cases} \tag{3.14}$$

be an indicator of whether we make progress in a given iteration. Moreover, for each $\ell+1 \in \mathbb{N}$, let $\mathfrak{Q}_\ell$ denote the collection of sets of vectors that are orthonormal and are a basis of $\text{col}(A^\intercal W_\ell \chi_\ell)$, and define $\mathcal{G}(Q_0, \ldots, Q_\ell)$ to be the set of matrices whose columns are maximal linearly independent subsets of $\cup_{s=0}^{\ell} Q_s$ where $Q_s \in \mathfrak{Q}_s$. With this notation, we have the following definition to control the distribution of $\{1-\gamma_\ell : \ell+1\}$.

[4] If $n_k > \text{rank}(A)$, then $\text{col}(A^\intercal W_k) = \text{row}(A)$ with probability one, which is covered by Corollary 3.8.

Definition 3.9 (uniformly nontrivial). *A row-action RBAS is uniformly nontrivial if for any* $\{\mathcal{A}_k : \mathbb{R}^d \times \mathfrak{Z} \to \mathcal{F}_{k+1}^k\}_{k+1\in\mathbb{N}}$ *such that* $\lim_{k\to\infty} \inf_{x_0 : x_0 \neq \mathcal{P}_{\mathcal{H}} x_0, \zeta_{-1}\in\mathfrak{Z}} \mathbb{P}[\mathcal{A}_k(x_0,\zeta_{-1}) | \mathcal{F}_1^0] = 1$, *there exists a* $g_{\mathcal{A}} \in (0,1]$ *such that*

$$\inf_{\substack{x_0 : x_0 \neq \mathcal{P}_{\mathcal{H}} x_0 \\ \zeta_{-1}\in\mathfrak{Z}}} \sup_{k\in\mathbb{N}\cup\{0\}} \mathbb{E}\left[\sup_{\substack{Q_s \in \mathfrak{Q}_s \\ s\in\{0,\ldots,k\}}} \min_{G\in\mathcal{G}(Q_0,\ldots,Q_k)} \det(G^\intercal G) \mathbf{1}\left[\mathcal{A}_k(x_0,\zeta_{-1})\right] \middle| \mathcal{F}_1^0 \right] \geq g_{\mathcal{A}}. \tag{3.15}$$

Before stating the result, we point out some important connections and features of Definition 3.9. First, so long as $G \in \mathcal{G}(Q_0,\ldots,Q_k)$ is nontrivial, $\det(G^\intercal G) > 0$ with probability one. Thus, for each x_0 such that $x_0 \neq \mathcal{P}_{\mathcal{H}} x_0$ and $\zeta_{-1} \in \mathfrak{Z}$, there exists a $k \in \mathbb{N} \cup \{0\}$ such that

$$\mathbb{E}\left[\sup_{\substack{Q_s \in \mathfrak{Q}_s \\ s\in\{0,\ldots,k\}}} \min_{G\in\mathcal{G}(Q_0,\ldots,Q_k)} \det(G^\intercal G) \mathbf{1}\left[\mathcal{A}_k(x_0,\zeta_{-1})\right] \middle| \mathcal{F}_1^0 \right] > 0. \tag{3.16}$$

Unfortunately, when we take the infimum over all allowed values of x_0 and ζ_{-1}, we can no longer guarantee that the lower bound is zero, as supplied by Definition 3.9.

Second, Definition 3.9 is closely related, yet complementary to the foundational notion of *uniformly integrable random variables.* To be specific, when a family of random variables is uniformly integrable, then the expected absolute value of the random variables in the family is uniformly bounded from above. Analogously and quite roughly, when we satisfy Definition 3.9, then the expected value of the random variables in the family are uniformly bounded from below.[5] Thus, we believe Definition 3.9 to be quite a foundational property and will need to be validated on a case-by-case basis (possibly with the help of tools such as analogues to the theorems of [27, 7]).

Finally, we are only controlling the expected behavior in Definition 3.9, and we do not need to make any statements about higher moments, which is surprising as $\{\gamma_\ell : \ell + 1\}$ is a dependent sequence, and usually dependencies require more complex moment statements (e.g., covariance relationships as in stationary processes). With these observations, we are ready for the next statement.

Corollary 3.10. *Let* $A \in \mathbb{R}^{n\times d}$ *and* $b \in \mathbb{R}^n$ *satisfy Assumption* 3.1. *Let* $x_0 \in \mathbb{R}^d$ *and* $\zeta_{-1} \in \mathfrak{Z}$. *Let* $\{x_k : k \in \mathbb{N}\}$ *be a sequence generated by* (3.3) *and* (3.4) *satisfying Definition* 3.5, *Definition* 3.7 *for some* $N \in \mathbb{N}$ *and* $\pi \in (0,1]$, *and Definition* 3.9. *One of the following is true.*

1. *There exists a stopping time* τ *with finite expectation such that* $x_\tau = \mathcal{P}_{\mathcal{H}} x_0$.
2. *There exists a sequence of nonnegative stopping times* $\{\tau_j : j + 1 \in \mathbb{N}\}$ *for which* $\mathbb{E}[\tau_j] \leq j[(\operatorname{rank}(A) - 1)(N/\pi) + 1]$, *there exists* $\bar{\gamma} \in (0,1)$, *and there exists a sequence of random variables* $\{\gamma_j : j + 1 \in \mathbb{N}\} \subset (0,1)$ *such that*

$$\mathbb{P}\left[\bigcap_{j=0}^{\infty} \left\{ \left\| x_{\tau_j} - \mathcal{P}_{\mathcal{H}} x_0 \right\|_2^2 \leq \left(\prod_{\ell=0}^{j-1} \gamma_\ell \right) \left\| x_0 - \mathcal{P}_{\mathcal{H}} x_0 \right\|_2^2 \right\} \right] = 1, \tag{3.17}$$

where for any $\gamma \in (\bar{\gamma}, 1)$, $\mathbb{P}[\cup_{L=0}^{\infty} \cap_{j=L}^{\infty} \{\prod_{\ell=0}^{j-1} \gamma_\ell \leq \gamma^j\}] = 1$.

[5] We say this roughly as we ignore the supremum over k to demonstrate the parallels between uniformly integrable families and a uniformly nontrivial RBAS.

Remark 3.11. $\mathbb{P}[\cup_{L=0}^{\infty} \cap_{j=L}^{\infty} \{\prod_{\ell=0}^{j-1} \gamma_\ell \leq \gamma^j\}] = 1$ is equivalent to, There exists a finite random variable, L, such that, for any $j \geq L$, $\prod_{\ell=0}^{j-1} \gamma_\ell \leq \gamma^j$ with probability one.

3.2. Column-action RBASs. In contrast to row-action RBASs, column-action RBASs *do not need* to assume that the system is consistent. Thus, we simply begin with an iterate $x_0 \in \mathbb{R}^d$ and some prior information, encapsulated by $\zeta_{-1} \in \mathfrak{Z}$, where $\mathfrak{Z}$ is finite in some sense. We then generate a sequence of iterates, $\{x_k : k \in \mathbb{N}\}$, according to

$$x_{k+1} = x_k - W_k (W_k^\intercal A^\intercal A W_k)^\dagger W_k^\intercal A^\intercal (A x_k - b), \tag{3.18}$$

where $\cdot^\dagger$ represents a pseudoinverse, and $\{W_k \in \mathbb{R}^{d \times n_k}\}$ are possibly random quantities (i.e., vectors or matrices) generated according to a possibly random, adaptive procedure, φ_C, which supplies

$$W_k, \zeta_k = \varphi_C(A, b, \{x_j : j \leq k\}, \{W_j : j < k\}, \{\zeta_j : j < k\}) \in \mathbb{R}^{d \times n_k} \times \mathfrak{Z}. \tag{3.19}$$

Note that our remarks about row-action RBASs apply here as well. We now present several examples.

Example 3.12 (cyclic vector coordinate descent). Let $\{e_i : i = 1, \ldots d\}$ denote the standard basis elements of $\mathbb{R}^d$. In this method, we update the iterate x_k to x_{k+1} by one coordinate at a time according to $x_{k+1} = x_k + e_i \alpha_k$ where α_k solves

$$\min_{\alpha \in \mathbb{R}} \left\| (b - A x_k) - A e_i \alpha \right\|_2, \tag{3.20}$$

which produces

$$x_{k+1} = x_k + e_i \frac{e_i^\intercal A^\intercal (b - A x_k)}{\|A e_i\|_2^2}. \tag{3.21}$$

The choice of e_i is determined by simply cycling through the basis elements in order. To rephrase this method within our formulation, we define $\mathfrak{Z} = \{0\} \cup \mathbb{N}$, $\zeta_{-1} = 0$, and

$$\varphi_C(A, b, \{x_j : j \leq k\}, \{W_j : j < k\}, \{\zeta_j : j < k\}) = (e_{\mathrm{rem}(\zeta_{k-1}, d) + 1}, \zeta_{k-1} + 1). \tag{3.22}$$

With this choice of (W_k, ζ_k), we see that the cyclic vector coordinate descent method is equivalent to (3.18). We underscore that φ_C only depends on ζ_{k-1} and the standard basis elements.

Example 3.13 (random permutation block coordinate descent). Let $\{E_i : i = 1, \ldots, \epsilon\}$ be matrices whose columns are generated by some partitioning of the $d \times d$ identity matrix. In this method, we have the update $x_{k+1} = x_k + E_i v_k$, where v_k solves

$$\min_{v} \left\| (b - A x_k) - A E_i v \right\|_2, \tag{3.23}$$

which produces the update

$$x_{k+1} = x_k + E_i (E_i^\intercal A^\intercal A E_i)^\dagger E_i^\intercal A^\intercal (b - A x_k). \tag{3.24}$$

To choose E_i, we begin by randomly permuting $\{E_i : i = 1, \ldots, \epsilon\}$, pass through this permutation until it is exhausted, select a new random permutation, pass through this permutation until it is exhausted, and repeat. By following the column-action analogue of Example 3.3, we can rephrase this method within our formulation.

Example 3.14 (block Gaussian column space descent). Let $\{W_k : k+1 \in \mathbb{N}\}$ be matrices with independent, identically distributed standard Gaussian components. In this method, we use the update $x_{k+1} = x_k + W_k v_k$, where v_k solves

$$\min_v \|(b - Ax_k) - AW_k v\|_2 , \tag{3.25}$$

which produces the update

$$x_{k+1} = x_k + W_k (W_k^\intercal A^\intercal A W_k)^\dagger W_k^\intercal A^\intercal (b - Ax_k). \tag{3.26}$$

It is clear that this update is exactly in the form of (3.18). Moreover, we can choose $\mathfrak{Z} = \{\emptyset\}$, $\zeta_{-1} = \emptyset$, and we can define

$$\varphi_C(A, b, \{x_j : j \le k\}, \{W_j : j < k\}, \{\zeta_j : j < k\}) = (W_k, \emptyset). \tag{3.27}$$

Thus, this method fits within our formulation.

As these examples demonstrate, column-action RBAS methods are also forgetful—that is, they satisfy the following analogue of Definition 3.5.

Definition 3.15 (Markovian). *A column-action RBAS is Markovian if there exists a finite $M \in \mathbb{N}$ such that for any measurable sets $\mathcal{W} \subset \mathbb{R}^{d \times n_k}$ and $\mathcal{Z} \subset \mathfrak{Z}$,*

$$\mathbb{P}\left[W_k \in \mathcal{W}, \zeta_k \in \mathcal{Z} \middle| \mathcal{F}_{k+1}^k\right] = \mathbb{P}\left[W_k \in \mathcal{W}, \zeta_k \in \mathcal{Z} \middle| \mathcal{F}_{\min\{M, k+1\}}^k\right]. \tag{3.28}$$

Remark 3.16. See Remark 3.6.

Similarly, just as with row-action methods, column-action RBASs are also N, π-exploratory. To state this definition, define $r^* = -\mathcal{P}_{\ker(A^\intercal)} b$.

Definition 3.17 (N, π-exploratory). *A column-action RBAS is N, π-exploratory for some $N \in \mathbb{N}$ and $\pi \in (0, 1]$ if*

$$\sup_{\substack{x_0 \in \mathbb{R}^d : Ax_0 - b \ne r^* \\ \zeta_{-1} \in \mathfrak{Z}}} \mathbb{P}\left[\bigcap_{j=0}^{N-1} \{\text{col}(AW_j) \perp Ax_0 - b\} \middle| \mathcal{F}_1^0 \right] \le 1 - \pi. \tag{3.29}$$

Note that the block Gaussian column space descent method, Example 3.14, is $1, 1$-exploratory.

Just as for row-action methods, we will have a bifurcation of the theory for the convergence of column-action methods based on whether the elements of $\{\text{col}(AW_k)\}$ take value in a finite set. In the case that they do, we have the following analogue of Corollary 3.8.

Corollary 3.18. *Let $A \in \mathbb{R}^{n \times d}$, $b \in \mathbb{R}^n$, and $r^* = -\mathcal{P}_{\ker(A^\intercal)} b$. Let $x_0 \in \mathbb{R}^d$ and $\zeta_{-1} \in \mathfrak{Z}$. Let $\{x_k : k \in \mathbb{N}\}$ be a sequence generated by* (3.18) *and* (3.19) *satisfying Definitions* 3.15 *and* 3.17 *for some $N \in \mathbb{N}$ and $\pi \in (0, 1]$. If the elements of $\{\text{col}(AW_k) : k + 1 \in \mathbb{N}\}$ take value in a finite set, then either*

1. *there exists a stopping time τ with finite expectation such that $Ax_\tau - b = r^*$, or*

2. *there exists a sequence of nonnegative stopping times* $\{\tau_j : j+1 \in \mathbb{N}\}$ *for which* $\mathbb{E}[\tau_j] \leq j[(\operatorname{rank}(A)-1)(N/\pi)+1]$, *and there exist* $\gamma \in (0,1)$ *and a sequence of random variables* $\{\gamma_j : j+1 \in \mathbb{N}\} \subset (0, \gamma]$ *such that*

$$\mathbb{P}\left[\bigcap_{j=0}^{\infty}\left\{\left\|Ax_{\tau_j} - b - r^*\right\|_2^2 \leq \left(\prod_{\ell=0}^{j-1}\gamma_\ell\right)\|Ax_0 - b - r^*\|_2^2\right\}\right] = 1. \tag{3.30}$$

The same comments for Corollary 3.8 apply to Corollary 3.18. Also, just as for Corollary 3.8, Corollary 3.18 does not cover Example 3.14 if $n_k < \operatorname{rank}(A)$. For the infinite set case, we will make use of the same notation as before with the following modifications. First,

$$\chi_\ell = \begin{cases} 1, & Ax_{\ell+1} - b \neq Ax_\ell - b, \\ 0, & Ax_{\ell+1} - b = Ax_\ell - b. \end{cases} \tag{3.31}$$

Second, let $\mathfrak{Q}_\ell$ denote the collection of sets of vectors that are orthonormal and are a basis of $\operatorname{col}(AW_\ell\chi_\ell)$. We now state the analogues of Definition 3.9 and Corollary 3.10.

Definition 3.19 (uniformly nontrivial). *A column-action RBAS is uniformly nontrivial if for any* $\{\mathcal{A}_k : \mathbb{R}^d \times \mathfrak{Z} \to \mathcal{F}_{k+1}^k\}_{k+1\in\mathbb{N}}$ *such that* $\lim_{k\to\infty}\inf_{x_0:Ax_0\neq b,\zeta_{-1}\in\mathfrak{Z}} \mathbb{P}[\mathcal{A}_k(x_0,\zeta_{-1})|\mathcal{F}_1^0] = 1$, *there exists a* $g_{\mathcal{A}} \in (0,1]$ *such that*

$$\inf_{\substack{x_0:Ax_0-b\neq r^*\\ \zeta_{-1}\in\mathfrak{Z}}} \sup_{k\in\mathbb{N}\cup\{0\}} \mathbb{E}\left[\sup_{\substack{Q_s\in\mathfrak{Q}_s\\ s\in\{0,\ldots,k\}}} \min_{G\in\mathcal{G}(Q_0,\ldots,Q_k)} \det(G^\intercal G)\mathbf{1}\left[\mathcal{A}_k(x_0,\zeta_{-1})\right]\middle| \mathcal{F}_1^0\right] \geq g_{\mathcal{A}}. \tag{3.32}$$

Corollary 3.20. *Let* $A \in \mathbb{R}^{n\times d}$, $b \in \mathbb{R}^n$, *and* $r^* = -\mathcal{P}_{\ker(A^\intercal)}b$. *Let* $x_0 \in \mathbb{R}^d$ *and* $\zeta_{-1} \in \mathfrak{Z}$. *Let* $\{x_k : k \in \mathbb{N}\}$ *be a sequence generated by* (3.18) *and* (3.19) *satisfying Definition* 3.15, *Definition* 3.17 *for some* $N \in \mathbb{N}$ *and* $\pi \in (0,1]$, *and Definition* 3.19. *One of the following is true.*

1. *There exists a stopping time* τ *with finite expectation such that* $Ax_\tau - b = r^*$.
2. *There exists a sequence of nonnegative stopping times* $\{\tau_j : j+1 \in \mathbb{N}\}$ *for which* $\mathbb{E}[\tau_j] \leq j[(\operatorname{rank}(A)-1)(N/\pi)+1]$, *there exists* $\bar{\gamma} \in (0,1)$, *and there exists a sequence of random variables* $\{\gamma_j : j+1 \in \mathbb{N}\} \subset (0,1)$ *such that*

$$\mathbb{P}\left[\bigcap_{j=0}^{\infty}\left\{\left\|Ax_{\tau_j} - b - r^*\right\|_2^2 \leq \left(\prod_{\ell=0}^{j-1}\gamma_\ell\right)\|Ax_0 - b - r^*\|_2^2\right\}\right] = 1, \tag{3.33}$$

where for any $\gamma \in (\bar{\gamma}, 1)$, $\mathbb{P}[\cup_{L=0}^{\infty}\cap_{j=L}^{\infty}\{\prod_{\ell=0}^{j-1}\gamma_\ell \leq \gamma^j\}] = 1$.

Remark 3.21. See Remark 3.11.

4. Convergence theory. We now prove Corollaries 3.8, 3.10, 3.18, and 3.20 by the following steps.

1. In subsection 4.1, we will write row-action and column-action methods using a common form, which reveals that the iterates (in the common form) are generated by products of orthogonal projections, which raises the questions, When will this sequence of products of orthogonal projections produce a reduction in the norms of the iterates and how big will this reduction be?

2. In subsection 4.2, we will answer this question by proving a generalized block Meany's inequality, which states that when the iterate is in a space generated by a sequence of projection matrices, we are guaranteed a certain amount of reduction in the norms of the iterates. Of course, this raises the question, When will the iterate be in this space?
3. In subsection 4.3, we define a stopping time for each iterate that, when finite, implies that the iterate will be in the aforementioned space. We show that when an RBAS is Markovian and N, π-exploratory, then, starting at any iterate, this stopping time is finite in expectation and we derive an explicit bound on this expectation.
4. Once we have established the finiteness of this stopping time, we can then apply our generalized block Meany's inequality to guarantee a reduction in the norm of the iterates. However, owing to the possible randomness of the procedure and the stopping times, we will need to find a deterministic control over the reduction constant provided by our generalized block Meany's inequality. In subsection 4.4, we will find this deterministic value by using the worst case over a finite set, which will prove Corollaries 3.8 and 3.18. In subsection 4.5, we will find this deterministic value by using the uniformly nontrivial property, which will prove Corollaries 3.10 and 3.20.

4.1. Common formulation. Our first step will be to rewrite row-action and column-action RBASs, and the corresponding definitions using a common formulation. To this end, we define

$$y_k = \begin{cases} x_k - \mathcal{P}_{\mathcal{H}} x_0 & \text{if (3.3) and Assumption 3.1,} \\ Ax_k - b - r^* & \text{if (3.18),} \end{cases} \tag{4.1}$$

where $r^* = -\mathcal{P}_{\ker(A^\intercal)} b$. Owing to this definition, the update y_k to y_{k+1} is

$$y_{k+1} = (I - \mathcal{P}_k) y_k, \tag{4.2}$$

where $\mathcal{P}_k$ are orthogonal projection matrices defined by

$$\mathcal{P}_k = \begin{cases} A^\intercal W_k (W_k^\intercal A A^\intercal W_k)^\dagger W_k^\intercal A & \text{if (3.3),} \\ A W_k (W_k^\intercal A^\intercal A W_k)^\dagger W_k^\intercal A^\intercal & \text{if (3.18).} \end{cases} \tag{4.3}$$

Thus, with these definitions, it is enough to prove convergence and rate of convergence results about $\{y_k\}$.

Remark 4.1. We can change the inner product space as done in [9], and we would still recover (4.2) with a simple change of variables. See [28].

To focus on $\{y_k\}$, we can update some of our definitions in terms of $\{y_k\}$ and $\{\mathcal{P}_k\}$.

Definition 4.2 (Markovian). *An RBAS (see* (4.2)*) is Markovian if there exists a finite $M \in \mathbb{N}$ such that for any measurable sets $\mathcal{W}$ and $\mathcal{Z} \subset \mathfrak{Z}$,*

$$\mathbb{P}\left[W_k \in \mathcal{W}, \zeta_k \in \mathcal{Z} \middle| \mathcal{F}_{k+1}^k\right] = \mathbb{P}\left[W_k \in \mathcal{W}, \zeta_k \in \mathcal{Z} \middle| \mathcal{F}_{\min\{M,k+1\}}^k\right]. \tag{4.4}$$

Definition 4.3 (N, π-exploratory). *An RBAS (see* (4.2)*) is N, π-exploratory for some $N \in \mathbb{N}$ and $\pi \in (0, 1]$ if*

$$\sup_{\substack{y_0 : y_0 \neq 0 \\ \zeta_{-1} \in \mathfrak{Z}}} \mathbb{P}\left[\bigcap_{j=0}^{N-1} \{\text{col}(\mathcal{P}_j) \perp y_0\} \middle| \mathcal{F}_1^0\right] \leq 1 - \pi. \tag{4.5}$$

Definition 4.4 (uniformly nontrivial). *An RBAS (see* (4.2)*) is uniformly nontrivial if for any* $\{\mathcal{A}_k : \mathbb{R}^d \times \mathfrak{Z} \to \mathcal{F}_{k+1}^k : k+1 \in \mathbb{N}\}$ *such that* $\lim_{k\to\infty} \inf_{y_0 : y_0 \neq 0, \zeta_{-1} \in \mathfrak{Z}} \mathbb{P}[\mathcal{A}_k(x_0,\zeta_{-1}) | \mathcal{F}_1^0] = 1$, *there exists a* $g_{\mathcal{A}} \in (0,1]$ *such that*

$$\inf_{\substack{y_0 : y_0 \neq 0 \\ \zeta_{-1} \in \mathfrak{Z}}} \sup_{k \in \mathbb{N} \cup \{0\}} \mathbb{E}\left[\sup_{\substack{Q_s \in \mathfrak{Q}_s \\ s \in \{0,\ldots,k\}}} \min_{G \in \mathcal{G}(Q_0,\ldots,Q_k)} \det(G^\intercal G) \mathbf{1}\left[\mathcal{A}_k(x_0,\zeta_{-1})\right] \middle| \mathcal{F}_1^0 \right] \geq g_{\mathcal{A}}. \tag{4.6}$$

4.2. Generalized block Meany's inequality. From (4.2), we see that $\{y_k\}$ are updated by applying a sequence of orthogonal projection matrices. We can now ask whether this application of projection matrices will drive $\{\|y_k\|_2\}$ to zero and at what rate. This question was first answered for products of projections of the form $I - qq^\intercal$ in [16], where the q's in the product are linearly independent—a result known as Meany's inequality. Meany's inequality has been generalized in two ways. First, Meany's inequality was extended to products of the form $I - QQ^\intercal$ in [2, Theorem 4.1], where each Q has orthonormal columns and the concatenation of all Q's in the product form a nonsingular matrix. Second, in [25, Theorem 4.1], Meany's inequality was generalized to the case in which there is a loss of independence between the q's. Here, we generalize all of these results.

To state our result, we will need to update some notation. For any k, let χ_k be 1 if $\mathcal{P}_k y_k \neq 0$ and zero otherwise, which we see is equivalent to (3.14) and (3.31) for the two different RBAS types. Let $\mathfrak{Q}_j$ be the set of orthogonal bases of $\mathrm{col}(\mathcal{P}_j)$ for all $j+1 \in \mathbb{N}$. Finally, let $\mathcal{C}_k^j = \mathrm{col}(\mathcal{P}_j\chi_j) + \cdots + \mathrm{col}(\mathcal{P}_{j+k}\chi_{j+k})$.

Theorem 4.5 (generalized block Meany's inequality). *Let* $j+1, k+1 \in \mathbb{N}$. *Then, for any* $y \in \mathcal{C}_k^j$ *with* $\|y\|_2 = 1$, $\|(I - \mathcal{P}_{j+k}\chi_{j+k}) \cdots (I - \mathcal{P}_j\chi_j)y\|_2^2$ *is no greater than* $1 - \sup_{Q_i \in \mathfrak{Q}_i, i \in \{j,\ldots,j+k\}} \min_{G \in \mathcal{G}(Q_j,\ldots,Q_{j+k})} \det(G^\intercal G)$.

Proof. Let $n_i = \dim(\mathrm{col}\,\mathcal{P}_i)$. Begin by fixing $Q_i \in \mathfrak{Q}_i$ for $i = j, \ldots, j+k$, and let $\{q_{i,\ell} : \ell = 1, \ldots, n_k\}$ denote the elements of Q_i. We can now follow the strategy of [2]. Letting the product notation indicate that terms with increasing index are being multiplied from the left, note that $\prod_{i=j}^{j+k}(I - \mathcal{P}_i\chi_i) = \prod_{i=j}^{k+j}[\prod_{\ell=1}^{n_i}(I - q_{i,\ell}q_{i,\ell}^\intercal\chi_i)]$. Therefore, [25, Thm. 4.1] provides

$$\|(I - \mathcal{P}_{j+k}\chi_{j+k}) \cdots (I - \mathcal{P}_j\chi_j)y\|_2^2 \leq \left(1 - \min_{G \in \mathcal{G}(Q_j,\ldots,Q_{j+k})} \det(G^\intercal G)\right) \|y\|_2^2. \tag{4.7}$$

This statement holds for every choice of $Q_i \in \mathfrak{Q}_i$. Therefore, the result follows. □

We pause for a moment to explain the importance of the supremum term in Theorem 4.5. We can first ask whether the choice of $Q_i \in \mathfrak{Q}_k$ will make any tangible difference. Consider the very simple situation of applying a block row selection method of a 4×3 matrix such that $A^\intercal W_1$ and $A^\intercal W_2$ generate

$$\begin{bmatrix} 2 & 1 & 0 \\ -1 & 2 & 3 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 1 & -3 & 6 \\ 0 & 1 & -5 \end{bmatrix}, \tag{4.8}$$

respectively. We can compute $\min_{G \in \mathcal{G}(Q_1,Q_2)} \det(G^\intercal G)$, which we will refer to as Meany's constant, from 10,000 uniformly sampled bases for the row spaces of these two matrices. The average value of Meany's constant is 0.12 with a standard deviation of 0.11, and the quantiles from this experiment are shown in Table 1. The supremum of Meany's constant is also included in Table 1. We see that the supremum is at least 8-fold larger than the average, and over 10^6-fold larger than the 0.001 quantile. Thus,

TABLE 1
Supremum and quantiles for randomly sampled Meany's constant for the example in (4.8).

Quantile	0.001	0.05	0.25	0.5	0.75	0.95	0.999	**Sup.**
Value	1.9×10^{-7}	6.5×10^{-4}	0.02	0.09	0.20	0.34	0.47	0.9995

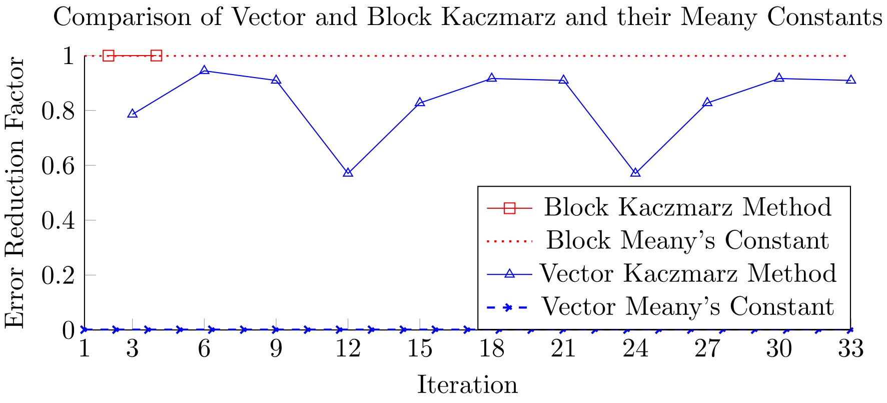


FIG. 2. *A plot of one less the ratio of norm errors squared between every three iterates for the vector method, and every two iterates for the block method. The horizontal lines correspond to the values of Meany's constant in Theorem* 4.5.

the supremum term is extremely important in finding better bounds on the rate of convergence.

Moreover, the supremum term also underscores the importance of block methods over vector methods (see [24]) from a theoretical perspective. For vector methods, there are only two choices in the set $\mathfrak{Q}_i$, and both produce the same value of Meany's constant. Thus, for vector methods, Meany's constant will only differ based on which vectors are seen. To demonstrate this, we run cyclic Kaczmarz and block cyclic Kacmzarz on the coefficient matrix in (4.8) until an absolute error of 10^{-4} is achieved. We plot one minus the ratio in norm error squared for each method and plot the corresponding Meany's constants in Figure 2. Clearly, we see that the block method is substantially superior to the corresponding vector method both in practice and in theory.

4.3. Stopping times. To apply Theorem 4.5, we need to determine for which k, $y_j \in \mathcal{C}_k^j$. Given that $\mathcal{C}_k^j$ is random, we will have to allow the time at which this occurs to be random, as follows.[6] For $j+1 \in \mathbb{N}$, let

$$\nu(j) = \min\left\{k \geq 0 : y_j \in \mathcal{C}_k^j\right\}. \tag{4.9}$$

Thus, when $\nu(j)$ is finite, Theorem 4.5 implies $\|y_{j+\nu(j)+1}\|_2^2/\|y_j\|_2^2$ is no greater than $1 - \sup_{Q_i \in \mathfrak{Q}_i, i \in \{j,\ldots,j+\nu(j)\}} \min_{G \in \mathcal{G}(Q_j,\ldots,Q_{j+\nu(j)})} \det(G^\intercal G)$. Hence, we need to determine whether $\nu(j)$ is finite for all j, and, ideally, we want to bound it, at the very least, in expectation. To this end, we will study another stopping time that is an upper bound on $\nu(j)$ and will find a bound on this new stopping time's expectation. We will begin by specifying this stopping time and showing that it is an upper bound on $\nu(j)$.

[6] It is understood that if the condition fails to occur, then the stopping time is infinite.

LEMMA 4.6. *For any $j+1 \in \mathbb{N}$, let $\nu(j)$ be defined as in* (4.9). *Then,* $\nu(j) \leq \min\{k \geq 0 : y_{j+k+1} \in \operatorname{span}[y_j, \ldots, y_{j+k}],\ \chi_{j+k} \neq 0\}$.

Proof. We begin with a key fact. By (4.2), $y_{k+1} - y_k \in \operatorname{col}(\mathcal{P}_k \chi_k)$ for any $k+1 \in \mathbb{N}$. It follows that $y_i \in \operatorname{span}[\mathcal{C}_k^j \cup \{y_\ell\}]$ for any $\ell, i \in [j, \ldots, j+k+1] \cap \mathbb{N}$.

Now, let $\nu(j)' = \min\{k \geq 0 : y_{j+k+1} \in \operatorname{span}[y_j, \ldots, y_{j+k}],\ \chi_{j+k} \neq 0\}$. Then, by the preceding fact, if $y_{j+\nu(j)'} \in \mathcal{C}_{\nu(j)'}^j$, then $y_j \in \mathcal{C}_{\nu(j)'}^j$. Thus, $\nu(j) \leq \nu(j)'$ by the minimality of $\nu(j)$. So it is enough to show $y_{j+\nu(j)'} \in \mathcal{C}_{\nu(j)'}^j$.

Let r denote the dimension of $\operatorname{span}[\mathcal{C}_{\nu(j)'}^j \cup \{y_{j+\nu(j)'}\}]$. Then, by the Gram–Schmidt procedure, there exist $\phi_1, \ldots, \phi_{r-1} \in \mathcal{C}_{\nu(j)'}^j$ such that the set of vectors $\{y_{j+\nu(j)'}, \phi_1, \ldots, \phi_{r-1}\}$ are an orthogonal basis for $\operatorname{span}[\mathcal{C}_{\nu(j)'}^j \cup \{y_{j+\nu(j)'}\}]$. Now, by the definition of $\nu(j)'$, there exist scalars $c_0, \ldots, c_{r-1}$ such that $y_{j+\nu(j)'+1} = c_0 y_{j+\nu(j)'} + c_1\phi_1 + \cdots c_{r-1}\phi_{r-1}$. Plugging this into (4.2),

$$c_0 y_{j+\nu(j)'} + c_1\phi_1 + \cdots c_{r-1}\phi_{r-1} = y_{j+\nu(j)'} - \mathcal{P}_{j+\nu(j)'} y_{j+\nu(j)'} \chi_{j+\nu(j)'}, \tag{4.10}$$

which gives rise to two cases. In the first case, we assume that $c_0 \neq 1$. Then, rearranging (4.10), we conclude $y_{j+\nu(j)'} \in \operatorname{span}[\phi_1, \ldots, \phi_{r-1}] + \operatorname{col}(\mathcal{P}_{j+\nu(j)'} \chi_{j+\nu(j)'}) = \mathcal{C}_{\nu(j)'}^j$. In the second case, $c_0 = 1$. Then, multiplying both sides of (4.10) by $y_{j+\nu(j)'}^\intercal$, $\sum_{i=1}^{r-1} c_i y_{j+\nu(j)'}^\intercal \phi_i = -\|\mathcal{P}_{j+\nu(j)'} y_{j+\nu(j)'}\|_2^2 \chi_{j+\nu(j)'}$. By the orthogonality of ϕ_i and $y_{j+\nu(j)'}$, the left-hand side is zero. The right-hand side can only be zero if $\chi_{j+\nu(j)'} = 0$, which contradicts the definition of $\nu(j)'$. To summarize these two cases, we showed that $y_{j+\nu(j)'} \in \mathcal{C}_{\nu(j)'}^j$. The result follows. □

THEOREM 4.7. *Let ξ be an arbitrary, finite stopping time with respect to $\{\mathcal{F}_{k+1}^k : k+1 \in \mathbb{N}\}$, and let $\mathcal{F}_{\xi+1}^\xi$ denote the stopped σ-algebra. Given that $\{y_k : k+1\}$ are well-defined (see* (4.1)*), let y_ξ be generated by an N, π-exploratory, Markovian RBAS. If $y_\xi \neq 0$, then $\nu(\xi)$ is finite, and $\mathbb{E}[\nu(\xi) | \mathcal{F}_{\xi+1}^\xi] \leq (\operatorname{rank}(A) - 1)(N/\pi)$.*

Proof. We need only bound the upper bound in Lemma 4.6. At any given $k \geq 0$, there are three possible cases, either (Case 1) $\chi_{\xi+k} = 0$; (Case 2) $\chi_{\xi+k} = 1$ and $y_{\xi+k+1} \notin \operatorname{span}[y_\xi, \ldots, y_{\xi+k}]$; or (Case 3) $\chi_{\xi+k} = 1$ and $y_{\xi+k+1} \in \operatorname{span}[y_\xi, \ldots, y_{\xi+k}]$. We will show that Cases 1 and 2 cannot hold for all $k \geq 0$ with probability one.

To this end, define $s(j) = \min\{k \geq 0 : \chi_{j+k} \neq 0\}$ and let $s_1 = s(\xi)$ and $s_{j+1} = s(\xi + s_1 + \cdots + s_j)$ for all $j \in \mathbb{N}$. With this notation, the Markovian property, and the N, π-exploratory property,

$$\begin{aligned}
&\mathbb{P}\left[s_1 \geq N \,|\, \mathcal{F}_{\xi+1}^\xi\right] \\
&= \mathbb{P}\left[\bigcap_{j=0}^{N-1} \{\chi_{\xi+j} = 0\} \middle| \mathcal{F}_{\xi+1}^\xi\right] = \mathbb{P}\left[\bigcap_{j=0}^{N-1} \{\operatorname{col}(\mathcal{P}_{\xi+j}) \perp y_\xi\} \middle| \mathcal{F}_{\xi+1}^\xi\right] &(4.11)\\
&= \mathbb{P}\left[\bigcap_{j=0}^{N-1} \{\operatorname{col}(\mathcal{P}_{\xi+j}) \perp y_\xi\} \middle| \mathcal{F}_1^\xi\right] \leq 1 - \pi, &(4.12)
\end{aligned}$$

where the last line is a consequence of Remark 3.6. Now, using induction and the Markovian property, $\mathbb{P}[s_1 \geq N\ell | \mathcal{F}_{\xi+1}^\xi] \leq (1-\pi)^\ell$ for all $\ell \in \mathbb{N}$. Therefore, s_1 is finite with probability one and $\mathbb{E}[s_1 | \mathcal{F}_{\xi+1}^\xi] \leq N/\pi$. Moreover, since ξ is an arbitrary stopping time, it follows that $\{s_j\}$ are finite with probability one and $\mathbb{E}[s_j | \mathcal{F}_{\xi+1}^\xi] \leq N/\pi$. Thus, Case 1 cannot occur for all $k \geq 0$, and Cases 2 or 3 must occur infinitely often.

Now, the dimension of span $\left[y_{\xi},\ldots,y_{\xi+s_1+\cdots+s_j}\right]$ is $j+1$. Since $\{y_k\}$ are either in row(A) or col(A), $j+1 \leq \mathrm{rank}\,(A)$. Thus, Case 2 cannot be the only situation to occur when $\chi_{\xi+k}\neq 0$. In conclusion, the largest value of j is $\mathrm{rank}\,(A)-1$, which implies $\nu(\xi)\leq s_1+\cdots+s_{\mathrm{rank}(A)-1}$, which are the sum of exponentially distributed random variables. The result follows by using $\mathbb{E}[s_j|\mathcal{F}_{\xi+1}^{\xi}]\leq N/\pi$. □

Now, putting together Theorems 4.5 and 4.7 supplies the following result.

COROLLARY 4.8. *Suppose $A\in\mathbb{R}^{n\times d}$ and $b\in\mathbb{R}^n$. Given that $\{y_k : k+1\}$ are well-defined (see* (4.1)*), suppose that they are generated by a Markovian, N,π-exploratory RBAS. Then, one of the following two cases occurs.*

1. *There exists a stopping time τ with finite expectation such that $y_\tau = 0$.*
2. *There exist stopping times $\{\tau_j : j+1\in\mathbb{N}\}$ such that $\mathbb{E}\,[\tau_j]\leq j[(\mathrm{rank}\,(A)-1)(N/\pi)+1]$ for all $j+1\in\mathbb{N}$, and $\mathbb{P}[\cap_{j=1}^{\infty}\{\|y_{\tau_j}\|_2^2\leq(\prod_{\ell=0}^{j-1}\gamma_\ell)\|y_0\|_2^2\}]=1$, where $\gamma_\ell = 1-\sup_{Q_i\in\mathfrak{Q}_i, i\in\{\tau_\ell,\ldots,\tau_\ell+\nu(\tau_\ell)\}}\min_{G\in\mathcal{G}(Q_{\tau_\ell},\ldots,Q_{\tau_\ell+\nu(\tau_\ell)})}\det(G^\intercal G)\in(0,1)$.*

Proof. The proof proceeds by induction. For $j=0$, recall $\tau_0=0$. Now, either $y_{\tau_0}=0$ or $y_{\tau_0}\neq 0$. In the former case, the statement of the result is true. In the latter case, define $\tau_1=\nu(\tau_0)+1$. Then, τ_1 is finite with probability one and $\mathbb{E}\,[\tau_1]\leq(\mathrm{rank}\,(A)-1)N/\pi+1$ by Theorem 4.7. Moreover, by Theorem 4.5, $\|y_{\tau_1}\|_2^2\leq\gamma_0\|y_{\tau_0}\|_2^2$. Thus, we have established the base case.

For the induction hypothesis, suppose that for $j\in\mathbb{N}$, $Ax_{\tau_{j-1}}\neq b$, $\mathbb{E}\,[\tau_k]\leq k[(\mathrm{rank}\,(A)-1)N/\pi+1]$ for $k\in[0,j-1]\cap\mathbb{N}$, and $\|y_{\tau_k}\|_2^2\leq\|y_0\|_2^2\prod_{\ell=0}^{k-1}\gamma_\ell$ for $k\in[1,j-1]\cap\mathbb{N}$.

To conclude, define $\tau_j=\tau_{j-1}+\nu(\tau_{j-1})+1$. By Theorem 4.7, τ_j is finite and $\mathbb{E}\,[\tau_j]\leq(j-1)[(\mathrm{rank}\,(A)-1)N/\pi+1]+(\mathrm{rank}\,(A)-1)N/\pi+1=j[(\mathrm{rank}\,(A)-1)N/\pi+1]$. Finally, either $y_{\tau_j}=0$ or $y_{\tau_j}\neq 0$. In the latter case, Theorem 4.5 implies $\|y_{\tau_j}\|_2^2\leq\gamma_{j-1}\|y_{\tau_{j-1}}\|_2^2$. The result follows. □

Our final task is to control the joint behavior of $\{\gamma_\ell : \ell+1\in\mathbb{N}\}\subset(0,1)$ in the latter case of Corollary 4.8. Depending on our goal, we could require two different types of control. For instance, to ensure convergence of $\{y_k\}$ to 0, we need to ensure that $\liminf_{\ell\to\infty}\gamma_\ell<1$. However, for a rate of convergence, we need to ensure that $\limsup_{\ell\to\infty}\gamma_\ell<1$. As the latter case is more desirable in practice, we will focus on ensuring that $\limsup_{\ell\to\infty}\gamma_\ell<1$. This will give rise to two separate cases in our theory of convergence of RBAS methods, which we now address one at a time.

4.4. Convergence for a finite set.

In the first case, we have that $\{\mathrm{col}(\mathcal{P}_k)\}$ take value in finite sets, as in Examples 3.2 to 3.4, 3.12 and 3.13.

THEOREM 4.9. *Let $A\in\mathbb{R}^{n\times d}$ and $b\in\mathbb{R}^n$. Given that $\{y_k : k+1\}$ are well-defined (see* (4.1)*), suppose $\{y_k : k+1\in\mathbb{N}\}$ are generated by a Markovian, N,π-exploratory RBAS. If the elements of $\{\mathrm{col}(\mathcal{P}_k) : k+1\in\mathbb{N}\}$ take value in a finite set, then either*

1. *there exists a stopping time τ with finite expectation such that $y_\tau=0$; or*
2. *there exist stopping times $\{\tau_j : j+1\in\mathbb{N}\}$ such that $\mathbb{E}\,[\tau_j]\leq j[(\mathrm{rank}\,(A)-1)(N/\pi)+1]$ for all $j+1\in\mathbb{N}$, and there exist $\gamma\in(0,1)$ and a sequence of random variables $\{\gamma_j : j+1\in\mathbb{N}\}\subset(0,\gamma]$ such that $\mathbb{P}[\cap_{j=1}^{\infty}\{\|y_{\tau_j}\|_2^2\leq(\prod_{\ell=0}^{j-1}\gamma_\ell)\|y_0\|_2^2\}]=1$.*

Proof. By Corollary 4.8, we can focus on the second case and we need only show that there exists a $\gamma\in(0,1)$ such that $\gamma_\ell\leq\gamma$. To this end, let $\{\mathcal{U}_i : i=1,\ldots,r\}$ denote the set of linear spaces in which $\{\mathrm{col}(\mathcal{P}_k)\}$ takes value. For each $\mathcal{U}_i$, we can define the set of all orthonormal bases of $\mathcal{U}_i$, denoted $\mathfrak{U}_i$. Let $\mathfrak{P}$ denote the power set of $\{\mathfrak{U}_i : i = 1,\ldots,r\}$. For a given element $\{\mathfrak{U}_{i_1},\ldots,\mathfrak{U}_{i_s}\} \in \mathfrak{P}$, we can choose a set $\{\cup_{j=1}^{s} U_j : U_j \in \mathfrak{U}_{i_j}\}$, and let $\mathcal{H}$ denote the set of all matrices whose columns are maximal linearly independent subsets of $\{\cup_{j=1}^{s} U_j : U_j \in \mathfrak{U}_{i_j}\}$. Finally, define

$$\Gamma = \left\{ 1 - \sup_{\{\cup_{j=1}^{s} U_j : U_j \in \mathfrak{U}_{i_j}\}} \min_{H \in \mathcal{H}} \det(H^\intercal H) : \{\mathfrak{U}_{i_1},\ldots,\mathfrak{U}_{i_s}\} \in \mathfrak{P},\ s = 1,\ldots,r \right\}. \tag{4.13}$$

Now, since $\{\mathfrak{Q}_k\}$ takes value in $\{\mathfrak{U}_i : i = 1,\ldots,r\}$, $\{\mathfrak{Q}_i : i = \tau_\ell,\ldots,\tau_\ell + \nu(\tau_\ell)\} \in \mathfrak{P}$ for all $\ell + 1 \in \mathbb{N}$. Therefore, $\gamma_\ell \in \Gamma$ for all $\ell + 1 \in \mathbb{N}$. Thus, $\gamma_\ell \leq \max\{\Gamma\} =: \gamma$. By Hadamard's inequality, each element of Γ is in $[0,1)$, which implies that $\gamma \in [0,1)$. As we are only proving the second case, $\gamma \neq 0$ (otherwise we would have converged finitely and would be in the first case), which implies $\gamma \in (0,1)$. □

We make two remarks. First, by substituting the appropriate definitions of $\{y_k : k+1 \in \mathbb{N}\}$ and $\{\mathcal{P}_k : k+1 \in \mathbb{N}\}$ into Theorem 4.9, then we have proven Corollaries 3.8 and 3.18. Second, the value of γ can vary depending on how the set to which $\{\operatorname{col}(\mathcal{P}_k)\}$ belongs is designed, which was a central point of discussion in [21]. For instance, consider the three unique partitions of the rows of the coefficient matrix

$$\begin{bmatrix} 1 & -1 & 1 \\ 1 & -1 & 1 + 10^{-5} \\ 3 & -1 & 3 \\ 0 & 1 & 6 \end{bmatrix} \tag{4.14}$$

such that each partition contains two rows. Now, consider a sampling scheme that selects a partition and cycles through the blocks in this partition. For such a method, we can compute γ. The results for each of the three partitions are presented in Table 2.

From Table 2, we see that to get the same guaranteed relative reduction in error from Partition I in comparison to Partition II or III requires over sevenfold more iterations. Indeed, as shown in Figure 3, we observe exactly this behavior when we implement cyclic block Kaczmarz on (4.14) for the three different partitions up to an absolute error of 10^{-4}.

4.5. Convergence for an infinite set. In the second case, $\{\operatorname{col}(\mathcal{P}_k)\}$ can take value over an infinite set, as in Example 3.14 with $n_k < \operatorname{rank}(A)$. Suppose we attempt to prove the convergence result as we did in subsection 4.4. Then, we would need to prove that $\sup\{\Gamma\} < 1$. However, when Γ is infinite, we could potentially have $\sup\{\Gamma\} = 1$. For instance, consider a 3×2 coefficient matrix whose first two rows are the first standard basis element of $\mathbb{R}^2$ and the last row is the second standard basis element, and a procedure that alternates between either choosing the first row of the matrix or taking a linear combination of the second row and a product of a standard Guassian random variable with the third row. If we let $\mathcal{N}(0,1)$ denote a standard Gaussian distribution, then Γ is made up of all possible values of $(Z^2+1)^{-1}$ with

TABLE 2
Estimates of the values of γ in Theorem 4.9 for the three unique equally sized partitions of (4.14).

Partition I	Partition II	Partition III
0.880	0.372	0.372

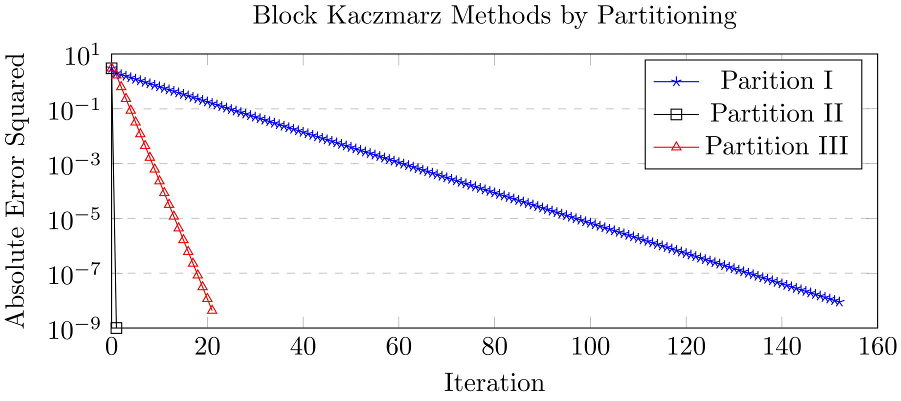


FIG. 3. *A comparison of the absolute errors of cyclic block Kaczmarz methods for the coefficient matrix in* (4.14)*. Corresponding to the theory, Partition* I *produces the worst convergence rate.*

$Z \sim \mathcal{N}(0,1)$. Since Z has nonzero density about 0, the supremum of Γ would be 1 in this case.

As this example suggests, it is possible to have arbitrarily poor values for $\{\gamma_\ell : \ell+1 \in \mathbb{N}\}$. However, this example also shows that the bulk of values of $\{\gamma_\ell : \ell + 1 \in \mathbb{N}\}$ are well-behaved (i.e., the mean and standard deviation of $(Z^2+1)^{-1}$ are approximately 0.66 and 0.26, respectively), which partially motivates Definition 4.4. Under Definition 4.4, we have the following result, from which Corollaries 3.10 and 3.20 follow immediately.

THEOREM 4.10. *Let $A \in \mathbb{R}^{n\times d}$ and $b \in \mathbb{R}^n$. Given that $\{y_k : k+1\}$ are well-defined (see* (4.1)*), suppose $\{y_k : k+1 \in \mathbb{N}\}$ are generated by a Markovian, N,π-exploratory, and uniformly nontrivial RBAS. One of the following is true.*

1. *There exists a stopping time τ with finite expectation such that $y_\tau = 0$.*
2. *There exists a sequence of nonnegative stopping times $\{\tau_j : j+1 \in \mathbb{N}\}$ for which $\mathbb{E}[\tau_j] \le j[(\operatorname{rank}(A)-1)(N/\pi)+1]$, there exists $\bar\gamma \in (0,1)$, and there exists a sequence of random variables $\{\gamma_j : j+1 \in \mathbb{N}\} \subset (0,1)$ such that $\mathbb{P}[\cap_{j=0}^\infty \{\|y_{\tau_j}\|_2^2 \le (\prod_{\ell=0}^{j-1}\gamma_\ell)\|y_0\|_2^2\}] = 1$, where for any $\gamma \in (\bar\gamma,1)$, $\mathbb{P}[\cup_{L=0}^\infty \cap_{j=L}^\infty \{\prod_{\ell=0}^{j-1}\gamma_\ell \le \gamma^j\}] = 1$.*

Proof. By Corollary 4.8, we can focus on the second case and we need only prove that there exists $\bar\gamma \in (0,1)$ such that for any $\gamma \in (\bar\gamma,1)$, $\mathbb{P}[\cup_{L=0}^\infty \cap_{j=L}^\infty \{\prod_{\ell=0}^{j-1}\gamma_\ell \le \gamma^j\}] = 1$. To show this, we need to prove $\mathbb{E}[\prod_{\ell=0}^{j-1}\gamma_\ell | \mathcal{F}_1^0] \le \bar\gamma^j$ for each j, which we will do by induction. For the base case, $j=0$,

$$\mathbb{E}\left[1-\gamma_0 | \mathcal{F}_1^0\right]$$

$$= \mathbb{E}\left[\sup_{Q_i \in \mathfrak{Q}_i, i \in \{\tau_0,\ldots,\tau_0+\nu(\tau_0)\}} \min_{G \in \mathcal{G}(Q_{\tau_0},\ldots,Q_{\tau_0+\nu(\tau_0)})} \det(G^\intercal G) \middle| \mathcal{F}_1^0\right] \tag{4.15}$$

$$= \mathbb{E}\left[\sum_{k=0}^\infty \mathbf{1}\left[\nu(\tau_0)=k\right] \sup_{Q_i \in \mathfrak{Q}_i, i\in\{\tau_0,\ldots,\tau_0+k\}} \min_{G \in \mathcal{G}(Q_{\tau_0},\ldots,Q_{\tau_0+k})} \det(G^\intercal G) \middle| \mathcal{F}_1^0\right]. \tag{4.16}$$

Since $\min_{G\in\mathcal{G}(Q_{\tau_0},\ldots,Q_{\tau_0+k})} \det(G^\intercal G) \ge \min_{G\in\mathcal{G}(Q_{\tau_0},\ldots,Q_{\tau_0+k},Q)} \det(G^\intercal G)$ for any $Q \in \mathfrak{Q}_{\tau_0+k+1}$ and any $k+1 \in \mathbb{N}$, then, for every $k+1\in\mathbb{N}$, $\mathbb{E}[1-\gamma_0|\mathcal{F}_1^0]$ is bounded below by

$$\mathbb{E}\left[\mathbf{1}\left[\nu(\tau_0) \leq k\right] \sup_{Q_i \in \mathfrak{Q}_i, i \in \{\tau_0,\ldots,\tau_0+k\}} \min_{G \in \mathcal{G}(Q_{\tau_0},\ldots,Q_{\tau_0+k})} \det(G^\intercal G) \middle| \mathcal{F}_1^0 \right]. \tag{4.17}$$

Now, by Theorem 4.7 and Markov's inequality, for any $y_0 \neq 0$ and any $\zeta_{-1} \in \mathfrak{Z}$, $\mathbb{P}[\nu(\tau_0) \leq k | \mathcal{F}_1^0] \geq 1 - N(\text{rank}\,(A) - 1)/(k\pi)$. Hence, we can apply Definition 4.4 to conclude that there exists a $g \in (0,1]$ such that $\mathbb{E}[1 - \gamma_0 | \mathcal{F}_1^0] \geq g$. If we let $\bar{\gamma} = 1 - g$, then $\mathbb{E}[\gamma_0 | \mathcal{F}_1^0] \leq \bar{\gamma}$. Now, for the induction hypothesis, suppose that $\mathbb{E}[\prod_{\ell=0}^{j-2} \gamma_\ell | \mathcal{F}_1^0] \leq \bar{\gamma}^{j-1}$. To conclude, we note that by the Markovian property and the base case, $\mathbb{E}[\gamma_{j-1} | \mathcal{F}_{\tau_j+1}^{\tau_j}] \leq \bar{\gamma}$. Therefore, $\mathbb{E}[\prod_{\ell=0}^{j-1} \gamma_\ell | \mathcal{F}_1^0] = \mathbb{E}[\mathbb{E}[\gamma_{j-1} | \mathcal{F}_{\tau_j+1}^{\tau_j}] \prod_{\ell=0}^{j-1} \gamma_\ell | \mathcal{F}_1^0] \leq \bar{\gamma}^j$.

Now, for any $\gamma \in (\bar{\gamma}, 1)$, the preceding proof and Markov's inequality provide

$$\sum_{j=1}^{\infty} \mathbb{P}\left[\prod_{\ell=0}^{j-1} \gamma_\ell > \gamma^j \middle| \mathcal{F}_1^0 \right] \leq \sum_{j=1}^{\infty} \left(\frac{\bar{\gamma}}{\gamma} \right)^j < \infty. \tag{4.18}$$

By the Borel–Cantelli lemma, $\mathbb{P}[\cup_{L=0}^{\infty} \cap_{j=L}^{\infty} \{\prod_{\ell=0}^{j-1} \gamma_\ell \leq \gamma^j\}] = 1$. □

Remark 4.11. Our proof readily allows us to bound the convergence rates of the moments of $\{y_k : k+1\}$.

5. Examples.

We provide a series of examples to demonstrate how we can apply our theory to a variety of methods. Of particular practical value, we will show how to verify the relevant properties (e.g., exploratory). We summarize these examples and references, and we reference the convergence result for the given method based on our theory in Table 3.

6. Conclusion.

In order to enable broader use of highly tailored randomized methods for solving linear systems, we began with the challenge of providing a unifying theory for randomized block adaptive solvers (RBASs) for linear systems—regardless of whether the linear systems are underdetermined, overdetermined, or rank deficient. To this end, we studied two archetypes of RBAS solvers—row-action methods for consistent linear systems and column-action methods for arbitrary linear systems—and showed that under very general conditions both archetypes will converge exponentially fast to a solution. Specifically, we had two results.

1. When an RBAS is Markovian, N, π-exploratory, and projects either the absolute error (for row-action methods) or residual (for column-action methods) onto only a finite number of spaces, then the RBAS will converge exponentially fast to a solution of the linear system.
2. When an RBAS is Markovian, N, π-exploratory, and uniformly nontrivial, then, after some finite number of iterations, the RBAS will converge exponentially fast to a solution of the linear system.

We further provided numerical evidence to elucidate key aspects of theory at key points. In particular, we demonstrated the value of the supremum in our generalization of Meany's inequality (see Theorem 4.5 and Figure 2), and we discussed the importance of finding appropriate partitions when using block cyclic solvers (see Figure 3), which was quite carefully studied in [21]. Finally, we provided a host of examples of how to apply our theory to existing methods and some novel methods, which we complemented with appropriate numerical experiments.

In completing the above tasks, we have provided practitioners with a powerful theory and demonstrations of how to use the theory to rigorously analyze a wide variety of RBASs. Thus, we hope that practitioners will be empowered to use this

TABLE 3
Summary of worked examples using our theory.

Method	**References**	**Details**	**Convergence result**
Cyclic vector Kaczmarz	[14, 13, 5]	Subsection SM1.1	Theorem SM1.3
Gaussian vector Kaczmarz	[11, 30]	Subsection SM1.2	Theorem SM1.7
Strohmer–Vershynin vector Kaczmarz	[34]	Subsection SM1.3	Theorem SM1.11
Steinerberger vector Kaczmarz	[32]	Subsection SM1.4	Theorem SM1.14
Motzkin's method	[18, 1]	Subsection SM1.5	Theorem SM1.17
Agmon's method	[1]	Subsection SM1.6	Theorem SM1.20
Greedy randomized vector Kaczmarz	[3]	Subsection SM1.7	Theorem SM1.23
Sampling Kaczmarz–Motzkin Method	[12]	Subsection SM1.8	Theorem SM1.27
Streaming vector Kaczmarz	[24]	Subsection SM1.9	Theorem SM1.31
Cyclic vector coordinate descent	[35]	Subsection SM1.10	Theorem SM1.34
Gaussian vector column space Descent		Subsection SM1.11	Theorem SM1.38
Zouzias–Freris vector coordinate Descent	[38]	Subsection SM1.12	Theorem SM1.41
Max residual coordinate descent		Subsection SM1.13	Theorem SM1.44
Max distance coordinate descent		Subsection SM1.14	Theorem SM1.47
Random permutation block Kaczmarz	[21, 19]	Subsection SM1.15	Theorem SM1.50
Steinerberger block Kaczmarz	[30, 10]	Subsection SM1.16	Theorem SM1.53
Motzkin's block Method		Subsection SM1.17	Theorem SM1.56
Agmon's block Method		Subsection SM1.18	Theorem SM1.59
Adaptive sketch-and-project	[10]	Subsection SM1.19	Theorem SM1.64
Greedy randomized block Kaczmarz		Subsection SM1.20	Theorem SM1.67
Streaming block Kaczmarz	[11, 29]	Subsection SM1.21	Theorem SM1.71
Random permutation block Coordinate descent	[20, 37]	Subsection SM1.22	Theorem SM1.74
Gaussian block column space descent		Subsection SM1.23	Theorem SM1.78
Zouzias–Freris block coordinate descent		Subsection SM1.24	Theorem SM1.81
Max residual block coordinate descent		Subsection SM1.25	Theorem SM1.84
Max distance block coordinate descent		Subsection SM1.26	Theorem SM1.87

theory and create novel RBASs that are optimized to their specific applications and computing environments.

REFERENCES

[1] S. Agmon, *The relaxation method for linear inequalities*, Canad. J. Math., 6 (1954), pp. 382–392.

[2] Z.-Z. Bai and X.-G. Liu, *On the Meany inequality with applications to convergence analysis of several row-action iteration methods*, Numer. Math., 124 (2013), pp. 215–236.

[3] 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.

[4] A. H. Baker, J. M. Dennis, and E. R. Jessup, *On improving linear solver performance: A block variant of GMRES*, SIAM J. Sci. Comput., 27 (2006), pp. 1608–1626.

[5] L. Dai and T. B. Schön, *On the exponential convergence of the Kaczmarz algorithm*, IEEE Signal Process. Lett., 22 (2015), pp. 1571–1574.

[6] P. J. Denning, *The locality principle*, Commun. ACM, 48 (2005), pp. 19–24.

[7] N. Dunford, *A mean ergodic theorem*, Duke Math. J., 5 (1939), pp. 635–646.

[8] G. H. Golub and C. F. Van Loan, *Matrix Computations*, Vol. 3, JHU Press, Baltimore, 2012.

[9] R. Gower, D. Molitor, J. Moorman, and D. Needell, *Adaptive Sketch-and-Project Methods for Solving Linear Systems*, preprint, https://arxiv.org/abs/1909.03604, 2019.

[10] R. M. Gower, D. Molitor, J. Moorman, and D. Needell, *On adaptive sketch-and-project for solving linear systems*, SIAM J. Matrix Anal. Appl., 42 (2021), pp. 954–989.

[11] R. M. Gower and P. Richtárik, *Randomized iterative methods for linear systems*, SIAM J. Matrix Anal. Appl., 36 (2015), pp. 1660–1690.

[12] J. Haddock and A. Ma, *Greed Works: An Improved Analysis of Sampling Kaczmarz-Motkzin*, preprint, https://arxiv.org/abs/1912.03544, 2019.

[13] S. Kaczmarz, *Approximate solution of systems of linear equations*, Internat. J. Control, 57 (1993), pp. 1269–1271.

[14] S. Kaczmarz, *Angenaherte auflosung von systemen linearer glei-chungen*, Bull. Int. Acad. Pol. Sci. Let., 35 (1937), pp. 355–357.

[15] P.-G. Martinsson and J. A. Tropp, *Randomized numerical linear algebra: Foundations and algorithms*, Acta Numer., 29 (2020), pp. 403–572.

[16] R. K. Meany, *A matrix inequality*, SIAM J. Numer. Anal., 6 (1969), pp. 104–107.

[17] S. P. Meyn and R. L. Tweedie, *Markov Chains and Stochastic Stability*, Springer, New York, 2012.

[18] T. S. Motzkin and I. J. Schoenberg, *The relaxation method for linear inequalities*, Canad. J. Math., 6 (1954), pp. 393–404.

[19] I. Necoara, *Faster randomized block Kaczmarz algorithms*, SIAM J. Matrix Anal. Appl., 40 (2019), pp. 1425–1452.

[20] I. Necoara and D. Clipici, *Parallel random coordinate descent method for composite minimization: Convergence analysis and error bounds*, SIAM J. Optim., 26 (2016), pp. 197–226.

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

[22] 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.

[23] J. Nutini, B. Sepehry, I. Laradji, M. Schmidt, H. Koepke, and A. Virani, *Convergence Rates for Greedy Kaczmarz Algorithms, and Faster Randomized Kaczmarz Rules Using the Orthogonality Graph*, preprint, https://arxiv.org/abs/1612.07838, 2016.

[24] V. Patel, M. Jahangoshahi, and D. A. Maldonado, *Convergence of Adaptive, Randomized, Iterative Linear Solvers*, preprint, https://arxiv.org/abs/2104.04816, 2021.

[25] V. Patel, M. Jahangoshahi, and D. A. Maldonado, *An implicit representation and iterative solution of randomly sketched linear systems*, SIAM J. Matrix Anal. Appl., 42 (2021), pp. 800–831.

[26] M. Pilanci and M. J. Wainwright, *Iterative Hessian sketch: Fast and accurate solution approximation for constrained least-squares*, J. Mach. Learn. Res., 17 (2016), pp. 1–38.

[27] C. D. L. V. Poussin, *Sur l'intégrale de lebesgue*, Trans. Amer. Math. Soc., 29 (1915), pp. 435–501.

[28] N. Pritchard and V. Patel, *Residual Tracking and Stopping for Iterative Random Sketching*, preprint, https://arxiv.org/abs/2201.05741, 2022.

[29] E. Rebrova and D. Needell, *On block Gaussian sketching for the Kaczmarz method*, Numer. Algorithms, 86 (2021), pp. 443–473.

[30] P. Richtárik and M. Takác, *Stochastic reformulations of linear systems: Algorithms and convergence theory*, SIAM J. Matrix Anal. Appl., 41 (2020), pp. 487–524.

[31] Y. Saad, *Iterative Methods for Sparse Linear Systems*, SIAM, Philadelphia, 2003.

[32] S. Steinerberger, *A weighted randomized Kaczmarz method for solving linear systems*, Math. Comp., 90 (2021), pp. 2815–2826.

[33] S. Steinerberger, *Approximate Solutions of Linear Systems at a Universal Rate*, preprint, https://arxiv.org/abs/2207.03388, 2022.

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

[35] J. Warga, *Minimizing certain convex functions*, J. SIAM, 11 (1963), pp. 588–593.

[36] D. P. Woodruff, *Sketching as a tool for numerical linear algebra*, Found. Trends Theor. Comput. Sci., 10 (2014), pp. 1–157.

[37] S. Wright and C.-P. Lee, *Analyzing random permutations for cyclic coordinate descent*, Math. Comp., 89 (2020), pp. 2217–2248.

[38] A. Zouzias and N. M. Freris, *Randomized extended Kaczmarz for solving least squares*, SIAM J. Matrix Anal. Appl., 34 (2013), pp. 773–793.