Convergence Rates for Greedy Kaczmarz Algorithms, and Faster Randomized Kaczmarz Rules Using the Orthogonality Graph

Julie Nutini[*]
jnutini@cs.ubc.ca

Behrooz Sepehry[*]
behrooz@cs.ubc.ca

Issam Laradji[*]
issamou@cs.ubc.ca

Mark Schmidt[*]
schmidtm@cs.ubc.ca

Hoyt Koepke[†]
hoytak@gmail.com

Alim Virani[*]
alim.virani@gmail.com

Abstract

The Kaczmarz method is an iterative algorithm for solving systems of linear equalities and inequalities, that iteratively projects onto these constraints. Recently, Strohmer and Vershynin [*J. Fourier Anal. Appl.,* 15(2):262-278, 2009] gave a non-asymptotic convergence rate analysis for this algorithm, spurring numerous extensions and generalizations of the Kaczmarz method. Rather than the randomized selection rule analyzed in that work, in this paper we instead discuss greedy and approximate greedy selection rules. We show that in some applications the computational costs of greedy and random selection are comparable, and that in many cases greedy selection rules give faster convergence rates than random selection rules. Further, we give the first multi-step analysis of Kaczmarz methods for a particular greedy rule, and propose a provably-faster randomized selection rule for matrices with many pairwise-orthogonal rows.

1 KACZMARZ METHOD

Solving large linear systems is a fundamental problem in machine learning. Applications range from least-squares problems to Gaussian processes to graph-based semi-supervised learning. All of these applications (and many others) benefit from advances in solving large-scale linear systems. The Kaczmarz method is a particular iterative algorithm suited for solving consistent linear systems of the form $Ax = b$. It was originally proposed by Polish mathematician Stefan Kaczmarz (1937) and later re-invented by Gordon et al. (1970) under the name *algebraic reconstruction technique* (ART). It has been used in numerous applications including image reconstruction and digital signal processing, and belongs to several general categories of methods including *row-action*, *component-solution*, *cyclic projection*, and *successive projection* methods (Censor, 1981).

At each iteration k, the Kaczmarz method uses a *selection rule* to choose some row i_k of A and then projects the current iterate x^k onto the corresponding hyperplane $a_{i_k}^T x^k = b_{i_k}$. Classically, the two categories of selection rules are *cyclic* and *random.* Cyclic selection repeatedly cycles through the coordinates in sequential order, making it simple to implement and computationally inexpensive. There are various linear convergence rates for cyclic selection (see Deutsch, 1985; Deutsch and Hundal, 1997; Galántai, 2005), but these rates are in terms of cycles through the entire dataset and involve constants that are not easily interpreted. Further, the performance of cyclic selection worsens if we have an undesirable ordering of the rows of A.

Randomized selection has recently become the default selection rule in the literature on Kaczmarz-type methods. Empirically, selecting i randomly often performs substantially better in practice than cyclic selection (Feichtinger et al., 1992; Herman and Meyer, 1993). Although a number of asymptotic convergence rates for randomized selection have been presented (Whitney and Meany, 1967; Tanabe, 1971; Censor et al., 1983; Hanke and Niethammer, 1990), the pivotal theoretical result supporting the use of randomized selection for the Kaczmarz method was given by Strohmer and Vershynin (2009). They proved a non-asymptotic linear convergence rate (in expectation) in terms of the number of iterations, when rows are selected proportional to their squared norms. This work spurred numerous extensions and generalizations of the randomized Kaczmarz method (Needell, 2010; Leventhal and Lewis, 2010; Zouzias and Freris, 2013; Lee and Sidford, 2013; Liu and Wright, 2014; Ma et al., 2015), including similar rates when we replace the equality constraints with inequality constraints.

Rather than cyclic or randomized, in this work we consider *greedy* selection rules. There are very few results

[*] Department of Computer Science, The University of British Columbia, Vancouver, BC, Canada

[†] DATO, Seattle, Washington, USA

in the literature that explore the use of greedy selection rules for Kaczmarz-type methods. Griebel and Oswald (2012) present the *maximum residual rule* for multiplicative Schwarz methods, for which the randomized Kaczmarz iteration is a special case. Their theoretical results show similar convergence rate estimates for both greedy and random methods, suggesting there is no advantage of greedy selection over randomized selection (since greedy selection has additional computational costs). Eldar and Needell (2011) propose a greedy *maximum distance rule*, which they approximate using the Johnson-Lindenstrauss (1984) transform to reduce the computation cost. They show that this leads to a faster algorithm in practice, and show that this rule may achieve more progress than random selection on certain iterations.

In the next section, we define several relevant problems of interest in machine learning that can be solved via Kaczmarz methods. Subsequently, we define the greedy selection rules and discuss cases where they can be computed efficiently. In Section 4 we give faster convergence rate analyses for both the maximum residual rule and the maximum distance rule, which clarify the relationship of these rules to random selection and show that greedy methods will typically have better convergence rates than randomized selection. Section 5 contrasts Kaczmarz methods with coordinate descent methods, Section 6 considers a simplified setting where we explicitly compute the constants in the convergence rates, Section 7 considers how these convergence rates are changed under *approximations* to the greedy rules, and Section 8 discusses the case of inequality constraints. We further give a non-trivial *multi-step* analysis of the maximal residual rule (Section 9), which is the first multi-step analysis of any Kaczmarz algorithm. By taking the multi-step perspective, we also propose provably-faster randomized selection rules for matrices A with pairwise-orthogonal rows by using the so-called “orthogonality graph”. Section 10 presents numerical experiments evaluating greedy Kaczmarz methods.

2 PROBLEMS OF INTEREST

We first consider systems of linear equations,

$$Ax = b, \tag{1}$$

where A is an $m \times n$ matrix and $b \in \mathbb{R}^m$. We assume the system is *consistent*, meaning a solution x^* exists. We denote the rows of A by $a_1^\top, \dots, a_m^\top$, where each $a_i \in R^n$, and use $b = (b_1, \dots, b_m)^\top$, where each $b_i \in \mathbb{R}$. One of the most important examples of a consistent linear system, and a fundamental model in machine learning, is the least squares problem,

$$\min_{x \in \mathbb{R}^n} \frac{1}{2}\|Ax - b\|^2.$$

An appealing way to write a least squares problem as a linear system is to solve the $(n + m)$-variable consistent system (see also Zouzias and Freris, 2013)

$$\begin{pmatrix} A & -I \\ \mathbf{0} & A^T \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} b \\ 0 \end{pmatrix},$$

Other applications in machine learning that involve solving consistent linear systems include: least-squares support vector machines, Gaussian processes, fitting the final layer of a neural network (using squared-error), graph-based semi-supervised learning or other graph-Laplacian problems (Bengio et al., 2006), and finding the optimal configuration in Gaussian Markov random fields (Rue and Held, 2005).

Kaczmarz methods can also be applied to solve consistent systems of linear *inequalities*,

$$Ax \leq b,$$

or combinations of linear equalities and inequalities. We believe there is a lot potential to use this application of Kaczmarz methods in machine learning. Indeed, a classic example of solving linear inequalities is finding a linear separator for a binary classification problem. The classic perceptron algorithm is a generalization of the Kaczmarz method, but unlike the classic sublinear rates of perceptron methods (Novikoff, 1962) we can show a linear rate for the Kaczmarz method.

Kaczmarz methods could also be used to solve the ℓ_1-regularized robust regression problem,

$$\min_x f(x) := \|Ax - b\|_1 + \lambda\|x\|_1,$$

for $\lambda \geq 0$. We can formulate finding an x with $f(x) \leq \tau$ for some constant τ as a set of linear inequalities. By doing a binary search for τ and using *warm-starting*, this can be substantially faster than existing approaches like stochastic subgradient methods (which have a sublinear convergence rate) or formulating as a linear program (which is not scaleable due to the super-linear cost). The above logic applies to many piecewise-linear problems in machine learning like variants of support vector machines/regression with the ℓ_1-norm, regression under the ℓ_∞-norm, and linear programming relaxations for decoding in graphical models.

3 KACZMARZ ALGORITHM AND GREEDY SELECTION RULES

The Kaczmarz algorithm for solving linear systems begins from an initial guess x^0, and each iteration k chooses a row i_k and projects the current iterate x^k onto the hyperplane defined by $a_{i_k}^T x^k = b_{i_k}$. This gives the iteration

$$x^{k+1} = x^k + \frac{b_{i_k} - a_{i_k}^T x^k}{\|a_{i_k}\|^2} a_{i_k}, \tag{2}$$

and the algorithm converges to a solution x^* under weak conditions (e.g., each i is visited infinitely often).

We consider two greedy selection rules: the *maximum residual* rule and the *maximum distance* rule. The maximum residual (MR) rule selects i_k according to

$$i_k = \underset{i}{\text{argmax}}\ |a_i^T x^k - b_i|, \tag{3}$$

which is the equation i_k that is 'furthest' from being satisfied. The maximum distance (MD) rule selects i_k according to

$$i_k = \underset{i}{\text{argmax}}\ \left|\frac{a_i^T x^k - b_i}{\|a_i\|}\right|, \tag{4}$$

which is the rule that maximizes the distance between iterations, $\|x^{k+1} - x^k\|$.

3.1 EFFICIENT CALCULATIONS FOR SPARSE A

In general, computing these greedy selection rules exactly is too computationally expensive, but in some applications we can compute them efficiently. For example, consider a *sparse* A with at most c non-zeros per column and at most r non-zeros per row. In this setting, we show in Appendix 3.1 that both rules can be computed exactly in $O(cr \log m)$ time, using that projecting onto row i does not change the residual of row j if a_i and a_j do not share a non-zero index.

The above sparsity condition guarantees that row i is orthogonal to row j, and indeed projecting onto row i will not change the residual of row j under the more general condition that a_i and a_j are orthogonal. Consider what we call the *orthogonality graph*: an undirected graph on m nodes where we place on edge between nodes i and j if a_i is not orthogonal to a_j. Given this graph, to update all residuals after we update a row i we only need to update the neighbours of node i in this graph. Even if A is dense ($r = n$ and $c = m$), if the maximum number of neighbours is g, then tracking the maximum residual costs $O(gr + g\log(m))$. If g is small, this could still be comparable to the $O(r + \log(m))$ cost of using existing randomized selection strategies.

3.2 APPROXIMATE CALCULATION

Many applications, particularly those arising from graphical models with a simple structure, will allow efficient calculation of the greedy rules using the method of the previous section. However, in other applications it will be too inefficient to calculate the greedy rules. Nevertheless, Eldar and Needell (2011) show that it's possible to efficiently select an i_k that *approximates* the greedy rules by making use of the dimensionality reduction technique of Johnson and Lindenstrauss (1984). Their experiments show that approximate greedy rules can be sufficiently accurate and that they still outperform random selection. After first analyzing exact greedy rules in the next section, we analyze the effect of using approximate rules in Section 7.

4 ANALYZING SELECTION RULES

All the convergence rates we discuss use the following relationship between $\|x^{k+1} - x^*\|$ and $\|x^k - x^*\|$:

$$\begin{aligned}&\|x^{k+1} - x^*\|^2 \\ &= \|x^k - x^*\|^2 - \|x^{k+1} - x^k\|^2 + 2\underbrace{\langle x^{k+1} - x^*, x^{k+1} - x^k\rangle}_{(=0,\text{ by orthogonality})}.\end{aligned}$$

Using the definition of x^{k+1} from (2) and simplifying, we obtain for the selected i_k that

$$\|x^{k+1} - x^*\|^2 = \|x^k - x^*\|^2 - \frac{\left(a_{i_k}^T x^k - b_{i_k}\right)^2}{\|a_{i_k}\|^2}. \tag{5}$$

4.1 RANDOMIZED AND MAXIMUM RESIDUAL

We first give an analysis of the Kaczmarz method with *uniform* random selection of the row to update i (which we abbreviate as 'U'). Conditioning on the σ-field $\mathcal{F}_{k-1}$ generated by the sequence $\{x^0, x^1, \ldots, x^{k-1}\}$, and taking expectations of both sides of (5), when i_k is selected using U we obtain

$$\begin{aligned}&\mathbb{E}[\|x^{k+1} - x^*\|^2] \\ &= \|x^k - x^*\|^2 - \mathbb{E}\left[\frac{\left(a_i^T x^k - b_i\right)^2}{\|a_i\|^2}\right] \\ &= \|x^k - x^*\|^2 - \sum_{i=1}^m \frac{1}{m}\frac{(a_i^\top(x^k - x^*))^2}{\|a_i\|^2} \\ &\le \|x^k - x^*\|^2 - \frac{1}{m\|A\|_{\infty,2}^2}\sum_{i=1}^m (a_i^\top(x^k - x^*))^2 \\ &= \|x^k - x^*\|^2 - \frac{1}{m\|A\|_{\infty,2}^2}\|A(x^k - x^*)\|^2 \\ &\le \left(1 - \frac{\sigma(A,2)^2}{m\|A\|_{\infty,2}^2}\right)\|x^k - x^*\|^2,\end{aligned} \tag{6}$$

where $\|A\|_{\infty,2}^2 := \max_i\{\|a_i\|^2\}$ and $\sigma(A,2)$ is the Hoffman (1952) constant. We've assumed that x^k is not a solution, allowing us to use Hoffman's bound. When A has independent columns, $\sigma(A,2)$ is the nth singular value of A and in general it is the smallest non-zero singular value.

The argument above is related to the analysis of Vishnoi (2013) but is simpler due to the use of the Hoffman bound. Further, this simple argument makes it straightforward to derive bounds on other rules. For example, we can derive the convergence rate bound of Strohmer and Vershynin (2009) by following the above steps but selecting i non-uniformly with probability $\|a_i\|^2/\|A\|_F^2$ (where $\|A\|_F$ is the Frobenius norm of A). We review these steps in Appendix 4.1, showing that this non-uniform (NU) selection strategy has

$$\mathbb{E}[\|x^{k+1} - x^*\|^2] \leq \left(1 - \frac{\sigma(A,2)^2}{\|A\|_F^2}\right)\|x^k - x^*\|^2. \quad (7)$$

This strategy requires prior knowledge of the row norms of A, but this is a one-time computation that can be reused for any linear system involving A. Because $\|A\|_F^2 \leq m\|A\|_{\infty,2}^2$, the NU rate (7) is at least as fast as the uniform rate (6).

While a trivial analysis shows that the MR rule also satisfies (6) in a deterministic sense, in Appendix 4.1 we give a tighter analysis of the MR rule showing it has the convergence rate

$$\|x^{k+1} - x^*\|^2 \leq \left(1 - \frac{\sigma(A,\infty)^2}{\|A\|_{\infty,2}^2}\right)\|x^k - x^*\|^2, \quad (8)$$

where the Hoffman-like constant $\sigma(A,\infty)$ satisfies the relationship

$$\frac{\sigma(A,2)}{\sqrt{m}} \leq \sigma(A,\infty) \leq \sigma(A,2).$$

Thus, at one extreme the maximum residual rule obtains the same rate as (6) for uniform selection when $\sigma(A,2)^2/m \approx \sigma(A,\infty)^2$. However, at the other extreme the maximum residual rule could be faster than uniform selection by a factor of m ($\sigma(A,\infty)^2 \approx \sigma(A,2)^2$). Thus, although the uniform and MR bounds are the same in the worst case, the MR rule can be superior by a large margin.

In contrast to comparing U and MR, the MR rate may be faster or slower than the NU rate. This is because

$$\|A\|_{\infty,2} \leq \|A\|_F \leq \sqrt{m}\|A\|_{\infty,2},$$

so these quantities and the relationship between $\sigma(A,2)$ and $\sigma(A,\infty)$ influence which bound is tighter.

4.2 TIGHTER UNIFORM AND MR ANALYSIS

In our derivations of rates (6) and (8), we use the following inequality

$$\|a_i\|^2 \leq \|A\|_{\infty,2}^2 \quad \forall\, i, \quad (9)$$

which leads to a simple result but could be very loose if the range of row norms is large. In this section, we give tighter analyses of the U and MR rules that are less interpretable but are tighter because they avoid this inequality.

In order to avoid using this inequality for our analysis of U, we can absorb the row norms of A into a row weighting matrix D, where $D = \text{diag}(\|a_1\|, \|a_2\|, \ldots, \|a_m\|)$. Defining $\bar{A} := D^{-1}A$, we show in Appendix 4.2 that this results in the following upper bound on the convergence rate for uniform random selection,

$$\mathbb{E}[\|x^{k+1} - x^*\|^2] \leq \left(1 - \frac{\sigma(\bar{A},2)^2}{m}\right)\|x^k - x^*\|^2. \quad (10)$$

A similar result is given by Needell et al. (2015) under the stronger assumption that A has independent columns. The rate in (10) is tighter than (6), since $\sigma(A,2)/\|A\|_{\infty,2} \leq \sigma(\bar{A},2)$ (van der Sluis, 1969). Further, *this rate can be faster than the non-uniform sampling method* of Strohmer and Vershynin (2009). For example, suppose row i is orthogonal to all other rows but has a significantly larger row norm than all other row norms. In other words, $\|a_i\| >> \|a_j\|$ for all $j \neq i$. In this case, NU selection will repeatedly select row i (even though it only needs to be selected once), whereas U will only select it on each iteration with probability $1/m$. It has been previously pointed out that Strohmer and Vershynin's method can perform poorly if you have a problem where one row norm is significantly larger than the other row norms (Censor et al., 2009). This result theoretically shows that U can have a tighter bound than the NU method of Strohmer and Vershynin.

In Appendix 4.2, we also give a simple modification of our analysis of the MR rule, which leads to the rate

$$\|x^{k+1} - x^*\|^2 \leq \left(1 - \frac{\sigma(A,\infty)^2}{\|a_{i_k}\|^2}\right)\|x^k - x^*\|^2. \quad (11)$$

This bound depends on the *specific* $\|a_{i_k}\|$ corresponding to the i_k selected at each iteration k. This convergence rate will be faster whenever we select an i_k with $\|a_{i_k}\| < \|A\|_{\infty,2}$. However, in the worst case we repeatedly select i_k values with $\|a_{i_k}\| = \|A\|_{\infty,2}$ so there is no improvement. In Section 9, we return to this issue and give tighter bounds on the *sequence* of $\|a_{i_k}\|$ values for problems with sparse orthogonality graphs.

4.3 MAXIMUM DISTANCE RULE

If we can only perform one iteration of the Kaczmarz method, the *optimal* rule (with respect to iteration progress) is in fact the MD rule. In Appendix 4.3, we show that this strategy achieves a rate of

$$\|x^{k+1} - x^*\|^2 \leq \left(1 - \sigma(\bar{A},\infty)^2\right)\|x^k - x^*\|^2, \quad (12)$$

where $\sigma(\bar{A},\infty)$ satisfies

$$\max\left\{\frac{\sigma(\bar{A},2)}{\sqrt{m}}, \frac{\sigma(A,2)}{\|A\|_F}, \frac{\sigma(A,\infty)}{\|A\|_{\infty,2}}\right\} \leq \sigma(\bar{A},\infty) \leq \sigma(\bar{A},2).$$

Thus, the maximum distance rule is at least as fast as the fastest among the U/NU/MR$_\infty$ rules, where MR$_\infty$ refers to rate (8). Further, in Appendix 7.3 we show that this new rate is not only simpler but is strictly tighter than the rate reported by Eldar and Needell (2011) for the exact MD rule.

Table 1: Comparison of Convergence Rates

	U_∞	U	NU	MR_∞	MR	MD
U_∞	$=$	$\leq$	$\leq$	$\leq$	$\leq$	$\leq$
U		$=$	P	P	P	$\leq$
NU			$=$	P	P	$\leq$
MR_∞				$=$	$\leq$	$\leq$
MR					$=$	$\leq$
MD						$=$

In Table 4.3, we summarize the relationships we have discussed in this section among the different selection rules. We use the following abbreviations: U_∞ - uniform (6), U - tight uniform (10), NU - non-uniform (7), MR_∞ - maximum residual (8), MR - tight maximum residual (11) and MD - maximum distance (12). The inequality sign ($\leq$) indicates that the rate for the selection rule listed in the row is slower or equal to the rule listed in the column, while we have written 'P' to indicate that the faster method is problem-dependent.

5 KACZMARZ AND COORDINATE DESCENT

With the exception of the tighter U and MR rate, the results of the previous section are analogous to the recent results of Nutini et al. (2015) for coordinate descent methods. Indeed, if we apply coordinate descent methods to minimize the squared error between Ax and b then we obtain similar-looking rates and analogous conclusions. With cyclic selection this is called the Gauss-Seidel method, and as discussed by Ma et al. (2015) there are several connections/differences between this method and Kaczmarz methods. In this section we highlight some key differences.

First, the previous work required strong-convexity which would require that A has independent columns. This is often unrealistic, and our results from the previous section hold for any A. Second, here our results are in terms of the iterates $\|x^k - x^*\|$, which is the natural measure for linear systems. The coordinate descent results are in terms of $f(x^k) - f(x^*)$ and although it's possible to use strong-convexity to turn this into a rate on $\|x^k - x^*\|$, this would result in a looser bound and would again require strong-convexity to hold (see Ma et al., 2015). On the other hand, coordinate descent gives the least squares solution for inconsistent systems. However, this is also true of Kaczmarz method using the formulation in Section 2. Another subtle issue is that the Kaczmarz rates depend on the row norms of A while the coordinate descent rates depend on the column norms. Thus, there are scenarios where we expect Kaczmarz methods to be much faster and vice versa. Finally, we note that Kaczmarz methods can be extended to allow inequality constraints (see Section 8).

As discussed by Wright (2015), Kaczmarz methods can also be interpreted as coordinate descent methods on the dual problem

$$\min_y \frac{1}{2}\|A^T y\|^2 - b^T y, \tag{13}$$

where $x = A^T y^*$ so that $Ax = AA^T y^* = b$. Applying the Gauss-Southwell rule in this setting yields the MR rule while applying the Gauss-Southwell-Lipschitz rule yields the MD rule (see Appendix 5 for details and numerical comparisons, indicating that in some cases Kaczmarz substantially outperforms CD). However, applying the analysis of Nutini et al. (2015) to this dual problem would require that A has independent rows and would only yield a rate on the dual objective, unlike the convergence rates in terms of $\|x^k - x^*\|$ that hold for general A from the previous section.

6 EXAMPLE: DIAGONAL A

To give a concrete example of these rates, we consider the simple case of a diagonal A. While such problems are not particularly interesting, this case provides a simple setting to understand these different rates without referring to Hoffman bounds.

Consider a square diagonal matrix A with $a_{ii} > 0$ for all i. In this case, the diagonal entries are the eigenvalues λ_i of the linear system. The convergence rate constants for this scenario are given in Table 2. We provide the details in Appendix 6 of the derivations for $\sigma(A, \infty)$ and $\sigma(\bar{A}, \infty)$, as well as substitutions for the uniform, non-uniform, and uniform tight rates to yield the above table. We note that the uniform tight rate follows from $\lambda_m^2(\bar{A})$ being equivalent to the minimum eigenvalue of the identity matrix.

Table 2: Convergence Rate Constants for Diagonal A

U_∞	$\left(1 - \frac{\lambda_m^2}{m\lambda_1^2}\right)$
U	$\left(1 - \frac{1}{m}\right)$
NU	$\left(1 - \frac{\lambda_m^2}{\sum_i \lambda_i^2}\right)$
MR_∞	$\left(1 - \frac{1}{\lambda_1^2}\left[\sum_i \frac{1}{\lambda_i^2}\right]^{-1}\right)$
MR	$\left(1 - \frac{1}{\lambda_{i_k}^2}\left[\sum_i \frac{1}{\lambda_i^2}\right]^{-1}\right)$
MD	$\left(1 - \frac{1}{m}\right)$

If we consider the most basic case when all the eigenvalues of A are equal, then all the selection rules yield the same rate of $(1 - 1/m)$ and the method converges in at most m steps for greedy selection rules and in at most $O(m \log m)$ steps (in expectation) for the random rules (due to the 'coupon collector' problem). Further, this is the worst situation for the greedy MR and MD rules since they satisfy their lower bounds on $\sigma(A, \infty)$ and $\sigma(\bar{A}, \infty)$.

Now consider the extreme case when all the eigenvalues are equal except for one. For example, consider when $\lambda_1 = \lambda_2 = \cdots = \lambda_{m-1} > \lambda_m$ with $m > 2$. Letting $\alpha = \lambda_i^2(A)$ for any $i = 1, \ldots, m-1$ and $\beta = \lambda_m^2(A)$, we have

$$\underbrace{\frac{\beta}{m\alpha}}_{\mathrm{U}_\infty} < \underbrace{\frac{\beta}{\alpha(m-1)+\beta}}_{\mathrm{NU}} < \underbrace{\frac{\beta}{\alpha+\beta(m-1)}}_{\mathrm{MR}_\infty}$$
$$\leq \underbrace{\frac{1}{\lambda_{i_k}^2}\frac{\alpha\beta}{\alpha+\beta(m-1)}}_{\mathrm{MR}} < \underbrace{\frac{1}{m}}_{\mathrm{U,\,MD}}.$$

Thus, Strohmer and Vershynin's NU rule would actually be the worst rule to use, whereas U and MD are the best. In this case $\sigma(A, \infty)^2$ is closer to its upper bound ($\approx \beta$) so we would expect greedy rules to perform well.

7 APPROXIMATE GREEDY RULES

In many applications, computing the exact MR or MD rule will be too inefficient, but we can always approximate it using a cheaper *approximate* greedy rule, as in the method of Eldar and Needell (2011). In this section we consider methods that compute the greedy rules up to multiplicative or additive errors.

7.1 MULTIPLICATIVE ERROR

Suppose we have approximated the MR rule such that there is a multiplicative error in our selection of i_k,

$$|a_{i_k}^T x^k - b_{i_k}| \geq \max_i |a_i^T x^k - b_i|(1 - \epsilon_k),$$

for some $\epsilon_k \in [0, 1)$. In this scenario, using the tight analysis for the MR rule, we show in Appendix 7.1 that

$$\|x^{k+1} - x^*\|^2 \leq \left(1 - \frac{(1-\epsilon_k)^2\sigma(A,\infty)^2}{\|a_{i_k}\|^2}\right)\|x^k - x^*\|^2.$$

Similarly, if we approximate the MD rule up to a multiplicative error,

$$\left|\frac{a_{i_k}^T x^k - b_{i_k}}{\|a_{i_k}\|}\right| \geq \max_i \left|\frac{a_i^T x^k - b_i}{\|a_i\|}\right|(1 - \bar{\epsilon}_k),$$

for some $\bar{\epsilon}_k \in [0, 1)$, then we show in Appendix 7.1 that the following rate holds,

$$\|x^{k+1} - x^*\|^2 \leq \left(1 - (1-\bar{\epsilon}_k)^2\sigma(\bar{A},\infty)^2\right)\|x^k - x^*\|^2.$$

These scenarios do not require the error to converge to 0. However, if ϵ_k or $\bar{\epsilon}_k$ is large, then the convergence rate will be slow.

7.2 ADDITIVE ERROR

Suppose we select i_k using the MR rule up to additive error,

$$|a_{i_k}^T x^k - b_{i_k}|^2 \geq \max_i |a_i^T x^k - b_i|^2 - \epsilon_k,$$

or similarly for the MD rule,

$$\left|\frac{a_{i_k}^T x^k - b_{i_k}}{\|a_{i_k}\|}\right|^2 \geq \max_i \left|\frac{a_i^T x^k - b_i}{\|a_i\|}\right|^2 - \bar{\epsilon}_k,$$

for some $\epsilon_k \geq 0$ or $\bar{\epsilon}_k \geq 0$, respectively. We show in Appendix 7.2 that this results in the following convergence rates for the MR and MD rules with additive error (respectively),

$$\|x^{k+1} - x^*\|^2 \leq \left(1 - \frac{\sigma(A,\infty)^2}{\|a_{i_k}\|^2}\right)\|x^k - x^*\|^2 + \frac{\epsilon_k}{\|a_{i_k}\|^2},$$

and

$$\|x^{k+1} - x^*\|^2 \leq \left(1 - \sigma(\bar{A},\infty)^2\right)\|x^k - x^*\|^2 + \bar{\epsilon}_k.$$

With an additive error, we need the errors to go to 0 in order for the algorithm to converge; if it does go to 0 fast enough, we obtain the same rate as if we were calculating the exact greedy rule. In the approximate greedy rule used by Eldar and Needell (2011), there is unfortunately a constant additive error. To address this, they compare the approximate greedy selection to a randomly selected i_k and take the one with the largest distance. This approach can be substantially faster when far from the solution, but may eventually revert to random selection. We give details comparing Eldar and Needell's rate to our above rate in Appendix 7.3, but here we note that the above bounds will typically be much stronger.

8 SYSTEMS OF LINEAR INEQUALITIES

Kaczmarz methods have been extended to systems of linear inequalities,

$$\begin{cases} a_i^T x \leq b_i & (i \in I_\leq) \\ a_i^T x = b_i & (i \in I_=). \end{cases} \tag{14}$$

where the disjoint index sets $I_\leq$ and $I_=$ partition the set $\{1, 2, \ldots, m\}$ (Leventhal and Lewis, 2010). In this setting the method takes the form

$$x^{k+1} = x^k - \frac{\beta^k}{\|a_i\|^2}a_i,$$

$$\text{with } \beta^k = \begin{cases} (a_i^T x^k - b_i)^+ & (i \in I_\leq) \\ a_i^T x^k - b_i & (i \in I_=), \end{cases}$$

where $(\gamma)^+ = \max\{\gamma, 0\}$. In Appendix 8 we derive analogous greedy rules and convergence results for this case. The main difference in this setting is that the rates are in terms of the distance of x^k to the feasible set S of (14),

$$d(x^k, S) = \min_{z \in S} \|x^k - z\|_2 = \|x^k - P_S(x^k)\|_2,$$

where $P_S(x)$ is the projection of x onto S. This generalization is needed because with inequality constraints the different iterates x^k may have different projections onto S.

9 MULTI-STEP ANALYSIS

All existing analyses of Kaczmarz methods consider convergence rates that depend on a *single* step (in the case of randomized/greedy selection rules) or a single cycle (in the cyclic case). In this section we derive the first tighter *multi*-step convergence rates for iterative Kaczmarz methods; we first consider the MR rule, and then we explore the potential of faster random selection rules. These new rates/rules depend on the orthogonality graph introduced in Section 3.1, and thus in some sense they depend on the 'angle' between rows. This dependence on the 'angle' is similar to the classic convergence rate analyses of cyclic Kaczmarz algorithms, and is a property that is not captured by existing randomized/greedy analyses (which only depend on the row norms).

9.1 MULTI-STEP MAXIMUM RESIDUAL BOUND

If two rows a_i and a_j are orthogonal, then if the equality $a_i^T x^k = b_i$ holds at iteration x^k and we select $i_k = j$, then we know that $a_i^T x^{k+1} = b_i$. More generally, updating i_k makes equality i_k satisfied but could make any equality j unsatisfied where a_j is not orthogonal to a_{i_k}. Thus, after we have selected row i_k, equation i_k will remain satisfied for all subsequent iterations until one of its neighbours is selected in the orthogonality graph. During these subsequent iterations, it cannot be selected by the MR rule since its residual is zero.

In Appendix 9.1, we show how the structure of the orthogonality graph can be used to derive a worst-case bound on the *sequence* of $\|a_{i_k}\|$ values that appear in the tighter analysis of the MR rule (11). In particular, we show that the MR rule achieves a convergence rate of

$$\|x^k - x^*\|^2 \leq$$
$$O(1)\left(\max_{S(G)}\left\{\sqrt[|S(G)|]{\prod_{j \in S(G)}\left(1 - \frac{\sigma(A,\infty)^2}{\|a_j\|^2}\right)}\right\}\right)^k R_0^2,$$

where $R_0 = \|x^0 - x^*\|$ and the maximum is taken over the geometric means of all the *star subgraphs* $S(G)$ of the orthogonality graph with at least two nodes (these are the connected subgraphs that have a diameter of 1 or 2). Although this result is quite complex, even to state, there is a simple implication of it: if the values of $\|a_i\|$ that are close to $\|A\|_{\infty,2}$ are all more than two edges away from each other in the orthogonality graph, then the MR rule converges substantially faster than the worst-case MR$_\infty$ bound (8) indicates.

A multi-step analysis of coordinate descent with the Gauss-Southwell rule and exact coordinate optimization was recently considered by Nutini et al. (2015). To derive this bound, they convert the problem to the same weighted graph construction we use in Appendix 9.1. However, they were only able to derive a bound on this construction in the case of chain-structured graphs. Our result is a generalization of their result to the case of general graphs, and indeed our result is *tighter* than the bound that they conjectured would hold for general graphs. Since the graph construction in this work is the same as in their work, our proof also gives the tightest known bound on coordinate descent with the Gauss-Southwell rule and exact coordinate optimization.

9.2 FASTER RANDOMIZED KACZMARZ RULES

The orthogonality graph can also be used to design faster randomized algorithms. To do this, we use the same property as in the previous section: after we have selected i_k, equality i_k will be satisfied on all subsequent iterations until we select one of its neighbours in the orthogonality graph. Based on this, we call a row i 'selectable' if i has never been selected or if a neighbour of i in the orthogonality graph has been selected since the last time i was selected.[1] We use the notation $s_i^k = 1$ to denote that row i is 'selectable' on iteration k, and otherwise we use $s_i^k = 0$ and say that i is 'not selectable' at iteration k. There is no reason to ever update a 'not selectable' row, because by definition the equality is already satisfied. Based on this, we propose two simple randomized schemes:

1. **Adaptive Uniform**: select i_k uniformly from the selectable rows.
2. **Adaptive Non-Uniform**: select i_k proportional to $\|a_i\|^2$ among the selectable rows.

Let $A_k/\bar{A}_k$ denote the sub-matrix of $A/\bar{A}$ formed by concatenating the selectable rows on iteration k, and let m_k denote the number of selectable rows. If we are given the set of selectable nodes at iteration k, then for adaptive uniform we obtain the bound

$$\mathbb{E}[\|x^{k+1} - x^*\|^2] \leq \left(1 - \frac{\sigma(\bar{A}_k, 2)^2}{m_k}\right)\|x^k - x^*\|^2,$$

[1] If we initialize with $x^0 = 0$, then instead of considering all nodes as initially selectable we can restrict to the nodes i with $b_i \neq 0$ since otherwise we have $a_i^T x^0 = b_i$ already.

while for adaptive non-uniform we obtain the bound

$$\mathbb{E}[\|x^{k+1} - x^*\|^2] \leq \left(1 - \frac{\sigma(A_k, 2)^2}{\|A_k\|_F^2}\right)\|x^k - x^*\|^2.$$

If we are not on the first iteration, then at least one node is not selectable and these are strictly faster than the previous bounds. The gain will be small if most nodes are selectable (which would be typical of dense orthogonality graphs), but the gain can be very large if only a few nodes are selectable (which would be typical of sparse orthogonality graphs).

Theoretical Rate: If we form a vector s^k containing the values s_i^k, it's possible (at least theoretically) to compute the expected value of s^k by viewing it as a Markov chain. In particular, s^0 is a vector of ones while $p(s^{k+1}|s^k)$ is equal to the normalized sum of all ways s^{k+1} could be the set of selectable nodes given the selectable nodes s^k and the orthogonality graph (most $p(s^{k+1}|s^k)$ values will be zero). Given this definition, we can express the probability of a particular s^k recursively using the Chapman-Kolmogorov equations,

$$p(s^{k+1}) = \sum_{s^k} p(s^{k+1}|s^k)p(s^k).$$

If we are interested in the probability that a particular $s_i^k = 1$, we can sum $p(s^k)$ over values s^k compatible with this event. Unfortunately, deriving tighter bound using these probabilities appears to be highly non-trivial.

Practical Issues: In order for the adaptive methods to be efficient, we must be able to efficiently form the orthogonality graph and update the set of selectable nodes. If each node has at most g neighbours in the orthogonality graph, then the cost of updating the set of selectable nodes and then sampling from the set of selectable nodes is $O(g\log(m))$ (we give details in Appendix 9.2). In order for this to not increase the iteration cost compared to the NU method, we only require the very-reasonable assumption that $g\log(m) = O(n+\log(m))$. In many applications where orthogonality is present, g will be far smaller than this.

However, forming the orthogonality graph at the start may be prohibitive: it would cost $O(m^2 n)$ in the worst case to test pairwise orthogonality of all nodes. In the sparse case where each column has at most c non-zeros, we can find an approximation to the orthogonality graph in $O(c^2 n)$ by assuming that all rows which share a non-zero are non-orthogonal. Alternately, in many applications the orthogonality graph is easily derived from the structure of the problem. For example, in graph-based semi-supervised learning where the graph is constructed based on the k-nearest neighbours, the orthogonality graph will simply be the given k-nearest neighbour graph as these correspond the columns that will be mutually non-zero in A.

10 EXPERIMENTS

Eldar and Needell (2011) have already shown that approximate greedy rules can outperform randomized rules for dense problems. Thus, in our experiments we focus on comparing the effectiveness of different rules on very sparse problems where our max-heap strategy allows us to efficiently compute the exact greedy rules. The first problem we consider is generating a dataset A with a 50 by 50 lattice-structured dependency (giving $n = 2500$). The corresponding A has the following non-zero elements: the diagonal elements $A_{i,i}$, the upper/lower diagonal elements $A_{i,i+1}$ and $A_{i+1,i}$ when i is not a multiple of 50 (horizontal edges), and the diagonal bands $A_{i,i+50}$ and $A_{i+50,i}$ (vertical edges). We generate these non-zero elements from a $\mathcal{N}(0,1)$ distribution and generate the target vector $b = Az$ using $z \sim \mathcal{N}(0, I)$. Each row in this problem has at most four neighbours, and this type of sparsity structure is typical of spatial Gaussian graphical models and linear systems that arise from discretizing two-dimensional partial differential equations.

The second problem we consider is solving an overdetermined consistent linear system with a very sparse A of size 2500×1000. We generate each row of A independently such that there are $\log(m)/2m$ non-zero entries per row drawn from a uniform distribution between 0 and 1. To explore how having different row norms affects the performance of the selection rules, we randomly multiply one out of every 11 rows by a factor of 10,000.

For the third problem, we solve a label propagation problem for semi-supervised learning in the 'two moons' dataset (Zhou et al., 2004). From this dataset, we generate 2000 samples and randomly label 100 points in the data. We then connect each node to its 5 nearest neighbours. Constructing a data set with such a high sparsity level is typical of graph-based methods for semi-supervised learning. We use a variant of the quadratic labelling criterion of Bengio et al. (2006),

$$\min_{y_i | i \notin S} \quad \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(y_i - y_j)^2,$$

where y is our label vector (each y_i can take one of 2 values), S is the set of labels that we do know and $w_{ij} \geq 0$ are the weights assigned to each y_i describing how strongly we want the label y_i and y_j to be similar. We can express this quadratic problem as a linear system that is consistent by construction (see Appendix 10), and hence apply Kaczmarz methods. As we labelled 100 points in our data, the resulting linear system has a matrix of size 1900×1900 while the number of neighbours g in the orthogonality graph is at most 5.

In Figure 1 we compare the normalized squared error and distance against the iteration number for 8 different selection rules: cyclic (C), random permutation (RP - where we

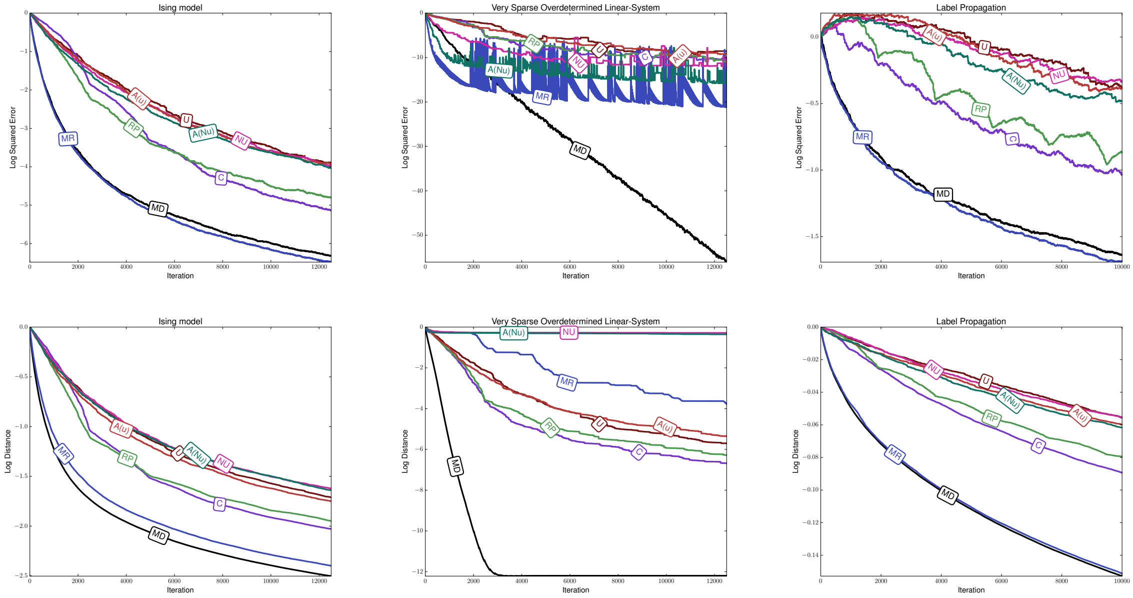

Figure 1: Comparison of Kaczmarz selection rules for squared error (top) and distance to solution (bottom).

change the cycle order after each pass through the rows), uniform random (U), adaptive uniform random (A(u)), non-uniform random NU, adaptive non-uniform random (A(Nu)), maximum residual (MR), and maximum distance (MD).

In experiments 1 and 3, MR performs similarly to MD and both outperform all other selection rules. For experiment 2, the MD rule outperforms all other selection rules in terms of distance to the solution although MR performs better on the early iterations in terms of squared error. In Appendix 10 we explore a 'hybrid' method on the overdetermined linear system problem that does well on both measures. In Appendix 10, we also plot the performance in terms of runtime.

The new randomized A(u) method did not give significantly better performance than the existing U method on any dataset. This agrees with our bounds which show that the gain of this strategy is modest. In contrast, the new randomized A(Nu) method performed much better than the existing NU method on the over-determined linear system in terms of squared error. This again agrees with our bounds which suggest that the A(Nu) method has the most to gain when the row norms are very different. Interestingly, in most experiments we found that *cyclic* selection worked better than any of the randomized algorithms. However, cyclic methods were clearly beaten by greedy methods.

11 DISCUSSION

In this work, we have proven faster convergence rate bounds for a variety of row-selection rules in the context of Kaczmarz methods for solving linear systems. We have also provided new randomized selection rules that make use of orthogonality in the data in order to achieve better theoretical and practical performance. While we have focused on the case of non-accelerated and single-variable variants of the Kaczmarz algorithm, we expect that all of our conclusions also hold for accelerated Kaczmarz and block Kaczmarz methods (Needell and Tropp, 2014; Lee and Sidford, 2013; Liu and Wright, 2014; Gower and Richtárik, 2015; Oswald and Zhou, 2015).

Acknowledgements

This research was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC RGPIN-06068-2015), and Julie Nutini is funded by an NSERC Canada Graduate Scholarship.

References

Y. Bengio, O. Delalleau, and N. Le Roux. Label propagation and quadratic criterion. In O. Chapelle, B. Schölkopf, and A. Zien, editors, *Semi-Supervised Learning*, chapter 11, pages 193–216. MIT Press, 2006.

Y. Censor. Row-action methods for huge and sparse systems and their applications. *SIAM Rev.*, 23(4):444–466, 1981.

Y. Censor, P. B. Eggermont, and D. Gordon. Strong underrelaxation in Kaczmarz's method for inconsistent systems. *Numer. Math.*, 41:83–92, 1983.

Y. Censor, G. T. Herman, and M. Jiang. A note on the behaviour of the randomized Kaczmarz algorithm of

Strohmer and Vershynin. *J. Fourier Anal. Appl.*, 15:431–436, 2009.

F. Deutsch. Rate of convergence of the method of alternating projections. *Internat. Schriftenreihe Numer. Math.*, 72:96–107, 1985.

F. Deutsch and H. Hundal. The rate of convergence for the method of alternating projections, II. *J. Math. Anal. Appl.*, 205:381–405, 1997.

Y. C. Eldar and D. Needell. Acceleration of randomized Kaczmarz methods via the Johnson-Lindenstrauss Lemma. *Numer. Algor.*, 58:163–177, 2011.

H. G. Feichtinger, C. Cenker, M. Mayer, H. Steier, and T. Strohmer. New variants of the POCS method using affine subspaces of finite codimension with applications to irregular sampling. *SPIE: VCIP*, pages 299–310, 1992.

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

R. Gordon, R. Bender, and G. T. Herman. Algebraic Reconstruction Techniques (ART) for three-dimensional electron microscopy and x-ray photography. *J. Theor. Biol.*, 29(3):471–481, 1970.

R. M. Gower and P. Richtárik. Randomized iterative methods for linear systems. *SIAM J. Matrix Anal. Appl.*, 36 (4):1660–1690, 2015.

M. Griebel and P. Oswald. Greedy and randomized versions of the multiplicative Schwartz method. *Lin. Alg. Appl.*, 437:1596–1610, 2012.

M. Hanke and W. Niethammer. On the acceleration of Kaczmarz's method for inconsistent linear systems. *Lin. Alg. Appl.*, 130:83–98, 1990.

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

A. J. Hoffman. On approximate solutions of systems of linear inequalities. *J. Res. Nat. Bur. Stand.*, 49(4):263–265, 1952.

W. B. Johnson and J. Lindenstrauss. Extensions of Lipchitz mappings into a Hilbert space. *Contemp. Math.*, 26:189–206, 1984.

S. Kaczmarz. Angenäherte Auflösung von Systemen linearer Gleichungen, Bulletin International de l'Académie Polonaise des Sciences et des Letters. *Classe des Sciences Mathématiques et Naturelles. Série A, Sciences Mathématiques*, 35:355–357, 1937.

Y. T. Lee and A. Sidford. Efficient accelerated coordinate descent methods and faster algorithms for solving linear systems. *arXiv:1305.1922v1*, 2013.

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

J. Liu and S. J. Wright. An accelerated randomized Kaczmarz method. *arXiv:1310.2887v2*, 2014.

A. Ma, D. Needell, and A. Ramdas. Convergence properties of the randomized extended Gauss-Seidel and Kaczmarz methods. *arXiv:1503.08235v2*, 2015.

D. Needell. Randomized Kaczmarz solver for noisy linear systems. *BIT Numer. Math.*, 50:395–403, 2010.

D. Needell and J. A. Tropp. Paved with good intentions: Analysis of a randomized block Kaczmarz method. *Lin. Alg. Appl.*, 441:199–221, 2014.

D. Needell, N. Srebro, and R. Ward. Stochastic gradient descent and the randomized Kaczmarz algorithm. *arXiv:1310.5715v5*, 2015.

A. B. J. Novikoff. On convergence proofs for perceptrons. *Symp. Math. Theory Automata*, 12:615–622, 1962.

J. Nutini, M. Schmidt, I. H. Laradji, M. Friedlander, and H. Koepke. Coordinate descent converges faster with the Gauss-Southwell rule than random selection. *ICML*, 2015.

P. Oswald and W. Zhou. Convergence analysis for Kaczmarz-type methods in a Hilbert space framework. *Lin. Alg. Appl.*, 478:131–161, 2015.

H. Rue and L. Held. *Gaussian Markov Random Fields: Theory and Applications*. CRC Press, 2005.

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

K. Tanabe. Projection method for solving a singular system of linear equations and its applications. *Numer. Math.*, 17:203–214, 1971.

A. van der Sluis. Condition numbers and equilibrium of matrices. *Numer. Math.*, 14:14–23, 1969.

N. K. Vishnoi. $Lx = b$ Laplacian solvers and their algorithmic applications. *Found. Trends Theoretical Computer Science*, 8(1-2):1–141, 2013.

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

S. J. Wright. Coordinate descent algorithms. *arXiv:1502.04759v1*, 2015.

D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf. Learning with local and global consistency. *NIPS*, 2004.

A. Zouzias and N. M. Freris. Randomized extended Kaczmarz for solving least-squares. *arXiv:1205.5770v3*, 2013.