Convergence Rates for Greedy Kaczmarz Algorithms

Julie Nutini[1], **Behrooz Sepehry**[1], **Alim Virani**[1], **Issam Laradji**[1], **Mark Schmidt**[1], **Hoyt Koepke**[2]
[1]University of British Columbia, [2]Dato

Abstract

We discuss greedy and approximate greedy selection rules within Kaczmarz algorithms for solving linear systems. We show that in some applications the costs of greedy and randomized rules are similar, and that greedy selection gives faster convergence rates. Further, we give a *multi-step* analysis of a particular greedy rule showing it can be much faster when many rows are orthogonal.

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. The Kaczmarz method was originally proposed by Polish mathematician Stefan Kaczmarz [1] and later re-invented by Gordon et. al. [2] under the name *algebraic reconstruction technique* (ART). At each iteration, the Kaczmarz method uses a *selection rule* to choose some row of the matrix and projects the current iterate onto the corresponding hyperplane. Classically, the two categories of selection rules are *cyclic* and *random*, but randomized selection is typically used due to its superior empirical performance. Recently, Strohmer and Vershynin [3] proved a non-asymptotic linear convergence rate (in expectation) when rows are sampled proportional to their squared row norms. This work spurred numerous extensions and generalizations of the randomized Kaczmarz method [4, 5, 6, 7, 8]. In this work we consider *greedy* selection rules. In particular, we consider two greedy rules and show that they can be computed exactly and efficiently for sufficiently sparse matrices A. Subsequently, in Section 3 we give faster convergence rates for both greedy rules, which clarify the relationship of these rules to random selection and show that greedy methods will typically have better convergence rates than randomized selection. In Section 4, we present convergence rates for *approximations* to the greedy rules, as in the greedy hybrid method of Eldar and Needell [9]. We further give a non-trivial *multi-step* bound for one of the rules (Section 5), which we believe is the first multi-step analysis of any Kaczmarz algorithm.

2 Problems of interest, greedy rules and efficient calculations

We consider a linear system of equations, $Ax = b$, where A is an $m \times n$ matrix and $b \in \mathbb{R}^m$. We assume the system is *consistent* (a solution x^* exists). We denote the rows of A by $a_1^T, \ldots, a_m^T$, where each $a_i \in R^n$, and use $b = (b_1, \ldots, b_m)^\top$ where each $b_i \in \mathbb{R}$. Applications in machine learning that involve solving linear systems include: least-squares, least-squares support vector machines, Gaussian processes, fitting the final layer of a neural network using the squared-error, graph-based semi-supervised learning and other graph-Laplacian problems [10], and finding the optimal configuration in Gaussian Markov random fields [11].

The Kaczmarz algorithm 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

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

and the algorithm converges under weak conditions (e.g., each i is visited infinitely often).

We consider two greedy selection rules: the *maximum residual* (MR) rule and the *maximum distance* (MD) rule, respectively:

$$i_k = \underset{i}{\text{argmax}} \ |a_i^T x^k - b_i| \quad \text{(MR)}, \quad i_k = \underset{i}{\text{argmax}} \ \left|a_i^T x^k - b_i\right| / \|a_i\| \quad \text{(MD)}.$$

In general, computing these greedy selection rules exactly is too computationally expensive, but in many applications we can compute them efficiently. For example, consider a *sparse* A with at most c non-zeroes per column and at most r non-zeroes per row. To compute the MR rule in this setting, we can use a max-heap structure. Initializing this structure requires $O(m)$ time (with $x^0 = 0$), computing the MR rule requires $O(1)$ time given the structure, and updating the structure costs $O(cr \log m)$: we need to update at most cr numbers in the heap, each at a cost of $O(\log(m))$. Thus, if c and r are sufficiently small, the MR rule is not much more expensive than random selection.

The reason the MR rule is efficient for sparse A is 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. However, 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 a graph on m nodes, where we place an edge between nodes i and j if a_i is not orthogonal to a_j. We call this the *orthogonality graph*. Given this graph, after we update a row i we only need to update the neighbours of node i in this graph. Even if A is dense, if the maximum number of neighbours is g, then computing the greedy rules costs $O(gn + g\log(m))$. If g is small, this could still be comparable to the $O(n + \log(m))$ cost of using existing randomized selection strategies.

3 Analyzing selection rules

In this section, we give single-step convergence rates for a variety of rules. Due to space restrictions, all proofs are relegated to the extended version of the paper. First, consider the Kaczmarz method with *uniform* (U) random selection. We can show that this selection strategy yields

$$\mathbb{E}[\|x^{k+1} - x^*\|^2] \ \leq \left(1 - \sigma(A)^2/m\|A\|_{\infty,2}^2\right) \|x^k - x^*\|^2, \tag{2}$$

where $\|A\|_{\infty,2}^2 := \max_i\{\|a_i\|^2\}$. When A has independent columns, $\sigma(A)$ is the nth singular value of A. Otherwise, $\sigma(A)$ is the smallest non-zero singular value of A [5].

Next, consider *non-uniform* (NU) random selection, where i is selected non-uniformly with probability $\|a_i\|^2/\|A\|_F^2$ (where $\|\cdot\|_F$ is the Frobenius norm). Strohmer and Vershynin [3] showed

$$\mathbb{E}[\|x^{k+1} - x^*\|^2] \leq \left(1 - \sigma(A)^2/\|A\|_F^2\right)\|x^k - x^*\|^2, \tag{3}$$

which is at least as fast as (2), since $\|A\|_F^2 \leq m\|A\|_{\infty,2}^2$.

For the greedy MR selection rule, we can show a rate of

$$\|x^{k+1} - x^*\|^2 \leq \left(1 - \kappa(A)/\|A\|_{\infty,2}^2\right)\|x^k - x^*\|^2, \quad \text{where } \sigma(A)^2/m \leq \kappa(A) \leq \sigma(A)^2. \tag{4}$$

At one extreme the MR rule obtains the same rate as (2) for U selection, while at the other extreme, the MR rule could be up to m times faster than U selection. In contrast, the MR rate may be faster or slower than the NU rate, as $\|A\|_{\infty,2} \leq \|A\|_F \leq \sqrt{m}\|A\|_{\infty,2}$. Hence, these quantities and the relationship between $\sigma(A)$ and $\kappa(A)$ influence which method is faster.

We can derive a tighter bound for U by absorbing the row norms of A into a row weighting matrix D, where $D = \text{diag}(\|a_1\|, \|a_2\|, \ldots, \|a_m\|)$. Using this, we obtain

$$\mathbb{E}[\|x^{k+1} - x^*\|^2] \leq \left(1 - \sigma(D^{-1}A)^2/m\right)\|x^k - x^*\|^2. \tag{5}$$

A similar result is presented in [12] that includes a relaxation parameter. This rate is tighter than (2), since $\sigma_i(A)/\|A\|_{\infty,2} \leq \sigma_i(D^{-1}A)$ for all i [13]. Further, *this rate can be faster than the NU sampling rate* of Strohmer and Vershynin [3]. For example, suppose row i is orthogonal to all other rows but has a significantly larger row norm than all other row norms, i.e., $\|a_i\| >> \|a_j\| \ \forall j \neq i$. In this case, NU selection will repeatedly select row i (even though it only needs to be selected once), whereas U selection will only select it on each iteration with probability $1/m$.

For one iteration of the Kaczmarz method, the *optimal* rule in terms of $\|x^k - x^*\|$ is in fact the MD rule, and we can show that this achieves

$$\|x^{k+1} - x^*\|^2 \leq (1 - \bar{\kappa}(A)) \|x^k - x^*\|^2, \tag{6}$$

where $\bar{\kappa}(A)$ satisfies $\max\left\{\sigma(D^{-1}A)^2/m, \sigma(A)^2/\|A\|_F^2, \kappa(A)/\|A\|_{\infty,2}^2\right\} \leq \bar{\kappa}(A) \leq \sigma(D^{-1}A)^2$. Thus, the MD rule is at least as fast as the fastest among (3), (4) and (5). This new rate is not only simpler but is strictly tighter than the rate reported by Eldar and Needell [9] for the exact MD rule.

4 Approximate greedy rules

In many applications, only approximate greedy rules will be feasible. If we consider an MD rule with a *multiplicative error*,

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

for some $\bar{\epsilon}_k \in [0, 1)$, then we obtain

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

A similar bound is achieved by the MR rule, and we note that this does not require the error to converge to 0. If we instead have an *additive error* ,

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

for some $\bar{\epsilon}_k \geq 0$, then we obtain

$$\|x^{k+1} - x^*\|^2 \leq \left(1 - \bar{\kappa}(A)\right) \|x^k - x^*\|^2 + \bar{\epsilon_k}.$$

With an additive error, $\bar{\epsilon}_k$ must to go to 0 in order for the algorithm to converge (though we can avoid this with the hybrid method of Eldar and Needell [9]); but if it does goes to 0 fast enough, we obtain the same rate as if we were calculating the exact greedy rule. In any case, this strategy can be substantially faster when far from the solution.

5 Multi-step maximum residual bound

A simple modification of the analysis for the MR rule leads to the rate

$$\|x^{k+1} - x^*\|^2 \leq \left(1 - \kappa(A)/\|a_{i_k}\|^2\right) \|x^k - x^*\|^2. \tag{7}$$

This bound depends on the *specific* $\|a_{i_k}\|$ corresponding to the i_k selected at each iteration k. It is strictly faster than (4) whenever $\|a_{i_k}\| < \|A\|_{\infty,2}$, but in the worst case we might have $\|a_{i_k}\| = \|A\|_{\infty,2}$. In this section, we give a tighter bound that holds across the iterations that depends on the *sequence* of i_k values that are chosen. We call this a *multi-step* analysis, as it contrasts with all existing analyses of Kaczmarz methods, which consider convergence rates that depend on the choice of i_k at each iteration (or a cycle of i_k values), but do not consider that the *sequence* of i_k could give a better bound than we obtain using the sequence of bounds.

Using a non-trivial proof, we show how the structure of the orthogonality graph (as presented in Section 2) can be used to derive a worst-case bound on the *sequence* of $\|a_{i_k}\|$ values that appear in our tighter analysis of the MR rule (7). In particular, we show that the MR rule achieves

$$\|x^k - x^*\|^2 \leq O(1) \left(\max_{S(G)} \left\{\sqrt[|S(G)|]{\prod_{j \in S(G)} \left(1 - \frac{\kappa(A)}{\|a_j\|^2}\right)}\right\}\right)^k \|x^0 - x^*\|^2,$$

where 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). The implication of this result is that 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 bound (4) indicates.

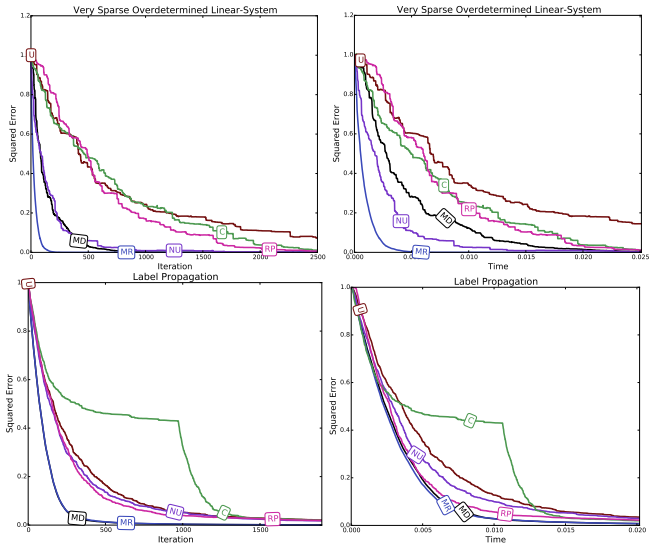

Figure 1: Comparison of Kaczmarz selection rules for both iteration and runtime.

6 Numerical experiments

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 solving is an overdetermined 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 the effect of wildly-different row norms, we randomly multiply one out of every 11 rows by a factor of 10,000. The second problem we consider is a label propagation problem for semi-supervised learning in the 'two moons' dataset [15]. From this dataset, we generate 2000 samples and randomly label 100 points. We then connect each node to its 5 nearest neighbours. We use a variant of the quadratic labelling criterion of Bengio et. al. [10], $\min_{y_i | i \notin S} \frac{1}{2} \sum_i \sum_j w_{ij}(y_i - y_j)^2$, where y is our label vector and S is the set of labels that we do know while $w_{ij} \geq 0$ 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 very sparse linear system, and hence apply Kaczmarz methods.

In Figure 1 we compare the average squared error against the iteration number and the runtime for the 4 previously presented rules, as well as cyclic (C) and random permutation (RP - where the cycle order is permuted after each pass through the rows) rules. In both cases, the MR rule is equal to or significantly better than all other rules in terms of both the number of iterations and the runtime. In contrast, the MD rule was effective than all other existing rules on the label-propagation dataset, but was less effective on the over-determined linear system. This seems paradoxical because we show that the MD rule is the optimal. However, this optimality only applies if we perform a single iteration of the method, and only applies to the distance to the solution. We tried plotting the distance to the solution on this problem and the MD rule does indeed perform better than the other methods. But if we are interested in the squared residual, the MR method seems to be a better choice.

7 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. If the matrix A is extremely low rank, an alternative to Kaczmarz methods are randomized low-rank matrix approximations [16]. However, real datasets are often not extremely low-rank while our methods are still applicable in these cases. 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 [17, 7, 8, 18, 19].

References

[1] 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.

[2] 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.

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

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

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

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

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

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

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

[10] Y. Bengio, O. Delalleau, and N. Le Roux. Label propagation and quadratic criterion. *Semi-Supervised Learning*, pages 193–216, 2006.

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

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

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

[14] 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.

[15] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf. Learning with local and global consistency. *Advances in Neural Information Processing Systems*, 2004.

[16] N. Halko, P. G. Martinsson, and J. A. Tropp. Finding structure with randomness: probabilistic algorithms for construction approximate matrix decompositions. *SIAM Rev.*, 53(2):217–288, 2011.

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

[18] R. M. Gower and P. Richtárik. Randomized iterative methods for linear systems. To appear in*: SIAM J. Matrix Anal. Appl., arXiv:1506.03296v4*, 2015.

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