Comments on the randomized Kaczmarz method

Thomas Strohmer and Roman Vershynin*
Department of Mathematics, University of California
Davis, CA 95616-8633, USA.
strohmer@math.ucdavis.edu, vershynin@math.ucdavis.edu

In this short note we respond to some concerns raised by Y. Censor, G. Herman, and M. Jiang about the randomized Kaczmarz method that we proposed in [5].

The Kaczmarz method is a well-known iterative algorithm for solving a linear system of equations $Ax = b$. For more than seven decades, this method was useful in practical applications, and it was studied in many research papers. Despite of this, little is known about the rate of convergence of this method. The classical scheme of Kaczmarz's method sweeps through the rows of A in a cyclic manner, projecting in each substep the last iterate orthogonally onto a hyperplane associated with a row of A. One variation of Kaczmarz's method consists of randomly choosing in each iteration the row for the projection. Our algorithm in [5] (labeled *Algorithm 1* there) is based on this approach. The idea of choosing the rows randomly is certainly not new. It has been mentioned for instance by Natterer [4], and later also by Feichtinger et al. [1] and by G. Herman and L. Meyer [3]. In these papers, improvement of performance is observed in numerical experiments, but none of these papers contain any proof of the rate of convergence.

We consider the main contribution of [5] that (i) it contains the first proof for a rate of convergence for the Kaczmarz's method that is applicable to general matrices (and not just to very restricted special cases); (ii) the algorithm achieves an exponential rate of convergence; and (iii) the rate

*T.S. was supported by NSF DMS grant 0511461. R.V. was supported by the Alfred P. Sloan Foundation and by NSF DMS grant 0401032.

of convergence is expressible in terms of standard quantities in numerical analysis (condition numbers of matrices).

Y. Censor, G. Herman, and M. Jiang (CHJ for short) claim in their note that *"the rule proposed in Algorithm 1 cannot be in general optimal (or even in any sense superior)"*. Here, "the rule" refers to choosing the rows for the projection steps randomly according to probabilities given by the row-norms of A. CHJ are correct that this rule is not optimal, and we actually never claimed it to be optimal. In fact, we even give a concrete example in Section 3.2 of [5] for which this rule (as indeed other Kaczmarz-type algorithms) *must* perform poorly. This example is presented in connection with a discussion of optimality of our convergence rate estimates. We prove in Section 3 of [5] that the estimate of the rate of convergence we derive for Algorithm 1 cannot be improved from below beyond a constant factor. Furthermore we show that the estimate cannot be improved from above, since there are matrices (well-conditioned ones and ill-conditioned ones) for which equality in our estimate is obtained. One should not confuse optimality of our estimates with optimality of "the rule" with which we pick the rows in our algorithm.

We prove in Theorem 2 of [5] that our randomized Kaczmarz algorithm achieves exponential rate of convergence (in expectation). We are not aware of any similar previous results for general matrices, for any Kaczmarz-based algorithms. CHJ also do not provide any reference for a Kaczmarz-based algorithm that provably achieves exponential rate of convergence (in expectation or otherwise) for general matrices. Our analysis further implies that the convergence rate of our algorithm does not even depend on the number of equations in the system. To the best of our knowledge such results have not been shown for any other Kaczmarz-based algorithm (nor do CHJ provide any reference that would demonstrate otherwise). Thus, to the best of our knowledge, our algorithm can claim superiority over other Kaczmarz methods in that sense (but not the optimality of the rule in which the probabilities are selected, which was never claimed).

Moreover, in the beginning of Section 5 we suggest, based on numerical simulations, how a certain choice of relaxation parameter can further improve the convergence – which would not make sense if we did believe that our algorithm were already optimal.

The value and motivation for choosing the probabilities according to the row-norms of the matrix lies in the following facts:

- It allows us to guarantee the exponential rate of convergence for Kacz-

marz's method;

- It is a computationally efficient strategy, and while not optimal, often provides very good results.

Choosing the probabilities according to the row-norms is related to the idea of preconditioning a matrix by row-scaling. From the viewpoint of preconditioning, it is clear that other methods of choosing a diagonal preconditioner will in general perform better. However, finding the optimal diagonal preconditioner for the system $Ax = b$ is an optimization problem whose complexity can easily exceed that of directly inverting the matrix A. Therefore a cheaper, suboptimal alternative is needed. Scaling by the inverses of the squared row-norms has been shown to be an efficient means to balance computational costs with optimality, see e.g. [**?**], and its suboptimality is well-known [**?**]. From a practical viewpoint, one may not even be willing to spend the computational effort to compute the row-norms of A, since the cost is still in the order of mn operations for an $m \times n$ matrix A.

In some (or many) specific applications better ways for choosing the probabilities can be obtained. This is essentially always possible when a simple way of choosing a better diagonal preconditioner can be found. One such example is the problem of reconstructing a bandlimited function or trigonometric polynomial from its nonuniformly spaced sampling values, that is presented in Section 4.1 of [5].

In connection with Section 4.1 of [5] CHJ state that "*A scaling of the equations will change the system matrix A and its scaled condition number $\kappa(A)$ and, in the light of [5] it might be tempting to think that it is possible to control in such a way the convergence rate of Kaczmarz's method. However, the geometric nature of Kaczmarz's method precludes such a possibility ...*".

CHJ are overlooking an important aspect of our algorithm here. It is definitely correct that a mere preconditioning of the system $Ax = b$ by a diagonal matrix D does not change the convergence rate of the standard Kaczmarz's method at all, since the angle between any two rows of DA is still the same as for A. However, a key element of our randomized Kaczmarz method is that the probabilities with which the rows are chosen are changing when one replaces A by DA. This is crucial, since now rows that are considered "more relevant" (expressed via the scaling by D) are chosen more often, and rows that have less relevance are chosen less often. And how often rows are chosen clearly will have a strong influence on the convergence.

With respect to the numerical example in Section 4.1, CHJ state that " *To avoid inferior behavior of the randomized Kaczmarz algorithm it is essential that the system of algebraic equations that represents the set of hyperplanes be carefully chosen. Indeed, this was done by Strohmer and Vershynin in their numerical simulation in Section 4.1, see their equation (18). Had they selected a different algebraic representation, they would have obtained a different convergence behavior.* " The representation that we supposedly "have carefully chosen" is a standard one, that is based on the well-established work by Feichtinger and Gröchenig [2]. In light of the preconditioner discussion before, the weights proposed by Feichtinger and Gröchenig play exactly the role of a simple diagonal preconditioner by assigning more weight to sampling points where the sampling density is low and less points to sampling points where the sampling density is high. This is a nice and simple way to incorporate geometrical information about the sampling pattern. Again, as mentioned above, CHJ seem to overlook that an important aspect of the randomized Kaczmarz method is that when we scale (i.e., diagonally precondition) the system $Ax = b$ we also change the probabilitites with which the rows of the scaled matrix are chosen during the projection steps. This is a simple, (suboptimal) yet efficient way to incorporate the geometry of the (sampling) problem into the algorithm.

CHJ claim that if we had scaled the equations given in equation (18), such that all rows of A would have norm equal to 1, then the difference between Algorithm 1 and the simple randomized Kaczmarz method (where every row is chosen with equal probability) "would have disappeared". This is certainly and obviously correct: in Algorithm 1, each row is selected with probability equal to its squared row-norm, whereas in the simple randomized Kaczmarz algorithm each row is selected with equal probability. However, such scaling can dramatically increase the condition number of matrix A. This, in turn, may lead to a poor performance of the randomized Kaczmarz algorithm on a system scaled in such "wrong" way.

To summarize, the randomized Kaczmarz method offers various advantages over the standard Kaczmarz method. When choosing the rows at random with probabilities equal to the squared row-norms, we are able to give a proof of the expected rate of convergence. This rate is actually exponential, and no comparable convergence rate for Kaczmarz has been proven before in the literature. A proper scaling of the system $Ax = b$ can indeed improve the convergence of the randomized Kaczmarz method (unlike the ordinary Kaczmarz method), since the probabilities with which rows are chosen is changing

with the scaling. Assigning probabilities corresponding to the row-norms is in general certainly not optimal (which was never claimed), but at least for well-conditioned matrices it is very efficient. And it is absolutely possible that somebody, maybe the reader of this note, will come up with a version of the randomized Kaczmarz method that is provably better than the one we proposed in [5].

References

[1] C. Cenker, H. G. Feichtinger, M. Mayer, H. Steier, and T. Strohmer. New variants of the POCS method using affine subspaces of finite codimension, with applications to irregular sampling. In *Proc. SPIE: Visual Communications and Image Processing*, pages 299–310, 1992.

[2] H.G. Feichtinger and K.H. Gröchenig. Theory and practice of irregular sampling. In J. Benedetto and M. Frazier, editors, *Wavelets: Mathematics and Applications*, pages 305–363. CRC Press, 1994.

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

[4] F. Natterer. *The Mathematics of Computerized Tomography*. Wiley, New York, 1986.

[5] A. Shapiro. Optimality bounds for diagonal scaling. *Numer. Math.*, 14:14–23, 1969.

[6] T. Strohmer and R. Vershynin. A randomized kaczmarz algorithm with exponential convergence. *J. Four. Anal. Appl.*, to appear.

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