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

Julie Nutini[1], Behrooz Sepehry[1], Issam Laradji[1],
Alim Virani[1], Mark Schmidt[1], and Hoyt Koepke[2]

[1]The University of British Columbia, [2]Dato

In this document, we derive several results omitted from the main paper due to space limitations. We have numbered the sections in this document according to the section numbers in the main paper.

3.1 Efficient Calculations for Sparse A

To compute the MR rule efficiently for sparse A, we need to store and update the residuals $r_i = (a_i^T x^k - b_i)$ for all i. If we initialize with $x^0 = 0$, then the initial values of the residuals are simply the corresponding b_i values. Given the initial residuals, we can construct a max-heap structure on these residuals in $O(m)$ time. The max-heap structure lets us compute the MR rule in $O(1)$ time. After an iteration of the Kaczmarz method, we can update the max-heap efficiently as follows:

For each j where $x_j^{k+1} \neq x_j^k$:

- **For each i with $a_{ij} \neq 0$:**
  - Update r_i using $r_i \leftarrow r_i - a_{ij}x_j^k + a_{ij}x_j^{k+1}$.
  - Update max-heap using the new value of $|r_i|$.

The cost of each update to an r_i is $O(1)$ and the cost of each heap update is $O(\log m)$. If each row of A has at most r non-zeroes and each column has at most c non-zeroes, then the outer loop is run at most r times while the inner loop is run at most c times for each outer loop iteration. Thus, in the worst case the total cost is $O(cr \log m)$, although it might be much faster if we have particularly sparse rows or columns. Thus, if c and r are sufficiently small, the MR rule is not much more expensive than non-uniform random selection which costs $O(\log m)$. For the MD rule, the cost is the same except there is an extra one-time cost to pre-compute the row norms $\|a_i\|$.
Now consider the case where A may be dense but each row is orthogonal to all but at most g other rows. In this setting it would be too slow to implement the above update of the residuals, since the cost would be $O(mn \log(m))$. In this setting, it makes more sense to use the following alternative approach to update the max-heap after we've updated row i_k:

For each i that is a neighbour of i_k in the orthogonality graph:

- Compute the residual $r_i = a_i^T x^k - b_i$.
- Update max-heap using the new value of $|r_i|$.

We can find the set of neighbours for each node in constant time by keeping a list of each node's neighbours. This loop would run at most g times and the cost of each iteration would be $O(n)$ to update the residual and $O(\log m)$ to update the heap. Thus, the cost to track the residuals using this alternative approach would be $O(gn + g\log(m))$ or the faster $O(gr + g\log(m))$ if each row has at most r non-zeros.

4.1 Randomized and Maximum Residual

In this section, we provide details of the convergence rate derivations for the non-uniform and maximum residual (MR) selection rules. All the convergence rates we discuss use the following relationship,

$$\begin{aligned}\|x^{k+1} - x^*\|^2 &= \|x^k - x^*\|^2 - \|x^{k+1} - x^k\|^2 \\ &= \|x^k - x^*\|^2 - \left\| \frac{(b_i - a_i^T x^k)}{\|a_i\|^2} \cdot a_i \right\|^2 \\ &= \|x^k - x^*\|^2 - \frac{\left(a_i^T x^k - b_i\right)^2}{\|a_i\|^2}, \end{aligned} \tag{1}$$

which is equation (5) in the main paper.

Non-Uniform

We review the steps discussed by Vishnoi (2013) that can be used to derive the convergence rate bound of Strohmer and Vershynin (2009) for non-uniform random selection when row i is chosen according to the probability distribution determined by $\|a_i\|/\|A\|_F$. Taking the expectation of (1) with respect to i, we have

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

where $\sigma(A, 2)$ is the Hoffman (1952) constant, which can be defined as the largest value such that for any x that is not a solution to the linear system we have

$$\sigma(A,2)\|x - x^*\| \leq \|A(x - x^*)\|, \tag{3}$$

where x^* is the projection of x onto the set of solutions S. In other words, we can write it as

$$\sigma(A,2) := \inf_{x \notin S} \frac{\|A(x - x^*)\|}{\|x - x^*\|}.$$

Strohmer and Vershynin (2009) consider the special case where A has independent columns, and this result yields their rate in this special case since under this assumption $\sigma(A,2)$ is given by the nth singular value of A. For general matrices, $\sigma(A,2)$ is given by the smallest non-zero singular value of A.

Maximum Residual

We use a similar analysis to prove a convergence rate bound for the MR rule,

$$i_k = \underset{i}{\operatorname{argmax}} \ |a_i^T x^k - b_i|. \tag{4}$$

Assuming that i is selected according to (4), then starting from (1) we have

$$\begin{aligned}
\|x^{k+1} - x^*\|^2 &= \|x^k - x^*\|^2 - \max_i \frac{(a_i^T x^k - b_i)^2}{\|a_i\|^2} \\
&\leq \|x^k - x^*\|^2 - \frac{1}{\|A\|_{\infty,2}^2} \max_i \ (a_i^T (x^k - x^*))^2 \\
&= \|x^k - x^*\|^2 - \frac{\|A(x^k - x^*)\|_\infty^2}{\|A\|_{\infty,2}^2} \\
&\leq \left(1 - \frac{\sigma(A,\infty)^2}{\|A\|_{\infty,2}^2}\right) \|x^k - x^*\|^2,
\end{aligned} \tag{5}$$

where $\|A\|_{\infty,2}^2 := \max_i\{\|a_i\|^2\}$ and $\sigma(A,\infty)$ is the largest value such that

$$\sigma(A,\infty)\|x - x^*\| \leq \|A(x - x^*)\|_\infty, \tag{6}$$

or equivalently

$$\sigma(A,\infty) := \inf_{x \notin S} \frac{\|A(x - x^*)\|_\infty}{\|x - x^*\|}.$$

The existence of such a Hoffman-like constant follows from the existence of the Hoffman constant and the equivalence between norms. Applying the norm equivalence $\|\cdot\|_\infty \geq \frac{1}{\sqrt{m}}\|\cdot\|$ to equation (3) we have

$$\sigma(A,2)\|x - x^*\| \leq \|A(x - x^*)\| \leq \sqrt{m}\|A(x - x^*)\|_\infty,$$

which implies that $\sigma(A,2)/\sqrt{m} \leq \sigma(A,\infty)$. Similarly, applying $\|\cdot\|_\infty \leq \|\cdot\|$ to (6) we have

$$\sigma(A,\infty)\|x - x^*\| \leq \|A(x - x^*)\|_\infty \leq \|A(x - x^*)\|,$$

which implies that $\sigma(A,\infty)$ cannot be larger than $\sigma(A,2)$. Thus, $\sigma(A,\infty)$ satisfies the relationship

$$\frac{\sigma(A,2)}{\sqrt{m}} \leq \sigma(A,\infty) \leq \sigma(A,2). \tag{7}$$

4.2 Tighter Uniform and MR Analysis

To avoid using the inequality $\|a_i\| \leq \|A\|_{\infty,2}$ for all i, we want to 'absorb' the individual row norms into the bound. We start with uniform selection.

Uniform

Consider the diagonal matrix $D = \text{diag}(\|a_1\|^2, \|a_2\|^2, \ldots, \|a_m\|^2)$. By taking the expectation of (1), we have

$$\begin{aligned}
\mathbb{E}[\|x^{k+1} - x^*\|^2] &= \|x^k - x^*\|^2 - \mathbb{E}\left[\frac{\left(a_i^T x^k - b_i\right)^2}{\|a_i\|^2}\right] \\
&= \|x^k - x^*\|^2 - \sum_{i=1}^{m} \frac{1}{m} \frac{(a_i^T x^k - b_i)^2}{\|a_i\|^2} \\
&= \|x^k - x^*\|^2 - \frac{1}{m} \sum_{i=1}^{m} \left(\left[\frac{a_i}{\|a_i\|}\right]^T (x^k - x^*) \right)^2 \\
&= \|x^k - x^*\|^2 - \frac{\|D^{-1}A(x^k - x^*)\|^2}{m} \\
&\leq \left(1 - \frac{\sigma(\bar{A}, 2)^2}{m}\right) \|x^k - x^*\|^2, \qquad (8)
\end{aligned}$$

where we used that $Ax = b$ and $\bar{A}x = b$ have the same solution set.

Maximum Residual

For the tighter analysis of the MR rule we do not want to alter the selection rule. Thus, we first evaluate the MR rule and then divide by the corresponding $\|a_{i_k}\|^2$ for the selected i_k at iteration k. Starting from (1), this gives us

$$\begin{aligned}
\|x^{k+1} - x^*\|^2 &= \|x^k - x^*\|^2 - \max_i \frac{(a_i^T x^k - b_i)^2}{\|a_i\|^2} \\
&= \|x^k - x^*\|^2 - \frac{1}{\|a_{i_k}\|^2} \max_i (a_i^T (x^k - x^*))^2 \\
&= \|x^k - x^*\|^2 - \frac{\|A(x^k - x^*)\|_\infty^2}{\|a_{i_k}\|^2} \\
&\leq \left(1 - \frac{\sigma(A, \infty)^2}{\|a_{i_k}\|^2}\right) \|x^k - x^*\|^2. \qquad (9)
\end{aligned}$$

Applying this recursively over all k iterations yields the rate

$$\|x^k - x^*\|^2 \leq \prod_{j=1}^{k} \left(1 - \frac{\sigma(A, \infty)^2}{\|a_{i_j}\|^2}\right) \|x^0 - x^*\|^2. \qquad (10)$$

4.3 Maximum Distance Rule

If we can only perform one iteration of the Kaczmarz method, the *optimal* rule with respect to iterate progress is the maximum distance (MD) rule,

$$i_k = \underset{i}{\text{argmax}} \left| \frac{a_i^T x^k - b_i}{\|a_i\|} \right|. \qquad (11)$$

Starting again from (1) and using D as defined in the tight analysis for the U rule, we have

$$\begin{aligned}\|x^{k+1}-x^*\|^2 &= \|x^k-x^*\|^2 - \max_i \left(\frac{a_i^T x^k - b_i}{\|a_i\|}\right)^2 \\ &= \|x^k-x^*\|^2 - \max_i \left(\left[\frac{a_i}{\|a_i\|}\right]^T (x^k-x^*)\right)^2 \\ &= \|x^k-x^*\|^2 - \|D^{-1}A(x^k-x^*)\|_\infty^2 \\ &\leq \left(1-\sigma(\bar{A},\infty)^2\right)\|x^k-x^*\|^2. \end{aligned} \tag{12}$$

We now show that

$$\max\left\{\frac{\sigma(\bar{A},2)}{\sqrt{m}}, \frac{\sigma(A,2)}{\|A\|_F}, \frac{\sigma(A,\infty)}{\|A\|_{\infty,2}}\right\} \leq \sigma(\bar{A},\infty) \leq \sigma(\bar{A},2). \tag{13}$$

To derive the upper bound on $\sigma(\bar{A},\infty)$, and to derive the lower bound in terms of $\sigma(\bar{A},2)$, we can use norm equivalence arguments as we did for $\sigma(A,\infty)$. This yields

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

The last argument in the maximum in (13), corresponding to the MR_∞ rate, holds because $\|A\|_{\infty,2} \geq \|a_i\|$ for all i so we have

$$\frac{\sigma(A,\infty)}{\|A\|_{\infty,2}}\|x-x^*\| \leq \frac{\|A(x-x^*)\|_\infty}{\|A\|_{\infty,2}} = \max_i\left\{\frac{|a_i^T(x-x^*)|}{\|A\|_{\infty,2}}\right\} \leq \max_i\left\{\frac{|a_i^T(x-x^*)|}{\|a_i\|}\right\} = \|\bar{A}(x-x^*)\|_\infty.$$

For the second argument in the maximum in (13), the NU rate, we have

$$\frac{\sigma(A,2)^2}{\|A\|_F^2}\|x-x^*\|^2 \leq \frac{\|A(x-x^*)\|^2}{\|A\|_F^2} = \frac{\sum_i (a_i^T(x-x^*))^2}{\sum_i \|a_i\|^2} \leq \max_i\left\{\frac{(a_i^T(x-x^*))^2}{\|a_i\|^2}\right\} = \|\bar{A}(x-x^*)\|_\infty.$$

The second inequality is true by noting that it is equivalent to the inequality

$$1 \leq \max_i\left\{\frac{(a_i^T(x-x^*)^2/\sum_j (a_j^T(x-x^*))^2}{\|a_i\|^2/\sum_j \|a_j\|^2}\right\},$$

and this true because the maximum ratio between two probability mass functions must be at least 1,

$$1 \leq \max_i \frac{p_i/\sum_j p_j}{q_i/\sum_j q_j}, \quad \text{with all } p_i \geq 0, q_i \geq 0.$$

Finally, we note that the MD rule obtains the tightest bound in terms of performing one step. This follows from (1),

$$\|x^{k+1}-x^*\|^2 = \|x^k-x^*\|^2 - \|x^{k+1}-x^k\|^2 = \|x^k-x^*\|^2 - \frac{\left(a_i^T x^k - b_i\right)^2}{\|a_i\|^2},$$

and noting that the MD rule maximizes $\|x^{k+1}-x^k\|$ and thus it maximizes how much smaller $\|x^{k+1}-x^*\|$ is than $\|x^k-x^*\|$.

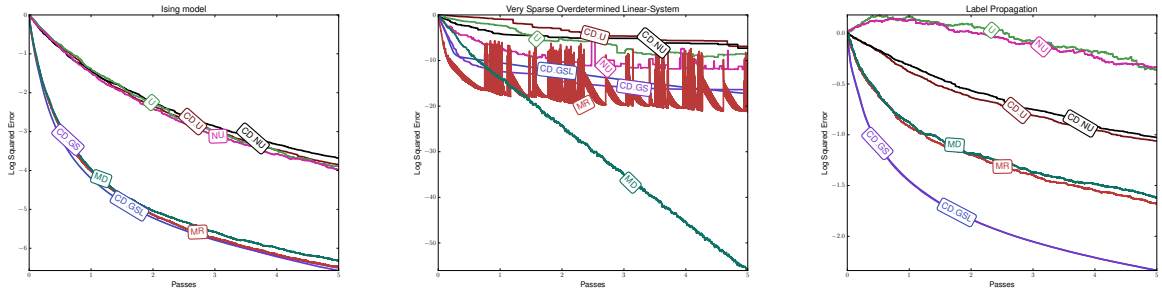

Figure 1: Comparison of Kaczmarz and Coordinate Descent.

5 Kaczmarz and Coordinate Descent

Consider the Kaczmarz update:

$$x^{k+1} = x^k - \frac{(a_i^T x^k - b_i)}{\|a_i\|^2} a_i.$$

This update is equivalent to one step of coordinate descent (CD) with step length $1/\|a_i\|^2$ applied to the dual problem,

$$\min_y \frac{1}{2}\|A^T y\|^2 - b^T y, \tag{14}$$

see Wright (2015). Using the primal-dual relationship $A^T y = x$, we can show the relationship between the greedy Kaczmarz selection rules and applying greedy coordinate descent rules to this dual problem. Consider the gradient of the dual problem,

$$\nabla f(y) = AA^T y - b.$$

The Gauss-Southwell (GS) rule for CD on the dual problem is equivalent to the MR rule for Kaczmarz on the primal problem since

$$i_k = \underbrace{\operatorname*{argmax}_i |\nabla_i f(y^k)|}_{\text{Gauss-Southwell rule}} = \operatorname*{argmax}_i |a_i^T (A^T y^k) - b_i| = \operatorname*{argmax}_i |a_i^T x^k - b_i|$$

where a_i^T is the ith row of A. Similarly, the Gauss-Southwell-Lipschitz (GSL) rule (Nutini et al., 2015) applied to the dual is equivalent to applying a Kaczmarz iteration with the MD rule,

$$i_k = \underbrace{\operatorname*{argmax}_i \frac{|\nabla_i f(y^k)|}{\sqrt{L_i}}}_{\text{Gauss-Southwell-Lipschitz rule}} = \operatorname*{argmax}_i \frac{|a_i^T (A^T y^k) - b_i|}{\|a_i\|} = \operatorname*{argmax}_i \left| \frac{a_i^T x^k - b_i}{\|a_i\|} \right|,$$

as the Lipschitz constants for the dual problem are $L_i = \|a_i\|^2$.

Figure 1 shows the results of running Kaczmarz compared to using CD (on the least-squares primal problem) for our 3 datasets from Section 10 of the main paper. In this figure we measure the performance in terms of the number of "effective passes" through the data (one "effective" pass would be the number of iterations needed for the cyclic variant of the algorithm to visit the entire dataset). In the first experiment Kaczmarz and CD methods perform similarly, while Kaczmarz methods work better in the second experiment and CD methods work better in the third experiment.

6 Example: Diagonal A

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 A and $\sigma(A,2) = \lambda_{\min}$. We give the convergence rate constants for such a diagonal A in Table 1, and in this section we show how to arrive at these rates. We use U_∞ for

Table 1: Convergence Rate Constants for Diagonal A

Rule	Rate	Diagonal A
U_∞	$\left(1 - \frac{\sigma(A,2)^2}{m\|A\|_{\infty,2}^2}\right)$	$\left(1 - \frac{\lambda_{\min}^2}{m\lambda_{\max}^2}\right)$
U	$\left(1 - \frac{\sigma(\bar{A},2)^2}{m}\right)$	$\left(1 - \frac{1}{m}\right)$
NU	$\left(1 - \frac{\sigma(A,2)^2}{\|A\|_F^2}\right)$	$\left(1 - \frac{\lambda_{\min}^2}{\sum_i \lambda_i^2}\right)$
MR_∞	$\left(1 - \frac{\sigma(A,\infty)^2}{\|A\|_{\infty,2}^2}\right)$	$\left(1 - \frac{1}{\lambda_1^2}\left[\sum_i \frac{1}{\lambda_i^2}\right]^{-1}\right)$
MR	$\left(1 - \frac{\sigma(A,\infty)^2}{\|a_{i_k}\|^2}\right)$	$\left(1 - \frac{1}{\lambda_{i_k}^2}\left[\sum_i \frac{1}{\lambda_i^2}\right]^{-1}\right)$
MD	$\left(1 - \sigma(\bar{A},\infty)^2\right)$	$\left(1 - \frac{1}{m}\right)$

the slower uniform rate to differentiate from U (tight uniform) for rate (8), and we use MR_∞ for rate (5) to differentiate it from MR (tight) rate (9).

For U_∞, the rate follows straight from $\|A\|_{\infty,2} = \max_i \|a_i\| = \max_i \lambda_i = \lambda_{\max}$. For U, we note that the weighted matrix $\bar{A} := D^{-1}A$ is simply the identity matrix. The NU rate uses that $\|A\|_F^2 = \sum_i \lambda_i^2$. For both MR_∞ and MR, we have

$$\sigma(A,\infty)^2 := \inf_{y \neq z} \frac{\|A(y-z)\|_\infty^2}{\|y-z\|^2} = \inf_{\|w\|=1} \|Aw\|_\infty^2.$$

Consider the equivalent problem

$$\begin{aligned} \min_{w \in \mathbb{R}^m, y \in R} \quad & y \\ \text{s.t.} \quad & -y \leq \lambda_i^2 w_i^2 \leq y \text{ for all } i, \\ & \|w\| = 1, \end{aligned}$$

From the first inequality, we get

$$-\frac{y}{\lambda_i^2} \leq w_i^2 \leq \frac{y}{\lambda_i^2} \quad \forall i \quad \Rightarrow \quad (w_i)^2 \leq \frac{y}{\lambda_i^2} \quad \forall i.$$

It follows that

$$\|w\|^2 = \sum_{i=1}^{m} w_i^2 \leq \sum_{i=1}^{m} \frac{y}{\lambda_i^2},$$

which is equivalent to

$$y \geq \frac{\|w\|^2}{\sum_{i=1}^{m} \frac{1}{\lambda_i^2}}.$$

Because we are minimizing y this must hold with equality at a solution, and because of the constraints $\|w\| = 1$ we have

$$\sigma(A, \infty)^2 = \left(\sum_i \frac{1}{\lambda_i^2} \right)^{-1}.$$

For the MR_∞ rate, we divide $\sigma(A, \infty)^2$ by the maximum eigenvalue squared. For the MR rate, we divide by the specific $\lambda_{i_k}^2$ corresponding to the row i_k selected at iteration k.
For the MD rule, following the argument we did to derive $\sigma(A, \infty)^2$ and using that $\bar{A} = I$ gives us

$$\sigma(\bar{A}, \infty)^2 = \frac{1}{m}.$$

7.1 Multiplicative Error

Suppose we have approximated the MR selection 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, we have

$$\begin{aligned}
\|x^{k+1} - x^*\|^2 &= \|x^k - x^*\|^2 - \frac{1}{\|a_{i_k}\|^2} \left(\left| a_{i_k}^T x^k - b_{i_k} \right|^2 \right) \\
&\leq \|x^k - x^*\|^2 - \frac{1}{\|a_{i_k}\|^2} \left(\max_i \left| a_i^T x^k - b_i \right| (1 - \epsilon_k) \right)^2 \\
&= \|x^k - x^*\|^2 - \frac{(1 - \epsilon_k)^2}{\|a_{i_k}\|^2} \|A(x^k - x^*)\|_\infty^2 \\
&\leq \left(1 - \frac{(1 - \epsilon_k)^2 \sigma(A, \infty)^2}{\|a_{i_k}\|^2} \right) \|x^k - x^*\|^2.
\end{aligned}$$

We define a multiplicative approximation to the MD rule as an i_k satisfying

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

for some $\bar{\epsilon}_k \in [0, 1)$. With such a rule we have

$$\begin{aligned}
\|x^{k+1} - x^*\|^2 &= \|x^k - x^*\|^2 - \left(\left| \frac{a_{i_k}^T x^k - b_{i_k}}{\|a_{i_k}\|} \right|^2 \right) \\
&\leq \|x^k - x^*\|^2 - \left(\max_i \left| \frac{a_i^T x^k - b_i}{\|a_i\|} \right| (1 - \bar{\epsilon}_k) \right)^2 \\
&= \|x^k - x^*\|^2 - (1 - \bar{\epsilon}_k)^2 \max_i \left| \frac{a_i^T (x^k - x^*)}{\|a_i\|} \right|^2 \\
&= \|x^k - x^*\|^2 - (1 - \bar{\epsilon}_k)^2 \|D^{-1} A (x^k - x^*)\|_\infty^2 \\
&\leq \left(1 - (1 - \bar{\epsilon}_k)^2 \sigma(\bar{A}, \infty)^2 \right) \|x^k - x^*\|^2.
\end{aligned}$$

7.2 Additive Error

Suppose we select i_k using an approximate MR rule where

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

for some $\epsilon_k \geq 0$. Then we have the following convergence rate,

$$\begin{aligned}
\|x^{k+1} - x^*\|^2 &= \|x^k - x^*\|^2 - \frac{1}{\|a_{i_k}\|^2} \left| a_{i_k}^T x^k - b_{i_k} \right|^2 \\
&\leq \|x^k - x^*\|^2 - \frac{1}{\|a_{i_k}\|^2} \left(\max_i \left| a_i^T x^k - b_i \right|^2 - \epsilon_k \right) \\
&= \|x^k - x^*\|^2 - \frac{\|A(x^k - x^*)\|_\infty^2}{\|a_{i_k}\|^2} + \frac{\epsilon_k}{\|a_{i_k}\|^2} \\
&\leq \left(1 - \frac{\sigma(A, \infty)^2}{\|a_{i_k}\|^2} \right) \|x^k - x^*\|^2 + \frac{\epsilon_k}{\|a_{i_k}\|^2}.
\end{aligned}$$

For the MD rule with additive error, i_k is selected such that

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

for some $\bar{\epsilon_k} \geq 0$. Then we have

$$\begin{aligned}
\|x^{k+1} - x^*\|^2 &= \|x^k - x^*\|^2 - \left| \frac{a_{i_k}^T x^k - b_{i_k}}{\|a_{i_k}\|} \right|^2 \\
&\leq \|x^k - x^*\|^2 - \left(\max_i \left| \frac{a_i^T x^k - b_i}{\|a_i\|} \right|^2 - \bar{\epsilon_k} \right) \\
&= \|x^k - x^*\|^2 - \|D^{-1} A (x^k - x^*)\|_\infty^2 + \bar{\epsilon_k} \\
&\leq \left(1 - \sigma(\bar{A}, \infty)^2 \right) \|x^k - x^*\|^2 + \bar{\epsilon_k}.
\end{aligned}$$

7.3 Comparison of Rates for the Maximum Distance Rule and the Randomized Kaczmarz via Johnson-Lindenstrauss Method

In Eldar and Needell (2011), the authors assume that the rows of A are normalized and that we are dealing with a homogeneous system ($Ax = 0$), which is not particularly interesting since we can solve it in $O(1)$ by setting $x = 0$. Their main convergence result is stated in Theorem 1. Note that *RKJL* stands for *Randomized Kaczmarz via Johnson-Lindenstrauss*, which is a hybrid technique using both random selection and an approximate MD rule using the dimensionality reduction technique of Johnson and Lindenstrauss (1984). In their work they give the result below.

Theorem 1 *Fix an estimation x^k and denote by x^{k+1} and x_{RK}^{k+1} the next estimations using the RKJL and the standard RK method, respectively. Define $\gamma_j = |\langle a_j, x^k\rangle|^2$ and ordering these so that $\gamma_1 \geq \gamma_2 \geq \cdots \geq \gamma_m$. Then, with δ being a constant affecting the error due to the JL approximation we have*

$$\mathbb{E}\|x^{k+1} - x^*\|^2 \leq \min\left[\mathbb{E}\|x_{RK}^{k+1} - x\|^2 - \sum_{j=1}^{m}\left(p_j - \frac{1}{m}\right)\gamma_j + 2\delta, \quad \mathbb{E}\|x_{RK}^{k+1} - x^*\|^2\right],$$

where

$$p_j = \begin{cases} \frac{\binom{m-j}{n-1}}{\binom{m}{n}}, & j \leq m - n + 1 \\ 0, & j > m - n + 1 \end{cases}$$

are non-negative values satisfying $\sum_{j=1}^{m} p_j = 1$ and $p_1 \geq p_2 \geq \cdots \geq p_m = 0$.

First, we simplify this bound. Applying the nonuniform random rate of Strohmer and Vershynin (2009) to the result of Theorem 1, we get

$$\begin{aligned}
&\mathbb{E}\left[\|x^{k+1} - x\|^2\right] \\
&\leq \min\left[\mathbb{E}\left[\|x_{RK}^{k+1} - x^*\|^2\right] - \sum_{j=1}^{m}\left(p_j - \frac{1}{m}\right)\gamma_j + 2\delta, \quad \mathbb{E}\left[\|x_{RK}^{k+1} - x^*\|^2\right]\right] \\
&= \min\left[\|x^k - x^*\|^2 - \frac{1}{\|A\|_F^2}\sum_{j=1}^{m}\gamma_j - \sum_{j=1}^{m} p_j\gamma_j + \sum_{j=1}^{m}\frac{1}{m}\gamma_j + 2\delta, \quad \|x^k - x^*\|^2 - \frac{1}{\|A\|_F^2}\sum_{j=1}^{m}\gamma_j\right] \\
&= \min\left[\|x^k - x^*\|^2 - \sum_{j=1}^{m} p_j\gamma_j + 2\delta, \quad \|x^k - x^*\|^2 - \frac{1}{m}\sum_{j=1}^{m}\gamma_j\right], \qquad (15)
\end{aligned}$$

where in the last line we use $\|A\|_F^2 = m$ for a matrix A with normalized rows (in this case of normalized rows non-uniform selection is simply uniform random selection). To compare this to our rate in the setting of an additive error, suppose we define ϵ_k such that the i_k selected satisfies

$$\gamma_{i_k} \geq \max_i \gamma_i - \bar{\epsilon}_k.$$

Then, noting that $\|a_i\| = 1$ for all i, our convergence rate with additive error is based on the bound

$$\begin{aligned}\|x^{k+1} - x^*\|^2 &= \|x^k - x^*\|^2 - \gamma_{i_k} \\ &\leq \|x^k - x^*\|^2 - \max_i \gamma_i + \bar{\epsilon}_k. \end{aligned} \tag{16}$$

Comparing the bounds (15) and (16), we see that our MD bound is always faster in the case of exact optimization ($\bar{\epsilon}_k = \delta = 0$), as the average and the weighted sum of the absolute inner products squared is less than the maximum inner product squared, $\max\{\frac{1}{m}\sum_{j=1}^m \gamma_j, \sum_{j=1}^m p_j\gamma_j\} \leq \max_i \gamma_i$. If there is error present, then our rate is faster when

$$\max_i \gamma_i - \epsilon_k \geq \max\left\{\frac{1}{m}\sum_{j=1}^m \gamma_j, \sum_{j=1}^m p_j\gamma_j - 2\delta\right\}.$$

We note that even if our approximation is worse than the error resulting from the RKJL method, $\epsilon_k \geq 2\delta$, it is possible that $\max_i \gamma_i$ is significantly larger than $\frac{1}{m}\sum_{j=1}^m \gamma_j$ and $\sum_{j=1}^m p_j\gamma_j$ and in this case our rate would be tighter. Further, our rate is more general as it does not specifically assume the Johnson-Lindenstrauss dimensionality reduction technique, that the rows of A are normalized, or that the linear system is homogeneous.

8 Systems of Linear Inequalities

Consider the system of linear equalities and inequalities,

$$\begin{cases} a_i^T x \leq b_i & (i \in I_{\leq}) \\ a_i^T x = b_i & (i \in I_{=}). \end{cases} \tag{17}$$

where the disjoint index sets $I_{\leq}$ and $I_{=}$ partition the set $\{1, 2, \ldots, m\}$. As presented by Leventhal and Lewis (2010), a generalization of the Kaczmarz algorithm that accommodates linear inequalities is given by

$$\begin{aligned}\beta_{i_k}^k &= \begin{cases} (a_{i_k}^T x^k - b_{i_k})^+ & (i_k \in I_{\leq}) \\ a_{i_k}^T x^k - b_{i_k} & (i_k \in I_{=}), \end{cases} \\ x^{k+1} &= x^k - \frac{\beta_{i_k}^k}{\|a_{i_k}\|^2} a_{i_k}, \end{aligned}$$

where for $x \in \mathbb{R}^n$ we define x^+ element-wise by

$$(x^+)_i = \max\{x_i, 0\}.$$

This leads to the following generalization of the MR and MD rules, respectively,

$$i_k = \max\left|\beta_i^k\right| = \|\beta^k\|_\infty, \quad \text{and} \quad i_k = \max\left|\frac{\beta_i^k}{\|a_i\|}\right| = \|D^{-1}\beta^k\|_\infty. \tag{18}$$

Unlike for equalities where the Kaczmarz method converges to the projection of the initial iterate x^0 onto the intersection of the constraints, for inequalities we can only guarantee that the Kaczmarz

method converges to a point in the feasible set. Thus, in convergence rates involving inequalities it is standard to use a bound for the distance from the current iterate x^k to the feasible region,

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

where $P_S(x)$ is the projection of x onto S.

Following closely the arguments of Leventhal and Lewis (2010) for systems of inequalities, we next give the following result which they credit to Hoffman (1952).

Theorem 1 *Let* (17) *be a consistent system of linear equalities and inequalities, then there exists a constant* $\sigma(A, \infty)$ *such that*

$$x \in \mathbb{R}^n \text{ and } S \neq \emptyset \quad \Rightarrow \quad d(x, S) \leq \frac{1}{\sigma(A, \infty)} \|e(Ax - b)\|_\infty,$$

where S *is the set of feasible solutions and where the function* $e : \mathbb{R}^m \mapsto \mathbb{R}^m$ *is defined by*

$$e(y)_i = \begin{cases} y_i^+ & (i \in I_{\leq}) \\ y_i & (i \in I_{=}). \end{cases}$$

From Leventhal and Lewis (2010), combining both cases ($i_k \in I_{\leq}$ or $i_k \in I_{=}$), the following relationship holds with respect to the distance measure $d(x, S)$,

$$d(x^{k+1}, S)^2 \leq d(x^k, S) - \frac{e(Ax^k - b)_{i_k}^2}{\|a_{i_k}\|^2}. \tag{19}$$

Following from this bound and Theorem 1, it is straightforward to derive analogous results for all greedy selection rates derived in the paper. For example, if we select i_k according to the generalized MR rule (18) then the analogous tight rate for the MR rule is given by

$$\begin{aligned} d(x^{k+1}, S)^2 &\leq d(x^k, S)^2 - \frac{e(Ax^k - b)_{i_k}^2}{\|a_{i_k}\|^2} \\ &= d(x^k, S)^2 - \frac{\|\beta^k\|_\infty^2}{\|a_{i_k}\|^2} \\ &\leq \left(1 - \frac{\sigma(A, \infty)^2}{\|a_{i_k}\|^2}\right) d(x^k, S)^2. \end{aligned}$$

9.1 Multi-Step Maximum Residual Bound

Recall the MR rate (10),

$$\|x^k - x^*\|^2 \leq \prod_{i=1}^{k} \left(1 - \frac{\sigma(A, \infty)^2}{\|a_{i_k}\|^2}\right) \|x^0 - x^*\|^2.$$

In the worst case this is no faster than the MR_∞ rate since we may have $\|a_{i_k}\| = \|A\|_{\infty,2}$ for all i. However, this rate is faster if we have $\|a_{i_k}\| < \|A\|_{\infty,2}$ for any i. In this section we derive a tighter bound that will typically be much tighter than MR_∞ by considering the sequence of $\|a_{i_k}\|$ values that are possible for problems with a sparse orthogonality graph. To derive an upper bound, we solve the problem below which was first introduced in Nutini et al. (2015).

Problem 1. *We are given a graph $G = (V, E)$, a weight M_i associated with each node i, and an iteration number k. Choose a sequence $\{i_t\}_{t=1}^{k}$ that maximizes the sum of the weights M_{i_t} subject to the following constraint: after each time node i has been chosen, it cannot be chosen again until after a neighbour of node i has been chosen.*

To map this problem to the problem of showing that the $\|a_{i_k}\|$ values are small when we use the MR rule, we the weights $M_{i_k} = \log\left(1 - \frac{\sigma(A,\infty)^2}{\|a_{i_k}\|^2}\right)$. The constraint in Problem 1 arises because the MR rule cannot choose i_k on any future iteration until after a neighbour of it is selected in the orthogonality graph.

Nutini et al. (2015) give a bound on the solution of this problem for the case of chain-structured graphs. In order to give a bound for general graphs, we first establish some notation.

Notation

We define an *optimal sequence* for this problem as a sequence with the highest sum of weights. As the total number of sequences with length k is finite, we know that for each $k \geq 1$, Problem 1 has at least one optimal sequence.

Without loss of generality, assume that the set of nodes is given by $V = \{1, 2, \ldots, |V|\}$. We define a binary vector $\mathbf{s}^t = (s_1^t, s_2^t, \ldots, s_{|V|}^t)$ as the state of our structure at time t such that

$$s_i^t = \begin{cases} 1 & \text{node } i \text{ is selectable,} \\ 0 & \text{node } i \text{ is not selectable.} \end{cases}$$

For an arbitrary finite sequence $\chi = \{x_t\}_{t=a}^{b}$, we define the average weight of the sequence as

$$M(\chi) = \frac{\sum_{t=a}^{b} M_{x_t}}{(b-a)}.$$

We define the maximum weight over all nodes by

$$M_{\max} = \max_{v \in V} M_v.$$

Observe that for any sequence χ of nodes in V, we have

$$M(\chi) \leq M_{\max} \tag{20}$$

We denote the number of appearances of a node v in sequence χ by $\text{count}(\chi, v)$. For two sequences χ and Υ, we denote the sequence obtained by concatenating the sequence Υ to χ as $\chi\Upsilon$.

We define $\mathbb{F}_G$ as the set of all *valid finite sequences* of nodes with respect to graph G, where a sequence is *valid* if we can begin from some state $\mathbf{s}$ and the sequence satisfies the constraints.

We define $\mathbb{O}_G^k \subseteq \mathbb{F}_G$ as the set of all *optimal* sequences (maximal sum of weights) with length k that can begin from the state $\mathbf{s} = \mathbf{1}$. We denote the (maximal) average weight for each optimal sequence $\mathcal{O}^k \in \mathbb{O}_G^k$ by $M(\mathcal{O}^k) = M_{\mathbb{O}_G^k}$.

We define $\mathbb{C}_G \subseteq \mathbb{F}_G$ as the set of all valid finite sequences that are *cyclical*, meaning from some state $\mathbf{s}$, we can begin and repeat the sequence indefinitely. We'll assume that the orthogonality graph has at least one edge, meaning that we have two rows of A that are not orthogonal (if there are no edges then the problem is solved exactly in most m iterations). Under this assumption that the orthogonality graph has at least one edge, there must exist a valid cyclic sequence. Restricting to cyclical sequences, we define the average weight of any cyclic sequence achieving the maximal average weight by $M_{\mathbb{C}_G^{\max}} = \max_{\mathcal{C}\in\mathbb{C}_G} M(\mathcal{C})$.

We define G_χ as the sub-graph of G whose nodes are the set $\{v|\,\mathrm{count}\,(\chi, v) > 0\}$. The diameter of graph G is denoted by $\mathrm{diam}(G)$ and the set of all neighbours of node v in graph G is denoted by $\delta_G(v)$. For a set $U \subseteq V$, we define $G(U)$ as the sub-graph of G whose nodes make up the set U.

Let $\mathbf{U}_2$ be the collection of all subsets of nodes such that the resulting sub-graph is connected and has diameter 1 or 2,

$$\mathbf{U}_2 = \{U \subseteq V \mid G(U) \text{ is connected, } 0 < \mathrm{diam}(G(U)) \leq 2\}. \tag{21}$$

Finally, for any $T \subseteq V$ we define a binary vector $\mathbf{e}_T$ denoting the membership in T:

$$\mathbf{e}_T = (e_1, e_2, \ldots, e_{|V|}), \text{ where } e_i = \begin{cases} 1 & i \in T, \\ 0 & i \notin T. \end{cases} \tag{22}$$

Armed with this notation, we proceed to the solution of Problem 1.

Upper Bound

We breakdown the solution of Problem 1 into several important results. The first result shows that if we have a long sequence $\mathcal{A}$ that visits nodes with the same frequencies as a union of sequences, then at least one of the sequences in the union must have an average weight at least as large as the long sequence.

Lemma 2 Consider sequences of nodes $\mathcal{A}$ and $\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_m$. If we have for all nodes v that

$$\mathrm{count}(\mathcal{A}, v) = \sum_{i=1}^{m} \mathrm{count}(\mathcal{A}_i, v), \tag{23}$$

then there exists some i such that

$$M(\mathcal{A}_i) \geq M(\mathcal{A}). \tag{24}$$

Proof. Assuming (23) holds for all nodes v, we have

$$M(\mathcal{A}) \sum_{i=1}^{m} |\mathcal{A}_i| = \sum_{i=1}^{m} M(\mathcal{A}_i)|\mathcal{A}_i| \quad \Rightarrow \quad \sum_{i=1}^{m} |\mathcal{A}_i| = \sum_{i=1}^{m} \frac{M(\mathcal{A}_i)}{M(\mathcal{A})}|\mathcal{A}_i|.$$

Suppose for all i we have $M(\mathcal{A}_i) < M(\mathcal{A})$. This yields $\sum_{i=1}^{m} |\mathcal{A}_i| < \sum_{i=1}^{m} |\mathcal{A}_i|$, which is a contradiction. Thus, there exists some i such that $M(\mathcal{A}_i) \geq M(\mathcal{A})$.

□

We next characterize how the constraint bounds the number of times a node can be visited in a sequence $\mathcal{A}$, based on whether it is initially selectable and the number of times its neighbours are visited in the sequence.

Lemma 3 Suppose $\mathcal{A}$ is a valid sequence that can begin from some state $\mathbf{s}$. Then for all nodes $v \in V$, we have

$$\text{count}(\mathcal{A}, v) \leq s_v + \sum_{u \in \delta_G(v)} \text{count}(\mathcal{A}, u).$$

Proof. We can only increase the count of node v if it is in the "selectable" state. This means each time we increment the count the node was either selectable from the beginning ($s_v = 1$) or one of its neighbours ($\delta_G(v)$) was selected. □

We next give a tighter bound on the counts in the case of cyclic sequences.

Lemma 4 If $\mathcal{C}$ is a cyclic sequence then for all nodes $v \in V$ we have

$$\text{count}(\mathcal{C}, v) \leq \sum_{u \in \delta_G(v)} \text{count}(\mathcal{C}, u). \tag{25}$$

Proof. Since $\mathcal{C} \in \mathbb{C}_G$ is a cyclic sequence, by definition it can be repeated indefinitely. Consider repeating a sequence $\mathcal{C}$ twice, beginning from some state $\mathbf{s}$. We denote this new sequence by $\mathcal{C}^2$. As $\mathcal{C}^2$ is a finite sequence, i.e., $\mathcal{C}^2 \in \mathbb{F}_G$, then because of Lemma 3, for all nodes $v \in V$ we have

$$\text{count}(\mathcal{C}^2, v) \leq s_v + \sum_{u \in \delta_G(v)} \text{count}(\mathcal{C}^2, u).$$

For any v, as $s_v \in \{0, 1\}$ and $\text{count}(\mathcal{C}^2, v) = 2 \cdot \text{count}(\mathcal{C}, v)$, then

$$\begin{aligned} 2 \cdot \text{count}(\mathcal{C}, v) &\leq 1 + \sum_{u \in \delta_G(v)} 2 \cdot \text{count}(\mathcal{C}, u) \\ &< 2 + \sum_{u \in \delta_G(v)} 2 \cdot \text{count}(\mathcal{C}, u). \end{aligned}$$

Dividing by 2, we have

$$\text{count}(\mathcal{C}, v) < 1 + \sum_{u \in \delta_G(v)} \text{count}(\mathcal{C}, u),$$

which yields our result, as $\text{count}(\mathcal{C}, u)$ will always be an integer. □

Lemma 5 Let c_v be a non-negative integer associated with each node $v \in V$, and let $\mathbf{c} = (c_1, c_2, \ldots, c_{|V|})$ be the associated vector. Suppose for all $v \in V$,

$$c_v \leq \sum_{u \in \delta_g(v)} c_u. \tag{26}$$

Let $\mathbf{U}_2$ be the set defined in (21) and let $\mathbf{e}_T$ be the vector defined in (22). Then we can assign a non-negative integer a_T to each $T \in \mathbf{U}_2$ such that

$$\mathbf{c} = \sum_{T \in \mathbf{U}_2} a_T \mathbf{e}_T. \tag{27}$$

Proof. If $\mathbf{c} = \mathbf{0}$, then for all $T \in \mathbf{U}_2$, we can set $a_T = 0$ to satisfy (27). Thus, we assume $\mathbf{c} \neq 0$. First we claim that because of (26) there must be an edge $\{u, v\} \in E$ such that $c_u > 0, c_v > 0$. As $\mathbf{c} \neq \mathbf{0}$, there must be a node u such that $c_u > 0$. Now because of (26) for at least one of the neighbours of u, such as v, we must have $c_v > 0$.

Consider

$$L = \sum_{v \in V} c_v. \tag{28}$$

We use induction on L to prove the lemma.

For $L = 2$, from our above argument, there are two neighbour nodes $u, v \in V$ such that $c_u > 0$ and $c_v > 0$. By (26) and since $L = 2$, we must have $c_u = 1, c_v = 1$ with all other nodes having a c value of zero. Let $T_1 = \{u, v\}$ and $\mathbf{c} = \mathbf{e}_{T_1}$. As $T_1 \in \mathbf{U}_2$, then (27) holds by setting $a_{T_1} = 1$ and all other $a_T = 0$.

Assume the result holds for $L = 3, \ldots, k-1$. We show the lemma holds for $L = k$.

For any vector $\mathbf{c}$, we can define a *remainder* vector $\mathbf{r} = (r_1, r_2, \ldots, r_{|V|})$ such that for all nodes $v \in V$ we have

$$r_v = -c_v + \sum_{u \in \delta_G(v)} c_u.$$

We can see that (26) is satisfied if and only if for all nodes $v \in V$, we have $r_v \geq 0$.

Let $V_1 = \{v | c_v > 0\}$. We define $r_{\min} = \min\{r_v | v \in V_1\}$ and divide the problem into different cases based on the value of $r_{\min}$. In each case we find some set T_1 such that the vector $\mathbf{c}' = \mathbf{c} - \mathbf{e}_{T_1}$ satisfies the constraint (26). To do this we show that all elements of the corresponding remainder vector $\mathbf{r}'$ are non-negative, where

$$c'_v = \begin{cases} c_v & v \notin T_1, \\ c_v - 1 & v \in T_1, \end{cases}$$

and thus,

$$r'_v = \begin{cases} r_v - |T_1 \cap \delta_G(v)| & v \in V - T_1, \\ r_v - |T_1 \cap \delta_G(v)| + 1 & v \in T_1. \end{cases} \tag{29}$$

For nodes v with $c_v = 0$, it is clear that $r'_v \geq 0$ is satisfied. Thus, we consider the nodes in V_1. For nodes $v \in V_1 - T_1$ that don't have a neighbour in T_1, we have $T_1 \cap \delta_G(v) = \emptyset$, so $r'_v = r_v \geq 0$. Thus, we only need to prove $r' \geq 0$ for the nodes of T_1 and neighbours of T_1 in V_1. We divide the problem into three cases: $r_{\min} = 0, r_{\min} = 1$ and $r_{\min} \geq 2$.

Case 1 ($r_{\min} = 0$):
Consider a node $x \in V_1$ such that $r_x = 0$. As $c_x > 0$, then because of (26), x should have some neighbour in V_1, say y. Define the set $N_0^y = \{v | v \in V_1 \cap \delta_G(y), r_v = 0\}$. We choose $T_1 = \{y\} \cup N_0^y$ which is in $\mathbf{U}_2$. Note that $x \in N_0^y$, so there are nodes other than y in T_1.

Claim: For all nodes $v \in N_0^y$, we have $r'_v \geq 0$.

Proof: First we prove that there are no two neighbour nodes $u, v \in N_0^y$.
By way of contradiction, assume $u, v \in N_0^y$ are neighbours. Since $r_u = 0$ and $v \in \delta_G(u)$, we have $c_u \geq c_v$. Since $r_v = 0$ and $u \in \delta_G(v)$, we have $c_v \geq c_u$. So $c_v = c_u$, but $u, v \in \delta_G(y)$, so we have $r_v > 0, r_u > 0$, which is a contradiction. So there are no two neighbour nodes $u, v \in N_0^y$. Thus, for all nodes $v \in N_0^y$, y is their only neighbour in T_1 and we have $|T_1 \cap \delta_G(v)| = 1$. As $N_0^y \subset T_1$, based on (29), we have $r'_v = r_v = 0$.

Claim: For all nodes $v \in V_1$ that have a neighbour in N_0^y, including y, we have $r'_v \geq 0$.

Proof: Assume a node $v \in V_1$ has a neighbour $u \in N_0^y$. As $r'_u = 0$, we have $c'_u \geq c'_v$, which by (29) implies $r'_v \geq 0$.

Claim: For all neighbours v of y with $v \notin N_0^y$, we have $r'_v \geq 0$.

Proof: Note that $r_v \geq 1$ because if $r_v = 0$, then based on the definition of N_0^y, we have $v \in N_0^y$, which contradicts our assumptions that $v \notin N_0^y$. We showed in the previous claim that if v has a neighbour in N_0^y, then $r_v \geq 0$. If v is not a neighbour of any nodes in N_0^y, then $|T_1 \cap \delta_G(v)| = 1$ and because $v \in V - T_1$, then based on (29), $r'_v = r_v - 1$ and because $r_v \geq 1$, we have $r'_v \geq 0$.

Case 2 ($r_{\min} = 1$):
We divide this case into different sub-cases.

Case A: There are no two neighbour nodes $u, v \in V_1$ such that $r_v = 1, r_u = 1$.

Approach: Pick some node x such that $r_x = 1$. Then because of (26), x has some neighbour y such that $c_y > 0$. We choose $T_1 = \{x, y\}$, which is in $\mathbf{U}_2$. Note that $r'_x = r_x \geq 0, r'_y = r_y \geq 0$. For all nodes outside of T_1 that are connected to T_1 such as v, if $|T_1 \cap \delta_G(v)| = 1$, then as $v \in V - T_1$ and $r_v \geq r_{\min} = 1$, then based on (29) we have $r'_v \geq 0$. If $|T_1 \cap \delta_G(v)| = 2$, then because v is a neighbour of x, then $r_v \geq 2$; otherwise our assumption will be violated. Thus, based on (29) we have $r'_v \geq 0$.

Case B: There are two neighbour nodes $x, y \in V_1$ such that $r_x = 1, r_y = 1$.

Case (i): For all $v \in V_1 - \{x, y\}$ connected to both of x, y, we have $r_v \geq 2$.

Approach: In this case we choose $T_1 = \{x, y\}$, which is in $\mathbf{U}_2$. We have $r'_x = r_x \geq 0$, and $r'_y = r_y \geq 0$. For all nodes v connected to one of x, y, as $r_v \geq r_{\min} = 1$ and $|T_1 \cap \delta_G(v)| = 1$, by (29) we have $r'_v \geq 0$. For nodes v connected to both x, y we have $r_v \geq 2$, and as $|T_1 \cap \delta_G(v)| = 2$, by (29) we have $r'_v \geq 0$.

– *Case (ii)*: There is some node $z \in V_1 - \{x, y\}$ connected to both of x, y such that $r_z = 1$.

Approach: In this case, using $r_x = 1$ and

$$c_x = -r_x + \sum_{u \in \delta_G(x)} c_u = -1 + c_y + c_z + \sum_{u \in \delta_G(x) - \{y,z\}} c_u,$$

as $c_z > 0$, we have $c_x \geq c_y$. Using a similar argument, we have $c_x \leq c_y$, so we have $c_x = c_y$. Similarly we can prove $c_x = c_z$. Thus, we have $c_x = c_y = c_z$.

We choose $T_1 = \{x, y, z\}$, which is in $\mathbf{U}_2$. We claim that $\{x, y, z\}$ are not connected to any other node in V_1. For the sake of contradiction, assume that there is a node v connected to x. So we have

$$r_x = -c_x + c_y + c_z + \sum_{u \in \delta_G(x) - \{y,z\}} c_u = c_z + \sum_{u \in \delta_G(x) - \{y,z\}} c_u. \tag{30}$$

As v is a neighbour of x, we have

$$\sum_{u \in \delta_G(x) - \{y,z\}} c_u > 0,$$

and as $c_z > 0$, based on (30) we have $r_x > 1$, which is a contradiction. So $\{x, y, z\}$ has no neighbour in V_1 and based on (29) we have $r'_x = r'_y = r'_z = 0$ because $r_x = r_y = r_z = 1$.

Case 3 ($r_{\min} \geq 2$):
As argued before, there are two neighbour nodes $x, y \in V_1$ because of (26). We choose $T_1 = \{x, y\}$, which is in $\mathbf{U}_2$. Then we have $r'_x = r_x \geq 0$ and $r'_y = r_y \geq 0$. For all other nodes $v \in V_1 - T_1$, we have $|T_1 \cap \delta_G(v)| \leq 2$. As $r_v \geq 2$, by (29) we have $r'_v \geq 0$.

So we proved that all nodes of T_1 and neighbours of T_1 in V_1 have non-negative r' value. Thus, we have shown that the vector $\mathbf{c}'$ satisfies the condition of (26). We assumed the lemma was true for $L = 3, \ldots, k-1$. As $\sum_{u \in V} c'_u < \sum_{v \in V} c_v = L$, the lemma is true for vector $\mathbf{c}'$, so we have $\mathbf{c}' = \sum_{T \in \mathbf{U}_2} a'_T \mathbf{e}_T$. As $\mathbf{c} = \mathbf{c}' + \mathbf{e}_T$, we have our result. □

Theorem 6 Assume $E \neq \emptyset$. There exists an optimal cycle $\mathcal{C}^* \in \mathbb{C}_G^{\max}$ such that $\text{diam}(G(\mathcal{C}^*)) \leq 2$, and for all nodes $v \in V$, we have $\text{count}(\mathcal{C}^*, v) \leq 1$.

Proof. As $\mathbb{C}_G^{\max} \neq \emptyset$, there exists some $\mathcal{C} \in \mathbb{C}_G^{\max} \subseteq \mathbb{C}_G$. From Lemma 4, this implies (25). Construct a vector $\mathbf{c} = (c_1, c_2, \ldots, c_{|V|})$ such that for all nodes $v \in V$, $c_v = \text{count}(\mathcal{C}, v)$. Under this construction, $\mathbf{c}$ satisfies (26) and from Lemma 5 we have (27). Note that for all $T \in \mathbf{U}_2$ we can find a valid cyclical sequence $\mathcal{C}_T$ such that each node of T appears once. So because of (27) we can find some sequences $\mathcal{A}_1, \ldots, \mathcal{A}_m$ such that if we define $\mathcal{A} = \mathcal{C}$ we have (23), which by Lemma 2 implies there is some $\mathcal{A}_j$ such that $M(\mathcal{A}_j) \geq M(\mathcal{A})$. Note that as $\mathcal{A} \in \mathbb{C}_G^{\max}$, for all i we have $M(\mathcal{A}_i) \leq M(\mathcal{A})$, but as $M(\mathcal{A}_j) \geq M(\mathcal{A})$ for some j, we must have $M(\mathcal{A}_j) = M(\mathcal{A})$. Note that as $T \in \mathbf{U}_2$, we have $\text{diam}(G(T)) \leq 2$ and thus, for all nodes $v \in V$, we have $\text{count}(\mathcal{A}_j, v) \leq 1$. So $\mathcal{A}_j \in \mathbb{C}_G^{\max}$ and the result holds for $\mathcal{C}^* = \mathcal{A}_j$. □

Based on Theorem 6, to find $M_{\mathbb{C}_G^{\max}}$ we search over all sub-graphs of graph G with diameter less than or equal to 2, and pick the one with the highest average weight. This can be done in $O(|E| + |V| \log |V|)$ time.

Theorem 7

$$\lim_{k \to \infty} M_{\mathbb{O}_G^k} = M_{\mathbb{C}_G^{\max}}. \tag{31}$$

Proof. Let $\mathcal{O}_0^k = \{i_t\}_{t=1}^k \in \mathbb{O}_G^k$ and $\{\mathbf{s}^t\}_{t=1}^k$ be the corresponding sequence of states, where $\mathbf{s}^1 = \mathbf{1}$. If $|\mathcal{O}_0^k| > 2^{|V|}$, then by the pigeon hole principle, there must be t_1 and t_2 such that $\mathbf{s}^{t_1} = \mathbf{s}^{t_2}$. Let

$\mathcal{A}_0^k = \{i_t\}_{t=1}^{t_1-1}$, $\mathcal{B}_0^k = \{i_t\}_{t=t_1}^{t_2-1}$ and $\mathcal{C}_0^k = \{i_t\}_{t=t_2}^{k}$, so that $\mathcal{O}_0^k = \mathcal{A}_0^k\mathcal{B}_0^k\mathcal{C}_0^k$. Now because $\mathbf{s}^{t_1} = \mathbf{s}^{t_2}$, $\mathcal{O}_1^k = \mathcal{A}_0^k\mathcal{C}_0^k$ is a valid sequence. Note that $\mathcal{B}_0^k = \mathbb{C}_G$, so $M(\mathcal{B}_0^k) \leq M_{\mathbb{C}_G^{\max}}$. If $|\mathcal{O}_1^k| > 2^{|V|}$, we repeat the process and obtain a new sequence $\mathcal{O}_2^k$. As long as $|\mathcal{O}_j^k| > 2^{|V|}$, we repeat this process until we obtain a sequence $\mathcal{O}_m^k$ such that

$$|\mathcal{O}_m^k| \leq 2^{|V|}. \tag{32}$$

We denote the omitted sub-sequence from $\mathcal{O}_j^k$ in step j as $\mathcal{B}_j^k$. As we argued,

$$M(\mathcal{B}_j^k) \leq M_{\mathbb{C}_G^{\max}}. \tag{33}$$

We have

$$M(\mathcal{O}_0^k) = \frac{1}{k}\left(|\mathcal{O}_m^k|M(\mathcal{O}_m^k) + \sum_{j=0}^{m-1}|\mathcal{B}_j^k|M(\mathcal{B}_j^k)\right). \tag{34}$$

Combining (32), (33) and (34) with equation (20), we have

$$M(\mathcal{O}_0^k) \leq \frac{1}{k}\left(2^{|V|}M_{\max} + kM_{\mathbb{C}_G^{\max}}\right). \tag{35}$$

Let $\mathcal{C}^*$ be a sequence satisfying the conditions of Theorem 6. We construct the new sequence $\mathcal{C}_\downarrow^*$ by sorting the elements of $\mathcal{C}^*$ by their weights in descending order. Because $\operatorname{diam}(G_{\mathcal{C}^*}) \leq 2$, $\mathcal{C}_\downarrow^*$ is also a valid cycle. Now we construct the sequence $\mathcal{Z}$ by repeating $\mathcal{C}_\downarrow^*$ until we obtain a sequence with length k. Note that in the last repeat of $\mathcal{C}_\downarrow^*$, all of it's elements may not be inserted. So

$$M_{\mathbb{C}_G^{\max}} \leq M(\mathcal{Z}). \tag{36}$$

And because $\mathcal{O}_0^k \in \mathbb{O}_G^k$ and $|\mathcal{Z}| = k$, we have

$$M(\mathcal{Z}) \leq M(\mathcal{O}_0^k). \tag{37}$$

Combining (35), (36) and (37), we get

$$M_{\mathbb{C}_G^{\max}} \leq M(\mathcal{O}_0^k) \leq \frac{1}{k}\left(2^{|V|}M_{\max} + kM_{\mathbb{C}_G^{\max}}\right).$$

Because

$$\lim_{k\to\infty} M_{\mathbb{C}_G^{\max}} = \lim_{k\to\infty}\frac{1}{k}\left(2^{|V|}M_{\max} + kM_{\mathbb{C}_G^{\max}}\right) = M_{\mathbb{C}_G^{\max}},$$

by the sandwich theorem we have our result. □

We can rewrite Theorem 7 as

$$M_{\mathbb{O}_G^k} = O(M_{\mathbb{C}_G^{\max}}). \tag{38}$$

Because we are interested in long sequences, and based on the structure of the graph G, the sequence $\{M_{\mathbb{O}_G^k}\}_{k=1}^{\infty}$ can have many oscillations, which makes finding the exact values of $M_{\mathbb{O}_G^k}$ hard. Thus, we acquiesce in asymptotic analysis of (31) and (38).

By mapping back to the original problem we obtain the rate stated in the main paper.

9.2 Faster Randomized Kaczmarz Using the Orthogonality Graph of A

In order for the adaptive methods to be efficient, we must be able to efficiently update the set of selectable nodes at each iteration. To do this we use a tree structure that keeps track of the number of selectable children in the tree (for uniform random selection) or the cumulative sums of the selectable row norms of A (for non-uniform random selection). A similar structure is used in the non-uniform sampling code of Schmidt et al. (2013).

Recall that the standard inverse-transform approach approach to sampling from a non-uniform discrete probability distribution over m variables:

1. Compute the cumulative probabilities, $c_i = \sum_{j=1}^{i} p_j$ for each i from 1 to m.
2. Generate a random number u uniformly distributed over $[0, 1]$.
3. Return the smallest i such that $c_i \geq u$.

We can compute all m values of c_i in Step 1 at a cost of $O(m)$ by maintaining the running sum. We'll assume that Step 2 costs $O(1)$ and we can implement Step 3 in $O(\log(m))$ using a binary search. If we are sampling from a fixed distribution, then we only need to perform Step 1 once and from that point we can generate samples from the distribution at a cost of $O(\log(m))$.

In the adaptive randomized selection rules, the probabilities p_j change at each iteration and hence the c_i values also change. This means we can't skip Step 1 as we can for fixed probabilities. However, if the orthogonality graph is sparse then it's still possible to efficiently implement these strategies. To do this, we consider a binary tree-structure that has the probabilities p_j as leaf nodes while each internal node is the *sum* of its two descendants (and thus the root node has a value of 1). Given this structure, we can find the smallest $c_i \geq u$ in $O(\log(m))$ by traversing the tree. Further, if we update one of the p_j values then we can update this data structure in $O(\log(m))$ time since this only requires changing one node at each depth of the tree. If each node has at most g neighbours in the orthogonality graph, then we need to update g probabilities in the binary tree, leading to a cost of $O(g \log(m))$ to update the tree structure on each iteration.

Note that the above structure can be modified to work with *unnormalized probabilities* at the leaf nodes, since the root node will contain the normalizing constant required to make these unnormalized probabilities into a valid probability mass function. Using this, we can implement the adaptive uniform method by setting the leaf nodes to 1 for selectable nodes and 0 for non-selectable nodes. To implement the adaptive non-uniform method, we set the leaf nodes to 0 for non-selectable nodes and $\|a_i\|^2$ for selectable nodes.

10 Experiments

Formulating the Semi-Supervised Label Propagation Problem as a Linear System

Our third experiment solves a label propagation problem for semi-supervised learning in the 'two moons' dataset (Zhou et al., 2004). We use a variant of the quadratic labelling criterion of Bengio

et al. (2006),

$$\min_{y_i \in S'} \quad f(y) \equiv \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(y_i - y_j)^2,$$

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, S' is the set of labels that we do not know and $w_{ij} \geq 0$ are the weights assigned to each y_i describing how strongly we want the labels y_i and y_j to be similar. We assume without loss of generality that $w_{ij} = w_{ji}$ for all i, j because the model only depends on these terms through $(w_{ij} + w_{ji})$. We can express this quadratic problem as a linear system that is consistent by construction. In other words, we can define A and b such that

$$\nabla f(y) = 0 \iff Ay = b, \quad \text{with } y \in S'.$$

Differentiating f with respect to some $y_k \in S'$, we have

$$\begin{aligned}
\nabla_k f(y) &= \underbrace{\sum_{j \neq k} w_{kj}(y_k - y_j)}_{i=k,\ j \neq k} - \underbrace{\sum_{i \neq k} w_{ik}(y_i - y_k)}_{i \neq k,\ j=k} + \underbrace{\sum_{i=k} w_{kk}(y_k - y_k)}_{i=k,\ j=k} \\
&= \sum_{i=1}^{n} w_{ki}(y_k - y_i) - \sum_{i=1}^{n} w_{ik}(y_i - y_k) \\
&= 2\sum_{i=1}^{n} w_{ki} y_k - 2\sum_{i=1}^{n} w_{ki} y_i.
\end{aligned}$$

Setting this equal to zero and splitting the summation over S and S' separately, we have

$$\sum_{i=1}^{n} w_{ki} y_k - \sum_{i \in S'} w_{ki} y_i = \sum_{i \in S} w_{ki} y_i.$$

Assuming the elements of S' form the first $|S'|$ elements of the matrix A, the above formulation yields the $|S'| \times |S'|$ matrix

$$A(k, i) = \begin{cases} \sum_{j=1}^{n} w_{kj} & \text{if } i = k, \\ -w_{ki} & \text{if } i \neq k, \end{cases}$$

where k and $i \in S'$ and

$$b(k) = \sum_{i \in S} w_{ki} y_i.$$

Time vs. Squared Error and Distance

Figure 2 compares the runtime results for our 3 experiments from the main paper using both squared error and distance (we made a reasonable effort to make the implementations of all methods as efficient as possible). We see that in the first experiment the greedy selection rules do not translate into gains in terms of runtime due to their higher iteration cost, while in the second and third experiments the greedy rules are still superior in terms of runtime.

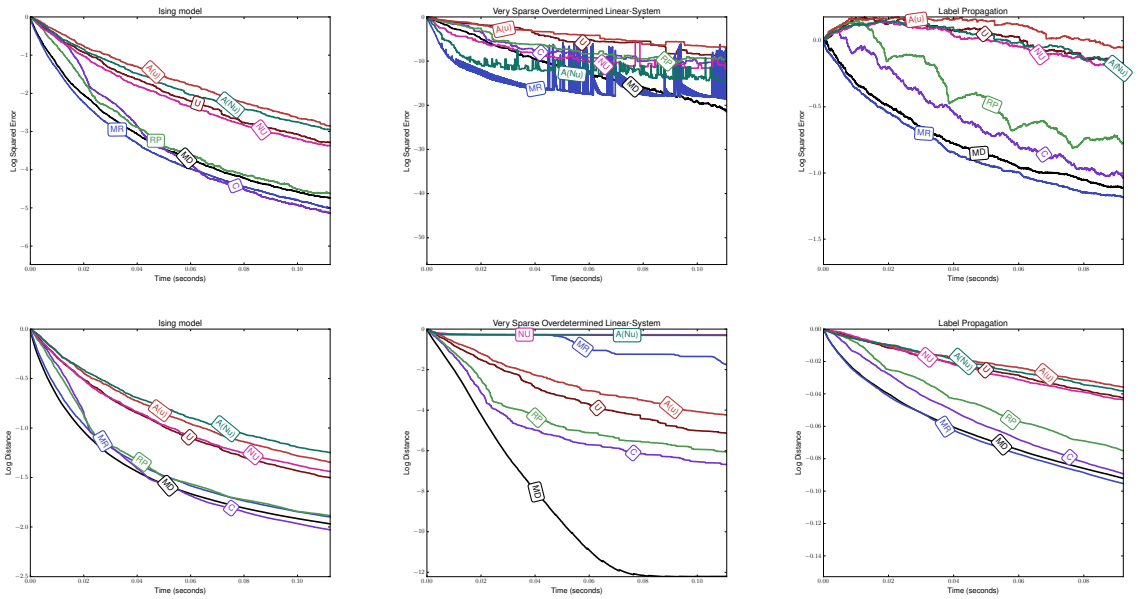

Figure 2: Runtime Comparisons of Kaczmarz Selection Rule.

Hybrid Methods

For the very sparse overdetermined dataset, we see very different performances between the MR and MD rules with respect to squared error and distance. We see that the MR rule outperforms the MD rule in the beginning with respect to squared-error and the MD rule outperforms the MR rule significantly with respect to distance. These observations align with the respective definitions of each greedy rule. However, if we want a method that converges well with respect to *both* of these objectives, then we could consider 'hybrid' greedy rule. For example, we could simply alternate between using the MR rule and the MD rule. As we see in Figure 3, this approach simultaneously exploits the convergence of the MR rule in terms of squared error and the MD rule in terms of distance to the solution. However, computationally this approach requires the maintenance of two max-heap structures.

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

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.

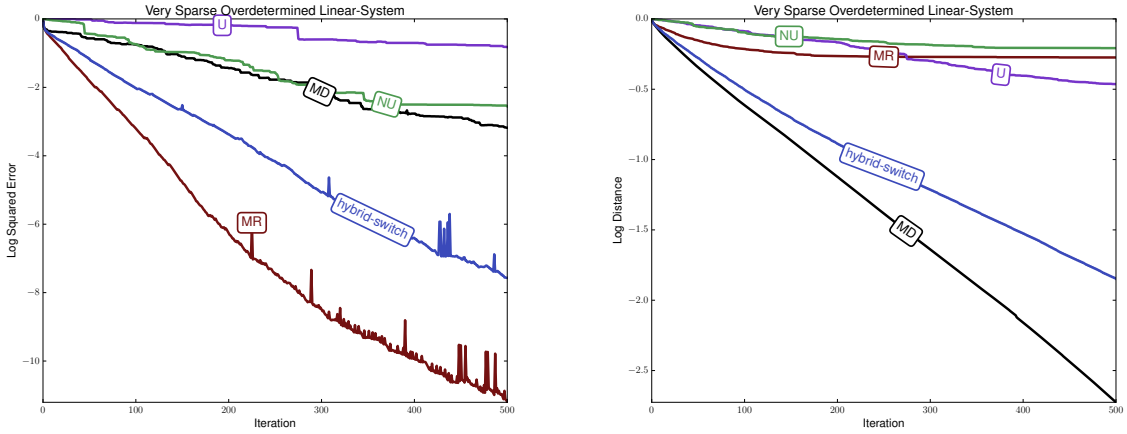

Figure 3: Comparison of MR, MD and Hybrid Method for Very Sparse Dataset.

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

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.

M. Schmidt, N. Le Roux, and F. Bach. Minimizing finite sums with the stochastic average gradient. *arXiv preprint*, 2013.

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

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

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.