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

Julie Nutini (UBC), Behrooz Sepehry (UBC), Issam Laradji (UBC), Mark Schmidt (UBC), Hoyt Koepke (Dato) and Alim Virani (UBC)

Overview: Greedy Selection for Kaczmarz Methods

- We consider solving linear systems with Kaczmarz methods.
- Strohmer & Vershynin [2009] show linear convergence with randomized row selection.
- Does it make sense to use greedy row selection?

- **Our contributions**:
  - ⋆ Efficient implementation of greedy rules for sparse A.
  - ⋆ Faster convergence rates for greedy selection rules.
  - ⋆ Analysis of approximate greedy selection rules.
  - ⋆ First multi-step analysis for Kaczmarz methods.
  - ⋆ Faster randomized selection rule with orthogonality.

Problems of Interest

We consider a consistent system of linear equalities/inequalities,

$$Ax = b \quad \text{and/or} \quad Ax \leq b,$$

where
- $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, and a solution x^* exists.

Applications in ML that involve solving linear systems:
1. Least squares:
$$\min_x \frac{1}{2}\|Ax - b\|^2 \iff \begin{pmatrix} A & -\mathbb{I} \\ \mathbf{0} & A^T \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} b \\ 0 \end{pmatrix}.$$
2. Least-squares support vector machines.
3. Gaussian processes.
4. Fitting final layer of neural network (squared-errors).
5. Graph-based semi-supervised learning.
6. Decoding of Gaussian Markov random fields.

The Kaczmarz Method

On each iteration of the Kaczmarz method:
- Choose row i_k and project x^k onto hyperplane $a_{i_k}^T x^k = b_{i_k}$,

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

- ⋆ Convergence under weak conditions.
- Usual rules are cyclic or random selection of i_k.

Greedy Selection Rules

- The maximum residual (MR) rule selects i_k according to
$$i_k = \operatorname{argmax}_i |a_i^T x^k - b_i|.$$
- ⋆ The equation i_k that is 'furthest' from being satisfied.
- The maximum distance (MD) rule selects i_k according to
$$i_k = \operatorname{argmax}_i \left|\frac{a_i^T x^k - b_i}{\|a_i\|}\right|.$$
- ⋆ Maximizing distance that iteration moves, $\|x^{k+1} - x^k\|$.

Kaczmarz vs. Coordinate Descent

Key differences between Kaczmarz and coordinate descent:

	Kaczmarz	CD
Problem	linear system	least-squares
Selects	rows of A	columns of A
Assumes	consistent system	linearly independent columns
Convergence	$\|x^k - x^*\|$	$f(x^k) - f(x^*)$

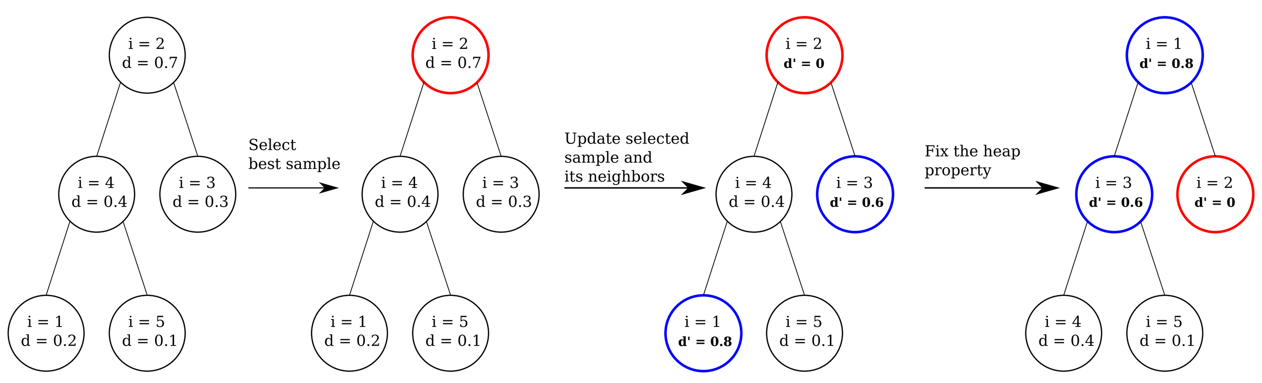
The Orthogonality Graph

Orthogonality graph G of the matrix A:
- Each row i is a node.
- Edge between nodes i and j if a_i is not orthogonal to a_j.

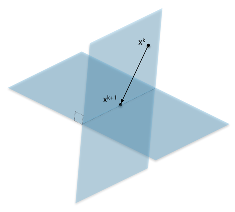
→ After selection i_k, equality i_k will be satisfied for all subsequent iterations until a neighbour in the orthogonality graph is selected.

Efficient Implementation of Greedy Rules

- If A has at most c non-zeros per column and r non-zeros per row:
  - Can compute greedy rules in $O(cr \log m)$ using max-heap.



- Use the orthogonality graph of A to track which rows to update:
  - For selected i, only update node i and neighbours of node i.
    → Projecting onto hyperplane does not affect sub-optimality of non-neighbours.
  - Costs $O(gn + g\log(m))$, where g is maximum number of neighbours of any node.
    → If g is small, comparable to $O(n + \log(m))$ of randomized strategies.



- Use an efficient approximation of the greedy rules:
  → e.g., Johnson-Lindenstrauss dimensionality reduction [Eldar & Needell, 2011].

Convergence Rates for Different Selection Rules

We use the following relationship between $\|x^{k+1} - x^*\|$ and $\|x^k - x^*\|$:

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

By the definition of the Kaczmarz update, we obtain for any selected i_k,

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

From (1), we can derive the following rates:
- For uniform random selection, we can show
$$\mathbb{E}\left[\|x^{k+1} - x^*\|^2\right] \leq \left(1 - \frac{\sigma(A,2)^2}{m\|A\|^2_{\infty,2}}\right)\|x^k - x^*\|^2, \qquad (\text{Uniform}_\infty)$$
where $\|A\|^2_{\infty,2} := \max_i\{\|a_i\|^2\}$ and $\sigma(A,2)$ is the Hoffman constant.
- Using $\bar{A} = D^{-1}A$, where $D = \operatorname{diag}(\|a_1\|, \ldots, \|a_m\|)$ gives tighter bound,
$$\mathbb{E}\left[\|x^{k+1} - x^*\|^2\right] \leq \left(1 - \frac{\sigma(\bar{A},2)^2}{m}\right)\|x^k - x^*\|^2. \qquad (\text{Uniform})$$
- Strohmer & Vershynin show that non-uniform selection with probability $\|a_i\|^2/\|A\|_F^2$ gives
$$\mathbb{E}\left[\|x^{k+1} - x^*\|^2\right] \leq \left(1 - \frac{\sigma(A,2)^2}{\|A\|_F^2}\right)\|x^k - x^*\|^2. \qquad (\text{Non-Uniform})$$
  - ⋆ Faster than Uniform$_\infty$ but not necessarily faster than Uniform.
- For the maximum residual selection rule we get
$$\|x^{k+1} - x^*\|^2 \leq \left(1 - \frac{\sigma(A,\infty)^2}{\|A\|^2_{\infty,2}}\right)\|x^k - x^*\|^2, \qquad (\text{Max Res}_\infty)$$
where
$$\frac{\sigma(A,2)}{\sqrt{m}} \leq \sigma(A,\infty) \leq \sigma(A,2).$$
  - ⋆ The MR rule is at least as fast as Uniform$_\infty$, could be up to m times faster.
- Using row norm $\|a_{i_k}\|$ gives tighter bound,
$$\|x^{k+1} - x^*\|^2 \leq \left(1 - \frac{\sigma(A,\infty)^2}{\|a_{i_k}\|^2}\right)\|x^k - x^*\|^2. \qquad (\text{Max Res})$$
  - ⋆ Faster when $\|a_{i_k}\| < \|A\|_{\infty,2}$, gives tighter rate with multi-step analysis.
- For the maximum distance rule, we can show a rate of
$$\|x^{k+1} - x^*\|^2 \leq \left(1 - \sigma(\bar{A},\infty)^2\right)\|x^k - x^*\|^2, \qquad (\text{Max Dist})$$
where
$$\max\left\{\frac{\sigma(\bar{A},2)}{\sqrt{m}}, \frac{\sigma(A,2)}{\|A\|_F}, \frac{\sigma(A,\infty)}{\|A\|_{\infty,2}}\right\} \leq \sigma(\bar{A},\infty) \leq \sigma(\bar{A},2).$$
  - ⋆ Faster than all other rules in terms of $\|x^{k+1} - x^*\|$.

Relationships Among Rules

	Uniform$_\infty$	Uniform	Non-Uniform	Max Res$_\infty$	Max Res	Max Dist
Uniform$_\infty$	$=$	$\leq$	$\leq$	$\leq$	$\leq$	$\leq$
Uniform		$=$	P	P	P	$\leq$
Non-Uniform			$=$	P	P	$\leq$
Max Res$_\infty$				$=$	$\leq$	$\leq$
Max Res					$=$	$\leq$
Max Dist						$=$

→ P: depends on problem.

Example: Diagonal A

For diagonal A, we can get explicit forms of constants.

Consider the case when all eigenvalues are equal except for one:

$$\lambda_1 = \lambda_2 = \cdots = \lambda_{m-1} > \lambda_m > 0.$$

Letting $\alpha = \lambda_i^2(A)$ for any $i = 1, \ldots, m-1$ and $\beta = \lambda_m^2(A)$, we have

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

⋆ Strohmer & Vershynin's NU is worst rule, greedy/uniform much faster.

Approximate Greedy Rules

- For multiplicative error in the MD rule,
$$\left|\frac{a_{i_k}^T x^k - b_{i_k}}{\|a_{i_k}\|}\right| \geq \max_i \left|\frac{a_i^T x^k - b_i}{\|a_i\|}\right|(1 - \bar{\epsilon}_k),$$
we show for some $\bar{\epsilon}_k \in [0,1)$,
$$\|x^{k+1} - x^*\|^2 \leq \left(1 - (1 - \bar{\epsilon}_k)^2\sigma(\bar{A},\infty)^2\right)\|x^k - x^*\|^2,$$
which does not require $\bar{\epsilon}_k \to 0$.
- For additive error in the MD rule,
$$\left|\frac{a_{i_k}^T x^k - b_{i_k}}{\|a_{i_k}\|}\right|^2 \geq \max_i \left|\frac{a_i^T x^k - b_i}{\|a_i\|}\right|^2 - \bar{\epsilon}_k,$$
we show for some $\bar{\epsilon}_k \geq 0$,
$$\|x^{k+1} - x^*\|^2 \leq (1 - \sigma(\bar{A},\infty)^2)\|x^k - x^*\|^2 + \bar{\epsilon}_k,$$
which requires $\bar{\epsilon}_k \to 0$ (avoid with hybrid of Eldar & Needell).
- ⋆ If $\bar{\epsilon}_k \to 0$ fast enough, we obtain the same rate of exact case.

Adaptive Randomized Rules

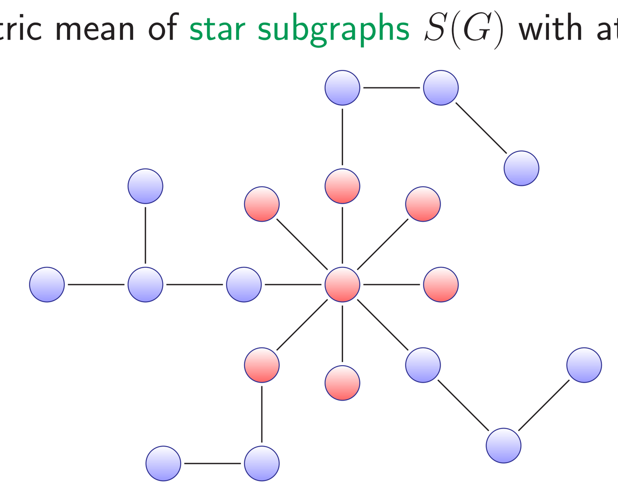
Define a sub-matrix A_k of *selectable* rows using orthogonality graph of A.
- For adaptive non-uniform, we obtain the bound
$$\mathbb{E}\left[\|x^{k+1} - x^*\|^2\right] \leq \left(1 - \frac{\sigma(A_k,2)^2}{\|A_k\|_F^2}\right)\|x^k - x^*\|^2.$$
- This bound is much tighter if you have one large $\|a_i\|$ and no neighbours have been selected since the last time row i was selected.
- ⋆ A similar bound is obtained for adaptive uniform selection.

Multi-Step Maximum Residual Bound

Using the orthogonality graph G of the matrix A, we obtain a tighter bound on the MR rule using sequence of $\|a_i\|$ values,

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

based on geometric mean of star subgraphs $S(G)$ with at least two nodes.



→ Much faster rate if large $\|a_i\|$ are more than 2 edges apart.

Experiments