Multi-Randomized Kaczmarz for Latent Class Regression

Erin George∗§, Yotam Yaniv∗§, Deanna Needell∗
∗Department of Mathematics, University of California, Los Angeles

Abstract—Linear regression is effective at identifying interpretable trends in a data set, but averages out potentially different effects on subgroups within data. We propose an iterative algorithm based on the randomized Kaczmarz (RK) method to automatically identify subgroups in data and perform linear regression on these groups simultaneously. We prove almost sure convergence for this method, as well as linear convergence in expectation under certain conditions. The result is an interpretable collection of different weight vectors for the regressor variables that capture the different trends within data. Furthermore, we experimentally validate our convergence results by demonstrating the method can successfully identify two trends within simulated data.

I. INTRODUCTION

Often, one needs to perform regression tasks on extremely large-scale data. Methods such as the randomized Kaczmarz method (RK) [1], [2] have gained recent attention for their ability to solve such systems with needing to only access a single row at a time rather than the full system in memory. However, in many settings, two or more population subgroups may be present in the data requiring multiple regressors. Often times, computing a single regressor will result in a minority group having far worse predictive power than the majority. Additionally, the minority group is not known a priori requiring that we both discover and regress on these subgroups on the fly. Here, we present a variant of RK that addresses this problem via multiple regressors.

Formally, given multiple consistent systems of equations

\[ M^{(i)} x^{(i)}_s = b^{(i)}, \ i \in \{0, 1, \ldots, n\} \]

we consider the combined matrices

\[ M = \begin{bmatrix} M^{(0)} \\ M^{(1)} \\ \vdots \\ M^{(n)} \end{bmatrix}, \quad b = \begin{bmatrix} b^{(0)} \\ b^{(1)} \\ \vdots \\ b^{(n)} \end{bmatrix} \]

with the goal of recovering \( x^{(0)}_s, x^{(1)}_s, \ldots, x^{(n)}_s \) where the rows of these matrices may be shuffled. Next we define the class of a set of rows and right hand side entries.

Definition 1 (Class). Given a regressor \( x^{(i)}_s \) a set of rows resulting in matrix \( M^{(j)} \) and right hand sides \( b^{(j)} \) are in class \( i \) if \( M^{(j)} x^{(i)}_s = b^{(j)} \).

We assume that the class of each row is not known beforehand. This task corresponds with uncovering multiple systems and their solutions. In the statistics literature, this problem can be framed as latent class linear regression where each class represents an overdetermined system of equations [3], [4]. Classically, this problem can be solved by using an expectation-maximization (EM) algorithm to iteratively fit the regressor coefficients and then classify the rows [5], [6]. The EM algorithm has been extensively studied in a statistics framework with convergence properties discussed in [7] and [8]. More recently the EM framework has been used to learn class data representations in unsupervised machine learning using neural networks [9].

We take a randomized numerical linear algebra approach to this problem by modifying the classical randomized Kaczmarz algorithm to this setting. This approach allows us to process very large data sets while only accessing single rows of our data set at a time.

II. MULTI-RANDOMIZED KACZMARZ METHOD

We propose a novel iterative method motivated by the randomized Kaczmarz (RK) algorithm for simultaneously solving all \( n + 1 \) systems, Algorithm 1. This approach is motivated by the assumption that the closer an iterate is to a hyperplane defined by a row of the combined system, the more likely that row belongs to the class of that iterate. Since Kaczmarz methods converge monotonically this is a reasonable assumption.

At each iteration, the multi-randomized Kaczmarz (MRK) method selects a hyperplane as in the standard RK algorithm. Then the Kaczmarz update for all iterates is computed. The update with the smallest magnitude is selected, denoted \( s_k \), with the respective magnitude denoted as \( c_{s_k} \). Given a swap probability \( r \) we then update iterate \( t_k \) chosen to be \( s_k \) with probability \( 1 - r \) and \( t_k \) chosen from all iterates uniformly at random with total probability \( r \). The selected iterate \( t_k \) is updated by the magnitude \( c_{s_k} \) in the direction the \( t_k \)-th iterate would have been updated given the standard Kaczmarz update.

We state two convergence results for this method, which we will prove in the following section. The first theorem, Theorem 1, proves a linear convergence result for the MRK algorithm in expectation under certain conditions. The second theorem, Theorem 2, is an almost sure convergence result for the MRK algorithm. Other almost sure convergence results have been shown for Kaczmarz type algorithms [10] under the assumption that measurements (rows of the matrix) are drawn from independent but not necessarily identical distributions.

To prove these theorems, we will make a uniqueness assumption on the problem.

978-1-6654-5906-8/22/$31.00 ©2022 IEEE 1367
Asilomar 2022
Algorithm 1 Multi-Randomized Kaczmarz (MRK) Algorithm

Input: System $M$, right hand side $b$, number of iterations $N$, initial iterates $x^{(0)}_0, x^{(1)}_0, \ldots, x^{(n)}_0$, swap probability $r$, sampling distribution $D$.

for $k$ from 0 to $N - 1$ do

Sample row $i_k \sim D$

$c_{i,k} = \frac{M_{i,k}x^{(i)}_k - b_k}{||M_{i,k}||^2}$, $i = 0, 1, \ldots, n$

$s_k = \text{argmin}_{i \in \{0, 1, \ldots, n\}} (c_{i,k})$

$t_k = \begin{cases} s_k & \text{with probability } 1 - r \\ t & \text{with probability } \frac{r}{n + r} \text{ for all } t \in \{0, 1, \ldots, n\} \end{cases}$

The total probability that $t_k = s_k$ is $1 - r + \frac{r}{n + r}$.

$x^{(t_k)}_{k+1} = x^{(t_k)}_k - |c_{s_k}| \text{sgn}(c_{s_k})M_{s_k}^T$

$x^{(j)}_{k+1} = x^{(j)}_k$, $j \neq t_k$

end for

Assumption 1. The solution to the set of systems is unique up to relabeling. That is, suppose there are $x_i$, $i \in \{0, 1, \ldots, n\}$ so that for each row in the combined system (indexed by $k$) there is $i_k$ where

$$M_kx_{i_k} = b_k.$$

Then there is a permutation $\sigma$ on $\{0, 1, \ldots, n\}$ so that $x^{(i)}_* = x^{(\sigma(i))}_*$ for all $i$.

In particular, this means all systems in the problem are full rank, even if rows which consistently belong to two or more classes are removed.

Theorem 1 (Conditional expected MRK Convergence). Define

$$e_k = \sum_{i=0}^n \|x^{(i)}_k - x^{(i)}_*\|^2.$$

Let $r \geq 0$ be sufficiently small. Choose $c \in (C_0, 1)$ and $\delta > 0$. There exists $\varepsilon > 0$ so that if $e_k < \varepsilon$ then

$$\mathbb{E}(e_{k+1} | A_k) \leq c^b e_k$$

where $A_k$ is an event that happens with probability at least $1 - \delta$. The constant $C_0 < 1$ depends on $M$, $b$, $n$, and $r$.

This theorem shows the convergence will be linear in expected squared error after a certain point. Limiting the initial squared error before convergence allows us to identify which solution each iterate is converging towards. The failure probability reflects the possibility that the iterates may still converge towards a different labeling of the solutions and iterates. In the case where the initial squared error is too large or the failure probability is triggered, we will still see convergence, as shown in the next theorem.

Theorem 2 (Almost sure MRK Convergence). There is $r' \in (0, 1)$ so that if $r \in (0, r')$, each iterate of the algorithm converges almost surely to a different solution of the subsystems.

The convergence rate given by the proof of this theorem is slow. In practice, we find the convergence rate quickly achieves the linear rate given in the previous theorem.

III. Proofs

A. Conditional Convergence in Expectation

Proof of Theorem 1. At the $k$-th iteration, we select a row from a system. Suppose we select a row $\ell_k$ from system $i$. There are three possibilities for how we update.

(a) We update $x^{(i)}_k$ fully, setting

$$\left\|x^{(i)}_{k+1} - x^{(i)}_*\right\| = C_k \left\|x^{(i)}_k - x^{(i)}_*\right\|^2$$

for some random variable $C_k$ taking value in the range $[0, 1]$. The expectation of $C_k$ is just the Kaczmarz constant for the subset of rows which we are allowed to make a full and correct update with.

(b) We update $x^{(i)}_k$ partially. We bound the error here as

$$\left\|x^{(i)}_{k+1} - x^{(i)}_*\right\| \leq \left\|x^{(i)}_k - x^{(i)}_*\right\|.$$

(c) We update $x^{(j)}_k$ for some $j \neq i$. Regardless of how this happens, we always update by a magnitude bounded above in norm by the correct update:

$$\left\|M_k x^{(i)}_k - b_k \right\| \leq \left\|x^{(i)}_k - x^{(i)}_*\right\|.$$

Therefore the new error satisfies

$$\left\|x^{(j)}_{k+1} - x^{(j)}_*\right\| \leq 2\left\|x^{(j)}_k - x^{(j)}_*\right\|^2 + 2\left\|x^{(i)}_k - x^{(j)}_k\right\|^2.$$

There are two ways for us to land in case (c). Either we trigger our swap probability and select iterate $j$, or we do not trigger our swap probability but we selected iterate $j$ anyway. The second happens only when

$$\frac{\left\|M_k x^{(j)}_k - b_k \right\|}{\left\|M_k \right\|} \leq \frac{\left\|M_k x^{(i)}_k - b_k \right\|}{\left\|M_k \right\|}$$

We can bound the left side below by

$$\frac{\left\|M_k \right\|}{\left\|M_k \right\|} - \left\|x^{(i)}_k - x^{(j)}_k\right\|$$

and the right side above by

$$\left\|x^{(j)}_k - x^{(j)}_*\right\|.$$

So this can only happen when

$$\frac{\left\|M_k \right\|}{\left\|M_k \right\|} \leq \left\|x^{(i)}_k - x^{(j)}_k\right\| + \left\|x^{(j)}_k - x^{(j)}_*\right\|$$

$$\leq \sum_{a=0}^n \left\|x^{(a)}_k - x^{(a)}\right\|$$

$$\leq \sqrt{(n + 1) \cdot e_k}.$$

We only need to consider the case when $M_k (x^{(i)}_k - x^{(j)}_k) \neq 0$, as otherwise we could consider this row $\ell_k$ as coming from
the $j$-th system anyway. Therefore, the probability of this happening goes to 0 as $e_k$ goes to 0. Suppose $\varepsilon$ is small enough so that whenever $e_k < \varepsilon$ the probability this mistake happens for any pair is less than $\eta$.

We will also assume that $\varepsilon$ is small enough so that, assuming we do not trigger our swap probability, there is a full rank set of rows for each system so that whenever $e_k < \varepsilon$ all these rows will make a correct update and that cannot be added to another system consistently. This is a consequence of Assumption 1. Then, for each system, the condition number for the set of rows that can make a correct update is bounded above, and the RK constant is bounded above by some value strictly less than 1. Let $c < 1$ bound above the RK constant for each system.

Now, whenever $e_k < \varepsilon$, we can bound
\[
\left(\begin{array}{c}
\|x_{k+1}^{(0)} - x_*^{(0)}\|^2 \\
\|x_{k+1}^{(1)} - x_*^{(1)}\|^2 \\
\|\ldots\| \\
\|x_{k+1}^{(n)} - x_*^{(n)}\|^2 \\
n\|x_{k+1}^{(0)} - x_*^{(0)}\|^2
\end{array}\right) \leq A \left(\begin{array}{c}
\|x_k^{(0)} - x_*^{(0)}\|^2 \\
\|x_k^{(1)} - x_*^{(1)}\|^2 \\
\|\ldots\| \\
\|x_k^{(n)} - x_*^{(n)}\|^2
\end{array}\right)
\]
where $A$ is interpreted component-wise and
\[
A_{ij} = 1 + \frac{m_j}{m} (c - 1) - \frac{nr}{n+1} + \frac{m - m_j}{m} (q + \frac{r}{n+1})
\]
Here $m_j$ is the number of rows in the $j$-th system and $m = \sum_j m_j$.

By induction,
\[
\left(\begin{array}{c}
\|x_{k+b}^{(0)} - x_*^{(0)}\|^2 \\
\|x_{k+b}^{(1)} - x_*^{(1)}\|^2 \\
\|\ldots\| \\
\|x_{k+b}^{(n)} - x_*^{(n)}\|^2
\end{array}\right) \leq A^b \left(\begin{array}{c}
\|x_k^{(0)} - x_*^{(0)}\|^2 \\
\|x_k^{(1)} - x_*^{(1)}\|^2 \\
\|\ldots\| \\
\|x_k^{(n)} - x_*^{(n)}\|^2
\end{array}\right)
\]
for all $b \in \mathbb{N}$ provided that $e_{k+a} < \varepsilon$ for all $a \in \{0, \ldots, b - 1\}$.

We wish to show the $\ell_1$ operator norm of $A$ is less than 1. This happens when
\[
\frac{m_j}{m} (c - 1) - \frac{nr}{n+1} + \left(q + \frac{r}{n+1}\right) \left(\frac{m - m_j}{m} + 2n \frac{m_j}{m}\right)
\]
is negative. This occurs when $q + \frac{nr}{n+1}$ is small enough. So then, for $r$ sufficiently small, we can choose $\varepsilon$ to make $\|A\|_{\ell_1} < 1$.

Suppose $e_k < \delta < \varepsilon$. By our previous bound, $e_{k+b} < d^b \delta$, conditioned on the intermediate values $e_{k+a} < \varepsilon$. By Markov’s inequality, the probability that $e_{k+a} \geq \varepsilon$ is at most $\frac{1}{\delta} d^a \delta$. The total probability of this happening is at most $\frac{d}{\delta} 1_{\delta < \varepsilon}$. Therefore our total error remains bounded above by $\varepsilon$ with probability at least $1 - \frac{d}{\varepsilon} 1_{\delta < \varepsilon}$, and in this case we have convergence in expectation.

\[
\square
\]

B. Convergence with full probability

An outline of the proof for Theorem 2:

1) We use Theorem 1 to define “convergence basins”: regions where, if the iterates fall into, there is some positive probability that they never escape and converge in expectation.

2) We show $\|x_k^{(i)} - x_*^{(i)}\|$ is bounded by some constant independent of $i$ and $k$.

3) We show we can bound the probability of falling into a basin eventually below by some positive number.

We will begin by proving the following lemma, which will be used in the second part of the outline above.

Lemma 1. Let $R$ be a sequence of rows from the problem. The sequence of Kaczmarz updates corresponding to $R$ defines an affine transformation $v \mapsto T_R v + v_R$. There are constants $c_r \in (0, 1), B_r \in \mathbb{R}_+$ for $r \in \{1, \ldots, d\}$ so that $\|v_R\| \leq B_{\dim \text{span } R}$ and $\|T_R\| \leq c_{\dim \text{span } R}$, where $\|\cdot\|_R$ is the $\ell_2$ operator norm when the operator is restricted to span $R$.

Proof of Lemma 1. We proceed by induction on $r$. The Kaczmarz update for the $\ell$-th row is
\[
K_{\ell} : v \mapsto \left( I - \frac{1}{\|M_{\ell}\|^2} M_{\ell}^T M_{\ell} \right) v + \frac{b_{\ell}}{\|M_{\ell}\|^2} M_{\ell}^T
\]
If $r = 1$, then $T_R$ is the zero operator restricted to span $R$. We can take $c_1 = 0$ and $B_1 = \max_i |b_i|/\|M_i\|^2$.

Now, assume the lemma is true for all $r < r'$. Let $R = (M_{t_1}, \ldots, M_{t_k})$ be a sequence of rows where $\dim \text{span } R = r'$. We can group
\[
K_{t_k} \circ \ldots \circ K_{t_1} = (K_{t_k} \circ \ldots \circ K_{t_N})
\]
so that $a_0 = 1$ and $a_i$ for $i \in \{1, \ldots, N\}$ is a strictly increasing sequence of rows where the following are true:

\[
\dim \text{span } \{M_{t_{i-1}}, \ldots, M_{t_i}\} = r' \quad \forall i \in \{1, \ldots, N\}
\]
\[
\dim \text{span } \{M_{t_{i-1}}, \ldots, M_{t_i}\} = r' - 1 \quad \forall i \in \{1, \ldots, N\}
\]
\[
\dim \text{span } \{M_{t_{i-1}}, \ldots, M_{t_i}\} < r'.
\]

Consider a grouping $(K_{t_{i-1}} \circ \ldots \circ K_{t_{i-1}})$, the linear part of the transformation is $A = \left( I - \frac{1}{\|M_{t_i}\|^2} M_{t_i}^T M_{t_i} \right)$ composed with an operator $T$ that sends $S = \text{span} \{M_{t_{i-1}}, \ldots, M_{t_i}\}$ to itself and has operator norm less than $c_{r' - 1}$ on this space. Consider a unit vector $v \in \text{span } R$. We can decompose $v = v_1 + v_2$ with $v_1 \in S$ and $v_2 \in S^2 \cap \text{span } R$. This allows us to bound
\[
\|ATv\| \leq \|Tv\| \leq \sqrt{c_{r'-1}^2 \|v_1\|^2 + \|v_2\|^2}
\]

1369
using that the operator norm of $A$ is at most 1 and that $T$ is the identity on $S^\perp$. Another bound is

$$\|ATv\| \leq \|ATv_1\| + \|Av_2\| \leq \|v_1\| + \left(1 - \frac{|M_{\ell a_i}v_2|}{\|M_{\ell a_i}\|}\right)$$

obtained with the triangle inequality and again $T$ being the identity on $S^\perp$.

These bounds combine to give a bound for the $\ell^2$ operator norm of $AT$ on $\text{span } R$ that depends only on $S$ and $\ell a_i$. There are finitely many possible choices for $S$ and $\ell a_i$, so there is some bound $c' < 1$ on the operator norm of $AT$ independent of what $S$ and $\ell a_i$ are. Next we turn to the affine part. By the induction hypothesis, this is a vector with norm at most $B' = B_{r-1} + \max_{\ell} \frac{|b_{\ell i}|}{\|M_{\ell i}\|}$.

Next we turn to the last grouping, which is not of this form. We will not analyze the linear part, and note the affine part $v'$ is bounded above in norm by $B'' = \max_{\ell, \ell' \neq \ell} B_{\ell'}$.

Since all linear operators here have operator norm at most 1, a final bound for the operator norm of $T_R$ restricted to $\text{span } R$ is $c'$, which we can take to be $c_{r'}. We bound

$$\|v_R\| \leq B'' + \sum_{i=1}^{\ell} c'_{\ell i}^{-1} B' \leq B'' + \frac{1}{1 - c'} B'.$$

So we can let $B_{r'}$ be this bound.

**Proof of Theorem 2.** We first bound the norm of the iterates above by some constant $D$.

Consider the evolution of a single iterate $x_k(i)$ after finitely many steps. At the last update for $x_k(i)$, we perform a Kaczmarz update with respect to some system, and move the iterate towards, but not past, the update. There is a line segment of possible choices for the next update once we have selected the row. The potential next iterate with largest norm is one of the end points of the line segment, corresponding to either doing a full update or no update at all. So we can either remove this last update or replace it with a full update to yield a final iterate with norm at least as large. We repeat with each update in reverse order, noting the image of a line segment after a series of affine transformations is still a line segment, choosing the one that will yield the largest norm at the end. Hence, to bound the norm of the iterates, we only need to consider sequences where we only ever make full Kaczmarz updates.

Using Lemma 1, we see that if the initial norm of the iterates are bounded above by $A$, after a sequence of rows $R$ the norm is at most $c_{\text{dim } R} \|A\| + B_{\text{dim } R}$, which is bounded above by $D = A + \max_{\ell} \frac{|b_{\ell i}|}{\|M_{\ell i}\|} B_{r'}$.

As given by Theorem 1, let $\epsilon$ be such that if $e_k \leq \epsilon$, we converge with some positive probability $t$.

Let $C$ be the set of all possible iterates such that $\frac{1}{n+1} \leq \left\|x_k(i) - x_{\ast}(i)\right\|^2 \leq D^2$ for all $i \in \{0, \ldots, n\}$ with $D$ given above. This set is compact. If our iterates do not lie in $C$, then already $e_k < \epsilon$, so we only need to look at what happens if our iterates lie in $C$.

Define

$$g(x^{(0)}, \ldots, x^{(n)}) = \max_{\ell \in \{1, \ldots, m\}} \min_{i \in \{0, \ldots, n\}} \frac{|M_{\ell i}x^{(i)} - b_{\ell}|}{\|M_{\ell i}\|}.$$

This function is the norm of the largest possible iterates in our algorithm if the iterates currently take the values $x^{(0)}, \ldots, x^{(n)}$. This is a continuous function on $\mathbb{R}^{d(n+1)}$, so it achieves a minimum $c$ on $C$. This minimum $c$ is positive, as by our assumption $g$ cannot be zero anywhere on $C$.

Now, for all value of iterates, we have probability at least $\frac{1}{m}$ of choosing a row where we will make an update with norm at least $c$. The probability of updating the correct system with a chosen row is at least $\frac{r}{m+1}$. The squared error of the corresponding iterate decreases by at least the norm squared of the update, which is at least $c^2$, because the resulting triangle between the previous iterate, next iterate, and solution is obtuse. We can keep doing this as long as our iterates remain in $C$. Therefore, if we make no more than

$$A = (n + 1) \left[ \frac{N - \frac{\epsilon}{n+1}}{c^2} \right]$$

of these updates, we have $e_k < \epsilon$. The probability of this happening is

$$\left( \frac{r}{m(n+1)} \right)^B,$$

where $m$ is the number of rows of $M$. This is a fixed positive value independent of the iterate. □

**IV. Experimental Results**

We test the MRK method on synthetic and real world data to verify the merits of the method. First we construct a problem in two-dimensional space and visualize how our iterates move in space in Figure 1. From Figure 1 we see how the iterates converge to solutions, moving closer with each projection. The problem is defined by two $10 \times 2$ systems where $M_1 \in \mathbb{R}^{10 \times 2}$ with entries drawn i.i.d. from $\mathcal{N}(0, 0.3)$ and $M_2 \in \mathbb{R}^{10 \times 2}$ with entries drawn i.i.d. from $\mathcal{N}(-0.8, 0.3)$.

Next, in Figure 2 we use the MRK method on a large synthetic data set. We plot the log norm squared error per iteration of the two iterates for the two class system. The system is defined by two matrices $M_1 \in \mathbb{R}^{1000 \times 10}$, $M_2 \in \mathbb{R}^{1000 \times 10}$ where the matrices have entries drawn i.i.d. from $\mathcal{N}(0, 1)$. Each initial iterate $x_0 \sim \mathcal{N}(0, 1)$ with zero swap probability. We plot the median and shade the interquartile range in Figure 2. We observe that both iterates converge to machine precision and the method succeeds at solving both systems simultaneously.

Finally, in Figure 3 we define a two problem system for the MRK method using real world data [11]. We construct two systems defined by matrices $M_1 \in \mathbb{R}^{300 \times 10}$, $M_2 \in \mathbb{R}^{399 \times 10}$ submatrices of the Wisconsin breast cancer data set. We start two initial iterates with standard normal entries and let our swap probability be zero. We observe that in Figure 3 both iterates converge to their respective solutions. We plot the median and interquartile range for the iterates' log error norm per iteration over 100 trials and observe that the method is able to solve this two system problem.
Fig. 1. Here we plot the evolution of two iterates in the two-dimensional plane. Our system is defined by two matrices $M_1 \in \mathbb{R}^{10 \times 2}$ with entries drawn i.i.d. from $\mathcal{N}(0,8,0.3)$ and $M_2 \in \mathbb{R}^{10 \times 2}$ with entries drawn i.i.d. from $\mathcal{N}(-0.8,0.3)$. Each initial iterate $x_0^1, x_0^2 \sim \mathcal{N}(0,1)$. We let our swap probability $r = 0$ and sample rows uniformly at random.

Fig. 2. System defined by two matrices $M_1 \in \mathbb{R}^{1000 \times 10}, M_2 \in \mathbb{R}^{1000 \times 10}$ where the matrices have entries distributed $M_1, M_2 \sim \mathcal{N}(0,1)$. Each initial iterate $x_0^1, x_0^2 \sim \mathcal{N}(0,1)$. We let our swap probability $r = 0$, sample rows uniformly at random and plot the median and interquartile range over the 100 trials.

Fig. 3. System defined by matrices $M_1 \in \mathbb{R}^{300 \times 10}, M_2 \in \mathbb{R}^{399 \times 10}$ submatrices of the Wisconsin breast cancer data set. Each initial iterate $x_0^1, x_0^2 \sim \mathcal{N}(0,1)$. We let our swap probability $r = 0$, sample rows uniformly at random and plot the median and interquartile range over the 100 trials.

V. CONCLUSION AND FUTURE WORK

In this paper we introduce the novel multi-randomized Kaczmarz algorithm, Algorithm 1, to solve the consistent latent class regression problem. We prove linear convergence for the algorithm in expectation with high probability under some contraints in Theorem 1 and almost surely in Theorem 2. Additionally, we observe promising results when applying the algorithm to test data sets. We plan on extending this work to inconsistent and noisy systems by leveraging the extended Kaczmarz method [12], which converges to the least squares solution [13]. Additionally, we would like to explore using Kaczmarz variants such as max distance [14], sampling Kaczmarz-Motzkin [15] and selectable set [16] methods in this setting. Finally, we are interested in adaptively marking and assigning which rows belong to which system in real time based on the iterate projection values.

REFERENCES

[1] S. Kaczmarz, “Angenaherte auflösung von systemen linearer gleichungen,” Bull. Internat. Acad. Polon. Sci. Lettres A, pp. 335–357, 1937.
[2] T. Strohmer and R. Vershynin, “A randomized Kaczmarz algorithm with exponential convergence,” Journal of Fourier Analysis and Applications, vol. 15, no. 2, pp. 262–278, 2009.
[3] M. Wedel and W. S. De Sarbo, “A review of recent developments in latent class regression models,” Advanced methods of marketing research, pp. 352–388, 1994.
[4] J. Magidson and J. K. Vermunt, “Latent class models,” The Sage handbook of quantitative methodology for the social sciences, pp. 175–198, 2004.
[5] T. K. Moon, “The expectation-maximization algorithm,” IEEE Signal processing magazine, vol. 13, no. 6, pp. 47–60, 1996.
[6] J. M. Kluowski, D. Yang, and W. Brinda, “Estimating the coefficients of a mixture of two linear regressions by expectation maximization,” IEEE Transactions on Information Theory, vol. 65, no. 6, pp. 3515–3524, 2019.
[7] C. J. Wu, “On the convergence properties of the em algorithm,” The Annals of statistics, pp. 95–103, 1983.
[8] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via the em algorithm,” Journal of the Royal Statistical Society: Series B (Methodological), vol. 39, no. 1, pp. 1–22, 1977.
[9] K. Greff, S. Van Steenkiste, and J. Schmidhuber, “Neural expectation maximization,” Advances in Neural Information Processing Systems, vol. 30, 2017.
[10] X. Chen and A. Powell, “Almost Sure Convergence of the Kaczmarz Algorithm with Random Measurements,” Journal of Fourier Analysis and Applications, vol. 18, 12 2012.
[11] W. Wolberg, W. Street, and O. Mangasarian, “Breast cancer Wisconsin (diagnostic) UCI machine learning repository,” Irvine, CA, USA, 1995.
[12] A. Zouzias and N. M. Freris, “Randomized extended kaczmarz for solving least squares,” SIAM Journal on Matrix Analysis and Applications, vol. 34, no. 2, pp. 773–793, 2013.
[13] D. Needell, “Randomized kaczmarz solver for noisy linear systems,” BIT Numerical Mathematics, vol. 50, no. 2, pp. 395–403, 2010.
[14] J. Nutini, B. Sepehrv, I. Laradji, M. Schmidt, H. Koepke, and A. Virani, “Convergence rates for greedy kaczmarz algorithms, and faster randomized kaczmarz rules using the orthogonality graph,” in Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence. AUAI Press, 2016, p. 547–556.
[15] J. A. De Loera, J. Haddock, and D. Needell, “A sampling kaczmarz-Motzkin algorithm for linear feasibility,” SIAM Journal on Scientific Computing, vol. 39, no. 5, pp. S66–S87, 2017.
[16] Y. Yaniv, J. D. Moorman, W. Swartworth, T. Tu, D. Landis, and D. Needell, “Selectable set randomized kaczmarz,” Numerical Linear Algebra with Applications, p. e2458, 2022.