Block Kaczmarz–Motzkin Method via Mean Shift Clustering

Yimou Liao 1, Tianxiu Lu 1,2,* and Feng Yin 3

1 College of Mathematical and Statistics, Sichuan University of Science and Engineering, Zigong 643000, China; liaoyimou596@163.com
2 Key Laboratory of Higher Education of Sichuan Province for Enterprise Informationalization and Internet of Things, Zigong 643000, China
3 College of Mathematics and Physics, Chengdu University of Technology, Chengdu 643000, China; fyin@cdut.edu.cn
* Correspondence: lubeeltx@suse.edu.cn; Tel.: +86-13408138464

Abstract: Solving systems of linear equations is a fundamental problem in mathematics. Combining mean shift clustering (MS) with greedy techniques, a novel block version of the Kaczmarz–Motzkin method (BKMS), where the blocks are predetermined by MS clustering, is proposed in this paper. Using a greedy strategy, which collects the row indices with the almost maximum distance of the linear subsystem per iteration, can be considered an efficient extension of the sampling Kaczmarz–Motzkin algorithm (SKM). The new method linearly converges to the least-norm solution when the system is consistent. Several examples show that the BKMS algorithm is more efficient compared with other methods (for example, RK, Motzkin, GRK, SKM, RBK, and GRBK).

Keywords: consistent linear system; Kaczmarz–Motzkin; clustering method; convergence property

MSC: 65F10; 65F25; 90C25; 15A06

1. Introduction

This paper is concerned with the approximate solution of a large-scale linear system of equations of the form

\[ Ax = b, \quad A \in \mathbb{R}^{m \times n}, \quad b \in \mathbb{R}^m, \]  

where \( x \in \mathbb{R}^n \) need to be determined. If \( m > n \), the linear systems (1) with such thin (or tall) coefficient matrices \( A \) are overdetermined and are often inconsistent. In this case, we are interested in the computation of the least-squares solution of systems (1) by

\[ x_{LS} := \arg \min \limits_x \| Ax - b \|^2_2, \]  

where \( \cdot \|_2 \) denotes the Euclidean norm. If \( m < n \), the linear system (1) with such a fat (or flat) matrix \( A \) is underdetermined. If the linear system is consistent and has many solutions, then the least Euclidean-norm solution

\[ x_{LN} := \arg \min \limits_x \| x \|_2, \quad Ax = b \]

will often be considered. In this paper, we focus on the approximate of the least-norm solution when (1) is consistent.

The Kaczmarz method in [1] is a popular algorithm for solving the linear system (1), which is called the algebraic reconstruction technique (ART), and has a large range of fields of applications, such as image reconstruction in computerized tomography [2,3], signal processing [4], and distributed computing [5,6]. In the classical Kaczmarz method [1], all rows of the matrix \( A \) are circularly passed according to the given order, and the current iteration point is orthogonally projected to the hyperplane formed by the index row for the
next iteration. For a given initial solution \( x_0 \), the iteration scheme of the Kaczmarz method can be presented as

\[
x_{k+1} = x_k + \frac{b_k - A^k x_k}{\|A^k\|^2} (A^k)^T,
\]

where \( i_k = \text{mod}(k, m) + 1, k = 0, 1, \cdots \), \( A^i \) denotes the \( i \)th row of \( A \), \( b_i \) presents the \( i \)th element of \( b \), and \( A^T \) denotes the transpose of \( A \).

For the consistent linear system (1), Strohmer and Vershynin [7] in 2009 presented a randomized Kaczmarz algorithm (RK) with expected exponential convergence. The RK algorithm selects each iteration row with the row index \( i_k = \|A^i\|^2 / \|A\|^2 \). A theoretical analysis showed that, compared with the selection of a row in the natural order, randomly selecting a row of iterations can greatly improve the convergence speed. In [8], Bai and Wu designed a greedy randomized Kaczmarz (GRK) algorithm. GRK uses a greedy probability strategy, which is derived from the maximum distance rule, to grasp larger entries of the residual vector, and the work row with a larger residual error is determined. The theoretical analysis shows that the GRK converges to the least norm solution with the expected exponential, and the numerical experiments illustrate that GRK converges much faster than RK. In [9], De Loera proposed a sampling Kaczmarz–Motzkin (SKM) for linear feasibility problems by combining the RK method with the Motzkin method [10]. This method mainly consists of two stages. The row index set \( \{1, 2, \cdots, m\} \) of matrix \( A \) is divided into multiple subsets \( T = \{\tau_1, \tau_2, \cdots, \tau_p\} \) for the first stage, and the second stage involves determining the working row \( i_k = \arg \max_{i \in \tau_k} \{b_i - A^i x_k\}^2 \) by uniformly extracting one \( (\tau_k, k \in [p]) \) in each iteration, and then updating \( x_k \) by (4). In particular, the SKM method overcomes the weakness of the RK and Motzkin methods. For example, the RK method may be slow to converge because it does not utilize a greedy strategy, and the calculations of all residual vectors are expensive for the Motzkin method. However, the convergence rate of the SKM method may be slow since it enforces only one constraint for the current iteration.

The block Kaczmarz methods, which utilize the multi-row constraints, have received extensive attention for their fast convergence properties. The block iterative methods were first presented by Elfving [11] and Eggermont et al. [12] to solve (1). Then, Needell et al. [13] proposed a randomized block Kaczmarz (RBK) algorithm to solve the least-squares problem by selecting a subsystem from the predetermined partition at random, which linearly converges to the least-norm solution in expectation. The randomized block Kaczmarz algorithm iterates the form

\[
x_{k+1} = x_k + A^\dagger_{\tau_k} (b_{\tau_k} - A_{\tau_k} x_k),
\]

where \( A_{\tau_k} \) and \( b_{\tau_k} \) are the submatrix and subvector of \( A \) and \( b \), respectively, \( \tau_k \subseteq \{1, 2, \cdots, m\} \) and \( A_{\tau_k}^\dagger \) represent the Moore–Penrose pseudoinverse of subsystem matrix \( A_{\tau_k} \). Although the existence of this “good” paving of all rows is theoretically guaranteed, it is not always effortless to find it. Liu and Gu in [14] designed a randomized greedy block Kaczmarz method to choose a “good” target block by embedding a greedy strategy in the predetermined subsystems, which was derived from the probability criterion proposed by Bai and Wu in [8]. The convergence theory of GRBK shows that the iterative solution sequence linearly converges to the least-norm solution in expectation. In addition, the partitioning method may not be a good partition since it only focuses on the average partitioning of the row index set. Many variants of the block Kaczmarz have recently received extensive attention. For example, literature studies [15–21] and references therein.

By exploring the structural properties of the coefficient matrix \( A \), some criteria can be used to determine the row partitions (see [13,14,16,22]). In the context of data science, the mean shift clustering (MS) is an iterative algorithm based on kernel density estimation [23], which is widely used for data clustering. So, the combination of the clustering algorithm and greedy technique is extremely meaningful. Inspired by this, to improve the efficiency of block Kaczmarz, we constructed a novel block Kaczmarz–Motzkin method (BKMS\( \delta \),
by determining the largest residual block, and then introducing an almost maximum distance criterion for the current block to collect the row indices, close to the largest entry (absolute value), as the iteration index set. The row partitions were divided based on MS clustering, and the row correlation coefficient of the system matrix A was considered the clustering criterion. Moreover, unlike k-means clustering \([16,24]\), it only needs to initialize the bandwidth’s \(d \in (0,1]\) automatic partitions instead of pre-specifying the number of partitions \(k\), which depends on the initialized cluster centers.

This paper is organized as follows. The RBK method in \([13]\) and the GRBK method in \([14]\) are summarized in Section 2. Section 3 introduces the MS clustering briefly, and then the BKMS method is proposed and its convergence theory is constructed when (1) is consistent. Section 4 reports several examples and applications. We complete this work with some conclusions and discuss future work in Section 5.

2. Block Kaczmarz Algorithm and Its Variants

For a matrix \(Q = (q_{ij}) \in \mathbb{R}^{m \times n}\), we indicate by \(Q^i, Q_{ij}, \|Q\|_2 := \max_{x \neq 0} \frac{\|Qx\|_2}{\|x\|_2}, \|Q\|_F := \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} |q_{ij}|^2}\) and \(Q^i\) the \(i\)th row and a row submatrix (\(\tau_\ell\) is a row indicator block), the spectral norm, Frobenius norm, and the transpose of the \(Q\), respectively. If any two rows of \(Q\) are independently and identically distributed (IID), the correlation coefficient of \(Q\) is given by

\[
\mu_k = \frac{\|Q^\ell(Q^p)^T\|}{\|Q^\ell\|_2 \|Q^p\|_2}, \quad l \neq p \in \{1,2,\ldots,m\}. 
\]

In addition, \(\sigma_{\text{max}}(Q)\) and \(\sigma_{\text{min}}(Q)\) are the largest and smallest positive singular values of the \(Q\), respectively. Moreover, \(p_i\) denotes the \(i\)th component of a vector \(p = (p_i) \in \mathbb{R}^m\). For constant \(c \in \mathbb{R}\), we call \(\lfloor c \rfloor\) the largest integer not exceeding \(c\).

2.1. The RBK Algorithm

The RBK method \([13]\) is summarized in Algorithm 1, which is described by Needell and Tropp to solve the overdetermined least-squares problem (2).

Algorithm 1 The randomized block Kaczmarz method (RBK)

Input: \(A \in \mathbb{R}^{m \times n}\) is standardized (i.e., \(\|A^i\|_2 = 1, i \in [m]\)), \(b \in \mathbb{R}^m, x_0 \in \mathbb{R}^n, l \in \mathbb{N}^+\) and partition \(\mathcal{T} = \{\tau_1, \tau_2, \ldots, \tau_p\}_{p \leq m}\)

Output: \(x_l\)

for \(k = 0,1,\ldots,l-1\) do

Select a block \(\tau_\ell\) uniformly at random from \(\mathcal{T}\)

Update \(x_{k+1}\) by (5)

end for

The restatement of Lemma 1 in \([13]\) guarantees the linear convergence of the RBK method.

Lemma 1. Assume that the overdetermined linear system (1) with \(A\) is of full rank and row normalized (i.e., \(\|A^i\|_2 = 1, i \in [m]\)). Let \(\alpha \leq \sigma_{\text{min}}^2(A^\tau)\) and \(\beta \geq \sigma_{\text{max}}^2(A^\tau)\) for each \(\tau \in \mathcal{T}\). Given an initial vector \(x_0 \in \mathbb{R}(A^\mathcal{T})\), then the sequence \(\{x_k\}_{k=0}^\infty\) generated by the RBK method exponentially converges to the least-squares solution \(x_* = A^\dagger b\) of (1) in expectation. Furthermore,

\[
\mathbb{E}\|x_k - x_*\|_2^2 \leq \left(1 - \frac{\sigma_{\text{min}}^2(A)}{\beta m}\right)^k \|x_0 - x_*\|_2^2 + \frac{\beta}{\alpha} \frac{\|Ax_* - b\|_2^2}{\sigma_{\text{min}}^2(A)}.
\]

According to Lemma 1, the convergence rate of RBK is limited by \(\sigma_{\text{min}}^2(A)\). If \(A\) is well-conditioned, the upper bound of the convergence error of RBK is large. In addition, we ignore the assumption that the linear system is inconsistent. The second term on the right-hand side of the inequality is zero. So, we remark the convergence factor of RBK as
\[
\rho_{\text{RBK}} = 1 - \frac{\sigma_{\text{min}}^2(A)}{\beta m}.
\]

2.2. The GRBK Algorithm

Based on the motivation of the GRK method, Liu and Gu in [14] applied the greedy sampling to the RBK method to replace uniform sampling, and proposed the GRBK method. Algorithm 2 summarizes the GRBK method to solve (1), as follows.

**Algorithm 2** The greedy randomized Kaczmarz block method (GRBK)

**Input:** \( A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^m, x_0 \in \mathbb{R}^n, l \in \mathbb{N}^+ \) and partition \( T = \{ \tau_1, \tau_2, \cdots, \tau_p \} \).

**Output:** \( x_l \)

for \( k = 0, 1, \ldots, l - 1 \) do

Compute

\[
\epsilon_k = \frac{1}{2} \left( \frac{1}{\| b - Ax_k \|_2^2} \max_{\tau_i \in T} \left\{ \frac{|b_{\tau_i} - A_{\tau_i} x_k|^2}{\| A_{\tau_i} \|_2^2} \right\} + \frac{1}{\| A \|_F^2} \right)
\]

Determine the partition of the indicator set

\[
U_k = \{ \tau_i \mid |b_{\tau_i} - A_{\tau_i} x_k|^2 \geq \epsilon_k \| b - Ax_k \|_2^2 \| A_{\tau_i} \|_2^2 \}
\]

Compute the block \( \tilde{p}_k^{\tau_i} \) of the entry \( \tilde{r}_i \) according to

\[
\tilde{p}_k^{\tau_i} = \begin{cases} 
\tilde{r}_i = b_{\tau_i} - A_{\tau_i} x_k, & \text{if } \tau_i \in U_k, \\
0, & \text{otherwise}
\end{cases}
\]

Select \( \tau_i \in U_k \) with probability

\[
P_r(\text{block} = \tau_i) = \frac{|\tilde{p}_k^{\tau_i}|_2^2}{\| \tilde{r}_i \|_2^2}
\]

Update \( x_{k+1} \) by (5)

end for

According to Algorithm 2, we consider that, if there is only a row in each block, then GRK in [8] is a special case of the GRBK method. Liu and Gu proved that the GRBK algorithm has a lower convergence upper bound than RBK and demonstrated the effectiveness of GRBK with several numerical experiments. The result in [14] is restated as follows.

**Lemma 2.** For a fix partition \( T = \{ \tau_1, \tau_2, \cdots, \tau_p \} \) of the row indices \( \{1, 2, \cdots, m\} \), we assume that \( \xi = \min_{\tau \in T} \| A \|_F^2 \) and \( \beta \geq \sigma_{\text{max}}^2(A_T) \) for each \( \tau \in T \). Given an initial vector \( x_0 \in \mathbb{R}(A^T) \), then the sequence \( \{x_k\}_{0}^{\infty} \) generated by the GRBK method exponentially converges to the least norm solution \( x_* = A^b \) in expectation when (1) is consistent. Moreover, it satisfies

\[
E \| x_1 - x_* \|_2^2 \leq \left( 1 - \frac{\xi^2 \sigma_{\text{min}}^2(A)}{\beta \| A \|_F^2} \right) \| x_0 - x_* \|_2^2
\]

and

\[
E \| x_{k+1} - x_* \|_2^2 \leq \left( 1 - \gamma \xi \frac{\sigma_{\text{min}}^2(A)}{\beta \| A \|_F^2} \right)^{k-1} \left( 1 - \frac{\xi^2 \sigma_{\text{min}}^2(A)}{\beta \| A \|_F^2} \right) \| x_0 - x_* \|_2^2,
\]

where \( \gamma = \frac{1}{2} \left( \frac{\| A \|_2^2}{\| A \|_F^2} + 1 \right) \).
By Lemma 2, the following estimates bound the convergence factor of the GRBK method

$$\rho_{GRBK} = 1 - \frac{\xi}{2} \left( \frac{\|A\|_F^2}{\|A\|_F^4 - \xi} + 1 \right) \frac{\sigma_{\min}^2(A)}{\beta \|A\|_F^2},$$

which is less than $\rho_{RBK}$ when $A$ is row normalized.

3. Block Kaczmarz–Motzkin Methods with Mean Shift Clustering

The greedy strategy in the SKM method ensures that the entry with the largest residual is prioritized. We hope to exploit the correlations between subsystem matrices to determine the row partition of the coefficient matrix $A$ in an efficient way. In the framework of big data, the mean shift method (MS) in [25] is one of the clustering algorithms that divides the data into different categories. Therefore, the combination of the MS clustering and the Kaczmarz method with greedy techniques comes naturally. Inspired by this, a novel block Kaczmarz–Motzkin method based on MS clustering is proposed. In this section, we first restate the sketch of the MS clustering, then design a new, more realistic block Kaczmarz–Motzkin method and construct its convergence theory.

3.1. The Mean Shift Clustering

The MS clustering is an iterative algorithm based on kernel density estimation. It mainly consists of three steps, as follows. The first step is to select a point $\tilde{x}_i, i \in \{1, 2, \cdots, m\}$ randomly from the unclassified data items $X = \{\tilde{x}_1, \tilde{x}_2, \cdots, \tilde{x}_m\}$ as the center point $y_j$. The second step is to find out all the data points in the circle entered on $y_j$ and the bandwidth $\delta/2$ as the radius. It can be written as

$$S_\delta(y_j) = \{\tilde{x}_i \mid \|y_j - \tilde{x}_i\|_2^2 < \delta/2\}, \quad j = 1.$$

Considering these points belong to the cluster $C_j$, calculate the vector from the center point $y_j$ to each element $x_i$ in the set $S_\delta(y_j)$, and add these vectors to obtain the mean shift vector $m(y_j)$, as follows

$$m(y_j) = \frac{\sum_{i=1}^n x_i g(\|y_j - x_i\|_2^2 / \delta^2)}{\sum_{i=1}^n g(\|y_j - x_i\|_2^2 / \delta^2)} - y_j,$$

where $\tilde{x}_i \in S_\delta(\tilde{x})$ and the Gaussian kernel function $g(\cdot)$ are designed to assign different weights to each data point. The center point is moved in the direction of the mean shift vector $m(y_j)$ by $|m(y_j)|$. Update the current center point by $y_{j+1} = m(y_j) + y_j (j = 1, 2, \cdots)$, then repeat the above process until the iterative sequence $\{y_j\}_{j \geq 1}$ satisfies $\|y_{j+1} - y_j\|_2^2 < \epsilon$, where $\epsilon$ is a user-preset threshold. The last step is to repeat the above two steps until all points are classified. Algorithm 3 summarizes the MS algorithm in [26] for data clustering.

According to Algorithm 3, one knows that, unlike the k-means algorithm in [16] and the random partition method in [13], the MS clustering algorithm does not require us to predetermine the number of clusters. For details, we refer to [25,26].
Algorithm 3 The mean shift method for clustering (MS)

**Input:** A dataset \( X = \{ \tilde{x}_1, \tilde{x}_2, \cdots, \tilde{x}_m \} \) containing \( m \) data items, the threshold \( \epsilon \) and the bandwidth \( \delta > 0 \).

**Output:** Estimate cluster centers \( y^*_1, y^*_2, \cdots, y^*_k \), where \( k \) is the number of clusters.

for \( i = 1, 2, \cdots, n \) do

Select an initial data point \( \tilde{x}_j \) randomly from the data set \( X \), let \( y_j = \tilde{x}_j \) and \( j \leftarrow 1 \).

repeat

Estimate the mean shift vector by

\[
m(y_j) = \frac{\sum_{i=1}^{n} \tilde{x}_i g(\frac{\| \tilde{x}_i - \tilde{x}_j \|_2}{\delta})}{\sum_{i=1}^{n} g(\frac{\| \tilde{x}_i - \tilde{x}_j \|_2}{\delta})} - y_j
\]

Update the cluster centers \( y_j+1 = m(y_j) + y_j \)

until \( \| y_{j+1} - y_j \|_2^2 \leq \epsilon \)

\( y^*_j \leftarrow y_j \)

end for

Merge the cluster centers with distances smaller than \( \delta \)

3.2. Block Kaczmarz–Motzkin Method with Mean Shift

The following describe in detail how to combine MS clustering and the Kaczmarz–Motzkin method. Considering the row partitions of \( A \), we first construct the dataset \( X = (F(row), \mu_k) \in \mathbb{R}^{m \times 2} \). The \( \mu_k \) is calculated by

\[
\mu_k = \frac{|A^i(A^I)^T|}{\|A^I\|_2 \|A^I\|_2},
\]

where \( i \neq I \in \{1, 2, \cdots, m\} \) and the superscript \( I \) refer to selecting an index randomly from the set formed by the maximum row norm of \( A \) (if there are multiple maximum row norm values). \( F(row) \) means linear mapping from the row index set of \( A \). By implementing the MS algorithm, we have partition \( \hat{T} = \{ \hat{\tau}_1, \hat{\tau}_2, \cdots, \hat{\tau}_p \} \) of \( \{1, 2, \cdots, m\} \), i.e.,

\[
A = [A_{\hat{\tau}_1}; A_{\hat{\tau}_2}; \cdots, A_{\hat{\tau}_p}] \quad \text{and} \quad b = [b_{\hat{\tau}_1}; b_{\hat{\tau}_2}; \cdots, b_{\hat{\tau}_p}].
\]

Similar to SKM, at the \( k \)th iteration, if \( \| b_{\hat{\tau}_s} - A_{\hat{\tau}_s}x_k \|_2 > \| b_{\hat{\tau}_s} - A_{\hat{\tau}_s}x_k \|_2 \), then the \( \hat{\tau}_s \) block is to be sampled rather than block \( \hat{\tau}_s \) for all \( i \neq s \in [p] \). Considering that a block extracted may be very large, we introduce an almost-maximum distance criterion in one iteration to control the iterative index set. Following this idea, the block Kaczmarz–Motzkin method with MS clustering is summarized in Algorithm 4.
Algorithm 4 Block Kaczmarz–Motzkin method with mean shift (BKMS).

**Input:** $A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^m, x_0 \in \mathbb{R}^n, I \in \mathbb{N}^+$, the bandwidth $\delta \in (0, 1]$, and the parameter $\eta \in (0, 1]$.

**Output:** $x_f$.

Data set $\mathcal{X}$ is obtained by data preprocessing on $A$.

Apply the MS$(\delta)$ method to $\mathcal{X}$ to obtain partition $\hat{\mathcal{F}} = \{\hat{\tau}_1, \hat{\tau}_2, \ldots, \hat{\tau}_p\}_{p \geq 1}$ from the set of row index sets $\{1, 2, \ldots, m\}$

for $k = 0, 1, \ldots, l - 1$ do

Determine the partition of the indicator set

$$\mathcal{R}_k = \{\tau_k | \tau_k = \arg \max_{1 \leq j \leq p} \|r^j_k\|_2^2\}$$

Compute $\epsilon_k = \eta \max_{j \in \tau_k} \frac{|b_j - A\tau_kx_k|^2}{\|A\tau_k\|^2_2}$

Choose $\hat{\tau}_k = \{i_k | \|b_{i_k} - A\hat{\tau}_kx_k\|^2 \geq \epsilon_k \|A\hat{\tau}_k\|^2_2\}$

Update $x_{k+1} = x_k + A_{\hat{\tau}_k}^\dagger(b_{\hat{\tau}_k} - A_{\hat{\tau}_k}x_k)$

end for

Remark 1. We compress the row index set of $A$ to the interval of $\mu_k$, i.e., $[1, m]$ $\xrightarrow{F} [\mu_k^{\min}, \mu_k^{\max}]$, and remark that the $F$ is a linear map. In fact, if $\forall$ $x, y \in \{1, 2, \cdots, m\}$ and $a, b \in \mathbb{R}$, the following conditions hold:

$$F(ax + by) = aF(x) + bF(y)$$

Now, the convergence of the BKMS algorithm is analyzed. The fact guaranteed in [2] is required.

Lemma 3. If $A \in \mathbb{R}^{m \times n}$ is a nonzero real matrix, then

$$\sigma_{\min}^2(A)\|u\|_2^2 \leq \|Au\|_2^2 \leq \sigma_{\max}^2(A)\|u\|_2^2$$

for any $u \in \text{range}(A^T)$.

The following theorem gives the convergence rate of the BKMS Algorithm 4 when (1) is consistent.

Theorem 1. Let $\{\hat{\tau}_p\}_{p \geq 1}$ be the target blocks generated by Algorithm 3. Assume that $\zeta \geq \max_{\tau} \|A_\tau\|_2^2$ and $\alpha \geq \max |\tau|$ for each $\tau \in \{\hat{\tau}_p\}_{p \geq 1}$, where $|\tau|$ is the cardinality of $\tau$. Given an initial guess $x_0 \in \mathcal{R}(A^T)$, then the sequence $\{x_k\}_{k \geq 0}$ generated by the BKMS Algorithm 4 converges linearly to the least norm solution $x_\ast = A^\dagger b$ of (1). Furthermore,

$$\|x_1 - x_\ast\|_2^2 \leq \left[1 - \eta \frac{\|A_{\hat{\tau}_0}\|_2^2 \sigma_{\min}^2(A)}{\alpha \zeta \sigma_{\max}^2(A_{\hat{\tau}_0})}\right] \|x_0 - x_\ast\|_2^2$$

and

$$\|x_{k+1} - x_\ast\|_2^2 \leq \left[1 - \eta \frac{\|A_{\hat{\tau}_k}\|_2^2 \sigma_{\min}^2(A)}{\alpha (m - N_{k-1}) \zeta \sigma_{\max}^2(A_{\hat{\tau}_k})}\right] \|x_k - x_\ast\|_2^2, \quad k = 1, 2, \ldots$$

Then,

$$\|x_{k+1} - x_\ast\|_2^2 \leq \left[1 - \eta \frac{\|A_{\hat{\tau}_k}\|_2^2 \sigma_{\min}^2(A)}{\alpha (m - N_{k-1}) \zeta \sigma_{\max}^2(A_{\hat{\tau}_k})}\right]^k \left[1 - \eta \frac{\|A_{\hat{\tau}_0}\|_2^2 \sigma_{\min}^2(A)}{\alpha \zeta \sigma_{\max}^2(A_{\hat{\tau}_0})}\right] \|x_0 - x_\ast\|_2^2, \quad (6)$$

where $\sigma_{\max}^2(A_{\hat{\tau}_k})$ is the largest nonzero singular value of $A_{\tau}$ and $N_{k-1}$ is the cardinality of the set $\hat{\tau}_{k-1}$. 

**Proof.** According to the (5) of Algorithm 4

\[ x_{k+1} - x_k = A^\dagger_{\tau_k} (b_{\tau_k} - A_{\tau_k} x_k). \]

Since \( A_{\tau_k} x_s = b_{\tau_k} \), one can see that

\[ x_{k+1} - x_k = A^\dagger_{\tau_k} (A_{\tau_k} x_s - A_{\tau_k} x_k) = A^\dagger_{\tau_k} A_{\tau_k} (x_s - x_k), \]

which implies that \( x_{k+1} - x_k \) belongs to the column space of matrix \( A^\dagger_{\tau_k} A_{\tau_k} \).

From the iteration scheme of the BKMS,

\[ A_{\tau_k} (x_{k+1} - x_s) = A_{\tau_k} [x_k + A^\dagger_{\tau_k} (b_{\tau_k} - A_{\tau_k} x_k)] - A_{\tau_k} x_s = A_{\tau_k} x_k + A_{\tau_k} A_{\tau_k} b_{\tau_k} - A_{\tau_k} A^\dagger_{\tau_k} A_{\tau_k} x_k - A_{\tau_k} x_s = A_{\tau_k} x_k + A_{\tau_k} x_s - A_{\tau_k} x_k - A_{\tau_k} x_s = 0. \]

(7)

One can obtain that \( A^\dagger_{\tau_k} A_{\tau_k} (x_{k+1} - x_s) = 0 \), which shows \( x_{k+1} - x_s \) is orthonormal to \( A^\dagger_{\tau_k} A_{\tau_k} \). So, \( (x_{k+1} - x_k)^T (x_{k+1} - x_s) = 0 \) is clearly established.

Thus, the following formula can be obtained

\[ \|x_{k+1} - x_s\|_2^2 = \|x_k - x_s\|_2^2 - \|x_{k+1} - x_k\|_2^2 = \|x_k - x_s\|_2^2 - \|A^\dagger_{\tau_k} A_{\tau_k} (x_k - x_s)\|_2^2. \]

(8)

On the other hand, according to the algorithm flow, we know that

\[ \|r^\tau_{\tau_k}\|_2^2 = \max_{i \in |p|} \|r^\tau_i\|_2^2 \quad \text{and} \quad \frac{\|r^\tau_{\tau_k}\|_2^2}{\|A_{\tau_k}\|_F} = \max_{\tau_k \in R_k} \frac{\|r^\tau_{\tau_k}\|_2^2}{\|A_{\tau_k}\|_F} \]

So,

\[ \|A^\dagger_{\tau_k} A_{\tau_k} (x_k - x_s)\|_2^2 \geq \sigma^2_{\min}(A^\dagger_{\tau_k}) \|A_{\tau_k} (x_k - x_s)\|_2^2 = \frac{1}{\sigma^2_{\max}(A_{\tau_k})} \sum_{i \in \tau_k} |A^i(x_k - x_s)|^2 \]

\[ = \frac{1}{\sigma^2_{\max}(A_{\tau_k})} \sum_{i \in \tau_k} |r^i_{\tau_k}|^2 \geq \frac{1}{\sigma^2_{\max}(A_{\tau_k})} \sum_{i \in \tau_k} \eta \max_{j \in \tau_k} \|r^{\tau_j}_{\tau_k}\|_2^2 \|A^i\|_2^2 \]

\[ \geq \eta \frac{\|A_{\tau_k}\|_F^2}{\sigma^2_{\max}(A_{\tau_k})} \max_{j \in \tau_k} \|r^j_{\tau_k}\|_2^2 \frac{\|r^i_{\tau_k}\|_2^2 \|A^i\|_2^2}{\|A_{\tau_k}\|_F^2}. \]

Let \( \alpha \geq \max_{i \in |p|} |\tau_i| \), where \( |\tau_i| \) is the cardinality of \( \tau_i \), then

\[ \max_{i \in \tau_k} |r^i_{\tau_k}|^2 = \frac{\max_{j \in \tau_k} |r^j_{\tau_k}|^2 \|r^\tau_{\tau_k}\|_2^2 \|r^i_{\tau_k}\|_2^2 \|A^i\|_2^2}{\sum_{j \in \tau_k} |r^j_{\tau_k}|^2 \|A^j\|_2^2} \geq \frac{1}{\eta \|A_{\tau_k}\|_F^2} \|r^\tau_{\tau_k}\|_2^2 \frac{\|r^i_{\tau_k}\|_2^2 \|A^i\|_2^2}{\|A_{\tau_k}\|_F^2}. \]

So, one can obtain

\[ \|A^\dagger_{\tau_k} A_{\tau_k} (x_k - x_s)\|_2^2 \geq \frac{\eta \|A_{\tau_k}\|_F^2}{\alpha \sigma^2_{\max}(A_{\tau_k})} \max_{i \in |p|} \|r^i_{\tau_k}\|_2^2 \]
For $k = 0$,
\[
\max_{i \in [p]} \| r_i^0 \|_2^2 = \max_{i \in [p]} \| r_i^0 \|_2^2 \cdot \frac{\| r_0 \|_2^2}{\sum_{i \in [p]} \| r_i^0 \|_2^2} \geq \frac{1}{p} \| r_0 \|_2^2,
\]
then,
\[
\| A_{10} \|_2 A_{10}(x_1 - x_0) \|_2^2 \geq \frac{\eta \| A_{10} \|_2^2 \sigma_{\min}^2(A)}{\alpha p \sigma_{\max}^2(A_{10}) \| A_{10} \|_F^2} \| x_0 - x_1 \|_2^2. \tag{9}
\]

Assume that $\zeta \geq \max_{\tau} \| A_\tau \|_2^2$ for each $\tau \in \{ \tau_p \}_{p \geq 1}$. Substituting (9) into (8), one has
\[
\| x_1 - x_\ast \|_2^2 \leq \left[ 1 - \frac{\eta \| A_{10} \|_2^2 \sigma_{\min}^2(A)}{\alpha p \sigma_{\max}^2(A_{10}) \| A_{10} \|_F^2} \right] \| x_0 - x_\ast \|_2^2. \tag{10}
\]

For $k = 1, 2, \ldots,$
\[
\max_{i \in [p]} \| r_i^k \|_2^2 = \frac{\max_{i \in [p]} \| r_i^k \|_2^2 \| r_k \|_2^2}{\sum_{i \in [p]} \| r_i^k \|_2^2 \| r_k \|_2^2} \geq \frac{1}{m - N_{k-1}} \| r_k \|_2^2,
\]
the establishment of the second equation above is guaranteed by the (7), then
\[
\| A_{10} \|_2 A_{10}(x_k - x_\ast) \|_2^2 \geq \frac{\eta \| A_{10} \|_2^2 \sigma_{\min}^2(A)}{\alpha (m - N_{k-1}) \zeta \sigma_{\max}^2(A_{10})} \| x_k - x_\ast \|_2^2. \tag{11}
\]

Substituting (11) into (8), one has
\[
\| x_{k+1} - x_\ast \|_2^2 \leq \left[ 1 - \frac{\eta \| A_{10} \|_2^2 \sigma_{\min}^2(A)}{\alpha (m - N_{k-1}) \zeta \sigma_{\max}^2(A_{10}) \| A_{10} \|_F^2} \right] \| x_k - x_\ast \|_2^2. \tag{12}
\]

The recurrence relation with the inequalities (10) and (12) results in (6).

**Remark 2.** The convergence rate of BKMS, which is denoted by $\rho_{BKMS} = \| x_{k+1} - x_\ast \|_2^2 / \| x_k - x_\ast \|_2^2$, is subject to $0 < \rho_{BKMS} < 1$ under the conditions of Theorem 1. In fact,
\[
1 - \frac{\eta \| A_{10} \|_2^2 \sigma_{\min}^2(A)}{\alpha (m - N_{k-1}) \zeta \sigma_{\max}^2(A_{10})} \leq 1 - \frac{\eta}{\alpha (m - N_{k-1}) \zeta \sigma_{\max}^2(A_{10}) \max_{i \in [p]} \| A_{10} \|_F^2} \sigma_{\min}^2(A) \\
\leq 1 - \frac{\eta}{\alpha (m - N_{k-1}) \max_{i \in [p]} \| A_{10} \|_F^2} \sigma_{\min}^2(A) \\
\leq 1 - \frac{\eta}{\alpha (m - N_{k-1}) \sigma_{\min}^2(A)}.
\]
The second inequality holds because $\frac{\| A_{10} \|_2^2}{\sigma_{\max}^2(A_{10})} \geq 1$. Moreover, $0 < \eta \leq 1$, $0 < N_{k-1}, \alpha \leq m$ and $0 < \frac{\sigma_{\min}^2(A)}{\sigma_{\max}^2(A)} \leq 1$. Then it holds that
\[
0 < \frac{\eta}{\alpha (m - N_{k-1}) \| A \|_F^2} \sigma_{\min}^2(A) < 1.
\]
Thus $0 < \rho_{BKMS} < 1$, which implies that the BKMS has a linear convergence.

4. Numerical Examples

In this section, the numerical performances of the RK, Motzkin, GRK, RBK, SKM, CRBK, and BKMS algorithms are compared in different settings of the least-squares problem (1). The operating environment of all experiments is the MATLAB 2020b version on a
private computer with AMD Ryzen 5 4600H and 16 GB of memory. The relative solution error is denoted by
\[ RSE := \frac{\|x_k - x_*\|_2^2}{\|x_*\|_2^2}, \]
which is used to measure the efficiency of all methods.

The initial vector is set as \( x_0 = (0, 0, \cdots, 0)^T \) for BKMS and other methods. We terminate the iteration process of each method once \( RSE \leq 10^{-6} \). The concept of a random partition sampling in [13, 14] is sampling \( \tau_k \) randomly from row partition \( T = \{ \tau_1, \tau_2, \cdots, \tau_p \} \) at the \( k \)-th iteration, and for each \( i \), the subset \( \tau_i \) can be denoted as
\[ \tau_i = \left\{ \left\lfloor \left( i - 1 \right) \frac{m}{p} \right\rfloor + 1, \left\lfloor \left( i - 1 \right) \frac{m}{p} \right\rfloor + 2, \cdots, \left\lfloor \frac{m}{p} \right\rfloor \right\}, \quad i = 1, 2, \cdots, p. \]

The iterative least-squares solver CGLS in [14] is utilized to solve the underdetermined linear subsystem at each iteration since it requires fewer floating-point numbers between matrix–vector products than the MATLAB function “pinv”.

This section consists of three parts. In the first part (Examples 1–3), we implement MS clustering to visualize the row partition of the coefficient matrix \( A \) and verify that BKMS is more competitive than other methods to solve consistent linear systems (1) with different \( A \). In the second part (Example 4), we consider the effects of parameters \( \delta \) and \( \eta \) on the performance of BKMS. In the last part (Examples 5 and 6), we provide two practical applications of the BKMS method.

**Example 1.** This example implements the partitions of different system matrices by adopting the MS Algorithm 3. It is known that an ill-conditioned matrix means a high linear correlation between the rows of the matrix [16]. In the left graph of Figure 1a, we tested a dense matrix \( A \in \mathbb{R}^{1000 \times 500} \) with the condition number (cond(\( A \))) being 2.60, in the middle graph of Figure 1b, we tested a sparse matrix named “Cities” with cond(\( A \)) being 207.15, and in the right graph of Figure 1c, we tested an ill-conditioned sparse matrix named “relat6” with the cond(\( A \)) being \( 4.03 \times 10^{16} \). Figure 1b plots the number of rows contained in each block of the matrices above, respectively.

As shown in Figure 1a, the rows with similar correlation coefficients \( \mu_k \) are likely to be allocated in a block. This may lead to the considerable control index set \( \tau_k \) in the BKMS method being very large, which is verified in Figure 1b. To alleviate this problem, we continue to partition the current block, so it is reasonable to introduce an almost-maximum distance rule in BKMS.

![Figure 1. Cont.](image-url)
Figure 1. Clustering performance of the MS(δ) Algorithm 3 on matrices $A \in \mathbb{R}^{1000 \times 500}$ with $\delta = 0.45$ (left), “Cities” with $\delta = 0.45$ (middle), and “relat6” with $\delta = 0.3$ (right). (a) The cluster visualizations of system matrix $A \in \mathbb{R}^{1000 \times 500}$ (left), “Cities” (middle), and “relat6” (right). (b) The number of rows contained in each subblock of the matrix $A \in \mathbb{R}^{1000 \times 500}$ (left), “Cities” (middle), and “relat6” (right).

Example 2. This example considers a linear system (1) with a tall (or fat) coefficient matrix $A \in \mathbb{R}^{m \times n}$, i.e., $m > n$ (or $m < n$). Its elements are pseudorandom numbers from a standard normal distribution generated by the MATLAB function randn. Table 1 reports the iteration steps (IT) and CPU time in seconds (CPU(s)) of RK, Motzkin, GRK, SKM, RBK($p$), and GRBK($p$) with the same number of blocks $p = 8$ and BKMS ($\delta_{\text{opt}}$, $\eta$). The $\delta_{\text{opt}}$ refers to the numerical optimum parameter rather than a theoretical one. We will further discuss the $\delta_{\text{opt}}$ in Example 4. Figure 2 shows the numerical performance of all methods, in particular, we control RBK($p$), GRBK($p$), and BKMS(0.5, 0.1) to have the same number of blocks $p = 5$. Figure 2a shows the number of rows contained in each subblock of BKMS. Figure 2b,c depict the curves of RSE versus IT and RSE versus CPU(s) of Algorithm 4, applied to solve (1) with the coefficient matrix $A \in \mathbb{R}^{1000 \times 500}$. In addition, we default $\eta = 0.01$ for BKMS($\delta$) in all examples unless otherwise specified.

It is observed from Table 1 that RBK($p$), GRBK($p$), and BKMS($\delta$) are almost superior to RK, Motzkin, GRK, and SKM in terms of iteration steps and CPU time, thanks to the fact that block iteration takes multiple constraints. In addition, in block Kaczmarz methods, we find that BKMS($\delta$) is more competitive than RBK($p$) and GRBK($p$). For example, the CPU time speed-up of the BKMS($\delta$) over GRBK($p$) ranges from 1.206 to 1.610. Specifically, the value 1.610 is calculated by a ratio of 2.790 (CPU time consumed by GRBK(5) with the size of $A$ is 30,000 $\times$ 5000) over 1.733 (CPU time used by BKMS(0.45, 0.1) for the same size of $A$). Furthermore, from Figure 2, BKMS(0.5) converges in less iterations and CPU time than other methods.

Example 3. This example executes the RK, Motzkin, GRK, SKM, RBK, GRBK, and BKMS methods to solve the sparse linear Equation (1). Table 2 summarizes different sparse matrices with different properties, which are taken from the Florida sparse matrix collection in [27]. The density of a matrix $A$ means the percentage of nonzero elements of $A$. The IT and CPU times for all methods are reported in Table 2. Similar to Example 2, Figure 3a shows the number of rows contained in each subblock of BKMS(0.5). Figure 3b,c display the plots of RSE versus IT and RSE versus the CPU times of all methods, which are applied to solve a sparse linear system (1) named “WorldCities”.

Example 3. This example executes the RK, Motzkin, GRK, SKM, RBK, GRBK, and BKMS methods to solve the sparse linear Equation (1). Table 2 summarizes different sparse matrices with different properties, which are taken from the Florida sparse matrix collection in [27]. The density of a matrix $A$ means the percentage of nonzero elements of $A$. The IT and CPU times for all methods are reported in Table 2. Similar to Example 2, Figure 3a shows the number of rows contained in each subblock of BKMS(0.5). Figure 3b,c display the plots of RSE versus IT and RSE versus the CPU times of all methods, which are applied to solve a sparse linear system (1) named “WorldCities”.

Figure 1. Clustering performance of the MS(δ) Algorithm 3 on matrices $A \in \mathbb{R}^{1000 \times 500}$ with $\delta = 0.45$ (left), “Cities” with $\delta = 0.45$ (middle), and “relat6” with $\delta = 0.3$ (right). (a) The cluster visualizations of system matrix $A \in \mathbb{R}^{1000 \times 500}$ (left), “Cities” (middle), and “relat6” (right). (b) The number of rows contained in each subblock of the matrix $A \in \mathbb{R}^{1000 \times 500}$ (left), “Cities” (middle), and “relat6” (right).
Table 1. IT and CPU of all methods for the \( m \)-by-\( n \) consistent linear system.

<table>
<thead>
<tr>
<th>Name</th>
<th>( m \times n )</th>
<th>IT 0.45</th>
<th>IT 0.40</th>
<th>IT 0.50</th>
<th>IT 0.55</th>
<th>IT 0.45</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10,000 \times 5000</td>
<td>&gt;2 \times 10^5</td>
<td>&gt;2 \times 10^5</td>
<td>&gt;2 \times 10^5</td>
<td>&gt;2 \times 10^5</td>
<td>&gt;2 \times 10^5</td>
</tr>
<tr>
<td></td>
<td>15,000 \times 5000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>20,000 \times 5000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>25,000 \times 5000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>30,000 \times 5000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Name</th>
<th>( m \times n )</th>
<th>CPU(s) 0.45</th>
<th>CPU(s) 0.40</th>
<th>CPU(s) 0.50</th>
<th>CPU(s) 0.55</th>
<th>CPU(s) 0.45</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10,000 \times 5000</td>
<td>670.145</td>
<td>432.064</td>
<td>414.186</td>
<td>448.784</td>
<td>407.672</td>
</tr>
<tr>
<td></td>
<td>15,000 \times 5000</td>
<td>710.741</td>
<td>462.309</td>
<td>431.177</td>
<td>421.825</td>
<td>427.727</td>
</tr>
<tr>
<td></td>
<td>20,000 \times 5000</td>
<td>55.246</td>
<td>320.554</td>
<td>408.865</td>
<td>406.814</td>
<td>406.462</td>
</tr>
<tr>
<td></td>
<td>25,000 \times 5000</td>
<td>665.832</td>
<td>418.554</td>
<td>408.865</td>
<td>406.814</td>
<td>406.462</td>
</tr>
<tr>
<td></td>
<td>30,000 \times 5000</td>
<td>665.832</td>
<td>418.554</td>
<td>408.865</td>
<td>406.814</td>
<td>406.462</td>
</tr>
</tbody>
</table>

Figure 2. (a) The number of rows contained in each partition in BKMS with \( \delta = 0.5 \). The RSE versus IT (b) and RSE versus CPU (c) of BKMS compared with other methods for a consistent linear system with \( A \in \mathbb{R}^{1000\times500} \).
Table 2. IT and CPU of all methods for the sparse consistent linear system.

<table>
<thead>
<tr>
<th>Name</th>
<th>abtaha1</th>
<th>relat6</th>
<th>Model1</th>
<th>mk11-b2</th>
</tr>
</thead>
<tbody>
<tr>
<td>m × n</td>
<td>14,596</td>
<td>2340</td>
<td>362</td>
<td>6930</td>
</tr>
<tr>
<td>density</td>
<td>2.21%</td>
<td>2.11%</td>
<td>1.05%</td>
<td>6.52%</td>
</tr>
<tr>
<td>cond(A)</td>
<td>12.23</td>
<td>4.03 × 10^16</td>
<td>17.57</td>
<td>1.05 × 10^16</td>
</tr>
<tr>
<td>full rank</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>p</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td>δ</td>
<td>0.5</td>
<td>0.6</td>
<td>0.4</td>
<td>0.45</td>
</tr>
<tr>
<td>IT</td>
<td>30,654</td>
<td>9717</td>
<td>22,196</td>
<td>13,093</td>
</tr>
<tr>
<td>CPU(s)</td>
<td>61.699</td>
<td>1.245</td>
<td>1.622</td>
<td>42.695</td>
</tr>
<tr>
<td>Motzkin</td>
<td>IT</td>
<td>704</td>
<td>1501</td>
<td>3908</td>
</tr>
<tr>
<td></td>
<td>CPU(s)</td>
<td>1.232</td>
<td>0.134</td>
<td>0.233</td>
</tr>
<tr>
<td>GRK</td>
<td>IT</td>
<td>608</td>
<td>1467</td>
<td>3492</td>
</tr>
<tr>
<td></td>
<td>CPU(s)</td>
<td>0.585</td>
<td>0.114</td>
<td>0.172</td>
</tr>
<tr>
<td>SKM</td>
<td>IT</td>
<td>2074</td>
<td>1389</td>
<td>5433</td>
</tr>
<tr>
<td></td>
<td>CPU(s)</td>
<td>1.795</td>
<td>0.072</td>
<td>0.181</td>
</tr>
<tr>
<td>RBK(p)</td>
<td>IT</td>
<td>167</td>
<td>205</td>
<td>849</td>
</tr>
<tr>
<td></td>
<td>CPU(s)</td>
<td>0.681</td>
<td>0.095</td>
<td>0.21</td>
</tr>
<tr>
<td>GRBK(p)</td>
<td>IT</td>
<td>70</td>
<td>142</td>
<td>772</td>
</tr>
<tr>
<td></td>
<td>CPU(s)</td>
<td>0.248</td>
<td>0.045</td>
<td>0.133</td>
</tr>
<tr>
<td>BKMS(δ)</td>
<td>IT</td>
<td>36</td>
<td>127</td>
<td>711</td>
</tr>
<tr>
<td></td>
<td>CPU(s)</td>
<td>0.071</td>
<td>0.024</td>
<td>0.076</td>
</tr>
</tbody>
</table>

From Table 2, RBK(p), GRBK(p), and BKMS(δ) require fewer iteration steps and CPU times. Similar to the block Kaczmarz methods above, we believe that the BKMS method has better performance. For example, we calculate the CPU time speed-up of the BKMS(δ) over GRBK(p) ranges from 1.692 to 3.493 for solving the (1) with a different sparse matrix A listed in Table 2. From Figure 3, we can see that BKMS converges faster than other methods. By carefully comparing Figure 3c, it is not difficult to find that SKM and RBK have similar time efficiencies. The reason is that calculating the Moore–Penrose pseudoinverse of large blocks reduces the efficiency of RBK. So, the number of coefficient matrix partitions affects the convergence speed of RBK, GRBK, and BKMS. Based on this, in Example 4, we will analyze the sensitivity of the parameters (δ, η) in BKMS.

Example 4. This example investigates the effects of the parameters δ and η on the BKMS method. We use BKMS to solve system (1) with \( A \in \mathbb{R}^{5000 \times 1000} \) and the sparse system (1) named “abtaha1”, respectively. Tables 3 and 4 list the iteration steps and CPU time of BKMS(δ, η) once the RSE < 10^-6. The “- -” means abandoning the introduction of the almost maximum distance rule (i.e., removing steps 5 and 6 in Algorithm 4). In fact, we usually care more about the CPU time, so we use enumeration to select optimal parameters δ and η. This tuning method is also called the grid
search in machine learning. That is, find the corresponding different parameter values by locking the optimal value. For example, we select initialization seeds $\delta = 0.7$ and $\eta = 0.1$ for BKMS($\delta, \eta$) in Table 3.

<table>
<thead>
<tr>
<th>$\delta$</th>
<th>$\eta$</th>
<th>IT</th>
<th>CPU(s)</th>
<th>IT</th>
<th>CPU(s)</th>
<th>IT</th>
<th>CPU(s)</th>
<th>IT</th>
<th>CPU(s)</th>
<th>IT</th>
<th>CPU(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.9</td>
<td>0.469</td>
<td>0.207</td>
<td>0.453</td>
<td>0.323</td>
<td>0.323</td>
<td>0.440</td>
<td>0.080</td>
<td>0.080</td>
<td>0.808</td>
<td>0.176</td>
</tr>
<tr>
<td>0.3</td>
<td>0.7</td>
<td>0.552</td>
<td>0.253</td>
<td>0.555</td>
<td>0.388</td>
<td>0.388</td>
<td>0.440</td>
<td>0.741</td>
<td>0.741</td>
<td>1.716</td>
<td>1.673</td>
</tr>
<tr>
<td>0.5</td>
<td>0.5</td>
<td>0.234</td>
<td>0.191</td>
<td>0.267</td>
<td>0.126</td>
<td>0.126</td>
<td>0.388</td>
<td>0.741</td>
<td>0.741</td>
<td>0.323</td>
<td>0.176</td>
</tr>
<tr>
<td>0.7</td>
<td>0.3</td>
<td>0.469</td>
<td>0.207</td>
<td>0.453</td>
<td>0.323</td>
<td>0.323</td>
<td>0.440</td>
<td>0.080</td>
<td>0.080</td>
<td>0.808</td>
<td>0.176</td>
</tr>
<tr>
<td>0.9</td>
<td>0.1</td>
<td>0.234</td>
<td>0.191</td>
<td>0.267</td>
<td>0.126</td>
<td>0.126</td>
<td>0.388</td>
<td>0.741</td>
<td>0.741</td>
<td>0.323</td>
<td>0.176</td>
</tr>
</tbody>
</table>

Table 3. Sensitivity analysis of BKMS($\delta, \eta$) for system (1) with $A \in \mathbb{R}^{5000 \times 1000}$.

As depicted in Table 3, fix $\delta = 0.3$, for the different $\eta$; although the number of iteration steps increases, the CPU time shows a trend of first falling and then rising. For each column, this law is roughly the same. Finding a suitable $\delta$ is crucial. When $\delta$ becomes larger, the number of blocks becomes smaller, which requires computing the Moore–Penrose pseudoinverse of a large block to be time-consuming. On the contrary, when $\delta$ becomes smaller, the number of blocks increases. Considering an extreme case, when $A$ is row-normalized and there is only one row index in each block, BKMS is mathematically equivalent to the Motzkin method. A similar result also occurs in Table 4. The initialization seeds $\delta = 0.3$ and $\eta = 0.1$ can be selected for BKMS($\delta, \eta$) in Table 4.

Example 5 (Discrete ill-posed problem). This example uses the BKMS method to the solution of Phillips ill-posed problem in [28] and compares it with RK, Motzkin, GRK, SKM, RBK, and GRBK
methods. The linear systems (1) are obtained by the discretization of the Fredholm integral equation of the first kind,

\[ \int_{-6}^{6} K(s,t) \phi(t) \, dt = f(s) \]

on the square \([-6,6] \times [-6,6]\), where the kernel function is presented by \(K(s,t) = \phi(s-t)\) with

\[ \phi(x) = \begin{cases} 
1 + \cos\left( \frac{\pi}{3} x \right), & |x| < 3, \\
0, & |x| \geq 3, 
\end{cases} \]

and the right-hand side

\[ f(s) = (6 - |s|)(1 + \frac{1}{2} \cos\left( \frac{s\pi}{3} \right)) + \frac{9}{2\pi} \sin\left( \frac{|s|\pi}{3} \right), \]

which is often severely ill-conditioned. In fact, ordinary linear systems (2) may yield poor approximate solutions in this case, then one of the effective ways to overcome this issue is ridge regression, also called the Tikhonov regularization, namely

\[ \min_{x \in \mathbb{R}^n} \{ \|Ax - b\|^2 + \lambda \|x\|^2 \}, \]

which can be restated as

\[ \min_{x \in \mathbb{R}^n} \{ \|Ax - b\|^2 + \lambda \|x\|^2 \} = \min_{x \in \mathbb{R}^n} \left\| \frac{A}{\sqrt{\lambda}} x - \frac{b}{\sqrt{\lambda}} \right\|^2 = \min_{x \in \mathbb{R}^n} \|\tilde{A}x - \tilde{b}\|^2, \]

where \(\tilde{A} \in \mathbb{R}^{(n+n) \times n}\) and \(\lambda > 0\) is a regularization parameter. The measurement matrix \(A \in \mathbb{R}^{n \times n}\), exact solution \(x^* \in \mathbb{R}^n\), and \(b \in \mathbb{R}^n\) are all calculated by the MATLAB function phillips(n) in [28].

Set \(n = 1024\) and \(\lambda = 0.01\). Then the size of the square matrix \(A\) is \(1024 \times 1024\) and the condition number of \(A\) is \(2.9043 \times 10^{10}\) while \(\tilde{A}\) is 58.04. In Figure 4, the (a) reports the approximation solution calculated by RK, Motzkin, GRK, SKM, RBK(4), GRBK(4), and BKMS(0.4, 0.8) together with the exact solution for the phillips test problem when the maximum number of iteration steps is 100. The (b) shows the partition performance of BKMS(0.4, 0.8). Table 5 lists the RSE, CPU time, and iteration of BKMS(0.4, 0.8) compared with RK, Motzkin, GRK, SKM, RBK(4), and GRBK(4) for the “phillips” problem when \(RSE \leq 10^{-10}\).

![Graphical representation of the approximation solutions and partition performance](image-url)
Table 5. Numerical results for the “phillips” test problem.

<table>
<thead>
<tr>
<th>Method</th>
<th>RSE</th>
<th>CPU Time</th>
<th>Iterations</th>
</tr>
</thead>
<tbody>
<tr>
<td>RK</td>
<td>$5.360 \times 10^{-3}$</td>
<td>$1.220 \times 10^{-1}$</td>
<td>100</td>
</tr>
<tr>
<td>Motzkin</td>
<td>$1.934 \times 10^{-7}$</td>
<td>$6.385 \times 10^{-2}$</td>
<td>100</td>
</tr>
<tr>
<td>GRK</td>
<td>$3.382 \times 10^{-5}$</td>
<td>$6.473 \times 10^{-2}$</td>
<td>100</td>
</tr>
<tr>
<td>SKM</td>
<td>$1.826 \times 10^{-7}$</td>
<td>$5.904 \times 10^{-2}$</td>
<td>100</td>
</tr>
<tr>
<td>RBK(4)</td>
<td>$3.880 \times 10^{-13}$</td>
<td>$7.183 \times 10^{-2}$</td>
<td>2</td>
</tr>
<tr>
<td>GRBK(4)</td>
<td>$3.821 \times 10^{-13}$</td>
<td>$6.999 \times 10^{-2}$</td>
<td>2</td>
</tr>
<tr>
<td>BKMS(0.4, 0.8)</td>
<td>$3.447 \times 10^{-15}$</td>
<td>$5.320 \times 10^{-3}$</td>
<td>2</td>
</tr>
</tbody>
</table>

Figure 4 shows that the recovered solutions by RBK(4), GRBK(4), and BKMS(0.4, 0.8) are much closer to the exact solution than RK, Motzkin, GRK, and SKM do. It can be seen from Table 5 that BKMS obtains less RSE than other methods. In addition, BKMS needs less CPU time and iterations to reach the stop rule.

Example 6 (CT reconstruction problem). This example implements the reconstruction of a computer tomography (CT) image. This test problem can be implemented by the MATLAB function `paralleltomo(N, θ, p)` in the ART Tools package in [29]. We set $N = 50$, $θ = 0° : 1° : 178°$ and $p = 60$, then resulting in the size of $A$ is $10,740 \times 2500$, the exact solution $x^*$ (See the (a) of Figure 5) and the $b$ we obtained by $b = Ax^*$. The BKMS(0.25, 0.3) is applied to reconstruct $x^*$ from $b$ and compared with the RK, SKM, RBK(7), and GRBK(7) methods.

![Figure 5](image_url)
Table 6. Numerical results for restoring the “phantom” image.

<table>
<thead>
<tr>
<th>Method</th>
<th>RSE</th>
<th>CPU Time</th>
<th>Iterations</th>
</tr>
</thead>
<tbody>
<tr>
<td>RK</td>
<td>$6.994 \times 10^{-2}$</td>
<td>104.653</td>
<td>8000</td>
</tr>
<tr>
<td>Motzkin</td>
<td>$2.681 \times 10^{-3}$</td>
<td>87.496</td>
<td>8000</td>
</tr>
<tr>
<td>SKM</td>
<td>$2.645 \times 10^{-3}$</td>
<td>82.193</td>
<td>8000</td>
</tr>
<tr>
<td>RBK(7)</td>
<td>$9.935 \times 10^{-7}$</td>
<td>136.756</td>
<td>221</td>
</tr>
<tr>
<td>GRBK(7)</td>
<td>$9.818 \times 10^{-7}$</td>
<td>72.386</td>
<td>118</td>
</tr>
<tr>
<td>BKMS(0.25, 0.3)</td>
<td>$9.740 \times 10^{-7}$</td>
<td>32.361</td>
<td>104</td>
</tr>
</tbody>
</table>

5. Conclusions

This work proposes a block Kaczmarz–Motzkin method with mean shift clustering as an efficient extension of the Kaczmarz algorithm to solve large consistent linear equations. The row correlation coefficient of the system matrix is utilized as the clustering criterion and this method does not need to specify the number of partitions in advance. The convergence theory of the BKMS method is established when the linear system is consistent. Several numerical examples demonstrate the effectiveness of the current method in this paper.

Although the effects of the clustering bandwidth $\delta$ and the parameter $\eta$ are investigated in the numerical example, the theoretical analysis of $\delta$ affecting the performance of BKMS needs further discussion. According to the properties of the linear system, choosing a more appropriate clustering criterion may improve the algorithm performance; we will use these problems as part of our future work.

Author Contributions: Y.L.: software, validation, formal analysis, investigation, writing—original draft. T.L.: conceptualization, validation, formal analysis, writing—review and editing, supervision, funding acquisition. F.Y.: validation, formal analysis, writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Project of the Department of Science and Technology of Sichuan Province (No. 2021ZYD0005), the Opening Project of the Key Laboratory of Higher Education of Sichuan Province (No. 2020WZJ01), the Scientific Project of SUSE (No. 2020RC24), and the Postgraduate Research Fund (No. y2021101).

Institutional Review Board Statement: Not applicable

Informed Consent Statement: Not applicable

Data Availability Statement: The data that support the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest: The authors have no financial or proprietary interests in any material discussed in this article.

References
2. Byrne, C. A unified treatment of some iterative algorithms in signal processing and image reconstruction. Inverse Probl. 2004, 20, 103–120. [CrossRef]