Greedy double subspaces coordinate descent method via orthogonalization

Li-Li Jin, Hou-Biao Li

School of Mathematical Sciences, University of Electronic Science and Technology of China, Chengdu, 611731, P. R. China

Abstract

The coordinate descent method is an effective iterative method for solving large linear least-squares problems. In this paper, for the highly coherent columns case, we construct an effective coordinate descent method which iteratively projects the estimate onto a solution space formed by two greedily selected hyperplanes via Gram-Schmidt orthogonalization. Our methods may be regarded as a simple block version of coordinate descent method which involves two active columns. The convergence analysis of this method is provided and numerical simulations also confirm the effectiveness for matrices with highly coherent columns.

Keywords: coordinate descent method, subspace, greedy rule, least-squares problem

1. Introduction

In practical applications, such as regression analysis and data fitting, we usually find \( x \in \mathbb{R}^n \) to minimize the function \( f(x) = \| b - Ax \|^2 \), namely,

\[
x_* := \arg \min_{x \in \mathbb{R}^n} \| b - Ax \|^2,
\]

where \( A \in \mathbb{R}^{m \times n} \) is a full column rank matrix with \( m > n \), \( b \in \mathbb{R}^m \), and \( x \in \mathbb{R}^n \) is the \( n \)-dimensional unknown vector. \( \| \cdot \| \) indicates the Euclidean norm of either a vector or a matrix throughout this paper. Iterative methods such as coordinate descent method \( [1, 2] \), which is also called the Gauss-Seidel method, are considered for solving linear least squares problem (1.1). The coordinate descent method, starting from an initial guess \( x_0 \), can be formulated as

\[
x_{k+1} = x_k + \frac{A^T_{(j_k)} (b - Ax_k)}{\| A_{(j_k)} \|^2} e_{j_k}, \quad k = 0, 1, 2, \ldots
\]

where \( A_{(j)} \) is the \( j \)th column of \( A \) and \( (\cdot)^T \) represents the transpose of either a vector or a matrix. Hence \( x_{k+1} \) is obtained by projecting the current iterate \( x_k \) to the hyperplane \( H_{j_k} = \{ x \mid A_{(j_k)}^T Ax = A_{(j_k)}^T b \} \). It is crucial to select a distinguished descent direction \( e_{j_k} \) in the \( k \)th iteration, that is, to select the column index \( j_k \). When \( j_k = (k \mod n) + 1 \), the coordinate descent method is the classical Gauss-Seidel method.
directly applied to the normal equation $A^T Ax = A^T b$, where $A^T A$ is symmetric positive definite. To speed up the convergence of the coordinate descent method, Strohmer and Vershynin [5] proposed randomized Kaczmarz method with expected exponential rate of convergence, and Leventhal and Lewis [6] constructed the randomized coordinate descent method by randomly choosing a coordinate direction as a search direction based on an appropriate probability distribution. Soon after that, Ramdas, Needell and Ma [7] provided an extended version of randomized Gauss-Seidel which converges linearly to the least norm solution in the under-determined case. In addition, Nutini et al. [8] argued that in these contexts greedy selection rules gave faster convergence rates than random selection rules, and analyzed Gauss-Southwell(GS) rule and Gauss-Southwell-Lipschitz (GSL) rule. Some randomized and greedy versions of the coordinate descent method can be found in [9–12]. In 2021, Liu, Jiang and Gu [13] proposed a two step Gauss-Seidel (2SGS) algorithm involved two active columns by the maximum residual rule. Motivated by Needell and Ward [14], here we consider the coordinate descent method in case of the matrix $A$ with highly coherent columns, which means that the angle between these hyperplanes $H_j = \{ x \mid A^T_{(j)}Ax = A^T_{(j)}b \} \ (j = 1, \ldots, n)$ is very small. By greedily selecting two linearly independent columns and introducing the auxiliary descent direction vector, we propose the greedy double subspaces coordinate descent method which adopts the Gram-Schmidt Orthogonalization [15] process to project the current iteration onto the corresponding solution space of two selected hyperplane.

This paper is organized as follows. In section 2, we present necessary notations. In section 3, we describe our new method and its convergence theory. In section 4, some numerical examples are provided to illustrate the effectiveness of our new methods for matrices with highly coherent columns.

2. Notation and preliminaries

For a vector $z \in \mathbb{R}^n$, $z^{(i)}$ denotes its $i$th entry. In addition, for a matrix $A$, $A^{(j)}$ represents its $j$th column. $\|A\|$, $\|A\|_{\infty}$ are used to indicate Euclidean norm and the infinite norm of $A$ respectively. In what follows, we denote the transpose of the matrix $A$ by $A^T$ and $\|x\|_G = \sqrt{x^T Gx}$ when $G$ is positive definite. $e_j$ is used to represent the column vector with a one in position $j$ and zero in all other positions. We use $x^* = A^T b$ to represent the unique least squares solution of (1.1). For $x, y \in \mathbb{R}^n$, $\langle x, y \rangle = \sum_{i=1}^{n} x^{(i)} y^{(i)}$ remarks standard Euclidean inner product of two vectors.

3. The greedy double subspaces coordinate descent method

We will assume throughout that the matrix $A$ of columns are normalized, meaning that its columns has unit Euclidean norm. In 2015, Nutini et al. [8] analyze Gauss-Southwell(GS) rule

$$j_k = \arg \max_{1 \leq j \leq n} |\nabla_j f(x_k)|$$
and Gauss-Southwell-Lipschitz (GSL) rule
\[ j_k = \arg \max_{1 \leq j \leq n} \frac{\left| \nabla_j f(x_k) \right|}{\sqrt{L_j}}. \]
If \( f(x) = \|b - Ax\|^2 \), and the residual vector of the normal equation \( s_k = A^T(b - Ax_k) \) is calculated in the \( k \)th iteration, the index \( j_k \) selected by GS rule satisfies
\[ j_k = \arg \max_{1 \leq j \leq n} \left| s_k^{(j)} \right| \]
and the index \( j_k \) selected by GSL rule satisfies
\[ j_k = \arg \max_{1 \leq j \leq n} \left\{ \left| s_k^{(j)} \right| \right\}. \]

Note that GS rule is the same as GSL rule when the columns of the matrix \( A \) are normalized. This greedy coordinate descent (GCD) method with GSL rule is described by Algorithm 1.

Algorithm 1 The GCD Algorithm [8]
\begin{itemize}
  \item \textbf{Input:} \( A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^m, \ell, \) initial estimate \( x_0 \).
  \item \textbf{Output:} \( x_\ell \).
  \item 1: \textbf{for} \( k = 0, 1, 2, \ldots, \ell - 1 \) \textbf{do}
  \item 2: \hspace{1em} Compute \( s_k = A^T(b - Ax_k) \)
  \item 3: \hspace{1em} Select \( j_k = \arg \max_{1 \leq j \leq n} \left| s_k^{(j)} \right| \)
  \item 4: \hspace{1em} Update \( x_{k+1} = x_k + s_k^{(j_k)} e_{j_k} \)
  \item 5: \hspace{1em} end for
\end{itemize}

In addition, Liu, Jiang and Gu [13] construct the two step Gauss-Seidel (2SGS) algorithm by GS rule as listed Algorithm 2.

Algorithm 2 The 2SGS Algorithm [13]
\begin{itemize}
  \item \textbf{Input:} \( A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^m, \ell, \) initial estimate \( x_0 \).
  \item \textbf{Output:} \( x_\ell \).
  \item 1: \textbf{for} \( k = 0, 1, 2, \ldots, \ell - 1 \) \textbf{do}
  \item 2: \hspace{1em} Compute \( s_k = A^T(b - Ax_k) \)
  \item 3: \hspace{1em} Select \( j_{k_1} = \arg \max_{1 \leq j \leq n} \left| s_k^{(j)} \right| ; \hspace{1em} j_{k_2} = \arg \max_{\substack{j \leq n \\mid j \neq j_{k_1}}} \left\{ \left| s_k^{(j)} \right| \right\} \)
  \item 4: \hspace{1em} Update \( x_{k+1} = x_k + s_k^{(j_{k_1})} e_{j_{k_1}} + s_k^{(j_{k_2})} e_{j_{k_2}} \)
  \item 5: \hspace{1em} end for
\end{itemize}

Considering the columns of the matrix \( A \) are highly coherent, which means that the angle between these hyperplanes corresponding to the normal equation is very small, the convergence speed of some versions
of coordinate descent with one working column will become slow. However, if the hyperplane that is the farthest and closest to the current iteration point can be selected in kth iteration,

\[ j_{k_1} = \arg \max_{1 \leq j \leq n} \left\{ \left| A_{(j)}^T (b - Ax) \right| \right\} \quad \text{and} \quad j_{k_2} = \arg \min_{1 \leq j \leq n} \left\{ \left| A_{(j)}^T (b - Ax) \right| \right\}, \]

which the angle between the two planes \( H_{j_{k_1}} \) and \( H_{j_{k_2}} \) is relatively large, it can expect to speed up the convergence of this method. Given \( x_0 \), we first initialize \( x_0 \) by one-step iterative GCD algorithm to get an approximate solution \( x_1 \), while reserve the selected column index marked as \( j_0 \). In the first iteration, since \( A_{(j_0)}^T (b - Ax_1) = 0 \), which means that the hyperplane \( H_{j_0} \) is the closest to \( x_1 \). We rewrite \( j_0 \) as the second index of the first iteration \( j_{i_1} \). In addition, we can use the GSL rule to select the hyperplane farthest from the current iteration point \( x_1 \), and remark the corresponding column index as the first index of the current iteration \( j_{i_1} \), and preserve it for next iteration. Here we construct the auxiliary descent direction vector \( w_1 \), and \( x_2 \) can be obtained by projecting \( x_1 \) onto the solution space formed by these two planes \( H_{j_{i_1}} \) and \( H_{j_{i_2}} \) after some simple algebraic calculations. In the second iteration, since \( A_{(j_{i_1})}^T (b - Ax_2) = 0 \), which means that \( x_2 \) falls on the hyperplane \( H_{j_{i_1}} \). We rewrite \( j_{i_1} \) as the second index of the second iteration \( j_{i_2} \). In addition, we select the hyperplane farthest from the current iteration point \( x_2 \) by GSL rule, and denote the corresponding column index as the first index of the second iteration \( j_{i_2} \), and keep it for next iteration. Analogously, we introduce the auxiliary descent direction vector \( w_2 \), and get \( x_3 \) by projecting \( x_2 \) onto the intersection of two hyperplane \( H_{j_{i_1}} \) and \( H_{j_{i_2}} \). Sequentially, we gain the sequence \( \{x_k\}_{k=0}^\infty \) in the same way. The specific algorithm is detailedly exhibited in the Algorithm 3.

**Algorithm 3** The greedy double subspaces coordinate descent (GDSCD) method

**Input:** \( A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^m, \ell \), initial estimate \( x_0 \).

**Output:** \( x_\ell \).

1. Compute \( s_0 = A^T (b - Ax_0) \), \( j_0 = \arg \max_{1 \leq j \leq n} \left\{ \left| s_0^{(j)} \right| \right\} \)
2. Update \( x_1 = x_0 + s_0^{(j_0)} e_{j_0} \)
3. for \( k = 1, 2, \ldots, \ell - 1 \) do
4. Select \( j_{k_1} = \arg \max_{1 \leq j \leq n} \left\{ \left| s_k^{(j)} \right| \right\} \), \( j_{k_2} = \arg \max_{1 \leq j \leq n} \left\{ \left| s_{k-1}^{(j)} \right| \right\} \)
5. Set \( \tilde{\mu}_k = \langle A_{(j_{k_1})}, A_{(j_{k_2})} \rangle \), \( \tilde{\gamma}_k = x_k + s_k^{(j_{k_1})} e_{j_{k_1}} \)
6. Set \( \tilde{\tilde{\gamma}}_k = \frac{A_{(j_{k_2})} - \tilde{\mu}_k A_{(j_{k_1})}}{\sqrt{1 - \tilde{\mu}_k^2}}, \beta_k = \frac{(A_{(j_{k_2})} - \tilde{\mu}_k A_{(j_{k_1})})^T b}{\sqrt{1 - \tilde{\mu}_k^2}}, w_k = \frac{e_{j_{k_2}} - \tilde{\mu}_k e_{j_{k_1}}}{\sqrt{1 - \tilde{\mu}_k^2}} \)
7. Update \( x_{k+1} = \tilde{\gamma}_k + \langle \tilde{\gamma}_k - \tilde{\tilde{\gamma}}_k A \tilde{\gamma}_k \rangle w_k \)
8. end for

For the convergence theory of the GDSCD algorithm, we have the following theorem.

**Theorem 1.** The iteration sequence \( \{x_k\}_{k=0}^\infty \), generated by the GDSCD method starting from any initial
guess \( x_0 \in \mathbb{R}^n \), linearly converges to the unique least-squares solution \( x_* = A^T b \) and satisfies
\[
\| x_* - x_1 \|^2_{A^T A} \leq (1 - \sigma(A^T, \infty)^2) \| x_* - x_0 \|^2_{A^T A},
\]
and
\[
\| x_* - x_{k+1} \|^2_{A^T A} \leq \left(1 - \frac{1}{1 - \mu_k^2} \sigma(A^T, \infty)^2\right) \| x_* - x_k \|^2_{A^T A}, \quad k = 1, 2, \ldots,
\]
where \( \tilde{\mu}_k = \left\langle A_{\{j_k\}}, A_{\{j_k\}} \right\rangle \), and \( \sigma(A^T, \infty) \) is the Hoffman-like constant as defined in [16].

**Proof.** Denote \( r_k = b - Ax_k, s_k = A^T r_k, e_k = x^* - x_k \). From Algorithm 3, for \( k \geq 1 \), we have
\[
x_{k+1} = \tilde{y}_k + (\tilde{\beta}_k - \tilde{v}_k^T A\tilde{y}_k)w_k
= x_k + s_k(j_k) e_{j_k} + (\tilde{\beta}_k - \tilde{v}_k^T A\tilde{y}_k)w_k.
\]
Hence, we can get
\[
A(x^* - x_{k+1}) = A(x^* - x_k) - s_k(j_k) A_{\{j_k\}} - (\tilde{\beta}_k - \tilde{v}_k^T A\tilde{y}_k) \tilde{v}_k.
\]
Since \( \tilde{v}_k \) is perpendicular to \( A_{\{j_k\}} \) for any \( k \geq 1 \), it holds that
\[
\| A e_{k+1} \|^2 = \| A e_k \|^2 - \left( s_k(j_k) \right)^2 + (\tilde{\beta}_k - \tilde{v}_k^T A\tilde{y}_k)^2 - 2(\tilde{\beta}_k - \tilde{v}_k^T A\tilde{y}_k) (\tilde{v}_k, A e_k).
\]
we can easily calculate
\[
\tilde{\beta}_k - \tilde{v}_k^T A\tilde{y}_k = \tilde{\beta}_k - \tilde{v}_k^T A\tilde{x}_k
= \frac{A_{\{j_k\}}^T b - A_{\{j_k\}}^T A x_k - \tilde{\mu}_k (A_{\{j_k\}}^T b - A_{\{j_k\}}^T A x_k)}{\sqrt{1 - \mu_k^2}}
= \frac{A_{\{j_k\}}^T b - \tilde{\mu}_k A_{\{j_k\}}^T A (x^* - x_k)}{\sqrt{1 - \mu_k^2}}
= \tilde{v}_k^T A e_k.
\]  
For \( k \geq 2 \),
\[
A_{\{j_k-1\}}^T A x_k = A_{\{j_k-1\}}^T A \left( x_{k-1} + s_{k-1}^{(j_k-1)} e_{j_k-1} + (\tilde{\beta}_{k-1} - \tilde{v}_{k-1}^T A\tilde{y}_{k-1}) w_{k-1} \right)
= A_{\{j_k-1\}}^T A x_{k-1} + s_{k-1}^{(j_k-1)} A_{\{j_k-1\}}^T A_{\{j_k-1\}} + (\tilde{\beta}_{k-1} - \tilde{v}_{k-1}^T A\tilde{y}_{k-1}) A_{\{j_k-1\}}^T \tilde{v}_{k-1}
= A_{\{j_k-1\}}^T A x_{k-1} + s_{k-1}^{(j_k-1)}
= A_{\{j_k-1\}}^T b,
\]
for \( k = 1 \),
\[
A_{\{j_1\}}^T A x_1 = A_{\{j_1\}}^T A (x_0 + s_0^{(j_1)} e_{j_1})
= A_{\{j_1\}}^T A x_0 + s_0^{(j_1)} A_{\{j_1\}}^T A_{\{j_1\}}
= A_{\{j_1\}}^T A x_0 + A_{\{j_1\}}^T (b - A x_0)
= A_{\{j_1\}}^T b.
\]
Thus, combining (3.1) and (3.3), we can further obtain

\[ \tilde{v}_k^T A e_k = \frac{A^T_{(j_{k+1})} - \tilde{\mu}_k A^T_{(j_{k+1})}}{\sqrt{1 - \tilde{\mu}_k^2}} A e_k \]
\[ = \frac{s_{k} - \tilde{\mu}_k s_{k}}{\sqrt{1 - \tilde{\mu}_k^2}} A e_k \]
\[ = \frac{s_{k} - \tilde{\mu}_k s_{k}}{\sqrt{1 - \tilde{\mu}_k^2}} A e_k \]
\[ = \frac{s_{k} - \tilde{\mu}_k s_{k}}{\sqrt{1 - \tilde{\mu}_k^2}} A e_k \]

Thus, combining (3.1) and (3.3), we can further obtain

\[ \|A e_{k+1}\|^2 = \|A e_k\|^2 - \left( s_{k} \right)^2 - 2(\tilde{\beta}_k - \tilde{v}_k^T A \tilde{y}_k)(\tilde{v}_k, A e_k) \]
\[ = \|A e_k\|^2 - \|A e_k\|^2 \frac{1}{1 - \tilde{\mu}_k^2} s_{k}^2 \]
\[ = \|A e_k\|^2 - \frac{1}{1 - \tilde{\mu}_k^2} \max \left\{ s_{k}^2 \right\} \]
\[ = \|A e_k\|^2 - \frac{1}{1 - \tilde{\mu}_k^2} \|A^T A (x^* - x_k)\|_\Sigma^2 \]
\[ \leq \|A e_k\|^2 - \frac{1}{1 - \tilde{\mu}_k^2} \sigma(A^T, \infty)^2 \|A e_k\|^2 \]
\[ \leq (1 - \frac{1}{1 - \tilde{\mu}_k^2} \sigma(A^T, \infty)^2) \|A e_k\|^2, \] (3.4)

where the equation (3.4) is achieved with the use of Hoffman-like constant \( \sigma(A^T, \infty) \) in [16].

In addition, for \( k = 0 \), we have

\[ \|A(x^* - x_0)\|^2 = \|A(x^* - x_0)\|^2 - \left( s_{0} \right)^2 \]
\[ = \|A(x^* - x_0)\|^2 - \|A^T A (x^* - x_0)\|^2 \]
\[ = \|A(x^* - x_0)\|^2 - \frac{1}{1 - \tilde{\mu}_k^2} \|A^T A (x^* - x_0)\|^2 \]
\[ \leq (1 - \sigma(A^T, \infty)^2) \|A(x^* - x_0)\|^2, \] (3.6)

where the equation (3.6) is obtained by making use of the Hoffman-like constant \( \sigma(A^T, \infty) \) in [16]. Therefore, we can get the estimate by combining the equation (3.5) and (3.6).

4. Numerical experiments

In this section we perform several experiments to compare the convergence rate of the GCD, 2SGS and GDSCD methods for some column normalized matrices in terms of the number of iteration steps (denoted
as "IT") and the computing time in seconds (denoted as "CPU"). Note that the IT and CPU listed in our numerical results denote the arithmetical averages of the required iteration numbers and the CPU times with respect to 30 times repeated runs of the corresponding methods. To be precise, we define the pairwise coherence of a column standardized matrix as follows by reference to the notations in [14]

\[ \delta = \delta(A) = \min_{i \neq j} |\langle A(i), A(j) \rangle| \quad \text{and} \quad \Delta = \Delta(A) = \max_{i \neq j} |\langle A(i), A(j) \rangle|, \]

In addition, we construct various types of \(500 \times 100\) and \(5000 \times 500\) matrices \(A\) by MATLAB function \textit{unifrnd} which generate continuous uniformly distributed random variables on the interval \([c,1]\). Thus matrices with different pairwise coherence can be constructed by changing the value of \(c\). Moreover, the corresponding linear system (1.1) could be either consistent or inconsistent. The solution vector \(x\) is generated by using MATLAB function \textit{randn}. As for \(b\), we take \(b = Ax\) when the systems is consistent, and \(b = Ax + b_0\) when the linear system is inconsistent, where \(b_0\) is a nonzero vector belonging to the null space of \(A^T\), and \textit{null}(\(A^T\)) is generated by using MATLAB function \textit{null}.

| Table 1 | IT and CPU for various \(500 \times 100\) matrices when the linear system is consistent. |
|---------|---------------------------------------------------------------|
|         | \(500 \times 100\) | \(A_1\) | \(A_2\) | \(A_3\) | \(A_4\) | \(A_5\) | \(A_6\) |
| \(c\)  | -0.8 | -0.1 | 0.8 | 0.85 | 0.9 | 0.95 |
| \(\delta\) | 4.0558e-06 | 0.6059 | 0.9951 | 0.9975 | 0.9989 | 0.9997 |
| \(\Delta\) | 0.1890 | 0.7318 | 0.9966 | 0.9982 | 0.9993 | 0.9998 |
| rank(\(A\)) | 100 | 100 | 100 | 100 | 100 | 100 |
| GCD IT | 494 | 1311 | 92067 | 197026 | -- | -- |
| CPU | 0.0270 | 0.0723 | 5.7447 | 10.8633 | -- | -- |
| 2SGS IT | 252 | 237 | 2262 | 4243 | 8768 | 40647 |
| CPU | 0.0170 | 0.0151 | 0.1357 | 0.2699 | 0.5342 | 2.3530 |
| GDSCD IT | 433 | 365 | 383 | 385 | 377 | 389 |
| CPU | 0.0393 | 0.0311 | 0.0332 | 0.0300 | 0.0309 | 0.0333 |
Table 2 IT and CPU for various 5000×500 matrices when the linear system is consistent.

| 5000 × 500 | $B_1$ | $B_2$ | $B_3$ | $B_4$ | $B_5$ | $B_6$ |
|------------|------|------|------|------|------|------|
| $c$        | -0.8 | -0.1 | 0.8  | 0.85 | 0.9  | 0.95 |
| $\delta$   | 2.6205e-06 | 0.6385 | 0.9956 | 0.9976 | 0.9991 | 0.9998 |
| $\Delta$   | 0.0946 | 0.6943 | 0.9962 | 0.9980 | 0.9990 | 0.9998 |
| rank($A$)  | 500  | 500  | 500  | 500  | 500  | 500  |
| GCD IT     | 1755 | 5001 | ---  | ---  | ---  | ---  |
| CPU        | 2.6047 | 8.3252 | ---  | ---  | ---  | ---  |
| 2SGS IT    | 803  | 800  | 2682 | 3758 | 11059 | 39232 |
| CPU        | 1.2081 | 1.4180 | 4.0388 | 5.6522 | 16.6704 | 40.4757 |
| GDSCD IT   | 1553 | 1731 | 2033 | 2043 | 2040 | 2050 |
| CPU        | 2.5811 | 3.2864 | 3.4455 | 3.2922 | 3.4014 | 3.3556 |

In all implementations, the initial point $x_0$ is set to be a zero vector and terminated once the relative solution error (RSE), defined by $\text{RSE} = \frac{\|x_k - x_*\|^2}{\|x_*\|^2}$ at the current iterate $x_k$, satisfies $\text{RSE} \leq 10^{-6}$, or the number of iteration steps exceeds 200,000. The latter is given a label “−−” in the numerical tables. All experiments are performed on a PC with Intel(R) Core(TM) i5-8250U CPU @ 1.60GHz 8.00GB using MATLAB R2016a.

For the consistent systems, Tables 1-2 show the number of iteration steps and the computing time of various 500 × 100 and 5000 × 500 matrices for the three methods. Here the matrix $A_5, A_6, B_5, B_6$ have highly coherent columns with $\delta \approx \Delta$ or $\delta = \Delta$. Note that when the value of $c$ is closer to 1, both $\delta$ and $\Delta$ are closer to 1, which means that the columns of these matrices are more coherent. Admittedly, the convergence rate of the 2SGS method is faster than both the GCD method and the GDSCD method in the

![Figure 1](image-url)  

**Figure 1** $\log_{10}(\text{RSE})$ versus CPU for 500 × 100 (left), 5000 × 500 (right) matrices when $c$ is -0.8.
case where $\delta$ and $\Delta$ are small (see Table 1 of $A_1$ and $A_2$, Table 2 of $B_1$ and $B_2$). However, it is true that GDSCD method outperforms the other two algorithms in both time and number of iteration steps with increasing the coherence of column of matrices (see Table 1 of $A_5$ and $A_6$, Table 2 of $B_5$ and $B_6$). What is better is that the GDSCD method is at least 5 times faster than the other two methods for matrices highly coherent columns (see Table 1 of $A_6$, Table 2 of $B_6$). In addition, Figures 1 and 2 depict the curves of the $\log_{10}(\text{RSE})$ versus CPU time of when the linear system is consistent with $c = -0.8$ and $c = 0.9$, which further shows the effectiveness of the proposed new method for highly column-coherent matrices.

Table 3 IT and CPU for various $500 \times 100$ matrices when the linear system is inconsistent.

| $500 \times 100$ | $A_1$ | $A_2$ | $A_3$ | $A_4$ | $A_5$ | $A_6$ |
|------------------|------|------|------|------|------|------|
| $c$              | -0.8 | -0.1 | 0.8  | 0.85 | 0.9  | 0.95 |
| $\delta$         | 4.0558e-06 | 0.6059 | 0.9951 | 0.9975 | 0.9989 | 0.9997 |
| $\Delta$         | 0.1890 | 0.7318 | 0.9966 | 0.9982 | 0.9993 | 0.9998 |
| rank(A)          | 100  | 100  | 100  | 100  | 100  | 100  |
| GCD IT           | 465  | 1261 | 105668 | 188781 | --   | --   |
| CPU              | 0.0281 | 0.0705 | 5.8241 | 10.3324 | --   | --   |
| 2SGS IT          | 243  | 219  | 2279 | 3518 | 8378 | 38576 |
| CPU              | 0.0167 | 0.0147 | 0.1469 | 0.2153 | 0.5080 | 2.2868 |
| GDSCD IT         | 443  | 360  | 410  | 384  | 400  | 388  |
| CPU              | 0.0338 | 0.0291 | 0.0329 | 0.0332 | 0.0327 | 0.0322 |

Similarly, Tables 3 and Tables 4 demonstrate the same conclusion for the inconsistent systems. We observe that when the column coherence of the matrix is not particularly large, the convergence efficiency of the 2SGS method is better than the other two methods (see Table 3 of $A_1$, $A_2$, Table 4 of $B_1$, $B_2$).

Figure 2 $\log_{10}(\text{RSE})$ versus CPU for $500 \times 100$ (left), $5000 \times 500$ (right) matrices when $c$ is 0.9.
Table 4 IT and CPU for various $500 \times 100$ matrices when the linear system is inconsistent.

| $5000 \times 500$ | $B_1$ | $B_2$ | $B_3$ | $B_4$ | $B_5$ | $B_6$ |
|-------------------|-------|-------|-------|-------|-------|-------|
| $c$               | -0.8  | -0.1  | 0.8   | 0.85  | 0.9   | 0.95  |
| $\delta$          | 2.6205e-06 | 0.6385 | 0.9956 | 0.9976 | 0.9991 | 0.9998 |
| $\Delta$          | 0.0946 | 0.6943 | 0.9962 | 0.9980 | 0.9990 | 0.9998 |
| rank(A)           | 500   | 500   | 500   | 500   | 500   | 500   |
| GCD IT            | 1430  | 4876  |       |       |       |       |
| CPU               | 2.5931| 8.2934|       |       |       |       |
| 2SGS IT           | 789   | 787   | 2676  | 3821  | 11068 | 39247 |
| CPU               | 1.1994| 1.3290| 4.0472| 5.8231| 16.6803| 41.0543|
| GDSCD IT          | 1550  | 1723  | 2042  | 2109  | 2036  | 2061  |
| CPU               | 2.4992| 3.1263| 3.4337| 3.1879| 3.4029| 3.3421|

Figure 3 $\log_{10}(RSE)$ versus CPU for $500 \times 100$ (left), $5000 \times 500$ (right) matrices when $c$ is -0.8.

Figure 4 $\log_{10}(RSE)$ versus CPU for $500 \times 100$ (left), $5000 \times 500$ (right) matrices when $c$ is 0.9.
However, as the value of $c$ getting closer to 1, that is, the coherence of columns of the constructed matrix increasing continuously, we see again that the GDSCD method significantly outperforms the GCD method and the 2SGS method for both iteration counts and CPU times (see Table 3 of $A_5$, $A_6$, Table 4 of $B_5$, $B_6$). In addition, we plot the curves of $\log_{10}(\text{RSE})$ versus CPU time of for $500 \times 100$ and $5000 \times 500$ matrices with $c = -0.8$ and $c = 0.9$ when the linear system inconsistent.

5. Conclusions

This paper introduces a greedy block version of the coordinate descent method for the linear least-squares problem. The main idea of our method is to greedily select two linearly independent columns at first, and then projects the estimate onto a solution space of the two selected hyperplanes by Gram-Schmidt orthogonalization process. The theoretical analysis and numerical results show that our proposed method is efficient for matrices with highly coherent columns.

Acknowledgements. The authors are very much indebted to the referees for their constructive comments and valuable suggestions. The authors are partially supported by National Natural Science Foundation of China (11101071, 11271001, 51175443).

References

[1] C.L. Byrne, Applied iterative methods, A. K. Peters, Wellesley, MA, 2008.
[2] Y. Saad, Iterative Methods for Sparse Linear Systems, SIAM, Philadelphia, 2003.
[3] W. Hackbusch, Iterative solution of large sparse systems of equations, Springer, New York, 1994.
[4] S. Wright, Coordinate descent algorithms, Mathematical Programming, 151(2015) 3-34.
[5] T. Strohmer, V. Vershynin, A randomized Kaczmarz algorithm with exponential convergence, Journal of Fourier Analysis and Applications, 15(2)(2009) 262-278.
[6] D. Leventhal, A.S. Lewis, Randomized methods for linear constraints: convergence rates and conditioning, Mathematics of Operations Research, 35(3)(2010) 641-654.
[7] A. Ma, D. Needell, A. Ramdas, Convergence properties of the randomized extended Gauss-Seidel and Kaczmarz methods, SIAM Journal on Matrix Analysis and Applications, 36(4)(2015) 1590-1604.
[8] J. Nutini, M. Schmidt, I.H. Laradji, M. Friedlander, H. Koepke, Coordinate descent converges faster with the Gauss-Southwell rule than random selection, International Conference on Machine Learning, 43(2015) 1632-1641.
[9] Z.Q. Luo, P. Tseng, On the convergence of the coordinate descent method for convex differentiable minimization, Journal of Optimization Theory and Applications, 72(1992) 7-35.
[10] Z.Z. Bai, W.W. Wu, On greedy randomized coordinate descent methods for solving large linear least-squares problems, Numerical Linear Algebra with Applications, 26(2019) 1-15.
[11] J. Zhang, J. Guo, On relaxed greedy randomized coordinate descent methods for solving large linear least-squares problems, Applied Numerical Mathematics, 157(2020) 372-384.
[12] Y.Q. Niu, B. Zheng, A new randomized Gauss-Seidel method for solving linear least-squares problems, Applied Mathematics Letters. 116(2021) 107057.
[13] Y. Liu, X.L. Jiang, C.Q. Gu, On maximum residual block and two-step Gauss-Seidel algorithms for linear least-squares problems, Calcolo, 58(2)(2021) 1-32.

[14] D. Needell, R. Ward, Two-subspace projection method for coherent overdetermined systems, Journal of Fourier Analysis and Applications, 19(2)(2013) 256-269.

[15] A. Ruhe, Numerical aspects of Gram-Schmidt orthogonalization of vectors, Linear Algebra and its Applications, 52-53(1983) 591-601.

[16] J. Nutini, B. Sepehry, I. Laradji, M. Schmidt, H. Koepke, A. Virani, Convergence rates for greedy Kaczmarz algorithms, and faster randomized Kaczmarz rules using the orthogonality graph, Jersey City, New Jersey, USA (2016) 547-556.