Greedy randomized sampling nonlinear Kaczmarz methods *

Yanjun Zhang, Hanyu Li*, Ling Tang

College of Mathematics and Statistics, Chongqing University, Chongqing 401331, P.R. China

Abstract

The nonlinear Kaczmarz method was recently proposed to solve the system of nonlinear equations. In this paper, we first discuss two greedy selection rules, i.e., the maximum residual and maximum distance rules, for the nonlinear Kaczmarz iteration. Then, based on them, two kinds of greedy randomized sampling methods are presented. Further, we also devise four corresponding greedy randomized block methods, i.e., the multiple samples-based methods. The linear convergence in expectation of all the proposed methods is proved. Numerical results show that, in some applications including brown almost linear function and generalized linear model, the greedy selection rules give faster convergence rates than the random ones, and the block methods outperform the single sample-based ones.

Keywords: Nonlinear Kaczmarz; greedy sampling; maximum residual rule; maximum distance rule; nonlinear problems

1. Introduction

We consider finding roots of system of nonlinear equations

\[ f(x) = 0, \quad (1) \]

where \( x \in \mathbb{R}^n \) is an unknown variable and \( f : \mathbb{R}^n \rightarrow \mathbb{R}^m \) is a continuously differentiable vector-valued function, which can be written as \( f(x) = [f_1(x), \cdots, f_m(x)]^T \in \mathbb{R}^m \) with \( f_i : \mathbb{R}^n \rightarrow \mathbb{R} \ (i = 1, \cdots, m) \) being a real-valued function. Throughout this paper, we assume that there exists a solution \( x^* \) satisfying the nonlinear equations \( (1) \), i.e., \( f(x^*) = 0 \).

*The work is supported by the National Natural Science Foundation of China (No. 11671060) and the Natural Science Foundation of Chongqing, China (No. cstc2019jcyj-msxmX0267)
*Corresponding author
Email address: lihy.hy@gmail.com or hyli@cqu.edu.cn (Hanyu Li)
Solving system of nonlinear equations is an ubiquitous problem that plays crucial role in a wide range of applied fields such as machine learning [1, 2], differential equations [3, 4], integral equations [5, 6] and optimization problems [7, 8]. Meanwhile, solving most convex optimization problems and non-convex optimization problems can also be converted into finding a good stationary point $x$ such that its gradient equals the zero vector. On the other hand, with the advent of the big data era, large-scale nonlinear problems are widely present in various fields. Therefore, it is very important to establish efficient methods for system of nonlinear equations both in practice and theory.

A classical algorithm for solving nonlinear problems (1) is the Newton-Raphson method [9] given by

$$x_{k+1} = x_k - (f'(x_k))^\dagger f(x_k),$$

where $f'(x) = [\nabla f_1(x), \ldots, \nabla f_m(x)]^T \in \mathbb{R}^{m \times n}$ is the Jacobian matrix of $f$ at $x$ and $\nabla f_i(x)^T$ is its $i$th row, and $(f'(x))^\dagger$ is the Moore-Penrose pseudoinverse of $f'(x)$. Noting that it involves calculating and storing the entire Jacobian matrix, as well as the computation of its inverse or Moore-Penrose pseudoinverse, the Newton-Raphson method is hard to be applied in large dimensional setting.

Recently, the nonlinear Kaczmarz method [10–12], which is also called Polyak step size method in [13–15], is proposed to solve nonlinear problems (1). Actually, the method can also be seen as a subsampled Newton Raphson method [10]. It consists of sequential orthogonal projections towards the linearization of one constraint set associated with a single nonlinear equation. More specifically, at each iteration, it first samples a single index $i$ from $\{1, \ldots, m\}$ and linearize $f_i(x)$ around a given $x_k \in \mathbb{R}^n$, and set the linearization of $f_i(x)$ to zero, that is,

$$f_i(x_k) + \nabla f_i(x_k)^T(x - x_k) = 0.$$

Then, the next iteration $x_{k+1}$ can be obtained in terms of orthogonal projection of the vector $x_k$ onto the above constraint, that is,

$$x_{k+1} = \arg\min_{x \in \mathbb{R}^n} \|x - x_k\|_2^2, \quad \text{s.t.} \quad f_i(x_k) + \nabla f_i(x_k)^T(x - x_k) = 0.$$

Finally, we can get the update formula of the nonlinear Kaczmarz method from (2) as follows:

$$x_{k+1} = x_k - \frac{f_i(x_k)}{\|\nabla f_i(x_k)\|_2} \nabla f_i(x_k).$$

From (3), we know that in each iteration the nonlinear Kaczmarz method only needs to compute one row of the Jacobian matrix instead of the entire matrix, which greatly reduces the amount of calculation and storage.
When \( f \) takes the form \( f(x) = Ax - b \), (1) reduces to a system of linear equations \( Ax = b \), and the nonlinear Kaczmarz method (3) reduces to the classical Kaczm arz method [16]:

\[
x_{k+1} = x_k - \frac{A^{(i)} x_k - b^{(i)}}{\|A^{(i)}\|_2^2} (A^{(i)})^T,
\]

where \( A^{(i)} \) denotes the \( i \)th row of the matrix \( A \) and \( b^{(i)} \) represents the \( i \)th entry of the vector \( b \). In 2009, Strohmer and Vershynin [17] showed that picking a row index with probability proportional to the square of the Euclidean norm of the row, the randomized Kaczmarz (RK) method converges linearly in expectation. Subsequently, to further improve its convergence rate, two greedy rules including maximum residual [18, 19], which is also known as the Motzkin method, and maximum distance [19, 20] are widely used in the literature [21–24]. In particular, combining the ideas of the RK and Motzkin methods, De Loera et al. [25] and Haddock et al. [26] presented the sampling Kaczmarz-Motzkin (SKM) method, which inherits the advantages of the previous two methods.

Considering the good performance and wide applications of greedy rules in classical Kaczmarz type methods, we discuss two greedy rules, i.e., the maximum residual and maximum distance rules, in detail for nonlinear Kaczmarz iteration in this paper, which can be regarded as a generalization from linear to nonlinear. Further, inspired by the algorithmic framework in the SKM method, we propose two different adaptive greedy randomized methods for solving the nonlinear equations (1). We prove that the two new methods converge linearly in expectation with convergence factors that are strictly smaller than that of the methods presented in [12]. Furthermore, four greedy block versions are also proposed to accelerate our new methods.

The rest of this paper is organized as follows. Section 2 provides some preliminaries. In Section 3 we first discuss two greedy rules for nonlinear Kaczmarz iteration and then propose two greedy sampling randomized nonlinear methods. The relevant convergence analysis are presented in Section 4. Section 5 gives four block greedy sampling nonlinear methods and their corresponding convergence theorems are provided in Section 6. Numerical experiments in Section 7 are devoted to illustrate our methods for solving different problems including brown almost linear function and generalized linear model. Finally, the concluding remarks of the whole paper are given in Section 8.
2. Preliminaries

2.1. Notation

For a matrix $A = (A_{i,j}) \in \mathbb{R}^{m \times n}$, $\sigma_{\text{max}}(A)$, $\|A\|_2$, $\|A\|_F$ and $A_\tau$ denote its largest singular value, spectral norm, Frobenius norm, and the restriction onto the row indices in the set $\tau$, respectively. If $A$ is a square matrix, i.e., $m = n$, $\lambda(A)$ stands for an eigenvalue of $A$ and we always let the eigenvalues of a symmetric matrix $A$ be arranged in algebraically nonincreasing order:

$$\lambda_{\text{max}}(A) = \lambda_1(A) \geq \lambda_2(A) \geq \cdots \geq \lambda_m(A).$$

If $A$ and $B$ are symmetric matrices, the relation $A \geq B$ means that $A - B$ is a positive-semidefinite matrix. We use $|\tau|$, $E^k$, and $E$ to denote the number of elements of a set $\tau$, the conditional expectation conditioned on the first $k$ iterations, and the full expected value, respectively, and let $[m] := \{1, \cdots, m\}$ for an integer $m \geq 1$. In addition, we also define

$$h_2(A) = \inf_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_2}, \quad \|A\|_{2,\infty} = \max_{1 \leq i \leq m} \|A^{(i)}\|_2,$$

$$u(x) = \begin{bmatrix} f_1(x) \\ \|\nabla f_1(x)\|_2 \\ \vdots \\ f_m(x) \\ \|\nabla f_m(x)\|_2 \end{bmatrix}^T \in \mathbb{R}^m,$$

and

$$G(x) = \begin{bmatrix} \nabla f_1(x) \\ \|\nabla f_1(x)\|_2 \\ \vdots \\ \nabla f_m(x) \\ \|\nabla f_m(x)\|_2 \end{bmatrix}^T \in \mathbb{R}^{m \times n}.$$

2.2. Previous results

We first list the nonlinear Kaczmarz method proposed in [12] in Algorithm 1. Three different rules selecting $i_k$ were considered in [12], which lead to the following three different methods.

\begin{algorithm}
\caption{Nonlinear Kaczmarz method for nonlinear equations (1)}
\begin{algorithmic}[1]
\State Input: The initial estimate $x_0 \in \mathbb{R}^n$.
\For{$k = 0, 1, 2, \cdots$ until convergence, do}
\State Select an index $i_k \in [m]$.
\State Update $x_{k+1} = x_k - \frac{f_{i_k}(x_k)}{\|\nabla f_{i_k}(x_k)\|_2^2} \nabla f_{i_k}(x_k)$.
\EndFor
\end{algorithmic}
\end{algorithm}

(1) Nonlinear randomized Kaczmarz (NRK) method: $i_k$ is randomly chosen from $[m]$ with probability of

$$p_i = \frac{|f_i(x_k)|^2}{\|f(x_k)\|_2^2}.$$
(2) Nonlinear Kaczmarz (NK) method: \( i_k \) cyclically picks value from \([m]\).

(3) Nonlinear uniformly randomized Kaczmarz (NURK) method: \( i_k \) is randomly sampled from \([m]\) with equal probability.

The convergence rate of the NK method is roughly equivalent to that of the NURK method when \( m \) is very large \([12]\), and the NRK and NURK methods also have the same convergence rate listed in Lemma 1.

Lemma 1 \([12]\). If \( f'(x) \) is a full column rank matrix and nonlinear function \( f \) satisfies the local tangential cone condition given in Definition \([7]\), then the NRK and NURK methods satisfy

\[
\mathbb{E} \left[ \| x_{k+1} - x_* \|_2^2 \right] \leq \left( 1 - \frac{1 - 2\eta}{(1 + \eta)^2 \| f'(x_k) \|_F^2 m} \right) \mathbb{E} \left[ \| x_k - x_* \|_2^2 \right].
\]

In addition, the following definition and lemmas are also necessary throughout the paper.

Definition 1 \([27]\). If for \( i \in [m] \) and \( \forall x_1, x_2 \in \mathbb{R}^n \), there exists \( \eta_i \in [0, \eta) \) satisfying \( \eta = \max_i \eta_i < \frac{1}{2} \) such that

\[
|f_i(x_1) - f_i(x_2) - \nabla f_i(x_1)^T (x_1 - x_2)| \leq \eta_i |f_i(x_1) - f_i(x_2)|,
\]

then the function \( f : \mathbb{R}^n \rightarrow \mathbb{R}^m \) is referred to satisfy the local tangential cone condition.

Lemma 2 \([12]\). If the function \( f \) satisfies the local tangential cone condition, then for \( i \in [m] \), \( \forall x_1, x_2 \in \mathbb{R}^n \) and the updating formula \([3]\), we have

\[
\| x_{k+1} - x_* \|_2^2 \leq \| x_k - x_* \|_2^2 - (1 - 2\eta_i) \frac{|f_i(x_k)|^2}{\| \nabla f_i(x_k) \|_2^2}.
\]

Lemma 3 \([28]\). If \( A \) and \( A + E \) are \( n \)-by-\( n \) symmetric matrices, then

\[
\lambda_k(A) + \lambda_n(E) \leq \lambda_k(A + E) \leq \lambda_k(A) + \lambda_1(E), \quad k = 1 : n.
\]

3. Single sample-based greedy randomized sampling methods

In this section, we mainly propose two single sample-based greedy randomized sampling methods. Before that, we first discuss two greedy rules in detail, which are the key of this paper.
3.1. Two greedy rules

In [11], Zeng et al. presented a greedy selection strategy as follows:

$$i_k = \arg \max_{1 \leq i \leq m} |f_i(x_k)|^2,$$  \hspace{1cm} (5)

which is aimed at choosing the maximum magnitude entry of the absolute residual vector. They also showed that this greedy strategy converges faster than the NK and NURK methods in both theoretical analysis and experimental results. When \( f(x) = Ax - b \), (5) reduces to:

$$i_k = \arg \max_{1 \leq i \leq m} |A^{(i)}x_k - b^{(i)}|^2,$$ \hspace{1cm} (6)

which is called the maximum residual rule \[18, 19\] or Motzkin method \[29, 30\] and widely used in various areas including linear system and linear feasibility problems.

According to Lemma 2, we know that the magnitude of \( \|x_{k+1} - x_*\|^2 \) is almost entirely determined by \( \|f_i(x_k)\|^2 \), because \( \|x_k - x_*\|^2 \) is a given value. Consequently, Lemma 2 implies that to make the most progress in one iteration, we should pick \( i_k \) corresponding to the largest \( \|f_i(x_k)\|^2 \) \( \|\nabla f_i(x_k)\|_2^2 \) for all \( i \in [m] \). Therefore, we propose a new greedy selection rule as follows:

$$i_k = \arg \max_{1 \leq i \leq m} \frac{|f_i(x_k)|^2}{\|\nabla f_i(x_k)\|_2^2}.$$  \hspace{1cm} (7)

When \( f(x) = Ax - b \), (7) reduces to:

$$i_k = \arg \max_{1 \leq i \leq m} \frac{|A^{(i)}x_k - b^{(i)}|^2}{\|A^{(i)}\|_2^2},$$  \hspace{1cm} (8)

which is called the maximum distance rule \[19, 20\] and also widely used in the Kaczmarz setting for solving linear problems.

From the above analysis, we find that (5) and (7) can be viewed as the generalization of (6) and (8) from linear problems to nonlinear problems, respectively. So, we still refer to the greedy selection strategies (5) and (7) as maximum residual rule and maximum distance rule, respectively.

3.2. Two greedy sampling methods

The maximum residual rule (5) shows that sampling larger entries of the residual should converge faster. However, calculating the whole residual is expensive when the number of samples is very large. From this point we give the maximum residual-based sampling nonlinear Kaczmarz (MR-SNK) method presented as the case 1 in Algorithm 2. Specifically, the MR-SNK method operates
by uniformly sampling an index subset from among $[m]$, computing the absolute subresidual vector of this subset, and projecting onto the linearization of a single nonlinear equation corresponding to the largest magnitude entry of this subresidual.

On the other hand, as the maximum distance selection rule leads to the best decrease in mean squared error in each iteration, similar to the MR-SNK method, the maximum distance-based sampling nonlinear Kaczmarz (MD-SNK) method is provided as the case 2 in Algorithm 2.

**Algorithm 2** The (MR/MD)-SNK method

1: Input: The initial estimate $x_0 \in \mathbb{R}^n$, parameter $\beta \in [m]$.
2: for $k = 0, 1, 2, \cdots$ until convergence, do
3: Choose an index subset $\tau_k$ of size $\beta$ uniformly at random from among $[m]$.
4: Switch the greedy sampling strategy to determine two different iterative methods, i.e., the MR-SNK and MD-SNK methods.
5: \hspace{0.5cm} $\triangleright$ case 1: maximum residual (MR) rule
6: \hspace{1cm} $i_k = \arg \max_{i \in \tau_k} |f_i(x_k)|^2$.
7: \hspace{0.5cm} $\triangleright$ case 2: maximum distance (MD) rule
8: \hspace{1cm} $i_k = \arg \max_{i \in \tau_k} \frac{|f_i(x_k)|^2}{\|\nabla f_i(x_k)\|^2}$.
9: Update $x_{k+1} = x_k - \frac{f_{i_k}(x_k)}{\|\nabla f_{i_k}(x_k)\|^2} \nabla f_{i_k}(x_k)$.
10: end for

**Remark 1.** The MR-SNK method has been studied for several specific choices of $\beta$ and the function $f$. If $\beta = 1$, the MR-SNK method reduces the NURK method \[12\]. If $\beta = m$, we call the MR-SNK method the maximum residual-based nonlinear Kaczmarz (MR-NK) method, which can be seen as a special variant of the GP method presented in \[11\]. If $f = Ax - b$, the MR-SNK method gives the SKM method \[25\], which can in turn recover a variant of randomized Kaczmarz method \[31\] and the Motzkin method \[19, 20\] by choosing different $\beta$.

The MD-SNK method also has been studied for several specific choices of $\beta$ and the function $f$. If $\beta = 1$, the MD-SNK method gives the NURK method \[12\]. The MD-SNK method with $\beta = m$ and $f = Ax - b$ recovers the maximal weighted residual Kaczmarz method \[19, 20\]. In addition, similar to the MR-NK method discussed above, we can obtain a new greedy projection selection method on the basis of the maximum distance rule by setting $\beta = m$ in the MD-SNK method. The new greedy method is called as the maximum distance-based nonlinear Kaczmarz (MD-NK)
method in this paper. Similar to the SKM method, we also obtain a new sampling type Kaczmarz method for solving linear problems by setting \( f = Ax - b \) in the MD-SNK method, and we call it the sampling distance-based Kaczmarz (SDK) method. Therefore, different parameters \( \beta \) and function \( f \) in the MR/MD-SNK method lead to different methods, which are summarized in Table 1 for convenience.

| Method | \( \beta \) | \( f \) |
|--------|-------------|---------|
| NURK [12] | 1 | \( f \) |
| MR-SNK | MR-NK | \( m \) | \( f \) |
| SKM [25] | \( \beta \in [m] \) | \( Ax - b \) |
| NURK [12] | 1 | \( f \) |
| MD-SNK | MD-NK | \( m \) | \( f \) |
| SDK | \( \beta \in [m] \) | \( Ax - b \) |

4. Convergence analysis

In this section, we show that the MR-SNK and MD-SNK methods enjoy a linear rate in expectation of convergence. Before that, we first present Lemmas 4 and 5 which are crucial to the convergence analysis in the following theorems.

**Lemma 4.** If the function \( f \) satisfies the local tangential cone condition given in Definition 1 for \( \forall x_1, x_2 \in \mathbb{R}^n \) and an index subset \( \tau \subseteq [m] \), we have

\[
\| f_\tau (x_1) - f_\tau (x_2) - f'_\tau (x_1) (x_1 - x_2) \|_2^2 \leq \eta^2 \| f_\tau (x_1) - f_\tau (x_2) \|_2^2 \tag{9}
\]

and

\[
\| f_\tau (x_1) - f_\tau (x_2) \|_2^2 \geq \frac{1}{1 + \eta^2} \| f'_\tau (x_1) (x_1 - x_2) \|_2^2. \tag{10}
\]

**Proof.** From (4) in Definition 1 we have

\[
\| f_\tau (x_1) - f_\tau (x_2) - f'_\tau (x_1) (x_1 - x_2) \|_2^2 = \sum_{i \in \tau} \left| f_i (x_1) - f_i (x_2) - \nabla f_i (x_1)^T (x_1 - x_2) \right|^2 \\
\leq \sum_{i \in \tau} \eta_i^2 \left| f_i (x_1) - f_i (x_2) \right|^2 \\
\leq \eta^2 \| f_\tau (x_1) - f_\tau (x_2) \|_2^2.
\]
which is the desired result (9).

According to (9), we get

\[ \eta^2 \| f_\tau(x_1) - f_\tau(x_2) \|_2^2 \geq \| f_\tau(x_1) - f_\tau(x_2) - f'_\tau(x_1)(x_1 - x_2) \|_2^2 \]
\[ \geq \| f'_\tau(x_1)(x_1 - x_2) \|_2^2 - \| f_\tau(x_1) - f_\tau(x_2) \|_2^2, \]

which implies

\[ (1 + \eta^2) \| f_\tau(x_1) - f_\tau(x_2) \|_2^2 \geq \| f'_\tau(x_1)(x_1 - x_2) \|_2^2, \]

that is,

\[ \| f_\tau(x_1) - f_\tau(x_2) \|_2^2 \geq \frac{1}{1 + \eta^2} \| f'_\tau(x_1)(x_1 - x_2) \|_2^2. \]

**Lemma 5.** If the function \( f \) satisfies the local tangential cone condition given in Definition 1 and a vector \( x_* \in \mathbb{R}^n \) satisfies \( f(x_*) = 0 \), then for \( \forall x \in \mathbb{R}^n \) and an index subset \( \tau \subseteq [m] \), we have

\[ \| u_\tau(x) - G_\tau(x)(x - x_*) \|_2^2 \leq \eta^2 \| u_\tau(x) \|_2^2 \]

and

\[ \| u_\tau(x) \|_2^2 \geq \frac{1}{1 + \eta^2} \| G_\tau(x)(x - x_*) \|_2^2. \]

**Proof.** Considering the definitions of \( u(x) \) and \( G(x) \) given in Sections 2.1 and 4 in Definition 1, we have

\[ \| u_\tau(x) - G_\tau(x)(x - x_*) \|_2^2 = \sum_{i \in \tau} \left| u^{(i)}(x) - G^{(i)}(x)(x - x_*) \right|_2^2 \]
\[ = \sum_{i \in \tau} \frac{|f_i(x) - \nabla f_i(x)^T(x - x_*)|^2}{\| \nabla f_i(x) \|_2^2} \]
\[ = \sum_{i \in \tau} \frac{|f_i(x) - f_i(x_*) - \nabla f_i(x)^T(x - x_*)|^2}{\| \nabla f_i(x) \|_2^2} \]
\[ \leq \sum_{i \in \tau} \eta^2 \frac{|f_i(x) - f_i(x_*)|^2}{\| \nabla f_i(x) \|_2^2} \]
\[ \leq \sum_{i \in \tau} \eta^2 \frac{|f_i(x)|^2}{\| \nabla f_i(x) \|_2^2} \]
\[ = \eta^2 \| u_\tau(x) \|_2^2, \]
which is the result (11).

On the other hand, by applying (11) we get
\[
\eta^2 \| u_\tau(x) \|_2^2 \geq \| u_\tau(x) - G_\tau(x) (x - x_*) \|_2^2 \\
\geq \| G_\tau(x) (x - x_*) \|_2^2 - \| u_\tau(x) \|_2^2 ,
\]
which implies
\[
(1 + \eta^2) \| u_\tau(x) \|_2^2 \geq \| G_\tau(x) (x - x_*) \|_2^2 ,
\]
that is,
\[
\| u_\tau(x) \|_2^2 \geq \frac{1}{1 + \eta^2} \| G_\tau(x) (x - x_*) \|_2^2 .
\]

**Theorem 1.** If the nonlinear function \( f \) satisfies the local tangential cone condition given in Definition 1, \( \eta = \max_i \eta_i < \frac{1}{2} \), and \( f(x_*) = 0 \), then the iterations of MR-SNK method, i.e., the first case in Algorithm 2, satisfy
\[
E \left[ \| x_{k+1} - x_* \|_2^2 \right] \leq \left( 1 - \frac{1 - 2\eta \beta}{1 + \eta^2 \xi_k} \right) \frac{\xi_k}{m} \| f'(x_k) \|_{2,\infty}^2 , 
\]
where
\[
\xi_k = \frac{\sum_{\tau_k \in \left( \eta_1 \right)} \| f_{\tau_k}(x_k) \|_2^2}{\sum_{\tau_k \in \left( \eta_1 \right)} \| f_{\tau_k}(x_k) \|_{\infty}^2} .
\]

**Proof.** According to the MR-SNK method in Algorithm 2 and Lemma 2, we know that
\[
\| x_{k+1} - x_* \|_2^2 \leq \| x_k - x_* \|_2^2 - \left( 1 - 2\eta \right) \frac{\| f_{i_k}(x_k) \|_2^2}{\| \nabla f_{i_k}(x_k) \|_2^2} \\
= \| x_k - x_* \|_2^2 - \left( 1 - 2\eta \right) \frac{\| f_{\tau_k}(x_k) \|_\infty^2}{\| \nabla f_{i_k}(x_k) \|_2^2} .
\]
Taking expectation of both sides conditioned on $x_k$ gives

$$
E^k \left[ \|x_{k+1} - x_*\|_2^2 \right] \leq \|x_k - x_*\|_2^2 - E^k \left[ (1 - 2\eta_k) \frac{\|f_{\tau_k}(x_k)\|_2^2}{\|\nabla f_{\tau_k}(x_k)\|_2} \right] 
$$

which together with $\xi_k$ defined in (14) gets

$$
E^k \left[ \|x_{k+1} - x_*\|_2^2 \right] \leq \|x_k - x_*\|_2^2 - \frac{(m)}{\beta} (1 - 2\eta) \frac{\|f_{\tau_k}(x_k)\|_2^2}{\|\nabla f_{\tau_k}(x_k)\|_2} \sum_{\tau_k \in (m)} |f_{\tau_k}(x_k)|_2^2 
$$

Further, considering (10) in Lemma 4 and the definition of $h_2(\cdot)$ in Section 2.1, we get

$$
E^k \left[ \|x_{k+1} - x_*\|_2^2 \right] \leq \|x_k - x_*\|_2^2 - \frac{1}{\beta} \frac{1 - 2\eta}{\|f'(x_k)\|_2^2} \frac{1}{m} \sum_{\tau_k \in (m)} |f_{\tau_k}(x_k)|_2^2 
$$

So, taking expectation on both sides and using the tower rule of expectation, we get the desired result (13).

**Remark 2.** Noting that $1 \leq \xi_k \leq \beta$ and $\beta h^2_2(f'(x_k)) \leq \|f'(x_k)\|_{2,\infty} m$, it is easy to obtain that

$$
0 < \frac{1 - 2\eta}{1 + \eta^2 \xi_k \|f'(x_k)\|_{2,\infty} m} < 1, 
$$

which indicates that the convergence factor of the MR-SNK method, i.e.,

$$
\rho_{MR-SNK} = 1 - \frac{1 - 2\eta \beta}{1 + \eta^2 \xi_k \|f'(x_k)\|_{2,\infty} m}, 
$$
is smaller than 1. Thus, the iteration of the MR-SNK method is contracted and convergent in expectation.

**Remark 3.** According to Lemma [1] we know that the convergence factors of the NRK and NURK methods are the same and

\[ \rho_{NRK} = \rho_{NURK} = 1 - \frac{1 - 2\eta_i h_i^2(f'(x_k))}{(1 + \eta_i)^2 \|f'(x_k)\|_2^2 m}. \]

Since \( \|f'(x_k)\|_F \geq \|f'(x_k)\|_2^2 \), \( \frac{1}{\xi_k} \geq 1 \) and \( 0 < \eta < \frac{1}{2} \), we have

\[ \frac{1 - 2\eta_i \beta h_i^2(f'(x_k))}{1 + \eta_i^2 \xi_k \|f'(x_k)\|_2^2 m} > \frac{1 - 2\eta_i h_i^2(f'(x_k))}{(1 + \eta_i)^2 \|f'(x_k)\|_2^2 m}, \]

which implies that the convergence factor of the MR-SNK method is smaller than that of the NRK and NURK methods. That is,

\[ \rho_{MR-SNK} < \rho_{NRK} = \rho_{NURK}. \]

**Theorem 2.** If the nonlinear function \( f \) satisfies the local tangential cone condition given in Definition [1] \( \eta = \max_i \eta_i < \frac{1}{2} \), and \( f(x_*) = 0 \), then the iterations of MD-SNK method, i.e., the second case in Algorithm [2] satisfy

\[ \mathbb{E}\left[ \|x_{k+1} - x_*\|_2^2 \right] \leq \left( 1 - \frac{1 - 2\eta_i \beta h_i^2(G(x_k))}{1 + \eta_i^2 \xi_k \|f'(x_k)\|_2^2 m} \right) \mathbb{E}\left[ \|x_k - x_*\|_2^2 \right], \quad (15) \]

where

\[ \xi_k = \frac{\sum_{\tau_k \in \binom{m}{i}} \|u_{\tau_k}(x_k)\|_2^2}{\sum_{\tau_k \in \binom{m}{i}} \|u_{\tau_k}(x_k)\|_\infty^2}. \quad (16) \]

**Proof.** According to Lemma [1] the definition of \( u(x) \) in Section [2.2] and the MD-SNK method in Algorithm [2] we have

\[ \|x_{k+1} - x_*\|_2^2 \leq \|x_k - x_*\|_2^2 - (1 - 2\eta_i h_i) \frac{|f_{i_k}(x_k)|^2}{\|\nabla f_{i_k}(x_k)\|_2^2} \]

\[ = \|x_k - x_*\|_2^2 - (1 - 2\eta_i h_i) |u_{i_k}(x_k)|^2 \]

\[ = \|x_k - x_*\|_2^2 - (1 - 2\eta_i h_i) \|u_{\tau_k}(x_k)\|_\infty^2. \]
Now, taking expectation of both sides conditioned on $x_k$ gives

$$
\mathbb{E}^k \left[ \|x_{k+1} - x_*\|_2^2 \right] \leq \|x_k - x_*\|_2^2 - \mathbb{E}^k \left[ (1 - 2\eta_k) \|u_{r_k}(x_k)\|_\infty^2 \right]
$$

$$
= \|x_k - x_*\|_2^2 - \sum_{r_k \in \binom{[m]}{\beta}} \frac{1}{\binom{m}{\beta}} (1 - 2\eta_k) \|u_{r_k}(x_k)\|_\infty^2
$$

$$
\leq \|x_k - x_*\|_2^2 - \frac{1}{\binom{m}{\beta}} (1 - 2\eta) \sum_{r_k \in \binom{[m]}{\beta}} \|u_{r_k}(x_k)\|_\infty^2,
$$

which together with the definition of $\zeta_k$ in (16) leads to

$$
\mathbb{E}^k \left[ \|x_{k+1} - x_*\|_2^2 \right] \leq \|x_k - x_*\|_2^2 - \left( 1 - \frac{2\eta}{\zeta_k} \right) \sum_{r_k \in \binom{[m]}{\beta}} \|u_{r_k}(x_k)\|_\infty^2
$$

$$
= \|x_k - x_*\|_2^2 - \frac{1}{\binom{m}{\beta}} (1 - 2\eta) \frac{\binom{m}{\beta}}{\zeta_k} \|u(x_k)\|_2^2
$$

$$
= \|x_k - x_*\|_2^2 - (1 - 2\eta) \frac{\beta}{\zeta_k} \|u(x_k)\|_2^2.
$$

Further, by making use of (12) in Lemma 5 and the definition of $h_2(\cdot)$ in Section 2.1, we can obtain

$$
\mathbb{E}^k \left[ \|x_{k+1} - x_*\|_2^2 \right] \leq \|x_k - x_*\|_2^2 - (1 - 2\eta) \frac{\beta}{\zeta_k} \|u(x_k)\|_2^2
$$

$$
\leq \left( 1 - \frac{1 - 2\eta \beta h_2^2(G(x_k))}{1 + \eta^2 \zeta_k} \frac{1}{m} \right) \|x_k - x_*\|_2^2.
$$

Finally, the desired result (15) can be obtained by taking the full expectation on both sides.

**Remark 4.** Considering that $1 \leq \zeta_k \leq \beta$ and $\beta h_2^2(G(x_k)) \leq m$, we get

$$
0 < \frac{1 - 2\eta \beta h_2^2(G(x_k))}{1 + \eta^2 \zeta_k} \frac{1}{m} < 1,
$$

which indicates that the convergence factor of the MD-SNK method, i.e.,

$$
\rho_{MD-SNK} = 1 - \frac{1 - 2\eta \beta h_2^2(G(x_k))}{1 + \eta^2 \zeta_k} \frac{1}{m},
$$

is smaller than 1.

**Remark 5.** Letting $D$ be an $m$-by-$m$ diagonal matrix in the following form

$$
D = \begin{bmatrix}
\|\nabla f_1(x_k)\|_2 \\
\vdots \\
\|\nabla f_m(x_k)\|_2
\end{bmatrix},
$$

we have
then we get a relation between \(f'(x_k)\) and \(G(x_k)\), that is,

\[
f'(x_k) = DG(x_k). \tag{17}
\]

Now, by setting \(A = G(x_k)G(x_k)^T\) and \(B = D^TD\), since both matrices \(A\) and \(B\) are symmetric positive-semidefinite, we have

\[
\lambda_1(B) I \geq B \geq \lambda_m(B) I
\]

and then obtain

\[
\lambda_1(B)A = A^{\frac{1}{2}}(\lambda_1(B)I - B)A^{\frac{1}{2}} + A^{\frac{1}{2}}BA^{\frac{1}{2}} \geq A^{\frac{1}{2}}BA^{\frac{1}{2}}.
\]

Further, by applying Lemma 3 we get

\[
\lambda_m \left( \lambda_1(B)A \right) = \lambda_m \left( A^{\frac{1}{2}}(\lambda_1(B)I - B)A^{\frac{1}{2}} + A^{\frac{1}{2}}BA^{\frac{1}{2}} \right) \\
\geq \lambda_m \left( A^{\frac{1}{2}}(\lambda_1(B)I - B)A^{\frac{1}{2}} \right) + \lambda_m \left( A^{\frac{1}{2}}BA^{\frac{1}{2}} \right) \\
\geq \lambda_m \left( A^{\frac{1}{2}}BA^{\frac{1}{2}} \right) \\
= \lambda_m \left( AB \right),
\]

which implies

\[
\lambda_1(B)\lambda_m (A) \geq \lambda_m (AB).
\]

That is,

\[
\lambda_1(D^TD)\lambda_m \left( G(x_k)G(x_k)^T \right) \geq \lambda_m \left( G(x_k)G(x_k)^TD^TD \right) = \lambda_m \left( DG(x_k)G(x_k)^TD^T \right),
\]

which together with the definitions of \(D\), \(G(x_k)\) and \(f'(x_k) = DG(x_k)\) in (17) indicates

\[
\|f'(x_k)\|_{2,\infty}^2 h_2^2(G(x_k)) \geq h_2^2 \left( f'(x_k) \right). \tag{18}
\]

Meanwhile, from Theorems 1 and 2 we know that the convergence factors of the MR-SNK and MD-SNK methods are

\[
\rho_{\text{MR-SNK}} \sim 1 - \frac{1 - 2\eta}{1 + \eta^2} \|f'(x_k)\|_{2,\infty}^2 m, \quad \text{and} \quad \rho_{\text{MD-SNK}} \sim 1 - \frac{1 - 2\eta}{1 + \eta^2} \left( G(x_k) \right). \frac{m}{m},
\]

respectively.
Thus, according to (18), we conclude that
\[
1 - \frac{1 - 2\eta h^2(G(x_k))}{1 + \eta^2} \leq 1 - \frac{1 - 2\eta h^2(f'(x_k))}{1 + \eta^2 \|f'(x_k)\|_{2,\infty}^2} m,
\]
which means that the convergence factor of the MD-SNK method is smaller than that of the MR-SNK method.

**Remark 6.** From Remark 1, we know that the MR-SNK and MD-SNK methods can recover the NURK method by setting \(\beta = 1\). It is convenient to obtain the convergence result of the NURK method by using Theorem 1 or Theorem 2. Specifically, substituting \(\beta = \xi_k = \zeta_k = 1\) into the convergence bounds (13) and (15) and combining with (19) yields
\[
\mathbb{E} \left[ \|x_{k+1} - x_*\|_2^2 \right] \leq \left( 1 - \frac{1 - 2\eta h^2(G(x_k))}{1 + \eta^2} \right) \mathbb{E} \left[ \|x_k - x_*\|_2^2 \right] \quad \text{(by Theorem 2)}
\]
\[
\leq \left( 1 - \frac{1 - 2\eta h^2(f'(x_k))}{1 + \eta^2 \|f'(x_k)\|_{2,\infty}^2} \right) \mathbb{E} \left[ \|x_k - x_*\|_2^2 \right] \quad \text{(by Theorem 1)}
\]
\[
< \left( 1 - \frac{1 - 2\eta h^2(f'(x_k))}{(1 + \eta^2) \|f'(x_k)\|_{\infty}^2} \right) \mathbb{E} \left[ \|x_k - x_*\|_2^2 \right] \quad \text{(by Theorem 3.1 in [12])},
\]
which indicates that our results are all tighter than the existing error estimate for the NURK method.

**Remark 7.** Theorems 1 and 2 show that the progress made by an iteration of the MR-SNK and MD-SNK methods depends on the value of \(\xi_k\) and \(\zeta_k\), respectively. Since \(1 \leq \xi_k \leq \beta\), we can conclude that the closer \(\xi_k\) approaches 1, the smaller the convergence factor of the MR-SNK method is, and we call it the “best case” when \(\xi_k = 1\). Conversely, when \(\xi_k\) is closer to \(\beta\), the convergence factor of MR-SNK method is bigger, and we call it the “worst case” when \(\xi_k = \beta\). Similarly, we can get other convergence results in the best and worst cases, which are summarised in Table 2. Moreover, we summarize the relationships of these best case convergence factors in Table 3. The inequality sign (\(\leq\)) denotes that the best case convergence factor of the method listed in the row is smaller or equal to the method listed in the column, and “P” indicates that the smaller convergence factor is problem-dependent.

5. Multiple samples-based greedy randomized sampling methods

In this section, on the basis of the MR/MD-SNK method, we further consider block version nonlinear iteration and propose block sampling nonlinear Kaczmarz (BSNK) method for solving
Thus, the solution of the projection is

\[ x_{k+1} = \text{argmin}_{x \in \mathbb{R}^n} \| x - x_k \|_2^2, \quad \text{s.t.} \quad f_{\tau}(x_k) + f'_{\tau}(x_k)(x - x_k) = 0. \]

Thus, the solution of the projection is

\[ x_{k+1} = x_k - (f'_{\tau}(x_k))^\dagger f_{\tau}(x_k). \] (20)

From the update formula (20) of the block method, we know that the choice of the index subset \( \tau \) greatly affects the performance of the algorithm. So, inspired by the work of Zhang and Li [32], we construct two algorithmic frameworks for determining the index subset, which combining with the greedy rules presented in Section 3.1 lead to four greedy BSNK methods shown in Algorithms 3 and 4 for the nonlinear problems (1). Algorithms 3 and 4 indicate that their iterations make faster progress than that of Algorithm 2. The main reason is that the iteration index \( i_k \) in Algorithm 2 belongs to the index set used in Algorithms 3 and 4. In addition, if \( \beta = m \) in Algorithm 3 and

Table 2: The best and worst case convergence factors of the MR-SNK, MR-NK, NURK-MR (result of the NURK method from Theorem 1), MD-SNK, MD-NK, NURK-MD (result of the NURK method from Theorem 2) methods.

| Method   | Worst case | Best case |
|----------|------------|-----------|
| MR-SNK   | \( 1 - \frac{1}{2} \eta^2 \frac{h^2_2(f_{\tau}(x_k))}{\| f(x_k) \|_\infty^2} \) | \( 1 - \frac{1}{2} \eta^2 \frac{h^2_2(f_{\tau}(x_k))}{\| f(x_k) \|_\infty^2} \) |
| MR-NK    | \( 1 - \frac{1}{2} \eta^2 \frac{h^2_2(f_{\tau}(x_k))}{\| f(x_k) \|_\infty^2} \) | \( 1 - \frac{1}{2} \eta^2 \frac{h^2_2(f_{\tau}(x_k))}{\| f(x_k) \|_\infty^2} \) |
| NURK-MR  | \( 1 - \frac{1}{2} \eta^2 \frac{h^2_2(f_{\tau}(x_k))}{\| f(x_k) \|_\infty^2} \) | \( 1 - \frac{1}{2} \eta^2 \frac{h^2_2(f_{\tau}(x_k))}{\| f(x_k) \|_\infty^2} \) |
| MD-SNK   | \( 1 - \frac{1}{2} \eta^2 \frac{h^2_2(G(x_k))}{m} \) | \( 1 - \frac{1}{2} \eta^2 \frac{h^2_2(G(x_k))}{m} \) |
| MD-NK    | \( 1 - \frac{1}{2} \eta^2 \frac{h^2_2(G(x_k))}{m} \) | \( 1 - \frac{1}{2} \eta^2 \frac{h^2_2(G(x_k))}{m} \) |
| NURK-MD  | \( 1 - \frac{1}{2} \eta^2 \frac{h^2_2(G(x_k))}{m} \) | \( 1 - \frac{1}{2} \eta^2 \frac{h^2_2(G(x_k))}{m} \) |

Table 3: Comparison of the best case convergence factors listed in Table 2.

| Method   | MD-NK | MR-NK | MD-SNK | MR-SNK | NURK-MD | NURK-MR |
|----------|-------|-------|--------|--------|---------|---------|
| MD-NK    | \( \leq \) | \( \leq \) | \( \leq \) | \( \leq \) | \( \leq \) |
| MR-NK    | \( \leq \) | \( \leq \) | \( \leq \) |
| MD-SNK   | \( \leq \) | \( \leq \) |
| MR-SNK   | \( \leq \) | \( \leq \) |
| NURK-MD  | \( \leq \) |
| NURK-MR  | \( \leq \) |
Algorithm 3 The (MR/MD)-BSNK1 method

1: Input: The initial estimate $x_0 \in \mathbb{R}^n$, parameter $\beta \in [m]$.

2: for $k = 0, 1, 2, \ldots$ until convergence do

3: Choose an index subset $\tau_k$ of size $\beta$ uniformly at random from among $[m]$.

4: Switch the greedy sampling strategy to determine two different iterative methods, i.e., the MR-BSNK1 and MD-BSNK1 methods.

5: case 1: maximum residual (MR) rule

   $i_k = \arg \max_{i \in \tau_k} |f_i(x_k)|^2$ and $\delta_k = \max_{i \in \tau_k} |f_i(x_k)|^2$.

   Determine the index set $\mathcal{I}_k = \{h_k||f_{h_k}(x_k)||^2 \geq \delta_k; h_k \in [m] \setminus \tau_k\} \cup \{i_k\}$.

6: case 2: maximum distance (MD) rule

   $i_k = \arg \max_{i \in \tau_k} \frac{|f_i(x_k)|^2}{\|\nabla f_i(x_k)\|_2^2}$ and $\delta_k = \max_{i \in \tau_k} \frac{|f_i(x_k)|^2}{\|\nabla f_i(x_k)\|_2^2}$.

   Determine the index set $\mathcal{I}_k = \{h_k||f_{h_k}(x_k)||^2 \geq \delta_k; h_k \in [m] \setminus \tau_k\} \cup \{i_k\}$.

7: Update $x_{k+1} = x_k - (f'_{\mathcal{I}_k}(x))^\dagger f_{\mathcal{I}_k}(x_k)$.

8: end for

6. Convergence analysis

We now present the convergence results for the four block methods proposed in Section 5, i.e., the MR-BSNK1, MD-BSNK1, MR-BSNK2 and MD-BSNK2 methods.

Lemma 6. If the nonlinear function $f$ satisfies the local tangential cone condition given in Definition 4 and a vector $x_* \in \mathbb{R}^n$ satisfies $f(x_*) = 0$, then from the iteration $x_{k+1} = x_k - (f'_{\tau_k}(x_k))^\dagger f_{\tau_k}(x_k)$ with $\tau_k \subseteq [m]$, we have

$$||x_{k+1} - x_*||^2 \leq ||x_k - x_*||^2 - (h_2^2((f'_{\tau_k}(x_k))^\dagger) - 2\eta \sigma^2_{\max} ((f'_{\tau_k}(x_k))^\dagger)) \|f_{\tau_k}(x_k)\|^2_2.$$

Proof. From the iteration formula $x_{k+1} = x_k - (f'_{\tau_k}(x_k))^\dagger f_{\tau_k}(x_k)$, we have

$$||x_{k+1} - x_*||^2 = ||x_k - (f'_{\tau_k}(x_k))^\dagger f_{\tau_k}(x_k) - x_*||^2_2$$

$$= ||x_k - x_*||^2_2 + \|(f'_{\tau_k}(x_k))^\dagger f_{\tau_k}(x_k)\|^2_2 - 2<(f'_{\tau_k}(x_k))^\dagger f_{\tau_k}(x_k), x_k - x_*>,$$

$$= ||x_k - x_*||^2_2 - \frac{1}{2}<\|f_{\tau_k}(x_k)\|^2_2, f_{\tau_k}(x_k), x_k - x_*>,$$

$$\leq ||x_k - x_*||^2_2 - \frac{1}{2}(h_2^2((f'_{\tau_k}(x_k))^\dagger) - 2\eta \sigma^2_{\max} ((f'_{\tau_k}(x_k))^\dagger)) \|f_{\tau_k}(x_k)\|^2_2.$$
Algorithm 4 The (MR/MD)-BSNK2 method

1: Input: The initial estimate $x_0 \in \mathbb{R}^n$, parameter $\nu \in [m]$, parameters $\beta_1, \beta_2, \cdots, \beta_\nu$ s.t. $\sum_{j=1}^{\nu} \beta_j = m$.

2: for $j = 1 : \nu$ do

3: Choose an index subset $\tau_j$ of size $\beta_j$ uniformly at random from among the remaining indices of $[m]$ without replacement.

4: Switch the greedy sampling strategy to determine two different iterative methods, i.e., the MR-BSNK2 and MD-BSNK2 methods.

5: \hspace{1em} \triangleright case 1: maximum residual (MR) rule

6: \hspace{2em} $t_j = \arg \max_{i \in \tau_j} |f_i(x_k)|^2$.

7: \hspace{1em} \triangleright case 2: maximum distance (MD) rule

8: \hspace{2em} $t_j = \arg \max_{i \in \tau_j} \|\nabla f_i(x_k)\|_2$.

9: end for

10: Determine the index set $\mathcal{J}_k = \{t_1, t_2, \cdots, t_\nu\}$.

11: Update $x_{k+1} = x_k - (f'_{\mathcal{J}_k}(x))' f_{\mathcal{J}_k}(x_k)$.

which together with the fact that $(f'_{\tau_k}(x_k))' = (f'_{\tau_k}(x_k))^T(f'_{\tau_k}(x_k))(f'_{\tau_k}(x_k))^T)$ leads to

$$
\|x_{k+1} - x_*\|_2^2 = \|x_k - x_*\|_2^2 + \| (f'_{\tau_k}(x_k))^T f_{\tau_k}(x_k)(f'_{\tau_k}(x_k))^T f_{\tau_k}(x_k)(x_k - x_*) \\
= \|x_k - x_*\|_2^2 + \| (f'_{\tau_k}(x_k))^T f_{\tau_k}(x_k)\|_2^2 \\
+ 2 f_{\tau_k}(x_k)(f'_{\tau_k}(x_k)(f'_{\tau_k}(x_k))^T f_{\tau_k}(x_k) - f_{\tau_k}(x_k) - f'_{\tau_k}(x_k)(x_k - x_*) \\
- 2 f_{\tau_k}(x_k)(f'_{\tau_k}(x_k)(f'_{\tau_k}(x_k))^T f_{\tau_k}(x_k) \\
= \|x_k - x_*\|_2^2 + \| (f'_{\tau_k}(x_k))^T f_{\tau_k}(x_k)\|_2^2 \\
+ 2 f_{\tau_k}(x_k)(f'_{\tau_k}(x_k)(f'_{\tau_k}(x_k))^T f_{\tau_k}(x_k) - f_{\tau_k}(x_k) - f'_{\tau_k}(x_k)(x_k - x_*) \\
- 2 f_{\tau_k}(x_k)(f'_{\tau_k}(x_k)(f'_{\tau_k}(x_k))^T f_{\tau_k}(x_k)
$$

Since

$$
\| (f'_{\tau_k}(x_k))^T f_{\tau_k}(x_k)\|_2^2 = f_{\tau_k}(x_k)(f'_{\tau_k}(x_k))^T f_{\tau_k}(x_k) \\
= f_{\tau_k}(x_k)(f'_{\tau_k}(x_k))^T f_{\tau_k}(x_k) \\
= f_{\tau_k}(x_k)(f'_{\tau_k}(x_k)(f'_{\tau_k}(x_k))^T)^T f_{\tau_k}(x_k)
$$

we get

$$
\|x_{k+1} - x_*\|_2^2 = \|x_k - x_*\|_2^2 - \| (f'_{\tau_k}(x_k))^T f_{\tau_k}(x_k)\|_2^2 \\
+ 2 f_{\tau_k}(x_k)(f'_{\tau_k}(x_k)(f'_{\tau_k}(x_k))^T f_{\tau_k}(x_k) - f_{\tau_k}(x_k) - f'_{\tau_k}(x_k)(x_k - x_*) \right).
$$
Furthermore, using Cauchy-Schwarz inequality, it is true that
\[ \|x_{k+1} - x_*\|_2^2 \leq \|x_k - x_*\|_2^2 - 2\|f'(x_k)\|_2^2 \]

Taking the expectation on both sides conditioned on \(x_k\) and combining with the definitions of \(|\mathcal{I}|\)

**Theorem 3.** If the nonlinear function \(f\) satisfies the local tangential cone condition given in Definition 4 \(\eta = \max_i \eta_i < \frac{1}{2}\),

\[ \varepsilon = h_2^2((f'_{x_k}(x_k))^\dagger) - 2\eta \sigma_{\text{max}}^2 ((f'_{x_k}(x_k))^\dagger) > 0, \tag{21} \]

where \(h_2^2((f'_{x_k}(x_k))^\dagger) = \min_{\mathcal{I}_k} h_2^2((f'_{x_k}(x_k))^\dagger) \) and \(\sigma_{\text{max}}^2 ((f'_{x_k}(x_k))^\dagger) = \max_{\mathcal{I}_k} \sigma_{\text{max}}^2 ((f'_{x_k}(x_k))^\dagger)\), and \(f(x_*) = 0\), then the iterations of the MR-BSNK1 method, i.e., the first case in Algorithm 5 satisfy

\[ \mathbb{E}\left[\|x_{k+1} - x_*\|_2^2\right] \leq \left(1 - \varepsilon|\mathcal{I}|\right) \frac{1}{\xi_k} m \frac{1}{1 + \eta^2} h_2^2(f'_{x_k}(x_k)) \mathbb{E}\left[\|x_k - x_*\|_2^2\right], \tag{22} \]

where

\[ |\mathcal{I}| = \min_{\mathcal{I}_k} |\mathcal{I}_k| \quad \text{and} \quad \xi_k = \frac{\sum_{\tau_k \in \{\mathcal{I}_k\}} \|f_{\tau_k}(x_k)\|_2}{\sum_{\tau_k \in \{\mathcal{I}_k\}} \|f_{\tau_k}(x_k)\|_\infty}. \tag{23} \]

**Proof.** According to Lemma 4 (21) and Algorithm 5, we have

\[ \|x_{k+1} - x_*\|_2^2 \leq \|x_k - x_*\|_2^2 - (h_2^2((f'_{x_k}(x_k))^\dagger) - 2\eta \sigma_{\text{max}}^2 ((f'_{x_k}(x_k))^\dagger)) \|f_{\tau_k}(x_k)\|_2^2 \]

Finally, we obtain the desired result by applying (9) in Lemma 4.
and $\xi_k$ in (23), we get
\[
\mathbb{E}^k \left[ \|x_{k+1} - x_*\|^2 \right] \leq \|x_k - x_*\|^2 - \varepsilon \mathbb{E}^k \left[ |I_k| \|f_{\tau_k}(x_k)\|_\infty^2 \right] \\
= \|x_k - x_*\|^2 - \varepsilon \sum_{\tau_k \in (\mathbb{I}_\beta)} \frac{1}{m} |I_k| \|f_{\tau_k}(x_k)\|_\infty^2 \\
\leq \|x_k - x_*\|^2 - \varepsilon |I| \sum_{\tau_k \in (\mathbb{I}_\beta)} \frac{1}{m} \|f_{\tau_k}(x_k)\|_2^2 \\
= \|x_k - x_*\|^2 - \varepsilon |I| \frac{1}{m} \xi_k \|f(x_k)\|_2^2 \\
= \|x_k - x_*\|^2 - \varepsilon |I| \frac{1}{m} \xi_k \|f(x_k) - f(x_\ast)\|_2^2.
\]
Moreover, noting that (10) in Lemma 4 and the definition of $h_2(\cdot)$ in Section 2.1, we get
\[
\mathbb{E}^k \left[ \|x_{k+1} - x_*\|^2 \right] \leq \|x_k - x_*\|^2 - \varepsilon |I| \frac{1}{m} \frac{1}{\xi_k} \frac{1}{m} \frac{1}{1 + \eta^2} \|f'(x_k)(x_k - x_\ast)\|_2^2 \\
\leq \|x_k - x_*\|^2 - \varepsilon |I| \frac{1}{m} \frac{1}{\xi_k} \frac{1}{m} \frac{1}{1 + \eta^2} h_k^2(f'(x_k)) \|x_k - x_\ast\|_2^2 \\
\leq \left( 1 - \varepsilon |I| \frac{1}{m} \frac{1}{\xi_k} \frac{1}{m} \frac{1}{1 + \eta^2} h_k^2(f'(x_k)) \right) \|x_k - x_\ast\|_2^2.
\]
Thus, taking expectation on both sides and using the tower rule of expectation, we get the desired result (22).

**Remark 8.** Since the iterations of the MR-BSNK1 method satisfy
\[
\mathbb{E} \left[ \|x_{k+1} - x_*\|^2 \right] \leq \left( 1 - \varepsilon |I| \frac{1}{m} \frac{1}{\xi_k} \frac{1}{m} \frac{1}{1 + \eta^2} h_k^2(f'(x_k)) \right) \mathbb{E} \left[ \|x_k - x_*\|^2 \right] \\
= \mathbb{E} \left[ \|x_k - x_*\|^2 \right] - \varepsilon |I| \frac{1}{m} \frac{1}{\xi_k} \frac{1}{m} \frac{1}{1 + \eta^2} h_k^2(f'(x_k)) \mathbb{E} \left[ \|x_k - x_*\|^2 \right],
\]
which together with the facts $0 \leq \mathbb{E} \left[ \|x_{k+1} - x_*\|^2 \right]$ and $\varepsilon |I| \frac{1}{m} \frac{1}{\xi_k} \frac{1}{m} \frac{1}{1 + \eta^2} h_k^2(f'(x_k)) > 0$ indicates
\[
0 \leq \mathbb{E} \left[ \|x_k - x_*\|^2 \right] - \varepsilon |I| \frac{1}{m} \frac{1}{\xi_k} \frac{1}{m} \frac{1}{1 + \eta^2} h_k^2(f'(x_k)) \mathbb{E} \left[ \|x_k - x_*\|^2 \right] < \mathbb{E} \left[ \|x_k - x_*\|^2 \right].
\]
Then we can obtain that the convergence factor of the MR-BSNK1 method is smaller than 1. Similarly, we can get that the convergence factors of the MD-BSNK1, MR-BSNK2 and MD-BSNK2 methods, which are respectively presented in Theorems 3 to 6, are also smaller than 1.

20
Theorem 4. If the nonlinear function $f$ satisfies the local tangential cone condition given in Definition 1, $\eta = \max \limits_i \eta_i < \frac{1}{2}$, $\varepsilon = h_2^2((f'_x(x_k))^\dagger) - 2\eta \sigma_{\max}^2 ((f'_x(x_k))^\dagger) > 0$, where $h_2^2((f'_x(x_k))^\dagger) = \min \limits_{I_k} h_2^2((f'_x(x_k))^\dagger)$ and $\sigma_{\max}^2 ((f'_x(x_k))^\dagger) = \max \limits_{I_k} \sigma_{\max}^2 ((f'_x(x_k))^\dagger)$, and $f(x_*) = 0$, then the iterations of MD-BSNK1 method, i.e., the second case in Algorithm 3, satisfy

$$\mathbb{E} \left[ \|x_{k+1} - x_*\|_2^2 \right] \leq \left( 1 - \varepsilon \min \limits_{i \in [m]} \|\nabla f_i(x_k)\|_2^2 |I| \right) \frac{\beta}{\zeta_k} h_2^2(G(x_k)) - \frac{1}{1 + \eta^2} \mathbb{E} \left[ \|x_k - x_*\|_2^2 \right],$$

where

$$|I| = \min \limits_{I_k} |I_k| \quad \text{and} \quad \zeta_k = \frac{\sum \limits_{\tau_k \in \binom{m}{\beta}} \|u_{\tau_k}(x_k)\|_2^2}{\sum \limits_{\tau_k \in \binom{m}{\beta}} \|u_{\tau_k}(x_k)\|_\infty^2}.$$ 

Proof. Similar to the proof of Theorem 3, we can obtain

$$\|x_{k+1} - x_*\|_2^2 \leq \|x_k - x_*\|_2^2 - \varepsilon \sum \limits_{j \in I_k} |f_j(x_k)|^2$$

$$= \|x_k - x_*\|_2^2 - \varepsilon \sum \limits_{j \in I_k} |f_j(x_k)|^2 \|\nabla f_j(x_k)\|_2^2$$

$$\leq \|x_k - x_*\|_2^2 - \varepsilon \min \limits_{i \in [m]} \|\nabla f_i(x_k)\|_2^2 \sum \limits_{j \in I_k} |f_j(x_k)|^2 \|\nabla f_j(x_k)\|_2^2,$$

which together with the definition of $u(x)$ in Section 2.1 and Algorithm 3 leads to

$$\|x_{k+1} - x_*\|_2^2 \leq \|x_k - x_*\|_2^2 - \varepsilon \min \limits_{i \in [m]} \|\nabla f_i(x_k)\|_2^2 \sum \limits_{j \in I_k} |u_j(x_k)|^2$$

$$\leq \|x_k - x_*\|_2^2 - \varepsilon \min \limits_{i \in [m]} \|\nabla f_i(x_k)\|_2^2 |I_k| \|\delta_k\|$$

$$= \|x_k - x_*\|_2^2 - \varepsilon \min \limits_{i \in [m]} \|\nabla f_i(x_k)\|_2^2 |I_k| \|u_{\tau_k}(x_k)\|_\infty^2.$$
which together with the definitions of $|Z|$ and $\zeta_k$ in (25), we have

$$E^k \left[ \|x_{k+1} - x_*\|_2^2 \right] \leq \|x_k - x_*\|_2^2 - \frac{1}{m \beta} \varepsilon \min_{i \in [m]} \|\nabla f_i(x_k)\|_2^2 \frac{1}{|Z|} \sum_{\tau \in \{-1, 1\}} \|u_{\tau_k}(x_k)\|_2^2$$

$$= \|x_k - x_*\|_2^2 - \frac{1}{m \beta} \varepsilon \min_{i \in [m]} \|\nabla f_i(x_k)\|_2^2 \frac{1}{|Z|} \frac{1}{\zeta_k} \|u(x_k)\|_2^2$$

Moreover, by making use of (12) in Lemma 5 and the definition of $h_2(\cdot)$ in Section 2.1, we can obtain

$$E^k \left[ \|x_{k+1} - x_*\|_2^2 \right] \leq \|x_k - x_*\|_2^2 - \varepsilon \min_{i \in [m]} \|\nabla f_i(x_k)\|_2^2 \frac{1}{|Z|} \frac{1}{\zeta_k} \|G(x_k)(x_k - x_*)\|_2^2$$

$$\leq \left( 1 - \varepsilon \min_{i \in [m]} \|\nabla f_i(x_k)\|_2^2 \frac{1}{|Z|} \frac{h_2^2(G(x_k))}{\zeta_k} \right) \|x_k - x_*\|_2^2.$$

Finally, the desired result (24) can be obtained by taking the full expectation on both sides.

**Remark 9.** Letting $D$ be an $m$-by-$m$ diagonal matrix of the following form

$$D = \begin{bmatrix} \|\nabla f_1(x_k)\|_2 & \cdots & \|\nabla f_m(x_k)\|_2 \end{bmatrix},$$

then we get a relation between $f'(x_k)$ and $G(x_k)$, that is,

$$f'(x_k) = DG(x_k). \quad (26)$$

Now, by setting $A = G(x_k)G(x_k)^T$ and $B = D^T D$, since both matrices $A$ and $B$ are symmetric positive-semidefinite, we have

$$\lambda_1(B) I \geq B \geq \lambda_m(B) I$$

and then obtain

$$A^T B A^T = A^T (B - \lambda_m(B) I) A^T + \lambda_m(B) A \geq \lambda_m(B) A.$$
which implies
\[
\lambda_m(AB) \geq \lambda_m(B)\lambda_m(A).
\]
That is,
\[
\lambda_m\left(G(x_k)G(x_k)^TD^TD\right) = \lambda_m\left(DG(x_k)G(x_k)^TD^T\right) \geq \lambda_m(D^TD)\lambda_m\left(G(x_k)G(x_k)^T\right),
\]
which together with the definitions of \(D\), \(G(x_k)\) and \(f'(x_k) = DG(x_k)\) in (26) indicates
\[
h_2^2(f'(x_k)) \geq \min_{i \in [m]} \|\nabla f_i(x_k)\|_2^2 \frac{\eta}{2} h_2^2(G(x_k)). \tag{27}
\]
Meanwhile, from Theorems 3 and 4 we know that the convergence factors of the MR-BSNK1 and MD-BSNK1 methods are
\[
\rho_{\text{MR-BSNK1}} \sim 1 - \frac{\varepsilon}{m 1 + \eta^2} h_2^2(f'(x_k)), \quad \text{and} \quad \rho_{\text{MD-BSNK1}} \sim 1 - \frac{\varepsilon}{m 1 + \eta^2} \min_{i \in [m]} \|\nabla f_i(x_k)\|_2^2 \frac{\eta}{2} h_2^2(G(x_k)),
\]
respectively. Thus, according to (27), we conclude that
\[
1 - \frac{\varepsilon}{m 1 + \eta^2} h_2^2(f'(x_k)) \leq 1 - \frac{\varepsilon}{m 1 + \eta^2} \min_{i \in [m]} \|\nabla f_i(x_k)\|_2^2 \frac{\eta}{2} h_2^2(G(x_k)),
\]
which means that the convergence factor of the MR-BSNK1 method is a smaller than that of the MD-BSNK1 method.

**Theorem 5.** If the nonlinear function \(f\) satisfies the local tangential cone condition given in Definition 4 \(\eta = \max_{i} \eta_i < \frac{1}{2}\), \(\alpha = h_2^2((f'_j(x_k))^i) - 2\eta \sigma_{\text{max}}^2 ((f'_j(x_k))^i) > 0\), where \(h_2^2((f'_j(x_k))^i) = \min_{J_k} h_2^2((f'_j(x_k))^i)\) and \(\sigma_{\text{max}}^2((f'_j(x_k))^i) = \max_{J_k} \sigma_{\text{max}}^2((f'_j(x_k))^i)\), \(f(x_*) = 0\), and \(\|f_{r_k}(x_k)\|_\infty = \min_{1 \leq i \leq \nu} \|f_{r_k}(x_k)\|_\infty\), then the iterations of the MR-BSNK2 method, i.e., the first case in Algorithm 4 satisfy
\[
\mathbb{E}\left[\|x_{k+1} - x_*\|_2^2\right] \leq \left(1 - \frac{\alpha \nu}{\tau_k h_2^2(f'_{r_k}(x_k))} \right) \mathbb{E}\left[\|x_k - x_*\|_2^2\right]. \tag{28}
\]

**Proof.** Following a similar argument as in Theorem 3 and from Algorithm 4 we get
\[
\|x_{k+1} - x_*\|_2^2 \leq \|x_k - x_*\|_2^2 - \alpha \sum_{j \in J_k} |f_j(x_k)|^2
\]
\[
= \|x_k - x_*\|_2^2 - \alpha \sum_{i=1}^{\nu} \|f_{r_i}(x_k)\|_\infty^2.
\]

23
Taking the expectation on both sides conditioned on $x_k$, we have

$$
E^k \left[ \|x_{k+1} - x_*\|_2^2 \right] \leq \|x_k - x_*\|_2^2 - \alpha E^k \left[ \sum_{i=1}^{\nu} \|f_{\tau_i}(x_k)\|_\infty^2 \right]
$$

$$
= \|x_k - x_*\|_2^2 - \alpha \sum_{i=1}^{\nu} E^k \left[ \|f_{\tau_i}(x_k)\|_\infty^2 \right].
$$

Noting that

$$
\sum_{i=1}^{\nu} E^k \left[ \|f_{\tau_i}(x_k)\|_\infty^2 \right] = E^k[\|f_{\tau_1}(x_k)\|_\infty^2] + E^k[\|f_{\tau_2}(x_k)\|_\infty^2] + \cdots + E^k[\|f_{\tau_\nu}(x_k)\|_\infty^2]
$$

which together with the fact $\|f_{\tau_i}(x_k)\|_\infty^2 = \min_{1 \leq i \leq \nu} \|f_{\tau_i}(x_k)\|_\infty^2$ gives

$$
\sum_{i=1}^{\nu} E^k \left[ \|f_{\tau_i}(x_k)\|_\infty^2 \right] \geq \sum_{\tau_1 \in \binom{[m]}{\beta_1}} \frac{1}{\binom{m}{\beta_1}} \|f_{\tau_1}(x_k)\|_\infty^2 + \sum_{\tau_2 \in \binom{[m]-\beta_1}{\beta_2}} \frac{1}{\binom{m-\beta_1}{\beta_2}} \|f_{\tau_2}(x_k)\|_\infty^2
$$

$$
+ \cdots + \sum_{\tau_\nu \in \binom{[m]-\beta_1-\cdots-\beta_{\nu-1}}{\beta_\nu}} \frac{1}{\binom{m-\beta_1-\cdots-\beta_{\nu-1}}{\beta_\nu}} \|f_{\tau_\nu}(x_k)\|_\infty^2
$$

$$
= \nu \|f_{\tau_n}(x_k)\|_\infty^2.
$$

Then, we have

$$
E^k \left[ \|x_{k+1} - x_*\|_2^2 \right] \leq \|x_k - x_*\|_2^2 - \alpha \nu \|f_{\tau_n}(x_k)\|_\infty^2
$$

$$
\leq \|x_k - x_*\|_2^2 - \alpha \nu \frac{1}{|\tau_n|} \|f_{\tau_n}(x_k)\|_2^2,
$$

which further together with (III) in Lemma 4 gets

$$
E^k \left[ \|x_{k+1} - x_*\|_2^2 \right] \leq \|x_k - x_*\|_2^2 - \alpha \nu \frac{1}{|\tau_n|} \frac{1}{1 + \eta^2} \|f_{\tau_n}'(x_k)(x_k - x_*)\|_2^2
$$

$$
\leq \|x_k - x_*\|_2^2 - \alpha \nu \frac{1}{|\tau_n|} \frac{1}{1 + \eta^2} h_2^2(f_{\tau_n}'(x_k)) \|x_k - x_*\|_2^2
$$

$$
= \left(1 - \frac{\alpha \nu \frac{h_2^2(f_{\tau_n}'(x_k))}{|\tau_n|}}{1 + \eta^2}\right) \|x_k - x_*\|_2^2.
$$

Thus, taking expectation on both sides and using the tower rule of expectation, we get the desired result (24).
Theorem 6. If the nonlinear function $f$ satisfies the local tangential cone condition given in Definition 4, $\eta = \max_i \eta_i < \frac{1}{2}$, $\alpha = h^2((f'_J(x_0))^\dagger) - 2\eta\sigma^2_{\max}((f'_J(x_0))^\dagger) > 0$, where $h^2((f'_J(x_0))^\dagger) = \min_{J_h} h^2((f'_J(x_0))^\dagger)$ and $\sigma^2_{\max}((f'_J(x_0))^\dagger) = \max_{J_h} \sigma^2_{\max}((f'_J(x_0))^\dagger)$, $f(x_*) = 0$, and $\|u_{\tau_i}(x_k)\|_\infty^2 = \min_{1 \leq i \leq \nu} \|u_{\tau_i}(x_k)\|_\infty^2$, then the iterations of the MD-BSNK2 method, i.e., the second case in Algorithm 3, satisfy

$$E \left[ \|x_{k+1} - x_*\|^2 \right] \leq \left( 1 - \min_{i \in [m]} \|\nabla f_i(x_k)\|_2^2 \frac{\alpha \nu \ h^2(G_{\tau_n}(x_k))}{\|\nabla f_{\dagger}(x_k)\|_2} \right) E \left[ \|x_k - x_*\|^2 \right].$$

(29)

PROOF. Following the proof of Theorem 3 and according to Algorithm 3, we have

$$\|x_{k+1} - x_*\|^2 \leq \|x_k - x_*\|^2 - \alpha \sum_{j \in \mathcal{J}_k} |f_j(x_k)|^2$$

$$= \|x_k - x_*\|^2 - \alpha \sum_{j \in \mathcal{J}_k} \frac{|f_j(x_k)|^2}{\|\nabla f_j(x_k)\|_2^2} \|\nabla f_j(x_k)\|_2^2$$

$$\leq \|x_k - x_*\|^2 - \alpha \min_{i \in [m]} \|\nabla f_i(x_k)\|_2^2 \sum_{j \in \mathcal{J}_k} \frac{|f_j(x_k)|^2}{\|\nabla f_j(x_k)\|_2^2}$$

$$\leq \|x_k - x_*\|^2 - \alpha \min_{i \in [m]} \|\nabla f_i(x_k)\|_2^2 \sum_{j \in \mathcal{J}_k} u_j(x_k)^2$$

$$= \|x_k - x_*\|^2 - \alpha \min_{i \in [m]} \|\nabla f_i(x_k)\|_2^2 \sum_{i=1}^\nu |u_{\tau_i}(x_k)|^2_\infty.$$

Now, taking the expectation on both sides conditioned on $x_k$, we have

$$E^k \left[ \|x_{k+1} - x_*\|^2 \right] \leq \|x_k - x_*\|^2 - \alpha \min_{i \in [m]} \|\nabla f_i(x_k)\|_2^2 \sum_{i=1}^\nu E^k \left[ |u_{\tau_i}(x_k)|^2_\infty \right]$$

$$= \|x_k - x_*\|^2 - \alpha \min_{i \in [m]} \|\nabla f_i(x_k)\|_2^2 \sum_{i=1}^\nu E^k \left[ |u_{\tau_i}(x_k)|^2_\infty \right].$$

Noting that

$$\sum_{i=1}^\nu E^k \left[ |u_{\tau_i}(x_k)|^2_\infty \right] = E^k[|u_{\tau_1}(x_k)|^2_\infty] + E^k[|u_{\tau_2}(x_k)|^2_\infty] + \cdots + E^k[|u_{\tau_\nu}(x_k)|^2_\infty]$$

$$= \sum_{\tau_1 \in (m_{\beta_1})} \frac{1}{m_{\beta_1}} |u_{\tau_1}(x_k)|^2_\infty + \sum_{\tau_2 \in (m_{\beta_2})} \frac{1}{m_{\beta_2}} |u_{\tau_2}(x_k)|^2_\infty + \cdots + \sum_{\tau_\nu \in (m_{\beta_\nu-1})} \frac{1}{m_{\beta_\nu-1}} |u_{\tau_\nu}(x_k)|^2_\infty,$$
which together with the fact \( \| u_{\tau_h}(x_k) \|_\infty^2 = \min_{1 \leq i \leq \nu} \| u_{\tau_i}(x_k) \|_\infty^2 \) gives

\[
\sum_{i=1}^{\nu} E^k \left[ \| u_{\tau_i}(x_k) \|_2^2 \right] \geq \sum_{\tau_1 \in (\beta_1)} \frac{1}{m} \| u_{\tau_h}(x_k) \|_2^2 + \sum_{\tau_2 \in (\beta_2)} \frac{1}{m} \| u_{\tau_h}(x_k) \|_2^2 + \cdots + \sum_{\tau_{\nu} \in (\beta_{\nu-1})} \frac{1}{m} \| u_{\tau_h}(x_k) \|_2^2
\]

\[
= \nu \| u_{\tau_h}(x_k) \|_2^2.
\]

Then, we have

\[
E^k \left[ \| x_{k+1} - x_* \|_2^2 \right] \leq \| x_k - x_* \|_2^2 - \alpha \min_{i \in [\nu]} \| \nabla f_i(x_k) \|_2^2 \nu \| u_{\tau_h}(x_k) \|_\infty^2
\]

\[
\leq \| x_k - x_* \|_2^2 - \alpha \min_{i \in [\nu]} \| \nabla f_i(x_k) \|_2^2 \nu \frac{1}{\min_{\tau_h}} \| u_{\tau_h}(x_k) \|_2^2,
\]

which further together with (12) in Lemma 5 gives

\[
E^k \left[ \| x_{k+1} - x_* \|_2^2 \right] \leq \| x_k - x_* \|_2^2 - \alpha \min_{i \in [\nu]} \| \nabla f_i(x_k) \|_2^2 \nu \frac{1}{\min_{\tau_h}} \frac{1}{1 + \eta^2} \| G_{\tau_h}(x_k) (x_k - x_*) \|_2^2
\]

\[
\leq \| x_k - x_* \|_2^2 - \alpha \min_{i \in [\nu]} \| \nabla f_i(x_k) \|_2^2 \nu \frac{1}{\min_{\tau_h}} \frac{1}{1 + \eta^2} \| h_{\tau}^2(G_{\tau_h}(x_k)) \| x_k - x_* \|_2^2
\]

\[
= \left( 1 - \min_{i \in [\nu]} \| \nabla f_i(x_k) \|_2^2 \frac{\alpha \nu \frac{h_{\tau}^2(G_{\tau_h}(x_k))}{\min_{\tau_h}}}{1 + \eta^2} \right) \| x_k - x_* \|_2^2.
\]

Thus, taking expectation on both sides and using the tower rule of expectation, we get the desired result [20].

**Remark 10.** According to Theorems 5 and 6 we know that the convergence factors of the MR-BSNK2 and MD-BSNK2 methods are

\[
\rho_{\text{MR-BSNK2}} = 1 - \frac{\alpha \nu \frac{h_{\tau}^2(f_{\tau}(x_k))}{\min_{\tau_h}}}{1 + \eta^2}, \quad \text{and} \quad \rho_{\text{MD-BSNK2}} = 1 - \frac{\alpha \nu \min_{i \in [\nu]} \| \nabla f_i(x_k) \|_2^2 h_{\tau}^2(G_{\tau_h}(x_k))}{\min_{\tau_h} \frac{1 + \eta^2}{1 + \eta^2}},
\]

respectively. Then, similar to the deduction in Remark 9 we can conclude that the convergence factor of the MR-BSNK2 method is smaller than that of the MD-BSNK2 method.

7. Experimental results

In this section, we investigate the convergence behavior of the proposed methods, i.e., the MR-SNK, MD-SNK, MR-BSNK1, MD-BSNK1, MR-BSNK2 and MD-BSNK2 methods, in solving the
problems including brown almost linear function and generalized linear model (GLM). We mainly compare our six methods with the existing methods in terms of the iteration numbers (denoted as “IT”) and computing time in seconds (denoted as “CPU”). Here, it should be pointed out that IT and CPU are the averages of IT and CPU of ten runs of algorithm. All experiments terminate once the residual at \(x_k\), defined by \(\|f(x_k)\|_2^2\) is less than \(10^{-6}\), or the number of iterations exceeds 200000.

7.1. Brown almost linear function

The function has the following form

\[
f_k(x) = x^{(k)} + \sum_{i=1}^{n} x^{(i)} - (n + 1), \quad 1 \leq k < n;
\]

\[
f_n(x) = (\prod_{i=1}^{n} x^{(i)}) - 1, \quad \text{for } k = n,
\]

which can be found in [12, 33]. Since the authors in [12] concluded that the convergence behavior of the NRK method outperforms the NURK and SGD methods, we only need to compare our methods with the NRK method and all experiments are initial from \(x_0 = 0.5 \times \text{ones}(n, 1)\).

7.1.1. The impact of parameters \(\beta\) and \(\nu\) on our methods

From Algorithms 2 to 4 we know that two parameters \(\beta\) and \(\nu\) affect the numerical results of our methods. With this in mind, we first show in Figure 1 how the parameter \(\beta\) impacts the MR-SNK, MD-SNK, MR-BSNK1 and MD-BSNK1 methods and show in Figure 2 how sensitive the MR-BSNK2 and MD-BSNK2 methods are for variation of parameter \(\nu\).

![Figure 1: IT (left) and CPU (right) versus \(\beta\) for the MR-SNK, MD-SNK, MR-BSNK1 and MD-BSNK1 methods.](image)

In Figure 1 both IT and CPU of the MR-SNK and MD-SNK methods drop as \(\beta\) increases from 5 to 50. Meanwhile, we find that the iteration numbers of the MR-BSNK1 and MD-BSNK1 methods
increase as $\beta$ increases, while their CPU times first reach their lowest point at $\beta = 10$ and $\beta = 15$, respectively, and then keep increasing as $\beta$ increases with some mild rebound. In addition, two multiple samples-based block methods outperform the two single sample-based methods in terms of iteration numbers and CPU time as expected.

Intuitively, as the sample size used in each iteration increases, the number of iteration steps is expected to decrease, which fits well with Figure 2 and is also our motivation for introducing block updating. However, the CPU times of the MR-BSNK2 and MD-BSNK2 methods exhibit a U-shaped curve as $\nu$ increases. That is, it first declines and then rises, albeit with some mild rebound. Actually, if the block size $\nu$ is small, then these two methods need more iteration numbers to achieve the desired accuracy, which in turn requires more CPU times. On the contrary, if the parameter $\nu$ is large, then more CPU time is needed to spend on each updating step. Hence although the number of iterations is relatively small, the total time is still high.

In a word, choosing a suitable parameter $\beta$ or $\nu$ is crucial for our methods. Thus, how to select a optimal parameter is a very interesting research topic, but the corresponding theoretical analysis is currently unavailable, which can be regarded as a future work.

7.1.2. Comparison of our methods and the NRK method

In this subsection, we demonstrate the effectiveness of our proposed methods by comparing them with the NRK method. The iteration numbers and CPU time for these methods are presented in Tables 4 and 5, respectively. As shown in Tables 4 and 5, although the optimal parameters are not always selected during the experiments, our methods, i.e., the MR-SNK, MD-SNK, MR-BSNK1, MD-BSNK1, MR-BSNK2 and MD-BSNK2 methods, usually require less iteration numbers and
computing time than that of the NRK method. Moreover, the block methods are always outperform the single sample-based methods in terms of iteration numbers and CPU time. Most interesting, the block methods even achieve a time speedup of 200 times compared to the NRK method, as shown in the last row of Table 5.

7.2. GLM

Here we compare our methods against the sketched Newton-Raphson (SNR) for solving regularized GLM, i.e., \( w^* = \arg\min_{w \in \mathbb{R}^d} P(w) \) where \( P(w) \) is the logistic loss, \( y_i \in \{-1, 1\} \) is the ith target value, \( a_i \in \mathbb{R}^d \) are data samples for \( i = 1, \ldots, p \), and \( w \in \mathbb{R}^d \) is the parameter to optimize. Using the processing procedure in [10], the original GLM can be equivalently transformed into solving the following nonlinear problem

\[
    f(x) \triangleq \begin{bmatrix}
        \frac{1}{p} A \alpha - w \\
        \alpha + \Phi(w)
    \end{bmatrix} = 0,
\]

where \( f : \mathbb{R}^{p+d} \to \mathbb{R}^{p+d} \), \( x = \begin{bmatrix} \alpha \\ w \end{bmatrix} \in \mathbb{R}^{p+d} \), \( A \triangleq \begin{bmatrix} a_1 & \cdots & a_p \end{bmatrix} \in \mathbb{R}^{d \times p} \), \( \Phi(w) \triangleq [\phi'_i(a_i^T w) \cdots \phi'_p(a_p^T w)]^T \in \mathbb{R}^p \) and \( \alpha = -\Phi(w) \). The advantage of this transformation is to ensure the sampling of a single sample and avoid a full passes over the data.

All datasets applied for GLM are obtained from \( \text{https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/} \) and the scaled versions are used if provided. These datasets are either ill or well conditioned, dense or sparse and their details including smoothness constant \( L \) of the model, density and condition

| \( m \times n \) | \( \beta \) | \( \nu \) | NRK | MR-SNK | MD-SNK | MR-BSNK1 | MD-BSNK1 | MR-BSNK2 | MD-BSNK2 |
|---------|-----|------|-----|--------|--------|----------|----------|----------|----------|
| 50 \times 50 | 5  | 5  | 4738.5 | 3545.1 | 3430.3 | 117.4 | 172.4 | 111.6 | 113.4 |
| 80 \times 80 | 5  | 5  | 1073.2 | 8327.1 | 8565.5 | 93  | 30.1  | 171.4 | 188.7 |
| 150 \times 150 | 10 | 10 | 33940 | 15093 | 15074 | 4.1  | 6.2   | 140.6 | 147   |
| 200 \times 200 | 10 | 10 | 56954 | 26135 | 26309 | 3.3  | 3.6   | 177.8 | 187.6 |
| 400 \times 400 | 20 | 20 | 197486 | 52110 | 52278 | 9.4  | 3     | 157   | 146.8 |

| \( m \times n \) | \( \beta \) | \( \nu \) | NRK | MR-SNK | MD-SNK | MR-BSNK1 | MD-BSNK1 | MR-BSNK2 | MD-BSNK2 |
|---------|-----|------|-----|--------|--------|----------|----------|----------|----------|
| 50 \times 50 | 5  | 5  | 0.3359 | 0.0688 | 0.0594 | 0.0500 | 0.0547 | 0.0422 | 0.0297 |
| 80 \times 80 | 5  | 5  | 0.5000 | 0.1437 | 0.1469 | 0.0453 | 0.0203 | 0.0531 | 0.0266 |
| 150 \times 150 | 10 | 10 | 1.3047 | 0.2422 | 0.2578 | 0.0297 | 0.0219 | 0.0422 | 0.0469 |
| 200 \times 200 | 10 | 10 | 2.5328 | 0.5656 | 0.4953 | 0.0047 | 0.0125 | 0.0578 | 0.0688 |
| 400 \times 400 | 20 | 20 | 8.0906 | 0.8313 | 0.9500 | 0.0484 | 0.0437 | 0.0750 | 0.0703 |
Table 6: Details of the data sets for GLM.

| dataset     | dimension (d) | samples (p) | $L$ condition number | density     |
|-------------|---------------|-------------|-----------------------|-------------|
| fourclass   | 2             | 862         | 0.0824                | 1.0757      | 0.9959      |
| german.numer| 24            | 1000        | 2.1113                | 15.4082     | 0.9584      |
| heart       | 13            | 270         | 0.6973                | 7.0996      | 0.9624      |
| ionosphere  | 34            | 351         | 1.5290                | 2.4485e+17  | 0.8841      |
| splice      | 60            | 1000        | 0.4349                | 2.6377      | 1           |
| sonar       | 60            | 208         | 3.2282                | 89.9388     | 0.9999      |
| w3a         | 300           | 4912        | 0.6436                | 3.9777e+33  | 0.0388      |
| w4a         | 300           | 7366        | 0.6385                | 1.0116e+34  | 0.0389      |

number of the data matrix $A \in \mathbb{R}^{d \times p}$ are provided in Table 6. Here smoothness constant $L$ is defined by $L \overset{\triangle}{=} \frac{\lambda_{\max}(AA^T)}{4p} + \lambda$ and density is computed by

$$\text{density} \overset{\triangle}{=} \frac{\text{number of nonzero of a } d \times p \text{ data matrix } A}{dp}.$$  

For all methods, we use $\lambda = \frac{1}{p}$ as the regularization parameter and let the initial value $x_0$ to be zero, i.e., $w_0 = 0 \in \mathbb{R}^d$ and $\alpha_0 = 0 \in \mathbb{R}^p$.

To make a more reasonable and fair comparison with the SNR method, we will discuss the convergence behavior of single-sample iteration and multiple-sample iteration respectively. Specifically, the iteration numbers and computing time of the MR-SNK, MD-SNK and Kaczmarz-TCS methods are compared in Figures 3 and 4, respectively. Here the Kaczmarz-TCS method is a special variation of the SNR method and its detailed implementation can be found in Algorithm 3 in [35]. In Figure 5, we compare the numerical results of block methods, i.e., the MR-BSNK1, MD-BSNK1, MR-BSNK2, MD-BSNK2 and Block TCS methods, where the last method is a block version of the Kaczmarz-TCS method; refer to Algorithm 4 in [35] for more details. In the specific experiments, for the Block TCS method, we set $\gamma = 1$, $\tau_d = d$, $\tau_p = 150$ and the Bernoulli parameter $b = \frac{p}{p+\tau_p} - 0.11$. Here, it is worth noting that, in order to demonstrate the advantage of greedy sampling over uniform sampling and the better use of the structure of the function $f$, our block methods adopts the same iterative framework as the Block TCS method. That is, for the first $d$ rows of $f$, we compute its least norm solution directly, which is the same as the Block TCS method because $\tau_d = d$; while for the last $p$ rows of $f$, we adopt greedy updating strategies discussed in Algorithms 3 and 4 but the Block TCS method uses uniform sampling.

Figures 3 to 5 show that our methods outperform the corresponding SNR method in terms of computing time and iteration numbers for all datasets, whether it is in the form of single-sample
iteration or multi-sample iteration.

Overall, these results confirm that our methods, i.e., the MR-SNK, MD-SNK, MR-BSNK1, MD-BSNK1, MR-BSNK2 and MD-BSNK2 methods are efficient for a wide variety of problems and datasets. This fact also further reflects the superiority and developability of greedy randomized sampling.

8. Concluding remarks

This paper mainly proposes six greedy randomized sampling nonlinear Kaczmarz methods for solving nonlinear problems. In theory, we analyze these methods in detail and rigorously compare the size of their convergence factors. Theoretical results show that the convergence factors of the MR-SNK and MD-SNK methods are smaller than those of the NRK and NURK methods. More interesting, for the NURK method, we obtain tighter upper bounds. Numerical results also show that our new methods are competitive. The above findings imply that these greedy strategies open up new venues for nonlinear problems.

Considering the superiority of greed, other greedy strategies can be further explored. In addition,
similar to \[36, 37\], by introducing an matrix $B$, more general updating formula can be deduced from the following relation:

$$x_{k+1} = \arg\min_{x \in \mathbb{R}^n} \|x - x_k\|_B^2, \quad s.t. \quad f_i(x_k) + \nabla f_i(x_k)^T (x - x_k) = 0.$$ 

So, we can further extend the greedy strategies discussed in this paper to the above iteration formula to obtain more efficient algorithms.

**References**

[1] Á. Björck, Numerical methods for least squares problems, SIAM, 1996.

[2] Q. Chen, W. Hao, A homotopy training algorithm for fully connected neural networks, Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci. 475 (2019) 20190662.

[3] J. E. Dennis Jr, R. B. Schnabel, Numerical methods for unconstrained optimization and nonlinear equations, SIAM, 1996.

[4] W. Hao, J. D. Hauenstein, B. Hu, A. J. Sommese, A bootstrapping approach for computing multiple solutions of differential equations, J. Comput. Appl. Math. 258 (2014) 181–190.
Figure 5: $\|f\|_2^2$ versus IT and CPU for the Block TCS, MR-BSNK1, MD-BSNK1, MR-BSNK2 and MD-BSNK2 methods with $\beta = 35$ and $\nu = \tau_p = 150$.

[5] K. E. Atkinson, A survey of numerical methods for solving nonlinear integral equations, J. Integral Equations Appl. 4 (1992) 15–46.

[6] W. Hao, J. Harlim, An equation-by-equation method for solving the multidimensional moment constrained maximum entropy problem, Commun. Appl. Math. Comput. Sci. 13 (2018) 189–214.

[7] W. Hao, A homotopy method for parameter estimation of nonlinear differential equations with multiple optima, J. Sci. Comput. 74 (2018) 1314–1324.

[8] C. T. Kelley, Iterative methods for optimization, SIAM, 1999.

[9] J. M. Ortega, W. C. Rheinboldt, Iterative solution of nonlinear equations in several variables, SIAM, 2000.

[10] R. Yuan, A. Lazaric, R. M. Gower, Sketched Newton–Raphson, SIAM J. Optim. 32 (2022) 1555–1583.

[11] W. Zeng, J. Ye, Successive projection for solving systems of nonlinear equations/inequalities, arXiv preprint arXiv:2012.07555 (2020).

[12] Q. Wang, W. Li, W. Bao, X. Gao, Nonlinear Kaczmarz algorithms and their convergence, J. Comput. Appl. Math. 399 (2022) 113720.

[13] B. T. Polyak, Introduction to Optimization. Optimization Software, New York 1 (1987) 32.
[14] M. Prazeres, A. M. Oberman, Stochastic gradient descent with polyak’s learning rate, J. Sci. Comput. 89 (2021) 1–16.

[15] N. Loizou, S. Vaswani, I. H. Laradji, S. Lacoste-Julien, Stochastic polyak step-size for sgd: An adaptive learning rate for fast convergence, in: International Conference on Artificial Intelligence and Statistics, PMLR, 2021, pp. 1306–1314.

[16] S. Kaczmarz, Angenäherte auflösung von systemen linearer gleichungen, Bull. Int. Acad. Pol. Sci. Lett. A 35 (1937) 355–357.

[17] T. Strohmer, R. Vershynin, A randomized Kaczmarz algorithm with exponential convergence, J. Fourier Anal. Appl. 15 (2009) 262–278.

[18] M. Griebel, P. Oswald, Greedy and randomized versions of the multiplicative schwarz method, Linear Algebra Appl. 437 (2012) 1596–1610.

[19] J. Nutini, B. Sepehry, A. Virani, I. Laradji, M. Schmidt, H. Koepke, Convergence rates for greedy Kaczmarz algorithms, in: 32nd Conference on Uncertainty in Artificial Intelligence, AUAI Press, 2016.

[20] K. Du, H. Gao, A new theoretical estimate for the convergence rate of the maximal weighted residual Kaczmarz algorithm, Numer. Math. Theor. Meth. Appl. 12 (2019) 627–639.

[21] Y. Eldar, D. Needell, Acceleration of randomized Kaczmarz method via the Johnson-Lindenstrauss lemma, Numer. Algor. 58 (2011) 163–177.

[22] Z. Bai, W. Wu, On greedy randomized Kaczmarz method for solving large sparse linear systems, SIAM J. Sci. Comput. 40 (2018) A592–A606.

[23] Y. Zhang, H. Li, A count sketch maximal weighted residual Kaczmarz method for solving highly overdetermined linear systems, Appl. Math. Comput. 410 (2021) 126486.

[24] Y. Zhang, H. Li, Greedy Motzkin–Kaczmarz methods for solving linear systems, Numer. Linear Algebra Appl. (2020) e2429.

[25] J. A. De Loera, J. Haddock, D. Needell, A sampling Kaczmarz-Motzkin algorithm for linear feasibility, SIAM J. Sci. Comput. 39 (5) (2017) S66–S87.
[26] J. Haddock, A. Ma, Greed works: an improved analysis of sampling Kaczmarz-Motzkin, SIAM J. Math. Data Sci. 3 (1) (2021) 342–368.

[27] O. Scherzer, Antonio, R. Kowar, M. Haltmeier, Kaczmarz methods for regularizing nonlinear ill-posed equations ii: Applications, Inverse Problems and Imaging 1 (2017) 507–523.

[28] G. H. Golub, C. F. Van Loan, Matrix Computations, JHU press, fourth edition ed., 2013.

[29] S. Agamon, The relaxation method for linear inequalities, Canad. J. Math. 6 (1954) 382–392.

[30] T. S. Motzkin, I. J. Schoenberg, The relaxation method for linear inequalities, Canad. J. Math. 6 (1954) 393–404.

[31] T. Strohmer, R. Vershynin, A randomized Kaczmarz algorithm with exponential convergence, J. Fourier Anal. Appl. 15 (2009) 262–278.

[32] Y. Zhang, H. Li, Block sampling Kaczmarz-Motzkin methods for consistent linear systems, Calcolo 58 (2021) 1–20.

[33] J. J. Moré, B. S. Garbow, K. E. Hillstrom, Testing unconstrained optimization software, ACM Transactions on Mathematical Software (TOMS) 7 (1981) 17–41.

[34] C. C. Chang, C. J. Lin, Libsvm: a library for support vector machines, ACM transactions on intelligent systems and technology (TIST) 2 (3) (2011) 1–27.

[35] R. Yuan, A. Lazaric, R. M. Gower, Sketched Newton-Raphson, arXiv preprint arXiv:2006.12120 (2020).

[36] R. M. Gower, P. Richtárik, Randomized iterative methods for linear systems, SIAM J. Matrix Anal. Appl. 36 (2015) 1660–1690.

[37] R. M. Gower, D. Molitor, J. Moorman, D. Needell, On adaptive sketch-and-project for solving linear systems, SIAM J. Matrix Anal. Appl. 42 (2) (2021) 954–989.